FUNDAMENTAL LIMITS OF LOW-RANK MATRIX ESTIMATION: INFORMATION-THEORETIC AND COMPUTATIONAL PERSPECTIVES

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF STATISTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Yuchen Wu

December 2023

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Andrea Montanari, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Iain Johnstone

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Tselil Schramm

Approved for the Stanford University Committee on Graduate Studies.

Stacey F. Bent, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format.

Preface

This dissertation explores several problems in the realm of low-rank matrix estimation. A primary focus is on understanding the statistical and computational limitations. From a practical perspective, understanding such limitations not only provides practitioners with guidance on algorithm selection, but also in some cases spurs the development of cutting-edge methodologies which improve on the state of the art. Within this theme, this dissertation explores and partially answers the following two questions: (1) Given a large-scale low-rank matrix corrupted by random noise, how much information can we accurately infer from the limited observations? (2) How do restrictions on computational resources affect information retrieval?

A secondary focus of this dissertation is on developing algorithms that sample from the posterior in the context of low-rank matrix estimation. A standard machinery to fulfill this task is based on Markov Chain Monte Carlo (MCMC) algorithms. However, rigorous guarantees are often difficult to obtain for MCMC algorithms of common use. This dissertation contributes to this line of work from an alternative perspective: We propose an alternative class of efficient algorithms based on diffusion processes that come with rigorous guarantee.

This dissertation is organized as follows: We describe the problem in Chapter 1. Chapter 2 studies low-rank matrix estimation from an information-theoretic perspective, and Chapter 3-4 analyzes the effects of limited computational resource. In Chapter 5, we design a sampling algorithm that works well with the low-rank model. Standalone versions of each chapter can be found in [154, 50, 155, 156].

Acknowledgments

I would like to express my acknowledgement to my mentors, friends, and family (listed and not listed below).

First and foremost, I want to thank my advisor Andrea for guiding me through my PhD. I could not have asked for a better advisor. In the past four years, Andrea has been extremely generous with his time and energy. I sincerely appreciate him for the countless hours we have spent together working on math equations in front of a white board, as well as for him sharing his remarkable intuition and insights which often inspire me to think from different perspectives. I also want to thank Andrea for his patience on improving my academic writing and research presentation skills, and for the inspiring conversations he has offered regarding career development. His encouragement and support have influenced me significantly and helped me transforming into a better version of myself.

I also benefitted greatly from Iain. Iain has overwhelming knowledge in statistics, and is always extremely kind and supportive. I want to thank him for introducing to me the beauty of random matrix theory and functional estimation as well as for numerous helpful literatures he has pointed me to. I also appreciate him for warmly encouraging me to reach out to potential collaborators, providing guidance on several research problems, as well as offering great career and personal advice.

I want to thank Tselil for working with me in the past year, especially for introducing new problems on tensor decomposition and tensor PCA. I have learned new ways to approach and tackle research problems from her. Apart from that, I have been greatly influenced by her passion and dedication towards research. I am also grateful for her efforts on offering a extremely nice and cozy office to its visitors, in which we have had many cheerful and stimulating discussions.

I want to express my gratitude to Aaron and Sourav, who generously serve on my oral exam committee. Besides, I also want to thank them for the outstanding lectures they have been providing to the Stanford community, which has helped me building the foundation for doing research in the early years of my PhD.

I would like to thank Ryan for shaping my perspectives on developing research ideas and evaluating research problems. I also want to thank him for offering conversations regarding the career queries I have come up in spite of his packed schedule.

I am thankful to my other brilliant collaborators André, Ashkan, Brandon, Daniel, Filipe, Jakab, Jenny, Kangjie, Michael, Miki, Mohammad, Pratik, Ran, Sarfaraz, Xinyi, Zihan, and Zhuoran. I appreciate them for accompanying me during the journey of exploration in the past few years, and I am grateful for benefiting from their wisdom and diligence. I would also like to thank everyone from Andrea's group and Tselil's group, for numerous wonderful group meetings and educative discussions in the past few years. My thanks goes to Basil, Chen, Frederick, Kabir, Kangjie, Issac, John, June, Leda, Michael, Misha, Raphel, Phan-Minh, Shuangping, Spencer, Song, Theodor, Will, and Yiqiao.

I would also like to thank my friends both inside and outside Sequoia, for the colorful memories we have created together. I want to thank my cohorts Ben, Dan, Han, Kevin Guo, Kevin Han, Hui, Marius, Samyak, and Souvik for the joyful cohort dinners and for supporting each other getting through the qualifying exam. I want to thank Stephen, Sky, and Zhimei for the wonderful qualifying exam lectures. I want to thank Daren, Fang, Feng, Han, Hui, Kevin Han, Lingfu, Ran, Shuangning, Shuangping, Song, Sophia, Ying, Zhihan, Zhimei, and Zitong for the board game nights and restaurant explorations. I want to thank Daren, Han, Kangjie, Kevin Han, Lingfu, Shuangning, and Shuangping for tolerating my awful tennis shots. I want to thank Kexin and Xinyi for the gatherings in many different cities across US. I want to thank Zijun for exchanging gossips and for generously letting me keep her car in the past year. I want to thank Pratik for all the cheerful “digressions”. A special thanks goes to Susie, for taking care of all my carelessness in all these years, without whom I will never finish my PhD.

I am fortunate to have met Kangjie here at Stanford, and I want to thank him for his company. I hope that our love will always remain in the many years to come.

Last but not least, I want to thank my family, especially my parents for their endless love and unwavering support. They have always been my greatest source of courage.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The problem of reconstructing a low-rank signal matrix observed through a noisy channel has received enormous attention from extensive body of literature within machine learning, signal processing and information theory. Various statistical and machine learning tasks can be reduced to this canonical problem, including but not limited to, sparse PCA [109, 113, 31, 66, 65], community detection [1, 64], submatrix localization [119, 98], and Gaussian mixture clustering [144, 173, 159, 44].

Among the numerous models developed for this purpose, the *spiked model* introduced by Johnstone [107] plays a fundamental role, especially for interpreting and understanding high-dimensional asymptotics. This model is also referred to as *deformed ensembles* in the literature of random matrix theory. Theoretical analysis of the spiked models has yielded a number of important statistical insights [14, 15, 108, 5, 109, 136, 66, 65, 28, 29, 111, 137].

This dissertation mainly focuses on the asymmetric version of the spiked model. In this model, we observe a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ which is given by the sum of a low-rank signal and random noise

$$\boldsymbol{A} = s_n \boldsymbol{\Lambda} \boldsymbol{\Theta}^{\mathsf{T}} + \boldsymbol{Z}\,, \tag{1.1}$$

where $\boldsymbol{\Lambda} \in \mathbb{R}^{n\times r}$ and $\boldsymbol{\Theta} \in \mathbb{R}^{d\times r}$ are the factors that we would like to estimate, $s_n > 0$ is the signal-to-noise ratio, and $\boldsymbol{Z} \in \mathbb{R}^{n\times d}$ consists of random noise. We will consider the high-dimensional asymptotics with a low-rank structure, whereby $d, n \to \infty$ and r remains fixed.

In what follows, we will denote by $\boldsymbol{a}_1, \dots, \boldsymbol{a}_n$ the rows of $\boldsymbol{A}$. Before stating any of our result, we describe a few applications of model (1.1) to motivate the study.

Example 1.0.1 (Sparse PCA)**.** In a simple model for sparse PCA [109], we observe vectors

$$\boldsymbol{a}_1, \dots, \boldsymbol{a}_n \sim_{iid} \mathsf{N}(0, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma} = s_n^2 \boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathsf{T}} + \boldsymbol{I}_d$ with $\boldsymbol{\Theta} \in \mathbb{R}^d$ a sparse vector that we would like to estimate.

This is the special case of model (1.1), if we let $r = 1$ and $\boldsymbol{\Lambda} \sim \mathsf{N}(0, \boldsymbol{I}_n)$.

Example 1.0.2 (Mixture of Gaussians with known covariance)**.** In a mixture of Gaussian model, we observe vectors $\bar{\boldsymbol{a}}_1, \dots, \bar{\boldsymbol{a}}_n \sim_{iid} p\,\mathsf{N}(\boldsymbol{\Theta}_1, \boldsymbol{\Sigma}_1) + (1-p)\,\mathsf{N}(\boldsymbol{\Theta}_2, \boldsymbol{\Sigma}_2)$. If the covariances coincide and are known: $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, and the population mean $\bar{\boldsymbol{\Theta}} := p\boldsymbol{\Theta}_1 + (1-p)\boldsymbol{\Theta}_2$ can be estimated accurately, then we can define

$\boldsymbol{a}_i = \boldsymbol{\Sigma}^{-1/2}(\bar{\boldsymbol{a}}_i - \bar{\boldsymbol{\Theta}})$. Hence the model is equivalent to observing $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n \sim_{iid} p\,\mathsf{N}((1-p)\boldsymbol{\Theta}, \boldsymbol{I}_d) + (1-p)\,\mathsf{N}(-p\boldsymbol{\Theta}, \boldsymbol{I}_d)$, with $\boldsymbol{\Theta} := \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)$.

This is another special case of model (1.1), with $r = 1$ and $(\Lambda_i)_{i \le n} \sim_{iid} p\delta_{(1-p)} + (1-p)\delta_{-p}$. Estimating $\boldsymbol{\Lambda}$ amounts to estimating the cluster labels.

1.1 An information-theoretic perspective

As demonstrated in these examples, it is often the case that the latent factors $\boldsymbol{\Lambda}$, $\boldsymbol{\Theta}$ have additional structure. In the first example $\boldsymbol{\Theta} \in \mathbb{R}^d$ is a sparse vector, while in the second one $\boldsymbol{\Lambda}$ is a vector with i.i.d. entries distributed according to a two-point mixture. This observation motivates us to assume a stylized model whereby the rows $\boldsymbol{\Lambda}$, $\boldsymbol{\Theta}$ are mutually independent (and independent of $\boldsymbol{Z}$) with $(\boldsymbol{\Lambda}_i)_{i \le n} \overset{iid}{\sim} \mu_\Lambda$ and $(\boldsymbol{\Theta}_j)_{j \le d} \overset{iid}{\sim} \mu_\Theta$. Here, μ_Λ and μ_Θ are fixed probability distributions on $\mathbb{R}^r$. For this part, we will also make the assumption that $Z_{ij} \overset{iid}{\sim} \mathsf{N}(0,1)$. We will assume an idealized setting where μ_Λ, μ_Θ and the signal-to-noise ratio s_n are known to the estimator. This setting was considered several times in recent past, see e.g. [142, 125, 152, 18]. Recent work addresses the assumption that $s_n, \mu_\Lambda, \mu_\Theta$ are known. Namely, [201] uses empirical Bayes techniques to show that $s_n, \mu_\Lambda, \mu_\Theta$ in many standard settings can be estimated consistently based on data.

Several natural questions that arise in this Bayesian setting are as follows: What is the minimum s_n that enables recovery of the low-rank factors, and what is the corresponding Bayes optimal estimation error? Closely related to our work are the results of [142, 125], who determined the precise asymptotics of mutual information and (certain) estimation error metrics when $n, d \to \infty$, in the proportional regime $n/d \to \delta \in (0, \infty)$. Our goal in this dissertation is to move beyond the proportional asymptotics and consider the cases $d/n \to \infty$ and $d/n \to 0$. We show that depending on the scaling of the signal-to-noise ratio s_n, there are two interesting regimes that control the behavior of the estimation problem.

1. **Strong signal regime:** When $s_n \asymp n^{-1/2}$, we show that $\boldsymbol{\Lambda}$ can be estimated consistently (possibly up to a rotation), while the minimum normalized estimation error of $\boldsymbol{\Theta}$ remains bounded away from 0. We characterize the limiting error for estimating $\boldsymbol{\Theta}$.

2. **Weak signal regime:** When $s_n \asymp (nd)^{-1/4}$, our results imply that non-trivial estimation of $\boldsymbol{\Theta}$ is impossible. As for the estimation of $\boldsymbol{\Lambda}$, we show that the current model (1.1) is equivalent to a symmetric spiked model of size $n \times n$. The minimum estimation error of the latter has been characterized in [125], which allows us to derive expression for the minimum estimation error of $\boldsymbol{\Lambda}$ under model (1.1).

Our results are insightful for at least two reasons: (1) Our study establishes optimal performance achieved by any algorithm in this context, which acts as an ideal benchmark in real-world applications; (2) The equivalence between our model (1.1) and an $n \times n$ symmetric model suggests that there is no substantial loss of accuracy in estimating $\boldsymbol{\Lambda}$ based on a smaller matrix, which yields a significant reduction of computational complexity if $n \ll d$. We illustrate our general theory by carrying out a numerical study on genomics data.

1.2 A computational perspective

High-dimensional statistical estimation problems are often addressed by constructing a suitable data-dependent cost function $\mathcal{L}(\boldsymbol{\vartheta})$, which encodes the statistician's knowledge of the problem. This cost is then minimized using an algorithm which scales well to large dimension. The most popular algorithms for high-dimensional statistical applications are first order methods, i.e., algorithms that query the cost $\mathcal{L}(\boldsymbol{\vartheta})$ by computing its gradient (or a subgradient) at a sequence of points $\boldsymbol{\Theta}^1$,. . . $\boldsymbol{\Theta}^t$. Examples include (projected) gradient descent, mirror descent, and accelerated gradient descent.

This raises a fundamental question: *What is the minimal statistical error achieved by first order methods?* In particular, we would like to understand in which cases these methods are significantly sub-optimal (in terms of estimation) with respect to statistically optimal but potentially intractable estimators, and what is the optimal tradeoff between number of iterations and estimation error.

These questions are relatively well understood only from the point of view of convex optimization, namely if estimation is performed by minimizing a convex cost function $\mathcal{L}(\boldsymbol{\vartheta})$, see e.g. [45, 34]. The seminal work of Nemirovsy and Yudin [161] characterizes the minimum gap to global optimality $\mathcal{L}(\boldsymbol{\Theta}^t) - \min_{\boldsymbol{\vartheta}} \mathcal{L}(\boldsymbol{\vartheta})$, where $\boldsymbol{\Theta}^t$ is the algorithm's output $\boldsymbol{\Theta}^t$ after t iterations (i.e., after t gradient evaluations). For instance, if $\mathcal{L}(\boldsymbol{\Theta})$ is a smooth convex function, there exists a first order algorithm which achieves $\mathcal{L}(\boldsymbol{\Theta}^t) \le \min_{\boldsymbol{\vartheta}} \mathcal{L}(\boldsymbol{\vartheta}) + O(t^{-2})$. At the same time, no algorithm can be guaranteed to achieve a better convergence rate over all functions in this class.

In contrast, if the cost $\mathcal{L}(\boldsymbol{\vartheta})$ is nonconvex, there cannot be general guarantees of global optimality. Substantial effort has been devoted to showing that –under suitable assumptions about the data distribution– certain nonconvex costs $\mathcal{L}(\boldsymbol{\Theta})$ can be minimized efficiently, e.g. by gradient descent [116, 132, 57]. This line of work resulted in upper bounds on the estimation error of first order methods. Unlike in the convex case, worst case lower bounds are typically overly pessimistic since non-convex optimization is NP-hard. Our work aims at developing precise average-case lower bounds for a restricted class of algorithms, which are applicable both to convex and nonconvex problems.

We are particularly interested in problems that exhibit an information-computation gap: we know that the optimal statistical estimator has high accuracy, but existing upper bounds on first order methods are substantially sub-optimal (see examples below). Is this a limitation of our analysis, of the specific algorithm under consideration, or of first order algorithms in general? The main result of this part is a tight asymptotic characterization of the minimum estimation error achieved by first order algorithms for two families of problems. This characterization can be used, in particular, to delineate information-computation gaps.

1.3 A sampling perspective

Sampling algorithms serve as one of the major building blocks of modern Bayesian inference. However, for many high-dimensional models of interest, analytical derivation of the posterior distributions is computationally intractable, thus creating challenges in designing efficient sampling methods. The third part of the dissertation is concerned with sampling from the posterior distribution of a low-rank signal that is corrupted by noise. For this part, we consider the symmetric spiked model. More precisely, for a given signal-to-noise parameter $\beta > 0$, we observe an $n \times n$ symmetric matrix $\boldsymbol{X}$ generated as follows:

$$\boldsymbol{X} = \frac{\beta}{n}\boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}} + \boldsymbol{W}. \tag{1.2}$$

In the above display, we assume $\boldsymbol{W} \sim \mathrm{GOE}(n)$, i.e., $\boldsymbol{W}$ is an $n \times n$ symmetric matrix with independently distributed entries above the diagonal: $\{W_{ii} : i \in [n]\} \overset{iid}{\sim} \mathsf{N}(0, 2/n)$, $\{W_{ij} : 1 \leq i < j \leq n\} \overset{iid}{\sim} \mathsf{N}(0, 1/n)$. The n-dimensional vector $\boldsymbol{\theta}$ follows a product prior: $\theta_i \overset{iid}{\sim} \pi_\Theta$, which is further independent of the additive noise $\boldsymbol{W}$. As always, we will assume β is fixed and is known to the estimator. We comment that if $\beta > 1$, then this parameter can be consistently estimated via inspecting the top eigenvalue of $\boldsymbol{X}$ [14].

Given observation $\boldsymbol{X}$, our goal is to establish efficient algorithms that sample from the posterior distribution $\mu_{\boldsymbol{X}}$. Namely, we aim to design a method that accepts as input $\boldsymbol{X}$, and outputs $\boldsymbol{\theta}^{\mathsf{alg}} \sim \mu_{\boldsymbol{X}}^{\mathsf{alg}}$, such that $\mathbb{E}\big[\mathsf{dist}(\mu_{\boldsymbol{X}}, \mu_{\boldsymbol{X}}^{\mathsf{alg}})\big] = o_n(1)$ for some distance measure $\mathsf{dist}(\cdot,\cdot)$.

If we take π_Θ to be the Rademacher distribution, then model (1.2) reduces to the problem of $\mathbb{Z}_2$-synchronization [176]. In this case, sampling from the posterior distribution of model (1.2) becomes a special case of sampling from the Ising model [103], which is a distribution over the hypercube $\{\pm 1\}^n$ that takes the form

$$\mu_{J,h}^{\mathsf{Ising}}(\sigma) = \frac{1}{Z}\exp\left(\frac{1}{2}\langle \sigma, J\sigma\rangle + \langle h, \sigma\rangle\right).$$

In the above expression, $Z > 0$ is an unknown normalizing constant. In our case, $h = 0$ and $J = \beta \cdot \boldsymbol{X}$.

One of the dominant approaches to approximately sample from such distributions is the Gibbs sampling algorithm, also known as the Glauber dynamics. This is a Markov chain that updates one index at each round according to its conditional probability distribution. Upper bounds on the mixing time of Glauber dynamics under the Ising model are, to our knowledge, only established in the high-temperature regime $\|J\|_{\mathsf{op}} < 1$ [**?**, 22, 82, 6], which corresponds to $\beta < 1/4$ in our model. Unfortunately, this is a regime in which it is information-theoretically impossible to recover $\boldsymbol{\theta}$ under the $\mathbb{Z}_2$-synchronization model [125]. On the other hand, [27] proves that at sufficiently low temperature, the mixing time of Glauber dynamics is exponentially large.

Another popular approach for conducting approximate Bayesian analysis is via variational inference [139, 36]. In general, variational inference attempts to compute the marginals of a high-dimensional distribution by optimizing a suitable "free energy" function. The most commonly used objective function under this category is the so-called "naive mean field" free energy. However, such approximation is incorrect for $\mathbb{Z}_2$-synchronization and returns an inconsistent estimation. This inaccuracy can be remedied by applying a simple TAP correction [186] to the free energy functional. In the case of $\mathbb{Z}_2$-synchronization and many other problems, minimizing the TAP free energy leads to a consistent estimation in the low-temperature regime [84, 49]. Despite its success, variational inference only provides estimates to the marginal posterior distributions, thus falling short of generating a sampling mechanism. [118] build a sampling algorithm for Ising model that fuses ideas from both MCMC (Markov Chain Monte Carlo) and variational inference, and their approach is able to move beyond the high-temperature regime $\|J\|_{\mathsf{op}} < 1$. However, they still require that the majority of the eigenvalues of J fall inside an interval of length one. More recently, [4] design a sampling method that is based on an algorithmic implementation of stochastic localization [80]. They focus on the Sherrington-Kirkpatrick model (this is Ising model with a random interaction matrix) at high-temperature with no external field. More precisely, they provide theoretical guarantee for $\|J\|_{\mathsf{op}} < 2$, and conjecture that their algorithm works in a broader setting $\|J\|_{\mathsf{op}} < 4$. Their results also apply to $\mathbb{Z}_2$-synchronization at high temperature.

In this part of the dissertation, we complement to prior works by designing a sampling algorithm that efficiently samples from the posterior distribution of the general low-rank matrix estimation problem (1.2) in the low-temperature regime (meaning that β is larger than some positive constant). We comment that this is also the regime in which non-trivial recovery is expected. Similar to [4], our approach is motivated by the stochastic localization process

Chapter 2

Low-rank matrix estimation with diverging aspect ratios

2.1 Summary of main results

In this chapter we state results from an information-theoretic perspective. Namely, we assume the observed matrix further follows a Bayesian model, and aim to give an exact characterization of the Bayesian minimum mean squared error under a diverging aspect ratio. For this part, we focus on the case $d/n \to \infty$: analogous statements for $d/n \to 0$ follow by interchanging n and d, as well as $\boldsymbol{\Lambda}$ and $\boldsymbol{\Theta}$. Our results reveal the following two regimes:

Strong signal regime. This is obtained for $s_n \asymp n^{-1/2}$, and is relatively easy to characterize analytically. Under this scaling, $\boldsymbol{\Lambda}$ can be estimated consistently (possibly up to a rotation), while the minimum normalized estimation error of $\boldsymbol{\Theta}$ remains bounded away from 0. We characterize the limiting error of estimating $\boldsymbol{\Theta}$.

Weak signal regime: Estimation of $\boldsymbol{\Theta}$. This regime corresponds to $s_n \asymp (nd)^{-1/4}$, and most of our technical work is devoted to its analysis. We prove that, in this regime, non-trivial estimation of $\boldsymbol{\Theta}$ is impossible: any estimator has asymptotically the same risk as the the null estimator $\hat{\boldsymbol{\Theta}}_0 = \mathbb{E}[\boldsymbol{\Theta}]$.

Weak signal regime: Mutual information. On the other hand, still in taking $s_n \asymp (nd)^{-1/4}$, estimation of $\boldsymbol{\Lambda}$ is non-trivial. As a first result in this direction, we characterize the asymptotic mutual information

$$\lim_{n,d\to\infty} \frac{1}{n} I(\boldsymbol{A}; \boldsymbol{\Lambda}),$$

and show that this is non-vanishing. Further, this mutual information is asymptotically the same as for a symmetric observation model in which instead of $\boldsymbol{A} \in \mathbb{R}^{n\times d}$, we observe $\boldsymbol{Y} \in \mathbb{R}^{n\times n}$ given by

$$\boldsymbol{Y} = \frac{q_\Theta}{n} \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}} + \boldsymbol{W}, \quad \boldsymbol{W} \sim \mathrm{GOE}(n)\,. \tag{2.1}$$

(Here, $q_\Theta := r^{-1} \int \|\boldsymbol{\theta}\|^2 \, \mu_\Theta(\mathrm{d}\boldsymbol{\theta})$, we take without loss of generality $s_n = (nd)^{-1/4}$, and $\mathrm{GOE}(n)$ denotes the distribution of a symmetric matrix with independent entries on or above the diagonal $(W_{ij})_{i\le j}$

such that $W_{ii} \sim \mathsf{N}(0, 2/n)$ and $W_{ij} \sim \mathsf{N}(0, 1/n)$ for $i < j$.)

Weak signal regime: Estimation error. We then proceed to study the asymptotics of the Bayes optimal matrix mean square error:

$$\mathrm{MMSE}(\mu_\Lambda, \mu_\Theta) := \lim_{n,d\to\infty} \frac{1}{n^2}\mathbb{E}\big\{\big\|\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\mathsf{T} - \mathbb{E}[\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\mathsf{T}|\boldsymbol{A}]\big\|_F^2\big\}\,. \tag{2.2}$$

We characterize this limit in two regimes: $n \ll d \ll n^{6/5}$ or $n^3 \ll d$, and in certain cases for all $n \ll d$ (here $\ll$ hides logarithmic factors.) In these cases we prove equivalence with model (2.1).

We believe that the conditions $n \ll d \ll n^{6/5}$ or $n^3 \ll d$ are artifacts of the proof. Indeed, the conclusion holds for all $n \ll d$ under a natural (unproven) continuity assumption. We leave it to future work to cover the intermediate range $n^{6/5} \lesssim d \lesssim n^3$.

Both the limiting mutual information and the asymptotic estimation error $\mathrm{MMSE}(\mu_\Lambda, \mu_\Theta)$ are given by explicit expressions known as 'replica symmetric formulas,' because they are correctly predicted by the replica method in spin glass theory [140, 151]. However, the asymptotic equivalence with the symmetric model (2.1) is insightful in itself (i.e., independently of the fact that we can give explicit formulas for the asymptotic error and mutual information):

1. The asymptotic equivalence between model (1.1) and model (2.1) implies that the optimal estimation depends on μ_Θ only though its second moment. In other words, no substantial improvement is achieved in the regime covered by this equivalence exploiting the knowledge of the distribution of $\boldsymbol{\Theta}$.

2. The symmetric matrix $\boldsymbol{Y}$ is closely related to the Gram matrix $\boldsymbol{Y}' = (\boldsymbol{A}\boldsymbol{A}^\mathsf{T} - d\boldsymbol{I}_n)/\sqrt{nd}$, an observation that is confirmed by inspecting the proof. This implies that there is no substantial loss of accuracy in estimating $\boldsymbol{\Lambda}$ uniquely on the basis of $\boldsymbol{Y}'$. This yields a substantial reduction in complexity for $n \ll d$.

We warn the reader that these conclusions do not apply in settings that are not captured here. For instance, in sparse PCA, cf. Example 1.0.1, one might be interested in cases in which the number of non-zeros of the principal component $\boldsymbol{\Theta}$ is sub-linear in the dimension d. This case cannot be modeled as above, and requires instead to consider μ_Θ dependent on n, d.

The rest of the paper is organized as follows. We briefly review related work in Section 2.2. We then present our results for the strong signal regime in Section 2.3 and the weak signal regime in Section 2.4. We finally apply the general theory to the case of Gaussian mixture models in Section 2.5 and compare it with analysis on real data in Section 2.6.

2.1.1 Notations and conventions

For $k \in \mathbb{N}$, we define the set $[k] := \{1, 2, \cdots, k\}$. We typically use lower case non-bold letters for scalars (m, n, j), and bold for vectors and matrices ($\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$, $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$). We use $\|\boldsymbol{v}\|$ to denote the Euclidean norm of a vector $\boldsymbol{v}$, and $\|\boldsymbol{M}\|_F$ to denote the Frobenius norm of a matrix $\boldsymbol{M}$. For $\{c_n\}_{n\in\mathbb{N}_+}, \{d_n\}_{n\in\mathbb{N}_+} \subseteq \mathbb{R}_+$, we say $c_n \gg d_n$ if and only if $c_n/d_n \to \infty$, and for $\{e_n\}_{n\in\mathbb{N}_+} \subseteq \mathbb{R}$, we say $e_n = o_n(1)$ if and only if $e_n \to 0$ as $n \to \infty$. We denote by p-lim convergence in probability.

For $k \in \mathbb{N}_+$, we denote by S_k^+ the set of positive semi-definite matrices in $\mathbb{R}^{k\times k}$, and denote by $\mathcal{O}(k)$ the set of orthogonal matrices in $\mathbb{R}^{k\times k}$. For $\boldsymbol{M} \in S_k^+$, we let $\boldsymbol{M}^{1/2} \in S_k^+$ be any positive semi-definite matrix such that $\boldsymbol{M} = \boldsymbol{M}^{1/2}\boldsymbol{M}^{1/2}$.

We denote the i-th row of the factors $\mathbf{\Lambda}$, $\mathbf{\Theta}$ by $\mathbf{\Lambda}_i$ and $\mathbf{\Theta}_i$, respectively. We use $\mathbf{\Lambda}_0$ and $\mathbf{\Theta}_0$ to represent length-r random vectors drawn from probability distributions μ_Λ and μ_Θ. We sometimes need to write the posterior distribution of $(\mathbf{\Lambda}, \mathbf{\Theta})$ given particular observations. In this case, we use the lower case letters $\boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\lambda}_i, \boldsymbol{\theta}_i$ to represent variables corresponding to $(\mathbf{\Theta}, \mathbf{\Lambda}, \mathbf{\Lambda}_i, \mathbf{\Theta}_i)$ in the posterior distribution.

Throughout the paper, we use capital letter C to represent various numerical constants.

2.2 Further related work

As mentioned in the introduction, most earlier work deriving sharp asymptotics results focuses on the proportional regime $n \asymp d$, $s_n = n^{-1/2}$. In particular, [127] first obtained the limiting expression for Bayesian mean square error using non-rigorous tools from statistical mechanics. The conjectured expression was rigorously justified for special distributions μ_Λ, μ_Θ in [65, 64]. However, the proof technique of [65, 64] relies on the fact that approximate message passing (AMP) algorithm achieves Bayes optimality and does not apply to the general case.

Several groups developed rigorous approaches to prove the asymptotic formulas in increasing degrees of generality: spatial coupling [68]; the cavity method [125, 142, 78]; adaptive interpolation [21]; partial differential equation techniques [70].

A different line of research uses the second moment method to derive upper and lower bounds on the information-theoretic thresholds [18, 171, 170] for partial or exact recovery. This approach typically yields non-asymptotic bounds, under a broader class of settings but the results only determine such thresholds up to undetermined multiplicative constants. In contrast, here we attempt to obtain a characterization that is accurate up to $(1 + o_n(1))$ factors.

From a computational viewpoint, AMP-based algorithms can be shown to achieve the Bayesian error for a large region of parameters [24, 152]. One appealing fact about the AMP is that its high-dimensional behavior can be sharply characterized by *state evolution*.

Minimax guarantees were obtained by a number of groups for special cases of the low-rank model (1.1). Sparse PCA and Gaussian mixtures are arguably the most studied models in the literature, see e.g., [174, 96, 160, 30, 43] . These works often yield characterizations that hold up to usually a constant or logarithmic multiplicative gap.

Gaussian mixture models (GMM) provide a useful context for evaluating and comparing various clustering algorithms. We will use it here to illustrate the applicability of our general results. The goal can be either estimating the centers [62, 63, 117, 143, 173], or recovering the underlying cluster assignments [190, 3, 39, 122, 10, 86]. As we will see, in the high-dimensional weak signal regime, the cluster centers cannot be estimated, but the cluster assignments can be estimated with non-trivial accuracy.

Several algorithms were studied in detail for clustering under GMM, including semi-definite programming (SDP) [169, 9, 86, 102, 129], iterative algorithms with spectral initialization [3, 190, 122, 10, 133], the method of moments [168, 87, 114, 144, 101, 26, 99], and EM-based algorithms [63, 16, 106, 44].

Finally, in concurrent work, Donoho and Feldman recently characterized the accuracy of eigenvalue shrinkage methods in the spiked model with diverging aspect ratio [88, 72].

2.3 Strong signal regime

We first consider the strong signal regime in which we set $s_n = 1/\sqrt{n}$, and therefore we have

$$\boldsymbol{A} = \frac{1}{\sqrt{n}}\boldsymbol{\Lambda}\boldsymbol{\Theta}^{\mathsf{T}} + \boldsymbol{Z} \in \mathbb{R}^{n\times d}. \tag{2.3}$$

We define $\boldsymbol{Q}_\Lambda := \mathbb{E}_{\boldsymbol{\Lambda}_0\sim\mu_\Lambda}[\boldsymbol{\Lambda}_0\boldsymbol{\Lambda}_0^{\mathsf{T}}] \in S_r^+$ and $\boldsymbol{Q}_\Theta := \mathbb{E}_{\boldsymbol{\Theta}_0\sim\mu_\Theta}[\boldsymbol{\Theta}_0\boldsymbol{\Theta}_0^{\mathsf{T}}] \in S_r^+$. Before we proceed, we establish the following conventions for the distributions μ_Λ, μ_Θ.

Remark 2.3.1. Without loss of generality we can and will assume that both $\boldsymbol{Q}_\Lambda$ and $\boldsymbol{Q}_\Theta$ are invertible. Furthermore, we can assume that $\boldsymbol{Q}_\Theta = q_\Theta \boldsymbol{I}_r$ for some $q_\Theta \in \mathbb{R}_{>0}$.

More precisely, we next show that —given arbitrary probability distributions $(\mu_\Theta, \mu_\Lambda)$— the conditions of Remark 2.3.1 can always be satisfied by a reparameterization.

For $\boldsymbol{\Lambda}_0 \sim \mu_\Lambda$ and $\boldsymbol{\Theta}_0 \sim \mu_\Theta$, if either $\boldsymbol{\Lambda}_0 \overset{a.s.}{=} 0$ or $\boldsymbol{\Theta}_0 \overset{a.s.}{=} 0$ then estimation becomes trivial. We can therefore assume that this is not the case. By eigendecomposition of $\boldsymbol{Q}_\Lambda$ and $\boldsymbol{Q}_\Theta$, there exist $0 < k_1, k_2 \le r$ and non-random matrices $\boldsymbol{M}_1 \in \mathbb{R}^{r\times k_1}$, $\boldsymbol{M}_2 \in \mathbb{R}^{r\times k_2}$ with full column ranks such that $\boldsymbol{\Lambda}_0 = \boldsymbol{M}_1\boldsymbol{\Lambda}_0'$, $\boldsymbol{\Theta}_0 = \boldsymbol{M}_2\boldsymbol{\Theta}_0'$, and $\mathbb{E}[\boldsymbol{\Lambda}_0'(\boldsymbol{\Lambda}_0')^{\mathsf{T}}] = \boldsymbol{I}_{k_1}$, $\mathbb{E}[\boldsymbol{\Theta}_0'(\boldsymbol{\Theta}_0')^{\mathsf{T}}] = \boldsymbol{I}_{k_2}$.

Assume $\boldsymbol{M}_1^{\mathsf{T}}\boldsymbol{M}_2$ has rank $k_3 \le \min(k_1, k_2)$, and let $\boldsymbol{M}_1^{\mathsf{T}}\boldsymbol{M}_2 = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathsf{T}}$ be its singular value decomposition (SVD) with $\boldsymbol{U} \in \mathbb{R}^{k_1\times k_3}$, $\boldsymbol{V} \in \mathbb{R}^{k_2\times k_3}$ having orthonormal columns. We then set $\bar{\boldsymbol{\Lambda}} \in \mathbb{R}^{n\times k_3}$ a matrix with i.i.d. rows that are copies of $\boldsymbol{S}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{\Lambda}_0'$ and $\bar{\boldsymbol{\Theta}} \in \mathbb{R}^{d\times k_3}$ a matrix with i.i.d. rows that are copies of $\boldsymbol{V}^{\mathsf{T}}\boldsymbol{\Theta}_0'$. We can then write $\boldsymbol{\Lambda}\boldsymbol{\Theta}^{\mathsf{T}} = \bar{\boldsymbol{\Lambda}}\bar{\boldsymbol{\Theta}}^{\mathsf{T}}$ and the latter satisfies the conditions of Remark 2.3.1.

Note that this argument shows that we could assume $q_\Theta = 1$ as well, but it is convenient to keep this as a free parameter.

2.3.1 Estimation of Λ

We first consider estimation of $\boldsymbol{\Lambda}$. We will show that a simple spectral estimator provides a consistent estimate up to a rotation in the r-dimensional Euclidean space. Consistency in terms of vector mean square error is not guaranteed due to potential non-identifiability issues. We propose sufficient conditions on $(\mu_\Lambda, \mu_\Theta)$, that imply consistency in terms of vector mean square error as well.

Denote by $\hat{\boldsymbol{\Lambda}}_s \in \mathbb{R}^{n\times r}$ the matrix whose columns are the top r eigenvectors of $\boldsymbol{A}\boldsymbol{A}^{\mathsf{T}}$, normalized so that $\hat{\boldsymbol{\Lambda}}_s^{\mathsf{T}}\hat{\boldsymbol{\Lambda}}_s/n = \boldsymbol{I}_r$. Denote by $\mathbf{P}, \hat{\mathbf{P}} \in \mathcal{O}(n)$ the projection matrices onto the column spaces of $\boldsymbol{\Lambda}$ and $\hat{\boldsymbol{\Lambda}}_s$, respectively. We use the following distance between the two subspaces as estimation loss

$$L^{\sin}(\hat{\boldsymbol{\Lambda}}_s, \boldsymbol{\Lambda}) := \|\mathbf{P}(\boldsymbol{I} - \hat{\mathbf{P}})\|_{\text{op}} = \|\hat{\mathbf{P}}(\boldsymbol{I} - \mathbf{P})\|_{\text{op}} = \sin\alpha(\hat{\boldsymbol{\Lambda}}_s, \boldsymbol{\Lambda}), \tag{2.4}$$

where $\alpha(\hat{\boldsymbol{\Lambda}}_s, \boldsymbol{\Lambda})$ is the principal angle between the two column spaces.

Theorem 2.3.1. *Assume μ_Λ, μ_Θ have finite non-singular second moments $\boldsymbol{Q}_\Lambda, \boldsymbol{Q}_\Theta$ with $\boldsymbol{Q}_\Theta = q_\Theta\boldsymbol{I}_r$ (with no loss of generality per Remark 2.3.1). If $n, d \to \infty$ with $d/n \to \infty$, then under the model of Eq. (2.3):*

1. $L^{\sin}(\hat{\boldsymbol{\Lambda}}_s, \boldsymbol{\Lambda}) \xrightarrow{P} 0.$

2. *If we further assume that for some $\varepsilon > 0$ we have $\mathbb{E}_{\boldsymbol{\Lambda}_0\sim\mu_\Lambda}[\|\boldsymbol{\Lambda}_0\|^{4+\varepsilon}] < \infty$, then there exists an estimator $\hat{\boldsymbol{L}} : \boldsymbol{A} \mapsto \hat{\boldsymbol{L}}(\boldsymbol{A}) \in \mathbb{R}^{n\times n}$, such that $\mathbb{E}\big[\|\hat{\boldsymbol{L}}(\boldsymbol{A}) - \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}}\|_F^2\big]/n^2 \to 0$ as $n, d \to \infty$.*

3. *Let $\mathbf{\Lambda}_0 \sim \mu_\Lambda$. If we further assume that there does not exist $\mathbf{\Omega} \in \mathcal{O}(r)$, such that $\mathbf{\Omega} \neq \boldsymbol{I}_r$ and $\mathbf{\Omega}\mathbf{\Lambda}_0 \overset{d}{=} \mathbf{\Lambda}_0$, then there exists $\hat{\mathbf{\Lambda}} : \boldsymbol{A} \mapsto \hat{\mathbf{\Lambda}}(\boldsymbol{A}) \in \mathbb{R}^{n\times r}$, such that $\mathbb{E}\big[\|\hat{\mathbf{\Lambda}}(\boldsymbol{A}) - \mathbf{\Lambda}\|_F^2\big]/n \to 0$ as $n, d \to \infty$.*

We delay the proof of Theorem 2.3.1 to Appendix A.3.1.

2.3.2 Estimation of $\mathbf{\Theta}$

Next, we turn to the estimation of $\mathbf{\Theta}$. According to Theorem 2.3.1, $\mathbf{\Lambda}$ can be estimated consistently under identifiability conditions. Therefore, a reasonable first step is to study the case in which $\mathbf{\Lambda}$ is given. This yields a lower bound on the Bayesian error of the original problem. We will see that this lower bound can be achieved asymptotically even if $\mathbf{\Lambda}$ must be estimated.

We can explicitly write the conditional distribution of $\mathbf{\Theta}$ given $(\mathbf{\Lambda}, \boldsymbol{A})$. Using the Gaussian density formula, we see that for all $j \in [d]$, the posterior distribution of $\mathbf{\Theta}_j$ is

$$p(\mathrm{d}\boldsymbol{\theta}_j|\mathbf{\Lambda}, \boldsymbol{A}) \propto \exp\left(-\frac{1}{2n}\sum_{i=1}^n \langle \mathbf{\Lambda}_i, \boldsymbol{\theta}_j\rangle^2 + \frac{1}{\sqrt{n}}\sum_{i=1}^n A_{ij}\langle \mathbf{\Lambda}_i, \boldsymbol{\theta}_j\rangle\right)\mu_\Theta(\mathrm{d}\boldsymbol{\theta}_j). \tag{2.5}$$

Eq. (2.5) leads to the following asymptotic lower bound:

Theorem 2.3.2. *Consider the strong signal model of Eq. (2.3), assuming, without loss of generality, the setting of Remark 2.3.1. We let $n, d \to \infty$ simultaneously with $d/n \to \infty$, then for any estimator $\hat{\boldsymbol{\theta}} : \boldsymbol{A} \mapsto \hat{\boldsymbol{\theta}}(\boldsymbol{A}) \in \mathbb{R}^{d\times r}$, we have*

$$\liminf_{n,d\to\infty} \frac{1}{d}\mathbb{E}\left[\|\hat{\boldsymbol{\theta}}(\boldsymbol{A}) - \boldsymbol{\theta}\|_F^2\right] \geq rq_\Theta - \mathbb{E}\left[\left\|\mathbb{E}[\mathbf{\Theta}_0|\boldsymbol{Q}_\Lambda^{1/2}\mathbf{\Theta}_0 + \boldsymbol{G}]\right\|^2\right], \tag{2.6}$$

where $\boldsymbol{G} \sim \mathsf{N}(0, \boldsymbol{I}_r)$, $\mathbf{\Theta}_0 \sim \mu_\Theta$ are mutually independent. Notice that the right hand side of Eq. (2.6) is independent of (n, d).

If we further assume $\mathbb{E}[\|\mathbf{\Theta}_0\|^4] < \infty$, then for any $\hat{\boldsymbol{M}} : \boldsymbol{A} \mapsto \hat{\boldsymbol{M}}(\boldsymbol{A}) \in \mathbb{R}^{d\times d}$, we have

$$\liminf_{n,d\to\infty} \frac{1}{d^2}\mathbb{E}\left[\|\hat{\boldsymbol{M}}(\boldsymbol{A}) - \mathbf{\Theta}\mathbf{\Theta}^\mathsf{T}\|_F^2\right] \geq rq_\Theta^2 - \left\|\mathbb{E}\left[\mathbb{E}[\mathbf{\Theta}_0|\boldsymbol{Q}_\Lambda^{1/2}\mathbf{\Theta}_0 + \boldsymbol{G}]\mathbb{E}[\mathbf{\Theta}_0|\boldsymbol{Q}_\Lambda^{1/2}\mathbf{\Theta}_0 + \boldsymbol{G}]^\mathsf{T}\right]\right\|_F^2. \tag{2.7}$$

We postpone the proof of Theorem 2.3.2 to Appendix A.3.2. Next, we show that the lower bound proposed in Theorem 2.3.2 can be achieved under identifiability conditions.

Theorem 2.3.3. *Under the conditions of Theorem 2.3.1, claim 3, there exist estimators $\hat{\boldsymbol{\theta}} : \boldsymbol{A} \mapsto \hat{\boldsymbol{\theta}}(\boldsymbol{A})$ and $\hat{\boldsymbol{M}} : \boldsymbol{A} \mapsto \hat{\boldsymbol{M}}(\boldsymbol{A})$, such that*

$$\begin{aligned}\lim_{n,d\to\infty} \frac{1}{d}\mathbb{E}\Big[\|\hat{\mathbf{\Theta}}(\boldsymbol{A}) - \mathbf{\Theta}\|_F^2\Big] &= rq_\Theta - \mathbb{E}\left[\left\|\mathbb{E}[\mathbf{\Theta}_0|\boldsymbol{Q}_\Lambda^{1/2}\mathbf{\Theta}_0 + \boldsymbol{G}]\right\|^2\right],\\ \lim_{n,d\to\infty} \frac{1}{d^2}\mathbb{E}\Big[\|\hat{\boldsymbol{M}}(\boldsymbol{A}) - \mathbf{\Theta}\mathbf{\Theta}^\mathsf{T}\|_F^2\Big] &= rq_\Theta^2 - \left\|\mathbb{E}\left[\mathbb{E}[\mathbf{\Theta}_0|\boldsymbol{Q}_\Lambda^{1/2}\mathbf{\Theta}_0 + \boldsymbol{G}]\mathbb{E}[\mathbf{\Theta}_0|\boldsymbol{Q}_\Lambda^{1/2}\mathbf{\Theta}_0 + \boldsymbol{G}]^\mathsf{T}\right]\right\|_F^2.\end{aligned}$$

We defer the proof of Theorem 2.3.3 to Appendix A.3.3. Theorems 2.3.2 and 2.3.3 together complete the analysis for the estimation of $\mathbf{\Theta}$ in the strong signal regime.

2.4 Weak signal regime

In this section, we consider the weak signal regime where $s_n = 1/\sqrt[4]{nd}$. Thus the model of interest is

$$\boldsymbol{A} = \frac{1}{\sqrt[4]{nd}}\boldsymbol{\Lambda}\boldsymbol{\theta}^{\mathsf{T}} + \boldsymbol{Z} \in \mathbb{R}^{n\times d}\,. \tag{2.8}$$

For convenience, we define $r_n := \sqrt[4]{d/n}$. By assumption, we see that $r_n \to \infty$ as $n, d \to \infty$.

2.4.1 Background: the symmetric spiked model

As mentioned in the introduction, our main technical result is that, in the weak signal regime, estimation under model (2.8) is equivalent to estimation under a symmetric spiked model. Under this model we observe $\boldsymbol{Y} \in \mathbb{R}^{n\times n}$ given by

$$\boldsymbol{Y} = \frac{q_\Theta}{n}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}} + \boldsymbol{W}, \tag{2.9}$$

where $\boldsymbol{W} \overset{d}{=} \mathrm{GOE}(n)$, and $\boldsymbol{\Lambda}_i \overset{iid}{\sim} \mu_\Lambda$, independent of each other. We view $q_\Theta > 0$ as a signal-to-noise ratio parameter.

We denote the Bayesian MMSE of model (2.9) by

$$\mathrm{MMSE}_n^{\mathrm{symm}}(\mu_\Lambda; q_\Theta) := \min_{\hat{\boldsymbol{M}}(\,\cdot\,)} \frac{1}{n^2}\mathbb{E}\left[\left\|\hat{\boldsymbol{M}}(\boldsymbol{Y}) - \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}}\right\|_F^2\right]. \tag{2.10}$$

Note that the Bayesian MMSE is achieved by the posterior expectation $\hat{\boldsymbol{M}}(\boldsymbol{Y}) = \mathbb{E}[\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}}|\boldsymbol{Y}]$. We also define the normalized mutual information

$$\mathrm{I}_n^{\mathrm{symm}}(\mu_\Lambda; q_\Theta) := \frac{1}{n}\mathbb{E}\log\frac{\mathrm{d}\mathbb{P}_{\boldsymbol{\Lambda},\boldsymbol{Y}}}{\mathrm{d}(\mathbb{P}_{\boldsymbol{\Lambda}}\times\mathbb{P}_{\boldsymbol{Y}})}(\boldsymbol{\Lambda},\boldsymbol{Y})\,. \tag{2.11}$$

A significant amount of rigorous information is available about this model. For $s > 0$ and $\boldsymbol{Q} \in S_r^+$, we define the free energy functional $\mathcal{F}(s, \boldsymbol{Q})$ and its maximizer $\boldsymbol{Q}^*(s) \in S_r^+$ via

$$\mathcal{F}(s,\boldsymbol{Q}) := -\frac{s}{4}\|\boldsymbol{Q}\|_F^2 + \mathbb{E}\left\{\log\left(\int \exp(\sqrt{s}\boldsymbol{z}^{\mathsf{T}}\boldsymbol{Q}^{1/2}\boldsymbol{\lambda} + s\boldsymbol{\lambda}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{\Lambda}_0 - \frac{s}{2}\boldsymbol{\lambda}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{\lambda})\mu_\Lambda(\mathrm{d}\boldsymbol{\lambda})\right)\right\}. \tag{2.12}$$

$$\boldsymbol{Q}^*(s) \in \mathrm{argmax}_{\boldsymbol{Q}\in S_r^+}\,\mathcal{F}(s,\boldsymbol{Q}). \tag{2.13}$$

In the above expression, expectation is taken over $\boldsymbol{\Lambda}_0 \sim \mu_\Lambda$ and $\boldsymbol{z} \sim \mathsf{N}(0, \boldsymbol{I}_r)$ independent of each other. These functionals are directly related to the mutual information and the Bayes MMSE of model (2.9), as stated below.

Theorem 2.4.1 ([125], Corollary 42, Proposition 43)**.** *There exists a deterministic countable set $\mathcal{D} \subseteq \mathbb{R}_{\geq 0}$ such that*

$$q_\Theta \in \mathbb{R}_{\geq 0} \;\Rightarrow\; \lim_{n\to\infty}\mathrm{I}_n^{\mathrm{symm}}(\mu_\Lambda; q_\Theta) = \frac{1}{4}q_\Theta^2\|\mathbb{E}_{\boldsymbol{\Lambda}_0\sim\mu_\Lambda}[\boldsymbol{\Lambda}_0\boldsymbol{\Lambda}_0^{\mathsf{T}}]\|_F^2 - \sup_{\boldsymbol{Q}\in S_r^+}\mathcal{F}(q_\Theta^2, \boldsymbol{Q})\,,$$

$$q_\Theta \in \mathbb{R}_{\geq 0}\setminus\mathcal{D} \;\Rightarrow\; \lim_{n\to\infty}\mathrm{MMSE}_n^{\mathrm{symm}}(\mu_\Lambda; q_\Theta) = \|\mathbb{E}_{\boldsymbol{\Lambda}_0\sim\mu_\Lambda}[\boldsymbol{\Lambda}_0\boldsymbol{\Lambda}_0^{\mathsf{T}}]\|_F^2 - \|\boldsymbol{Q}^*(s)\|_F^2.$$

2.4.2 Estimation of $\boldsymbol{\Theta}$

We first consider estimation of $\boldsymbol{\Theta}$. We claim that in this case no estimator outperforms a naive one.

Theorem 2.4.2. *Consider the weak signal model of Eq. (2.8), assuming, without loss of generality, the setting of Remark 2.3.1. Let $n, d \to \infty$ simultaneously with $d/n \to \infty$. Then for any estimator $\hat{\boldsymbol{\Theta}} : \boldsymbol{A} \mapsto \hat{\boldsymbol{\Theta}}(\boldsymbol{A}) \in \mathbb{R}^{d\times r}$, we have*

$$\liminf_{n,d\to\infty} \frac{1}{d}\mathbb{E}[\|\hat{\boldsymbol{\Theta}}(\boldsymbol{A}) - \boldsymbol{\theta}\|_F^2] \geq rq_\Theta - \|\mathbb{E}_{\boldsymbol{\Theta}_0\sim\mu_\Theta}[\boldsymbol{\Theta}_0]\|^2.$$

If we further assume μ_Θ has bounded fourth moment, then for any $\hat{\boldsymbol{M}} : \boldsymbol{A} \mapsto \hat{\boldsymbol{M}}(\boldsymbol{A}) \in \mathbb{R}^{d\times d}$, we have

$$\liminf_{n,d\to\infty} \frac{1}{d^2}\mathbb{E}[\|\hat{\boldsymbol{M}}(\boldsymbol{A}) - \boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T}\|_F^2] \geq rq_\Theta^2 - \|\mathbb{E}_{\boldsymbol{\Theta}_0\sim\mu_\Theta}[\boldsymbol{\Theta}_0]\|^4 .$$

Notice that the above lower bounds are achieved by the null estimators $\hat{\boldsymbol{\Theta}}(\boldsymbol{A}) = \mathbb{E}[\boldsymbol{\Theta}] \in \mathbb{R}^{d\times r}$ and $\hat{\boldsymbol{M}}(\boldsymbol{A}) = \mathbb{E}[\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T}] \in \mathbb{R}^{d\times d}$.

The proof of this statement is similar to the one of Theorem 2.3.2. Namely, we will prove that the mean square error achieved by simply taking the prior mean asymptotically agrees with the Bayesian MMSE for an estimator that has access to $\boldsymbol{\Lambda}$ as additional information. The argument is summarized in Appendix A.4.1.

2.4.3 Estimation of $\boldsymbol{\Lambda}$

We finally consider the technically most interesting case, namely the estimation of $\boldsymbol{\Lambda}$ in the weak signal regime. For simplicity, we will restrict ourselves to studying the matrix mean square error:

$$\mathrm{MMSE}_n^{\mathrm{asym}}(\mu_\Lambda, \mu_\Theta) := \inf_{\hat{M}(\,\cdot\,)} \frac{1}{n^2}\mathbb{E}\left[\left\|\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\mathsf{T} - \hat{\boldsymbol{M}}(\boldsymbol{A})\right\|_F^2\right], \tag{2.14}$$

where the infimum is taken over all estimators (measurable functions) $\hat{\boldsymbol{M}} : \boldsymbol{A} \mapsto \hat{\boldsymbol{M}}(\boldsymbol{A}) \in \mathbb{R}^{n\times n}$. Of course $\mathrm{MMSE}_n^{\mathrm{asym}}$ depends on the distributions μ_Λ, μ_Θ.

In the rank-one case $r = 1$, if $\mathbb{E}_{\boldsymbol{\Theta}_0\sim\mu_\Theta}[\boldsymbol{\Theta}_0] \neq 0$, then the naive estimator $r_n^{-1}\boldsymbol{A}\boldsymbol{y}$ with $\boldsymbol{y} = \mathbb{E}_{\mu_\Theta}[\boldsymbol{\Theta}_0]^{-1}\mathbf{1}_d/\sqrt{d}$ is consistent:

$$\frac{1}{n}\left\|r_n^{-1}\boldsymbol{A}\boldsymbol{y} - \boldsymbol{\Lambda}\right\|^2 \xrightarrow{P} 0.$$

In this case, a consistent estimate of $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\mathsf{T}$ naturally follows. Therefore, if $\boldsymbol{\Theta}_0$ has non-vanishing expectation, the estimation problem is significantly easier.

When $r \geq 2$, if μ_Θ has non-zero mean, the same construction leads to consistent estimation of the projection of $\boldsymbol{\Lambda}$ onto the direction determined by $\mathbb{E}[\boldsymbol{\Theta}_0]$ (for $\boldsymbol{\Theta}_0 \sim \mu_\Theta$). Once this component is subtracted, the problem is effectively reduced to one in which μ_Θ has zero mean.

For the remainder of this section, we focus on the more challenging case $\mathbb{E}[\boldsymbol{\Theta}_0] = \mathbf{0}_r$. In addition, for technical reasons we will require μ_Θ to have vanishing third moment.

Assumption 2.4.1. *We assume* $\mathbb{E}[\mathbf{\Theta}_0] = \mathbf{0}_r$, $\mathbb{E}[\mathbf{\Theta}_0 \otimes \mathbf{\Theta}_0 \otimes \mathbf{\Theta}_0] = \mathbf{0}_{r\times r\times r}$, *where* $\otimes$ *denotes the tensor product. Furthermore, we assume that* μ_Θ, μ_Λ *are sub-Gaussian.*

Our main results establish that, according to several criteria, estimation in the asymmetric model (2.8) with $n, d \to \infty$, $d/n \to \infty$ is equivalent to estimation in the symmetric spiked model (2.9).

Our first result on the relation between these models is in terms of mutual information.

Theorem 2.4.3. *Define the mutual information per coordinate in asymmetric model of Eqs. (2.8), via*

$$\mathrm{I}_n^{\mathrm{asym}}(\mu_\Lambda, \mu_\Theta) := \frac{1}{n}\mathbb{E}\log\frac{\mathrm{d}\mathbb{P}_{\mathbf{\Lambda},\boldsymbol{A}}}{\mathrm{d}(\mathbb{P}_{\mathbf{\Lambda}}\times\mathbb{P}_{\boldsymbol{A}})}(\mathbf{\Lambda},\boldsymbol{A})\,. \tag{2.15}$$

Further recall the definition of mutual information in the symmetric model (2.9) given by Eq. (2.11). Within the setting of Assumption 2.4.1, we let $n, d \to \infty$ *simultaneously with* $d/n \to \infty$. *In addition, we require without loss of generality* $\boldsymbol{Q}_\Theta = q_\Theta \boldsymbol{I}_r$, *cf. Remark 2.3.1. Then the following limits exist and are equal*

$$\lim_{n,d\to\infty} \mathrm{I}_n^{\mathrm{asym}}(\mu_\Lambda, \mu_\Theta) = \lim_{n\to\infty} \mathrm{I}_n^{\mathrm{symm}}(\mu_\Lambda; q_\Theta)\,.$$

The proof of Theorem 2.4.3 is presented in Appendix A.4.2 for μ_Λ with bounded support. The generalization to μ_Λ with unbounded support is discussed in Appendix A.4.4.

As mentioned above, earlier work determined the asymptotics of the mutual information for the symmetric model $\mathrm{I}_n^{\mathrm{symm}}(\mu_\Lambda; q_\Theta)$. In particular, the next corollary follows directly from Theorem 2.4.3 and Theorem 2.4.1.

Corollary 2.4.1. *Recall that* S_r^+ *denotes the set of* $r\times r$ *positive semidefinite matrices, and* $\mathcal{F} : \mathbb{R}_{\geq 0}\times S_r^+ \to \mathbb{R}$ *is defined in Eq. (2.12). Under the conditions of Theorem 2.4.3, we have*

$$\begin{aligned}\lim_{n,d\to\infty} \mathrm{I}_n^{\mathrm{asym}}(\mu_\Lambda, \mu_\Theta) &= \frac{1}{4}q_\Theta^2 \|\mathbb{E}_{\mathbf{\Lambda}_0\sim\mu_\Lambda}[\mathbf{\Lambda}_0\mathbf{\Lambda}_0^\mathsf{T}]\|_F^2 - \mathcal{F}_*(q_\Theta)\\ &:= \frac{1}{4}q_\Theta^2 \|\mathbb{E}_{\mathbf{\Lambda}_0\sim\mu_\Lambda}[\mathbf{\Lambda}_0\mathbf{\Lambda}_0^\mathsf{T}]\|_F^2 - \sup_{\boldsymbol{Q}\in S_r^+} \mathcal{F}(q_\Theta^2, \boldsymbol{Q})\,.\end{aligned}$$

Recall the de Bruijn identity relating mutual information and minimum mean square error, see [183, 97, 64]:

$$\frac{1}{4}\mathrm{MMSE}_n^{\mathrm{symm}}(\mu_\Lambda; \sqrt{s}) = \frac{\mathrm{d}}{\mathrm{d}s}\mathrm{I}_n^{\mathrm{symm}}(\mu_\Lambda; \sqrt{s})\,. \tag{2.16}$$

Since $\mathrm{MMSE}_n^{\mathrm{symm}}(\mu_\Lambda; \sqrt{s})$ is non-increasing in s, the asymptotics of $\mathrm{I}_n^{\mathrm{symm}}(\mu_\Lambda; \sqrt{s})$ essentially determines the asymptotics of $\mathrm{MMSE}_n^{\mathrm{symm}}(\mu_\Lambda; \sqrt{s})$. Namely, we have $\lim_{n\to\infty}\mathrm{MMSE}_n^{\mathrm{symm}}(\mu_\Lambda; \sqrt{s}) = \|\mathbb{E}_{\mathbf{\Lambda}_0\sim\mu_\Lambda}[\mathbf{\Lambda}_0\mathbf{\Lambda}_0^\mathsf{T}]\|_F^2 - 4\frac{\partial}{\partial s}\mathcal{F}_*(\sqrt{s})$ for almost all values of s.

It would be tempting to conclude that Theorem 2.4.3 and Corollary 2.4.1 lead directly to analogous theorems relating $\mathrm{MMSE}_n^{\mathrm{asym}}(\mu_\Lambda, \mu_\Theta)$ and $\mathrm{MMSE}_n^{\mathrm{symm}}(\mu_\Lambda; q_\Theta)$. Establishing such a consequence is more challenging than one would naively expect because we do not have an identity analogous[1] to Eq. (2.16) for the asymmetric model. We can nevertheless establish the following, via a perturbation argument.

[1] One could differentiate the mutual information $\mathrm{I}_n^{\mathrm{asym}}(\mu_\Lambda, \mu_\Theta)$ with respect to the signal-to-noise ratio parameter q_Θ, but the result is related to error in estimating $\mathbf{\Lambda}\mathbf{\Theta}^\mathsf{T}$ instead of $\mathbf{\Lambda}\mathbf{\Lambda}^\mathsf{T}$.

Theorem 2.4.4. *Under the conditions of Theorem 2.4.3, for all but countably many values of $q_\Theta > 0$, we have*

$$\liminf_{n,d\to\infty} \mathrm{MMSE}_n^{\text{asym}}(\mu_\Lambda, \mu_\Theta) \geq \lim_{n\to\infty} \mathrm{MMSE}_n^{\text{symm}}(\mu_\Lambda; q_\Theta). \tag{2.17}$$

Further, consider a modified model in which the statistician observes $(\boldsymbol{A}, \boldsymbol{Y}'(\varepsilon))$, where $\boldsymbol{A}$ is given by Eq. (2.8), and

$$\boldsymbol{Y}'(\varepsilon) := \frac{\sqrt{\varepsilon}}{n} \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}} + \boldsymbol{W}', \quad \boldsymbol{W}' \sim \mathrm{GOE}(n)\,.$$

Here, we assume $\boldsymbol{W}'$ is independent of everything else. Denote by $\mathrm{MMSE}_n^{\text{asym}}(\mu_\Lambda, \mu_\Theta; \varepsilon)$ the corresponding matrix mean square error. Then, for all but countably many values of $q_\Theta > 0$, we have

$$\lim_{\varepsilon\to 0+} \limsup_{n,d\to\infty} \mathrm{MMSE}_n^{\text{asym}}(\mu_\Lambda, \mu_\Theta; \varepsilon) \leq \lim_{n\to\infty} \mathrm{MMSE}_n^{\text{symm}}(\mu_\Lambda; q_\Theta). \tag{2.18}$$

The proof of Theorem 2.4.4 is outlined in Appendix A.4.3 (for distributions with bounded support) and A.4.4 (for the general case).

Remark 2.4.1. In particular, Theorem 2.4.4 establishes that the estimation errors under the symmetric and asymmetric models coincide asymptotically, provided that the error in the perturbed model $\mathrm{MMSE}_n^{\text{asym}}(\mu_\Lambda, \mu_\Theta; \varepsilon)$ is uniformly continuous (in n) as $\varepsilon \downarrow 0$. We expect this to be generically the case, but proving this remains an open problem.

The next theorem establishes a sequence of sufficient conditions under which we can prove asymptotic equivalence of estimation errors in the asymmetric and symmetric models.

Theorem 2.4.5. *Under the conditions of Theorem 2.4.3, we further assume at least one of the following conditions holds:*

(*a*) $dn^{-3}(\log n)^{-6} \to \infty$.

(*b*) $d(\log d)^{8/5}/n^{6/5} \to 0$ *and μ_Λ has bounded support.*

(*c*) *For the case $r = 1$, define $Y = \sqrt{\gamma}\Lambda_0 + G$ with $G \sim \mathsf{N}(0,1)$ independent of $\Lambda_0 \sim \mu_\Lambda$, and define $\mathsf{I}(\gamma) = \mathbb{E} \log \frac{\mathrm{d}p_{Y|\Lambda_0}}{\mathrm{d}p_Y}(Y, \Lambda_0)$. Let*

$$\Psi(\gamma, s) = \frac{s^2}{4} + \frac{\gamma^2}{4s} - \frac{\gamma}{2} + \mathsf{I}(\gamma).$$

Assume that the global maximum of $\gamma \mapsto \Psi(\gamma, q_\Theta)$ over $(0, \infty)$ is also the first stationary point of the same function.

Then, we have

$$\lim_{n,d\to\infty} \mathrm{MMSE}_n^{\text{asym}}(\mu_\Lambda, \mu_\Theta) = \lim_{n\to\infty} \mathrm{MMSE}_n^{\text{symm}}(\mu_\Lambda; q_\Theta). \tag{2.19}$$

(For condition (b), the conclusion is guaranteed to hold for all but countably many values of $q_\Theta > 0$.)

We defer the proof of Theorem 2.4.5 to Appendix A.4.5.

As anticipated in the introduction, the results presented in Section 2.4.2 and Section 2.4.3 support two key statistical insights, which we next summarize:

1. In the weak signal regime, it is possible to partially recover $\mathbf{\Lambda}$ while impossible to recover $\mathbf{\Theta}$ in any non-trivial sense. For instance, in the high-dimensional Gaussian mixture model, we might be able to estimate the labels, even if it is impossible to estimate the cluster centers.

   In the next section, we will further explore the application of these results to Gaussian mixture models, while in Section 2.6 we will investigate such asymmetry in real world datasets.

2. In this regime, ideal estimation accuracy is asymptotically independent of the distribution of the high-dimensional factor $\mathbf{\Theta}$. As demonstrated, for instance, by Eq. (2.19), the only dependence on μ_Θ is through its second moment.

The three sufficient conditions given in Theorem 2.4.5 correspond to three different arguments.

The most straightforward case is the one of condition (a). We use the fact that

$$\frac{1}{\sqrt{nd}}\big(\boldsymbol{A}\boldsymbol{A}^{\mathsf{T}} - d\boldsymbol{I}_n\big) = \frac{1}{n}\mathbf{\Lambda}\hat{\boldsymbol{Q}}_\Theta\mathbf{\Lambda}^{\mathsf{T}} + \frac{1}{\sqrt{nd}}\big(\boldsymbol{Z}\boldsymbol{Z}^{\mathsf{T}} - d\boldsymbol{I}_n\big) + \text{cross terms}\,, \tag{2.20}$$

where $\hat{\boldsymbol{Q}}_\Theta := \mathbf{\Theta}^{\mathsf{T}}\mathbf{\Theta}/d \approx q_\Theta \boldsymbol{I}_r$. For $d \gg n^3$, [40] proved that the total variation distance between the distribution of the Wishart matrix $\big(\boldsymbol{Z}\boldsymbol{Z}^{\mathsf{T}} - d\boldsymbol{I}_n\big)/\sqrt{nd}$ and the one of $\boldsymbol{W} \sim \mathrm{GOE}(n)$ converges to 0. While we still have to deal with the cross terms, under this condition the two models are close to each other.

For $d \ll n^3$ the Wishart and GOE distributions are asymptotically mutually singular [40], and therefore proving asymptotic equality of the mean square error has to rely on a more carefully analysis. In fact, the proof of part (b) follows a different path and relies heavily on Theorem 2.4.4.

Finally, part (c) combines the bound of Theorem 2.4.4 with a matching bound that is based on the analysis of a Bayesian approximate message passing (AMP) algorithm [152]. Indeed, the sufficient condition of part (c) coincides with the condition that Bayes AMP achieves Bayes optimal estimation error.

2.5 Clustering under the Gaussian mixture model

As an application of our theory, we consider clustering under Gaussian mixture model (GMM). Throughout, we will assume that all Gaussian components have equal covariance $\mathbf{\Sigma}$, and that $\mathbf{\Sigma}$ is known. Without loss of generality, we can therefore assume that data are preprocessed so that $\mathbf{\Sigma} = \boldsymbol{I}_d$. We will focus on the weak signal regime, because it is mathematically the most interesting regime.

The Gaussian mixture model fits our general framework, with the $\mathbf{\Lambda}_i$'s encoding the data point labels: $\mathbf{\Lambda}_i$ takes k possible values, with k being the number of clusters. We will measure estimation accuracy using the overlap

$$\mathrm{Overlap}_n := \max_{\pi\in\mathfrak{S}_k} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{\hat{\mathbf{\Lambda}}_i = \mathbf{\Lambda}_i^\pi\right\}\,.$$

Here, $\mathfrak{S}_k$ denotes the group of permutations over k elements, and $\mathbf{\Lambda}_i^\pi$ denotes the action of this group on the cluster label encodings of the i-th sample. (We will work with slightly different encodings for the cases

$k = 2$ and $k \geq 3$ below.)

2.5.1 Two clusters with symmetric centers

As a warm-up example, we consider the case of $k = 2$ clusters with equal weights. For $i \in \{1, \dots, n\}$, we observe an independent sample

$$\boldsymbol{a}_i \sim \frac{1}{2}\,\mathsf{N}(\boldsymbol{\Theta}/\sqrt[4]{nd}, \boldsymbol{I}_d) + \frac{1}{2}\,\mathsf{N}(-\boldsymbol{\Theta}/\sqrt[4]{nd}, \boldsymbol{I}_d)\,. \tag{2.21}$$

Here, $\pm\boldsymbol{\Theta}/\sqrt[4]{nd} \subseteq \mathbb{R}^d$ are the cluster centers. Denoting by $(\Lambda_i)_{i\leq n} \overset{iid}{\sim} \text{Unif}(\{-1,+1\})$ the cluster labels, and by $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ the matrix whose i-th row corresponds to the i-th sample, we have that $\boldsymbol{A}$ follows model (2.8) with $r = 1$.

We further assume $\boldsymbol{\Theta}$ has independent coordinates: $(\Theta_j)_{j\leq d} \overset{iid}{\sim} \mu_\Theta$, where μ_Θ is a centered sub-Gaussian distribution and has zero third moment.

Remark 2.5.1. We note that, for $r = 1$, there is no real loss of generality in assuming $(\Theta_j)_{j\leq d}$ to be i.i.d. sub-Gaussian. Indeed, model (2.8) is equivariant under rotations $\boldsymbol{\Theta} \mapsto \boldsymbol{\Omega}\boldsymbol{\Theta}$, $\boldsymbol{A} \mapsto \boldsymbol{A}\boldsymbol{\Omega}^{\mathsf{T}}$, where $\boldsymbol{\Omega} \in \mathbb{R}^{d\times d}$ is an orthogonal matrix. Further, any loss function that depends uniquely on $\boldsymbol{\Lambda}$ is also invariant under the same group. Consider minimax estimation when $\boldsymbol{\Theta}$ belongs to the sphere: $\|\boldsymbol{\Theta}\|_2^2 = d\,q_\Theta$. As a consequence of the Hunt-Stein theorem, the least favorable prior is the uniform distribution over the same sphere. We expect the asymptotic Bayes risk under this prior (and therefore the minimax risk) to be the same as the risk under the prior $(\Theta_j)_{j\leq d} \overset{iid}{\sim} \mathsf{N}(0, q_\Theta)$.

Proposition 2.5.1. *Consider the Gaussian mixture model as in Eq. (2.21). Assume $n, d \to \infty$ simultaneously and $d/n \to \infty$, then the following results hold:*

(*a*) *If $q_\Theta \leq 1$, then, for any clustering estimator $\hat{\boldsymbol{\Lambda}}$, as $n, d \to \infty$ we have*

$$\text{Overlap}_n \overset{P}{\to} \frac{1}{2}\,. \tag{2.22}$$

(*b*) *If $q_\Theta > 1$, let s_* be the largest non-negative solution of*

$$s = q_\Theta^2\mathbb{E}\big\{\tanh\big(s + \sqrt{s}G\big)^2\big\}\,, \tag{2.23}$$

where $G \sim \mathsf{N}(0,1)$. Then $s_* > 0$ and there exists an estimator achieving

$$\text{Overlap}_n \overset{P}{\to} \Phi(\sqrt{s_*}), \tag{2.24}$$

where Φ denotes the cumulative distribution function for standard Gaussian distribution.

The proof of this result uses the characterization of optimal estimation in the corresponding symmetric model proven in [64], and we present the proof of point (a) in Appendix A.7.1. The overlap in point (b) can be achieved using orthogonal invariant Bayes AMP with spectral initialization on $(\boldsymbol{A}\boldsymbol{A}^{\mathsf{T}} - d\boldsymbol{I}_n)/\sqrt{nd}$. This algorithm is described and analyzed in [148].

2.5.2 Two or more clusters with orthogonal centers

We next consider the case of $k \geq 2$ clusters with approximately orthogonal centers. We denote by $\{\boldsymbol{\Theta}_{\cdot i}/\sqrt[4]{nd} : i \in [k]\} \subseteq \mathbb{R}^d$ the cluster centers. Let $\boldsymbol{\Theta} \in \mathbb{R}^{d\times k}$ with the i-th column given by $\boldsymbol{\Theta}_{\cdot i}$. For $j \in [d]$, we let $\boldsymbol{\Theta}_j \in \mathbb{R}^k$ be the j-th row of $\boldsymbol{\Theta}$. We assume $\boldsymbol{\Theta}_j \overset{iid}{\sim} \mu_\Theta$, where μ_Θ is sub-Gaussian with vanishing first and third moments and diagonal covariance: $\mathrm{Cov}(\boldsymbol{\Theta}_1) = q_\Theta \boldsymbol{I}_k$.

Let $\boldsymbol{e}_j$ be the j-th standard basis vector in $\mathbb{R}^k$. We encode the data point labels by setting $\boldsymbol{\Lambda}_i = \boldsymbol{e}_j$ if and only if the i-th sample belongs to the j-th cluster, and consider the case of equal proportions, so that $(\boldsymbol{\Lambda}_i)_{i\leq n} \overset{iid}{\sim} \mathrm{Unif}(\{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_k\})$.

As before, we let $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ be the matrix whose rows are i.i.d. samples $\boldsymbol{a}_i$ from the Gaussian mixture model with centers $\{\boldsymbol{\Theta}_{\cdot i}/\sqrt[4]{nd} : i \in [k]\}$. With these definitions, the matrix $\boldsymbol{A}$ is distributed according to model (2.8).

Remark 2.5.2. While we state our results for random $\boldsymbol{\Theta}$, we can generalize Remark 2.5.1 to the present setting. This argument implies that the results of this section also characterize the minimax estimation error over the class of problems with orthogonal centers $\boldsymbol{\Theta}^\mathsf{T}\boldsymbol{\Theta} = d\, q_\Theta \boldsymbol{I}_k$.

Our next theorem establishes the threshold for weak recovery of the cluster labels in the high-dimensional regime $d/n \to \infty$. Recall the function $\mathcal{F}(s, \boldsymbol{Q})$ is defined in Eq. (2.12), where we take $\mu_\Lambda = \sum_{i=1}^k \delta_{\boldsymbol{e}_i}/k$. We let $\boldsymbol{Q}_0 := \mathbf{1}_k\mathbf{1}_k^\mathsf{T}/k^2$ and define the threshold

$$q_\Theta^{\mathsf{info}}(k) := \inf\left\{q_\Theta \geq 0 : \sup_{\boldsymbol{Q}\in S_k^+} \mathcal{F}(q_\Theta^2, \boldsymbol{Q}) > \mathcal{F}(q_\Theta^2, \boldsymbol{Q}_0)\right\}. \tag{2.25}$$

Theorem 2.5.1. *Consider the Gaussian mixture model with k components of equal weights, in the high-dimensional asymptotics $d, n \to \infty$, $d/n \to \infty$. Under the above assumptions on the centers $\boldsymbol{\Theta}$, the following results hold:*

(*a*) *If $q_\Theta < q_\Theta^{\mathsf{info}}(k)$, then for any estimator $\hat{\boldsymbol{\Lambda}} : \mathbb{R}^{n\times d} \to \{\boldsymbol{e}_j : j \in [k]\}^n$ that is a measurable function of the input $\boldsymbol{A}$, we have*

$$\operatorname*{p-lim}_{n,d\to\infty} \mathrm{Overlap}_n = \frac{1}{k}.$$

(*b*) *Assume either $dn^{-3}(\log n)^{-6} \to \infty$ or $dn^{-6/5}(\log d)^{8/5} \to 0$. If $q_\Theta > q_\Theta^{\mathsf{info}}(k)$, then there exists an estimator $\hat{\boldsymbol{\Lambda}} : \mathbb{R}^{n\times d} \to \{\boldsymbol{e}_j : j \in [k]\}^n$ that is a measurable function of the input $\boldsymbol{A}$, such that*

$$\liminf_{n,d\to\infty} \mathbb{E}[\mathrm{Overlap}_n] > \frac{1}{k}.$$

We defer the proofs of parts (a) and (b) of Theorem 2.5.1 to Appendices A.7.2 and A.7.3, respectively. Note that [18, Theorem 2] implies $q_\Theta^{\mathsf{info}}(k) = 2\sqrt{k\log k}\cdot(1+o_k(1))$ as $k \to \infty$ (however, [18] does not establish a sharp threshold). In contrast, Theorem 2.5.1 derives the exact threshold for every k.

Table 2.1 collects values for the thresholds $q_\Theta^{\mathsf{info}}(k)$ for a few values of k, as obtained by numerically evaluating Eq. (2.25). For $k \leq 4$, this is expected to coincide with the spectral threshold, namely $q_\Theta^{\mathsf{info}}(k) = k$ [126]. (Notice that the apparent discrepancy with the threshold for $k = 2$ in the previous section is due to the different normalization adopted here.)

k	$q_\Theta^{\mathsf{info}}(k)$
2	2
3	3
4	4
5	4.95
6	5.81
7	6.61
8	7.36

Table 2.1: Information-theoretic thresholds $q_\Theta^{\mathsf{info}}(k)$ for $2 \le k \le 8$.

2.5.3 Numerical experiments

We present in this section numerical experiments suggesting that the theory of the last sections is already relevant at moderate values of d, n. We consider several clustering methods, and compare their performances with the threshold $q_\Theta^{\mathsf{info}}(k)$.

For our experiment we use built-in functions in Python3, for the following clustering methods:

(1) Lloyd's algorithm, as implemented by the function KMeans() in the scikit-learn module with option algorithm = "lloyd".

(2) Agglomerative clustering, implemented by the function AgglomerativeClustering() in the scikit-learn module with default parameters.

(3) EM algorithm, implemented by the function GaussianMixture() in the scikit-learn module with default parameters.

(4) A semidefinite programming (SDP) relaxation described in [169]. We use the cvxpy module for the optimization steps.

In Figure 2.1 we present results for these algorithms for $n = 100$, $d = 2000$, and $\mu_\Theta = \mathsf{N}(0, q_\Theta)$. For each value of the pair (k, q_Θ), we run 100 independent trials, and plot the average overlap versus q_Θ. For the case $k = 2$, we consider two slightly different settings: "Symmetric=True" corresponds to the case of two centers symmetric around the origin, as in Section 2.5.1, and "Symmetric=False" corresponds to the case of two approximately orthogonal centers as per Section 2.5.2. We also report the threshold $q_\Theta^{\mathsf{info}}(k)$, its large k approximation $2\sqrt{k \log k}$, and the algorithmic threshold $q_\Theta^{\mathsf{algo}}(k) = k$ (this is the conjectured threshold for efficient recovery, which coincides with the spectral threshold [126, 152]).

Despite the small sample size, we observe that $q_\Theta^{\mathsf{info}}(k)$ appears to capture the onset of non-trivial clustering accuracy across multiple algorithms.

2.6 Asymmetry in factors estimation: real world datasets

Previous sections imply the existence of gaps in the estimation of $\mathbf{\Lambda}$ and $\mathbf{\Theta}$, in the high-dimensional asymptotics $d/n \to \infty$. In summary, in the strong signal regime, $\mathbf{\Lambda}$ can be estimated consistently up to a potential rotation, while $\mathbf{\Theta}$ can only be partially recovered. On the other hand, in the weak signal regime, $\mathbf{\Lambda}$ can be partially recovered, while no estimator achieves better asymptotic performance than a naive one in terms of the estimation of $\mathbf{\Theta}$.

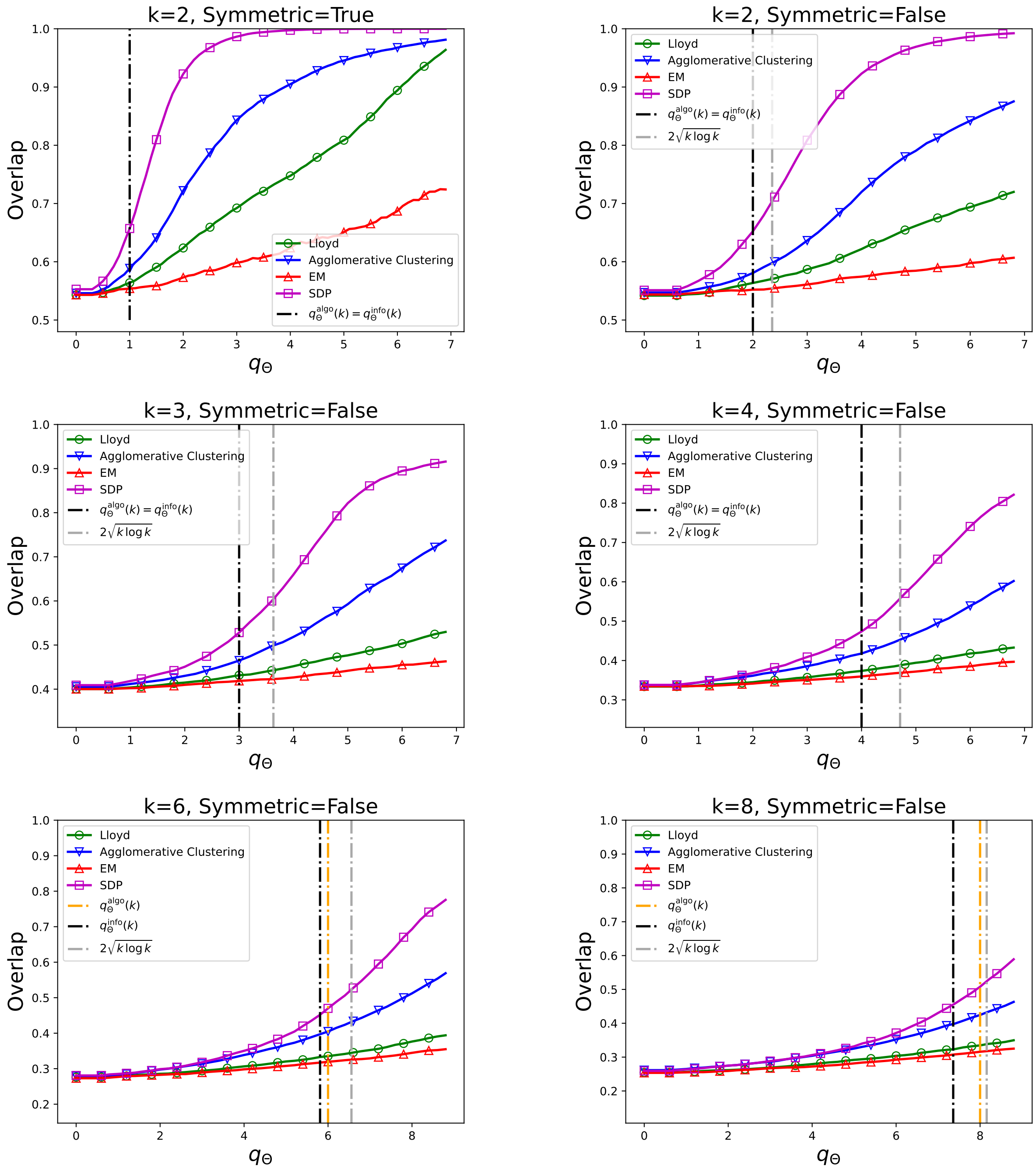


Figure 2.1: Average overlap achieved by several clustering algorithms on the Gaussian mixture model with $n = 100$ datapoints, $d = 2000$ dimensions, averaged over 100 instances. The black vertical line corresponds to the information-theoretic threshold for identifying clusters significantly better than random guessing; the orange vertical line corresponds to the spectral or algorithmic threshold; the grey vertical line corresponds to the approximated information-theoretic threshold $2\sqrt{k \log k}$.

In this section, we investigate this asymmetry in real world datasets. We focus on a problem that can be modeled as clustering with $k = 2$ clusters (this can be modeled as a GMM model, leading to Eq. (2.1) with $r = 1$ as described in the previous section).

2.6.1 1000 Genomes Project

Our first experiment involves genotype data from the 1000 Genomes Project [59]. This provides genotypes for $n = 2,504$ individuals grouped in five population groups (corresponding to their geographic origins). For our experiments, we extract $d = 100,000$ common single-nucleotide polymorphisms (SNPs). Our preprocessing steps follow from [200]. After preprocessing, we add independent Gaussian noise with variance 5 to the data matrix, to make the problem more challenging.

Principal component analysis (PCA) is often used in genome-wide association studies, in particular to explore the genetic structure of human populations [163, 164]. As a first step of our experiment, for each pair of population groups, we randomly extract 30 subjects from each group without replacement. The subsampled observations form a $60 \times 100,000$ genotype matrix, the columns of which are then centered and rescaled. We next run PCA on this subset, and plot the projections onto the top 2 principal components. We display one typical outcome of PCA in Figure 2.2. From the figures, we see that despite the high-dimensionality, PCA still reflects the underlying population structure. We interpret this as indicating that non-trivial clustering can be achieved on these data.

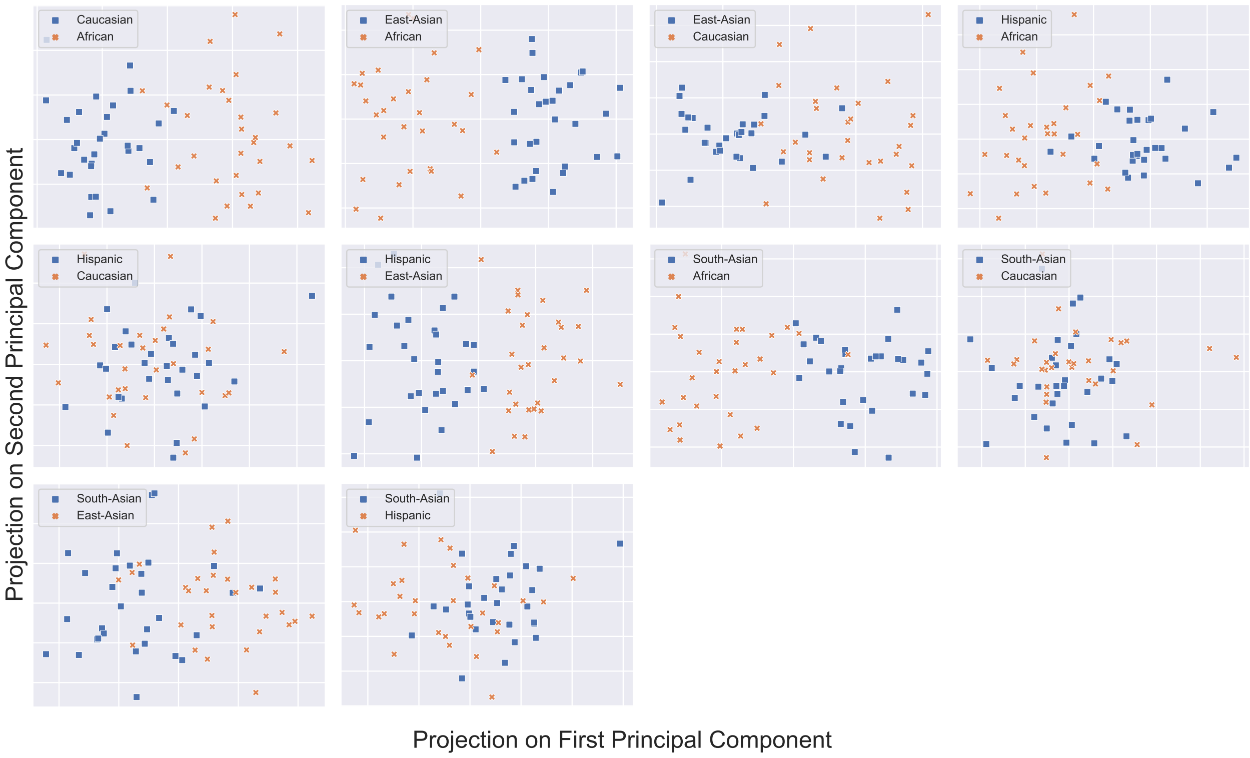


Figure 2.2: Illustration of PCA on a subset of 1000 Genomes Project data. In these plots, the x axis represents the projection onto the first principal component, and the y axis represents projection onto the second principal component. Point colors and shapes correspond to population groups. Each experiment involves 60 individuals in total, with 30 individuals from each of the two population groups.

To further support our conclusion, we run K-means clustering on the subsampled datasets (using sklearn.cluster.KMeans in Python 3 with default parameters). We then compute the overlap between the true and estimated labels (in this example labels correspond to population groups). We repeat this procedure independently 1000 times on randomly selected subsets of the data. The outcomes are recorded and displayed in the lower triangle of Figure 2.3. From the figures, we see that K-means clustering estimates the labels significantly better than random guessing (i.e., better than 50% accuracy) and achieves near-perfect recovery for certain pairs of population groups.

We next estimate the cluster centers $\boldsymbol{\Theta}$, for each pair of population groups. We take two non-overlapping subsets of the data (each with size 60) and run K-means on each subset: this leads to two distinct estimates of the cluster centers $(\hat{\boldsymbol{\Theta}}_i^{(1)})_{i\in\{1,2\}}$ for data subset 1, and $(\hat{\boldsymbol{\Theta}}_i^{(2)})_{i\in\{1,2\}}$ for data subset 2. We then compute the maximum normalized inner product

$$\max_{i,j\le 2} |\langle \hat{\boldsymbol{\Theta}}_i^{(1)}, \hat{\boldsymbol{\Theta}}_j^{(2)} \rangle| / (\|\hat{\boldsymbol{\Theta}}_i^{(1)}\|_2 \|\hat{\boldsymbol{\Theta}}_j^{(2)}\|_2)$$

between the estimated cluster centers obtained via K-means from these two subsets of data.

This procedure is again repeated for 1000 times independently, and the distributions of the maximum normalized inner products are displayed in the upper triangle of Figure 2.3. We observe that, for several population pairs, the estimates $\hat{\boldsymbol{\Theta}}^{(1)}$, $\hat{\boldsymbol{\Theta}}^{(2)}$ are not significantly correlated (using initials, this is the case for the pairs C-H, C-SA, EA-H, EA-SA, H-SA). Since these estimates are obtained based on independent samples from the same population, we conclude that they are also not significantly correlated with the true centers. When this happens, the behavior of this clustering problem seems to be captured by the weak signal regime analyzed in the previous sections: clusters can be estimated in a non-trivial way, but cluster centers cannot be estimated.

For the other population pairs, the cluster centers estimates are correlated, and clustering accuracy is very high (this is the case for pairs A-C, A-EA, A-H, A-SA, with C-EA not as clear a case). This is analogous to what we observe in our model in the strong signal regime.

2.6.2 RNA-Seq gene expression

We carry out a similar experiment on gene expression data for different types of cancers from the UCI Machine Learning Repository[2] [73]. The dataset contains 801 samples and 20531 attributes, with the predictors being RNA-Seq gene expression levels measured by the Illumina HiSeq platform. Before proceeding, again we apply additive Gaussian noise to the data matrix, with mean zero and variance 5. We consider five different cancer types, denoted by “COAD”, “BRCA”, “KIRC”, “LUAD” and “PRAD”.

For each pair of cancer groups we subsample 30 subjects from each group, to construct a 60×20531 data matrix. We then center and rescale the columns of this matrix to unit norms. A typical outcome of PCA is presented in Figure 2.4. We observe that clusters corresponding to different cancer groups are well separated for each of the pairs. In Figure 2.5, we report the overlaps between the labels obtained from K-means clustering and the ground truth labels. The overlaps are very high for all pairs. These plots summarize the results of 1000 independent repetitions of this experiment.

In the upper half of the same figure, we present the maximum normalized inner products between the

[2] https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq

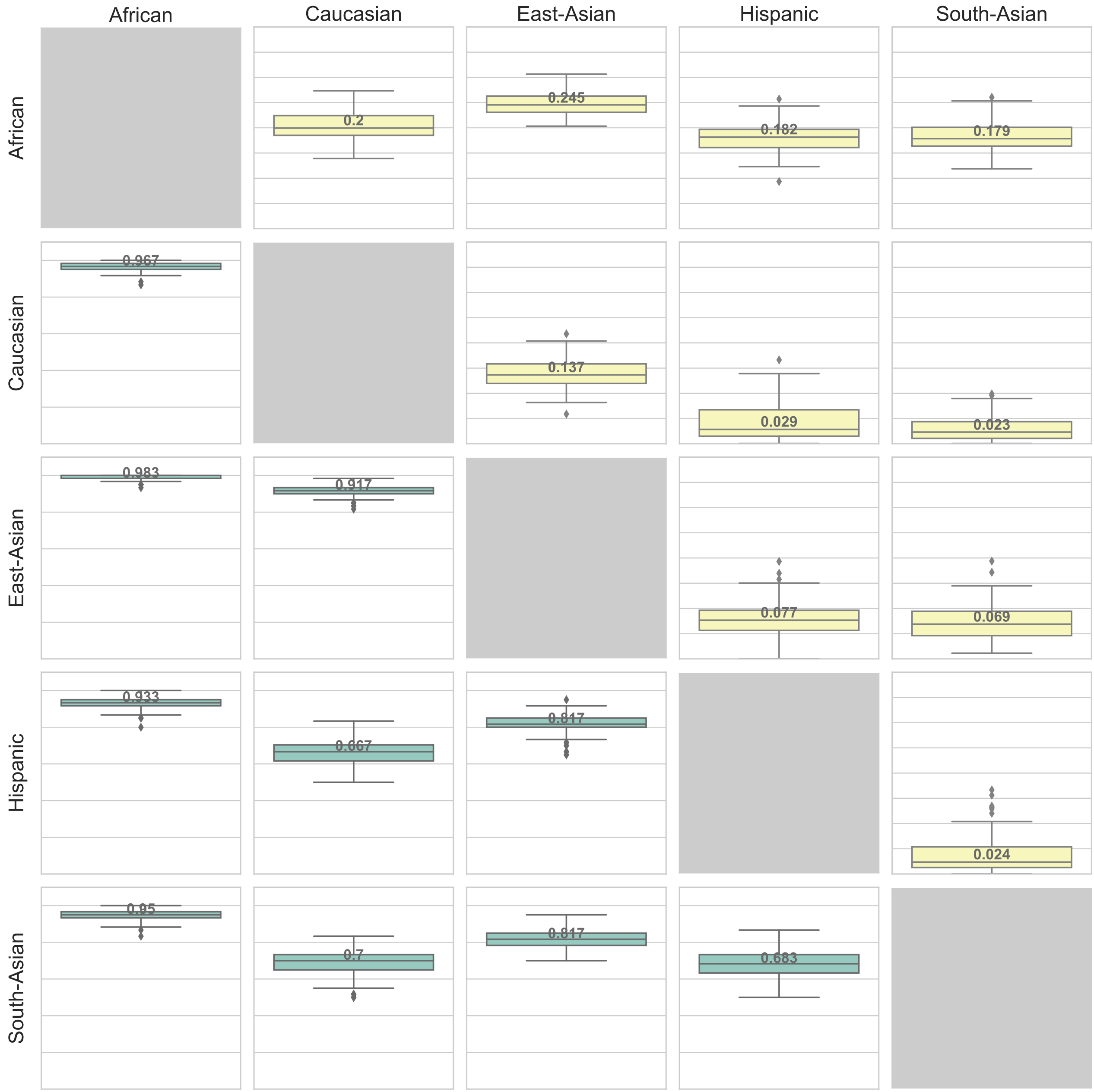


Figure 2.3: Simulation results derived from 1000 independent experiments for the 1000 Genomes Project dataset. The boxplots in the upper triangle display the quantiles of the normalized inner products between the estimated cluster centers. The boxplots in the lower triangle display the quantiles of overlaps between true labels and labels obtained via K-means clustering. We annotate the medians in the corresponding figures for readers' convenience.

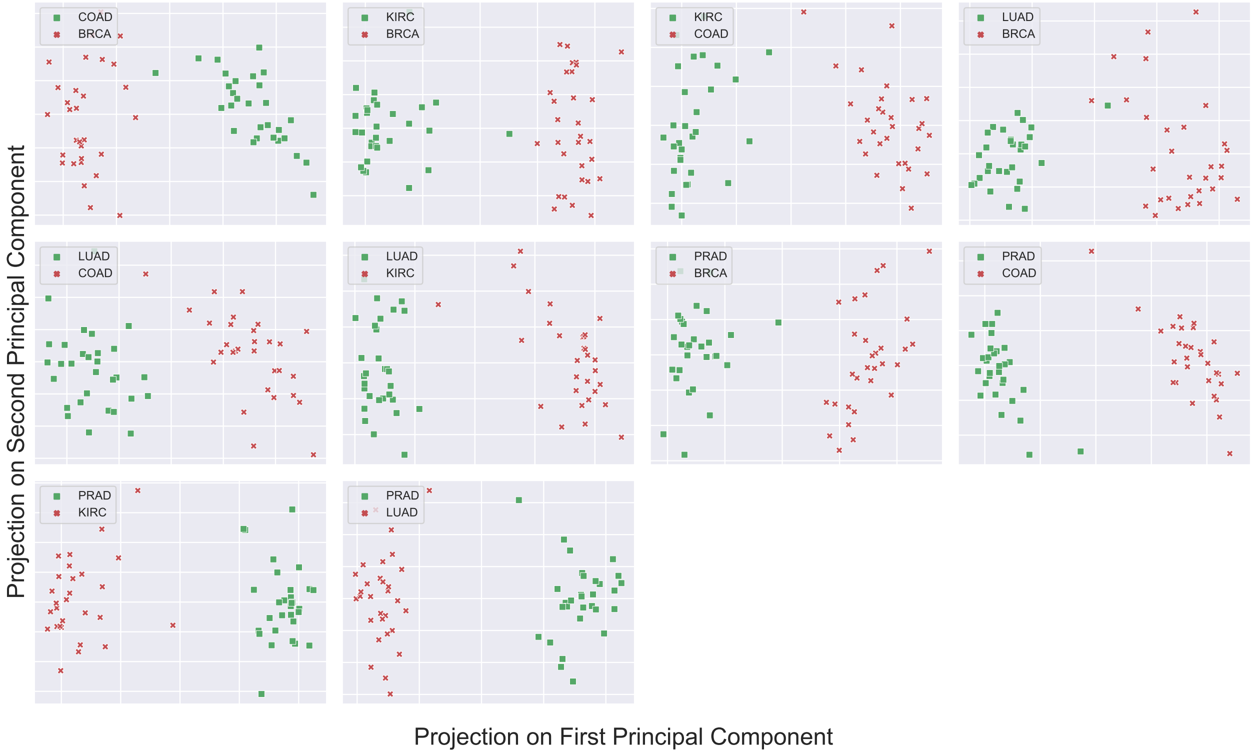


Figure 2.4: PCA on subsets of RNA-Seq gene expression data: each time we select 30 datapoints from each of the two cancer groups at random. Point colors and shapes stand for different cancer groups. We plot the projections of these datapoints onto the subspace defined by their first two principal components.

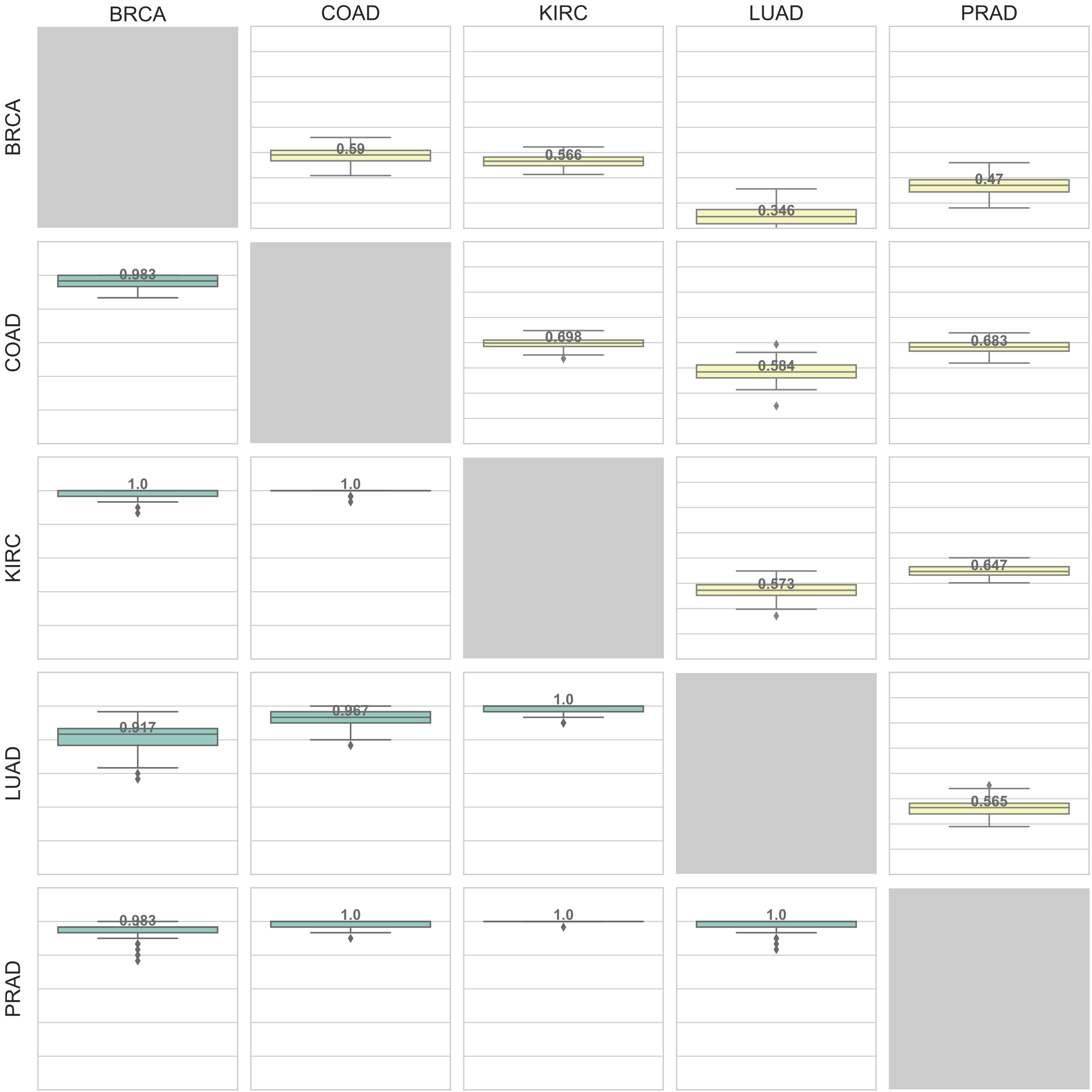


Figure 2.5: Simulation results derived from 1000 independent experiments for the UCI gene expression dataset. The boxplots in the upper triangle display the quantiles of the normalized inner products between estimated cluster centers. The boxplots in the lower triangle display the quantiles of the accuracy (overlap) in reconstructing the true clusters. Medians are annotated in the figures.

estimated cluster centers on two independent subsamples. The correlation is significantly different from zero, but far from being close to one. Once more, this is analogous to the strong signal regime in our analysis.

Chapter 3

The estimation error of general first order methods

3.1 Introduction

For this part, we not only consider low-rank matrix estimation, but also apply our results to high-dimensional regression. More precisely, we study the following two models:

High-dimensional regression. Data are i.i.d. pairs $\{(y_i, \boldsymbol{x}_i)\}_{i\le n}$, where $y_i \in \mathbb{R}$ is a label and $\boldsymbol{x}_i \in \mathbb{R}^p$ is a feature vector. We assume $\boldsymbol{x}_i \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_p/n)$ and $y_i|\boldsymbol{x}_i \sim \mathbb{P}(y_i \in \cdot|\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\theta})$ for a vector $\boldsymbol{\theta} \in \mathbb{R}^p$. Our objective is to estimate the coefficients θ_j from data $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ (the matrix whose i-th row is vector $\boldsymbol{x}_i$) and $\boldsymbol{y} \in \mathbb{R}^n$ (the vector whose i-th entry is label y_i).

Low-rank matrix estimation. Data consist of a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ where $x_{ij} = \frac{1}{n}\boldsymbol{\Lambda}_i^{\mathsf{T}}\boldsymbol{\theta}_j + z_{ij}$ with $\boldsymbol{\Lambda}_i, \boldsymbol{\theta}_j \in \mathbb{R}^r$ and $z_{ij} \overset{\text{iid}}{\sim} \mathsf{N}(0, 1/n)$. We denote by $\boldsymbol{\Lambda} \in \mathbb{R}^{n\times r}$ and $\boldsymbol{\theta} \in \mathbb{R}^{p\times r}$ the matrices whose rows are $\boldsymbol{\Lambda}_i^{\mathsf{T}}$ and $\boldsymbol{\theta}_j^{\mathsf{T}}$ respectively. Our objective is to to estimate $\boldsymbol{\Lambda}, \boldsymbol{\theta}$ from data $\boldsymbol{X}$.

In order to discuss these two examples in a unified fashion, we will introduce a dummy vector $\boldsymbol{y}$ (e.g., the all-zeros vector) as part of the data in the low-rank matrix estimation problem. Let us point out that our normalizations are somewhat different from, but completely equivalent to, the traditional ones in statistics.

The first question to address is how to properly define ‘first order methods.’ A moment of thought reveals that the above discussion in terms of a cost function $\mathcal{L}(\boldsymbol{\theta})$ needs to be revised. Indeed, given either of the above statistical models, there is no simple way to construct a ‘statistically optimal’ cost function.[1] Further, it is not clear that using a faster optimization algorithm for that cost will result in faster decrease of the estimation error.

We follow instead a different strategy and introduce the class of *general first order methods* (GFOM). In words, these include all algorithms that keep as state sequences of matrices $\boldsymbol{u}^1, \dots, \boldsymbol{u}^t \in \mathbb{R}^{n\times r}$, and $\boldsymbol{v}^1, \dots, \boldsymbol{v}^t \in \mathbb{R}^{p\times r}$, which are updated by two types of operations: row-wise application of a function, or multiplication by $\boldsymbol{X}$ or $\boldsymbol{X}^{\mathsf{T}}$. We will then show that standard first order methods, for common choices of the cost $\mathcal{L}(\boldsymbol{\theta})$, are in fact special examples of GFOMs.

[1] In particular, maximum likelihood is not statistically optimal in high dimension [25].

Formally, a GFOM is defined by sequences of functions $F_t^{(1)}, G_t^{(2)} : \mathbb{R}^{r(t+1)+1} \to \mathbb{R}^r$, $F_t^{(2)}, G_t^{(1)} : \mathbb{R}^{r(t+1)} \to \mathbb{R}^r$, with the F's indexed by $t \geq 0$ and the G's indexed by $t \geq 0$. In the high-dimensional regression problem, we set $r = 1$. The algorithm produces two sequences of matrices (vectors for $r = 1$) $(\boldsymbol{u}^t)_{t\geq 1}$, $\boldsymbol{u}^t \in \mathbb{R}^{n\times r}$, and $(\boldsymbol{v}^t)_{t\geq 1}$, $\boldsymbol{v}^t \in \mathbb{R}^{p\times r}$,

$$\boldsymbol{v}^{t+1} = \boldsymbol{X}^\top F_t^{(1)}(\boldsymbol{u}^1, \dots, \boldsymbol{u}^t; \boldsymbol{y}, \boldsymbol{u}) + F_t^{(2)}(\boldsymbol{v}^1, \dots, \boldsymbol{v}^t; \boldsymbol{v}) \tag{3.1a}$$

$$\boldsymbol{u}^t = \boldsymbol{X} G_t^{(1)}(\boldsymbol{v}^1, \dots, \boldsymbol{v}^t; \boldsymbol{v}) + G_t^{(2)}(\boldsymbol{u}^1, \dots, \boldsymbol{u}^{t-1}; \boldsymbol{y}, \boldsymbol{u}), \tag{3.1b}$$

where it is understood that each function is applied row-wise. For instance

$$F_t^{(1)}(\boldsymbol{u}^1, \dots, \boldsymbol{u}^t; \boldsymbol{u}) = (F_t^{(1)}(\boldsymbol{u}_i^1, \dots, \boldsymbol{u}_i^t; \boldsymbol{u}_i))_{i\leq n} \in \mathbb{R}^{n\times r},$$

where $(\boldsymbol{u}_i^s)^\top$ is the i^{th} row of $\boldsymbol{u}^s$. Here $\boldsymbol{u}, \boldsymbol{v}$ are either deterministic or random and independent of everything else. In particular, the iteration is initialized with $\boldsymbol{v}^1 = \boldsymbol{X}^\top F_0^{(1)}(\boldsymbol{y}, \boldsymbol{u}) + F_0^{(2)}(\boldsymbol{v})$. The unknown matrices (or vectors) $\boldsymbol{\theta}$ and $\boldsymbol{\Lambda}$ are estimated after t_* iterations by $\hat{\boldsymbol{\theta}} = G_*(\boldsymbol{v}^1, \cdots, \boldsymbol{v}^{t_*}; \boldsymbol{v})$ and $\hat{\boldsymbol{\Lambda}} = F_*(\boldsymbol{u}^1, \dots, \boldsymbol{u}^{t_*}; \boldsymbol{y}, \boldsymbol{u})$, where the latter only applies in the low-rank matrix estimation problem. Let us point out that the update also depend on additional information encoded in the two vectors $\boldsymbol{u} \in \mathbb{R}^n$, $\boldsymbol{v} \in \mathbb{R}^p$. This enables us to model side information provided to the statistician (e.g., an 'initialization' correlated with the true signal) or auxiliary randomness.

We study the regime in which $n, p \to \infty$ with $n/p \to \delta \in (0, \infty)$ and r is fixed. We assume the number of iterations t_* is fixed, or potentially $t_* \to \infty$ after $n \to \infty$. In other words, we are interested in linear-time or nearly linear-time algorithms (complexity being measured relative to the input size np). As mentioned above, our main result is a general lower bound on the minimum estimation error that is achieved by any GFOM in this regime.

This chapter is organized as follows: Section 3.2 illustrates the setting introduced above in two examples; Section 3.3 contains the statement of our general lower bounds; Section 3.4 applies these lower bounds to the two examples; Section 3.5 presents an outline of the proof, deferring technical details to appendices.

3.2 Two examples

Example #1: M-estimation in high-dimensional regression and phase retrieval

Consider the high-dimensional regression problem. Regularized M-estimators minimize a cost

$$\mathcal{L}_n(\boldsymbol{\vartheta}) := \sum_{i=1}^n \ell(y_i; \langle \boldsymbol{x}_i, \boldsymbol{\vartheta}\rangle) + \Omega_n(\boldsymbol{\vartheta}) = \hat{\ell}_n(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{\vartheta}) + \Omega_n(\boldsymbol{\vartheta}), \tag{3.2}$$

Here $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a loss function, $\hat{\ell}_n(\boldsymbol{y}, \hat{\boldsymbol{y}}) := \sum_{i=1}^n \ell(y_i, \hat{y}_i)$ is its empirical average, and $\Omega_n : \mathbb{R}^p \to \mathbb{R}$ is a regularizer. It is often the case that ℓ is smooth and Ω_n is separable, i.e., $\Omega_n(\boldsymbol{\vartheta}) = \sum_{i=1}^p \Omega_1(\vartheta_i)$. We will assume this to be the case in our discussion.

The prototypical first order method is proximal gradient [166]:

$$\boldsymbol{\theta}^{t+1} = \mathsf{Prox}_{\gamma_t\Omega_1}\big(\boldsymbol{\theta}^t - \gamma_t \nabla_{\boldsymbol{\vartheta}} \hat{\ell}_n(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{\theta}^t)\big),$$

$$\mathsf{Prox}_{\gamma\Omega_1}(y) := \arg\min_{\theta\in\mathbb{R}} \left\{ \frac{1}{2}(y-\theta)^2 + \gamma\Omega_1(\theta) \right\} .$$

Here $(\gamma_t)_{t\geq 0}$ is a sequence of step sizes and $\mathsf{Prox}_{\gamma\Omega_1}$ acts on a vector coordinate-wise. Notice that

$$\nabla_{\boldsymbol{\vartheta}}\hat{\ell}_n(\boldsymbol{y},\boldsymbol{X}\boldsymbol{\theta}^t) = \boldsymbol{X}^{\mathsf{T}} s(\boldsymbol{y},\boldsymbol{X}\boldsymbol{\theta}^t) , \quad s(\boldsymbol{h},\hat{\boldsymbol{y}})_i \equiv \frac{\partial\ell}{\partial\hat{y}_i}(y,\hat{y}_i) . \tag{3.3}$$

Therefore proximal gradient –for the cost function (3.2)– is an example of a GFOM. Similarly, mirror descent with a separable Bregman divergence and accelerated proximal gradient methods are easily shown to fit in the same framework.

Among the countless applications of regularized M-estimation, we will focus on the sparse phase retrieval problem. We want to reconstruct a sparse signal $\boldsymbol{\theta}\in\mathbb{R}^p$ but only have noisy measurements of the modulus $|\langle\boldsymbol{\theta},\boldsymbol{x}_i\rangle|$; that is, we lose the 'phase' of these projections. (We will consider for simplicity the case of a real-valued signal, but the generalization of our results to the complex case should be immediate.)

As a concrete model, we will assume that number of non-zero entries of $\boldsymbol{\theta}$ is $\|\boldsymbol{\theta}\|_0 \leq s_0$. From an information-theoretic viewpoint, it is known that $\boldsymbol{\theta}$ can be reconstructed accurately as soon as the number of measurements satisfies $n \geq Cs_0\log(p/s_0)$, with C a sufficiently large constant [130]. Several groups have investigated practical reconstruction algorithms by exploiting either semidefinite programming relaxations [130] or first order methods [175, 46, 42]. A standard approach would be to apply a proximal gradient algorithm to the cost function (3.2) with $\Omega_n(\boldsymbol{\vartheta}) = \lambda\|\boldsymbol{\vartheta}\|_1$. However, all existing global convergence guarantees for these methods require $n \geq Cs_0^2\log p$. Is the dependence on s_0^2 due to a fundamental computational barrier or an artifact of the theoretical analysis? Recently [179] presented partial evidence towards the possibility of 'breaking' this barrier, by proving that a first order method can accurately reconstruct the signal for $n \geq Cs_0\log(p/s_0)$, if it is initialized close enough to the true signal $\boldsymbol{\theta}$.

Example #2: Sparse PCA

In a simple model for sparse principal component analysis (PCA), we observe a matrix $\boldsymbol{X} = \frac{1}{n}\boldsymbol{\Lambda}\boldsymbol{\theta}^{\mathsf{T}} + \boldsymbol{Z} \in \mathbb{R}^{n\times p}$, where $\boldsymbol{\Lambda}\in\mathbb{R}^n$ has entries $(\lambda_i)_{i\leq n} \overset{\text{iid}}{\sim} \mathsf{N}(0,1)$, $\boldsymbol{\theta}\in\mathbb{R}^p$ is a sparse vector with $s_0 \ll p$ non-zero entries, and $\boldsymbol{Z}$ is a noise matrix with entries $(z_{ij})_{i\leq n, j\leq p} \overset{\text{iid}}{\sim} \mathsf{N}(0,1/n)$. Given data $\boldsymbol{X}$, we would like to reconstruct the signal $\boldsymbol{\theta}$. From an information-theoretic viewpoint, it is known that accurate reconstruction of $\boldsymbol{\theta}$ is possible if $n \geq Cs_0\log(p/s_0)$, with C a sufficiently large constant [5].

A number of polynomial time algorithms have been studied, ranging from simple thresholding algorithms [110, 67] to sophisticated convex relaxations [5, 135]. Among other approaches, one natural idea is to modify the power iteration algorithm of standard PCA by computing

$$\boldsymbol{\theta}^{t+1} = c_t\,\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\eta(\boldsymbol{\theta}^t;\gamma_t) . \tag{3.4}$$

Here $(c_t)_{t\geq 0}$ is a deterministic normalization, and $\eta(\,\cdot\,;\gamma)$ is a thresholding function at level γ, e.g., soft thresholding $\eta(x;\gamma) = \text{sign}(x)(|x|-\gamma)_+$. It is immediate to see that this algorithm is a GFOM. More elaborate versions of non-linear power iteration were developed, for example, by [113, 136], and are typically equivalent to suitable GFOMs.

Despite these efforts, no algorithm is known to succeed unless $n \geq Cs_0^2$. Is this a fundamental barrier or a limitation of present algorithms or analysis? Evidence towards intractability was provided by [31, 38] via reduction from the planted clique problem. Our analysis provides new evidence towards the same conclusion.

3.3 Main results

In this section we state formally our general results about high-dimensional regression and low-rank matrix estimation. The next section will apply these general results to concrete instances. Throughout we make the following assumptions:

A1. The functions $F_t^{(1)}, G_t^{(2)}, F_* : \mathbb{R}^{r(t+1)+1} \to \mathbb{R}$, $F_t^{(2)}, G_t^{(1)}, G_* : \mathbb{R}^{r(t+1)} \to \mathbb{R}$, are Lipschitz continuous, with the F's indexed by $t \geq 0$ and the G's indexed by $t \geq 0$.

A2. The covariates matrix $\boldsymbol{X}$ (for high-dimensional regression) or the noise matrix $\boldsymbol{Z}$ (for low-rank estimation) have entries $x_{ij} \overset{\text{iid}}{\sim} \mathsf{N}(0, 1/n)$, $z_{ij} \overset{\text{iid}}{\sim} \mathsf{N}(0, 1/n)$.

Also, we denote by $\mathscr{P}_q(\mathbb{R}^k)$ the set of probability distributions with finite q-th moment on $\mathbb{R}^k$ and $\mathscr{P}_{\mathrm{c}}(\mathbb{R}^k)$ those with compact support. We say a function $f : \mathbb{R}^k \to \mathbb{R}$ is *pseudo-Lipschitz of order 2* if there exists constant C such that $|f(\boldsymbol{x}) - f(\boldsymbol{x}')| \leq C(1 + \|\boldsymbol{x}\| + \|\boldsymbol{x}'\|)\|\boldsymbol{x} - \boldsymbol{x}'\|$ for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^k$. We call a function $\ell : (\mathbb{R}^k)^2 \to \mathbb{R}$ a *quadratically-bounded loss* if it is non-negative and pseudo-Lipschitz of order 2 and there exists $C > 0$ such that for all $\boldsymbol{x}, \boldsymbol{x}', \boldsymbol{d} \in \mathbb{R}^k$ we have $|\ell(\boldsymbol{x}, \boldsymbol{d}) - \ell(\boldsymbol{x}', \boldsymbol{d})| \leq C(1 + \sqrt{\ell(\boldsymbol{x}, \boldsymbol{d})} + \sqrt{\ell(\boldsymbol{x}', \boldsymbol{d})})\|\boldsymbol{x} - \boldsymbol{x}'\|$.

3.3.1 High-dimensional regression

We make the following additional assumptions for the regression problem:

R1. We sample $\{(w_i, u_i)\}_{i \leq n} \overset{\text{iid}}{\sim} \mu_{W,U}$, $\{(\theta_i, v_i)\}_{i \leq p} \overset{\text{iid}}{\sim} \mu_{\Theta,V}$ for $\mu_{\Theta,V}, \mu_{W,U} \in \mathscr{P}_2(\mathbb{R}^2)$.

R2. There exists a measurable function $h : \mathbb{R}^2 \to \mathbb{R}$ such that $y_i = h(\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\theta}, w_i)$. Moreover, there exists constant C such that $|h(x, w)| \leq C(1 + |x| + |w|)$ for all x, w.

Notice that the description in terms of a probability kernel $\mathbb{P}(y_i \in \cdot\, | \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\theta})$ is equivalent to the one in terms of a 'noisy' function $y_i = h(\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\theta}, w_i)$ in most cases of interest.

Our lower bound is defined in terms of a one-dimensional recursion. Let $(\Theta, V) \sim \mu_{\Theta,V}$. Let $\mathsf{mmse}_{\Theta,V}(\tau^2)$ be the minimum mean square error for estimation of Θ given observations V and $\Theta + \tau G$ where $G \sim \mathsf{N}(0, 1)$ independent of Θ. Set $\tau_\Theta^2 = \mathbb{E}[\Theta^2]$ and $\tau_0^2 = \infty$, and define recursively

$$\begin{gathered} \tilde{\tau}_s^2 = \frac{1}{\delta}\,\mathsf{mmse}_{\Theta,V}(\tau_s^2), \qquad \sigma_s^2 = \frac{1}{\delta}(\tau_\Theta^2 - \mathsf{mmse}_{\Theta,V}(\tau_s^2))\,, \\ \frac{1}{\tau_{s+1}^2} = \frac{1}{\tilde{\tau}_s^2}\mathbb{E}\left[\mathbb{E}[G_1|Y, G_0, U]^2\right], \end{gathered} \tag{3.5}$$

where $Y = h(\sigma_s G_0 + \tilde{\tau}_s G_1, W)$ and the expectation is with respect to $G_0, G_1 \overset{\text{iid}}{\sim} \mathsf{N}(0, 1)$ and $(W, U) \sim \mu_{W,U}$ independent.

Theorem 3.3.1. *Under assumptions A1, A2, R1, R2 in the high-dimensional regression model and under the asymptotics $n, p \to \infty$, $n/p \to \delta \in (0, \infty)$, let $\hat{\boldsymbol{\theta}}^t$ be output of any GFOM after t iterations ($2t - 1$*

matrix-vector multiplications). Then

$$\lim_{n\to\infty} \frac{1}{p}\|\hat{\boldsymbol{\theta}}^t - \boldsymbol{\theta}\|_2^2 \geq \mathsf{mmse}_{\Theta,V}(\tau_t^2)\,.$$

More generally, for any quadratically-bounded loss $\ell : \mathbb{R}^2 \to \mathbb{R}_{\geq 0}$,

$$\lim_{n\to\infty} \frac{1}{p}\sum_{j=1}^{p} \ell(\theta_j, \hat{\theta}_j^t) \geq \inf_{\hat{\theta}(\,\cdot\,)} \mathbb{E}\big\{\ell(\Theta, \hat{\theta}(\Theta + \tau_t G, V))\big\}\,, \tag{3.6}$$

where $(\Theta, V) \sim \mu_{\Theta,V}$ independent of $G \sim \mathsf{N}(0,1)$, and the infimum on the right-hand side is over measurable functions $\hat{\theta} : \mathbb{R}^2 \to \mathbb{R}$. The limits are in probability and to a constant, and they are guaranteed to exist. For all $\epsilon > 0$, there exist GFOMs which satisfy these bounds to within tolerance ϵ.

3.3.2 Low-rank matrix estimation

We make the following additional assumption:

M1. We sample $\{(\boldsymbol{\Lambda}_i, \boldsymbol{u}_i)\}_{i\leq n} \overset{\text{iid}}{\sim} \mu_{\boldsymbol{\Lambda},\boldsymbol{U}}$ and $\{(\boldsymbol{\theta}_j, \boldsymbol{v}_j)\}_{j\leq p} \overset{\text{iid}}{\sim} \mu_{\boldsymbol{\Theta},\boldsymbol{V}}$ for $\mu_{\boldsymbol{\Lambda},\boldsymbol{U}}, \mu_{\boldsymbol{\Theta},\boldsymbol{V}} \in \mathscr{P}_2(\mathbb{R}^{2r})$.

Again, our lower bound is defined in terms of recursion, which this time is defined over positive semidefinite matrices $\boldsymbol{Q}_t, \hat{\boldsymbol{Q}}_t \in \mathbb{R}^{r\times r}$, $\boldsymbol{Q}_t, \hat{\boldsymbol{Q}}_t \succeq \boldsymbol{0}$. Set $\hat{\boldsymbol{Q}}_0 = \boldsymbol{0}$, and define recursively

$$\boldsymbol{Q}_{t+1} = \boldsymbol{V}_{\boldsymbol{\Lambda},\boldsymbol{U}}(\hat{\boldsymbol{Q}}_t)\,, \qquad \hat{\boldsymbol{Q}}_t = \frac{1}{\delta}\boldsymbol{V}_{\boldsymbol{\Theta},\boldsymbol{V}}(\boldsymbol{Q}_t)\,, \tag{3.7}$$

where we define the second moment of the conditional expectation $\boldsymbol{V}_{\boldsymbol{\Theta},\boldsymbol{V}} : \mathbb{R}^{r\times r} \to \mathbb{R}^{r\times r}$ by

$$\boldsymbol{V}_{\boldsymbol{\Theta},\boldsymbol{V}}(\boldsymbol{Q}) := \mathbb{E}\Big\{\mathbb{E}[\boldsymbol{\Theta}|\boldsymbol{Q}^{1/2}\boldsymbol{\Theta} + \boldsymbol{G} = \boldsymbol{Y}; \boldsymbol{V}]\mathbb{E}[\boldsymbol{\Theta}|\boldsymbol{Q}^{1/2}\boldsymbol{\Theta} + \boldsymbol{G} = \boldsymbol{Y}; \boldsymbol{V}]^{\mathsf{T}}\Big\},$$

and analogously for $\boldsymbol{V}_{\boldsymbol{\Lambda},\boldsymbol{U}}(\hat{\boldsymbol{Q}})$. Here the expectation is with respect to $(\boldsymbol{\Theta}, \boldsymbol{V}) \sim \mu_{\boldsymbol{\Theta},\boldsymbol{V}}$ and an independent Gaussian vector $\boldsymbol{G} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_r)$. Notice in particular that $\mathbb{E}\{\boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathsf{T}}\} - \boldsymbol{V}_{\boldsymbol{\Theta},\boldsymbol{V}}(\boldsymbol{Q})$ is the vector minimum mean square error when $\boldsymbol{\Theta}$ is observed in Gaussian noise with covariance $\boldsymbol{Q}^{-1}$. For $r = 1$, Eq. (3.7) is a simple scalar recursion.

Theorem 3.3.2. *Under assumptions A1, A2, M1 in the low-rank matrix estimation model and under the under the asymptotics $n, p \to \infty$, $n/p \to \delta \in (0, \infty)$, let $\hat{\boldsymbol{\theta}}^t$ be output of any GFOM after t iterations ($2t - 1$ matrix-vector multiplications). Then*

$$\lim_{n\to\infty} \frac{1}{p}\|\hat{\boldsymbol{\theta}}^t - \boldsymbol{\theta}\|_{\mathsf{F}}^2 \geq \mathbb{E}\{\|\boldsymbol{\Theta}\|^2\} - \operatorname{Tr}\boldsymbol{V}_{\boldsymbol{\Theta},\boldsymbol{V}}(\boldsymbol{Q}_t)\,.$$

More generally, for any quadratically-bounded loss $\ell : \mathbb{R}^{2r} \to \mathbb{R}_{\geq 0}$,

$$\lim_{n\to\infty} \frac{1}{p}\sum_{j=1}^{p} \ell(\boldsymbol{\theta}_j, \hat{\boldsymbol{\theta}}_j^t) \geq \inf_{\hat{\boldsymbol{\theta}}(\,\cdot\,)} \mathbb{E}\big\{\ell(\boldsymbol{\Theta}, \hat{\boldsymbol{\theta}}(\boldsymbol{Q}_t^{1/2}\boldsymbol{\Theta} + \boldsymbol{G}, \boldsymbol{V}))\big\}\,, \tag{3.8}$$

where the infimum on the right-hand side is over functions $\hat{\boldsymbol{\theta}} : \mathbb{R}^r \to \mathbb{R}^r$. The limits are in probability and to a constant, and they are guaranteed to exist. As above, for all $\epsilon > 0$ there exist GFOMs which satisfy these *bounds to within tolerance ϵ.*

3.3.3 Discussion

Our motivations are similar to the ones for statistical query (SQ) lower bounds [89, 90]: we want to provide estimation lower bounds under a restricted computational model, that are sensitive to the data distribution. However the scope of our approach is significantly different from SQ algorithms: the latter can query data distributions and compute approximate expectations with respect to that distribution. In contrast, our algorithms work with a fixed sample (the data matrix $\boldsymbol{X}$ and responses $\boldsymbol{y}$), which is queried multiple times. These queries can be thought as weighted averages of *both rows and columns* of $\boldsymbol{X}$ and, as such, cannot be simulated by the SQ oracle. For instance, the proximal gradient method or the nonlinear power iteration of Section 3.2 cannot be framed as a SQ algorithms.

The lower bounds of Theorems 3.3.1 and 3.3.2 are satisfied with equality by a specific first order method that is an approximate message passing (AMP) algorithm, with Bayes updates. This can be regarded as a version of belief propagation (BP) for densely connected graphs [120], or an iterative implementation of the TAP equations from spin glass theory [140].

Our proof builds on the asymptotically exact analysis of AMP algorithms developed in [37, 24, 105, 32]. However we need to overcome three technical obstacles: (1) Show that any GFOM can be reduced (in a suitable sense) to a certain AMP algorithms, whose behavior can be exactly tracked. (2) Show that Bayes-AMP is optimal among all AMP algorithms. We achieve this goal by considering an estimation problem on trees and showing that, in a suitable large degree limit, it has the same asymptotic behavior as AMP on the complete graph. On trees it is immediate to see that BP is the optimal local algorithm. (3) We need to prove that the asymptotic behavior of BP for trees of large degree is equivalent to the one of Bayes-AMP on the original problem. This amounts to proving a Gaussian approximation theorem for BP. While similar results were obtained in the past for discrete models [177, 158], the current setting is technically more challenging because the underlying variables θ_i are continuous and unbounded.

While the line of argument above is –in hindsight– very natural, the conclusion is broadly useful. For instance, [8] study a class of of message passing algorithms inspired to replica symmetry breaking and survey propagation [141], and observe that they do not perform better than Bayes AMP. These algorithms are within the scope of our Theorem 3.3.2, which implies that indeed they cannot outperform Bayes AMP, for any constant number of iterations.

Finally, a sequence of recent papers characterize the asymptotics of the Bayes-optimal estimation error in the two models described above [125, 20]. It was conjectured that, in this context, no polynomial-time algorithm can outperform Bayes AMP, provided these algorithms have access to an arbitrarily small amount of side information.[2] Theorems 3.3.1 and 3.3.2 establish this result within the restricted class of GFOMs.

3.4 Applying the general lower bounds

In our two examples, we will refer to the sets $B_0^p(k) \subset \mathbb{R}^p$ of k-sparse vectors and $B_2^p(R) \subset \mathbb{R}^p$ of vectors with ℓ_2-norm bounded by R.

[2] Concretely, side information can take the form $\boldsymbol{v} = \eta\boldsymbol{\theta} + \boldsymbol{g}$ for $\eta > 0$ arbitrarily small, $\boldsymbol{g} \sim \mathsf{N}(0, \boldsymbol{I}_p)$

Example #1: Sparse phase retrieval

For the reader's convenience, we follow the standard normalization in phase retrieval, whereby the 'sensing vectors' (i.e. the rows of the design matrix) have norm concentrated around one. In other words, we observe $y_i \sim p(\,\cdot\,|\tilde{\boldsymbol{x}}_i^{\mathsf{T}}\boldsymbol{\theta})\mathrm{d}y$, where $\tilde{\boldsymbol{x}}_i \sim \mathsf{N}(0, \boldsymbol{I}_p/p)$.

In order to model the phase retrieval problem, we assume that the conditional density $p(\,\cdot\,|\,\cdot\,)$ satisfies the symmetry condition $p(y|x) = p(y|-x)$. In words: we only observe a noisy version of the absolute value $|\langle\tilde{\boldsymbol{x}}_i, \boldsymbol{\theta}\rangle|$. An important role is played by the following critical value of the number of observations per dimension

$$\delta_{\rm sp} := \left(\int_{\mathbb{R}} \frac{\mathbb{E}_G[p(y|G)(G^2-1)]}{\mathbb{E}_G[p(y|G)]}\,\mathrm{d}y\right)^{-1}. \tag{3.9}$$

Here expectation is with respect to $G \sim \mathsf{N}(0,1)$. It was proved in [146] that, if $\|\boldsymbol{\theta}\|_2 = \sqrt{p}$ and $n > (\delta_{\rm sp} + \eta)p$, for some η bounded away from zero, then there exists a simple spectral estimator $\hat{\boldsymbol{\theta}}_{\rm sp}$ that achieves weak recovery, i.e., a positive correlation with the true signal. Namely, $\frac{|\langle\hat{\boldsymbol{\theta}}_{\rm sp}, \boldsymbol{\theta}\rangle|}{\|\hat{\boldsymbol{\theta}}_{\rm sp}\|_2\|\boldsymbol{\theta}\|_2}$ is bounded away from zero as $p, n \to \infty$.

In the case of a dense signal $\boldsymbol{\theta}$ and observation model $y_i = |\tilde{\boldsymbol{x}}_i^{\mathsf{T}}\boldsymbol{\theta}| + w_i$, $w_i \sim \mathsf{N}(0, \sigma^2)$, the oversampling ratio $\delta_{\rm sp}$ is known to be information-theoretically optimal: for $n < (\delta_{\rm sp} - \eta)p$ no estimator can achieve a correlation that is bounded away from 0 [146]. On the other hand, if $\boldsymbol{\theta}$ has at most $p\varepsilon$ nonzero entries, it is information-theoretically possible to reconstruct it from $\delta > C\varepsilon\log(1/\varepsilon)$ phaseless measurements per dimension [130].

Our next result implies that no GFOM can achieve reconstruction from $O(\varepsilon\log(1/\varepsilon))$ measurements per dimension, unless it is initialized close enough to the true signal. In order to model the additional information provided by the initialization we assume to be given

$$\overline{\boldsymbol{v}} = \sqrt{\alpha}\,\boldsymbol{\theta}/\|\boldsymbol{\theta}\|_2 + \sqrt{1-\alpha}\tilde{\boldsymbol{g}}, \qquad (\tilde{g}_i)_{i\le p} \overset{\text{iid}}{\sim} \mathsf{N}(0, 1/p),. \tag{3.10}$$

Notice that with this normalization $\|\overline{\boldsymbol{v}}\|_2$ concentrates tightly around 1, and $\sqrt{\alpha}$ can be interpreted as the cosine of the angle between $\boldsymbol{\theta}$ and $\overline{\boldsymbol{v}}$.

Corollary 3.4.1. *Consider the phase retrieval model, for a sequence of deterministic signals $\boldsymbol{\theta} \in \mathbb{R}^p$, and let $\mathscr{T}(\varepsilon, R) := B_0^p(p\varepsilon) \cap B_2^p(R)$. Assume the noise kernel $p(\,\cdot\,|x)$ to satisfy the conditions of Theorem 3.3.1 and to be be twice differentiable with respect to x.*

Then, for any $\delta < \delta_{\rm sp}$, there exists $\alpha_* = \alpha_*(\delta, \varepsilon) > 0$ and $C_* = C_*(\delta, \varepsilon)$ such that, if $\alpha \le \alpha_*$, then

$$\sup_{t\ge 0} \lim_{n,p\to\infty} \inf_{\boldsymbol{\theta}\in\mathscr{T}(\varepsilon,\sqrt{p})} \mathbb{E}\frac{\langle\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^t\rangle}{\|\boldsymbol{\theta}\|_2\|\hat{\boldsymbol{\theta}}^t\|_2} \le C_*\sqrt{\alpha}\,. \tag{3.11}$$

The same conclusion holds if $\boldsymbol{\theta}$ is drawn randomly with i.i.d. entries $\theta_i \sim \mu_\theta := (1-\varepsilon)\delta_0 + (\varepsilon/2)(\delta_\mu + \delta_{-\mu})$, $\mu = 1/\sqrt{\varepsilon}$.

Example #2: Sparse PCA

For ease of interpretation, we assume the observation model $\tilde{\boldsymbol{X}} = \boldsymbol{\Lambda}\overline{\boldsymbol{\theta}}^{\mathsf{T}} + \tilde{\boldsymbol{Z}}$, where $(\tilde{z}_{ij})_{i\le n, j\le p} \sim \mathsf{N}(0,1)$ and $(\lambda_i)_{i\le n} \sim \mathsf{N}(0,1)$. Equivalently, conditional on $\overline{\boldsymbol{\theta}}$, the rows of $\tilde{\boldsymbol{X}}$ are i.i.d. samples $\tilde{\boldsymbol{x}}_i \sim \mathsf{N}(0, \boldsymbol{\Sigma})$,

$\boldsymbol{\Sigma} = \boldsymbol{I}_p + \overline{\boldsymbol{\theta}}\overline{\boldsymbol{\theta}}^{\mathsf{T}}$. We also assume to have access to an initialization $\overline{\boldsymbol{v}}$ correlated with $\overline{\boldsymbol{\theta}}$, as per Eq. (3.10). In order to apply Theorem 3.3.2, we choose a specific distribution for the spike. Defining $\boldsymbol{\theta} = \overline{\boldsymbol{\theta}}\sqrt{p}$, we assume that the entries of $\boldsymbol{\theta}$ follow a three-points sparse distribution $(\theta_i)_{i\le p} \sim \mu_\theta := (1-\varepsilon)\delta_0 + (\varepsilon/2)(\delta_{+\mu} + \delta_{-\mu})$. The next lemma specializes Theorem 3.3.2.

Lemma 3.4.1. *Assume the sparse PCA model with the distribution of $\overline{\boldsymbol{\theta}}$ given above. Define $(q_t)_{t\ge 0}$ by*

$$q_{t+1} = \frac{V_\pm(q_t + \tilde{\alpha})}{1 + V_\pm(q_t + \tilde{\alpha})}\,, \qquad q_0 = 0\,, \tag{3.12}$$

$$V_\pm(q) := e^{-\delta q \mu^2}\mu^2\varepsilon^2\mathbb{E}\left\{\frac{\sinh(\mu\sqrt{\delta q}G)^2}{1-\varepsilon+\varepsilon e^{-\delta q\mu^2/2}\cosh(\mu\sqrt{\delta q}G)}\right\}\,, \tag{3.13}$$

where $\tilde{\alpha} = \alpha/(\mu^2\varepsilon(1-\alpha))$. Then, for any GFOM

$$\lim_{n,p\to\infty}\frac{\langle\overline{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}^t\rangle}{\|\overline{\boldsymbol{\theta}}\|_2\|\hat{\boldsymbol{\theta}}^t\|_2} \le \sqrt{\frac{V_\pm(q_t+\tilde{\alpha})}{\mu^2\varepsilon}}\,. \tag{3.14}$$

The bound in the last lemma holds for random vectors $\overline{\boldsymbol{\theta}}$ with i.i.d. entries from the three-points distribution. As a consequence, it implies a minimax bound for non-random vectors $\overline{\boldsymbol{\theta}}$ with given ℓ_2-norm and sparsity. We state this bound in the corollary below. In order to develop explicit expressions, we analyze the recursion of Eqs. (3.12), (3.13).

Corollary 3.4.2. *Assume the sparse PCA model, for $\overline{\boldsymbol{\theta}} \in \mathbb{R}^p$ a deterministic vector and $\boldsymbol{\Lambda}$, $\tilde{\boldsymbol{Z}}$ random, and consider the parameter space $\mathscr{T}(\varepsilon, R) := B_0^p(p\varepsilon)\cap B_2^p(R)$.*

(a) If $R^2 < 1/\sqrt{\delta}$, then there exists $\alpha_* = \alpha_*(R,\delta,\varepsilon), C_* = C_*(R,\delta,\varepsilon)$ such that, for $\alpha < \alpha_*$, and any GFOM

$$\sup_{t\ge 0}\lim_{n,p\to\infty}\inf_{\overline{\boldsymbol{\theta}}\in\mathscr{T}(\varepsilon,R)}\mathbb{E}\frac{\langle\overline{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}^t\rangle}{\|\overline{\boldsymbol{\theta}}\|_2\|\hat{\boldsymbol{\theta}}^t\|_2} \le C_*\sqrt{\alpha}\,. \tag{3.15}$$

(b) If $R^2 < \sqrt{(1-\varepsilon)/4\delta}$, then the above statement holds with $\alpha_* = \left(\frac{\varepsilon}{4\delta}\wedge\frac{1}{2}\right)$, $C_* = 3/R^2$.

In words, the last corollary implies that for $R^2\delta < 1$, no estimator achieves a non-vanishing correlation with the true signal $\overline{\boldsymbol{\theta}}$, unless sufficient side information about $\overline{\boldsymbol{\theta}}$ is available. Notice that for $R^2\delta = 1$ is the threshold above which the principal eigenvector of the empirical covariance $\tilde{\boldsymbol{X}}^{\mathsf{T}}\tilde{\boldsymbol{X}}/n$ becomes correlated with $\overline{\boldsymbol{\theta}}$. Hence, our result implies that, simple PCA fails, then every GFOM will fail.

Viceversa, if simple PCA succeed, then it can be implemented via a GFOM, provided arbitrarily weak side information if available. Indeed, assume side information $\boldsymbol{v} = \eta\boldsymbol{\theta} + \boldsymbol{g}$, with $\boldsymbol{g} \sim \mathsf{N}(0, \boldsymbol{I}_p)$, and an η arbitrarily small constant. Then the power method initialized at $\boldsymbol{v}$ converges to an estimate that has correlation with $\boldsymbol{\theta}$ bounded away from zero in $O(\log(1/\eta))$ iterations.

3.5 Proof of main results

In this section, we prove Theorems 3.3.1 and 3.3.2 under stronger assumptions than in their statements. In the high-dimensional regression model, these assumptions are as follows.

R3. Given $\mu_{\Theta,V} \in \mathscr{P}_{\rm c}(\mathbb{R}^2)$ and $\mu_{\boldsymbol{W},U} \in \mathscr{P}_4(\mathbb{R}^k \times \mathbb{R})$ for some $k \geq 1$, we sample $\{(\theta_i, v_i)\}_{i\leq p} \stackrel{\rm iid}{\sim} \mu_{\Theta,V}$, $\{(\boldsymbol{w}_i, u_i)\}_{i\leq n} \stackrel{\rm iid}{\sim} \mu_{\boldsymbol{W},U}$.

R4. There exists Lipschitz function $h : \mathbb{R} \times \mathbb{R}^k \to \mathbb{R}$ such that $y_i = h(\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\theta}, \boldsymbol{w}_i)$. Measure $\mu_{\boldsymbol{W},U}$ has regular conditional probability distribution $\mu_{\boldsymbol{W}|U}(u, \cdot)$ such that, for all fixed x, u, the distribution of $h(x, \boldsymbol{W})$ when $\boldsymbol{W} \sim \mu_{\boldsymbol{W}|u}(u, \cdot)$ has positive and bounded density $p(y|x, u)$ with respect Lebesgue measure. Further, $\partial_x^k \log p(y|x, u)$ for $1 \leq k \leq 5$ exists and is bounded.

In the low-rank matrix estimation model, this assumption is as follows.

M2. Given $\mu_{\boldsymbol{\Lambda},\boldsymbol{U}}, \mu_{\boldsymbol{\Theta},\boldsymbol{V}} \in \mathscr{P}_{\rm c}(\mathbb{R}^{2r})$, we sample $\{(\boldsymbol{\Lambda}_i, \boldsymbol{u}_i)\}_{i\leq n} \stackrel{\rm iid}{\sim} \mu_{\boldsymbol{\Lambda},\boldsymbol{U}}$, $\{(\boldsymbol{\theta}_j, \boldsymbol{v}_j)\}_{j\leq p} \stackrel{\rm iid}{\sim} \mu_{\boldsymbol{\Theta},\boldsymbol{V}}$.

In Appendix B.5, we show that Theorem 3.3.1 (resp. Theorem 3.3.2) under assumptions R3 and R4 (resp. M2) implies the theorem under the weaker assumptions R1 and R2 (resp. M1).

3.5.1 Reduction of GFOMs to approximate message passing algorithms

Approximate message passing (AMP) algorithms are a special class of GFOMs that admit an asymptotic characterization called *state evolution* [24]. We show that, in both models we consider, any GFOM is equivalent to an AMP algorithm after a change of variables.

An AMP algorithm is defined by sequences of Lipschitz functions $(f_t : \mathbb{R}^{r(t+1)+1} \to \mathbb{R}^r)_{t\geq 0}$, $(g_t : \mathbb{R}^{r(t+1)} \to \mathbb{R}^r)_{t\geq 1}$. It generates sequences $(\boldsymbol{a}^t)_{t\geq 1}$, $(\boldsymbol{b}^t)_{t\geq 1}$ of matrices in $\mathbb{R}^{p\times r}$ and $\mathbb{R}^{n\times r}$, respectively, according to

$$\begin{aligned}\boldsymbol{a}^{t+1} &= \boldsymbol{X}^\mathsf{T} f_t(\boldsymbol{b}^1, \dots, \boldsymbol{b}^t; \boldsymbol{y}, \boldsymbol{u}) - \sum_{s=1}^{t} g_s(\boldsymbol{a}^1, \dots, \boldsymbol{a}^s; \boldsymbol{v})\boldsymbol{\xi}_{t,s}^\mathsf{T}, \\ \boldsymbol{b}^t &= \boldsymbol{X} g_t(\boldsymbol{a}^1, \dots, \boldsymbol{a}^t; \boldsymbol{v}) - \sum_{s=0}^{t-1} f_s(\boldsymbol{b}^1, \dots, \boldsymbol{b}^s; \boldsymbol{y}, \boldsymbol{u})\boldsymbol{\zeta}_{t,s}^\mathsf{T},\end{aligned} \tag{3.16}$$

with initialization $\boldsymbol{a}^1 = \boldsymbol{X}^\mathsf{T} f_0(\boldsymbol{y}, \boldsymbol{u})$. Here $(\boldsymbol{\xi}_{t,s})_{1\leq s\leq t}$, $(\boldsymbol{\zeta}_{t,s})_{0\leq s<t}$ are deterministic $r \times r$ matrices. The we refer to the recursion (3.16) as to an AMP algorithm if only if the matrices $(\boldsymbol{\xi}_{t,s})_{1\leq s\leq t}$, $(\boldsymbol{\zeta}_{t,s})_{0\leq s<t}$ are determined by the functions $(f_t)_{t\geq 0}$, $(g_t)_{t\geq 1}$ in a specific way, which depends on the model under consideration, and we describe in Appendix B.2. For this special choice of the matrices $(\boldsymbol{\xi}_{t,s})_{1\leq s\leq t}$, $(\boldsymbol{\zeta}_{t,s})_{0\leq s<t}$, the iterates $\boldsymbol{a}^t, \boldsymbol{b}^t$ are asymptotically Gaussian, with a covariance that can be determined via the state evolution recursion.

The next lemma, proved in Appendix B.2, makes this precise and describes the state evolution of the resulting AMP algorithm.

Lemma 3.5.1. *Under assumptions A1, A2, R3, R4 (for high-dimensional regression) or assumptions A1, A2, M2 (for low-rank matrix estimation), there exist Lipschitz functions $(f_t)_{t\geq 0}, (g_t)_{t\geq 1}$ as above and $(\varphi_t : \mathbb{R}^{r(t+1)} \to \mathbb{R})_{t\geq 1}, (\phi_t : \mathbb{R}^{r(t+1)+1} \to \mathbb{R})_{t\geq 1}$, such that the following holds. Let $(\boldsymbol{\xi}_{t,s})_{1\leq s\leq t}, (\boldsymbol{\zeta}_{t,s})_{0\leq s<t}$ be $r \times r$ matrices determined by the general AMP prescription (see Appendix B.2), and define $\{\boldsymbol{a}^s, \boldsymbol{b}^s\}_{s\geq 0}$ via the AMP algorithm (3.16). Then we have*

$$\boldsymbol{v}^t = \varphi_t(\boldsymbol{a}^1, \dots, \boldsymbol{a}^t; \boldsymbol{v}), \quad t \geq 1,$$

$$\boldsymbol{u}^t = \phi_t(\boldsymbol{b}^1, \ldots, \boldsymbol{b}^t; \boldsymbol{y}, \boldsymbol{u}), \quad t \geq 1.$$

Further, state evolution determines two collections of of $r \times r$ matrices $(\boldsymbol{T}_{s,t})_{s,t\geq 1}, (\boldsymbol{\alpha}_t)_{t\geq 1}$ such that for all pseudo-Lipschitz functions $\psi : \mathbb{R}^{r(t+2)} \to \mathbb{R}$ of order 2,

$$\frac{1}{p}\sum_{j=1}^{p} \psi(\boldsymbol{a}_j^1, \ldots, \boldsymbol{a}_j^t, \boldsymbol{v}_j, \boldsymbol{\theta}_j) \xrightarrow{\mathrm{p}} \mathbb{E}[\psi(\boldsymbol{\alpha}_1\boldsymbol{\Theta} + \boldsymbol{Z}^1, \ldots, \boldsymbol{\alpha}_t\boldsymbol{\Theta} + \boldsymbol{Z}^t, \boldsymbol{V}, \boldsymbol{\Theta})], \tag{3.17}$$

where $(\boldsymbol{\Theta}, \boldsymbol{V}) \sim \mu_{\boldsymbol{\Theta},\boldsymbol{V}}$ independent of $(\boldsymbol{Z}^1, \ldots, \boldsymbol{Z}^t) \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{T}_{[1:t]})$. Here $\boldsymbol{T}_{[1:t]} \in \mathbb{R}^{tr\times tr}$ is a positive semi-definite block matrix with block (s, s') given by $\boldsymbol{T}_{s,s'}$.[3]

Lemma 3.5.1 implies that the estimator $\hat{\boldsymbol{\theta}}^t$ in Theorem 3.3.1 and 3.3.2 can alternatively be viewed as a Lipschitz function $g_* : \mathbb{R}^{r(t+1)} \to \mathbb{R}^r$ of the AMP iterates $(\boldsymbol{a}^s)_{s\leq t}$ and side information $\boldsymbol{v}$, applied row-wise. Thus, $\ell(\boldsymbol{\theta}_j, \hat{\boldsymbol{\theta}}_j^t)$ can be viewed as a pseudo-Lipschitz function of order 2 applied to $(\boldsymbol{a}_j^s)_{s\leq t}, \boldsymbol{v}_j, \boldsymbol{\theta}_j$; namely, $\ell(\boldsymbol{\theta}_j, g_*((\boldsymbol{a}_j^s)_{s\leq t}, \boldsymbol{v}_j))$. Then, Lemma 3.5.1 implies that the limits in Theorems 3.3.1 and 3.3.2 exist and have lower bound

$$\inf R_\ell(g_*, (\boldsymbol{\alpha}_s), (\boldsymbol{T}_{s,s'})) := \inf \mathbb{E}[\ell(\boldsymbol{\Theta}, g_*(\boldsymbol{\alpha}_1\boldsymbol{\Theta} + \boldsymbol{Z}^1, \ldots, \boldsymbol{\alpha}_t\boldsymbol{\Theta} + \boldsymbol{Z}^t, \boldsymbol{V}))], \tag{3.18}$$

where the infimum is taken over Lipschitz functions g_* and matrices $(\boldsymbol{\alpha}_s), (\boldsymbol{T}_{s,s'})$ generated by the state evolution of *some* AMP algorithm. This lower bound is characterized in the following sections.

3.5.2 Models and message passing on the computation tree

We introduce two statistical models on trees and a collection of algorithms which correspond, in a sense we make precise, to the high-dimensional regression and low-rank matrix estimation models, and AMP algorithms. We derive lower bounds on the estimation error in these models using information-theoretic, rather than algorithmic, techniques. We then transfer these to lower bounds on (3.18). The models are defined using an infinite connected tree $\mathcal{T} = (\mathcal{V}, \mathcal{F}, \mathcal{E})$ consisting of infinite collections of variable nodes $\mathcal{V}$, factor nodes $\mathcal{F}$, and edges $\mathcal{E}$. Factor nodes have degree p and have only variables nodes as neighbors, and variable nodes have degree n and have only factor nodes as neighbors. These properties define the tree uniquely up to isomorphism. We denote the set of neighbors of a variable v by ∂v, and similarly define ∂f. We call $\mathcal{T}$ the *computation tree.*

The statistical models are joint distributions over random variables associated to the nodes and edges of the computation tree.

High-dimensional regression on the computation tree. The random variables $\{(\theta_v, v_v)\}_{v\in\mathcal{V}} \overset{\text{iid}}{\sim} \mu_{\Theta,V}$, $\{(\boldsymbol{w}_f, u_f)\}_{f\in\mathcal{F}} \overset{\text{iid}}{\sim} \mu_{\boldsymbol{W},U}$, and $\{x_{fv}\}_{(f,v)\in\mathcal{E}} \overset{\text{iid}}{\sim} \mathsf{N}(0, 1/n)$ are generated independently. We assume $\mu_{\Theta,V}$, $\mu_{\boldsymbol{W},U}$ are as in assumption R3. We define $y_f = h(\sum_{v\in\partial f} x_{fv}\theta_v, \boldsymbol{w}_f)$ for h as in assumption R4. For each $v \in \mathcal{V}$, our objective is to estimate the coefficient θ_v from data $(y_f, u_f)_{f\in\mathcal{F}}$, $(v_v)_{v\in\mathcal{V}}$, and $(x_{fv})_{(f,v)\in\mathcal{E}}$.

Low-rank matrix estimation on the computation tree. The random variables $\{(\boldsymbol{\theta}_v, \boldsymbol{v}_v)\}_{v\in\mathcal{V}} \overset{\text{iid}}{\sim} \mu_{\boldsymbol{\Theta},\boldsymbol{V}}$, $\{(\boldsymbol{\Lambda}_f, \boldsymbol{u}_f)\}_{f\in\mathcal{F}}$, and $\{z_{fv}\}_{(f,v)\in\mathcal{E}} \overset{\text{iid}}{\sim} \mathsf{N}(0, 1/n)$ are generated independently. We assume $\mu_{\boldsymbol{\Lambda},\boldsymbol{U}}$, $\mu_{\boldsymbol{\Theta},\boldsymbol{V}}$ are as in assumption M2. For each $v \in \mathcal{V}$, our objective is to estimate $\boldsymbol{\theta}_v$ from data $(x_{fv})_{(f,v)\in\mathcal{E}}$, $(\boldsymbol{v}_v)_{v\in\mathcal{V}}$, and $(\boldsymbol{u}_f)_{f\in\mathcal{F}}$.

[3] We emphasize that the construction of all relevant functions and matrices depend on the model. We describe these constructions and prove Lemma 3.5.1 in Appendix B.2.

When ambiguity will result, we will refer to the models of Section 3.3 as high-dimensional regression and low-rank matrix estimation *on the graph*.[4] As on the graph, we introduce dummy variables $(y_f)_{f\in\mathcal{F}}$ in the low-rank matrix estimation problem on the computation tree.

To estimate $\boldsymbol{\theta}_v$, we introduce the class of *message passing algorithms*. A message passing algorithm is defined by sequences of Lipschitz functions $(f_t : \mathbb{R}^{r(t+1)+1} \to \mathbb{R}^r)_{t\geq 0}$, $(g_t : \mathbb{R}^{r(t+1)} \to \mathbb{R}^r)_{t\geq 1}$. For each edge $(f, v) \in \mathcal{E}$, it generates sequences $(\boldsymbol{a}^t_{v\to f})_{t\geq 1}$, $(\boldsymbol{q}^t_{v\to f})_{t\geq 1}$, $(\boldsymbol{b}^t_{f\to v})_{t\geq 1}$, and $(\boldsymbol{r}^t_{f\to v})_{t\geq 0}$ of vectors in $\mathbb{R}^r$, called *messages*, according to

$$\begin{aligned} \boldsymbol{a}^{t+1}_{v\to f} &= \sum_{f'\in\partial v\setminus f} x_{f'v}\boldsymbol{r}^t_{f'\to v}, \qquad & \boldsymbol{r}^t_{f\to v} &= f_t(\boldsymbol{b}^1_{f\to v},\ldots,\boldsymbol{b}^t_{f\to v}; y_f, \boldsymbol{u}_f),\\ \boldsymbol{b}^t_{f\to v} &= \sum_{v'\in\partial f\setminus v} x_{fv'}\boldsymbol{q}^t_{v'\to f}, \qquad & \boldsymbol{q}^t_{v\to f} &= g_t(\boldsymbol{a}^1_{v\to f},\ldots,\boldsymbol{a}^t_{v\to f}; \boldsymbol{v}_v), \end{aligned} \tag{3.19}$$

with initialization $\boldsymbol{r}^0_{f\to v} = f_0(y_f, \boldsymbol{u}_f)$ and $\boldsymbol{a}^1_{v\to f} = \sum_{f'\in\partial v\setminus f} x_{f'v}\boldsymbol{r}^0_{f'\to v}$. We also define for every variable and factor node the vectors

$$\boldsymbol{a}^{t+1}_v = \sum_{f\in\partial v} x_{fv}\boldsymbol{r}^t_{f\to v}, \qquad \boldsymbol{b}^t_f = \sum_{v\in\partial f} x_{fv}\boldsymbol{q}^t_{v\to f}. \tag{3.20}$$

These are called *beliefs*. The vector $\boldsymbol{\theta}_v$ is estimated after t iterations by $\hat{\boldsymbol{\theta}}^t_v = g_*(\boldsymbol{a}^1_v,\ldots,\boldsymbol{a}^t_v;\boldsymbol{v}_v)$.

Message passing algorithms on the computation tree correspond to AMP algorithms on the graph in the sense that their iterates are asymptotically characterized by the same state evolution.

Lemma 3.5.2. *In both the high-dimensional regression and low-rank matrix estimation problems on the tree, the following is true. For any Lipschitz functions $(f_t)_{t\geq 0}$, $(g_t)_{t\geq 1}$, there exist collections of $r\times r$ matrices $(\boldsymbol{T}_{s,t})_{s,t\geq 1}, (\boldsymbol{\alpha}_t)_{t\geq 1}$ such that for any node v chosen independently of the randomness on the model, fixed $t \geq 1$, and under the asymptotics $n, p \to \infty$, $n/p \to \delta \in (0,\infty)$, the message passing algorithm (3.19) generates beliefs at v satisfying*

$$(\boldsymbol{a}^1_v,\ldots,\boldsymbol{a}^t_v,\boldsymbol{v}_v,\boldsymbol{\theta}_v) \xrightarrow{\mathrm{W}} (\boldsymbol{\alpha}_1\boldsymbol{\Theta}+\boldsymbol{Z}^1,\ldots,\boldsymbol{\alpha}_t\boldsymbol{\Theta}+\boldsymbol{Z}^t,\boldsymbol{V},\boldsymbol{\Theta}),$$

where $(\boldsymbol{\Theta},\boldsymbol{V}) \sim \mu_{\boldsymbol{\Theta},\boldsymbol{V}}$ independent of $(\boldsymbol{Z}^1,\ldots,\boldsymbol{Z}^t) \sim \mathsf{N}(\boldsymbol{0},\boldsymbol{T}_{[1:t]})$, and $\xrightarrow{\mathrm{W}}$ denotes convergence in the Wasserstein metric of order 2 (see Appendix B.1). Moreover, the matrices $(\boldsymbol{T}_{s,t})_{s,t\geq 1}, (\boldsymbol{\alpha}_t)_{t\geq 1}$ agree with those in Lemma 3.5.1 when the functions $(f_t)_{t\geq 0}$, $(g_t)_{t\geq 1}$ also agree.

We prove Lemma 3.5.2 in Appendix B.3. Lemma 3.5.2 and the properties of convergence in the Wasserstein metric of order 2 (see Lemma B.1.2, Appendix B.1) imply that for any message passing estimator $\hat{\boldsymbol{\theta}}^t_v$ and loss ℓ, the risk $\mathbb{E}[\ell(\boldsymbol{\theta}_v,\hat{\boldsymbol{\theta}}^t_v)] = \mathbb{E}[\ell(\boldsymbol{\theta}_v, g_*(\boldsymbol{a}^1_v,\ldots,\boldsymbol{a}^t_v;\boldsymbol{v}_v)]$ converges to $R_\ell(g_*,(\boldsymbol{\alpha}_s),(\boldsymbol{T}_{s,s'}))$, in agreement with the asymptotic error of the corresponding AMP estimator on the graph.

On the computation tree, we may lower bound this limiting risk by information-theoretic techniques, as we now explain. By induction, the estimate $\hat{\boldsymbol{\theta}}^t_v$ is a function only of observations corresponding to edges and nodes in the ball of radius $2t-1$ centered at v on the computation tree. We denote the observations in this local neighborhood by $\mathcal{T}_{v,2t-1}$. We lower bound the risk of $\hat{\boldsymbol{\theta}}^t_v$ by the optimal risk of any measurable

[4] This terminology is motivated by viewing the models of Section 3.3 as equivalent to the tree-based models except that they are defined with respect to a finite complete bipartite graph between factor and variable nodes.

estimator, possibly intractable, which depends only on $\mathcal{T}_{v,2t-1}$; we call this the *local Bayes risk*. The following lemma characterizes the local Bayes risk.

Lemma 3.5.3. *Consider a quadratically-bounded loss $\ell : \mathbb{R}^{2r} \to \mathbb{R}_{\geq 0}$. In the high-dimensional regression (resp. low-rank matrix estimation) model on the computation tree and under the asymptotics $n, p \to \infty$, $n/p \to \delta \in (0, \infty)$,*

$$\liminf_{n\to\infty} \inf_{\hat{\boldsymbol{\theta}}(\cdot)} \mathbb{E}[\ell(\boldsymbol{\theta}_v, \hat{\boldsymbol{\theta}}(\mathcal{T}_{v,2t-1}))] \geq R^*,$$

where the infimum is over all measurable functions of $\mathcal{T}_{v,2t-1}$, and R^* is equal to the right-hand side of Eq. (3.6) (resp. Eq. (3.8)).

We prove Lemma 3.5.3 in Appendix B.4. Combining Lemma 3.5.3 with the preceding discussion, we conclude that $R_\ell(g_*, (\boldsymbol{\alpha}_s), (\boldsymbol{T}_{s,s'})) \geq R_*$ for all Lipschitz functions g_* and matrices $(\boldsymbol{\alpha}_s), (\boldsymbol{T}_{s,s'})$ generated by the state evolution of some message passing or, equivalently, by some AMP algorithm. The bounds (3.6) and (3.8) now follow. Moreover, as we show in Appendix B.6, the bounds (3.6) and (3.8) are achieved by a certain AMP algorithm. The proof is complete.

Chapter 4

A proof for GFOM via orthogonalization

4.1 Summary of contribution

In Chapter 3 we introduced a class of 'generalized first order methods' (GFOM) to perform estimation efficiently. Informally, GFOMs proceed iteratively. At time t, the state of the algorithm is given by order t vectors of dimension n or d (which we can think of as estimates of $\boldsymbol{\theta}$). A new vector is computed by applying a nonlinear function to these vectors (independent of the data) and then multiplying the result by $\boldsymbol{X}$ or $\boldsymbol{X}^{\mathsf{T}}$. This class of algorithm is broad enough to include classical first order methods from optimization theory [162], such as gradient descent, accelerated gradient descent, and mirror descent with respect to a broad class of objective functions (both convex and nonconvex).

Given this setting, a natural question is:

What is the optimal estimation algorithm among all GFOMs?

This question was answered in the last chapter under the assumption that the noise matrix $\boldsymbol{W}$ (in the case of low-rank matrix estimation) or the covariates matrix $\boldsymbol{X}$ (for regression in generalized linear models) has i.i.d. normal entries, and under some regularity assumptions on the algorithm iterations. Namely, in Chapter 3 we proves that in the proportional asymptotics $n, d \to \infty$, $n/d \to \delta \in (0, \infty)$, optimal estimation error is achieved, for any fixed number of iterations t, by the Bayes approximate message passing (AMP) algorithm. Also this algorithm choice is unique up to reparametrizations.

The proof of Chapter 3 was based on three steps:

(I) *Reduction.* Any GFOM can be simulated by a certain AMP algorithm, with the same number of matrix-vector multiplications, plus (eventually) a post-processing step that is independent of data $\boldsymbol{X}$.

(II) *Tree model.* The estimation error achieved by an AMP algorithm after t iterations is asymptotically equivalent to the one achieved by a corresponding message passing algorithm for a certain estimation problem on a tree graphical model T after t-iterations (this algorithm is t-local on the tree).

(III) *Optimality on trees.* Belief propagation is the optimal t-local algorithm for the estimation problem on T. As a consequence, Bayes AMP is the optimal first order method in the original problem (since it achieves the same accuracy as belief propagation in the tree model).

The main objective of this note is to present a simpler proof of the optimality of Bayes AMP that does not take the detour of constructing the equivalent tree model. Namely, steps (II) and (III) are replaced by the following.

(II') *Reduction to orthogonal AMP.* Any AMP algorithm can be simulated by a certain orthogonal AMP algorithm, which, after t iterations, generates t vectors in $\mathbb{R}^d$ or $\mathbb{R}^n$ whose projections orthogonal to $\boldsymbol{\theta}$ are orthonormal. The algorithm output at iteration t is a function of these t vectors, which is independent of data $\boldsymbol{X}$.

(III') *Optimality of Bayes AMP.* The asymptotic estimation error of the orthogonal AMP estimator is characterized via state evolution [24]. By minimizing this error among orthogonal AMP algorithms, we obtain the error of Bayes AMP.

This proof strategy avoids several technicalities that arise because of the tree equivalence steps and the analysis of belief propagation. Also, it is easier to generalize to different settings, and indeed we establish the following generalizations of the result of Chapter 3:

- We treat the case of noise matrices $\boldsymbol{W}$ (for low-rank matrix estimation) or $\boldsymbol{X}$ (for regression) with independent entries, satisfying a bound on the fourth moment. In contrast, the results of [50] were limited to Gaussian matrices.
- In the Gaussian case, we cover the case in which the first order method applies, at each iteration, a general Lipschitz continuous nonlinearity to previous iterates. The only limitation is that this nonlinearity should be independent from the data matrix $\boldsymbol{X}$. In contrast, the results of [50] were limited to separable nonlinearities (i.e. nonlinearities that act row-wise to the previous iterates, see below).

In order to motivate our work, we will begin in Section 4.2 by presenting a numerical experiment. We will carry out this experiment in the context of phase retrieval, since a large number of first order methods have been developed for this problem.

We will next pass to explaining our new optimality results. In order to present the new proof technique in the most transparent fashion, we will devote most of the main text to the simplest possible example, namely estimating a rank-one symmetric matrix from a noisy observation. We will describe the setting and state our results in this context in Section 4.3. We then prove this result in Section 4.4 for the case of separable nonlinearities. Finally section 4.5 presents our results for the case of regression. The appendices presents technical proofs for non-separable nonlinearities and for the regression setting. These follow the same strategy as the proof in the main text with some modifications.

4.2 An experiment: benchmarking algorithms for phase retrieval

As a motivating example, we consider noiseless phase retrieval, in which we take measurements y_i of an unknown signal $\boldsymbol{\theta} \in \mathbb{R}^d$ according to:

$$y_i = \langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle^2, \qquad i \in \{1, \cdots, n\}.$$

We let $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ with the i-th row being $\boldsymbol{x}_i$ and $\boldsymbol{y} \in \mathbb{R}^n$ with the i-th coordinate being y_i. We will consider the simple example of random measurements $\boldsymbol{x}_i \overset{iid}{\sim} \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_d/n)$ and assume the normalization $\|\boldsymbol{\theta}\|^2/d =$

$1 + o_d(1)$. Given $(\boldsymbol{y}, \boldsymbol{X})$, our goal is to recover $\boldsymbol{\theta}$. Since the signal $\boldsymbol{\theta}$ is real, 'sign retrieval' would be a more appropriate name here. We expect that an experiment with complex signal would yield similar results.

Needless to say, first order methods (with spectral initialization or not) were studied in a substantial body of work, see among others [175, 46, 58, 42, 196, 74, 145, 195, 134, 138, 85, 147].

Apart from illustrating the content of our results, this section also demonstrates a practical use of these results to benchmarking algorithms.

4.2.1 Spectral initialization

As is common in the literature, we consider first order methods with a spectral initialization. Since our main objective is to compare various first order methods, we will use a common spectral initialization developed in [145], which is defined as follows.

We define $\boldsymbol{D}_n \in \mathbb{R}^{d\times d}$ as follows:

$$\boldsymbol{D}_n := \sum_{i=1}^{n} \mathcal{T}(y_i)\boldsymbol{x}_i\boldsymbol{x}_i^{\mathsf{T}},$$

where $\mathcal{T} : \mathbb{R} \to \mathbb{R}$ is a preprocessing function given in [145, Eq. (137)]:

$$\mathcal{T}(y) = \frac{y-1}{y+\sqrt{1+\varepsilon}-1}\,. \tag{4.1}$$

Here, $\varepsilon > 0$ can be taken arbitrarily, but in simulations we fix $\varepsilon = 10^{-3}$. We then use the initialization $\boldsymbol{\theta}^0 := \sqrt{d}\boldsymbol{v}_1(\boldsymbol{D}_n)$, where $\boldsymbol{v}_1(\boldsymbol{D}_n)$ denotes the leading eigenvector of $\boldsymbol{D}_n$. Without loss of generality, we assume $\langle\boldsymbol{\theta}^0, \boldsymbol{\theta}\rangle \geq 0$ (the overall sign of $\boldsymbol{\theta}$ cannot be estimated). As shown in [145], this initialization is optimal in the following sense. Consider $n, d \to \infty$, with $n/d \to \delta$. For $\delta > 1 + \varepsilon$, $\boldsymbol{\theta}^0$ achieves a positive correlation with $\boldsymbol{\theta}$, with probability converging to one as $n, d \to \infty$. For $\delta < 1$, no estimator can achieve a positive correlation.

In fact, for any $\delta > 1$, the correlation between $\boldsymbol{\theta}^0$ and $\boldsymbol{\theta}$ converges in probability to a deterministic value that is given as follows. For $\lambda \in (1, \infty)$, we define the functions

$$\phi(\lambda) := \lambda\mathbb{E}\left[\frac{\mathcal{T}(G^2)G^2}{\lambda - \mathcal{T}(G^2)}\right], \qquad \psi(\lambda) := \frac{\lambda}{\delta} + \lambda\mathbb{E}\left[\frac{\mathcal{T}(G^2)}{\lambda - \mathcal{T}(G^2)}\right],$$

where expectation is with respect to $G \sim \mathsf{N}(0, 1)$. We let $\bar{\lambda} = \mathrm{argmin}_{\lambda\geq 1}\,\psi(\lambda)$ and, for $\lambda \in (1, \infty)$, define $\zeta(\lambda) := \psi(\max(\lambda, \bar{\lambda}))$. Denote by λ^* the unique solution of the equation $\zeta(\lambda) = \phi(\lambda)$ on $(1, \infty)$. Finally, let $a \geq 0$ be given by

$$a^2 = \frac{\frac{1}{\delta} - \mathbb{E}\left[\frac{\mathcal{T}(G^2)^2}{(\lambda^* - \mathcal{T}(G^2))^2}\right]}{\frac{1}{\delta} + \mathbb{E}\left[\frac{\mathcal{T}(G^2)^2(G^2-1)}{(\lambda^* - \mathcal{T}(G^2))^2}\right]}.$$

Then, [145, Lemma 2] proves that $|\langle\boldsymbol{\theta}, \boldsymbol{\theta}^0\rangle|/d$ converges to a as $n, d \to \infty$. Further, the approximate joint distribution of these vectors is given by $\boldsymbol{\theta}^0 \approx a\boldsymbol{\theta} + \sqrt{1-a^2}\boldsymbol{g}$, in the sense that, for any Lipschitz function

$\psi : \mathbb{R} \to \mathbb{R}$,

$$\operatorname*{p-lim}_{n,d\to\infty} \frac{1}{d}\sum_{i=1}^{d} \psi\big(\theta_i^0 - s\, a\theta_i\big) = \mathbb{E}\big\{\psi(\sqrt{1-a^2}G)\big\}\,. \tag{4.2}$$

(This follows from the convergence of the correlation $|\langle \boldsymbol{\theta}, \boldsymbol{\theta}^0\rangle|/d$, together with rotational invariance.). Here, p-lim denotes convergence in probability, $\boldsymbol{g} \sim \mathsf{N}(0, \boldsymbol{I}_d)$ and is independent of $\boldsymbol{\theta}$. Finally, [147] shows that initializing AMP at $\boldsymbol{\theta}^0$ is (asymptotically) equivalent to running a first order method from a warm start initialization independent of $\boldsymbol{\theta}^0$, and hence the analysis of the next sections apply to the present case.

4.2.2 First order methods

We will consider three specific GFOMs for phase retrieval. GFOMs will only be introduced formally in Section 4.3 (for low-rank matrix estimation) and Section 4.5 (for regression, including phase retrieval as a special case). For this section, it is sufficient to say that GFOMs operate at each iteration by performing multiplication by $\boldsymbol{X}$ or $\boldsymbol{X}^{\mathsf{T}}$ plus, eventually, applying a suitable nonlinear operation that is independent of $\boldsymbol{X}$.

In the next subsection we will implement the algorithms listed below and compare their estimation error with the minimum error among all GFOMs.

Bayes AMP

Bayes AMP is a special type of AMP algorithm and fits the general framework of [24]. The theory presented in Section 4.5 suggests that it is indeed optimal among all GFOMs. A detailed description and analysis of the Bayes AMP for phase retrieval is carried out in [147]. Since the precise definition is somewhat technical and not needed for the rest of the paper, we omit it here and refer to [147].

Remark 4.2.1. It is worth clarifying that —despite the name— Bayes AMP does not rely on Bayesian assumptions.

More precisely, the definition Bayes AMP requires specifying a nominal distribution $\mu_\Theta^{\rm AMP}$ for the entries of the true signal $\boldsymbol{\theta}$. Here, we are assuming $\boldsymbol{\theta}$ arbitrary (either deterministic or random) and such that $\|\boldsymbol{\theta}\|_2^2/d = 1 + o_d(1)$. By rotational invariance of the distribution of the covariates $\boldsymbol{x}_i$, we can achieve at any such $\boldsymbol{\theta}$ the same error as if $\boldsymbol{\theta}$ was uniformly distributed over the sphere of radius $\|\boldsymbol{\theta}\|_2$. For large d, this is achieved by setting $\mu_\Theta^{\rm AMP}$ the standard normal distribution, which is what we do here.

Gradient descent

If we attempt to minimize the ℓ_2 loss on the training dataset, we can derive the corresponding gradient descent algorithm:

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \frac{4\eta\delta^2}{n}\boldsymbol{X}^{\mathsf{T}}(\boldsymbol{y} - |\boldsymbol{X}\boldsymbol{\theta}^t|^2) \odot (\boldsymbol{X}\boldsymbol{\theta}^t),$$

where $\eta > 0$ is the step size, $|\boldsymbol{X}\boldsymbol{\theta}^t|^2 \in \mathbb{R}^n$ is the vector whose i-th coordinate is $\langle \boldsymbol{x}_i, \boldsymbol{\theta}^t\rangle^2$, and $\odot$ denotes entrywise multiplication.

	Bayes AMP	Gradient descent	Prox-linear	1 step prox-linear	TAF
Wall clock time	1.83×10^{-2}	6.63×10^{-3}	$\mathbf{5.87 \times 10^{1}}$	6.23×10^{-3}	7.43×10^{-3}

Table 4.1: Averaged wall clock time for different algorithms.

Prox-linear algorithm

The prox-linear algorithm was proposed in [74]. The original algorithm sets $L := 2\|\boldsymbol{X}\|_{\rm op}^2$ and proceeds by solving a sequence of sub-problems:

$$\boldsymbol{\theta}^{t+1} = \text{argmin}_{\boldsymbol{\vartheta}\in\mathbb{R}^d} \left\{ \frac{L}{2}\|\boldsymbol{\vartheta} - \boldsymbol{\theta}^t\|_2^2 + \sum_{i=1}^n \left| \langle \boldsymbol{x}_i, \boldsymbol{\theta}^t\rangle^2 + 2\langle \boldsymbol{x}_i, \boldsymbol{\theta}^t\rangle\langle \boldsymbol{x}_i, \boldsymbol{\vartheta} - \boldsymbol{\theta}^t\rangle - y_i \right| \right\}. \tag{4.3}$$

Notice that this is *not* a GFOM, since each iteration requires solving an optimization problem, and does not reduce to a pair of matrix-vector multiplications by $\boldsymbol{X}^{\mathsf{T}}$ and $\boldsymbol{X}$.

In order to obtain a first order algorithm we replace the full optimization of the subproblem by a single gradient step, with stepsize ξ:

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + 2\xi \boldsymbol{X}^{\mathsf{T}}(\boldsymbol{s}^t \odot \boldsymbol{X}\boldsymbol{\theta}^t), \quad s_i^t := \text{sign}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\theta}^t\rangle^2)\,. \tag{4.4}$$

We will carry out simulations both with the prox-linear algorithm and the 1-step prox-linear algorithm. It is however important to keep in mind that the comparison between prox-linear algorithm and GFOMs is unfair to GFOMs because each prox-linear step potentially requires a large number of matrix-vector multiplications.

Truncated amplitude flow (TAF)

Truncated amplitude flow (TAF) was proposed in [196], which claimed superior statistical performances with respect to state of the art. Following [196], we fix parameters $\alpha = 0.6$, $\gamma = 0.7$. For $t \in \mathbb{N}$, we define the set

$$\mathcal{I}_t := \left\{ i \in [n] : |\langle \boldsymbol{x}_i, \boldsymbol{\theta}^t\rangle| \geq (1+\gamma)^{-1}\sqrt{y_i} \right\}.$$

At the $(t+1)$-th iteration, we perform the following update:

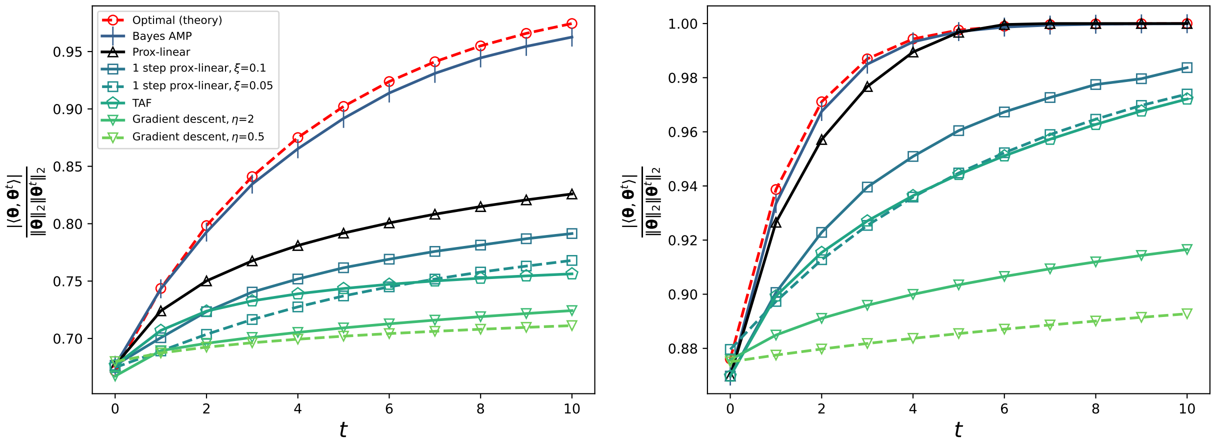
$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \alpha \sum_{i\in\mathcal{I}_t} \left(\langle \boldsymbol{x}_i, \boldsymbol{\theta}^t\rangle - \sqrt{y_i}\,\text{sign}(\langle \boldsymbol{x}_i, \boldsymbol{\theta}^t\rangle) \right) \boldsymbol{x}_i.$$

4.2.3 Simulation results

In our first set of simulations, we take $d = 400$, $n \in \{600, 1000\}$, and run reconstruction experiments using each of the algorithms described above, averaging results over 50 independent trials. We compute the correlation between the estimates produced by these algorithms and the true signal $\boldsymbol{\theta}$, and plot the results in Figure 4.1, as a function of the number of iterations $t \in \{0, 1, \cdots, 10\}$. We also plot the theoretical prediction (cf. Theorem 4.5.1) for the maximum achievable correlation by any GFOM.

A few remarks are in order:

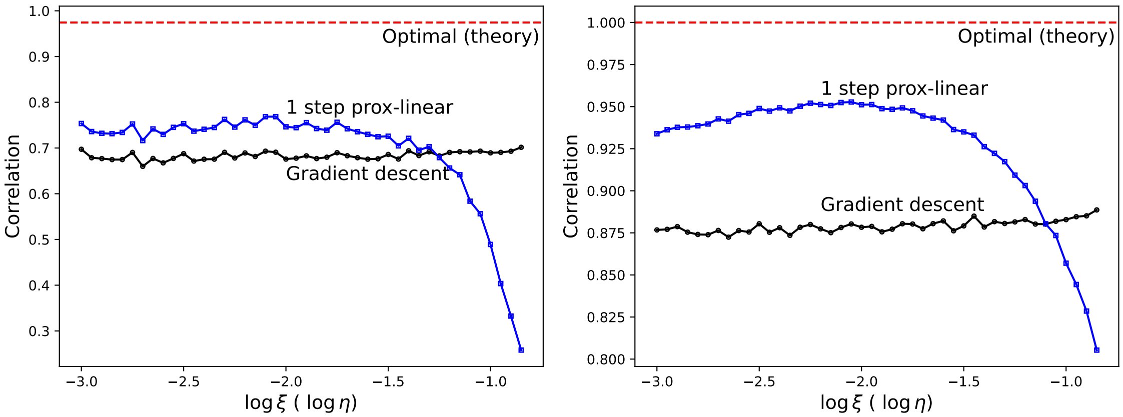
- While the theory developed below applies to $n, d \to \infty$, $n/d \to \delta$, it appears to be fairly accurate



(a) $n = 600$.

(b) $n = 1000$.

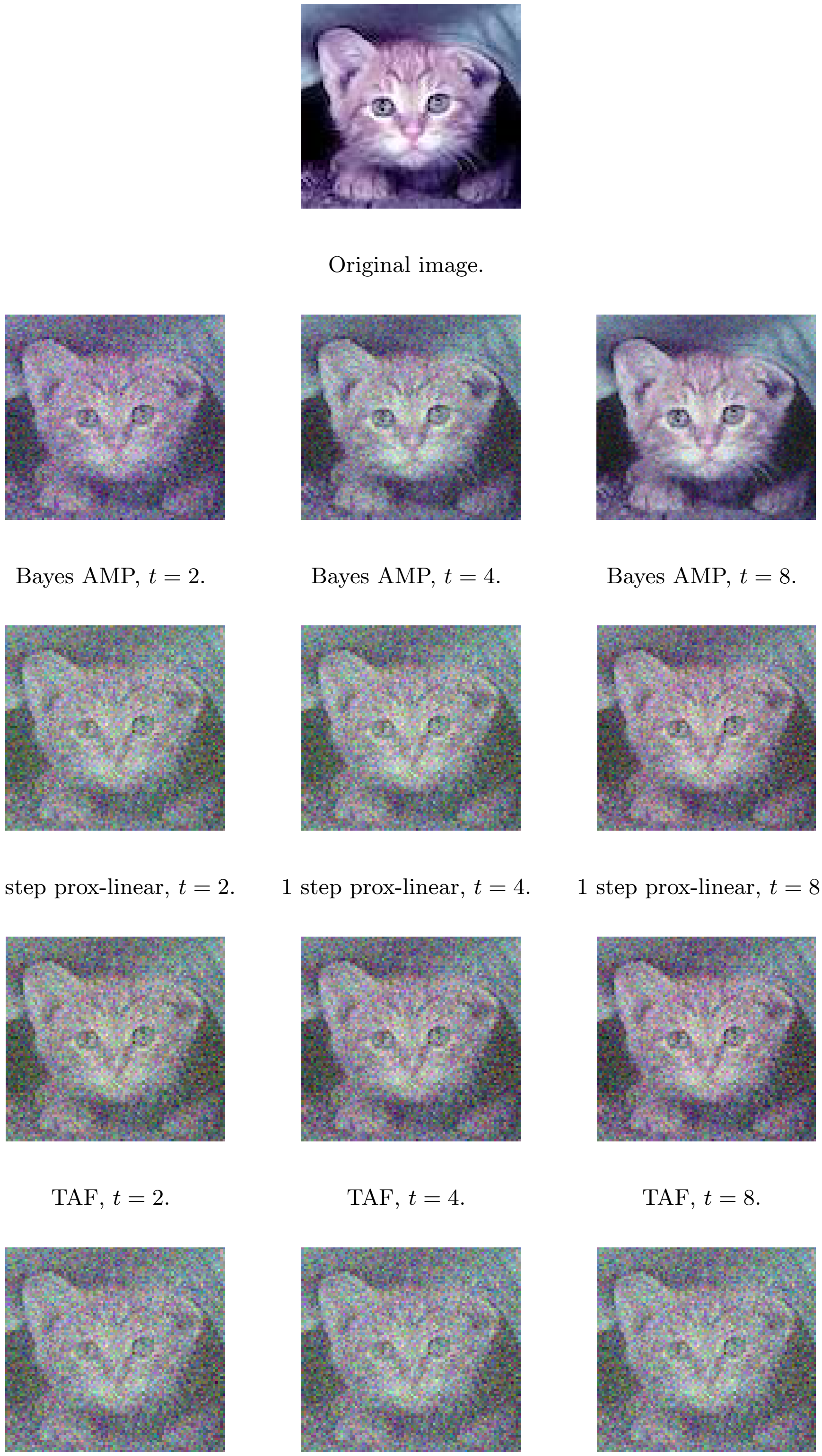
Figure 4.1: Correlation $|\langle \boldsymbol{\theta}^t, \boldsymbol{\theta} \rangle| / \|\boldsymbol{\theta}^t\|_2 \|\boldsymbol{\theta}\|_2$ for various algorithms, as a function of the number of iterations, for $d = 400$. All algorithms are GFOMs with the exception of prox-linear. Red dashed lines represent the optimal correlation of Theorem 4.5.1.



(a) $n = 600$.

(b) $n = 1000$.

Figure 4.2: Performance of gradient descent and the one step prox-linear algorithm with $t = 10$ iterations as a function of the step sizes. The x axis is the logarithm of the step size η (for gradient descent) or ξ (for one step prox-linear algorithm). The y axis is the correlation $|\langle \boldsymbol{\theta}^t, \boldsymbol{\theta} \rangle| / \|\boldsymbol{\theta}^t\|_2 \|\boldsymbol{\theta}\|_2$. Red dashed lines represent the optimal correlation of Theorem 4.5.1. Results are averaged over 50 independent trials.



Original image.

Bayes AMP, $t = 2$.

Bayes AMP, $t = 4$.

Bayes AMP, $t = 8$.

1 step prox-linear, $t = 2$.

1 step prox-linear, $t = 4$.

1 step prox-linear, $t = 8$.

TAF, $t = 2$.

TAF, $t = 4$.

TAF, $t = 8$.

Gradient descent, $t = 2$.

Gradient descent, $t = 4$.

Gradient descent, $t = 8$.

Figure 4.3: Performance comparison between various GFOMs in noiseless phase retrieval (all algorithms use the same spectral initialization).

already at moderate values of n, d. This is not surprising given past results on AMP theory.

- All GFOMs are substantially sub-optimal with the exception of Bayes AMP that appears to achieve the upper bound correlation, as predicted by the theory.
- The prox-linear algorithm (black lines) appears to be nearly optimal for the largest sample size, at $n/d = 2.5$.

  However, as emphasized above, prox-linear algorithm is not a GFOM. In each round of iteration, we use `cvxpy` in Python with the default solver to solve the optimization problem (4.3). In Table 4.1, we report the averaged wall clock time in seconds for the algorithms listed in Section 4.2.2 with 10 iterations. All experiments were conducted on a personal computer with 8GB memory and 2 cores.

The step sizes for gradient descent and one-step prox-linear were chosen in Figure 4.1 via trial and error as to optimize the performance of each algorithm. In Figure 4.2 we plot accuracy as a function of step size parameter for each algorithm, in the same setting as Figure 4.1. Our findings appear to be robust to the choice of this parameter.

In order to further illustrate the difference in performance and the optimality of Bayes AMP, we test the algorithms on a real image in Figure 4.3. The measurement matrix $\boldsymbol{X}$ is random as above. The image contains $d = 7560$ pixels and we used $n = 12000$ (hence $\delta = n/d \approx 1.6$), and we treated each of the 3 color channels separately. The step sizes were chosen for gradient descent and one step prox-linear algorithm as to maximize reconstruction accuracy.

4.3 Symmetric rank-one matrix estimation

We observe a symmetric matrix $\boldsymbol{X} \in \mathbb{R}^{n\times n}$ given by

$$\boldsymbol{X} = \frac{1}{n}\boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}} + \boldsymbol{W}\,, \tag{4.5}$$

where $\boldsymbol{W} = \boldsymbol{W}^{\mathsf{T}}$ is a matrix with independent entries above the diagonal, $(W_{ij})_{1\le i\le j\le n}$ such that $\mathbb{E}\{W_{ij}\} = 0$, $\mathbb{E}\{W_{ij}^2\} = 1/n$ for $1 \le i < j \le n$, and $\mathbb{E}\{W_{ii}^2\} = C/n$ for $1 \le i \le n$. In addition, we observe a vector $\boldsymbol{u} \in \mathbb{R}^n$ that could provide side information about $\boldsymbol{\theta}$. The case in which this side information is not available is covered by setting $\boldsymbol{u} = \boldsymbol{0}$. Given $\mu_{\Theta,U}$, which is a fixed probability distribution over $\mathbb{R}^2$ with finite second moment, we assume $\{(\theta_i, u_i)\}_{i\le n} \overset{iid}{\sim} \mu_{\Theta,U}$. Our objective is to estimate $\boldsymbol{\theta}$ from observations $(\boldsymbol{X}, \boldsymbol{u})$.

4.3.1 General first order methods (GFOM)

A GFOM is an iterative algorithm. At the t-th iteration performs the following update:

$$\begin{aligned} &\boldsymbol{u}^{t+1} = \boldsymbol{X} F_t(\boldsymbol{u}^{\le t}; \boldsymbol{u}) + G_t(\boldsymbol{u}^{\le t}; \boldsymbol{u})\,, \\ &F_t(\boldsymbol{u}^{\le t}; \boldsymbol{u}) := F_t(\boldsymbol{u}^1, \cdots, \boldsymbol{u}^t; \boldsymbol{u})\,, \quad G_t(\boldsymbol{u}^{\le t}; \boldsymbol{u}) := G_t(\boldsymbol{u}^1, \cdots, \boldsymbol{u}^t; \boldsymbol{u})\,. \end{aligned} \tag{4.6}$$

where $F_t, G_t : \mathbb{R}^{n(t+1)} \to \mathbb{R}^n$ are functions indexed by $t \in \mathbb{N}$. After s iterations, the algorithm estimates $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}^s = F_*^{(s)}(\boldsymbol{u}^{\le s}; \boldsymbol{u})$, where $F_*^{(s)} : \mathbb{R}^{n(s+1)} \to \mathbb{R}^n$ is a continuous function. Notice that a GFOM is uniquely determined by the choice of nonlinearities $\{F_t, G_t, F_*^{(t)}\}_{t\in\mathbb{N}}$.

We will consider two specific settings for the functions $\{F_t, G_t, F_*^{(t)}\}_{t\in\mathbb{N}}$, and the noise $\boldsymbol{W}$. The choice of these settings is dictated by the cases in which an asymptotic characterization of the AMP algorithms, known as 'state evolution' [24, 104] has been established rigorously. Namely, for Setting 1 we will leverage the results of [33], while for Setting 2 we will use the results of [23, 55].

Setting 1.

- *The matrix $\boldsymbol{W}$ has entries $(W_{ij})_{i<j} \sim_{iid} \mathsf{N}(0, 1/n)$, and $\mathbb{E}W_{ii}^2 \le C/n$ for a constant C.*
- *The probability measure $\mu_{\Theta,U}$ is sub-Gaussian.*
- *The functions $F_t, G_t, F_*^{(t)} : \mathbb{R}^{n(t+1)} \to \mathbb{R}^n$ are uniformly Lipschitz*[1]*. Further, for any fixed $\boldsymbol{\mu} \in \mathbb{R}^{\mathbb{N}}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{\mathbb{N}\times\mathbb{N}}$ positive semi-definite and $(b_{ij})_{i,j\in\mathbb{N}_{>0}}$, letting $(\boldsymbol{g}_t)_{t\in\mathbb{N}_{>0}}$ be a sequence of centered Gaussian vectors with $\mathbb{E}[\boldsymbol{g}_s(\boldsymbol{g}_t)^{\mathsf{T}}] = \Sigma_{s,t}\boldsymbol{I}_n$, the following limits exist and is finite for all $s \le t$:*

$$\operatorname*{p-lim}_{n\to\infty} \frac{1}{n}\langle F_s(\boldsymbol{y}^1, \dots, \boldsymbol{y}^s; \boldsymbol{u}), F_t(\boldsymbol{y}^1, \dots, \boldsymbol{y}^t; \boldsymbol{u})\rangle ,$$

where p-lim *denotes limit in probability and $\{\boldsymbol{y}^t\}_{t\ge1}$ is defined recursively as follows:*

$$\begin{aligned} \boldsymbol{y}^1 &= \mu_1\boldsymbol{\theta} + \boldsymbol{g}_1 + G_0(\boldsymbol{u}), \\ \boldsymbol{y}^{t+1} &= \mu_{t+1}\boldsymbol{\theta} + \boldsymbol{g}_{t+1} + G_t(\boldsymbol{y}^1, \cdots, \boldsymbol{y}^t; \boldsymbol{u}) + \sum_{s=1}^{t} b_{ts}F_{s-1}(\boldsymbol{y}^1, \cdots, \boldsymbol{y}^{s-1}; \boldsymbol{u}). \end{aligned} \tag{4.7}$$

Since F_s is uniformly Lipschitz and the input random vectors are all sub-Gaussian, one can verify that $\{\|F_s(\boldsymbol{y}^1, \dots, \boldsymbol{y}^s; \boldsymbol{u})\|_2^2/n : n \in \mathbb{N}^+\}$ is uniformly integrable. As a consequence, $\mathbb{E}\langle F_s, F_t\rangle/n$ converges to the same limit. The analogous limits for $\langle F_s, G_t\rangle/n$, $\langle G_s, G_t\rangle/n$, $\langle F_s^*, G_t\rangle/n$, $\langle F_s^*, F_t\rangle/n$ $\langle F_s^*, F_t^*\rangle/n$, $\langle F_t, \boldsymbol{\theta}\rangle/n$, $\langle G_t, \boldsymbol{\theta}\rangle/n$, $\langle F_t^*, \boldsymbol{\theta}\rangle/n$ are also assumed to exist. Similarly, the limits of their expectations also exist.

Setting 2.

- *The matrix $\boldsymbol{W}$ has independent entries on and above the diagonal with $W_{ij} = \overline{W}_{ij}/\sqrt{n}$ where $(\overline{W}_{ij})_{i<j\le n}$ is a collection of i.i.d. random variables with distribution independent of n, such that $\mathbb{E}\overline{W}_{ij} = 0$, $\mathbb{E}\overline{W}_{ij}^2 = 1$, and $\mathbb{E}\overline{W}_{ij}^4 < \infty$. Further, there exists an absolute constant $C > 0$, such that $\mathbb{E}\{W_{ii}^4\} \le C/n^2$ for all $i \le n$.*
- *The probability measure $\mu_{\Theta,U}$ is sub-Gaussian.*
- *Fixed (n-independent) functions $F_t, G_t, F_*^{(t)} : \mathbb{R}^{t+1} \to \mathbb{R}$ are given. We overload this notation by letting $F_t(\boldsymbol{u}^1, \dots, \boldsymbol{u}^t; \boldsymbol{u}) \in \mathbb{R}^n$ be the vector with the i-th component $F_t(\boldsymbol{u}^1, \dots, \boldsymbol{u}^t; \boldsymbol{u})_i = F_t(u_i^1, \dots, u_i^t; u_i)$. Either of the following is assumed:*
  - *(a) The functions F_t, G_t, F_t^* are Lipschitz continuous.*
  - *(b) The functions F_t, G_t, F_t^* are polynomials, and in addition the entries of $\boldsymbol{W}$ are sub-Gaussian $\mathbb{E}\{\exp(\lambda W_{ij})\} \le \exp(C\lambda^2/n)$ for some n-independent constant C.*

[1] We say that sequence of functions $\{f_n : \mathbb{R}^{a_n} \to \mathbb{R}^{b_n}\}_{n\ge1}$ is *uniformly Lipschitz* if there exists n-independent constant $L > 0$, such that for all n and all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^{a_n}$, $\|f_n(\boldsymbol{x}) - f_n(\boldsymbol{y})\|_2/\sqrt{b_n} \le L\|\boldsymbol{x} - \boldsymbol{y}\|_2/\sqrt{a_n}$ and $\|f_n(\boldsymbol{0})\|_2/\sqrt{b_n} \le L$.

4.3.2 Main result for rank-one matrix estimation

In this section we state our optimality result for the case of rank-one matrix estimation. We refer to the appendices for similar statements in the case of generalized linear models.

Let $(\Theta, U) \sim \mu_{\Theta,U}$, $G \sim \mathsf{N}(0,1)$, independent of each other. Define the minimum mean square error function $\mathsf{mmse}_{\Theta,U} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ via

$$\begin{aligned}\mathsf{mmse}_{\Theta,U}(\gamma) &:= \inf_{\hat{\theta}:\mathbb{R}^2\to\mathbb{R}} \mathbb{E}\big\{\big[\Theta - \hat{\theta}(\gamma\Theta + G, U)\big]^2\big\}\\ &= \mathbb{E}[\Theta^2] - \mathbb{E}[\mathbb{E}[\Theta \mid \gamma\Theta + G, U]^2]\,.\end{aligned}$$

Define the sequence $(\gamma_t)_{t\in\mathbb{N}}$ via the following *state evolution* recursion:

$$\gamma_{t+1}^2 = \mathbb{E}[\Theta^2] - \mathsf{mmse}_{\Theta,U}(\gamma_t)\,, \qquad \gamma_0 = 0\,. \tag{4.8}$$

The following theorem establishes that no GFOM can achieve mean square error below $\mathsf{mmse}_{\Theta}(\gamma_t)$ after t iterations.

Theorem 4.3.1. *For $t \in \mathbb{N}_{\geq 0}$, let $\hat{\boldsymbol{\theta}}^t \in \mathbb{R}^n$ be the output of any GFOM after t iterations, under either of Setting 1 or Setting 2. Then the following holds*

$$\operatorname*{p-lim}_{n\to\infty} \frac{1}{n}\|\hat{\boldsymbol{\theta}}^t - \boldsymbol{\theta}\|_2^2 \geq \mathsf{mmse}_{\Theta,U}(\gamma_t)\,. \tag{4.9}$$

Further there exists a GFOM which satisfies the above bound with equality.

In this statement $\text{p-lim}_{n\to\infty}$ denotes limit in probability.

In the next section we will prove Eq. (4.9). We refer to [50] for a proof of the fact this lower bound is achieved. The proof given there implies that the algorithm achieving the lower bound is essentially unique and coincides with Bayes AMP.

Remark 4.3.1. The sequence $(\gamma_t)_{t\geq 0}$ is easily seen to be non-degreasing in t, whence the sequence of lower bounds $\mathsf{mmse}_{\Theta,U}(\gamma_t)$ is non-increasing and converging to $\mathsf{mmse}_{\Theta,U}(\gamma_\infty)$. The latter quantity therefore provides the optimal error achieved by first order methods with $O(1)$ matrix-vector multiplications.

In some cases, $\mathsf{mmse}_{\Theta,U}(\gamma_\infty)$ is conjectured to be the optimal error achieved by polynomial-time algorithms [125, 152]. More precisely, this is expected to be the case if the noise $\boldsymbol{W}$ is Gaussian and $\mathbb{E}[\mathbb{E}[\Theta \mid U]^2] > 0$ (which is the case for instance if $\mathbb{E}[\Theta] \neq 0$). If these conditions are violated, better estimation can be achieved by the following approaches:

- If $\boldsymbol{W}$ has i.i.d. but non-Gaussian entries, applying a nonlinear function entrywise to $\boldsymbol{X}$, and then using a spectral or first order method can improve estimation, see [150] and references therein.

- If $\mathbb{E}[\mathbb{E}[\Theta \mid U]^2] = 0$, then using a spectral initialization improves estimation, see e.g. [152].

Refined versions of the conjecture mentioned above can be formulated in these cases.

4.4 Proof of Theorem 4.3.1

In this section we prove Theorem 4.3.1 under Setting 2. Additionally, we will assume $\boldsymbol{W}$ to have sub-Gaussian entries, namely $\mathbb{E}\{\exp(\lambda W_{ij})\} \le \exp(C\lambda^2/n)$ for all $i, j \le n$ and some n-independent constant C. The proof under Setting 1 is given in Appendix C.1, and the generalization to Setting 2 without sub-Gaussian assumption is carried out in Appendix C.4.

Throughout the proof $(\Theta, U) \sim \mu_{\Theta,U}$ are random variables independent of other random variables unless explicitly stated.

4.4.1 Approximate message passing algorithms

As mentioned above, an important role in the proof is played by approximate message passing (AMP) algorithms. These are GFOMs that enjoy special properties: here we limit ourselves to giving a definition for the problem of symmetric rank-one matrix estimation, in the context of Setting 2.

An AMP algorithm is defined by a sequence of continuous functions $\{f_t : \mathbb{R}^{t+1} \to \mathbb{R}\}_{t\ge 0}$ (also termed the nonlinearities of the AMP algorithm), and produces a sequence of vectors $\{\boldsymbol{a}^t\}_{t\ge 1} \subseteq \mathbb{R}^n$ via the following iteration

$$\boldsymbol{a}^{t+1} = \boldsymbol{X} f_t(\boldsymbol{a}^{\le t}; \boldsymbol{u}) - \sum_{s=1}^{t} b_{t,s} f_{s-1}(\boldsymbol{a}^{\le s-1}; \boldsymbol{u}) \,. \tag{4.10}$$

Here $\boldsymbol{a}^{\le t} = (\boldsymbol{a}^1, \dots, \boldsymbol{a}^t)$ and, as before, nonlinearities are applied entrywise. The term subtracted on the right-hand side is known as Onsager correction term, and we will introduce the notation

$$\mathsf{OC}^t_{\text{AMP}}(\boldsymbol{a}^{\le t-1}; \boldsymbol{u}) := \sum_{s=1}^{t} b_{t,s} f_{s-1}(\boldsymbol{a}^{\le s-1}; \boldsymbol{u}) \tag{4.11}$$

The coefficients $(b_{t,s})_{1\le s\le t}$ are deterministic. Before defining them, we introduce the following state evolution recursion to construct the sequences $\boldsymbol{\mu} = (\mu_t)_{t\ge 1}$, $\boldsymbol{\Sigma} = (\Sigma_{s,t})_{s,t\ge 1}$, where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{\mathsf{T}}$:

$$\begin{aligned} \mu_{t+1} &= \mathbb{E}\big\{\Theta\, f_t(\boldsymbol{\mu}_{\le t}\Theta + \boldsymbol{G}_{\le t}; U)\big\} \,, \\ \Sigma_{s+1,t+1} &= \mathbb{E}\big\{f_s(\boldsymbol{\mu}_{\le s}\Theta + \boldsymbol{G}_{\le s}; U) f_t(\boldsymbol{\mu}_{\le t}\Theta + \boldsymbol{G}_{\le t}; U)\big\} \,, \\ \boldsymbol{G}_{\le t} &:= (G_1, \cdots, G_t) \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\le t}) \,. \end{aligned} \tag{4.12}$$

In the above equations $\boldsymbol{\Sigma}_{\le t} := (\Sigma_{ij})_{i,j\le t}$ and $\boldsymbol{\mu}_{\le t} := (\mu_i)_{i\le t}$, and it is understood that $\boldsymbol{\mu}_{\le s}\Theta + \boldsymbol{G}_{\le s} := (\mu_1\Theta + G_1, \cdots, \mu_t\Theta + G_t)$. Note that f_0 only depends on U and therefore the above recursion does not need any specific initialization. In terms of the above, we define:

$$b_{t,s} = \mathbb{E}\big\{\partial_s f_t(\boldsymbol{\mu}_{\le t}\Theta + \boldsymbol{G}_{\le t}; U)\big\} \,, \tag{4.13}$$

where $\partial_s f_t$ denotes s-th entry of the weak derivative of f.

After t iterations as in Eq. (4.10), AMP estimates $\boldsymbol{\theta}$ by applying a function $F^*_t : \mathbb{R}^{t+1} \to \mathbb{R}$ entrywise:

$$\hat{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{u}) := F^*_t(\boldsymbol{a}^1, \dots, \boldsymbol{a}^t; \boldsymbol{u}) \,. \tag{4.14}$$

For $k, m \in \mathbb{N}_{>0}$, we say a function $\phi : \mathbb{R}^m \to \mathbb{R}$ is *pseudo-Lipschitz of order* k if there exists a constant $L > 0$, such that for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^m$,

$$|\phi(\boldsymbol{x}) - \phi(\boldsymbol{y})| \le L(1 + \|\boldsymbol{x}\|_2^{k-1} + \|\boldsymbol{y}\|_2^{k-1})\|\boldsymbol{x} - \boldsymbol{y}\|_2.$$

Notice that if $f_1, f_2 : \mathbb{R}^m \to \mathbb{R}$ are pseudo-Lipschitz of order k_1 and k_2 respectively, then their product $f_1 f_2$ is pseudo-Lipschitz of order $k_1 + k_2$.

The following theorem characterizes the asymptotics of the AMP iteration (4.10) for Wigner matrices. It was established in [24, 104] for Gaussian matrices, in [23] for Wigner matrices with sub-Gaussian entries and polynomials nonlinearities and in [55] for Wigner matrices with sub-Gaussian entries and Lipschitz nonlinearities. (Some small adaptations are required in the last two cases to get the next statement in its full generality. These are carried out in the appendix.)

Theorem 4.4.1. *Assume the matrix $\boldsymbol{W}$, and nonlinearities f_t satisfy the same assumptions as $\boldsymbol{W}$ and F_t in Setting 2. Then, for any $t \in \mathbb{N}_{>0}$, and any $\psi : \mathbb{R}^{t+2} \to \mathbb{R}$ be a pseudo-Lipschitz function of order 2, the AMP algorithm (4.10) satisfies*

$$\operatorname*{p-lim}_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \psi(\boldsymbol{a}_i^{\le t}, \theta_i, u_i) = \mathbb{E}\{\psi(\boldsymbol{\mu}_{\le t}\Theta + \boldsymbol{G}_{\le t}, \Theta, U)\}\,, \qquad \boldsymbol{G}_{\le t} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\le t})\,. \tag{4.15}$$

(Here p-lim *denotes limit in probability.)*

Remark 4.4.1. Theorem 4.4.1 under Setting 2.(b) is a modified version of [23, Theorem 4], but follows from the latter through a standard argument. More precisely:

- In [23, Theorem 4], the nonlinearity f_t depends only on $\boldsymbol{a}^t$, while here we allow it to depend on all previous iterates and the initialization $(\boldsymbol{a}^{\le t}, \boldsymbol{u})$. However [23, Theorem 4] covers the case in which iterates $\boldsymbol{x}^t$ are matrices $\boldsymbol{x}^t \in \mathbb{R}^{n\times q}$. We can easily reduce the treatment of nonlinearities that depend on all previous times to this one [104, 149]. Fix a time horizon t and choose $q > t$ (independent of n): by suitably choosing the nonlinearities in the algorithm that defines $\boldsymbol{x}^t$, we can ensure that $(\boldsymbol{x}_s^t)_{1\le s\le t}$ coincides with $(\boldsymbol{a}^s)_{1\le s\le t}$.
- In [23, Theorem 4], the matrix $\boldsymbol{X}$ has independent centered entries (up to symmetries). The case of rank-one plus noise matrix $\boldsymbol{X}$ can be reduced to this one as in [65, 64, 152].

4.4.2 Any generalized first order method can be reduced to an AMP algorithm

Following [50], we first show that any GFOM of the form (4.6) can be reduced to an AMP algorithm by a change of variables.

Lemma 4.4.1. *Assume the matrix $\boldsymbol{W}$, the measure $\mu_{\Theta,U}$, and the nonlinearities $(F_s, G_s, F_s^*)_{s\ge0}$ satisfy the assumptions of Setting 2. Then, there exist non-random functions $\{\varphi_s : \mathbb{R}^{s+1} \to \mathbb{R}^s\}_{s\ge1}$ and $\{f_s : \mathbb{R}^{s+1} \to \mathbb{R}\}_{s\ge0}$, satisfying the same assumptions (and independent of $(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{W})$) such that the following holds. Letting $\{\boldsymbol{a}^s\}_{s\ge1}$ be the sequence of vectors produced by the AMP iteration (4.10) with non-linearities $\{f_s\}_{s\ge0}$, we have, for any $t \in \mathbb{N}_{>0}$,*

$$\boldsymbol{u}^{\le t} = \varphi_t(\boldsymbol{a}^{\le t}; \boldsymbol{u}).$$

Proof. The proof is by induction over t. For the base case $t = 1$, we may simply take $f_0(u) = F_0(u)$ and $\varphi_1(\boldsymbol{a}^1; \boldsymbol{u}) := \boldsymbol{a}^1 + G_0(\boldsymbol{u})$.

Suppose the claim holds for the first t iterations. We prove that it holds for iteration $t+1$. By the induction hypothesis,

$$\boldsymbol{u}^{t+1} = \boldsymbol{X} F_t(\varphi_t(\boldsymbol{a}^{\leq t}; \boldsymbol{u}); \boldsymbol{u}) + G_t(\varphi_t(\boldsymbol{a}^{\leq t}; \boldsymbol{u}); \boldsymbol{u}).$$

Let $f_t(x^{\leq t}; u) = F_t(\varphi_t(x^{\leq t}; u); u)$. Since the composition of Lipschitz functions is still Lipschitz, we may conclude that f_t is a Lipschitz function under Setting 2.(a). Analogously, it is a polynomial under Setting 2.(b). Based on the choice of $\{f_s\}_{0\leq s\leq t}$, we compute the coefficients for the Onsager correction term $\{b_{t,j}\}_{1\leq j\leq t}$, as per Eq. (4.13). We then define $\boldsymbol{a}^{t+1}$ via Eq. (4.10), which yields

$$\boldsymbol{a}^{t+1} = \boldsymbol{u}^{t+1} - G_t(\varphi_t(\boldsymbol{a}^{\leq t}; \boldsymbol{u}); \boldsymbol{u}) - \sum_{j=1}^{t} b_{t,j} f_{j-1}(\boldsymbol{a}^{\leq j-1}; \boldsymbol{u})\,.$$

We can therefore define φ_{t+1} via

$$\varphi_{t+1}(\boldsymbol{a}^{\leq t+1}; \boldsymbol{u}) = (\varphi_t(\boldsymbol{a}^{\leq t}; \boldsymbol{u}); \boldsymbol{a}^{t+1} + G_t(\varphi_t(\boldsymbol{a}^{\leq t}; \boldsymbol{u}) + \sum_{j=1}^{t} b_{t,j} f_{j-1}(\boldsymbol{a}^{\leq j-1}; \boldsymbol{u})).$$

(Here note that $\varphi_{t+1}(\boldsymbol{a}^{\leq t+1}; \boldsymbol{u}) \in \mathbb{R}^{n\times(t+1)}$, and $(\boldsymbol{A}; \boldsymbol{B})$ denotes concatenation by columns.)

As above, we see immediately that φ_{t+1} is Lipschitz under Setting 2.(a), and a polynomial under Setting 2.(b). This completes the proof by induction.

□

As an immediate consequence of the last lemma, AMP algorithms achieve the same error as GFOMs, for the same number of iterations, under any loss. (In this statement $\text{p-liminf}_{n\to\infty}$ denotes $\liminf$ in probability. Namely, given a sequence of random variables Z_n, and $z \in \mathbb{R}$, we write $\text{p-liminf}_{n\to\infty} Z_n \geq z$ if, for any $\varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(Z_n \leq z - \varepsilon) = 0$.)

Corollary 4.4.1. *Let $\mathcal{A}^t_{\text{GFOM}}$ be the class of GFOM estimators with t iterations, and $\mathcal{A}^t_{\text{AMP}}$ be the class of AMP algorithms with t iterations (under the assumptions of either Setting 2.(a), or Setting 2.(b)). (In particular $\hat{\boldsymbol{\theta}}(\,\cdot\,) \in \mathcal{A}^t_{\text{GFOM}}$ is defined by a set of n-independent functions $\{F_t, G_t, F_*^{(t)}\}_{t\in\mathbb{N}}$, and similarly for $\hat{\boldsymbol{\theta}}(\,\cdot\,) \in \mathcal{A}^t_{\text{AMP}}$.)*

Then for any loss function $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_{\geq 0}$:

$$\inf_{\hat{\boldsymbol{\theta}}(\,\cdot\,)\in\mathcal{A}^t_{\text{GFOM}}} \operatorname*{p-liminf}_{n\to\infty} \mathcal{L}(\hat{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{u}), \boldsymbol{\theta}) = \inf_{\hat{\boldsymbol{\theta}}(\,\cdot\,)\in\mathcal{A}^t_{\text{AMP}}} \operatorname*{p-liminf}_{n\to\infty} \mathcal{L}(\hat{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{u}), \boldsymbol{\theta})\,. \tag{4.16}$$

Proof. The left-hand side of Eq. (4.16) is smaller or equal than the right-hand side because $\mathcal{A}^t_{\text{AMP}} \subseteq \mathcal{A}^t_{\text{GFOM}}$. To show that they are equal, let $\hat{\boldsymbol{\theta}}(\,\cdot\,) \in \mathcal{A}^t_{\text{GFOM}}$ be any GFOM that achieves the infimum on the left with tolerance ε. By Lemma 4.4.1 we can construct $\hat{\boldsymbol{\theta}}'(\,\cdot\,) \in \mathcal{A}^t_{\text{AMP}}$ achieving the same loss.

□

Remark 4.4.2. Note that throughout this section we are assuming $\{F_t, G_t, F_*^{(t)}\}_{t\in\mathbb{N}}$ to be n-independent. However, standard compactness arguments allows to extend the present treatment to n-dependent nonlinearities as long as the constants implicit in the definitions of Setting 2 (Lipschitz constant, maximum polynomial degree, and so on) are uniformly bounded.

Appendix C.1 will treat the case of nonlinearities that are non-separable and hence necessarily n-dependent.

4.4.3 Any AMP algorithm can be reduced to an orthogonal AMP algorithm

In the previous section we reduced GFOMs to AMP algorithms. We next show that we can in fact limit ourselves to the analysis of a special subset of AMP algorithms, whose iterates are approximately orthogonal, after we subtract their components along $\boldsymbol{\theta}$. We refer to this special subset as orthogonal AMP (OAMP) algorithms.

Lemma 4.4.2. *Let $\{\boldsymbol{a}^t\}_{t\geq 1}$ be a sequence generated by the AMP iteration (4.10), under either of Setting 2.(a) or Setting 2.(b). Then there exist functions $\{\phi_t : \mathbb{R}^{t+1} \to \mathbb{R}^t\}_{t\geq 1}$, $\{g_t : \mathbb{R}^{t+1} \to \mathbb{R}\}_{t\geq 0}$, satisfying the same assumptions (and independent of $(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{W})$) such that the following holds. Let $\{\boldsymbol{v}^t\}_{t\geq 1}$ be the sequence generated by an AMP algorithm with non-linearities $\{g_t\}_{t\geq 0}$ (and same matrix $\boldsymbol{X}$ as for $\{\boldsymbol{a}^t\}_{t\geq 1}$), namely*

$$\boldsymbol{v}^{t+1} = \boldsymbol{X} g_t(\boldsymbol{v}^{\leq t}; \boldsymbol{u}) - \sum_{s=1}^{t} b'_{t,s} g_{s-1}(\boldsymbol{v}^{\leq s-1}; \boldsymbol{u}) \,, \tag{4.17}$$

with deterministic coefficients $(b'_{t,s})$ determined by the analogous of Eq. (4.13), with f_t replaced by g_t. Then we have:

(i) For all $t \geq 1$,

$$\boldsymbol{a}^{\leq t} = \phi_t(\boldsymbol{v}^{\leq t}; \boldsymbol{u}).$$

(ii) For any pseudo-Lipschitz function $\psi : \mathbb{R}^{t+2} \to \mathbb{R}$ of order 2,

$$\operatorname*{p-lim}_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \psi(\boldsymbol{v}_i^{\leq t}, \theta_i, u_i) = \mathbb{E}\{\psi(V_1, \dots, V_t, \Theta, U)\}, \tag{4.18}$$

where $V_i := x_{i-1}(\alpha_i \Theta + Z_i)$, with $(x_0, \dots, x_{t-1}) \in \{0,1\}^t$, $(\alpha_1, \dots, \alpha_t) \in \mathbb{R}^t$, and $\{Z_i\}_{i\in\mathbb{N}_{\geq 1}} \overset{iid}{\sim} \mathsf{N}(0,1)$ standard random variables independent of (Θ, U).

Proof. Throughout this proof, given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we denote by $L^2(\mathbb{P}) = L^2(\Omega, \mathcal{F}, \mathbb{P})$ the space of random variables with finite second moment. Given a closed linear subspace $\mathcal{S} \subseteq L^2(\mathbb{P})$ and a random variable $T \in L^2(\mathbb{P})$, we denote by $\Pi_{\mathcal{S}}(T)$ the projection of T onto $\mathcal{S}$ (i.e. the unique minimizer of $\|S - T\|_{L^2}^2 = \mathbb{E}\{(S-T)^2\}$ over $S \in \mathcal{S}$). We denote by $\Pi_{\mathcal{S}}^{\perp} = I - \Pi_{\mathcal{S}}$ the projector onto its orthogonal complement.

Given $(\mu_t)_{t\geq 1}$, and $(\Sigma_{s,t})_{s,t\geq 1}$ defined via state evolution, see Eq. (4.12), let $\boldsymbol{G}$ be a centered Gaussian process with covariance $\boldsymbol{\Sigma}$, and define the random variables and subspaces

$$Y_t := f_t(\boldsymbol{\mu}_{\leq t}\Theta + \boldsymbol{G}_{\leq t}; U), \qquad \mathcal{S}_t := \mathrm{span}(Y_k : 0 \leq k \leq t) \,.$$

Note that by state evolution $\langle Y_t, Y_s\rangle_{L^2} = \Sigma_{t+1,s+1}$.

By linear algebra, there exist deterministic constants $\{c_{ts}\}_{0\le s\le t}$, $x_t \in \{0,1\}$, such that $c_{tt}\neq 0$, and

$$R_t := c_{tt}\Pi^{\perp}_{\mathcal{S}_{t-1}}(Y_t) = \sum_{s=0}^{t} c_{ts}Y_s, \qquad \mathbb{E}[R_tR_s] = \mathbb{1}_{s=t}x_t,$$

Indeed if Y_t does not belong to $\mathcal{S}_{t-1}$ we can simply take $x_t = 1$ and $c_{tt} = \|\Pi^{\perp}_{\mathcal{S}_{t-1}}(Y_t)\|^{-1}_{L^2}$. Otherwise we take $R_t = 0$, $c_{tt} = 1$, $x_t = 0$.

We prove the lemma by induction. For the base case $t = 1$, we set $g_0(u) = c_{00}f_0(u)$ whence the claim (i) follows trivially. For claim (ii) there are two cases. Either $\mathbb{E}\{f_0(U)^2\} = 0$, whence $x_0 = 0$ and therefore (ii) holds with $V_1 = 0$ almost surely, or $\mathbb{E}\{f_0(U)^2\} > 0$ whence $x_0 = 1$, $c_{00} = \mathbb{E}\{f_0(U)^2\}^{-1/2}$, and therefore the claim follows by state evolution, where

$$\alpha_1 = \frac{\mathbb{E}[\Theta f_0(U)]}{\mathbb{E}[f_0(U)^2]^{1/2}}. \tag{4.19}$$

Suppose the lemma holds for the first t iterations. We prove it also holds for the $(t+1)$-th iteration. Define

$$g_t(\boldsymbol{v}^{\le t}; u) = \sum_{s=0}^{t} c_{ts}f_s(\phi_s(\boldsymbol{v}^{\le s}; u); u). \tag{4.20}$$

Then by the assumptions and the induction hypothesis, g_t is Lipschitz under Setting 2.(a), and is a polynomial under Setting 2.(b). Given the nonlinearities $\{g_t\}_{s\le t}$, we can compute the coefficients $(b'_{s,j})_{1\le j\le s\le t}$. We denote the Onsager term for this new iteration by $\mathsf{OC}^t_{\text{OAMP}}(\boldsymbol{v}^{\le t-1};\boldsymbol{u}) := \sum_{j=1}^{t} b'_{t,j}g_{j-1}(\boldsymbol{v}^{\le j-1};\boldsymbol{u})$. With this notation, Eq. (4.17) can be rewritten as:

$$\boldsymbol{v}^{t+1} = \sum_{s=0}^{t} c_{ts}\boldsymbol{X} f_s(\phi_s(\boldsymbol{v}^{\le s};\boldsymbol{u});\boldsymbol{u}) - \mathsf{OC}^t_{\text{OAMP}}(\boldsymbol{v}^{\le t-1};\boldsymbol{u})\,.$$

Using the AMP iteration that defines $\{\boldsymbol{a}^s\}_{s\ge 1}$, we get:

$$\boldsymbol{v}^{t+1} = \sum_{s=0}^{t} c_{ts}(\boldsymbol{a}^{s+1} + \mathsf{OC}^s_{\text{AMP}}(\boldsymbol{a}^{\le s-1};\boldsymbol{u})) - \mathsf{OC}^t_{\text{OAMP}}(\boldsymbol{v}^{\le t-1};\boldsymbol{u}).$$

Solving for $\boldsymbol{a}^{t+1}$ and expressing $\boldsymbol{a}^{\le t+1} = \phi_t(\boldsymbol{v}^{\le t+1};\boldsymbol{u})$ (recall that c_{tt} is always non-vanishing) we obtain the desired mapping ϕ_{t+1} thus proving claim (i).

In order to prove claim (ii), we distinguish two cases. In the first case $x_t = 0$ and $R_t \overset{a.s.}{=} 0$. Using the state evolution for the orthogonal AMP iteration (4.17) and the definition (4.20) we obtain that claim (ii) folds with $V_{t+1} \overset{a.s.}{=} 0$.

In the second case $x_t = 1$, then again by state evolution we obtain that the claim holds with $V_{t+1} \overset{\mathrm{d}}{=} \alpha_{t+1}\Theta + Z_{t+1}$, where

$$\alpha_{t+1} = \frac{\mathbb{E}[\Theta\,\Pi^{\perp}_{\mathcal{S}_{t-1}}(Y_t)]}{\mathbb{E}[\Pi^{\perp}_{\mathcal{S}_{t-1}}(Y_t)^2]^{1/2}}, \tag{4.21}$$

this completes the proof.

□

Considering the case in which $x_t \neq 0$ for all t (i.e., each new non-linearity is 'non-degenerate'), Eq. (4.18) implies

$$\boldsymbol{v}^t = \alpha_t \boldsymbol{\theta} + \boldsymbol{z}^t \,, \qquad \frac{1}{n}\langle \boldsymbol{z}^t, \boldsymbol{z}^s \rangle = \mathbb{1}_{s=t} + o_n(1)\,, \qquad \frac{1}{n}\langle \boldsymbol{z}^t, \boldsymbol{\theta} \rangle = o_n(1)\,. \tag{4.22}$$

In other words, the iterates are approximately orthonormal along the subspace orthogonal to $\boldsymbol{\theta}$. This justifies the name 'orthogonal AMP' (OAMP).

Remark 4.4.3. In the following we can and will restrict ourselves to the case in which, in the notation of Eq. (4.18), $x_t = 1$ for all t. Indeed if $x_t = 0$ for some t, we can set to zero the corresponding AMP iterate $\boldsymbol{v}_t = 0$ (i.e. set $g_{t-1} = 0$), and the resulting algorithm will asymptotically have the same state evolution. By removing this iteration altogether, we obtain an algorithm with same accuracy and one less iteration.

4.4.4 Optimal orthogonal AMP

By Lemma 4.4.1 and 4.4.2 in order to derive a lower bound of estimation error achieved by GFOMs with t iterations, it is sufficient to restrict ourselves to the class of orthogonal AMP algorithms (it is understood that the latter can be followed by entrywise post processing).

We therefore have the following consequence of the previous results (see also Remark 4.4.3).

Corollary 4.4.2. *Let $\hat{\boldsymbol{\theta}} : (\boldsymbol{X}, \boldsymbol{u}) \mapsto \hat{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{u})$ be a t-iterations GFOM estimator under the assumptions of either Setting 2.(a), or Setting 2.(b). Then for any loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$, pseudo-Lipschitz of order 2, we have*

$$\operatorname*{p-lim}_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \ell(\hat{\theta}_i(\boldsymbol{X}, \boldsymbol{u}), \theta_i) \geq \inf_{(\{g_\ell\}, \varphi) \in \mathcal{A}^t_{\mathrm{OAMP}}} \mathbb{E}\big\{\ell(\varphi(\boldsymbol{\alpha}_{\leq t}\Theta + \boldsymbol{Z}_{\leq t}, U), \Theta)\big\}\,. \tag{4.23}$$

Here the infimum is over all sequences of Lipschitz (Setting 2.(a)) or polynomial (Setting 2.(b)) nonlinearities for an orthogonal AMP algorithm, and over all functions $\varphi : \mathbb{R}^{t+1} \to \mathbb{R}$ with the same properties.

Recall that a sufficient statistics for $\boldsymbol{\Theta}$ given $\boldsymbol{V}_{\leq t} := \boldsymbol{\alpha}_{\leq t}\Theta + \boldsymbol{Z}_{\leq t}$ is $T_0 := \langle \boldsymbol{\alpha}_{\leq t}, \boldsymbol{V}_{\leq t} \rangle / \|\boldsymbol{\alpha}_{\leq t}\|_2$, and T_0 can be rewritten as:

$$T_0 = \|\boldsymbol{\alpha}_{\leq t}\|_2 \Theta + G\,, \qquad G \sim \mathsf{N}(0,1)\,, \quad G \perp \Theta\,. \tag{4.24}$$

Since in addition U is conditionally independent of $\boldsymbol{V}_{\leq t}$ given Θ, the function φ in Eq. (4.23) can be taken to be a function of (U, T_0), and precisely the function that minimizes the risk of estimating Θ with respect to the loss ℓ. The minimization on the right-hand side of Eq. (4.23) reduces to the maximization of $\|\boldsymbol{\alpha}_{\leq t}\|_2$, which is solved by the next lemma.

Lemma 4.4.3. *Recall the definition of $(\gamma_s)_{s\geq 0}$ in Eq. (4.8). Then, for all $t \in \mathbb{N}_{>0}$, and all choices of nonlinearities $g_0, \ldots, g_t$, we have $\|\boldsymbol{\alpha}_{\leq t}\|_2 \leq \gamma_t$.*

Proof. The proof is by induction over t. For the base case $t = 1$, using equation (4.19), we have

$$\alpha_1^2 \le \sup_{f_0} \frac{\mathbb{E}[\Theta f_0(U)]^2}{\mathbb{E}[f_0(U)^2]} = \sup_{f_0} \frac{\mathbb{E}\big\{\mathbb{E}[\Theta|U] f_0(U)\big\}^2}{\mathbb{E}[f_0(U)^2]} \le \mathbb{E}\big\{\mathbb{E}[\Theta \mid U]^2\big\}\,.$$

The last step holds by Cauchy-Schwarz inequality.

We next assume that the claim holds for iteration t, and will prove it also holds for iteration $t+1$. Let $\hat{\Theta}_t := \mathbb{E}[\Theta \mid U, V_1, \cdots, V_t]$. Using equation (4.21), we have

$$\begin{aligned}
\alpha_{t+1}^2 &= \frac{\mathbb{E}\big\{\hat{\Theta}_t\, \Pi^{\perp}_{\mathcal{S}_{t-1}}(Y_t)\big\}^2}{\mathbb{E}[\Pi^{\perp}_{\mathcal{S}_{t-1}}(Y_t)^2]} \\
&\overset{(a)}{\le} \mathbb{E}\{\Pi^{\perp}_{\mathcal{S}_{t-1}}(\hat{\Theta}_t)^2\} \\
&\overset{(b)}{=} \mathbb{E}\{\hat{\Theta}_t^2\} - \mathbb{E}\{\Pi_{\mathcal{S}_{t-1}}(\hat{\Theta}_t)^2\}\,,
\end{aligned}$$

where (a) follows by Cauchy-Schwarz and (b) by Pythagora's theorem. By construction $\{\Pi^{\perp}_{\mathcal{S}_{s-1}}(Y_s)/\mathbb{E}[\Pi^{\perp}_{\mathcal{S}_{s-1}}(Y_s)^2]^{1/2} : 0 \le s \le t-1\}$ is an orthonormal basis for $\mathcal{S}_{t-1}$, whence

$$\begin{aligned}
\alpha_{t+1}^2 &\le \mathbb{E}[\hat{\Theta}_t^2] - \sum_{s=0}^{t-1} \frac{\mathbb{E}[\Theta \Pi^{\perp}_{\mathcal{S}_{s-1}}(Y_s)]^2}{\mathbb{E}[\Pi^{\perp}_{\mathcal{S}_{s-1}}(Y_s)^2]} \\
&= \mathbb{E}[\hat{\Theta}_t^2] - \sum_{s=1}^{t} \alpha_s^2\,,
\end{aligned}$$

Therefore $\|\boldsymbol{\alpha}_{\le t+1}\|_2^2 \le \mathbb{E}[\hat{\Theta}_t^2]$. Further

$$\begin{aligned}
\mathbb{E}[\hat{\Theta}_t^2] &= \mathbb{E}[\mathbb{E}[\Theta \mid U, V_1, \cdots, V_t]^2] \\
&\overset{(a)}{=} \mathbb{E}[\mathbb{E}[\Theta \mid U, \|\boldsymbol{\alpha}_{\le t}\|_2 \Theta + G]] \\
&\overset{(b)}{\le} \mathbb{E}[\mathbb{E}[\Theta \mid U, \gamma_t \Theta + G]^2] \\
&\overset{(c)}{=} \gamma_{t+1}^2,
\end{aligned}$$

where (a) follows because, as pointed above, $T_0 = \langle \boldsymbol{\alpha}_{\le t}, \boldsymbol{V}_{\le t}\rangle / \|\boldsymbol{\alpha}_{\le t}\|_2$ is a sufficient statistics for $\boldsymbol{\Theta}$ given $\boldsymbol{V}_{\le t} = \boldsymbol{\alpha}_{\le t}\Theta + \boldsymbol{Z}_{\le t}$, and is distributed as in Eq. (4.24). Further, (b) follows by Jensen's inequality since, by the induction hypothesis, $\|\boldsymbol{\alpha}_{\le t}\|_2 \le \gamma_t$, and (c) by the definition of γ_{t+1}. This completes the induction.

□

The proof of Theorem 4.3.1 follows immediately from Corollary 4.4.2 and Lemma 4.4.3.

4.5 High-dimensional regression

In this section, we generalize our results to regression in generalized linear models. We observe a vector of responses $\boldsymbol{y} \in \mathbb{R}^n$ and a matrix of covariates $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ which are related according to

$$\boldsymbol{y} = h(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{w}),$$

Here $\boldsymbol{w} \in \mathbb{R}^n$ is a noise vector, $\boldsymbol{\theta} \in \mathbb{R}^d$ is a vector of parameters, and $h : \mathbb{R}^2 \to \mathbb{R}$ is a continuous function which we apply to vectors entrywise. Namely, denoting by $\boldsymbol{x}_i \in \mathbb{R}^d$ the i-th row of $\boldsymbol{X}$, the above equation is equivalent to $y_i = h(\langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle, w_i)$ for $i \leq n$.

We assume that $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ has i.i.d. entries with $\mathbb{E}[X_{ij}] = 0$ and $\mathbb{E}[X_{ij}^2] = 1/n$ for all $1 \leq i \leq n$ and $1 \leq j \leq d$. In addition, we observe side information $\boldsymbol{u} \in \mathbb{R}^n$ and $\boldsymbol{v} \in \mathbb{R}^d$. Given $\mu_{W,U}$ and $\mu_{\Theta,V}$ two fixed probability distributions over $\mathbb{R}^2$, we assume $\{(w_i, u_i)\}_{i\leq n} \overset{iid}{\sim} \mu_{W,U}$ and $\{(\theta_i, v_i)\}_{i\leq d} \overset{iid}{\sim} \mu_{\Theta,V}$. We consider the asymptotic setting where we have fixed asymptotic aspect ratio: $n/d \to \delta \in (0, \infty)$. The goal is to estimate $\boldsymbol{\theta}$ given $(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{u}, \boldsymbol{v})$.

4.5.1 General first order methods

In this section we introduce our notations for GFOMs for generalized linear models. At the t-th iteration, GFOM performs the following updates:

$$\begin{aligned}\boldsymbol{v}^t :=& \boldsymbol{X}^\mathsf{T} F_{t-1}^{(1)}(\boldsymbol{u}^{\leq t-1}; \boldsymbol{y}, \boldsymbol{u}) + F_{t-1}^{(2)}(\boldsymbol{v}^{\leq t-1}; \boldsymbol{v}),\\ \boldsymbol{u}^t :=& \boldsymbol{X} G_t^{(1)}(\boldsymbol{v}^{\leq t}; \boldsymbol{v}) + G_t^{(2)}(\boldsymbol{u}^{\leq t-1}; \boldsymbol{y}, \boldsymbol{u}),\end{aligned} \tag{4.25}$$

where we use the shorthands $F_s^{(\ell)}(\boldsymbol{u}^{\leq s}; \boldsymbol{y}, \boldsymbol{u}) := F_s^{(\ell)}(\boldsymbol{u}^1, \cdots, \boldsymbol{u}^s; \boldsymbol{y}, \boldsymbol{u})$ and $G_s^{(\ell)}(\boldsymbol{v}^{\leq s}; \boldsymbol{v}) := G_s^{(\ell)}(\boldsymbol{v}^1, \cdots, \boldsymbol{v}^s; \boldsymbol{v})$, where $F_t^{(1)}, G_{t+1}^{(2)} : \mathbb{R}^{n(t+2)} \to \mathbb{R}^n$, $F_t^{(2)}, G_t^{(1)} : \mathbb{R}^{d(t+1)} \to \mathbb{R}^d$ are continuous functions with the F's indexed by $t \in \mathbb{N}$ and G's indexed by $t \in \mathbb{N}_{>0}$. After s iterations, the algorithm estimates $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}^s = G_*^{(s)}(\boldsymbol{v}^{\leq s}; \boldsymbol{v})$, where $G_*^{(s)} : \mathbb{R}^{d(s+1)} \to \mathbb{R}^d$ is a continuous function. In this setting, a GFOM is uniquely determined by the set of nonlinearities $\{F_{t-1}^{(1)}, F_{t-1}^{(2)}, G_t^{(1)}, G_t^{(2)}, G_*^{(t)}\}_{t\in\mathbb{N}_{>0}}$.

As in the case of low-rank matrix estimation, we consider two settings for the random matrix $\boldsymbol{X}$, and the nonlinearities $\{F_{t-1}^{(1)}, F_{t-1}^{(2)}, G_t^{(1)}, G_t^{(2)}, G_*^{(t)}\}_{t\in\mathbb{N}_{>0}}$.

Setting 3.

- *The matrix $\boldsymbol{X}$ has entries $X_{ij} \overset{iid}{\sim} \mathsf{N}(0, 1/n)$.*
- *The probability measures $\mu_{\Theta,V}$, $\mu_{W,U}$ are sub-Gaussian.*
- *The functions $F_t^{(1)}, F_t^{(2)}, G_t^{(1)}, G_t^{(2)}, G_*^{(t)}$ are uniformly Lipschitz. Further, for any $\boldsymbol{\mu} \in \mathbb{R}^{\mathbb{N}}$, $\boldsymbol{\Sigma}, \bar{\boldsymbol{\Sigma}} \in \mathbb{R}^{\mathbb{N}\times\mathbb{N}}$ positive semi-definite and $(b_{ij})_{1\leq i,j\leq t}$, $(\bar{b}_{ij})_{1\leq i,j\leq t}$ n-independent constants, we let $(\boldsymbol{g}_t)_{t\in\mathbb{N}_{>0}}$ and $(\bar{\boldsymbol{g}}_t)_{t\in\mathbb{N}}$ be centered Gaussian processes with $\mathbb{E}[\boldsymbol{g}_s \boldsymbol{g}_t^\mathsf{T}] = \Sigma_{st}\boldsymbol{I}_d$ and $\mathbb{E}[\bar{\boldsymbol{g}}_s \bar{\boldsymbol{g}}_t{}^\mathsf{T}] = \bar{\Sigma}_{st}\boldsymbol{I}_n$, we assume the following limits exist for all $s \leq t$,*

$$\begin{aligned}&\operatorname*{p-lim}_{n,d\to\infty} \frac{1}{d}\langle F_t^{(2)}(\boldsymbol{y}^1, \cdots, \boldsymbol{y}^t; \boldsymbol{v}), F_s^{(2)}(\boldsymbol{y}^1, \cdots, \boldsymbol{y}^s; \boldsymbol{v})\rangle,\\ &\operatorname*{p-lim}_{n,d\to\infty} \frac{1}{n}\langle F_t^{(1)}(\bar{\boldsymbol{y}}^1, \cdots, \bar{\boldsymbol{y}}^t; h(\bar{\boldsymbol{g}}_0, \boldsymbol{w}), \boldsymbol{u}), F_s^{(1)}(\bar{\boldsymbol{y}}^1, \cdots, \bar{\boldsymbol{y}}^s; h(\bar{\boldsymbol{g}}_0, \boldsymbol{w}), \boldsymbol{u})\rangle,\end{aligned}$$

where $\{\boldsymbol{y}^t\}_{t\geq 1}$, $\{\bar{\boldsymbol{y}}_t\}_{t\geq 1}$ are defined recursively as follows:

$$\begin{aligned}&\boldsymbol{y}^1 = \mu_1\boldsymbol{\theta} + \boldsymbol{g}_1 + F_0^{(2)}(\boldsymbol{v}),\\ &\boldsymbol{y}^{t+1} = \mu_{t+1}\boldsymbol{\theta} + \boldsymbol{g}_{t+1} + F_t^{(2)}(\boldsymbol{y}^{\leq t}; \boldsymbol{v}) + \sum_{s=1}^{t} b_{ts} G_s^{(1)}(\boldsymbol{y}^{\leq s}; \boldsymbol{v}),\\ &\bar{\boldsymbol{y}}^1 = \bar{\boldsymbol{g}}_1 + G_1^{(2)}(h(\bar{\boldsymbol{g}}_0, \boldsymbol{w}), \boldsymbol{u}) + \bar{b}_{11} F_0^{(1)}(h(\bar{\boldsymbol{g}}_0, \boldsymbol{w}), \boldsymbol{u}),\end{aligned}$$

$$\bar{\boldsymbol{y}}^{t+1} = \bar{\boldsymbol{g}}_{t+1} + G_{t+1}^{(2)}(\bar{\boldsymbol{y}}^1, \cdots, \bar{\boldsymbol{y}}^t; h(\bar{\boldsymbol{g}}_0, \boldsymbol{w}), \boldsymbol{u}) + \sum_{s=1}^{t+1} \bar{b}_{t+1,s} F_{s-1}^{(1)}(\bar{\boldsymbol{y}}^1, \cdots, \bar{\boldsymbol{y}}^{s-1}; h(\bar{\boldsymbol{g}}_0, \boldsymbol{w}), \boldsymbol{u}).$$

The analogous limits for $\langle G_t^{(1)}, G_s^{(1)}\rangle/d$, $\langle G_t^{(1)}, F_s^{(2)}\rangle/d$, $\langle G_*^{(t)}, G_s^{(1)}\rangle/d$, $\langle G_*^{(t)}, F_s^{(2)}\rangle/d$, $\langle G_*^{(t)}, G_*^{(s)}\rangle/d$, $\langle \boldsymbol{\theta}, G_t^{(1)}\rangle/d$, $\langle \boldsymbol{\theta}, F_t^{(2)}\rangle/d$, $\langle \boldsymbol{\theta}, G_*^{(t)}\rangle/d$, $\langle G_t^{(2)}, G_s^{(2)}\rangle/n$, $\langle G_t^{(2)}, F_s^{(1)}\rangle/n$, $\langle F_t^{(1)}, \bar{\boldsymbol{g}}_s\rangle/n$ and $\langle G_t^{(1)}, \boldsymbol{g}_s\rangle/d$ are also assumed to exist.

Setting 4.

- *The matrix $\boldsymbol{X}$ has independent entries with $X_{ij} = \overline{X}_{ij}/\sqrt{n}$ where $(\overline{X}_{ij})_{i\le n, j\le d}$ is a collection of i.i.d. random variables with distribution independent of (n, d), such that $\mathbb{E}\overline{X}_{ij} = 0$, $\mathbb{E}\overline{X}_{ij}^2 = 1$, and $\mathbb{E}\overline{X}_{ij}^4 < \infty$.*
- *The probability measures $\mu_{\Theta,V}$, $\mu_{W,V}$ are sub-Gaussian.*
- *We have n-independent functions $F_{t-1}^{(1)}, F_t^{(2)}, G_t^{(1)}, G_t^{(2)}, G_*^{(t)} : \mathbb{R}^{t+1} \to \mathbb{R}$. We overload these notations by letting $F_t^{(1)}(\boldsymbol{u}^1, \cdots, \boldsymbol{u}^t; \boldsymbol{y}, \boldsymbol{u}) \in \mathbb{R}^n$ be the vector with the i-th component $F_t(\boldsymbol{u}^1, \cdots, \boldsymbol{u}^t; \boldsymbol{y}, \boldsymbol{u})_i = F_t(u_i^1, \cdots, u_i^t; y_i, u_i)$. Similar notations apply for $F_t^{(2)}, G_t^{(1)}, G_t^{(2)}$ and $G_*^{(t)}$. We assume either of the following conditions:*
  - *(a) The functions $F_{t-1}^{(1)}, F_{t-1}^{(2)}, G_t^{(1)}, G_t^{(2)}, G_*^{(t)}$ are Lipschitz continuous.*
  - *(b) The functions $F_{t-1}^{(1)}, F_{t-1}^{(2)}, G_t^{(1)}, G_t^{(2)}, G_*^{(t)}$ are polynomial, and in addition the entries of $\boldsymbol{X}$ are sub-Gaussian $\mathbb{E}[\exp(\lambda X_{ij})] \le \exp(C\lambda^2/n)$ for some n-independent constant C.*

4.5.2 Main result for generalized linear models

Unless explicitly stated, in the rest parts of the proof we let $(\Theta, V) \sim \mu_{\Theta,V}$, $(W, U) \sim \mu_{W,U}$ and $Z, Z_0, Z_1 \overset{iid}{\sim} \mathsf{N}(0,1)$ independent of each other. We define the minimum mean squared error function $\mathsf{mmse}_{\Theta,V}\colon \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ via

$$\begin{aligned}\mathsf{mmse}_{\Theta,V}(\alpha) :=& \inf_{\hat{\Theta}:\mathbb{R}^2\to\mathbb{R}^2} \mathbb{E}\big\{[\Theta - \hat{\Theta}(\alpha\Theta + Z, V)]^2\big\}\\ =&\mathbb{E}[\Theta^2] - \mathbb{E}\big\{\mathbb{E}[\Theta \mid \alpha\Theta + Z, V]^2\big\}.\end{aligned}$$

We let $\beta_0 := 0$, $\sigma_1 := \delta^{-1/2}\mathbb{E}[\Theta^2]^{1/2}$ and $\tilde{\sigma}_1 := 0$. Then for $s \in \mathbb{N}^+$, we define the following quantities recursively:

$$\begin{aligned}&\beta_s^2 = \frac{1}{\sigma_s^2}\mathbb{E}[\mathbb{E}[Z_0 \mid h(\sigma_s Z_0 + \tilde{\sigma}_s Z_1, W), U, Z_1]^2], \qquad \beta_s \ge 0,\\ &\sigma_{s+1}^2 = \frac{1}{\delta}\mathsf{mmse}_{\Theta,V}(\beta_s), \qquad \tilde{\sigma}_{s+1}^2 = \frac{1}{\delta}(\mathbb{E}[\Theta^2] - \mathsf{mmse}_{\Theta,V}(\beta_s)).\end{aligned} \tag{4.26}$$

The following theorem establishes that no GFOM can achieve mean squared error below $\mathsf{mmse}_{\Theta,V}(\beta_t)$ after t iterations.

Theorem 4.5.1. *For $t \in \mathbb{N}_{>0}$, let $\hat{\boldsymbol{\theta}}^t \in \mathbb{R}^d$ be the output of any GFOM after t iterations, then under either Setting 3 or 4, the following holds:*

$$\operatorname*{p-lim}_{n,d\to\infty} \frac{1}{d}\|\hat{\boldsymbol{\theta}}^t - \boldsymbol{\theta}\|_2^2 \ge \mathsf{mmse}_{\Theta,V}(\beta_t). \tag{4.27}$$

Further, there exists a GFOM which satisfies the above bound with equality.

The proof of the lower bound (4.27) is presented in Appendix C.2 under Setting 4 and in Appendix C.3 under Setting 3. We refer to [50] for a proof that there exists a GFOM achieving the bound with equality.

Chapter 5

Sampling from the posterior via diffusion processes

5.1 Introduction

Consider the standard setup of Bayesian inference, whereby the joint distribution of observed data $\boldsymbol{X}$ and unobserved parameter $\boldsymbol{\theta}$ is defined by

$$\boldsymbol{\theta} \sim \pi(\,\cdot\,), \qquad \boldsymbol{X} \sim \mathrm{P}(\,\cdot\,\mid \boldsymbol{\theta}), \tag{5.1}$$

where $\pi(\,\cdot\,)$ is the *prior distribution*. Bayesian techniques draw inferences from the *posterior distribution*[1]

$$\mu_{\boldsymbol{X}}(\mathrm{d}\boldsymbol{\theta}) := \mathbb{P}(\mathrm{d}\boldsymbol{\theta}|\boldsymbol{X}) \;\propto\; \mathrm{P}(\boldsymbol{X}\mid\boldsymbol{\theta})\pi(\mathrm{d}\boldsymbol{\theta})\,. \tag{5.2}$$

A substantial amount of research has been devoted to developing approximation methods [60, 36] and sampling algorithms for the Bayes posterior. Among these, Markov Chain Monte Carlo (MCMC) [95, 92, 185, 121] methods play a special role because of their versatility. However, MCMC often suffers from slow mixing [157, 198], especially when the posterior is multi-modal. In these circumstances, it might be impossible to run the chain long enough to produce a sample with approximately correct distribution, and this can lead to erroneous inference. Rigorous upper bounds for the mixing times have been established under various settings [128, 199, 61, 77, 75], but proving such bounds is very challenging and existing guarantees only cover a small fraction of practical applications.

In this paper we study a different approach to posterior sampling. In a nutshell, we construct a non-homogeneous stochastic process (more precisely a diffusion process). The distribution of the state at time t converges, as $t \to \infty$, to $\mu_{\boldsymbol{X}}(\mathrm{d}\boldsymbol{\theta})$. This diffusion process uses as drift the posterior expectation of $\boldsymbol{\theta}$ given Gaussian observations. Hence, the whole construction can be thought as a way to reduce sampling to regression.

This approach was first proposed within generative modeling in machine learning [180, 100, 181, 182]. The construction of the stochastic process was based on time-reversal. Here we follow a different viewpoint,

[1] In the application of Bayes formula below, we are identifying $\mathrm{P}(\,\cdot\,\mid\boldsymbol{\theta})$ with its density with respect to a reference measure.

initiated in [4] and based on the idea of *stochastic localization*, originally introduced by Eldan as a proof technique [79, 80, 81, 56].

The paper [4] developed this construction to address a problem from statistical physics: sampling from the Sherrington-Kirkpatrick Gibbs measure at a high-temperature. The present paper describes the first rigorous application of the same approach to sampling problems in high-dimensional statistics. We outline a general analysis technique that relies on establishing certain properties for the posterior expectation estimator. We then apply this approach to posterior sampling in symmetric spiked models, and check the validity of the stated conditions.

5.1.1 Construction of the diffusion process

We begin by outlining the construction of the diffusion process via stochastic localization. Stochastic localization [80, 56] defines (random) sequence $\mu_{\boldsymbol{X},t}(\mathrm{d}\boldsymbol{\theta})$ of probability measure indexed by $t \in [0, \infty)$. This sequence has the following properties:

1. The initial condition is the actual posterior $\mu_{\boldsymbol{X},0} = \mu_{\boldsymbol{X}}$.
2. The final point is a point mass at a random $\boldsymbol{\theta}_*$, $\mu_{\boldsymbol{X},\infty} = \delta_{\boldsymbol{\theta}_*}$.
3. The process is a martingale with respect to a filtration $\mathcal{F}_t$. Namely, for any Borel set A, and any $t_1 \leq t_2$, $\mathbb{E}[\mu_{\boldsymbol{X},t_2}(A)|\mathcal{F}_{t_1}] = \mu_{\boldsymbol{X},t_1}(A)$.

As a consequence of the last two points, $\boldsymbol{\theta}_*$ is a sample form the posterior $\mu_{\boldsymbol{X}}$.

In this paper we use one specific construction for such process (throughout, we assume $\mu_{\boldsymbol{X}}$ to have finite second moment.) Let $(\boldsymbol{G}(t))_{t\geq 0}$ be a standard n-dimensional Brownian motion independent of $(\boldsymbol{\theta}, \boldsymbol{X})$ with distribution (5.1). We then define the *observation process* $(\boldsymbol{y}(t))_{t\geq 0}$ by

$$\boldsymbol{y}(t) = t\,\boldsymbol{\theta} + \boldsymbol{G}(t)\,, \tag{5.3}$$

and define $\mu_{\boldsymbol{X},t}(\mathrm{d}\boldsymbol{\theta})$ to be the posterior distribution of $\boldsymbol{\theta}$ given $\boldsymbol{X}$ and $\boldsymbol{y}(t)$:

$$\begin{align} \mu_{\boldsymbol{X},t}(\mathrm{d}\boldsymbol{\theta}) &= \mathbb{P}(\mathrm{d}\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}(t)) \tag{5.4}\\ &= \frac{1}{Z(\boldsymbol{X},t)} \exp\left\{\langle \boldsymbol{y}(t), \boldsymbol{\theta}\rangle - \frac{t}{2}\|\boldsymbol{\theta}\|_2^2\right\} \mu_{\boldsymbol{X}}(\mathrm{d}\boldsymbol{\theta}). \tag{5.5} \end{align}$$

It is easy to check that this construction satisfies the conditions 1—3 introduced above. (For other constructions of statistical interest, see Section 5.4.)

It follows from Eq. (5.5) that it is sufficient to track the process $(\boldsymbol{y}(t))_{t\geq 0}$ in order to determine the measure $\mu_{\boldsymbol{X},t}(\mathrm{d}\boldsymbol{\theta})$. The process $(\boldsymbol{y}(t))_{t\geq 0}$ admits an alternative characterization as the unique solution of the following stochastic differential equation (see, e.g., [4] or [131, Theorem 7.1], for a derivation):

$$\mathrm{d}\boldsymbol{y}(t) = \boldsymbol{m}(\boldsymbol{y}(t), t)\mathrm{d}t + \mathrm{d}\boldsymbol{B}(t), \qquad \boldsymbol{y}(0) = \boldsymbol{0}. \tag{5.6}$$

Here $(\boldsymbol{B}(t))_{t\geq 0}$ is a standard Brownian motion in $\mathbb{R}^n$, and

$$\boldsymbol{m}(\boldsymbol{y}, t) := \mathbb{E}[\boldsymbol{\theta}|\boldsymbol{X}, t\boldsymbol{\theta} + \boldsymbol{G}(t) = \boldsymbol{y}]\,, \tag{5.7}$$

is the posterior expectation function (with $\boldsymbol{G}(t) \sim \mathsf{N}(0, t\boldsymbol{I}_n)$.)

In order to sample from $\mu_{\boldsymbol{X}}$, it is sufficient to closely track process (5.6). To this end we proceed as follows:

1. We discretize the SDE (5.6) in time, using a straightforward Euler scheme.
2. We construct an efficient algorithm that provides an approximation $\hat{\boldsymbol{m}}(\boldsymbol{y}, t)$ of the posterior expectation $\boldsymbol{m}(\boldsymbol{y}, t)$.

We will prove a general bound on a certain distance between the distribution sampled by this sampling method and the target, under two conditions: (1) A bound on the mean square distance between and $\boldsymbol{m}(\boldsymbol{y}, t)$, and $\hat{\boldsymbol{m}}(\boldsymbol{y}, t)$, and (2) Lipschitz continuity of the map $\boldsymbol{y} \mapsto \hat{\boldsymbol{m}}(\boldsymbol{y}, t)$.

We will then focus on spiked matrix models, and introduce an approximate message passing (AMP) algorithm that satisfies the conditions above.

5.1.2 Posterior sampling from spiked models

Most of our technical effort will be devoted to a canonical problem in high-dimensional statistics: making inferences about a low-rank signal that is corrupted by noise. Given signal-to-noise parameter $\beta > 0$, we observe an $n \times n$ symmetric matrix $\boldsymbol{X}$ generated as follows:

$$\boldsymbol{X} = \frac{\beta}{n}\boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}} + \boldsymbol{W}. \tag{5.8}$$

Here $\boldsymbol{W} \sim \mathrm{GOE}(n)$, independent of $\boldsymbol{\theta}$, i.e., $\boldsymbol{W}$ is an $n \times n$ symmetric matrix with independently distributed entries above the diagonal: $\{W_{ii} : i \in [n]\} \overset{iid}{\sim} \mathsf{N}(0, 2/n)$, $\{W_{ij} : 1 \le i < j \le n\} \overset{iid}{\sim} \mathsf{N}(0, 1/n)$. For $\boldsymbol{\theta}$ we use a product prior:

$$(\theta_i)_{i \le n} \overset{iid}{\sim} \pi_\Theta\,. \tag{5.9}$$

(Throughout, we will denote by $\pi_\Theta^{\otimes n}(\mathrm{d}\boldsymbol{\theta}) = \pi_\Theta(\mathrm{d}\theta_1) \cdots \pi_\Theta(\mathrm{d}\theta_n)$ the product distribution over $\mathbb{R}^n$ with marginal π_Θ.)

Model (5.8) is the symmetric version of the *spiked model* first introduced in [107]. In the asymmetric (rectangular) version, data takes the form $\tilde{\boldsymbol{X}} = \boldsymbol{u}\boldsymbol{\theta}^{\mathsf{T}} + \boldsymbol{Z} \in \mathbb{R}^{n \times p}$, where $(u_i)_{i \le n} \overset{iid}{\sim} \pi_U$, $(\theta_i)_{i \le p} \overset{iid}{\sim} \pi_\Theta$, and $\boldsymbol{Z}$ is a noise matrix. While we carry out our analysis in the symmetric setting for simplicity, the generalization to asymmetric matrices is straightforward. We refer to Section 5.4 for a discussion of the general sampling problem.

The symmetric spiked model (5.8) has been used as an idealized setting to understand low-rank matrix estimation. Consider, to be definite, the case $\mathbb{E}\{\Theta\} := \int \theta \, \pi_\Theta(\mathrm{d}\theta) = 0$. Then, it is known that non-trivial estimation of $\boldsymbol{\theta}$ is possible if $\beta > \beta_{\text{IT}}$ with β_{IT} a constant first rigorously characterized in [125]. On the other hand, for $\beta < \beta_{\text{IT}}$ the posterior measure is very close to the prior and hence sampling from the posterior is not interesting. Finally, polynomial-time algorithms that achieve non-trivial estimation are known to exist for $\beta > \beta_{\text{alg}} := 1/\mathbb{E}\{\Theta^2\}$, while they are conjectured not to exist below that threshold [152]. (See Section 5.2 for further background on this model.)

Throughout the paper, we will assume β is fixed and is known to the statistician. If $\beta > \beta_{\text{alg}}$ is unknown, then β can be consistently estimated from the top eigenvalue of $\boldsymbol{X}$, $\lambda_1(\boldsymbol{X})$ [14]. If $\beta \le \beta_{\text{alg}}$, then either the posterior is close —in a suitable sense— to the prior (for $\beta < \beta_{\rm IT}$) or estimation (hence sampling) is conjectured to be hard (for $\beta_{\rm IT} < \beta < \beta_{\rm alg}$).

Given observation $\boldsymbol{X}$, the posterior takes the form:

$$\mu_{\boldsymbol{X}}({\rm d}\boldsymbol{\theta}) \propto \exp\left(\frac{\beta}{2}\langle\boldsymbol{\theta}, \boldsymbol{X}\boldsymbol{\theta}\rangle - \frac{\beta^2}{4n}\|\boldsymbol{\theta}\|_2^4\right)\pi_\Theta^{\otimes n}({\rm d}\boldsymbol{\theta})\,. \tag{5.10}$$

The problem of sampling from $\mu_{\boldsymbol{X}}$ is already interesting if the prior is the uniform distribution over $\{+1, -1\}$, i.e. $\pi_\Theta = {\rm Unif}(\{+1, -1\})$. In this case, model (5.8) is also known as $\mathbb{Z}_2$-synchronization. This is a prototype of the broader group-synchronization problems, and is closely related to the two-groups stochastic block model for random graphs [176, 64, 2].

For $\mathbb{Z}_2$-synchronization, the measure (5.10) takes the form of an Ising model [103], with 'coupling matrix' $\boldsymbol{X}$. For a general deterministic $\boldsymbol{X}$, sampling (even approximately) from the Ising measure is known to be #-P complete, see [178, 91] and references therein.

5.1.3 Contributions

In the present paper, we establish the following results.

Posterior sampling for spiked models. We describe an algorithm to sample from the posterior distribution (5.10) under the spiked model (5.8). The algorithm has complexity $O(n^2)$.

We establish a rigorous guarantee for this algorithm. Namely, we prove that there exists a constant β_*, such that, for $\beta \geq \beta_*$, the sample generated by the algorithm has distribution that is close to the actual posterior (5.10). Here, "close" means that the Wasserstein distance between these two distributions is $W_2(\mu_{\boldsymbol{X}}, \mu^{\rm alg}_{\boldsymbol{X}}) = o(\sqrt{n})$.

As mentioned above, β larger than a constant is the regime in which non-trivial estimation is possible. In this regime, the measure $\mu_{\boldsymbol{X}}$ is strongly correlated, and non-trivially aligned with the true $\boldsymbol{\theta}$. Note that this is different from the regime studied in [4], which corresponds to weak correlations.

A general framework to prove approximate sampling. Moving beyond the spiked model setting, we describe a general algorithm that, given an oracle that approximates the posterior mean of $\boldsymbol{\theta}$ (given additional observation $\boldsymbol{y}(t)$), samples from the posterior. We give a general approximation bound for samples generated by this algorithm, provided the oracle satisfies these conditions.

Posterior mean via approximate message passing (AMP). Our construction of the posterior mean estimator is based on the algorithm of [152], which uses approximate message passing in conjunction with a spectral initialization. We have to extend the analysis of [152] to prove Bayes optimality in presence of the additional observation process $\boldsymbol{y}(t)$. However, the most challenging part of the proof is to prove that the estimator is a Lipschitz function of $\boldsymbol{y}(t)$. We believe this is a result of independent interest. We prove it by building on recent work by Celentano, Fan, Mei [49].

The $\mathbb{Z}_2$-synchronization example is useful to emphasize an important technical subtlety. If the prior π_Θ is symmetric around the origin (i.e. $\pi_\Theta(A) = \pi_\Theta(-A)$), then the posterior is also symmetric under flips $\boldsymbol{\theta} \mapsto -\boldsymbol{\theta}$. In particular, the posterior mean of Eq. (5.7) vanishes at $t = 0$, $\boldsymbol{m}(\boldsymbol{y}(0), 0) = \boldsymbol{0}$. In the sampling algorithm, we need to break this symmetry. We achieve this by letting $\boldsymbol{\nu}$ be a top eigenvector of $\boldsymbol{X}$ (almost surely the top eigenvalue is non-degenerate and hence there are only two choices for this eigenvector), and replacing $\mu_{\boldsymbol{X},t}$ by the truncation $\mu^{+}_{\boldsymbol{X},t}(\mathrm{d}\boldsymbol{\theta}) \propto \mu_{\boldsymbol{X},t}(\mathrm{d}\boldsymbol{\theta})\,\mathbf{1}_{\langle\boldsymbol{\nu},\boldsymbol{\theta}\rangle\geq 0}$.

We expect our sampling algorithm to be successful in a broader range of values of β, see Remark 5.3.2 below. Proving Lipschitz continuity of the AMP estimator is the current bottleneck towards reaching this goal.

The rest of the paper is organized as follows. We survey related work in Section 5.2 and state our results in Section 5.3. In Section 5.4 we describe a general guarantee for sampling via stochastic localization. Some numerical experiments are presented in Section 5.5, and Section 5.6 presents the proof for sampling in the spiked model, with most technical details deferred to the appendices.

5.1.4 Notations

For $n \in \mathbb{N}_+$, we define the set $[n] := \{1, 2, \cdots, n\}$. We denote by $\mathbb{S}_n$ the collection of all $n \times n$ symmetric real matrices, and let $\mathcal{O}(n)$ be the collection of $n \times n$ orthogonal matrices. We denote by $\boldsymbol{A}^+$ the pseudoinverse of matrix $\boldsymbol{A}$.

For random variables X and Y, we write $X \perp Y$ if and only if X is independent of Y. We denote by p-lim convergence in probability. The set of probability measures over the measurable space $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ is denoted by $\mathscr{P}(\mathbb{R}^n)$, and the set of probability measures with finite second moment by $\mathscr{P}_2(\mathbb{R}^n)$. We write $\mathrm{Law}(X)$ for the probability distribution of the random variable (or vector) X.

The W_2 Wasserstein distance of two measures μ, ν on $\mathbb{R}^n$ is denoted by $W_2(\mu,\nu) := \inf_{P\in\mathcal{C}(\mu,\nu)} \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}}\{\|\boldsymbol{X}-\boldsymbol{Y}\|_2^2\}^{1/2}$, the infimum being taken over all couplings of μ and ν. We will also consider the scaled distance $W_{2,n}(\mu,\nu) := W_2(\mu,\nu)/n$.

5.2 Further related work

As mentioned in the introduction, Markov Chain Monte Carlo is the dominant approach to sampling from Bayes posteriors. In the case of non-convex and possibly discrete distributions such as Eq. (5.10) of interest here, Gibbs sampling (a.k.a. Glauber dynamics) would probably be the method of choice. This Markov chain updates one coordinate at each step according to its conditional probability distribution given the other coordinates. Classical methods to bound the mixing time of such a Markov chain are based on the so-called Dobrushin condition [69] and require (in the present case) $\beta \lesssim 1/\sqrt{n}$. Over the last two years, remarkable breakthroughs were achieved establishing Markov Chain mixing under much weaker 'spectral mixing conditions' [22, 82, 6]. Existing results only apply to the case of $\mathbb{Z}_2$-synchronization (i.e., $\boldsymbol{\theta} \in \{+1,-1\}^n$). In that setting, they yield mixing under the condition $\beta(\lambda_{\max}(\boldsymbol{X}) - \lambda_{\min}(\boldsymbol{X})) \leq 1$. Using classical results about extremal eigenvalues of spiked random matrices [14], this condition amounts to $\beta < 1/4$. This is not a regime of statistical interest, since for $\beta < 1$ it is impossible to produce an estimator with non-vanishing correlation with the true signal [64].

Variational methods [112, 194, 139, 36] provide another popular approach to approximate Bayesian analysis. In the challenging regime of constant signal-to-noise ratio treated here, naive mean field methods are known to yield incorrect inference [94]. However, for $\mathbb{Z}_2$-synchronization, asymptotically correct inference can be achieved using the so-called Thouless-Anderson-Palmer (TAP) approach [84, 49, 172]. Let us emphasize that these methods do not yield a sampling procedure.

The recent work [118] develops a sampling algorithm for Ising models that merges ideas from MCMC and variational inference. However, even for the case of $\mathbb{Z}_2$-synchronization, this approach falls short of providing an algorithm that is successful in the regime of β that is of statistical interest.

A substantial literature characterize optimal estimation error, efficient algorithms and computational barriers for the spiked model (5.8) and its generalizations to rank larger than one or to asymmetric matrices. A subset of these works include [64, 68, 142, 78, 125, 19, 21, 52, 152, 154].

As already mentioned, the present work is closely related to [4], which first uses an algorithmic implementation of stochastic localization, in conjunction with an AMP approximation of the posterior expectation. However, [4] focuses uniquely on the Sherrington-Kirkpatrick model, i.e. the measure over $\boldsymbol{\theta} \in \{+1, -1\}^n$, given by $\mu_{\boldsymbol{W}}(\boldsymbol{\theta}) \propto \exp(\beta\langle\boldsymbol{\theta}, \boldsymbol{W}\boldsymbol{\theta}\rangle/2)$, whereby $\boldsymbol{W} \sim \mathrm{GOE}(n)$ (no spike). The paper [4] establishes sampling guarantees for $\beta < 1/2$, a result improved to $\beta < 1$ by Michael Celentano [48]. Hardness for $\beta > 1$ was established in [4] for the class of 'stable' algorithms.

The Sherrington-Kirkpatrick model is not of statistical interest: the data is "pure noise", and the probability measure is not a Bayes posterior. To the best of our knowledge, the present paper is the first application of diffusion-based methods to statistical inference.

In concurrent work, guarantees for sampling via diffusions were recently established in [123, 54, 53] (which followed [4]). Two important differences with respect to the present paper are: (1) The results of [123, 54, 53] *assume* the existence of an accurate posterior mean estimator, while we construct it; (2) The posterior mean is assumed to be very accurate, and the resulting sampling guarantee is, as a consequence, stronger.

We note that the latter point is related to a difference in the proof technique and it is crucial because at the moment we are not aware of posterior mean estimators with the accuracy required by [123, 54, 53].

5.3 Main results

Recall the data distribution model of Eq. (5.8). We assume $\boldsymbol{\theta}$ to be distributed according to the prior of Eq. (5.9), where π_Θ is independent of n and known. Without loss of generality, we will assume π_Θ to have unit second moment $\int \theta^2 \, \pi_\Theta(\mathrm{d}\theta) = 1$.

As mentioned several times, we need to construct an approximation $\hat{\boldsymbol{m}}(\boldsymbol{y}, t)$ of the posterior mean function $\boldsymbol{m}(\boldsymbol{y}, t) := \mathbb{E}[\boldsymbol{\theta}|\boldsymbol{X}, t\boldsymbol{\theta} + \boldsymbol{G}(t) = \boldsymbol{y}]$, to be used in Eq. (5.6). For this purpose, we use the Bayes AMP algorithm with spectral initialization [152]. Namely, define the scalar denoiser $\mathsf{F}(\,\cdot\,;\gamma) : \mathbb{R} \to \mathbb{R}$ via

$$\mathsf{F}(z;\gamma) := \mathbb{E}[\Theta|\gamma\Theta + \sqrt{\gamma}G = z]\,, \tag{5.11}$$

where $(\Theta, G) \sim \pi_\Theta \otimes \mathsf{N}(0,1)$. Of course, the function $\mathsf{F}(z;\gamma)$ can be evaluated efficiently numerically, via a onre-dimensional integral (or a sum if π_Θ is discrete). As an example, in the case of $\mathbb{Z}_2$-synchronization (i.e., $\pi_\Theta = (\delta_{+1} + \delta_{-1})/2$), we have $\mathsf{F}(z;\gamma) = \tanh(z)$.

For $\boldsymbol{z} \in \mathbb{R}^n$, we denote by $\mathsf{F}(\boldsymbol{z};\gamma)$ the entrywise action of $\mathsf{F}(\,\cdot\,;\gamma)$, i.e.,

$$\mathsf{F}(\boldsymbol{z};\gamma) = \big(\mathsf{F}(z_1;\gamma), \dots, \mathsf{F}(z_n;\gamma)\big)^{\mathsf{T}}.$$

Given input $(\boldsymbol{X}, \boldsymbol{y}(t))$, we compute $\boldsymbol{\nu} \in \mathbb{R}^n$ that is a top eigenvector of $\boldsymbol{X}$, rescaled such that $\|\boldsymbol{\nu}\|_2 = \sqrt{n\beta^2(\beta^2-1)}$. The Bayes AMP estimate at time t is then computed recursively as follows:

$$\begin{aligned} \boldsymbol{z}_t^0 &= \boldsymbol{\nu}\,, \\ \hat{\boldsymbol{m}}_t^k &= \mathsf{F}(\boldsymbol{z}_t^k; \alpha_t^k)\,, \\ \boldsymbol{z}_t^{k+1} &= \beta \boldsymbol{X} \hat{\boldsymbol{m}}_t^k + \boldsymbol{y}(t) - b_t^k \hat{\boldsymbol{m}}_t^{k-1}\,, \end{aligned} \tag{5.12}$$

Here the sequence α_t^k is defined by the state evolution recursion

$$\begin{aligned} \alpha_t^0 &= \beta^2 - 1, \\ \alpha_t^{k+1} &= \beta^2 \mathbb{E}[\mathbb{E}[\Theta \mid \alpha_t^k \Theta + (\alpha_t^k)^{1/2} G]^2] + t, \end{aligned} \tag{5.13}$$

and the memory (Onsager) coefficient is given by

$$b_t^k = \beta^2 \mathbb{E}[(\Theta - \mathbb{E}[\Theta \mid \alpha_t^k \Theta + (\alpha_t^k)^{1/2} G])^2]\,. \tag{5.14}$$

Earlier literature characterizes optimality of Bayes AMP in the case $t = 0$, see [152] and references therein. Namely, under suitable conditions on β, Bayes AMP computes an estimator that is asymptotically equivalent, as $n \to \infty$, to the Bayes-optimal estimator.

In order to be more precise, we introduce the following free energy functional:

$$\Phi(\gamma, \beta, t) := \frac{\gamma^2}{4\beta^2} - \frac{\gamma}{2} + \mathsf{I}(\gamma + t), \tag{5.15}$$

where $\mathsf{I}(\gamma) := I(\Theta; Y)$ is the mutual information between Θ and $Y = \sqrt{\gamma}\Theta + G$ when $(\Theta, G) \sim \pi_\Theta \otimes \mathsf{N}(0,1)$. Explicitly, $\mathsf{I}(\gamma) = \mathbb{E} \log \frac{\mathrm{d}p_{Y|\Theta}}{\mathrm{d}p_Y}(Y, \Theta)$.

In analogy with Corollary 2.3 in [152], we expect the Bayes AMP algorithm of Eq. (5.12) to achieve Bayes optimality for all $t \in \mathbb{R}_{\geq 0}$, if and only if the following condition is satisfied.

Condition 5.3.1. *The global minimum of $\gamma \mapsto \Phi(\gamma, \beta, t)$ over $\gamma \in (0, \infty)$ is also the first stationary point of the same function on $(0, \infty$, for all $t \geq 0$.*

By evaluating the function $\Phi(\gamma, \beta, t)$ for specific priors π_Θ, we observe that typically there exists $\beta_*(\pi_\Theta) < \infty$ such that this condition holds for all $\beta \geq \beta_*(\pi_\Theta)$ (and possibly other intervals of β as well.) In particular, this is true when π_Θ is supported on finitely many points (see Proposition D.4.1). In our running example, $\mathbb{Z}_2$-synchronization, the situation is even simpler: Bayes AMP is asymptotically Bayes optimal for all $\beta > 0$ [64].

The algorithm proceeds in slightly different ways depending on whether π_Θ is symmetric around 0 or not. If it is symmetric, then posterior $\mu_{\boldsymbol{X}}$ is also symmetric under reflection $\boldsymbol{\theta} \mapsto -\boldsymbol{\theta}$: we take account of this by explicitly symmetrizing at the end.

If π_Θ is not symmetric, we use the next lemma to align the initial spectral estimate with $\boldsymbol{\theta}$.

Lemma 5.3.1. *If π_Θ is not symmetric about the origin, then for any $\beta > 1$, there exists an algorithm $\mathcal{A} : \mathbb{R}^n \to \{+1, -1\}$ with complexity $O(n)$, such that with probability $1 - o_n(1)$*

$$\lim_{n \to \infty} \mathbb{P}\big(\mathcal{A}(\boldsymbol{\nu}) = \operatorname{sign}(\langle \boldsymbol{\theta}, \boldsymbol{\nu} \rangle)\big) = 1\,.$$

We postpone the proof of Lemma 5.3.1 to Appendix D.3.

Finally, our sampling algorithm is defined by Algorithm 1.

Algorithm 1 Diffusion-based sampling for spiked models

```
Input: Data $\boldsymbol{X}$, parameters $(\beta, K_{\mathsf{AMP}}, L, \delta)$;
Set $\hat{\boldsymbol{y}}_0 = \mathbf{0}_n$;
Compute $\boldsymbol{\nu}$, a uniformly random leading eigenvector of $\boldsymbol{X}$;
Normalize $\|\boldsymbol{\nu}\|_2^2 = n\beta^2(\beta^2 - 1)$;
if $\pi_\Theta$ is not symmetric then
    $\boldsymbol{\nu} \leftarrow \mathcal{A}(\boldsymbol{\nu})\boldsymbol{\nu}$;
end if
for $\ell = 0, 1, \cdots, L-1$ do
    Draw $\boldsymbol{w}_{\ell+1} \sim \mathsf{N}(0, \boldsymbol{I}_n)$ independent of everything so far;
    Let $\hat{\boldsymbol{m}}(\hat{\boldsymbol{y}}_\ell, \ell\delta)$, be the output of Bayes AMP algorithm (5.12) with $K_{\mathsf{AMP}}$ iterations;
    Update $\hat{\boldsymbol{y}}_{\ell+1} = \hat{\boldsymbol{y}}_\ell + \hat{\boldsymbol{m}}(\hat{\boldsymbol{y}}_\ell, \ell\delta)\delta + \sqrt{\delta}\boldsymbol{w}_{\ell+1}$;
end for
if $\pi_\Theta$ is symmetric then
    Set $s \sim \mathrm{Unif}(\{+1, -1\})$;
else
    Set $s = +1$;
end if
return $\boldsymbol{\theta}^{\mathrm{alg}} = s \cdot \hat{\boldsymbol{m}}(\hat{\boldsymbol{y}}_L, L\delta)$;
```

We will call $\mu_{\boldsymbol{X}}^{\mathrm{alg}} := \mathrm{Law}(\boldsymbol{\theta}^{\mathrm{alg}})$ the distribution of the algorithm output $\boldsymbol{\theta}^{\mathrm{alg}}$. Our main result is the following guarantee for Algorithm 1.

Theorem 5.3.1. *Assume that π_Θ is supported on finitely many points. Then there exists a constant $\beta_0(\pi_\Theta) \geq \beta_*(\pi_\Theta)$ depending only on π_Θ, such that for all $\beta \geq \beta_0(\pi_\Theta)$, the following holds.*

For any $\xi > 0$, there exist $K_{\mathsf{AMP}}, L \in \mathbb{N}$ and $\delta \in \mathbb{R}_{>0}$ that depend uniquely on (β, ξ, π_Θ), such that if Algorithm 1 takes as input $(\boldsymbol{X}, \beta, K_{\mathsf{AMP}}, L, \delta)$, then, with probability $1 - o_n(1)$ with respect to the choice of $\boldsymbol{X}$,

$$W_{2,n}(\mu_{\boldsymbol{X}}, \mu_{\boldsymbol{X}}^{\mathrm{alg}}) \leq \xi\,.$$

(We recall that $W_{2,n}(\mu, \nu) = W_2(\mu, \nu)/\sqrt{n}$ is the scaled Wasserstein distance.)

The proof of this theorem is given in Section 5.6.

Remark 5.3.1. The assumption that π_Θ is supported on a finite number of points is likely to be a proof artifact, and is only used in showing that the AMP approximation of the posterior mean function is Lipschitz continuous. In the all other lemmas we assume the weaker **??** 5.3.1. Weakening this finite support assumption is an interesting direction for future work.

Remark 5.3.2. We expect our sampling algorithm to be successful in a broader range of values of β. Namely, we expect it to succeed for all $\beta \in (\beta_{\mathrm{AMP}}^+, \infty)$ where β_{AMP}^+ is the supremum value of β at which AMP does not achieve (asymptotically) Bayes optimal estimation.

In the other direction, it has been conjectured that if Bayes AMP does not reach the Bayes optimal estimation error, then no polynomial time algorithm does. The paper [153] provides some rigorous support for this conjecture. Of course if Bayes optimal estimation is impossible, so is sampling. Therefore, under the mentioned conjecture, $(\beta^+_{\rm AMP}, \infty)$ is the largest interval of β over which efficient sampling is possible.

Remark 5.3.3. Note that if the prior is discrete, Algorithm 1 may return a vector $\boldsymbol{\theta}^{\rm alg}$ that is outside the support of the prior. This does not contradicts Theorem 5.3.1, which guarantees that $\boldsymbol{\theta}^{\rm alg}$ is close in ℓ_2 distance to a sample $\boldsymbol{\theta} \sim \mu_{\boldsymbol{X}}$ but not necessarily in the support.

If necessary, this can be remedied by a simple rounding procedure. Simply, replace the last line of Algorithm 1 by the following. Compute $\boldsymbol{m}^{\rm alg} := s \cdot \hat{\boldsymbol{m}}(\hat{\boldsymbol{y}}_L, L\delta)$, and return a vector with conditionally independent coordinates given $\boldsymbol{m}^{\rm alg}$ such that, for each $i \le n$ $\theta^{\rm alg}_i$ is in the support of π_Θ and $\mathbb{E}[\theta^{\rm alg}_i | \boldsymbol{m}^{\rm alg}] = m^{\rm alg}_i$. By convexity of the W_2 distance, this modification enjoys the same guarantees.

In fact the properties of AMP can be used to construct an even better rounding procedure, but we defer this to future work.

5.4 Variants and generalizations

The technique developed in this paper (which in turn builds on [4]) is applicable to other posterior sampling problems beyond the spiked model treated in the previous section. In this section, we consider the general Bayesian posterior sampling problem as introduced in Section 5.1 and describe a sampling algorithm based on a 'linear observation process,' whereby the $\boldsymbol{y}(t)$ of Eq. (5.3) is replaced by a noisy linear function of the unknown parameters. We then further extend this setting in Section 5.4.2 to the case of non-linear observations, illustrating this generalization in the context of the spiked model.

5.4.1 Stochastic localization and sampling via linear observations

Consider the general posterior[2] $\mu_{\boldsymbol{X}}$ of Eq. (5.2). For $\boldsymbol{\theta} \sim \mu_{\boldsymbol{X}}$, we define the linear observation process $\{\boldsymbol{y}(t)\}_{t \ge 0}$ via

$$\boldsymbol{y}(t) = t\, \boldsymbol{H}\, \boldsymbol{\theta} + \boldsymbol{G}(t)\,, \tag{5.16}$$

where $\{\boldsymbol{G}(t)\}_{t\ge 0}$ is a standard Brownian motion and the matrix $\boldsymbol{H} \in \mathbb{R}^{m\times n}$ is to be designed by the statistician. The posterior distribution of $\boldsymbol{\theta}$ given $\boldsymbol{X}$ and $\boldsymbol{y}(t)$ takes the form:

$$\begin{aligned} \mu_{\boldsymbol{X},t}({\rm d}\boldsymbol{\theta}) &= \mathbb{P}({\rm d}\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{y}(t)) && (5.17)\\ &= \frac{1}{Z(\boldsymbol{X},t)} \exp\left\{ \langle \boldsymbol{y}(t), \boldsymbol{H\theta}\rangle - \frac{t}{2}\|\boldsymbol{H\theta}\|_2^2 \right\} \mu_{\boldsymbol{X}}({\rm d}\boldsymbol{\theta})\,. && (5.18) \end{aligned}$$

We define two posterior mean functions

$$\boldsymbol{m}(\boldsymbol{y},t) := \mathbb{E}[\boldsymbol{H\theta}|\boldsymbol{X}, t\boldsymbol{H\theta} + \boldsymbol{G}(t) = \boldsymbol{y}]\,, \qquad \boldsymbol{m}_{\boldsymbol{\theta}}(\boldsymbol{y},t) := \mathbb{E}[\boldsymbol{\theta}|\boldsymbol{X}, t\boldsymbol{H\theta} + \boldsymbol{G}(t) = \boldsymbol{y}]\,. \tag{5.19}$$

Of course $\boldsymbol{m}(\boldsymbol{y},t) = \boldsymbol{H}\boldsymbol{m}_{\boldsymbol{\theta}}(\boldsymbol{y},t)$, but distinguishing the two functions is useful when considering estimators

[2] Of course, any probability distribution over $\mathbb{R}^n$ also fits the same framework, since we can always take $\boldsymbol{X} = \emptyset$

that approximate them. By a straightforward generalization of Eq. (5.6), $(\boldsymbol{y}(t))_{t\geq 0}$ satisfies the SDE

$$\boldsymbol{y}(t) = \boldsymbol{m}(\boldsymbol{y}(t), t)\mathrm{d}t + \mathrm{d}\boldsymbol{B}(t), \qquad \boldsymbol{y}(0) = \boldsymbol{0}_m\,,$$

with $\{\boldsymbol{B}(t)\}_{t\geq 0}$ being a standard Brownian motion in $\mathbb{R}^m$.

Algorithm 2 defines the sampling procedure associated to this stochastic process. This makes use of an oracle $\hat{\boldsymbol{m}}(\boldsymbol{y}, t)$ that approximates $\boldsymbol{m}(\boldsymbol{y}, t)$, and an oracle $\hat{\boldsymbol{m}}_{\boldsymbol{\theta}}(\boldsymbol{y}, t)$ that approximates $\boldsymbol{m}_{\boldsymbol{\theta}}(\boldsymbol{y}, t)$.

Algorithm 2 Approximate sampling

```
Input: Parameters $(L, \delta, \bar{R})$;
Set $\hat{\boldsymbol{y}}_0 = \boldsymbol{0}$;
for $\ell = 0, 1, \cdots, L-1$ do
    Draw $\boldsymbol{w}_{\ell+1} \sim \mathsf{N}(0, \boldsymbol{I}_m)$, independent of everything so far;
    Update $\hat{\boldsymbol{y}}_{\ell+1} = \hat{\boldsymbol{y}}_\ell + \delta\hat{\boldsymbol{m}}(\hat{\boldsymbol{y}}_\ell, \delta\ell) + \sqrt{\delta}\boldsymbol{w}_{\ell+1}$;
end for
return $\boldsymbol{\theta}^{\rm alg} = \hat{\boldsymbol{m}}_{\boldsymbol{\theta}}(\hat{\boldsymbol{y}}_L, L\delta)\mathbb{1}\{\|\hat{\boldsymbol{m}}_{\boldsymbol{\theta}}(\hat{\boldsymbol{y}}_L, L\delta)\|_2/\sqrt{n} \leq \bar{R}\}$;
```

Remark 5.4.1. Note that the oracle $\hat{\boldsymbol{m}}_{\boldsymbol{\theta}}(\boldsymbol{y}, t)$ is only used at the final time $t = T := L\delta$. Hence we only need to approximate the posterior expectation of $\boldsymbol{\theta}$ given $\boldsymbol{y}(T)/T = \boldsymbol{H}\boldsymbol{\theta} + (\boldsymbol{g}/\sqrt{T})$, where $\boldsymbol{g} \sim \mathsf{N}(0, \boldsymbol{I}_m)$. For large T, this corresponds to very low noise. Constructing such an oracle is relatively easy in a number of circumstances, as illustrated by two examples:

- $\boldsymbol{H}$ has full column rank (in particular, $m \geq n$). We can then approximate $\hat{\boldsymbol{m}}_{\boldsymbol{\theta}}(\boldsymbol{y}; T) = \boldsymbol{H}^{+}\boldsymbol{y}/T$.
- $\boldsymbol{H}$ does not have full column rank (e.g., $m < n$), but π_Θ is supported on sparse vectors. In this case, standard techniques from compressed sensing and high-dimensional regression can be brought to bear [187, 71, 47, 41].

Remark 5.4.2. In contrast with the previous remark, the oracle $\hat{\boldsymbol{m}}(\boldsymbol{y}, t)$ is required for all t. However, one can hope to exploit the freedom to choose the matrix $\boldsymbol{H}$ to simplify this task.

We provide a theoretical guarantee for Algorithm 2 under the following assumptions (throughout $\ell \in \mathbb{N}$):

(A1) (Posterior mean consistency) With probability at least $1-\eta$, it holds that

$$\frac{1}{\sqrt{m}}\max_{\ell\leq L}\big\|\boldsymbol{m}(\boldsymbol{y}(\ell\delta), \ell\delta) - \hat{\boldsymbol{m}}(\boldsymbol{y}(\ell\delta), \ell\delta)\big\|_2 \leq \varepsilon_1.$$

Further, with the same probability, $\|\boldsymbol{m}_{\boldsymbol{\theta}}(\boldsymbol{y}(T), T) - \hat{\boldsymbol{m}}_{\boldsymbol{\theta}}(\boldsymbol{y}(T), T)\big\|_2 \leq \varepsilon_1\sqrt{n}$.

(A2) (Path regularity) With probability at least $1-\eta$, it holds that

$$\max_{\ell\leq L}\sup_{t\in[\ell\delta,(\ell+1)\delta]}\frac{1}{\sqrt{m}}\|\boldsymbol{m}(\boldsymbol{y}(t), t) - \boldsymbol{m}(\boldsymbol{y}(\ell\delta), \ell\delta)\|_2 \leq C_1\sqrt{\delta} + \varepsilon_2.$$

(A3) (Lipschitz continuity)) There exists a sequence $\{r_\ell\}_{1\leq\ell\leq L} \subseteq \mathbb{R}_+$ such that, letting $B(\ell) := \{\boldsymbol{y} \in \mathbb{R}^m : \|\boldsymbol{y} - \boldsymbol{y}(\ell\delta)\| \leq r_\ell\sqrt{m}\}$, then the following holds with probability at least $1-\eta$:

$$\max_{\ell\leq L}\sup_{\boldsymbol{y}_1\neq\boldsymbol{y}_2\in B(\ell)}\frac{\|\hat{\boldsymbol{m}}(\boldsymbol{y}_1, \ell\delta) - \hat{\boldsymbol{m}}(\boldsymbol{y}_2, \ell\delta)\|_2}{\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2} \leq C_2\,.$$

Further $r(\ell) > (C_1\sqrt{\delta}+\varepsilon_1+\varepsilon_2)e^{C_2\ell\delta}/C_2$ for all $\ell \le L$. Finally, with the same probability, $\|\boldsymbol{m_\theta}(\boldsymbol{y}_1, T) - \hat{\boldsymbol{m}}_{\boldsymbol{\theta}}(\boldsymbol{y}_2, T)\|_2/\sqrt{n} \le C_2\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2/\sqrt{m}$ for all $\boldsymbol{y}_1, \boldsymbol{y}_2$.

The dependence on constants C_1, C_2 will be tracked in the statement below.

Theorem 5.4.1. *Assume $\mu_{\boldsymbol{X}}$ to be such that $\int(\|\boldsymbol{\theta}\|_2^2/n)^{c_0}\mu_{\boldsymbol{X}}(\mathrm{d}\boldsymbol{\theta}) \le R^{2c_0}$ for some $c_0 > 1$, and that assumptions (A1), (A2), (A3) hold. Then, letting $\mu_{\boldsymbol{X}}^{\mathrm{alg}} := \mathrm{Law}(\boldsymbol{\theta}^{\mathrm{alg}})$ be the distribution of the output of Algorithm 2, we have:*

$$W_{2,n}(\mu_{\boldsymbol{X}}, \mu_{\boldsymbol{X}}^{\mathrm{alg}}) \le \varepsilon_1 + (C_1\sqrt{\delta} + \varepsilon_1 + \varepsilon_2)\cdot e^{C_2 T} + R^{c_0}/\bar{R}^{c_0-1} + 10\bar{R}\eta + W_{2,n}\big(\mu_{\boldsymbol{X}}, \mathrm{Law}(\boldsymbol{m_\theta}(\boldsymbol{y}(T), T))\big). \tag{5.20}$$

In addition, there exists $\bar{R} > 0$, such that

$$W_{2,n}(\mu_{\boldsymbol{X}}, \mu_{\boldsymbol{X}}^{\mathrm{alg}}) \le \varepsilon_1 + (C_1\sqrt{\delta} + \varepsilon_1 + \varepsilon_2)\cdot e^{C_2 T} + C(c_0)R\eta^{(c_0-1)/c_0} + W_{2,n}\big(\mu_{\boldsymbol{X}}, \mathrm{Law}(\boldsymbol{m_\theta}(\boldsymbol{y}(T), T))\big), \tag{5.21}$$

where $C(c_0)$ is a constant depending uniquely on c_0. If in addition $\boldsymbol{H}$ has full column rank, then

$$W_{2,n}(\mu_{\boldsymbol{X}}, \mu_{\boldsymbol{X}}^{\mathrm{alg}}) \le \varepsilon_1 + (C_1\sqrt{\delta} + \varepsilon_1 + \varepsilon_2)\cdot e^{C_2 T} + C(c_0)R\eta^{(c_0-1)/c_0} + \frac{1}{nT}\mathrm{Tr}\left((\boldsymbol{H}^\mathsf{T}\boldsymbol{H})^{-1}\right). \tag{5.22}$$

Proof. We couple $\{\boldsymbol{w}_\ell\}_{1\le\ell\le L}$ and $\{\boldsymbol{B}(t)\}_{0\le t\le T}$ by letting $\boldsymbol{w}_\ell = \boldsymbol{B}(\ell\delta) - \boldsymbol{B}((\ell-1)\delta)$, and define $A_\ell = \|\hat{\boldsymbol{y}}_\ell - \boldsymbol{y}(\ell\delta)\|_2/\sqrt{m}$. We also write $t_\ell := \ell\delta$.

Let Ω be the intersection of the events at points (A1), (A2), (A3). By union bound $\mathbb{P}(\Omega) \ge 1 - 5\eta$. We will prove by induction that, on Ω, the following holds for $\ell \le L$:

$$\hat{\boldsymbol{y}}_\ell \in B(\ell), \quad \text{and} \quad A_\ell \le \frac{C_1\sqrt{\delta}+\varepsilon_1+\varepsilon_2}{C_2}\cdot\left(e^{C_2\ell\delta} - 1\right). \tag{5.23}$$

Indeed, by definition, $A_0 = 0$ and $\hat{\boldsymbol{y}}_0 = \boldsymbol{y}(0) = \boldsymbol{0} \in B(0)$. Next, assume that the induction hypothesis holds up to step $\ell - 1$. On the event Ω:

$$\begin{aligned}
A_\ell - A_{\ell-1} &\le \frac{1}{\sqrt{m}}\int_{t_{\ell-1}}^{t_\ell} \|\hat{\boldsymbol{m}}(\hat{\boldsymbol{y}}_{\ell-1}, t_{\ell-1}) - \boldsymbol{m}(\boldsymbol{y}(t), t)\|_2 \mathrm{d}t \\
&\le \frac{\delta}{\sqrt{m}}\|\hat{\boldsymbol{m}}(\boldsymbol{y}(t_{\ell-1}), t_{\ell-1}) - \boldsymbol{m}(\boldsymbol{y}(t_{\ell-1}), t_{\ell-1})\|_2 \\
&\quad + \sup_{t\in[t_{\ell-1}, t_\ell]}\frac{\delta}{\sqrt{m}}\|\boldsymbol{m}(\boldsymbol{y}(t), t) - \boldsymbol{m}(\boldsymbol{y}(t_{\ell-1}), t_{\ell-1})\|_2 \\
&\quad + \frac{\delta}{\sqrt{m}}\|\hat{\boldsymbol{m}}(\hat{\boldsymbol{y}}_{\ell-1}, t_{\ell-1}) - \hat{\boldsymbol{m}}(\boldsymbol{y}((\ell-1)\delta), t_{\ell-1})\|_2 \\
&\le \delta\cdot\Big(\varepsilon_1 + C_1\sqrt{\delta} + \varepsilon_2 + C_2 A_{\ell-1}\Big).
\end{aligned}$$

Substituting in the induction hypothesis, we obtain $A_\ell \le \frac{C_1\sqrt{\delta}+\varepsilon_1+\varepsilon_2}{C_2}\cdot\left(e^{C_2\ell\delta} - 1\right)$ as desired. The claim $\hat{\boldsymbol{y}}_\ell \in B(\ell)$ follows from the stated condition on r_ℓ. This complete the induction proof.

Applying the bound (5.23) to $\ell = L$ and using once more assumptions (A1) and (A3), we have that on

Ω $(T = L\delta)$

$$\begin{aligned}\frac{1}{\sqrt{n}}\|\boldsymbol{m}_{\boldsymbol{\theta}}(\boldsymbol{y}(T),T)-\hat{\boldsymbol{m}}_{\boldsymbol{\theta}}(\hat{\boldsymbol{y}}_L,T)\|_2 \leq & \frac{1}{\sqrt{n}}\|\boldsymbol{m}_{\boldsymbol{\theta}}(\boldsymbol{y}(T),T)-\hat{\boldsymbol{m}}_{\boldsymbol{\theta}}(\boldsymbol{y}(T),T)\|_2\\ &+\frac{1}{\sqrt{n}}\|\hat{\boldsymbol{m}}_{\boldsymbol{\theta}}(\boldsymbol{y}(T),T)-\hat{\boldsymbol{m}}_{\boldsymbol{\theta}}(\hat{\boldsymbol{y}}_L,T)\|_2\\ \leq & \varepsilon_1 + C_2 A_L\\ \leq & \varepsilon_1 + (C_1\sqrt{\delta}+\varepsilon_1+\varepsilon_2)\cdot e^{C_2 T} =: \Delta\,. \end{aligned} \tag{5.24}$$

This implies

$$W_{2,n}(\mu_{\boldsymbol{X}},\mu_{\boldsymbol{X}}^{\rm alg}) \leq \Delta + \mathbb{E}\Big(\|\boldsymbol{\theta}\|_2^2/n\mathbb{1}_{\|\boldsymbol{\theta}\|_2\geq \overline{R}\sqrt{n}}\Big)^{1/2} + 10\overline{R}\eta + W_{2,n}\big(\mu_{\boldsymbol{X}},\mathrm{Law}(\boldsymbol{m}_{\boldsymbol{\theta}}(\boldsymbol{y}(T),T))\big)\,. \tag{5.25}$$

This in turns implies Eqs. (5.20) and (5.21) by using the moment assumption to bound the expectation on the right-hand side and optimizing over $\overline{R}$.

Equation (5.22) follows by applying Lemma D.2.1.

□

Remark 5.4.3. We formulated assumptions (A1), (A2), (A3) as conditions that hold with-high probability. For this reason, we need to assume a bound on the $2+\varepsilon$ moment of $\|\boldsymbol{\theta}\|_2$ in the statement of Theorem 5.4.1.

The same argument also yields a guarantee under the minimal condition $\int(\|\boldsymbol{\theta}\|_2^2/n)\mu_{\boldsymbol{X}}(\mathrm{d}\boldsymbol{\theta}) \leq R^2$, if we replace (A1), (A2), (A3) by similar conditions that hold in expectation.

5.4.2 The use of nonlinear observations

We can further generalize the linear observation model (5.16), by admitting non-linear observations and non-isotropic noise. Such an observation process takes the form:

$$\boldsymbol{y}(t) = \int_0^t \boldsymbol{Q}(t)\boldsymbol{F}(\boldsymbol{\theta},t)\,\mathrm{d}t + \int_0^t \boldsymbol{Q}(s)^{1/2}\mathrm{d}\boldsymbol{G}(s)\,. \tag{5.26}$$

Here $\boldsymbol{Q} : \mathbb{R}_{\geq 0} \to \mathbb{S}_k$ is a function taking values in the cone of positive semidefinite matrices, and $\boldsymbol{F} : \mathbb{R}^n \times \mathbb{R}_{\geq 0} \to \mathbb{R}^k$.

It is straightforward to generalize the algorithm and analysis of the previous section to this case. Instead of doing this, we discuss a specific construction that is well suited to the spiked model analyzed in this paper. We let

$$\overline{\boldsymbol{Y}}(t) = \frac{t}{n}\boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}} + \overline{\boldsymbol{G}}(t)\,, \tag{5.27}$$

where $\overline{\boldsymbol{G}}(t)$ is a *symmetric Brownian motion*, i.e., a stochastic process taking values in $\mathbb{R}^{n\times n}$, with $\overline{\boldsymbol{G}}(t) = \overline{\boldsymbol{G}}(t)^{\mathsf{T}}$ and such that $(\overline{G}_{ij}(t))_{1\leq i\leq j\leq n}$ is a collection of independent Brownian motions (independent of $\boldsymbol{\theta}$), which are time scaled so that $\mathbb{E}\{\overline{G}_{ii}(t)^2\} = 2t/n$, $\mathbb{E}\{\overline{G}_{ij}(t)^2\} = t/n$ for $i < j$.

This observation process is of the same nature as the original observation that defines the model, cf. Eq. (5.8). In particular, the process (5.27) does not break the symmetry $\boldsymbol{\theta} \to -\boldsymbol{\theta}$. These remarks can be further formalized by noting that $\boldsymbol{Y}(\beta^2+t) := \beta\boldsymbol{X} + \overline{\boldsymbol{Y}}(t)$ is a sufficient statistics for $\boldsymbol{\theta}$ given $\boldsymbol{X}, \overline{\boldsymbol{Y}}(t)$. Of course, $\{\boldsymbol{Y}(t)\}_{t\geq\beta^2}$ takes the form

$$\boldsymbol{Y}(t)=\frac{t}{n}\boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}}+\boldsymbol{G}(t)\,, \tag{5.28}$$

where $\{\boldsymbol{G}(t)\}_{t\geq 0}$ is again a symmetric Brownian motion, except that it is initialized at $\boldsymbol{G}(\beta^2)=\beta\boldsymbol{W}$.

As for similar observation processes derived in the previous pages, $\boldsymbol{Y}(t)$ satisfies an SDE, namely

$$\mathrm{d}\boldsymbol{Y}(t)=\boldsymbol{M}(\boldsymbol{Y}(t);t)\mathrm{d}t+\mathrm{d}\boldsymbol{B}(t)\,, \tag{5.29}$$

$$\boldsymbol{M}(\boldsymbol{Y};t):=\mathbb{E}\Big\{\frac{1}{n}\boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}}\Big|\frac{t}{n}\boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}}+\boldsymbol{G}(t)=\boldsymbol{Y}\Big\}\,, \tag{5.30}$$

since we incorporated the observation $\boldsymbol{X}$ in $\boldsymbol{Y}(t)$, this SDE has to be solved with initialization at $t=\beta^2$:

$$\boldsymbol{Y}(\beta^2)=\beta\boldsymbol{X}\,. \tag{5.31}$$

Algorithm 3 Diffusion-based sampling for spiked models

```
Input: Data $\boldsymbol{X}$, parameters $(\beta, K_{\mathsf{AMP}}, L, \delta)$;
Set $\hat{\boldsymbol{Y}}_0=\beta\boldsymbol{X}\in\mathbb{R}^{n\times n}$;
for $\ell=0,1,\cdots,L-1$ do
    Draw $\boldsymbol{W}_{\ell+1}\sim\mathrm{GOE}(n)$ independent of everything so far;
    Let $\hat{\boldsymbol{M}}(\hat{\boldsymbol{Y}}_\ell,\ell\delta+\beta^2)$, be Bayes AMP estimate of $\boldsymbol{M}(\hat{\boldsymbol{Y}}_\ell,\ell\delta+\beta^2)$ (see main text);
    Update $\hat{\boldsymbol{Y}}_{\ell+1}=\hat{\boldsymbol{Y}}_\ell+\hat{\boldsymbol{M}}(\hat{\boldsymbol{Y}}_\ell,\ell\delta+\beta^2)\delta+\sqrt{\delta}\boldsymbol{W}_{\ell+1}$;
end for
Compute $\boldsymbol{X}^{\rm alg}=\hat{\boldsymbol{M}}(\hat{\boldsymbol{Y}}_L,L\delta+\beta^2)$, and let $\lambda_1(\boldsymbol{X}^{\rm alg})$, $\boldsymbol{v}_1(\boldsymbol{X}^{\rm alg})$ be its top eigenvalue/eigenvector
if $\pi_\Theta$ is symmetric then Draw $s\sim\mathrm{Unif}(\{+1,-1\})$
else Compute $s=\tilde{\mathcal{A}}(\boldsymbol{v}_1(\boldsymbol{X}^{\rm alg}))$
end if
return $\boldsymbol{\theta}^{\rm alg}=s\sqrt{\lambda_1(\boldsymbol{X}^{\rm alg})}\boldsymbol{v}_1(\boldsymbol{X}^{\rm alg})$
```

The resulting sampling procedure is outlined as Algorithm 3. Two components are unspecified: (i) An algorithm $\tilde{\mathcal{A}}:\mathbb{R}^n\to\{+1,-1\}$ such that, with high probability, $\tilde{\mathcal{A}}(\boldsymbol{v}_1)=\mathrm{sign}\langle\boldsymbol{v}_1,\boldsymbol{\theta}\rangle$ when $\boldsymbol{v}_1=\boldsymbol{v}_1(\boldsymbol{X}^{\rm alg})$; (ii) An algorithm to compute an approximation $\hat{\boldsymbol{M}}(\boldsymbol{Y},t)$ for the conditional expectation $\boldsymbol{M}(\boldsymbol{Y};t)$. The first algorithm is completely analogous to $\mathcal{A}$ introduced in Lemma 5.3.1.

Finally, in order to compute an approximation of $\boldsymbol{M}(\boldsymbol{Y},t)$ we use once more AMP, whereby we replace $\boldsymbol{X}$ by $\boldsymbol{Y}(t)$:

$$\boldsymbol{z}_t^0=\boldsymbol{\nu}_t\,, \tag{5.32}$$

$$\hat{\boldsymbol{m}}_t^k=\mathsf{F}(\boldsymbol{z}_t^k;\tilde{\alpha}_t^k)\,, \tag{5.33}$$

$$\boldsymbol{z}_t^{k+1}=\boldsymbol{Y}\hat{\boldsymbol{m}}_t^k-\tilde{b}_t^k\hat{\boldsymbol{m}}_t^{k-1}\,, \tag{5.34}$$

where $\boldsymbol{\nu}_t$ is a randomly selected top eigenvector of $\boldsymbol{Y}$, normalized such that $\|\boldsymbol{\nu}_t\|_2=\sqrt{nt(t-1)}$,

$$\tilde{\alpha}_t^0=t-1, \tag{5.35}$$

$$\tilde{\alpha}_t^{k+1}=t\mathbb{E}[\mathbb{E}[\Theta\mid\tilde{\alpha}_t^k\Theta+(\tilde{\alpha}_t^k)^{1/2}G]^2], \tag{5.36}$$

$$\tilde{b}_t^k=t\mathbb{E}[(\Theta-\mathbb{E}[\Theta\mid\tilde{\alpha}_t^k\Theta+(\tilde{\alpha}_t^k)^{1/2}G])^2]. \tag{5.37}$$

We then let $\hat{\boldsymbol{m}}(\boldsymbol{Y};t) = \hat{\boldsymbol{m}}_t^{K_{\mathsf{AMP}}}$ and $\hat{\boldsymbol{M}}(\boldsymbol{Y};t) = \hat{\boldsymbol{m}}(\boldsymbol{Y};t)\hat{\boldsymbol{m}}(\boldsymbol{Y};t)^{\mathsf{T}}$.

5.5 Numerical experiments

In this section we present numerical experiments in which we use Algorithm 1 to sample from the Bayes posterior of the spiked model. In these experiments we focus on $\mathbb{Z}_2$-synchronization, i.e., $\pi_\Theta = (\delta_{+1}+\delta_{-1})/2$. Since the prior is discrete, several standard techniques (e.g., Langevin or Hamiltonian Monte Carlo) do not apply. No guarantee exists for Gibbs sampling (also known as 'Glauber dynamics' in this context.)

In our first experiment, we set $L = 500$, $\delta = 0.02$, and $n = 1000$. In Figure 5.1, we plot the trajectories of the first and second coordinates of the mean vectors generated by the algorithm: $\hat{\boldsymbol{m}}(\hat{\boldsymbol{y}}_\ell, \ell\delta)$, $\ell \in \{0, \dots, L\}$. For each realization of the data, we run five independent experiments and plot the resulting trajectories.

We see that both $\hat{m}_1$ and $\hat{m}_2$ converge to either $+1$ or -1 as $t \to \infty$, regardless of the value of β. When the signal-to-noise ratio is below the information-theoretic threshold (that is to say, $\beta \le 1$), the trajectory appears to converge to an arbitrary corner. On the contrary, when the signal-to-noise ratio is above the information-theoretic threshold ($\beta > 1$), most trajectories that correspond to the same data $\boldsymbol{X}$ consistently converge to the same corner, which is correlated with the actual signal $\boldsymbol{\theta}$.

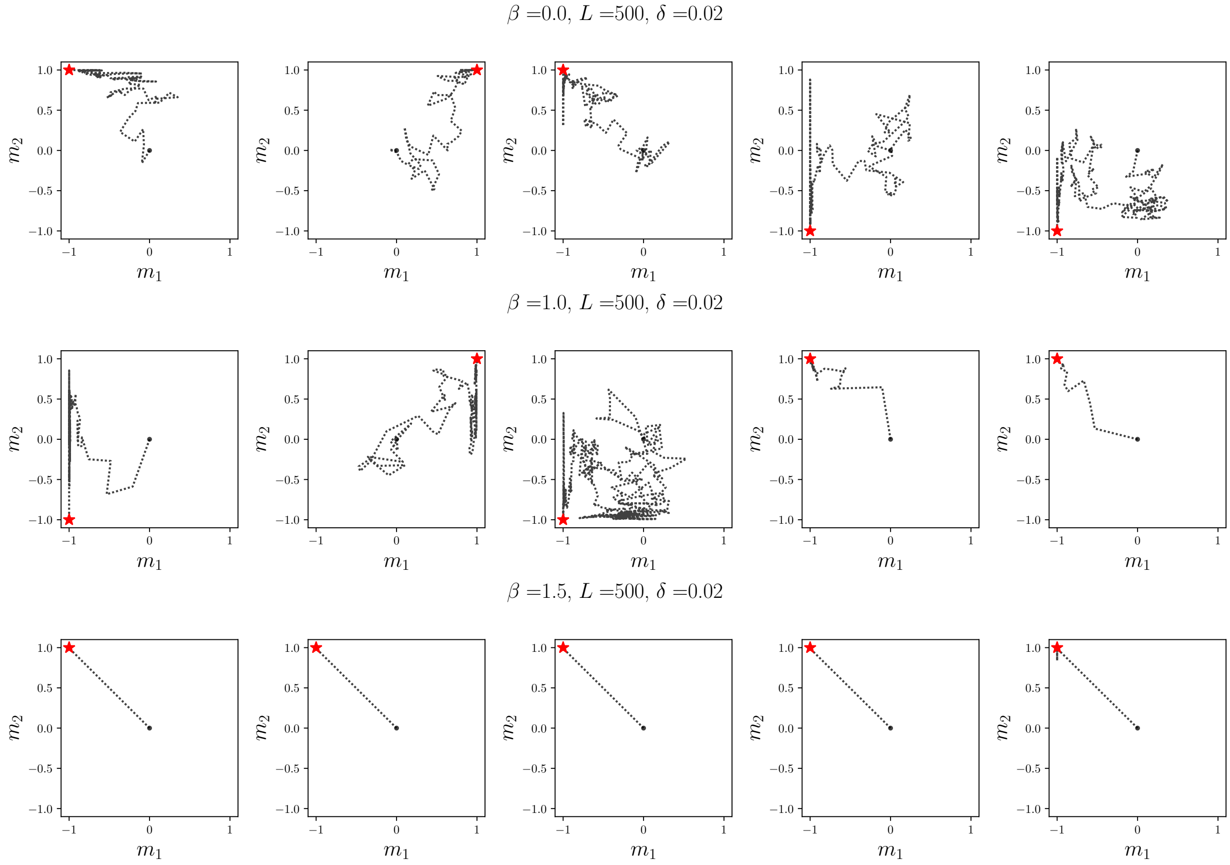


Figure 5.1: Trajectories of the first and second coordinates of the estimated mean vectors computed by Algorithm 3, in the case of $\mathbb{Z}_2$-synchronization. For this experiment we set $n = 1000$, $L = 500$, and $\delta = 0.02$.

In our second experiment, we consider the inner product $|\langle \boldsymbol{\theta}_{\le 10}, \boldsymbol{\theta}^{\mathrm{alg}}_{\le 10}\rangle|$, where $\boldsymbol{\theta}_{\le 10} \in \mathbb{R}^{10}$ is the vector comprising the first ten coordinates of $\boldsymbol{\theta}$ and similarly for $\boldsymbol{\theta}^{\rm alg}_{\le 10} \in \mathbb{R}^{10}$. This inner product takes values $\{-10, -8, \cdots, 8, 10\}$. Let $\boldsymbol{\theta}'$ be a sample from the posterior distribution $\mu_{\boldsymbol{X}}$. If $\boldsymbol{\theta}^{\rm alg}$ has distribution close to the target posterior $\mu_{\boldsymbol{X}}$, we expect the distribution of $|\langle \boldsymbol{\theta}_{\le 10}, \boldsymbol{\theta}^{\rm alg}_{\le 10}\rangle|$ to be close to the one of $|\langle \boldsymbol{\theta}_{\le 10}, \boldsymbol{\theta}'_{\le 10}\rangle|$. Let us emphasize that this is not a consequence of Theorem 5.3.1, since $\boldsymbol{\theta}' \mapsto \langle \boldsymbol{\theta}_{\le 10}, \boldsymbol{\theta}'_{\le 10}\rangle$ is not $O(1/\sqrt{n})$-Lipschitz.

We can compute an asymptotically exact prediction for the distribution of $|\langle \boldsymbol{\theta}_{\le 10}, \boldsymbol{\theta}'_{\le 10}\rangle|$ as follows. For large n, projection of $\boldsymbol{\theta}' \sim \mu_{\boldsymbol{X}}$ onto an $O(1)$ subset of coordinates has approximately independent entries (modulo an overall sign), with marginals given by the AMP estimates [93, 64]. More explicitly, the distribution of $|\langle \boldsymbol{\theta}_{\le 10}, \boldsymbol{\theta}'_{\le 10}\rangle|$ is expected to be approximately the same as the sum of 10 independent Rademacher random variables $Z_1, \dots, Z_{10}$, with $\mathbb{E}\{Z_i\} = \hat{\boldsymbol{m}}_i(\boldsymbol{0}, 0)$. This prediction can be easily evaluated numerically.

In Figure 5.2 we compare the empirical distributions of $|\langle \boldsymbol{\theta}_{\le 10}, \boldsymbol{\theta}^{\rm alg}_{\le 10}\rangle|$ with the the theoretical prediction just described. Here, we take $n = 1000$, $L = 500$, $\delta = 0.02$ and multiple values of β. For each value of β, we draw a single realization $(\boldsymbol{X}, \boldsymbol{\theta})$, and generate 1000 samples via Algorithm 1 for those data. The experimental outcomes match well with the theoretical predictions, witnessing that the algorithm behaves better than what is guaranteed by our theory.

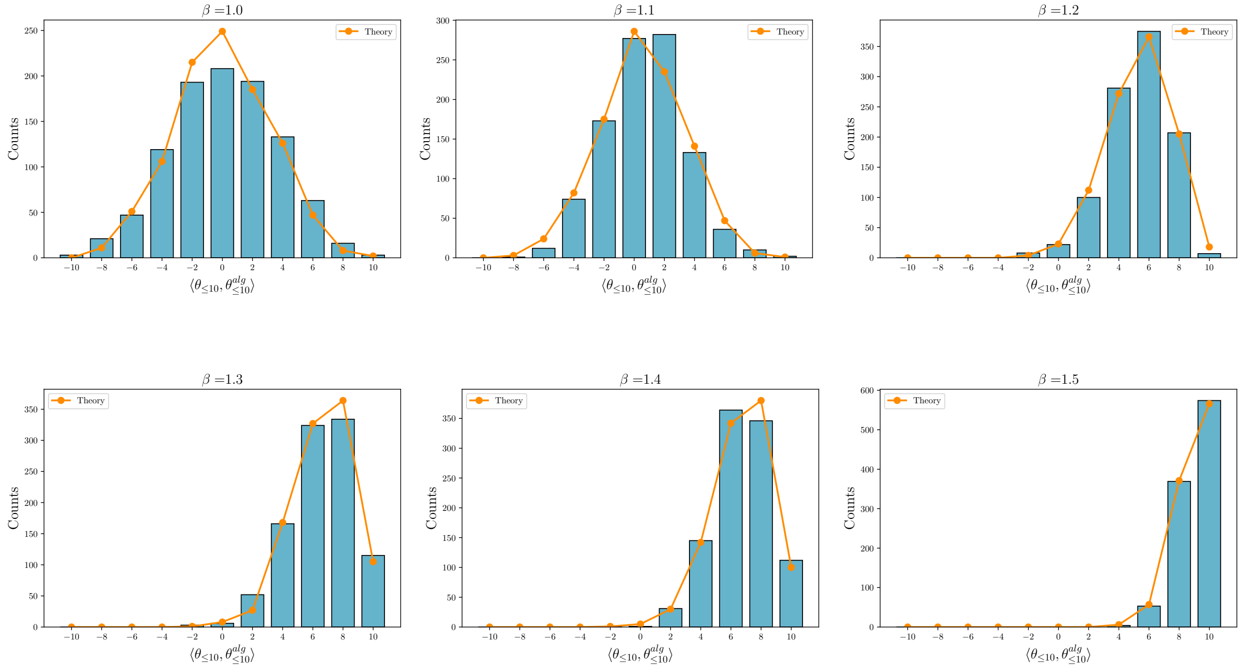


Figure 5.2: Histograms: empirical distributions of $|\langle \boldsymbol{\theta}_{\le 10}, \boldsymbol{\theta}^{\rm alg}_{\le 10}\rangle$ for a samples generated by Algorithm 1 for a single realization of the data $(\boldsymbol{X}, \boldsymbol{\theta})$ at each value of β. Continuous line: theoretical prediction approximating the distribution of $|\langle \boldsymbol{\theta}_{\le 10}, \boldsymbol{\theta}^{\rm alg}_{\le 10}\rangle|$ with the true posterior.

We just mentioned that, under the posterior, the joint distribution of a small subset of the coordinates of $\boldsymbol{\theta}$ is expected to be well approximated by a product form. Let us emphasize that this does not mean that the distribution of the whole vector $\boldsymbol{\theta}$ is close in $W_{2,n}$ to a product distribution.

In order to highlight the nontrivial correlations in $\mu_{\boldsymbol{X}}^{\rm alg}$, we consider the normalized log-likelihood:

$$\mathcal{L}(\boldsymbol{X}, \boldsymbol{\theta}^{\rm alg}) := \frac{\beta}{2n}\langle \boldsymbol{\theta}^{\rm alg}, \boldsymbol{X}\boldsymbol{\theta}^{\rm alg}\rangle .$$

As $T \to \infty$, our theory implies that the distribution of $\boldsymbol{\theta}^{\rm alg}$ is close to the true posterior. For $\boldsymbol{\theta} \sim \mu_{\boldsymbol{X}}(\,\cdot\,)$, we have

$$\begin{aligned} \operatorname*{p-lim}_{n\to\infty} \mathcal{L}(\boldsymbol{X}, \boldsymbol{\theta}) &= \lim_{n\to\infty} \mathbb{E}\left\{ \int \mathcal{L}(\boldsymbol{X}, \boldsymbol{\theta}) \mu_{\boldsymbol{X}}(\mathrm{d}\boldsymbol{\theta}) \right\} \\ &= \lim_{n\to\infty} \frac{\beta}{2n} \mathbb{E}\left\{ \mathbb{E}\{\langle \boldsymbol{\theta}, \boldsymbol{X}\boldsymbol{\theta}\rangle | \boldsymbol{\theta}\} \right\} \\ &= \frac{\beta^2}{2}\,. \end{aligned} \tag{5.38}$$

Note that the function $\boldsymbol{\theta} \mapsto \mathcal{L}(\boldsymbol{X}, \boldsymbol{\theta})$ is Lipschitz (over the domain of interest $\|\boldsymbol{\theta}\|_2 \le \sqrt{n}$) with Lipschitz constant $(\beta/n) \sup_{\|\boldsymbol{\theta}\|_2 \le \sqrt{n}} \|\boldsymbol{X}\boldsymbol{\theta}\|_2 \le C(\beta)/\sqrt{n}$ (with high probability with respect to the choice of $\boldsymbol{X}$). Hence Theorem 5.3.1 implies that Algorithm 1 will produce samples with the correct expectation for the value of this function.

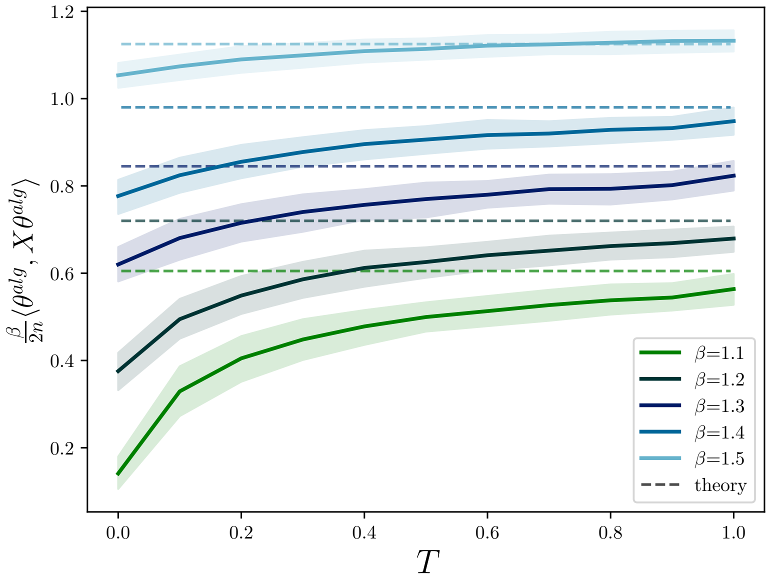


Figure 5.3: Bands: normalized log-likelihood achieved by Algorithm 3. Dashed lines: theoretical predictions.

The simple calculation in the last display also shows that the posterior $\mu_{\boldsymbol{X}}$ cannot be approximated in $W_{2,n}$ by a product measure. Indeed, it is possible to show that, at least for certain values of β,

$$\operatorname*{p-lim}_{n\to\infty} \frac{\beta}{2n}\langle \boldsymbol{m}(\boldsymbol{0}, 0), \boldsymbol{X}\boldsymbol{m}(\boldsymbol{0}, 0)\rangle < \frac{\beta^2}{2}\,, \tag{5.39}$$

with strict inequality[3].

[3] Consider for instance $\beta = 1 + \varepsilon$. Then, the results of [64] imply $\operatorname{p-lim}_{n\to\infty} \|\boldsymbol{m}(\boldsymbol{0}, 0)\|_2^2/n \le C\varepsilon$, and therefore

In Figure 5.3 we compare the theoretical prediction of Eq. (5.38) with numerical results for $\mathcal{L}(\boldsymbol{X}, \boldsymbol{\theta}^{\rm alg})$. We fix $\delta = 0.01$, $n = 1000$ and consider several values of the signal-to-noise ratio β and the number of steps L. For each value of the parameters, we repeat the experiment independently for 300 times, and display our results. The bands represent the 10% and 90% quantiles. From the figure, we see that the likelihood increases with T, which agrees with the expectation that at larger T, $\mu_{\boldsymbol{X}}^{\rm alg}$ matches better with the actual posterior. The agreement with the theoretical prediction is again excellent.

5.6 Proof of Theorem 5.3.1

This section is devoted to analyzing Algorithm 1. In particular, we will outline the proof of Theorem 5.3.1, while delaying most of the technical details to the appendices.

5.6.1 The tilted measure

As discussed in the introduction, and further explained in Section 5.4, after ℓ steps, the state of the algorithm is given by vectors $\hat{\boldsymbol{y}}_\ell$ and $\hat{\boldsymbol{m}}(\hat{\boldsymbol{y}}_\ell, \ell\delta)$. In particular, $\hat{\boldsymbol{m}}(\hat{\boldsymbol{y}}_\ell, \ell\delta)$ is interpreted as an estimate the posterior mean of $\boldsymbol{\theta}$ given $\boldsymbol{y}(t) = \hat{\boldsymbol{y}}_\ell$.

This interpretation has to be slightly modified if π_Θ is symmetric. To be explicit, for any $\boldsymbol{y} \in \mathbb{R}^n$ and $t > 0$, we define the 'tilted measure' $\mu_{\boldsymbol{X},\boldsymbol{y},t}$ as follows (in the formulas below $Z(\boldsymbol{X}, \boldsymbol{y}, t)$ are normalization constants defined by $\int \mu_{\boldsymbol{X},\boldsymbol{y},t}({\rm d}\boldsymbol{\theta}) = 1$):

- **If π_Θ is not a symmetric distribution,** then $\mu_{\boldsymbol{X},\boldsymbol{y},t}$ is the posterior distribution of $\boldsymbol{\theta}$ given $\boldsymbol{X}$ and $\boldsymbol{y}(t) = \boldsymbol{y}$:

$$\mu_{\boldsymbol{X},\boldsymbol{y},t}({\rm d}\boldsymbol{\theta}) := \frac{1}{Z(\boldsymbol{X}, \boldsymbol{y}, t)} \exp\left\{\frac{\beta}{2}\langle\boldsymbol{\theta}, \boldsymbol{X}\boldsymbol{\theta}\rangle - \frac{\beta^2}{4n}\|\boldsymbol{\theta}\|_2^4 + \langle\boldsymbol{y}, \boldsymbol{\theta}\rangle - \frac{t}{2}\|\boldsymbol{\theta}\|_2^2\right\} \pi_\Theta^{\otimes n}({\rm d}\boldsymbol{\theta}) \, .$$

- **If π_Θ is a symmetric distribution,** we let $\boldsymbol{v}_1(\boldsymbol{X})$ be a uniformly[4] random leading eigenvector of $\boldsymbol{X}$. Then we break the symmetry by conditioning on the sign of $\langle\boldsymbol{\theta}, \boldsymbol{v}_1(\boldsymbol{X})\rangle$ as well. Namely,

$$\mu_{\boldsymbol{X},\boldsymbol{y},t}({\rm d}\boldsymbol{\theta}) := \frac{1}{Z(\boldsymbol{X}, \boldsymbol{y}, t)} \exp\left\{\frac{\beta}{2}\langle\boldsymbol{\theta}, \boldsymbol{X}\boldsymbol{\theta}\rangle - \frac{\beta^2}{4n}\|\boldsymbol{\theta}\|_2^4 + \langle\boldsymbol{y}, \boldsymbol{\theta}\rangle - \frac{t}{2}\|\boldsymbol{\theta}\|_2^2\right\} \mathbb{1}_{\langle\boldsymbol{\theta}, \boldsymbol{v}_1(\boldsymbol{X})\rangle \geq 0} \, \pi_\Theta^{\otimes n}({\rm d}\boldsymbol{\theta}).$$

In the symmetric case, the relation to the posterior distribution is given by $\mathbb{P}(\boldsymbol{\theta} \in A | \boldsymbol{X}, \boldsymbol{y}(t) = \boldsymbol{y}) = (1/2)\mu_{\boldsymbol{X},\boldsymbol{y},t}(A) + (1/2)\mu_{\boldsymbol{X},\boldsymbol{y},t}(-A)$. Of course, if we can approximately sample $\boldsymbol{\theta} \sim \mu_{\boldsymbol{X},\boldsymbol{0},0}$, then we can sample from $\boldsymbol{\theta} \sim \mathbb{P}(\boldsymbol{\theta} \in \cdot | \boldsymbol{X})$ with same approximation guarantees in W_2 distance. Therefore, it is sufficient to generate $\boldsymbol{\theta} \sim \mu_{\boldsymbol{X},\boldsymbol{0},0}$ and then flip its sign with probability 1/2, which is what we do in Algorithm 1. Hereafter we will focus on $\mu_{\boldsymbol{X},\boldsymbol{y},t}$.

Throughout the proof, we use $\boldsymbol{m}(\boldsymbol{y}, t)$ to denote the mean of the tilted measure $\boldsymbol{m}(\boldsymbol{y}, t) := \int \boldsymbol{\theta} \, \mu_{\boldsymbol{X},\boldsymbol{y},t}({\rm d}\boldsymbol{\theta})$.

$\text{p-lim}_{n\to\infty}(\beta/2n)\langle\boldsymbol{m}(\boldsymbol{0}, 0), \boldsymbol{X}\boldsymbol{m}(\boldsymbol{0}, 0)\rangle \leq C\varepsilon$.

[4] Almost surely, the leading eigenvalue of $\boldsymbol{X}$ is non-degenerate, and therefore there are two choices of $\{+\boldsymbol{v}, -\boldsymbol{v}\}$ for the normalized leading eigenvector. We let $\boldsymbol{v}_1(\boldsymbol{X}) \sim \text{Unif}(\{+\boldsymbol{v}, -\boldsymbol{v}\})$ independently of $\boldsymbol{X}$, $\boldsymbol{y}(t)$.

5.6.2 Proof outline

The proof consists in checking the assumptions of Theorem 5.4.1. Namely, in Section 5.6.3 we prove that the AMP estimate is close to the expectation, thus verifying (A1); in Section 5.6.4 we verify the path-regularity assumption (A2); finally, in Section 5.6.5 we establish the Lipschitz continuity of the AMP estimate, verifying assumption (A3).

The proof is completed in Section 5.6.6. Throughout the proof, the prior distribution π_Θ is fixed and hence we will not note the dependency of various quantities on π_Θ. We will on the other hand track dependencies on other quantities by noting them in parentheses. We will use $\hat{\boldsymbol{m}}^k(\boldsymbol{y},t)$ to denote the estimate produced by the AMP algorithm of Eq. (5.12) after k iterations, on input $\boldsymbol{X},\boldsymbol{y}$.

5.6.3 AMP achieves Bayes optimality

The analysis of AMP maes use on the following characterization in terms of state evolution, which is adapted from [152]. Here, we refer to a function $\psi:\mathbb{R}^m\to\mathbb{R}$ as *pseudo-Lipschitz* if $|\psi(\boldsymbol{x}_1)-\psi(\boldsymbol{x}_2)|\le C(1+\|\boldsymbol{x}_1\|_2+\|\boldsymbol{x}_2\|_2)\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2$.

Proposition 5.6.1 ([152])**.** *Consider the Bayes AMP algorithm with spectral initialization, defined in Eq.* (5.12)*, and the state evolution recursion of Eqc.* (5.13)*. Assume π_Θ to be sub-Gaussian. Then, for any fixed k, $t\ge 0$ and any pseudo-Lipschitz test function $\psi:\mathbb{R}^2\to\mathbb{R}$, we have*

$$\operatorname*{p-lim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\psi(\theta_i,z^k_{t,i})=\mathbb{E}\big\{\psi(\Theta,\alpha^k_t\Theta+(\alpha^k_t)^{1/2}G)\big\}\,, \tag{5.40}$$

where expectation is with respect $(\Theta,G)\sim\pi_\Theta\otimes\mathsf{N}(0,1)$.

Building on state evolution, we prove Bayes optimality when the signal-to-noise ratio is above a suitable constant β_0.

Lemma 5.6.1. *Assume there exists $\beta_*<\infty$ depending only on π_Θ, such that Condition 5.3.1 holds for all $\beta>\beta_*$. Then there exists a constant $\beta_0\ge\beta_*$ that depends uniquely on π_Θ, such that the following hold.*

For any $\beta\ge\beta_0$ and $\varepsilon,T>0$ there exists $K(\beta,T,\varepsilon)\in\mathbb{N}$, such that for any $t\le T$:

$$\operatorname*{p-limsup}_{n\to\infty}\frac{1}{\sqrt{n}}\|\boldsymbol{m}(\boldsymbol{y}(t),t)-\hat{\boldsymbol{m}}^{K(\beta,T,\varepsilon)}(\boldsymbol{y}(t),t)\|_2\le\varepsilon,$$

where $\hat{\boldsymbol{m}}^k(\boldsymbol{y},t)=\hat{\boldsymbol{m}}^k_t$ is the output of AMP (5.12) *at time t after k iterations, and $\boldsymbol{m}(\boldsymbol{y},t)$ is the mean vector of the tilted measure as defined in Section 5.6.1.*

The proof of Lemma 5.6.1 is deferred to Appendix D.4.

Remark 5.6.1. Denote by $\mathscr{E}^{(1)}_{\beta,L,\delta,\varepsilon,n}$ the event that AMP returns an accurate approximation of the posterior mean for all $t\in\{0,\delta,\dots,L\delta\}$. Namely, we define by Lemma 5.6.1:

$$\mathscr{E}^{(1)}_{L,\delta,\varepsilon,n}:=\left\{\frac{1}{\sqrt{n}}\big\|\boldsymbol{m}(\boldsymbol{y}(\ell\delta),\ell\delta)-\hat{\boldsymbol{m}}^{K(\beta,T,\varepsilon)}(\boldsymbol{y}(\ell\delta),\ell\delta)\big\|_2\le\varepsilon\ \ \forall\ell\in\{0,1,\cdots,L\}\right\}. \tag{5.41}$$

By Lemma 5.6.1, we have $\mathbb{P}(\mathscr{E}^{(1)}_{L,\delta,\varepsilon,n})=1-o_n(1)$.

5.6.4 Path regularity

We next consider assumption (A2) of Theorem 5.4.1: namely we show that the path $t \mapsto \boldsymbol{m}(\boldsymbol{y}(t), t)$ is sufficiently regular.

Lemma 5.6.2. *Assume there exists $\beta_* < \infty$ such that Condition 5.3.1 holds for all $\beta > \beta_*$. Then there exists a constant $\beta_0 > \beta_*$ that depends only on π_Θ, such that the following holds. For fixed $\beta > \beta_0, T \in \mathbb{R}_{\geq 0}$ there exists a constant $C_{\rm reg} = C_{\rm reg}(\beta) > 1$, such that for all $0 \leq t_1 < t_2$,*

$$\begin{aligned}\operatorname*{p-lim}_{n\to\infty} \sup_{t\in[t_1,t_2]} \frac{1}{n}\big\|\boldsymbol{m}(\boldsymbol{y}(t),t) - \boldsymbol{m}(\boldsymbol{y}(t_1),t_1)\big\|_2^2 &= \operatorname*{p-lim}_{n\to\infty} \frac{1}{n}\big\|\boldsymbol{m}(\boldsymbol{y}(t_1),t_1) - \boldsymbol{m}(\boldsymbol{y}(t_2),t_2)\big\|_2^2 \\ &\leq \frac{C_{\rm reg}}{2} \cdot |t_1 - t_2|.\end{aligned}$$

The proof of Lemma 5.6.2 is deferred to Appendix D.5.

Remark 5.6.2. Define $\mathscr{E}^{(2)}_{\beta,L,\delta,n}$ to be the event

$$\mathscr{E}^{(2)}_{\beta,L,\delta,n} := \left\{ \sup_{t\in[\ell\delta,(\ell+1)\delta]} \frac{1}{\sqrt{n}}\big\|\boldsymbol{m}(\boldsymbol{y}(t),t) - \boldsymbol{m}(\boldsymbol{y}(\ell\delta),\ell\delta)\big\|_2 \leq C_{\rm reg}(\beta,T)\sqrt{\delta} \;\; \forall \ell \leq L \right\}. \tag{5.42}$$

Lemma 5.6.2 implies that $\mathbb{P}(\mathscr{E}^{(2)}_{\beta,L,\delta,n}) = 1 - o_n(1)$.

5.6.5 AMP is Lipschitz continuous

The crucial technical step is to prove that Bayes AMP is Lipschitz continuous in a neighborhood of $\boldsymbol{y}(t)$, thus establishing Assumption (A2) of Theorem 5.4.1. Namely, we will use a change of variables introduced in [49] to prove that Bayes AMP is a contraction in the new variables, for β above a threshold.

In order to define change of variables, for $\gamma > 0$, we define $\Gamma_\gamma, \Psi_\gamma : \mathbb{R} \to \mathbb{R}$ by

$$\begin{aligned}\Gamma_\gamma(h) &:= \int_0^h \mathrm{Var}[\Theta \mid \gamma\Theta + \sqrt{\gamma}G = s]^{1/2}\mathrm{d}s, \\ \Psi_\gamma(p) &:= \mathbb{E}[\Theta \mid \gamma\Theta + \sqrt{\gamma}G = \Gamma_\gamma^{-1}(p)],\end{aligned} \tag{5.43}$$

where $\Theta \sim \pi_\Theta$, $G \sim \mathsf{N}(0,1)$ and $\Theta \perp G$. Note that $\Gamma_\gamma, \Psi_\gamma$ are both strictly increasing. The mappings $\Gamma_\gamma, \Psi_\gamma$ are specifically designed such that if we let $p = \Gamma_\gamma(h)$ and $m = \Psi_\gamma(p)$, then the following factorization equality holds:

$$\mathrm{Var}[\Theta \mid \gamma\Theta + \sqrt{\gamma}G = h]^{1/2} = \frac{\mathrm{d}m}{\mathrm{d}p} = \frac{\mathrm{d}p}{\mathrm{d}h}.$$

One can verify that both Γ_γ and Ψ_γ are strictly increasing. Furthermore, both Γ_γ and Ψ_γ are M_Θ-Lipschitz continuous, where $M_\Theta = \|\pi_\Theta\|_\infty$.

Recall the posterior expectation function F is defined in Eq. (5.11). For $t \in \mathbb{R}_{\geq 0}$, $k \in \mathbb{N}$, we define AMP mapping $T^{(t,k)}_{\mathsf{AMP}} : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ via

$$T^{(t,k)}_{\mathsf{AMP}}(\boldsymbol{m}, \overline{\boldsymbol{m}}, \boldsymbol{y}) := \mathsf{F}(\beta\boldsymbol{X}\boldsymbol{m} + \boldsymbol{y} - b_t^k\overline{\boldsymbol{m}}, \alpha_t^{k+1})\,.$$

The AMP iteration Eq. (5.12) can therefore be rewritten as

$$\hat{\boldsymbol{m}}_t^{k+1} = T_{\mathsf{AMP}}^{(t,k)}(\hat{\boldsymbol{m}}_t^k, \hat{\boldsymbol{m}}_t^{k-1}, \boldsymbol{y}(t)) \,.$$

Let $\hat{\boldsymbol{p}}_t^k := \Psi_{\alpha_t^k}^{-1}(\hat{\boldsymbol{m}}_t^k)$ and define the AMP mapping in $\boldsymbol{p}$-domain by

$$\tilde{T}_{\mathsf{AMP}}^{(t,k)}(\boldsymbol{p}, \overline{\boldsymbol{p}}, \boldsymbol{y}) = \Psi_{\alpha_t^{k+1}}^{-1}(\mathsf{F}(\beta \boldsymbol{X} \Psi_{\alpha_t^k}(\boldsymbol{p}) + \boldsymbol{y} - b_t^k \Psi_{\alpha_t^{k-1}}(\overline{\boldsymbol{p}}), \alpha_t^{k+1})).$$

We immediately see that the vectors $\hat{\boldsymbol{p}}_t^k$ satisfy the recursion

$$\hat{\boldsymbol{p}}_t^{k+1} = \tilde{T}_{\mathsf{AMP}}^{(t,k)}(\hat{\boldsymbol{p}}_t^k, \hat{\boldsymbol{p}}_t^{k-1}, \boldsymbol{y}(t)).$$

Note that the range of Ψ_γ is (a_Θ, b_Θ), where $a_\Theta = \inf \operatorname{supp}(\pi_\Theta)$, $b_\Theta = \sup \operatorname{supp}(\pi_\Theta)$. For $\boldsymbol{m} \in (a_\Theta, b_\Theta)^n$, we define $\boldsymbol{D}_\gamma(\boldsymbol{m}) := \operatorname{diag}\{\operatorname{Var}[\Theta \mid \gamma\Theta + \sqrt{\gamma} G = \Gamma_\gamma^{-1}(\Psi_\gamma^{-1}(\boldsymbol{m}))]^{1/2}\} \in \mathbb{R}^{n \times n}$, where, by convention, the conditional variance operator applies on vectors entrywise.

We compute the Jacobian matrices of the AMP mappings $T_{\mathsf{AMP}}^{(t,k)}$ and $\tilde{T}_{\mathsf{AMP}}^{(t,k)}$:

$$\frac{\mathrm{d}T_{\mathsf{AMP}}^{(t,k)}(\boldsymbol{m}, \boldsymbol{m}^-, \boldsymbol{y})}{\mathrm{d}(\boldsymbol{m}, \boldsymbol{m}^-, \boldsymbol{y})} = \Big(\beta \boldsymbol{D}_{\alpha_t^{k+1}}(\boldsymbol{m}^+)^2 \boldsymbol{X};\ -b_t^k \boldsymbol{D}_{\alpha_t^{k+1}}(\boldsymbol{m}^+)^2;\ \boldsymbol{D}_{\alpha_t^{k+1}}(\boldsymbol{m}^+)^2\Big)\,, \tag{5.44}$$

$$\frac{\mathrm{d}\tilde{T}_{\mathsf{AMP}}^{(t,k)}(\boldsymbol{p}, \overline{\boldsymbol{p}}, \boldsymbol{y})}{\mathrm{d}(\boldsymbol{p}, \overline{\boldsymbol{p}}, \boldsymbol{y})} = \Big(\beta \boldsymbol{D}_{\alpha_t^{k+1}}(\boldsymbol{m}^+) \boldsymbol{X} \boldsymbol{D}_{\alpha_t^k}(\boldsymbol{m});\ -b_t^k \boldsymbol{D}_{\alpha_t^{k+1}}(\boldsymbol{m}^+) \boldsymbol{D}_{\alpha_t^{k-1}}(\boldsymbol{m}^-), \boldsymbol{D}_{\alpha_t^{k+1}}(\boldsymbol{m}^+)\Big)\,. \tag{5.45}$$

(Here it is understood that $\boldsymbol{m}^+ = T_{\mathsf{AMP}}^{(t,k)}(\boldsymbol{m}, \boldsymbol{m}^-, \boldsymbol{y})$, $\boldsymbol{m} = \Psi_{\alpha_t^k}(\boldsymbol{p})$, and $\boldsymbol{m}^- = \Psi_{\alpha_t^{k-1}}(\boldsymbol{p}^-)$.

Roughly speaking, we will show that, if the signal strength β is large, and after large number of iterations k, then most elements $\boldsymbol{D}_{\alpha_t^k}(\hat{\boldsymbol{m}}_k^t)$, $\boldsymbol{D}_{\alpha_t^{k+1}}(\hat{\boldsymbol{m}}_t^{k+1})$, $\boldsymbol{D}_{\alpha_t^{k-1}}(\hat{\boldsymbol{m}}_t^{k-1})$ become small. This in turn will imply that the operator norms of the Jacobian matrices in Eqs. (5.44) and (5.45) are small. Finally, this can be used to prove that the AMP mapping is contractive.

The next two lemmas formalize this argument. In the first lemma, we provide an upper bound on $\|\boldsymbol{D}_{\alpha_{l\delta}^k}(\hat{\boldsymbol{m}}_{l\delta}^k)\|_F^2/n$ with a function of β. In the same lemma, we also show that AMP is with high probability Lipschitz continuous if we allow the Lipschitz constant to depend on (β, π_Θ) and the number of iterations.

Lemma 5.6.3. *Assume there exists $\beta_* < \infty$ such that Condition 5.3.1 holds for all $\beta > \beta_*$ and further assume π_Θ is supported on finitely many points.Let $K(\beta, T, \varepsilon)$ be the constant of Lemma 5.6.1.*

Then there exist constants $\beta_0, C_{\rm conv} > 0$ that depend uniquely on π_Θ, such that the following hold: For all $\beta \ge \beta_0$, there exist $k_0(\beta) \in \mathbb{N}$, $\mathrm{Lip}_0(\beta)$ which are functions of (π_Θ, β) only, such that for all $\varepsilon > 0$, $t \in [0, T]$, the following hold with probability $1 - o_n(1)$:

1. *For all $k_0(\beta) \le k \le K(\beta, T, \varepsilon)$,*

$$\frac{1}{n}\|\boldsymbol{D}_{\alpha_t^k}(\hat{\boldsymbol{m}}^k(\boldsymbol{y}(t), t))\|_F^2 \le C_{\rm conv}^{-1} \exp(-C_{\rm conv}\beta^2), \tag{5.46}$$

$$b_t^k \le C_{\rm conv}^{-1} \exp(-C_{\rm conv}\beta^2)\,. \tag{5.47}$$

2. *For $k \in \{k_0(\beta)-1, k_0(\beta), k_0(\beta)+1\}$:*

$$\sup_{\boldsymbol{y}_1 \neq \boldsymbol{y}_2} \frac{\|\hat{\boldsymbol{m}}^k(\boldsymbol{y}_1, t) - \hat{\boldsymbol{m}}^k(\boldsymbol{y}_2, t)\|_2}{\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2} \leq \mathrm{Lip}_0(\beta)\,, \tag{5.48}$$

$$\sup_{\boldsymbol{y}_1 \neq \boldsymbol{y}_2} \frac{\|\hat{\boldsymbol{p}}^k(\boldsymbol{y}_1, t) - \hat{\boldsymbol{p}}^k(\boldsymbol{y}_2, t)\|_2}{\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2} \leq \mathrm{Lip}_0(\beta)\,, \tag{5.49}$$

where $\hat{\boldsymbol{p}}^k(\boldsymbol{y}, t) := \Psi^{-1}_{\alpha^k_t}(\hat{\boldsymbol{m}}^k(\boldsymbol{y}, t))$.

We postpone the proof of Lemma 5.6.3 to Appendix D.6.

By classical estimates on the norm of spiked random matrices [28], with probability $1 - o_n(1)$ we have $\|\boldsymbol{X}\|_{\mathrm{op}} \leq \beta + \beta^{-1} + 1$. We denote by $\mathscr{E}^{(3)}_{\beta,L,\delta,\varepsilon,n}$ the intersection of this event and the one of Lemma 5.6.3. Namely

$$\begin{aligned}\mathscr{E}^{(3)}_{\beta,L,\delta,\varepsilon,n} :=\Big\{&\text{Eq. (5.46) holds for all } k_0(\beta) \leq k \leq K(\beta, L\delta, \varepsilon) \text{ and all } t/\delta \in \{0\} \cup [L],\\ &\text{Eq. (5.48) and Eq. (5.49) hold for all } k \in \{k_0(\beta), k_0(\beta) \pm 1\} \text{ and all } t/\delta \in \{0\} \cup [L]\ ,\\ &\text{and } \|\boldsymbol{X}\|_{\mathrm{op}} \leq 1 + \beta + \beta^{-1}\Big\}.\end{aligned} \tag{5.50}$$

By the last lemma and a union bound, we have $\mathbb{P}(\mathscr{E}^{(3)}_{\beta,L,\delta,\varepsilon,n}) = 1 - o_n(1)$. In what follows, we will be mainly working on the set $\mathscr{E}^{(1)}_{L,\delta,\varepsilon,n} \cap \mathscr{E}^{(2)}_{\beta,L,\delta,n} \cap \mathscr{E}^{(3)}_{\beta,L,\delta,\varepsilon,n}$, which occurs with probability $1 - o_n(1)$ by the lemmas we establish.

The next lemma from [49] is useful for bounding the operator norms of the Jacobian matrices.

Lemma 5.6.4 (Lemma C.2. in [49])**.** *For $\boldsymbol{t} \in [0,1]^n$ and $\xi > 0$, denote by $S(\boldsymbol{t}, \xi)$ the subset of indices $i \in \{1, \cdots, n\}$ for which $t_i \geq \xi$. Then there exist universal constants $C, C', c > 0$ such that for $\boldsymbol{W} \sim \mathrm{GOE}(n)$, any $\xi > 0$ and $0 < q < 1$,*

$$\mathbb{P}\left(\sup_{\substack{\boldsymbol{t}_1, \boldsymbol{t}_2 \in [0,1]^n:\\ |S(\boldsymbol{t}_1,\xi)| \vee |S(\boldsymbol{t}_2,\xi)| \leq nq}} \|\operatorname{diag}(\boldsymbol{t}_1)\boldsymbol{W}\operatorname{diag}(\boldsymbol{t}_2)\|_{\mathrm{op}} \geq C'(\xi + \sqrt{q\log(e/q)})\right) \leq Ce^{-cqn}. \tag{5.51}$$

Lemmas 5.6.3 and 5.6.4 together imply that AMP is a contraction, whence Lipschitz, in a neighborhood of $\boldsymbol{y}(t)$.

Lemma 5.6.5. *Under the assumptions of Lemma 5.6.3, let $k_0(\beta), K(\beta, T, \varepsilon)$ be as defined there.*

Then there exists $\beta_0 > 0$ that depends uniquely on π_Θ, such that the following hold: For all $\beta \geq \beta_0$, there exists $r(\beta), \mathrm{Lip}_*(\beta) > 0$ depending uniquely on (π_Θ, β) such that, for all $t \in [0, T]$, the following holds with probability $1 - o_n(1)$ for all $k_0(\beta) \leq k \leq K(\beta, T, \varepsilon)$:

$$\sup_{\boldsymbol{y}_1 \neq \boldsymbol{y}_2 \in B^n(\boldsymbol{y}(t), r(\beta))} \frac{\|\hat{\boldsymbol{m}}^k(\boldsymbol{y}_1, t) - \hat{\boldsymbol{m}}^k(\boldsymbol{y}_2, t)\|_2}{\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2} \leq 2\,\mathrm{Lip}_*(\beta)\,. \tag{5.52}$$

(Here $B^n(\boldsymbol{x}_0; r) := \{\boldsymbol{x} \in \mathbb{R}^n :\ \|\boldsymbol{x} - \boldsymbol{x}_0\| \leq r\}$.)

The proof of Lemma 5.6.5 can be found in Appendix D.7.

5.6.6 Completing the proof of Theorem 5.3.1

We are now in position to apply Theorem 5.4.1. First of all notice that in the present case $\boldsymbol{H} = \boldsymbol{I}_n$ and therefore $\boldsymbol{m}(\boldsymbol{y}, t) = \boldsymbol{m}_{\boldsymbol{\theta}}(\boldsymbol{y}, t)$. We set T, ε and δ as follows

$$T = \frac{2}{\xi}\,, \tag{5.53}$$

$$\varepsilon = \frac{r(\beta) \wedge \xi}{8}\, e^{-4\mathrm{Lip}_*(\beta)/\xi}\,, \tag{5.54}$$

$$\sqrt{\delta} = \frac{r(\beta) \wedge \xi}{8C_{\rm reg}(\beta, 2/\xi)}\, e^{-4\mathrm{Lip}_*(\beta)/\xi}\,. \tag{5.55}$$

and set $K_{\mathsf{AMP}} = K(\beta, T, \varepsilon)$, where $K(\beta, T, \varepsilon)$ is defined by Lemma 5.6.1.

We next check that assumptions (A1), (A2), (A3), hold (with $\eta = o_n(1)$):

(A1) By Remark 5.6.1, this assumption holds with $\varepsilon_1 = \varepsilon$.

(A2) By Remark 5.6.2, this assumption holds with $C_1 = C_{\rm reg}(\beta, 2/\xi)$, $\varepsilon_2 = 0$.

(A3) By Lemma 5.6.5, this assumption holds with $C_2 = 2\mathrm{Lip}_0(\beta)$, $r_\ell = r(\beta)$. We need to check the lower bound on r_ℓ that is required by assumption (A3). For that purpose, note that

$$(C_1\sqrt{\delta} + \varepsilon_1 + \varepsilon_2)\frac{e^{C_2 L\delta}}{C_2} \le \big(C_{\rm reg}(\beta, 2/\xi)\sqrt{\delta} + \varepsilon\big)e^{4\mathrm{Lip}_*(\beta)/\xi} \tag{5.56}$$

$$\le \frac{r(\beta)}{2} < r_\ell\,. \tag{5.57}$$

where in the first step we used $L\delta = T = 2/\xi$ and, without loss of generality, $\mathrm{Lip}_*(\beta) \ge 1$. In the second inequality, we used the choices for ε, δ given in Eqs. (5.54), (5.55).

Note that, since $\|\pi_\Theta\|_\infty < \infty$, we then have $\int(\|\boldsymbol{\theta}\|^2/n)^2 \mu_{\boldsymbol{X}}(\mathrm{d}\boldsymbol{\theta}) \le R^4$, where $R > 0$ is a constant depending only on π_Θ.

By applying (5.22) from Theorem 5.4.1, we obtain that, for any $\eta > 0$, the following holds with probability $1 - o_n(1)$ with respect to the choice of $\boldsymbol{X}$

$$W_{2,n}(\mu_{\boldsymbol{X}}, \mu^{\rm alg}_{\boldsymbol{X}}) \le \varepsilon + \big(C_{\rm reg}(\beta, 2/\xi)\sqrt{\delta} + \varepsilon\big)e^{4\mathrm{Lip}_*(\beta)/\xi} + CR\eta^{1/2} + \frac{1}{T} \tag{5.58}$$

$$\le \frac{7\xi}{8} + CR\eta^{1/2} \le \frac{9}{10}\xi\,. \tag{5.59}$$

where $C > 0$ is a numerical constant, and the last inequality follows by choosing a suitably small η.

Appendix A

Low-rank matrix estimation with diverging aspect ratios

A.1 Preliminaries

A.1.1 Further notations and conventions

In this section, we present an incomplete summary of the notations and conventions that will be applied throughout the appendix.

For two sequences of random vectors $\{\boldsymbol{X}_n\}_{n\in\mathbb{N}_+} \subseteq \mathbb{R}^k$ and $\{\boldsymbol{Y}_n\}_{n\in\mathbb{N}_+} \subseteq \mathbb{R}^k$, we say $\boldsymbol{X}_n \overset{P}{\simeq} \boldsymbol{Y}_n$ if and only if $\|\boldsymbol{X}_n - \boldsymbol{Y}_n\| = o_p(1)$. For $n, k \in \mathbb{N}_+$ and matrix $\boldsymbol{X} \in \mathbb{R}^{n\times k}$ with the i-th row denoted by $\boldsymbol{x}_i \in \mathbb{R}^k$, we let $\hat{p}_{\boldsymbol{X}}$ be the empirical distribution of the $\boldsymbol{x}_i$'s:

$$\hat{p}_{\boldsymbol{X}} := \frac{1}{n}\sum_{i=1}^{n} \delta_{\boldsymbol{x}_i} \,,$$

where $\delta_{\boldsymbol{x}_i}$ is the point mass at $\boldsymbol{x}_i$.

A.1.2 Wasserstein distance

For two probability distributions μ_1, μ_2 over $\mathbb{R}^r$, recall that the Wasserstein distance between μ_1 and μ_2 is defined as

$$W_2(\mu_1, \mu_2) := \left(\inf_{\gamma\in\Gamma(\mu_1,\mu_2)} \int_{\mathbb{R}^r\times\mathbb{R}^r} \|x - y\|^2 \mathrm{d}\gamma(x, y)\right)^{1/2}, \tag{A.1}$$

where $\Gamma(\mu_1, \mu_2)$ denotes the collection of all probability distributions over $\mathbb{R}^r \times \mathbb{R}^r$ with marginals μ_1 and μ_2 on the first and last r coordinates, respectively. One observation is that for matrices $\boldsymbol{L}_1, \boldsymbol{L}_2 \in \mathbb{R}^{n\times r}$, the W_2 distance between the empirical distributions of their rows is upper bounded by the Frobenius norm of their difference: $W_2(\hat{p}_{\boldsymbol{L}_1}, \hat{p}_{\boldsymbol{L}_2}) \le \frac{1}{\sqrt{n}}\|\boldsymbol{L}_1 - \boldsymbol{L}_2\|_F$.

A.2 Technical lemmas

Lemma A.2.1. *Let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \in \mathbb{R}^p$ be independent τ^2 sub-Gaussian random vectors, with mean 0 and covariance $\mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^T] = \boldsymbol{\Sigma}$. We define the sample covariance matrix $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^T$. Then for any $s \geq 100$, with probability at least $1 - 2e^{-ps^2/3200}$ we have*

$$\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_{\mathrm{op}} \leq s\tau^2 \sqrt{\frac{p}{n}},$$

provided that $s\sqrt{p/n} \leq 25$.

Lemma A.2.2 (Wedin's theorem [197]). *Let $\boldsymbol{A}_0$, $\boldsymbol{A}_1 \in \mathbb{R}^{m \times n}$ have singular value decomposition (for $a \in \{0, 1\}$)*

$$\boldsymbol{A}_a = \boldsymbol{U}_a \boldsymbol{\Sigma}_a \boldsymbol{V}_a^T,$$

with $\boldsymbol{\Sigma}_a$ containing the singular values of $\boldsymbol{A}_a$ in decreasing order. Furthermore, we let $\boldsymbol{U}_{a,+} \in \mathbb{R}^{m \times k(a)}$, $\boldsymbol{V}_{a,+} \in \mathbb{R}^{n \times k(a)}$, be formed by the first $k(a)$ columns of $\boldsymbol{U}_a$, $\boldsymbol{V}_a$, respectively, such that

$$\boldsymbol{U}_a = [\boldsymbol{U}_{a,+} | \boldsymbol{U}_{a,-}], \qquad \boldsymbol{V}_a = [\boldsymbol{V}_{a,+} | \boldsymbol{V}_{a,-}].$$

Let $\sigma_k(\cdot)$ denote the k-th largest singular value of a matrix. Finally assume $\Delta \equiv \sigma_{k(1)}(\boldsymbol{A}_1) - \sigma_{k(0)+1}(\boldsymbol{A}_0) > 0$. Let $\boldsymbol{P}_a = \boldsymbol{V}_{a,+}\boldsymbol{V}_{a,+}^T$(respectively, $\boldsymbol{Q}_a = \boldsymbol{U}_{a,+}\boldsymbol{U}_{a,+}^T$) denote the projector onto the right singular space (left singular space) corresponding to the top $k(a)$ singular values of $\boldsymbol{A}_a$. Then we have

$$\|(\boldsymbol{I} - \boldsymbol{P}_0)\boldsymbol{P}_1\|_{\mathrm{op}} \leq \frac{1}{\Delta}\left\{\|(\boldsymbol{I} - \boldsymbol{Q}_0)(\boldsymbol{A}_0 - \boldsymbol{A}_1)\boldsymbol{P}_1\|_{\mathrm{op}} \vee \|\boldsymbol{Q}_1(\boldsymbol{A}_0 - \boldsymbol{A}_1)(\boldsymbol{I} - \boldsymbol{P}_0)\|_{\mathrm{op}}\right\}.$$

If instead we have $\Delta \equiv \sigma_{k(0)}(\boldsymbol{A}_0) - \sigma_{k(1)+1}(\boldsymbol{A}_1) > 0$, then

$$\|\boldsymbol{P}_0(\boldsymbol{I} - \boldsymbol{P}_1)\|_{\mathrm{op}} \leq \frac{1}{\Delta}\left\{\|(\boldsymbol{I} - \boldsymbol{Q}_1)(\boldsymbol{A}_0 - \boldsymbol{A}_1)\boldsymbol{P}_0\|_{\mathrm{op}} \vee \|\boldsymbol{Q}_0(\boldsymbol{A}_0 - \boldsymbol{A}_1)(\boldsymbol{I} - \boldsymbol{P}_1)\|_{\mathrm{op}}\right\}.$$

Lemma A.2.3 (Nishimori identity, Proposition 16 in [125]). *Let $(\boldsymbol{X}, \boldsymbol{Y})$ be a couple of random variables on a polish space. Let $k \geq 1$ and let $\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(k)}$ be k i.i.d. samples (given $\boldsymbol{Y}$) from the distribution $\mathbb{P}(\boldsymbol{X} = \cdot \mid \boldsymbol{Y})$, independently of every other random variables. Let us denote $\langle \cdot \rangle$ the expectation with respect to $\mathbb{P}(\boldsymbol{X} = \cdot \mid \boldsymbol{Y})$ and $\mathbb{E}$ the expectation with respect to $(\boldsymbol{X}, \boldsymbol{Y})$. Then for all continuous bounded function f,*

$$\mathbb{E}\langle f(\boldsymbol{Y}, \boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(k)})\rangle = \mathbb{E}\langle f(\boldsymbol{Y}, \boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(k-1)}, \boldsymbol{X})\rangle.$$

Lemma A.2.4. *Let $\{f_n\}_{n \in \mathbb{N}_+}$ be a sequence of convex differentiable functions on $\mathbb{R}$, and $f_n(x) \to f(x)$ for all $x \in \mathbb{R}$. Let $D_f = \{x \in \mathbb{R} : f$ is differentiable at $x\}$, then $f_n'(x) \to f'(x)$ for all $x \in D_f$.*

Lemma A.2.5. *If f and g are two differentiable convex functions, then for any $b > 0$,*

$$|f'(a) - g'(a)| \leq g'(a + b) - g'(a - b) + \frac{d}{b},$$

where $d = |f(a + b) - g(a + b)| + |f(a - b) - g(a - b)| + |f(a) - g(a)|$.

Proof. See [165], Lemma 3.2.

□

Lemma A.2.6. *For $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, we define $\|\boldsymbol{X}\|_1 = \sum_{i\in[n],j\in[d]} |X_{ij}|$. Then for $\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathbb{R}^{n\times d}$, we have $\|\boldsymbol{X}_1\boldsymbol{X}_1^\mathsf{T} - \boldsymbol{X}_2\boldsymbol{X}_2^\mathsf{T}\|_1 \leq n\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|_F(\|\boldsymbol{X}_1\|_F + \|\boldsymbol{X}_2\|_F)$.*

Proof. we let $\boldsymbol{x}_i^1 \in \mathbb{R}^d$ be the i-th row of $\boldsymbol{X}_1$ and we let $\boldsymbol{x}_i^2 \in \mathbb{R}^d$ be the i-th row of $\boldsymbol{X}_2$. Then by triangle inequality,

$$\begin{aligned}
\|\boldsymbol{X}_1\boldsymbol{X}_1^\mathsf{T} - \boldsymbol{X}_2\boldsymbol{X}_2^\mathsf{T}\|_1 \leq & \|\boldsymbol{X}_1(\boldsymbol{X}_1 - \boldsymbol{X}_2)^\mathsf{T}\|_1 + \|\boldsymbol{X}_2(\boldsymbol{X}_1 - \boldsymbol{X}_2)^\mathsf{T}\|_1 \\
\leq & \sum_{i,j\in[n]} (\|\boldsymbol{x}_i^1\|_2 + \|\boldsymbol{x}_i^2\|_2) \times \|\boldsymbol{x}_j^1 - \boldsymbol{x}_j^2\|_2 \\
\leq & \sum_{i\in[n]} \|\boldsymbol{x}_i^1\|_2 \times \sqrt{n\sum_{j\in[n]} \|\boldsymbol{x}_j^1 - \boldsymbol{x}_j^2\|_2^2} + \sum_{i\in[n]} \|\boldsymbol{x}_i^2\|_2 \times \sqrt{n\sum_{j\in[n]} \|\boldsymbol{x}_j^1 - \boldsymbol{x}_j^2\|_2^2} \\
\leq & \sqrt{n\sum_{i\in[n]} \|\boldsymbol{x}_i^1\|_2^2} \times \sqrt{n\sum_{j\in[n]} \|\boldsymbol{x}_j^1 - \boldsymbol{x}_j^2\|_2^2} + \sqrt{n\sum_{i\in[n]} \|\boldsymbol{x}_i^1\|_2^2} \times \sqrt{n\sum_{j\in[n]} \|\boldsymbol{x}_j^1 - \boldsymbol{x}_j^2\|_2^2} \\
= & n\|\boldsymbol{X}_1 - \boldsymbol{X}_2\|_F(\|\boldsymbol{X}_1\|_F + \|\boldsymbol{X}_2\|_F).
\end{aligned}$$

□

A.3 Proofs for the strong signal regime

A.3.1 Proof of Theorem 2.3.1

Proof of claim 1

The top r eigenvectors of $\boldsymbol{A}\boldsymbol{A}^\mathsf{T}$ are also the top r eigenvectors of the following matrix

$$\frac{1}{d}\boldsymbol{A}\boldsymbol{A}^\mathsf{T} - \boldsymbol{I}_n = \frac{\boldsymbol{\Lambda}\boldsymbol{\Theta}^\mathsf{T}\boldsymbol{\Theta}\boldsymbol{\Lambda}^\mathsf{T}}{nd} + \frac{1}{d\sqrt{n}}\left(\boldsymbol{\Lambda}\boldsymbol{\theta}^\mathsf{T}\boldsymbol{Z}^\mathsf{T} + \boldsymbol{Z}\boldsymbol{\theta}\boldsymbol{\Lambda}^\mathsf{T}\right) + \frac{1}{d}\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T} - \boldsymbol{I}_n.$$

We let $\boldsymbol{X} = \left(\boldsymbol{\Lambda}\boldsymbol{\theta}^\mathsf{T}\boldsymbol{Z}^\mathsf{T} + \boldsymbol{Z}\boldsymbol{\theta}\boldsymbol{\Lambda}^\mathsf{T}\right)/d\sqrt{n} + \boldsymbol{Z}\boldsymbol{Z}^\mathsf{T}/d - \boldsymbol{I}_n$. Applying in sequence triangle inequality, Lemma A.2.1 and the law of large numbers, we conclude that there exists a constant $C > 0$, such that with probability $1 - o_n(1)$

$$\begin{aligned}
\|\boldsymbol{X}\|_{\mathrm{op}} \leq & \frac{2}{d\sqrt{n}}\|\boldsymbol{\Lambda}\boldsymbol{\theta}^T\boldsymbol{Z}^T\|_{\mathrm{op}} + \|\frac{1}{d}\boldsymbol{Z}\boldsymbol{Z}^T - \boldsymbol{I}_n\|_{\mathrm{op}} \\
\leq & \frac{2}{d\sqrt{n}}\|\boldsymbol{\Lambda}\|_F\|\boldsymbol{Z}\boldsymbol{\theta}\|_F + \|\frac{1}{d}\boldsymbol{Z}\boldsymbol{Z}^T - \boldsymbol{I}_n\|_{\mathrm{op}} \\
\leq & C\sqrt{\frac{n}{d}}.
\end{aligned} \tag{A.2}$$

Using Lemma A.2.2, we see that there exists another constant $\tilde{C} > 0$, such that with probability $1 - o_n(1)$ we have

$$L^{\sin}(\hat{\boldsymbol{\Lambda}}_s, \boldsymbol{\Lambda}) \leq \tilde{C}\sqrt{\frac{n}{d}}.$$

This completes the proof of the first claim of the theorem since by assumption $d/n \to \infty$.

Proof of claim 2

We write $\boldsymbol{A}\boldsymbol{A}^{\mathsf{T}}/d - \boldsymbol{I}_n = q_\Theta \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}}/n + \boldsymbol{W}$, where $\boldsymbol{W}$ is an $n \times n$ symmetric matrix. Using Eq. (A.2) and the law of large numbers, we find out that $\|\boldsymbol{W}\|_{\mathrm{op}} = o_P(1)$.

We denote the unique eigenvalues of $q_\Theta \boldsymbol{Q}_\Lambda$ by $u_1 > u_2 > \cdots > u_k > 0$, where $k \leq r$. The corresponding geometric multiplicities are denoted by $s_1, \cdots, s_k \in \mathbb{N}_+$. We let δ_l, $\hat{\delta}_l$ be the l-th largest eigenvalues of $q_\Theta \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}}/n$ and $q_\Theta \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}}/n + \boldsymbol{W}$, respectively. We then see immediately that $\delta_l, \hat{\delta}_l \xrightarrow{P} \sum_{i=1}^k u_i \mathbb{1}\{\sum_{j=1}^{i-1} s_j + 1 \leq l \leq \sum_{j=1}^{i} s_j\}$ as $n, d \to \infty$. Let

$$\boldsymbol{D}_i = \mathrm{diag}\left(\delta_{\sum_{j=1}^{i-1} s_j + 1}, \cdots, \delta_{\sum_{j=1}^{i} s_j}\right) \in \mathbb{R}^{s_i \times s_i},$$
$$\hat{\boldsymbol{D}}_i = \mathrm{diag}\left(\hat{\delta}_{\sum_{j=1}^{i-1} s_j + 1}, \cdots, \hat{\delta}_{\sum_{j=1}^{i} s_j}\right) \in \mathbb{R}^{s_i \times s_i}.$$

The above arguments imply that $\boldsymbol{D}_i, \hat{\boldsymbol{D}}_i \xrightarrow{P} u_i \boldsymbol{I}_{s_i}$.

For $i \in [k]$, we define the matrices $\boldsymbol{V}_i, \hat{\boldsymbol{V}}_i \in \mathbb{R}^{n \times s_i}$, such that the columns of $\boldsymbol{V}_i/\sqrt{n}$, $\hat{\boldsymbol{V}}_i/\sqrt{n}$ are the eigenvectors of $q_\Theta \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}}/n$, $q_\Theta \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}}/n + \boldsymbol{W}$ that correspond to the top $\sum_{j=1}^{i-1} s_j + 1$ to $\sum_{j=1}^{i} s_j$ eigenvalues, respectively. By Wedin's theorem (Lemma A.2.2), we see that $L^{\sin}(\boldsymbol{V}_i, \hat{\boldsymbol{V}}_i) \xrightarrow{P} 0$ for all $i \in [k]$. Combining all arguments derived, we conclude that

$$\Big\| \frac{1}{n} \sum_{i=1}^{r} \boldsymbol{V}_i \boldsymbol{D}_i \boldsymbol{V}_i^{\mathsf{T}} - \frac{1}{n} \sum_{i=1}^{r} \hat{\boldsymbol{V}}_i \hat{\boldsymbol{D}}_i \hat{\boldsymbol{V}}_i^{\mathsf{T}} \Big\|_F^2 \xrightarrow{P} 0.$$

Note that $\sum_{i=1}^{r} \boldsymbol{V}_i \boldsymbol{D}_i \boldsymbol{V}_i^{\mathsf{T}}/n = q_\Theta \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}}/n$ and $\hat{\boldsymbol{D}}_i$, $\hat{\boldsymbol{V}}_i$ are functions of $\boldsymbol{A}$, thus we have found an estimator $\hat{\boldsymbol{L}} \in \mathbb{R}^{n \times n}$ such that $\|\hat{\boldsymbol{L}} - \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}}\|_F^2/n^2 = o_P(1)$. Based on this convergence, we only need to apply a standard truncation argument to show the expected mean square error vanishes. We skip the details here for the sake of simplicity.

Proof of claim 3

By claim 2 of the theorem, we see that there exists an estimate $\hat{\boldsymbol{L}}$ of $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}}$ which achieves consistency: $\|\hat{\boldsymbol{L}} - \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}}\|_F/n = o_P(1)$. Let

$$\boldsymbol{R} := \mathrm{argmin}_{\boldsymbol{X} \in \mathbb{R}^{n \times r}} \|\hat{\boldsymbol{L}} - \boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}\|_F.$$

By definition, $\|\hat{\boldsymbol{L}} - \boldsymbol{R}\boldsymbol{R}^{\mathsf{T}}\|_F \leq \|\hat{\boldsymbol{L}} - \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}}\|_F$, thus $\|\hat{\boldsymbol{L}} - \boldsymbol{R}\boldsymbol{R}^{\mathsf{T}}\|_F/n = o_P(1)$. By triangle inequality we see that $\|\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}} - \boldsymbol{R}\boldsymbol{R}^{\mathsf{T}}\|_F/n = o_P(1)$, from which we conclude that there exists $\boldsymbol{\Omega}_0 \in \mathcal{O}(r)$ such that

$$\frac{1}{\sqrt{n}} \|\boldsymbol{R}\boldsymbol{\Omega}_0 - \boldsymbol{\Lambda}\|_F = o_P(1).$$

We define

$$\boldsymbol{\Omega}_* := \mathrm{argmin}_{\boldsymbol{\Omega} \in \mathcal{O}(r)} W_2(\hat{p}_{\boldsymbol{R}\boldsymbol{\Omega}}, \mu_\Lambda), \qquad \boldsymbol{R}_* := \boldsymbol{R}\boldsymbol{\Omega}_*.$$

Denote by $\mu_{\boldsymbol{\Omega}\Lambda}$ the distribution of $\boldsymbol{\Omega}\boldsymbol{\Lambda}_0$ for $\boldsymbol{\Omega} \in \mathcal{O}(r)$ and $\boldsymbol{\Lambda}_0 \sim \mu_\Lambda$. We then have $\|\boldsymbol{\Lambda}\boldsymbol{\Omega}_0^{\mathsf{T}}\boldsymbol{\Omega}_* - \boldsymbol{R}_*\|_F/\sqrt{n} =$

$o_P(1)$, which implies $W_2(\hat{p}_{\boldsymbol{\Lambda}\boldsymbol{\Omega}_0^{\mathsf{T}}\boldsymbol{\Omega}_*}, \hat{p}_{\boldsymbol{R}_*}) = o_P(1)$. Furthermore,

$$\begin{aligned} W_2(\hat{p}_{\boldsymbol{R}_*}, \mu_\Lambda) \leq & W_2(\hat{p}_{\boldsymbol{R}\boldsymbol{\Omega}_0}, \mu_\Lambda) \\ \leq & W_2(\hat{p}_{\boldsymbol{R}\boldsymbol{\Omega}_0}, \hat{p}_{\boldsymbol{\Lambda}}) + W_2(\hat{p}_{\boldsymbol{\Lambda}}, \mu_\Lambda) \\ \leq & \frac{1}{\sqrt{n}} \|\boldsymbol{R}\boldsymbol{\Omega}_0 - \boldsymbol{\Lambda}\|_F + W_2(\hat{p}_{\boldsymbol{\Lambda}}, \mu_\Lambda) = o_P(1). \end{aligned}$$

Invoking triangle inequality, we have

$$\begin{aligned} W_2(\mu_{\boldsymbol{\Omega}_*^{\mathsf{T}}\boldsymbol{\Omega}_0\Lambda}, \mu_\Lambda) \leq & W_2(\mu_\Lambda, \hat{p}_{\boldsymbol{R}_*}) + W_2(\hat{p}_{\boldsymbol{R}_*}, \hat{p}_{\boldsymbol{\Lambda}\boldsymbol{\Omega}_0^{\mathsf{T}}\boldsymbol{\Omega}_*}) + W_2(\hat{p}_{\boldsymbol{\Lambda}\boldsymbol{\Omega}_0^{\mathsf{T}}\boldsymbol{\Omega}_*}, \mu_{\boldsymbol{\Omega}_*^{\mathsf{T}}\boldsymbol{\Omega}_0\Lambda}) \\ = & W_2(\mu_\Lambda, \hat{p}_{\boldsymbol{R}_*}) + W_2(\hat{p}_{\boldsymbol{R}_*}, \hat{p}_{\boldsymbol{\Lambda}\boldsymbol{\Omega}_0^{\mathsf{T}}\boldsymbol{\Omega}_*}) + W_2(\hat{p}_{\boldsymbol{\Lambda}}, \mu_\Lambda). \end{aligned}$$

Combining the above results, we see that $W_2(\mu_{\boldsymbol{\Omega}_*^{\mathsf{T}}\boldsymbol{\Omega}_0\Lambda}, \mu_\Lambda) = o_P(1)$. Notice that the mapping $\boldsymbol{\Omega} \mapsto W_2(\mu_\Lambda, \mu_{\boldsymbol{\Omega}\Lambda})$ is continuous on $\mathcal{O}(r)$, and $W_2(\mu_1, \mu_2) = 0$ if and only if $\mu_1 = \mu_2$. Therefore, by assumption we obtain that $\|\boldsymbol{\Omega}_*^{\mathsf{T}}\boldsymbol{\Omega}_0 - \boldsymbol{I}_r\|_F = o_P(1)$, thus $\|\boldsymbol{R}_* - \boldsymbol{\Lambda}\|_F/\sqrt{n} = o_P(1)$. Notice that $\boldsymbol{R}$ is a function of the observation $\boldsymbol{A}$, thus $\boldsymbol{R}_*$ is a function of $\boldsymbol{A}$ as well. Therefore, we have constructed a consistent estimator for $\boldsymbol{\Lambda}$ under the metric of vector mean square error. The rest parts of the proof again follow from a standard truncation argument.

A.3.2 Proof of Theorem 2.3.2

Proof of claim 1

We first prove Eq. (2.6). Define $\boldsymbol{A}_0 := (\sum_{i=1}^n \boldsymbol{\Lambda}_i \boldsymbol{\Lambda}_i^{\mathsf{T}}/n)^{-1/2} \in S_r^+$. We note that under the assumptions of remark 2.3.1, with high probability $\boldsymbol{A}_0$ is well-defined. For $j \in [d]$, we let $\boldsymbol{B}_0^j := \frac{1}{\sqrt{n}} \boldsymbol{A}_0^2 \sum_{i=1}^n A_{ij} \boldsymbol{\Lambda}_i \in \mathbb{R}^r$. We denote by $\boldsymbol{M}_r$ the set of symmetric invertible matrices in $\mathbb{R}^{r\times r}$. Let $\boldsymbol{G} \sim \mathsf{N}(0, \boldsymbol{I}_r)$, independent of $\boldsymbol{\Theta}_0 \sim \mu_\Theta$. We define the mapping $f_\Theta : \boldsymbol{M}_r \times \mathbb{R}^r \to \mathbb{R}^r$ such that

$$f_\Theta(\boldsymbol{A}, \boldsymbol{B}) := \mathbb{E}\left[\boldsymbol{\Theta}_0 \mid \boldsymbol{\Theta}_0 + \boldsymbol{A}\boldsymbol{G} = \boldsymbol{B}\right] = \frac{\int \boldsymbol{\theta} \exp\left(-\frac{1}{2}\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{A}^{-2}\boldsymbol{\theta} + \boldsymbol{B}^{\mathsf{T}}\boldsymbol{A}^{-2}\boldsymbol{\theta}\right) \mu_\Theta(\mathrm{d}\boldsymbol{\theta})}{\int \exp\left(-\frac{1}{2}\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{A}^{-2}\boldsymbol{\theta} + \boldsymbol{B}^{\mathsf{T}}\boldsymbol{A}^{-2}\boldsymbol{\theta}\right) \mu_\Theta(\mathrm{d}\boldsymbol{\theta})}.$$

Let $\hat{\boldsymbol{\Theta}}_j^B := \mathbb{E}[\boldsymbol{\Theta}_j \mid \boldsymbol{A}, \boldsymbol{\Lambda}] = f_\Theta(\boldsymbol{A}_0, \boldsymbol{B}_0^j)$, then $\hat{\boldsymbol{\Theta}}_j^B$ achieves Bayesian mean square error.

Dominated convergence theorem reveals that $f_\Theta(\cdot,\cdot)$ is continuous. By the law of large numbers and central limit theorem, we see that $(\boldsymbol{A}_0, \boldsymbol{B}_0^j) \overset{d}{\to} (\boldsymbol{Q}_\Lambda^{-1/2}, \boldsymbol{\Theta}_0 + \boldsymbol{Q}_\Lambda^{-1/2}\boldsymbol{G})$ as $n, d \to \infty$. Using Skorokhod's representation theorem, there exist $(\boldsymbol{A}_n, \boldsymbol{B}_n^j)$ and $(\boldsymbol{A}_\infty, \boldsymbol{B}_\infty^j)$ being random vectors defined on the same probability space, such that $(\boldsymbol{A}_n, \boldsymbol{B}_n^j) \overset{a.s.}{\to} (\boldsymbol{A}_\infty, \boldsymbol{B}_\infty^j)$, $(\boldsymbol{A}_n, \boldsymbol{B}_n^j) \overset{d}{=} (\boldsymbol{A}_0, \boldsymbol{B}_0^j)$, and $\boldsymbol{A}_\infty = \boldsymbol{Q}_\Lambda^{-1/2}$, $\boldsymbol{B}_\infty^j \overset{d}{=} \boldsymbol{\Theta}_0 + \boldsymbol{Q}_\Lambda^{-1/2}\boldsymbol{G}$. Therefore, $f_\Theta(\boldsymbol{A}_0, \boldsymbol{B}_0^j) \overset{d}{=} f_\Theta(\boldsymbol{A}_n, \boldsymbol{B}_n^j) \overset{a.s.}{\to} f_\Theta(\boldsymbol{A}_\infty, \boldsymbol{B}_\infty^j)$. Since $\|f_\Theta(\boldsymbol{A}_n, \boldsymbol{B}_n^j)\|^2 \overset{d}{=} \|\mathbb{E}[\boldsymbol{\Theta}_j \mid \boldsymbol{A}, \boldsymbol{\Lambda}]\|^2$, we conclude that the set of random variables $\left\{\|f_\Theta(\boldsymbol{A}_n, \boldsymbol{B}_n^j)\|^2 : n \in \mathbb{N}_+\right\}$ is uniformly integrable. Therefore, we have $\|f_\Theta(\boldsymbol{A}_n, \boldsymbol{B}_n^j)\|^2 \overset{L_1}{\to} \|f_\Theta(\boldsymbol{A}_\infty, \boldsymbol{B}_\infty^j)\|^2$. This further implies that as $n, d \to \infty$

$$\begin{aligned} \mathbb{E}[\|\boldsymbol{\Theta}_j - \hat{\boldsymbol{\Theta}}_j^B\|_F^2] = \mathbb{E}_{\boldsymbol{\Theta}_0\sim\mu_\Theta}[\|\boldsymbol{\Theta}_0\|^2] - \mathbb{E}\left[\|f_\Theta(\boldsymbol{A}_n, \boldsymbol{B}_n^j)\|^2\right] \to & \mathbb{E}[\|\boldsymbol{\Theta}_0\|^2] - \mathbb{E}\left[\|f_\Theta(\boldsymbol{A}_\infty, \boldsymbol{B}_\infty)\|^2\right] \\ = & \mathbb{E}[\|\boldsymbol{\Theta}_0\|^2] - \mathbb{E}\left[\|\mathbb{E}[\boldsymbol{\Theta}_0 | \boldsymbol{Q}_\Lambda^{1/2}\boldsymbol{\Theta}_0 + \boldsymbol{G}]\|^2\right], \end{aligned}$$

thus completing the proof of the first claim.

Proof of claim 2

Next, we prove Eq. (2.7). For $k, j \in [d]$, $k \neq j$, notice that $\mathbf{\Theta}_k$ and $\mathbf{\Theta}_j$ are conditionally independent conditioning on $(\boldsymbol{A}, \mathbf{\Lambda})$. Then we have $\mathbb{E}[\mathbf{\Theta}_k^{\mathsf{T}}\mathbf{\Theta}_j \mid \boldsymbol{A}, \mathbf{\Lambda}] = \mathbb{E}[\mathbf{\Theta}_k \mid \boldsymbol{A}, \mathbf{\Lambda}]^{\mathsf{T}}\mathbb{E}[\mathbf{\Theta}_j \mid \boldsymbol{A}, \mathbf{\Lambda}] = f_\Theta(\boldsymbol{A}_0, \boldsymbol{B}_0^k)^{\mathsf{T}} f_\Theta(\boldsymbol{A}_0, \boldsymbol{B}_0^j)$, thus

$$\mathbb{E}[(\mathbf{\Theta}_k^{\mathsf{T}}\mathbf{\Theta}_j - \mathbb{E}[\mathbf{\Theta}_k^{\mathsf{T}}\mathbf{\Theta}_j \mid \boldsymbol{A}, \mathbf{\Lambda}])^2] = rq_\Theta^2 - \mathbb{E}[(f_\Theta(\boldsymbol{A}_0, \boldsymbol{B}_0^k)^{\mathsf{T}} f_\Theta(\boldsymbol{A}_0, \boldsymbol{B}_0^j))^2].$$

By the law of large numbers and the central limit theorem, we have $(\boldsymbol{A}_0, \boldsymbol{B}_0^j, \boldsymbol{B}_0^k) \xrightarrow{d} (\boldsymbol{Q}_\Lambda^{-1/2}, \mathbf{\Theta}_1 + \boldsymbol{Q}_\Lambda^{-1/2}\boldsymbol{G}_1, \mathbf{\Theta}_2 + \boldsymbol{Q}_\Lambda^{-1/2}\boldsymbol{G}_2)$, where $\mathbf{\Theta}_1, \mathbf{\Theta}_2 \sim \mu_\Theta$, $\boldsymbol{G}_1, \boldsymbol{G}_2 \sim \mathsf{N}(0, \boldsymbol{I}_r)$ are mutually independent. By Skorokhod's representation theorem, there exist $(\boldsymbol{A}_n, \boldsymbol{B}_n^j, \boldsymbol{B}_n^k)$ and $(\boldsymbol{A}_\infty, \boldsymbol{B}_\infty^j, \boldsymbol{B}_\infty^k)$ being random vectors on the same probability space, such that $(\boldsymbol{A}_n, \boldsymbol{B}_n^j, \boldsymbol{B}_n^k) \overset{a.s.}{\to} (\boldsymbol{A}_\infty, \boldsymbol{B}_\infty^j, \boldsymbol{B}_\infty^k)$, $(\boldsymbol{A}_n, \boldsymbol{B}_n^j, \boldsymbol{B}_n^k) \overset{d}{=} (\boldsymbol{A}_0, \boldsymbol{B}_0^j, \boldsymbol{B}_0^k)$, and $\boldsymbol{A}_\infty = \boldsymbol{Q}_\Lambda^{-1/2}$, $(\boldsymbol{B}_\infty^j, \boldsymbol{B}_\infty^k) \overset{d}{=} (\mathbf{\Theta}_1 + \boldsymbol{Q}_\Lambda^{-1/2}\boldsymbol{G}_1, \mathbf{\Theta}_2 + \boldsymbol{Q}_\Lambda^{-1/2}\boldsymbol{G}_2)$. Therefore, as $n, d \to \infty$

$$\begin{aligned}&(f_\Theta(\boldsymbol{A}_0, \boldsymbol{B}_0^k)^{\mathsf{T}} f_\Theta(\boldsymbol{A}_0, \boldsymbol{B}_0^j))^2 \\ \overset{d}{=}&(f_\Theta(\boldsymbol{A}_n, \boldsymbol{B}_n^k)^{\mathsf{T}} f_\Theta(\boldsymbol{A}_n, \boldsymbol{B}_n^j))^2 \overset{a.s.}{\to} (f_\Theta(\boldsymbol{A}_\infty, \boldsymbol{B}_\infty^k)^{\mathsf{T}} f_\Theta(\boldsymbol{A}_\infty, \boldsymbol{B}_\infty^j))^2.\end{aligned}$$

Notice that

$$\begin{aligned}(f_\Theta(\boldsymbol{A}_0, \boldsymbol{B}_0^k)^{\mathsf{T}} f_\Theta(\boldsymbol{A}_0, \boldsymbol{B}_0^j))^2 \leq& \|f_\Theta(\boldsymbol{A}_0, \boldsymbol{B}_0^k)\|^2 \|f_\Theta(\boldsymbol{A}_0, \boldsymbol{B}_0^j)\|^2 \\ \leq& \mathbb{E}[\|\mathbf{\Theta}_k\|^2 \mid \boldsymbol{A}, \mathbf{\Lambda}]\mathbb{E}[\|\mathbf{\Theta}_j\|^2 \mid \boldsymbol{A}, \mathbf{\Lambda}] \\ =& \mathbb{E}[\|\mathbf{\Theta}_k\|^2\|\mathbf{\Theta}_j\|^2 \mid \boldsymbol{A}, \mathbf{\Lambda}].\end{aligned}$$

Therefore, the set of random variables $\{(f_\Theta(\boldsymbol{A}_0, \boldsymbol{B}_0^k)^{\mathsf{T}} f_\Theta(\boldsymbol{A}_0, \boldsymbol{B}_0^j))^2 : n \in \mathbb{N}_+\}$ is uniformly integrable. This further implies that $(f_\Theta(\boldsymbol{A}_0, \boldsymbol{B}_0^k)^{\mathsf{T}} f_\Theta(\boldsymbol{A}_0, \boldsymbol{B}_0^j))^2 \overset{L_1}{\to} (f_\Theta(\boldsymbol{A}_\infty, \boldsymbol{B}_\infty^k)^{\mathsf{T}} f_\Theta(\boldsymbol{A}_\infty, \boldsymbol{B}_\infty^j))^2$, thus as $n, d \to \infty$

$$\mathbb{E}[(\mathbf{\Theta}_k^{\mathsf{T}}\mathbf{\Theta}_j - \mathbb{E}[\mathbf{\Theta}_k^{\mathsf{T}}\mathbf{\Theta}_j \mid \boldsymbol{A}, \mathbf{\Lambda}])^2] \to rq_\Theta^2 - \left\|\mathbb{E}\left[\mathbb{E}[\mathbf{\Theta}_0 \mid \boldsymbol{Q}_\Lambda^{1/2}\mathbf{\Theta}_0 + \boldsymbol{G}]\mathbb{E}[\mathbf{\Theta}_0 \mid \boldsymbol{Q}_\Lambda^{1/2}\mathbf{\Theta}_0 + \boldsymbol{G}]^{\mathsf{T}}\right]\right\|_F^2,$$

which concludes the proof of the second claim of the theorem.

A.3.3 Proof of Theorem 2.3.3

By Theorem 2.3.1 claim 3, we see that there exists estimate $\hat{\mathbf{\Lambda}}$ of $\mathbf{\Lambda}$, such that $\|\mathbf{\Lambda} - \hat{\mathbf{\Lambda}}\|_F/\sqrt{n} \overset{P}{\to} 0$. Notice with high probability $\|\boldsymbol{A}\|_{\rm op} \leq C\sqrt{d}$ for some constant $C > 0$ that depends uniquely on $(\mu_\Lambda, \mu_\Theta)$, we then conclude that

$$\frac{1}{\sqrt{nd}}\|\boldsymbol{A}^{\mathsf{T}}\hat{\mathbf{\Lambda}} - \boldsymbol{A}^{\mathsf{T}}\mathbf{\Lambda}\|_F \leq \frac{\|\boldsymbol{A}\|_{op}}{\sqrt{d}} \cdot \frac{\|\mathbf{\Lambda} - \hat{\mathbf{\Lambda}}\|_F}{\sqrt{n}} = o_P(1).$$

Since $\boldsymbol{Z}$ is independent of $\mathbf{\Lambda}$, we immediately see that there exists $\boldsymbol{g} \in \mathbb{R}^{d\times r}$ that has i.i.d. standard Gaussian entries and is independent of $(\mathbf{\Lambda}, \mathbf{\Theta})$, such that

$$\frac{1}{\sqrt{d}}\left\|\frac{1}{\sqrt{n}}\boldsymbol{A}^{\mathsf{T}}\hat{\mathbf{\Lambda}}\boldsymbol{Q}_\Lambda^{-1/2} - \mathbf{\Theta}\boldsymbol{Q}_\Lambda^{1/2} - \boldsymbol{g}\right\|_F = o_P(1). \tag{A.3}$$

Proof of the first result

We let $\boldsymbol{G} \sim \mathsf{N}(0, \boldsymbol{I}_r)$, $\boldsymbol{\Theta}_0 \sim \mu_\Theta$, independent of each other. Define the mapping $F : \mathbb{R}^r \to \mathbb{R}^r$, such that

$$F(\boldsymbol{y}) := \mathbb{E}[\boldsymbol{\Theta}_0 \mid \boldsymbol{Q}_\Lambda^{1/2}\boldsymbol{\Theta}_0 + \boldsymbol{G} = \boldsymbol{y}].$$

Dominated convergence theorem straightforwardly implies that F is continuous on $\mathbb{R}^r$. Therefore, for any $w \in (0,1)$, we see that there exists a mapping $F^w : \mathbb{R}^r \to \mathbb{R}^r$, such that F^w is Lipschitz continuous. In addition,

$$\mathbb{E}\left[\|F(\boldsymbol{Q}_\Lambda^{1/2}\boldsymbol{\Theta}_0 + \boldsymbol{G}) - F^w(\boldsymbol{Q}_\Lambda^{1/2}\boldsymbol{\Theta}_0 + \boldsymbol{G})\|^2\right] \leq w^2$$

We denote the Lipschitz constant of F^w by $L_w > 0$. Let $\boldsymbol{g}_i \in \mathbb{R}^r$ be the i-th row of $\boldsymbol{g}$. The law of large numbers gives the following convergence:

$$\frac{1}{d}\sum_{i=1}^d \|\boldsymbol{\Theta}_i - F(\boldsymbol{Q}_\Lambda^{1/2}\boldsymbol{\Theta}_i + \boldsymbol{g}_i)\|^2 \xrightarrow{P} \mathbb{E}[\|\boldsymbol{\Theta}_0\|^2] - \mathbb{E}\left[\left\|\mathbb{E}[\boldsymbol{\Theta}_0 \mid \boldsymbol{Q}_\Lambda^{1/2}\boldsymbol{\Theta}_0 + \boldsymbol{G}]\right\|^2\right].$$

Therefore, as $n, d \to \infty$

$$\left|\frac{1}{d}\sum_{i=1}^d \|\boldsymbol{\Theta}_i - F^w(\boldsymbol{Q}_\Lambda^{1/2}\boldsymbol{\Theta}_i + \boldsymbol{g}_i)\|^2 - \mathbb{E}[\|\boldsymbol{\Theta}_0\|^2] + \mathbb{E}\left[\left\|\mathbb{E}[\boldsymbol{\Theta}_0 \mid \boldsymbol{Q}_\Lambda^{1/2}\boldsymbol{\Theta}_0 + \boldsymbol{G}]\right\|^2\right]\right| \leq Cw + o_P(1),$$

where $C > 0$ is a constant depending only on μ_Θ. We denote by $\boldsymbol{v}_i$ the i-th row of $\boldsymbol{A}^\mathsf{T}\hat{\boldsymbol{\Lambda}}\boldsymbol{Q}_\Lambda^{-1/2}/\sqrt{n}$. By assumption we have

$$\frac{1}{d}\sum_{i=1}^d \|F^w(\boldsymbol{Q}_\Lambda^{1/2}\boldsymbol{\Theta}_i + \boldsymbol{g}_i) - F^w(\boldsymbol{v}_i)\|^2 \leq \frac{L_w^2}{d}\sum_{i=1}^d \|\boldsymbol{Q}_\Lambda^{1/2}\boldsymbol{\Theta}_i + \boldsymbol{g}_i - \boldsymbol{v}_i\|^2,$$

which is $o_P(1)$ according to Eq. (A.3). Combining the above analysis, we conclude that for any $w \in (0,1)$, there exists $n_w \in \mathbb{N}_+$, such that for $n \geq n_w$, there exists estimator $\hat{\boldsymbol{\Theta}}_w \in \mathbb{R}^{d\times r}$, such that with probability at least $1 - w$

$$\frac{1}{d}\sum_{i=1}^d \|\boldsymbol{\Theta}_i - F^w(\boldsymbol{v}_i)\|^2 \leq \mathbb{E}[\|\boldsymbol{\Theta}_0\|^2] - \mathbb{E}\left[\left\|\mathbb{E}[\boldsymbol{\Theta}_0 \mid \boldsymbol{Q}_\Lambda^{1/2}\boldsymbol{\Theta}_0 + \boldsymbol{G}]\right\|^2\right] + 2w + o_P(1),$$

Since w is arbitrary, the rest parts of the proof follow from a simple truncation argument.

Proof of the second result

By analyzing the second moment we obtain that

$$\begin{aligned}&\frac{1}{d^2}\sum_{i,j\in[d]} |\boldsymbol{\Theta}_i^\mathsf{T}\boldsymbol{\Theta}_j - F(\boldsymbol{Q}_\Lambda^{1/2}\boldsymbol{\Theta}_i + \boldsymbol{g}_i)^\mathsf{T}F(\boldsymbol{Q}_\Lambda^{1/2}\boldsymbol{\Theta}_j + \boldsymbol{g}_j)|^2 \\ =& rq_\Theta^2 - \left\|\mathbb{E}\left[\mathbb{E}[\boldsymbol{\Theta}_0 \mid \boldsymbol{Q}_\Lambda^{1/2}\boldsymbol{\Theta}_0 + \boldsymbol{G}]\mathbb{E}[\boldsymbol{\Theta}_0 \mid \boldsymbol{Q}_\Lambda^{1/2}\boldsymbol{\Theta}_0 + \boldsymbol{G}]^\mathsf{T}\right]\right\|_F^2 + o_P(1).\end{aligned}$$

Since F is continuous, then for any $w \in (0,1)$, there exists $\tilde{F}^w : \mathbb{R}^{2r} \to \mathbb{R}$ such that $\tilde{F}^w$ is $\tilde{L}_w$-Lipschitz continuous. Furthermore,

$$\mathbb{E}\left[|F(\boldsymbol{Q}_\Lambda^{1/2}\boldsymbol{\Theta}_i + \boldsymbol{g}_i)^\mathsf{T} F(\boldsymbol{Q}_\Lambda^{1/2}\boldsymbol{\Theta}_j + \boldsymbol{g}_j) - \tilde{F}^w(\boldsymbol{Q}_\Lambda^{1/2}\boldsymbol{\Theta}_i + \boldsymbol{g}_i, \boldsymbol{Q}_\Lambda^{1/2}\boldsymbol{\Theta}_j + \boldsymbol{g}_j)|^2\right] \le w^2.$$

Again through analysis of the second moment we have

$$\begin{aligned}&\frac{1}{d^2}\sum_{i,j\in[d]} |\boldsymbol{\Theta}_i^\mathsf{T}\boldsymbol{\Theta}_j - \tilde{F}^w(\boldsymbol{Q}_\Lambda^{1/2}\boldsymbol{\Theta}_i + \boldsymbol{g}_i, \boldsymbol{Q}_\Lambda^{1/2}\boldsymbol{\Theta}_j + \boldsymbol{g}_j)|^2 \\ \le & r q_\Theta^2 - \left\| \mathbb{E}\left[\mathbb{E}[\boldsymbol{\Theta}_0 \mid \boldsymbol{Q}_\Lambda^{1/2}\boldsymbol{\Theta}_0 + \boldsymbol{G}]\mathbb{E}[\boldsymbol{\Theta}_0 \mid \boldsymbol{Q}_\Lambda^{1/2}\boldsymbol{\Theta}_0 + \boldsymbol{G}]^\mathsf{T}\right]\right\|_F^2 + \tilde{C}w + o_P(1),\end{aligned}$$

where $\tilde{C} > 0$ is a constant depending uniquely on μ_Θ. By Lipschitzness we have

$$\begin{aligned}&\frac{1}{d^2}\sum_{i,j\in[d]} |\tilde{F}^w(\boldsymbol{Q}_\Lambda^{1/2}\boldsymbol{\Theta}_i + \boldsymbol{g}_i, \boldsymbol{Q}_\Lambda^{1/2}\boldsymbol{\Theta}_j + \boldsymbol{g}_j) - \tilde{F}^w(\boldsymbol{v}_i, \boldsymbol{v}_j)|^2 \\ \le & \frac{\tilde{L}_w^2}{d^2}\sum_{i,j\in[d]} \left\{\|\boldsymbol{Q}_\Lambda^{1/2}\boldsymbol{\Theta}_i + \boldsymbol{g}_i - \boldsymbol{v}_i\|^2 + \|\boldsymbol{Q}_\Lambda^{1/2}\boldsymbol{\Theta}_j + \boldsymbol{g}_j - \boldsymbol{v}_j\|^2\right\},\end{aligned}$$

which by Eq. (A.3) is $o_P(1)$. Since w is arbitrary, again the claim follows by applying standard truncation argument.

A.4 Proof outlines for the weak signal regime

A.4.1 Proof of Theorem 2.4.2

Assume $\boldsymbol{\Lambda}$ is given, then for any $j \in [d]$, the posterior distribution of $\boldsymbol{\Theta}_j$ given $(\boldsymbol{A}, \boldsymbol{\Lambda})$ can be expressed as

$$p(\mathrm{d}\boldsymbol{\theta}_j | \boldsymbol{\Lambda}, \boldsymbol{A}) \propto \exp\left(-\frac{1}{2\sqrt{nd}}\sum_{i=1}^n \langle \boldsymbol{\Lambda}_i, \boldsymbol{\theta}_j\rangle^2 + \frac{1}{\sqrt[4]{nd}}\sum_{i=1}^n A_{ij}\langle \boldsymbol{\Lambda}_i, \boldsymbol{\theta}_j\rangle^2\right)\mu_\Theta(\mathrm{d}\boldsymbol{\theta}_j).$$

From the above equation we see that the posterior of $\boldsymbol{\Theta}$ given $(\boldsymbol{A}, \boldsymbol{\Lambda})$ is a product distribution over $\mathbb{R}^d$, thus greatly simplifies the analysis. The rest of the proof is similar to that of Theorem 2.3.2, and we skip it for simplicity.

A.4.2 Proof outline of Theorem 2.4.3

In this section we outline the proof of Theorem 2.4.3. We leave the proofs of technical lemmas to Appendix A.5. For the sake of simplicity, here we consider only $r = 1$. We comment that cases with $r \ge 2$ can be proven similarly.

Free energy density

Note that the posterior distributions that correspond to the symmetric and asymmetric models can be expressed as follows:

$$\mathrm{d}\mathbb{P}(\boldsymbol{\Lambda} = \boldsymbol{\lambda} \mid \boldsymbol{Y}) = \frac{e^{H_{s,n}(\boldsymbol{\lambda})}\mu_\Lambda^{\otimes n}(\mathrm{d}\boldsymbol{\lambda})}{\int e^{H_{s,n}(\boldsymbol{\lambda})}\mu_\Lambda^{\otimes n}(\mathrm{d}\boldsymbol{\lambda})},$$
$$\mathrm{d}\mathbb{P}(\boldsymbol{\Lambda} = \boldsymbol{\lambda}, \boldsymbol{\Theta} = \boldsymbol{\theta} \mid \boldsymbol{A}) = \frac{e^{H_n(\boldsymbol{\lambda},\boldsymbol{\theta})}\mu_\Theta^{\otimes d}(\mathrm{d}\boldsymbol{\theta})\mu_\Lambda^{\otimes n}(\mathrm{d}\boldsymbol{\lambda})}{\int e^{H_n(\boldsymbol{\lambda},\boldsymbol{\theta})}\mu_\Theta^{\otimes d}(\mathrm{d}\boldsymbol{\theta})\mu_\Lambda^{\otimes n}(\mathrm{d}\boldsymbol{\lambda})},$$

where $\mu_\Lambda^{\otimes n}(\mathrm{d}\boldsymbol{\lambda})$ ($\mu_\Theta^{\otimes d}(\mathrm{d}\boldsymbol{\theta})$) is the product distribution over $\mathbb{R}^n$ ($\mathbb{R}^d$) with each coordinate having marginal distribution μ_Λ (μ_Θ), and $H_{s,n}, H_n$ are the Hamiltonians that correspond to models (2.9) and (2.8), respectively:

$$\begin{aligned} H_{s,n}(\boldsymbol{\lambda}) &:= \frac{q_\Theta^2}{2n}\langle \boldsymbol{\Lambda}, \boldsymbol{\lambda}\rangle^2 + \frac{q_\Theta}{2}\boldsymbol{\lambda}^\intercal \boldsymbol{W}\boldsymbol{\lambda} - \frac{q_\Theta^2}{4n}\|\boldsymbol{\lambda}\|^4, \\ H_n(\boldsymbol{\lambda}, \boldsymbol{\theta}) &:= \frac{1}{\sqrt{nd}}\langle \boldsymbol{\Lambda}, \boldsymbol{\lambda}\rangle\langle \boldsymbol{\Theta}, \boldsymbol{\theta}\rangle + \frac{1}{\sqrt[4]{nd}}\boldsymbol{\lambda}^\intercal \boldsymbol{Z}\boldsymbol{\theta} - \frac{1}{2\sqrt{nd}}\|\boldsymbol{\lambda}\|^2\|\boldsymbol{\theta}\|^2. \end{aligned} \tag{A.4}$$

Following the terminology of statistical mechanics, the *free energy density* is defined as the expected log-partition function (also known as log normalizing constant):

$$\Psi_n^s := \frac{1}{n}\mathbb{E}\log\int e^{H_{s,n}(\boldsymbol{\lambda})}\mu_\Lambda^{\otimes n}(\mathrm{d}\boldsymbol{\lambda}),$$
$$\Psi_n := \frac{1}{n}\mathbb{E}\log\int e^{H_n(\boldsymbol{\lambda},\boldsymbol{\theta})}\mu_\Lambda^{\otimes n}(\mathrm{d}\boldsymbol{\lambda})\mu_\Theta^{\otimes d}(\mathrm{d}\boldsymbol{\theta}).$$

The lemma below connects free energy densities with the corresponding mutual informations.

Lemma A.4.1. *The following equations hold:*

$$\Psi_n^s = \frac{q_\Theta^2\mathbb{E}[\boldsymbol{\Lambda}_0^2]^2}{4} - \mathrm{I}_n^{\mathrm{symm}}(\mu_\Lambda; q_\Theta) + o_n(1),$$
$$\Psi_n = \frac{q_\Theta^2\mathbb{E}[\boldsymbol{\Lambda}_0^2]^2}{4} - \mathrm{I}_n^{\mathrm{asym}}(\mu_\Lambda, \mu_\Theta) + o_n(1).$$

Proof. By definition, the mutual information that corresponds to the symmetric model can be reformulated as

$$\begin{aligned}\mathrm{I}_n^{\mathrm{symm}}(\mu_\Lambda; q_\Theta) =& \frac{1}{n}\mathbb{E}\left\{\log\frac{\mathrm{d}\mu_\Lambda^{\otimes n}(\boldsymbol{\Lambda})\cdot\exp(H_{s,n}(\boldsymbol{\Lambda}))}{\mathrm{d}\mu_\Lambda^{\otimes n}(\boldsymbol{\Lambda})\cdot\int\exp(H_{s,n}(\boldsymbol{\lambda}))\mathrm{d}\mu_\Lambda^{\otimes n}(\boldsymbol{\lambda})}\right\} \\ =& \frac{q_\Theta^2\mathbb{E}[\boldsymbol{\Lambda}_0^2]^2}{4} - \Psi_n^s.\end{aligned}$$

The asymmetric mutual information is a slightly more complicated, which we write below

$$\begin{aligned}&\mathrm{I}_n^{\mathrm{asym}}(\mu_\Lambda, \mu_\Theta) \\ =& \frac{1}{n}\mathbb{E}\left\{\log\frac{\mathrm{d}\mu_\Lambda^{\otimes n}(\boldsymbol{\Lambda})\cdot\int\exp(H_n(\boldsymbol{\Lambda},\boldsymbol{\theta}))\mathrm{d}\mu_\Theta^{\otimes n}(\boldsymbol{\theta})}{\mathrm{d}\mu_\Lambda^{\otimes n}(\boldsymbol{\Lambda})\cdot\int\exp(H_n(\boldsymbol{\lambda},\boldsymbol{\theta}))\mathrm{d}\mu_\Lambda^{\otimes n}(\boldsymbol{\lambda})\mathrm{d}\mu_\Theta^{\otimes n}(\boldsymbol{\theta})}\right\}\end{aligned}$$

$$
\begin{aligned}
=&\frac{1}{n}\sum_{i=1}^{d}\mathbb{E}\Big\{\log\int\exp\Big(\frac{1}{\sqrt{nd}}\|\boldsymbol{\Lambda}\|^2\boldsymbol{\Theta}_i\boldsymbol{\theta}_i+\frac{1}{\sqrt[4]{nd}}\langle\boldsymbol{Z}_{\cdot i},\boldsymbol{\Lambda}\rangle\boldsymbol{\theta}_i-\frac{1}{2\sqrt{nd}}\|\boldsymbol{\Lambda}\|^2\boldsymbol{\theta}_i^2\Big)\mu_\Theta(\mathrm{d}\boldsymbol{\theta}_i)\Big\}-\Psi_n\\
=&\frac{d}{n}\,\mathbb{E}\left\{\log\int\exp\left(\frac{1}{\sqrt{nd}}\|\boldsymbol{\Lambda}\|^2\boldsymbol{\Theta}_1\boldsymbol{\theta}_1+\frac{1}{\sqrt[4]{nd}}\langle\boldsymbol{Z}_{\cdot 1},\boldsymbol{\Lambda}\rangle\boldsymbol{\theta}_1-\frac{1}{2\sqrt{nd}}\|\boldsymbol{\Lambda}\|^2\boldsymbol{\theta}_1^2\right)\mu_\Theta(\mathrm{d}\boldsymbol{\theta}_1)\right\}-\Psi_n.
\end{aligned}
$$

Define

$$
F(q)=\mathbb{E}\left\{\log\int\exp\left(q\boldsymbol{\Theta}_0\boldsymbol{\theta}+\sqrt{q}\boldsymbol{G}\boldsymbol{\theta}-\frac{1}{2}q\boldsymbol{\theta}^2\right)\mu_\Theta(\mathrm{d}\boldsymbol{\theta})\right\},
$$

where the expectation is taken over $\boldsymbol{\Theta}_0\sim\mu_\Theta,\boldsymbol{G}\sim\mathsf{N}(0,1)$ that are independent of each other. Applying Stein's lemma, we obtain that for $q>0$, $F(q)$ has second order continuous derivatives satisfying

$$
\begin{aligned}
F'(q)=&\frac{1}{2}\mathbb{E}\left\{\mathbb{E}[\boldsymbol{\Theta}_0\mid\sqrt{q}\boldsymbol{\Theta}_0+\boldsymbol{G}]^2\right\},\\
F''(q)=&\mathbb{E}\left\{\frac{1}{2}\mathbb{E}[\boldsymbol{\Theta}_0^2\mid\sqrt{q}\boldsymbol{\Theta}_0+\boldsymbol{G}]^2+\frac{1}{2}\mathbb{E}[\boldsymbol{\Theta}_0\mid\sqrt{q}\boldsymbol{\Theta}_0+\boldsymbol{G}]^4-\mathbb{E}[\boldsymbol{\Theta}_0\mid\sqrt{q}\boldsymbol{\Theta}_0+\boldsymbol{G}]^2\mathbb{E}[\boldsymbol{\Theta}_0^2\mid\sqrt{q}\boldsymbol{\Theta}_0+\boldsymbol{G}]\right\}.
\end{aligned}
$$

Since μ_Θ has mean zero, we conclude that F also has second order continuous derivatives at zero, and $F'(0)=0$, $F''(0)=q_\Theta^2/2$. These arguments imply that

$$
\mathrm{I}_n^{\text{asym}}(\mu_\Lambda,\mu_\Theta)=\frac{d}{n}\mathbb{E}\left\{F\left(\frac{1}{\sqrt{nd}}\|\boldsymbol{\Lambda}\|^2\right)\right\}-\Psi_n=\frac{q_\Theta^2\mathbb{E}[\boldsymbol{\Lambda}_0^2]^2}{4}-\Psi_n+o_n(1),
$$

which concludes the proof of the lemma.

□

From Lemma A.4.1 we see that in order to prove the theorem, it suffices to show that the free energy densities agree asymptotically:

$$
\lim_{n\to\infty}\Psi_n^s=\lim_{n,d\to\infty}\Psi_n. \tag{A.5}
$$

Asymptotic equivalence of free energy densities

We then proceed to prove Eq. (A.5). We will start with the additional constraint that μ_Λ has bounded support. Later in Appendix A.4.4, we show that proofs for general μ_Λ can be reduced to the bounded case.

Assumption A.4.1. *We assume that support$(\mu_\Lambda)\subseteq[-K,K]$, with $K>0$ being some fixed constant that is independent of n,d.*

For $h,s\geq 0$, we define the perturbations

$$
\begin{aligned}
\boldsymbol{Y}'(h)&=\frac{\sqrt{h}}{n}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}}+\boldsymbol{W}',\\
\boldsymbol{x}'(s)&=\sqrt{s}\boldsymbol{\Lambda}+\boldsymbol{g}',
\end{aligned}
$$

where $\boldsymbol{W}'\overset{d}{=}\mathrm{GOE}(n)$ and $\boldsymbol{g}'\overset{d}{=}\mathsf{N}(0,\boldsymbol{I}_n)$. Furthermore, we require that $(\boldsymbol{W}',\boldsymbol{g}',\boldsymbol{\Lambda},\boldsymbol{\Theta},\boldsymbol{Z},\boldsymbol{W})$ are mutually independent. We define the Hamiltonians associated with the perturbations $\boldsymbol{Y}'(h)$ and $\boldsymbol{x}'(s)$ respectively as follows:

$$H_n(\boldsymbol{\lambda}; \boldsymbol{Y}'(h)) := \frac{h}{2n}\langle \boldsymbol{\Lambda}, \boldsymbol{\lambda}\rangle^2 + \frac{\sqrt{h}}{2}\boldsymbol{\lambda}^\mathsf{T}\boldsymbol{W}'\boldsymbol{\lambda} - \frac{h}{4n}\|\boldsymbol{\lambda}\|^4, \tag{A.6}$$

$$H_n(\boldsymbol{\lambda}; \boldsymbol{x}'(s)) := \sqrt{s}\langle \boldsymbol{\lambda}, \boldsymbol{g}'\rangle + s\langle \boldsymbol{\Lambda}, \boldsymbol{\lambda}\rangle - \frac{s}{2}\|\boldsymbol{\lambda}\|^2. \tag{A.7}$$

The posterior distribution of $\boldsymbol{\Lambda}$ given $(\boldsymbol{A}, \boldsymbol{Y}'(h), \boldsymbol{x}'(s))$ can be expressed as

$$\mu(\mathrm{d}\boldsymbol{\lambda}) = \frac{1}{Z_n(h,s)}\mu_\Lambda^{\otimes n}(\mathrm{d}\boldsymbol{\lambda})\int \exp\left(H_n(\boldsymbol{\lambda}, \boldsymbol{\theta}) + H_n(\boldsymbol{\lambda}; \boldsymbol{Y}'(h)) + H_n(\boldsymbol{\lambda}; \boldsymbol{x}'(s))\right)\mu_\Theta^{\otimes d}(\mathrm{d}\boldsymbol{\theta}),$$

where $Z_n(h,s)$ is the normalizing constant:

$$Z_n(h,s) = \int \exp\left(H_n(\boldsymbol{\lambda}, \boldsymbol{\theta}) + H_n(\boldsymbol{\lambda}; \boldsymbol{Y}'(h)) + H_n(\boldsymbol{\lambda}; \boldsymbol{x}'(s))\right)\mu_\Theta^{\otimes d}(\mathrm{d}\boldsymbol{\theta})\mu_\Lambda^{\otimes n}(\mathrm{d}\boldsymbol{\lambda}).$$

Note that $Z_n(h,s)$ is random and depends on $(\boldsymbol{A}, \boldsymbol{Y}'(h), \boldsymbol{x}'(s))$. We define the free energy density that corresponds to observations $(\boldsymbol{A}, \boldsymbol{Y}'(h), \boldsymbol{x}'(s))$ as

$$\Phi_n(h,s) := \frac{1}{n}\mathbb{E}\left[\log Z_n(h,s)\right]. \tag{A.8}$$

The next equation follows from Gaussian integration by parts and Nishimori identity (Lemma A.2.3):

$$\frac{\partial}{\partial h}\Phi_n(h,s) = \frac{1}{4n^2}\mathbb{E}\left[\langle \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\mathsf{T}, \mathbb{E}[\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\mathsf{T} \mid \boldsymbol{A}, \boldsymbol{Y}'(h), \boldsymbol{x}'(s)]\rangle\right]. \tag{A.9}$$

Eq. (A.9) holds for all $h > 0, s \geq 0$, and is directly related to the MMSE in the perturbed model. The rest parts of the proof will be devoted to proving convergence of $\Phi_n(h,s)$ as $n, d \to \infty$.

To this end, we first show that asymptotically speaking, the free energy density depends on μ_Θ only through its second moment. More precisely, we can replace μ_Θ with a Gaussian distribution which has mean zero and variance q_Θ. This vastly simplifies further computation.

Lemma A.4.2. *For $k \in [d]$, we let $P_{\Theta,k}$ be a distribution over $\mathbb{R}^d$ with independent coordinates, such that $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_d) \sim P_{\Theta,k}$ if and only if $\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_k \overset{iid}{\sim} \mu_\Theta$ and $\boldsymbol{\theta}_{k+1}, \cdots, \boldsymbol{\theta}_d \overset{iid}{\sim} \mathsf{N}(0, q_\Theta)$. We define*

$$\Phi_n^{(k)}(h,s) := \frac{1}{n}\mathbb{E}\left[\log\left(\int \exp(H_n(\boldsymbol{\lambda}, \boldsymbol{\theta}) + H_n(\boldsymbol{\lambda}; \boldsymbol{Y}'(h)) + H_n(\boldsymbol{\lambda}; \boldsymbol{x}'(s)))\mu_\Lambda^{\otimes n}(\mathrm{d}\boldsymbol{\lambda})\mathrm{d}P_{\Theta,k}(\boldsymbol{\theta})\right)\right].$$

In the above expression, the expectation is taken over $(\boldsymbol{\Lambda}, \boldsymbol{\Theta}, \boldsymbol{Z}, \boldsymbol{W}', \boldsymbol{g}')$. Notice that by definition $\Phi_n(h,s) = \Phi_n^{(d)}(h,s)$. Then under the conditions of Theorem 2.4.3 and in addition Assumption A.4.1, as $n, d \to \infty$ we have $\Phi_n(h,s) - \Phi_n^{(0)}(h,s) = o_n(1)$ for all fixed $h, s \geq 0$.

Lemma A.4.2 can be proved via a Lindeberg type argument, and we postpone the details to Appendix A.5.1. According to Lemma A.4.2, in order to derive the limiting expression of $\Phi_n(h,s)$, it suffices to compute the limit of $\Phi_n^{(0)}(h,s)$ instead, which can be done via Gaussian integration techniques.

Lemma A.4.3. *For fixed $h, s \geq 0$, we define*

$$\tilde{H}_n(\boldsymbol{\lambda}; \boldsymbol{Y}'(h), \boldsymbol{x}'(s)) := \frac{q_\Theta^2}{2n}\langle \boldsymbol{\Lambda}, \boldsymbol{\lambda}\rangle^2 + \frac{q_\Theta}{2\sqrt{nd}}\|\boldsymbol{Z}^\mathsf{T}\boldsymbol{\lambda}\|^2 - \frac{dq_\Theta}{2\sqrt{nd}}\|\boldsymbol{\lambda}\|^2 - \frac{q_\Theta^2}{4n}\|\boldsymbol{\lambda}\|^4$$

$$+ H_n(\boldsymbol{\lambda}; \boldsymbol{Y}'(h)) + H_n(\boldsymbol{\lambda}; \boldsymbol{x}'(s)),$$
$$\tilde{\Phi}_n(h,s) := \frac{1}{n}\mathbb{E}\Big[\log\Big(\int \exp\Big(\tilde{H}_n(\boldsymbol{\lambda}; \boldsymbol{Y}'(h), \boldsymbol{x}'(s))\Big)\mu_\Lambda^{\otimes n}(\mathrm{d}\boldsymbol{\lambda})\Big)\Big].$$

Then under the conditions of Theorem 2.4.3 and Assumption A.4.1, as $n, d \to \infty$, we have $\tilde{\Phi}_n(h,s) - \Phi_n^{(0)}(h,s) = o_n(1)$.

We defer the proof of Lemma A.4.3 to Appendix A.5.2. Under the asymptotics $n, d \to \infty$, $d/n \to \infty$, according to [13], the matrix $\left(\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T} - d\boldsymbol{I}_n\right)/\sqrt{nd}$ behaves like a GOE(n) matrix. Replacing $\left(\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T} - d\boldsymbol{I}_n\right)/\sqrt{nd}$ with a GOE(n) matrix in the definition of $\tilde{\Phi}_n(h,s)$, we see that this allows us to approximate $\tilde{\Phi}_n(h,s)$ via the free energy density of the symmetric model (2.9). Such heuristics can be made rigorous via the following lemma:

Lemma A.4.4. *Recall that $\boldsymbol{Y}$ is defined in Eq. (2.9). For $h, s \geq 0$, we define the free energy density $\Phi_n^Y(h,s)$ that corresponds to the observations $(\boldsymbol{Y}, \boldsymbol{Y}'(h), \boldsymbol{x}'(s))$ as*

$$\Phi_n^Y(h,s) := \frac{1}{n}\mathbb{E}\left[\log\left(\int \exp\left(H_n^Y(\boldsymbol{\lambda}) + H_n(\boldsymbol{\lambda}; \boldsymbol{Y}'(h)) + H_n(\boldsymbol{\lambda}; \boldsymbol{x}'(s))\right)\mu_\Lambda^{\otimes n}(\mathrm{d}\boldsymbol{\lambda})\right)\right], \tag{A.10}$$
$$H_n^Y(\boldsymbol{\lambda}) := \frac{q_\Theta^2}{2n}\langle \boldsymbol{\Lambda}, \boldsymbol{\lambda}\rangle^2 + \frac{q_\Theta}{2}\boldsymbol{\lambda}^\mathsf{T}\boldsymbol{W}\boldsymbol{\lambda} - \frac{q_\Theta^2}{4n}\|\boldsymbol{\lambda}\|^4.$$

Then under the conditions of Theorem 2.4.3 and Assumption A.4.1, as $n, d \to \infty$, we have $\Phi_n^Y(h,s) - \tilde{\Phi}_n(h,s) = o_n(1)$.

We defer the proof of Lemma A.4.4 to Appendix A.5.3. Combining Lemmas A.4.2 to A.4.4, we conclude that as $n, d \to \infty$, for all fixed $h, s \geq 0$, we have $\Phi_n(h,s) - \Phi_n^Y(h,s) = o_n(1)$. This relates the asymmetric model to the symmetric model through their free energy densities. The following lemma summarizes this result and lists several additional useful properties for future reference.

Lemma A.4.5. *Under the conditions of Theorem 2.4.3 and Assumption A.4.1, for all fixed $h, s \geq 0$, the following claims hold:*

1. *As $n, d \to \infty$, we have $\Phi_n(h,s) = \Phi_n^Y(h,s) + o_n(1)$.*
2. *The following mappings $x \mapsto \Phi_n(h,x)$, $x \mapsto \Phi_n(x,s)$, $x \mapsto \Phi_n^Y(h,x)$, $x \mapsto \Phi_n^Y(x,s)$ are all convex on $[0,\infty)$ and differentiable on $(0,\infty)$.*
3. $\lim_{n\to\infty}\Phi_n^Y(0,s)$ *exists for all*

$$(q_\Theta^2, s) \in \{(tx, (1-t)xq^*(x)) : x \geq 0, q^*(x) \text{ exists and is unique}, t \in [0,1]\},$$

where $\mathcal{F}(\cdot,\cdot)$ is defined in Eq. (2.12) and

$$q^*(x) := \mathrm{argmax}_{q\geq 0}\,\mathcal{F}(x,q). \tag{A.11}$$

Remark A.4.1. By [125, Proposition 17], $q^*(x)$ exists and is unique for all but countably many $x > 0$. Claims 2 and 3 do not rely on Assumption A.4.1.

We delay the proof of Lemma A.4.5 to Appendix A.5.4. We note that Theorem 2.4.3 is an immediate consequence of Lemma A.4.5.

A.4.3 Proof of Theorem 2.4.4

In this section, we will apply Lemma A.4.5 to prove Theorem 2.4.4. Using Lemma A.2.3 and Gaussian integration by parts, for all $h > 0$ we have

$$\begin{aligned} \frac{\partial}{\partial h}\Phi_n(h,0) &= \frac{1}{4n^2}\mathbb{E}\left[\langle \mathbf{\Lambda}\mathbf{\Lambda}^\mathsf{T}, \mathbb{E}[\mathbf{\Lambda}\mathbf{\Lambda}^\mathsf{T} \mid \boldsymbol{A}, \boldsymbol{Y}'(h)]\right], \\ \frac{\partial}{\partial h}\Phi_n^Y(h,0) &= \frac{1}{4n^2}\mathbb{E}\left[\langle \mathbf{\Lambda}\mathbf{\Lambda}^\mathsf{T}, \mathbb{E}[\mathbf{\Lambda}\mathbf{\Lambda}^\mathsf{T} \mid \boldsymbol{Y}, \boldsymbol{Y}'(h)]\rangle\right]. \end{aligned} \tag{A.12}$$

Recall that $\mathcal{F}$ is defined in Eq. (2.12). We let

$$D := \{s > 0 \mid \mathcal{F}(s,\cdot) \text{ has a unique maximizer } q^*(s)\}. \tag{A.13}$$

By proposition 17 in [125], D is equal to $(0, +\infty)$ minus a countable set, and is precisely the set of $s > 0$ at which the function $\phi : s \mapsto \sup_{q\geq 0}\mathcal{F}(s,q)$ is differentiable. Furthermore, by [125, Theorem 13], for all $h \geq 0$,

$$\lim_{n\to\infty}\Phi_n^Y(h,0) = \sup_{q\geq 0}\mathcal{F}(q_\Theta^2 + h, q). \tag{A.14}$$

By the first claim of Lemma A.4.5, $\Phi_n(h,0) = \Phi_n^Y(h,0) + o_n(1)$, thus $\lim_{n\to\infty}\Phi_n(h,0) = \sup_{q\geq 0}\mathcal{F}(q_\Theta^2 + h, q)$. By the second claim of Lemma A.4.5, the mappings $h \mapsto \Phi_n(h,0)$, $h \mapsto \Phi_n^Y(h,0)$ are convex and differentiable on $(0,\infty)$. Next, we apply Lemma A.2.4 to function sequences $\{h \mapsto \Phi_n(h,0)\}_{n\geq 1}$, $\{h \mapsto \Phi_n^Y(h,0)\}_{n\geq 1}$, and conclude that for all but countably many values of $h > 0$,

$$\lim_{n\to\infty}\frac{\partial}{\partial h}\Phi_n(h,0) = \lim_{n\to\infty}\frac{\partial}{\partial h}\Phi_n^Y(h,0) = \phi'(h + q_\Theta^2). \tag{A.15}$$

From the above equation we see that the mapping $\lambda \mapsto \phi'(\lambda)$ is non-decreasing on D. Therefore, for all but countably many $q_\Theta > 0$, ϕ' is continuous at $q_\Theta^2 \in D$. For these q_Θ, we immediately see that for any $\varepsilon > 0$, there exists $h_\varepsilon > 0$ depending uniquely on $(q_\Theta, \varepsilon, \mu_\Lambda)$, such that $\phi'(q_\Theta^2 + h_\varepsilon) \leq \phi'(q_\Theta^2) + \varepsilon$, and ϕ is differentiable at $q_\Theta^2 + h_\varepsilon$. According to Eq. (A.15), there exists $n_\varepsilon \in \mathbb{N}_+$, such that for all $n \geq n_\varepsilon$,

$$\left|\frac{\partial}{\partial h}\Phi_n(h_\varepsilon,0) - \phi'(h_\varepsilon + q_\Theta^2)\right| \leq \varepsilon.$$

According to Eq. (A.12),

$$\begin{aligned} \frac{\partial}{\partial h}\Phi_n(h,0) \geq& \frac{1}{4n^2}\mathbb{E}\left[\langle \mathbf{\Lambda}\mathbf{\Lambda}^\mathsf{T}, \mathbb{E}[\mathbf{\Lambda}\mathbf{\Lambda}^\mathsf{T} \mid \boldsymbol{A}]\rangle\right] \\ =& \frac{1}{4n^2}\left(n\mathbb{E}_{\mathbf{\Lambda}_0\sim\mu_\Lambda}[\mathbf{\Lambda}_0^4] + n(n-1)\mathbb{E}_{\mathbf{\Lambda}_0\sim\mu_\Lambda}[\mathbf{\Lambda}_0^2]^2 - n^2\mathrm{MMSE}_n^{\text{asym}}(\mu_\Lambda, \mu_\Theta)\right). \end{aligned}$$

Invoking Proposition 17 and Corollary 18 from [125], for all $q_\Theta^2 \in D$

$$\phi'(q_\Theta^2) = \frac{1}{4}q^*(q_\Theta^2)^2 = \frac{1}{4}\mathbb{E}_{\mathbf{\Lambda}_0\sim\mu_\Lambda}[\mathbf{\Lambda}_0^2]^2 - \frac{1}{4}\lim_{n\to\infty}\mathrm{MMSE}_n^{\text{symm}}(\mu_\Lambda; q_\Theta) + o_n(1).$$

Combining all arguments above, we obtain that

$$\liminf_{n,d\to\infty} \mathrm{MMSE}_n^{\mathrm{asym}}(\mu_\Lambda,\mu_\Theta) \geq \lim_{n\to\infty} \mathrm{MMSE}_n^{\mathrm{symm}}(\mu_\Lambda;q_\Theta) - 8\varepsilon.$$

Since ε is arbitrary, we then complete the proof of the first claim of the theorem.

We then proceed to prove the second claim. For $q_\Theta^2, q_\Theta^2+\eta \in D$ satisfying $0<\eta<\varepsilon$, by Eqs. (A.12) and (A.15) we have

$$\begin{aligned}\lim_{n,d\to\infty} \mathrm{MMSE}_n^{\mathrm{asym}}(\mu_\Lambda,\mu_\Theta;\eta) =& \mathbb{E}_{\boldsymbol{\Lambda}_0\sim\mu_\Lambda}[\boldsymbol{\Lambda}_0^2]^2 - 4\phi'(q_\Theta^2+\eta)\\ \leq& \mathbb{E}_{\boldsymbol{\Lambda}_0\sim\mu_\Lambda}[\boldsymbol{\Lambda}_0^2]^2 - 4\phi'(q_\Theta^2)\\ =& \lim_{n\to\infty} \mathrm{MMSE}_n^{\mathrm{symm}}(\mu_\Lambda;q_\Theta).\end{aligned}$$

Note that $\lim_{n,d\to\infty}\mathrm{MMSE}_n^{\mathrm{asym}}(\mu_\Lambda,\mu_\Theta;\eta) \geq \limsup_{n,d\to\infty}\mathrm{MMSE}_n^{\mathrm{asym}}(\mu_\Lambda,\mu_\Theta;\varepsilon)$, the proof of the second claim immediately follows.

A.4.4 Reduction to bounded prior

In this section, we show that in order to prove Theorem 2.4.3 and Theorem 2.4.4, it suffices to prove the theorems under Assumption A.4.1.

Since μ_Λ is sub-Gaussian, for any $\varepsilon>0$, there exists $K_\varepsilon>0$, such that if we let $\bar{\boldsymbol{\Lambda}}_0 := \boldsymbol{\Lambda}_0 \mathbb{1}\{|\boldsymbol{\Lambda}_0| \leq K_\varepsilon\}$, then $\mathbb{E}_{\boldsymbol{\Lambda}_0\sim\mu_\Lambda}[(\boldsymbol{\Lambda}_0-\bar{\boldsymbol{\Lambda}}_0)^4] < \varepsilon$ and $\mu_\Lambda([-K_\varepsilon,K_\varepsilon]) > 1-\varepsilon^2$. For all $i\in[n]$, we define $\bar{\boldsymbol{\Lambda}}_i := \boldsymbol{\Lambda}_i\mathbb{1}\{|\boldsymbol{\Lambda}_i|\leq K_\varepsilon\}$ and $\bar{\boldsymbol{\lambda}}_i := \boldsymbol{\lambda}_i\mathbb{1}\{|\boldsymbol{\lambda}_i|\leq K_\varepsilon\}$. Let $\bar{\boldsymbol{\Lambda}} = (\bar{\boldsymbol{\Lambda}}_i)_{i\leq n}\in\mathbb{R}^n$ and $\bar{\boldsymbol{\lambda}} = (\bar{\boldsymbol{\lambda}}_i)_{i\leq n}\in\mathbb{R}^n$. We introduce the truncated Hamiltonians:

$$\begin{aligned}\bar{H}_n^{\varepsilon}(\bar{\boldsymbol{\lambda}},\boldsymbol{\theta}) &:= \frac{1}{\sqrt{nd}}\langle\bar{\boldsymbol{\Lambda}},\bar{\boldsymbol{\lambda}}\rangle\langle\boldsymbol{\Theta},\boldsymbol{\theta}\rangle + \frac{1}{\sqrt[4]{nd}}\bar{\boldsymbol{\lambda}}^\mathsf{T}\boldsymbol{Z}\boldsymbol{\theta} - \frac{1}{2\sqrt{nd}}\|\bar{\boldsymbol{\lambda}}\|^2\|\boldsymbol{\theta}\|^2,\\ \bar{H}_n^{Y,\varepsilon}(\bar{\boldsymbol{\lambda}}) &:= \frac{q_\Theta^2}{2n}\langle\bar{\boldsymbol{\Lambda}},\bar{\boldsymbol{\lambda}}\rangle^2 + \frac{q_\Theta}{2}\bar{\boldsymbol{\lambda}}^\mathsf{T}\boldsymbol{W}\bar{\boldsymbol{\lambda}} - \frac{q_\Theta^2}{4n}\|\bar{\boldsymbol{\lambda}}\|^4.\end{aligned}$$

Recall that $\boldsymbol{W}' \overset{d}{=} \mathrm{GOE}(n)$ and is independent of $(\boldsymbol{W},\boldsymbol{Z})$. For $s,q,h\geq 0$, we define the truncated versions of Φ_n^Y, Φ_n and $\mathcal{F}$ as

$$\begin{aligned}\bar{\Phi}_n^{Y,\varepsilon}(h) &:= \frac{1}{n}\mathbb{E}\left[\log\left(\int\exp\left(\bar{H}_n^{Y,\varepsilon}(\bar{\boldsymbol{\lambda}}) + \frac{h}{2n}\langle\bar{\boldsymbol{\Lambda}},\bar{\boldsymbol{\lambda}}\rangle^2 + \frac{\sqrt{h}}{2}\bar{\boldsymbol{\lambda}}^\mathsf{T}\boldsymbol{W}'\bar{\boldsymbol{\lambda}} - \frac{h}{4n}\|\bar{\boldsymbol{\lambda}}\|^4\right)\mu_{\bar{\Lambda}}^{\otimes n}(\mathrm{d}\bar{\boldsymbol{\lambda}})\right)\right],\\ \bar{\Phi}_n^{\varepsilon}(h) &:= \frac{1}{n}\mathbb{E}\left[\log\left(\int\exp\left(\bar{H}_n^{\varepsilon}(\bar{\boldsymbol{\lambda}},\boldsymbol{\theta}) + \frac{h}{2n}\langle\bar{\boldsymbol{\Lambda}},\bar{\boldsymbol{\lambda}}\rangle^2 + \frac{\sqrt{h}}{2}\bar{\boldsymbol{\lambda}}^\mathsf{T}\boldsymbol{W}'\bar{\boldsymbol{\lambda}} - \frac{h}{4n}\|\bar{\boldsymbol{\lambda}}\|^4\right)\mu_{\bar{\Lambda}}^{\otimes n}(\mathrm{d}\bar{\boldsymbol{\lambda}})\mu_\Theta^{\otimes d}(\mathrm{d}\boldsymbol{\theta})\right)\right],\\ \bar{\mathcal{F}}^{\varepsilon}(s,q) &:= -\frac{s}{4}q^2 + \mathbb{E}_{Z\sim\mathsf{N}(0,1),\boldsymbol{\Lambda}_0\sim\mu_\Lambda}\left[\log\left(\int\exp\left(\sqrt{sq}Z\bar{\lambda} + sq\bar{\lambda}\bar{\boldsymbol{\Lambda}}_0 - \frac{s}{2}q\bar{\lambda}^2\right)\mu_{\bar{\Lambda}}(\mathrm{d}\bar{\lambda})\right)\right].\end{aligned}$$

In the above display, $\mu_{\bar{\Lambda}}$ stands for the law of $\bar{\boldsymbol{\Lambda}}_0$. The following lemma states that $\bar{\Phi}_n^{\varepsilon}(h)$ is close to $\Phi_n(h,0)$ for small ε.

Lemma A.4.6. *Under the conditions of Theorem 2.4.3, there exists a constant $C_0>0$, which is a function* *of $(\mu_\Lambda, \mu_\Theta)$ only, such that for n, d large enough, the following inequality holds for all $h \in [0,1]$:*

$$\left|\Phi_n(h,0) - \bar{\Phi}_n^\varepsilon(h)\right| \leq C_0 \sqrt[4]{\varepsilon}.$$

The proof of Lemma A.4.6 is deferred to Appendix A.5.5. Furthermore, according to Lemma 46 from [125], $\mathcal{F}$ is also close to $\bar{\mathcal{F}}^\varepsilon$ for ε small.

Lemma A.4.7 (Lemma 46 from [125])**.** *Under the conditions of Theorem 2.4.3, there exists a constant $K' > 0$ that depends only on μ_Λ, such that*

$$\left|\sup_{q\geq 0} \mathcal{F}(s,q) - \sup_{q\geq 0} \bar{\mathcal{F}}^\varepsilon(s,q)\right| \leq sK'\varepsilon.$$

Invoking the convergence results of free energy density for the symmetric spiked model [125], we have

$$\left|\bar{\Phi}_n^{Y,\varepsilon}(h) - \sup_{q\geq 0} \bar{\mathcal{F}}^\varepsilon(q_\Theta^2 + h, q)\right| = o_n(1).$$

Applying Lemma A.4.5 to the truncated distribution $\mu_{\bar{\Lambda}}$, we obtain that $|\bar{\Phi}_n^{Y,\varepsilon}(h) - \bar{\Phi}_n^\varepsilon(h)| = o_n(1)$ for all $h \geq 0$. Using this result and Lemma A.4.6, A.4.7, we derive that for all $h \in [0,1]$,

$$\begin{aligned}
&\left|\Phi_n(h,0) - \Phi_n^Y(h,0)\right| \\
\leq &\left|\Phi_n(h,0) - \bar{\Phi}_n^\varepsilon(h)\right| + \left|\bar{\Phi}_n^\varepsilon(h) - \bar{\Phi}_n^{Y,\varepsilon}(h)\right| + \left|\bar{\Phi}_n^{Y,\varepsilon}(h) - \sup_{q\geq 0} \bar{\mathcal{F}}^\varepsilon(q_\Theta^2 + h, q)\right| \\
&+ \left|\sup_{q\geq 0} \bar{\mathcal{F}}^\varepsilon(q_\Theta^2 + h, q) - \sup_{q\geq 0} \mathcal{F}(q_\Theta^2 + h, q)\right| + \left|\sup_{q\geq 0} \mathcal{F}(q_\Theta^2 + h, q) - \Phi_n^Y(h,0)\right| \\
\leq &C_0 \sqrt[4]{\varepsilon} + (q_\Theta^2 + 1)K'\varepsilon + o_n(1).
\end{aligned}$$

Since ε is arbitrary, we then have the following lemma:

Lemma A.4.8. *Under the conditions of Theorem 2.4.3, for all $h \in [0,1]$, as $n, d \to \infty$ we have*

$$\lim_{n,d\to\infty} |\Phi_n(h,0) - \Phi_n^Y(h,0)| = 0.$$

Theorem 2.4.3 is a direct consequence of Lemma A.4.8. The remainder proof of Theorem 2.4.4 follows exactly the same procedure as stated in Appendix A.4.3, and here we skip it for the sake of simplicity.

A.4.5 Proof outline of Theorem 2.4.5

We state the proof outline of Theorem 2.4.5 in this section. The proofs of supporting lemmas are delayed to Appendix A.6. For the sake of simplicity, here we only consider the rank-one case $r = 1$. We comment that proof for $r \geq 2$ can be conducted analogously.

Proof outline of Theorem 2.4.5 under condition (a)

Since μ_Λ is sub-Gaussian, there exists a constant $K_0 > 0$ depending only on μ_Λ, such that for all $x > 0$

$$\mathbb{P}(|\mathbf{\Lambda}_0| \geq x) \leq 2\exp(-x^2/K_0^2). \tag{A.16}$$

For all $i \in [n]$, we define $\bar{\mathbf{\Lambda}}_i := \mathbf{\Lambda}_i \mathbb{1}\{|\mathbf{\Lambda}_i| \leq 2K_0\sqrt{\log n}\}$, $\bar{\mathbf{\Lambda}} := (\bar{\mathbf{\Lambda}}_1, \cdots, \bar{\mathbf{\Lambda}}_n)^{\mathsf{T}} \in \mathbb{R}^n$ and $\bar{\boldsymbol{A}} := \bar{\mathbf{\Lambda}}\mathbf{\Theta}^{\mathsf{T}}/\sqrt[4]{nd} + \boldsymbol{Z} \in \mathbb{R}^{n\times d}$. The next lemma says that truncation does not decrease the MMSE too much.

Lemma A.4.9. *Under the conditions of Theorem 2.4.5 (a) , as $n, d \to \infty$ we have*

$$\frac{1}{n^2}\mathbb{E}\left[\left\|\mathbf{\Lambda}\mathbf{\Lambda}^{\mathsf{T}} - \mathbb{E}[\mathbf{\Lambda}\mathbf{\Lambda}^{\mathsf{T}} \mid \boldsymbol{A}]\right\|_F^2\right] \leq \frac{1}{n^2}\mathbb{E}\left[\left\|\bar{\mathbf{\Lambda}}\bar{\mathbf{\Lambda}}^{\mathsf{T}} - \mathbb{E}[\bar{\mathbf{\Lambda}}\bar{\mathbf{\Lambda}}^{\mathsf{T}} \mid \bar{\boldsymbol{A}}]\right\|_F^2\right] + o_n(1).$$

We leave the proof of Lemma A.4.9 to Appendix A.6.1. By Lemma A.4.9, in order to prove the theorem, it suffices to show that under the current conditions, for all but countably many values of $q_\Theta > 0$,

$$\limsup_{n,d\to\infty} \frac{1}{n^2}\mathbb{E}\left[\left\|\bar{\mathbf{\Lambda}}\bar{\mathbf{\Lambda}}^{\mathsf{T}} - \mathbb{E}[\bar{\mathbf{\Lambda}}\bar{\mathbf{\Lambda}}^{\mathsf{T}} \mid \bar{\boldsymbol{A}}]\right\|_F^2\right] \leq \lim_{n\to\infty} \mathrm{MMSE}_n^{\mathrm{symm}}(\mu_\Lambda; q_\Theta). \tag{A.17}$$

We let $\boldsymbol{Z}\mathbf{\Theta} = \|\mathbf{\Theta}\| \cdot \boldsymbol{g}$, where $\boldsymbol{g} \sim \mathsf{N}(0, \boldsymbol{I}_n)$ is independent of $(\mathbf{\Theta}, \mathbf{\Lambda})$. We denote by $\boldsymbol{Z}'$ an independent copy of $\boldsymbol{Z}$, such that $\boldsymbol{Z}'$ is further independent of $(\mathbf{\Theta}, \mathbf{\Lambda}, \boldsymbol{g})$. Furthermore, we can choose $\boldsymbol{Z}'$ such that $(\mathbf{\Theta}, \mathbf{\Lambda}, \boldsymbol{g}, \boldsymbol{Z}\mathbf{P}_{\mathbf{\Theta}}^{\perp}\boldsymbol{Z}^{\mathsf{T}}) = (\mathbf{\Theta}, \mathbf{\Lambda}, \boldsymbol{g}, \boldsymbol{Z}'\mathbf{P}_{\mathbf{\Theta}}^{\perp}\boldsymbol{Z}'^{\mathsf{T}})$, where $\mathbf{P}_{\mathbf{\Theta}}^{\perp}$ denotes the projection onto the null space of $\mathbf{\Theta}$.

We define $\boldsymbol{Y}_1 := (\bar{\boldsymbol{A}}\bar{\boldsymbol{A}}^{\mathsf{T}} - d\boldsymbol{I}_n)/\sqrt{d}$. Then $\boldsymbol{Y}_1$ admits the following decomposition:

$$\begin{aligned}
\boldsymbol{Y}_1 =& \frac{\|\mathbf{\Theta}\|^2}{\sqrt{nd}}\bar{\mathbf{\Lambda}}\bar{\mathbf{\Lambda}}^{\mathsf{T}} + \frac{\|\mathbf{\Theta}\|}{n^{1/4}d^{3/4}}\bar{\mathbf{\Lambda}}\boldsymbol{g}^{\mathsf{T}} + \frac{\|\mathbf{\Theta}\|}{n^{1/4}d^{3/4}}\boldsymbol{g}\bar{\mathbf{\Lambda}}^{\mathsf{T}} + \frac{1}{\sqrt{d}}(\boldsymbol{Z}\mathbf{P}_{\mathbf{\Theta}}^{\perp}\boldsymbol{Z}^{\mathsf{T}} - d\boldsymbol{I}_n) + \frac{1}{\sqrt{d}}\boldsymbol{g}\boldsymbol{g}^{\mathsf{T}} \\
=& \frac{1}{\sqrt{n}}\left(q_\Theta^{1/2}\bar{\mathbf{\Lambda}} + r_n^{-1}\boldsymbol{g}\right)\left(q_\Theta^{1/2}\bar{\mathbf{\Lambda}} + r_n^{-1}\boldsymbol{g}\right)^{\mathsf{T}} + \frac{1}{\sqrt{d}}(\boldsymbol{Z}'\boldsymbol{Z}'^{\mathsf{T}} - d\boldsymbol{I}_n) + \boldsymbol{E}.
\end{aligned}$$

In the above display, the $n \times n$ symmetric matrix $\boldsymbol{E}$ is defined as follows:

$$\begin{aligned}
\boldsymbol{E} :=& -\frac{1}{\sqrt{d}\|\mathbf{\Theta}\|^2}\boldsymbol{Z}'\mathbf{\Theta}\mathbf{\Theta}^{\mathsf{T}}\boldsymbol{Z}'^{\mathsf{T}} + \frac{1}{\sqrt{n}}\left(\frac{\|\mathbf{\Theta}\|^2}{d} - q_\Theta\right)\bar{\mathbf{\Lambda}}\bar{\mathbf{\Lambda}}^{\mathsf{T}} \\
&+ \frac{1}{n^{1/4}d^{1/4}}\left(\frac{\|\mathbf{\Theta}\|}{\sqrt{d}} - q_\Theta^{1/2}\right)\left(\bar{\mathbf{\Lambda}}\boldsymbol{g}^{\mathsf{T}} + \boldsymbol{g}\bar{\mathbf{\Lambda}}^{\mathsf{T}}\right) + \frac{1}{\sqrt{d}}\boldsymbol{g}\boldsymbol{g}^{\mathsf{T}}.
\end{aligned}$$

We define the set

$$\Omega_1 := \left\{\left|\frac{1}{d}\|\mathbf{\Theta}\|^2 - q_\Theta\right| \leq \frac{C_1\sqrt{\log n}}{\sqrt{d}}, \frac{1}{\sqrt{d}}|(\boldsymbol{Z}'\mathbf{\Theta})_i| \leq C_1\sqrt{\log n}, \frac{1}{\sqrt{d}}|g_i| \leq C_1\sqrt{\log n} \text{ for all } i \in [n]\right\},$$

where $C_1 > 0$ is a constant depending only on μ_Θ. Since μ_Θ, μ_Λ are sub-Gaussian distributions, if we choose C_1 large enough, then we have $\mathbb{P}(\Omega_1) = 1 - o_n(1)$. Using the definition of Ω_1, we conclude that there exists a constant $C_2 > 0$ that depends only on (C_1, K_0, μ_Θ), such that on Ω_1 we have $|E_{ij}| \leq C_2 \log n/\sqrt{d}$ for all $i, j \in [n]$. For some absolute constant $C_3 > 0$, we let $\bar{\boldsymbol{g}} \in \mathbb{R}^n$ such that $\bar{g}_i := g_i\mathbb{1}\{|g_i| \leq C_3\sqrt{\log n}\}$ for all $i \in [n]$. Direct computation reveals that for C_3 large enough, we have $\mathbb{P}(\bar{\boldsymbol{g}} \neq \boldsymbol{g}) \to 0$ as $n, d \to \infty$.

Define

$$\boldsymbol{Y}_2 := \frac{1}{\sqrt{n}}\left(q_\Theta^{1/2}\bar{\mathbf{\Lambda}} + r_n^{-1}\bar{\boldsymbol{g}}\right)\left(q_\Theta^{1/2}\bar{\mathbf{\Lambda}} + r_n^{-1}\bar{\boldsymbol{g}}\right)^{\mathsf{T}} + \boldsymbol{G},$$

where $\boldsymbol{G} \sim \sqrt{n}\mathrm{GOE}(n)$ and is independent of $(\mathbf{\Lambda}, \bar{\mathbf{\Lambda}}, \mathbf{\Theta}, \boldsymbol{g}, \bar{\boldsymbol{g}})$. By [40, Theorem 4], under condition (a), there exists a coupling such that as $n, d \to \infty$, with probability $1 - o_n(1)$ we have $(\boldsymbol{Z}'\boldsymbol{Z}'^{\mathsf{T}} - d\boldsymbol{I}_n)/\sqrt{d} = \boldsymbol{G}$. We define $\Omega_2 := \Omega_1 \cap \{\boldsymbol{Y}_2 = \boldsymbol{Y}_1 - \boldsymbol{E}\}$, then we see that $\mathbb{P}(\Omega_2) \to 1$ as $n, d \to \infty$.

For $\boldsymbol{X} \in \mathbb{R}^{n\times n}$, we define

$$\boldsymbol{M}_n(\boldsymbol{X}) := \frac{1}{n}\mathbb{E}\left[\left(q_\Theta^{1/2}\bar{\boldsymbol{\Lambda}} + r_n^{-1}\bar{\boldsymbol{g}}\right)\left(q_\Theta^{1/2}\bar{\boldsymbol{\Lambda}} + r_n^{-1}\bar{\boldsymbol{g}}\right)^\intercal \big| \boldsymbol{Y}_2 = \boldsymbol{X}\right].$$

Then for any $i,j,k,s \in [n]$, we have

$$\frac{\partial \boldsymbol{M}_n(\boldsymbol{X})_{ks}}{\partial X_{ij}} = \frac{1}{n^{3/2}}\mathbb{E}\big[(q_\Theta^{1/2}\bar{\boldsymbol{\Lambda}}_i + r_n^{-1}\bar{g}_i)(q_\Theta^{1/2}\bar{\boldsymbol{\Lambda}}_j + r_n^{-1}\bar{g}_j)(q_\Theta^{1/2}\bar{\boldsymbol{\Lambda}}_k + r_n^{-1}\bar{g}_k)(q_\Theta^{1/2}\bar{\boldsymbol{\Lambda}}_s + r_n^{-1}\bar{g}_s) \mid \boldsymbol{Y}_2 = \boldsymbol{X}\big]$$
$$-\frac{1}{n^{3/2}}\mathbb{E}\big[(q_\Theta^{1/2}\bar{\boldsymbol{\Lambda}}_i + r_n^{-1}\bar{g}_i)(q_\Theta^{1/2}\bar{\boldsymbol{\Lambda}}_j + r_n^{-1}\bar{g}_j) \mid \boldsymbol{Y}_2 = \boldsymbol{X}\big]\mathbb{E}\big[(q_\Theta^{1/2}\bar{\boldsymbol{\Lambda}}_k + r_n^{-1}\bar{g}_k)(q_\Theta^{1/2}\bar{\boldsymbol{\Lambda}}_s + r_n^{-1}\bar{g}_s) \mid \boldsymbol{Y}_2 = \boldsymbol{X}\big].$$

Note that on Ω_2 we have $|\bar{\boldsymbol{\Lambda}}_i| \leq 2K_0\sqrt{\log n}$, $|\bar{g}_i| \leq C_3\sqrt{\log n}$, and $|E_{ij}| \leq C_2 \log n/\sqrt{d}$ for all $i,j \in [n]$. Therefore, we conclude that there exists a constant $C_4 > 0$ depending only on $(K_0, \mu_\Theta, C_1, C_2, C_3)$, such that on Ω_2 we have

$$|\boldsymbol{M}_n(\boldsymbol{Y}_1)_{ks} - \boldsymbol{M}_n(\boldsymbol{Y}_1 - \boldsymbol{E})_{ks}| \leq \frac{C_4\sqrt{n}(\log n)^3}{\sqrt{d}}.$$

Therefore, we obtain that

$$\mathbb{E}\left[\|\boldsymbol{M}_n(\boldsymbol{Y}_1 - \boldsymbol{E}) - \boldsymbol{M}_n(\boldsymbol{Y}_1)\|_F^2 \mathbb{1}_{\Omega_2}\right] \leq \frac{C_4^2 n^3(\log n)^6}{d}.$$

The right hand side of the above equation vanishes as $n, d \to \infty$ under condition (a). Therefore, we derive that $\|\boldsymbol{M}_n(\boldsymbol{Y}_2) - \boldsymbol{M}_n(\boldsymbol{Y}_1)\|_F = o_P(1)$. Note that $\boldsymbol{Y}_1$ is a function of $\bar{\boldsymbol{A}}$. Using standard truncation argument, we conclude that in order to prove Eq. (A.17), it suffices to show

$$\limsup_{n,d\to\infty} \mathbb{E}\left[\left\|q_\Theta^{-1}\boldsymbol{M}_n(\boldsymbol{Y}_2) - \frac{1}{n}\bar{\boldsymbol{\Lambda}}\bar{\boldsymbol{\Lambda}}^\intercal\right\|_F^2\right] \leq \lim_{n\to\infty} \mathrm{MMSE}_n^{\mathrm{symm}}(\mu_\Lambda; q_\Theta),$$

which we prove in Lemma A.4.10 below. The proof of this lemma is deferred to Appendix A.6.2.

Lemma A.4.10. *Under the conditions of Theorem 2.4.5 (a), for all but countably many values of $q_\Theta > 0$, we have*

$$\lim_{n\to\infty} \mathbb{E}\left[\left\|q_\Theta^{-1}\boldsymbol{M}_n(\boldsymbol{Y}_2) - \frac{1}{n}\bar{\boldsymbol{\Lambda}}\bar{\boldsymbol{\Lambda}}^\intercal\right\|_F^2\right] = \lim_{n\to\infty} \mathrm{MMSE}_n^{\mathrm{symm}}(\mu_\Lambda; q_\Theta).$$

Proof outline of Theorem 2.4.5 under condition (b)

Truncation

By assumption, there exists $0 < K_1 < \infty$ such that $\mathrm{support}(\mu_\Lambda) \subseteq [-K_1, K_1]$. Since μ_Θ is sub-Gaussian, there exists $K_2 > 0$ which depends only on μ_Θ, such that for all $x > 0$

$$\mathbb{P}_{\boldsymbol{\Theta}_0\sim\mu_\Theta}(|\boldsymbol{\Theta}_0| \geq x) \leq 2\exp(-x^2/K_2^2). \tag{A.18}$$

For $j \in \{0\}\cup[d]$, we define $\bar{\boldsymbol{\Theta}}_j := \boldsymbol{\Theta}_j \mathbb{1}\{|\boldsymbol{\Theta}_j| \leq 2K_2\sqrt{\log d}\}$, $\bar{\boldsymbol{\Theta}} := (\bar{\boldsymbol{\Theta}}_1, \cdots, \bar{\boldsymbol{\Theta}}_d)^\intercal \in \mathbb{R}^d$ and $\bar{\boldsymbol{A}} := \boldsymbol{\Lambda}\bar{\boldsymbol{\Theta}}^\intercal/\sqrt[4]{nd} + \boldsymbol{Z}$. Note that $\bar{\boldsymbol{A}}$ defined here is not to be confused with $\bar{\boldsymbol{A}}$ defined in Appendix A.4.5. The following lemma states that truncation does not decrease the asymptotic matrix MMSE.

Lemma A.4.11. *Under the conditions of Theorem 2.4.5 (b), as $n, d \to \infty$ we have*

$$\frac{1}{n^2}\mathbb{E}\left[\left\|\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}} - \mathbb{E}[\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}} \mid \boldsymbol{A}]\right\|_F^2\right] \leq \frac{1}{n^2}\mathbb{E}\left[\left\|\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}} - \mathbb{E}[\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}} \mid \bar{\boldsymbol{A}}]\right\|_F^2\right] + o_n(1).$$

We postpone the proof of the lemma to Appendix A.6.3. By Lemma A.4.11, in order to prove Theorem 2.4.5 under condition (b), we only need to show for all but countably many values of $q_\Theta > 0$,

$$\limsup_{n,d\to\infty} \frac{1}{n^2}\mathbb{E}\left[\left\|\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}} - \mathbb{E}[\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}} \mid \bar{\boldsymbol{A}}]\right\|_F^2\right] \leq \lim_{n\to\infty} \mathrm{MMSE}_n^{\mathrm{symm}}(\mu_\Lambda; q_\Theta).$$

Model with extra perturbation

For $s, h, a, a' \geq 0$ and $\{\varepsilon_n\}_{n\geq 1}, \{\varepsilon_n'\}_{n\geq 1} \subseteq \mathbb{R}_+$, we introduce a perturbed model sequence, such that for each n, we observe $(\bar{\boldsymbol{A}}(s), \boldsymbol{x}'(a'), \bar{\boldsymbol{x}}(a), \boldsymbol{Y}'(h))$ defined as follows:

$$\bar{\boldsymbol{A}}(s) := \frac{\sqrt{s}}{\sqrt[4]{nd}}\boldsymbol{\Lambda}\bar{\boldsymbol{\Theta}}^{\mathsf{T}} + \boldsymbol{Z}, \tag{A.19}$$

$$\boldsymbol{x}'(a') := a'\sqrt{\varepsilon_n'}\boldsymbol{\Lambda} + \boldsymbol{g}', \tag{A.20}$$

$$\bar{\boldsymbol{x}}(a) := a\sqrt{\frac{n\varepsilon_n}{d}}\bar{\boldsymbol{\Theta}} + \boldsymbol{g}, \tag{A.21}$$

$$\boldsymbol{Y}'(h) := \frac{\sqrt{h}}{n}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}} + \boldsymbol{W}', \tag{A.22}$$

where $\boldsymbol{g} \sim \mathsf{N}(0, \boldsymbol{I}_d)$, $\boldsymbol{g}' \sim \mathsf{N}(0, \boldsymbol{I}_n)$, $\boldsymbol{W}' \sim \mathrm{GOE}(n)$, mutually independent and are independent of everything else. Note that $\bar{\boldsymbol{A}}(1) = \bar{\boldsymbol{A}}$. Furthermore, we assume that $\varepsilon_n, \varepsilon_n' \to 0^+$ as $n, d \to \infty$. We can associate to the observation (A.19) the Hamiltonian

$$\bar{H}_n^{[s]}(\boldsymbol{\lambda}, \bar{\boldsymbol{\theta}}) = \sum_{i\in[n], j\in[d]} \left\{\frac{s}{\sqrt{nd}}\boldsymbol{\Lambda}_i\boldsymbol{\lambda}_i\bar{\boldsymbol{\Theta}}_j\bar{\boldsymbol{\theta}}_j + \frac{\sqrt{s}}{\sqrt[4]{nd}}Z_{ij}\boldsymbol{\lambda}_i\bar{\boldsymbol{\theta}}_j - \frac{s}{2\sqrt{nd}}\boldsymbol{\lambda}_i^2\bar{\boldsymbol{\theta}}_j^2\right\}, \tag{A.23}$$

where $\bar{\boldsymbol{\theta}}_j = \boldsymbol{\theta}_j\mathbb{1}\{|\boldsymbol{\theta}_j| \leq 2K_2\sqrt{\log d}\}$. For the sake of simplicity, we let $\bar{H}_n(\boldsymbol{\lambda}, \bar{\boldsymbol{\theta}}) = \bar{H}_n^{[1]}(\boldsymbol{\lambda}, \bar{\boldsymbol{\theta}})$. Similarly, we can associate to the observations (A.20) and (A.21) the following Hamiltonians, respectively:

$$\bar{H}_{n,\lambda}^{(pert)}(\boldsymbol{\lambda}) = \sum_{i=1}^n \left\{\sqrt{\varepsilon_n'}a'\boldsymbol{\lambda}_i g_i' + {a'}^2\varepsilon_n'\boldsymbol{\Lambda}_i\boldsymbol{\lambda}_i - \frac{{a'}^2\varepsilon_n'}{2}\boldsymbol{\lambda}_i^2\right\},$$

$$\bar{H}_{n,\theta}^{(pert)}(\bar{\boldsymbol{\theta}}) = \sum_{j=1}^d \left\{\sqrt{\frac{n\varepsilon_n}{d}}a\bar{\boldsymbol{\theta}}_j g_j + \frac{a^2 n\varepsilon_n}{d}\bar{\boldsymbol{\Theta}}_j\bar{\boldsymbol{\theta}}_j - \frac{a^2 n\varepsilon_n}{2d}\bar{\boldsymbol{\theta}}_j^2\right\}.$$

We then define the “total” Hamiltonian, which corresponds to all observations in the perturbed model as

$$\bar{H}_n^{(tot)}(\boldsymbol{\lambda}, \bar{\boldsymbol{\theta}}) := \bar{H}_n^{[s]}(\boldsymbol{\lambda}, \bar{\boldsymbol{\theta}}) + \bar{H}_{n,\lambda}^{(pert)}(\boldsymbol{\lambda}) + \bar{H}_{n,\theta}^{(pert)}(\bar{\boldsymbol{\theta}}) + H_n(\boldsymbol{\lambda}; \boldsymbol{Y}'(h)),$$

where we recall that $H_n(\boldsymbol{\lambda}; \boldsymbol{Y}'(h))$ is defined in Eq. (A.6). The posterior distribution of $(\boldsymbol{\Lambda}, \bar{\boldsymbol{\Theta}})$ given observations $(\bar{\boldsymbol{A}}(s), \boldsymbol{x}'(a'), \bar{\boldsymbol{x}}(a), \boldsymbol{Y}'(h))$ can be expressed as

$$\mu(\mathrm{d}\boldsymbol{\lambda}, \mathrm{d}\bar{\boldsymbol{\theta}} \mid \bar{\boldsymbol{A}}(s), \boldsymbol{x}'(a'), \bar{\boldsymbol{x}}(a), \boldsymbol{Y}'(h)) \propto \exp(\bar{H}_n^{(tot)}(\boldsymbol{\lambda}, \bar{\boldsymbol{\theta}}))\mu_\Lambda^{\otimes n}(\mathrm{d}\boldsymbol{\lambda})\mu_{\bar{\Theta}}^{\otimes d}(\mathrm{d}\bar{\boldsymbol{\theta}}).$$

We define the free energy functionals corresponding to the total Hamiltonian as

$$\begin{aligned} \bar{\phi}_n(s, a, a', h) &:= \frac{1}{n} \log \int \exp(\bar{H}_n^{(tot)}(\boldsymbol{\lambda}, \bar{\boldsymbol{\theta}})) \mu_\Lambda^{\otimes n}(\mathrm{d}\boldsymbol{\lambda}) \mu_{\bar{\Theta}}^{\otimes d}(\mathrm{d}\bar{\boldsymbol{\theta}}), \\ \bar{\Phi}_n(s, a, a', h) &:= \mathbb{E}\left[\bar{\phi}_n(s, a, a', h)\right], \end{aligned} \tag{A.24}$$

where $\mu_{\bar{\Theta}}^{\otimes d}$ is the product distribution over $\mathbb{R}^d$ with each coordinate distributed as $\bar{\boldsymbol{\Theta}}_0$.

Truncation does not change the asymptotic free energy density

Next, we show that truncation does not change the asymptotic free energy density.

Lemma A.4.12. *For $s, h \geq 0$, we define the following unperturbed Hamiltonian and free energy density*

$$\begin{aligned} H_n^{[s]}(\boldsymbol{\lambda}, \boldsymbol{\theta}) &:= \sum_{i \in [n], j \in [d]} \left\{ \frac{s}{\sqrt{nd}} \boldsymbol{\Lambda}_i \boldsymbol{\lambda}_i \boldsymbol{\Theta}_j \boldsymbol{\theta}_j + \frac{\sqrt{s}}{\sqrt[4]{nd}} Z_{ij} \boldsymbol{\lambda}_i \boldsymbol{\theta}_j - \frac{s}{2\sqrt{nd}} \boldsymbol{\lambda}_i^2 \boldsymbol{\theta}_j^2 \right\}, \\ \Phi_n(s, 0, 0, h) &:= \frac{1}{n} \mathbb{E}\left[\log \int \exp(H_n^{[s]}(\boldsymbol{\lambda}, \boldsymbol{\theta}) + H_n(\boldsymbol{\lambda}; \boldsymbol{Y}'(h))) \mu_\Lambda^{\otimes n}(\mathrm{d}\boldsymbol{\lambda}) \mu_\Theta^{\otimes d}(\mathrm{d}\boldsymbol{\theta}) \right]. \end{aligned}$$

Then for all fixed $S_0 > 0$, under the conditions of Theorem 2.4.5 (b), as $n, d \to \infty$ we have

$$\sup_{h \geq 0, S_0 \geq s \geq 0} |\Phi_n(s, 0, 0, h) - \bar{\Phi}_n(s, 0, 0, h)| = o_n(1).$$

The proof of Lemma A.4.12 is given in Appendix A.6.4. We further characterize the convergence of free energy density in Lemma A.4.13, again postponing the proof to Appendix A.6.5.

Lemma A.4.13. *Recall that $\mathcal{F}(\cdot, \cdot)$ is defined in Eq. (2.12). Under the conditions of Theorem 2.4.5 (b), if we further assume that $\varepsilon_n, \varepsilon_n' \to 0^+$, then for all fixed $s, h \geq 0$, as $n, d \to \infty$*

$$\lim_{n,d \to \infty} \sup_{a, a' \in [0,10]} \left| \bar{\Phi}_n(s, a, a', h) - \sup_{q \geq 0} \mathcal{F}(q_\Theta^2 s^2 + h, q) \right| = 0.$$

Recall that D is defined in Eq. (A.13). For all $h + q_\Theta^2 \in D$, the mapping $s \mapsto \sup_{q \geq 0} \mathcal{F}(q_\Theta^2 s^2 + h, q)$ is differentiable at $s = 1$. Furthermore, for all fixed $a, a', h \geq 0$, the mappings $s \mapsto \bar{\Phi}_n(s, a, a', h)$, $s \mapsto \bar{\Phi}_n(s, 0, 0, h)$ are convex differentiable on $(0, \infty)$. By Lemmas A.4.12 and A.4.13, as $n, d \to \infty$, $\bar{\Phi}_n(s, a, a', h) \to \sup_{q \geq 0} \mathcal{F}(q_\Theta^2 s^2 + h, q)$ and $\bar{\Phi}_n(s, 0, 0, h) \to \sup_{q \geq 0} \mathcal{F}(q_\Theta^2 s^2 + h, q)$. Then we apply Lemma A.2.4 and conclude that for $h + q_\Theta^2 \in D$, we have

$$\left| \frac{\partial}{\partial s} \bar{\Phi}_n(s, a, a', h) \Big|_{s=1} - \frac{\partial}{\partial s} \bar{\Phi}_n(s, 0, 0, h) \Big|_{s=1} \right| = o_n(1).$$

Using Gaussian integration by parts and Nishimori identity, we further derive that for $h + q_\Theta^2 \in D$

$$\lim_{n,d \to \infty} \frac{1}{2n\sqrt{nd}} \mathbb{E}\left[\| \mathbb{E}[\boldsymbol{\Lambda} \bar{\boldsymbol{\Theta}}^\mathsf{T} \mid \bar{\boldsymbol{A}}(1), \bar{\boldsymbol{x}}(a), \boldsymbol{x}'(a'), \boldsymbol{Y}'(h)] - \mathbb{E}[\boldsymbol{\Lambda} \bar{\boldsymbol{\Theta}}^\mathsf{T} \mid \bar{\boldsymbol{A}}(1), \boldsymbol{Y}'(h)] \|_F^2 \right] = 0. \tag{A.25}$$

By Jensen's inequality, for all $a, a' \in [1,2]$

$$
\begin{aligned}
&\frac{1}{2n\sqrt{nd}}\mathbb{E}\left[\|\mathbb{E}[\boldsymbol{\Lambda}\bar{\boldsymbol{\Theta}}^{\mathsf{T}} \mid \bar{\boldsymbol{A}}(1), \bar{\boldsymbol{x}}(a), \boldsymbol{x}'(a'), \boldsymbol{Y}'(h)] - \mathbb{E}[\boldsymbol{\Lambda}\bar{\boldsymbol{\Theta}}^{\mathsf{T}} \mid \bar{\boldsymbol{A}}(1), \boldsymbol{Y}'(h)]\|_F^2\right] \\
\leq &\frac{1}{2n\sqrt{nd}}\mathbb{E}\left[\|\mathbb{E}[\boldsymbol{\Lambda}\bar{\boldsymbol{\Theta}}^{\mathsf{T}} \mid \bar{\boldsymbol{A}}(1), \bar{\boldsymbol{x}}(2), \boldsymbol{x}'(2), \boldsymbol{Y}'(h)] - \mathbb{E}[\boldsymbol{\Lambda}\bar{\boldsymbol{\Theta}}^{\mathsf{T}} \mid \bar{\boldsymbol{A}}(1), \boldsymbol{Y}'(h)]\|_F^2\right].
\end{aligned}
$$

By Eq. (A.25), the last line above converges to zero as $n, d \to \infty$ for all $q_\Theta^2 + h \in D$. In this case, we have

$$
\lim_{n,d\to\infty}\int_1^2\int_1^2 \frac{1}{2n\sqrt{nd}}\mathbb{E}\left[\|\mathbb{E}[\boldsymbol{\Lambda}\bar{\boldsymbol{\Theta}}^{\mathsf{T}} \mid \bar{\boldsymbol{A}}(1), \bar{\boldsymbol{x}}(a), \boldsymbol{x}'(a'), \boldsymbol{Y}'(h)] - \mathbb{E}[\boldsymbol{\Lambda}\bar{\boldsymbol{\Theta}}^{\mathsf{T}} \mid \bar{\boldsymbol{A}}(1), \boldsymbol{Y}'(h)]\|_F^2\right] \mathrm{d}a\mathrm{d}a' = 0. \tag{A.26}
$$

We define

$$
\begin{aligned}
\boldsymbol{\Lambda}_{s,a,a',h} &:= \mathbb{E}[\boldsymbol{\Lambda} \mid \bar{\boldsymbol{A}}(1), \bar{\boldsymbol{x}}(a), \boldsymbol{x}'(a'), \boldsymbol{Y}'(h)] \in \mathbb{R}^n, \\
\bar{\boldsymbol{\Theta}}_{s,a,a',h} &:= \mathbb{E}[\bar{\boldsymbol{\Theta}} \mid \bar{\boldsymbol{A}}(1), \bar{\boldsymbol{x}}(a), \boldsymbol{x}'(a'), \boldsymbol{Y}'(h)] \in \mathbb{R}^d, \\
\boldsymbol{M}_{s,a,a',h} &:= \mathbb{E}[\boldsymbol{\Lambda}\bar{\boldsymbol{\Theta}}^{\mathsf{T}} \mid \bar{\boldsymbol{A}}(1), \bar{\boldsymbol{x}}(a), \boldsymbol{x}'(a'), \boldsymbol{Y}'(h)] \in \mathbb{R}^{n\times d}.
\end{aligned}
$$

Invoking Stein's lemma, we see that the following equation holds:

$$
\frac{\partial}{\partial a}\mathbb{E}\left[\|\boldsymbol{\Lambda}_{1,a,a',h}\|^2\right] = \frac{2an\varepsilon_n}{d}\mathbb{E}\left[\|\boldsymbol{M}_{1,a,a',h} - \boldsymbol{\Lambda}_{1,a,a',h}\bar{\boldsymbol{\Theta}}_{1,a,a',h}^{\mathsf{T}}\|_F^2\right].
$$

Using the above equation, we obtain that

$$
\begin{aligned}
&\int_1^2\int_1^2 \frac{1}{2n\sqrt{nd}}\mathbb{E}\left[\|\boldsymbol{M}_{1,a,a',h} - \boldsymbol{\Lambda}_{1,a,a',h}\bar{\boldsymbol{\Theta}}_{1,a,a',h}\|_F^2\right]\mathrm{d}a\mathrm{d}a' \\
\leq &\frac{d^{1/2}}{2\varepsilon_n n^{5/2}}\int_1^2 \left\{\mathbb{E}[\|\boldsymbol{\Lambda}_{1,2,a',h}\|^2] - \mathbb{E}[\|\boldsymbol{\Lambda}_{1,1,a',h}\|^2]\right\}\mathrm{d}a' \\
\leq &\frac{d^{1/2}\mathbb{E}_{\boldsymbol{\Lambda}_0\sim\mu_\Lambda}[\boldsymbol{\Lambda}_0^2]}{\varepsilon_n n^{3/2}}.
\end{aligned} \tag{A.27}
$$

Overlap concentration

The next lemmas show that if we draw two independent samples from the posterior distribution of $\bar{\boldsymbol{\Theta}}$ ($\boldsymbol{\Lambda}$) given ($\bar{\boldsymbol{A}}(1), \bar{\boldsymbol{x}}(a), \boldsymbol{x}'(a'), \boldsymbol{Y}'(h)$), then their normalized inner product concentrates. This phenomenon is referred to as *overlap concentration* in the literature of statistical mechanics. In what follows, we prove overlap concentration for $\bar{\boldsymbol{\Theta}}$. Since the proof is similar, in order to avoid redundancy, we skip the counterpart proof for $\boldsymbol{\Lambda}$. For the sake of simplicity, we denote by $\langle\cdot\rangle_{s,a,a',h}$ the expectation with respect to the posterior distribution $\mathbb{P}(\cdot \mid \bar{\boldsymbol{A}}(s), \boldsymbol{x}'(a'), \bar{\boldsymbol{x}}(a), \boldsymbol{Y}'(h))$.

Lemma A.4.14. *For $\bar{\boldsymbol{\theta}} \in \mathbb{R}^d$, we define*

$$
U(\bar{\boldsymbol{\theta}}) := \sum_{j=1}^d \left\{\frac{1}{\sqrt{\varepsilon_n nd}}\bar{\boldsymbol{\theta}}_j g_j + \frac{2a}{d}\bar{\boldsymbol{\theta}}_j\bar{\boldsymbol{\Theta}}_j - \frac{a}{d}\bar{\boldsymbol{\theta}}_j^2\right\}.
$$

Let $\bar{\boldsymbol{\theta}}^{(1)}$, $\bar{\boldsymbol{\theta}}^{(2)}, \bar{\boldsymbol{\theta}} \in \mathbb{R}^d$ be independent samples drawn from the posterior distribution $\mathbb{P}(\bar{\boldsymbol{\Theta}} = \cdot \mid \bar{\boldsymbol{A}}(1), \bar{\boldsymbol{x}}(a), \boldsymbol{x}'(a'), \boldsymbol{Y}'(h))$.

Then under the conditions of Theorem 2.4.5 (b), for all $a, a' \in [0.1, 10]$ and $h \geq 0$, we have

$$\begin{aligned}&\mathbb{E}[\langle((\bar{\boldsymbol{\theta}}^{(1)})^{\mathsf{T}}\bar{\boldsymbol{\theta}}^{(2)}/d - \mathbb{E}[\langle(\bar{\boldsymbol{\theta}}^{(1)})^{\mathsf{T}}\bar{\boldsymbol{\theta}}^{(2)}/d\rangle_{1,a,a',h}])^2\rangle_{1,a,a',h}] \\ \leq & 40K_2^2 \log d \mathbb{E}[\langle|U(\bar{\boldsymbol{\theta}}) - \mathbb{E}[\langle U(\bar{\boldsymbol{\theta}})\rangle_{1,a,a',h}]|\rangle_{1,a,a',h}]. \end{aligned} \tag{A.28}$$

Using Eq. (A.28), we see that in order to prove overlap concentration, we only need to show that the right hand side of Eq. (A.28) is sufficiently small, which is accomplished via the following lemmas:

Lemma A.4.15. *We let $\bar{\boldsymbol{\theta}} \in \mathbb{R}^d$ be a sample drawn from the posterior distribution $\mathbb{P}(\bar{\boldsymbol{\Theta}} = \cdot \mid \bar{\boldsymbol{A}}(1), \bar{\boldsymbol{x}}(a), \boldsymbol{x}'(a'), \boldsymbol{Y}'(h))$. For $h \geq 0$, we define*

$$v_n(h) := \sup_{1/2 \leq a, a' \leq 3} \mathbb{E}[|\bar{\phi}_n(1, a, a', h) - \mathbb{E}[\bar{\phi}_n(1, a, a', h)]|],$$

where we recall that $\bar{\phi}_n$ is defined in Eq. (A.24). Then under the conditions of Theorem 2.4.5 (b), if we further assume that $\varepsilon_n \to 0^+$ and $nd^{-1/2}\varepsilon_n \to \infty$ as $n, d \to \infty$, then there exists a numerical constant $C > 0$, such that for n, d large enough

$$\int_1^2 \int_1^2 \mathbb{E}[\langle|U(\bar{\boldsymbol{\theta}}) - \mathbb{E}[\langle U(\bar{\boldsymbol{\theta}})\rangle_{1,a,a',h}]|\rangle_{1,a,a',h}]\mathrm{d}a\mathrm{d}a' \leq CK_2\sqrt{(v_n(h) + n^{-1})\varepsilon_n^{-1}\log d}.$$

Lemma A.4.16. *Under the conditions of Theorem 2.4.5 (b), if we further assume $\varepsilon_n, \varepsilon_n' \to 0^+$ as $n, d \to \infty$, then there exists a numerical constant $C_1 > 0$ such that for all n, d large enough and $0 \leq h \leq 1$*

$$v_n(h) \leq C_1 K_1^2 K_2^2 d^{1/2} n^{-1} \log d.$$

We defer the proofs of Lemmas A.4.14 to A.4.16 to Appendices A.6.6 to A.6.8, respectively. Combining Lemmas A.4.14 to A.4.16, we deduce that under the conditions of these lemmas, for $\bar{\boldsymbol{\theta}}^{(1)}, \bar{\boldsymbol{\theta}}^{(2)} \in \mathbb{R}^d$ that are two independent samples from the posterior distribution $\mathbb{P}(\bar{\boldsymbol{\Theta}} = \cdot \mid \bar{\boldsymbol{A}}(1), \boldsymbol{x}'(a'), \bar{\boldsymbol{x}}(a), \boldsymbol{Y}'(h))$, there exists a numerical constant $C > 0$, such that for n, d large enough

$$\frac{d}{n}\int_1^2 \int_1^2 \mathbb{E}[\langle(\bar{\boldsymbol{\theta}}_1^{\mathsf{T}}\bar{\boldsymbol{\theta}}_2/d - \mathbb{E}[\langle\bar{\boldsymbol{\theta}}_1^{\mathsf{T}}\bar{\boldsymbol{\theta}}_2/d\rangle_{1,a,a',h}])^2\rangle_{1,a,a',h}]\mathrm{d}a\mathrm{d}a' \leq CK_1K_2^4\varepsilon_n^{-1/2}(\log d)^2 d^{5/4} n^{-3/2}.$$

Under condition (b), we see that there exists $\varepsilon_n \to 0^+$, such that $\varepsilon_n^{-1/2}(\log d)^2 d^{5/4} n^{-3/2} \to 0$ and $nd^{-1/2}\varepsilon_n \to \infty$ as $n, d \to \infty$. We summarize the overlap concentration results in Theorem A.4.1 below, which also contains concentration argument for $\boldsymbol{\Lambda}$ (that we skip the proof).

Theorem A.4.1 (Overlap concentration)**.** *Let $\bar{\boldsymbol{\theta}}^{(1)}, \bar{\boldsymbol{\theta}}^{(2)} \in \mathbb{R}^d$ be two independent samples drawn from the posterior distribution $\mathbb{P}(\bar{\boldsymbol{\Theta}} = \cdot \mid \bar{\boldsymbol{A}}(1), \boldsymbol{x}'(a'), \bar{\boldsymbol{x}}(a), \boldsymbol{Y}'(h))$, and $\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)} \in \mathbb{R}^n$ be two independent samples drawn from the posterior distribution $\mathbb{P}(\boldsymbol{\Lambda} = \cdot \mid \bar{\boldsymbol{A}}(1), \boldsymbol{x}'(a'), \bar{\boldsymbol{x}}(a), \boldsymbol{Y}'(h))$. Under the conditions of Theorem 2.4.5 (b), there exist $\varepsilon_n, \varepsilon_n' \to 0^+$ such that for all $h \in [0, 1]$, as $n, d \to \infty$*

$$\frac{(\log d)^2 d^{5/4}}{n^{3/2}\varepsilon_n^{1/2}} \to 0, \qquad \frac{n\varepsilon_n}{d^{1/2}} \to \infty, \qquad \frac{(\log d)^{1/2} d^{1/4}}{n^{1/2}\varepsilon_n'^{\,1/2}} \to 0, \qquad \frac{n\varepsilon_n'}{d^{1/2}\log d} \to \infty,$$

$$\frac{d}{n}\int_1^2 \int_1^2 \mathbb{E}[\langle(\bar{\boldsymbol{\theta}}_1^{\mathsf{T}}\bar{\boldsymbol{\theta}}_2/d - \mathbb{E}[\langle\bar{\boldsymbol{\theta}}_1^{\mathsf{T}}\bar{\boldsymbol{\theta}}_2/d\rangle_{1,a,a',h}])^2\rangle_{1,a,a',h}]\mathrm{d}a\mathrm{d}a' \to 0,$$

$$\int_1^2\int_1^2 \mathbb{E}[\langle(\boldsymbol{\lambda}_1^\mathsf{T}\boldsymbol{\lambda}_2/n - \mathbb{E}[\langle\boldsymbol{\lambda}_1^\mathsf{T}\boldsymbol{\lambda}_2/n\rangle_{1,a,a',h}])^2\rangle_{1,a,a',h}]\mathrm{d}a\mathrm{d}a' \to 0.$$

Corollary A.4.1. *Under the conditions of Theorem A.4.1, for all $h \in [0,1]$, as $n,d\to\infty$ we have*

$$\frac{d}{n}\int_1^2\int_1^2 \mathbb{E}[\langle(\langle\bar{\boldsymbol{\theta}}_1^\mathsf{T}\bar{\boldsymbol{\theta}}_2/d\rangle_{1,a,a',h} - \mathbb{E}[\langle\bar{\boldsymbol{\theta}}_1^\mathsf{T}\bar{\boldsymbol{\theta}}_2/d\rangle_{1,a,a',h}])^2\rangle_{1,a,a',h}]\mathrm{d}a\mathrm{d}a' \to 0,$$
$$\int_1^2\int_1^2 \mathbb{E}[\langle(\langle\boldsymbol{\lambda}_1^\mathsf{T}\boldsymbol{\lambda}_2/n\rangle_{1,a,a',h} - \mathbb{E}[\langle\boldsymbol{\lambda}_1^\mathsf{T}\boldsymbol{\lambda}_2/n\rangle_{1,a,a',h}])^2\rangle_{1,a,a',h}]\mathrm{d}a\mathrm{d}a' \to 0.$$

Remark A.4.2. In Theorem A.4.1 and Corollary A.4.1, we can replace the interval $[1,2]$ with $[a,b]$ for any fixed $0<a<b<\infty$.

Proof of the theorem

In the rest parts of the proof, we always assume that $\{\varepsilon_n\}_{n\in\mathbb{N}_+}$ and $\{\varepsilon_n'\}_{n\in\mathbb{N}_+}$ are chosen as in Theorem A.4.1. Under this assumption we have $\varepsilon_n^{-1}(\log d)^4 d^{5/2}n^{-3}\to 0$ and $d^{-2}n^{3/2}(\log d)^{-4}\to 0$, thus $d^{1/2}n^{-3/2}\varepsilon_n^{-1}\to 0$ as $n,d\to\infty$. Plugging this result into Eqs. (A.26) and (A.27), we obtain that

$$\lim_{n,d\to\infty}\int_1^2\int_1^2 \frac{1}{2n\sqrt{nd}}\mathbb{E}\left[\|\boldsymbol{\Lambda}_{1,a,a',h}\bar{\boldsymbol{\Theta}}_{1,a,a',h} - \boldsymbol{M}_{1,0,0,h}\|_F^2\right]\mathrm{d}a\mathrm{d}a' = 0. \tag{A.29}$$

By Lemma A.4.12 and Lemma A.4.13 we see that $|\Phi_n(s,0,0,h) - \sup_{q\geq 0}\mathcal{F}(q_\Theta^2 s^2 + h, q)| = o_n(1)$ for all fixed $s,h\geq 0$. Notice that $s\mapsto\Phi_n(s,0,0,h)$ is convex and differentiable. Furthermore, for all but countably many values of $q_\Theta^2 + h > 0$, the mapping $s\mapsto\sup_{q\geq 0}\mathcal{F}(q_\Theta^2 s^2 + h, q)$ is differentiable at $s=1$. Using Gaussian integration by parts, Lemmas A.2.3 and A.2.4 we conclude that for these $q_\Theta^2 + h$ we have

$$\lim_{n,d\to\infty}\frac{1}{2n\sqrt{nd}}\mathbb{E}\left[\|\mathbb{E}[\boldsymbol{\Lambda}\bar{\boldsymbol{\Theta}}^\mathsf{T}\mid\bar{\boldsymbol{A}}(1),\boldsymbol{Y}'(h)]\|_F^2\right] = \frac{\partial}{\partial s}\sup_{q\geq 0}\mathcal{F}(s^2q_\Theta^2 + h, q)\Big|_{s=1}. \tag{A.30}$$

Define $D_\Theta(h) := \frac{\partial}{\partial s}\sup_{q\geq 0}\mathcal{F}(s^2q_\Theta^2 + h, q)\Big|_{s=1}$. For all but countably many $q_\Theta > 0$ the mapping $s\mapsto\sup_{q\geq 0}\mathcal{F}(q_\Theta^2 s^2, q)$ is differentiable at $s=1$, thus $D_\Theta(0)$ is well-defined. In this case, if $D_\Theta(0)=0$, then by [125] we have $\lim_{n\to\infty}\mathrm{MMSE}_n^{\mathrm{symm}}(\mu_\Lambda, q_\Theta) = \mathbb{E}_{\boldsymbol{\Lambda}_0\sim\mu_\Lambda}[\boldsymbol{\Lambda}_0^2]^2$, which is achieved by $\mathbf{0}_{n\times n}$. Using Theorem 2.4.4 we deduce that $\lim_{n,d\to\infty}\mathrm{MMSE}_n^{\mathrm{asym}}(\mu_\Lambda,\mu_\Theta) = \mathbb{E}_{\boldsymbol{\Lambda}_0\sim\mu_\Lambda}[\boldsymbol{\Lambda}_0^2]^2$, which concludes the proof of the theorem.

In the following parts of the proof we will assume $D_\Theta(0) > 0$. Let $a_1 < a_2$, $a_1' < a_2'$, and $a\sim\mathrm{Unif}[a_1,a_2]$ and $a'\sim\mathrm{Unif}[a_1',a_2']$. Similar to the derivation of Eq. (A.29) we have

$$\lim_{n,d\to\infty}\frac{1}{2n\sqrt{nd}}\mathbb{E}\left[\|\boldsymbol{\Lambda}_{1,a,a',h}\bar{\boldsymbol{\Theta}}_{1,a,a',h} - \boldsymbol{M}_{1,0,0,h}\|_F^2\right] = 0, \tag{A.31}$$

where the expectation is taken over $a\sim\mathrm{Unif}[a_1,a_2]$ and $a'\sim\mathrm{Unif}[a_1',a_2']$. By Eq. (A.30), Eq. (A.31) and triangle inequality, for all but countably many q_Θ^2, $q_\Theta^2 + h$ we have

$$\limsup_{n,d\to\infty}\left\{\frac{1}{2\sqrt{n^3d}}\mathbb{E}\left[\|\boldsymbol{\Lambda}_{1,a,a',h}\bar{\boldsymbol{\Theta}}_{1,a,a',h} - \boldsymbol{M}_{1,0,0,0}\|_F^2\right]\right\}^{1/2}$$
$$\leq\frac{1}{2\sqrt{n^3d}}\left(\limsup_{n,d\to\infty}\left\{\mathbb{E}\left[\|\boldsymbol{\Lambda}_{1,a,a',h}\bar{\boldsymbol{\Theta}}_{1,a,a',h} - \boldsymbol{M}_{1,0,0,h}\|_F^2\right]\right\}^{1/2} + \limsup_{n,d\to\infty}\left\{\mathbb{E}\left[\|\boldsymbol{M}_{1,0,0,h} - \boldsymbol{M}_{1,0,0,0}\|_F^2\right]\right\}^{1/2}\right)$$

$$= (D_\Theta(h) - D_\Theta(0))^{1/2}. \tag{A.32}$$

The following lemmas establish several useful properties of $\boldsymbol{\Lambda}_{1,a,a',h}$ and $\bar{\boldsymbol{\Theta}}_{1,a,a',h}$.

Lemma A.4.17. *Recall that D is defined in Eq. (A.13). For all fixed $a_*, a'_* \in (0,5]$, $h \in [0,1]$, under the assumptions of Theorem 2.4.5 (b), if we further assume that $h + q_\Theta^2 \in D$, then as $n, d \to \infty$ we have*

$$\frac{1}{n}\mathbb{E}\left[\left\|\mathbb{E}[\boldsymbol{\Lambda} \mid \bar{\boldsymbol{A}}(1), \boldsymbol{x}'(a'_*), \bar{\boldsymbol{x}}(a_*), \boldsymbol{Y}'(h)]\right\|^2\right] = 2\left(\frac{\partial}{\partial h}\sup_{q \geq 0}\mathcal{F}(q_\Theta^2 + h, q)\right)^{1/2} + o_n(1),$$
$$\frac{1}{\sqrt{nd}}\mathbb{E}\left[\|\mathbb{E}[\bar{\boldsymbol{\Theta}} \mid \bar{\boldsymbol{A}}(1), \boldsymbol{x}'(a'_*), \bar{\boldsymbol{x}}(a_*), \boldsymbol{Y}'(h)]\|^2\right] = 2q_\Theta^2\left(\frac{\partial}{\partial h}\sup_{q \geq 0}\mathcal{F}(q_\Theta^2 + h, q)\right)^{1/2} + o_n(1).$$

The proof of Lemma A.4.17 is deferred to Appendices A.6.9 and A.6.10, respectively. By Lemma A.4.17 and Corollary A.4.1, we conclude that for all $0 < a_1 < a_2 < 5$ and $0 < a'_1 < a'_2 < 5$, if we let $a \sim \mathrm{Unif}[a_1, a_2]$ and $a' \sim \mathrm{Unif}[a'_1, a'_2]$, then for all $h + q_\Theta^2 \in D$ we have

$$\frac{1}{\sqrt{nd}}\left\|\mathbb{E}[\bar{\boldsymbol{\Theta}} \mid \bar{\boldsymbol{A}}(1), \boldsymbol{x}'(a'), \bar{\boldsymbol{x}}(a), \boldsymbol{Y}'(h)]\right\|^2 = 2q_\Theta^2\left(\frac{\partial}{\partial h}\sup_{q \geq 0}\mathcal{F}(q_\Theta^2 + h, q)\right)^{1/2} + o_P(1). \tag{A.33}$$

Let $C_0(q_\Theta, h) := 2q_\Theta^2\left(\frac{\partial}{\partial h}\sup_{q\geq 0}\mathcal{F}(q_\Theta^2 + h, q)\right)^{1/2}$. We define the mapping $\boldsymbol{M} : \mathbb{R}^{n\times n} \times \mathbb{R}^+ \mapsto \mathbb{R}^{n \times n}$, such that $\boldsymbol{M}(\boldsymbol{X}; b)_{ij} = X_{ij}\mathbb{1}\{|X_{ij}| \leq b\}$. We further define $\boldsymbol{M}_0 : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times n}$ such that for $\boldsymbol{X} \in \mathbb{R}^{n \times d}$

$$\boldsymbol{M}_0(\boldsymbol{X}) = \boldsymbol{M}\left(\tfrac{1}{\sqrt{nd}}\mathbb{E}[\boldsymbol{\Lambda}\bar{\boldsymbol{\Theta}}^\mathsf{T} \mid \bar{\boldsymbol{A}} = \boldsymbol{X}]\mathbb{E}[\bar{\boldsymbol{\Theta}}\bar{\boldsymbol{\Lambda}}^\mathsf{T} \mid \bar{\boldsymbol{A}} = \boldsymbol{X}];\, 2K_1^2 C_0(q_\Theta, h)\right).$$

By triangle inequality

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}\left[\|\boldsymbol{M}_0(\bar{\boldsymbol{A}}) - C_0(q_\Theta, h)\mathbb{E}[\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\mathsf{T} \mid \bar{\boldsymbol{A}}, \boldsymbol{Y}'(h)]\|_F^2\right]^{1/2} \\
\leq &\frac{C_0(q_\Theta, h)}{n}\mathbb{E}\left[\|\boldsymbol{\Lambda}_{1,a,a',h}\boldsymbol{\Lambda}_{1,a,a',h}^\mathsf{T} - \mathbb{E}[\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\mathsf{T} \mid \bar{\boldsymbol{A}}, \boldsymbol{Y}'(h)]\|_F^2\right]^{1/2} \\
&+ \frac{1}{n}\mathbb{E}\left[\|C_0(q_\Theta, h)\boldsymbol{\Lambda}_{1,a,a',h}\boldsymbol{\Lambda}_{1,a,a',h}^\mathsf{T} - \boldsymbol{M}_0(\bar{\boldsymbol{A}})\|_F^2\right]^{1/2},
\end{aligned} \tag{A.34}$$

where the expectation is taken over $(a, a', \boldsymbol{\Lambda}, \bar{\boldsymbol{\Theta}}, \boldsymbol{Z}, \boldsymbol{g}, \boldsymbol{g}', \boldsymbol{W}')$. Using Gaussian integration by parts we obtain

$$\frac{\partial}{\partial a'}\mathbb{E}\left[\|\boldsymbol{\Lambda}_{1,a,a',h}\|^2\right] = \varepsilon'_n a'\mathbb{E}\left[\|\mathbb{E}[\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\mathsf{T} \mid \bar{\boldsymbol{A}}, \bar{\boldsymbol{x}}(a), \bar{\boldsymbol{x}}'(a'), \boldsymbol{Y}'(h)] - \boldsymbol{\Lambda}_{1,a,a',h}\boldsymbol{\Lambda}_{1,a,a',h}^\mathsf{T}\|_F^2\right]. \tag{A.35}$$

By Eq. (A.35) we have

$$\begin{aligned}
&\frac{1}{n^2}\mathbb{E}\left[\|\boldsymbol{\Lambda}_{1,a,a',h}\boldsymbol{\Lambda}_{1,a,a',h} - \mathbb{E}[\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\mathsf{T} \mid \bar{\boldsymbol{A}}, \boldsymbol{x}'(a'), \bar{\boldsymbol{x}}(a), \boldsymbol{Y}'(h)]\|_F^2\right] \\
= &\frac{1}{n^2(a_2 - a_1)(a'_2 - a'_1)}\int_{a_1}^{a_2}\int_{a'_1}^{a'_2}(\varepsilon'_n a')^{-1}\frac{\partial}{\partial a'}\mathbb{E}\left[\|\boldsymbol{\Lambda}_{1,a,a',h}\|^2\right]\mathrm{d}a'\mathrm{d}a \\
\leq &\frac{1}{n^2\varepsilon'_n a'_1(a_2 - a_1)(a'_2 - a'_1)}\int_{a_1}^{a_2}\mathbb{E}\left[\|\boldsymbol{\Lambda}_{1,a,a'_2,h}\|^2\right]\mathrm{d}a \\
\leq &\frac{K_1^2}{n(a'_2 - a'_1)\varepsilon'_n a'_1},
\end{aligned} \tag{A.36}$$

which vanishes as $n, d \to \infty$. For $\boldsymbol{X} \in \mathbb{R}^{n \times n}$, we define $\|\boldsymbol{X}\|_1 := \sum_{i,j \in [n]} |X_{ij}|$. Then we have

$$
\begin{aligned}
&\frac{1}{n^2}\mathbb{E}\left[\|C_0(q_\Theta, h)\boldsymbol{\Lambda}_{1,a,a',h}\boldsymbol{\Lambda}_{1,a,a',h}^\mathsf{T} - \boldsymbol{M}_0(\bar{\boldsymbol{A}})\|_F^2\right] \\
&\overset{(i)}{\leq} \frac{3K_1^2 C_0(q_\Theta, h)}{n^2}\mathbb{E}\left[\|C_0(q_\Theta, h)\boldsymbol{\Lambda}_{1,a,a',h}\boldsymbol{\Lambda}_{1,a,a',h}^\mathsf{T} - \boldsymbol{M}_0(\bar{\boldsymbol{A}})\|_1\right] \\
&\overset{(ii)}{\leq} \frac{3K_1^2 C_0(q_\Theta, h)}{n^2}\mathbb{E}\left[\left\|C_0(q_\Theta, h)\boldsymbol{\Lambda}_{1,a,a',h}\boldsymbol{\Lambda}_{1,a,a',h}^\mathsf{T} - \frac{1}{\sqrt{nd}}\boldsymbol{M}_{1,0,0,0}\boldsymbol{M}_{1,0,0,0}^\mathsf{T}\right\|_1\right] \\
&\overset{(iii)}{\leq} \frac{3K_1^2 C_0(q_\Theta, h)}{n^2}\mathbb{E}\left[\left\|\frac{\|\bar{\boldsymbol{\Theta}}_{1,a,a',h}\|^2}{\sqrt{nd}}\boldsymbol{\Lambda}_{1,a,a',h}\boldsymbol{\Lambda}_{1,a,a',h}^\mathsf{T} - \frac{1}{\sqrt{nd}}\boldsymbol{M}_{1,0,0,0}\boldsymbol{M}_{1,0,0,0}^\mathsf{T}\right\|_1\right] &\text{(A.37)} \\
&\quad + 3K_1^4 C_0(q_\Theta, h)\mathbb{E}\left[\left|\frac{\|\bar{\boldsymbol{\Theta}}_{1,a,a',h}\|^2}{\sqrt{nd}} - C_0(q_\Theta, h)\right|\right], &\text{(A.38)}
\end{aligned}
$$

where in *(i)* we use the assumption that μ_Λ has bounded support, and *(ii)* is by the fact that for all $|x| \leq C_0(q_\Theta, h)K_1^2$, $y \in \mathbb{R}$,

$$
\big|x - y\mathbb{1}\{|y| \leq 2K_1^2 C_0(q_{\Theta,h})\}\big| \leq \big|x - y\big|.
$$

Lastly, *(iii)* is by triangle inequality. Applying Lemma A.2.6 and Hölder's inequality we see that

$$
\begin{aligned}
&\frac{3K_1^2 C_0(q_\Theta, h)}{n^2} \times \mathbb{E}\Big[\Big\|\frac{\|\bar{\boldsymbol{\Theta}}_{1,a,a',h}\|^2}{\sqrt{nd}}\boldsymbol{\Lambda}_{1,a,a',h}\boldsymbol{\Lambda}_{1,a,a',h}^\mathsf{T} - \frac{1}{\sqrt{nd}}\boldsymbol{M}_{1,0,0,0}\boldsymbol{M}_{1,0,0,0}^\mathsf{T}\Big\|_1\Big] \\
\leq &\frac{6K_1^2 C_0(q_\Theta, h)}{n\sqrt{nd}} \times \mathbb{E}\left[\left\|\boldsymbol{M}_{1,0,0,0} - \boldsymbol{\Lambda}_{1,a,a',h}\bar{\boldsymbol{\Theta}}_{1,a,a',h}^\mathsf{T}\right\|_F^2\right]^{1/2} \\
&\times \left(\mathbb{E}\left[\left\|\boldsymbol{M}_{1,0,0,0}\right\|_F^2\right]^{1/2} + \mathbb{E}\left[\left\|\boldsymbol{\Lambda}_{1,a,a',h}\bar{\boldsymbol{\Theta}}_{1,a,a',h}^\mathsf{T}\right\|_F^2\right]^{1/2}\right). &\text{(A.39)}
\end{aligned}
$$

By Lemma A.4.13, for all $a, a' \in [0, 5]$, as $n, d \to \infty$ we have $\bar{\Phi}_n(1, a, a', h) \to \sup_{q \geq 0} \mathcal{F}(q_\Theta^2 + h, q)$. Notice that the mapping $h \mapsto \bar{\Phi}_n(1, a, a', h)$ is convex and differentiable, thus for $q_\Theta^2 + h \in D$ we can apply Lemma A.2.4 and conclude that

$$
\lim_{n,d\to\infty} \frac{1}{4n^2}\mathbb{E}[\|\mathbb{E}[\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\mathsf{T} \mid \bar{\boldsymbol{A}}, \boldsymbol{x}'(a'), \bar{\boldsymbol{x}}(a), \boldsymbol{Y}'(h)]\|_F^2] = \frac{\partial}{\partial h}\sup_{q \geq 0}\mathcal{F}(q_\Theta^2 + h, q). \tag{A.40}
$$

Leveraging triangle inequality, Eqs. (A.30) and (A.32), we obtain that for all $a, a' \in [0, 5]$, $q_\Theta^2 + h, q_\Theta^2 \in D$

$$
\begin{aligned}
&\limsup_{n,d\to\infty} \frac{1}{n\sqrt{nd}}\mathbb{E}\big[\big\|\boldsymbol{M}_{1,0,0,0}\big\|_F^2\big] = 2D_\Theta(0), \\
&\limsup_{n,d\to\infty} \frac{1}{n\sqrt{nd}}\mathbb{E}\big[\big\|\boldsymbol{\Lambda}_{1,a,a',h}\bar{\boldsymbol{\Theta}}_{1,a,a',h}^\mathsf{T}\big\|_F^2\big] = 2D_\Theta(h), &\text{(A.41)} \\
&\lim_{n,d\to\infty} \frac{1}{n\sqrt{nd}}\mathbb{E}\big[\big\|\boldsymbol{M}_{1,0,0,0} - \boldsymbol{\Lambda}_{1,a,a',h}\bar{\boldsymbol{\Theta}}_{1,a,a',h}\big\|_F^2\big] = 2(D_\Theta(h) - D_\Theta(0)).
\end{aligned}
$$

By Theorem A.4.1 and Eq. (A.33) we have

$$
\limsup_{n,d\to\infty} \mathbb{E}\left[\left|\frac{1}{\sqrt{nd}}\|\bar{\boldsymbol{\Theta}}_{1,a,a',h}\|^2 - C_0(q_\Theta, h)\right|\right] = 0. \tag{A.42}
$$

We plug Eqs. (A.39), (A.41) and (A.42) into Eq. (A.37) and obtain that

$$\begin{aligned}&\limsup_{n,d\to\infty}\frac{1}{n^2}\mathbb{E}\left[\|C_0(q_\Theta,h)\boldsymbol{\Lambda}_{1,a,a',h}\boldsymbol{\Lambda}_{1,a,a',h}^\mathsf{T}-\boldsymbol{M}_0(\bar{\boldsymbol{A}})\|_F^2\right]\\ \le& 24K_1^2C_0(q_\Theta,h)\times\left(D_\Theta(h)-D_\Theta(0)\right)^{1/2}.\end{aligned}\tag{A.43}$$

Using triangle inequality

$$\begin{aligned}&\frac{1}{n}\mathbb{E}\left[\|\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\mathsf{T}-C_0(q_\Theta,h)^{-1}\boldsymbol{M}_0(\bar{\boldsymbol{A}})\|_F^2\right]^{1/2}\\ \le&\frac{1}{n}\mathbb{E}\left[\|\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\mathsf{T}-\mathbb{E}[\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\mathsf{T}\mid\bar{\boldsymbol{A}},\boldsymbol{x}'(a'),\bar{\boldsymbol{x}}(a),\boldsymbol{Y}'(h)]\|_F^2\right]^{1/2}+\frac{1}{n}\mathbb{E}\left[\|\boldsymbol{\Lambda}_{1,a,a',h}\boldsymbol{\Lambda}_{1,a,a',h}^\mathsf{T}-C_0(q_\Theta,h)^{-1}\boldsymbol{M}_0(\bar{\boldsymbol{A}})\|_F^2\right]^{1/2}\\ &+\frac{1}{n}\mathbb{E}\left[\|\mathbb{E}[\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\mathsf{T}\mid\bar{\boldsymbol{A}},\boldsymbol{x}'(a'),\bar{\boldsymbol{x}}(a),\boldsymbol{Y}'(h)]-\boldsymbol{\Lambda}_{1,a,a',h}\boldsymbol{\Lambda}_{1,a,a',h}^\mathsf{T}\|_F^2\right]^{1/2}.\end{aligned}$$

We plug Eqs. (A.36), (A.40) and (A.43) into the above equation and conclude that

$$\begin{aligned}&\limsup_{n,d\to\infty}\frac{1}{n^2}\mathbb{E}\left[\|\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\mathsf{T}-C_0(q_\Theta,h)^{-1}\hat{\boldsymbol{M}}(\bar{\boldsymbol{A}})\|_F^2\right]\\ \le&\mathbb{E}_{\boldsymbol{\Lambda}_0\sim\mu_\Lambda}[\boldsymbol{\Lambda}_0^2]^2-4\frac{\partial}{\partial h}\sup_{q\ge0}\mathcal{F}(q_\Theta^2+h,q)+2K_1^2\sqrt{24K_1^2C_0(q_\Theta,h)^{-1}(D_\Theta(h)-D_\Theta(0))^{1/2}}\\ &+24K_1^2C_0(q_\Theta,h)^{-1}(D_\Theta(h)-D_\Theta(0))^{1/2},\end{aligned}$$

which is an upper bound for $\limsup_{n,d\to\infty}\mathbb{E}\left[\left\|\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\mathsf{T}-\mathbb{E}[\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\mathsf{T}\mid\bar{\boldsymbol{A}}]\right\|_F^2\right]/n^2$. Recall that $D_\Theta(0)>0$, thus $C_0(q_\Theta,h)^{-1}<C_0(q_\Theta,0)^{-1}<\infty$. For all but countably many $q_\Theta>0$ the mapping $h\mapsto D_\Theta(h)$ is continuous at 0. For these q_Θ, if we take $h\to0^+$ while maintaining $h+q_\Theta^2\in D$ then we derive that

$$\limsup_{n,d\to\infty}\frac{1}{n^2}\mathbb{E}\left[\left\|\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\mathsf{T}-\mathbb{E}[\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\mathsf{T}\mid\bar{\boldsymbol{A}}]\right\|_F^2\right]\le\mathbb{E}_{\boldsymbol{\Lambda}_0\sim\mu_\Lambda}[\boldsymbol{\Lambda}_0^2]^2-4\frac{\partial}{\partial h}\sup_{q\ge0}\mathcal{F}(q_\Theta^2+h,q)\Big|_{h=0},$$

thus concludes the proof of the theorem using Lemma A.4.11.

Proof of Theorem 2.4.5 under condition (c)

We define $\tilde{\boldsymbol{Y}}=(\boldsymbol{A}\boldsymbol{A}^\mathsf{T}-d\boldsymbol{I}_n)/\sqrt{nd}$ and $\tilde{\boldsymbol{Y}}'=q_\Theta\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\mathsf{T}/n+(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}-d\boldsymbol{I}_n)/\sqrt{nd}$. One can verify that as $n,d\to\infty$ we have $\|\tilde{\boldsymbol{Y}}-\tilde{\boldsymbol{Y}}'\|_{op}=o_P(1)$. We then run rotationally invariant AMP with spectral initialization based on $\tilde{\boldsymbol{Y}}$. According to [148], for large enough number of iterations this algorithm achieves Bayesian MMSE, thus completing the proof of the theorem under condition (c).

A.5 Convergence of free energy density

A.5.1 Proof of Lemma A.4.2

For $k\in[d]$, we define

$$H_n^{(k)}(\boldsymbol{\lambda},\boldsymbol{\theta}):=\frac{1}{\sqrt{nd}}\sum_{i=1}^n\sum_{j=1,j\neq k}^d\boldsymbol{\Lambda}_i\boldsymbol{\lambda}_i\boldsymbol{\Theta}_j\boldsymbol{\theta}_j+\frac{1}{\sqrt[4]{nd}}\sum_{i=1}^n\sum_{j=1,j\neq k}^d Z_{ij}\boldsymbol{\lambda}_i\boldsymbol{\theta}_j-\frac{1}{2\sqrt{nd}}\sum_{i=1}^n\sum_{j=1,j\neq k}^d\boldsymbol{\lambda}_i^2\boldsymbol{\theta}_j^2$$

$$+ H_n(\boldsymbol{\lambda}; \boldsymbol{Y}'(h)) + H_n(\boldsymbol{\lambda}; \boldsymbol{x}'(s)).$$

Furthermore, we introduce the following distributions

$$\begin{aligned}
\mu_n^{(k,+)}(\mathrm{d}\boldsymbol{\lambda}, \mathrm{d}\boldsymbol{\theta}) &:= \frac{1}{Z_n^{(k,+)}} \exp\left(H_n^{(k)}(\boldsymbol{\lambda}, \boldsymbol{\theta})\right) \mu_\Lambda^{\otimes n}(\mathrm{d}\boldsymbol{\lambda}) P_{\Theta,k}(\mathrm{d}\boldsymbol{\theta}), \\
Z_n^{(k,+)} &:= \int \exp\left(H_n^{(k)}(\boldsymbol{\lambda}, \boldsymbol{\theta})\right) \mu_\Lambda^{\otimes n}(\mathrm{d}\boldsymbol{\lambda}) P_{\Theta,k}(\mathrm{d}\boldsymbol{\theta}), \\
\mu_n^{(k,-)}(\mathrm{d}\boldsymbol{\lambda}, \mathrm{d}\boldsymbol{\theta}) &:= \frac{1}{Z_n^{(k,-)}} \exp\left(H_n^{(k)}(\boldsymbol{\lambda}, \boldsymbol{\theta})\right) \mu_\Lambda^{\otimes n}(\mathrm{d}\boldsymbol{\lambda}) P_{\Theta,k-1}(\mathrm{d}\boldsymbol{\theta}), \\
Z_n^{(k,-)} &:= \int \exp\left(H_n^{(k)}(\boldsymbol{\lambda}, \boldsymbol{\theta})\right) \mu_\Lambda^{\otimes n}(\mathrm{d}\boldsymbol{\lambda}) P_{\Theta,k-1}(\mathrm{d}\boldsymbol{\theta}).
\end{aligned}$$

Note that $\mu_n^{(k,+)}$, $\mu_n^{(k,-)}$, $Z_n^{(k,+)}$ and $Z_n^{(k,-)}$ are all random objects. The following lemma is a straightforward consequence of the above definitions.

Lemma A.5.1. *The following statements are true for all $k \in [d]$:*

1. $Z_n^{(k,+)} = Z_n^{(k,-)}$.
2. *We denote by $\boldsymbol{Z}_{\cdot k} \in \mathbb{R}^n$ the k-th column of $\boldsymbol{Z}$, then $(\mu_n^{(k,+)}, \mu_n^{(k,-)}, Z_n^{(k,+)}, Z_n^{(k,-)})$ are independent of $(\boldsymbol{Z}_{\cdot k}, \boldsymbol{\Theta}_k)$.*
3. *We let $\boldsymbol{\theta}_{-k} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_{k-1}, \boldsymbol{\theta}_{k+1}, \cdots, \boldsymbol{\theta}_d) \in \mathbb{R}^{d-1}$. For $(\boldsymbol{\lambda}, \boldsymbol{\theta}) \sim \mu_n^{(k,+)}$, we have $\boldsymbol{\theta}_k \sim \mu_\Theta$ and is independent of $(\boldsymbol{\theta}_{-k}, \boldsymbol{\lambda})$. Similarly, for $(\boldsymbol{\lambda}, \boldsymbol{\theta}) \sim \mu_n^{(k,-)}$, we have $\boldsymbol{\theta}_k \sim \mathsf{N}(0, q_\Theta)$, and is independent of $(\boldsymbol{\theta}_{-k}, \boldsymbol{\lambda})$.*

We define

$$h_n^{(k)}(\boldsymbol{\lambda}, \boldsymbol{\theta}) := \frac{1}{\sqrt{nd}} \sum_{i=1}^n \boldsymbol{\Lambda}_i \boldsymbol{\lambda}_i \boldsymbol{\Theta}_k \boldsymbol{\theta}_k + \frac{1}{\sqrt[4]{nd}} \sum_{i=1}^n Z_{ik} \boldsymbol{\lambda}_i \boldsymbol{\theta}_k - \frac{1}{2\sqrt{nd}} \sum_{i=1}^n \boldsymbol{\lambda}_i^2 \boldsymbol{\theta}_k^2.$$

For some random variable X, we denote by $\mu_n^{(k,+)}[X]$, $\mu_n^{(k,-)}[X]$ the expectations of X evaluated under distributions $\mu_n^{(k,+)}$ and $\mu_n^{(k,-)}$, respectively. Using Lemma A.5.1, we obtain that

$$\begin{aligned}
&\Phi_n^{(k)} - \Phi_n^{(k-1)} \\
=&\frac{1}{n}\mathbb{E}\left[\log\left(\int \exp(h_n^{(k)}(\boldsymbol{\lambda}, \boldsymbol{\theta}))\mu_n^{(k,+)}(\mathrm{d}\boldsymbol{\lambda}, \mathrm{d}\boldsymbol{\theta}) Z_n^{(k,+)}\right)\right] - \frac{1}{n}\mathbb{E}\left[\log\left(\int \exp(h_n^{(k)}(\boldsymbol{\lambda}, \boldsymbol{\theta}))\mu_n^{(k,-)}(\mathrm{d}\boldsymbol{\lambda}, \mathrm{d}\boldsymbol{\theta}) Z_n^{(k,-)}\right)\right] \\
\overset{(i)}{=}&\frac{1}{n}\mathbb{E}\left[\log\left(\mu_n^{(k,+)}\left[\exp(h_n^{(k)}(\boldsymbol{\lambda}, \boldsymbol{\theta}))\right]\right)\right] - \frac{1}{n}\mathbb{E}\left[\log\left(\mu_n^{(k,-)}\left[\exp(h_n^{(k)}(\boldsymbol{\lambda}, \boldsymbol{\theta}))\right]\right)\right], \qquad \text{(A.44)}
\end{aligned}$$

where *(i)* is by result 1 in Lemma A.5.1. We consider the following Taylor expansion:

$$\exp\left(h_n^{(k)}(\boldsymbol{\lambda}, \boldsymbol{\theta})\right) = \sum_{l=0}^\infty \frac{1}{l!} h_n^{(k)}(\boldsymbol{\lambda}, \boldsymbol{\theta})^l = 1 + \sum_{l=1}^4 c_l^{(k)} \boldsymbol{\theta}_k^l + R^{(k)} + \sum_{l=5}^\infty \frac{1}{l!} h_n^{(k)}(\boldsymbol{\lambda}, \boldsymbol{\theta})^l, \qquad \text{(A.45)}$$

where

$$c_1^{(k)} = \frac{1}{\sqrt{nd}} \langle \boldsymbol{\Lambda}, \boldsymbol{\lambda} \rangle \boldsymbol{\Theta}_k + \frac{1}{\sqrt[4]{nd}} \langle \boldsymbol{Z}_{\cdot k}, \boldsymbol{\lambda} \rangle,$$

$$
\begin{aligned}
c_2^{(k)} =& \frac{1}{2}\left(\frac{1}{\sqrt{nd}}\langle \boldsymbol{\Lambda}, \boldsymbol{\lambda}\rangle \boldsymbol{\Theta}_k + \frac{1}{\sqrt[4]{nd}}\langle \boldsymbol{Z}_{\cdot k}, \boldsymbol{\lambda}\rangle\right)^2 - \frac{1}{2\sqrt{nd}}\langle \boldsymbol{\lambda}, \boldsymbol{\lambda}\rangle, \\
c_3^{(k)} =& -\frac{1}{2nd}\langle \boldsymbol{\Lambda}, \boldsymbol{\lambda}\rangle\langle \boldsymbol{\lambda}, \boldsymbol{\lambda}\rangle \boldsymbol{\Theta}_k - \frac{1}{2n^{3/4}d^{3/4}}\langle \boldsymbol{Z}_{\cdot k}, \boldsymbol{\lambda}\rangle\langle \boldsymbol{\lambda}, \boldsymbol{\lambda}\rangle + \frac{1}{6}\left(\frac{1}{\sqrt{nd}}\langle \boldsymbol{\Lambda}, \boldsymbol{\lambda}\rangle \boldsymbol{\Theta}_k + \frac{1}{\sqrt[4]{nd}}\langle \boldsymbol{Z}_{\cdot k}, \boldsymbol{\lambda}\rangle\right)^3, \\
c_4^{(k)} =& \frac{1}{8nd}\langle \boldsymbol{\lambda}, \boldsymbol{\lambda}\rangle^2 - \frac{1}{4\sqrt{nd}}\langle \boldsymbol{\lambda}, \boldsymbol{\lambda}\rangle\left(\frac{1}{\sqrt{nd}}\langle \boldsymbol{\Lambda}, \boldsymbol{\lambda}\rangle \boldsymbol{\Theta}_k + \frac{1}{\sqrt[4]{nd}}\langle \boldsymbol{Z}_{\cdot k}, \boldsymbol{\lambda}\rangle\right)^2 \\
&+ \frac{1}{24}\left(\frac{1}{\sqrt{nd}}\langle \boldsymbol{\Lambda}, \boldsymbol{\lambda}\rangle \boldsymbol{\Theta}_k + \frac{1}{\sqrt[4]{nd}}\langle \boldsymbol{Z}_{\cdot k}, \boldsymbol{\lambda}\rangle\right)^4, \\
R^{(k)} =& -\frac{\langle \boldsymbol{\lambda}, \boldsymbol{\lambda}\rangle^3}{48(nd)^{3/2}}\boldsymbol{\theta}_k^6 + \frac{1}{8nd}\langle \boldsymbol{\lambda}, \boldsymbol{\lambda}\rangle^2\left(\frac{1}{\sqrt{nd}}\langle \boldsymbol{\Lambda}, \boldsymbol{\lambda}\rangle \boldsymbol{\Theta}_k + \frac{1}{\sqrt[4]{nd}}\langle \boldsymbol{Z}_{\cdot k}, \boldsymbol{\lambda}\rangle\right)\boldsymbol{\theta}_k^5 + \frac{1}{384n^2d^2}\langle \boldsymbol{\lambda}, \boldsymbol{\lambda}\rangle^4\boldsymbol{\theta}_k^8 \\
&- \frac{1}{12\sqrt{nd}}\langle \boldsymbol{\lambda}, \boldsymbol{\lambda}\rangle\left(\frac{1}{\sqrt{nd}}\langle \boldsymbol{\Lambda}, \boldsymbol{\lambda}\rangle \boldsymbol{\Theta}_k + \frac{1}{\sqrt[4]{nd}}\langle \boldsymbol{Z}_{\cdot k}, \boldsymbol{\lambda}\rangle\right)^3\boldsymbol{\theta}_k^5 \\
&+ \frac{1}{16nd}\langle \boldsymbol{\lambda}, \boldsymbol{\lambda}\rangle^2\left(\frac{1}{\sqrt{nd}}\langle \boldsymbol{\Lambda}, \boldsymbol{\lambda}\rangle \boldsymbol{\Theta}_k + \frac{1}{\sqrt[4]{nd}}\langle \boldsymbol{Z}_{\cdot k}, \boldsymbol{\lambda}\rangle\right)^2\boldsymbol{\theta}_k^6 \\
&- \frac{1}{48(nd)^{3/2}}\langle \boldsymbol{\lambda}, \boldsymbol{\lambda}\rangle^3\left(\frac{1}{\sqrt{nd}}\langle \boldsymbol{\Lambda}, \boldsymbol{\lambda}\rangle \boldsymbol{\Theta}_k + \frac{1}{\sqrt[4]{nd}}\langle \boldsymbol{Z}_{\cdot k}, \boldsymbol{\lambda}\rangle\right)\boldsymbol{\theta}_k^7.
\end{aligned}
$$

The next lemma characterizes convergence of power series:

Lemma A.5.2. *For n, d large enough, the following quantities almost surely exist and are finite:*

$$
\sum_{l=5}^{\infty}\frac{1}{l!}\mu_n^{(k,+)}\left[|h_n^{(k)}(\boldsymbol{\lambda}, \boldsymbol{\theta})|^l\right], \qquad \sum_{l=5}^{\infty}\frac{1}{l!}\mu_n^{(k,-)}\left[|h_n^{(k)}(\boldsymbol{\lambda}, \boldsymbol{\theta})|^l\right].
$$

Proof. We will only prove the lemma for $\mu_n^{(k,-)}$. Proof for $\mu_n^{(k,+)}$ is analogous and we skip it for simplicity. By the power mean inequality we have

$$
\begin{aligned}
&\sum_{l=5}^{\infty}\frac{1}{l!}|h_n^{(k)}(\boldsymbol{\lambda}, \boldsymbol{\theta})|^l \\
\leq& \sum_{l=5}^{\infty}\frac{3^l}{l!}\left\{\left|\frac{1}{\sqrt{nd}}\langle \boldsymbol{\Lambda}, \boldsymbol{\lambda}\rangle \boldsymbol{\Theta}_k\boldsymbol{\theta}_k\right|^l + \left|\frac{1}{\sqrt[4]{nd}}\langle \boldsymbol{Z}_{\cdot k}, \boldsymbol{\lambda}\rangle \boldsymbol{\theta}_k\right|^l + \left|\frac{1}{2\sqrt{nd}}\langle \boldsymbol{\lambda}, \boldsymbol{\lambda}\rangle \boldsymbol{\theta}_k^2\right|^l\right\}.
\end{aligned}
$$

Next, we take the expectation of the last line above with respect to $\mu_n^{(k,-)}$, which gives

$$
\begin{aligned}
&\sum_{l=5}^{\infty}\frac{1}{l!}\mu_n^{(k,-)}\left[|h_n^{(k)}(\boldsymbol{\lambda}, \boldsymbol{\theta})|^l\right] \\
\overset{(i)}{\leq}& \sum_{l=5}^{\infty}\left\{\frac{3^l n^{l/2}K^{2l}|\boldsymbol{\Theta}_k|^l l!!q_\Theta^{l/2}}{l!d^{l/2}} + \frac{3^l n^{3l/4}\|\boldsymbol{Z}_{\cdot k}\|_\infty^l K^l l!!q_\Theta^{l/2}}{l!d^{l/4}} + \frac{3^l n^{l/2}K^{2l}(2l)!!q_\Theta^l}{l!d^{l/2}}\right\} \overset{(ii)}{<} \infty,
\end{aligned}
$$

where *(i)* is by Assumption A.4.1 and the third result of Lemma A.5.1. In order to prove *(ii)*, we only need to use the following fact: For n, d large enough we have

$$
\frac{6q_\Theta K^2 n^{1/2}}{d^{1/2}} < 1.
$$

□

According to Lemma A.5.2, we can take the expectation of Eq. (A.45) with respect to $\mu_n^{(k,+)}$, which gives

$$\begin{aligned}&\mu_n^{(k,+)}\left[\exp\left(h_n^{(k)}(\boldsymbol{\lambda},\boldsymbol{\theta})\right)\right]\\ =&1+\mu_n^{(k,+)}[c_2^{(k)}]q_\Theta+\mu_n^{(k,+)}[c_4^{(k)}]\mathbb{E}[\boldsymbol{\Theta}_0^4]+\mu_n^{(k,+)}[R^{(k)}]+\mu_n^{(k,+)}\Big[\sum_{l=5}^{\infty}\frac{1}{l!}h_n^{(k)}(\boldsymbol{\lambda},\boldsymbol{\theta})^l\Big].\end{aligned} \tag{A.46}$$

In the above derivation, we use the fact that under $\mu_n^{(k,+)}$, we have $\boldsymbol{\theta}_k \overset{d}{=} \mu_\Theta$, which has zero first and third moments. Using Assumption A.4.1, we conclude that $c_2^{(k)} \geq -\frac{1}{2}\sqrt{\frac{n}{d}}K^2$. Furthermore, notice that

$$\frac{3}{8nd}\langle\boldsymbol{\lambda},\boldsymbol{\lambda}\rangle^2-\frac{1}{4\sqrt{nd}}\langle\boldsymbol{\lambda},\boldsymbol{\lambda}\rangle\left(\frac{1}{\sqrt{nd}}\langle\boldsymbol{\Lambda},\boldsymbol{\lambda}\rangle\boldsymbol{\Theta}_k+\frac{1}{\sqrt[4]{nd}}\langle\boldsymbol{Z}_{\cdot k},\boldsymbol{\lambda}\rangle\right)^2+\frac{1}{24}\left(\frac{1}{\sqrt{nd}}\langle\boldsymbol{\Lambda},\boldsymbol{\lambda}\rangle\boldsymbol{\Theta}_k+\frac{1}{\sqrt[4]{nd}}\langle\boldsymbol{Z}_{k\cdot},\boldsymbol{\lambda}\rangle\right)^4$$

is non-negative, thus we have $c_4^{(k)} \geq -\frac{1}{4nd}\langle\boldsymbol{\lambda},\boldsymbol{\lambda}\rangle^2 \geq -\frac{n}{4d}K^4$. Since $\boldsymbol{\theta}_k$ has zero expectation under $\mu_n^{(k,+)}$, we then conclude that $\mu_n^{(k,+)}[h_n^{(k)}(\boldsymbol{\lambda},\boldsymbol{\theta})] = -\frac{1}{2\sqrt{nd}}\mu_n^{(k,+)}[\langle\boldsymbol{\lambda},\boldsymbol{\lambda}\rangle\boldsymbol{\theta}_k^2] \geq -\frac{1}{2}\sqrt{\frac{n}{d}}K^2 q_\Theta$. By Jensen's inequality

$$\mu_n^{(k,+)}\left[\exp\left(h_n^{(k)}(\boldsymbol{\lambda},\boldsymbol{\theta})\right)\right] \geq \exp\left(\mu_n^{(k,+)}\left[h_n^{(k)}(\boldsymbol{\lambda},\boldsymbol{\theta})\right]\right) \geq \exp\left(-\frac{1}{2}\sqrt{\frac{n}{d}}K^2 q_\Theta\right).$$

Note that the function $x \mapsto \log(x)$ is concave. We next plug the lower bounds derived above into Eq. (A.46), and obtain

$$\begin{aligned}&\left|\log\left(\mu_n^{(k,+)}\left[\exp\left(h_n^{(k)}(\boldsymbol{\lambda},\boldsymbol{\theta})\right)\right]\right)-\log\left(1+\mu_n^{(k,+)}[c_2^{(k)}]q_\Theta+\mu_n^{(k,+)}[c_4^{(k)}]\mathbb{E}[\boldsymbol{\Theta}_0^4]\right)\right|\\ \leq&\left|\sum_{l=5}^{\infty}\mu_n^{(k,+)}\left[\frac{1}{l!}h_n^{(k)}(\boldsymbol{\lambda},\boldsymbol{\theta})^l\right]+\mu_n^{(k,+)}[R^{(k)}]\right|\times\\ &\max\left\{\mu_n^{(k,+)}\left[\exp\left(h_n^{(k)}(\boldsymbol{\lambda},\boldsymbol{\theta})\right)\right]^{-1},\left(1+\mu_n^{(k,+)}[c_2^{(k)}]q_\Theta+\mu_n^{(k,+)}[c_4^{(k)}]\mathbb{E}[\boldsymbol{\Theta}_0^4]\right)^{-1}\right\}\\ \leq&\underbrace{\left|\sum_{l=5}^{\infty}\mu_n^{(k,+)}\left[\frac{1}{l!}h_n^{(k)}(\boldsymbol{\lambda},\boldsymbol{\theta})^l\right]+\mu_n^{(k,+)}[R^{(k)}]\right|}_{I}\times\\ &\underbrace{\max\left\{\exp\left(\frac{1}{2}\sqrt{\frac{n}{d}}K^2q_\Theta\right),\left(1-\frac{1}{2}\sqrt{\frac{n}{d}}K^2q_\Theta-\frac{n}{4d}K^4\mathbb{E}[\boldsymbol{\Theta}_0^4]\right)^{-1}\right\}}_{II}.\end{aligned} \tag{A.47}$$

Since $d \gg n$ and $(K, q_\Theta, \mathbb{E}[\boldsymbol{\Theta}_0^4])$ are independent of (n,d), we obtain that term II above converges to 1 as $n, d \to \infty$. Next, we will provide an upper bound for term I. To this end, we upper bound $\mathbb{E}[|\mu_n^{(k,+)}[\sum_{l=5}^{\infty}\frac{1}{l!}h_n^{(k)}(\boldsymbol{\lambda},\boldsymbol{\theta})^l]|]$ and $\mathbb{E}[|\mu_n^{(k,+)}[R^{(k)}]|]$ in Appendix A.5.1 and Appendix A.5.1, respectively, and combine them to finish the proof in Appendix A.5.1.

Upper bounding $\mathbb{E}[|\mu_n^{(k,+)}[\sum_{l=5}^{\infty}\frac{1}{l!}h_n^{(k)}(\boldsymbol{\lambda},\boldsymbol{\theta})^l]|]$

Since μ_Θ is sub-Gaussian, there exists a constant $C > 0$ depending only on μ_Θ, such that for all $p \in \mathbb{N}_+$, $\mathbb{E}_{\boldsymbol{\Theta}_0\sim\mu_\Theta}[|\boldsymbol{\Theta}_0|^p] \leq C^p p^{p/2}$ and $\mathbb{E}_{G\sim\mathsf{N}(0,1)}[|G|^p] \leq C^p p^{p/2}$. Then for n, d large enough, we have

$$\begin{aligned}
&\mathbb{E}\left[\left|\sum_{l=5}^{\infty}\frac{1}{l!}\mu_n^{(k,+)}\left[h_n^{(k)}(\boldsymbol{\lambda},\boldsymbol{\theta})^l\right]\right|\right] \\
&\overset{(i)}{\leq}\sum_{l=5}^{\infty}\frac{1}{l!}\mathbb{E}\left[\mu_n^{(k,+)}\left[\left|\frac{1}{\sqrt{nd}}\langle\boldsymbol{\Lambda},\boldsymbol{\lambda}\rangle\boldsymbol{\Theta}_k\boldsymbol{\theta}_k+\frac{1}{\sqrt[4]{nd}}\langle\boldsymbol{Z}_{\cdot k},\boldsymbol{\lambda}\rangle\boldsymbol{\theta}_k-\frac{1}{2\sqrt{nd}}\langle\boldsymbol{\lambda},\boldsymbol{\lambda}\rangle\boldsymbol{\theta}_k^2\right|^l\right]\right] \\
&\overset{(ii)}{\leq}\sum_{l=5}^{\infty}\frac{1}{l!}\mathbb{E}\left[\mu_n^{(k,+)}\left[\left|\frac{1}{\sqrt{nd}}\langle\boldsymbol{\Lambda},\boldsymbol{\lambda}\rangle\boldsymbol{\Theta}_k\boldsymbol{\theta}_k+\frac{1}{\sqrt[4]{nd}}\langle\boldsymbol{Z}_{\cdot k},\boldsymbol{\lambda}\rangle\boldsymbol{\theta}_k-\frac{1}{2\sqrt{nd}}\langle\boldsymbol{\lambda},\boldsymbol{\lambda}\rangle\boldsymbol{\theta}_k^2\right|^{2l}\right]\right]^{1/2} \\
&\overset{(iii)}{\leq}\sum_{l=5}^{\infty}\frac{3^l}{l!}\times\left\{\mathbb{E}\left[\mu_n^{(k,+)}\left[\left|\frac{1}{\sqrt{nd}}\langle\boldsymbol{\Lambda},\boldsymbol{\lambda}\rangle\boldsymbol{\Theta}_k\boldsymbol{\theta}_k\right|^{2l}\right]\right]^{1/2}+\mathbb{E}\left[\mu_n^{(k,+)}\left[\left|\frac{1}{\sqrt[4]{nd}}\langle\boldsymbol{Z}_{\cdot k},\boldsymbol{\lambda}\rangle\boldsymbol{\theta}_k\right|^{2l}\right]\right]^{1/2}+\right. \\
&\qquad\left.\mathbb{E}\left[\mu_n^{(k,+)}\left[\left|\frac{1}{2\sqrt{nd}}\langle\boldsymbol{\lambda},\boldsymbol{\lambda}\rangle\boldsymbol{\theta}_k^2\right|^{2l}\right]\right]^{1/2}\right\} \\
&\overset{(iv)}{\leq}\sum_{l=5}^{\infty}\frac{3^l}{l!}\times\left\{\frac{K^{2l}C^{2l}n^{l/2}}{d^{l/2}}\times(2l)^l+\frac{K^lC^{2l}n^{l/4}}{d^{l/4}}\times(2l)^l+\frac{K^{2l}C^{2l}n^{l/2}}{d^{l/2}}\times(2l)^l\right\} \\
&\overset{(v)}{\leq}F_{K,C}\times\frac{n^{5/4}}{d^{5/4}},
\end{aligned}\tag{A.48}$$

where $F_{K,C} > 0$ is a constant that depends only on K and C. In the above inequalities, *(i)* is by triangle inequality, *(ii)* is by Hölder's inequality and *(iii)* is by power mean inequality. Argument *(iv)* is via a combination of the following facts: (1) Support$(\Lambda) \subseteq [-K, K]$, (2) μ_Θ is sub-Gaussian, (3) the random distribution $\mu_n^{(k,+)}$ is independent of $(\boldsymbol{Z}_{\cdot k}, \boldsymbol{\Theta}_k)$.

For illustration, in the following parts of the proof we upper bound the second summand in the second to last line of Eq. (A.48). The proofs for the first and third summands follow analogously.

By Lemma A.5.1, we see that $\mu_n^{(k,+)}$ is independent of $\boldsymbol{Z}_{\cdot k}$, and $\boldsymbol{\theta}_k$ is independent of $\boldsymbol{\lambda}$ under $\mu_n^{(k,+)}$. Therefore, we have

$$\mathbb{E}\left[\mu_n^{(k,+)}\left[\left|\frac{1}{\sqrt[4]{nd}}\langle\boldsymbol{Z}_{\cdot k},\boldsymbol{\lambda}\rangle\boldsymbol{\theta}_k\right|^{2l}\right]\right]=\sum_{s_1=1}^{n}\cdots\sum_{s_{2l}=1}^{n}\mathbb{E}\left[Z_{s_1k}Z_{s_2k}\cdots Z_{s_{2l}k}\right]C_{s_1,s_2,\cdots,s_{2l}},\tag{A.49}$$

where $C_{s_1,s_2,\cdots,s_{2l}} = n^{-l/2}d^{-l/2}\mathbb{E}[\mu_n^{(k,+)}[\lambda_{s_1}\lambda_{s_2}\cdots\lambda_{s_{2l}}]]\mathbb{E}[\boldsymbol{\Theta}_0^{2l}]$. By sub-Gaussian property we have

$$|C_{s_1,s_2,\cdots,s_{2l}}| \leq K^{2l}C^{2l}(2l)^l n^{-l/2}d^{-l/2}.$$

Consider all terms that take the form of $\mathbb{E}\left[Z_{s_1k}Z_{s_2k}\cdots Z_{s_{2l}k}\right]$. If such term is non-zero, then it must be positive. Using property of Gaussian distribution, we have

$$\sum_{s_1=1}^{n}\cdots\sum_{s_{2l}=1}^{n}\mathbb{E}\left[Z_{s_1k}Z_{s_2k}\cdots Z_{s_{2l}k}\right]=\mathbb{E}[(Z_{1k}+\cdots+Z_{nk})^{2l}]=n^l(2l-1)!!\leq n^lC^{2l}(2l)^l.$$

Therefore, the right hand side of Eq. (A.49) has value no larger than

$$n^l C^{2l}(2l)^l \times K^{2l} C^{2l}(2l)^l n^{-l/2} d^{-l/2} = K^{2l} C^{4l}(2l)^{2l} n^{l/2} d^{-l/2},$$

which leads to the desired upper bound for the second summand. Finally, we use Stirling formula and the assumption that $d \gg n$ to prove argument *(v)*.

Upper bounding $\mathbb{E}[|\mu_n^{(k,+)}[R^{(k)}]|]$

Similar to the proof in Appendix A.5.1, we conclude that for n, d large enough

$$\mathbb{E}[|\mu_n^{(k,+)}[R^{(k)}]|] \leq F'_{K,C} \times \frac{n^{5/4}}{d^{5/4}}, \tag{A.50}$$

where $F'_{K,C} > 0$ is a constant depending only on K and C. The derivation of the above upper bound is similar to the derivation of the upper bound for $\mathbb{E}[|\mu_n^{(k,+)}[\sum_{l=5}^{\infty} \frac{1}{l!} h_n^{(k)}(\boldsymbol{\lambda}, \boldsymbol{\theta})^l]|]$ given in Eq. (A.48), and we skip the details here for the sake of simplicity.

Combining the upper bounds

Combining Eqs. (A.47), (A.48) and (A.50), we obtain that for n, d large enough

$$\begin{aligned} &\left| \sum_{k=1}^{d} \frac{1}{n} \mathbb{E}\left[\log\left(\mu_n^{(k,+)}\left[\exp\left(h_n^{(k)}(\boldsymbol{\lambda}, \boldsymbol{\theta}) \right) \right] \right) - \log\left(1 + \mu_n^{(k,+)}[c_2^{(k)}] q_\Theta + \mu_n^{(k,+)}[c_4^{(k)}] \mathbb{E}[\boldsymbol{\Theta}_0^4] \right) \right] \right| \\ \leq & 2(F_{K,C} + F'_{K,C}) \times \frac{n^{1/4}}{d^{1/4}}. \end{aligned} \tag{A.51}$$

In what follows, we show that the following quantity is small:

$$\left| \sum_{k=1}^{d} \frac{1}{n} \mathbb{E}\left[\log\left(1 + \mu_n^{(k,+)}[c_2^{(k)}] q_\Theta + \mu_n^{(k,+)}[c_4^{(k)}] \mathbb{E}[\boldsymbol{\Theta}_0^4] \right) \right] - \sum_{k=1}^{d} \frac{1}{n} \mathbb{E}\left[\log\left(1 + \mu_n^{(k,+)}[c_2^{(k)}] q_\Theta \right) \right] \right|. \tag{A.52}$$

Again we use the concavity of the mapping $x \mapsto \log(x)$, which gives

$$\begin{aligned} & \log\left(1 + \mu_n^{(k,+)}[c_2^{(k)}] q_\Theta \right) + \mu_n^{(k,+)}[c_4^{(k)}] \mathbb{E}[\boldsymbol{\Theta}_0^4] - \frac{\left| \mu_n^{(k,+)}[c_4^{(k)}] \mathbb{E}[\boldsymbol{\Theta}_0^4] \left(\mu_n^{(k,+)}[c_2^{(k)}] q_\Theta + \mu_n^{(k,+)}[c_4^{(k)}] \mathbb{E}[\boldsymbol{\Theta}_0^4] \right) \right|}{1 + \mu_n^{(k,+)}[c_2^{(k)}] q_\Theta + \mu_n^{(k,+)}[c_4^{(k)}] \mathbb{E}[\boldsymbol{\Theta}_0^4]} \\ \leq & \log\left(1 + \mu_n^{(k,+)}[c_2^{(k)}] q_\Theta \right) + \frac{\mu_n^{(k,+)}[c_4^{(k)}] \mathbb{E}[\boldsymbol{\Theta}_0^4]}{1 + \mu_n^{(k,+)}[c_2^{(k)}] q_\Theta + \mu_n^{(k,+)}[c_4^{(k)}] \mathbb{E}[\boldsymbol{\Theta}_0^4]} \\ \leq & \log\left(1 + \mu_n^{(k,+)}[c_2^{(k)}] q_\Theta + \mu_n^{(k,+)}[c_4^{(k)}] \mathbb{E}[\boldsymbol{\Theta}_0^4] \right) \leq \log\left(1 + \mu_n^{(k,+)}[c_2^{(k)}] q_\Theta \right) + \frac{\mu_n^{(k,+)}[c_4^{(k)}] \mathbb{E}[\boldsymbol{\Theta}_0^4]}{1 + \mu_n^{(k,+)}[c_2^{(k)}] q_\Theta} \\ \leq & \log\left(1 + \mu_n^{(k,+)}[c_2^{(k)}] q_\Theta \right) + \mu_n^{(k,+)}[c_4^{(k)}] \mathbb{E}[\boldsymbol{\Theta}_0^4] + \frac{|\mu_n^{(k,+)}[c_2^{(k)}] \mu_n^{(k,+)}[c_4^{(k)}] q_\Theta \mathbb{E}[\boldsymbol{\Theta}_0^4]|}{1 + \mu_n^{(k,+)}[c_2^{(k)}] q_\Theta}. \end{aligned} \tag{A.53}$$

By Lemma A.5.1, $\mu_n^{(k,+)}$ and $\boldsymbol{Z}_{\cdot k}$ are independent of each other, thus

$$\mathbb{E}\left[\mu_n^{(k,+)}\left[\frac{1}{8nd} \langle \boldsymbol{\lambda}, \boldsymbol{\lambda} \rangle^2 - \frac{1}{4nd} \langle \boldsymbol{\lambda}, \boldsymbol{\lambda} \rangle \langle \boldsymbol{Z}_{\cdot k}, \boldsymbol{\lambda} \rangle^2 + \frac{1}{24nd} \langle \boldsymbol{Z}_{\cdot k}, \boldsymbol{\lambda} \rangle^4 \right] \right] = 0. \tag{A.54}$$

Next, we plug Eq. (C.9) into the definition of $c_4^{(k)}$, then apply Lemma A.5.1 claim 3, which gives

$$\begin{aligned}
&\left|\mathbb{E}[\mu_n^{(k,+)}[c_4^{(k)}]]\right| \\
=&\left| -\frac{\mathbb{E}\left[\mu_n^{(k,+)}\left[\langle\boldsymbol{\lambda},\boldsymbol{\lambda}\rangle\langle\boldsymbol{\lambda},\boldsymbol{\Lambda}\rangle^2\boldsymbol{\Theta}_k^2\right]\right]}{4n^{3/2}d^{3/2}} + \frac{\mathbb{E}\left[\mu_n^{(k,+)}\left[\langle\boldsymbol{\Lambda},\boldsymbol{\lambda}\rangle^4\boldsymbol{\Theta}_k^4\right]\right]}{24n^2d^2} + \frac{\mathbb{E}\left[\mu_n^{(k,+)}\left[\langle\boldsymbol{\lambda},\boldsymbol{\Lambda}\rangle^2\langle\boldsymbol{Z}_{\cdot k},\boldsymbol{\lambda}\rangle^2\boldsymbol{\Theta}_k^2\right]\right]}{4n^{3/2}d^{3/2}}\right| \\
\leq&\frac{n^{3/2}}{2d^{3/2}}K^6 q_\Theta + \frac{n^2}{24d^2}K^8\mathbb{E}[\boldsymbol{\Theta}_0^4]. \qquad (A.55)
\end{aligned}$$

In addition, we have the following lemma:

Lemma A.5.3. *There exist constants* $A_1(K,\mu_\Theta), A_2(K,\mu_\Theta) > 0$*, which are functions of* (K,μ_Θ) *only, such that*

$$\mathbb{E}[\mu_n^{(k,+)}[|c_2^{(k)}|^2]] \leq A_1(K,\mu_\Theta) \times \frac{n}{d}, \qquad \mathbb{E}[\mu_n^{(k,+)}[|c_4^{(k)}|^2]] \leq A_2(K,\mu_\Theta) \times \frac{n^2}{d^2}.$$

Proof. Straightforward computation reveals that there exist $A_1'(K,\mu_\Theta), A_2'(K,\mu_\Theta) > 0$ depending only on (K,μ_Θ), such that

$$\begin{aligned}
\mathbb{E}[\mu_n^{(k,+)}[|c_2^{(k)}|^2]] \leq& A_1'(K,\mu_\Theta)\mathbb{E}\left[\mu_n^{(k,+)}\left[\frac{1}{n^2d^2}\langle\boldsymbol{\Lambda},\boldsymbol{\lambda}\rangle^4\Theta_k^4 + \frac{1}{nd}\langle\boldsymbol{Z}_{\cdot k},\boldsymbol{\lambda}\rangle^4 + \frac{1}{nd}\langle\boldsymbol{\lambda},\boldsymbol{\lambda}\rangle^2\right]\right] \\
\mathbb{E}[\mu_n^{(k,+)}[|c_4^{(k)}|^2]] \leq& A_2'(K,\mu_\Theta)\mathbb{E}\left[\mu_n^{(k,+)}\left[\frac{1}{n^2d^2}\langle\boldsymbol{\lambda},\boldsymbol{\lambda}\rangle^4 + \frac{1}{n^3d^3}\langle\boldsymbol{\lambda},\boldsymbol{\lambda}\rangle^2\langle\boldsymbol{\Lambda},\boldsymbol{\lambda}\rangle^4\boldsymbol{\Theta}_k^4 + \frac{1}{n^2d^2}\langle\boldsymbol{\lambda},\boldsymbol{\lambda}\rangle^2\langle\boldsymbol{Z}_{\cdot k},\boldsymbol{\lambda}\rangle^4 + \right.\right. \\
&\left.\left.\frac{1}{n^4d^4}\langle\boldsymbol{\Lambda},\boldsymbol{\lambda}\rangle^8\boldsymbol{\Theta}_k^8 + \frac{1}{n^2d^2}\langle\boldsymbol{Z}_{\cdot k},\boldsymbol{\lambda}\rangle^8\right]\right].
\end{aligned}$$

The rest of the proof follows from Lemma A.5.1 and the assumption that $d \gg n$.

□

Recall that $c_2^{(k)} \geq -\frac{1}{2}\sqrt{\frac{n}{d}}K^2$ and $c_4^{(k)} \geq -\frac{n}{4d}K^4$. Then for n, d large enough, using Lemma A.5.3, we obtain

$$\begin{aligned}
\mathbb{E}\left[\frac{|\mu_n^{(k,+)}[c_2^{(k)}]\mu_n^{(k,+)}[c_4^{(k)}]q_\Theta\mathbb{E}[\boldsymbol{\Theta}_0^4]|}{1+\mu_n^{(k,+)}[c_2^{(k)}]q_\Theta}\right] \leq& \frac{q_\Theta\mathbb{E}[\boldsymbol{\Theta}_0^4]}{1-\frac{1}{2}\sqrt{\frac{n}{d}}K^2q_\Theta} \times \mathbb{E}[\mu_n^{(k,+)}[|c_2^{(k)}|^2]]^{1/2}\mathbb{E}[\mu_n^{(k,+)}[|c_4^{(k)}|^2]]^{1/2}, \\
\leq& \frac{2q_\Theta\mathbb{E}[\boldsymbol{\Theta}_0^4]A_1(K,\mu_\Theta)^{1/2}A_2(K,\mu_\Theta)^{1/2}n^{3/2}}{d^{3/2}}, \qquad (A.56)
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}\left[\frac{|\mu_n^{(k,+)}[c_4^{(k)}]\mathbb{E}[\boldsymbol{\Theta}_0^4]\left(\mu_n^{(k,+)}[c_2^{(k)}]q_\Theta + \mu_n^{(k,+)}[c_4^{(k)}]\mathbb{E}[\boldsymbol{\Theta}_0^4]\right)|}{1+\mu_n^{(k,+)}[c_2^{(k)}]q_\Theta + \mu_n^{(k,+)}[c_4^{(k)}]\mathbb{E}[\boldsymbol{\Theta}_0^4]}\right] \\
\leq&\frac{\mathbb{E}[\boldsymbol{\Theta}_0^4]^2\mathbb{E}[\mu_n^{(k,+)}[|c_4^{(k)}|^2]]}{1-\frac{1}{2}\sqrt{\frac{n}{d}}K^2q_\Theta - \frac{n}{4d}K^4\mathbb{E}[\boldsymbol{\Theta}_0^4]} + \frac{\mathbb{E}[\boldsymbol{\Theta}_0^4]q_\Theta\mathbb{E}[\mu_n^{(k,+)}[|c_2^{(k)}|^2]]^{1/2}\mathbb{E}[\mu_n^{(k,+)}[|c_4^{(k)}|^2]]^{1/2}}{1-\frac{1}{2}\sqrt{\frac{n}{d}}K^2q_\Theta - \frac{n}{4d}K^4\mathbb{E}[\boldsymbol{\Theta}_0^4]} \\
\leq&\frac{2\mathbb{E}[\boldsymbol{\Theta}_0^4]^2A_2(K,\mu_\Theta)n^2}{d^2} + \frac{2\mathbb{E}[\boldsymbol{\Theta}_0^4]q_\Theta A_1(K,\mu_\Theta)^{1/2}A_2(K,\mu_\Theta)^{1/2}n^{3/2}}{d^{3/2}}. \qquad (A.57)
\end{aligned}$$

Next, we plug Eqs. (A.55) to (A.57) into Eq. (A.53), then sum over $k \in [d]$. This implies the existence of $C(K, \mu_\Theta) > 0$, which is a constant depending only on (K, μ_Θ), such that for n, d large enough

$$\begin{aligned}
&\sum_{k=1}^{d} \frac{1}{n} \mathbb{E}\left[\log\left(1 + \mu_n^{(k,+)}[c_2^{(k)}] q_\Theta\right)\right] - \frac{C(K, \mu_\Theta) n^{1/2}}{d^{1/2}} \\
\leq &\sum_{k=1}^{d} \frac{1}{n} \mathbb{E}\left[\log\left(1 + \mu_n^{(k,+)}[c_2^{(k)}] q_\Theta + \mu_n^{(k,+)}[c_4^{(k)}] \mathbb{E}[\boldsymbol{\Theta}_0^4]\right)\right] \\
\leq &\sum_{k=1}^{d} \frac{1}{n} \mathbb{E}\left[\log\left(1 + \mu_n^{(k,+)}[c_2^{(k)}] q_\Theta\right)\right] + \frac{C(K, \mu_\Theta) n^{1/2}}{d^{1/2}}.
\end{aligned} \tag{A.58}$$

Combining Eqs. (A.51) and (A.58), we derive that

$$\sum_{k=1}^{d} \frac{1}{n} \mathbb{E}\left[\log\left(1 + \mu_n^{(k,+)}[c_2^{(k)}] q_\Theta\right)\right] - \sum_{k=1}^{d} \frac{1}{n} \mathbb{E}\left[\log\left(\mu_n^{(k,+)}\left[\exp\left(h_n^{(k)}(\boldsymbol{\lambda}, \boldsymbol{\theta})\right)\right]\right)\right] = o_n(1). \tag{A.59}$$

Similarly, we can prove that

$$\sum_{k=1}^{d} \frac{1}{n} \mathbb{E}\left[\log\left(1 + \mu_n^{(k,-)}[c_2^{(k)}] q_\Theta\right)\right] - \sum_{k=1}^{d} \frac{1}{n} \mathbb{E}\left[\log\left(\mu_n^{(k,-)}\left[\exp\left(h_n^{(k)}(\boldsymbol{\lambda}, \boldsymbol{\theta})\right)\right]\right)\right] = o_n(1). \tag{A.60}$$

Since $c_2^{(k)}$ is independent of $\boldsymbol{\theta}_k$, by Lemma A.5.1 we have $\mu_n^{(k,+)}[c_2^{(k)}] = \mu_n^{(k,-)}[c_2^{(k)}]$. Finally, we combine Eqs. (A.44), (A.60) and (C.14), which gives $\Phi_n^{(d)} - \Phi_n^{(0)} = o_n(1)$. Thus, we have completed the proof of Lemma A.4.2.

A.5.2 Proof of Lemma A.4.3

In this section we prove Lemma A.4.3. Applying Gaussian integration by parts, we obtain that

$$\Phi_n^{(0)}(h, s) = \frac{1}{n} \mathbb{E}\Big[\log\Big(\int \exp\Big(\tilde{H}_n'(\boldsymbol{\lambda}; \boldsymbol{Y}'(h), \boldsymbol{x}'(s)) + H_n(\boldsymbol{\lambda}; \boldsymbol{Y}'(h)) + H_n(\boldsymbol{\lambda}; \boldsymbol{x}'(s))\Big) \mu_\Lambda^{\otimes n}(\mathrm{d}\boldsymbol{\lambda})\Big)\Big],$$

where

$$\begin{aligned}
\tilde{H}_n'(\boldsymbol{\lambda}; \boldsymbol{Y}'(h), \boldsymbol{x}'(s)) = &\frac{\langle \boldsymbol{\Lambda}, \boldsymbol{\lambda}\rangle^2 \|\boldsymbol{\Theta}\|^2/(nd) + 2\langle \boldsymbol{\Lambda}, \boldsymbol{\lambda}\rangle \langle \boldsymbol{\lambda}, \boldsymbol{Z}\boldsymbol{\Theta}\rangle/(nd)^{3/4} + \|\boldsymbol{Z}^\mathsf{T}\boldsymbol{\lambda}\|^2/\sqrt{nd}}{2q_\Theta^{-1} + 2\|\boldsymbol{\lambda}\|^2/\sqrt{nd}} \\
&- \frac{d}{2} \log\left(1 + \frac{q_\Theta}{\sqrt{nd}} \|\boldsymbol{\lambda}\|^2\right).
\end{aligned}$$

By triangle inequality,

$$\begin{aligned}
&\sup_{\|\boldsymbol{\lambda}\|_\infty \leq K} \left| \frac{\langle \boldsymbol{\Lambda}, \boldsymbol{\lambda}\rangle^2 \|\boldsymbol{\Theta}\|^2/(nd)}{2q_\Theta^{-1} + 2\|\boldsymbol{\lambda}\|^2/\sqrt{nd}} - \frac{q_\Theta^2}{2n} \langle \boldsymbol{\Lambda}, \boldsymbol{\lambda}\rangle^2 \right| \\
\leq &\sup_{\|\boldsymbol{\lambda}\|_\infty \leq K} \left| \frac{\langle \boldsymbol{\Lambda}, \boldsymbol{\lambda}\rangle^2 \left(\|\boldsymbol{\Theta}\|^2 - dq_\Theta\right)/(nd)}{2q_\Theta^{-1} + 2\|\boldsymbol{\lambda}\|^2/\sqrt{nd}} \right| + \sup_{\|\boldsymbol{\lambda}\|_\infty \leq K} \left| \frac{q_\Theta^2 \langle \boldsymbol{\Lambda}, \boldsymbol{\lambda}\rangle^2 \|\boldsymbol{\lambda}\|^2/(n^{3/2} d^{1/2})}{2q_\Theta^{-1} + 2\|\boldsymbol{\lambda}\|^2/\sqrt{nd}} \right| \\
\leq &\frac{nK^4 q_\Theta}{2d} \left|\|\boldsymbol{\Theta}\|^2 - dq_\Theta\right| + \frac{q_\Theta^3 n^{3/2} K^6}{2d^{1/2}}.
\end{aligned} \tag{A.61}$$

Furthermore, notice that the following inequalities hold:

$$\sup_{\|\boldsymbol{\lambda}\|_\infty \le K} \left| \frac{\langle \boldsymbol{\Lambda}, \boldsymbol{\lambda}\rangle \langle \boldsymbol{\lambda}, \boldsymbol{Z\Theta}\rangle/(nd)^{3/4}}{q_\Theta^{-1} + \|\boldsymbol{\lambda}\|^2/\sqrt{nd}} \right| \le \frac{n^{3/4} q_\Theta K^3}{d^{3/4}} \|\boldsymbol{Z\Theta}\|, \tag{A.62}$$

$$\sup_{\|\boldsymbol{\lambda}\|_\infty \le K} \left| \frac{\left\|\boldsymbol{Z}^\mathsf{T}\boldsymbol{\lambda}\right\|^2/\sqrt{nd}}{2q_\Theta^{-1} + 2\|\boldsymbol{\lambda}\|^2/\sqrt{nd}} - \frac{\left\|\boldsymbol{Z}^\mathsf{T}\boldsymbol{\lambda}\right\|^2/\sqrt{nd}}{2q_\Theta^{-1}} + \frac{\left\|\boldsymbol{Z}^\mathsf{T}\boldsymbol{\lambda}\right\|^2 \|\boldsymbol{\lambda}\|^2/(nd)}{2q_\Theta^{-2}} \right| \le \frac{n^2 K^6 q_\Theta^3}{2d\sqrt{nd}} \|\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T}\|_{\mathrm{op}}, \tag{A.63}$$

$$\sup_{\|\boldsymbol{\lambda}\|_\infty \le K} \left| \frac{d}{2}\log\left(1 + \frac{q_\Theta}{\sqrt{nd}}\|\boldsymbol{\lambda}\|^2\right) - \frac{dq_\Theta}{2\sqrt{nd}}\|\boldsymbol{\lambda}\|^2 + \frac{q_\Theta^2}{4n}\|\boldsymbol{\lambda}\|^4 \right| \le \frac{q_\Theta^3 n^{3/2} K^6}{6d^{1/2}}, \tag{A.64}$$

$$\sup_{\|\boldsymbol{\lambda}\|_\infty \le K} \left| \frac{q_\Theta^2 \|\boldsymbol{Z}^\mathsf{T}\boldsymbol{\lambda}\|^2 \|\boldsymbol{\lambda}\|^2}{2nd} - \frac{q_\Theta^2}{2n}\|\boldsymbol{\lambda}\|^4 \right| \le \frac{nq_\Theta^2 K^4}{2d}\|\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T} - d\boldsymbol{I}_n\|_{\mathrm{op}}, \tag{A.65}$$

where in Eq. (A.64), we use the fact that for all $x \in [0, \infty)$, there exists $y \in [0, x]$ such that

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3(1+y)^3}.$$

Next, we combine Eqs. (A.61) to (A.65), and conclude that

$$\begin{aligned} &\left|\Phi_n^{(0)}(h,s) - \tilde{\Phi}_n(h,s)\right| \\ \le& \frac{1}{n}\mathbb{E}\left[\frac{nK^4 q_\Theta \left|\|\boldsymbol{\Theta}\|^2 - dq_\Theta\right|}{2d} + \frac{q_\Theta^3 n^{3/2} K^6}{2d^{1/2}} + \frac{n^{3/4} q_\Theta K^3 \|\boldsymbol{Z\Theta}\|}{d^{3/4}} + \frac{n^{3/2}K^6 q_\Theta^3 \|\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T}\|_{\mathrm{op}}}{2d^{3/2}}\right. \\ &\left. + \frac{q_\Theta^3 n^{3/2} K^6}{6d^{1/2}} + \frac{nq_\Theta^2 K^4 \|\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T} - d\boldsymbol{I}_n\|_{\mathrm{op}}}{2d}\right]. \end{aligned} \tag{A.66}$$

Using Lemma A.2.1, we see that

$$\begin{aligned} \mathbb{E}\left[\left\|\frac{1}{d}\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T} - \boldsymbol{I}_n\right\|_{\mathrm{op}}\right] \le& 100\sqrt{\frac{n}{d}} + \int_{100\sqrt{n/d}}^{\infty} \exp\left(-\frac{dx^2}{3200}\right)\mathrm{d}x \\ \le& 100\sqrt{\frac{n}{d}} + \frac{40}{\sqrt{d}}\int_0^\infty \exp\left(-\frac{y^2}{2}\right)\mathrm{d}y. \end{aligned} \tag{A.67}$$

Finally, we combine Eq. (A.66), (A.67), the assumption that $d \gg n$, and conclude that as $n, d \to \infty$, $|\Phi_n^{(0)}(h,s) - \tilde{\Phi}_n(h,s)| = o_n(1)$, thus completing the proof of Lemma A.4.3.

A.5.3 Proof of Lemma A.4.4

Recall that $\boldsymbol{W} \overset{d}{=} \mathrm{GOE}(n)$. Then for all fixed orthogonal matrix $\boldsymbol{O} \in \mathbb{R}^{n\times n}$, $\boldsymbol{O}^\mathsf{T}\boldsymbol{W}\boldsymbol{O} \overset{d}{=} \boldsymbol{W}$ and $\boldsymbol{O}^\mathsf{T}(\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T} - d\boldsymbol{I}_n)\boldsymbol{O} \overset{d}{=} (\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T} - d\boldsymbol{I}_n)$. By orthogonal invariance, we can couple $\left(\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T} - d\boldsymbol{I}_n\right)/\sqrt{nd}$ with $\boldsymbol{W}$ such that they admit the following eigen-decomposition:

$$\frac{1}{\sqrt{nd}}\left(\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T} - d\boldsymbol{I}_n\right) = \boldsymbol{\Omega}^\mathsf{T}\boldsymbol{S}_1\boldsymbol{\Omega}, \qquad \boldsymbol{W} = \boldsymbol{\Omega}^\mathsf{T}\boldsymbol{S}_2\boldsymbol{\Omega}. \tag{A.68}$$

In the above display, $\boldsymbol{\Omega}$ is Haar-distributed on the orthogonal matrix group $\mathcal{O}(n)$, $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are diagonal matrices containing ascendingly ordered eigenvalues of matrices $\left(\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T} - d\boldsymbol{I}_n\right)/\sqrt{nd}$ and $\boldsymbol{W}$, respectively.

Furthermore, $\boldsymbol{S}_1, \boldsymbol{S}_2$ are both independent of $\boldsymbol{\Omega}$. Direct computation implies the following inequality:

$$\begin{aligned}\left|\tilde{\Phi}_n(h,s) - \Phi_n^Y(h,s)\right| \leq & \frac{1}{n}\mathbb{E}\left[\sup_{\|\boldsymbol{\lambda}\|_\infty \leq K}\left|\frac{q_\Theta}{2}\boldsymbol{\lambda}^\mathsf{T}\left(\boldsymbol{W} - \frac{1}{\sqrt{nd}}\left(\boldsymbol{Z}\boldsymbol{Z}^\mathsf{T} - d\boldsymbol{I}_n\right)\right)\boldsymbol{\lambda}\right|\right] \\ \leq & \frac{q_\Theta K^2}{2}\mathbb{E}\left[\|\boldsymbol{S}_1 - \boldsymbol{S}_2\|_{\mathrm{op}}\right].\end{aligned}$$

Let $\sigma_i(\boldsymbol{S}_j)$ be the i-th largest eigenvalue of $\boldsymbol{S}_j$ for $j \in [2]$, then $\|\boldsymbol{S}_1 - \boldsymbol{S}_2\|_{\mathrm{op}} = \max_{i\in[n]}|\sigma_i(\boldsymbol{S}_1) - \sigma_i(\boldsymbol{S}_2)|$. We denote by $\mathrm{ESD}(\boldsymbol{M})$ the empirical spectral distribution of matrix $\boldsymbol{M}$. Then using random matrix theory, $\mathrm{ESD}(\boldsymbol{S}_1)$ and $\mathrm{ESD}(\boldsymbol{S}_2)$ both converge almost surely to the semicircle law (see [13]). Furthermore, according to the results in [11, 167, 115], asymptotically speaking, we have $\sigma_1(\boldsymbol{S}_1), \sigma_1(\boldsymbol{S}_2) \overset{a.s.}{\to} 2$ and $\sigma_n(\boldsymbol{S}_1), \sigma_n(\boldsymbol{S}_2) \overset{a.s.}{\to} -2$. Therefore, we see that $\|\boldsymbol{S}_1 - \boldsymbol{S}_2\|_{\mathrm{op}} \overset{a.s.}{\to} 0$ as $n, d \to \infty$.

By Theorem 1.1 in [17], for all $0 < \varepsilon \leq 1/2$

$$\mathbb{E}[\|\boldsymbol{W}\|_{\mathrm{op}}] \leq (1+\varepsilon)\left\{2\sqrt{1+\frac{1}{n}} + \frac{6}{\sqrt{\log(1+\varepsilon)}}\sqrt{\frac{2\log n}{n}}\right\}.$$

In the above equation, we first let $n \to \infty$ then let $\varepsilon \to 0^+$, which gives $\limsup_{n\to\infty}\mathbb{E}[\|\boldsymbol{S}_2\|_{\mathrm{op}}] \leq 2$. By Fatou's lemma, we further have $\liminf_{n\to\infty}\mathbb{E}[\|\boldsymbol{S}_2\|_{\mathrm{op}}] \geq 2$, thus $\lim_{n\to\infty}\mathbb{E}[\|\boldsymbol{S}_2\|_{\mathrm{op}}] = 2$. By Lemma A.2.1, for any $\varepsilon > 0$, there exists $M > 0$, such that for n, d large enough

$$\mathbb{E}\left[\|\boldsymbol{S}_1\|_{\mathrm{op}}\mathbb{1}\left\{\|\boldsymbol{S}_1\|_{\mathrm{op}} \geq M\right\}\right] < \varepsilon.$$

Dominated convergence theorem gives $\limsup_{n\to\infty}\mathbb{E}[\|\boldsymbol{S}_1\|_{\mathrm{op}}\mathbb{1}\{\|\boldsymbol{S}_1\|_{\mathrm{op}} < M\}] \leq 2$, thus we have $\limsup_{n\to\infty}\mathbb{E}[\|\boldsymbol{S}_1\|_{\mathrm{op}}] \leq 2+\varepsilon$. On the other hand, Fatou's lemma implies $\liminf_{n\to\infty}\mathbb{E}[\|\boldsymbol{S}_1\|_{\mathrm{op}}] \geq 2$, thus $\lim_{n\to\infty}\mathbb{E}[\|\boldsymbol{S}_1\|_{\mathrm{op}}] = 2$. Finally, notice that $\|\boldsymbol{S}_1\|_{\mathrm{op}} + \|\boldsymbol{S}_2\|_{\mathrm{op}} - \|\boldsymbol{S}_1 - \boldsymbol{S}_2\|_{\mathrm{op}} \geq 0$. We then apply Scheffé's lemma to both $\|\boldsymbol{S}_1 - \boldsymbol{S}_2\|_{\mathrm{op}}$ and $\|\boldsymbol{S}_1\|_{\mathrm{op}} + \|\boldsymbol{S}_2\|_{\mathrm{op}} - \|\boldsymbol{S}_1 - \boldsymbol{S}_2\|_{\mathrm{op}}$, which gives $\mathbb{E}[\|\boldsymbol{S}_1 - \boldsymbol{S}_2\|_{\mathrm{op}}] \to 0$. This concludes the proof of Lemma A.4.4.

A.5.4 Proof of Lemma A.4.5

The first claim is a direct consequence of Lemmas A.4.2 to A.4.4. As for the second claim, it is straightforward that the free energy densities $\Phi_n(h,s)$ and $\Phi_n^Y(h,s)$ are well-defined on $[0,\infty)\times[0,\infty)$ and differentiable for all $h, s \in (0,\infty)$.

By Nishimori identity (Lemma A.2.3) and Gaussian integration by parts, we see that for $h, s > 0$,

$$\begin{aligned}\frac{\partial}{\partial h}\Phi_n(h,s) &= \frac{1}{4n^2}\mathbb{E}\left[\langle\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\mathsf{T}, \mathbb{E}[\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\mathsf{T} \mid \boldsymbol{A}, \boldsymbol{Y}'(h), \boldsymbol{x}'(s)]\rangle\right], \\ \frac{\partial}{\partial s}\Phi_n(h,s) &= \frac{1}{2n}\mathbb{E}\left[\langle\boldsymbol{\Lambda}, \mathbb{E}[\boldsymbol{\Lambda} \mid \boldsymbol{A}, \boldsymbol{Y}'(h), \boldsymbol{x}'(s)]\rangle\right].\end{aligned}$$

Recall that h, s stand for the signal-to-noise ratios in the perturbed model. Therefore, we obtain that for fixed $s \geq 0$, $\frac{\partial}{\partial h}\Phi_n(h,s)$ is increasing in h and for fixed $h \geq 0$, $\frac{\partial}{\partial s}\Phi_n(h,s)$ is increasing in s.

As a result, for all fixed $h, s \geq 0$, the mappings $x \mapsto \Phi_n(h,x)$, $x \mapsto \Phi_n(x,s)$ are convex on $(0,\infty)$. Since these mappings are obviously continuous, we obtain that they are convex functions on $[0,\infty)$. Similarly, we can show that for all fixed $h, s \geq 0$, $x \mapsto \Phi_n^Y(h,x)$, $x \mapsto \Phi_n^Y(x,s)$ are convex functions on $[0,\infty)$. This concludes the proof of the second claim.

Finally, we prove the third claim. This proof is based on Guerra's interpolation technique. For $t \in [0,1], x, q \in \mathbb{R}_+$, we define

$$\begin{aligned} H_n^G(\boldsymbol{\lambda}; t, x, q) :=& \frac{xt}{2n}\langle \boldsymbol{\Lambda}, \boldsymbol{\lambda}\rangle^2 + \frac{\sqrt{xt}}{2}\boldsymbol{\lambda}^\mathsf{T}\boldsymbol{W}'\boldsymbol{\lambda} - \frac{xt}{4n}\|\boldsymbol{\lambda}\|^4 + \sqrt{(1-t)xq}\langle \boldsymbol{g}', \boldsymbol{\lambda}\rangle \\ &+ (1-t)xq\langle \boldsymbol{\Lambda}, \boldsymbol{\lambda}\rangle - \frac{(1-t)xq}{2}\|\boldsymbol{\lambda}\|^2, \\ \Psi_n^G(t, x, q) :=& \frac{1}{n}\mathbb{E}\left[\log \int \exp\left(H_n^G(\boldsymbol{\lambda}; t, x, q)\right)\mu_\Lambda^{\otimes n}(\mathrm{d}\boldsymbol{\lambda})\right]. \end{aligned}$$

By [125, Theorem 13], we see that $\lim_{n\to\infty}\Psi_n^G(1, x, q) = \sup_{y \geq 0}\mathcal{F}(x, y)$. Using Lemma A.2.3 and Gaussian integration by parts, we see that

$$\begin{aligned} \frac{\partial}{\partial t}\Psi_n^G(t, x, q) =& \frac{1}{n}\mathbb{E}\left[\frac{x}{4n}\langle \boldsymbol{\Lambda}, \boldsymbol{\lambda}\rangle^2 - \frac{xq}{2}\langle \boldsymbol{\Lambda}, \boldsymbol{\lambda}\rangle \,\Big|\, \frac{\sqrt{xt}}{n}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\mathsf{T} + \boldsymbol{W}', \sqrt{(1-t)xq}\boldsymbol{\Lambda} + \boldsymbol{g}'\right] \\ =& \frac{x}{4}\mathbb{E}\left[\left(\frac{1}{n}\langle \boldsymbol{\Lambda}, \boldsymbol{\lambda}\rangle - q\right)^2 \,\Big|\, \frac{\sqrt{xt}}{n}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\mathsf{T} + \boldsymbol{W}', \sqrt{(1-t)xq}\boldsymbol{\Lambda} + \boldsymbol{g}'\right] - \frac{xq^2}{4} \\ \geq& -\frac{xq^2}{4}. \end{aligned}$$

Direct computation reveals that

$$\Psi_n^G(0, x, q) = \mathcal{F}(x, q) + \frac{xq^2}{4}.$$

Therefore, for all $x, q \geq 0$,

$$\Psi_n^G(1, x, q) = \Psi_n^G(0, x, q) + \int_0^1 \frac{\partial}{\partial t}\Psi_n^G(t, x, q)\mathrm{d}t \geq \mathcal{F}(x, q). \tag{A.69}$$

According to [125, Proposition 17], for all but countably many $x > 0$, $\mathcal{F}(x, \cdot)$ has a unique maximizer $q^*(x)$. For these x, we plug $q = q^*(x)$ into Eq. (A.69), which implies for all but countably many $x > 0$ and all $t \in [0,1]$, $\lim_{n\to\infty}\Psi_n^G(t, x, q^*(x)) = \mathcal{F}(x, q^*(x)) + xq^*(x)^2(1-t)/4$. Notice that $\Psi_n^G(t, x, q^*(x)) = \Phi_n^Y(0, s)$ if $xt = q_\Theta^2$ and $(1-t)xq^*(x) = s$. This concludes the proof of the third claim of the lemma.

A.5.5 Proof of Lemma A.4.6

For $t \in [0,1]$, we define the interpolated Hamiltonian as

$$\begin{aligned} H_{n,t}^\varepsilon(\boldsymbol{\lambda}, \boldsymbol{\theta}; h) :=& \frac{t}{\sqrt{nd}}\langle \boldsymbol{\Lambda}, \boldsymbol{\lambda}\rangle\langle \boldsymbol{\Theta}, \boldsymbol{\theta}\rangle + \frac{\sqrt{t}}{\sqrt[4]{nd}}\boldsymbol{\lambda}^\mathsf{T}\boldsymbol{Z}\boldsymbol{\theta} - \frac{t}{2\sqrt{nd}}\|\boldsymbol{\lambda}\|^2\|\boldsymbol{\theta}\|^2 + \\ & \frac{1-t}{\sqrt{nd}}\langle \bar{\boldsymbol{\Lambda}}, \bar{\boldsymbol{\lambda}}\rangle\langle \boldsymbol{\Theta}, \boldsymbol{\theta}\rangle + \frac{\sqrt{1-t}}{\sqrt[4]{nd}}\bar{\boldsymbol{\lambda}}^\mathsf{T}\boldsymbol{Z}'\boldsymbol{\theta} - \frac{1-t}{2\sqrt{nd}}\|\bar{\boldsymbol{\lambda}}\|^2\|\boldsymbol{\theta}\|^2 + \\ & \frac{ht}{2n}\langle \boldsymbol{\Lambda}, \boldsymbol{\lambda}\rangle^2 + \frac{\sqrt{ht}}{2}\boldsymbol{\lambda}^\mathsf{T}\boldsymbol{W}\boldsymbol{\lambda} - \frac{ht}{4n}\|\boldsymbol{\lambda}\|^4 + \\ & \frac{h(1-t)}{2n}\langle \bar{\boldsymbol{\Lambda}}, \bar{\boldsymbol{\lambda}}\rangle^2 + \frac{\sqrt{h(1-t)}}{2}\bar{\boldsymbol{\lambda}}^\mathsf{T}\boldsymbol{W}'\bar{\boldsymbol{\lambda}} - \frac{h(1-t)}{4n}\|\bar{\boldsymbol{\lambda}}\|^4. \end{aligned}$$

where $\boldsymbol{Z}' = (Z'_{ij})_{i\in[n],j\in[d]}$ is an independent copy of $\boldsymbol{Z}$ and is independent of everything else. We emphasize that $\boldsymbol{Z}, \boldsymbol{Z}', \boldsymbol{W}, \boldsymbol{W}', \boldsymbol{\Lambda}, \boldsymbol{\Theta}$ are mutually independent. Notice that $H^{\varepsilon}_{n,t}(\boldsymbol{\lambda}, \boldsymbol{\theta}; h)$ is the Hamiltonian that corresponds to observations $(\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{Y}_1, \boldsymbol{Y}_2)$

$$\boldsymbol{A}_1 = \frac{\sqrt{t}}{\sqrt[4]{nd}}\boldsymbol{\Lambda}\boldsymbol{\Theta}^{\mathsf{T}} + \boldsymbol{Z}, \qquad \boldsymbol{A}_2 = \frac{\sqrt{1-t}}{\sqrt[4]{nd}}\bar{\boldsymbol{\Lambda}}\boldsymbol{\Theta}^{\mathsf{T}} + \boldsymbol{Z}',$$
$$\boldsymbol{Y}_1 = \frac{\sqrt{ht}}{n}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}} + \boldsymbol{W}, \qquad \boldsymbol{Y}_2 = \frac{\sqrt{h(1-t)}}{n}\bar{\boldsymbol{\Lambda}}\bar{\boldsymbol{\Lambda}}^{\mathsf{T}} + \boldsymbol{W}'.$$

We define the corresponding free energy density

$$\Phi^{\varepsilon}_{n,t}(h) := \frac{1}{n}\mathbb{E}\left[\log\left(\int \exp\left(H^{\varepsilon}_{n,t}(\boldsymbol{\lambda}, \boldsymbol{\theta}; h)\right)\mu_{\Lambda}^{\otimes n}(\mathrm{d}\boldsymbol{\lambda})\mu_{\Theta}^{\otimes d}(\mathrm{d}\boldsymbol{\theta})\right)\right].$$

At the endpoints, we have $\Phi^{\varepsilon}_{n,0}(h) = \bar{\Phi}^{\varepsilon}_{n}(h)$ and $\Phi^{\varepsilon}_{n,1}(h) = \Phi_n(h, 0)$. For simplicity, we denote by $\langle\cdot\rangle_{h,\varepsilon,t}$ the expectation with respect to the posterior distribution $\mathbb{P}(\boldsymbol{\Lambda} = \cdot, \boldsymbol{\Theta} = \cdot \mid \boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{Y}_1, \boldsymbol{Y}_2)$. Using Gaussian integration by parts and Nishimori identity (Lemma A.2.3), we have

$$\begin{aligned}\frac{\partial}{\partial t}\Phi^{\varepsilon}_{n,t}(h) =& \frac{1}{2n\sqrt{nd}}\sum_{i\in[n],j\in[d]}\mathbb{E}\left[\langle(\boldsymbol{\Lambda}_i\boldsymbol{\lambda}_i - \bar{\boldsymbol{\Lambda}}_i\bar{\boldsymbol{\lambda}}_i)\boldsymbol{\Theta}_j\boldsymbol{\theta}_j\rangle_{h,\varepsilon,t}\right] + \frac{h}{4n^2}\mathbb{E}[\langle(\boldsymbol{\Lambda}^{\mathsf{T}}\boldsymbol{\lambda})^2\rangle_{h,\varepsilon,t}] - \frac{h}{4n^2}\mathbb{E}[\langle(\bar{\boldsymbol{\Lambda}}^{\mathsf{T}}\bar{\boldsymbol{\lambda}})^2\rangle_{h,\varepsilon,t}]\\ =& \frac{1}{2\sqrt{nd}}\mathbb{E}\left[\left\langle(\boldsymbol{\Lambda}_1\boldsymbol{\lambda}_1 - \bar{\boldsymbol{\Lambda}}_1\bar{\boldsymbol{\lambda}}_1)\langle\boldsymbol{\Theta}, \boldsymbol{\theta}\rangle\right\rangle_{h,\varepsilon,t}\right] + \frac{h}{4n^2}\sum_{i\in[n],j\in[n]}\mathbb{E}\left[\langle\boldsymbol{\Lambda}_i\boldsymbol{\Lambda}_j\boldsymbol{\lambda}_i\boldsymbol{\lambda}_j - \bar{\boldsymbol{\Lambda}}_i\bar{\boldsymbol{\Lambda}}_j\bar{\boldsymbol{\lambda}}_i\bar{\boldsymbol{\lambda}}_j\rangle_{h,\varepsilon,t}\right].\end{aligned}$$

Next, we provide upper bound for the above partial derivative. Invoking Holder's inequality, we see that

$$\begin{aligned}&\left|\frac{\partial}{\partial t}\Phi^{\varepsilon}_{n,t}(h)\right|\\ \leq& \frac{1}{2\sqrt{nd}}\mathbb{E}\left[\left\langle(\boldsymbol{\Lambda}_1\boldsymbol{\lambda}_1 - \bar{\boldsymbol{\Lambda}}_1\bar{\boldsymbol{\lambda}}_1)^2\right\rangle^{1/2}_{h,\varepsilon,t}\left\langle\langle\boldsymbol{\Theta}, \boldsymbol{\theta}\rangle^2\right\rangle^{1/2}_{h,\varepsilon,t}\right] + \frac{h}{2n^2}\sum_{i\in[n],j\in[n]}\mathbb{E}[(\boldsymbol{\Lambda}_i\boldsymbol{\Lambda}_j - \bar{\boldsymbol{\Lambda}}_i\bar{\boldsymbol{\Lambda}}_j)^2]^{1/2}\mathbb{E}[\boldsymbol{\Lambda}_i^2\boldsymbol{\Lambda}_j^2]^{1/2}\\ \leq& \frac{1}{2\sqrt{nd}}\mathbb{E}\left[\left\langle(\boldsymbol{\Lambda}_1\boldsymbol{\lambda}_1 - \bar{\boldsymbol{\Lambda}}_1\bar{\boldsymbol{\lambda}}_1)^2\right\rangle_{h,\varepsilon,t}\right]^{1/2}\mathbb{E}\left[\left\langle\langle\boldsymbol{\Theta}, \boldsymbol{\theta}\rangle^2\right\rangle_{h,\varepsilon,t}\right]^{1/2} + h\mathbb{E}[\boldsymbol{\Lambda}_1^4]^{3/4}\mathbb{E}[(\boldsymbol{\Lambda}_1 - \bar{\boldsymbol{\Lambda}}_1)^4]^{1/4}.\end{aligned}$$

We denote by $\langle\cdot\rangle_{h,\varepsilon,t,*}$ the expectation with respect to the posterior distribution

$$\mathbb{P}(\boldsymbol{\Theta} = \cdot \mid \boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{\Lambda}, \bar{\boldsymbol{\Lambda}}).$$

Direct computation gives the following inequality:

$$\begin{aligned}\mathbb{E}\left[\left\langle\langle\boldsymbol{\Theta}, \boldsymbol{\theta}\rangle^2\right\rangle_{h,\varepsilon,t,*}\right] =& d\mathbb{E}\left[\left\langle\boldsymbol{\Theta}_1^2\boldsymbol{\theta}_1^2\right\rangle_{h,\varepsilon,t,*}\right] + d(d-1)\mathbb{E}\left[\langle\boldsymbol{\Theta}_1\boldsymbol{\Theta}_2\boldsymbol{\theta}_1\boldsymbol{\theta}_2\rangle_{h,\varepsilon,t,*}\right]\\ \leq& d\mathbb{E}[\boldsymbol{\Theta}_1^4] + d(d-1)\mathbb{E}\left[\langle\boldsymbol{\Theta}_1\boldsymbol{\Theta}_2\boldsymbol{\theta}_1\boldsymbol{\theta}_2\rangle_{h,\varepsilon,t,*}\right].\end{aligned} \tag{A.70}$$

Recall that $r_n = d^{1/4}n^{-1/4}$. We define the mapping

$$F_{\Theta}(\delta) := r_n^2\mathbb{E}[\mathbb{E}[\boldsymbol{\Theta}_0 \mid r_n^{-1}\delta\boldsymbol{\Theta}_0 + \boldsymbol{G}]^2],$$

where $\boldsymbol{\Theta}_0 \sim \mu_\Theta$, $\boldsymbol{G} \sim \mathsf{N}(0,1)$ and $\boldsymbol{\Theta}_0 \perp \boldsymbol{G}$. Notice that

$$\mathbb{E}\Big[\langle \boldsymbol{\Theta}_1\boldsymbol{\Theta}_2\boldsymbol{\theta}_1\boldsymbol{\theta}_2\rangle_{h,\varepsilon,t,*}\Big] \le \frac{n}{d}\mathbb{E}\Big[F_\Theta\Big(\sqrt{(t\|\boldsymbol{\Lambda}\|_2^2+(1-t)\|\bar{\boldsymbol{\Lambda}}\|_2^2)/n}\Big)^2\Big]. \tag{A.71}$$

Direct computation gives

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\delta}\mathbb{E}\left[\boldsymbol{\Theta}_0 \mid \delta\boldsymbol{\Theta}_0+\boldsymbol{G}\right] \\
=&(2\delta\boldsymbol{\Theta}_0+\boldsymbol{G})\operatorname{Var}[\boldsymbol{\Theta}_0 \mid \delta\boldsymbol{\Theta}_0+\boldsymbol{G}] - \delta\mathbb{E}[\boldsymbol{\Theta}_0^3 \mid \delta\boldsymbol{\Theta}_0+\boldsymbol{G}] + \delta\mathbb{E}[\boldsymbol{\Theta}_0^2 \mid \delta\boldsymbol{\Theta}_0+\boldsymbol{G}]\mathbb{E}[\boldsymbol{\Theta}_0 \mid \delta\boldsymbol{\Theta}_0+\boldsymbol{G}],
\end{aligned} \tag{A.72}$$

Next, we apply triangle inequality to upper bound the right hand side of Eq. (A.72), which gives

$$\begin{aligned}
&\left|\frac{\mathrm{d}}{\mathrm{d}\delta}\mathbb{E}\left[\boldsymbol{\Theta}_0 \mid r_n^{-1}\delta\boldsymbol{\Theta}_0+\boldsymbol{G}\right]\right| \\
\le& r_n^{-1} \times \left\{(2r_n^{-1}\delta|\boldsymbol{\Theta}_0|+|\boldsymbol{G}|)\operatorname{Var}\left[\boldsymbol{\Theta}_0 \mid r_n^{-1}\delta\boldsymbol{\Theta}_0+\boldsymbol{G}\right]\right. \\
&\left.+r_n^{-1}\delta\mathbb{E}[|\boldsymbol{\Theta}_0|^3 \mid r_n^{-1}\delta\boldsymbol{\Theta}_0+\boldsymbol{G}] + r_n^{-1}\delta\mathbb{E}[\boldsymbol{\Theta}_0^2 \mid r_n^{-1}\delta\boldsymbol{\Theta}_0+\boldsymbol{G}]\mathbb{E}[|\boldsymbol{\Theta}_0| \mid r_n^{-1}\delta\boldsymbol{\Theta}_0+\boldsymbol{G}]\right\}.
\end{aligned}$$

Leveraging the above formulas and Hölder's inequality, we obtain that for n, d large enough

$$\begin{aligned}
F_\Theta(\delta) =&\sqrt{\frac{d}{n}}\mathbb{E}\left[\mathbb{E}[\boldsymbol{\Theta}_0 \mid r_n^{-1}\delta\boldsymbol{\Theta}_0+\boldsymbol{G}]^2\right] \\
\le&\sqrt{\frac{d}{n}}\mathbb{E}\left[\left(\int_0^\delta\left|\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{E}\left[\boldsymbol{\Theta}_0 \mid r_n^{-1}x\boldsymbol{\Theta}_0+\boldsymbol{G}\right]\right|\mathrm{d}x\right)^2\right] \\
\le&\sqrt{\frac{d}{n}}\mathbb{E}\left[\delta\int_0^\delta\left|\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{E}\left[\boldsymbol{\Theta}_0 \mid r_n^{-1}x\boldsymbol{\Theta}_0+\boldsymbol{G}\right]\right|^2\mathrm{d}x\right] \\
\le&4\delta\mathbb{E}\left[\int_0^\delta 4r_n^{-2}x^2\boldsymbol{\Theta}_0^2\operatorname{Var}\left[\boldsymbol{\Theta}_0 \mid r_n^{-1}x\boldsymbol{\Theta}_0+\boldsymbol{G}\right]^2+\boldsymbol{G}^2\operatorname{Var}\left[\boldsymbol{\Theta}_0 \mid r_n^{-1}x\boldsymbol{\Theta}_0+\boldsymbol{G}\right]^2\right. \\
&\left.+r_n^{-2}x^2\mathbb{E}[|\boldsymbol{\Theta}_0|^3 \mid r_n^{-1}x\boldsymbol{\Theta}_0+\boldsymbol{G}]^2+r_n^{-2}x^2\mathbb{E}[\boldsymbol{\Theta}_0^2 \mid r_n^{-1}x\boldsymbol{\Theta}_0+\boldsymbol{G}]^2\mathbb{E}[|\boldsymbol{\Theta}_0| \mid r_n^{-1}x\boldsymbol{\Theta}_0+\boldsymbol{G}]^2\mathrm{d}x\right] \\
\le&C_{\mu_\Theta}(\delta^4+1).
\end{aligned} \tag{A.73}$$

In the above display, $C_{\mu_\Theta} > 0$ is a constant that depends only on μ_Θ.

We define the set $S = \left\{\|\boldsymbol{\Lambda}\|_2^2 \le n\mathbb{E}[\boldsymbol{\Lambda}_0^2] + n\|\boldsymbol{\Lambda}_0^2\|_{\Psi_1}\right\}$, where $\|\cdot\|_{\Psi_1}$ is the sub-exponential norm of $\boldsymbol{\Lambda}_0^2$. Then by Bernstein's inequality [192, Theorem 2.8.1], we can conclude that there exists a constant $C_{\mu_\Lambda} > 0$ depending only on μ_Λ, such that for all $s \ge 1$,

$$\mathbb{P}\left(\|\boldsymbol{\Lambda}\|_2^2 \ge n\mathbb{E}[\boldsymbol{\Lambda}_0^2] + sn\|\boldsymbol{\Lambda}_0^2\|_{\Psi_1}\right) \le 2\exp\left(-C_{\mu_\Lambda}ns\right).$$

Therefore, for n, d large enough we have

$$\begin{aligned}
&\frac{n}{d}\mathbb{E}\left[F_\Theta\left(\sqrt{(t\|\boldsymbol{\Lambda}\|_2^2+(1-t)\|\bar{\boldsymbol{\Lambda}}\|_2^2)/n}\right)^2\right] \\
=&\frac{n}{d}\mathbb{E}\left[F_\Theta\left(\sqrt{(t\|\boldsymbol{\Lambda}\|_2^2+(1-t)\|\bar{\boldsymbol{\Lambda}}\|_2^2)/n}\right)^2\mathbb{1}_S\right] + \frac{n}{d}\mathbb{E}\left[F_\Theta\left(\sqrt{(t\|\boldsymbol{\Lambda}\|_2^2+(1-t)\|\bar{\boldsymbol{\Lambda}}\|_2^2)/n}\right)^2\mathbb{1}_{S^c}\right]
\end{aligned}$$

$$\begin{aligned}
&\overset{(i)}{\leq}\frac{4C_{\mu_\Theta}^2 n}{d}+\frac{2C_{\mu_\Theta}^2 n}{d}\left(\mathbb{E}[\mathbf{\Lambda}_0^2]+\|\mathbf{\Lambda}_0^2\|_{\Psi_1}\right)^4+\frac{2nC_{\mu_\Theta}^2}{d}\mathbb{E}\left[\left(\|\mathbf{\Lambda}\|_2^2/n\right)^4\mathbb{1}_{S^c}\right]\\
&\leq\frac{4C_{\mu_\Theta}^2 n}{d}+\frac{2C_{\mu_\Theta}^2 n}{d}\left(\mathbb{E}[\mathbf{\Lambda}_0^2]+\|\mathbf{\Lambda}_0^2\|_{\Psi_1}\right)^4+ \qquad (A.74)\\
&\quad\frac{2nC_{\mu_\Theta}^2}{d}\int_1^\infty 4\mathbb{P}\left(\|\mathbf{\Lambda}\|_2^2/n\geq\mathbb{E}[\mathbf{\Lambda}_0^2]+s\|\mathbf{\Lambda}_0^2\|_{\Psi_1}\right)\left(\mathbb{E}[\mathbf{\Lambda}_0^2]+s\|\mathbf{\Lambda}_0^2\|_{\Psi_1}\right)^3\|\mathbf{\Lambda}_0^2\|_{\Psi_1}\mathrm{d}s\\
&\leq\frac{4C_{\mu_\Theta}^2 n}{d}+\frac{2C_{\mu_\Theta}^2 n}{d}\left(\mathbb{E}[\mathbf{\Lambda}_0^2]+\|\mathbf{\Lambda}_0^2\|_{\Psi_1}\right)^4+\frac{2nC_{\mu_\Theta}^2}{d}\int_1^\infty 8\exp\left(-C_{\mu_\Lambda}ns\right)\left(\mathbb{E}[\mathbf{\Lambda}_0^2]+s\|\mathbf{\Lambda}_0^2\|_{\Psi_1}\right)^3\|\mathbf{\Lambda}_0^2\|_{\Psi_1}\mathrm{d}s\\
&\leq\frac{C_1 n}{d}, \qquad (A.75)
\end{aligned}$$

where $C_1 > 0$ is a constant depending only on $(\mu_\Theta, \mu_\Lambda)$, and in *(i)* we use Eq. (A.73). Furthermore,

$$\begin{aligned}
\mathbb{E}\left[\left\langle(\mathbf{\Lambda}_1\boldsymbol{\lambda}_1-\bar{\mathbf{\Lambda}}_1\bar{\boldsymbol{\lambda}}_1)^2\right\rangle_{h,\varepsilon,t}\right]\leq&2\mathbb{E}\left[\mathbf{\Lambda}_1^2\langle(\boldsymbol{\lambda}_1-\bar{\boldsymbol{\lambda}}_1)^2\rangle_{h,\varepsilon,t}\right]+2\mathbb{E}\left[(\mathbf{\Lambda}_1-\bar{\mathbf{\Lambda}}_1)^2\langle\bar{\boldsymbol{\lambda}}_1^2\rangle_{h,\varepsilon,t}\right]\\
\leq&2\mathbb{E}\left[\mathbf{\Lambda}_1^4\right]^{1/2}\mathbb{E}\left[\langle(\boldsymbol{\lambda}_1-\bar{\boldsymbol{\lambda}}_1)^2\rangle_{h,\varepsilon,t}^2\right]^{1/2}+2\mathbb{E}\left[(\mathbf{\Lambda}_1-\bar{\mathbf{\Lambda}}_1)^4\right]^{1/2}\mathbb{E}\left[\langle\bar{\boldsymbol{\lambda}}_1^2\rangle_{h,\varepsilon,t}^2\right]^{1/2}\\
\leq&2\mathbb{E}\left[\mathbf{\Lambda}_1^4\right]^{1/2}\mathbb{E}\left[(\mathbf{\Lambda}_1-\bar{\mathbf{\Lambda}}_1)^4\right]^{1/2}+2\mathbb{E}\left[(\mathbf{\Lambda}_1-\bar{\mathbf{\Lambda}}_1)^4\right]^{1/2}\mathbb{E}\left[\bar{\mathbf{\Lambda}}_1^4\right]\\
\leq&C_2\sqrt{\varepsilon}, \qquad (A.76)
\end{aligned}$$

where $C_2 > 0$ is a constant depending only on μ_Λ. Finally, we combine Eqs. (A.70), (A.71), (A.75) and (A.76), and conclude that

$$\left|\frac{\partial}{\partial t}\Phi_{n,t}^{\varepsilon}(h)\right|\leq C_0\sqrt[4]{\varepsilon}$$

for all $t, h \in [0, 1]$, where $C_0 > 0$ is a constant depending only on $(\mu_\Theta, \mu_\Lambda)$. This concludes the proof of the lemma.

A.6 Achieving the Bayesian MMSE

In this section we prove the technical lemmas required to prove Theorem 2.4.5.

A.6.1 Proof of Lemma A.4.9

We define the set

$$\Omega := \left\{|\mathbf{\Lambda}_i|\leq 2K_0\sqrt{\log n} : i\in[n]\right\}.$$

By Eq. (A.16) we see that $\mathbb{P}(\Omega^c) \leq 2n^{-3}$. Furthermore, on Ω we have $\boldsymbol{A} = \bar{\boldsymbol{A}}$. For matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, we define the mapping $\bar{\boldsymbol{M}} : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times n}$ such that $\bar{\boldsymbol{M}}(\boldsymbol{X}) = \mathbb{E}[\bar{\mathbf{\Lambda}}\bar{\mathbf{\Lambda}}^\mathsf{T} \mid \bar{\boldsymbol{A}} = \boldsymbol{X}]$. Then we have

$$\begin{aligned}
\frac{1}{n}\mathbb{E}\left[\left\|\mathbf{\Lambda}\mathbf{\Lambda}^\mathsf{T}-\mathbb{E}[\mathbf{\Lambda}\mathbf{\Lambda}^\mathsf{T}\mid\boldsymbol{A}]\right\|_F^2\right]^{1/2}\overset{(i)}{\leq}&\frac{1}{n}\mathbb{E}\left[\left\|\mathbf{\Lambda}\mathbf{\Lambda}^\mathsf{T}-\bar{\boldsymbol{M}}(\boldsymbol{A})\right\|_F^2\right]^{1/2}\\
\overset{(ii)}{\leq}&\frac{1}{n}\mathbb{E}\left[\left\|\bar{\mathbf{\Lambda}}\bar{\mathbf{\Lambda}}^\mathsf{T}-\bar{\boldsymbol{M}}(\boldsymbol{A})\right\|_F^2\right]^{1/2}+\frac{1}{n}\mathbb{E}\left[\left\|\bar{\mathbf{\Lambda}}\bar{\mathbf{\Lambda}}^\mathsf{T}-\mathbf{\Lambda}\mathbf{\Lambda}^\mathsf{T}\right\|_F^2\right]^{1/2}, \qquad (A.77)
\end{aligned}$$

where *(i)* is by the fact that the posterior expectation achieves Bayesian MMSE, and *(ii)* is by triangle inequality. Applying triangle inequality and Hölder's inequality, we have

$$
\begin{aligned}
\frac{1}{n}\mathbb{E}\left[\left\|\bar{\boldsymbol{\Lambda}}\bar{\boldsymbol{\Lambda}}^{\mathsf{T}} - \bar{\boldsymbol{M}}(\boldsymbol{A})\right\|_F^2\right]^{1/2} \leq & \frac{1}{n}\mathbb{E}\left[\left\|\bar{\boldsymbol{\Lambda}}\bar{\boldsymbol{\Lambda}}^{\mathsf{T}} - \bar{\boldsymbol{M}}(\boldsymbol{A})\right\|_F^2 \mathbb{1}_{\Omega}\right]^{1/2} + \frac{1}{n}\mathbb{E}\left[\left\|\bar{\boldsymbol{\Lambda}}\bar{\boldsymbol{\Lambda}}^{\mathsf{T}} - \bar{\boldsymbol{M}}(\boldsymbol{A})\right\|_F^2 \mathbb{1}_{\Omega^c}\right]^{1/2} \\
\leq & \frac{1}{n}\mathbb{E}\left[\left\|\bar{\boldsymbol{\Lambda}}\bar{\boldsymbol{\Lambda}}^{\mathsf{T}} - \bar{\boldsymbol{M}}(\bar{\boldsymbol{A}})\right\|_F^2\right]^{1/2} + \frac{1}{n}\mathbb{E}\left[\left\|\bar{\boldsymbol{\Lambda}}\bar{\boldsymbol{\Lambda}}^{\mathsf{T}} - \bar{\boldsymbol{M}}(\boldsymbol{A})\right\|_F^4\right]^{1/4} \mathbb{P}(\Omega^c)^{1/4}. \quad \text{(A.78)}
\end{aligned}
$$

Direct computation reveals that $\mathbb{E}[\|\bar{\boldsymbol{\Lambda}}\bar{\boldsymbol{\Lambda}}^{\mathsf{T}} - \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}}\|_F^2]^{1/2}/n = o_n(1)$ as $n, d \to \infty$, and $\mathbb{E}[\|\bar{\boldsymbol{\Lambda}}\bar{\boldsymbol{\Lambda}}^{\mathsf{T}} - \bar{M}(\boldsymbol{A})\|_F^4]^{1/4}/n \leq 8K_0^2 \log n$. As a result, we conclude that $\mathbb{E}[\|\bar{\boldsymbol{\Lambda}}\bar{\boldsymbol{\Lambda}}^{\mathsf{T}} - \bar{M}(\boldsymbol{A})\|_F^4]^{1/4}\mathbb{P}(\Omega^c)^{1/4}/n = o_n(1)$. Combining these analysis with Eqs. (A.77) and (A.78) concludes the proof of the lemma.

A.6.2 Proof of Lemma A.4.10

Let $\boldsymbol{W}, \boldsymbol{W}' \overset{iid}{\sim} \mathrm{GOE}(n)$ that are independent of $\boldsymbol{\Lambda}$. For $t \in [0,1]$, $s \geq 0$, we define

$$
\begin{aligned}
\boldsymbol{Y}_{a,t}^{(s)} &:= \frac{q_\Theta \sqrt{(1-t)s}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}}}{n} + \boldsymbol{W}, \\
\boldsymbol{Y}_{b,t}^{(s)} &:= \frac{\sqrt{ts}}{n}(q_\Theta^{1/2}\bar{\boldsymbol{\Lambda}} + r_n^{-1}\bar{\boldsymbol{g}})(q_\Theta^{1/2}\bar{\boldsymbol{\Lambda}} + r_n^{-1}\bar{\boldsymbol{g}})^{\mathsf{T}} + \boldsymbol{W}'.
\end{aligned}
$$

For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we define the corresponding truncated vectors $\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}} \in \mathbb{R}^n$ such that $\bar{x}_i = x_i \mathbb{1}\{|x_i| \leq 2K_0\sqrt{\log n}\}$ and $\bar{y}_i = y_i \mathbb{1}\{|y_i| \leq C_3\sqrt{\log n}\}$ for all $i \in [n]$. The Hamiltonian that corresponds to $(\boldsymbol{Y}_{a,t}^{(s)}, \boldsymbol{Y}_{b,t}^{(s)})$ can be expressed as

$$
\begin{aligned}
& H_{n,t}^{(s)}(\boldsymbol{x}, \boldsymbol{y}) \\
:= & \frac{ts}{2n}((q_\Theta^{1/2}\bar{\boldsymbol{\Lambda}} + r_n^{-1}\bar{\boldsymbol{g}})^{\mathsf{T}}(q_\Theta^{1/2}\bar{\boldsymbol{x}} + r_n^{-1}\bar{\boldsymbol{y}}))^2 + \frac{\sqrt{ts}}{2}(q_\Theta^{1/2}\bar{\boldsymbol{x}} + r_n^{-1}\bar{\boldsymbol{y}})^{\mathsf{T}}\boldsymbol{W}'(q_\Theta^{1/2}\bar{\boldsymbol{x}} + r_n^{-1}\bar{\boldsymbol{y}}) - \frac{ts}{4n}\|q_\Theta^{1/2}\bar{\boldsymbol{x}} + r_n^{-1}\bar{\boldsymbol{y}}\|^4 \\
& + \frac{(1-t)sq_\Theta^2}{2n}(\boldsymbol{\Lambda}^{\mathsf{T}}\boldsymbol{x})^2 + \frac{\sqrt{(1-t)s}q_\Theta}{2}\boldsymbol{x}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{x} - \frac{(1-t)sq_\Theta^2}{4n}\|\boldsymbol{x}\|^4.
\end{aligned}
$$

The corresponding free energy density can be written as

$$
G_n(t,s) := \frac{1}{n}\mathbb{E}\left[\log\left(\int \exp\left(H_{n,t}^{(s)}(\boldsymbol{x}, \boldsymbol{y})\right) P_{\boldsymbol{\Lambda}}^{\otimes n}(\mathrm{d}\boldsymbol{x}) P_{\mathsf{N}(0,1)}^{\otimes n}(\mathrm{d}\boldsymbol{y})\right)\right],
$$

where $P_{\mathsf{N}(0,1)}^{\otimes n}$ is the distribution of $\mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. Invoking Lemma A.2.3 and Gaussian integration by parts, we obtain that

$$
\begin{aligned}
\frac{\partial}{\partial t}G_n(t,s) = & \frac{s}{4n^2}\mathbb{E}\left[\left\|\mathbb{E}\left[(q_\Theta^{1/2}\bar{\boldsymbol{\Lambda}} + r_n^{-1}\bar{\boldsymbol{g}})(q_\Theta^{1/2}\bar{\boldsymbol{\Lambda}} + r_n^{-1}\bar{\boldsymbol{g}})^{\mathsf{T}} \mid \boldsymbol{Y}_{a,t}^{(s)}, \boldsymbol{Y}_{b,t}^{(s)}\right]\right\|_F^2\right] \\
& - \frac{sq_\Theta^2}{4n^2}\mathbb{E}\left[\left\|\mathbb{E}[\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathsf{T}} \mid \boldsymbol{Y}_{a,t}^{(s)}, \boldsymbol{Y}_{b,t}^{(s)}]\right\|_F^2\right].
\end{aligned}
$$

Leveraging Hölder's inequality, we see that

$$
\left|\frac{\partial}{\partial t}G_n(t,s)\right|
$$

$$
\begin{aligned}
&\leq \left| \mathbb{E}\left[\frac{s}{4} \mathbb{E}\left[(q_\Theta^{1/2} \bar{\mathbf{\Lambda}}_1 + r_n^{-1} \bar{g}_1)(q_\Theta^{1/2} \bar{\mathbf{\Lambda}}_2 + r_n^{-1} \bar{g}_2) \mid \boldsymbol{Y}_{a,t}^{(s)}, \boldsymbol{Y}_{b,t}^{(s)} \right]^2 - \frac{s q_\Theta^2}{4} \mathbb{E}\left[\mathbf{\Lambda}_1 \mathbf{\Lambda}_2 \mid \boldsymbol{Y}_{a,t}^{(s)}, \boldsymbol{Y}_{b,t}^{(s)} \right]^2 \right] \right| + o_n(1) \\
&\leq \frac{s}{4} \mathbb{E}\left[\left(q_\Theta \bar{\mathbf{\Lambda}}_1 \bar{\mathbf{\Lambda}}_2 + r_n^{-1} q_\Theta^{1/2} \bar{\mathbf{\Lambda}}_1 \bar{g}_2 + r_n^{-1} q_\Theta^{1/2} \bar{\mathbf{\Lambda}}_2 \bar{g}_1 + r_n^{-2} \bar{g}_1 \bar{g}_2 - q_\Theta \mathbf{\Lambda}_1 \mathbf{\Lambda}_2 \right)^2 \right]^{1/2} \times \\
&\quad \mathbb{E}\left[\left(q_\Theta \bar{\mathbf{\Lambda}}_1 \bar{\mathbf{\Lambda}}_2 + r_n^{-1} q_\Theta^{1/2} \bar{\mathbf{\Lambda}}_1 \bar{g}_2 + r_n^{-1} q_\Theta^{1/2} \bar{\mathbf{\Lambda}}_2 \bar{g}_1 + r_n^{-2} \bar{g}_1 \bar{g}_2 + q_\Theta \mathbf{\Lambda}_1 \mathbf{\Lambda}_2 \right)^2 \right]^{1/2} + o_n(1). \qquad \text{(A.79)}
\end{aligned}
$$

The upper bound given in the last line of Eq. (A.79) is independent of t and converges to 0 as $n, d \to \infty$. Therefore, we conclude that as $n, d \to \infty$

$$
\sup_{t \in (0,1), s \in [0,2]} \left| \frac{\partial}{\partial t} G_n(t, s) \right| \to 0,
$$

which further implies that $|G_n(1, s) - G_n(0, s)| = o_n(1)$ for all $s \in [0, 2]$. Recall that $\mathcal{F}(\cdot, \cdot)$ is defined in Eq. (2.12). Using [125, Theorem 13], we have $G_n(0, s) = \sup_{q \geq 0} \mathcal{F}(q_\Theta^2 s, q) + o_n(1)$. Observe that $s \mapsto G_n(1, s)$ is convex differentiable on $(0, \infty)$, and converges point-wisely to $\sup_{q \geq 0} \mathcal{F}(q_\Theta^2 s, q)$ as $n, d \to \infty$, the later is differentiable at $s = 1$ for all but countably many values of $q_\Theta > 0$ according to [125, Proposition 17]. Invoking Lemma A.2.4, Lemma A.2.3 and Gaussian integration by parts, we conclude that for all but countably many $q_\Theta > 0$

$$
\lim_{n \to \infty} \frac{1}{n^2} \mathbb{E}\left[\left\| q_\Theta^{-1} \boldsymbol{M}_n(\boldsymbol{Y}_2) - q_\Theta^{-1} \big(q_\Theta^{1/2} \bar{\mathbf{\Lambda}} + r_n^{-1} \bar{\boldsymbol{g}} \big) \big(q_\Theta^{1/2} \bar{\mathbf{\Lambda}} + r_n^{-1} \bar{\boldsymbol{g}} \big)^\mathsf{T} \right\|_F^2 \right] = \lim_{n \to \infty} \mathrm{MMSE}_n^{\mathrm{symm}}(\mu_\Lambda; q_\Theta).
$$

Notice that $\lim_{n \to \infty} \mathbb{E}[\| (q_\Theta^{1/2} \bar{\mathbf{\Lambda}} + r_n^{-1} \bar{\boldsymbol{g}})(q_\Theta^{1/2} \bar{\mathbf{\Lambda}} + r_n^{-1} \bar{\boldsymbol{g}})^\mathsf{T} / n - q_\Theta \bar{\mathbf{\Lambda}} \bar{\mathbf{\Lambda}}^\mathsf{T} / n \|_F^2] = 0$, then the proof of the lemma follows immediately from triangle inequality.

A.6.3 Proof of Lemma A.4.11

We define the set

$$
\Omega := \left\{ |\mathbf{\Theta}_j| \leq 2 K_2 \sqrt{\log d} : j \in [d] \right\}.
$$

By Eq. (A.18) we have $\mathbb{P}(\Omega^c) \leq 2 d^{-3}$. Furthermore, on the set Ω we have $\boldsymbol{A} = \bar{\boldsymbol{A}}$. For $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, we define the mapping $\boldsymbol{M}(\boldsymbol{X}) := \mathbb{E}[\mathbf{\Lambda} \mathbf{\Lambda}^\mathsf{T} \mid \bar{\boldsymbol{A}} = \boldsymbol{X}]$. Leveraging triangle inequality and Hölder's inequality, we obtain that

$$
\begin{aligned}
\frac{1}{n} \mathbb{E}\left[\left\| \mathbf{\Lambda} \mathbf{\Lambda}^\mathsf{T} - M(\boldsymbol{A}) \right\|_F^2 \right]^{1/2} &\leq \frac{1}{n} \mathbb{E}\left[\left\| \mathbf{\Lambda} \mathbf{\Lambda}^\mathsf{T} - M(\boldsymbol{A}) \right\|_F^2 \mathbb{1}_\Omega \right]^{1/2} + \frac{1}{n} \mathbb{E}\left[\left\| \mathbf{\Lambda} \mathbf{\Lambda}^\mathsf{T} - M(\boldsymbol{A}) \right\|_F^2 \mathbb{1}_{\Omega^c} \right]^{1/2} \\
&\leq \frac{1}{n} \mathbb{E}\left[\left\| \mathbf{\Lambda} \mathbf{\Lambda}^\mathsf{T} - M(\bar{\boldsymbol{A}}) \right\|_F^2 \right]^{1/2} + \frac{1}{n} \mathbb{E}\left[\left\| \mathbf{\Lambda} \mathbf{\Lambda}^\mathsf{T} - M(\boldsymbol{A}) \right\|_F^4 \right]^{1/4} \mathbb{P}(\Omega^c)^{1/4}.
\end{aligned}
$$

Since the posterior expectation minimizes the expected l^2 risk, we then have

$$
\frac{1}{n^2} \mathbb{E}\left[\left\| \mathbf{\Lambda} \mathbf{\Lambda}^\mathsf{T} - M(\boldsymbol{A}) \right\|_F^2 \right] \geq \frac{1}{n^2} \mathbb{E}\left[\left\| \mathbf{\Lambda} \mathbf{\Lambda}^\mathsf{T} - \mathbb{E}[\mathbf{\Lambda} \mathbf{\Lambda}^\mathsf{T} \mid \boldsymbol{A}] \right\|_F^2 \right].
$$

By the bounded-support assumption, we see that $\mathbb{E}[\|\mathbf{\Lambda}\mathbf{\Lambda}^\mathsf{T} - M(\boldsymbol{A})\|_F^4]^{1/4}/n \leq 2K_1^2$. Thus, as $n, d \to \infty$,

$$\frac{1}{n}\mathbb{E}\left[\left\|\mathbf{\Lambda}\mathbf{\Lambda}^\mathsf{T} - M(\boldsymbol{A})\right\|_F^4\right]^{1/4} \mathbb{P}(\Omega^c)^{1/4} \to 0,$$

which completes the proof of the lemma.

A.6.4 Proof of Lemma A.4.12

For $t \in [0,1]$, we define the interpolated Hamiltonian as

$$\begin{aligned} H_{n,t}^{[s]}(\boldsymbol{\lambda}, \boldsymbol{\theta}; h) := &\sum_{i\in[n], j\in[d]} \left\{ \frac{ts}{\sqrt{nd}} \mathbf{\Lambda}_i \boldsymbol{\lambda}_i \mathbf{\Theta}_j \boldsymbol{\theta}_j + \frac{\sqrt{ts}}{\sqrt[4]{nd}} Z_{ij} \boldsymbol{\lambda}_i \boldsymbol{\theta}_j - \frac{ts}{2\sqrt{nd}} \boldsymbol{\lambda}_i^2 \boldsymbol{\theta}_j^2 \right\} + \\ &\sum_{i\in[n], j\in[d]} \left\{ \frac{s(1-t)}{\sqrt{nd}} \mathbf{\Lambda}_i \boldsymbol{\lambda}_i \bar{\mathbf{\Theta}}_j \bar{\boldsymbol{\theta}}_j + \frac{\sqrt{s(1-t)}}{\sqrt[4]{nd}} Z'_{ij} \boldsymbol{\lambda}_i \bar{\boldsymbol{\theta}}_j - \frac{s(1-t)}{2\sqrt{nd}} \boldsymbol{\lambda}_i^2 \bar{\boldsymbol{\theta}}_j^2 \right\} + H_n(\boldsymbol{\lambda}; \boldsymbol{Y}'(h)), \end{aligned}$$

where $\boldsymbol{Z}' = (Z'_{ij})_{i\in[n], j\in[d]}$ is an independent copy of $\boldsymbol{Z}$ and is independent of everything else. Note that $H_{n,t}^{[s]}(\boldsymbol{\lambda}, \boldsymbol{\theta}; h)$ is the Hamiltonian corresponding to the observations $(\boldsymbol{A}_1^{(s,t)}, \boldsymbol{A}_2^{(s,t)}, \boldsymbol{Y}'(h))$, where $\boldsymbol{A}_1^{(s,t)} = \sqrt{ts}\mathbf{\Lambda}\mathbf{\Theta}^\mathsf{T}/\sqrt[4]{nd} + \boldsymbol{Z}$, $\boldsymbol{A}_2^{(s,t)} = \sqrt{(1-t)s}\mathbf{\Lambda}\bar{\mathbf{\Theta}}^\mathsf{T}/\sqrt[4]{nd} + \boldsymbol{Z}'$ and $\boldsymbol{Y}'(h) = \sqrt{h}\mathbf{\Lambda}\mathbf{\Lambda}^\mathsf{T}/n + \boldsymbol{W}'$. Here, we recall that $\boldsymbol{W}' \sim \mathrm{GOE}(n)$, and $\boldsymbol{W}', \boldsymbol{Z}, \boldsymbol{Z}'$ are mutually independent. We define the free energy density corresponding to the Hamiltonian $H_{n,t}^{[s]}(\boldsymbol{\lambda}, \boldsymbol{\theta}; h)$ as

$$\Phi_{n,t}^{[s]}(h) := \frac{1}{n}\mathbb{E}\left[\log\left(\int \exp\left(H_{n,t}^{[s]}(\boldsymbol{\lambda}, \boldsymbol{\theta}; h)\right) \mu_\Lambda^{\otimes n}(\mathrm{d}\boldsymbol{\lambda}) \mu_\Theta^{\otimes d}(\mathrm{d}\boldsymbol{\theta})\right)\right].$$

At the endpoints, we have $\Phi_{n,0}^{[s]}(h) = \bar{\Phi}_n(s,0,0,h)$ and $\Phi_{n,1}^{[s]}(h) = \Phi_n(s,0,0,h)$. We denote by $\langle\cdot\rangle_{t,h}^{[s]}$ the expectation with respect to the posterior distribution $\mathbb{P}(\cdot \mid \boldsymbol{A}_1^{(s,t)}, \boldsymbol{A}_2^{(s,t)}, \boldsymbol{Y}'(h))$. Then we have

$$\begin{aligned} \left|\frac{\partial}{\partial t}\Phi_{n,t}^{[s]}(h)\right| &\overset{(i)}{=} \left|\frac{s}{2n\sqrt{nd}} \sum_{i\in[n], j\in[d]} \mathbb{E}\left[\langle \mathbf{\Lambda}_i \boldsymbol{\lambda}_i (\mathbf{\Theta}_j \boldsymbol{\theta}_j - \bar{\mathbf{\Theta}}_j \bar{\boldsymbol{\theta}}_j)\rangle_{t,h}^{[s]}\right]\right| \\ &\overset{(ii)}{\leq} \frac{s}{2n\sqrt{nd}} \sum_{i\in[n], j\in[d]} \mathbb{E}\left[\langle \mathbf{\Lambda}_i^2 \boldsymbol{\lambda}_i^2\rangle_{t,h}^{[s]}\right]^{1/2} \mathbb{E}\left[\langle (\mathbf{\Theta}_j \boldsymbol{\theta}_j - \bar{\mathbf{\Theta}}_j \bar{\boldsymbol{\theta}}_j)^2\rangle_{t,h}^{[s]}\right]^{1/2}, \end{aligned} \tag{A.80}$$

where *(i)* is by Gaussian integration by parts and Nishimori identity (Lemma A.2.3), and *(ii)* is by Hölder's inequality. For all $j \in [d]$, using power mean inequality and Hölder's inequality, we have

$$\begin{aligned} \mathbb{E}\left[\langle (\mathbf{\Theta}_j \boldsymbol{\theta}_j - \bar{\mathbf{\Theta}}_j \bar{\boldsymbol{\theta}}_j)^2\rangle_{t,h}^{[s]}\right] &\leq 2\mathbb{E}[\mathbf{\Theta}_j^2 \langle (\boldsymbol{\theta}_j - \bar{\boldsymbol{\theta}}_j)^2\rangle_{t,h}^{[s]}] + 2\mathbb{E}[(\mathbf{\Theta}_j - \bar{\mathbf{\Theta}}_j)^2 \langle \bar{\boldsymbol{\theta}}_j^2\rangle_{t,h}^{[s]}] \\ &\leq 2\mathbb{E}[\mathbf{\Theta}_j^4]^{1/2} \mathbb{E}[\langle (\boldsymbol{\theta}_j - \bar{\boldsymbol{\theta}}_j)^2\rangle_{t,h}^{[s]2}]^{1/2} + 2\mathbb{E}[(\mathbf{\Theta}_j - \bar{\mathbf{\Theta}}_j)^4]^{1/2} \mathbb{E}[\langle \bar{\boldsymbol{\theta}}_j^2\rangle_{t,h}^{[s]2}]^{1/2} \\ &\leq 2\mathbb{E}[\mathbf{\Theta}_j^4]^{1/2} \mathbb{E}[(\mathbf{\Theta}_j - \bar{\mathbf{\Theta}}_j)^4]^{1/2} + 2\mathbb{E}[(\mathbf{\Theta}_j - \bar{\mathbf{\Theta}}_j)^4]^{1/2} \mathbb{E}[\bar{\mathbf{\Theta}}_j^4]^{1/2}. \end{aligned} \tag{A.81}$$

Notice that

$$\mathbb{E}[(\mathbf{\Theta}_j - \bar{\mathbf{\Theta}}_j)^4] \leq \int_{2K_2\sqrt{\log d}}^\infty 4x^3 \mathbb{P}(|\mathbf{\Theta}_0| \geq x)\mathrm{d}x \leq 4\int_{4K_2^2 \log d}^\infty y \exp\Big(-\frac{y}{K_2^2}\Big)\mathrm{d}y$$

$$= -4(yK_2^2 + K_2^4)\exp\Big(-\frac{y}{K_2^2}\Big)\Big|_{4K_2^2\log d}^{\infty} = \frac{4K_2^4 + 16K_2^4\log d}{d^4}. \tag{A.82}$$

Combining Eqs. (A.80) to (A.82), we obtain that $\sup_{t\in[0,1],h\geq 0,S_0\geq s\geq 0}\left|\frac{\partial}{\partial t}\Phi_{n,t}^{(s)}(h)\right| \to 0$ as $n, d \to \infty$, thus completing the proof of the lemma.

A.6.5 Proof of Lemma A.4.13

Using Lemma A.4.12, we have $\lim_{n,d\to\infty}\left|\bar{\Phi}_n(s,0,0,h) - \Phi_n(s,0,0,h)\right| = 0$. Similar to the proof of Lemma A.4.8, we can conclude that $\lim_{n,d\to\infty}|\Phi_n(s,0,0,h) - \sup_{q\geq 0}\mathcal{F}(q_\Theta^2 s^2 + h, q)| = 0$ as $n, d \to \infty$. Therefore, in order to prove the lemma, it suffices to show

$$\lim_{n,d\to\infty}\sup_{a,a'\in[0,10]}\left|\bar{\Phi}_n(s,a,a',h) - \bar{\Phi}_n(s,0,0,h)\right| = 0.$$

Using Gaussian integration by parts and Nishimori identity (Lemma A.2.3), we obtain that for all $a, a' \in [0, 10]$,

$$\begin{aligned}\frac{\partial}{\partial\varepsilon_n}\bar{\Phi}_n(s,a,a',h) &= \frac{a^2}{2d}\mathbb{E}\left[\bar{\boldsymbol{\Theta}}^{\mathsf{T}}\mathbb{E}[\bar{\boldsymbol{\Theta}} \mid \bar{\boldsymbol{A}}(s), \boldsymbol{x}'(a'), \bar{\boldsymbol{x}}(a), \boldsymbol{Y}'(h)]\right] \leq 50\mathbb{E}_{\boldsymbol{\Theta}_0\sim\mu_\Theta}[\bar{\boldsymbol{\Theta}}_0^2],\\ \frac{\partial}{\partial\varepsilon_n'}\bar{\Phi}_n(s,a,a',h) &= \frac{a'^2}{2n}\mathbb{E}\left[\boldsymbol{\Lambda}^{\mathsf{T}}\mathbb{E}[\boldsymbol{\Lambda} \mid \bar{\boldsymbol{A}}(s), \boldsymbol{x}'(a'), \bar{\boldsymbol{x}}(a), \boldsymbol{Y}'(h)]\right] \leq 50\mathbb{E}_{\boldsymbol{\Lambda}_0\sim\mu_\Lambda}[\boldsymbol{\Lambda}_0^2].\end{aligned} \tag{A.83}$$

Notice that if $\varepsilon_n = \varepsilon_n' = 0$, then $\bar{\Phi}_n(s,a,a',h) = \bar{\Phi}_n(s,0,0,h)$. Therefore, by Eq. (C.11), we conclude that as $n, d \to \infty$,

$$\sup_{a,a'\in[0,10]}\left|\bar{\Phi}_n(s,a,a',h) - \bar{\Phi}_n(s,0,0,h)\right| \leq 50\big(\mathbb{E}_{\boldsymbol{\Lambda}_0\sim\mu_\Lambda}[\boldsymbol{\Lambda}_0^2] + \mathbb{E}_{\boldsymbol{\Theta}_0\sim\mu_\Theta}[\bar{\boldsymbol{\Theta}}_0^2]\big)(\varepsilon_n + \varepsilon_n') \to 0,$$

thus completing the proof of the lemma.

A.6.6 Proof of Lemma A.4.14

Since $|\bar{\boldsymbol{\Theta}}_0| \leq 2K_2\sqrt{\log d}$, we then have

$$\begin{aligned}&\left|\mathbb{E}[\langle U(\bar{\boldsymbol{\theta}}^{(1)})(\bar{\boldsymbol{\theta}}^{(1)})^{\mathsf{T}}\bar{\boldsymbol{\theta}}^{(2)}/d\rangle_{1,a,a',h}] - \mathbb{E}[\langle(\bar{\boldsymbol{\theta}}^{(1)})^{\mathsf{T}}\bar{\boldsymbol{\theta}}^{(2)}/d\rangle_{1,a,a',h}]\mathbb{E}[\langle U(\bar{\boldsymbol{\theta}}^{(1)})\rangle_{1,a,a',h}]\right|\\ \leq& 4K_2^2\log d\,\mathbb{E}[\langle|U(\bar{\boldsymbol{\theta}}) - \mathbb{E}[\langle U(\bar{\boldsymbol{\theta}})\rangle_{1,a,a',h}]|\rangle_{1,a,a',h}].\end{aligned} \tag{A.84}$$

Using Gaussian integration by parts and Nishimori identity (Lemma A.2.3), we have

$$\mathbb{E}[\langle(\bar{\boldsymbol{\theta}}^{(1)})^{\mathsf{T}}\bar{\boldsymbol{\theta}}^{(2)}/d\rangle_{1,a,a',h}]\mathbb{E}[\langle U(\bar{\boldsymbol{\theta}}^{(1)})\rangle_{1,a,a',h}] = a\mathbb{E}[\langle(\bar{\boldsymbol{\theta}}^{(1)})^{\mathsf{T}}\bar{\boldsymbol{\theta}}^{(2)}/d\rangle_{1,a,a',h}]^2, \tag{A.85}$$

$$\mathbb{E}[\langle U(\bar{\boldsymbol{\theta}}^{(1)})(\bar{\boldsymbol{\theta}}^{(1)})^{\mathsf{T}}\bar{\boldsymbol{\theta}}^{(2)}/d\rangle_{1,a,a',h}] = a\mathbb{E}[\langle((\bar{\boldsymbol{\theta}}^{(1)})^{\mathsf{T}}\bar{\boldsymbol{\theta}}^{(2)}/d)^2\rangle_{1,a,a',h}]. \tag{A.86}$$

Next, we combine Eqs. (A.84) to (A.86), and conclude that for all $a \in [10^{-1}, 10]$,

$$\mathbb{E}[\langle((\bar{\boldsymbol{\theta}}^{(1)})^{\mathsf{T}}\bar{\boldsymbol{\theta}}^{(2)}/d - \mathbb{E}[\langle(\bar{\boldsymbol{\theta}}^{(1)})^{\mathsf{T}}\bar{\boldsymbol{\theta}}^{(2)}/d\rangle_{1,a,a',h}])^2\rangle_{1,a,a',h}]$$

$$\leq 40K_2^2 \log d \mathbb{E}[\langle |U(\bar{\boldsymbol{\theta}}) - \mathbb{E}[\langle U(\bar{\boldsymbol{\theta}})\rangle_{1,a,a',h}]|\rangle_{1,a,a',h}].$$

A.6.7 Proof of Lemma A.4.15

One can verify that $\bar{\phi}_n(1,a,a',h)$ is twice differentiable for $a, a' \in (0,10)$. Using Gaussian integration by parts and Nishimori identity, we can compute its partial derivatives:

$$\frac{\partial}{\partial a}\bar{\phi}_n(1,a,a',h) = \varepsilon_n \langle U(\bar{\boldsymbol{\theta}})\rangle_{1,a,a',h}, \tag{A.87}$$

$$\frac{\partial^2}{\partial a^2}\bar{\phi}_n(1,a,a',h) = n\varepsilon_n^2 \langle (U(\bar{\boldsymbol{\theta}}) - \langle U(\bar{\boldsymbol{\theta}})\rangle_{1,a,a',h})^2\rangle_{1,a,a',h} + \varepsilon_n \langle 2\bar{\boldsymbol{\Theta}}^{\mathsf{T}}\bar{\boldsymbol{\theta}}/d - \bar{\boldsymbol{\theta}}^{\mathsf{T}}\bar{\boldsymbol{\theta}}/d\rangle_{1,a,a',h}. \tag{A.88}$$

Notice that $|\bar{\boldsymbol{\Theta}}_0| \leq 2K_2\sqrt{\log d}$, then $\left|\mathbb{E}[\frac{\partial}{\partial a}\bar{\phi}_n(1,a,a',h)]\right| = \varepsilon_n a\mathbb{E}[\langle \bar{\boldsymbol{\Theta}}^{\mathsf{T}}\bar{\boldsymbol{\theta}}/d\rangle_{1,a,a',h}] \leq 40\varepsilon_n K_2^2 \log d$ for all $a, a' \in (0,10)$. Using these results, we further obtain that

$$\begin{aligned}
&\langle (U(\bar{\boldsymbol{\theta}}) - \langle U(\bar{\boldsymbol{\theta}})\rangle_{1,a,a',h})^2\rangle_{1,a,a',h} \leq \frac{1}{n\varepsilon_n^2}\left(\frac{\partial^2}{\partial a^2}\bar{\phi}_n(1,a,a',h) + 12K_2^2\varepsilon_n \log d\right), \\
&\int_1^2\int_1^2 \mathbb{E}[\langle (U(\bar{\boldsymbol{\theta}}) - \langle U(\bar{\boldsymbol{\theta}})\rangle_{1,a,a',h})^2\rangle_{1,a,a',h}]\mathrm{d}a\mathrm{d}a' \\
\leq &\int_1^2 \mathbb{E}\left[\frac{1}{n\varepsilon_n^2}\left(\frac{\partial}{\partial a}\bar{\phi}_n(1,a,a',h)\Big|_{a=2} - \frac{\partial}{\partial a}\bar{\phi}_n(1,a,a',h)\Big|_{a=1} + 12K_2^2\varepsilon_n\log d\right)\right]\mathrm{d}a' \\
\leq &CK_2^2 n^{-1}\varepsilon_n^{-1}\log d,
\end{aligned} \tag{A.89}$$

where $C > 0$ is a numerical constant. Leveraging Eq. (A.88), we conclude that the following two functions are convex for all fixed $a' \in (0,10)$ and $h \geq 0$:

$$\begin{aligned}
&a \mapsto \bar{\phi}_n(1,a,a',h) + 6\varepsilon_n K_2^2 a^2 \log d, \\
&a \mapsto \mathbb{E}[\bar{\phi}_n(1,a,a',h)] + 6\varepsilon_n K_2^2 a^2 \log d.
\end{aligned}$$

By Lemma A.2.5, for all $a \in [1,2]$, $b \in (0,1/2)$, $a' \in [1/2,3]$ and $h \geq 0$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{\partial}{\partial a}\bar{\phi}_n(1,a,a',h) - E[\frac{\partial}{\partial a}\bar{\phi}_n(1,a,a',h)]\right|\right] \\
\leq &\mathbb{E}\left[\frac{\partial}{\partial a}\bar{\phi}_n(1,a+b,a',h) - \frac{\partial}{\partial a}\bar{\phi}_n(1,a-b,a',h)\right] + 24\varepsilon_n K_2^2 b\log d + \frac{3v_n(h)}{b}.
\end{aligned} \tag{A.90}$$

Again we use the fact that $\left|\mathbb{E}[\frac{\partial}{\partial a}\bar{\phi}_n(1,a,a',h)]\right| \leq 40\varepsilon_n K_2^2 \log d$ for all $a, a' \in (0,10)$, and conclude that

$$\begin{aligned}
&\int_1^2 \mathbb{E}\left[\frac{\partial}{\partial a}\bar{\phi}_n(1,a+b,a',h) - \frac{\partial}{\partial a}\bar{\phi}_n(1,a-b,a',h)\right]\mathrm{d}a \\
= &\mathbb{E}\left[\bar{\phi}_n(1,b+2,a',h) - \bar{\phi}_n(1,b+1,a',h) - \bar{\phi}_n(1,2-b,a',h) + \bar{\phi}_n(1,1-b,a',h)\right] \\
\leq &C'K_2^2 b\varepsilon_n \log d,
\end{aligned} \tag{A.91}$$

where $C' > 0$ is a numerical constant. Then we combine Eqs. (A.90) and (A.91) and obtain that

$$\int_1^2\int_1^2 \mathbb{E}\left[\left|\frac{\partial}{\partial a}\bar{\phi}_n(1,a,a',h) - E[\frac{\partial}{\partial a}\bar{\phi}_n(1,a,a',h)]\right|\right]\mathrm{d}a\mathrm{d}a' \leq C''\left(b\varepsilon_n K_2^2\log d + \frac{v_n(h)}{b}\right), \tag{A.92}$$

where $C'' > 0$ is another numerical constant. Later in Lemma A.4.16 we will see that under the current conditions, for n, d large enough we have $v_n(h) < \frac{1}{4}K_2^2\varepsilon_n \log d$. Since b is arbitrary in $(0, 1/2)$, we can then take $b = \sqrt{v_n(h)/(\varepsilon_n K_2^2 \log d)}$ in Eq. (A.92) and apply this to Eq. (A.87), which gives

$$\int_1^2 \int_1^2 \mathbb{E}\left[\left|\langle U(\bar{\boldsymbol{\theta}})\rangle_{1,a,a',h} - \mathbb{E}[\langle U(\bar{\boldsymbol{\theta}})\rangle_{1,a,a',h}]\right|\right] \mathrm{d}a\mathrm{d}a' \leq 2C''K_2\sqrt{v_n(h)\varepsilon_n^{-1}\log d}. \tag{A.93}$$

Finally, we combine Hölder's inequality, Eqs. (A.89) and (A.93) and concludes the proof of the lemma.

A.6.8 Proof of Lemma A.4.16

Conditioning on $(\boldsymbol{\Lambda}, \bar{\boldsymbol{\Theta}})$, we consider the mapping

$$f : (\boldsymbol{Z}, \boldsymbol{g}, \boldsymbol{g}', \sqrt{n}\boldsymbol{W}') \mapsto \bar{\phi}_n(1, a, a', h).$$

For n, d large enough, the following inequality holds for all $a, a' \in [0, 10]$.

$$\|\nabla f\|^2 \leq CK_1^2K_2^2 d^{1/2} n^{-3/2} \log d,$$

where $C > 0$ is a numerical constant. By Gaussian Poincaré inequality [189], we conclude that for n, d large enough

$$\mathbb{E}_{\boldsymbol{Z},\boldsymbol{g},\boldsymbol{g}',\boldsymbol{W}'}\left[\left(\bar{\phi}_n(1, a, a', h) - \mathbb{E}_{\boldsymbol{Z},\boldsymbol{g},\boldsymbol{g}',\boldsymbol{W}'}[\bar{\phi}_n(1, a, a', h)]\right)^2\right] \leq CK_1^2K_2^2 d^{1/2} n^{-3/2} \log d. \tag{A.94}$$

In the above display, the expectations are taken over $(\boldsymbol{Z}, \boldsymbol{g}, \boldsymbol{g}', \boldsymbol{W}')$.

Next, we show that $\mathbb{E}_{\boldsymbol{Z},\boldsymbol{g},\boldsymbol{g}',\boldsymbol{W}'}[\bar{\phi}_n(1, a, a', h)]$ (as a function of $(\boldsymbol{\Lambda}, \bar{\boldsymbol{\Theta}})$), concentrates around its expectation. Notice that for n, d large enough, for all $i \in [n], j \in [d]$ we have

$$\begin{aligned}
\left|\frac{\partial}{\partial \boldsymbol{\Lambda}_i}\mathbb{E}_{\boldsymbol{Z},\boldsymbol{g},\boldsymbol{g}',\boldsymbol{W}'}[\bar{\phi}_n(1, a, a', h)]\right| &\leq C'K_1K_2^2 d^{1/2} n^{-3/2} \log d, \\
\left|\frac{\partial}{\partial \bar{\boldsymbol{\Theta}}_j}\mathbb{E}_{\boldsymbol{Z},\boldsymbol{g},\boldsymbol{g}',\boldsymbol{W}'}[\bar{\phi}_n(1, a, a', h)]\right| &\leq C''K_1^2K_2 d^{-1/2} n^{-1/2} (\log d)^{1/2},
\end{aligned}$$

where $C', C'' > 0$ are numerical constants. By Efron-Stein inequality [189], we see that there exists a numerical constant $C''' > 0$, such that for n, d large enough

$$\mathbb{E}\left[\left(\mathbb{E}_{\boldsymbol{Z},\boldsymbol{g},\boldsymbol{g}',\boldsymbol{W}'}[\bar{\phi}_n(1, a, a', h)] - \mathbb{E}[\bar{\phi}_n(1, a, a', h)]\right)^2\right] \leq C'''K_1^4K_2^4 dn^{-2}(\log d)^2. \tag{A.95}$$

Finally, we combine Eqs. (A.94) and (A.95) and conclude that for n, d large enough, there exists a numerical constant $C_1 > 0$, such that for all $a, a' \in [0, 10]$

$$\mathbb{E}\left[\left(\bar{\phi}_n(1, a, a', h) - \mathbb{E}[\bar{\phi}_n(1, a, a', h)]\right)^2\right] \leq C_1^2K_1^4K_2^4 dn^{-2}(\log d)^2,$$

which concludes the proof of the lemma using Cauchy–Schwarz inequality.

A.6.9 Proof of the first claim of Lemma A.4.17

Invoking Lemma A.4.13, as $n, d \to \infty$ we have

$$\sup_{a\in[0,2a_*],a'\in[0,2a'_*]} \left|\bar{\Phi}_n(1, a, a', h) - \sup_{q\geq 0} \mathcal{F}(q_\Theta^2 + h, q)\right| = o_n(1). \tag{A.96}$$

By Jensen's inequality, for $a \sim \text{Unif}[a_*/2, a_*]$ and $a' \sim \text{Unif}[a'_*/2, a'_*]$ we have

$$\begin{aligned}&\frac{1}{n^2}\mathbb{E}\left[\left\|\mathbb{E}[\mathbf{\Lambda}\mathbf{\Lambda}^\mathsf{T} \mid \bar{\boldsymbol{A}}(1), \boldsymbol{Y}'(h)] - \mathbb{E}[\mathbf{\Lambda}\mathbf{\Lambda}^\mathsf{T} \mid \bar{\boldsymbol{A}}(1), \boldsymbol{x}'(a'), \bar{\boldsymbol{x}}(a), \boldsymbol{Y}'(h)]\right\|_F^2\right]\\ \leq&\frac{1}{n^2}\mathbb{E}\left[\left\|\mathbb{E}[\mathbf{\Lambda}\mathbf{\Lambda}^\mathsf{T} \mid \bar{\boldsymbol{A}}(1), \boldsymbol{Y}'(h)] - \mathbb{E}[\mathbf{\Lambda}\mathbf{\Lambda}^\mathsf{T} \mid \bar{\boldsymbol{A}}(1), \boldsymbol{x}'(a'_*), \bar{\boldsymbol{x}}(a_*), \boldsymbol{Y}'(h)]\right\|_F^2\right].\end{aligned} \tag{A.97}$$

Notice that the mapping $h \mapsto \bar{\Phi}_n(1, a_*, a'_*, h)$ is convex and differentiable, and $h \mapsto \sup_{q\geq 0} \mathcal{F}(q_\Theta^2 + h, q)$ is differentiable for all $q_\Theta^2 + h \in D$. Therefore, using Gaussian integration by parts, Lemmas A.2.3 and A.2.4, we conclude that for $h + q_\Theta^2 \in D$, as $n, d \to \infty$ the right hand side of Eq. (A.97) converges to 0 as $n, d \to \infty$. Furthermore,

$$\frac{\partial}{\partial a'}\mathbb{E}\left[\|\mathbf{\Lambda}_{1,a,a',h}\|^2\right] = \varepsilon'_n a' \mathbb{E}\left[\|\mathbb{E}[\mathbf{\Lambda}\mathbf{\Lambda}^\mathsf{T} \mid \bar{\boldsymbol{A}}(1), \bar{\boldsymbol{x}}(a), \boldsymbol{x}'(a'), \boldsymbol{Y}'(h)] - \mathbf{\Lambda}_{1,a,a',h}\mathbf{\Lambda}_{1,a,a',h}^\mathsf{T}\|_F^2\right] \leq 4K_1^4 n^2 \varepsilon'_n a'.$$

Therefore, for $a \sim \text{Unif}[a_*/2, a_*]$ and $a' \sim \text{Unif}[a'_*/2, a'_*]$, using the above equation we have

$$\begin{aligned}&\frac{1}{n^2}\mathbb{E}\left[\|\mathbb{E}[\mathbf{\Lambda}\mathbf{\Lambda}^\mathsf{T} \mid \bar{\boldsymbol{A}}(1), \boldsymbol{x}'(a'), \bar{\boldsymbol{x}}(a), \boldsymbol{Y}'(h)] - \mathbf{\Lambda}_{1,a,a',h}\mathbf{\Lambda}_{1,a,a',h}^\mathsf{T}\|_F^2\right]\\ \leq&\frac{8}{n^2 a_* (a'_*)^2 \varepsilon'_n}\int_{a_*/2}^{a_*}\int_{a'_*/2}^{a'_*}\frac{\partial}{\partial a'}\mathbb{E}\left[\|\mathbf{\Lambda}_{1,a,a',h}\|^2\right]\mathrm{d}a'\mathrm{d}a\\ \leq&\frac{4K_1^2}{n\varepsilon'_n (a'_*)^2},\end{aligned} \tag{A.98}$$

which converges to zero as $n, d \to \infty$ under the current assumptions. Eqs. (A.97) and (C.24) imply

$$\frac{1}{n^2}\|\mathbb{E}[\mathbf{\Lambda}\mathbf{\Lambda}^\mathsf{T} \mid \bar{\boldsymbol{A}}(1), \boldsymbol{Y}'(h)]\|_F^2 = \frac{1}{n^2}\|\mathbf{\Lambda}_{1,a,a',h}\|^4 + o_P(1).$$

By Corollary A.4.1 we have

$$\frac{1}{n}\|\mathbf{\Lambda}_{1,a,a',h}\|^2 = \frac{1}{n}\mathbb{E}\left[\|\mathbf{\Lambda}_{1,a,a',h}\|^2 \mid a, a'\right] + o_P(1).$$

By Jensen's inequality, for all $a \in [a_*/2, a_*]$ and $a' \in [a'_*/2, a'_*]$ we have

$$\mathbb{E}\left[\|\mathbf{\Lambda}_{1,a,a',h}\|^2 \mid a, a'\right] \leq \mathbb{E}\left[\|\mathbf{\Lambda}_{1,a_*,a'_*,h}\|^2\right].$$

Next, we combine the above equations and obtain that as $n, d \to \infty$

$$\frac{1}{n^2}\|\mathbb{E}[\mathbf{\Lambda}\mathbf{\Lambda}^\mathsf{T} \mid \bar{\boldsymbol{A}}(1), \boldsymbol{Y}'(h)]\|_F^2 \leq \frac{1}{n^2}\mathbb{E}\left[\|\mathbf{\Lambda}_{1,a_*,a'_*,h}\|^2\right]^2 + o_P(1).$$

Similarly, if we consider $a \sim \mathrm{Unif}[a_*, 2a_*]$ and $a' \sim \mathrm{Unif}[a'_*, 2a'_*]$, then we have

$$\frac{1}{n^2}\|\mathbb{E}[\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\mathsf{T} \mid \bar{\boldsymbol{A}}(1), \boldsymbol{Y}'(h)]\|_F^2 \geq \frac{1}{n^2}\mathbb{E}\left[\|\boldsymbol{\Lambda}_{1,a_*,a'_*,h}\|^2\right]^2 + o_P(1).$$

Since $\mathrm{support}(\boldsymbol{\Lambda}_0) \subseteq [-K_1, K_1]$, by Lebesgue dominated convergence theorem we have

$$\lim_{n,d\to\infty} \frac{1}{n^2}\mathbb{E}\left[\|\boldsymbol{\Lambda}_{1,a_*,a'_*,h}\|^2\right]^2 = \lim_{n,d\to\infty} \frac{1}{n^2}\mathbb{E}[\|\mathbb{E}[\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\mathsf{T} \mid \bar{\boldsymbol{A}}(1), \boldsymbol{Y}'(h)]\|_F^2]. \tag{A.99}$$

By Lemma A.4.13, as $n, d \to \infty$ we have $\bar{\Phi}_n(1,0,0,h) \to \sup_{q\geq 0} \mathcal{F}(q_\Theta^2 + h, q)$. Furthermore, the mapping $h \mapsto \bar{\Phi}_n(1,0,0,h)$ is convex and differentiable. Therefore, if $q_\Theta^2 + h \in D$, then by Lemma A.2.4 and Gaussian integration by parts we have

$$\lim_{n,d\to\infty} \frac{1}{4n^2}\mathbb{E}[\|\mathbb{E}[\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\mathsf{T} \mid \bar{\boldsymbol{A}}(1), \boldsymbol{Y}'(h)]\|_F^2] = \frac{\partial}{\partial h}\sup_{q\geq 0} \mathcal{F}(q_\Theta^2 + h, q). \tag{A.100}$$

The proof of the first claim follows immediately from Eqs. (A.99) and (A.100).

A.6.10 Proof of the second claim of Lemma A.4.17

We let $a \sim \mathrm{Unif}[a_*/2, a_*]$ and $a' \sim \mathrm{Unif}[a'_*/2, a'_*]$, then by Corollary A.4.1, we have

$$\begin{aligned}
&\frac{1}{n}\|\boldsymbol{\Lambda}_{1,a,a',h}\|^2 = \frac{1}{n}\mathbb{E}\left[\|\boldsymbol{\Lambda}_{1,a,a',h}\|^2 \mid a, a'\right] + \delta_{n,1},\\
&\frac{1}{\sqrt{nd}}\|\bar{\boldsymbol{\Theta}}_{1,a,a',h}\|^2 = \frac{1}{\sqrt{nd}}\mathbb{E}\left[\|\bar{\boldsymbol{\Theta}}_{1,a,a',h}\|^2 \mid a, a'\right] + \delta_{n,2},
\end{aligned}$$

where $\mathbb{E}[\delta_{n,1}^2]$ and $\mathbb{E}[\delta_{n,2}^2]$ are random variables that converge to 0 as $n, d \to \infty$. By Eq. (A.31) we have

$$\frac{1}{2n\sqrt{nd}}\|\boldsymbol{M}_{1,0,0,h}\|_F^2 = \frac{1}{2n\sqrt{nd}}\|\boldsymbol{\Lambda}_{1,a,a',h}\|^2\|\bar{\boldsymbol{\Theta}}_{1,a,a',h}\|^2 + \delta_{n,0}.$$

In the above equation, $\delta_{n,0}$ is a random variable satisfying $\mathbb{E}[|\delta_{n,0}|] \to 0$ as $n, d \to \infty$. Therefore, for all $a \in [a_*/2, a_*]$ and $a' \in [a'_*/2, a'_*]$

$$\begin{aligned}
&\frac{1}{2n\sqrt{nd}}\|\boldsymbol{M}_{1,0,0,h}\|_F^2\\
\leq &\frac{1}{2n\sqrt{nd}}\mathbb{E}\left[\|\boldsymbol{\Lambda}_{1,a,a',h}\|^2\right]\mathbb{E}\left[\|\bar{\boldsymbol{\Theta}}_{1,a,a',h}\|^2\right] + \frac{1}{2}|\delta_{n,1}||\delta_{n,2}| \qquad \text{(A.101)}\\
&+ \frac{|\delta_{n,1}|}{2\sqrt{nd}}\mathbb{E}\left[\|\bar{\boldsymbol{\Theta}}_{1,a,a',h}\|^2\right] + \frac{|\delta_{n,2}|}{2n}\mathbb{E}\left[\|\boldsymbol{\Lambda}_{1,a,a',h}\|^2\right] + \delta_{n,0}, \qquad \text{(A.102)}
\end{aligned}$$

Notice that

$$\limsup_{n,d\to\infty} \frac{1}{\sqrt{nd}}\mathbb{E}\left[\|\bar{\boldsymbol{\Theta}}_{1,a_*,a'_*,h}\|^2\right] < \infty, \qquad \limsup_{n,d\to\infty} \frac{1}{n}\mathbb{E}\left[\|\boldsymbol{\Lambda}_{1,a_*,a'_*,h}\|^2\right] < \infty. \tag{A.103}$$

We plug Eq. (A.103) into Eq. (A.102) then take the expectation, which implies that as $n, d \to \infty$ we have

$$\begin{aligned}&\liminf_{n,d\to\infty} \frac{1}{2n\sqrt{nd}}\mathbb{E}\left[\|\boldsymbol{\Lambda}_{1,a_*,a'_*,h}\|^2\right]\mathbb{E}\left[\|\bar{\boldsymbol{\Theta}}_{1,a_*,a'_*,h}\|^2\right]\\ \geq &\lim_{n,d\to\infty} \frac{1}{2n\sqrt{nd}}\mathbb{E}\left[\|\boldsymbol{M}_{1,0,0,h}\|^2\right] = D_\Theta(h). \end{aligned} \tag{A.104}$$

Similarly, if we let $a \sim \text{Unif}[a_*, 2a_*]$ and $a' \sim \text{Unif}[a'_*, 2a'_*]$, then we can conclude that

$$\begin{aligned}&\limsup_{n,d\to\infty} \frac{1}{2n\sqrt{nd}}\mathbb{E}\left[\|\boldsymbol{\Lambda}_{1,a_*,a'_*,h}\|^2\right]\mathbb{E}\left[\|\bar{\boldsymbol{\Theta}}_{1,a_*,a'_*,h}\|^2\right]\\ \leq &\lim_{n,d\to\infty} \frac{1}{2n\sqrt{nd}}\mathbb{E}\left[\|\boldsymbol{M}_{1,0,0,h}\|^2\right] = D_\Theta(h). \end{aligned} \tag{A.105}$$

Notice that $D_\Theta(h) = 2q_\Theta^2 \frac{\partial}{\partial h}\sup_{q\geq 0}\mathcal{F}(q_\Theta^2 + h, q)$, then the proof of the second claim follows from Eq. (A.104), Eq. (A.105) and the first claim.

A.7 Proofs for the Gaussian mixture clustering example

A.7.1 Proof of Proposition 2.5.1 claim (a)

We define the pairwise overlap achieved by estimator $\hat{\boldsymbol{\Lambda}}$ as

$$\text{PairOverlap}_n := \frac{2}{n^2}\sum_{i<j}\mathbb{1}\left\{\mathbb{1}\{\boldsymbol{\Lambda}_i = \boldsymbol{\Lambda}_j\} = \mathbb{1}\{\hat{\boldsymbol{\Lambda}}_i = \hat{\boldsymbol{\Lambda}}_j\}\right\}.$$

We notice that

$$\begin{aligned}\text{PairOverlap}_n =&\frac{2}{n^2}\sum_{i<j}\Big(\mathbb{1}\{\boldsymbol{\Lambda}_i = \hat{\boldsymbol{\Lambda}}_i\}\mathbb{1}\{\boldsymbol{\Lambda}_j = \hat{\boldsymbol{\Lambda}}_j\} + (1-\mathbb{1}\{\boldsymbol{\Lambda}_i = \hat{\boldsymbol{\Lambda}}_i\})(1-\mathbb{1}\{\boldsymbol{\Lambda}_j = \hat{\boldsymbol{\Lambda}}_j\})\Big)\\ =&2\text{Overlap}_n^2 + 1 - 2\text{Overlap}_n + o_n(1). \end{aligned}\tag{A.106}$$

According to [125, Section 2.3], under the symmetric model (2.9), if $q_\Theta \leq 1$, then we have $\lim_{n\to\infty}\text{MMSE}_n^{\text{symm}}(\mu_\Lambda; q_\Theta) = 1$. This is also the mean square error achieved by the constant estimator $\boldsymbol{0}_{n\times n}$. For an estimate of the labels $\hat{\boldsymbol{\Lambda}} \in \{-1, +1\}^n$ and $a \in (0,1)$, we define the rescaled vector $\hat{\boldsymbol{\Lambda}}_a := \sqrt{a}\hat{\boldsymbol{\Lambda}} \in \{-\sqrt{a}, +\sqrt{a}\}^n$. Then by Theorem 2.4.4, we have

$$\begin{aligned}\frac{1}{n^2}\mathbb{E}\left[\left\|\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\intercal - \hat{\boldsymbol{\Lambda}}_a\hat{\boldsymbol{\Lambda}}_a^\intercal\right\|_F^2\right] =&\mathbb{E}\left[(1-a)^2(\text{PairOverlap}_n + n^{-1}) + (1+a)^2(1 - n^{-1} - \text{PairOverlap}_n)\right]\\ \geq&\text{MMSE}_n^{\text{asym}}(\mu_\Lambda, \mu_\Theta).\end{aligned}$$

Theorem 2.4.4 implies that $\lim_{n,d\to\infty}\text{MMSE}_n^{\text{asym}}(\mu_\Lambda, \mu_\Theta) = 1$, thus

$$\limsup_{n,d\to\infty}\mathbb{E}\left[\text{PairOverlap}_n\right] \leq \frac{a^2 + 2a}{4a},$$

which holds for every $a \in (0,1)$. Let $a \to 0^+$, we then conclude that $\limsup_{n,d\to\infty} \mathbb{E}[\mathrm{PairOverlap}_n] \leq 1/2$. Next, we plug this result into Eq. (A.106) then apply dominated convergence theorem, which gives $\mathrm{Overlap}_n \overset{P}{\to} 1/2$. In summary, partial recovery of component identity is impossible in the current setting.

A.7.2 Proof of Theorem 2.5.1 part (a)

Let $\hat{\mathbf{\Lambda}} \in \mathbb{R}^{n\times k}$ be any estimator of the cluster assignments constructed based on data matrix $\boldsymbol{A}$. For $a > 0$, we define $\hat{\mathbf{\Lambda}}_a := \sqrt{a}\hat{\mathbf{\Lambda}}$. Under the current conditions, as $n, d \to \infty$ we have

$$\begin{aligned}
&\mathrm{MMSE}_n^{\mathrm{asym}}(\mu_\Lambda, \mu_\Theta) \\
=&k^{-2}(k-1) + o_n(1) \\
\leq&\frac{1}{n^2}\mathbb{E}\left[\left\|\mathbf{\Lambda}\mathbf{\Lambda}^{\mathsf{T}} - k^{-1}\mathbf{1}_{n\times n} - \hat{\mathbf{\Lambda}}_a\hat{\mathbf{\Lambda}}_a^{\mathsf{T}}\right\|_F^2\right] \\
=&k^{-2}(k-1) + \frac{a^2}{n^2}\sum_{i,j\in[n]}\mathbb{E}\left[\mathbb{1}\{\hat{\mathbf{\Lambda}}_i = \hat{\mathbf{\Lambda}}_j\}\right] - \frac{2a}{n^2}\sum_{i,j\in[n]}\mathbb{E}\left[\mathbb{1}\{\hat{\mathbf{\Lambda}}_i = \hat{\mathbf{\Lambda}}_j\}(\mathbb{1}\{\mathbf{\Lambda}_i = \mathbf{\Lambda}_j\} - k^{-1})\right] + o_n(1) \\
\leq&k^{-2}(k-1) + a^2 - \frac{2a}{n^2}\sum_{i,j\in[n]}\mathbb{E}\left[\mathbb{1}\{\hat{\mathbf{\Lambda}}_i = \hat{\mathbf{\Lambda}}_j\}(\mathbb{1}\{\mathbf{\Lambda}_i = \mathbf{\Lambda}_j\} - k^{-1})\right] + o_n(1).
\end{aligned}$$

Using the above equation we can conclude that

$$\limsup_{n,d\to\infty}\frac{1}{n^2}\sum_{i,j\in[n]}\mathbb{E}\left[\mathbb{1}\{\hat{\mathbf{\Lambda}}_i = \hat{\mathbf{\Lambda}}_j\}(\mathbb{1}\{\mathbf{\Lambda}_i = \mathbf{\Lambda}_j\} - k^{-1})\right] \leq \frac{a}{2}.$$

Since $a > 0$ is arbitrary, we then have

$$\limsup_{n,d\to\infty}\frac{1}{n^2}\sum_{i,j\in[n]}\mathbb{E}\left[\mathbb{1}\{\hat{\mathbf{\Lambda}}_i = \hat{\mathbf{\Lambda}}_j\}(\mathbb{1}\{\mathbf{\Lambda}_i = \mathbf{\Lambda}_j\} - k^{-1})\right] \leq 0. \tag{A.107}$$

For $s, r \in [k]$, we define

$$C_{sr} := \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{\hat{\mathbf{\Lambda}}_i = \boldsymbol{e}_s, \mathbf{\Lambda}_i = \boldsymbol{e}_r\}. \tag{A.108}$$

We immediately see that $C_{sr} \geq 0$ and $\sum_{s\in[k]}\sum_{r\in[k]} C_{sr} = 1$. Furthermore, notice that

$$\begin{aligned}
\frac{1}{n^2}\sum_{i,j\in[n]}\mathbb{1}\{\hat{\mathbf{\Lambda}}_i = \hat{\mathbf{\Lambda}}_j, \mathbf{\Lambda}_i = \mathbf{\Lambda}_j\} =&\frac{1}{n^2}\sum_{i,j\in[n]}\sum_{s,r\in[k]}\mathbb{1}\{\hat{\mathbf{\Lambda}}_i = \hat{\mathbf{\Lambda}}_j = \boldsymbol{e}_s, \mathbf{\Lambda}_i = \mathbf{\Lambda}_j = \boldsymbol{e}_r\} \\
=&\sum_{s,r\in[k]}\Big(\frac{1}{n}\sum_{i\in[n]}\mathbb{1}\{\hat{\mathbf{\Lambda}}_i = \boldsymbol{e}_s, \mathbf{\Lambda}_i = \boldsymbol{e}_r\}\Big)^2 \\
=&\sum_{s,r\in[k]} C_{sr}^2,
\end{aligned} \tag{A.109}$$

$$\frac{1}{kn^2}\sum_{i,j\in[n]}\mathbb{1}\{\hat{\mathbf{\Lambda}}_i = \hat{\mathbf{\Lambda}}_j\} =\frac{1}{kn^2}\sum_{i,j\in[n]}\sum_{s\in[k]}\mathbb{1}\{\hat{\mathbf{\Lambda}}_i = \boldsymbol{e}_s\}\mathbb{1}\{\hat{\mathbf{\Lambda}}_j = \boldsymbol{e}_s\}$$

$$=\frac{1}{k}\sum_{s\in[k]}\Big(\frac{1}{n}\sum_{i\in[n]}\sum_{r\in[k]}\mathbb{1}\{\hat{\boldsymbol{\Lambda}}_i=\boldsymbol{e}_s,\boldsymbol{\Lambda}_i=\boldsymbol{e}_r\}\Big)^2 \tag{A.110}$$
$$=\frac{1}{r}\sum_{s\in[k]}(C_{s1}+\cdots+C_{sk})^2.$$

Next, we subtract Eq. (A.109) by Eq. (A.110) and apply Eq. (A.107), which gives

$$\lim_{n\to\infty}\mathbb{E}\Big[\sum_{s\in[k]}\sum_{1\le r_1<r_2\le k}(C_{sr_1}-C_{sr_2})^2\Big]=0.$$

Note that for all $s\in[k]$, there exists $r_s\in[k]$, such that $[k]=\{r_s:s\in[k]\}$ and $\mathrm{Overlap}_n=\sum_{s\in[k]}C_{sr_s}$. Combining the above results, we conclude that $\mathrm{Overlap}_n=k^{-1}+o_P(1)$, thus completing the proof of part (a).

A.7.3 Proof of Theorem 2.5.1 part (b)

Suppose the statement is not true, then for any $\hat{\boldsymbol{\Lambda}}$ and any subsequence of $\mathbb{N}_+$, there further exists a subsequence $\{n_i\}_{i\in\mathbb{N}_+}\subseteq\mathbb{N}_+$ of the previous subsequence, such that $n_i<n_{i+1}$ and $\lim_{i\to\infty}\mathbb{E}[\mathrm{Overlap}_{n_i}]=k^{-1}$. Therefore, $\mathrm{Overlap}_{n_i}\overset{P}{\to}k^{-1}$. In the following parts of the proof, we will restrict to this subsequence $\{n_i\}_{i\in\mathbb{N}_+}$.

We assume $\hat{\boldsymbol{\Lambda}}\in\mathbb{R}^{n\times k}$ such that $\hat{\boldsymbol{\Lambda}}\overset{d}{=}\mu(\boldsymbol{\Lambda}=\cdot\mid\boldsymbol{A})$. Recall that for $s,r\in[k]$, C_{sr} is defined in Eq. (A.108). Furthermore, notice that $\hat{\boldsymbol{\Lambda}}\overset{d}{=}\boldsymbol{\Lambda}$, then by the law of large numbers, for all $s,r\in[k]$ we have

$$C_{s1}+C_{s2}+\cdots+C_{sk}\overset{P}{\to}\frac{1}{k},\qquad C_{1r}+C_{2r}+\cdots+C_{kr}\overset{P}{\to}\frac{1}{k}. \tag{A.111}$$

For $\delta>0$, if $C_{11}>k^{-2}+\delta$, then by Eq. (A.111) $\sum_{2\le s,r\le k}C_{sr}>(1-k^{-1})^2+\delta+o_P(1)$. As a result, we conclude that there exists a permutation π of $\{2,3,\cdots,k\}$, such that $C_{2\pi(2)}+C_{3\pi(3)}+\cdots+C_{k\pi(k)}\ge k^{-2}(k-1)+(r-1)^{-1}\delta+o_P(1)$. Therefore, $\mathrm{Overlap}_n\ge C_{11}+C_{2\pi(2)}+C_{3\pi(3)}+\cdots+C_{k\pi(k)}>k^{-1}+k(k-1)^{-1}\delta+o_P(1)$. For $s,r\in[k]$, we define the set $S_{sr}^{\delta}:=\{C_{sr}>k^{-2}+\delta\}$. Since $\mathrm{Overlap}_{n_i}\overset{P}{\to}k^{-1}$, using the above analysis we derive that $\lim_{i\to\infty}\mathbb{P}(S_{11}^{\delta})=0$. Indeed, we can repeat such analysis for all $s,r\in[k]$ and conclude that $\lim_{i\to\infty}\mathbb{P}(S_{sr}^{\delta})=0$, thus $C_{sr}\le k^{-2}+\delta+o_P(1)$ along the subsequence $\{n_i\}_{i\in\mathbb{N}_+}$. Since $\sum_{s,r\in[k]}C_{sr}=1$ and δ is arbitrary, we deduce that $C_{sr}\overset{P}{\to}k^{-2}$ along the subsequence $\{n_i\}_{i\in\mathbb{N}_+}$. This further implies that $C_{sr}\overset{P}{\to}k^{-2}$.

However, according to [18, Theorem 2] and Theorem 2.4.5, we see that under the conditions of this part

$$\liminf_{n,d\to\infty}\frac{1}{n^2}\sum_{i,j\in[n]}\mathbb{E}\left[\mathbb{1}\{\hat{\boldsymbol{\Lambda}}_i=\hat{\boldsymbol{\Lambda}}_j\}\mathbb{1}\{\boldsymbol{\Lambda}_i=\boldsymbol{\Lambda}_j\}\right]=\liminf_{n\to\infty}\frac{1}{n^2}\sum_{i,j\in[n]}\mathbb{E}\left[\mathbb{E}[\mathbb{1}\{\boldsymbol{\Lambda}_i=\boldsymbol{\Lambda}_j\}\mid\boldsymbol{A}]^2\right]>k^{-2}.$$

Finally, we plug Eq. (A.109) into the formula above, which leads to $\liminf_{n,d\to\infty}\mathbb{E}[\sum_{s,k\in[r]}C_{sk}^2]>k^{-2}$. This is in contradiction with the previously established claim that $C_{sr}\overset{P}{\to}k^{-2}$ for all $s,r\in[k]$, thus completing the proof of part (b).

Appendix B

The estimation error of general first order methods

B.1 Technical definitions and lemmas

We collect some useful technical definitions and lemmas, some of which we state without proof. First, we recall the definition of the Wasserstein metric of order 2 on the space $\mathscr{P}_2(\mathbb{R}^k)$:

$$W_2(\mu, \mu')^2 = \inf_{\Pi} \mathbb{E}_{(\boldsymbol{A}, \boldsymbol{A}') \sim \Pi}[\|\boldsymbol{A} - \boldsymbol{A}'\|^2] ,$$

where the infimum is over couplings Π between μ and μ'. That is, $\Pi \in \mathscr{P}_2(\mathbb{R}^k \times \mathbb{R}^k)$ whose first and second marginals are μ (where a marginal here involves a block of k coordinates). It is well known that $W_2(\mu, \mu')$ is a metric on $\mathscr{P}_2(\mathbb{R}^k)$ [193, pg. 94]. When a sequence of probability distributions μ_n converges to μ in the Wasserstein metric of order 2, we write $\mu_n \overset{\mathrm{W}}{\to} \mu$. We also write $\boldsymbol{A}_n \overset{\mathrm{W}}{\to} \boldsymbol{A}$ when $\boldsymbol{A}_n \sim \mu_n$, $\boldsymbol{A} \sim \mu$ for such a sequence.

Lemma B.1.1. *If $f : \mathbb{R}^r \to \mathbb{R}$ and $g : \mathbb{R}^r \to \mathbb{R}$ are pseudo-Lipschitz of order k_1 and k_2, respectively, then their product is pseudo-Lipschitz of order $k_1 + k_2$.*

Lemma B.1.2. *If a sequence of random vectors $\boldsymbol{X}_n \overset{\mathrm{W}}{\to} \boldsymbol{X}$, then for any pseudo-Lipschitz function f of order 2 we have $\mathbb{E}[f(\boldsymbol{X}_n)] \to \mathbb{E}[f(\boldsymbol{X})]$.*

Lemma B.1.3. *Consider a sequence of random variables $(A_n, \boldsymbol{B}_n) \overset{\mathrm{d}}{\to} (A, \boldsymbol{B})$ with values in $\mathbb{R} \times \mathbb{R}^k$ such that $(A_n, \boldsymbol{B}_n) \overset{\mathrm{d}}{\to} (A, \boldsymbol{B})$ and $A_n \overset{\mathrm{d}}{=} A$ for all n. Then, for any bounded measurable function $f : \mathbb{R} \times \mathbb{R}^k \to \mathbb{R}$ for which $\boldsymbol{b} \mapsto f(a, \boldsymbol{b})$ is continuous for all a, we have $\mathbb{E}[f(A_n, \boldsymbol{B}_n)] \to \mathbb{E}[f(A, \boldsymbol{B})]$.*

Further, for any function $\phi : \mathbb{R} \times \mathbb{R}^k \to \mathbb{R}^{k'}$ (possibly unbounded) which is continuous in all but the first coordinate, we have $\phi(A_n, \boldsymbol{B}_n) \overset{\mathrm{d}}{\to} \phi(A, \boldsymbol{B})$.

Proof.[Proof of Lemma B.1.3] Without loss of generality, f takes values in $[0, 1]$. First we show that for any set $S \times I$ where $S \subset \mathbb{R}$ is measurable and $I \subset \mathbb{R}^k$ is a rectangle whose boundary has probability 0 under $\boldsymbol{B}$ that

$$\mu_{A_n, \boldsymbol{B}_n}(S \times I) \to \mu_{A, \boldsymbol{B}}(S \times I) . \tag{B.1}$$

First, we show this is true for $S = K$ a closed set. Fix $\epsilon > 0$. Let $\phi_K^\epsilon : \mathbb{R} \to [0,1]$ be a continuous function which is 1 on K and 0 for all points separted from K by distance ϵ. Similarly define $\phi_I^\epsilon : \mathbb{R}^k \to \mathbb{R}$. Then

$$\mathbb{E}[\phi_K^\epsilon(A_n)\phi_I^\epsilon(\boldsymbol{B}_n)] \geq \mu_{A_n,\boldsymbol{B}_n}(K \times I) \geq \mathbb{E}[\phi_K^\epsilon(A_n)\phi_I^\epsilon(\boldsymbol{B}_n)] - \epsilon - \mu_{\boldsymbol{B}_n}(\mathsf{spt}(\phi_I^\epsilon) \setminus I)\,.$$

Because the boundary of I has measure 0 under $\mu_{\boldsymbol{B}}$, we have $\lim_{\epsilon\to 0}\limsup_{n\to\infty}\mu_{\boldsymbol{B}_n}(\mathsf{spt}(\phi_I^\epsilon)\setminus I) = 0$. Also, $\lim_{\epsilon\to 0}\lim_{n\to\infty}\mathbb{E}[\phi_K^\epsilon(A_n)\phi_I^\epsilon(\boldsymbol{B}_n)] = \lim_{\epsilon\to 0} = \mathbb{E}[\phi_K^\epsilon(A)\phi_I^\epsilon(\boldsymbol{B})] = \mu_{A,\boldsymbol{B}}(K\times I)$. Thus, taking $\epsilon \to \infty$ after $n \to \infty$, the previous display gives $\mu_{A_n,\boldsymbol{B}_n}(K \times I) \to \mu_{A,\boldsymbol{B}}(K\times I)$. For $S = G$ an open set, we can show $\mu_{A_n,\boldsymbol{B}_n}(G\times I) \to \mu_{A,\boldsymbol{B}}(G \times I)$ by a similar argument: take instead ϕ_K^ϵ to be 0 outside of G and 1 for all points in G separated from the boundary by at least ϵ, and likewise for ϕ_I^ϵ. By Theorem 12.3 of [35], we can construct $K \subset S \subset G$ such that K is closed and G is open, and $\mu_A(K) > \mu_A(S) - \epsilon$, $\mu_A(G) < \mu_A(S) + \epsilon$. The previous paragraph implies that

$$\begin{aligned}\mu_{A,\boldsymbol{B}}(S\times I) - \epsilon \leq \mu_{A,\boldsymbol{B}}(K\times I) &\leq \liminf_{n\to\infty}\mu_{A_n,\boldsymbol{B}_n}(S\times I)\\ &\leq \limsup_{n\to\infty}\mu_{A_n,\boldsymbol{B}_n}(S\times I) \leq \mu_{A,\boldsymbol{B}}(G\times I) \leq \mu_{A,\boldsymbol{B}}(S\times I) + \epsilon\,.\end{aligned}$$

Taking $\epsilon \to 0$, we conclude (B.1).

We now show (B.1) implies the lemma. Fix $\epsilon > 0$. Let M be such that $\mathbb{P}(\boldsymbol{B}_n \in [-M,M]^k) > 1-\epsilon$ for all n and $\mathbb{P}(\boldsymbol{B} \in [-M,M]^k) > 1-\epsilon$, which we may do by tightness. For each a, let $\delta(a,\epsilon) = \sup\{0 < \Delta \leq M \mid \|\boldsymbol{b}-\boldsymbol{b}'\|_\infty < \Delta \Rightarrow |f(a,\boldsymbol{b}) - f(a,\boldsymbol{b}')| < \epsilon\}$. Because continuous functions are uniformly continuous on compact sets, the supremum is over a non-empty, bounded set. Thus, $\delta(a,\epsilon)$ is positive and bounded above by M for all a. Further, $\delta(a,\epsilon)$ is measurable and non-decreasing in ϵ. Pick δ_* such that $\mathbb{P}(\delta(A,\epsilon) < \delta_*) < \epsilon$, which we may do because $\delta(a,\epsilon)$ is positive for all a. We can partition $[-M,M]^k$ into rectangles with side-widths smaller than δ_* such that the probability that $\boldsymbol{B}$ lies on the boundary of one of the partitioning rectangles is 0. Define $f_-(a,\boldsymbol{b}) := \sum_\iota \mathbf{1}\{\boldsymbol{b} \in I_\iota\}\inf_{\boldsymbol{b}'\in I_\iota} f(a,\boldsymbol{b}')$ and $f_+(a,\boldsymbol{b}) := \sum_\iota \mathbf{1}\{\boldsymbol{b}\in I_\iota\}\sup_{\boldsymbol{b}'\in I_\iota} f(a,\boldsymbol{b}')$, and note that on $\{\delta(a,\epsilon) < \delta^*\} \times [-M,M]^k$, we have $f_-(a,\boldsymbol{b}) \leq f(a,\boldsymbol{b}) \leq f_+(a,\boldsymbol{b})$ and $|f(a,\boldsymbol{b}) - f_-(a,\boldsymbol{b})| < \epsilon$ and $|f(a,\boldsymbol{b}) - f_+(a,\boldsymbol{b})| < \epsilon$. Thus, by the boundedness of f and the high-probability bound on $\{\delta(a,\epsilon) < \delta^*\} \times [-M,M]^k$

$$\begin{gathered}\mathbb{E}[f_-(A_n,\boldsymbol{B}_n)] - 2\epsilon < \mathbb{E}[f(A_n,\boldsymbol{B}_n)] < \mathbb{E}[f_+(A_n,\boldsymbol{B}_n)] + 2\epsilon\,,\\ \mathbb{E}[f_-(A,\boldsymbol{B})] - 2\epsilon < \mathbb{E}[f(A,\boldsymbol{B})] < \mathbb{E}[f_+(A,\boldsymbol{B})] + 2\epsilon\,.\end{gathered} \tag{B.2}$$

We show that $\mathbb{E}[f_-(A_n,\boldsymbol{B}_n)] \to \mathbb{E}[f_-(A,\boldsymbol{B})]$. Fix $\xi > 0$. Take $0 = x_0 \leq \ldots \leq x_N = 1$ such that $x_{j+1} - x_j < \xi$ for all j. Let $S_{j\iota} = \{a \mid \inf_{\boldsymbol{b}'\in I_\iota} f(a,\boldsymbol{b}') \in [x_j, x_{j+1})\}$. Then

$$\sum_{\iota,j} x_j\mathbf{1}\{a \in S_{j\iota}, \boldsymbol{b}\in I_\iota\} + \xi \geq f_-(a,\boldsymbol{b}) \geq \sum_{\iota,j} x_j \mathbf{1}\{a\in S_{j\iota},\boldsymbol{b}\in I_\iota\}\,.$$

By (B.1), we conclude $\mathbb{E}[\sum_{\iota,j} x_j\mathbf{1}\{A_n \in S_{j\iota}, \boldsymbol{B}_n \in I_\iota\}] \to \mathbb{E}[\sum_{\iota,j} x_j \mathbf{1}\{A\in S_{j\iota},\boldsymbol{B}\in I_\iota\}]$. Combined with the previous display and taking $\xi \to 0$, we conclude that $\mathbb{E}[f_-(A_n,\boldsymbol{B}_n)] \to \mathbb{E}[f_-(A,\boldsymbol{B})]$. Similarly, we may argue that $\mathbb{E}[f_+(A_n,\boldsymbol{B}_n)] \to \mathbb{E}[f_+(A,\boldsymbol{B})]$. The first statment in the lemma now follows from taking $\epsilon \to 0$ after $n \to \infty$ in (B.2).

The second statement in the lemma follows by observing that for any bounded continuous function $f : \mathbb{R}^{k'} \to \mathbb{R}$, we have that $f \circ \phi$ is bounded and is continuous in all but the first coordinate, so that we may apply the first part of the lemma to conclude $\mathbb{E}[f(\phi(A_n, \boldsymbol{B}_n))] \to \mathbb{E}[f(\phi(A, \boldsymbol{B}))]$.

□

We will sometimes use the following alternative form of recursion (3.5) defining the lower bound in the high-dimensional regression model.

Lemma B.1.4. *Consider a family, indexed by $x \in \mathbb{R}$, of bounded probability densities $p(\cdot|x, u)$ with respect to some base measure μ_Y. Then for $\tilde{\tau} > 0$ and $\sigma \geq 0$ we have that*

$$\frac{1}{\tilde{\tau}^2}\mathbb{E}[\mathbb{E}[G_1|Y, G_0, U]^2] = \mathbb{E}_{G_0,Y}\left[\left(\frac{\mathrm{d}}{\mathrm{d}x}\log \mathbb{E}_{G_1} p(Y|x + \sigma G_0 + \tilde{\tau} G_1, U)\Big|_{x=0}\right)^2\right],$$

where $G_0, G_1 \overset{\text{iid}}{\sim} \mathsf{N}(0,1)$ and $Y|G_0, G_1, U$ had density $p(\cdot|\sigma G_0 + \tilde{\tau} G_1, U)$ with respect to μ_Y. In particular, the derivatives exist. (In this case, we may equivalently generate $Y = h(\sigma G_0 + \tilde{\tau} G_1, \boldsymbol{W})$ for $(\boldsymbol{W}, U) \sim \mu_{\boldsymbol{W},U}$).

The preceding lemma applies, in particular, for p as in R4. It then provides an alternative form of the second equation in recursion (3.5).

Proof.[Lemma B.1.4] We have

$$\mathbb{E}_{G_1} p(Y|x + \sigma G_0 + \tilde{\tau} G_1, U) = \int p(Y|\sigma G_0 + s, U)\frac{1}{\sqrt{2\pi\tilde{\tau}}}e^{-\frac{1}{2\tilde{\tau}^2}(s-x)^2}\mathrm{d}g\,,$$

so that

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{E}_{G_1} p(Y|x + \sigma G_0 + \tilde{\tau} G_1, U) = \frac{1}{\tilde{\tau}^2}\int p(Y|\sigma G_0 + s, U)\frac{(s-x)}{\sqrt{2\pi\tilde{\tau}}}e^{-\frac{1}{2\tilde{\tau}^2}(s-x)^2}\mathrm{d}g\,,$$

where the boundedness of of p allows us to exchange integration and differentition. Thus,

$$\frac{\mathrm{d}}{\mathrm{d}x}\log \mathbb{E}_{G_1} p(Y|x + \sigma G_0 + \tilde{\tau} G_1, U) = \frac{1}{\tilde{\tau}}\mathbb{E}[G_1|Y, G_0, U]\,.$$

The result follows.

□

Finally, we collect some results on the Bayes risk with respect to quadratically-bounded losses $\ell : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}_{\geq 0}$. Recall that quadratically-bounded means that ℓ is pseudo-Lipschitz of order 2 and also satisfies

$$|\ell(\boldsymbol{\vartheta}, \boldsymbol{d}) - \ell(\boldsymbol{\vartheta}', \boldsymbol{d})| \leq C\left(1 + \sqrt{\ell(\boldsymbol{\vartheta}, \boldsymbol{d})} + \sqrt{\ell(\boldsymbol{\vartheta}', \boldsymbol{d})}\right)\|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}'\|\,. \tag{B.3}$$

We consider a setting $(\boldsymbol{\Theta}, \boldsymbol{V}) \sim \mu_{\boldsymbol{\Theta},\boldsymbol{V}} \in \mathscr{P}_2(\mathbb{R}^k \times \mathbb{R}^k)$, $\boldsymbol{Z} \sim \mathsf{N}(0, \boldsymbol{I}_k)$ independent and $\tau, K, M \geq 0$. Define $\boldsymbol{\Theta}^{(K)}$ by $\Theta_i^{(K)} = \Theta_i \mathbf{1}\{|\Theta_i| \leq K\}$. Denote by $\mu_{\boldsymbol{\Theta}^{(K)},\boldsymbol{V}}$ the joint distribution of $\boldsymbol{\Theta}^{(K)}$ and $\boldsymbol{V}$, and by $\mu_{\boldsymbol{\Theta}^{(K)}|\boldsymbol{V}} : \mathbb{R}^k \times \mathcal{B} \to [0,1]$ a regular conditional probability distribution for $\boldsymbol{\Theta}^{(K)}$ conditioned on $\boldsymbol{V}$. Define the posterior Bayes risk

$$R(\boldsymbol{y}, \tau, \boldsymbol{v}, K, M) := \inf_{\|\boldsymbol{d}\|_\infty \leq M}\int \frac{1}{Z}\ell(\boldsymbol{\vartheta}, \boldsymbol{d})e^{-\frac{1}{2\tau^2}\|\boldsymbol{y}-\boldsymbol{\vartheta}\|^2}\mu_{\boldsymbol{\Theta}^{(K)}|\boldsymbol{V}}(\boldsymbol{v}, \mathrm{d}\boldsymbol{\vartheta})\,, \tag{B.4}$$

where $Z = \int e^{-\frac{1}{2\tau^2}\|\boldsymbol{y}-\boldsymbol{\vartheta}\|^2}\mu_{\boldsymbol{\Theta}^{(K)}|\boldsymbol{V}}(\boldsymbol{v},\mathrm{d}\boldsymbol{\vartheta})$ is a normalization constant. It depends on $\boldsymbol{y},\tau,\boldsymbol{v},K$. When required for clarity, we write $Z(\boldsymbol{y},\tau,\boldsymbol{v},K)$.

Lemma B.1.5. *The following properties hold for the Bayes risk with respect to pseudo-Lipschitz losses of order 2 satsifying* (B.3)*.*

(a) For any τ, K, M, with K, M possibly equal to infinity, the Bayes risk is equal to the expected posterior Bayes risk. That is,

$$\inf_{\hat{\boldsymbol{\theta}}(\cdot)} \mathbb{E}[\ell(\boldsymbol{\Theta}^{(K)},\hat{\boldsymbol{\theta}}(\boldsymbol{\Theta}^{(K)}+\tau\boldsymbol{z},\boldsymbol{V})] = \mathbb{E}[R(\boldsymbol{Y}^{(K)},\tau,\boldsymbol{V},K,M)]\,, \tag{B.5}$$

where $\boldsymbol{Y}^{(K)} = \boldsymbol{\Theta}^{(K)} + \tau\boldsymbol{Z}$ with $\boldsymbol{Z}\sim\mathsf{N}(\boldsymbol{0},\boldsymbol{I}_k)$ independent of $\boldsymbol{\Theta}^{(K)}$ and the infimum is taken over all measurable functions $(\mathbb{R}^k)^2\to[-M,M]^k$. Moreover,

$$\mathbb{E}[R(\boldsymbol{\Theta}^{(K)}+\tau\boldsymbol{Z},\tau,\boldsymbol{V},K,\infty)] = \lim_{M\to\infty}\mathbb{E}[R(\boldsymbol{\Theta}^{(K)}+\tau\boldsymbol{Z},\tau,\boldsymbol{V},K,M)]\,. \tag{B.6}$$

(b) For a fixed $K<\infty$, the posterior Bayes risk is bounded: $R(\boldsymbol{y},\tau,\boldsymbol{v},K,M)\le\bar{R}(K)$ for some function $\bar{R}$ which does not depend on $\boldsymbol{y},\tau,\boldsymbol{v},M$. Further, for $K<\infty$ the function $(\boldsymbol{y},\tau)\mapsto R(\boldsymbol{y},\tau,\boldsymbol{v},K,M)$ is continuous on $\mathbb{R}^k\times\mathbb{R}_{>0}$.

(c) The Bayes risk is jointly continuous in truncation level K and noise variance τ. This is true also at $K=\infty$:

$$\mathbb{E}[R(\boldsymbol{\Theta}^{(K)}+\tau\boldsymbol{Z},\tau,\boldsymbol{V},K,\infty)] = \lim_{\substack{K\to\infty\\ \tau'\to\tau}}\mathbb{E}[R(\boldsymbol{\Theta}^{(K)}+\tau'\boldsymbol{Z},\tau',\boldsymbol{V},K,\infty)]\,, \tag{B.7}$$

where the limit holds for any way of taking K,τ' to their limits (ie., sequentially or simultaneously).

Proof.[Proof of Lemma B.1.5(a)] For any measurable $\hat{\boldsymbol{\theta}}:\mathbb{R}^k\times\mathbb{R}^k\to[-M,M]^k$,

$$\begin{aligned}\mathbb{E}[\ell(\boldsymbol{\Theta}^{(K)},\hat{\boldsymbol{\theta}}(\boldsymbol{\Theta}^{(K)}+\tau\boldsymbol{Z},\boldsymbol{V}))] &= \mathbb{E}[\mathbb{E}[\ell(\boldsymbol{\Theta}^{(K)},\hat{\boldsymbol{\theta}}(\boldsymbol{\Theta}^{(K)}+\tau\boldsymbol{Z},\boldsymbol{V}))|\boldsymbol{\Theta}^{(K)}+\tau\boldsymbol{Z},\boldsymbol{V}]]\\ &\ge \mathbb{E}[R(\boldsymbol{\Theta}^{(K)}+\tau\boldsymbol{Z},\tau,\boldsymbol{V},K,M)]\,.\end{aligned} \tag{B.8}$$

For $M<\infty$, equality obstains. Indeed, we may define

$$\hat{\boldsymbol{\theta}}^{(M)}(\boldsymbol{y},\boldsymbol{v};\tau) = \arg\min_{\|\boldsymbol{d}\|_\infty\le M}\int\frac{1}{Z}\ell(\boldsymbol{\vartheta},\boldsymbol{d})e^{-\frac{1}{2\tau^2}\|\boldsymbol{y}-\boldsymbol{\vartheta}\|^2}\mu_{\boldsymbol{\Theta}^{(K)}|\boldsymbol{V}}(\boldsymbol{v},\mathrm{d}\boldsymbol{\vartheta})\,, \tag{B.9}$$

because the integral is continuous in $\boldsymbol{d}$ by dominated convergence. Then $\mathbb{E}[\ell(\boldsymbol{\Theta}^{(K)},\hat{\boldsymbol{\theta}}^{(M)}(\boldsymbol{Y},\boldsymbol{V};\tau))] = \mathbb{E}[R(\boldsymbol{Y},\tau,\boldsymbol{V},K,M)]$ when $\boldsymbol{Y}=\boldsymbol{\Theta}^{(K)}+\tau\boldsymbol{Z}$. Observe $R(\boldsymbol{y},\tau,\boldsymbol{v},K,M)\downarrow R(\boldsymbol{y},\tau,\boldsymbol{v},K,\infty)$ as $M\to\infty$ with the other arguments fixed. Thus, $\mathbb{E}[R(\boldsymbol{\Theta}^{(K)}+\tau\boldsymbol{Z},\tau,\boldsymbol{V},K,M)]\downarrow\mathbb{E}[R(\boldsymbol{\Theta}^{(K)}+\tau\boldsymbol{Z},\tau,\boldsymbol{V},K,\infty)]$ in this limit. Because $\mathbb{E}[R(\boldsymbol{\Theta}^{(K)}+\tau\boldsymbol{Z},\tau,\boldsymbol{V},K,\infty)]$ is a lower bound on the Bayes risk at $M=\infty$ by (B.8) and we may achieve risk arbitrarily close to this lower bound by taking $M\to\infty$ in (B.9), we conclude (B.5) at $M=\infty$ as well.

□

Proof.[Proof of Lemma B.1.5(b)] The quantity $R(\boldsymbol{y},\tau,\boldsymbol{v},K,M)$ is non-negative. Define $\bar{R}(K)=\max_{\|\boldsymbol{\vartheta}\|_\infty\le K}\ell(\boldsymbol{\vartheta},\boldsymbol{0})$. Observe that $R(\boldsymbol{y},\tau,\boldsymbol{v},K,M)\le\bar{R}(K)$ for all $\boldsymbol{y},\tau,\boldsymbol{v},K,M$. Let $p^*(\boldsymbol{\theta}|\boldsymbol{y},\tau,\boldsymbol{v},K)=\frac{1}{Z}e^{-\frac{1}{2\tau^2}\|\boldsymbol{y}-\boldsymbol{\vartheta}\|^2}$. For any fixed $\boldsymbol{d}$, we have

$$\begin{aligned}\left\|\nabla_{\boldsymbol{y}}\int\ell(\boldsymbol{\vartheta},\boldsymbol{d})p^*(\boldsymbol{\vartheta}|\boldsymbol{y},\tau,\boldsymbol{v},K)\mu_{\boldsymbol{\Theta}^{(K)}|\boldsymbol{V}}(\boldsymbol{v},\mathrm{d}\boldsymbol{\vartheta})\right\|&\le\int\ell(\boldsymbol{\vartheta},\boldsymbol{d})p^*(\boldsymbol{\vartheta}|\boldsymbol{y},\tau,v)\left\|\nabla_{\boldsymbol{y}}\log p^*(\boldsymbol{\vartheta}|\boldsymbol{y},\tau,\boldsymbol{v})\right\|\mu_{\boldsymbol{\Theta}^{(K)}|\boldsymbol{V}}(\boldsymbol{v},\mathrm{d}\boldsymbol{\vartheta})\\&\le\frac{2K\sqrt{k}}{\tau^2}\int\ell(\boldsymbol{\vartheta},\boldsymbol{d})p^*(\boldsymbol{\vartheta}|\boldsymbol{y},\tau,\boldsymbol{v})\mu_{\boldsymbol{\Theta}^{(K)}|\boldsymbol{V}}(\boldsymbol{v},\mathrm{d}\boldsymbol{\vartheta})\,,\end{aligned}$$

where we have used that $\|\nabla_{\boldsymbol{y}}\log p^*(\boldsymbol{\vartheta}|\boldsymbol{y},\tau,\boldsymbol{v})\|=\frac{1}{\tau^2}(\boldsymbol{\vartheta}-\mathbb{E}_{\boldsymbol{\Theta}^{(K)}}[\boldsymbol{\Theta}^{(K)}])\le 2K\sqrt{k}/\tau^2$, and the expectation is taken with respect to $\boldsymbol{\Theta}^{(K)}$ having density $p^*(\boldsymbol{\vartheta}|\boldsymbol{y},\tau,\boldsymbol{v})$ with respect to $\mu_{\boldsymbol{\Theta}^{(K)}|V}(\boldsymbol{v},\cdot)$. Thus, for fixed $\tau,\boldsymbol{d},\boldsymbol{v}$ satisfying $\int\ell(\boldsymbol{\vartheta},\boldsymbol{d})p^*(\boldsymbol{\vartheta}|\boldsymbol{y},\tau,\boldsymbol{v})\mu_{\Theta|V}(\boldsymbol{v},\mathrm{d}\boldsymbol{\vartheta})\le\bar{R}$, the function $\boldsymbol{y}\mapsto\int\ell(\boldsymbol{\vartheta},\boldsymbol{d})p^*(\boldsymbol{\vartheta}|\boldsymbol{y},\tau,\boldsymbol{v})\mu_{\Theta|V}(\boldsymbol{v},\mathrm{d}\boldsymbol{\vartheta})$ is $2K\sqrt{k}\bar{R}/\tau^2$-Lipschitz. Because the infimum defining R can be taken over such $\boldsymbol{d}$ and infima retain a uniform Lipschitz property, $R(\boldsymbol{y},\tau,\boldsymbol{v},K,M)$ is $2K\sqrt{k}\bar{R}/\tau^2$-Lipschitz in $\boldsymbol{y}$ for fixed $\tau,\boldsymbol{v},K,M$. By a similar argument, we can establish that $R(\boldsymbol{y},\tau,\boldsymbol{v},K,M)$ is $2(K^2k+2\|\boldsymbol{y}\|K\sqrt{k})/\bar{\tau}^3$-Lipschitz in τ on the set $\tau>\bar{\tau}$ for any fixed $\bar{\tau}>0$ and any fixed $\boldsymbol{y},\boldsymbol{v},K,M$. We conclude $(\boldsymbol{y},\tau)\mapsto R(\boldsymbol{y},\tau,\boldsymbol{v},K,M)$ is continuous on $\mathbb{R}^k\times\mathbb{R}_{>0}$. Lemma B.1.5(b) has been shown.

□

Proof.[Proof of Lemma B.1.5(c)] Finally, we prove (B.7). For any $K>0$, we may write[1]

$$\mu_{\boldsymbol{\Theta}^{(K)}|\boldsymbol{V}}(\boldsymbol{v},\cdot)=\mu_{\boldsymbol{\Theta}|\boldsymbol{V}}(\boldsymbol{v},\cdot)|_{[-K,K]^k}+\mu_{\boldsymbol{\Theta}|\boldsymbol{V}}(\boldsymbol{v},([-K,K]^k)^c)\delta_{\boldsymbol{0}}(\cdot)\,. \tag{B.10}$$

Choose $\bar{K},\epsilon'>0$ such that $|\tau'-\tau|<\epsilon'$ implies

$$\int_{[-\bar{K},\bar{K}]^k}\frac{1}{Z(\boldsymbol{y},\tau',\boldsymbol{v},\infty)}e^{-\frac{1}{2\tau'^2}\|\boldsymbol{y}-\boldsymbol{\vartheta}\|^2}\mu_{\boldsymbol{\Theta}|\boldsymbol{V}}(\boldsymbol{v},\mathrm{d}\boldsymbol{\vartheta})\ge\frac{1}{2}\int\frac{1}{Z(\boldsymbol{y},\tau,\boldsymbol{v},\infty)}e^{-\frac{1}{2\tau^2}\|\boldsymbol{y}-\boldsymbol{\vartheta}\|^2}\mu_{\boldsymbol{\Theta}|\boldsymbol{V}}(\boldsymbol{v},\mathrm{d}\boldsymbol{\vartheta})\,.$$

Fix $\epsilon>0$ and $K'>K>0$ with K' possibly equal to infinity. By (B.4), we may choose $\boldsymbol{d}^*$ such that

$$\int\frac{1}{Z(\boldsymbol{y},\tau,\boldsymbol{v},K)}\ell(\boldsymbol{\vartheta},\boldsymbol{d}^*)e^{-\frac{1}{2\tau^2}\|\boldsymbol{y}-\boldsymbol{\vartheta}\|^2}\mu_{\boldsymbol{\Theta}^{(K)}|\boldsymbol{V}}(\boldsymbol{v},\mathrm{d}\boldsymbol{\vartheta})\le(1+\epsilon)R(\boldsymbol{y},\tau,\boldsymbol{v},K,\infty)\,. \tag{B.11}$$

By the definition of $\bar{K}$, there exists $\boldsymbol{\vartheta}^*\in[-\bar{K},\bar{K}]^k$ such that

$$\ell(\boldsymbol{\vartheta}^*,\boldsymbol{d}^*)\le 2(1+\epsilon)R(\boldsymbol{y},\tau,\boldsymbol{v},K,\infty)\,.$$

By (B.3), we conclude that

$$\ell(\boldsymbol{\vartheta},\boldsymbol{d}^*)\le C\left(1+\sqrt{2(1+\epsilon)R(\boldsymbol{y},\tau,\boldsymbol{v},K,\infty)}+\sqrt{\ell(\boldsymbol{\vartheta},\boldsymbol{d}^*)}\right)\|\boldsymbol{\vartheta}-\boldsymbol{\vartheta}^*\|\,,$$

whence

$$\ell(\boldsymbol{\vartheta},\boldsymbol{d}^*)\le\left(1+\sqrt{2(1+\epsilon)R(\boldsymbol{y},\tau,\boldsymbol{v},K,\infty)}+3C\|\boldsymbol{\vartheta}-\boldsymbol{\vartheta}^*\|\right)^2\,. \tag{B.12}$$

[1] Precisely, for any regular conditional probability distribution $\mu_{\boldsymbol{\Theta}|\boldsymbol{V}}$ for $\boldsymbol{\Theta}$ given $\boldsymbol{V}$, this formula gives a valid version of a regular conditional probability distribution for $\boldsymbol{\Theta}^{(K)}$ given $\boldsymbol{V}$. We assume we use this version throughout our proof.

Then

$$\begin{aligned}
&\left|\int \ell(\boldsymbol{\vartheta},\boldsymbol{d}^*)e^{-\frac{1}{2\tau'^2}\|\boldsymbol{y}-\boldsymbol{\vartheta}\|^2}\mu_{\boldsymbol{\Theta}^{(K')}|\boldsymbol{V}}(\boldsymbol{v},\mathrm{d}\boldsymbol{\vartheta}) - \int \ell(\boldsymbol{\vartheta},\boldsymbol{d}^*)e^{-\frac{1}{2\tau^2}\|\boldsymbol{y}-\boldsymbol{\vartheta}\|^2}\mu_{\boldsymbol{\Theta}^{(K)}|\boldsymbol{V}}(\boldsymbol{v},\mathrm{d}\boldsymbol{\vartheta})\right| \\
&\quad\le \left|\int \ell(\boldsymbol{\vartheta},\boldsymbol{d}^*)e^{-\frac{1}{2\tau'^2}\|\boldsymbol{y}-\boldsymbol{\vartheta}\|^2}\mu_{\boldsymbol{\Theta}^{(K')}|\boldsymbol{V}}(\boldsymbol{v},\mathrm{d}\boldsymbol{\vartheta}) - \int \ell(\boldsymbol{\vartheta},\boldsymbol{d}^*)e^{-\frac{1}{2\tau'^2}\|\boldsymbol{y}-\boldsymbol{\vartheta}\|^2}\mu_{\boldsymbol{\Theta}^{(K)}|\boldsymbol{V}}(\boldsymbol{v},\mathrm{d}\boldsymbol{\vartheta})\right| \\
&\qquad + \left|\int \ell(\boldsymbol{\vartheta},\boldsymbol{d}^*)e^{-\frac{1}{2\tau'^2}\|\boldsymbol{y}-\boldsymbol{\vartheta}\|^2}\mu_{\boldsymbol{\Theta}^{(K)}|\boldsymbol{V}}(\boldsymbol{v},\mathrm{d}\boldsymbol{\vartheta}) - \int \ell(\boldsymbol{\vartheta},\boldsymbol{d}^*)e^{-\frac{1}{2\tau^2}\|\boldsymbol{y}-\boldsymbol{\vartheta}\|^2}\mu_{\boldsymbol{\Theta}^{(K)}|\boldsymbol{V}}(\boldsymbol{v},\mathrm{d}\boldsymbol{\vartheta})\right| \\
&\quad\le \int_{([-K,K]^k)^c} \ell(\boldsymbol{\vartheta},\boldsymbol{d}^*)e^{-\frac{1}{2\tau'^2}\|\boldsymbol{y}-\boldsymbol{\vartheta}\|^2}\mu_{\boldsymbol{\Theta}|\boldsymbol{V}}(\boldsymbol{v},\mathrm{d}\boldsymbol{\vartheta}) + \ell(\boldsymbol{0},\boldsymbol{d}^*)e^{-\frac{1}{2\tau'^2}\|\boldsymbol{y}\|^2}\mu_{\boldsymbol{\Theta}|\boldsymbol{V}}(\boldsymbol{v},([-K,K]^k)^c) \\
&\qquad + \left|\int \ell(\boldsymbol{\vartheta},\boldsymbol{d}^*)e^{-\frac{1}{2\tau'^2}\|\boldsymbol{y}-\boldsymbol{\vartheta}\|^2}\mu_{\boldsymbol{\Theta}^{(K)}|\boldsymbol{V}}(\boldsymbol{v},\mathrm{d}\boldsymbol{\vartheta}) - \int \ell(\boldsymbol{\vartheta},\boldsymbol{d}^*)e^{-\frac{1}{2\tau^2}\|\boldsymbol{y}-\boldsymbol{\vartheta}\|^2}\mu_{\boldsymbol{\Theta}^{(K)}|\boldsymbol{V}}(\boldsymbol{v},\mathrm{d}\boldsymbol{\vartheta})\right| \\
&\quad\le \xi(K,\tau')(1+R(\boldsymbol{y},\tau,\boldsymbol{v},K,\infty))\,,
\end{aligned}$$

for some $\xi(K,\tau') \to 0$ as $K\to\infty$, $\tau'\to\tau$ because the conditional measure $\mu_{\boldsymbol{\Theta}|\boldsymbol{V}}(\boldsymbol{v},\cdot)$ has finite second moment and ℓ is bounded by (B.12). Then, by (B.11),

$$\begin{aligned}
Z(\boldsymbol{y},\tau',\boldsymbol{v},K')R(\boldsymbol{y},\tau',\boldsymbol{v},K',\infty) &\le \int \ell(\boldsymbol{\vartheta},\boldsymbol{d}^*)e^{-\frac{1}{2\tau^2}\|\boldsymbol{y}-\boldsymbol{\vartheta}\|^2}\mu_{\boldsymbol{\Theta}^{(K')}|\boldsymbol{V}}(\boldsymbol{v},\mathrm{d}\boldsymbol{\vartheta}) \\
&\le (1+\epsilon)Z(\boldsymbol{y},\tau,\boldsymbol{v},K)R(\boldsymbol{y},\tau,\boldsymbol{v},K,\infty) + \xi(K,\tau')(1+R(\boldsymbol{y},\tau,\boldsymbol{v},K,\infty))\,.
\end{aligned}$$

By dominated convergence, we have that $Z(\boldsymbol{y},\tau',\boldsymbol{v},K') \to Z(\boldsymbol{y},\tau,\boldsymbol{v},\infty)$ as $\tau'\to\tau, K'\to\infty$. Also, $\bar{R}(K) = \max_{\|\boldsymbol{\vartheta}\|_\infty \le K} \ell(\boldsymbol{\vartheta},\boldsymbol{0})$ cannot diverge at finite K. Thus, applying the previous display with K,ϵ fixed allows us to conclude that $R(\boldsymbol{y},\tau,\boldsymbol{v},K',\infty)$ is uniformly bounded over $K' > K$ and τ' in a neighborhood of τ. Then, taking $K' = \infty$ and $K\to\infty$, $\tau'\to\tau$ followed by $\epsilon\to 0$ allows us to conclude that

$$\lim_{\substack{K\to\infty\\ \tau'\to\tau}} R(\boldsymbol{y},\tau',\boldsymbol{v},K,\infty) = R(\boldsymbol{y},\tau,\boldsymbol{v},\infty,\infty)\,. \tag{B.13}$$

for every fixed $\boldsymbol{y},\boldsymbol{v}$. Moreover,

$$\begin{aligned}
R(\boldsymbol{y},\tau,\boldsymbol{v},K,M) &= \inf_{\|\boldsymbol{d}\|_\infty\le M}\int \frac{1}{Z}\ell(\boldsymbol{\vartheta},\boldsymbol{d})e^{-\frac{1}{2\tau^2}\|\boldsymbol{y}-\boldsymbol{\vartheta}\|^2}\mu_{\boldsymbol{\Theta}^{(K)}|\boldsymbol{V}}(\boldsymbol{v},\mathrm{d}\boldsymbol{\vartheta}) \\
&\le \int \frac{1}{Z}\ell(\boldsymbol{\vartheta},\boldsymbol{0})e^{-\frac{1}{2\tau^2}\|\boldsymbol{y}-\boldsymbol{\vartheta}\|^2}\mu_{\boldsymbol{\Theta}^{(K)}|\boldsymbol{V}}(\boldsymbol{v},\mathrm{d}\boldsymbol{\vartheta}) \\
&\le \int \frac{1}{Z}C(1+\|\boldsymbol{\Theta}^{(K)}\|^2)e^{-\frac{1}{\tau^2}(\boldsymbol{y}-\boldsymbol{\vartheta})^2}\mu_{\boldsymbol{\Theta}^{(K)}|\boldsymbol{V}}(\boldsymbol{v},\mathrm{d}\boldsymbol{\vartheta}) \\
&= C(1+\mathbb{E}[\|\boldsymbol{\Theta}^{(K)}\|^2|\boldsymbol{\Theta}^{(K)}+\tau\boldsymbol{G}=\boldsymbol{y},\boldsymbol{V}=\boldsymbol{v}])\,.
\end{aligned}$$

Thus, $R(\boldsymbol{\Theta}^{(K)}+\tau\boldsymbol{Z},\tau,\boldsymbol{V},K,M)$ is uniformly integrable as we vary τ,K,M. Because the total variation distance between $(\boldsymbol{\Theta}^{(K)}+\tau'\boldsymbol{Z},\boldsymbol{V})$ and $(\boldsymbol{\Theta}+\tau\boldsymbol{Z},\boldsymbol{V}))$ goes to 0 as $K\to\infty$ and $\tau'\to\tau$, for any discrete sequence $(K,\tau')\to(\infty,\tau)$, there exists a probability space containing variables $\tilde{\boldsymbol{Y}}^{(K,\tau')},\tilde{\boldsymbol{V}},\tilde{\boldsymbol{Y}}$ such that $(\tilde{\boldsymbol{Y}}^{(K,\tau')},\tilde{\boldsymbol{V}}) = (\tilde{\boldsymbol{Y}},\tilde{\boldsymbol{V}})$ eventually. Thus, Eq. (B.13) and uniform integrability imply (B.7).

□

B.2 Proof for reduction from GFOMs to AMP (Lemma 3.5.1)

In this section, we prove Lemma 3.5.1.

B.2.1 A general change of variables

For any GFOM (3.1), there is a collection of GFOMs to which it is, up to a change of variabes, equivalent. In this section, we specify these GFOMs and the corresponding changes of variables.

The change of variables is determined by a collection of $r \times r$ matrices $(\boldsymbol{\xi}_{t,s})_{t\geq 1, 1\leq s\leq t}$, $(\boldsymbol{\zeta}_{t,s})_{t\geq 1, 0\leq s<t}$. We will often omit subscripts outside of the parentheses. Define recursively the functions $(f_t)_{t\geq 0}$, $(\phi_t)_{t\geq 1}$

$$\begin{aligned} f_t(\boldsymbol{b}^1,\dots,\boldsymbol{b}^t;y,\boldsymbol{u}) &= F_t^{(1)}(\phi_1(\boldsymbol{b}^1;y,\boldsymbol{u}),\dots,\phi_t(\boldsymbol{b}^1,\dots,\boldsymbol{b}^t;y,\boldsymbol{u});y,\boldsymbol{u}) \\ \phi_t(\boldsymbol{b}^1,\dots,\boldsymbol{b}^t;y,\boldsymbol{u}) &= \boldsymbol{b}^t + \sum_{s=0}^{t-1} f_s(\boldsymbol{b}^1,\dots,\boldsymbol{b}^s;y,\boldsymbol{u})\boldsymbol{\zeta}_{t,s}^\mathsf{T} \\ &\qquad + G_t^{(2)}(\phi_1(\boldsymbol{b}^1;y,\boldsymbol{u}),\dots,\phi_{t-1}(\boldsymbol{b}^1,\dots,\boldsymbol{b}^{t-1};y,\boldsymbol{u});y,\boldsymbol{u}), \end{aligned} \tag{B.14a}$$

initialized by $f_0(y,\boldsymbol{u}) = F_0^{(1)}(y,\boldsymbol{u})$ (here $\boldsymbol{b}^s,\boldsymbol{u}\in\mathbb{R}^r$), and define recursively the functions $(g_t)_{t\geq 1}$, $(\varphi_t)_{t\geq 1}$

$$\begin{aligned} \varphi_{t+1}(\boldsymbol{a}^1,\dots,\boldsymbol{a}^{t+1};\boldsymbol{v}) &= \boldsymbol{a}^{t+1} + \sum_{s=1}^{t} g_s(\boldsymbol{a}^1,\dots,\boldsymbol{a}^{t+1};\boldsymbol{v})\boldsymbol{\xi}_{t,s}^\mathsf{T} \\ &\qquad + F_t^{(2)}(\phi_1(\boldsymbol{a}^1;\boldsymbol{v}),\dots,\phi_t(\boldsymbol{a}^1,\dots,\boldsymbol{a}^t;\boldsymbol{v});\boldsymbol{v}), \\ g_t(\boldsymbol{a}^1,\dots,\boldsymbol{a}^t;\boldsymbol{v}) &= G_t^{(1)}(\varphi_1(\boldsymbol{a}^1;\boldsymbol{v}),\dots,\varphi_t(\boldsymbol{a}^1,\dots,\boldsymbol{a}^t;\boldsymbol{v});\boldsymbol{v}), \end{aligned} \tag{B.14b}$$

initialized by $\varphi_1(\boldsymbol{a}^1;\boldsymbol{v}) = a^1 + F_0^{(2)}(\boldsymbol{v})$ (here $\boldsymbol{a}^s,\boldsymbol{v}\in\mathbb{R}^r$). Algebraic manipulation verifies that the iteration

$$\begin{aligned} \boldsymbol{a}^{t+1} &= \boldsymbol{X}^\mathsf{T} f_t(\boldsymbol{b}^1,\dots,\boldsymbol{b}^t;y,\boldsymbol{u}) - \sum_{s=1}^{t} g_s(\boldsymbol{a}^1,\dots,\boldsymbol{a}^s;\boldsymbol{v})\boldsymbol{\xi}_{t,s}^\mathsf{T}, \\ \boldsymbol{b}^t &= \boldsymbol{X} g_t(\boldsymbol{a}^1,\dots,\boldsymbol{a}^t;\boldsymbol{v}) - \sum_{s=0}^{t-1} f_s(\boldsymbol{b}^1,\dots,\boldsymbol{b}^s;y,\boldsymbol{u})\boldsymbol{\zeta}_{t,s}^\mathsf{T} \end{aligned} \tag{B.15}$$

initialized by $\boldsymbol{a}^1 = \boldsymbol{X}^\mathsf{T} f_0(y,\boldsymbol{u})$ generates sequences $(\boldsymbol{a}^t)_{t\geq 1}$, $(\boldsymbol{b}^t)_{t\geq 1}$ which satisfy

$$\begin{aligned} \boldsymbol{v}^t &= \varphi_t(\boldsymbol{a}^1,\dots,\boldsymbol{a}^t;\boldsymbol{v}), \quad t\geq 1, \\ \boldsymbol{u}^t &= \phi_t(\boldsymbol{b}^1,\dots,\boldsymbol{b}^t;y,\boldsymbol{u}), \quad t\geq 1. \end{aligned}$$

Thus, $(\boldsymbol{\xi}_{t,s})$, $(\boldsymbol{\zeta}_{t,s})$ index a collection of GFOMs which, up to a change of variables, are equivalent.

B.2.2 Approximate message passing and state evolution

We call the iteration (B.15) an approximate message passing algorithm if the matrices $(\boldsymbol{\xi}_{t,s}),(\boldsymbol{\zeta}_{t,s})$ satisfy a certain model-specific recursion involving the functions f_t, g_t. The state evolution characterization of the iterates (see Eq. (3.17)) holds whenever the matrices $\boldsymbol{\xi}_{t,s}$, $\boldsymbol{\zeta}_{t,s}$ satisfy this recursion. In this section, we specify this recursion and the parameters $(\boldsymbol{\alpha}_s),(\boldsymbol{T}_{s,s'})$ in both the high-dimensional regression and low-rank matrix estimation models.

High-dimensional regression AMP

In the high-dimensional regression model, $r = 1$ and $\xi_{t,s}$, $\zeta_{t,s}$, α_t, and $T_{s,s'}$ will be scalars (hence, written with non-bold font). The recursion defining $\xi_{t,s}$, $\zeta_{t,s}$ also defines (α_t), $T_{s,s'}$ as well as a collection of scalars $(\Sigma_{s,t})_{s,t\geq 0}$ which did not appear in the statement of Lemma 3.5.1. The recursion, whose lines are implemented in the order in which they appear, is

$$\begin{aligned}
\xi_{t,s} &= \mathbb{E}[\partial_{B^s} f_t(B^1,\dots,B^t; h(B^0,W),U)], \quad 1\leq s\leq t,\\
\alpha_{t+1} &= \mathbb{E}[\partial_{B^0} f_t(B^1,\dots,B^t; h(B^0,W),U)],\\
T_{s+1,t+1} &= \mathbb{E}[f_s(B^1,\dots,B^s; h(B^0,W),U) f_t(B^1,\dots,B^t; h(B^0,W),U)], \quad 0\leq s\leq t,\\
\zeta_{t,s} &= \frac{1}{\delta}\mathbb{E}[\partial_{Z^{s+1}} g_t(\alpha_1\Theta + Z^1,\dots,\alpha_t\Theta + Z^t; V)], \quad 0\leq s\leq t-1,\\
\Sigma_{0,t} &= \frac{1}{\delta}\mathbb{E}[\Theta g_t(\alpha_1\Theta + Z^1,\dots,\alpha_t\Theta + Z^t; V)],\\
\Sigma_{s,t} &= \frac{1}{\delta}\mathbb{E}[g_s(\alpha_1\Theta + Z^1,\dots,\alpha_t\Theta + Z^s; V) g_t(\alpha_1\Theta + Z^1,\dots,\alpha_t\Theta + Z^t; V)], \quad 1\leq s\leq t,
\end{aligned} \tag{B.16}$$

where $\Theta\sim\mu_\Theta$, $U\sim\mu_U$, $V\sim\mu_V$, $W\sim\mu_W$, $(B^0,\dots,B^t)\sim\mathsf{N}(\mathbf{0},\mathbf{\Sigma}_{[0:t]})$, $(Z^1,\dots,Z^t)\sim\mathsf{N}(\mathbf{0},\boldsymbol{T}_{[1:t]})$, all independent. We initialize just before the second line with $\Sigma_{0,0} = \mathbb{E}[\Theta^2]$.

Eq. (3.17) for $(\alpha_s),(T_{s,s'})$ defined in this way is a special case of Proposition 5 of [104], as we now explain. We fix iteration t design an algorithm that agrees, after a change of variables, with iteration (3.16) up to iteration t and to which we can apply the results of [104]. Because we take $n,p\to\infty$ before $t\to\infty$, this establishes the result.

We view the first t iterations of (3.16) as acting on matrices $\tilde{\boldsymbol{a}}^s\in\mathbb{R}^{p\times(t+1)}$ and $\tilde{\boldsymbol{b}}^s\in\mathbb{R}^{n\times(t+1)}$ as follows. Define $\tilde{\boldsymbol{a}}^s$ to be the matrix whose first column is $\boldsymbol{\theta}$ and whose i^{th} column is $\boldsymbol{a}^{i-1}$ for $2\leq i\leq s+1$ and is $\mathbf{0}$ for $i>s+1$; define $\tilde{\boldsymbol{b}}^s$ to be the matrix whose first column is $\boldsymbol{X\theta}$ and whose i^{th} column is $\boldsymbol{b}^{i=1}$ for $2\leq i\leq s+1$ and is $\mathbf{0}$ for $i>s+1$. The following change of variables transforms (3.16) into equations (28) and (29) of Proposition 5 in [104]. Our notation is on the right and is separated from the notation of [104] by the symbol "←".

$$\tilde{A}\leftarrow X,$$

$$u^s(i)\leftarrow\begin{cases}\boldsymbol{X\theta} & i=1,\\ \boldsymbol{b}^{i-1} & 2\leq i\leq s+1,\\ \mathbf{0} & \text{otherwise,}\end{cases}\qquad\text{and}\qquad v^s(i)\leftarrow\begin{cases}\boldsymbol{a}^{i-1}-\alpha_{i-1}\boldsymbol{\theta} & 2\leq i\leq s+1,\\ \mathbf{0} & \text{otherwise,}\end{cases}$$

$$y(i)\leftarrow\begin{cases}\boldsymbol{v} & i=1,\\ \boldsymbol{\theta} & i=2,\\ \mathbf{0} & \text{otherwise,}\end{cases}\qquad\text{and}\qquad w(i)\leftarrow\begin{cases}\boldsymbol{u} & i=1,\\ \boldsymbol{w} & i=2,\\ \mathbf{0} & \text{otherwise,}\end{cases}$$

$$\widehat{e}(v, y; s)(i) \leftarrow \begin{cases} y(2) & i = 1, \\ g_{i-1}(v(2) + \alpha_1 y(2), \dots, v(i+1) + \alpha_i y(2); y(1)) & 2 \le i \le s+1, \\ \mathbf{0} & \text{otherwise}, \end{cases}$$

$$\widehat{h}(u, w; s)(i) \leftarrow \begin{cases} f_{i-1}(u(2), \dots, u(i+1); h(u(1), w(2)), w(1)), & 1 \le i \le s+1, \\ \mathbf{0} & \text{otherwise}, \end{cases}$$

where the "(i)" notation indexes columns of a matrix. The Onsager correction coefficients $(\xi_{t,s})$ and $(\zeta_{t,s})$ correspond, after a change of variables, to entries in the matrices D_s and B_s in [104].

$$(\mathsf{D}_s)_{i,j} = \mathbb{E}[\partial_{u(j)} \widehat{h}(U, W; s)] \leftarrow \begin{cases} \mathbb{E}[\partial_{B^{j-1}} f_{i-1}(B^1, \dots, B^i; h(B^0, W), U)] & 1 \le j-1 \le i \le s+1, \\ 0 & \text{otherwise}, \end{cases},$$

$$(\mathsf{B}_s)_{i,j} = \frac{1}{\delta} \mathbb{E}[\partial_{v(j)} \widehat{e}(V, Y; i)] \leftarrow \begin{cases} 0 & i = 1 \text{ or } j = 1, \\ \frac{1}{\delta} \mathbb{E}[\partial_{Z^{j-1}} g_i(\alpha_1 \Theta + Z^1, \dots, \alpha_i \Theta + Z^i; V)] & 2 \le j \le i+1 \le s+2, \\ 0 & \text{otherwise}. \end{cases}$$

The Onsager coefficients and state evolution coefficients are arrived at through the change of variables:

$$(\mathsf{B}_s)_{s+1,s'+2} \leftarrow \zeta_{s,s'}, \quad (\mathsf{D}_s)_{s+1,s'+1} \leftarrow \xi_{s,s'} \quad (\mathsf{D}_s)_{s+1,1} \leftarrow \alpha_s.$$

We remark that in [104] the quantities $(\mathsf{B}_s)_{s+1,s'+2}$, $(\mathsf{D}_s)_{s+1,s'+1}$, and $(\mathsf{D}_s)_{s+1,1}$ are empirical averages. Because they concentration well on their population averages, we may replace them with their population averages, as we do here, without affecting the validity of state evolution. This observation is common in the AMP literature: see, for example, the relationship between Theorem 1 and Corollary 2 of [32]. The state evolution matrices now correspond to

$$\begin{aligned} \mathbb{E}[V^{s+1}(s+1) V^{s+1}(s'+1)] &= \mathbb{E}[\widehat{h}(U, W; s)(s) \widehat{h}(U, W; s)(s')] \\ &\leftarrow \mathbb{E}[f_{s-1}(B^1, \dots, B^{s-1}; h(B^0, W), U) f_{s'-1}(B^1, \dots, B^{s'-1}; h(B^0, W), U)] \\ &= T_{s,s'}, \\ \mathbb{E}[U^{s+1}(s+1) U^{s+1}(s'+1)] &= \frac{1}{\delta} \mathbb{E}[\widehat{e}(V, Y; s+1)(s+1) \widehat{e}(V, Y; s+1)(s'+1)] \\ &\leftarrow \frac{1}{\delta} \mathbb{E}[g_s(\alpha_1 \Theta + Z^1, \dots, \alpha_s \Theta + Z^s; V) g_{s'}(\alpha_1 \Theta + Z^1, \dots, \alpha_{s'} \Theta + Z^{s'}; V)] \\ &= \Sigma_{s,s'}. \end{aligned}$$

From these changes of variables, Eq. (3.17) holds in the high-dimensional regression model from Theorem 1 and Proposition 5 of [104].

Low-rank matrix estimation AMP

In the low-ank matrix estimation model, the recrusion defining $(\boldsymbol{x}_{t,x})$, $(\boldsymbol{\zeta}_{t,s})$ also defines $(\boldsymbol{\alpha}_t)$, $(\boldsymbol{T}_{s,t})_{s,t \ge 1}$ as well as collections of $r \times r$ matrices $(\boldsymbol{\gamma}_t)_{t \ge 1}$, $(\boldsymbol{\Sigma}_{s,t})_{s,t \ge 0}$ which did not appear in Lemma 3.5.1. The recursion, whose lines are implemented in the order in which they appear, is

$$\begin{aligned}
\boldsymbol{\xi}_{t,s} &= \mathbb{E}[\nabla_{\tilde{\boldsymbol{Z}}^s} f_t(\boldsymbol{\gamma}_1\boldsymbol{\Lambda} + \tilde{\boldsymbol{Z}}^1, \ldots, \boldsymbol{\gamma}_t\boldsymbol{\Lambda} + \tilde{\boldsymbol{Z}}^t; 0, \boldsymbol{U})], \quad 1 \le s \le t, \\
\boldsymbol{\alpha}_{t+1} &= \mathbb{E}[f_t(\boldsymbol{\gamma}_1\boldsymbol{\Lambda} + \tilde{\boldsymbol{Z}}^1, \ldots, \boldsymbol{\gamma}_t\boldsymbol{\Lambda} + \tilde{\boldsymbol{Z}}^t; 0, \boldsymbol{U})\boldsymbol{\Lambda}^\mathsf{T}], \\
\boldsymbol{T}_{s+1,t+1} &= \mathbb{E}[f_s(\boldsymbol{\gamma}_1\boldsymbol{\Lambda} + \tilde{\boldsymbol{Z}}^1, \ldots, \boldsymbol{\gamma}_t\boldsymbol{\Lambda} + \tilde{\boldsymbol{Z}}^s; 0, \boldsymbol{U}) f_t(\boldsymbol{\gamma}_1\boldsymbol{\Lambda} + \tilde{\boldsymbol{Z}}^1, \ldots, \boldsymbol{\gamma}_t\boldsymbol{\Lambda} + \tilde{\boldsymbol{Z}}^t; 0, \boldsymbol{U})^\mathsf{T}], \ s \le t, \\
\boldsymbol{\zeta}_{t,s} &= \frac{1}{\delta}\mathbb{E}[\nabla_{\boldsymbol{Z}^{s+1}} g_t(\boldsymbol{\alpha}_1\boldsymbol{\Theta} + Z^1, \ldots, \boldsymbol{\alpha}_t\boldsymbol{\Theta} + \boldsymbol{Z}^t; \boldsymbol{V})], \quad 0 \le s \le t-1, \\
\boldsymbol{\gamma}_t &= \frac{1}{\delta}\mathbb{E}[g_t(\boldsymbol{\alpha}_1\boldsymbol{\Theta} + \boldsymbol{Z}^1, \ldots, \boldsymbol{\alpha}_t\boldsymbol{\Theta} + \boldsymbol{Z}^t; \boldsymbol{V})\boldsymbol{\Theta}^\mathsf{T}], \\
\boldsymbol{\Sigma}_{s,t} &= \frac{1}{\delta}\mathbb{E}[g_s(\boldsymbol{\alpha}_1\boldsymbol{\Theta} + \boldsymbol{Z}^1, \ldots, \boldsymbol{\alpha}_t\boldsymbol{\Theta} + \boldsymbol{Z}^s; \boldsymbol{V}) g_t(\boldsymbol{\alpha}_1\boldsymbol{\Theta} + \boldsymbol{Z}^1, \ldots, \boldsymbol{\alpha}_t\boldsymbol{\Theta} + \boldsymbol{Z}^t; \boldsymbol{V})^\mathsf{T}], \quad 1 \le s \le t,
\end{aligned} \tag{B.17}$$

where $\boldsymbol{\Lambda} \sim \mu_{\boldsymbol{\Lambda}}$ $\boldsymbol{U} \sim \mu_{\boldsymbol{U}}$, $\boldsymbol{\Theta} \sim \mu_{\boldsymbol{\Theta}}$, $\boldsymbol{V} \sim \mu_{\boldsymbol{V}}$, $(\tilde{\boldsymbol{Z}}^1, \ldots, \tilde{\boldsymbol{Z}}^t) \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{[1:t]})$, and $(\boldsymbol{Z}^1, \ldots, \boldsymbol{Z}^t) \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{T}_{[1:t]})$, all independent. Here ∇ denotes the Jacobian with respect to subscripted (vectorial) argument, which exists almost everywhere because the functions involved are Lipschitz and the random variables have density with respect to Lebesgue measure [83, pg. 81]. As with $\boldsymbol{T}_{[1:t]}$, we define $\boldsymbol{\Sigma}_{[1:t]}$ to be the $rt \times rt$ block matrix with block (s,t) given by $\boldsymbol{\Sigma}_{s,t}$. We initialize at the second line with $\boldsymbol{\alpha}_1 = \mathbb{E}[f_0(0, \boldsymbol{U})\boldsymbol{\Lambda}^\mathsf{T}]$. In addition to (3.17), we have

$$\frac{1}{n}\sum_{i=1}^n \psi(\boldsymbol{b}_i^1, \ldots, \boldsymbol{b}_i^t, \boldsymbol{u}_i, \boldsymbol{\Lambda}_i) \xrightarrow{\mathrm{p}} \mathbb{E}[\psi(\boldsymbol{\gamma}_1\boldsymbol{\Lambda} + \tilde{\boldsymbol{Z}}^1, \ldots, \boldsymbol{\gamma}_t\boldsymbol{\Lambda} + \tilde{\boldsymbol{Z}}^t, \boldsymbol{U}, \boldsymbol{\Lambda})],$$

where we remind the reader that $\psi : \mathbb{R}^{r(t+2)} \to \mathbb{R}$ is any pseudo-Lipschitz function of order 2.

We now show Eq. (3.17) for $(\alpha_s), (T_{s,s'})$ defined in this way. We consider the $r = 1$ case, as $r > 1$ is similar by requires more notational overhead. Because $\boldsymbol{X} = \frac{1}{n}\boldsymbol{\Lambda}\boldsymbol{\theta}^\mathsf{T} + \boldsymbol{Z}$, we have

$$\begin{aligned}
\boldsymbol{a}^{t+1} - \frac{1}{n}\langle \boldsymbol{\Lambda}, f_t(\boldsymbol{b}^1, \ldots, \boldsymbol{b}^t, 0, \boldsymbol{u})\rangle \boldsymbol{\theta} &= \boldsymbol{Z}^\mathsf{T} f_t(\boldsymbol{b}^1, \ldots, \boldsymbol{b}^t, 0, \boldsymbol{u}) - \sum_{s=1}^t \xi_{t,s} g_s(\boldsymbol{a}^1, \ldots, \boldsymbol{a}^s, \boldsymbol{v}), \\
\boldsymbol{b}^t - \frac{1}{n}\langle \boldsymbol{\theta}, g_t(\boldsymbol{a}^1, \ldots, \boldsymbol{a}^t, \boldsymbol{v})\rangle \boldsymbol{\Lambda} &= \boldsymbol{Z} g_t(\boldsymbol{a}^1, \ldots, \boldsymbol{a}^t, \boldsymbol{v}) - \sum_{s=0}^{t-1} \zeta_{t,s} f_s(\boldsymbol{b}^1, \ldots, \boldsymbol{b}^s, \boldsymbol{y}, \boldsymbol{u}).
\end{aligned}$$

We introduce a change of variables:

$$\begin{aligned}
\hat{f}_t(d^1, \ldots, d^t, u, \lambda) &\triangleq f_t(d^1 + \gamma_1\lambda, \ldots, d^t + \gamma_t\lambda, 0, u), & \boldsymbol{d}^t &= \boldsymbol{b}^t - \gamma_t\boldsymbol{\Lambda} \in \mathbb{R}^n, \\
\hat{g}_t(c^1, \ldots, c^t, v, \theta) &\triangleq g_t(c^1 + \alpha_1\theta, \ldots, c^t + \alpha_t\theta, v), & \boldsymbol{c}^t &= \boldsymbol{a}^t - \alpha_t\boldsymbol{\theta} \in \mathbb{R}^p.
\end{aligned}$$

Because f_t, g_t are Lipschitz continuous, so too are $\hat{f}_t$, $\hat{g}_t$. We have

$$\begin{aligned}
\boldsymbol{a}^{t+1} - \frac{1}{n}\langle \boldsymbol{\Lambda}, \hat{f}_t(\boldsymbol{d}^1, \ldots, \boldsymbol{d}^t, \boldsymbol{u}, \boldsymbol{\Lambda})\rangle \boldsymbol{\theta} &= \boldsymbol{Z}^\mathsf{T} \hat{f}_t(\boldsymbol{d}^1, \ldots, \boldsymbol{d}^t, \boldsymbol{u}, \boldsymbol{\Lambda}) - \sum_{s=1}^t \xi_{t,s} \hat{g}_s(\boldsymbol{c}^1, \ldots, \boldsymbol{c}^s, \boldsymbol{v}, \boldsymbol{\theta}), \\
\boldsymbol{b}^t - \frac{1}{n}\langle \boldsymbol{\theta}, \hat{g}_t(\boldsymbol{c}^1, \ldots, \boldsymbol{c}^t, \boldsymbol{v}, \boldsymbol{\theta})\rangle \boldsymbol{\Lambda} &= \boldsymbol{Z} \hat{g}_t(\boldsymbol{c}^1, \ldots, \boldsymbol{c}^t, \boldsymbol{v}, \boldsymbol{\theta}) - \sum_{s=0}^{t-1} \zeta_{t,s} \hat{f}_s(\boldsymbol{b}^1, \ldots, \boldsymbol{b}^s, \boldsymbol{u}, \boldsymbol{\Lambda}).
\end{aligned}$$

Define

$$\hat{\boldsymbol{c}}^{t+1} = \boldsymbol{Z}^{\mathsf{T}} \hat{f}_t(\hat{\boldsymbol{d}}^1, \dots, \hat{\boldsymbol{d}}^t, \boldsymbol{u}, \boldsymbol{\Lambda}) - \sum_{s=1}^{t} \xi_{t,s} \hat{g}_s(\hat{\boldsymbol{c}}^1, \dots, \hat{\boldsymbol{c}}^t, \boldsymbol{v}, \boldsymbol{\theta}),$$
$$\hat{\boldsymbol{d}}^t = \boldsymbol{Z} \hat{g}_t(\hat{\boldsymbol{c}}^1, \dots, \hat{\boldsymbol{c}}^t, \boldsymbol{v}, \boldsymbol{\theta}) - \sum_{s=0}^{t-1} \zeta_{t,s} \hat{f}_s(\hat{\boldsymbol{d}}^1, \dots, \hat{\boldsymbol{d}}^t, \boldsymbol{u}, \boldsymbol{\Lambda}).$$

We can analzye this iteration via the same techniques we used to analyze AMP in the high-dimensional regression model in the previous section [104]. In particular, for any pseudo-Lipschitz function $\psi : \mathbb{R}^{t+2} \to \mathbb{R}$ of order 2, we have

$$\begin{aligned} \frac{1}{p} \sum_{j=1}^{p} \psi(\hat{c}_j^1, \dots, \hat{c}_j^t, v_j, \theta_j) &\xrightarrow{\mathrm{p}} \mathbb{E}[\psi(Z^1, \dots, Z^t, V, \Theta)], \\ \frac{1}{n} \sum_{i=1}^{n} \psi(\hat{d}_i^1, \dots, \hat{d}_i^t, u_i, \lambda_i) &\xrightarrow{\mathrm{p}} \mathbb{E}[\psi(\tilde{Z}^1, \dots, \tilde{Z}^t, U, \Lambda)]. \end{aligned} \tag{B.18}$$

Now, tøestablish (3.17), it suffices to show

$$\frac{1}{n} \|\hat{\boldsymbol{c}}^t - \boldsymbol{c}^t\|_2^2 \xrightarrow{\mathrm{p}} 0, \qquad \frac{1}{n} \|\hat{\boldsymbol{d}}^t - \boldsymbol{d}^t\|_2^2 \xrightarrow{\mathrm{p}} 0. \tag{B.19}$$

We proceed by induction. By the weak law of large numbers, we have that $\frac{1}{n}\langle \boldsymbol{\Lambda}, \hat{f}_0(\boldsymbol{\Lambda}, \boldsymbol{u})\rangle = \frac{1}{n}\langle \boldsymbol{\Lambda}, f_0(0, \boldsymbol{u})\rangle \xrightarrow{\mathrm{p}} \alpha_1$. Therefore, $\boldsymbol{c}^1 = \boldsymbol{Z}^{\mathsf{T}} \hat{f}_0(\boldsymbol{\Lambda}, \boldsymbol{u}) + o_p(1)\boldsymbol{\theta} = \hat{\boldsymbol{c}}^1 + o_p(1)\boldsymbol{\theta}$. Since $\frac{1}{p}\|\boldsymbol{\theta}\|_2^2 \xrightarrow{\mathrm{p}} \mathbb{E}[\Theta^2]$, we have that $\frac{1}{n}\|\boldsymbol{c}^1 - \hat{\boldsymbol{c}}^1\|_2^2 \xrightarrow{\mathrm{p}} 0$.

Because $\hat{g}_1$ is Lipschitz and $\frac{1}{p}\|\boldsymbol{\theta}\|^2 = O_p(1)$, we have $|\frac{1}{n}\langle \boldsymbol{\theta}, \hat{g}_1(\boldsymbol{c}^1, \boldsymbol{\theta}, \boldsymbol{v})\rangle - \frac{1}{n}\langle \boldsymbol{\theta}, \hat{g}_1(\hat{\boldsymbol{c}}^1, \boldsymbol{\theta}, \boldsymbol{v})\rangle| \xrightarrow{\mathrm{p}} 0$. By (B.18), we have that $\frac{1}{n}\langle \boldsymbol{\theta}, \hat{g}_1(\hat{\boldsymbol{c}}^1, \boldsymbol{\theta}, \boldsymbol{v})\rangle \xrightarrow{\mathrm{p}} \gamma_1$. We have

$$\frac{1}{n} \|\hat{g}_1(\boldsymbol{c}^1, \boldsymbol{v}, \boldsymbol{\theta}) - \hat{g}_1(\hat{\boldsymbol{c}}^1, \boldsymbol{v}, \boldsymbol{\theta})\|_2^2 \leq \frac{1}{n} L^2 \|\boldsymbol{c}^1 - \hat{\boldsymbol{c}}^1\|_2^2 \xrightarrow{\mathrm{p}} 0,$$

wehre L is a Lipschitz constant for $\hat{g}_1$. By [12], the maximal singular value of $\boldsymbol{Z}^T\boldsymbol{Z}$ is $O_p(1)$. Therefore, $\frac{1}{n}\|\boldsymbol{Z}\hat{g}_1(\boldsymbol{c}^1, \boldsymbol{v}, \boldsymbol{\theta}) - \boldsymbol{Z}\hat{g}_1(\hat{\boldsymbol{c}}^1, \boldsymbol{v}, \boldsymbol{\theta})\|_2^2 \xrightarrow{\mathrm{p}} 0$. As a result, and using that $\frac{1}{n}\|\boldsymbol{\Lambda}\|_2^2$ converges almost surely to a constant,

$$\frac{1}{n}\|\hat{\boldsymbol{d}}^1 - \boldsymbol{d}^1\|_2^2 = \frac{1}{n}\|\boldsymbol{Z}\hat{g}_t(\hat{\boldsymbol{c}}^1, \boldsymbol{v}, \boldsymbol{\theta}) - \boldsymbol{Z}\hat{g}_t(\boldsymbol{c}^1, \boldsymbol{v}, \boldsymbol{\theta}) + (\frac{1}{n}\langle \boldsymbol{\theta}, \hat{g}_1(\boldsymbol{c}^1, \boldsymbol{\theta}, \boldsymbol{v})\rangle - \gamma_1)\boldsymbol{\Lambda}\|_2^2 \xrightarrow{\mathrm{p}} 0.$$

Now assume that (B.19) holds for $1, 2, \dots, t$. For the $(t+1)$-th iteration, we have

$$|\frac{1}{n}\langle \boldsymbol{\Lambda}, \hat{f}_t(\boldsymbol{d}^1, \dots, \boldsymbol{d}^t, \boldsymbol{u}, \boldsymbol{\Lambda})\rangle - \frac{1}{n}\langle \boldsymbol{\Lambda}, \hat{f}_t(\hat{\boldsymbol{d}}^1, \dots, \hat{\boldsymbol{d}}^t, \boldsymbol{u}, \boldsymbol{\Lambda})\rangle| \leq \frac{L}{n}\|\boldsymbol{\Lambda}\|_2 \sum_{s=1}^{t} \|\boldsymbol{d}^s - \hat{\boldsymbol{d}}^s\|_2 \xrightarrow{\mathrm{p}} 0.$$

where L is a Lipschitz constant for $\hat{f}$. By (B.18), we have $\frac{1}{n}\langle \boldsymbol{\Lambda}, \hat{f}_t(\hat{\boldsymbol{d}}^1, \dots, \hat{\boldsymbol{d}}^t, \boldsymbol{\Lambda}, \boldsymbol{u})\rangle \xrightarrow{\mathrm{p}} \alpha_{t+1}$. As a result, we have $\frac{1}{n}\langle \boldsymbol{\Lambda}, \hat{f}_t(\boldsymbol{d}^1, \dots, \boldsymbol{d}^t, \boldsymbol{u}, \boldsymbol{\Lambda})\rangle \xrightarrow{\mathrm{p}} \alpha_{t+1}$. Furthermore, for any $1 \leq s \leq t$, we have

$$\frac{1}{n}\|\hat{f}_s(\boldsymbol{d}^1, \dots, \boldsymbol{d}^s, \boldsymbol{u}, \boldsymbol{\Lambda}) - \hat{f}_s(\hat{\boldsymbol{d}}^1, \dots, \hat{\boldsymbol{d}}^s, \boldsymbol{u}, \boldsymbol{\Lambda})\|_2^2 \leq \frac{\hat{L}_t^2}{n} \sum_{i=1}^{s} \|\boldsymbol{d}^i - \hat{\boldsymbol{d}}^i\|_2^2 \xrightarrow{\mathrm{p}} 0,$$
$$\frac{1}{n}\|\hat{g}_s(\boldsymbol{c}^1, \dots, \boldsymbol{c}^s, \boldsymbol{v}, \boldsymbol{\theta}) - \hat{g}_s(\hat{\boldsymbol{c}}^1, \dots, \hat{\boldsymbol{c}}^s, \boldsymbol{v}, \boldsymbol{\theta})\|_2^2 \leq \frac{\hat{L}_t^2}{n} \sum_{i=1}^{s} \|\boldsymbol{c}^i - \hat{\boldsymbol{c}}^i\|_2^2 \xrightarrow{\mathrm{p}} 0.$$

Again using that the maximal singular value of $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$ is $O_p(1)$, we have

$$\frac{1}{n}\|\boldsymbol{Z}^\mathsf{T}\hat{f}_t(\hat{\boldsymbol{d}}^1,\ldots,\hat{\boldsymbol{d}}^t,\boldsymbol{u},\boldsymbol{\Lambda})-\boldsymbol{Z}^\mathsf{T}\hat{f}_t(\boldsymbol{d}^1,\ldots,\boldsymbol{d}^t,\boldsymbol{u},\boldsymbol{\Lambda})\|_2^2\xrightarrow{\mathrm{p}}0\,.$$

As a result, we have

$$\begin{aligned}&\frac{1}{n}\|\hat{\boldsymbol{c}}^{t+1}-\boldsymbol{c}^{t+1}\|_2^2\\ =\;&\frac{1}{n}\|(\frac{1}{n}\langle\boldsymbol{\Lambda},\hat{f}_t(\boldsymbol{d}^1,\ldots,\boldsymbol{d}^t,\boldsymbol{u},\boldsymbol{\Lambda})-\alpha_{t+1})\boldsymbol{\theta}+\boldsymbol{Z}^\mathsf{T}(\hat{f}_t(\hat{\boldsymbol{d}}^1,\ldots,\hat{\boldsymbol{d}}^t,\boldsymbol{u},\boldsymbol{\Lambda})-\hat{f}_t(\boldsymbol{d}^1,\ldots,\boldsymbol{d}^t,\boldsymbol{u},\boldsymbol{\Lambda}))-\\ &\sum_{s=1}^t\xi_{t,s}(\hat{g}_s(\hat{\boldsymbol{c}}^1,\ldots,\hat{\boldsymbol{c}}^s,\boldsymbol{v},\boldsymbol{\theta})-\hat{g}_s(\boldsymbol{c}^1,\ldots,\boldsymbol{c}^s,\boldsymbol{\theta},\boldsymbol{v}))\|_2^2\xrightarrow{\mathrm{p}}0.\end{aligned}$$

Similarly, we have

$$|\frac{1}{n}\langle\boldsymbol{\theta},\hat{g}_{t+1}(\hat{\boldsymbol{c}}^1,\ldots,\hat{\boldsymbol{c}}^{t+1},\boldsymbol{v},\boldsymbol{\theta})\rangle-\frac{1}{n}\langle\boldsymbol{\theta},\hat{g}_{t+1}(\boldsymbol{c}^1,\ldots,\boldsymbol{c}^{t+1},\boldsymbol{v},\boldsymbol{\theta})\rangle|\le\frac{L}{n}\|\boldsymbol{\theta}\|_2\sum_{s=1}^{t+1}\|\hat{\boldsymbol{c}}^{t+1}-\boldsymbol{c}^{t+1}\|_2\xrightarrow{\mathrm{p}}0,$$

where L is a Lipschitz constant for $\hat{g}_{t+1}$. By (B.18), we have that $\frac{1}{n}\langle\boldsymbol{\theta},\hat{g}_{t+1}(\hat{\boldsymbol{c}}^1,\ldots,\hat{\boldsymbol{c}}^{t+1},\boldsymbol{v},\boldsymbol{\theta})\rangle\xrightarrow{\mathrm{p}}\gamma_{t+1}$. As a result, we have that $\frac{1}{n}\langle\boldsymbol{\theta},\hat{g}_{t+1}(\boldsymbol{c}^1,\ldots,\boldsymbol{c}^{t+1},\boldsymbol{v},\boldsymbol{\theta})\rangle\xrightarrow{\mathrm{p}}\gamma_{t+1}$. Furthermore, for any $1\le s\le t$, we have

$$\frac{1}{n}\|\hat{f}_s(\boldsymbol{d}^1,\ldots,\boldsymbol{d}^s,\boldsymbol{u},\boldsymbol{\Lambda})-\hat{f}_s(\hat{\boldsymbol{d}}^1,\ldots,\hat{\boldsymbol{d}}^s,\boldsymbol{u},\boldsymbol{\Lambda})\|_2^2\le\frac{L^2}{n}\sum_{i=1}^s\|\boldsymbol{d}^i-\hat{\boldsymbol{d}}^i\|_2^2\xrightarrow{\mathrm{p}}0.$$

Also, for any $1\le s\le t+1$, we have

$$\frac{1}{n}\|\hat{g}_s(\boldsymbol{c}^1,\ldots,\boldsymbol{c}^s,\boldsymbol{v},\boldsymbol{\theta})-\hat{g}_s(\hat{\boldsymbol{c}}^1,\ldots,\hat{\boldsymbol{c}}^s,\boldsymbol{v},\boldsymbol{\theta})\|_2^2\le\frac{L^2}{n}\sum_{i=1}^s\|\boldsymbol{c}^i-\hat{\boldsymbol{c}}^i\|_2^2\xrightarrow{\mathrm{p}}0.$$

Then $\frac{1}{n}\|\boldsymbol{Z}\hat{g}_{t+1}(\hat{\boldsymbol{c}}^1,\ldots,\hat{\boldsymbol{c}}^{t+1},\boldsymbol{v},\boldsymbol{\theta})-\boldsymbol{Z}\hat{g}_{t+1}(\boldsymbol{c}^1,\ldots,\boldsymbol{c}^{t+1},\boldsymbol{v},\boldsymbol{\theta})\|_2^2\xrightarrow{\mathrm{p}}0$. As a result, we have

$$\begin{aligned}&\frac{1}{n}\|\hat{\boldsymbol{d}}^{t+1}-\boldsymbol{d}^{t+1}\|_2^2\\ &\quad=\frac{1}{n}\|(\frac{1}{n}\langle\boldsymbol{\theta},\hat{g}_{t+1}(\boldsymbol{c}^1,\ldots,\boldsymbol{c}^{t+1},\boldsymbol{v},\boldsymbol{\theta})\rangle-\gamma_{t+1})\boldsymbol{\Lambda}\\ &\quad+\boldsymbol{Z}(\hat{g}_{t+1}(\hat{\boldsymbol{c}}^1,\ldots,\hat{\boldsymbol{c}}^{t+1},\boldsymbol{v},\boldsymbol{\theta})-\hat{g}_{t+1}(\boldsymbol{c}^1,\ldots,\boldsymbol{c}^{t+1},\boldsymbol{v},\boldsymbol{\theta}))\\ &\quad-\sum_{s=0}^t\zeta_{t,s}(\hat{f}_s(\hat{\boldsymbol{d}}^1,\ldots,\hat{\boldsymbol{d}}^s,\boldsymbol{u},\boldsymbol{\Lambda})-\hat{f}_s(\boldsymbol{d}^1,\ldots,\boldsymbol{d}^s,\boldsymbol{u},\boldsymbol{\Lambda}))\|_2^2\xrightarrow{\mathrm{p}}0.\end{aligned}$$

Thus, we have proved (B.19). Therefore, for all pseudo-Lipschitz function ψ of order 2, we have that there exists a numerical constant C such that

$$\begin{aligned}&\left|\frac{1}{p}\sum_{j=1}^p\psi(c_j^1+\alpha_1\theta_j,\ldots,c_j^t+\alpha_t\theta_j,v_j,\theta_j)-\frac{1}{p}\sum_{j=1}^p\psi(\hat{c}_j^1+\alpha_1\theta_j,\ldots,\hat{c}_j^t+\alpha_t\theta_j,v_j,\theta_j)\right|\\ &\qquad\le L_\psi(1+\sum_{s=1}^t\|\boldsymbol{a}^s\|_2+\|\boldsymbol{\theta}\|_2+\|\boldsymbol{v}\|_2)\sum_{s=1}^t\|\hat{\boldsymbol{c}}^s-\boldsymbol{c}^s\|_2\xrightarrow{\mathrm{p}}0.\end{aligned}$$

By (B.18),

$$\frac{1}{p}\sum_{j=1}^{p}\psi(\hat{c}_j^1+\alpha_1\theta_j,\ldots,\hat{c}_j^t+\alpha_t\theta_j,v_j,\theta_j)\xrightarrow{\mathrm{p}}\mathbb{E}[\psi(\boldsymbol{\alpha}_1\boldsymbol{\Theta}+\boldsymbol{Z}^1,\ldots,\boldsymbol{\alpha}_t\boldsymbol{\Theta}+\boldsymbol{Z}^t,\boldsymbol{V},\boldsymbol{\Theta})].$$

Therefore, $\frac{1}{p}\sum_{j=1}^{p}\psi(a_j^1,\ldots,a_j^t,v_j,\theta_j)\xrightarrow{\mathrm{p}}\mathbb{E}[\psi(\boldsymbol{\alpha}_1\boldsymbol{\Theta}+\boldsymbol{Z}^1,\ldots,\boldsymbol{\alpha}_t\boldsymbol{\Theta}+\boldsymbol{Z}^t,\boldsymbol{V},\boldsymbol{\Theta})]$. Similarly, we can show that $\frac{1}{n}\sum_{i=1}^{n}\psi(\boldsymbol{b}_i^1,\ldots,\boldsymbol{b}_i^t,\boldsymbol{u}_i,\boldsymbol{\Lambda}_i)\xrightarrow{\mathrm{p}}\mathbb{E}[\psi(\boldsymbol{\gamma}_1\boldsymbol{\Lambda}+\tilde{\boldsymbol{Z}}^1,\ldots,\boldsymbol{\gamma}_t\boldsymbol{\Lambda}+\tilde{\boldsymbol{Z}}^t,\boldsymbol{U},\boldsymbol{\Lambda})]$. Thus we have finished the proof.

B.2.3 The AMP change of variables

To prove Lemma 3.5.1, all that remains is to show that for any GFOM (3.1), at least one of the change-of-variables in Eqs. (B.14) generates an iteration (B.15) which is an AMP iteration. That is, in addition to satisfying Eq. (B.14), the matrices $(\boldsymbol{\xi}_{t,s})$, $(\boldsymbol{\zeta}_{t,s})$ and functions (f_t), (g_t) satisfy Eqs. (B.16) and (B.17) in the high-dimensional regression and low-rank matrix estimation models respectively.

To construct such a choice of scalars, we may define $(\boldsymbol{\xi}_{t,s}),(\boldsymbol{\zeta}_{t,s}),(f_t),(g_t)$ in a single recursion by interlacing definition (B.14) with either (B.16) or (B.17). Specifically, in the high-dimensional regression model, we place (B.14a) before the first line of (B.16) and (B.14b) before the fourth line of (B.16). In the combined recursion, all quantities are defined in terms of previously defined quantities, yielding choices for $(\boldsymbol{\xi}_{t,s}),(\boldsymbol{\zeta}_{t,s}),(f_t),(g_t)$ which simultaneously satisfy (B.14) and (B.16). Thus, in the high-dimensional regression model every GFOM is equivalent, up to a change of variables, to a certain AMP algorithm. The construction in the low-rank matrix estimation model is analogous: we place (B.14a) before the first line of (B.17) and (B.14b) before the fourth line of (B.17).

The proof of Lemma 3.5.1 is complete.

B.3 Proof of state evolution for message passing (Lemma 3.5.2)

In this section, we prove Lemma 3.5.2. We restrict ourselves to the case $r=1$ and $k=1$ (with k the dimensionality of $\boldsymbol{W}$) because the proof for $r>1$ or $k>1$ is completely analogous but would complicate notation.

Let $\mathcal{T}_{v\to f}=(\mathcal{V}_{v\to f},\mathcal{F}_{v\to f},\mathcal{E}_{v\to f})$ be the tree consisting of edges and nodes in $\mathcal{T}$ which are separated from f by v. By convention, $\mathcal{T}_{v\to f}$ will also contain the node v. In particular, $f\notin\mathcal{F}_{v\to f}$ and $(f,v)\notin\mathcal{E}_{v\to f}$, but $v\in\mathcal{V}_{v\to f}$, and $f'\in\mathcal{F}_{v\to f}$ and $(v,f')\in\mathcal{E}_{v\to f}$ for $f'\in\partial v\setminus f$. We define $\mathcal{T}_{f\to v},\mathcal{V}_{f\to v},\mathcal{F}_{f\to v},\mathcal{E}_{f\to v}$ similarly. With some abuse of notation, we will sometimes use $\mathcal{T}_{f\to v},\mathcal{V}_{f\to v},\mathcal{F}_{f\to v},\mathcal{E}_{f\to v}$ to denote either the collection of observations corresponding to nodes and edges in these sets or the σ-algebra generated by these obervations. No confusion should result. Which random variables we consider to be "observed" will vary with the model, and will be explicitly described in each part of the proof to avoid potential ambiguity.

B.3.1 Gaussian message passing

We first introduce a message passing algorithm whose behavior is particularly easy to analyze. We call this message passing algorithm a *Gaussian message passing* algorithm. We will see that in both the high-dimensional regression and low-rank matrix estimation models, the message passing algorithm (3.19) approximates a certain Gaussian message passing algorithm.

Gaussian message passing algorithms operate on a computation tree with associated random variables $\{(\theta_v, v_v)\}_{v\in\mathcal{V}} \overset{\text{iid}}{\sim} \mu_{\Theta,V}$, $\{(w_f, u_f)\}_{f\in\mathcal{F}} \overset{\text{iid}}{\sim} \mu_{W,U}$, and $\{z_{fv}\}_{(f,v)\in\mathcal{E}} \overset{\text{iid}}{\sim} \mathsf{N}(0, 1/n)$, all independent, where $\mu_{\Theta,V}, \mu_{W,U} \in \mathscr{P}_4(\mathbb{R}^2)$.[2] Gaussian message passing algorithms access all these random variables, so that all are considered to be "observed." Thus, for example, $\mathcal{V}_{f\to v}$ contains $\theta_{v'}, v_{v'}$ for all nodes v' separated from f by v (including, by convention, v).

Gaussian message passing algorithms are defined by sequences of Lipschitz functions $(\tilde{f}_t : \mathbb{R}^{t+3} \to \mathbb{R})_{t\geq 0}$, $(\tilde{g}_t : \mathbb{R}^{t+2} \to \mathbb{R})_{t\geq 0}$. We initialize the indexing differently than with Gaussian message passing algorithms than with the message passing algorithms in Section 3.5 in anticipation of notational simplifications that will occur later. For every pair of neighboring nodes v, f, we generate sequences of messages $(\tilde{a}^t_{v\to f})_{t\geq 1}$, $(\tilde{q}^t_{v\to f})_{t\geq 0}$, $(\tilde{b}^t_{f\to v})_{t\geq 0}$, $(\tilde{r}^t_{f\to v})_{t\geq 0}$ according to the iteration

$$\tilde{a}^{t+1}_{v\to f} = \sum_{f'\in\partial v\setminus f} z_{f'v}\tilde{r}^t_{f'\to v}, \qquad \tilde{r}^t_{f\to v} = \tilde{f}_t(\tilde{b}^0_{f\to v}, \dots, \tilde{b}^t_{f\to v}; w_f, u_f), \tag{B.20a}$$

$$\tilde{b}^t_{f\to v} = \sum_{v'\in\partial f\setminus v} z_{fv'}\tilde{q}^t_{v'\to f}, \qquad \tilde{q}^t_{v\to f} = \tilde{g}_t(\tilde{a}^1_{v\to f}, \dots, \tilde{a}^t_{v\to f}; \theta_v, v_v), \tag{B.20b}$$

with initialization $\tilde{q}^0_{v\to f} = g_0(\theta_v, v_v)$. For $t \geq 0$, define the node beliefs

$$\tilde{a}^{t+1}_v = \sum_{f\in\partial v} z_{fv}\tilde{r}^t_{f\to v}, \qquad \tilde{b}^t_f = \sum_{v\in\partial f} z_{fv}\tilde{q}^t_{v\to f}. \tag{B.21}$$

To compactify notation, denote $\tilde{\boldsymbol{a}}^t_v = (\tilde{a}^1_v, \dots, \tilde{a}^t_v)^\mathsf{T}$, and likewise for $\tilde{\boldsymbol{a}}^t_{v\to f}$, $\tilde{\boldsymbol{q}}^t_{v\to f}$, $\tilde{\boldsymbol{b}}^t_f$, $\tilde{\boldsymbol{b}}^t_{f\to v}, \tilde{\boldsymbol{r}}^t_{f\to v}$ (where the first two of these are t-dimensional, and the last three are $(t+1)$-dimensional). We will often write $\tilde{f}_t(\tilde{\boldsymbol{b}}^t_{f\to v}; w_f, u_f)$ in place of $\tilde{f}_t(\tilde{b}^0_{f\to v}, \dots, b^t_{f\to v}; w_f, u_f)$, and similarly for $\tilde{g}_t$. The reader should not confuse the bold font here with that in Section 3.5, in which, for example, $\boldsymbol{a}^t_{v\to f}$ denotes the vectorial message at time t rather than the collection of scalar messages prior to and including time t.

Gaussian message passing obeys a Gaussian state evolution, defined by covariance matrices

$$\Sigma_{s,s'} = \mathbb{E}[\tilde{g}_s(\tilde{\boldsymbol{A}}^s; \Theta, V)\tilde{g}_{s'}(\tilde{\boldsymbol{A}}^{s'}; \Theta, V)], \quad T_{s+1,s'+1} = \mathbb{E}[\tilde{f}_s(\tilde{\boldsymbol{B}}^s; W, U)\tilde{f}_{s'}(\tilde{\boldsymbol{B}}^{s'}; W, U)], \tag{B.22}$$

where $s, s' \geq 0$, $\tilde{\boldsymbol{A}}^s \sim \mathsf{N}(\boldsymbol{0}_s, \boldsymbol{T}_{[1:s]})$, $\tilde{\boldsymbol{B}}^s \sim \mathsf{N}(\boldsymbol{0}_{s+1}, \boldsymbol{\Sigma}_{[0:s]})$, and $(\Theta, V) \sim \mu_{\Theta,V}$, $(W, U) \sim \mu_{W,U}$ independent of $\tilde{\boldsymbol{A}}^s, \tilde{\boldsymbol{B}}^s$. The iteration is initialized by $\Sigma_{0,0} = \mathbb{E}[\tilde{g}_0(\Theta, V)^2]$.

Lemma B.3.1. *If we choose a variable node v and factor node f independently of the randomness in our model, then for fixed t and for $n, p \to \infty$, $n/p \to \delta$ we have*

$$(\tilde{\boldsymbol{a}}^t_v, \theta_v, v_v) \xrightarrow{\text{W}} \mathsf{N}(\boldsymbol{0}_t, \boldsymbol{T}_{[1:t]}) \otimes \mu_{\Theta,V} \;\; \textit{and} \;\; (\tilde{\boldsymbol{a}}^t_{v\to f}, \theta_v, v_v) \xrightarrow{\text{W}} \mathsf{N}(\boldsymbol{0}_t, \boldsymbol{T}_{[1:t]}) \otimes \mu_{\Theta,V}, \tag{B.23a}$$

[2] We believe that only $\mu_{\Theta,V}, \mu_{W,U} \in \mathscr{P}_2(\mathbb{R}^2)$ is needed, but the analysis under this weaker assumption would be substantially more complicated, and the weaker assumptions are not necessary for our purposes.

$$(\tilde{\boldsymbol{b}}_f^t, w_f, u_f) \xrightarrow{\mathrm{W}} \mathsf{N}(\mathbf{0}_{t+1}, \boldsymbol{\Sigma}_{[0:t]}) \otimes \mu_{W,U} \quad \textit{and} \quad (\tilde{\boldsymbol{b}}_{f\to v}^t, w_f, u_f) \xrightarrow{\mathrm{W}} \mathsf{N}(\mathbf{0}_{t+1}, \boldsymbol{\Sigma}_{[0:t]}) \otimes \mu_{W,U}. \tag{B.23b}$$

Further, all the random variables in the preceding displays have bounded fourth moments and $\mathbb{E}[\|\tilde{\boldsymbol{a}}_v^t - \tilde{\boldsymbol{a}}_{v\to f}^t\|^2] \to 0$ *and* $\mathbb{E}[\|\tilde{\boldsymbol{b}}_f^t - \tilde{\boldsymbol{b}}_{f\to v}^t\|^2] \to 0$.

The analysis of message passing on the tree is facilitated by the many independence relationships between messages, which follow from the following lemma.

Lemma B.3.2. *For all* $(f, v) \in \mathcal{E}$ *and all* t, *the messages* $\tilde{r}_{f\to v}^t, \tilde{b}_{f\to v}^t$ *are* $\mathcal{T}_{f\to v}$*-measurable, and the messages* $\tilde{q}_{v\to f}^t, \tilde{a}_{a\to f}^t$ *is* $\mathcal{T}_{v\to f}$*-measurable.*

Proof.[Lemma B.3.2] The proof is by induction. The base case is that $\tilde{q}_{v\to f}^0 = g_0(\theta_v, v_v)$ is $\mathcal{T}_{v\to f}$-measurable. Then, if $\tilde{q}_{v\to f}^s$ are $\mathcal{T}_{v\to f}$-measurable and $\tilde{b}_{f\to v}^s$ are $\mathcal{T}_{f\to v}$-measurable for $0 \le s \le t$ and all $(f, v) \in \mathcal{E}$, then $\tilde{b}_{f\to v}^t, \tilde{r}_{f\to v}^t$ are $\mathcal{T}_{f\to v}$-measurable by (B.20). Similarly, if $\tilde{r}_{f\to v}^s$ are $\mathcal{T}_{f\to v}$-measurable and $\tilde{a}_{v\to f}^s$ are $\mathcal{T}_{v\to r}$-measurable for $0 \le s \le t$ and all $(f, v) \in \mathcal{E}$, then $\tilde{a}_{f\to v}^{t+1}, \tilde{r}_{f\to v}^{t+1}$ are $\mathcal{T}_{v\to f}$-measurable by (B.20). The induction is complete.

□

We now prove Lemma B.3.1.

Proof.[Lemma B.3.1] The proof is by induction.

Base case: $(\theta_v, v_f) \xrightarrow{\mathrm{W}} \mu_{\Theta,V}$.

This is the exact distribution in finite samples by assumption.

Inductive step 1: Eq. (B.23a) *at* t, *bounded fourth moments of* $\tilde{\boldsymbol{a}}_v^t, \tilde{\boldsymbol{a}}_{v\to f}^t$, *and* $\mathbb{E}[\|\tilde{\boldsymbol{a}}_v^t - \tilde{\boldsymbol{a}}_{v\to f}^t\|^2] \to 0$ *imply Eq.* (B.23b) *at* t, *bounded fourth moments of* $\tilde{\boldsymbol{b}}_f^t, \tilde{\boldsymbol{b}}_{f\to v}^t$, *and* $\mathbb{E}[\|\tilde{\boldsymbol{b}}_f^t - \tilde{\boldsymbol{b}}_{f\to v}^t\|^2] \to 0$.

The σ-algebras $(\mathcal{T}_{v\to f})_{v\in\partial f}$ are independent of $(z_{fv})_{v\in\partial f}$, which are mutually independent of each other. Thus, by (B.21), conditional on $\sigma((\mathcal{T}_{v\to f})_{v\in\partial f})$ the beliefs $\tilde{\boldsymbol{b}}_f^t$ are jointly normal with covariance $\widehat{\boldsymbol{\Sigma}}_{[0:t]} := \frac{1}{n}\sum_{v\in\partial f} \tilde{\boldsymbol{q}}_{v\to f}^t (\tilde{\boldsymbol{q}}_{v\to f}^t)^\mathsf{T}$. That is,

$$\tilde{\boldsymbol{b}}_f^t \;\big|\; \sigma((\mathcal{T}_{v\to f})_{v\in\partial f}) \sim \mathsf{N}(\mathbf{0}_{t+1}, \widehat{\boldsymbol{\Sigma}}_{[0:t]}).$$

Because $(\tilde{\boldsymbol{a}}_{v\to f}^t, \theta_v, v_v) \mapsto \tilde{g}_s(\tilde{\boldsymbol{a}}_{v\to f}^s; \theta_v, v_v)\tilde{g}_{s'}(\tilde{\boldsymbol{a}}_{v\to f}^{s'}; \theta_v, v_v)$ is uniformly pseudo-Lipschitz of order 2 by Lemma B.1.1, we have $\mathbb{E}[\widehat{\Sigma}_{s,s'}] = \mathbb{E}[\tilde{q}_{v\to f}^s \tilde{q}_{v\to f}^{s'}] = \mathbb{E}[\tilde{g}_s(\tilde{\boldsymbol{a}}_{v\to f}^s; \theta_v, v_v)\tilde{g}_{s'}(\tilde{\boldsymbol{a}}_{v\to f}^{s'}; \theta_v, v_v)] \to \Sigma_{s,s'}$ by the inductive hypothesis, Lemma B.1.2, and (B.22). The terms in the sum defining $\widehat{\boldsymbol{\Sigma}}_{[0:t]}$ are mutually independent by Lemma B.3.2 and have bounded second moments by the inductive hypothesis and the Lipschitz continuity of the functions $(\tilde{g}_s)_{0\le s\le t}$. By the weak law of large numbers, $\widehat{\boldsymbol{\Sigma}}_{[0:t]} \xrightarrow{L_1} \boldsymbol{\Sigma}_{[0:t]}$, whence by Slutsky's theorem, $\tilde{\boldsymbol{b}}_f^t \xrightarrow{\mathrm{d}} \mathsf{N}(\mathbf{0}_{t+1}, \boldsymbol{\Sigma}_{[0:t]})$. Further, $\mathbb{E}[\tilde{\boldsymbol{b}}_f^t(\tilde{\boldsymbol{b}}_f^t)^\mathsf{T}] = \mathbb{E}[\widehat{\boldsymbol{\Sigma}}_{[0:t]}] \to \boldsymbol{\Sigma}_{[0:t]}$. Convergence in distribution and in second moment implies convergence in the Wasserstein space of order 2 [193, Theorem 6.9], so $\tilde{\boldsymbol{b}}_f^t \xrightarrow{\mathrm{W}} \mathsf{N}(\mathbf{0}_{t+1}, \boldsymbol{\Sigma}_{[0:t]})$.

To bound the fourth moments of $\tilde{b}_f^t$, we compute

$$\mathbb{E}[(\tilde{b}_f^t)^4] = \mathbb{E}[\widehat{\Sigma}_{t,t}^2] = \frac{1}{n^2}\sum_{v\in\partial f}\mathbb{E}[(\tilde{q}_{v\to f}^t)^4] + \frac{1}{n^2}\sum_{v\neq v'\in\partial f}\mathbb{E}[(\tilde{q}_{v\to f}^t)^2]\mathbb{E}[(\tilde{q}_{v'\to f}^t)^2] \to \Sigma_{t,t},$$

where the first term goes to 0 because the fourth moments of $\tilde{q}_{v\to f}^t$ are bounded by the inductive hypothesis and Lipschitz continuity of $\tilde{g}_t$, and the second term goes to $\mathbb{E}[(\tilde{q}_{v\to f}^t)^2]$ by the same argument in the preceding paragraph. The boundedness of the fourth moments of $\tilde{b}_f^s$ holds similarly (and, anyway, will have been established earlier in the induction).

Finally, observe $\tilde{b}_f^t - \tilde{b}_{f\to v}^t = z_{fv}\tilde{q}_{v\to f}^t$ and $\mathbb{E}[(z_{fv}\tilde{q}_{v\to f}^t)^2] = \mathbb{E}[\tilde{q}_{v\to f}^t)^2]/n \to 0$, where $\mathbb{E}[\tilde{q}_{v\to f}^t)^2]$ is bounded by the inductive hypothesis and Lipschitz continuity of $\tilde{g}_t$. The convergence $\mathbb{E}[(\tilde{b}_f^t - \tilde{b}_{f\to v}^s)^2] \to 0$ for $s < t$ holds similarly (and, anyway, will have been established earlier in the induction). The Wasserstein convergence of $(\tilde{\boldsymbol{b}}_{v\to f}^t, \theta_v, v_v)$ now follows. The bounded fourth moments of $\tilde{\boldsymbol{b}}_{v\to f}^t$ hold similarly.

Inductive step 2: Eq. (B.23) *at t, bounded fourth moments of $\tilde{\boldsymbol{b}}_f^t, \tilde{\boldsymbol{b}}_{f\to v}^t$, and $\mathbb{E}[\|\tilde{\boldsymbol{b}}_f^t - \tilde{\boldsymbol{b}}_{f\to v}^t\|^2] \to 0$ imply Eq.* (B.23) *at $t+1$, bounded fourth moments of $\tilde{\boldsymbol{a}}_v^t, \tilde{\boldsymbol{a}}_{v\to f}^{t+1}$, and $\mathbb{E}[\|\tilde{\boldsymbol{a}}_v^{t+1} - \tilde{\boldsymbol{a}}_{v\to f}^{t+1}\|^2] \to 0$.*

This follows by exactly the same argument as in inductive step 1.

The induction is complete, and Lemma B.3.1 follows.

□

B.3.2 Message passing in the high-dimensional regression model

We prove Lemma 3.5.2 for the high-dimensional regression model by showing that the iteration (3.19) is well approximated by a Gaussian message passing algorithm after a change of variables. The functions $\tilde{f}_t, \tilde{g}_t$ in the Gaussian message passing algorithm are defined in terms of the functions f_t, g_t of the original message passing algorithm (3.19) and the function h used to define the high-dimensional regression model.

$$\tilde{f}_t(\tilde{b}^0, \cdots, \tilde{b}^t, w, u) := f_t(\tilde{b}^1, \cdots, \tilde{b}^t; h(\tilde{b}^0, w), u), \quad t \geq 0,$$
$$\tilde{g}_0(\theta, v) = \theta, \quad \tilde{g}_t(\tilde{a}^1, \cdots, \tilde{a}^t; \theta, v) := g_t(\alpha_1\theta + \tilde{a}^1, \cdots, \alpha_1\theta + \tilde{a}^t; v), \quad t \geq 1.$$

Define $(\tilde{a}_{v\to f}^t)_{t\geq 1}$, $(\tilde{a}_v^t)_{t\geq 1}$, $(\tilde{q}_{v\to f}^t)_{t\geq 0}$, $(\tilde{b}_{f\to v}^t)_{t\geq 0}$, $(\tilde{b}_f^t)_{t\geq 0}$, $(\tilde{r}_{f\to v}^t)_{t\geq 0}$ via the Gaussian message passing algorithm (B.20) with initial data θ_v, v_v, w_f, u_f and with $z_{fv} = x_{fv}$. Because f_t, g_t, and h are Lipschitz, so too are $\tilde{f}_t$ and $\tilde{g}_t$. Under the function definitions $\tilde{f}_t, \tilde{g}_t$ given above, the definitions of $\Sigma_{s,s}$ and $T_{s,s'}$ in (B.22) and (B.16) are equivalent. Thus, Lemma B.3.1 holds for the iterates of this Gaussian message passing algorithm with the $\boldsymbol{T}_{[1:t]}$, $\boldsymbol{\Sigma}_{[0:t]}$ defined by (B.16).

We claim that for fixed $s \geq 1$, as $n \to \infty$ we have

$$\mathbb{E}[(\alpha_s\theta_v + \tilde{a}_{v\to f}^s - a_{v\to f}^s)^2] \to 0 \text{ and } \mathbb{E}[(\tilde{b}_{f\to v}^s - b_{f\to v}^s)^2] \to 0, \tag{B.24a}$$

and

$$\mathbb{E}[(a_{v\to f}^s)^4] \text{ and } \mathbb{E}[(b_{f\to v}^s)^4] \text{ are uniformly bounded with respect to } n, \tag{B.24b}$$

where (α_s) are defined by (B.16). These are the same coefficients appearing in the AMP state evolution (Lemma 3.5.1), as claimed. We show (B.24) by induction. There is no base case because the inductive steps work for $t = 0$ as written.

Inductive step 1: If (B.24) *holds for $1 \leq s \leq t$, then* (B.24a) *holds for $s = t+1$.*

We expand

$$\begin{aligned}\alpha_{t+1}\theta_v + \tilde{a}^{t+1}_{v\to f} - a^{t+1}_{v\to f} &= \alpha_{t+1}\theta_v + \sum_{f'\in\partial v\setminus f} z_{f'v}(\tilde{f}_t(\tilde{\boldsymbol{b}}^t_{f'\to v}; w_{f'}, u_{f'}) - f_t(\boldsymbol{b}^t_{f'\to v}; y_{f'}, u_{f'}))\\ &= \alpha_{t+1}\theta_v + \sum_{f'\in\partial v\setminus f} z_{f'v}(\tilde{f}_t(\tilde{\boldsymbol{b}}^t_{f'\to v}; w_{f'}, u_{f'}) - \tilde{f}_t(\tilde{b}^0_{f'\to v}, \boldsymbol{b}^t_{f'\to v}; w_{f'}, u_{f'}))\\ &\quad + \sum_{f'\in\partial v\setminus f} z_{f'v}(\tilde{f}_t(\tilde{b}^0_{f'\to v}, \boldsymbol{b}^t_{f'\to v}; w_{f'}, u_{f'}) - \tilde{f}_t(\tilde{b}^0_{f'}, \boldsymbol{b}^t_{f'\to v}; w_{f'}, u_{f'}))\\ &=: \alpha_{t+1}\theta_v + \mathsf{I} + \mathsf{II}.\end{aligned}$$

(Note that $\tilde{\boldsymbol{b}}^t_{f'\to v}$ is $(t+1)$-dimensional and $\boldsymbol{b}^t_{f'\to v}$ is t-dimensional). First we analyze I. We have

$$|\tilde{f}_t(\tilde{\boldsymbol{b}}^t_{f'\to v}; w_{f'}, u_{f'}) - \tilde{f}_t(\tilde{b}^0_{f'\to v}, \boldsymbol{b}^t_{f'\to v}; w_{f'}, u_{f'})| \le L\sum_{s=1}^t |\tilde{b}^s_{f'\to v} - b^s_{f'\to v}|,$$

where L is a Lipschitz constant of $\tilde{f}_t$. The terms in the sum defining I are mutually independent, and $\tilde{b}^s_{f'\to v}, b^s_{f'\to v}$ are independent of $z_{f'v}$. Thus,

$$\begin{aligned}\mathbb{E}[\mathsf{I}^2] &= \frac{n-1}{n}\mathbb{E}[(\tilde{f}_t(\tilde{\boldsymbol{b}}^t_{f'\to v}; w_{f'}, u_{f'}) - \tilde{f}_t(\tilde{b}^0_{f'\to v}, \boldsymbol{b}^t_{f'\to v}; w_{f'}, u_{f'}))^2]\\ &\le \frac{L^2(n-1)t}{n}\sum_{s=1}^t \mathbb{E}[(\tilde{b}^s_{f'\to v} - b^s_{f'\to v})^2] \to 0,\end{aligned}$$

by the inductive hypothesis.

Next we analyze II. Note that all arguments to the functions in the sum defining II are independent of $z_{f'v}$ and θ_v except for $\tilde{b}^0_{f'} = z_{f'v}\theta_v + \sum_{v'\in\partial f'\setminus v} z_{f'v'}\theta_{v'}$. Because $\tilde{f}_t$ is Lipschitz, we may apply Stein's lemma (ie., Gaussian integration by parts) [184] to get

$$\begin{aligned}&\mathbb{E}[\alpha_{t+1}\theta_v + \mathsf{II} \mid \theta_v, \sigma((\mathcal{T}_{v''\to f'})_{v''\in\partial f'\setminus v})]\\ &= \alpha_{t+1}\theta_v + (n-1)\mathbb{E}\big[z_{f'v}(\tilde{f}_t(\tilde{b}^0_{f'\to v}, \boldsymbol{b}^t_{f'\to v}; w_{f'}, u_{f'}) - \tilde{f}_t(\tilde{b}^0_{f'}, \boldsymbol{b}^t_{f'\to v}; w_{f'}, u_{f'})) \mid \theta_v\big]\\ &= \theta_v\left(\alpha_{t+1} - \frac{n-1}{n}\mathbb{E}[\partial_{\tilde{b}^0}\tilde{f}_t(\tilde{b}^0_{f'}, \boldsymbol{b}^t_{f'\to v}; w_{f'}, u_{f'}) \mid \theta_v]\right),\end{aligned}$$

where $\partial_{\tilde{b}^0}\tilde{f}_t$ is the weak-derivative of $\tilde{f}_t$ with respect to its first argument, which is defined almost everywhere with respect to Lebesgue measure because $\tilde{f}_t$ is Lipschitz [83, pg. 81].

We claim the right-hand side of the preceding display converges in L_2 to 0, as we now show. The random variable $\mathbb{E}[\partial_{\tilde{b}^0}\tilde{f}_t(\tilde{b}^0_{f'}, \boldsymbol{b}^t_{f'\to v}; w_{f'}, u_{f'})|\theta_v, (\mathcal{T}_{v''\to f'})_{v''\in\partial f'\setminus v}]$ is almost-surely bounded because $\tilde{f}_t$ is Lipschitz. It converges in probability to α_{t+1}. The random vector $(\tilde{b}^0_{f'}, \boldsymbol{b}^t_{f'\to v})$ has a Gaussian distribution conditional on $\sigma((\mathcal{T}_{v''\to f'})_{v''\in\partial f'\setminus v})$ and θ_v; in particular,

$$(\tilde{b}^0_{f'\to v} + z_{f'v}\theta_v, \boldsymbol{b}^t_{f'\to v})|\theta_v, \sigma((\mathcal{T}_{v''\to f'})_{v''\in\partial f'\setminus v}) \overset{\mathrm{d}}{=} \mathsf{N}(\boldsymbol{0}, \widehat{\boldsymbol{\Sigma}}),$$

where we define $\widehat{\boldsymbol{\Sigma}} \in \mathbb{R}^{(t+1)\times(t+1)}$ by

$$\widehat{\Sigma}_{0,0} = \frac{1}{n}\sum_{v'\in\partial f'} \theta_{v'}^2 \quad\text{and}\quad \widehat{\Sigma}_{s,s'} = \frac{1}{n}\sum_{v'\in\partial f'\setminus v} q_{v'\to f'}^s q_{v'\to f'}^{s'} \quad \text{for } s\geq 1 \text{ or } s'\geq 1,$$

where for the purposes of the preceding display we set $q_{v'\to f'}^0 = \theta_{v'}$. By the Lipschitz continuity of the functions (g_s), Lemmas B.1.1 and B.1.2, and the inductive hypothesis, we have $\mathbb{E}[\widehat{\boldsymbol{\Sigma}}] \to \boldsymbol{\Sigma}_{[0:t]}$. The terms in the sums in the previous display have bounded second moments by the inductive hypthesis (B.24b) and the Lipschitz continuity of the functions (g_s). By the weak law of large numbers, we conclude $\widehat{\boldsymbol{\Sigma}} \xrightarrow{\mathrm{p}} \boldsymbol{\Sigma}_{[0:t+1]}$.

Observe that $\mathbb{E}[\partial_{\tilde{b}^0} \tilde{f}_t(\tilde{b}_{f'}^0, \boldsymbol{b}_{f'\to v}^t; w_{f'}, u_{f'})|\theta_v, (\mathcal{T}_{v''\to f'})_{v''\in\partial f'\setminus v}] = \mathbb{E}[\partial_{\tilde{b}^0}\tilde{f}_t(\widehat{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{Z}; W, U)]$, where on the right-hand side the expectation is with respect to $(W,U)\sim\mu_{W,U}$ and $\boldsymbol{Z}\sim\mathsf{N}(\boldsymbol{0}_{t+1}, \boldsymbol{I}_{t+1})$ independent. Because $\partial_{\tilde{b}^0}\tilde{f}_t$ is almost surely bounded, by the dominated convergence theorem, the right-hand side is continuous in $\widehat{\boldsymbol{\Sigma}}$. By the continuous mapping theorem and (B.16), we conclude $\mathbb{E}[\partial_{\tilde{b}^0} \tilde{f}_t(\tilde{b}_{f'}^0, \boldsymbol{b}_{f'\to v}^t; w_{f'}, u_{f'})|\theta_v, (\mathcal{T}_{v''\to f'})_{v''\in\partial f'\setminus v}] \xrightarrow{\mathrm{p}} \alpha_{t+1}$. Then, by dominated convergence, $\mathbb{E}[\alpha_{t+1}\theta_v + \mathsf{II} \mid \theta_v] \xrightarrow{L_2} 0$. Moreover, because the terms in the sum defining II are mutually independent given θ_v

$$\begin{aligned}\mathrm{Var}(\alpha_{t+1}\theta_v + \mathsf{II} \mid \theta_v) &\leq (n-1)\mathbb{E}\left[z_{f'v}^2(\tilde{f}_t(\tilde{b}_{f'\to v}^0, \boldsymbol{b}_{f'\to v}^t; w_{f'}, u_{f'}) - \tilde{f}_t(\tilde{b}_{f'}^0, \boldsymbol{b}_{f'\to v}^t; w_{f'}, u_{f'}))^2 \mid \theta_v\right]\\ &\leq L^2(n-1)\mathbb{E}[z_{f'v}^4\theta_v^2 \mid \theta_v] \leq 3\theta_v^2/n,\end{aligned}$$

where L is the Lipschitz constant of $\tilde{f}_t$. We conclude that $\mathbb{E}[\mathrm{Var}(\alpha_{t+1}\theta_v + \mathsf{II} \mid \theta_v)] \to 0$. Combined with $\mathbb{E}[\alpha_{t+1}\theta_v + \mathsf{II} \mid \theta_v] \xrightarrow{L_2} 0$, we get $\mathrm{Var}(\alpha_{t+1}\theta_v + \mathsf{II}) = \mathrm{Var}(\mathbb{E}[\alpha_{t+1}\theta_v + \mathsf{II}|\theta_v]) + \mathbb{E}[\mathrm{Var}(\alpha_{t+1}\theta_v + \mathsf{II}|\theta_v)] \to 0$, so that $\alpha_{t+1}\theta_v + \mathsf{II} \xrightarrow{L_2} 0$. Combining $\mathsf{I} \xrightarrow{L_2} 0$ and $\alpha_{t+1}\theta_v + \mathsf{II} \xrightarrow{L_2} 0$ gives $\mathbb{E}[(\alpha_{t+1}\theta_v + \tilde{a}_{v\to f}^{t+1} - a_{v\to f}^{t+1})^2] \to 0$, as desired.

We now expand

$$\tilde{b}_{f\to v}^{t+1} - b_{f\to v}^{t+1} = \sum_{v'\in\partial f\setminus v} z_{fv'}(g_t(\boldsymbol{\alpha}_{t+1}\theta_{v'} + \tilde{\boldsymbol{a}}_{v'\to f}^{t+1}; v_{v'}) - g_t(\boldsymbol{a}_{v'\to f}^{t+1}; v_{v'})).$$

The terms in this sum are mutually independent, and $\tilde{\boldsymbol{a}}_{v'\to f}^{t+1}, \boldsymbol{a}_{v'\to f}^{t+1}, \theta_{v'}$ are independent of $z_{f'v}$. Thus,

$$\begin{aligned}\mathbb{E}[(\tilde{b}_{f\to v}^{t+1} - b_{f\to v}^{t+1})^2] &= \frac{p-1}{n}\mathbb{E}[(g_t(\boldsymbol{\alpha}_{t+1}\theta_{v'} + \tilde{\boldsymbol{a}}_{v'\to f}^{t+1}; v_{v'}) - g_t(\boldsymbol{a}_{v'\to f}^{t+1}; v_{v'}))^2]\\ &\leq \frac{L^2(p-1)(t+1)}{n}\sum_{s=1}^{t+1}\mathbb{E}[(\alpha_s\theta_{v'} + \tilde{a}_{v'\to f}^s - a_{v'\to f}^s)^2] \to 0.\end{aligned}$$

This completes the proof of (B.24a) at $s = t+1$.

Inductive step 2: If (B.24) *holds for* $1 \leq s \leq t$*, then* (B.24b) *holds for* $s = t+1$*.*

By Lipschitz continuity,

$$\left| a_{v\to f}^{t+1} - \sum_{f'\in\partial v\setminus f} z_{f'v}\tilde{f}_t(\tilde{b}_{f'\to v}^0, \boldsymbol{b}_{f'\to v}^t, u_{f'}, w_{f'})\right| \leq L|\theta_v|\sum_{f'\in\partial v\setminus f}|z_{f'v}|,$$

where L is a Lipschitz constant for $\tilde{f}_t$. The right-hand side has bounded fourth moment, so we must only show that the sum in the previous display has bounded fourth moment. The quantity $\tilde{f}_t(\tilde{b}_{f'\to v}^0, \boldsymbol{b}_{f'\to v}^t, u_{f'}, w_{f'})$

has bounded fourth moment by the inductive hypothesis and Lipschitz continuity of $\tilde{f}_t$. Because $z_{f'v}$ is independent of the argument to $\tilde{f}_t$ and has fourth moment $3/n^2$, the product $z_{f'v}\tilde{f}_t(\tilde{b}^0_{f'\to v}, \boldsymbol{b}^t_{f'\to v}, u_{f'}, w_{f'})$ has mean 0 and fourth moment $O(1/n^2)$. Because these products are mean zero and independent across f', their sum has bounded fourth moment. We conclude $a^{t+1}_{v\to f}$ has bounded fourth moment as well.

Recall $b^{t+1}_{f\to v} = \sum_{v'\in\partial f\setminus v} z_{fv'} g_t(\boldsymbol{a}^{t+1}_{v'\to f}; v'_v)$. The terms in the sum are independent, and $z_{fv'}$ is independent of $\boldsymbol{a}^{t+1}_{v'\to f}; v'_v$. Using the Lipschitz continuity of g_t and the inductive hypothesis, we conclude $b^{t+1}_{f\to v}$ has bounded fourth moment by the same argument as in the preceding paragraph.

We conclude (B.24b) at $s = t+1$.

The induction is complete, and (B.24a) holds for all $s \geq 1$. Lemma 3.5.2 follows by combining Lemma B.3.1 and Eq. (B.24a).

B.3.3 Message passing in the low-rank matrix estimation model

Like in the preceding section, we prove Lemma 3.5.2 for the low-rank matrix estimation model by showing that the iteration (3.19) is well approximated by a Gaussian message passing algorithm after a change of variables. The functions in the Gaussian message passing algorithm are defined in terms of the functions f_t, g_t of the original message passing algorithm (3.19).

$$\tilde{f}_t(\tilde{b}^0, \cdots, \tilde{b}^t, w, u) := f_t(\tilde{b}^1 + \gamma_1 w, \cdots, \tilde{b}^t + \gamma_t w; 0, u),$$
$$\tilde{g}_t(\tilde{a}^1, \cdots, \tilde{a}^t; \theta, v) := g_t(\tilde{a}^1 + \alpha_1\theta, \cdots, \tilde{a}^t + \alpha_t\theta; v).$$

Note that here $\tilde{f}_t$ does not depend on $\tilde{b}^0$ is never used, and we may define $\tilde{g}_0$ arbitrarily without affecting later iterates.[3] Define $(\tilde{a}^t_{v\to f})_{t\geq 1}$, $(\tilde{a}^t_v)_{t\geq 1}$, $(\tilde{q}^t_{v\to f})_{t\geq 0}$, $(\tilde{b}^t_{f\to v})_{t\geq 0}$, $(\tilde{b}^t_f)_{t\geq 0}$, $(\tilde{r}^t_{f\to v})_{t\geq 0}$ via the Gaussian message passing algorithm (B.20) with initial data $\theta_v, v_v, u_f, z_{fv}$ and $w_f = \lambda_f$. Because f_t, g_t, and h are Lipschitz, so too are $\tilde{f}_t$ and $\tilde{g}_t$. Under the function definitions $\tilde{f}_t, \tilde{g}_t$ given above and the change of variables $w_f = \lambda_f$, the definitions of $\Sigma_{s,s}$ and $T_{s,s'}$ in (B.22) and (B.17) are equivalent. Thus, Lemma B.3.1 holds for the iterates of this Gaussian message passing algorithm with the $\boldsymbol{T}_{[1:t]}$, $\boldsymbol{\Sigma}_{[0:t]}$ defined by (B.17).

We claim that for fixed $s \geq 1$, as $n \to \infty$ we have

$$\mathbb{E}[(\alpha_s\theta_v + \tilde{a}^s_{v\to f} - a^s_{v\to f})^2] \to 0 \;\text{ and }\; \mathbb{E}[(\gamma_s\lambda_f + \tilde{b}^s_{f\to v} - b^s_{f\to v})^2] \to 0, \tag{B.25a}$$

and

$$\mathbb{E}[\theta_v^2(a^s_{v\to f})^2] \;\text{ and }\; \mathbb{E}[\lambda_f^2(b^s_{f\to v})^2] \;\text{ are bounded for fixed } s. \tag{B.25b}$$

We show this by induction. There is no base case because the inductive step works for $t = 0$ as written.

Inductive step: If (B.25) *holds for* $1 \leq s \leq t$, *then* (B.25) *holds for* $s = t+1$.

We expand

$$\alpha_{t+1}\theta_v + \tilde{a}^{t+1}_{v\to f} - a^{t+1}_{v\to f} = \alpha_{t+1}\theta_v + \sum_{f'\in\partial v\setminus f} z_{f'v}(f_t(\tilde{\boldsymbol{b}}^t_{f'\to v} + \boldsymbol{\gamma}_t\lambda_{f'}; 0, u_{f'}) - f_t(\boldsymbol{b}^t_{f'\to v}; 0, u_{f'}))$$

[3] The iterate $\tilde{b}^0$ only played a role in approximating the high-dimensional regression message passing algorithm by a Gaussian message passing algorithm.

$$
\begin{aligned}
&- \frac{1}{n}\theta_v \sum_{f' \in \partial v \setminus f} \lambda_{f'} f_t(\boldsymbol{b}^t_{f' \to v}; 0, u_{f'}) \\
&=: \alpha_{t+1}\theta_v + \mathsf{I} + \mathsf{II},
\end{aligned}
$$

where $\tilde{\boldsymbol{b}}^t_{f' \to v} = (\tilde{b}^1_{f' \to v}, \ldots, \tilde{b}^t_{f' \to v})$ and $\boldsymbol{\gamma}_t = (\gamma_1, \ldots, \gamma_t)$ (note that $\tilde{b}^0_{f' \to v}$ is excluded, which differs from the notation used in the proof of Lemma 3.5.2).

First we analyze I. The terms in the sum defining I are mutually independent, and $\tilde{b}^s_{f' \to v}$, $b^s_{f' \to v}$, $\lambda_{f'}$, $u_{f'}$ are independent of $z_{f'v}$. Thus,

$$
\begin{aligned}
\mathbb{E}[\mathsf{I}^2] &= \frac{n-1}{n}\mathbb{E}[(f_t(\tilde{\boldsymbol{b}}^t_{f' \to v} + \boldsymbol{\gamma}_t \lambda_{f'}; 0, u_{f'}) - f_t(\boldsymbol{b}^t_{f' \to v}; 0, u_{f'})^2] \\
&\leq \frac{L^2(n-1)t}{n} \sum_{s=1}^t \mathbb{E}[(\tilde{b}^s_{f' \to v} + \gamma_s \lambda_{f'} - b^s_{f' \to v})^2] \to 0,
\end{aligned}
$$

by the inductive hypothesis, where L is a Lipschitz constant of f_t. Moreover, because θ_v is independent of I and has bounded fourth moment, $\mathbb{E}[\theta_v^2 \mathsf{I}^2] \to 0$ as well.

Next we analyze II. By the inductive hypothesis and Lemma B.3.1,

$$
(\boldsymbol{b}^t_{f' \to v}, \lambda_{f'}, u_{f'}) \xrightarrow{\mathrm{W}} (\boldsymbol{\gamma}_t \Lambda + \tilde{B}^t, \Lambda, U),
$$

where $(\Lambda, U) \sim \mu_{\Lambda,U}$ and $\tilde{B}^t \sim \mathsf{N}(\mathbf{0}_t, \boldsymbol{\Sigma}_{[1:t]})$ independent. Because $(\boldsymbol{b}^t, \lambda, u) \mapsto \lambda f_t(\boldsymbol{b}^t; 0, u)$ is uniformly pseudo-Lipschitz of order 2 by Lemma B.1.1, we have $\mathbb{E}[\lambda_{f'} f_t(\boldsymbol{b}^t_{f' \to v}; 0, u_{f'})] \to \alpha_{t+1}$ by Lemma B.1.2 and the state evolution recursion (B.17). Moreover, because f_t is Lipschitz, for some constant C

$$
\begin{aligned}
\mathbb{E}[\lambda_{f'}^2 f_t(\boldsymbol{b}^t_{f' \to v}; 0, u_{f'})^2] &\leq C\mathbb{E}\left[\lambda_{f'}^2 \left(1 + \sum_{s=1}^t (b^s_{f' \to v})^2 + u_{f'}^2\right)\right] \\
&= C\left(\mathbb{E}[\lambda_{f'}^2] + \sum_{s=1}^t \mathbb{E}[\lambda_{f'}^2 (b^s_{f' \to v})^2] + \mathbb{E}[\lambda_{f'}^2 u_{f'}^2]\right),
\end{aligned}
$$

which bounded by the inductive hypothesis and the fourth moment assumption on $\mu_{\Lambda,U}$. Because the terms in the sum defining II are mutually independent, by the weak law of large numbers the preceding observations imply

$$
\frac{1}{n} \sum_{f' \in \partial v \setminus f} \lambda_{f'} f_t(\boldsymbol{b}^t_{f' \to v}; 0, u_{f'}) \xrightarrow{L_2} \alpha_{t+1}.
$$

Because θ_v is independent of this sum and has bounded second moment, we conclude that

$$
\alpha_{t+1}\theta_v + \mathsf{II} = \theta_v \left(\alpha_{t+1} - \frac{1}{n} \sum_{f' \in \partial v \setminus f} \lambda_{f'} f_t(\boldsymbol{b}^t_{f' \to v}; 0, u_{f'})\right) \xrightarrow{L_2} 0.
$$

Moreover, because θ_v is independent of the term in parentheses and has bounded fourth moment, $\mathbb{E}[\theta_v^2(\alpha_{t+1}\theta_v + \mathsf{II})^2] \to 0$.

Combining the preceding results, we have that $\mathbb{E}[(\alpha_{t+1}\theta_v + \tilde{a}^{t+1}_{v \to f} - a^{t+1}_{v \to f})^2] \to 0$ and $\mathbb{E}[\theta_v^2(\alpha_{t+1}\theta_v + \tilde{a}^{t+1}_{v \to f} - a^{t+1}_{v \to f})^2]$ is bounded. Because θ_v is independent of $\tilde{a}^{t+1}_{v \to f}$, the term $\mathbb{E}[\theta_v^2(\tilde{a}^{t+1}_{v \to f})^2]$ is bounded, so also $\mathbb{E}[\theta_v^2(a^{t+1}_{v \to f})^2]$ is bounded, as desired.

The argument establishing that $\mathbb{E}[(\gamma_{t+1}\lambda_f + \tilde{b}^{t+1}_{f\to v} - b^{t+1}_{f\to v})^2] \to 0$ and that $\mathbb{E}[\lambda_f^2(b^{t+1}_{f\to v})^2]$ is bounded is equivalent. The induction is complete, and (B.25) holds for all s.

Lemma 3.5.2 follows by combining Lemma B.3.1 and Eq. (B.25).

B.4 Proof of information-theoretic lower bounds on the computation tree (Lemma 3.5.3)

In this section, we prove Lemma 3.5.3 in both the high-dimensional regression and low-rank matrix estimation models. We restrict ourselves to the case $r = 1$ and $k = 1$ (with k the dimensionality of $\boldsymbol{W}$) because the proof for $r > 1$ or $k > 1$ is completely analogous but would complicate notation.

For any pair of nodes u, u' in the tree $\mathcal{T}$, let $d(u, u')$ denote the length (number of edges) of the shortest path between nodes u and u' in the tree. Let $\mathcal{T}_{u,k} = (\mathcal{V}_{u,k}, \mathcal{F}_{u,k}, \mathcal{E}_{u,k})$ be the radius-k neighborhood of node u; that is,

$$\begin{aligned}\mathcal{V}_{u,k} &= \{v \in \mathcal{V} \mid d(u, v) \le k\},\\ \mathcal{F}_{u,k} &= \{f \in \mathcal{F} \mid d(u, f) \le k\},\\ \mathcal{E}_{u,k} &= \{(f, v) \in \mathcal{E} \mid \max\{d(u, f), d(u, v)\} \le k\}.\end{aligned}$$

With some abuse of notation, we will often use $\mathcal{T}_{u,k}, \mathcal{V}_{u,k}, \mathcal{F}_{u,k}, \mathcal{E}_{u,k}$ to denote either the collection of observations corresponding to nodes and edges in these sets or the σ-algebra generated by these obervations. No confusion should result. Note, our convention is that when used to denote a σ-algebra or collection of random variables, only observed random variables are in include. Thus, in the high-dimensional regression model, $\mathcal{T}_{u,k}$ is the σ-algebra generated by the local observations x_{fv}, y_f, v_v, and u_f; in the low-rank matrix estimation, it is the σ-algebra genreated by the local observations x_{fv}, v_v, and u_f. We also denote by $\mathcal{T}^{t,k}_{v\to f}$ the collection of observations associated to edges or nodes of $\mathcal{T}$ which are separated from f by v by at least k intervening edges and at most t intervening edges. For example, $\mathcal{T}^{1,1}_{v\to f}$ contains only $(y_{f'})_{f'\in\partial v\setminus f}$, and $\mathcal{T}^{2,1}_{v\to f}$ contains additional the observations $v_{v'}$ and $x_{f'v'}$ for $v' \in \partial f' \setminus v$ for some some $f' \in \partial v \setminus f$. The collections (or σ-algebras) $\mathcal{V}^{t,k}_{v\to f}$, $\mathcal{F}^{t,k}_{v\to f}$, $\mathcal{E}^{t,k}_{v\to f}$ are defined similarly, as are the versions of these where the roles of v and f are reversed.

B.4.1 Information-theoretic lower bound in the high-dimensional regression model

In this section, we prove Lemma 3.5.3 in the high-dimensional regression model.

Note that conditions on the conditional density in assumption R4 are equivalent positivity, boundedness, and the existence finite, non-negative constants q_k' such that $\frac{|\partial_x^k p(y|x)|}{p(y|x)} \le q_k'$ for $1 \le k \le 5$. We will often use this form of the assumption without further comment. This implies that for any random variable A

$$\frac{|\partial_x^k \mathbb{E}[p(y|x+A)]|}{\mathbb{E}[p(y|x+A)]} \le \int \frac{|\partial_x^k p(y|x+a)|}{p(y|x+a)} \frac{p(y|x+a)}{\mathbb{E}[p(y|x+A)]} \mu_A(\mathrm{d}a) \le q_k', \tag{B.26}$$

because $p(y|x+a)/\mathbb{E}[p(y|x+A)]$ is a probability density with respect to μ_A, the distribution of A.

Denote the regular conditional probability of Θ conditional on V for the measure $\mu_{\Theta,V}$ by $\mu_{\Theta|V} : \mathbb{R}\times\mathcal{B} \to [0,1]$, where $\mathcal{B}$ denotes the Borel σ-algebra on $\mathbb{R}$. The posterior of θ_v given $\mathcal{T}_{v,2t}$ has density with respect to $\mu_{\Theta|V}(v_v,\cdot)$ given by

$$p_v(\vartheta|\mathcal{T}_{v,2t}) \propto \int \prod_{f\in\mathcal{F}_{v,2t}} p(y_f \mid \sum_{v'\in\partial f} \vartheta_{v'} X_{v'f}, u_f) \prod_{v'\in\mathcal{V}_{v,2t}\setminus v} \mu_{\Theta|V}(v_{v'}, \mathrm{d}\vartheta_{v'}).$$

Asymptotically, the posterior density with respect to $\mu_{\Theta|V}(v_v,\cdot)$ behaves like that produced by a Gaussian observation of θ_v with variance τ_t^2, where τ_t is defined by (3.5).

Lemma B.4.1. *In the high-dimensional regression model, there exist $\mathcal{T}_{v,2t}$-measurable random variables $\tau_{v,t}, \chi_{v,t}$ such that*

$$p_v(\vartheta|\mathcal{T}_{v,2t}) \propto \exp\left(-\frac{1}{2\tau_{v,t}^2}(\chi_{v,t}-\vartheta)^2 + o_p(1)\right),$$

where $o_p(1)$ has no ϑ dependence. Moreover, $(\chi_{v,t}, \tau_{v,t}, \theta_v, v_v) \xrightarrow{\mathrm{d}} (\Theta + \tau_t G, \tau_t, \Theta, V)$ where $(\Theta, V) \sim \mu_{\Theta,V}$, $G \sim \mathsf{N}(0,1)$ independent of Θ, V, and τ_t is given by (3.5).

Proof.[Lemma B.4.1] We compute the posterior density $p_v(\vartheta|\mathcal{T}_{v,2t})$ via an iteration called belief propagation. For each edge $(v,f)\in\mathcal{E}$, belief propagation generates a pair of sequences of real-valued functions $(m_{v\to f}^t(\vartheta))_{t\geq 0}, (m_{f\to v}^t(\vartheta))_{t\geq 0}$. The iteration is

$$\begin{gathered} m_{v\to f}^0(\vartheta) = 1, \\ m_{f\to v}^s(\vartheta) \propto \int p(y_f|X_{fv}\vartheta + \sum_{v'\in\partial f\setminus v} X_{fv'}\vartheta_{v'}, u_f) \prod_{v'\in\partial f\setminus v} m_{v'\to f}^s(\vartheta_{v'}) \prod_{v'\in\partial f\setminus v} \mu_{\Theta|V}(v_{v'}, \mathrm{d}\vartheta_{v'}), \\ m_{v\to f}^{s+1}(\vartheta) \propto \prod_{f'\in\partial v\setminus f} m_{f'\to v}^s(\vartheta), \end{gathered}$$

with normalization $\int m_{f\to v}^t(\vartheta)\mu_{\Theta|V}(v_v,\mathrm{d}\vartheta) = \int m_{v\to f}^t(\vartheta)\mu_{\Theta|V}(v_v,\mathrm{d}\vartheta) = 1$. For any variable node v,

$$p_v(\vartheta|\mathcal{T}_{v,2t}) \propto \prod_{f\in\partial v} m_{f\to v}^{t-1}(\vartheta). \tag{B.27}$$

This equation is exact.

We define several quantities related to the belief propagation iteration.

$$\begin{aligned} \mu_{v\to f}^s &= \int \vartheta m_{v\to f}^s(\vartheta)\mu_{\Theta|V}(v_v,\mathrm{d}\vartheta), & (\tilde{\tau}_{v\to f}^s)^2 &= \int \vartheta^2 m_{v\to f}^s(\vartheta)\mu_{\Theta|V}(v_v,\mathrm{d}\vartheta) - (\mu_{v\to f}^s)^2, \\ \mu_{f\to v}^s &= \sum_{v'\in\partial f\setminus v} x_{fv'}\mu_{v'\to f}^s, & (\tilde{\tau}_{f\to v}^s)^2 &= \sum_{v'\in\partial f\setminus v} x_{fv'}^2(\tilde{\tau}_{v'\to f}^s)^2, \\ a_{f\to v}^s &= \frac{1}{x_{fv}}\frac{\mathrm{d}}{\mathrm{d}\vartheta}\log m_{f\to v}^s(\vartheta)\Big|_{\vartheta=0}, & b_{f\to v}^s &= -\frac{1}{x_{fv}^2}\frac{\mathrm{d}^2}{\mathrm{d}\vartheta^2}\log m_{f\to v}^s(\vartheta)\Big|_{\vartheta=0}, \\ a_{v\to f}^s &= \frac{\mathrm{d}}{\mathrm{d}\vartheta}\log m_{v\to f}^s(\vartheta)\Big|_{\vartheta=0}, & b_{v\to f}^s &= -\frac{\mathrm{d}^2}{\mathrm{d}\vartheta^2}\log m_{v\to f}^s(\vartheta)\Big|_{\vartheta=0}, \\ \chi_{v\to f}^s &= a_{v\to f}^s/b_{v\to f}^s, & (\tau_{v\to f}^s)^2 &= 1/b_{v\to f}^s. \end{aligned}$$

Lemma B.4.1 follows from the following asymptotic characterization of the quantities in the preceding display in the limit $n, p \to \infty$, $n/p \to \delta$:

$$\begin{gathered}
\mathbb{E}[(\mu^s_{v\to f})^2] \to \delta\sigma_s^2, \qquad \mathbb{E}[(\tilde{\tau}^s_{v\to f})^2] \to \delta\tilde{\tau}_s^2, \\
(\mu^s_{f\to v}, u_f) \xrightarrow{\mathrm{d}} \mathsf{N}(0,\sigma_s^2)\otimes\mu_U, \qquad (\tilde{\tau}^s_{f\to v})^2 \xrightarrow{\mathrm{p}} \tilde{\tau}_s^2, \\
(\theta_v, v_v, a^s_{v\to f}/b^s_{v\to f}, b^s_{v\to f}) \xrightarrow{\mathrm{d}} (\Theta, V, \Theta + \tau_s G, 1/\tau_s^2),
\end{gathered} \tag{B.28}$$

where in the last line $\Theta \sim \mu_\Theta$, $G \sim \mathsf{N}(0,1)$ independent, and σ_s^2, τ_s^2 are defined in (3.5). By symmetry, the distribution of these quantities does not depend upon v or f, so that the limits holds for all v, f once we establish them for any v, f. We establish the limits inductively in s.

Base case: $\mathbb{E}[(\mu^0_{v\to f})^2] \to \delta\sigma_0^2$ *and* $\mathbb{E}[(\tilde{\tau}^0_{v\to f})^2] \to \delta\tilde{\tau}_0^2$.

Observe that $\mu^s_{v\to f} = \int \vartheta\mu_{\Theta|V}(v_v, \mathrm{d}\vartheta) = \mathbb{E}_{\Theta,V}[\Theta|V = v_v]$. Because $v_v \sim \mu_V$, we have $\mathbb{E}[(\mu^1_{v\to f})^2] = \mathbb{E}_{\Theta,V}[\mathbb{E}_{\Theta,V}[\Theta|V]^2] = \mathbb{E}[\Theta^2] - \mathsf{mmse}_{\Theta,V}(\infty) = \delta\sigma_1^2$. Similarly, $(\tilde{\tau}^1_{v\to f})^2 = \mathrm{Var}_{\Theta,V}(\Theta|V = v_v)$, so that $\mathbb{E}[(\tilde{\tau}^1_{v\to f})^2] = \mathsf{mmse}_{\Theta,V}(\infty) = \delta\tilde{\tau}_0^2$.

Inductive step 1: If $\mathbb{E}[(\mu^s_{v\to f})^2] \to \delta\sigma_s^2$, *then* $(\mu^s_{f\to v}, u_f) \xrightarrow{\mathrm{d}} \mathsf{N}(0,\sigma_s^2)\otimes\mu_U$.

The quantity $\mu^s_{v'\to f}$ is $\mathcal{T}^{2s,0}_{v'\to f}$-measurable, whence it is independent of $x_{fv'}$ and u_f. Moreover, $(\mu_{v'\to f}, x_{fv})$ are independent as we vary $v' \in \partial f \setminus v$. Thus, $\mu^s_{f\to v}|\mathcal{T}^{2s+1,1}_{f\to v} \sim \mathsf{N}(0, \frac{1}{n}\sum_{v'\in\partial f\setminus v}(\mu^s_{v'\to f})^2)$. Note that $\mathbb{E}[\frac{1}{n}\sum_{v'\in\partial f\setminus v}(\mu^s_{v'\to f})^2] = (p-1)\mathbb{E}[(\mu^s_{v\to f})^2]/n \to \sigma_s^2$ by the inductive hypothesis. Moreover, $\mu^s_{v\to f}$ has bounded fourth moments because it is bounded by M. By the weak law of large numbers, $\frac{1}{n}\sum_{v'\in\partial f\setminus v}(\mu^s_{v'\to f})^2 \xrightarrow{\mathrm{p}} \sigma_s^2$. We conclude by Slutsky's theorem and independence that $(\mu^s_{f\to v}, u_f) \xrightarrow{\mathrm{d}} \mathsf{N}(0,\sigma_s^2)\otimes\mu_U$.

Inductive step 2: If $\mathbb{E}[(\tilde{\tau}^s_{v\to f})^2] \to \delta\tilde{\tau}_s^2$, *then* $(\tilde{\tau}^s_{f\to v})^2 \xrightarrow{\mathrm{p}} \tilde{\tau}_s^2$.

The quantity $\tilde{\tau}^s_{v'\to f}$ is $\mathcal{T}^{2s,0}_{v'\to f}$-measurable, whence it is independent of $x_{fv'}$. Therefore,

$$\mathbb{E}[\sum_{v'\in\partial f\setminus v} x^2_{fv'}(\tilde{\tau}^s_{v'\to f})^2] = (p-1)\mathbb{E}[(\tilde{\tau}^s_{v\to f})^2]/n \to \tilde{\tau}_s^2.$$

Moreover, $(\tilde{\tau}_{v'\to f}, x_{fv})$ are mutually independent as we vary $v' \in \partial f \setminus v$, and because $\tilde{\tau}^s_{v\to f}$ is bounded by M, the terms $nx^2_{fv'}(\sigma^s_{v'\to f})^2$ have bounded fourth moments. By the weak law of large numbers, $(\tilde{\tau}^s_{f\to v})^2 \xrightarrow{\mathrm{p}} \tilde{\tau}_s^2$.

Inductive step 3: If $(\mu^s_{f\to v}, u_f, \tilde{\tau}^s_{f\to v}) \xrightarrow{\mathrm{d}} \mathsf{N}(0,\sigma_s^2)\otimes\mu_U\otimes\delta_{\tilde{\tau}_s}$, *then* $(\theta_v, v_v, a^{s+1}_{v\to f}/b^{s+1}_{v\to f}, b^{s+1}_{v\to f}) \xrightarrow{\mathrm{d}} (\Theta, V, \Theta + \tau_{s+1}G, 1/\tau_{s+1}^2)$ *where* $G \sim \mathsf{N}(0,1)$ *independent of* $(\Theta, V) \sim \mu_{\Theta,V}$.

For all $(f, v) \in \mathcal{E}$ and $s \geq 1$, define

$$p^s_{f\to v}(y;x) = \int p(y|x + \sum_{v'\in\partial f\setminus v} x_{fv'}\vartheta_{v'}, u_f) \prod_{v'\in\partial f\setminus v} m^s_{v'\to f}(\vartheta_{v'}) \prod_{v'\in\partial f\setminus v} \mu_{\Theta|V}(v_{v'}, \mathrm{d}\vartheta_{v'}).$$

More compactly, we may write $p^s_{f\to v}(y;x,u_f) = \mathbb{E}_{\{\Theta_{v'}\}}[p(y|x+\sum_{v'\in\partial f\setminus v} x_{fv'}\Theta_{v'}, u_f)]$, where it is understood that the expectation is taken over $\Theta_{v'}$ independent with densities $m^s_{v'\to f}$ with respect to $\mu_{\Theta|V}(v_{v'},\cdot)$. Note that for all x, we have

$$\int p^s_{f\to v}(y;x)\mathrm{d}y = 1$$

everywhere. That is, $p^s_{f\to v}(\cdot;x)$ is a probability density with respect to Lebesgue measure. We will denote by $\dot p^s_{f\to v}(y;x) = \frac{\mathrm{d}}{\mathrm{d}\xi} p^s_{f\to v}(y;x)\big|_{\xi=x}$, and likewise for higher derivatives. These derivatives exist and may be taken under the integral by $\mathsf{R4}$. Define

$$a^s_{f\to v}(y) = \frac{\mathrm{d}}{\mathrm{d}x}\log p^s_{f\to v}(y;x)\Big|_{x=0} \qquad \text{and} \qquad b^s_{f\to v}(y) = -\frac{\mathrm{d}^2}{\mathrm{d}x^2}\log p^s_{f\to v}(y;x)\Big|_{x=0}.$$

For fixed y, the quantity $a^s_{f'\to v}(y)$ is independent of $x_{f'v}$, and $(a^s_{f'\to v}(y), x_{f'v})$ are mutually independent for $f' \in \partial v \setminus f$. Observe that

$$\begin{aligned} a^s_{f\to v} &= a^s_{f\to v}(y_f) \qquad \text{and} \qquad a^{s+1}_{v\to f} = \sum_{f'\in\partial v\setminus f} x_{f'v} a^s_{f'\to v}(y_{f'}),\\ b^s_{f\to v} &= b^s_{f\to v}(y_f) \qquad \text{and} \qquad b^{s+1}_{v\to f} = \sum_{f'\in\partial v\setminus f} x^2_{f'v} b^s_{f'\to v}(y_{f'}). \end{aligned}$$

We will study the distributions of $a^s_{f\to v}, a^{s+1}_{v\to f}, b^s_{f\to v}$, and $b^{s+1}_{v\to f}$ under several measures, which we now introduce. Define $P_{v,\vartheta}$ to be the distribution of the regression model with θ_v forced to be θ and v_v forced to be 0. That is, under $P_{v,\theta}$, we have $(\theta_{v'}, v_{v'}) \overset{\text{iid}}{\sim} \mu_{\Theta,V}$ for $v' \neq v$, $v_v = 0$ and $\theta_v = \theta$, the features are distributed independently $x_{fv'} \overset{\text{iid}}{\sim} \mathsf{N}(0,1/n)$ for all f, v', and the observations y_f are drawn independently from $p(\cdot|\sum_{v'\in\partial f} x_{fv'}\theta_{v'})$ for all f. We will consider the distribution of $a^s_{f\to v}, a^{s+1}_{v\to f}, b^s_{f\to v}$, and $b^{s+1}_{v\to f}$ under $P_{v,\theta}$ for $\theta \in [-M, M]$.

We require the following lemmas, whose proofs are deferred to Section B.4.1.

Lemma B.4.2. *Under $P_{v,\theta}$ for any $\theta \in [-M,M]$, we have for all fixed y that*

$$\begin{aligned} p^s_{f\to v}(y;0) - \mathbb{E}_{G_1}[p(y|\mu^s_{f\to v} + \tilde\tau^s_{f\to v}G_1, u_f)] &= o_p(1),\\ \dot p^s_{f\to v}(y;0) - \mathbb{E}_{G_1}[\dot p(y|\mu^s_{f\to v} + \tilde\tau^s_{f\to v}G_1, u_f)] &= o_p(1),\\ \ddot p^s_{f\to v}(y;0) - \mathbb{E}_{G_1}[\ddot p(y|\mu^s_{f\to v} + \tilde\tau^s_{f\to v}G_1, u_f)] &= o_p(1), \end{aligned}$$

where the expectation is over $G_1 \sim \mathsf{N}(0,1)$. Further, for any u, the functions $(\mu,\tilde\tau) \mapsto \mathbb{E}_{G_1}[p(y|\mu+\tilde\tau G_1, u)]$, $(\mu,\tilde\tau) \mapsto \mathbb{E}_{G_1}[\dot p(y|\mu+\tilde\tau G_1, u)]$, and $(\mu,\tilde\tau) \mapsto \mathbb{E}_{G_1}[\ddot p(y|\mu+\tilde\tau G_1, u)]$ are continuous.

Lemma B.4.3. *Under $P_{v,\theta}$ for any $\theta \in [-M,M]$, we have for any fixed s*

$$\log\frac{m^{s+1}_{v\to f}(\vartheta)}{m^{s+1}_{v\to f}(0)} = \vartheta a^{s+1}_{v\to f} - \frac{1}{2}\vartheta^2 b^{s+1}_{v\to f} + O_p(n^{-1/2}),$$

where $O_p(n^{-1/2})$ has no ϑ dependence, and the statement holds for $\vartheta \in [-M,M]$.

First we study the distribution of $a^{s+1}_{v\to f}, b^{s+1}_{v\to f}$ under $P_{v,0}$. Because $\mu^s_{f'\to v}, \tilde\tau^s_{f'\to v}$ is independent of θ_v, v_v for all $f' \in \partial v$, its distribution is the same under $P_{v,\theta}$ for all $\theta \in [-M,M]$ and is equal to its distribution under the original model. Thus, the inductive hypothesis implies $(\mu^s_{f\to v}, \tilde\tau^s_{f\to v}) \xrightarrow[P_{v,0}]{\mathrm{d}} \mathsf{N}(0,\sigma^2_s) \times \delta_{\tilde\tau_s}$.

By Lemma B.4.2, the inductive hypothesis, and Lemma B.1.3, we have for fixed y

$$\begin{pmatrix} \mathbb{E}_{G_1}[p(y|\mu^s_{f\to v} + \tilde{\tau}^s_{f\to v}G_1, u_f)] \\ \mathbb{E}_{G_1}[\dot{p}(y|\mu^s_{f\to v} + \tilde{\tau}^s_{f\to v}G_1, u_f)] \\ \mathbb{E}_{G_1}[\ddot{p}(y|\mu^s_{f\to v} + \tilde{\tau}^s_{f\to v}G_1, u_f)] \end{pmatrix} \xrightarrow[P_{v,0}]{\mathrm{d}} \begin{pmatrix} \mathbb{E}_{G_1}[p(y|\sigma_s G_0 + \tilde{\tau}_s G_1, U] \\ \mathbb{E}_{G_1}[\dot{p}(y|\sigma_s G_0 + \tilde{\tau}_s G_1, U)] \\ \mathbb{E}_{G_1}[\ddot{p}(y|\sigma_s G_0 + \tilde{\tau}_s G_1, U)] \end{pmatrix},$$

where and $G_0, G_1 \sim \mathsf{N}(0,1)$ and $U \sim \mu_U$ independent. Applying Lemma B.4.2 and Slutsky's Theorem, we have that

$$\begin{pmatrix} p^s_{f\to v}(y;0) \\ \dot{p}^s_{f\to v}(y;0) \\ \ddot{p}^s_{f\to v}(y;0) \end{pmatrix} \xrightarrow[P_{v,0}]{\mathrm{d}} \begin{pmatrix} \mathbb{E}_{G_1}[p(y|\sigma_s G_0 + \tilde{\tau}_s G_1, U)] \\ \mathbb{E}_{G_1}[\dot{p}(y|\sigma_s G_0 + \tilde{\tau}_s G_1, U)] \\ \mathbb{E}_{G_1}[\ddot{p}(y|\sigma_s G_0 + \tilde{\tau}_s G_1, U)] \end{pmatrix}.$$

By the Continuous Mapping Theorem,

$$\begin{aligned} p^s_{f\to v}(y;0) &\xrightarrow[P_{v,0}]{\mathrm{d}} \mathbb{E}_{G_1}[p(y|\sigma_s G_0 + \tilde{\tau}_s G_1, U)], \\ a^s_{f\to v}(y) &\xrightarrow[P_{v,0}]{\mathrm{d}} \frac{\mathrm{d}}{\mathrm{d}x} \log \mathbb{E}_{G_1}[\dot{p}(y|\sigma_s G_0 + \tilde{\tau}_s G_1, U)]\Big|_{x=0}, \\ b^s_{f\to v}(y) &\xrightarrow[P_{v,0}]{\mathrm{d}} -\frac{\mathrm{d}^2}{\mathrm{d}x^2} \log \mathbb{E}_{G_1}[\dot{p}(y|\sigma_s G_0 + \tilde{\tau}_s G_1, U)]\Big|_{x=0}. \end{aligned}$$

Because the quantity $p(y|x)$ is bounded (assumption R4) and the quantities $a^s_{f\to v}(y), b^s_{f\to v}(y)$ are bounded by (B.26), we have

$$\begin{aligned} \mathbb{E}_{P_{v,0}}[p^s_{f\to v}(y|0)] &\to \mathbb{E}_{G_0,G_1,U}[p(y|\sigma_s G_0 + \tilde{\tau}_s G_1, U)], \\ \mathbb{E}_{P_{v,0}}[a^s_{f\to v}(y)^2] &\to \mathbb{E}_{G_0,U}\left[\left(\frac{\mathrm{d}}{\mathrm{d}x} \log \mathbb{E}_{G_1}[\dot{p}(y|\sigma_s G_0 + \tilde{\tau}_s G_1, U)]\Big|_{x=0}\right)^2\right], \\ \mathbb{E}_{P_{v,0}}[b^s_{f\to v}] &\to -\mathbb{E}_{G_0,U}\left[\frac{\mathrm{d}^2}{\mathrm{d}x^2} \log \mathbb{E}_{G_1}[\dot{p}(y|\sigma_s G_0 + \tilde{\tau}_s G_1, U)]\Big|_{x=0}\right]. \end{aligned}$$

Under $P_{v,0}$, we have for all $f' \in \partial v$ that the random variable $y_{f'}$ is independent of $x_{f'v}$. Thus, conditional on $\mathcal{T}^{2s+2,1}_{v\to f}$, the random variable $\sum_{f'\in\partial v\setminus f} x_{f'v} a^s_{f'\to v}(y_{f'})$ is normally distributed. Specifically,

$$\sum_{f'\in\partial v\setminus f} x_{f'v} a^s_{f'\to v}(y_{f'}) \mid \mathcal{T}^{2s+2,1}_{v\to f} \underset{P_{v,0}}{\sim} \mathsf{N}\left(0, \frac{1}{n}\sum_{f'\in\partial v\setminus f} (a^s_{f'\to v}(y_{f'}))^2\right).$$

Because $(a^s_{f'\to v}(y_{f'}))^2$ is bounded by (B.26), if we show $\mathbb{E}_{P_{v,0}}[(a^s_{f\to v}(y_f))^2] \to 1/\tau^2_{s+1}$, then the weak law of large numbers and Slutsky's theorem will imply that

$$a^{s+1}_{v\to f} = \sum_{f'\in\partial v\setminus f} x_{f'v} a^s_{f'\to v}(y_{f'}) \xrightarrow[P_{v,0}]{\mathrm{d}} \mathsf{N}\left(0, 1/\tau^2_{s+1}\right). \tag{B.29}$$

We compute

$$\mathbb{E}_{P_{v,0}}[(a^s_{f\to v}(y_f))^2] = \mathbb{E}_{P_{v,0}}[\mathbb{E}_{P_{v,0}}[(a^s_{f\to v}(y_f))^2|\sigma(\mathcal{T}^{2s+1,1}_{f\to v}, (x_{fv'})_{v'\in\partial f\setminus v}), u_f]]$$

$$
\begin{aligned}
&= \mathbb{E}_{P_{v,0}}\left[\int a^s_{f\to v}(y)^2 p^s_{f\to v}(y;0)\mathrm{d}y\right] \\
&= \int \mathbb{E}_{P_{v,0}}\left[a^s_{f\to v}(y)^2 p^s_{f\to v}(y;0)\right]\mathrm{d}y.
\end{aligned}
$$

where the second equation holds because under $P_{v,0}$ we have $y_f \mid \sigma(\mathcal{T}^{2s+1,1}_{f\to v}, (x_{fv'})_{v'\in\partial f\setminus v}, u_f)$ has density $p^s_{f\to v}(\cdot;0)$ with respect to Lebesgue measure, and the last equation follows by Fubini's theorem (using the non-negativity of the integrand). Because $a^s_{f\to v}(y)^2 \le (q_1')^2$ and $\mathbb{E}_{P_{v,0}}[p^s_{f\to v}(y;0)]$ are probability densities which converge pointwise to $\mathbb{E}_{G_1}[p(y|\sigma_s G_0 + \tilde{\tau}_s G_1)]$, we conclude that

$$
\begin{aligned}
\mathbb{E}_{P_{v,0}}[(a^s_{f\to v}(y_f))^2] &\to \int \mathbb{E}_{G_0,U}\left[\frac{\mathbb{E}_{G_1}[\dot{p}(y|\sigma_s G_0 + \tilde{\tau}_s G_1, U)]^2}{\mathbb{E}_{G_1}[p(y|\sigma_s G_0 + \tilde{\tau}_s G_1, U)]}\right]\mathrm{d}y \\
&= \mathbb{E}_{G_0,U}\left[\int \frac{\mathbb{E}_{G_1}[\dot{p}(y|\sigma_s G_0 + \tilde{\tau}_s G_1, U)]^2}{\mathbb{E}_{G_1}[p(y|\sigma_s G_0 + \tilde{\tau}_s G_1, U)]}\mathrm{d}y\right] = \frac{1}{\tau^2_{s+1}},
\end{aligned}
$$

where we have used the alternative characterization of the recursion (3.5) from Lemma B.1.4. We conclude (B.29).

Now we compute the asymptotic behavior of $b^{s+1}_{v\to f}$ under $P_{v,0}$. Under $P_{v,0}$, $x_{f'v}$ is independent of $y_{f'}$, and $(x_{f'v}, b^s_{f'\to v}(y_{f'}))$ are mutually independent for $f' \in \partial v\setminus f$. Thus, $\mathbb{E}_{P_{v,0}}[x^2_{f'v} b^s_{f'\to v}(y_{f'})] = \mathbb{E}_{P_{v,0}}[b^s_{f'\to v}(y_{f'})]/n$. Because $b^s_{f'\to v}(y_{f'})$ is bounded by (B.26), if we can show that $\mathbb{E}_{P_{v,0}}[b^s_{f'\to v}(y_{f'})] \to 1/\tau^2_{s+1}$, then $b^{s+1}_{v\to f} \xrightarrow[P_{v,0}]{\mathrm{p}} 1/\tau^2_{s+1}$ will follow by the weak law of large numbers. We compute

$$
\begin{aligned}
\mathbb{E}_{P_{v,0}}[b^s_{f\to v}(y_f)] &= \mathbb{E}_{P_{v,0}}[\mathbb{E}_{P_{v,0}}[b^s_{f\to v}(y_f)|\sigma(\mathcal{T}^{2s+1,1}_{f\to v}, (x_{fv'})_{v'\in\partial f\setminus v}, u_f)]] \\
&= \mathbb{E}_{P_{v,0}}\left[\int b^s_{f\to v}(y) p^s_{f\to v}(y;0)\mathrm{d}y\right] \\
&= \int \mathbb{E}_{P_{v,0}}\left[b^s_{f\to v}(y) p^s_{f\to v}(y;0)\right]\mathrm{d}y
\end{aligned}
$$

where the last equation follows by Fubini's theorem (using that the integrand is bounded by the integrable function $q_2\mathbb{E}_{P_{v,0}}[p^s_{f\to v}(y;0)]$). The integrands converge point-wise, so that

$$
\begin{aligned}
&\mathbb{E}_{P_{v,0}}[b^s_{f\to v}(y_f)] \\
&\quad\to \mathbb{E}_{G_0,U}\left[\int \frac{\mathbb{E}_{G_1}[\dot{p}(y|\sigma_s G_0 + \tilde{\tau}_s G_1, U)]^2}{\mathbb{E}_{G_1}[p(y|\sigma_s G_0 + \tilde{\tau}_s G_1, U)]}\mathrm{d}y\right] - \int \mathbb{E}_{G_0,G_1,U}[\ddot{p}(y|\sigma_s G_0 + \tilde{\tau}_s G_1, U)]\mathrm{d}y \\
&\quad= \frac{1}{\tau^2_{s+1}},
\end{aligned}
$$

where we have concluded that the second integral is zero because $x \mapsto \mathbb{E}_{G_0,G_1,U}[p(y|\sigma_s G_0 + \tilde{\tau}_s G_1, U)]$ parameterizes a statistical model whose scores up to order 3 are bounded by (B.26). Thus, we conclude that $b^{s+1}_{v\to f} \xrightarrow[P_{v,0}]{\mathrm{p}} 1/\tau^2_{s+1}$.

Now we compute the asymptotic distribution of $(a^{s+1}_{v\to f}, b^{s+1}_{v\to f})$ under $P_{v,\theta}$ for any $\theta \in [-M, M]$. The log-likelihood ratio between $P_{v,\theta}$ and $P_{v,0}$ is

$$
\sum_{f'\in\partial v} \log \frac{p^s_{f'\to v}(y_{f'}|x_{f'v}\theta)}{p^s_{f'\to v}(y_{f'}|0)} = \log \frac{m^{s+1}_{v\to f}(\theta)}{m^{s+1}_{v\to f}(0)} + \log \frac{p^s_{f\to v}(y_f|x_{fv}\theta)}{p^s_{f\to v}(y_f|0)}
$$

$$= \theta a_{v\to f}^{s+1} - \frac{1}{2}\theta^2 b_{v\to f}^{s+1} + O_p(n^{-1/2}),$$

where we have used Lemma B.4.3 and that $\left|\log \frac{p_{f\to v}^s(y_f|x_{fv}\theta)}{p_{f\to v}^s(y_f|0)}\right| \le Mq_1|x_{fv}| = O_p(n^{-1/2})$. Thus,

$$\left(a_{v\to f}^{s+1}, b_{v\to f}^{s+1}, \log\frac{P_{v,\theta}}{P_{v,0}}\right) \xrightarrow[P_{v,0}]{\mathrm{p}} \left(Z, \frac{1}{\tau_{s+1}^2}, \theta Z - \frac{1}{2}\frac{\theta^2}{\tau_{s+1}^2}\right),$$

where $Z \sim \mathsf{N}(0, 1/\tau_{s+1}^2)$. By Le Cam's third lemma [188, Example 6.7], we have

$$(a_{v\to f}^{s+1}, b_{v\to f}^{s+1}) \xrightarrow[P_{v,\theta}]{\mathrm{d}} \left(Z', \frac{1}{\tau_{s+1}^2}\right).$$

where $Z' \sim \mathsf{N}(\theta/\tau_{s+1}^2, 1/\tau_{s+1}^2)$. By the Continuous Mapping Theorem [188, Theorem 2.3], we conclude $(a_{v\to f}^{s+1}/b_{v\to f}^{s+1}, b_{v\to f}^{s+1}) \xrightarrow[P_{v,\theta}]{\mathrm{d}} \mathsf{N}(\theta, \tau_{s+1}^2) \otimes \delta_{1/\tau_{s+1}^2}$.

Denote by P^* the distribution of the the original model. Consider a continuous bounded function $f : (\theta, \nu, \chi, b) \mapsto \mathbb{R}$, and define $\hat{f}_n(\theta, \nu) = \mathbb{E}_{P_{v,\theta}}[f(\theta, \nu, a_{v\to f}^{s+1}/b_{v\to f}^{s+1}, b_{v\to f}^{s+1})]$. Under P^*, the random variables $a_{v\to f}^{s+1}, b_{v\to f}^{s+1}$ are functions are θ_v and random vectors $\boldsymbol{D} := \mathcal{T}_{v,2t} \setminus \{\theta_v, v_v\}$, which is independent of θ_v, v_v. In particular, we may write

$$\mathbb{E}_{P^*}[f(\theta_v, v_v, a_{v\to f}^{s+1}/b_{v\to f}^{s+1}, b_{v\to f}^{s+1})] = \mathbb{E}_{P^*}[f(\theta_v, v_v, \chi(\theta_v, \boldsymbol{D}), B(\theta_v, \boldsymbol{D}))],$$

for some measurable functions χ, B. We see that

$$\mathbb{E}_{P^*}[f(\theta_v, v_v, a_{v\to f}^{s+1}/b_{v\to f}^{s+1}, b_{v\to f}^{s+1}) \mid \theta_v, v_v] = \hat{f}_n(\theta_v, v_v)$$

where

$$\hat{f}_n(\theta, \nu) = \mathbb{E}_{\boldsymbol{D}}[f(\theta, \nu, \chi(\theta, \boldsymbol{D}), B(\theta, \boldsymbol{D}))],$$

with $\boldsymbol{D}$ distributed as it is under P^* (see e.g., [76, Example 5.1.5]). Because $\boldsymbol{D}$ has the same distribution on P^* as under $P_{v,\theta}$, we see that in fact $\hat{f}_n(\theta, \nu) = \mathbb{E}_{P_{v,\theta}}[f(\theta, \nu, a_{v\to f}^{s+1}/b_{v\to f}^{s+1}, b_{v\to f}^{s+1})]$. Because $(a_{v\to f}^{s+1}/b_{v\to f}^{s+1}, b_{v\to f}^{s+1}) \xrightarrow[P_{v,\theta}]{\mathrm{d}} \mathsf{N}(\theta, \tau_{s+1}^2) \otimes \delta_{1/\tau_{s+1}^2}$, we conclude that $\hat{f}_n(\theta, \nu) \to \mathbb{E}_G[f(\theta, \nu, \theta + \tau_{s+1}G, \tau_{s+1}^{-2})]$ for all θ, ν. By bounded convergence and the tower property, $\mathbb{E}_{\Theta,V}[\hat{f}_n(\Theta, V)] \to \mathbb{E}_{\Theta,V,G}[f(\theta, \nu, \theta + \tau_{s+1}G, \tau_{s+1}^{-2})]$ where $(\Theta, V) \sim \mu_{\Theta,V}$ independent of $G \sim \mathsf{N}(0,1)$. Also by the tower property, we have

$$\mathbb{E}_{\Theta,V}[\hat{f}_n(\Theta, V)] = \mathbb{E}_{P^*}[f(\theta_v, v_v, \chi(\theta_v, \boldsymbol{D}), B(\theta_v, \boldsymbol{D}))] = \mathbb{E}_{P^*}[f(\theta_v, v_v, a_{v\to f}^{s+1}/b_{v\to f}^{s+1}, b_{v\to f}^{s+1})].$$

We conclude

$$\mathbb{E}_{P^*}[f(\theta_v, v_v, a_{v\to f}^{s+1}/b_{v\to f}^{s+1}, b_{v\to f}^{s+1})] \to \mathbb{E}_{\Theta,V,G}[f(\Theta, V, \Theta + \tau_{s+1}G, \tau_{s+1}^{-2})].$$

Thus, we conclude that $(\theta_v, v_v, a_{v\to f}^{s+1}/b_{v\to f}^{s+1}, b_{v\to f}^{s+1}) \xrightarrow[P^*]{\mathrm{d}} (\Theta, V, \Theta + \tau_{s+1}G, 1/\tau_{s+1}^2)$, as desired.

Inductive step 4: If $(\theta_v, v_v, a_{v\to f}^{s+1}/b_{v\to f}^{s+1}, b_{v\to f}^{s+1}) \xrightarrow{\mathrm{d}} (\Theta, V, \Theta + \tau_{s+1}G, 1/\tau_{s+1}^2)$ *where* $G \sim \mathsf{N}(0,1)$ *independent of* $(\Theta, V) \sim \mu_{\Theta,V}$*, then* $\mathbb{E}[(\mu_{v\to f}^s)^2] \to \delta\sigma_s^2$ *and* $\mathbb{E}[(\tilde{\tau}_{v\to f}^s)^2] \to \mathsf{mmse}_{\Theta,V}(\tau_s^2)$.

Define

$$\epsilon^s_{v\to f} = \sup_{\vartheta\in[-M,M]} \left| \log \frac{m^s_{v\to f}(\vartheta)}{m^s_{v\to f}(0)} - \left(\vartheta a^s_{v\to f} - \frac{1}{2}\vartheta^2 b^s_{v\to f}\right)\right|,$$

where because all the terms are continuous in ϑ, the random variable $\epsilon^s_{v\to f}$ is measurable and finite. We have that

$$\mu^s_{v\to f} \geq \frac{\int \vartheta \exp(\vartheta a^s_{v\to f} - \vartheta^2 b^s_{v\to f}/2 - \epsilon^s_{v\to f})\mu_\Theta(v_v, \mathrm{d}\vartheta)}{\int \exp(\vartheta a^s_{v\to f} - \vartheta^2 b^s_{v\to f}/2 + \epsilon^s_{v\to f})\mu_\Theta(v_v, \mathrm{d}\vartheta)} \geq e^{-2\epsilon^s_{v\to f}}\eta_{\Theta,V}(a^s_{v\to f}/b^s_{v\to f}, v_v; 1/b^s_{v\to f})$$

where $\eta_{\Theta,V}(y, v; \tau^2) = \mathbb{E}_{\Theta,V,G}[\Theta|\Theta + \tau G = y; V = v]$ where $(\Theta, V) \sim \mu_{\Theta,V}$, $G \sim \mathsf{N}(0, 1)$ independent. Likewise,

$$\mu^s_{v\to f} \leq e^{2\epsilon^s_{v\to f}}\eta_{\Theta,V}(a^s_{v\to f}/b^s_{v\to f}, v_v; 1/b^s_{v\to f}).$$

Because $\eta_{\Theta,V}$ takes values in the bounded interval $[-M, M]$ and $\epsilon_{v\to f} = o_p(1)$ by Lemma B.4.3, we conclude that

$$\mu^s_{v\to f} = \eta_{\Theta,V}(a^s_{v\to f}/b^s_{v\to f}, v_v; 1/b^s_{v\to f}) + o_p(1).$$

For a fixed v_v, the Bayes estimator $\eta_{\Theta,V}$ is continuous in the observation and the noise variance on $\mathbb{R}\times\mathbb{R}_{>0}$.[4] Thus, by the inductive hypothesis and the fact that $v_v \sim \mu_V$ for all n, we have $\mathbb{E}[\eta_{\Theta,V}(a^s_{v\to f}/b^s_{v\to f}, v_v; 1/b^s_{v\to f})^2] = \mathbb{E}[\eta_{\Theta,V}(a^s_{v\to f}/b^s_{v\to f}, v_v; 1/b^s_{v\to f})^2 \vee M^2] \to \mathbb{E}_{\Theta,V,G}[\eta_{\Theta,V}(\Theta + \tau_s G, V; \tau_s^2)] = \mathbb{E}[\Theta^2] - \mathsf{mmse}_{\Theta,V}(\tau_s^2) = \delta\sigma_s^2$ by Lemma B.1.3. By the previous display and the boundedness of $\mu^s_{v\to f}$ and $\eta_{\Theta,V}$, we conclude $\mathbb{E}[(\mu^s_{v\to f})^2] \to \delta\sigma_s^2$, as desired.

Similarly, we may derive that

$$\begin{aligned} e^{-2\epsilon^s_{v\to f}} s^2_{\Theta,V}(a^s_{v\to f}/b^s_{v\to f}, v_v; 1/b^s_{v\to f}) &\leq \int \vartheta^2 m^s_{v\to f}(\vartheta)\mu_\Theta(\mathrm{d}\vartheta) \\ &\leq e^{2\epsilon^s_{v\to f}} s^2_{\Theta,V}(a^s_{v\to f}/b^s_{v\to f}, v_v; 1/b^s_{v\to f}), \end{aligned}$$

where $s^2_{\Theta,V}(y, v; \tau^2) = \mathbb{E}_{\Theta,V,G}[\Theta^2|\Theta + \tau G = y, V = v]$ where $(\Theta, V) \sim \mu_{\Theta,V}$, $G \sim \mathsf{N}(0, 1)$ independent. For fixed v_v, the the posterior second moment is continuous in the observation and the noise variance. Further, it is bounded by M^2. Thus, by exactly the same argument as in the previous paragraph, we have that $\mathbb{E}[(\tilde{\tau}^s_{v\to f})^2] \to \mathbb{E}_{\Theta,V,G}[s^2_{\Theta,V}(\Theta + \tau_s G, V; \tau_s^2) - \eta_{\Theta,V}(\Theta + \tau_s G, V; \tau_s^2)^2] = \mathsf{mmse}_{\Theta,V}(\tau_s^2)$, as desired.

The inductive argument is complete, and (B.28) is established.

To complete the proof of Lemma B.4.1, first observe by (B.27) that we may express $\log p_v(\vartheta|\mathcal{T}_{v,2t})$ as, up to a constant, $\log \frac{m^t_{v\to f}(\vartheta)}{m^t_{v\to f}(0)} + \log \frac{m^{t-1}_{f\to v}(\vartheta)}{m^{t-1}_{f\to v}(0)}$. Note that

$$\left|\log \frac{m^{t-1}_{f\to v}(\vartheta_v)}{m^{t-1}_{f\to v}(0)}\right| \leq M|x_{fv}| \sup_{x\in\mathbb{R}} \left|\frac{\dot{p}^{t-1}_{f\to v}(y_f; x)}{p^{t-1}_{f\to v}(y_f; x)}\right| \leq Mq_1|x_{fv}| = o_p(1).$$

By Lemma B.4.3, we have that, up to a constant, $\log \frac{m^t_{v\to f}(\vartheta)}{m^t_{v\to f}(0)} = -\frac{1}{2}b^s_{v\to f}\left(a^t_{v\to f}/b^t_{v\to f} - \vartheta\right)^2 + o_p(1)$. The lemma follows from (B.28).

□

[4] This commonly known fact holds, for example, by [124, Theorem 2.7.1] because the posterior mean can be viewed as the mean in an exponential family paramterized by the observation and noise variance.

We complete the proof of Lemma 3.5.3 for the high-dimensional regression model. Consider any estimator $\hat{\theta} : \mathcal{T}_{v,2t} \mapsto [-M, M]$ on the computation tree. We compute

$$\begin{aligned}
\mathbb{E}[\ell(\theta_v, \hat{\theta}(\mathcal{T}_{v,2t}))] &= \mathbb{E}[\mathbb{E}[\ell(\theta_v, \hat{\theta}(\mathcal{T}_{v,2t}))|\mathcal{T}_{v,2t}]] \\
&= \mathbb{E}\left[\int \ell(\vartheta, \hat{\theta}(\mathcal{T}_{v,2t}))\frac{1}{Z(\mathcal{T}_{v,2t})}\exp\left(-\frac{1}{2\tau_{v,t}^2}(\chi_{v,t}-\vartheta)^2 + o_p(1)\right)\mu_{\Theta|V}(v_v, \mathrm{d}\vartheta)\right] \\
&\geq \mathbb{E}\left[\exp(-2\epsilon_v)\int \ell(\vartheta, \hat{\theta}(\mathcal{T}_{v,2t}))\frac{1}{Z(\chi_{v,t}, \tau_{v,t}, v_v)}\exp\left(-\frac{1}{2\tau_{v,t}^2}(\chi_{v,t}-\vartheta)^2\right)\mu_{\Theta|V}(v_v, \mathrm{d}\vartheta)\right] \\
&\geq \mathbb{E}\left[\exp(-2\epsilon_v)R(\chi_{v,2t}, \tau_{v,2t}, v_v)\right],
\end{aligned}$$

where $Z(\mathcal{T}_{v,2t}) = \int \exp\left(-\frac{1}{2\tau_{v,t}^2}(\chi_{v,t}-\vartheta)^2 + o_p(1)\right)\mu_{\Theta|V}(v_v, \mathrm{d}\vartheta)$,

$$R(\chi, \tau, v) := \inf_{d\in\mathbb{R}} \int \frac{1}{Z}\ell(\vartheta, d)e^{-\frac{1}{2\tau^2}(\chi-\vartheta)^2}\mu_{\Theta|V}(v, \mathrm{d}\vartheta)\,,$$

and

$$\epsilon_v = \sup_{\vartheta\in[-M,M]}\left|\log\frac{p(\vartheta|\mathcal{T}_{v,2t})}{p(0|\mathcal{T}_{v,2t})} + \vartheta\chi_{v,t}/\tau_{v,t}^2 - \vartheta^2/(2\tau_{v,t}^2)\right|.$$

Because Θ is bounded support, by Lemma B.1.5(b), $R(\chi, \tau, v)$ is continuous in (χ, τ) on $\mathbb{R}\times\mathbb{R}_{>0}$. By Lemma B.4.1, $\epsilon_v = o_p(1)$. The quantity on the right-hand side does not depend on $\hat{\theta}$, so provides a uniform lower bound over the performance of any estimator. Because $(v_v, \chi_{v,2t}, \tau_{v,2t}, \epsilon_v) \xrightarrow{\mathrm{d}} (V, \Theta+\tau_t G, \tau_t, 0)$, $v_v \overset{\mathrm{d}}{=} V$ for all n, and $\tau_t > 0$, we have $\mathbb{E}\left[\exp(-2\epsilon_v)R(\chi_{v,2t}, \tau_{v,2t}, v_v)\right] \to \mathbb{E}[R(\Theta+\tau_t G, \tau_t, V)] = \inf_{\hat{\theta}(\cdot)}\mathbb{E}[\ell(\Theta, \hat{\theta}(\Theta+\tau_t G, V))]$, where the convergence holds by Lemma B.1.3 and the equality holds by Lemma B.1.5(a). Thus,

$$\liminf_{n\to\infty}\inf_{\hat{\theta}(\cdot)}\mathbb{E}[\ell(\theta_v, \hat{\theta}(\mathcal{T}_{v,2t}))] \geq \inf_{\hat{\theta}(\cdot)}\mathbb{E}[\ell(\Theta, \hat{\theta}(\Theta+\tau_t G))].$$

The proof of Lemma 3.5.3 in the high-dimensional regression model is complete.

Technical tools

Proof.[Lemma B.4.2] By Lindeberg's principle (see, e.g., [51]) and using that μ_Θ is supported on $[-M, M]$, we have

$$\begin{aligned}
|p^s_{f\to v}(y;0) - \mathbb{E}_{G_1}[p(y|\mu^s_{f\to v} + \tilde{\tau}^s_{f\to v}G_1, u_f)]| &\leq \frac{M^3\sup_{x\in\mathbb{R}}|\partial_x^3 p(y|x,u_f)|}{3}\sum_{v'\in\partial f\setminus v}|x_{fv'}|^3, \\
|\dot{p}^s_{f\to v}(y;0) - \mathbb{E}_{G_1}[\dot{p}(y|\mu^s_{f\to v} + \tilde{\tau}^s_{f\to v}G_1, u_f)]| &\leq \frac{M^3\sup_{x\in\mathbb{R}}|\partial_x^4 p(y|x,u_f)|}{3}\sum_{v'\in\partial f\setminus v}|x_{fv'}|^3, \\
|\ddot{p}^s_{f\to v}(y;0) - \mathbb{E}_{G_1}[\ddot{p}(y|\mu^s_{f\to v} + \tilde{\tau}^s_{f\to v}G_1, u_f)]| &\leq \frac{M^3\sup_{x\in\mathbb{R}}|\partial_x^5 p(y|x,u_f)|}{3}\sum_{v'\in\partial f\setminus v}|x_{fv'}|^3.
\end{aligned}$$

Using the $\sup_{x\in\mathbb{R}}|\partial_x^k p(y|x,u)| \leq q_k' \sup_{x\in\mathbb{R}}|p(y|x,u)| < \infty$ for $k = 3, 4, 5$ by R4, we have that for fixed y the expectations on the right-hand side go to 0 as $n\to\infty$, whence the required expessions are $o_p(1)$.

Further, $|\mathbb{E}_{G_1}[p(y|\mu+\tilde{\tau}G_1, u)] - \mathbb{E}_{G_1}[p(y|\mu'+\tilde{\tau}'G_1, u)]| \leq (|\mu-\mu'| + |\tilde{\tau}-\tilde{\tau}'|\sqrt{2/\pi})\sup_{x\in\mathbb{R}}|\dot{p}(y|x,u)|$, whence $\mathbb{E}_{G_1}[p(y|\mu+\tilde{\tau}G_1, u)]$ is continuous in $(\mu, \tilde{\tau})$ by R4. The remaining continuity results follow similarly.

□

Proof.[Lemma B.4.3] Fix any $\vartheta \in [-M, M]$. By Taylor's theorem, there exist $\vartheta_i \in [-M, M]$ (in fact, between 0 and ϑ) such that

$$\begin{aligned}\log \frac{m^{s+1}_{v\to f}(\vartheta)}{m^{s+1}_{v\to f}(0)} &= \sum_{f' \in \partial v \setminus f} \log \frac{m^{s}_{f'\to v}(\vartheta)}{m^{s}_{f'\to v}(0)} \\ &= \vartheta a^{s+1}_{v\to f} - \frac{1}{2}\vartheta^2 b^{s+1}_{v\to f} + \frac{1}{6}\vartheta^3 \sum_{f' \in \partial v \setminus f} \left(\frac{\mathrm{d}^3}{\mathrm{d}\vartheta^3} \log \mathbb{E}_{\hat{G}_{f'}}[p(y_{f'}|x_{f'v}\vartheta + \hat{G}_{f'}, u_{f'})] \Big|_{\vartheta = \vartheta_i} \right).\end{aligned}$$

where it is understood that expectation is taken with respect to $\hat{G}_{f'} \overset{\mathrm{d}}{=} \sum_{v' \in \partial f' \setminus v} x_{f'v'} \Theta_{v' \to f'}$ where $x_{f'v'}$ is considered fixed and $\Theta_{v'\to f'}$ are drawn independently with densities $m^s_{v' \to f'}$ with respect to $\mu_{\Theta|V}(v_{v'}, \cdot)$. We bound the sum using assumption R4:

$$\left| \sum_{f' \in \partial v \setminus f} \left(\frac{\mathrm{d}^3}{\mathrm{d}\vartheta^3} \log \mathbb{E}_{\hat{G}_{f'}}[p(y_f|x_{fv}\vartheta + \hat{G}_{f'}, u_{f'})] \Big|_{\vartheta = \vartheta_i} \right) \right| \le q_3 \sum_{f' \in \partial v \setminus f} |x_{f'v}|^3 = O_p(n^{-1/2}).$$

The proof is complete.

□

B.4.2 Information-theoretic lower bound in the low-rank matrix estimation model

In this section, we prove Lemma 3.5.3 in the low-rank matrix estimation model.

Recall that conditions on the conditional density in assumption R4 are equivalent positivity, boundedness, and the existence finite, non-negative constants q'_k such that $\frac{|\partial_x^k p(y|x)|}{p(y|x)} \le q'_k$ for $1 \le k \le 5$. In particular, we have (B.26) for any random variable A.

Denote the regular conditional probability of Θ conditional on V for the measure $\mu_{\Theta,V}$ by $\mu_{\Theta|V} : \mathbb{R} \times \mathcal{B} \to [0, 1]$, where $\mathcal{B}$ denotes the Borel σ-algebra on $\mathbb{R}$, similarly for $\mu_{\Lambda|U}$. The posterior density of θ_v given $\mathcal{T}_{v,2t-1}$ has density respect to $\mu_{\Theta|V}(v_v, \cdot)$ given by

$$p_v(\vartheta_v|\mathcal{T}_{v,2t-1}) \propto \int \prod \exp\left(-\frac{n}{2}(x_{f'v'} - \frac{1}{n}\ell_{f'}\vartheta_{v'})^2\right) \prod \mu_\Lambda(u_f, \mathrm{d}\ell_f) \prod \mu_\Theta(v_{v'}, \mathrm{d}\vartheta_{v'}),$$

where the produces are over $(f', v') \in \mathcal{E}_{v,2t-1}$, $f \in \mathcal{F}_{v,2t-1}$, and $v' \in \mathcal{V}_{v,2t-1}$, respectively. Asymptotically, the posterior behaves like that produced by a Gaussian observation of θ_v with variance τ_t^2.

Lemma B.4.4. *In the low-rank matrix estimation model, there exist $\mathcal{T}_{v,2t-1}$-measurable random variables $q_{v,t}, \chi_{v,t}$ such that for fixed $t \ge 1$*

$$p_v(\vartheta|\mathcal{T}_{v,2t-1}) \propto \exp\left(-\frac{1}{2}(\chi_{v,t} - q_{v,t}^{1/2}\vartheta)^2 + o_p(1)\right),$$

where $o_p(1)$ has no ϑ dependence. Moreover, $(\theta_v, v_v, \chi_{v,t}, q_{v,t}) \xrightarrow{\mathrm{d}} (\Theta, V, q_t^{1/2}\Theta + G, q_t)$ where $(\Theta, V) \sim \mu_{\Theta,V}$, $G \sim \mathsf{N}(0, 1)$ independent of Θ, V, and q_t is given by (3.7).

Proof.[Lemma B.4.4] As in the proof of Lemma B.4.1, we compute the posterior density $p_v(\vartheta|\mathcal{T}_{v,2t-1})$ via belief propogation. The belief propagation iteration is

$$\begin{aligned} m^0_{f\to v}(\ell) &= 1, \\ m^{s+1}_{v\to f}(\vartheta) &\propto \int \prod_{f'\in\partial v\setminus f}\left(\exp\left(-\frac{n}{2}(x_{f'v}-\frac{1}{n}\ell_{f'}\vartheta)^2\right)m^s_{f'\to v}(\ell_{f'})\mu_{\Lambda|U}(u_{f'},\mathrm{d}\ell_{f'})\right), \\ m^s_{f\to v}(\ell) &\propto \int \prod_{v'\in\partial f\setminus v}\left(\exp\left(-\frac{n}{2}(x_{fv'}-\frac{1}{n}\ell\vartheta_{v'})^2\right)m^s_{v'\to f}(\vartheta_{v'})\mu_{\Theta|V}(v_{v'},\mathrm{d}\vartheta_{v'})\right), \end{aligned}$$

with normalization $\int m^s_{f\to v}(\ell)\mu_{\Lambda|U}(u_f,\mathrm{d}\ell) = \int m^s_{v\to f}(\vartheta)\mu_{\Theta|V}(v_v,\mathrm{d}\vartheta) = 1$. For $t \geq 1$

$$p_v(\vartheta|\mathcal{T}_{v,2t-1}) \propto \int \prod_{f\in\partial v}\left(\exp\left(-\frac{n}{2}(x_{fv}-\frac{1}{n}\ell_f\vartheta)^2\right)m^{t-1}_{f\to v}(\ell_f)\mu_{\Lambda|U}(u_f,\mathrm{d}\ell_f)\right),$$

This equation is exact.

We define several quantities related to the belief propagation iteration.

$$\begin{aligned} \mu^s_{f\to v} &= \int \ell m^s_{f\to v}(\ell)\mu_{\Lambda|U}(u_f,\mathrm{d}\ell), & s^s_{f\to v} &= \int \ell^2 m^s_{f\to v}(\ell)\mu_{\Lambda|U}(u_f,\mathrm{d}\ell), \\ \alpha^{s+1}_{v\to f} &= \frac{1}{n}\sum_{f'\in\partial v\setminus f}\mu^s_{f'\to v}\lambda_{f'}, & (\tau^{s+1}_{v\to f})^2 &= \frac{1}{n}\sum_{f'\in\partial v\setminus f}(\mu^s_{f'\to v})^2, \\ a^s_{v\to f} &= \frac{\mathrm{d}}{\mathrm{d}\vartheta}\log m^s_{v\to f}(\vartheta)\Big|_{\vartheta=0}, & b^s_{v\to f} &= -\frac{\mathrm{d}^2}{\mathrm{d}\vartheta^2}\log m^s_{v\to f}(\vartheta)\Big|_{\vartheta=0}, \\ \mu^s_{v\to f} &= \int \vartheta m^s_{v\to f}(\vartheta)\mu_{\Theta|V}(v_v,\mathrm{d}\vartheta), & s^s_{v\to f} &= \int \vartheta^2 m^s_{v\to f}(\vartheta)\mu_{\Theta|V}(v_v,\mathrm{d}\vartheta), \\ \alpha^s_{f\to v} &= \frac{1}{n}\sum_{v'\in\partial f\setminus v}\mu^s_{v'\to f}\theta_{v'}, & (\hat\tau^s_{f\to v})^2 &= \frac{1}{n}\sum_{v'\in\partial f\setminus v}(\mu^s_{v'\to f})^2, \\ a^s_{f\to v} &= \frac{\mathrm{d}}{\mathrm{d}\ell}\log m^s_{f\to v}(\ell)\Big|_{\ell=0}, & b^s_{f\to v} &= -\frac{\mathrm{d}^2}{\mathrm{d}\ell^2}\log m^s_{f\to v}(\ell)\Big|_{\ell=0}, \end{aligned}$$

Lemma B.4.4 follows from the following asymptotic characterization of the quantities in the preceding display in the limit $n, p \to \infty$, $n/p \to \delta$:

$$\begin{gathered} \mathbb{E}[\mu^s_{f\to v}\lambda_f] \to q_{s+1}, \qquad \mathbb{E}[(\mu^s_{f\to v})^2] \to q_{s+1}, \\ \alpha^{s+1}_{v\to f} \xrightarrow{\mathrm{p}} q_{s+1}, \qquad (\tau^{s+1}_{v\to f}) \xrightarrow{\mathrm{p}} q_{s+1}, \\ (\theta_v, v_v, a^s_{v\to f}, b^s_{v\to f}) \xrightarrow{\mathrm{d}} (\Theta, V, q_s\Theta + q_s^{1/2}G_2, q_s), \\ \mathbb{E}[\mu^s_{v\to f}\theta_v] \to \delta\hat q_s, \qquad \mathbb{E}[(\mu^s_{v\to f})^2] \to \delta\hat q_s, \\ \alpha^s_{f\to v} \xrightarrow{\mathrm{p}} \hat q_s, \qquad (\hat\tau^{s+1}_{f\to v})^2 \xrightarrow{\mathrm{p}} \hat q_s, \\ (\lambda_f, u_f, a^s_{f\to v}, b^s_{f\to v}) \xrightarrow{\mathrm{d}} (\Lambda, U, \hat q_s\Lambda + \hat q_s^{1/2}G, \hat q_s). \end{gathered} \tag{B.30}$$

As in the proof of Lemma B.4.1, the distribution of these quantities does not depend upon v or f, so that the limits hold for all v, f once we establish them for any v, f. We establish the limits inductively in s.

Base case: $\mathbb{E}[\mu^0_{f\to v}\lambda_f] \to q_1$ *and* $\mathbb{E}[(\mu^0_{f\to v})^2] \to q_1$*.*

Note $\mu^0_{f\to v} = \mathbb{E}[\lambda_f | u_f]$. Thus $\mathbb{E}[\mu^0_{f\to v}\lambda_f] = \mathbb{E}[\mathbb{E}[\lambda_f|u_f]^2] = V_{\Lambda,U}(0) = q_1$ exactly in finite samples, so also asymptotically. The expectation $\mathbb{E}[(\mu^0_{f\to v})^2]$ has the same value.

Inductive step 1: If $\mathbb{E}[\mu^s_{f\to v}\lambda_f] \to q_{s+1}$ *and* $\mathbb{E}[(\mu^s_{f\to v})^2] \to q_{s+1}$*, then* $\alpha^{s+1}_{v\to f} \xrightarrow{\mathrm{p}} q_{s+1}$ *and* $(\tau^{s+1}_{v\to f})^2 \xrightarrow{\mathrm{p}} q_{s+1}$*.*

By the inductive hypothesis, $\mathbb{E}[\alpha^{s+1}_{v\to f}] = (n-1)\mathbb{E}[\mu^s_{f\to v}\lambda_f]/n \to q_{s+1}$ and $\mathbb{E}[(\tau^{s+1}_{v\to f})^2] = (n-1)\mathbb{E}[(\mu^s_{f\to v})^2]/n \to q_{s+1}$. Moreover, $\mu^s_{f'\to v}\lambda_{f'}$ are mutually independent as we vary $f' \in \partial v \setminus f$, and likewise for $\mu^s_{f'\to v}$. We have $\mathbb{E}[(\mu^s_{f'\to v}\lambda_{f'})^2] \le M^4$ and $\mathbb{E}[(\mu^s_{f'\to v})^4] \le M^4$ because the integrands are bounded by M^4. By the weak law of large numbers, $\alpha^{s+1}_{v\to f} \xrightarrow{\mathrm{p}} q_{s+1}$ and $(\tau^{s+1}_{v\to f})^2 \xrightarrow{\mathrm{p}} q_{s+1}$.

Inductive step 2: If $\alpha^{s+1}_{v\to f} \xrightarrow{\mathrm{p}} q_{s+1}$ *and* $(\tau^{s+1}_{v\to f})^2 \xrightarrow{\mathrm{p}} q_{s+1}$*, then* $(\theta_v, v_v, a^{s+1}_{v\to f}, b^{s+1}_{v\to f}) \xrightarrow{\mathrm{d}} (\Theta, V, q_{s+1}\Theta + q^{1/2}_{s+1}G, q_{s+1})$*.*

We may express

$$\log m^{s+1}_{v\to f}(\vartheta) = \mathsf{const} + \sum_{f'\in\partial v\setminus f} \log \mathbb{E}_{\Lambda_{f'}}\left[\exp\left(-\frac{1}{2n}\Lambda^2_{f'}\vartheta^2 + x_{f'v}\Lambda_{f'}\vartheta\right)\right],$$

where $\Lambda_{f'}$ has density $m^s_{f'\to v}$ with respect to $\mu_{\Lambda|U}(u_{f'},\cdot)$. We compute

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\vartheta}\mathbb{E}_{\Lambda_{f'}}\left[\exp\left(-\frac{1}{2n}\Lambda^2_{f'}\vartheta^2 + x_{f'v}\Lambda_{f'}\vartheta\right)\right]\Big|_{\vartheta=0} &= \mathbb{E}_{\Lambda_{f'}}\left[x_{f'v}\Lambda_{f'}\right] = x_{f'v}\mu^s_{f'\to v},\\
\frac{\mathrm{d}^2}{\mathrm{d}\vartheta^2}\mathbb{E}_{\Lambda_{f'}}\left[\exp\left(-\frac{1}{2n}\Lambda^2_{f'}\vartheta^2 + x_{f'v}\Lambda_{f'}\vartheta\right)\right]\Big|_{\vartheta=0} &= \mathbb{E}_{\Lambda_{f'}}\left[x^2_{f'v}\Lambda^2_{f'} - \frac{1}{n}\Lambda^2_{f'}\right] = \left(x^2_{f'v} - \frac{1}{n}\right)s^s_{f'\to v}.
\end{aligned}$$

Then

$$a^{s+1}_{v\to f} = \sum_{f'\in\partial v\setminus f} x_{f'v}\mu^s_{f'\to v} \quad\text{and}\quad b^{s+1}_{v\to f} = \sum_{f'\in\partial v\setminus f'}\left(x^2_{f'v}(\mu^s_{f'\to v})^2 - \left(x^2_{f'v} - \frac{1}{n}\right)s^s_{f'\to v}\right).$$

We compute

$$a^{s+1}_{v\to f} = \left(\frac{1}{n}\sum_{f'\in\partial v\setminus f}\mu^s_{f'\to v}\lambda_{f'}\right)\theta_v + \sum_{f'\in\partial v\setminus f} z_{f'v}\mu^s_{f'\to v}.$$

Because $(z_{f'v})_{f'\in\partial v\setminus f}$ are independent of $\mu^2_{f'\to v}$ and are mutually independent from each other, conditional on $\mathcal{T}^1_{v\to f}$ the quantity $\sum_{f'\in\partial v\setminus f} z_{f'v}\mu^s_{f'\to v}$ is distributed $\mathsf{N}(0,(\tau^{s+1}_{v\to f})^2)$. By the inductive hypothesis, $(\tau^{s+1}_{f\to v})^2 \xrightarrow{\mathrm{p}} q_{s+1}$, so that $\sum_{f'\in\partial v\setminus f} z_{f'v}\mu^s_{f'\to v} \xrightarrow{\mathrm{d}} \mathsf{N}(0, q_{s+1})$. Further, $z_{f'v}$ and $\mu^s_{f'\to v}$ are independent of θ_v, and by the inductive hypothesis, the coefficient of θ_v converges in probability to q_{s+1}. By the Continuous Mapping Theorem [188, Theorem 2.3], we conclude that $(\theta_v, v_v, a^{s+1}_{v\to f}) \xrightarrow{\mathrm{d}} (\Theta, V, q_{s+1}\Theta + q^{1/2}_{s+1}G)$ where $G \sim \mathsf{N}(0,1)$ independent of Θ, as desired.

Now we show that $b^{s+1}_{v\to f} \xrightarrow{\mathrm{d}} q_{s+1}$. We expand $b^{s+1}_{v\to f} = A - B$ where $A = \sum_{f'\in\partial v\setminus f} x^2_{f'v}(\mu^s_{f'\to v})^2$ and

$B = \sum_{f'\in\partial v\setminus f}(x_{f'v}^2 - 1/n)s_{f'\to v}^s$. We have

$$A = \frac{1}{n^2}\sum_{v'\in\partial f\setminus v}\lambda_{f'}^2\theta_v^2(\mu_{f'\to v}^s)^2 + \frac{2}{n}\sum_{f'\in\partial v\setminus f}\lambda_{f'}\theta_v z_{f'v}(\mu_{f'\to v}^s)^2 + \sum_{v'\in\partial f\setminus v} z_{f'v}^2(\mu_{f'\to v}^s)^2.$$

Observe $\mathbb{E}[\lambda_{f'}^2\theta_v^2(\mu_{f'\to v}^s)^2] \le M^6$, so that the expectation of the first term is bounded by $M^6(p-1)/n^2 \to 0$. Thus, the first term converges to 0 in probability. Because $z_{f'v}$ is independent of $\mu_{f'\to v}^s$, $\mathbb{E}[|\lambda_{f'}\theta_v z_{f'v}(\mu_{f'\to v}^s)^2|] \le M^4\sqrt{2/(\pi n)}$, so that the absolute value of the expectation of the second term is bounded by $2M^4\sqrt{2/(\pi n)} \to 0$. Thus, the second term converges to 0 in probability. Because $\mu_{f'\to v}^s$ is independent of $z_{f'v}$, the expectation of the last term is $(n-1)\mathbb{E}[(\mu_{f'\to v})^2]/n \to q_{s+1}$ (we have used here the assumption of inductive step 1). The terms $(z_{f'v}^2(\mu_{f'\to v}^s)^2)_{f'\in\partial v\setminus f}$ are mutually independent and $\mathbb{E}[z_{f'v}^4(\mu_{f'\to v}^s)^4] \le 3M^4/n^2$, so that by the weak law of large numbers we have that the last term converges to q_{s+1} in probability. Thus, $A \xrightarrow{\text{p}} q_{s+1}$.

We have

$$B = \frac{1}{n^2}\sum_{v'\in\partial f\setminus v}\lambda_f^2\theta_{v'}^2 s_{v'\to f}^s + \frac{2}{n}\sum_{v'\in\partial f\setminus v}\lambda_f\theta_{v'}s_{v'\to f}^s + \sum_{v'\in\partial f\setminus v}(z_{f'v}^2 - 1/n)s_{v'\to f}^s.$$

As in the analysis of the first two terms of A, we may use that $s_{v'\to f}^s \le M^2$ to argue that the first two terms of B converge to 0 in probability. Further, because $z_{f'v}$ is independent of $s_{v'\to f}^s$, the expectation of the last term is 0. Further, $\mathbb{E}[(z_{f'v}^2 - 1/n)^2(s_{v'\to f}^s)^2] \le 2\mathbb{E}[(z_{f'v}^4 + 1/n^2)]\mathbb{E}[(s_{v'\to f}^s)^2] \le 8M^4/n^2$, so that by the weak law of large numbers, the final term converges to 0 in probability. Thus, $B \xrightarrow{\text{p}} 0$. Because, as we have shown, $A \xrightarrow{\text{p}} q_{s+1}$, we conclude $b_{v\to f}^{s+1} \xrightarrow{\text{p}} q_{s+1}$.

Combining with $(\theta_v, v_v, a_{v\to f}^{s+1}) \xrightarrow{\text{d}} (\Theta, V, q_{s+1}\Theta + q_{s+1}^{1/2}G)$ and applying the Continuous Mapping Theorem [188, Theorem 2.3], we have $(\theta_v, a_{v\to f}^{s+1}, b_{v\to f}^{s+1}) \xrightarrow{\text{d}} (\Theta, q_{s+1}\Theta + q_{s+1}^{1/2}G, q_{s+1})$.

Inductive step 3: If $(\theta_v, v_v, a_{v\to f}^s, b_{v\to f}^s) \xrightarrow{\text{d}} (\Theta, V, q_s\Theta + q_s^{1/2}G_1, q_s)$*, then* $\mathbb{E}[\mu_{v\to f}^s\theta_v] \to \delta\hat{q}_s$ *and* $\mathbb{E}[(\mu_{v\to f}^s)^2] \to \delta\hat{q}_s$*.*

We will require the following lemma, whose proof is deferred to section B.4.2.

Lemma B.4.5. *For any fixed* s*, we have* $\vartheta, \ell \in [-M, M]$

$$\log\frac{m_{v\to f}^s(\vartheta)}{m_{v\to f}^s(0)} = \vartheta a_{v\to f}^s - \frac{1}{2}\vartheta^2 b_{v\to f}^s + O_p(n^{-1/2}),$$
$$\log\frac{m_{f\to v}^s(\ell)}{m_{f\to v}^s(0)} = \ell a_{f\to v}^s - \frac{1}{2}\ell^2 b_{f\to v}^s + O_p(n^{-1/2}),$$

where $O_p(n^{-1/2})$ *has no* ϑ *(or* ℓ*) dependence.*

Define

$$\epsilon_{f\to v}^s = \sup_{\vartheta\in[-M,M]}\left|\log\frac{m_{v\to f}^s(\vartheta)}{m_{v\to f}^s(0)} - \left(\vartheta a_{v\to f}^s - \frac{1}{2}\vartheta^2 b_{v\to f}^s\right)\right|.$$

By Lemma B.4.5, we have $\epsilon_{v\to f}^s = o_p(1)$. Moreover, using the same argument as in inductive step 4 of the proof of Theorem B.4.1, we have that

$$e^{-2\epsilon_{v\to f}^s}\eta_{\Theta,V}(a_{v\to f}^s(b_{v\to f}^s)^{-1/2}, v_v; b_{v\to f}^s) \le \mu_{v\to f}^s$$

$$\leq e^{2\epsilon^s_{v\to f}}\eta_{\Theta,V}(a^s_{v\to f}(b^s_{v\to f})^{1/2}, v_v; b^s_{v\to f}),$$

where $\eta_{\Theta,V}(y,v;q) = \mathbb{E}_{\Theta,V,G}[\Theta | q^{1/2}\Theta + \tau G = y; V = v]$. Because $\eta_{\Theta,V}$ takes values in the bounded interval $[-M,M]$ and $\epsilon^s_{v\to f} = o_p(1)$ by Lemma B.4.5, we conclude that

$$\mu^s_{v\to f} = \eta_{\Theta,V}(a^s_{v\to f}/b^s_{v\to f}, v_v; b^s_{v\to f}) + o_p(1).$$

For a fixed v_v, the Bayes estimator in the observation and coefficient q. Thus, by the inductive hypothesis and the fact that $v_v \sim \mu_V$ for all n, we have that $\mathbb{E}[\Theta\eta_{\Theta,V}(a^s_{v\to f}(b^s_{v\to f})^{1/2}, v_v; b^s_{v\to f})]$ has limit $\mathbb{E}[\Theta\eta_{\Theta,V}(q_s^{1/2}\Theta + G, V; q_s)] = \delta\hat{q}_s$ and $\mathbb{E}[\eta_{\Theta,V}(q_s^{1/2}\Theta + G, V; q_s)^2]$ has limit $\mathbb{E}_{\Theta,V,G}[\eta_{\Theta,V}(q_s^{1/2}\Theta + G, V; q_s)^2] = \delta\hat{q}_s$. Because $|\theta_v|, |\mu^s_{v\to f}|, |\eta_{\Theta,V}(a^s_{v\to f}/b^s_{v\to f}, v_v; b^s_{v\to f})| \leq M$, by bounded convergence, we conclude $\mathbb{E}[\mu^s_{v\to f}\theta_v] \to \delta\hat{q}_s$ and $\mathbb{E}[(\mu^s_{f\to v})^2] \to \delta\hat{q}_s$.

The remaining inductive steps are completely analagous to those already shown. We list them here for completeness.

Inductive step 4: If $\mathbb{E}[\mu^s_{v\to f}\theta_v] \to \delta\hat{q}_s$ *and* $\mathbb{E}[(\mu^s_{v\to f})^2] \to \delta\hat{q}_s$*, then* $\alpha^s_{f\to v} \xrightarrow{\mathrm{p}} \hat{q}_s$ *and* $(\hat{\tau}^{s+1}_{f\to v})^2 \xrightarrow{\mathrm{p}} \hat{q}_s$*.*

Inductive step 5: If $\alpha^s_{f\to v} \xrightarrow{\mathrm{p}} \hat{q}_s$ *and* $(\hat{\tau}^s_{f\to v})^2 \xrightarrow{\mathrm{p}} \hat{q}_s$*, then* $(\lambda_f, u_f, a^s_{f\to v}, b^s_{f\to v}) \xrightarrow{\mathrm{d}} (\Lambda, U, \hat{q}_s\Lambda + \hat{q}_s^{1/2}G, \hat{q}_s)$*.*

Inductive step 6: If $(\lambda_f, u_f, a^s_{f\to v}, b^s_{f\to v}) \xrightarrow{\mathrm{d}} (\Lambda, U, \hat{q}_s\Lambda + \hat{q}_s^{1/2}G, \hat{q}_s)$*, then* $\mathbb{E}[\mu^s_{f\to v}\lambda_f] \to q_{s+1}$ *and* $\mathbb{E}[(\mu^s_{f\to v})^2] \to q_{s+1}$*.*

The induction is complete, and we conclude (B.30).

To complete the proof of Lemma B.4.4, first observe that we may express $\log\frac{p_v(\vartheta|\mathcal{T}_{v,2t-1})}{p_v(0|\mathcal{T}_{v,2t-1})}$ as $\log\frac{m^t_{v\to f}(\vartheta)}{m^t_{v\to f}(\vartheta)} + \log\mathbb{E}_{\Lambda_f}[\exp(\vartheta x_{fv}\Lambda_f - \vartheta^2\Lambda_f^2/(2n))]$. Note that

$$\left|\log\mathbb{E}_{\Lambda_f}[\exp(\vartheta x_{fv}\Lambda_f - \vartheta^2\Lambda_f^2/(2n))]\right| \leq M^2|x_{fv}| + M^4/2n = o_p(1).$$

By Lemma B.4.5, we have that, up to a constant, $\log\frac{m^t_{v\to f}(\vartheta)}{m^t_{v\to f}(\vartheta)} = -\frac{1}{2}((a^t_{v\to f}(b^t_{v\to f})^{-1/2} - b^t_{v\to f})^{1/2}\vartheta)^2 + o_p(1)$. The lemma follows from (B.30) and Slutsky's theorem.

□

Lemma 3.5.3 in the low-rank matrix estimation model follows from Lemma B.4.4 by exactly the same argument that derived Lemma 3.5.3 in the high-dimensional regression model from Lemma B.4.1.

Technical tools

Proof.[Lemma B.4.5] Fix any $\vartheta \in [-M,M]$. By Taylor's theorem, there exist $\vartheta_{f'} \in [-M,M]$ (in fact, between 0 and ϑ) such that

$$\begin{aligned}\log\frac{m^s_{v\to f}(\vartheta)}{m^s_{v\to f}(0)} &= \sum_{f'\in\partial v\setminus f}\log\frac{\mathbb{E}_{\Lambda_{f'}}[\exp(-n(x_{f'v} - \Lambda_{f'}\vartheta/n)^2/2)]}{\mathbb{E}_{\Lambda_{f'}}[\exp(-nx^2_{f'v}/2)]}\\ &= \vartheta a^{s+1}_{v\to f} - \frac{1}{2}\vartheta^2 b^{s+1}_{v\to f} + \frac{1}{6}\vartheta^3\sum_{f'\in\partial v\setminus f}\frac{\mathrm{d}^3}{\mathrm{d}\vartheta^3}\log\mathbb{E}_{\Lambda_{f'}}[\exp(-n(x_{f'v} - \Lambda_{f'}\vartheta/n)^2/2)]\Big|_{\vartheta=\vartheta_{f'}},\end{aligned}$$

where it is understood that $\Lambda_{f'} \sim \mu_{\Lambda|U}(u_{f'}, \cdot)$. Denote $\psi(\vartheta, \ell, x) = -n(x_{f'v} - \ell\vartheta/n)^2/2$. By the same argument that allowed us to derive (B.26) from R4 in the proof of Lemma 3.5.3(a), we conclude

$$\begin{aligned}
&\frac{\mathrm{d}^3}{\mathrm{d}\vartheta^3} \log \mathbb{E}_\Lambda[\exp(\psi(\vartheta, \Lambda, x))]\Big|_{\vartheta=\vartheta_{f'}} \\
&\qquad \le C \sup_{\ell,\vartheta \in [-M,M]} \max\{|\partial_\vartheta \psi(\vartheta, \ell, x)|^3, |\partial_\vartheta \psi(\vartheta, \ell, x)\partial_\vartheta^2 \psi(\vartheta, \ell, x)|, |\partial_\vartheta^3 \psi(\vartheta, \ell, x)|\} \\
&\qquad \le C \max\left\{M^3|M^2/n + x_{f'v}|^3, (M^2/n)M|M^2/n + x_{f'v}|, 0\right\},
\end{aligned}$$

where C is a universal constant. The expectaton of the right-hand side is $O(n^{-3/2})$, whence we get

$$\frac{1}{6}\vartheta^3 \sum_{f' \in \partial v \setminus f} \frac{\mathrm{d}^3}{\mathrm{d}\vartheta^3} \log \mathbb{E}_{\Lambda_{f'}}[\exp(-n(x_{f'v} - \Lambda_{f'}\vartheta/n)^2/2)]\Big|_{\vartheta=\vartheta_{f'}} = O_p(n^{-1/2}),$$

where because $\vartheta \in [-M, M]$, we may take $O_p(n^{-1/2})$ to have no ϑ-dependence.

The expansion of $\log \frac{m^s_{f\to v}(\ell)}{m^s_{f\to v}(0)}$ is proved similarly.

□

B.5 Weakening the assumptions

Section 3.5 and the preceding appendices establish under the assumptions A1, A2 and either R3, R4 or M2 all claims in Theorems 3.3.1 and 3.3.2 except that the lower bound may be achieved. In this section we show that if these claims hold under assumptions A1, A2, R3, R4, then they also hold under assumptions A1, A2, R1, R2 in the high-dimensional regression model; and similarly for the low-rank matrix estimation model. In the next section we prove we can achieve the lower bounds under the weaker assumptions A1, A2 and either R1, R2 or M1.

B.5.1 From strong to weak assumptions in the high-dimensional regression model

To prove the reduction from the stronger assumptions in the high-dimensional regression model, we need the following lemma, whose proof is given at the end of this section.

Lemma B.5.1. *Consider on a single probability space random variables $A, B, (B_n)_{n\ge1}$, and $Z \sim \mathsf{N}(0,1)$ independent of the A's and B's, all with finite second moment. Assume $\mathbb{E}[(B - B_n)^2] \to 0$. Let $Y = B + \tau Z$ and $Y_n = B_n + \tau Z$ for $\tau > 0$. Then*

$$\mathbb{E}[\mathbb{E}[A|Y_n]^2] \to \mathbb{E}[\mathbb{E}[A|Y]^2].$$

We now establish the reduction.

Consider $\mu_{W,U}$, $\mu_{\Theta,V}$, and h satisfying R1 and R2. For any $\epsilon > 0$, we construct $\mu_{\tilde{\boldsymbol{W}},\tilde{U}}$, $\mu_{\tilde{\Theta},\tilde{V}}$, and $\tilde{h}$ satisfying R3 and R4 for $k = 3$ as well as data $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}, \boldsymbol{v}, \tilde{\boldsymbol{v}} \in \mathbb{R}^p$, and $\boldsymbol{y}, \tilde{\boldsymbol{y}}, \boldsymbol{w}, \boldsymbol{u}, \tilde{\boldsymbol{u}} \in \mathbb{R}^n$ and $\tilde{\boldsymbol{w}} \in \mathbb{R}^{n\times 3}$ such that the following all hold.

1. $(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{v}, \boldsymbol{u}, \boldsymbol{w}, \boldsymbol{y})$ and $(\boldsymbol{X}, \tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{v}}, \tilde{\boldsymbol{u}}, \tilde{\boldsymbol{w}}, \tilde{\boldsymbol{y}})$ are generated according to their respective regression models: namely, $(\theta_j, v_j) \overset{\text{iid}}{\sim} \mu_{\Theta,V}$ and $(w_i, u_i) \overset{\text{iid}}{\sim} \mu_{W,U}$ independent; $(\tilde{\theta}_j, \tilde{v}_j) \overset{\text{iid}}{\sim} \mu_{\tilde{\Theta},\tilde{V}}$ and $(\tilde{\boldsymbol{w}}_i, \tilde{\boldsymbol{u}}_i) \overset{\text{iid}}{\sim} \mu_{\tilde{\boldsymbol{W}},\tilde{U}}$ independent; $x_{ij} \overset{\text{iid}}{\sim} \mathsf{N}(0, 1/n)$ independent of everything else; and $\boldsymbol{y} = h(\boldsymbol{X\theta}, \boldsymbol{w})$ and $\tilde{\boldsymbol{y}} = \tilde{h}(\boldsymbol{X}\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{v}})$. Here $\tilde{\boldsymbol{w}}_i^{\mathsf{T}}$ is the i^{th} row of $\tilde{\boldsymbol{w}}$. We emphasize that the data from the two models are not independent.

2. We have

$$\mathbb{P}\left(\frac{1}{n}\|\boldsymbol{y} - \tilde{\boldsymbol{y}}\|^2 > \epsilon\right) \to 0,\ \mathbb{P}\left(\frac{1}{p}\|\boldsymbol{v} - \tilde{\boldsymbol{v}}\|^2 > \epsilon\right) \to 0,\ \mathbb{P}\left(\frac{1}{n}\|\boldsymbol{u} - \tilde{\boldsymbol{u}}\|^2 > \epsilon\right) \to 0\,. \tag{B.31}$$

   Note that because in any GFOM the functions $F_t^{(1)}, F_t^{(2)}, G_t^{(1)}, G_t^{(2)}, G_*$ are Lipschitz and $\|\boldsymbol{X}\|_{\mathsf{op}} \xrightarrow{\text{p}} C_\delta < \infty$ as $n, p \to \infty, n/p \to 0$ [191, Theorem 5.31], the previous display and the iteration (3.1) imply

$$\mathbb{P}\left(\frac{1}{p}\|\hat{\boldsymbol{\theta}}^t - \tilde{\hat{\boldsymbol{\theta}}}^t\|^2 > c(\epsilon, t)\right) \to 0\,, \tag{B.32}$$

   for some $c(\epsilon, t) < \infty$ which goes to 0 as $\epsilon \to 0$ for fixed t.

3. We have

$$|\mathsf{mmse}_{\Theta,V}(\tau_s^2) - \mathsf{mmse}_{\tilde{\Theta},\tilde{V}}(\tau_s)^2| < \epsilon, \tag{B.33}$$

$$\left|\mathbb{E}\left[\mathbb{E}[G_1|h(G,W) + \epsilon^{1/2}Z, G_0]^2\right] - \mathbb{E}\left[\mathbb{E}[G_1|\tilde{h}(G,\tilde{\boldsymbol{W}}), G_0]^2\right]\right| < \tilde{\tau}_s^2\epsilon\,, \tag{B.34}$$

   for all $s \le t$ where $G_0, G_1, Z \overset{\text{iid}}{\sim} \mathsf{N}(0,1)$, $W \sim \mu_W$, and $\tilde{\boldsymbol{W}} \sim \mu_{\tilde{\boldsymbol{W}}}$ independent, and $G = \sigma_s G_0 + \tilde{\tau}_s G_1$.

We now describe the construction described and prove it has the desired properties. Let μ_A be a smoothed Laplace distribution with mean zero and variance 1; namely, μ_A has a C_∞ positive density $p_A(\cdot)$ with respect to Lebesgue measure which satisfies $\partial_a \log p_A(a) = c \cdot \mathsf{sgn}(a)$ when $|x| > 1$ for some positive constant c. This implies that $|\partial_a^k \log p_A(a)| \le q_k$ for all k and some constants q_k, and that μ_A has moments of all orders.

First we construct $\tilde{h}$ and $\tilde{\boldsymbol{W}}$. For a $\xi > 0$ to be chosen, let $\hat{h}$ be a Lipschitz function such that $\mathbb{E}[(\hat{h}(G,W) - h(G,W))^2] < \xi$ for (G,W) as above, which is permitted by assumption R2. Let $L > 0$ be a Lipschitz constant for $\hat{h}$. Choose $M > 0$ such that $\mathbb{E}[W^2\mathbf{1}\{|W| > M\}] < \xi/L^2$. Define $\bar{W} = W\mathbf{1}\{|W| \le M\}$. Note that $\mathbb{E}[(h(G,W) - \hat{h}(G + \xi^{1/2}A, \bar{W}))^2] \le 2\mathbb{E}[(h(G,W) - \hat{h}(G,W))^2] + 2\mathbb{E}[(\hat{h}(G,W) - \hat{h}(G + \xi^{1/2}A, \bar{W}))^2] < 4\xi$. By Lemma B.5.1, we may pick $0 < \xi < \min\{\epsilon/4, \epsilon/L^2\}$ sufficiently small that

$$\left|\mathbb{E}\left[\mathbb{E}[G_1|h(G,W) + \epsilon^{1/2}Z, G_0]^2\right] - \mathbb{E}\left[\mathbb{E}[G_1|\hat{h}(G + \xi^{1/2}A, \bar{W}) + \epsilon^{1/2}Z, G_0]^2\right]\right| < \tilde{\tau}_s^2\epsilon\,.$$

In fact, because t is finite, we may choose $\xi > 0$ small enough that this holds for all $s \le t$. Define $\tilde{\boldsymbol{W}} = (\bar{W}, A, Z)$ and $\tilde{h}(x, \tilde{\boldsymbol{w}}) = \hat{h}(x + \xi^{1/2}a, \bar{w}) + \epsilon^{1/2}z$ where $\tilde{\boldsymbol{w}} = (\bar{w}, a, z)$. Then $\tilde{h}$ is Lipschitz, Eq. (B.34) holds for all $s \le t$, and $\mathbb{E}[(h(G,W) - \tilde{h}(G, \tilde{\boldsymbol{W}}))^2] < \epsilon$ (the last because $\xi < \epsilon/4$).

Now choose $K > 0$ large enough that

$$\mathbb{E}[\Theta^2\mathbf{1}\{|\Theta| > K\}] < \delta\epsilon/L^2,\ \ \mathbb{E}[U^2\mathbf{1}\{|U| > K\}] < \epsilon/2,\ \ \mathbb{E}[V^2\mathbf{1}\{|V| > K\}] < \epsilon/2\,. \tag{B.35}$$

Define $\tilde{\Theta} = \bar{\Theta} = \Theta\mathbf{1}\{|\Theta| \le K\}$, $\tilde{V} = \bar{V} = V\mathbf{1}\{|V| \le K\}$, $\tilde{U} = \bar{U} = U\mathbf{1}\{|U| \le K\}$, and let $\mu_{\tilde{\Theta},\tilde{V}}, \mu_{\tilde{\boldsymbol{W}},\tilde{U}}$

be the corresponding distributions; namely, $\mu_{\tilde{\Theta},\tilde{V}}$ is the distribution of $(\Theta\mathbf{1}\{|\Theta| \le K\}, V\mathbf{1}\{|V| \le K\})$ when $(\Theta, V) \sim \mu_{\Theta,V}$, and $\mu_{\tilde{\boldsymbol{W}},\tilde{U}}$ is the distribution of $(W\mathbf{1}\{|W| \le M\}, A, Z)$ when $(W, U) \sim \mu_{W,U}$ and $(A, Z) \sim \mu_A \otimes \mathsf{N}(0, 1)$ independent. Because the Bayes risk converges as $K \to \infty$ to the Bayes risk with respect to the untruncated prior, we may choose K large enough that also (B.33) holds for these truncated distributions.

The distributions $\mu_{\tilde{\Theta},\tilde{V}}, \mu_{\tilde{\boldsymbol{W}},\tilde{U}}$ satisfy assumption R3. We now show that $\tilde{h}$ and $\tilde{\boldsymbol{W}}$ constructed in this way satisfy assumption R4. The function $\tilde{h}$ is Lipschitz because $\hat{h}$ is Lipschitz. The random variable $\tilde{Y} := \hat{h}(x + \xi^{1/2}A, \bar{W}) + \epsilon^{1/2}Z$ has density with respect to Lebesgue measure given by

$$p(y|x) = \int\int p_{\xi^{1/2}A}(s - x)\, p_{\mathsf{N}(0,\epsilon)}(y - \hat{h}(s, \bar{w}))\mu_{\bar{W}}(\mathrm{d}\bar{w})\mathrm{d}s,$$

where $p_{\mathsf{N}(0,\epsilon)}$ is the density of $\mathsf{N}(0, \epsilon)$ and $p_{\xi^{1/2}A}(s - x)$ the density of $\xi^{1/2}A$ with respect to Lebesgue measure. We have $p(y|x) \le \sup_y p_{\mathsf{N}(0,\epsilon)}(y) = 1/\sqrt{2\pi\epsilon}$, so is bounded, as desired. Moreover

$$\left|\frac{\int\int \partial_x p_{\xi^{1/2}A}(s - x)\, p_{\mathsf{N}(0,\epsilon)}(y - \hat{h}(s, \bar{w}))\mu_{\bar{W}}(\mathrm{d}\bar{w})\mathrm{d}s}{p(y|x)}\right| \le \sup_s \left|\frac{\dot{p}_{\xi^{1/2}A}(s)}{p_{\xi^{1/2}A}(s)}\right|.$$

Because A has a smoothed Laplace distribution, the right-hand side is finite. Thus, by bounded convergence, we may exchange differentiation and integration and the preceding display is equal to $\partial_x \log p(y|x)$. We conclude that $|\partial_x \log p(y|x)|$ is bounded. The boundededness of all higher derivatives holds similarly. Thus, R4 holds.

We now generate the appropriate joint distribution over $(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{v}, \boldsymbol{u}, \boldsymbol{w}, \boldsymbol{y})$ and $(\boldsymbol{X}, \tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{v}}, \tilde{\boldsymbol{u}}, \tilde{\boldsymbol{w}}, \tilde{\boldsymbol{y}})$. First, generate $(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{v}, \boldsymbol{u}, \boldsymbol{w}, \boldsymbol{y})$ from original the high-dimensional regression model. Then generate $\boldsymbol{a}, \boldsymbol{z}$ independent and with entries $a_i \overset{\text{iid}}{\sim} \mu_A$ and $z_i \overset{\text{iid}}{\sim} \mathsf{N}(0, 1)$. Define $\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{v}}, \tilde{\boldsymbol{u}}$ by truncating $\boldsymbol{\theta}, \boldsymbol{v}, \boldsymbol{u}$ at threshold K; define $\tilde{\boldsymbol{w}}$ by truncating $\boldsymbol{w}$ at threshold M to form $\bar{\boldsymbol{w}}$ and concatenating to it the vectors $\boldsymbol{a}, \boldsymbol{z}$ to form a matrix in $\mathbb{R}^{n\times 3}$; and define $\tilde{\boldsymbol{y}} = \tilde{h}(\boldsymbol{X}\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{w}})$.

All that remains is to show (B.31) holds for the model generated in this way. The bounds on $\|\boldsymbol{v} - \tilde{\boldsymbol{v}}\|^2$ and $\|\boldsymbol{u} - \tilde{\boldsymbol{u}}\|^2$ hold by the weak law of large numbers and (B.35). To control $\|\boldsymbol{y} - \tilde{\boldsymbol{y}}\|$, we bound

$$\begin{aligned}
\|\boldsymbol{y} - \tilde{\boldsymbol{y}}\| &= \|h(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{w}) - \tilde{h}(\boldsymbol{X}\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{w}})\| \\
&\le \|h(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{w}) - \hat{h}(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{w})\| + \|\hat{h}(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{w}) - \hat{h}(\boldsymbol{X}\tilde{\boldsymbol{\theta}}, \boldsymbol{w})\| + \|\hat{h}(\boldsymbol{X}\tilde{\boldsymbol{\theta}}, \boldsymbol{w}) - \tilde{h}(\boldsymbol{X}\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{w}}) \\
&\le \|h(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{w}) - \hat{h}(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{w})\| + L\|\boldsymbol{X}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})\| + L\xi^{1/2}\|\boldsymbol{a}\| + L\|\boldsymbol{w} - \bar{\boldsymbol{w}}\| + \epsilon^{1/2}\|\boldsymbol{z}\|\,.
\end{aligned}$$

Because $|h(x, w)| \le C(1 + |x| + |w|)$ by R2 and $\hat{h}$ is Lipschitz, there exist $C > 0$ such that $|h(x, w) - \hat{h}(x, w)| \le C(1 + |x| + |w|)$. Then, $\mathbb{E}[(h(\tau Z, w) - \hat{h}(\tau Z, w))^2] = \int (h(x, w) - \hat{h}(x, w))^2 \frac{1}{\sqrt{2\pi\tau}} e^{-\frac{1}{2\tau^2}x^2}\mathrm{d}x < C(1 + \tau^2 + w^2)$ and is continuous in τ^2 for $\tau > 0$ by dominated convergence convergence, and is uniformly continuous for τ bounded away from 0 and infinity and w_i restricted to a compact set. Because $\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\theta}|\boldsymbol{\theta} \sim \mathsf{N}(0, \|\boldsymbol{\theta}\|^2/n)$ and $\|\boldsymbol{\theta}\|^2/n \xrightarrow{\text{p}} \tau_\Theta^2/\delta$, we have that

$$\mathbb{E}[(h(\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\theta}, w_i) - \hat{h}(\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\theta}, w_i))^2|\boldsymbol{\theta}, w_i] = \mathbb{E}[(h(\tau_\Theta\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\theta}/\|\boldsymbol{\theta}\|, w_i) - \hat{h}(\tau_\Theta\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\theta}/\|\boldsymbol{\theta}\|, w_i))^2|\boldsymbol{\theta}, w_i] + o_p(1)\,.$$

The right-hand side is a constant equal to $\mathbb{E}[(h(G, W) - \hat{h}(G, W))^2]$ and the left-hand side is uniformly integrable. Thus,

$$\limsup_{n\to\infty} \mathbb{E}[(h(\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\theta}, w_i) - \hat{h}(\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\theta}, w_i))^2] \le \mathbb{E}[(h(G, w_i) - \hat{h}(G, w_i))^2] < \xi\,.$$

Markov's inequality proves the the first convergence in (B.31) because $\xi < \epsilon$. Further, by the weak law of large numbers

$$\frac{L^2}{n}\|\boldsymbol{X}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})\|^2 \le \frac{L^2\|\boldsymbol{X}\|_{\mathsf{op}}^2}{n}\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|^2 \xrightarrow{\mathrm{p}} L^2 C_\delta \delta^{-1}\mathbb{E}[\Theta^2 \mathbf{1}\{|\Theta| > M\}] < C_\delta \epsilon\,,$$

where C_δ is the constant satisfying $\|\boldsymbol{X}\|_{\mathsf{op}}^2 \xrightarrow{\mathrm{p}} C_\delta$ [191, Theorem 5.31]. Similarly, by the weak law of large numbers

$$\frac{L^2\xi}{n}\|\boldsymbol{a}\|^2 \xrightarrow{\mathrm{p}} L^2\xi < \epsilon, \quad \frac{L^2}{n}\|\boldsymbol{w} - \bar{\boldsymbol{w}}\|^2 \xrightarrow{\mathrm{p}} L^2\mathbb{E}[W^2\mathbf{1}\{|W| > M\}] < \xi < \epsilon, \quad \frac{\epsilon}{n}\|\boldsymbol{z}\|^2 \xrightarrow{\mathrm{p}} \epsilon\,.$$

We conclude that

$$\mathbb{P}\left(\frac{1}{n}\|\boldsymbol{y} - \tilde{\boldsymbol{y}}\|^2 > 5(C_\delta + 4)\epsilon\right) \to 0.$$

Becuse ϵ was arbitrary, we can in fact achieve (B.31) by considering a smaller ϵ (without affecting the validity of (B.33)).

This completes the construction. To summarize, we have two models: the first satisfying R1 and R2, and the second satisfying R3 and R4.

With the construction now complete, we explain why it establishes the reduction. Let $\tau_s^{(\epsilon)}, \tilde{\tau}_s^{(\epsilon)}$ be the state evolution parameters generated by (3.5) with $\mu_{\tilde{\boldsymbol{W}},\tilde{U}}$, $\mu_{\tilde{\Theta},\tilde{V}}$, and $\tilde{h}$ in place of $\mu_{W,U}, \mu_{\Theta,V}$, and h. First, we claim that Eqs. (B.33) and (B.34) imply, by induction, that as $\epsilon \to 0$, we have

$$\tau_t^{(\epsilon)} \to \tau_t.$$

Indeed, to show this, we must only establish that $\mathbb{E}\left[\mathbb{E}[G_1|h(G,W) + \epsilon^{1/2}Z, G_0]^2\right]$ converges to $\mathbb{E}[\mathbb{E}[G_1|h(G,W), G_0]^2]$ as $\epsilon \to 0$. Without loss of generality, we may assume that on the same probability space there exists a Brownian motion $(B_\epsilon)_{\epsilon>0}$ independent of everything else. We see that $\mathbb{E}[G_1|h(G,W) + \epsilon^{1/2}Z, G_0]^2] \overset{\mathrm{d}}{=} \mathbb{E}[G_1|h(G,W) + B_\epsilon, G_0] = \mathbb{E}[G_1|(h(G,W) + B_s)_{s\ge\epsilon}, G_0]$. By Lévy's upward theorem [76, Theorem 5.5.7], we have that $\mathbb{E}[G_1|(h(G,W) + B_s)_{s\ge\epsilon}, G_0]$ converges to $\mathbb{E}[G_1|(h(G,W) + B_s)_{s\ge 0}, G_0] = \mathbb{E}[G_1|h(G,W), G_0]$ almost surely. By uniform integrability, we conclude that $\mathbb{E}[\mathbb{E}[G_1|(h(G,W) + B_s)_{s\ge\epsilon}, G_0]^2] \to \mathbb{E}[\mathbb{E}[G_1|h(G,W), G_0]^2]$, as claimed. Thus, we conclude the previous display.

We now show that as $\epsilon \to 0$, we have

$$\inf_{\hat{\theta}(\cdot)} \mathbb{E}[\ell(\tilde{\Theta}, \hat{\theta}(\tilde{\Theta} + \tau_t^{(\epsilon)}G, V))] \to \inf_{\hat{\theta}(\cdot)} \mathbb{E}[\ell(\Theta, \hat{\theta}(\Theta + \tau_t G, V))]\,.$$

Because the truncation level K can be taken to ∞ as $\epsilon \to 0$, this holds by combining Lemma B.1.5(a) and (c), and specifically, Eqs. (B.5) and (B.7).

Because the lower bound of Theorem 3.3.1 holds under assumptions R3 and R4, which are satisfied by

$\mu_{\tilde{\boldsymbol{W}},\tilde{U}}$, $\mu_{\tilde{\Theta},\tilde{V}}$, and $\tilde{h}$, we conclude that

$$\lim_{n\to\infty}\frac{1}{p}\sum_{j=1}^{p}\ell(\theta_j,\hat{\theta}_j^t) \geq \inf_{\hat{\theta}(\cdot)}\mathbb{E}[\ell(\tilde{\Theta},\hat{\theta}(\tilde{\Theta}+\tau_t^{(\epsilon)}G,V))].$$

Taking $\epsilon \to 0$ and applying (B.32), we conclude that (3.6) holds for $\hat{\boldsymbol{\theta}}^t$, as desired.

The reduction in the high-dimensional regression model is complete.

Proof.[Lemma B.5.1] It is enough to prove the result for $\tau = 1$. Note

$$\mathbb{E}[A|Y=y] = \frac{\int ae^{-(y-b)^2}\mu(\mathrm{d}a,\mathrm{d}b)}{\int e^{-(y-b)^2}\mu(\mathrm{d}a,\mathrm{d}b)}, \qquad \mathbb{E}[A|Y_n=y] = \frac{\int ae^{-(y-b)^2}\mu_n(\mathrm{d}a,\mathrm{d}b)}{\int e^{-(y-b)^2}\mu_n(\mathrm{d}a,\mathrm{d}b)}.$$

Because $\mu_n \overset{\mathrm{W}}{\to} \mu$, we have

$$\frac{\int ae^{-(y-b)^2}\mu_n(\mathrm{d}a,\mathrm{d}b)}{\int e^{-(y-b)^2}\mu_n(\mathrm{d}a,\mathrm{d}b)} \to \frac{\int ae^{-(y-b)^2}\mu(\mathrm{d}a,\mathrm{d}b)}{\int e^{-(y-b)^2}\mu(\mathrm{d}a,\mathrm{d}b)},$$

for all y, and moreover, this convergence is uniform on compact sets. Moreover, one can check that the stated functions are Lipschitz (with uniform Lipschitz constant) in y on compact sets. This implies that $\mathbb{E}[A|Y_n] \to \mathbb{E}[A|Y]$ almost surely. Because the $\mathbb{E}[A|Y_n]^2$ are uniformly integrable, the lemma follows.

□

B.5.2 From strong to weak assumptions in the low-rank matrix estimation model

Consider $\mu_{\boldsymbol{\Lambda},\boldsymbol{U}}, \mu_{\boldsymbol{\Theta},\boldsymbol{V}}$ satisfying $\mathsf{M1}$. Fix $M > 0$. For $(\boldsymbol{\Lambda},\boldsymbol{U}) \sim \mu_{\boldsymbol{\Lambda},\boldsymbol{U}}$, define $\tilde{\Lambda}$ by setting $\tilde{\Lambda}_i = \Lambda_i \mathbf{1}\{|\Lambda_i| \leq M\}$ for $1 \leq i \leq k$. Define $\tilde{\boldsymbol{U}}$ similarly, and let $\mu_{\tilde{\boldsymbol{\Lambda}},\tilde{\boldsymbol{U}}}$ be the distribution of $(\tilde{\boldsymbol{\Lambda}},\tilde{\boldsymbol{U}})$ so constructed. Define $\mu_{\tilde{\boldsymbol{\Theta}},\tilde{\boldsymbol{V}}}$ similarly.

Consider $\{(\boldsymbol{\Lambda}_i,\boldsymbol{u}_i)\}_{i\leq n} \overset{\mathrm{iid}}{\sim} \mu_{\boldsymbol{\Lambda},\boldsymbol{U}}$ and $\{(\boldsymbol{\theta}_j,\boldsymbol{v}_j)\}_{j\leq p} \overset{\mathrm{iid}}{\sim} \mu_{\boldsymbol{\Theta},\boldsymbol{V}}$ and $\boldsymbol{Z} \in \mathbb{R}^{n\times p}$ independent with $z_{ij} \overset{\mathrm{iid}}{\sim} \mathsf{N}(0,1/n)$. Constructe $\tilde{\boldsymbol{\Lambda}}_i, \tilde{\boldsymbol{u}}_i, \tilde{\boldsymbol{\theta}}_j, \tilde{\boldsymbol{v}}_j$ by truncated each coordinate at level M as above. Define $\boldsymbol{X}, \tilde{\boldsymbol{X}} \in \mathbb{R}^{n\times p}$ by $x_{ij} = \frac{1}{n}\boldsymbol{\Lambda}_i^\top\boldsymbol{\theta}_j + z_{ij}$ and $\tilde{z}_{ij} = \frac{1}{n}\tilde{\boldsymbol{\Lambda}}_i^\top\tilde{\boldsymbol{\theta}} + z_{ij}$. As in the previous section, we have for any $\epsilon > 0$ that

$$\mathbb{P}(\|\boldsymbol{X}-\tilde{\boldsymbol{X}}\|_{\mathsf{op}} > \epsilon) \to 0,\ \ \mathbb{P}\left(\frac{1}{p}\|\boldsymbol{v}-\tilde{\boldsymbol{v}}\|^2 > \epsilon\right) \to 0,\ \ \mathbb{P}\left(\frac{1}{p}\|\boldsymbol{u}-\tilde{\boldsymbol{u}}\|^2 > \epsilon\right) \to 0.$$

As in the previous section, this implies that the iterates of the GFOMs before and after the truncation become arbitrarily close with high probability at a fixed iterate t as we take $M \to \infty$.

Further, as $M \to \infty$ we have $\boldsymbol{V}_{\tilde{\boldsymbol{\Theta}},\tilde{\boldsymbol{V}}}(\boldsymbol{Q}) \to \boldsymbol{V}_{\boldsymbol{\Theta},\boldsymbol{V}}(\boldsymbol{Q})$ for all $\boldsymbol{Q}$, and likewise for $\tilde{\boldsymbol{\Lambda}},\tilde{\boldsymbol{U}}$. Further, $\boldsymbol{V}_{\tilde{\boldsymbol{\Theta}},\tilde{\boldsymbol{V}}}(\boldsymbol{Q})$ is jointly continuous in $\boldsymbol{Q}$ and M (where M is implicit in the truncation used to generate $\tilde{\boldsymbol{\Theta}},\tilde{\boldsymbol{V}}$). Thus, as we take $M \to \infty$, the state evolution (3.7) after the truncation converges to the state evolution with no truncation.

The reduction now occurs exactly as in the previous section.

B.6 Achieving the bound

All that remains to prove Theorems 3.3.1 and 3.3.2 under assumptions A1, A2 and either R1, R2 or M1, respectively, is to show that the lower bounds in Eqs. (3.6) and (3.8) can be achieved. In both cases, we can achieve the bound up to tolerance ϵ using a certain AMP algorithm.

B.6.1 Achieving the bound in the high-dimensional regression model

We first derive certain monotonicity properies of the parameters $\tau_s, \sigma_s, \tilde{\tau}_s$ defined in the state evolution recursion (3.5). As we saw in Appendix B.4.1 and in particular, in Lemma B.4.1, the posterior of θ_v on the computation tree given observations in the local neighborhood $T_{v,2s}$ behaves like that from an observation under Gaussian noise with variance τ_s^2. This is made precise in Lemma B.4.1. Moreover, we saw in the same section that a consequence of Lemma B.4.1 is that the asymptotic limiting Bayes risk with respect to loss ℓ for estimation θ_v given observations in $\mathcal{T}_{v,2s}$ is given by the corresponding risk for estimating Θ given $\Theta + \tau_s G$, V with $(\Theta, V) \sim \mu_{\Theta,V}$ and $G \sim \mathsf{N}(0,1)$ independent. In particular, this applies to the minimum mean square error. On the computation tree, minimum mean square error can only decrease as s grows because as s grows we receive strictly more information. If $\mathbb{E}[\mathrm{Var}(\Theta|V)] > 0$, then $\mathsf{mmse}_{\Theta,V}(\tau^2)$ is strictly increasing in τ, so that we conclude that τ_s is non-increasing in s. Thus, by (3.5), we have also $\tilde{\tau}_s$ is non-increasing in s and σ_s is non-decreasing in s. In the complementary case that $\mathbb{E}[\mathrm{Var}(\Theta|V)] = 0$, we compute $\sigma_s^2 = \tau_\Theta^2/\delta$ and $\tilde{\tau}_s^2 = 0$ for all $s \geq 0$, and $\tau_s^2 = 0$ for all $s \geq 1$. Thus, the same monotoncity results hold in this case. These monotonicity results will imply the needed structural properties of the state evolution matrices $(T_{s,s'}), (\Sigma_{s,s'})$ used below.

For all $s \leq t$, define

$$\alpha_s = \frac{1}{\tilde{\tau}_s}\mathbb{E}[\mathbb{E}[G_1|Y, G_0, U]^2], \quad T_{s,t} = \mathbb{E}[\mathbb{E}[G_1|Y, G_0, U]^2], \quad \Sigma_{s,t} = \sigma_t^2,$$

where $Y = h(\sigma_s G_0 + \tilde{\tau}_s G_1, W)$ and $G_0, G_1 \overset{\text{iid}}{\sim} \mathsf{N}(0,1)$ and $W \sim \mu_W$ independent. By the monotoncity properties stated, $(T_{s,t}), (\Sigma_{s,t})$ define positive definite arrays. Define

$$\begin{gathered} f_t(b^t; y, u) = \mathbb{E}[B^0 - B^t | h(B^0, W) = y,\, B^t = b^t,\, U = u]/\tilde{\tau}_t, \\ g_t(a^t; v) = \mathbb{E}[\Theta | V = v,\, \alpha_t \Theta + Z^t = a^t], \end{gathered}$$

where $(\Theta, V) \sim \mu_{\Theta,V})$, $(W, U) \sim \mu_{W,U}$, $(B^0, \dots, B^t) \sim \mathsf{N}(\mathbf{0}, \mathbf{\Sigma}_{[0:t]})$, $(Z^1, \dots, Z^t) \sim \mathsf{N}(\mathbf{0}, \boldsymbol{T}_{[1:t]})$, all independent. With these definitions, $(B^t, B^0 - B^t) \overset{\text{d}}{=} (\sigma_t G_0, \tilde{\tau}_t G_1)$ where $G_0, G_1 \overset{\text{iid}}{\sim} \mathsf{N}(0,1)$. In particular, (B^t) form a backwards Gaussian random walk. We thus compute

$$\begin{aligned} &\mathbb{E}[(B^0 - B^t) f_t(B^t; h(B^0, W), U)]/\tilde{\tau}_t^2 = \mathbb{E}[(\mathbb{E}[B^0 - B^t | Y, B^t, U]/\tilde{\tau}_t)^2]/\tilde{\tau}_t = \alpha_t, \\ &\mathbb{E}[f_s(B^s; h(B^0, W), U) f_t(B^t; h(B^0, W), U)] \\ &\qquad\quad = \mathbb{E}[\mathbb{E}[B^0 - B^s | Y, B^s, U]\mathbb{E}[B^0 - B^t | Y, B^t, U]]/\tilde{\tau}_t^2 \\ &\qquad\quad = \mathbb{E}[(B^0 - B^t)^2 | Y, B^t, U]/\tilde{\tau}_t^2 = T_{s,t}, \\ &\frac{1}{\delta}\mathbb{E}[\Theta g_t(\alpha_t \Theta + Z^t; V)] = \frac{1}{\delta}\mathbb{E}[\mathbb{E}[\Theta | \Theta + Z^t/\alpha_t, V]^2] = \sigma_t^2, \end{aligned}$$

$$\frac{1}{\delta}\mathbb{E}[g_s(\alpha_s\Theta + Z^s; V)g_t(\alpha_t\Theta + Z^t; V)] = \frac{1}{\delta}\mathbb{E}[\mathbb{E}[\Theta|\Theta + Z^t/\alpha_t, V]^2].$$

If f_t, g_t are Lipschitz, then, because h is also Lipschitz, Stein's lemma [184] implies that the first line is equivalent to $\mathbb{E}[\partial_{B^0} f_t(B^t; h(B^0, W), U)] = \alpha_t$. (Here, we have used that $B^0 - B^t$ is independent of B^t). Thus, $(\alpha_s), (T_{s,t}), (\Sigma_{s,t})$ are exactly the state evolution parameters determined by (B.16), and Lemma 3.5.1 implies that AMP with these $(f_s), (g_s)$ achieves the lower bound.

If the f_t, g_t are not Lipschitz, we proceed as follows. Fix $\epsilon > 0$. First, pick Lipschitz $\hat{f}_0$ such that $\mathbb{E}[(\hat{f}_0(B^0, W) - f_0(B^0, W))^2] < \epsilon$, which is possibly because Lipschitz functions are dense in L_2. Define $\hat{\alpha}_0$ and $\hat{T}_{1,1}$ via (B.16) with $\hat{f}_0$ in place of f_0. Note that $\lim_{\epsilon\to 0}\hat{\alpha}_0 = \alpha_0$ and $\lim_{\epsilon \to 0}\hat{T}_{1,1} = T_{1,1}$. Next, pick Lipschitz $\hat{g}_0$ such that $\mathbb{E}[(\hat{g}_0(\hat{\alpha}_0\Theta + \hat{T}_{1,1}^{1/2}G; V) - \mathbb{E}[\Theta|\hat{\alpha}_0 + \Theta + \hat{T}_{1,1}^{1/2}G; V]])^2] < \epsilon$, which is again possibly because Lipschitz functions are dense in L_2. Define $\hat{\Sigma}_{0,1} = \frac{1}{\delta}\mathbb{E}[\Theta\hat{g}_t(\hat{\alpha}\Theta + \hat{T}_{1,1}^{1/2}G; V)]$ and $\hat{\Sigma}_{1,1} = \frac{1}{\delta}\mathbb{E}[\hat{g}_t(\hat{\alpha}\Theta + \hat{T}_{1,1}^{1/2}G; V)^2]$. Because as $\alpha \to \alpha_0$ and $\tau \to T_{0,0}^{1/2}$, we have $\mathbb{E}[\Theta|\alpha\Theta + \tau G; V)] \overset{L_2}{\to} \mathbb{E}[\Theta|\alpha_0\Theta + T_{0,0}^{1/2}G; V)]$, we conclude that as $\epsilon \to 0$ that $\hat{\Sigma}_{0,1} \to \Sigma_{1,1}$ and $\hat{\Sigma}_{1,1} \to \Sigma_{1,1}$. Continuing in this way, we are able to by taking ϵ sufficiently small construct Lipschitz functions $(\hat{f}_t), (\hat{g}_t)$ which track the state evolutoin of the previous paragraph arbitrarily closely up to a fixed time t^*. Thus, we may come arbitrarily close to achieving the lower bound of Theorem 3.3.1.

B.6.2 Achieving the bound in the low-rank matrix estimation model

Let $\boldsymbol{\gamma}_t = \hat{\boldsymbol{Q}}_t$ for $t \geq 0$ and $\boldsymbol{\alpha}_t = \boldsymbol{Q}_t$, $\boldsymbol{\Sigma}_{t,t} = \hat{\boldsymbol{Q}}_t$, $\boldsymbol{T}_{t,t} = \boldsymbol{Q}_t$ for $t \geq 1$. Define

$$\begin{aligned} f_t(\boldsymbol{b}^t; \boldsymbol{u}) &= \mathbb{E}[\boldsymbol{\Lambda}|\boldsymbol{\gamma}_t\boldsymbol{\Lambda} + \boldsymbol{\Sigma}_{t,t}^{1/2}\boldsymbol{G} = \boldsymbol{b}^t; \boldsymbol{U}], \\ g_t(\boldsymbol{a}^t; \boldsymbol{v}) &= \mathbb{E}[\boldsymbol{\Theta}|\boldsymbol{\alpha}_t\boldsymbol{\Theta} + \boldsymbol{T}_{t,t}^{1/2}\boldsymbol{G} = \boldsymbol{a}^t; \boldsymbol{V}]. \end{aligned}$$

We check that the parameters so defined satisfy the AMP state evolution (B.17). Note that by (3.7),

$$\begin{aligned} \boldsymbol{T}_{t+1,t+1} = \boldsymbol{Q}_{t+1} &= \mathbb{E}[\mathbb{E}[\boldsymbol{\Lambda}|\hat{\boldsymbol{Q}}_t^{1/2}\boldsymbol{\Lambda} + \boldsymbol{G}; \boldsymbol{U}]\mathbb{E}[\boldsymbol{\Lambda}|\hat{\boldsymbol{Q}}_t^{1/2}\boldsymbol{\Lambda} + \boldsymbol{G}; \boldsymbol{U}]^\mathsf{T}] \\ &= \mathbb{E}[\mathbb{E}[\boldsymbol{\Lambda}|\hat{\boldsymbol{Q}}_t\boldsymbol{\Lambda} + \hat{\boldsymbol{Q}}_t^{1/2}\boldsymbol{G}; \boldsymbol{U}]\mathbb{E}[\boldsymbol{\Lambda}|\hat{\boldsymbol{Q}}_t\boldsymbol{\Lambda} + \hat{\boldsymbol{Q}}_t^{1/2}\boldsymbol{G}; \boldsymbol{U}]^\mathsf{T}] \\ &= \mathbb{E}[\mathbb{E}[\boldsymbol{\Lambda}|\boldsymbol{\gamma}_t\boldsymbol{\Lambda} + \boldsymbol{\Sigma}_{t,t}^{1/2}\boldsymbol{G}; \boldsymbol{U}]\mathbb{E}[\boldsymbol{\Lambda}|\boldsymbol{\gamma}_t\boldsymbol{\Lambda} + \boldsymbol{\Sigma}_{t,t}^{1/2}\boldsymbol{G}; \boldsymbol{U}]^\mathsf{T}], \\ \boldsymbol{\alpha}_{t+1} &= \mathbb{E}[\mathbb{E}[\boldsymbol{\Lambda}|\hat{\boldsymbol{Q}}_t^{1/2}\boldsymbol{\Lambda} + \boldsymbol{G}; \boldsymbol{U}]\mathbb{E}[\boldsymbol{\Lambda}|\hat{\boldsymbol{Q}}_t^{1/2}\boldsymbol{\Lambda} + \boldsymbol{G}; \boldsymbol{U}]^\mathsf{T}] \\ &= \mathbb{E}[\mathbb{E}[\boldsymbol{\Lambda}|\boldsymbol{\gamma}_t\boldsymbol{\Lambda} + \boldsymbol{\Sigma}_{t,t}^{1/2}\boldsymbol{G}; \boldsymbol{U}]\boldsymbol{\Lambda}^\mathsf{T}] \end{aligned}$$

where $(\boldsymbol{\Theta}, \boldsymbol{V}) \sim \mu_{\boldsymbol{\Theta},\boldsymbol{V}}$ and $(\boldsymbol{\Lambda}, \boldsymbol{U}) \sim \mu_{\boldsymbol{\Lambda},\boldsymbol{U}}$. The state evolution equations (3.7) for $\boldsymbol{\Sigma}_{t,t}$ and $\boldsymbol{\gamma}_t$ hold similarly.

If f_t, g_t so defined are Lipschitz, then $(\alpha_s), (\boldsymbol{T}_{s,t}), (\boldsymbol{\Sigma}_{s,t})$ are exactly the state evolution parameters determined by (B.16), and Lemma 3.5.1 implies that AMP with these $(f_s), (g_s)$ achieves the lower bound. If the f_t, g_t so defined are not Lipschitz, then the same strategy used in the previous section allows us to achieve the lower bound within tolerance $\epsilon > 0$.

B.7 Proofs for sparse phase retrieval and sparse PCA

B.7.1 Proof of Lemma 3.4.1

Note that $\|\overline{\boldsymbol{\theta}}_0\|_2$ is tightly concentrated around $\mu^2\varepsilon$. As a consequence, we can replace the side information $\overline{\boldsymbol{v}}$ by $\boldsymbol{v} = \sqrt{\tilde{\alpha}}\boldsymbol{\theta}_0 + \boldsymbol{g}$. We apply Theorem 3.3.2 with $r = 1$, and loss $\ell_\lambda(\theta, \hat{\theta}) = (\hat{\theta} - \theta_0/\lambda)^2$, where $\lambda \in \mathbb{R}_{\geq 0}$ will be adjusted below. Setting $\boldsymbol{Q}_t = q_t$, $\hat{\boldsymbol{Q}}_t = \hat{q}_t$, we obtain the iteration

$$q_{t+1} = \frac{\hat{q}_t}{1+\hat{q}_t}\,, \quad \hat{q}_t = \frac{1}{\delta}\mathbb{E}\big\{\mathbb{E}[\sqrt{\delta}\Theta_0|(\delta q_t)^{1/2}\Theta_0 + G; V]^2\big\}\,, \tag{B.36}$$

where $\Theta_0 \sim \mu_\theta$,and $V = \sqrt{\delta\tilde{\alpha}} + G'$, $G' \sim \mathsf{N}(0,1)$. Notice that the additional factors $\sqrt{\delta}$ are due to the different normalization of the vector $\boldsymbol{\theta}_0$ with respect to the statement in Theorem 3.3.2. Also note that the second moment of the conditional expectation bove is equal to $\mathbb{E}\big\{\mathbb{E}[\sqrt{\delta}\Theta_0|(\delta(q_t + \tilde{\alpha}))^{1/2}\Theta_0 + G]^2\big\}$ and a simple calculation yields

$$\hat{q}_{t+1} = V_\pm(q_t + \tilde{\alpha})\,, \quad q_t = \frac{\hat{q}_t}{1+\hat{q}_t}\,, \tag{B.37}$$

which is equivalent to Eqs. (3.12), (3.13).

Let $Y = \sqrt{\delta(q_t + \tilde{\alpha})}\Theta_0 + G$, $G \sim \mathsf{N}(0,1)$. Theorem 3.3.2 then yields

$$\frac{1}{p}\|\hat{\boldsymbol{\theta}}^t - \boldsymbol{\theta}_0/\lambda\|_2^2 \geq \inf_{\hat{\theta}(\,\cdot\,)} \mathbb{E}\big\{\big(\hat{\theta}(Y) - \Theta_0/\lambda\big)^2\big\} + o_p(1) \tag{B.38}$$

$$= \frac{1}{\lambda^2}\mathbb{E}\big\{\big(\mathbb{E}(\Theta_0|Y) - \Theta_0\big)^2\big\} + o_p(1)\,. \tag{B.39}$$

In order to prove the upper bound (3.14), it is sufficient to consider $\|\hat{\boldsymbol{\theta}}^t\|_2^2 \leq p$. Then, for any $\lambda \geq 0$,

$$\frac{1}{p}\langle\hat{\boldsymbol{\theta}}^t, \boldsymbol{\theta}_0\rangle \leq \frac{1}{p}\langle\hat{\boldsymbol{\theta}}^t, \boldsymbol{\theta}_0\rangle - \frac{\lambda}{2p}(\|\hat{\boldsymbol{\theta}}^t\|_2^2 - p) \tag{B.40}$$

$$= \frac{\lambda}{2} + \frac{1}{2\lambda p}\|\boldsymbol{\theta}_0\|_2^2 - \frac{\lambda}{2p}\|\hat{\boldsymbol{\theta}}^t - \boldsymbol{\theta}_0/\lambda\|_2^2 \tag{B.41}$$

$$\leq \frac{\lambda}{2} + \frac{1}{2\lambda}\mathbb{E}\{\Theta_0^2\} - \frac{1}{2\lambda}\mathbb{E}\big\{\big(\mathbb{E}(\Theta_0|Y) - \Theta_0\big)^2\big\} + o(1) \tag{B.42}$$

$$\leq \frac{\lambda}{2} + \frac{1}{2\lambda}V_\pm(q_t + \tilde{\alpha}) + o(1)\,. \tag{B.43}$$

The claim follows by choosing $\lambda = V_\pm(q_t + \tilde{\alpha})^{1/2}$, and noting that $\|\boldsymbol{\theta}_0\|_2^2/p \to \mu^2\varepsilon$, almost surely.

B.7.2 Proof of Corollary 3.4.2

Choose $\mu = R/\sqrt{\varepsilon}$, and let $\mu' < \mu$, $\varepsilon' < \varepsilon$, $R' = \mu'\sqrt{\varepsilon'}$. Draw the coordinates of $\boldsymbol{\theta}_0 = \overline{\boldsymbol{\theta}}_0\sqrt{p}$ according to the three points distribution with parameters μ', ε'. Then, with probability one, we have $\overline{\boldsymbol{\theta}}_0 \in \mathscr{T}(\varepsilon, R)$ for all n large enough. Applying Lemma 3.4.1, we get

$$\lim_{n\to\infty} \inf_{\overline{\boldsymbol{\theta}}_0 \in \mathscr{T}(\varepsilon,R)} \mathbb{E}\left\{\frac{\langle\overline{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}^t\rangle}{\|\overline{\boldsymbol{\theta}}_0\|_2\|\hat{\boldsymbol{\theta}}^t\|_2}\right\} \leq \sqrt{\frac{V_\pm(q_t' + \tilde{\alpha}')}{(\mu')^2\varepsilon'}}\,, \tag{B.44}$$

ahere we used dominated convergence to pass from the limit in probability to limit in expectation, and $q_t', \tilde{\alpha}'$ are computed with parameters μ', ε'. By letting $\varepsilon' \to \varepsilon$, $\mu' \to \mu$, and since $\tilde{\alpha}', q_t'$ are continuous in these parameters by an induction argument, Eq. (B.44) also holds with μ', ε', q_t' replaced by μ, ε, q_t:

$$\lim_{n\to\infty} \inf_{\overline{\boldsymbol{\theta}}_0 \in \mathscr{T}(\varepsilon,R)} \mathbb{E}\left\{ \frac{\langle \overline{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}^t \rangle}{\|\overline{\boldsymbol{\theta}}_0\|_2 \|\hat{\boldsymbol{\theta}}^t\|_2} \right\} \le \sqrt{\frac{V_\pm(q_t+\tilde{\alpha})}{\mu^2\varepsilon}}\,, \tag{B.45}$$

Claims (a) and (b) follow by upper bounding the right-hand side of the last equation.

First notice that $V_\pm(q) = \mu^4\varepsilon^2\delta\, q + O(q^2)$ and hence Eqs. (3.12), (3.13) imply that, for any $\eta > 0$ there exists $q_* > 0$ such that, if $q_t + \tilde{\alpha} \le q_*$, then

$$q_{t+1} \le (\mu^4\varepsilon^2\delta + \eta)(q_t + \tilde{\alpha})\,. \tag{B.46}$$

If $\mu^4\varepsilon^2\delta < 1$, choosing $\eta = (1-\mu^4\varepsilon^2\delta)/2$, this inequality implies $q_t \le 2\tilde{\alpha}/(1-\mu^4\varepsilon^2\delta)$, which proves claim (a).

For the second claim, we use the bounds $e^{-\delta q\mu^2/2}\cosh(\mu\sqrt{\delta q}G) \ge 0$ and $x/(1+x) \le x$ in Eq. (3.13) to get $q_t \le \overline{q}_t$ for all t, where $\overline{q}_0 = 0$ and

$$\overline{q}_{t+1} = F_0(\overline{q}_t + \tilde{\alpha})\,, \qquad F_0(q) := \frac{\mu^2\varepsilon^2}{1-\varepsilon}\sinh(\mu^2\delta q)\,. \tag{B.47}$$

Further Eq. (B.45) implies

$$\lim_{n\to\infty} \inf_{\overline{\boldsymbol{\theta}}_0 \in \mathscr{T}(\varepsilon,R)} \mathbb{E}\left\{ \frac{\langle \overline{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}^t \rangle}{\|\overline{\boldsymbol{\theta}}_0\|_2 \|\hat{\boldsymbol{\theta}}^t\|_2} \right\} \le \sqrt{\frac{\overline{q}_{t+1}}{\mu^2\varepsilon}}\,. \tag{B.48}$$

Define $x_t := \mu^2\delta\overline{q}_t$, $a := \mu^4\varepsilon^2\delta/(1-\varepsilon)$, $b := \mu^2\delta\tilde{\alpha} = (\delta/\varepsilon)(\alpha/(1-\alpha))$. Then x_t obeys the recursion

$$x_{t+1} = a\sinh(x_t + b)\,. \tag{B.49}$$

Since $a = R^4\delta/(1-\varepsilon)$, we know that $a < 1/4$. Using the fact that $\sinh(u) \le 2u$ for $u \le 1$, this implies $x_t \le b$ for all t provided $b < 1/2$. Subsitiuting this bound in Eq. (B.48), we obtain the desired claim.

B.7.3 Proof of Corollary 3.4.1

Consider first the case of a random vector $\boldsymbol{\theta}_0$ with i.i.d. entries $\theta_{0,i} \sim \mu_\theta$. Define, for $\Theta_0 \sim \mu_\theta$,

$$F_\varepsilon(q) := \mathbb{E}\big\{\mathbb{E}[\Theta_0|\sqrt{q}\Theta_0 + G]^2\big\} \tag{B.50}$$

$$= e^{-q\mu^2}\mu^2\varepsilon^2\mathbb{E}\left\{\frac{\sinh(\mu\sqrt{q}G)^2}{1-\varepsilon+\varepsilon e^{-q\mu^2/2}\cosh(\mu\sqrt{q}G)}\right\}\,. \tag{B.51}$$

Setting $q_t = \tau_t^{-2}$, $\hat{q}_t = \sigma_t^2$, and $\tilde{\alpha} = \alpha/(1-\alpha)$, and referring to Lemma B.1.4, the state evolution recursion (3.5) takes the form

$$\hat{q}_t = F_\varepsilon(q_t + \tilde{\alpha})\,, \quad q_{t+1} = \delta\, H(\hat{q}_t)\,, \tag{B.52}$$

$$H(q) := \mathbb{E}_{G_0,Y}\left[\left(\frac{\mathbb{E}_{G_1}\partial_x p(Y|\sqrt{q}\,G_0+\sqrt{1-q}G_1)}{\mathbb{E}_{G_1}p(Y|\sqrt{q}\,G_0+\sqrt{1-q}G_1}\right)^2\right]. \tag{B.53}$$

Notice the change in factors δ with respect to Eq. (3.5), which is due to the different normalization of the design matrix.

By the same argument used in the proof of Lemma 3.4.1, Theorem 3.3.1 implies that, for any GFOM, with output $\hat{\boldsymbol{\theta}}_t$, we have

$$\lim_{n,p\to\infty}\mathbb{E}\frac{\langle\overline{\boldsymbol{\theta}}_0,\hat{\boldsymbol{\theta}}^t\rangle}{\|\overline{\boldsymbol{\theta}}_0\|_2\|\hat{\boldsymbol{\theta}}^t\|_2}\leq\sqrt{\hat{q}_t}\,. \tag{B.54}$$

We next compute the first order Taylor-expansion of the iteration (B.52), and obtain $F_\varepsilon(q) = q + O(q^2)$, $H(q) = q/\delta_{\rm sp} + O(q^2)$ (the first order Taylor expanson of $H(q)$ was already computed in [146]). As a consequence, for any $\eta > 0$, there exists α_0 such that, if $\tilde{\alpha} < \alpha_0$, $q_t < \alpha_0$, then

$$q_{t+1} \leq (\frac{\delta}{\delta_{\rm sp}}+\eta)(q_t+\tilde{\alpha})\,.$$

The claim follows by taking $\eta = \eta(\delta) := (\delta_{\rm sp}-\delta)/(2\delta_{\rm sp})$, whence $q_t \leq \tilde{\alpha}/\eta(\delta)$ for all t, provided $\tilde{\alpha} < \alpha_* := \alpha_0\eta(\delta)$. The deterministic argument follows in the same way as Corollary 3.4.2.

Appendix C

A proof for GFOM via orthogonalization

C.1 Proof of Theorem 4.3.1 under Setting 1

In this section we prove Theorem 4.3.1 in the context of Setting 1. Therefore, $F_t, G_t, F_*^{(t)}$ are non-separable, namely they do not necessarily act on vectors entrywise.

Before we proceed, we first generalize the definition of pseudo-Lipschitz functions given in the main text. For any $m, l, k \in \mathbb{N}_{>0}$, a function $\phi : \mathbb{R}^l \to \mathbb{R}^m$ is called a pseudo-Lipschitz function of order k if there exists a constant $L > 0$, such that for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^l$,

$$\frac{1}{\sqrt{m}}\|\phi(\boldsymbol{x}) - \phi(\boldsymbol{y})\|_2 \leq L\left(1 + \left(\frac{\|\boldsymbol{x}\|_2}{\sqrt{l}}\right)^{k-1} + \left(\frac{\|\boldsymbol{y}\|_2}{\sqrt{l}}\right)^{k-1}\right)\frac{\|\boldsymbol{x}-\boldsymbol{y}\|_2}{\sqrt{l}}, \tag{C.1}$$

$$\frac{1}{\sqrt{m}}\|\phi(\boldsymbol{x})\|_2 \leq L\left(1 + \left(\frac{\|\boldsymbol{x}\|_2}{\sqrt{l}}\right)^{k}\right). \tag{C.2}$$

In what follows, we will often consider sequences of functions $\phi_n : \mathbb{R}^{l_n} \to \mathbb{R}^{m_n}$ indexed by n (even if we often do not write explicitly that we are considering a sequence). We say that such a sequence $\{\phi_n\}_{n\geq 1}$ is *uniformly pseudo-Lipschitz* of order k if Eqs. (C.1), (C.2) hold with L a constant that is independent of n.

C.1.1 Approximate message passing algorithms

As before, the first step is to define the AMP algorithm for this setting. An AMP algorithm is defined by Lipschitz non-linearities $\{f_t : \mathbb{R}^{n(t+1)} \to \mathbb{R}^n\}_{t\geq 0}$, and produces vectors $\{\boldsymbol{a}^t\}_{t\geq 1} \subseteq \mathbb{R}^n$ via the following iteration:

$$\boldsymbol{a}^{t+1} = \boldsymbol{X} f_t(\boldsymbol{a}^{\leq t}; \boldsymbol{u}) - \sum_{s=1}^{t} b_{t,s} f_{s-1}(\boldsymbol{a}^{\leq s-1}; \boldsymbol{u}). \tag{C.3}$$

For each $t \in \mathbb{N}$, f_t stands for a sequence of functions which are uniformly Lipschitz continuous. As before, we introduce the notation $\mathsf{OC}_{\text{AMP}}(\boldsymbol{a}^{\leq t-1}; \boldsymbol{u}) := \sum_{s=1}^{t} b_{t,s} f_{s-1}(\boldsymbol{a}^{\leq s-1}; \boldsymbol{u})$. Under Setting 1, the state evolution recursion to construct $\boldsymbol{\mu} = (\mu_t)_{t\geq 1}$ and $\boldsymbol{\Sigma} = (\Sigma_{s,t})_{s,t\geq 1}$ is defined as follows:

$$\begin{aligned}
\mu_{t+1} &= \lim_{n\to\infty} \frac{1}{n}\mathbb{E}[\boldsymbol{\theta}^\mathsf{T} f_t(\boldsymbol{\mu}_{\leq t}\boldsymbol{\theta} + \boldsymbol{g}_{\leq t}; \boldsymbol{u})]\,, \\
\Sigma_{s+1,t+1} &= \lim_{n\to\infty} \frac{1}{n}\mathbb{E}[f_s(\boldsymbol{\mu}_{\leq s}\boldsymbol{\theta} + \boldsymbol{g}_{\leq s}; \boldsymbol{u})^\mathsf{T} f_t(\boldsymbol{\mu}_{\leq t}\boldsymbol{\theta} + \boldsymbol{g}_{\leq t}; \boldsymbol{u})]\,, \\
\boldsymbol{g}_{\leq t} &:= (\boldsymbol{g}_1, \cdots, \boldsymbol{g}_t) \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\leq t} \otimes \boldsymbol{I}_n)\,,
\end{aligned} \tag{C.4}$$

where we adopted the notation $\boldsymbol{\mu}_{\leq t}\boldsymbol{\theta} + \boldsymbol{g}_{\leq t} := (\mu_1\boldsymbol{\theta} + \boldsymbol{g}_1, \cdots, \mu_t\boldsymbol{\theta} + \boldsymbol{g}_t)$ and we assume the above limits exist. Given $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, we define

$$b_{t,s} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\partial_{i,s} f_{t,i}(\boldsymbol{\mu}_{\leq t}\boldsymbol{\theta} + \boldsymbol{g}_{\leq t}; \boldsymbol{u})], \tag{C.5}$$

where $f_{t,i}$ is the i-th coordinate of f_t, and $\partial_{i,s}$ denotes the weak derivative with respect to the s-th variable of the i-th row of the input matrix. To give an example, for variables $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t \in \mathbb{R}^n$ and a function $f(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t)$ mapping from $\mathbb{R}^{nt}$ to $\mathbb{R}$, we have $\partial_{i,s} f(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t) = \partial_{(\boldsymbol{x}_s)_i} f(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t)$. Notice that here $b_{t,s}$ depends on n. Since f_t is uniformly Lipschitz in terms of n, for all $t, s \in \mathbb{N}_{>0}$, $b_{t,s}$ is uniformly bounded as a sequence in n.

After t iterations as in Eq. (C.3), the AMP algorithm estimates $\boldsymbol{\theta}$ by applying a uniformly Lipschitz function $f_t^* : \mathbb{R}^{n(t+1)} \to \mathbb{R}^n$ to $(\boldsymbol{a}^{\leq t}, \boldsymbol{u})$:

$$\hat{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{u}) = f_t^*(\boldsymbol{a}^{\leq t}; \boldsymbol{u}).$$

The following theorem characterizes the asymptotic performance of the AMP algorithm (C.3).

Theorem C.1.1. *Assume that $\{(\theta_i, u_i)\}_{i\leq n} \overset{iid}{\sim} \mu_{\Theta,U}$, and $\boldsymbol{W}$ satisfies the same assumption as $\boldsymbol{W}$ under Setting 1. For all $t \in \mathbb{N}$, assume f_t is uniformly Lipschitz. Furthermore, we assume the limits*

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n}\mathbb{E}[\boldsymbol{\theta}^\mathsf{T} f_t(\boldsymbol{\mu}_{\leq t}\boldsymbol{\theta} + \boldsymbol{g}_{\leq t}; \boldsymbol{u})], \\
&\lim_{n\to\infty} \frac{1}{n}\mathbb{E}[f_s(\boldsymbol{\mu}_{\leq s}\boldsymbol{\theta} + \boldsymbol{g}_{\leq s}; \boldsymbol{u})^\mathsf{T} f_t(\boldsymbol{\mu}_{\leq t}\boldsymbol{\theta} + \boldsymbol{g}_{\leq t}; \boldsymbol{u})], \\
&\lim_{n\to\infty} \frac{1}{n}\mathbb{E}[\boldsymbol{\theta}^\mathsf{T} f_t^*(\boldsymbol{\mu}_{\leq t}\boldsymbol{\theta} + \boldsymbol{g}_{\leq t}; \boldsymbol{u})], \\
&\lim_{n\to\infty} \frac{1}{n}\mathbb{E}[f_s^*(\boldsymbol{\mu}_{\leq s}\boldsymbol{\theta} + \boldsymbol{g}_{\leq s}; \boldsymbol{u})^\mathsf{T} f_t^*(\boldsymbol{\mu}_{\leq t}\boldsymbol{\theta} + \boldsymbol{g}_{\leq t}; \boldsymbol{u})]
\end{aligned}$$

exist for all n-independent $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $t, s \in \mathbb{N}$. Then, for any $t \in \mathbb{N}_{>0}$ and $\{\psi_n : \mathbb{R}^{n(t+1)} \to \mathbb{R}\}_{n\geq 1}$ uniformly pseudo-Lipschitz of order 2,

$$\operatorname*{p-lim}_{n\to\infty} \left|\psi_n(\boldsymbol{a}^{\leq t}; \boldsymbol{u}) - \mathbb{E}[\psi_n(\boldsymbol{\mu}_{\leq t}\boldsymbol{\theta} + \boldsymbol{g}_{\leq t}; \boldsymbol{u})]\right| = 0\,.$$

Remark C.1.1. Theorem C.1.1 is a generalized version of [33, Theorem 1]. In [33] the non-linearity f_t only depends on $(\boldsymbol{a}^t, \boldsymbol{u})$, while here we allow it to depend on all previous iterates $(\boldsymbol{a}^{\leq t}, \boldsymbol{u})$.

This generalization can be conducted through the following steps: (1) Replace the vectors $f_t(\boldsymbol{a}^t; \boldsymbol{u}), \boldsymbol{a}^t \in \mathbb{R}^n$ by matrices $f_t(\boldsymbol{a}^t; \boldsymbol{u}), \boldsymbol{a}^t \in \mathbb{R}^{n\times q}$, and replace the coefficients for the Onsager correction term $b_{t,t}$ by $q \times q$ matrices (see, e.g., [104]). Such generalization follows exactly by the same proof as in [33]. (2) Fix a time horizon t, and choose an n-independent q such that $q \geq t$. With initialization $\boldsymbol{x}_1^0 = \cdots = \boldsymbol{x}_q^0 = \boldsymbol{0}$, we set the non-linearity corresponding to the $(s+1)$-th iteration as

$$(\boldsymbol{x}_1^s, \cdots, \boldsymbol{x}_q^s, \boldsymbol{u}) \mapsto (f_0(\boldsymbol{u}), \cdots, f_s(\boldsymbol{x}_1^s, \cdots, \boldsymbol{x}_s^s; \boldsymbol{u}), \boldsymbol{0}, \cdots, \boldsymbol{0}) \in \mathbb{R}^{n \times q}.$$

In this way, the vectors $(\boldsymbol{x}_s^t)_{1 \leq s \leq t}$ coincides with $(\boldsymbol{a}^s)_{1 \leq s \leq t}$.

C.1.2 Any GFOM can be reduced to an AMP algorithm

In this section we show that, under Setting 1, any GFOM can be reduced to an AMP algorithm via a change of variables.

Lemma C.1.1. *Under the assumptions of Setting 1, for all $t \in \mathbb{N}_{>0}$, there exist uniformly Lipschitz functions $\varphi_t : \mathbb{R}^{n(t+1)} \to \mathbb{R}^{nt}$ and $f_{t-1} : \mathbb{R}^{nt} \to \mathbb{R}^n$ that are independent of $(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{W})$, such that the following holds. Let $\{\boldsymbol{a}^t\}_{t \geq 1}$ be the sequence of vectors produced by the AMP iteration (C.3) with non-linearities $\{f_s\}_{s \geq 0}$, then for any $t \in \mathbb{N}_{>0}$, we have*

$$\boldsymbol{u}^{\leq t} = \varphi_t(\boldsymbol{a}^{\leq t}; \boldsymbol{u}), \qquad f_{t-1}(\boldsymbol{a}^{\leq t-1}; \boldsymbol{u}) = F_{t-1}(\varphi_t(\boldsymbol{a}^{\leq t-1}; \boldsymbol{u}); \boldsymbol{u}).$$

Furthermore, $\{\varphi_t\}_{t \geq 1}$ satisfies the following conditions. Let $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the state evolution of the AMP algorithm defined in Eq. (C.4). For any $t \in \mathbb{N}_{>0}$, there exist uniformly bounded numbers $(b_{ij})_{1 \leq i,j \leq t}$ (which depend on n), such that for $\boldsymbol{y}_{\leq t}$ defined in Eq. (4.7), we have $\boldsymbol{y}_{\leq t} = \varphi_t(\boldsymbol{\mu}_{\leq t}\boldsymbol{\theta} + \boldsymbol{g}_{\leq t}; \boldsymbol{u})$.

Proof. We prove the lemma by induction over t. For the base case $t = 1$, we may simply take $f_0(\boldsymbol{u}) = F_0(\boldsymbol{u})$ and $\varphi_1(\boldsymbol{a}^1; \boldsymbol{u}) := \boldsymbol{a}^1 + G_0(\boldsymbol{u})$. Then $\boldsymbol{y}^1 = \varphi_1(\mu_1\boldsymbol{\theta} + \boldsymbol{g}_1; \boldsymbol{u})$ by definition.

Suppose the claim holds for the first t iterations, then we prove it holds for the $(t+1)$-th iteration. By the induction hypothesis,

$$\boldsymbol{u}^{t+1} = \boldsymbol{X} F_t(\varphi_t(\boldsymbol{a}^{\leq t}; \boldsymbol{u}); \boldsymbol{u}) + G_t(\varphi_t(\boldsymbol{a}^{\leq t}; \boldsymbol{u}); \boldsymbol{u}).$$

Let $f_t(\boldsymbol{x}^{\leq t}; \boldsymbol{u}) = F_t(\varphi_t(\boldsymbol{x}^{\leq t}; \boldsymbol{u}); \boldsymbol{u})$. The composite of uniformly Lipschitz functions is still uniformly Lipschitz, thus, we conclude that f_t is uniformly Lipschitz. Based on the choice of $\{f_s\}_{0 \leq s \leq t}$, we compute the coefficients for the Onsager correction term $\{b_{t,s}\}_{1 \leq s \leq t}$ according to Eq. (C.5). Then we define $\boldsymbol{a}^{t+1}$ via Eq. (C.3), which gives

$$\boldsymbol{a}^{t+1} = \boldsymbol{u}^{t+1} - G_t((\varphi_t(\boldsymbol{a}^{\leq t}; \boldsymbol{u}); \boldsymbol{u}) - \sum_{s=1}^{t} b_{t,s} f_{s-1}(\boldsymbol{a}^{s-1}; \boldsymbol{u}).$$

Therefore, we define φ_{t+1} as

$$\varphi_{t+1}(\boldsymbol{a}^{\leq t+1}; \boldsymbol{u}) = (\varphi_t(\boldsymbol{a}^{\leq t}; \boldsymbol{u}); \boldsymbol{a}^{t+1} + G_t(\varphi_t(\boldsymbol{a}^{\leq t}; \boldsymbol{u}); \boldsymbol{u}) + \sum_{s=1}^{t} b_{t,s} f_{s-1}(\boldsymbol{a}^{\leq s-1}; \boldsymbol{u})).$$

By induction hypothesis and the fact that $b_{t,s}$ is uniformly bounded with respect to n for all fixed $t, s \in \mathbb{N}_{>0}$, we have that φ_{t+1} is uniformly Lipschitz. Furthermore,

$$
\begin{aligned}
&\varphi_{t+1}(\boldsymbol{\mu}_{\le t+1}\boldsymbol{\theta}+\boldsymbol{g}_{\le t+1};\boldsymbol{u})\\
=&(\varphi_t(\boldsymbol{\mu}_{\le t}\boldsymbol{\theta}+\boldsymbol{g}_{\le t};\boldsymbol{u}),\mu_{t+1}\boldsymbol{\theta}+\boldsymbol{g}_{t+1}+G_t(\varphi_t(\boldsymbol{\mu}_{\le t}\boldsymbol{\theta}+\boldsymbol{g}_{\le t};\boldsymbol{u});\boldsymbol{u})+\sum_{s=1}^{t}b_{t,s}f_{s-1}(\boldsymbol{\mu}_{\le s-1}\boldsymbol{\theta}+\boldsymbol{g}_{\le s-1};\boldsymbol{u}))\\
=&(\boldsymbol{y}^{\le t},\boldsymbol{y}^{t+1}),
\end{aligned}
$$

thus completes the proof of the lemma by induction.

□

The next lemma enables us to check the conditions of Theorem C.1.1.

Lemma C.1.2. *Under the assumptions of Setting 1, let $\{f_{t-1},\varphi_t\}_{t\in\mathbb{N}^+}$ be the functions defined in Lemma C.1.1. For any $\boldsymbol{\mu}=(\mu_i)_{i\ge1}$, $\boldsymbol{\Sigma}=(\Sigma_{ij})_{i,j\ge1}\succeq\boldsymbol{0}$, let $(\boldsymbol{g}_t)_{t>0}$ be a centered Gaussian process with covariance $\mathbb{E}\{\boldsymbol{g}_s\boldsymbol{g}_t^{\mathsf{T}}\}=\Sigma_{s,t}\boldsymbol{I}_n$. Then, for any $t\in\mathbb{N}$ and any infinite subsequence $\mathcal{S}\subseteq\mathbb{N}_{>0}$ there exists a further subsequence $\mathcal{S}'\subseteq\mathcal{S}$ along which the following limits exist for all $0\le s\le r\le t$:*

$$
\begin{aligned}
&\lim_{n\to\infty;n\in S'}\frac{1}{n}\mathbb{E}[f_r(\boldsymbol{\mu}_{\le r}\boldsymbol{\theta}+\boldsymbol{g}_{\le r};\boldsymbol{u})^{\mathsf{T}}f_s(\boldsymbol{\mu}_{\le s}\boldsymbol{\theta}+\boldsymbol{g}_{\le s};\boldsymbol{u})],\\
&\lim_{n\to\infty;n\in S'}\frac{1}{n}\mathbb{E}[\boldsymbol{\theta}^{\mathsf{T}}f_s(\boldsymbol{\mu}_{\le s}\boldsymbol{\theta}+\boldsymbol{g}_{\le s};\boldsymbol{u})],\\
&\lim_{n\to\infty;n\in S'}\frac{1}{n}\mathbb{E}[F_*^{(r)}(\varphi_r(\boldsymbol{\mu}_{\le r}\boldsymbol{\theta}+\boldsymbol{g}_{\le r};\boldsymbol{u});\boldsymbol{u})^{\mathsf{T}}F_*^{(s)}(\varphi_s(\boldsymbol{\mu}_{\le s}\boldsymbol{\theta}+\boldsymbol{g}_{\le s};\boldsymbol{u});\boldsymbol{u})],\\
&\lim_{n\to\infty;n\in S'}\frac{1}{n}\mathbb{E}[\boldsymbol{\theta}^{\mathsf{T}}F_*^{(s)}(\varphi_s(\boldsymbol{\mu}_{\le s}\boldsymbol{\theta}+\boldsymbol{g}_{\le s};\boldsymbol{u});\boldsymbol{u})].
\end{aligned}
\tag{C.6}
$$

Proof. We can assume that the subsequence $\mathcal{S}$ does coincide with the whole sequence, i.e. $\mathcal{S}=\mathbb{N}_{>0}$, as the general case follows by a simple change of notations.

Fix $t\in\mathbb{N}$. Since $(b_{i,j})_{1\le i,j\le t}$ are uniformly bounded, there exists a subsequence $\{n_k\}_{k>0}$ of $\mathbb{N}_{>0}$, such that for all $1\le s,r\le t$, $b_{s,r}$ converges to limit $b^*_{s,r}$. Suppose we replace $(b_{i,j})_{1\le i,j\le t}$ with $(b^*_{i,j})_{1\le i,j\le t}$ in Eq. (4.7), and we denote the resulting vectors by $(\boldsymbol{y}^*_t)_{t\ge1}$. It follows by induction and using the uniform Lipschitz property that for all $0\le s,r\le t$, along $\{n_k\}_{k>0}$,

$$
\begin{aligned}
&\frac{1}{n}F_r(\boldsymbol{y}^*_{\le r};\boldsymbol{u})^{\mathsf{T}}F_s(\boldsymbol{y}^*_{\le s};\boldsymbol{u})-\frac{1}{n}F_r(\boldsymbol{y}_{\le r};\boldsymbol{u})^{\mathsf{T}}F_s(\boldsymbol{y}_{\le s};\boldsymbol{u})\xrightarrow{P}0,\\
&\frac{1}{n}F_*^{(r)}(\boldsymbol{y}^*_{\le r};\boldsymbol{u})^{\mathsf{T}}F_*^{(s)}(\boldsymbol{y}^*_{\le s};\boldsymbol{u})-\frac{1}{n}F_*^{(r)}(\boldsymbol{y}_{\le r};\boldsymbol{u})^{\mathsf{T}}F_*^{(s)}(\boldsymbol{y}_{\le s};\boldsymbol{u})\xrightarrow{P}0,\\
&\frac{1}{n}\boldsymbol{\theta}^{\mathsf{T}}F_s(\boldsymbol{y}^*_{\le s};\boldsymbol{u})-\frac{1}{n}\boldsymbol{\theta}^{\mathsf{T}}F_s(\boldsymbol{y}_{\le s};\boldsymbol{u})\xrightarrow{P}0.\\
&\frac{1}{n}\boldsymbol{\theta}^{\mathsf{T}}F_*^{(s)}(\boldsymbol{y}^*_{\le s};\boldsymbol{u})-\frac{1}{n}\boldsymbol{\theta}^{\mathsf{T}}F_*^{(s)}(\boldsymbol{y}_{\le s};\boldsymbol{u})\xrightarrow{P}0.
\end{aligned}
$$

By the third assumption of Setting 1, the limits of $F_r(\boldsymbol{y}^*_{\le r};\boldsymbol{u})^{\mathsf{T}}F_s(\boldsymbol{y}^*_{\le s};\boldsymbol{u})/n$, $F_*^{(r)}(\boldsymbol{y}^*_{\le r};\boldsymbol{u})^{\mathsf{T}}F_*^{(s)}(\boldsymbol{y}^*_{\le s};\boldsymbol{u})/n$, $\boldsymbol{\theta}^{\mathsf{T}}F_*^{(s)}(\boldsymbol{y}^*_{\le s};\boldsymbol{u})/n$ and $\boldsymbol{\theta}^{\mathsf{T}}F_s(\boldsymbol{y}^*_{\le s};\boldsymbol{u})/n$ exist in probability as $n,d\to\infty$. Combining these results and the results of Lemma C.1.1, we conclude that the limits of Eqs. (C.6) exist along $\{n_k\}_{k\in\mathbb{N}_{>0}}$:

□

The following corollary is an immediate consequence of Lemma C.1.1.

Corollary C.1.1. *Under the assumptions of Setting 1, let $\mathcal{A}^t_{\rm GFOM}(L)$ be the class of GFOM estimators with t iterations and uniform Lipschitz constant L, and $\mathcal{A}^t_{\rm AMP}(L')$ be the class of AMP algorithms with t iterations and uniform Lipschitz constant L'. Then for any $L < \infty$ there exist $L' < \infty$ (independent of n), such that the following holds. For any $z \in \mathbb{R}$ and any loss function $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_{\geq 0}$:*

$$\inf_{\hat{\boldsymbol{\theta}}(\cdot)\in\mathcal{A}^t_{\rm GFOM}(L)} \mathbb{P}\Big(\mathcal{L}(\hat{\boldsymbol{\theta}}(\boldsymbol{X},\boldsymbol{u}),\boldsymbol{\theta}) \leq z\Big) \leq \inf_{\hat{\boldsymbol{\theta}}(\cdot)\in\mathcal{A}^t_{\rm AMP}(L')} \mathbb{P}\Big(\mathcal{L}(\hat{\boldsymbol{\theta}}(\boldsymbol{X},\boldsymbol{u}),\boldsymbol{\theta}) \leq z\Big)\,.$$

Notice that in this corollary $\hat{\boldsymbol{\theta}}(\,\cdot\,) \in \mathcal{A}^t_{\rm GFOM}(L)$ is (implicitly) a sequence of estimators indexed by n, which is uniformly Lipschitz with constant L. The corollary also implies an asymptotic statement. Namely, write $\mathcal{A}^t_{\rm GFOM} := \cup_{L\geq 1}\mathcal{A}^t_{\rm GFOM}(L)$ for the class of (sequences of) GFOM estimators with t iterations and any uniform Lipschitz constant L, and similarly for $\mathcal{A}^t_{\rm AMP}$. Then we have

$$\inf_{\hat{\boldsymbol{\theta}}(\cdot)\in\mathcal{A}^t_{\rm GFOM}} \operatorname*{p-liminf}_{n\to\infty} \mathcal{L}(\hat{\boldsymbol{\theta}}(\boldsymbol{X},\boldsymbol{u}),\boldsymbol{\theta}) = \inf_{\hat{\boldsymbol{\theta}}(\cdot)\in\mathcal{A}^t_{\rm AMP}} \operatorname*{p-liminf}_{n\to\infty} \mathcal{L}(\hat{\boldsymbol{\theta}}(\boldsymbol{X},\boldsymbol{u}),\boldsymbol{\theta})\,. \tag{C.7}$$

Here equality holds because $\mathcal{A}^t_{\rm AMP} \subseteq \mathcal{A}^t_{\rm GFOM}$.

C.1.3 Any AMP algorithm can be reduced to an orthogonal AMP algorithm

By Corollary C.1.1, and in particular Eq. (C.7), we can limit ourselves to lower-bounding the error of AMP algorithms. By Lemma C.1.2 we can assume —possibly taking subsequences— that such algorithm satisfies the conditions of Theorem C.1.1. To simplify notations, we will assume hereafter that these conditions are satisfied along $n \in \mathbb{N}$. There is no loss of generality in this.

Here we show that it is in fact sufficient to lower bound the error for OAMP algorithms.

Lemma C.1.3. *Let $\{\boldsymbol{a}^t\}_{t\geq 1}$ be a sequence generated by the AMP iteration (C.3) under the conditions of Theorem C.1.1. Then for all $t \in \mathbb{N}^+$, there exist uniformly Lipschitz functions $\phi_t : \mathbb{R}^{n(t+1)} \to \mathbb{R}^{nt}$, $g_{t-1} : \mathbb{R}^{nt} \to \mathbb{R}^n$ such that the following holds. Let $\{\boldsymbol{v}^t\}_{t\geq 1}$ be the sequence of vectors produced by AMP iteration with non-linearities $\{g_t\}_{t\geq 0}$ (and the same matrix $\boldsymbol{X}$ as for $\{\boldsymbol{a}^t\}_{t\geq 1}$). Namely,*

$$\boldsymbol{v}^{t+1} = \boldsymbol{X} g_t(\boldsymbol{v}^{\leq t};\boldsymbol{u}) - \sum_{s=1}^{t} b'_{t,s} g_{s-1}(\boldsymbol{v}^{\leq s-1};\boldsymbol{u}) \tag{C.8}$$

with deterministic coefficients $(b'_{t,s})$ determinied by the analogous of Eq. (C.5), with f_t replaced by g_t. Then we have

- *(i) For all $t \in \mathbb{N}_{>0}$, $\boldsymbol{a}^{\leq t} = \phi_t(\boldsymbol{v}^{\leq t};\boldsymbol{u})$. Further, there exists n-independent constants $\{c_{ts}\}_{0\leq s\leq t}$, such that we can write $\boldsymbol{v}^t = \sum_{s=0}^{t-1} c_{t-1,s}\boldsymbol{a}^{s+1}$.*
- *(ii) For all $t \in \mathbb{N}_{>0}$, there exist $(x_0, \cdots, x_{t-1}) \in \{0,1\}^t$ and $(\alpha_1, \cdots, \alpha_t) \in \mathbb{R}^t$, such that for any $\{\psi_n : \mathbb{R}^{n(t+2)} \to \mathbb{R}\}_{n\geq 1}$ uniformly pseudo-Lipschitz of order 2,*

$$\psi_n(\boldsymbol{v}^{\leq t},\boldsymbol{\theta},\boldsymbol{u}) = \mathbb{E}[\psi_n(\boldsymbol{\nu}^{\leq t},\boldsymbol{\theta},\boldsymbol{u})] + o_P(1),$$

where $\boldsymbol{\nu}^i = x_{i-1}(\alpha_i\boldsymbol{\theta} + \boldsymbol{z}_i)$ and $\{\boldsymbol{z}_i\}_{i\geq 1} \overset{iid}{\sim} \mathsf{N}(\boldsymbol{0},\boldsymbol{I}_n)$ independent of $(\boldsymbol{\theta},\boldsymbol{u})$.

Proof. Recall that, as in the proof of Lemma 4.4.2, $\Pi_{\mathcal{S}}$ denotes the orthogonal projection onto the closed linear subspace $\mathcal{S} \subseteq L^2(\mathbb{P})$, and $\Pi_{\mathcal{S}}^{\perp} := I - \Pi_{\mathcal{S}}$.

We denote by $(\mu_t)_{t\geq 1}$, $(\Sigma_{s,t})_{s,t\geq 1}$ the state evolution sequence corresponding to $\{\boldsymbol{a}^t\}_{t\geq 1}$, defined via Eq. (C.4). Let $(\boldsymbol{g}_t)_{t\geq 1}$ be a centered Gaussian process in $\mathbb{R}^n$ such that $\mathrm{Cov}(\boldsymbol{g}_s, \boldsymbol{g}_t) = \Sigma_{s,t}\boldsymbol{I}_n$. We define the following random vectors and subspaces:

$$\boldsymbol{h}_t = f_t(\boldsymbol{\mu}_{\leq t}\boldsymbol{\theta} + \boldsymbol{g}_{\leq t}; \boldsymbol{u}), \qquad \mathcal{S}_t = \mathrm{span}(\boldsymbol{h}_k : 0 \leq k \leq t).$$

By assumption, for all $s, t \in \mathbb{N}$,

$$\frac{1}{n}\mathbb{E}\langle \boldsymbol{h}_s, \boldsymbol{h}_t\rangle \to \Sigma_{s+1,t+1}, \qquad \frac{1}{n}\mathbb{E}\langle \boldsymbol{\theta}, \boldsymbol{h}_t\rangle \to \mu_{t+1}. \tag{C.9}$$

By linear algebra, there exist deterministic n-independent constants $\{c_{ts}\}_{t,s\in\mathbb{N}}$, $\{x_t\}_{t\in\mathbb{N}} \in \{0,1\}^{\mathbb{N}}$, such that $c_{tt} \neq 0$ and

$$\sum_{i=0}^{t}\sum_{j=0}^{s} c_{ti}c_{sj}\Sigma_{i+1,j+1} = \mathbb{1}_{s=t}x_t.$$

If we let $\boldsymbol{r}_t = \sum_{s=0}^{t} c_{ts}\boldsymbol{h}_s$, then by the convergence of second moments given in Eq. (C.9), for all $s, t \in \mathbb{N}$

$$\lim_{n\to\infty}\frac{1}{n}\mathbb{E}\langle \boldsymbol{r}_t, \boldsymbol{r}_s\rangle = \mathbb{1}_{s=t}x_t.$$

Then we prove the lemma by induction. For the base case $t = 1$, we let $g_0(\boldsymbol{u}) = c_{00}f_0(\boldsymbol{u})$, thus $\boldsymbol{v}^1 = c_{00}\boldsymbol{a}^1$ and claim (i) follows trivially. As for claim (ii), first notice that the limits exist for both $\mathbb{E}\langle g_0(\boldsymbol{u}), g_0(\boldsymbol{u})\rangle/n$ and $\mathbb{E}\langle g_0(\boldsymbol{u}), \boldsymbol{\theta}\rangle/n$ by the assumption on the original AMP iteration. Then we consider two cases. In the first case $x_0 = 0$, thus $\Sigma_{11} = 0$, $\mu_1^2 \leq c_{00}^{-2}\mathbb{E}[\|\boldsymbol{\theta}\|_2^2/n]\mathbb{E}[\|g_0(\boldsymbol{u})\|_2^2/n] \to 0$, and (ii) holds with $\boldsymbol{\nu}^1 = \boldsymbol{0}$ by Theorem C.1.1. In the second case $x_0 = 1$, whence $c_{00} = \Sigma_{11}^{-1/2}$, and claim (ii) again follows from state evolution. Furthermore,

$$\alpha_1 = \lim_{n\to\infty}\frac{\mathbb{E}[\langle f_0(\boldsymbol{u}), \boldsymbol{\theta}\rangle]}{\sqrt{n}\mathbb{E}[\langle f_0(\boldsymbol{u}), f_0(\boldsymbol{u})\rangle]^{1/2}}. \tag{C.10}$$

Suppose the lemma holds for the first t iterations. We prove it also holds for the $(t+1)$-th iteration. We let

$$g_t(\boldsymbol{v}^{\leq t}; \boldsymbol{u}) = \sum_{s=0}^{t} c_{ts}f_s(\phi_s(\boldsymbol{v}^{\leq s}; \boldsymbol{u}); \boldsymbol{u}).$$

By induction hypothesis and assumptions, g_t is uniformly Lipschitz. Given $\{g_s\}_{0\leq s\leq t}$, we can derive the coefficients $(b'_{s,j})_{1\leq j\leq s\leq t}$ via Eq. (C.5), and we denote the Onsager correction term of this new AMP iteration by $\mathsf{OC}_{\mathrm{OAMP}}^{t}(\boldsymbol{v}^{\leq t-1}; \boldsymbol{u}) = \sum_{s=1}^{t} b'_{t,s}g_{s-1}(\boldsymbol{v}^{\leq s-1}; \boldsymbol{u})$. Then Eq. (C.8) can be rewritten as

$$\boldsymbol{v}^{t+1} = \sum_{s=0}^{t} c_{ts}\boldsymbol{X}f_s(\phi_s(\boldsymbol{v}^{\leq s}; \boldsymbol{u}); \boldsymbol{u}) - \mathsf{OC}_{\mathrm{OAMP}}^{t}(\boldsymbol{v}^{\leq t-1}; \boldsymbol{u}).$$

Plugging in the AMP iteration that defines $\{\boldsymbol{a}^t\}_{t\geq 1}$, we have

$$\boldsymbol{v}^{t+1} = \sum_{s=0}^{t} c_{ts}(\boldsymbol{a}^{s+1} + \mathsf{OC}^{s}_{\rm AMP}(\boldsymbol{a}^{\leq s-1};\boldsymbol{u})) - \mathsf{OC}^{t}_{\rm OAMP}(\boldsymbol{v}^{\leq t-1};\boldsymbol{u}). \tag{C.11}$$

Recall that c_{tt} is non-vanishing, thus, we can solve for $\boldsymbol{a}^{t+1}$ and express $\boldsymbol{a}^{\leq t+1}$ as a function of $(\boldsymbol{v}^{\leq t+1};\boldsymbol{u})$. We denote this function by ϕ_{t+1}. By induction hypothesis, ϕ_{t+1} is uniformly Lipschitz. Plugging the definition of $\mathsf{OC}^{s}_{\rm AMP}$ and $\mathsf{OC}^{t}_{\rm OAMP}$ into Eq. (C.11) gives

$$\boldsymbol{v}^{t+1} = \sum_{s=0}^{t} c_{ts}\boldsymbol{a}^{s+1} + \sum_{i=1}^{t}\big(\sum_{s=i}^{t} c_{ts}b_{si} - \sum_{s=i}^{t} b'_{ts}c_{s-1,i-1}\big) f_{i-1}(\boldsymbol{a}^{\leq i-1};\boldsymbol{u}). \tag{C.12}$$

By induction hypothesis, $g_t(c_{00}\boldsymbol{x}^1, \cdots, \sum_{s=0}^{t-1} c_{t-1,s}\boldsymbol{x}^{s+1};\boldsymbol{u}) = \sum_{s=0}^{t} c_{ts} f_s(\boldsymbol{x}^{\leq s};\boldsymbol{u})$. Taking the gradient on both sides with respect to $\boldsymbol{x}^i$, then compute the expected average of the coordinates of the gradient with respect to the distribution $\boldsymbol{x}^{\leq t} \overset{d}{=} \boldsymbol{\mu}^{\leq t}\boldsymbol{\theta} + \boldsymbol{g}^{\leq t}$ gives $\sum_{s=i}^{t} c_{ts}b_{si} - \sum_{s=i}^{t} b'_{ts}c_{s-1,i-1} = 0$. Plugging this into Eq. (C.12) finishes the proof of claim (i).

One can verify that the non-linearities $\{g_s\}_{0\leq s\leq t}$ defined in this way satisfy the conditions of Theorem C.1.1, thus the asymptotics of OAMP can be characterized by state evolution. As for the proof of claim (ii), again we consider two cases. If $x_t = 0$, then $\mathbb{E}\langle \boldsymbol{r}_t, \boldsymbol{r}_t\rangle/n \to 0$, and $\mathbb{E}\langle \boldsymbol{r}_t, \boldsymbol{\theta}\rangle/n \to 0$. Using the state evolution for OAMP (C.8), we obtain that (ii) holds with $\boldsymbol{\nu}^{t+1} = \boldsymbol{0}$. If $x_t = 1$, then again by state evolution for OAMP, claim (ii) holds with $\boldsymbol{\nu}^{t+1} = \alpha_{t+1}\boldsymbol{\theta} + \boldsymbol{z}_{t+1}$ where

$$\alpha_{t+1} = \lim_{n\to\infty} \frac{\mathbb{E}\langle \boldsymbol{\theta}, \Pi^{\perp}_{\mathcal{S}_{t-1}}(\boldsymbol{h}_t)\rangle}{\sqrt{n}\mathbb{E}[\langle \Pi^{\perp}_{\mathcal{S}_{t-1}}(\boldsymbol{h}_t), \Pi^{\perp}_{\mathcal{S}_{t-1}}(\boldsymbol{h}_t)\rangle]^{1/2}}, \tag{C.13}$$

thus completes the proof by induction.

□

C.1.4 Optimal orthogonal AMP

Following the same reasoning of Remark 4.4.3, in the following we will restrict to the cases in which $x_t = 1$ for all $t \in \mathbb{N}$.

Combining Lemma C.1.1 and C.1.3, we conclude that it is sufficient to lower bound the error of OAMP algorithms. The following corollary is a direct consequence of the proceeding results, and extends Eq. (C.7).

Corollary C.1.2. *Under the assumptions of Setting 1, recall that $\mathcal{A}^t_{\rm GFOM}$ denotes the class of uniformly Lipschitz GFOM estimators with t iterations, and denote by $\mathcal{A}^t_{\rm OAMP}$ the class of OAMP estimators with t iterations (i.e., AMP estimators whose state evolution yields $\Sigma_{s,t} = \mathbf{1}_{s=t}$).*

Then we have

$$\inf_{\hat{\boldsymbol{\theta}}(\,\cdot\,)\in\mathcal{A}^t_{\rm GFOM}} \operatorname*{p-liminf}_{n\to\infty} \frac{1}{n}\big\|\hat{\boldsymbol{\theta}}(\boldsymbol{X},\boldsymbol{u}) - \boldsymbol{\theta}\big\|_2^2 = \inf_{\hat{\boldsymbol{\theta}}(\,\cdot\,)\in\mathcal{A}^t_{\rm OAMP}} \operatorname*{p-liminf}_{n\to\infty} \big\|\hat{\boldsymbol{\theta}}(\boldsymbol{X},\boldsymbol{u}) - \boldsymbol{\theta}\big\|_2^2. \tag{C.14}$$

Notice that a sufficient statistics for $\boldsymbol{\theta}$ given $\boldsymbol{\alpha}_{\leq t}\boldsymbol{\theta} + \boldsymbol{z}_{\leq t}$ is $T_0 := \|\boldsymbol{\alpha}_{\leq t}\|_s\boldsymbol{\theta} + \boldsymbol{z}$ with $\boldsymbol{z} \overset{d}{=} \mathsf{N}(\vec{0}, \boldsymbol{I}_n)$ independent of $\boldsymbol{\theta}$. Therefore, in order to derive the minimum of the right hand side of Eq. (C.14), it is sufficient to compute the maximum value of $\|\boldsymbol{\alpha}_{\leq t}\|_2$, which is provided by the following lemma. The proof of Theorem 4.3.1 under Setting 1 directly follows.

Lemma C.1.4. *Recall that $(\gamma_s)_{s\geq 0}$ is defined in Eq. (4.8). Then, for all $t \in \mathbb{N}$ and all choice of non-linearities $g_0, \cdots, g_t$, we have $\|\boldsymbol{\alpha}_{\leq t}\|_2 \leq \gamma_t$.*

Proof. The proof is by induction over t. For the base case $t = 1$, notice that

$$\sup_{f_0} \frac{\mathbb{E}[\langle f_0(\boldsymbol{u}), \boldsymbol{\theta}\rangle]^2}{n\mathbb{E}[\langle f_0(\boldsymbol{u}), f_0(\boldsymbol{u})\rangle]} = \frac{\mathbb{E}[\langle f_0(\boldsymbol{u}), \mathbb{E}[\boldsymbol{\theta} \mid \boldsymbol{u}]\rangle]^2}{n\mathbb{E}[\langle f_0(\boldsymbol{u}), f_0(\boldsymbol{u})\rangle]} \leq \gamma_1^2.$$

The last step above is via application of Cauchy-Schwarz inequality. Then the base case holds by taking the limit $n \to \infty$ in Eq. (C.10).

We assume that the claim holds for the first t iterations, and we prove by induction that it also holds for iteration $t+1$. We let $\hat{\boldsymbol{\theta}}_t := \mathbb{E}[\boldsymbol{\theta} \mid \boldsymbol{r}_1, \cdots, \boldsymbol{r}_t, \boldsymbol{u}]$, then

$$\begin{aligned}
\frac{\mathbb{E}[\langle \boldsymbol{\theta}, \Pi^{\perp}_{\mathcal{S}_{t-1}}(\boldsymbol{h}_t)\rangle]^2}{n\mathbb{E}[\langle \Pi^{\perp}_{\mathcal{S}_{t-1}}(\boldsymbol{h}_t), \Pi^{\perp}_{\mathcal{S}_{t-1}}(\boldsymbol{h}_t)\rangle]} &= \frac{\mathbb{E}[\langle \hat{\boldsymbol{\theta}}_t, \Pi^{\perp}_{\mathcal{S}_{t-1}}(\boldsymbol{h}_t)\rangle]^2}{n\mathbb{E}[\langle \Pi^{\perp}_{\mathcal{S}_{t-1}}(\boldsymbol{h}_t), \Pi^{\perp}_{\mathcal{S}_{t-1}}(\boldsymbol{y}_t)\rangle]} \\
&\overset{(a)}{\leq} \frac{1}{n}\mathbb{E}[\|\Pi^{\perp}_{\mathcal{S}_{t-1}}(\hat{\boldsymbol{\theta}}_t)\|_2^2] \\
&\overset{(b)}{=} \frac{1}{n}\mathbb{E}[\|\hat{\boldsymbol{\theta}}_t\|_2^2] - \frac{1}{n}\mathbb{E}[\|\Pi_{\mathcal{S}_{t-1}}(\hat{\boldsymbol{\theta}}_t)\|_2^2],
\end{aligned}$$

where (a) follows from Cauchy-Schwartz inequality and (b) from Pythagora's theorem. Notice that

$$\{\Pi_{\mathcal{S}_{s-1}}(\boldsymbol{h}_s)/\mathbb{E}[\|\Pi_{\mathcal{S}_{s-1}}(\boldsymbol{h}_s)\|_2^2]^{1/2} : 0 \leq s \leq t-1\}$$

is an orthonormal basis for $\mathcal{S}_{t-1}$, thus,

$$\frac{\mathbb{E}[\langle \boldsymbol{\theta}, \Pi^{\perp}_{\mathcal{S}_{t-1}}(\boldsymbol{h}_t)\rangle]^2}{n\mathbb{E}[\langle \Pi^{\perp}_{\mathcal{S}_{t-1}}(\boldsymbol{h}_t), \Pi^{\perp}_{\mathcal{S}_{t-1}}(\boldsymbol{h}_t)\rangle]} \leq \frac{1}{n}\mathbb{E}[\|\hat{\boldsymbol{\theta}}_t\|_2^2] - \sum_{s=0}^{t-1} \frac{\mathbb{E}[\langle \boldsymbol{\theta}, \Pi^{\perp}_{\mathcal{S}_{s-1}}(\boldsymbol{h}_s)\rangle]^2}{n\mathbb{E}[\|\Pi^{\perp}_{\mathcal{S}_{s-1}}(\boldsymbol{h}_s)\|_2^2]}.$$

Taking the limits on both sides of the above inequality gives $\alpha_{t+1}^2 \leq \mathbb{E}[\|\hat{\boldsymbol{\theta}}_t\|_2^2]/n - \sum_{s=1}^{t} \alpha_s^2$. By induction,

$$\begin{aligned}
\frac{1}{n}\mathbb{E}[\|\hat{\boldsymbol{\theta}}_t\|_2^2] &= \frac{1}{n}\mathbb{E}[\|\mathbb{E}[\boldsymbol{\theta} \mid \boldsymbol{r}_1, \cdots, \boldsymbol{r}_t, \boldsymbol{u}]\|_2^2] \\
&\overset{(a)}{=} \frac{1}{n}\mathbb{E}[\|\mathbb{E}[\boldsymbol{\theta} \mid \|\boldsymbol{\alpha}_{\leq t}\|_2\boldsymbol{\theta} + \boldsymbol{z}, \boldsymbol{u}]\|_2^2] \\
&\overset{(b)}{\leq} \frac{1}{n}\mathbb{E}[\|\mathbb{E}[\boldsymbol{\theta} \mid \gamma_t\boldsymbol{\theta} + \boldsymbol{z}, \boldsymbol{u}]\|_2^2] \\
&\overset{(c)}{=} \gamma_{t+1}^2,
\end{aligned}$$

where (a) follows because T_0 is a sufficient statistics for $\boldsymbol{\theta}$, (b) is by induction hypothesis and Jensen's inequality, and (c) is by the definition of γ_{t+1}. This concludes the proof of the lemma.

□

C.2 Proof of Theorem 4.5.1 under Setting 4

In this section we prove Theorem 4.5.1 under the assumptions of Setting 4. As in Section 4.4 in the main text, we will additionally assume $\boldsymbol{X}$ has sub-Gaussian entries, and relax this assumption in Appendix C.4. Namely, in this section we assume $\mathbb{E}[\exp(\lambda X_{ij})] \le \exp(C\lambda^2/n)$ for all $i \in [n]$, $j \in [d]$ and some n-independent constant C.

C.2.1 AMP algorithm

As before, the first step of our proof is to define the class of AMP algorithms for the current setting. An AMP algorithm for solving generalized linear models under Setting 4 is defined by a sequence of continuous functions (also known as the non-linearities) $\{f_t : \mathbb{R}^{t+2} \to \mathbb{R}\}_{t\ge 0}$ and $\{g_t : \mathbb{R}^{t+1} \to \mathbb{R}\}_{t\ge 1}$, and produces vectors $\{\boldsymbol{b}^t\}_{t\ge 1} \subseteq \mathbb{R}^d$ and $\{\boldsymbol{a}^t\}_{t\ge 1} \subseteq \mathbb{R}^n$ via the following iteration:

$$\begin{cases} \boldsymbol{b}^{t+1} = \boldsymbol{X}^{\mathsf{T}} f_t(\boldsymbol{a}^{\le t}; \boldsymbol{y}, \boldsymbol{u}) - \sum_{s=1}^{t} \xi_{t,s} g_s(\boldsymbol{b}^{\le s}; \boldsymbol{v}), \\ \boldsymbol{a}^{t} = \boldsymbol{X} g_t(\boldsymbol{b}^{\le t}; \boldsymbol{v}) - \sum_{s=1}^{t} \eta_{t,s} f_{s-1}(\boldsymbol{a}^{\le s-1}; \boldsymbol{y}, \boldsymbol{u}). \end{cases} \tag{C.15}$$

As before, non-linearities are applied entrywise. We denote the Onsager terms by

$$\begin{aligned} \mathsf{OC}^{t}_{\text{AMP}}(\boldsymbol{a}^{\le t-1}; \boldsymbol{y}, \boldsymbol{u}) &:= \sum_{s=1}^{t} \eta_{t,s} f_{s-1}(\boldsymbol{a}^{\le s-1}; \boldsymbol{y}, \boldsymbol{u}), \\ \mathsf{OC}^{t+1}_{\text{AMP}}(\boldsymbol{b}^{\le t}; \boldsymbol{v}) &:= \sum_{s=1}^{t} \xi_{t,s} g_s(\boldsymbol{b}^{\le s}; \boldsymbol{v}). \end{aligned}$$

The coefficients $(\xi_{t,s})_{1\le s\le t}$ and $(\eta_{t,s})_{1\le s\le t}$ are deterministic, defined via:

$$\begin{aligned} \xi_{t,s} &= \mathbb{E}\big[\partial_s f_t(\bar{\boldsymbol{G}}_{\le t}; Y, U)\big], \qquad Y := h(\bar{G}_0, W) \\ \eta_{t,s} &= \frac{1}{\delta}\mathbb{E}\big[\partial_s g_t(\boldsymbol{\mu}_{\le t}\Theta + \boldsymbol{G}_{\le t}; V)\big], \end{aligned} \tag{C.16}$$

where we use the notations $\bar{\boldsymbol{G}}_{\le t} := (\bar{G}_1, \cdots, \bar{G}_t)$, $\boldsymbol{G}_{\le t} := (G_1, \cdots, G_t)$, the joint distributions of $(\bar{\boldsymbol{G}}_{\le t}, Y, U)$ and of $(\boldsymbol{G}_{\le t}, \Theta, V)$ is defined via the following state evolution recursion

$$\begin{aligned} &(\bar{G}_0, \bar{\boldsymbol{G}}_t) \sim \mathsf{N}(\boldsymbol{0}_{t+1}, \bar{\boldsymbol{\Sigma}}_{\le t}), \qquad \boldsymbol{G}_{\le t} \sim \mathsf{N}(\boldsymbol{0}_t, \boldsymbol{\Sigma}_{\le t}), \\ &\bar{\Sigma}_{ij} = \frac{1}{\delta}\mathbb{E}[g_i(\boldsymbol{\mu}_{\le i}\Theta + \boldsymbol{G}_{\le i}; V) g_j(\boldsymbol{\mu}_{\le j}\Theta + \boldsymbol{G}_{\le j}; V)], \qquad i, j \ge 1, \\ &\bar{\Sigma}_{i0} = \bar{\Sigma}_{0i} = \frac{1}{\delta}\mathbb{E}[g_i(\boldsymbol{\mu}_{\le i}\Theta + \boldsymbol{G}_{\le i}; V)\Theta], \qquad \bar{\Sigma}_{00} = \frac{1}{\delta}\mathbb{E}[\Theta^2], \qquad i \ge 1, \\ &\Sigma_{ij} = \mathbb{E}[f_{i-1}(\bar{\boldsymbol{G}}_{\le i-1}; Y, U) f_{j-1}(\bar{\boldsymbol{G}}_{\le j-1}; Y, U)], \qquad i, j \ge 1, \\ &\mu_{t+1} = \mathbb{E}\big[\partial_{\bar{G}_0} f_t(\bar{\boldsymbol{G}}_{\le t}; Y, U)\big]. \end{aligned} \tag{C.17}$$

Here it is understood that $(\Theta, V) \sim \mu_{\Theta,V}$ is independent of $(G_i)_{i\ge 1}$ and $(W, U) \sim \mu_{W,U}$ is independent of $(\bar{G}_i)_{i\ge 0}$. Further, $\bar{\boldsymbol{\Sigma}}_{\le t} = (\bar{\Sigma}_{ij})_{0\le i,j\le t}$, $\boldsymbol{\Sigma}_{\le t} = (\Sigma_{ij})_{1\le i,j\le t}$ and $\boldsymbol{\mu}_{\le t} = (\mu_i)_{1\le i\le t}$. Here, ∂_s refers to the partial derivative with respect to the s-th variable, and $\partial_{\bar{G}_0}$ refers to the partial derivative with respect to $\bar{G}_0$. To be precise, $\partial_{\bar{G}_0} f_t(\boldsymbol{x}_{\leq t}; h(x_0, w), u) = \partial_{x_0} f_t(\boldsymbol{x}_{\leq t}; h(x_0, w), u)$. Note that f_0 depends only on (Y, U). Thus, the above recursion does not need any specific initialization. After t iterations as in Eq. (C.15), the AMP algorithm estimates $\boldsymbol{\theta}$ by applying a Lipschitz function $g_t^* : \mathbb{R}^{t+1} \to \mathbb{R}$ row-wise to $(\boldsymbol{b}^{\leq t}, \boldsymbol{v})$:

$$\hat{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{u}, \boldsymbol{v}) = g_t^*(\boldsymbol{b}^{\leq t}; \boldsymbol{v}).$$

The following theorem characterizes the asymptotic performance of the AMP iteration (C.15):

Theorem C.2.1. *Assume the matrix $\boldsymbol{X}$ and non-linearities (f_t, g_t) satisfy the same assumptions as $\boldsymbol{X}$ and $(F_t^{(1)}, G_t^{(1)})$ under either Setting 4.(a) or Setting 4.(b). Then for any $t \in \mathbb{N}_{>0}$, and any $\psi : \mathbb{R}^{t+2} \to \mathbb{R}$ pseudo-Lipschitz of order 2, the AMP iteration (C.15) satisfies*

$$\operatorname*{p-lim}_{n,d\to\infty} \frac{1}{d} \sum_{i=1}^{d} \psi(\boldsymbol{b}_i^{\leq t}, \theta_i, v_i) = \mathbb{E}[\psi(\boldsymbol{\mu}_{\leq t}\Theta + \boldsymbol{G}_{\leq t}, \Theta, V)], \qquad \boldsymbol{G}_{\leq t} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\leq t}).$$

C.2.2 Any GFOM can be reduced to an AMP algorithm

As for the case of low-rank matrix estimation, we first show that any GFOM (4.25) can be reduced to an AMP algorithm via a change of variables. The proof of the next lemma is very similar to the one of Lemma 4.4.1 and we omit it.

Lemma C.2.1. *Assume the matrix $\boldsymbol{X}$ and non-linearities $(F_t^{(1)}, F_t^{(2)}, G_t^{(1)}, G_t^{(2)}, G_*^{(t)})$ satisfy the assumptions of either Setting 4.(a) or Setting 4.(b). Then there exist functions $\{\varphi_t : \mathbb{R}^{t+1} \to \mathbb{R}\}_{t\geq 1}$, $\{\bar{\varphi}_t : \mathbb{R}^{t+2} \to \mathbb{R}\}_{t\geq 1}$, $\{f_t : \mathbb{R}^{t+2} \to \mathbb{R}\}_{t\geq 0}$ and $\{g_t : \mathbb{R}^{t+1} \to \mathbb{R}\}_{t\geq 1}$ satisfying the same assumptions such that the following holds. Let $\{\boldsymbol{a}^t\}_{t\geq 1}$ and $\{\boldsymbol{b}^t\}_{t\geq 1}$ be sequences of vectors produced by the AMP iteration (C.15) with non-linearities $\{f_t\}_{t\geq 0}$ and $\{g_t\}_{t\geq 1}$. Then for any $t \in \mathbb{N}_{>0}$, we have*

$$\boldsymbol{u}^{\leq t} = \bar{\varphi}_t(\boldsymbol{a}^{\leq t}; \boldsymbol{y}, \boldsymbol{u}), \qquad \boldsymbol{v}^{\leq t} = \varphi_t(\boldsymbol{b}^{\leq t}; \boldsymbol{v}).$$

Lemma C.2.1 implies that the class of AMP algorithms achieve the same minimum expected error as the class of GFOM for the same number of iterations under any loss. This is formalized by the next corollary, which is analogous to Corollary 4.4.1.

Corollary C.2.1. *Let $\mathcal{A}^t_{\rm GFOM}$ be the class of GFOM estimators with t iterations, and $\mathcal{A}^t_{\rm AMP}$ be the class of AMP algorithms with t iterations (under the assumptions of either Setting 4.(a), or Setting 4.(b)). (In particular $\hat{\boldsymbol{\theta}}(\,\cdot\,) \in \mathcal{A}^t_{\rm GFOM}$ is defined by a set of n-independent functions $\{F_t^{(1)}, F_t^{(2)}, G_{t+1}^{(1)}, G_{t+1}^{(2)}, G_*^{(t+1)}\}_{t\in\mathbb{N}}$, and similarly for $\hat{\boldsymbol{\theta}}(\,\cdot\,) \in \mathcal{A}^t_{\rm GFOM}$.)*

Then for any loss function $\mathcal{L} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$:

$$\inf_{\hat{\boldsymbol{\theta}}(\,\cdot\,)\in\mathcal{A}^t_{\rm GFOM}} \operatorname*{p-liminf}_{n\to\infty} \mathcal{L}(\hat{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{u}, \boldsymbol{v}), \boldsymbol{\theta}) = \inf_{\hat{\boldsymbol{\theta}}(\,\cdot\,)\in\mathcal{A}^t_{\rm AMP}} \operatorname*{p-liminf}_{n\to\infty} \mathcal{L}(\hat{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{u}, \boldsymbol{v}), \boldsymbol{\theta})\,. \tag{C.18}$$

C.2.3 Orthogonalization

In this section we show that we can further restrict ourselves to lower bounding the error of orthogonal AMP (OAMP) algorithms.

Lemma C.2.2. *Let $\{\boldsymbol{a}^t\}_{t\geq 1}$, $\{\boldsymbol{b}^t\}_{t\geq 1}$ be sequences produced by the AMP iteration (C.15) under either Setting 4.(a) or Setting 4.(b). Then there exist functions $\{\phi_t : \mathbb{R}^{t+1} \to \mathbb{R}^t\}_{t\geq 1}$ satisfying the same assumptions as the non-linearities in the AMP iteration, such that the following holds:*

(i) *For all $t \in \mathbb{N}_{>0}$ we have $\boldsymbol{b}^{\leq t} = \phi_t(\boldsymbol{q}^{\leq t}; \boldsymbol{v})$.*

(ii) *For any $\psi : \mathbb{R}^{t+2} \to \mathbb{R}$ pseudo-Lipschitz of order 2,*

$$\operatorname*{p-lim}_{n,d\to\infty} \frac{1}{d}\sum_{i=1}^{d} \psi(q_i^1, \cdots, q_i^t, v_i, \theta_i) = \mathbb{E}[\psi(Q_1, \cdots, Q_t, V, \Theta)],$$

where $Q_i = x_{i-1}(\alpha_i \Theta + Z_i)$ with $(x_0, \cdots, x_{t-1}) \in \{0,1\}^t$ and $(\alpha_1, \cdots, \alpha_t) \in \mathbb{R}^t$ deterministic vectors, and $(Z_i)_{i\geq 1} \overset{iid}{\sim} \mathsf{N}(0,1)$ independent of (Θ, V).

Proof. Given the state evolution of the AMP iteration defined via Eq. (C.17), we let

$$Y_t := f_t(\bar{\boldsymbol{G}}_{\leq t}; Y, U), \qquad \mathcal{S}_t = \mathrm{span}(Y_k : 0 \leq k \leq t)\,, \quad Y = h(\bar{G}_0; W).$$

Note that by state evolution, $\mathbb{E}[Y_t Y_s] = \Sigma_{t+1,s+1}$. By linear algebra, for all $t \in \mathbb{N}$, there exist deterministic constants $\{c_{ts}\}_{0\leq s\leq t}$ and $x_t \in \{0,1\}$, such that $c_{tt} \neq 0$ and

$$R_t := c_{tt}\Pi^{\perp}_{\mathcal{S}_{t-1}}(Y_t) = \sum_{s=0}^{t} c_{ts} Y_s, \qquad \mathbb{E}[R_t R_s] = \mathbb{1}_{s=t} x_t.$$

Indeed, proceeding by induction, if Y_t does not belong to $\mathcal{S}_{t-1}$, then we can take $x_t = 1$ and $c_{tt} = \|\Pi^{\perp}_{\mathcal{S}_{t-1}}(Y_t)\|_{L^2}^{-1}$. Otherwise we take $R_t = 0$, $c_{tt} = 1$ and $x_t = 0$.

We prove the lemma by induction. For the base case $t = 1$, we let $\boldsymbol{q}^1 = c_{00}\boldsymbol{b}^1$, thus, claim *(i)* follows. As for claim *(ii)*, we consider two cases. If $x_0 = 0$, then $\mathbb{E}[f_0(Y,U)^2] = 0$. By Stein's lemma, $\mathbb{E}[\partial_{\bar{G}_0} f_0(h(\bar{G}_0, W), U)] = \mathbb{E}[\bar{G}_0 f_0(h(\bar{G}_0, W), U)]/\operatorname{Var}[\bar{G}_0] = 0$. Thus, claim *(ii)* holds with $Q_1 \equiv 0$. If $x_0 = 1$, then $c_{00} = \mathbb{E}[f_0(Y,U)^2]^{1/2}$, and claim (ii) follows from state evolution (C.17) with

$$\alpha_1 = \frac{\mathbb{E}[\partial_{\bar{G}_0} f_0(h(\bar{G}_0, W), U)]}{\mathbb{E}[f_0(h(\bar{G}_0, W), U)^2]^{1/2}} \overset{(a)}{=} \frac{\mathbb{E}[\bar{G}_0 f_0(h(\bar{G}_0, W), U)]}{\operatorname{Var}[\bar{G}_0]\mathbb{E}[f_0(h(\bar{G}_0, W), U)^2]^{1/2}}. \tag{C.19}$$

where (a) holds by Stein's lemma.

Suppose the lemma holds for the first t iterations, then we prove it also holds for the $(t+1)$-th iteration. We let $\boldsymbol{q}^{t+1} = \sum_{s=0}^{t} c_{ts}\boldsymbol{b}^{s+1}$. Since $c_{tt} \neq 0$, we can solve for $\boldsymbol{b}^{t+1}$. Thus, we obtain the transformation ϕ_{t+1} that satisfies the desired properties. As a consequence, claim *(i)* follows.

As for claim *(ii)*, first notice that the mapping

$$(b_1, \cdots, b_t, v, \theta) \mapsto \psi(c_{00}b_1, \cdots, \textstyle\sum_{s=0}^{t-1} c_{t-1,s} b_{s+1}, v, \theta)$$

is pseudo-Lipschitz of order two. Then we consider two cases. In the first case $x_t = 0$, then $R_t \overset{a.s.}{=} 0$. By state evolution (C.17) and an application of Stein's lemma, we obtain that *(ii)* holds with $Q_{t+1} \overset{a.s.}{=} 0$. In the second case, $x_t = 1$, then again by the state evolution (C.17), $Q_{t+1} \overset{d}{=} \alpha_{t+1}\Theta + Z_{t+1}$, where

$$\alpha_{t+1} = \frac{\mathbb{E}[\partial_{\bar{G}_0}\Pi^{\perp}_{\mathcal{S}_{t-1}}(Y_t)]}{\mathbb{E}[\Pi^{\perp}_{\mathcal{S}_{t-1}}(Y_t)^2]^{1/2}} \overset{(b)}{=} \frac{\mathbb{E}[\bar{G}_0^{\perp,t}\Pi^{\perp}_{\mathcal{S}_{t-1}}(Y_t)]}{\mathrm{Var}[\bar{G}_0^{\perp,t}]\mathbb{E}[\Pi^{\perp}_{\mathcal{S}_{t-1}}(Y_t)^2]^{1/2}}. \tag{C.20}$$

Here, $\bar{G}_0^{\perp,t} = \Pi^{\perp}_{\bar{\mathcal{G}}_t}(\bar{G}_0)$ with $\bar{\mathcal{G}}_t = \mathrm{span}(\bar{G}_i : 1 \le i \le t)$ and (b) follows from Stein's lemma. Thus, we complete the proof by induction.

□

By similar arguments as discussed in Remark 4.4.3, in the following parts of the paper, we will set $x_t = 1$ for all $t \in \mathbb{N}$ without loss of generality.

C.2.4 Optimal orthogonal AMP

Recall that a sufficient statistics for $\boldsymbol{\Theta}$ given $\boldsymbol{S}_{\le t} := \boldsymbol{\alpha}_{\le t}\Theta + \boldsymbol{Z}_{\le t}$ is $T_0 := \langle \boldsymbol{\alpha}_{\le t}, \boldsymbol{S}_{\le t}\rangle / \|\boldsymbol{\alpha}_{\le t}\|_2$, and T_0 can be rewritten as:

$$T_0 = \|\boldsymbol{\alpha}_{\le t}\|_2\Theta + G\,, \qquad G \sim \mathsf{N}(0,1)\,, \quad G \perp \Theta\,. \tag{C.21}$$

Further $\boldsymbol{S}_{\le t}$ and V are conditionally independent, given Θ. Hence, the proof of Theorem 4.5.1 follows exactly as for Theorem 4.3.1, once we upper bound the value of $\|\boldsymbol{\alpha}_{\le t}\|_2$ achieved by any OAMP algorithm. Before proving such a bound, we establish some useful identities.

Lemma C.2.3. *Recall that* $(\bar{G}_0, \bar{\boldsymbol{G}}_{\le t}) \sim \mathsf{N}(\mathbf{0}_{t+1}, \bar{\boldsymbol{\Sigma}}_{\le t})$*, where*

$$\bar{\Sigma}_{ij} = \frac{1}{\delta}\mathbb{E}[g_i(\phi_i(\boldsymbol{\alpha}_{\le i}\Theta + \boldsymbol{Z}_{\le i}; V); V)g_j(\phi_j(\boldsymbol{\alpha}_{\le j}\Theta + \boldsymbol{Z}_{\le j}; V); V)] \tag{C.22}$$

with $(Z_i)_{i\ge 1} \sim_{i.i.d.} \mathsf{N}(0,1)$*. Further recall that* $\bar{G}_0^{\perp,t} = \Pi^{\perp}_{\bar{\mathcal{G}}_t}(\bar{G}_0)$ *with* $\bar{\mathcal{G}}_t = span(\bar{G}_i : 1 \le i \le t)$*. Define*

$$\omega_t^2 := \mathrm{Var}[\bar{G}_0^{\perp,t}], \qquad \zeta_t^2 := \frac{1}{\delta}(\mathbb{E}[\Theta^2] - \omega_t^2). \tag{C.23}$$

Then, the following holds for all $s, t \in \mathbb{N}$ *with* $s \le t$*,*

$$\mathbb{E}[\bar{G}_0^{\perp,t} \mid h(\bar{G}_0, W), U, \bar{\boldsymbol{G}}_{\le t}] \overset{d}{=} \mathbb{E}[\omega_t Z_0 \mid h(\omega_t Z_0 + \zeta_t Z_1, W), U, Z_1],$$
$$\mathbb{E}[\bar{G}_0^{\perp,t} \mid h(\bar{G}_0, W), U, \bar{\boldsymbol{G}}_{\le s}] = \frac{\omega_t^2}{\omega_s^2}\mathbb{E}[\bar{G}_0^{\perp,s} \mid h(\bar{G}_0, W), U, \bar{\boldsymbol{G}}_{\le s}],$$

where $Z_0, Z_1 \overset{iid}{\sim} \mathsf{N}(0,1)$*,*

Proof. We let $\bar{G}_0^{\|,t} := \bar{G}_0 - \bar{G}_0^{\perp,t}$, then we can write $\bar{G}_0^{\|,t}$ as a deterministic function of $\bar{\boldsymbol{G}}_{\le t}$, and we denote this function by $\bar{G}_0^{\|,t} = c_t(\bar{\boldsymbol{G}}_{\le t})$. For $s \le t$, we observe that $(\bar{G}_0^{\perp,t}, \bar{G}_0^{\|,t} - \bar{G}_0^{\|,s}, \bar{G}_0^{\|,s}) \sim \mathsf{N}(\mathbf{0}, \mathrm{diag}((\omega_t^2, \omega_s^2 - \omega_t^2, \zeta_s^2)))$. In the following parts, with a slight abuse of notations, we use p to represent probability density functions for various distributions. Then the following formula regarding the conditional probability density holds:

$$p(\bar{G}_0^{\perp,t} = z \mid h(\bar{G}_0, W) = h, U = u, \bar{\boldsymbol{G}}_{\le s} = z_{\le s})$$

$$
\begin{aligned}
&\propto \int p(\bar{\boldsymbol{G}}_{\le s} = z_{\le s}) p(\bar{G}_0^{\perp,t} = z) \mathbb{1}\{h(z + c_s(z_{\le s}) + y, w) = h\} \mu_{W|U=u}(\mathrm{d}w)\phi(y)\mathrm{d}y \\
&\propto \int p(\bar{G}_0^{\perp,t} = z) \mathbb{1}\{h(z + c_s(z_{\le s}) + y, w) = h\} \mu_{W|U=u}(\mathrm{d}w)\phi(y)\mathrm{d}y \\
&\propto \int p(\bar{G}_0^{\|,s} = c_s(z_{\le s})) p(\bar{G}_0^{\perp,t} = z) \mathbb{1}\{h(z + c(z_{\le t}) + y, w) = h\} \mu_{W|U=u}(\mathrm{d}w)\phi(y)\mathrm{d}y \\
&\propto p(\bar{G}_0^{\perp,t} = z \mid h(\bar{G}_0, W) = h, U = u, \bar{G}_0^{\|,s} = c_s(z_{\le s})), \qquad \text{(C.24)}
\end{aligned}
$$

where ϕ is the probability density function for $\mathsf{N}(0, \omega_s^2 - \omega_t^2)$. Notice that $(\bar{G}_0^{\perp,t}, \bar{G}_0^{\|,t}, U, W) \overset{d}{=} (\omega_t Z_0, \zeta_t Z_1, U, W)$, therefore, we take $s = t$ in Eq. (C.24) and conclude that

$$
\mathbb{E}[\bar{G}_0^{\perp,t} \mid h(\bar{G}_0, W), U, \bar{\boldsymbol{G}}_{\le t}] = \mathbb{E}[\bar{G}_0^{\perp,t} \mid h(\bar{G}_0, W), U, \bar{G}_0^{\|,t}] \overset{d}{=} \mathbb{E}[\omega_t Z_0 \mid h(\omega_t Z_0 + \zeta_t Z_1, W), U, Z_1],
$$

which completes the proof of the first claim.

As for the second claim, notice that there exists $Z_2, Z_3, Z_4 \overset{iid}{\sim} \mathsf{N}(0,1)$, such that $(\bar{G}_0^{\perp,t}, \bar{G}_0^{\|,t} - \bar{G}_0^{\|,s}, \bar{G}_0^{\|,s}) = (\omega_t Z_2, \sqrt{\omega_s^2 - \omega_t^2} Z_3, \zeta_s Z_4)$. Therefore, using Eq. (C.24), we have

$$
\begin{aligned}
\mathbb{E}[\bar{G}_0^{\perp,t} \mid h(\bar{G}_0, W), U, \bar{\boldsymbol{G}}_{\le s}] =& \mathbb{E}[\bar{G}_0^{\perp,t} \mid h(\bar{G}_0, W), U, \bar{G}_0^{\|,s}] \\
=& \mathbb{E}[\omega_t Z_2 \mid h(\omega_t Z_2 + \sqrt{\omega_s^2 - \omega_t^2} Z_3 + \zeta_s Z_4, W), U, Z_4] \\
\overset{(a)}{=}& \frac{\omega_t^2}{\omega_s^2} \mathbb{E}\big[\omega_t Z_2 + \sqrt{\omega_s^2 - \omega_t^2} Z_3 \mid h(\omega_t Z_2 + \sqrt{\omega_s^2 - \omega_t^2} Z_3 + \zeta_s Z_4, W), U, Z_4\big] \\
=& \frac{\omega_t^2}{\omega_s^2} \mathbb{E}[\bar{G}_0^{\perp,s} \mid h(\bar{G}_0, W), U, \bar{G}_0^{\|,s}] \\
\overset{(b)}{=}& \frac{\omega_t^2}{\omega_s^2} \mathbb{E}[\bar{G}_0^{\perp,s} \mid h(\bar{G}_0, W), U, \bar{\boldsymbol{G}}_{\le s}],
\end{aligned}
$$

where (a) is by Lemma C.2.6, and (b) is by Eq. (C.24). Thus, we complete the proof of the lemma.

□

The next lemma proves the desired upper bound on $\|\boldsymbol{\alpha}_{\le t}\|_2$.

Lemma C.2.4. *Recall the definition of $\{\beta_t\}$ in Eq. (4.26). Then for all $t \in \mathbb{N}_{>0}$ and all AMP algorithms we have $\|\boldsymbol{\alpha}_{\le t}\|_2 \le \beta_t$.*

Proof. Recall the definition of ω_t, ζ_t in Eq. (C.23), and of $(\sigma_t)_{t \in \mathbb{N}_{>0}}$ in Eq. (4.26). We will prove the following claims by induction over t: $\|\boldsymbol{\alpha}_{\le t}\|_2 \le \beta_t$ and $\omega_{t-1} \ge \sigma_t$.

For the base case $t = 1$, $\omega_0 \ge \sigma_1$ holds by definition. Using Eq. (C.19) we have

$$
\alpha_1^2 = \frac{\mathbb{E}[\bar{G}_0 f_0(h(\bar{G}_0, W), U)]^2}{\mathrm{Var}[\bar{G}_0]^2 \mathbb{E}[f_0(h(\bar{G}_0, W), U)^2]} \le \sup_{X \in \sigma\{h(\bar{G}_0, W), U\}} \frac{\mathbb{E}[\bar{G}_0 X]^2}{\mathrm{Var}[\bar{G}_0]^2 \mathbb{E}[X^2]} \le \frac{1}{\sigma_1^2} \mathbb{E}[\mathbb{E}[Z_0 \mid h(\sigma_1 Z_0, W), U]^2],
$$

where $Z_0 \sim \mathsf{N}(0,1)$ and the last step follows from Cauchy-Schwarz inequality.

Next we assume the induction claim holds for the first t iterations, and we prove it holds for the $(t+1)$-th iteration. Notice that the random variables $\{Y_0/\mathbb{E}[Y_0^2]^{1/2}, \cdots, \Pi_{\mathcal{S}_{t-1}}^{\perp}(Y_t)/\mathbb{E}[\Pi_{\mathcal{S}_{t-1}}^{\perp}(Y_t)^2]^{1/2}\}$ are orthonormal.

Then we have:

$$\begin{aligned}
\alpha_{t+1}^2 =& \frac{\mathbb{E}[\mathbb{E}[\bar{G}_0^{\perp,t} \mid h(\bar{G}_0, W), \bar{\boldsymbol{G}}_{\le t}, U]\, \Pi^{\perp}_{\mathcal{S}_{t-1}}(Y_t)]^2}{\omega_t^4 \mathbb{E}[\Pi^{\perp}_{\mathcal{S}_{t-1}}(Y_t)^2]} \\
\overset{(a)}{\le}& \frac{1}{\omega_t^4}\mathbb{E}[\mathbb{E}[\bar{G}_0^{\perp,t} \mid h(\bar{G}_0, W), \bar{\boldsymbol{G}}_{\le t}, U]^2] - \sum_{s=0}^{t-1} \frac{\mathbb{E}[\bar{G}_0^{\perp,t}\Pi^{\perp}_{\mathcal{S}_{s-1}}(Y_s)]^2}{\omega_t^4 \mathbb{E}[\Pi^{\perp}_{\mathcal{S}_{s-1}}(Y_s)^2]} \\
\overset{(b)}{=}& \frac{1}{\omega_t^2}\mathbb{E}[\mathbb{E}[Z_0 \mid h(\omega_t Z_0 + \zeta_t Z_1, W), U, Z_1]^2] - \sum_{s=0}^{t-1} \frac{\mathbb{E}[\bar{G}_0^{\perp,s}\Pi^{\perp}_{\mathcal{S}_{s-1}}(Y_s)]^2}{\omega_s^4 \mathbb{E}[\Pi^{\perp}_{\mathcal{S}_{s-1}}(Y_s)^2]} \\
\overset{(c)}{\le}& \frac{1}{\sigma_{t+1}^2}\mathbb{E}[\mathbb{E}[Z_0 \mid h(\sigma_{t+1} Z_0 + \tilde{\sigma}_{t+1} Z_1, W), U, Z_1]^2] - \sum_{s=1}^{t} \alpha_s^2,
\end{aligned}$$

where (a) holds by Eq. (C.20) and Pythagora's theorem, (b) by Lemma C.2.3, and (c) is by induction hypothesis and Lemma C.2.5. The last inequality above gives $\sum_{s=1}^{t+1} \alpha_s^2 \le \beta_{t+1}^2$.

For $t \in \mathbb{N}_{>0}$ we define

$$Y_t' := g_t(\phi_t(\boldsymbol{\alpha}_{\le t}\Theta + \boldsymbol{Z}_{\le t}; V); V), \qquad \mathcal{S}_t' := \operatorname{span}(Y_i' : 1 \le i \le t).$$

By state evolution (C.17), $\omega_{t+1}^2 = \mathbb{E}[\Pi^{\perp}_{\mathcal{S}'_{t+1}}(\Theta)^2]/\delta$. Further we have

$$\begin{aligned}
\omega_{t+1}^2 \overset{(d)}{=}& \frac{1}{\delta}\mathbb{E}[\Theta^2] - \frac{1}{\delta}\mathbb{E}[\Pi_{\mathcal{S}'_{t+1}}(\Theta)^2] \\
\overset{(e)}{\ge}& \frac{1}{\delta}\mathbb{E}[\Theta^2] - \frac{1}{\delta}\mathbb{E}[\mathbb{E}[\Theta \mid \boldsymbol{\alpha}_{\le t+1}\Theta + \boldsymbol{Z}_{\le t+1}, V]^2] \\
\overset{(f)}{=}& \frac{1}{\delta}\mathbb{E}[\Theta^2] - \frac{1}{\delta}\mathbb{E}[\mathbb{E}[\Theta \mid \|\boldsymbol{\alpha}_{\le t+1}\|_2\Theta + G, V]^2] \\
\overset{(g)}{\ge}& \frac{1}{\delta}\mathbb{E}[\Theta^2] - \frac{1}{\delta}\mathbb{E}[\mathbb{E}[\Theta \mid \beta_{t+1}\Theta + G, V]^2] = \sigma_{t+2}^2,
\end{aligned}$$

where (d) holds by Pythagora's theorem, (e) by Jensen's inequality, (f) by property of sufficient statistics and (g) is by induction hypothesis and Jensen's inequality.

This completes the proof of the lemma by induction.

□

Lemma C.2.5. *Let* $Z_0, Z_1 \overset{iid}{\sim} \mathsf{N}(0,1)$*. For any fixed* $\omega_0^2 \ge 0$*, the following function is non-increasing in* $a \in (0, \omega_0^2])$*:*

$$a \mapsto \frac{1}{a^2}\mathbb{E}[\mathbb{E}[Z_0 \mid h(aZ_0 + (\omega_0^2 - a^2)^{1/2} Z_1, W), U, Z_1]^2].$$

Proof. For $\delta > 0$, we introduce the decomposition $Z_1 = \delta Z_2 + \sqrt{1-\delta^2} Z_3$, with $Z_2, Z_3 \overset{iid}{\sim} \mathsf{N}(0,1)$ that are independent of Z_0. Then by Jensen's inequality,

$$\begin{aligned}
&\frac{1}{a^2}\mathbb{E}[\mathbb{E}[Z_0 \mid h(aZ_0 + (\omega_0^2 - a^2)^{1/2} Z_1, W), U, Z_1]^2] \\
=& \frac{1}{a^2}\mathbb{E}[\mathbb{E}[Z_0 \mid h(aZ_0 + (\omega_0^2 - a^2)^{1/2}\delta Z_2 + ((\omega_0^2 - a^2)(1-\delta^2))^{1/2} Z_3, W), U, Z_2, Z_3]^2] \\
\ge& \frac{1}{a^2}\mathbb{E}[\mathbb{E}[Z_0 \mid h(aZ_0 + (\omega_0^2 - a^2)^{1/2}\delta Z_2 + ((\omega_0^2 - a^2)(1-\delta^2))^{1/2} Z_3, W), U, Z_3]^2]
\end{aligned}$$

$$=\frac{1}{a^2+\delta^2(\omega_0^2-a^2)}\mathbb{E}[\mathbb{E}[Z_0 \mid h((a^2+\delta^2(\omega_0^2-a^2))^{1/2}Z_0+((\omega_0^2-a^2)(1-\delta^2))^{1/2}Z_3,W),U,Z_3]^2].$$

The above inequality holds for all $\delta \in [0,1]$, thus completes the proof of the lemma.

□

Lemma C.2.6. *We let Z_1, Z_2 be independent mean-zero Gaussian random variables with variance σ_1^2 and σ_2^2, respectively. For $\sigma_1^2 \ge q \ge 0$, we let G_q be a mean-zero Gaussian random variable such that $\mathrm{Cov}(G_q, Z_2) = 0$ and $\mathrm{Var}(G_q) = \mathrm{Cov}(G_q, Z_1) = q$. Then for all $h : \mathbb{R}^2 \to \mathbb{R}$, we have*

$$f_h(q) := \mathbb{E}[G_q \mid h(Z_1+Z_2,W),Z_2] = \frac{q}{\sigma_1^2}\mathbb{E}[Z_1 \mid h(Z_1+Z_2,W),Z_2].$$

Proof. For $q_1, q_2 \ge 0$ with $q_1 + q_2 \le \sigma_1^2$, there exist G_{q_1}, G_{q_2} independent of each other, and satisfy the above constraints. Then, we have $\mathrm{Cov}(G_{q_1}+G_{q_2}, Z_2) = 0$, $\mathrm{Cov}(G_{q_1}+G_{q_2}, Z_1) = \mathrm{Var}(G_{q_1}+G_{q_2}) = q_1 + q_2$. Therefore,

$$f_h(q_1+q_2) = \mathbb{E}[G_{q_1}+G_{q_2} \mid h(Z_1+Z_2,W),Z_2] = f_h(q_1)+f_h(q_2).$$

For all fixed $(h(Z_1+Z_2,W),Z_2)$, f_h is continuous, thus the lemma follows from Cauchy's equation.

□

C.3 Proof of Theorem 4.5.1 under Setting 3

In this section we prove Theorem 4.5.1 under the assumptions of Setting 3.

C.3.1 AMP algorithm

As in previous proofs, we start with the definition of AMP algorithms with non-separable non-linearities. Under Setting 3, an AMP algorithm for solving generalized linear models is defined by a sequence of uniformly Lipschitz functions $\{f_t : \mathbb{R}^{n(t+2)} \to \mathbb{R}^n\}_{t\ge 0}$ and $\{g_t : \mathbb{R}^{d(t+1)} \to \mathbb{R}^d\}_{t\ge 1}$, and produces $\{\boldsymbol{b}^t\}_{t\ge1} \subseteq \mathbb{R}^d$ and $\{\boldsymbol{a}^t\}_{t\ge1} \subseteq \mathbb{R}^n$ via the following iteration:

$$\begin{cases} \boldsymbol{b}^{t+1} = \boldsymbol{X}^{\mathsf{T}} f_t(\boldsymbol{a}^{\le t}; \boldsymbol{y}, \boldsymbol{u}) - \sum\limits_{s=1}^{t} \xi_{t,s} g_s(\boldsymbol{b}^{\le s}; \boldsymbol{v}), \\ \boldsymbol{a}^{t} = \boldsymbol{X} g_t(\boldsymbol{b}^{\le t}; \boldsymbol{v}) - \sum\limits_{s=1}^{t} \eta_{t,s} f_{s-1}(\boldsymbol{a}^{\le s-1}; \boldsymbol{y}, \boldsymbol{u}). \end{cases} \tag{C.25}$$

Here, $(\xi_{t,s})_{1\le s\le t}$ and $(\eta_{t,s})_{1\le s\le t}$ are deterministic coefficients defined via

$$\begin{aligned} \xi_{t,s} &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[\partial_{i,s} f_{t,i}(\bar{\boldsymbol{g}}_{\le t}; \boldsymbol{y}_*, \boldsymbol{u})\big], \qquad \boldsymbol{y}_* := h(\bar{\boldsymbol{g}}_0, \boldsymbol{w}) \\ \eta_{t,s} &= \frac{1}{n}\sum_{i=1}^{d} \mathbb{E}\big[\partial_{i,s} g_{t,i}(\boldsymbol{\mu}_{\le t}\boldsymbol{\theta} + \boldsymbol{g}_{\le t}; \boldsymbol{v})\big]. \end{aligned} \tag{C.26}$$

Here we introduced the notations $\bar{\boldsymbol{g}}_{\le t} := (\bar{\boldsymbol{g}}_1, \cdots, \bar{\boldsymbol{g}}_t) \in \mathbb{R}^{n\times t}$, $\boldsymbol{g}_{\le t} := (\boldsymbol{g}_1, \cdots, \boldsymbol{g}_t) \in \mathbb{R}^{d\times t}$, and the joint distributions of $(\boldsymbol{\theta}, \boldsymbol{v}, (\boldsymbol{g}_i)_{i\ge 1})$ and of $(\boldsymbol{y}_*, \boldsymbol{u}, \boldsymbol{w}, (\bar{\boldsymbol{g}}_i)_{i\ge 0})$ are determined by the following state evolution recursions

$$\begin{aligned}
&(\bar{\boldsymbol{g}}_0, \bar{\boldsymbol{g}}_{\le t}) \sim \mathsf{N}(\boldsymbol{0}, \bar{\boldsymbol{\Sigma}}_{\le t+1} \otimes \boldsymbol{I}_n), \qquad \boldsymbol{g}_{\le t} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\le t} \otimes \boldsymbol{I}_d),\\
&\bar{\Sigma}_{ij} = \lim_{n,d\to\infty} \frac{1}{n}\mathbb{E}[g_i(\boldsymbol{\mu}_{\le i}\boldsymbol{\theta} + \boldsymbol{g}_{\le i}; \boldsymbol{v})^{\mathsf{T}} g_j(\boldsymbol{\mu}_{\le j}\boldsymbol{\theta} + \boldsymbol{g}_{\le j}; \boldsymbol{v})], \qquad i, j \ge 1,\\
&\bar{\Sigma}_{i0} = \bar{\Sigma}_{0i} = \lim_{n,d\to\infty} \frac{1}{n}\mathbb{E}[g_i(\boldsymbol{\mu}_{\le i}\boldsymbol{\theta} + \boldsymbol{g}_{\le i}; \boldsymbol{v})^{\mathsf{T}} \boldsymbol{\theta}], \qquad \bar{\Sigma}_{00} = \frac{1}{\delta}\mathbb{E}[\Theta^2], \qquad i \ge 1.\\
&\Sigma_{ij} = \lim_{n,d\to\infty} \frac{1}{n}\mathbb{E}[f_{i-1}(\bar{\boldsymbol{g}}_{\le i-1}; \boldsymbol{y}_*, \boldsymbol{u})^{\mathsf{T}} f_{j-1}(\bar{\boldsymbol{g}}_{\le j-1}; \boldsymbol{y}_*, \boldsymbol{u})],\\
&\mu_{t+1} = \lim_{n,d\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[\partial_{\bar{g}_{0,i}} f_{t,i}(\bar{\boldsymbol{g}}_{\le t}; \boldsymbol{y}_*, \boldsymbol{u})\big].
\end{aligned} \tag{C.27}$$

In the above equations $\boldsymbol{\Sigma}_{\le t} = (\Sigma_{ij})_{1\le i,j\le t}$, $\bar{\boldsymbol{\Sigma}}_{\le t} = (\bar{\Sigma}_{ij})_{0\le i,j\le t}$ and $\boldsymbol{\mu}_{\le t} = (\mu_i)_{1\le i\le t}$, and the limits are assumed to exist. Here, $\partial_{i,s}$ refers to the partial derivative with respect to the s-th variable of the i-th row of the input matrix, and $\partial_{\bar{g}_{0,i}}$ refers to the partial derivative with respect to $\bar{g}_{0,i}$. Note that f_0 depends only on $(\boldsymbol{y}_*, \boldsymbol{u})$, thus, the state evolution does not need any specific initialization. After t iterations as in Eq. (C.25), the AMP algorithm estimates $\boldsymbol{\theta}$ by applying a uniformly Lipschitz function $g_t^* : \mathbb{R}^{d(t+1)} \to \mathbb{R}^d$ to $(\boldsymbol{b}^{\le t}, \boldsymbol{v})$:

$$\hat{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{u}, \boldsymbol{v}) = g_t^*(\boldsymbol{b}^{\le t}; \boldsymbol{v}).$$

The following theorem describes the state evolution of the AMP iteration (C.25).

Theorem C.3.1. *Assume $X_{ij} \overset{iid}{\sim} \mathsf{N}(0, 1/n)$ for all $i \in [n]$ and $j \in [d]$, $(\theta_i, v_i)_{i\le d} \overset{iid}{\sim} \mu_{\Theta,V}$, $(w_i, u_i)_{i\le n} \overset{iid}{\sim} \mu_{W,U}$, and for all $t \in \mathbb{N}$, the non-linearities (f_t, g_{t+1}) are uniformly Lipschitz. Furthermore, we assume the following limits exist for all $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \bar{\boldsymbol{\Sigma}})$:*

$$\begin{aligned}
&\lim_{n,d\to\infty} \frac{1}{n}\mathbb{E}[f_t(\bar{\boldsymbol{g}}_{\le t}; \boldsymbol{y}_*, \boldsymbol{u})^{\mathsf{T}} f_s(\bar{\boldsymbol{g}}_{\le s}; \boldsymbol{y}_*, \boldsymbol{u})],\\
&\lim_{n,d\to\infty} \frac{1}{n}\mathbb{E}[f_t(\bar{\boldsymbol{g}}_{\le t}; \boldsymbol{y}_*, \boldsymbol{u})^{\mathsf{T}} \bar{\boldsymbol{g}}_0],\\
&\lim_{n,d\to\infty} \frac{1}{d}\mathbb{E}[g_t(\boldsymbol{\mu}_{\le t}\boldsymbol{\theta} + \boldsymbol{g}_{\le t}; \boldsymbol{v})^{\mathsf{T}} g_s(\boldsymbol{\mu}_{\le s}\boldsymbol{\theta} + \boldsymbol{g}_{\le s}; \boldsymbol{v})],\\
&\lim_{n,d\to\infty} \frac{1}{d}\mathbb{E}[g_t(\boldsymbol{\mu}_{\le t}\boldsymbol{\theta} + \boldsymbol{g}_{\le t}; \boldsymbol{v})^{\mathsf{T}} \boldsymbol{\theta}],\\
&\lim_{n,d\to\infty} \frac{1}{d}\mathbb{E}[g_t^*(\boldsymbol{\mu}_{\le t}\boldsymbol{\theta} + \boldsymbol{g}_{\le t}; \boldsymbol{v})^{\mathsf{T}} g_s^*(\boldsymbol{\mu}_{\le s}\boldsymbol{\theta} + \boldsymbol{g}_{\le s}; \boldsymbol{v})],\\
&\lim_{n,d\to\infty} \frac{1}{d}\mathbb{E}[g_t^*(\boldsymbol{\mu}_{\le t}\boldsymbol{\theta} + \boldsymbol{g}_{\le t}; \boldsymbol{v})^{\mathsf{T}} \boldsymbol{\theta}].
\end{aligned}$$

Then for $\{\psi_n : \mathbb{R}^{d(t+2)} \to \mathbb{R}\}_{n\ge 1}$ uniformly pseudo-Lipschitz of order 2,

$$\psi_n(\boldsymbol{b}^{\le t}, \boldsymbol{\theta}, \boldsymbol{v}) = \mathbb{E}[\psi_n(\boldsymbol{\mu}_{\le t}\boldsymbol{\theta} + \boldsymbol{g}_{\le t}, \boldsymbol{\theta}, \boldsymbol{v})] + o_P(1).$$

C.3.2 Any GFOM can be reduced to an AMP algorithm

Again we show that GFOM (4.25) can be reduced to an AMP algorithm (C.25) under Setting 3. To be specific, we have the following lemma:

Lemma C.3.1. *Under the assumptions of Setting 3, for all $t \in \mathbb{N}_{>0}$, there exist uniformly Lipschitz functions $\varphi_t : \mathbb{R}^{d(t+1)} \to \mathbb{R}^{dt}$, $\bar{\varphi}_t : \mathbb{R}^{n(t+2)} \to \mathbb{R}^{nt}$, $f_{t-1} : \mathbb{R}^{n(t+1)} \to \mathbb{R}^n$ and $g_t : \mathbb{R}^{d(t+1)} \to \mathbb{R}^d$ that satisfy the following conditions. We let $\{\boldsymbol{a}^t\}_{t\geq 1}$ and $\{\boldsymbol{b}^t\}_{t\geq 1}$ be sequences of vectors produced by the AMP iteration (C.25) with non-linearities $\{f_t\}_{t\geq 0}$ and $\{g_t\}_{t\geq 1}$. Then for any $t \in \mathbb{N}_{>0}$, we have*

$$\begin{aligned}
&\boldsymbol{u}^{\leq t} = \bar{\varphi}_t(\boldsymbol{a}^{\leq t}; \boldsymbol{y}, \boldsymbol{u}), \qquad \boldsymbol{v}^{\leq t} = \varphi_t(\boldsymbol{b}^{\leq t}; \boldsymbol{v}),\\
&f_{t-1}(\boldsymbol{a}^{\leq t-1}; \boldsymbol{y}, \boldsymbol{u}) = F_{t-1}^{(1)}(\bar{\varphi}_{t-1}(\boldsymbol{a}^{\leq t-1}; \boldsymbol{y}, \boldsymbol{u}); \boldsymbol{y}, \boldsymbol{u}), \qquad g_t(\boldsymbol{b}^{\leq t}; \boldsymbol{v}) = G_t^{(1)}(\varphi_t(\boldsymbol{b}^{\leq t}; \boldsymbol{v}); \boldsymbol{v}).
\end{aligned}$$

Furthermore, $\{\varphi_t\}_{t\geq 1}$ and $\{\bar{\varphi}_t\}_{t\geq 1}$ satisfy the following conditions. For any $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \bar{\boldsymbol{\Sigma}})$ and $t \in \mathbb{N}_{>0}$, there exist uniformly bounded $(b_{ij})_{1\leq j\leq i\leq t}$, $(\bar{b}_{ij})_{1\leq j\leq i\leq t}$, which are sequences with respect to n, such that for $\boldsymbol{y}_{\leq t}$, $\bar{\boldsymbol{y}}_{\leq t}$ as defined in Setting 3, we have $\bar{\boldsymbol{y}}_{\leq t} = \bar{\varphi}_t(\bar{\boldsymbol{g}}_{\leq t}; \boldsymbol{y}_*, \boldsymbol{u})$ and $\boldsymbol{y}_{\leq t} = \varphi_t(\boldsymbol{\mu}_{\leq t}\boldsymbol{\theta} + \boldsymbol{g}_{\leq t}; \boldsymbol{v})$.

Remark C.3.1. For all $t \in \mathbb{N}_{>0}$, since $(b_{ij})_{1\leq j\leq i\leq t}$ and $(\bar{b}_{ij})_{1\leq j\leq i\leq t}$ are uniformly bounded, there exists a subsequence of $\mathbb{N}_{>0}$, which we denote by $\{n_k\}_{k\in\mathbb{N}_{>0}}$, such that for all $s, r \leq t$, $b_{s,t}$ and $\bar{b}_{s,r}$ converge to n-independent limits along $\{n_k\}_{k\in\mathbb{N}_{>0}}$. As a consequence, the following limits exist in probability along the subsequence $\{n_k\}_{k\in\mathbb{N}_{>0}}$ by the third assumption of Setting 3:

$$\begin{aligned}
&\lim_{n,d\to\infty} \frac{1}{n} f_t(\bar{\boldsymbol{g}}_{\leq t}; \boldsymbol{y}_*, \boldsymbol{u})^\mathsf{T} f_s(\bar{\boldsymbol{g}}_{\leq s}; \boldsymbol{y}_*, \boldsymbol{u}), && \lim_{n,d\to\infty} \frac{1}{n} f_t(\bar{\boldsymbol{g}}_{\leq t}; \boldsymbol{y}_*, \boldsymbol{u})^\mathsf{T} \bar{\boldsymbol{g}}_0,\\
&\lim_{n,d\to\infty} \frac{1}{d} g_t(\boldsymbol{\mu}_{\leq t}\boldsymbol{\theta} + \boldsymbol{g}_{\leq t}; \boldsymbol{v})^\mathsf{T} g_s(\boldsymbol{\mu}_{\leq s}\boldsymbol{\theta} + \boldsymbol{g}_{\leq s}; \boldsymbol{v}), && \lim_{n,d\to\infty} \frac{1}{d} g_t(\boldsymbol{\mu}_{\leq t}\boldsymbol{\theta} + \boldsymbol{g}_{\leq t}; \boldsymbol{v})^\mathsf{T} \boldsymbol{\theta},\\
&\lim_{n,d\to\infty} \frac{1}{d} g_t^*(\boldsymbol{\mu}_{\leq t}\boldsymbol{\theta} + \boldsymbol{g}_{\leq t}; \boldsymbol{v})^\mathsf{T} g_s^*(\boldsymbol{\mu}_{\leq s}\boldsymbol{\theta} + \boldsymbol{g}_{\leq s}; \boldsymbol{v}), && \lim_{n,d\to\infty} \frac{1}{d} g_t^*(\boldsymbol{\mu}_{\leq t}\boldsymbol{\theta} + \boldsymbol{g}_{\leq t}; \boldsymbol{v})^\mathsf{T} \boldsymbol{\theta}.
\end{aligned}$$

As a consequence, the new AMP iteration satisfies all assumptions of Theorem C.3.1, thus, its asymptotics can be characterized by the state evolution (C.27) along the subsequence.

Proof. We prove the lemma by induction over t. For the base case $t = 1$, we set $f_0(\boldsymbol{y}, \boldsymbol{u}) := F_0^{(1)}(\boldsymbol{y}, \boldsymbol{u})$, $\varphi_1(\boldsymbol{b}^1; \boldsymbol{v}) := \boldsymbol{b}^1 + F_0^{(2)}(\boldsymbol{v})$, $g_1(\boldsymbol{b}^1; \boldsymbol{v}) := G_1^{(1)}(\varphi_1(\boldsymbol{b}^1; \boldsymbol{v}); \boldsymbol{v})$ and $\bar{\varphi}_1(\boldsymbol{a}^1; \boldsymbol{y}, \boldsymbol{u}) := \boldsymbol{a}^1 + G_1^{(2)}(\boldsymbol{y}, \boldsymbol{u}) + \eta_{1,1} f_0(\boldsymbol{y}, \boldsymbol{u})$, where $\eta_{1,1}$ is defined via state evolution (C.27). Notice that $\eta_{1,1}$ is a function of n. By the uniform Lipschitzness assumption, $\eta_{1,1}$ is uniformly bounded as a sequence in n. Thus, $\varphi_1, \bar{\varphi}_1$ are uniformly Lipschitz. By definition, $\boldsymbol{y}^1 = \varphi_1(\mu_1\boldsymbol{\theta} + \boldsymbol{g}_1; \boldsymbol{v})$ and $\bar{\boldsymbol{y}}^1 = \bar{\varphi}_1(\bar{\boldsymbol{g}}_1; \boldsymbol{y}_*, \boldsymbol{u})$ with $\bar{b}_{11} = \eta_{1,1}$, which completes the proof for the base case.

Next, suppose the lemma holds for the first t iterations, we then prove it holds for the $(t+1)$-th iteration. By induction hypothesis,

$$\begin{aligned}
\boldsymbol{v}^{t+1} &= \boldsymbol{X}^\mathsf{T} F_t^{(1)}(\bar{\varphi}_t(\boldsymbol{a}^{\leq t}; \boldsymbol{y}, \boldsymbol{u}); \boldsymbol{y}, \boldsymbol{u}) + F_t^{(2)}(\varphi_t(\boldsymbol{b}^{\leq t}; \boldsymbol{v}); \boldsymbol{v}),\\
\boldsymbol{u}^{t+1} &= \boldsymbol{X} G_{t+1}^{(1)}(\varphi_t(\boldsymbol{b}^{\leq t+1}; \boldsymbol{v}); \boldsymbol{v}) + G_{t+1}^{(2)}(\bar{\varphi}_t(\boldsymbol{a}^{\leq t}; \boldsymbol{y}, \boldsymbol{u}); \boldsymbol{y}, \boldsymbol{u}).
\end{aligned}$$

We let $f_t(\boldsymbol{x}^{\leq t}; \boldsymbol{y}, \boldsymbol{u}) := F_t^{(1)}(\bar{\varphi}_t(\boldsymbol{x}^{\leq t}; \boldsymbol{y}, \boldsymbol{u}); \boldsymbol{y}, \boldsymbol{u})$ and $g_{t+1}(\boldsymbol{x}^{\leq t+1}; \boldsymbol{v}) := G_{t+1}^{(1)}(\varphi_{t+1}(\boldsymbol{x}^{\leq t+1}; \boldsymbol{v}); \boldsymbol{v})$. The composition of uniformly Lipschitz functions is still uniformly Lipschitz. As a consequence, we can conclude that

f_t, g_{t+1} are uniformly Lipschitz functions. Based on the choice of $\{f_s\}_{0\le s\le t}$ and $\{g_s\}_{1\le s\le t+1}$, we can compute the coefficients for the Onsager correction terms $\{\xi_{t,s}\}_{1\le s\le t}$ and $\{\eta_{t+1,s}\}_{1\le s\le t+1}$, which are uniformly bounded as sequences in n.

Then we define $\boldsymbol{a}^{t+1}$, $\boldsymbol{b}^{t+1}$ via the AMP iteration (C.25), which gives

$$\begin{aligned}
\boldsymbol{b}^{t+1} &= \boldsymbol{v}^{t+1} - F_t^{(2)}(\varphi_t(\boldsymbol{b}^{\le t};\boldsymbol{v});\boldsymbol{v}) - \sum_{s=1}^{t}\xi_{t,s}G_s^{(1)}(\varphi_t(\boldsymbol{b}^{\le s};\boldsymbol{v});\boldsymbol{v}),\\
\boldsymbol{a}^{t+1} &= \boldsymbol{u}^{t+1} - G_{t+1}^{(2)}(\bar{\varphi}_t(\boldsymbol{a}^{\le t};\boldsymbol{y},\boldsymbol{u});\boldsymbol{y},\boldsymbol{u}) - \sum_{s=1}^{t+1}\eta_{t+1,s}F_{s-1}^{(1)}(\bar{\varphi}_{s-1}(\boldsymbol{a}^{\le s-1};\boldsymbol{y},\boldsymbol{u});\boldsymbol{y},\boldsymbol{u}).
\end{aligned}$$

Solving for $\boldsymbol{u}^{t+1}$ and $\boldsymbol{v}^{t+1}$ leads to the definition of φ_{t+1} and $\bar{\varphi}_{t+1}$. Furthermore, by setting $b_{ts} = \xi_{t,s}$ and $\bar{b}_{t+1,s} = \eta_{t+1,s}$, we have

$$\begin{aligned}
&\varphi_{t+1}(\boldsymbol{\mu}_{\le t+1}\boldsymbol{\theta} + \boldsymbol{g}_{\le t+1};\boldsymbol{v})\\
=&(\varphi_t(\boldsymbol{\mu}_t\boldsymbol{\theta} + \boldsymbol{g}_{\le t};\boldsymbol{v}), \mu_{t+1}\boldsymbol{\theta} + \boldsymbol{g}_{t+1} + F_t^{(2)}(\varphi_t(\boldsymbol{\mu}_{\le t}\boldsymbol{\theta} + \boldsymbol{g}_{\le t};\boldsymbol{v});\boldsymbol{v}) + \sum_{s=1}^{t}\xi_{t,s}G_s^{(1)}(\varphi_t(\boldsymbol{\mu}_{\le s}\boldsymbol{\theta} + \boldsymbol{g}_{\le s};\boldsymbol{v});\boldsymbol{v}))\\
=&(\boldsymbol{y}^{\le t}, \boldsymbol{y}^{t+1}),\\
&\bar{\varphi}_{t+1}(\bar{\boldsymbol{g}}_{\le t+1};\boldsymbol{y}_*,\boldsymbol{u})\\
=&(\bar{\varphi}_t(\bar{\boldsymbol{g}}_{\le t};\boldsymbol{y}_*,\boldsymbol{u}), \bar{\boldsymbol{g}}_{t+1} + G_{t+1}^{(2)}(\bar{\varphi}_t(\bar{\boldsymbol{g}}_{\le t};\boldsymbol{y}_*,\boldsymbol{u});\boldsymbol{y}_*,\boldsymbol{u}) + \sum_{s=1}^{t+1}\eta_{t+1,s}F_{s-1}^{(1)}(\bar{\varphi}_{s-1}(\bar{\boldsymbol{g}}_w\le s-1;\boldsymbol{y}_*,\boldsymbol{u});\boldsymbol{y}_*,\boldsymbol{u}))\\
=&(\bar{\boldsymbol{y}}^{\le t}, \bar{\boldsymbol{y}}^{t+1}),
\end{aligned}$$

thus completes the proof of the lemma by induction.

□

As an immediate consequence of Lemma C.3.1, Corollary C.2.1 holds true under Setting 3 as well.

C.3.3 Orthogonalization

By linear algebra, $\{\boldsymbol{b}^t\}_{t\ge1}$ derived via AMP iteration (C.25) can be further reduced to a set of vectors that are approximately orthogonal after subtracting the component along $\boldsymbol{\theta}$, which leads to the following lemma:

Lemma C.3.2. *Let $\{\boldsymbol{a}^t\}_{t\ge1}$, $\{\boldsymbol{b}^t\}_{t\ge1}$ be sequences produced by the AMP iteration (C.25) under Setting 3. Then there exist functions $\{\phi_t : \mathbb{R}^{d(t+1)} \to \mathbb{R}^{dt}\}_{t\ge1}$ which are uniformly Lipschitz, such that the following holds:*

(i) For all $t \in \mathbb{N}_{>0}$, there exist n-independent constants $\{c_{ts}\}_{0\le s\le t}$ such that $c_{tt} \neq 0$ and $\boldsymbol{q}^{t+1} = \sum_{s=0}^{t} c_{ts}\boldsymbol{b}^{s+1}$. We write $\boldsymbol{q}^{\le t} = \phi_t(\boldsymbol{b}^{\le t})$, and ϕ_t as a sequence in n is uniformly Lipschitz.

(ii) For all $t \in \mathbb{N}_{>0}$, there exist $(x_0, \cdots, x_{t-1}) \in \{0,1\}^t$ and $(\alpha_1, \cdots, \alpha_t) \in \mathbb{R}^t$, such that for any $\{\psi_n : \mathbb{R}^{n(t+2)} \to \mathbb{R}^n\}$ uniformly pseudo-Lipschitz of order 2,

$$\psi_n(\boldsymbol{q}^{\le t};\boldsymbol{\theta},\boldsymbol{v}) = \mathbb{E}[\psi_n(\boldsymbol{q}^{\le t};\boldsymbol{\theta},\boldsymbol{v})] + o_P(1),$$

where $\boldsymbol{q}^i = x_{i-1}(\alpha_i \boldsymbol{\theta} + \boldsymbol{z}_i)$, *with* $\{\boldsymbol{z}_i\}_{i\ge 1} \overset{iid}{\sim} \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ *independent of* $(\boldsymbol{\theta}, \boldsymbol{v})$.

Proof. Recall that $\boldsymbol{y}_* = h(\bar{\boldsymbol{g}}_0, \boldsymbol{w})$. Given the state evolution (C.27) of the AMP iteration, we define

$$\boldsymbol{h}_t := f_t(\bar{\boldsymbol{g}}_{\le t}; \boldsymbol{y}_*, \boldsymbol{u}), \qquad \mathcal{S}_t := \operatorname{span}(\boldsymbol{h}_k : 0 \le k \le t).$$

Note that by state evolution, $\lim_{n,d\to\infty} \mathbb{E}\langle \boldsymbol{h}_t, \boldsymbol{h}_s \rangle / n = \Sigma_{s+1,t+1}$. By linear algebra, for all $t \in \mathbb{N}$, there exist deterministic constants $\{c_{ts}\}_{0\le s\le t}$ and $x_t \in \{0,1\}$, such that $c_{tt} \neq 0$ and

$$\sum_{i=0}^{t} \sum_{j=0}^{s} c_{ti} c_{sj} \Sigma_{i+1,j+1} = \mathbb{1}_{s=t} x_t.$$

We define $\boldsymbol{r}_t := \sum_{s=0}^{t} c_{ts} \boldsymbol{h}_s$, then $\lim_{n\to\infty} \mathbb{E}\langle \boldsymbol{r}_t, \boldsymbol{r}_s \rangle / n = \mathbb{1}_{s=t} x_t$ for all $s, t \in \mathbb{N}$. Next, we prove the lemma by induction. For the base case $t = 1$, we let $\boldsymbol{q}^1 = c_{00} \boldsymbol{b}^1$, thus, claim (i) follows. As for claim (ii), we consider two cases. In the first case, $x_0 = 0$, then $\mathbb{E}\langle \boldsymbol{h}_0, \boldsymbol{h}_0 \rangle / n \to 0$. By state evolution (C.27),

$$\begin{aligned}
\mu_1 &\overset{(a)}{=} \lim_{n,d\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{\delta \mathbb{E}[\bar{g}_{0,i} f_{0,i}(h(\bar{\boldsymbol{g}}_0, \boldsymbol{w}), \boldsymbol{u})]}{\mathbb{E}[\Theta^2]}, \\
&\overset{(b)}{\le} \limsup_{n,d\to\infty} \frac{1}{\sqrt{n}} \frac{\delta^{1/2}}{\mathbb{E}[\Theta^2]^{1/2}} \mathbb{E}[\|f_0(\boldsymbol{y}_*, \boldsymbol{u})\|_2^2]^{1/2} \to 0,
\end{aligned}$$

where (a) holds by Stein's lemma, and (b) holds by Cauchy-Schwartz inequality. Thus, claim (ii) holds with $\boldsymbol{q}^1 \equiv \boldsymbol{0}$. In the second case, $x_0 = 1$, whence $c_{00} = \Sigma_{11}^{-1/2}$, and claim (ii) holds by the state evolution (C.27). Moreover,

$$\alpha_1 = \lim_{n,d\to\infty} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\mathbb{E}[\partial_{\bar{g}_{0,i}} f_{0,i}(h(\bar{\boldsymbol{g}}_0, \boldsymbol{w}), \boldsymbol{u})]}{\mathbb{E}[\|f_0(h(\bar{\boldsymbol{g}}_0, \boldsymbol{w}), \boldsymbol{u})\|_2^2]^{1/2}}. \tag{C.28}$$

Suppose the lemma holds for the first t iterations, then we prove it holds for the $(t+1)$-th iteration as well. We let $\boldsymbol{q}^{t+1} = \sum_{s=0}^{t} c_{ts} \boldsymbol{b}^{s+1}$, and the definition of ϕ_{t+1} together with claim (i) follows immediately. As for claim (ii), first notice that the following mapping is uniformly Lipschitz of order 2:

$$(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{t+1}, \boldsymbol{\theta}, \boldsymbol{v}) \mapsto \psi_n(\phi_{t+1}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{t+1}); \boldsymbol{\theta}, \boldsymbol{v}).$$

Again we consider two cases. In the first case, $x_t = 0$, thus by state evolution (C.27), (ii) holds with $\boldsymbol{q}^{t+1} = \boldsymbol{0}$. In the second case, $x_t = 1$, then again by state evolution recursion, we can set $\boldsymbol{q}^{t+1} = \alpha_{t+1} \boldsymbol{\theta} + \boldsymbol{z}_{t+1}$, with

$$\alpha_{t+1} = \lim_{n,d\to\infty} \frac{\sqrt{n} \mathbb{E}[\langle \bar{\boldsymbol{g}}_0^{\perp,t}, \Pi_{\mathcal{S}_{t-1}}^{\perp}(\boldsymbol{h}_t) \rangle]}{\mathbb{E}[\|\Pi_{\mathcal{S}_{t-1}}^{\perp}(\boldsymbol{h}_t)\|_2^2]^{1/2} \mathbb{E}[\|\bar{\boldsymbol{g}}_0^{\perp,t}\|_2^2]}, \tag{C.29}$$

where $\bar{\boldsymbol{g}}_0^{\perp,t} := \Pi_{\bar{\mathcal{G}}_t}^{\perp}(\bar{\boldsymbol{g}}_0)$ with $\bar{\mathcal{G}}_t := \operatorname{span}(\bar{\boldsymbol{g}}_i : 1 \le i \le t)$. Therefore, we complete the proof of the lemma by induction.

□

C.3.4 Optimality analysis

As before, we restrict to the case with $x_t = 1$ for all $t \in \mathbb{N}$. Given $(\boldsymbol{v}, \boldsymbol{\alpha}_{\le t}\boldsymbol{\theta} + \boldsymbol{g}_{\le t})$, a sufficient statistics of $\boldsymbol{\theta}$ is $(\boldsymbol{v}, \|\boldsymbol{\alpha}_{\le t}\|_2 \boldsymbol{\theta} + \boldsymbol{g})$ with $\boldsymbol{g} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ independent of $\boldsymbol{\theta}$. Therefore, by Lemma C.3.1 and C.3.2, in order to derive the minimum estimation error achieved by any GFOM with t iterations, it suffices to study the maximum value of $\|\boldsymbol{\alpha}_{\le t}\|_2$, which leads to the following lemma:

Lemma C.3.3. *For all $t \in \mathbb{N}_{>0}$ and all AMP iterations (C.25), we have $\|\boldsymbol{\alpha}_{\le t}\|_2^2 \le \beta_t^2$.*

Proof. Recall that $\bar{\boldsymbol{g}}_0^{\perp,t} := \Pi_{\bar{\mathcal{G}}_t}^{\perp}(\bar{\boldsymbol{g}}_0)$ with $\bar{\mathcal{G}}_t := \mathrm{span}(\bar{\boldsymbol{g}}_i : 1 \le i \le t)$. We define:

$$\omega_t^2 := \lim_{n,d\to\infty} \frac{1}{n}\mathbb{E}[\|\bar{\boldsymbol{g}}_0^{\perp,t}\|_2^2], \qquad \zeta_t^2 := \frac{1}{\delta}\mathbb{E}[\Theta^2] - \omega_t^2.$$

The above limit exists by the assumption of the AMP algorithm. Here, we will prove a stronger result. To be precise, we will establish that the following two claims hold for all $t \in \mathbb{N}^+$: (1) $\omega_{t-1} \ge \sigma_t$; (2) $\|\boldsymbol{\alpha}_{\le t}\|_2^2 \le \beta_t^2$. We prove the claims via induction. By definition, $\omega_0 = \sigma_1$. Furthermore, by Eq. (C.28),

$$\begin{aligned}
\alpha_1^2 &= \lim_{n,d\to\infty} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\mathbb{E}[\partial_{\bar{g}_{0,i}} f_{0,i}(h(\bar{\boldsymbol{g}}_0, \boldsymbol{w}), \boldsymbol{u})]}{\mathbb{E}[\|f_0(h(\bar{\boldsymbol{g}}_0, \boldsymbol{w}), \boldsymbol{u})\|_2^2]^{1/2}} \right\}^2 \\
&\overset{(a)}{=} \lim_{n,d\to\infty} \frac{\delta^2 \mathbb{E}[\langle f_0(h(\bar{\boldsymbol{g}}_0, \boldsymbol{w}), \boldsymbol{u}), \bar{\boldsymbol{g}}_0 \rangle]^2}{n\mathbb{E}[\|f_0(h(\bar{\boldsymbol{g}}_0, \boldsymbol{w}), \boldsymbol{u})\|_2^2]\mathbb{E}[\Theta^2]^2} \\
&= \lim_{n,d\to\infty} \frac{\delta^2 \mathbb{E}[\langle f_0(h(\bar{\boldsymbol{g}}_0, \boldsymbol{w}), \boldsymbol{u}), \mathbb{E}[\bar{\boldsymbol{g}}_0 \mid h(\bar{\boldsymbol{g}}_0, \boldsymbol{w}), \boldsymbol{u}] \rangle]^2}{n\mathbb{E}[\|f_0(h(\bar{\boldsymbol{g}}_0, \boldsymbol{w}), \boldsymbol{u})\|_2^2]\mathbb{E}[\Theta^2]^2} \\
&\overset{(b)}{\le} \lim_{n,d\to\infty} \frac{\delta^2 \mathbb{E}[\|\mathbb{E}[\bar{\boldsymbol{g}}_0 \mid h(\bar{\boldsymbol{g}}_0, \boldsymbol{w}), \boldsymbol{u}]\|_2^2]}{n\mathbb{E}[\Theta^2]^2} = \beta_1^2,
\end{aligned}$$

where (a) is by Stein's lemma, and (b) is by Cauchy-Schwartz inequality. Then we assume the lemma holds for the first t iterations, and we prove by induction that it also holds for iteration $(t+1)$. For $t \in \mathbb{N}_{>0}$, we let

$$\boldsymbol{k}_t := g_t(\boldsymbol{\mu}_{\le t}\boldsymbol{\theta} + \boldsymbol{g}_{\le t}; \boldsymbol{v}), \qquad \mathcal{S}_t' := \mathrm{span}(\boldsymbol{k}_i : 1 \le i \le t).$$

By the state evolution of the AMP algorithm, $\omega_t^2 = \lim_{n,d\to\infty} \mathbb{E}[\|\Pi_{\mathcal{S}_t'}^{\perp}(\boldsymbol{\theta})\|_2^2]/n$. Thus, we have

$$\begin{aligned}
\omega_t^2 &\overset{(d)}{=} \frac{1}{\delta}\mathbb{E}[\Theta^2] - \lim_{n,d\to\infty} \frac{1}{n}\mathbb{E}[\|\Pi_{\mathcal{S}_t'}(\boldsymbol{\theta})\|_2^2] \\
&\overset{(e)}{\ge} \frac{1}{\delta}\mathbb{E}[\Theta^2] - \lim_{n,d\to\infty} \frac{1}{n}\mathbb{E}[\|\mathbb{E}[\boldsymbol{\theta} \mid \boldsymbol{\alpha}_{\le t}\boldsymbol{\theta} + \boldsymbol{z}_{\le t}, \boldsymbol{v}]\|_2^2] \\
&\overset{(f)}{=} \frac{1}{\delta}\mathbb{E}[\Theta^2] - \lim_{n,d\to\infty} \frac{1}{n}\mathbb{E}[\|\mathbb{E}[\boldsymbol{\theta} \mid \|\boldsymbol{\alpha}_{\le t}\|_2\boldsymbol{\theta} + \boldsymbol{z}, \boldsymbol{v}]\|_2^2] \\
&\overset{(g)}{\ge} \frac{1}{\delta}\mathbb{E}[\Theta^2] - \frac{1}{\delta}\mathbb{E}[\mathbb{E}[\Theta \mid \beta_t\Theta + G, V]^2] = \sigma_{t+1}^2,
\end{aligned}$$

where (d) is by Pythagora's theorem, (e) is by Jensen's inequality, (f) is by property of sufficient statistics, and (g) is by induction hypothesis. Thus, we have completed the proof of claim (1).

Then we prove claim (2). By Eq. (C.29),

$$
\begin{aligned}
\alpha_{t+1}^2 &= \lim_{n,d\to\infty} \frac{n\mathbb{E}[\langle \mathbb{E}[\bar{\boldsymbol{g}}_0^{\perp,t} \mid \bar{\boldsymbol{g}}_{\le t}, \boldsymbol{u}, h(\bar{\boldsymbol{g}}_0, \boldsymbol{w})], \Pi_{\mathcal{S}_{t-1}}^{\perp}(\boldsymbol{h}_t)\rangle]^2}{\mathbb{E}[\|\Pi_{\mathcal{S}_{t-1}}^{\perp}(\boldsymbol{h}_t)\|_2^2]\mathbb{E}[\|\bar{\boldsymbol{g}}_0^{\perp,t}\|_2^2]^2} \\
&\overset{(a)}{\le} \lim_{n,d\to\infty} \frac{\mathbb{E}[\|\mathbb{E}[\bar{\boldsymbol{g}}_0^{\perp,t} \mid \bar{\boldsymbol{g}}_{\le t}, \boldsymbol{u}, h(\bar{\boldsymbol{g}}_0, \boldsymbol{w})]\|_2^2]}{n\omega_t^4} - \lim_{n,d\to\infty} \sum_{s=0}^{t-1} \frac{\mathbb{E}[\langle \Pi_{\mathcal{S}_{s-1}}^{\perp}(\boldsymbol{h}_s), \mathbb{E}[\bar{\boldsymbol{g}}_0^{\perp,t} \mid \bar{\boldsymbol{g}}_{\le s}, \boldsymbol{u}, h(\bar{\boldsymbol{g}}_0, \boldsymbol{w})]\rangle]^2}{n\omega_t^4\mathbb{E}[\|\Pi_{\mathcal{S}_{s-1}}^{\perp}(\boldsymbol{h}_s)\|_2^2]} \\
&\overset{(b)}{=} \lim_{n,d\to\infty} \frac{1}{\omega_t^2}\mathbb{E}[\mathbb{E}[Z_0 \mid h(\omega_t Z_0 + \zeta_t Z_1, W), U, Z_1]^2] - \lim_{n,d\to\infty} \sum_{s=0}^{t-1} \frac{\mathbb{E}[\langle \Pi_{\mathcal{S}_{s-1}}^{\perp}(\boldsymbol{h}_s), \mathbb{E}[\bar{\boldsymbol{g}}_0^{\perp,s} \mid \bar{\boldsymbol{g}}_{\le s}, \boldsymbol{u}, h(\bar{\boldsymbol{g}}_0, \boldsymbol{w})]\rangle]^2}{n\omega_s^4\mathbb{E}[\|\Pi_{\mathcal{S}_{s-1}}^{\perp}(\boldsymbol{h}_s)\|_2^2]} \\
&= \lim_{n,d\to\infty} \frac{1}{\omega_t^2}\mathbb{E}[\mathbb{E}[Z_0 \mid h(\omega_t Z_0 + \zeta_t Z_1, W), U, Z_1]^2] - \lim_{n,d\to\infty} \sum_{s=0}^{t-1} \frac{n\mathbb{E}[\langle \bar{\boldsymbol{g}}_0^{\perp,s}, \Pi_{\mathcal{S}_{s-1}}^{\perp}(\boldsymbol{h}_s)\rangle]^2}{\mathbb{E}[\|\Pi_{\mathcal{S}_{s-1}}^{\perp}(\boldsymbol{h}_s)\|_2^2]\mathbb{E}[\|\bar{\boldsymbol{g}}_0^{\perp,s}\|_2^2]^2} \\
&\overset{(c)}{\le} \frac{1}{\sigma_{t+1}^2}\mathbb{E}[\mathbb{E}[Z_0 \mid h(\sigma_{t+1} Z_0 + \tilde{\sigma}_{t+1} Z_1, W), U, Z_1]^2] - \sum_{s=1}^{t} \alpha_s^2,
\end{aligned}
$$

where (a) is by Pythagora's theorem, (b) is by Lemma C.2.3, and (c) is by induction hypothesis and Lemma C.2.5. The last inequality above gives $\sum_{s=1}^{t+1} \alpha_s^2 \le \beta_{s+1}^2$. Thus, we have completed the proof of the lemma by induction.

□

C.4 Reduction to matrices with sub-Gaussian entries

In this section, we show that in order to prove Theorem 4.3.1 under Setting 2.(a) (or to prove Theorem 4.5.1 under Setting 4.(a)), it suffices to consider cases where the matrix $\boldsymbol{W}$(or $\boldsymbol{X}$) has sub-Gaussian entries. Here, we prove this claim for Theorem 4.3.1 under Setting 2.(a). Proof of the claim for Theorem 4.5.1 under Setting 4.(a) follows by the same argument, with notational adaptations.

By assumption, $\mathbb{E}[W_{ij}^4] \le C/n^2$ and $\mathbb{E}[W_{ij}] = 0$. Thus, we claim that for all $\epsilon > 0$ and $i, j \in [n]$, there exists decomposition $W_{ij} = W_{ij}^{(1)} + W_{ij}^{(2)}$, such that $\mathbb{E}[W_{ij}^{(1)}] = \mathbb{E}[W_{ij}^{(2)}] = 0$, $\operatorname{ess\,sup}_n \sqrt{n}|W_{ij}^{(1)}| < \infty$, $\sup_n n^2\mathbb{E}[(W_{ij}^{(2)})^4] < \infty$ and $n\operatorname{Var}[W_{ij}^{(2)}] \le \epsilon$. Furthermore, $(W_{ij}^{(1)})_{i<j\le n}$ are independent and identically distributed random variables, and the same property holds for $(W_{ij}^{(2)})_{i<j\le n}$. To prove this claim, we let $\xi_\epsilon > 0$ such that $C/\xi_\epsilon^2 < \epsilon$. We define

$$
\begin{aligned}
W_{ij}^{(1)} &:= W_{ij}\mathbb{1}_{\sqrt{n}|W_{ij}|\le\xi_\epsilon} - \mathbb{E}[W_{ij}\mathbb{1}_{\sqrt{n}|W_{ij}|\le\xi_\epsilon}], \\
W_{ij}^{(2)} &:= W_{ij}\mathbb{1}_{\sqrt{n}|W_{ij}|>\xi_\epsilon} - \mathbb{E}[W_{ij}\mathbb{1}_{\sqrt{n}|W_{ij}|>\xi_\epsilon}].
\end{aligned}
$$

Then $\sqrt{n}|W_{ij}^{(1)}| \le 2\xi_\epsilon$, $\mathbb{E}[W_{ij}^{(1)}] = \mathbb{E}[W_{ij}^{(2)}] = 0$, $\sup_n n^2\mathbb{E}[(W_{ij}^{(1)})^4] < \infty$ and $\sup_n n^2\mathbb{E}[(W_{ij}^{(2)})^4] < \infty$. Furthermore, $n\operatorname{Var}[W_{ij}^{(2)}] \le n\mathbb{E}[W_{ij}^2\mathbb{1}_{\sqrt{n}|W_{ij}|>\xi_\epsilon}] \le C/\xi_\epsilon^2 < \epsilon$, thus completes the proof of the claim.

With the above decomposition, we let $\boldsymbol{W}^{(1)} = (W_{ij}^{(1)})_{i,j\le n}$ and $\boldsymbol{W}^{(2)} = (W_{ij}^{(2)})_{i,j\le n}$ be $n \times n$ matrices. By the Bai-Yin law [192], we have $\|\boldsymbol{W}^{(2)}\|_{\mathrm{op}} \le 2\sqrt{\epsilon} + o_P(1)$. If we replace $\boldsymbol{W}$ with $\boldsymbol{W}^{(1)}$ in model definition (4.5), and denote the iterates obtained by GFOM (4.6) by $\{\tilde{\boldsymbol{u}}^t\}_{t\ge1}$, then we can prove by induction that for all $t \in \mathbb{N}_{>0}$, with probability $1 - o_n(1)$,

$$
\frac{1}{\sqrt{n}}\|\boldsymbol{u}^t - \tilde{\boldsymbol{u}}^t\|_2 \le F(\epsilon, t).
$$

Here, $F(\epsilon, t) \to 0$ as $\epsilon \to 0^+$. The proof is via simple application of the Lipschitz assumption and the upper bound of the spectral norm of $\boldsymbol{W}^{(2)}$ we have just derived. Since ϵ is arbitrary, we conclude that if Theorem 4.3.1 holds for sub-Gaussian distributions, then it also holds for distributions with bounded fourth moments.

Appendix D

Sampling from the posterior via diffusion processes

D.1 Technical preliminaries

This section summarize some technical facts that will be useful in the proof.

Remark D.1.1. Let $\boldsymbol{\theta}, \boldsymbol{X}$ be a couple of random variables (vectors) whose joint distribution is given by the general Bayesian model (5.1). Let $\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(k)}$ be i.i.d. samples from the posterior $\mu_{\boldsymbol{X}}(\,\cdot\,)x := \mathbb{P}(\,\cdot\,|\boldsymbol{X})$, independent of $\boldsymbol{\theta}$. Then

$$\boldsymbol{X}, \boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(k)} \overset{\mathrm{d}}{=} \boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(k-1)}\,. \tag{D.1}$$

(Here $\overset{\mathrm{d}}{=}$ denotes equality in distribution.)

This fact is immediate (just write the joint distribution) and is known in physics as the "Nishimori identity."

Lemma D.1.1 (Lemma 3.2 in [165])**.** *If f and g are two differentiable convex functions, then for any $b > 0$,*

$$|f'(a) - g'(a)| \le g'(a+b) - g'(a-b) + \frac{d}{b},$$

where $d = |f(a+b) - g(a+b)| + |f(a-b) - g(a-b)| + |f(a) - g(a)|$.

Lemma D.1.2 (Lemma 4.15 in [4])**.** *Suppose probability distributions μ_1, μ_2 on $[-1,1]^n$ are given. Sample $\boldsymbol{m}_1 \sim \mu_1$ and $\boldsymbol{m}_2 \sim \mu_2$ and let $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \{-1,+1\}^n$ be standard randomized roundings, respectively of $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$. (Namely, the coordinates of $\boldsymbol{\theta}_i$ are conditionally independent given $\boldsymbol{m}_i$, with $\mathbb{E}[\boldsymbol{\theta}_i \mid \boldsymbol{m}_i] = \boldsymbol{m}_i$.) Then*

$$W_{2,n}(\mathcal{L}(\boldsymbol{\theta}_1), \mathcal{L}(\boldsymbol{\theta}_2)) \le 2\sqrt{W_{2,n}(\mu_1, \mu_2)}.$$

Lemma D.1.3 (Proposition 2.1.2 in [192])**.** *Let $g \sim \mathsf{N}(0,1)$. Then for all $t > 0$, it holds that*

$$\left(\frac{1}{t} - \frac{1}{t^3}\right) \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \le \mathbb{P}(g \ge t) \le \frac{1}{t} \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2}.$$

D.2 Concentration of the stochastic localization process

The next lemma is a slight generalization of analogous results in [80, 4].

Lemma D.2.1. *Let $\mu \in \mathscr{P}_2(\mathbb{R}^n)$ be a probability measure with finite second moment, and $\boldsymbol{y}(t) = t\boldsymbol{H}\boldsymbol{\theta} + \sqrt{t}\boldsymbol{g}$ for $(\boldsymbol{\theta}, \boldsymbol{g}) \sim \mu \otimes \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. Further, let $\mu_t(\,\cdot\,) := \mathbb{P}(\boldsymbol{\theta} \in \,\cdot\, | \boldsymbol{y}(t))$, and $\boldsymbol{m}(\boldsymbol{y}(t); t) := \mathbb{E}[\boldsymbol{\theta} | \boldsymbol{y}(t)]$. Finally, denote by $\boldsymbol{P}_{\ker(\boldsymbol{H})}$ the projector onto the null space of $\boldsymbol{H}$.*

Then the following inequalities hold for all $t > 0$:

$$\mathbb{E}\,\mathrm{Cov}(\mu_t) \preceq \boldsymbol{P}_{\ker(\boldsymbol{H})}\,\mathrm{Cov}(\mu)\boldsymbol{P}_{\ker(\boldsymbol{H})} + \frac{1}{t}\boldsymbol{H}^+(\boldsymbol{H}^+)^\mathsf{T}\,, \tag{D.2}$$

$$W_{2,n}(\mu, \mathrm{Law}(\boldsymbol{m}(\boldsymbol{y}(t); t)))^2 \le \frac{1}{n}\,\mathrm{Tr}(\boldsymbol{P}_{\ker(\boldsymbol{H})}\,\mathrm{Cov}(\mu)) + \frac{1}{nt}\,\mathrm{Tr}\big(\boldsymbol{H}^+(\boldsymbol{H}^+)^\mathsf{T}\big)\,. \tag{D.3}$$

Proof. By rescaling $\boldsymbol{H}$, we can assume without loss of generality $t = 1$. We can also center μ so that $\mathbb{E}(\boldsymbol{\theta}) = \boldsymbol{0}$. We will write, for simplicity, $\boldsymbol{y} = \boldsymbol{y}(1)$. Note that

$$\begin{aligned} \mathbb{E}\,\mathrm{Cov}(\mu_t) &= \mathbb{E}\Big\{\big(\boldsymbol{\theta} - \mathbb{E}(\boldsymbol{\theta}|\boldsymbol{y})\big)\big(\boldsymbol{\theta} - \mathbb{E}(\boldsymbol{\theta}|\boldsymbol{y})\big)^\mathsf{T}\Big\} && \text{(D.4)}\\ &\le \mathbb{E}\Big\{\big(\boldsymbol{\theta} - \boldsymbol{H}^+\boldsymbol{y}\big)\big(\boldsymbol{\theta} - \boldsymbol{H}^+\boldsymbol{y}\big)^\mathsf{T}\Big\} && \text{(D.5)}\\ &= \boldsymbol{P}_{\ker(\boldsymbol{H})}\,\mathrm{Cov}(\mu)\boldsymbol{P}_{\ker(\boldsymbol{H})} + \boldsymbol{H}^+(\boldsymbol{H}^+)^\mathsf{T}\,, && \text{(D.6)} \end{aligned}$$

where the inequality follows by the optimality of posterior expectation under quadratic losses. This proves the first claim (D.2).

In order to prove the second one, denote by $B(n,t)$ the right-hand side of Eq. (D.3). By taking the trace of the former inequality, we obtain

$$\mathbb{E}\big\{W_{2,n}(\mu_t, \delta_{\boldsymbol{m}(\boldsymbol{y}(t),t)})^2\big\} = \frac{1}{n}\mathbb{E}\,\mathrm{Tr}\,\mathrm{Cov}(\mu_t) \le B(n,t)\,. \tag{D.7}$$

Since $(\mu, \nu) \mapsto W_2(\mu, \nu)^2$ is jointly convex in (μ, ν), Jensen's inequality implies

$$\mathbb{E}\big\{W_{2,n}(\mu_t, \delta_{\boldsymbol{m}(\boldsymbol{y}(t),t)})^2\big\} \ge W_{2,n}(\mathbb{E}\mu_t, \mathbb{E}\delta_{\boldsymbol{m}(\boldsymbol{y}(t),t)})^2 = W_{2,n}(\mu, \mathrm{Law}(\boldsymbol{m}(\boldsymbol{y}(t); t)))^2\,, \tag{D.8}$$

which completes our proof.

□

D.3 Proof of Lemma 5.3.1

We will use the following lemma, which is a straightforward consequence of the fact that the characteristic function uniquely identifies the corresponding probability measure.

Lemma D.3.1. *Let* P *be a probability measure on* $\mathbb{R}$*. Then* P *is symmetric (i.e.* $\mathrm{P}(A) = \mathrm{P}(-A)$ *for every Borel set* A*) if and only if its characteristic function* φ_{P} *is real values or, equivalently, if and only if* $\varphi_{\mathrm{P}}(t) = \varphi_{\mathrm{P}}(-t)$ *for every* $t \in \mathbb{R}$*.*

Recall that $\boldsymbol{\nu}$ is a top eigenvector of $\boldsymbol{X}$ with norm $\|\boldsymbol{\nu}\|_2^2 = n\beta^2(\beta^2-1)$. We denote by λ_1 the corresponding eigenvalue, and note that this is almost surely non-degenerate (because the law of $\boldsymbol{X}$ is absolutely continuous with respect to Lebesgue). Let

$$\boldsymbol{\nu}_+ = s\boldsymbol{\nu}\,, \qquad s := \mathrm{sign}\langle \boldsymbol{\nu}, \boldsymbol{\theta}\rangle\,. \tag{D.9}$$

Note that we can assume s independent of $\boldsymbol{\theta}, \boldsymbol{W}$ (because we can define $\boldsymbol{\nu}$ to be taken uniformly at random among the two eigenvectors with given norm.)

For any $\boldsymbol{\Omega} \in \mathcal{O}(n)$ satisfying $\boldsymbol{\Omega\theta} = \boldsymbol{\theta}$ and is independent of $\boldsymbol{W}$, we have $\boldsymbol{\Omega W \Omega}^{\mathsf{T}} \overset{d}{=} \boldsymbol{W}$. Moreover, if we replace $\boldsymbol{W}$ by $\boldsymbol{\Omega W \Omega}^{\mathsf{T}}$, then λ_X is the top eigenvalue of $\boldsymbol{\Omega X \Omega}^{\mathsf{T}} = \beta\boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}}/n + \boldsymbol{\Omega W \Omega}^{\mathsf{T}}$ and $\boldsymbol{\Omega\nu}$ is the corresponding eigenvector. As a result, we can conclude that the following two conditional distributions are equal:

$$\boldsymbol{\Omega W \Omega}^{\mathsf{T}}, \boldsymbol{\Omega\nu}_+ \mid \boldsymbol{\theta}, \boldsymbol{\Omega} \overset{\mathrm{d}}{=} \boldsymbol{W}, \boldsymbol{\nu}_+ \mid \boldsymbol{\theta}, \boldsymbol{\Omega}\,.$$

Let $\boldsymbol{P}_{\boldsymbol{\theta}}^{\perp}$ be the projector orthogonal to $\boldsymbol{\theta}$. The above invariance implies that, conditioning on $\boldsymbol{\theta}$, $\langle \boldsymbol{\theta}, \boldsymbol{\nu}_+\rangle_+/\|\boldsymbol{\theta}\|_2 = \rho_{\|}$ and $\|\boldsymbol{P}_{\boldsymbol{\theta}}^{\perp}\boldsymbol{\nu}_+\|_2 = \rho_{\perp}$, we have (on $\boldsymbol{\theta} \neq \boldsymbol{0}$):

$$\boldsymbol{\nu}_+ = \rho_{\|}\frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|_2} + \rho_{\perp}\boldsymbol{u}\,,$$

where $\boldsymbol{u}$ is a uniformly random unit vector orthogonal to $\boldsymbol{\theta}$. Equivalently,

$$\boldsymbol{\nu}_+ = \rho_{\|}\frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|_2} + \rho_{\perp}\frac{\boldsymbol{P}_{\boldsymbol{\theta}}^{\perp}\boldsymbol{g}}{\|\boldsymbol{P}_{\boldsymbol{\theta}}^{\perp}\boldsymbol{g}\|_2}\,,$$

where $\boldsymbol{g} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ independent of $\boldsymbol{\theta}$.

By [28] $\rho_{\|}/\sqrt{n} = (\beta^2-1) + o_{n,P}(1)$ and therefore, using the normalization of $\boldsymbol{\nu}_+$, $\rho_{\perp}/\sqrt{n} = (\beta^2-1)^{1/2} + o_{n,P}(1)$. Using the fact that $\|\boldsymbol{\theta}\|_2 = \sqrt{n} + on, P(1)$ and $\|\boldsymbol{P}_{\boldsymbol{\theta}}^{\perp}\boldsymbol{g}\|_2 = \sqrt{n} + on, P(1)$ (both hold by the law of large numbers, since $\int \theta^2\,\pi_\Theta(\mathrm{d}\theta) = 1$), we obtain

$$\operatorname*{p-lim}_{n\to\infty} \frac{1}{n}\big\|\boldsymbol{\nu}_+ - (\beta^2-1)\boldsymbol{\theta} - \sqrt{\beta^2-1}\boldsymbol{g}\big\|_2^2 = 0\,.$$

By Lemma D.3.1 we can assume without loss of generality $\Im\varphi_{\pi_\Theta}((\beta^2-1)t_0) > \delta_0 > 0$ for some $delta_0, t_0 > 0$. We then define

$$\mathcal{A}(\boldsymbol{\nu}) := \mathrm{sign}\, T_n(\boldsymbol{\nu})\,, \qquad T_n(\boldsymbol{\nu}) := \frac{1}{n}\sum_{i=1}^n \sin(t_0\nu_i)\,. \tag{D.10}$$

Note that $T_n(\boldsymbol{\nu}) = s\,T_n(\boldsymbol{\nu}_+)$ and therefore the proof is completed by showing that, with high probability $T_n(\boldsymbol{\nu}_+) > 0$. Indeed, let $\delta_1 := \exp(-(\beta^2-1)t_0^2/2)\delta_0$. Then, letting $\overline{\boldsymbol{\nu}}_+ := (\beta^2-1)\boldsymbol{\theta} - \sqrt{\beta^2-1}\boldsymbol{g}$, by

Eq (D.10), we have, with high probability

$$\big|T_n(\overline{\boldsymbol{\nu}}_+) - T_n(\boldsymbol{\nu}_+)\big| \le \frac{\delta_1}{2}\,. \tag{D.11}$$

On the other hand, by the law of large numbers (for $(\Theta, G) \sim \pi_\Theta \otimes \mathsf{N}(0,1)$)

$$\begin{aligned}\operatorname*{p-lim}_{n\to\infty} T_n(\overline{\boldsymbol{\nu}}_+) &= \mathbb{E}\sin((\beta^2-1)\Theta - \sqrt{\beta^2-1}G) &\text{(D.12)}\\ &= \exp\big(-(\beta^2-1)t_0^2/2\big)\,\Im\varphi_{\pi_\Theta}((\beta^2-1)t_0) \ge \delta_1\,. &\text{(D.13)}\end{aligned}$$

Together with the previous display, this completes the proof.

D.4 Proof of Lemma 5.6.1

The claim of Lemma 5.6.1 can be restated as follows. For all $\beta \ge \beta_0$ and $T > t > 0$, there exists $K(\beta, T, \varepsilon) \in \mathbb{N}_{>0}$ depending only on $(\beta, T, \varepsilon, \pi_\Theta)$, such that with probability $1 - o_n(1)$ as $n \to \infty$, it holds that

$$\frac{1}{\sqrt{n}}\|\boldsymbol{m}(\boldsymbol{y}(t), t) - \hat{\boldsymbol{m}}^{K(\beta,T,\varepsilon)}(\boldsymbol{y}(t), t)\|_2 \le \varepsilon. \tag{D.14}$$

We will consider separately the case of non-symmetric and symmetric prior π_Θ. The latter case is more challenging and requires new ideas with respect to earlier work, e.g. [4].

D.4.1 Proof for non-symmetric π_Θ

Recall the state evolution sequence $(\alpha_t^k)_{k\ge0}$ is defined by Eq. (5.13). For $k \in \mathbb{N}$ and $t \in \mathbb{R}_{\ge0}$, we define $\gamma_k(\beta, t) = \alpha_t^k - t$. Hence $\gamma_k(\beta, t)$ is defined by the recursion:

$$\begin{aligned}\gamma_{k+1}(\beta,t) &= \beta^2(1 - \mathsf{mmse}(\gamma_k(\beta,t)+t)), \qquad \gamma_0(\beta,t) = \beta^2 - 1 - t\,.\\ \mathsf{mmse}(\gamma) &:= \mathbb{E}[(\Theta - \mathbb{E}[\Theta \mid \gamma\Theta + \sqrt{\gamma}G])^2]\\ &= \inf_{f:\mathbb{R}\to\mathbb{R}} \mathbb{E}[(\Theta - f(\gamma\Theta + \sqrt{\gamma}G))^2]\end{aligned}$$

(Expectation is with respect to $(\Theta, G) \sim \pi_\Theta \otimes \mathsf{N}(0,1)$.)

By using $f(y) = y/(\gamma+1)$ in the expression for mmse, we get

$$\beta^2(1 - \mathsf{mmse}(\gamma+t)) \ge \beta^2(1 - (\gamma+t+1)^{-1})\,,$$

with the inequality strict unless π_Θ is the Gaussian measure, and $t = 0$. This in turns implies $\beta^2(1 - \mathsf{mmse}(\gamma+t)) \ge \gamma$ for $\gamma = \gamma_0(\beta, t)$.

Recall that Φ is defined in Eq. (5.15), and $\gamma_*(\beta, t)$ is the unique global maximizer of $\gamma \mapsto \Phi(\gamma, \beta, t)$ over $\gamma \in (0, \infty)$. Taking the partial derivative of $\Phi(\gamma, \beta, t)$ with respect to γ, we obtain

$$\frac{\partial}{\partial\gamma}\Phi(\gamma,\beta,t) = \frac{\gamma}{2\beta^2} - \frac{1}{2} + \frac{1}{2}\mathsf{mmse}(\gamma+t),$$

Therefore, $\gamma_*(\beta,t)$ the first strictly positive solution of

$$\gamma_* = \beta^2(1-\mathsf{mmse}(\gamma_*+t))\,. \tag{D.15}$$

Note that $\gamma_* < \infty$ because the right-hand side is bounded by β^2. Since by definition $\beta^2(1-\mathsf{mmse}(\gamma_*(\beta,t)+t)) > \gamma$ for $\gamma \in [0,\gamma_*(\beta,t))$, we have

$$\lim_{k\to\infty} \gamma_k(\beta,t) = \gamma_*(\beta,t)\,. \tag{D.16}$$

Our next proposition provides a more quantitative estimate on this convergence.

Proposition D.4.1. *Assume π_Θ to have support in $[-M_\Theta, M_\Theta]$ (not necessarily symmetric). Then there exists $\beta_0 = \beta_0(\pi_\Theta) > 0$ depending only on π_Θ, such that for all $\beta \geq \beta_0(\pi_\Theta)$ and $t \geq 0$, $\gamma_*(\beta,t)$ is the unique positive solution of the fixed point equation D.15.*

Furthermore, for all $k \geq 0$,

$$1-2^{-k} \leq \frac{\gamma_k(\beta,t)}{\gamma_*(\beta,t)} \leq 1.$$

Proof.[Proof of Proposition D.4.1] We define the non-decreasing function

$$\begin{aligned} H(\gamma) &:= \beta^2(1-\mathsf{mmse}(\gamma)) && \text{(D.17)}\\ &= \beta^2\Big(1-\mathbb{E}\big[(\Theta - f_{\rm B}(Y;\gamma))^2\big]\Big)\,, && \text{(D.18)}\end{aligned}$$

where $Y = \gamma\Theta + \sqrt{\gamma}G$ and

$$f_{\rm B}(y;\gamma) := \mathbb{E}[\Theta|\gamma\Theta+\sqrt{\gamma}G = y] = \mathsf{F}\Big(\frac{y}{\sqrt{\gamma}};\gamma\Big)\,.$$

We then have (here we repeatedly use the fact that $\mathbb{E}[(\Theta - f_{\rm B}(Y;\gamma))\,h(Y)] = 0$ for any function h such that the expectation exists)

$$\begin{aligned} H'(\gamma) &= 2\beta^2\mathbb{E}\Big[\big(\Theta - f_{\rm B}(Y;\gamma)\big)\partial_\gamma f_{\rm B}(Y;\gamma)\Big] + 2\beta^2\mathbb{E}\Big[\Big(\Theta - f_{\rm B}(Y;\gamma)\Big)\partial_Y f_{\rm B}(Y;\gamma)\big(\Theta + (G/2)\gamma^{-1/2}\big)\Big]\\ &= -\beta^2\gamma^{-1/2}\mathbb{E}\big[\big(\Theta - f_{\rm B}(Y;\gamma)\big)\,\mathrm{Var}(\Theta|Y)G\big]\\ &\leq \beta^2\gamma^{-1/2}\mathsf{mmse}(\gamma)^{1/2}\mathbb{E}\big[\,\mathrm{Var}(\Theta|Y)^2G^2\big]^{1/2}\\ &\leq 2\beta^2\gamma^{-1/2}\mathsf{mmse}(\gamma)^{1/2}\mathbb{E}\big[\,\mathrm{Var}(\Theta|Y)^4\big]^{1/4}\\ &\leq 2M_\Theta^{3/2}\beta^2\gamma^{-1/2}\mathsf{mmse}(\gamma)^{3/4}\,,\end{aligned}$$

where the last inequalities follow from Cauchy-Schwartz. Recalling that $\mathsf{mmse}(\gamma) \leq 1/\gamma$, we obtain

$$0 \leq H'(\gamma+t) \leq 2M_\Theta^{3/4}\frac{\beta^2}{(\gamma+t)^{5/4}}\,. \tag{D.19}$$

And therefore, for $\gamma \geq \gamma_0(\beta,t)$, and all $\beta \geq \beta_0(\pi_\Theta)$, we proved that $0 \leq H'(\gamma+t) \leq 1/2$

This implies

$$|\gamma_*(\beta,t)-\gamma_{k+1}(\beta,t)| = |H(t+\gamma_*(\beta,t)) - H(t+\gamma_k(\beta,t))| \le \frac{1}{2}\left|\gamma_*(\beta,t)-\gamma_k(\beta,t)\right| \,,$$

which concludes the proof of the proposition.

□

Proposition D.4.2. *For any $\beta \ge \beta_*$ and $t \ge 0$, we have*

$$\lim_{n\to\infty}\frac{1}{n}\mathbb{E}\left[\|\boldsymbol{\theta}-\boldsymbol{m}(\boldsymbol{y}(t),t)\|_2^2\right] = 1-\frac{\gamma_*(\beta,t)}{\beta^2}.$$

Proof.[Proof of Proposition D.4.2] This is a direct consequence of discussions in Section 2.4 of [152]. In particular, Proposition 2.2.

□

We finally notice that

$$\frac{1}{n}\mathbb{E}\Big\{\|\boldsymbol{m}(\boldsymbol{y}(t),t)-\hat{\boldsymbol{m}}^k(\boldsymbol{y}(t),t)\|_2^2\Big\} = \frac{1}{n}\mathbb{E}\Big\{\|\boldsymbol{\theta}-\hat{\boldsymbol{m}}^k(\boldsymbol{y}(t),t)\|_2^2\Big\} - \frac{1}{n}\mathbb{E}\Big\{\|\boldsymbol{\theta}-\boldsymbol{m}(\boldsymbol{y}(t),t)\|_2^2\Big\}\,. \tag{D.20}$$

Therefore taking the limit $n\to\infty$, using Proposition D.4.2, and noticing that all the expectations are of bounded random variables, we get

$$\operatorname*{p-lim}_{n\to\infty}\frac{1}{n}\|\boldsymbol{m}(\boldsymbol{y}(t),t)-\hat{\boldsymbol{m}}^k(\boldsymbol{y}(t),t)\|_2^2 \le \frac{\gamma_*(\beta,t)}{\beta^2}-\frac{\gamma_k(\beta,t)}{\beta^2}\,. \tag{D.21}$$

The proof is completed by applying Proposition D.4.1.

D.4.2 Proof for symmetric π_Θ

We now consider the case of symmetric π_Θ. In this case, the posterior $\mu_{\boldsymbol{X},0}(\mathrm{d}\boldsymbol{\theta})$ is symmetric under flip $\boldsymbol{\theta}\to-\boldsymbol{\theta}$, and the original vector $\boldsymbol{\theta}$ is identifiable only up to a global sign. We let $\boldsymbol{v}_1=\boldsymbol{v}_1(\boldsymbol{X})$ be a uniformly random eigenvector of $\boldsymbol{X}$, and denote by $\mathbb{P}_0$ the joint distribution of $\boldsymbol{\theta},\boldsymbol{X},\boldsymbol{v}_1$. (Since the top eigenvalue is almost surely non-degenerate, there are two possible choices for $\boldsymbol{v}_1$ given $\boldsymbol{X}$.) We denote by $\mathbb{P}_+$ the same distribution, conditioned to $\langle\boldsymbol{v}_1,\boldsymbol{\theta}\rangle>0$:

$$\begin{aligned}\mathbb{P}_+(\mathrm{d}\boldsymbol{\theta},\mathrm{d}\boldsymbol{X},\mathrm{d}\boldsymbol{v}_1) &= \frac{1}{\mathbb{P}_0(\langle\boldsymbol{v}_1,\boldsymbol{\theta}\rangle>0)}\mathbb{P}_0(\mathrm{d}\boldsymbol{\theta},\mathrm{d}\boldsymbol{X},\mathrm{d}\boldsymbol{v}_1)\,\mathbf{1}\{\langle\boldsymbol{v}_1,\boldsymbol{\theta}\rangle\ge 0\} &\text{(D.22)}\\ &= 2\,\mathbb{P}_0(\mathrm{d}\boldsymbol{\theta},\mathrm{d}\boldsymbol{X},\mathrm{d}\boldsymbol{v}_1)\,\mathbf{1}\{\langle\boldsymbol{v}_1,\boldsymbol{\theta}\rangle\ge 0\}\,. &\text{(D.23)}\end{aligned}$$

Note that under $\mathbb{P}_0$, $\boldsymbol{v}_1$, and $\boldsymbol{\theta}$ are conditionally independent given $\boldsymbol{X}$, while they are not under $\mathbb{P}_+$. Also, the marginal law of $\boldsymbol{X},\boldsymbol{v}_1$ is the same under the two distributions.

Conditionally on $\boldsymbol{X},\boldsymbol{v}_1\sim\mathbb{P}_0$, we let $\boldsymbol{\theta}_1^+,\boldsymbol{\theta}_2^+,\dots$ be i.i.d. vectors with distribution $\mathbb{P}_+(\boldsymbol{\theta}\in\ \cdot\ |\boldsymbol{X},\boldsymbol{v}_1)$ and $\boldsymbol{\theta}_1^0,\boldsymbol{\theta}_2^0,\dots$ be i.i.d. vectors with distribution $\mathbb{P}_0(\boldsymbol{\theta}\in\ \cdot\ |\boldsymbol{X},\boldsymbol{v}_1)$ (independent of the $\boldsymbol{\theta}_i^+$'s). We will use the fact that, by an application of Remark D.1.1, if $\boldsymbol{X}=\beta\boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}}/n+\boldsymbol{W}$,

$$\Big(\boldsymbol{X},\boldsymbol{v}_1,\boldsymbol{\theta}_1^0,\dots,\boldsymbol{\theta}_k^0,\boldsymbol{\theta}_1^+,\dots,\boldsymbol{\theta}_k^+\Big) \overset{\mathrm{d}}{=} \Big(\boldsymbol{X},\boldsymbol{v}_1,\boldsymbol{\theta},\boldsymbol{\theta}_1^0\dots,\boldsymbol{\theta}_{k-1}^0,\boldsymbol{\theta}_1^+,\dots,\boldsymbol{\theta}_k^+\Big)\,. \tag{D.24}$$

We first prove a concentration result $\langle \boldsymbol{\theta}_1^+, \boldsymbol{\theta}_2^+ \rangle$.

Lemma D.4.1. *Let $\mathscr{D}(\beta)$ be the set of discontinuity points of $t \mapsto \gamma_*(t, \beta)$. Then, for all $t \in \mathbb{R}_{\geq 0} \setminus \mathscr{D}(\beta)$, we have*

$$\lim_{n\to\infty} \frac{1}{n^2} \mathbb{E}[(\langle \boldsymbol{\theta}_1^+, \boldsymbol{\theta}_2^+ \rangle - \mathbb{E}[\langle \boldsymbol{\theta}_1^+, \boldsymbol{\theta}_2^+ \rangle])^2] = 0\,, \tag{D.25}$$

$$\lim_{n\to\infty} \frac{1}{n} \mathbb{E}[\langle \boldsymbol{\theta}_1^+, \boldsymbol{\theta}_2^+ \rangle] = \frac{\gamma_*(\beta, t)}{\beta^2}\,. \tag{D.26}$$

Remark D.4.1. Note that $\mathscr{D}(\beta)$ is countable by monotonicity of $\gamma_*(\,\cdot\,, \beta)$, and further it is empty for all $\beta \geq \beta_*(\pi_\Theta)$. Hence, in applying this lemma, we will disregard the set of exceptional points $\mathscr{D}(\beta)$.

Proof.[Proof of Lemma D.4.1] We divide the proof into two parts depending on the value of t.

Case I: $\boldsymbol{t} > 0$. We begin with a useful concentration result.

Lemma D.4.2. *For all $0 < t_1 < t_2$, we have*

$$\lim_{n\to\infty} \frac{1}{n^2} \int_{t_1}^{t_2} \mathbb{E}\Big[\Big(\langle \boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2^0 \rangle - \mathbb{E}[\langle \boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2^0 \rangle]\Big)^2\Big]\, \mathrm{d}t = 0\,. \tag{D.27}$$

We present the proof of this fact in Appendix D.7.1. A similar statement is proven in [125].

Let us next show that this implies the desired concentration result:

$$\begin{aligned}
\mathbb{E}\left[(\langle \boldsymbol{\theta}_1^+, \boldsymbol{\theta}_2^+ \rangle - \mathbb{E}[\langle \boldsymbol{\theta}_1^+, \boldsymbol{\theta}_2^+ \rangle])^2\right] \leq & \mathbb{E}\left[(\langle \boldsymbol{\theta}_1^+, \boldsymbol{\theta}_2^+ \rangle - \mathbb{E}[\langle \boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2^0 \rangle])^2\right] \\
= & \frac{\mathbb{E}\left[(\langle \boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2^0 \rangle - \mathbb{E}[\langle \boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2^0 \rangle])^2 \mathbf{1}_{\langle \boldsymbol{v}_1, \boldsymbol{\theta}_0^1 \rangle \geq 0, \langle \boldsymbol{v}_1, \boldsymbol{\theta}_2^0 \rangle \geq 0)}\right]}{\mathbb{P}(\langle \boldsymbol{v}_1, \boldsymbol{\theta}_1^0 \rangle \geq 0, \langle \boldsymbol{v}_1, \boldsymbol{\theta}_2^0 \rangle \geq 0)} \\
\overset{(i)}{\leq} & 4\mathbb{E}[(\langle \boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2^0 \rangle - \mathbb{E}[\langle \boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2^0 \rangle])^2],
\end{aligned} \tag{D.28}$$

Here in *(i)* we made use of the fact that $\boldsymbol{v}_1, \boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2^0$ are conditionally independent given $\boldsymbol{X}$, implying

$$\mathbb{P}(\langle \boldsymbol{v}_1, \boldsymbol{\theta}_1^0 \rangle \geq 0, \langle \boldsymbol{v}_1, \boldsymbol{\theta}_2^0 \rangle \geq 0 | \boldsymbol{X}) = \mathbb{P}(\langle \boldsymbol{v}_1, \boldsymbol{\theta}_1^0 \rangle \geq 0 | \boldsymbol{X})^2 = \frac{1}{4}\,, \tag{D.29}$$

and therefore the same identity holds unconditionally.

Recall that, by [125], it holds that (for any $t \geq 0$)

$$\lim_{n\to\infty} \frac{1}{n^2} \mathbb{E}[\langle \boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2^0 \rangle^2] = \lim_{n\to\infty} \frac{1}{n^2} \mathbb{E}\Big\{\big\|\mathbb{E}\big[\boldsymbol{\theta}\boldsymbol{\theta}^{\mathsf{T}} | \boldsymbol{X}\big]\big\|_F^2\Big\} = \frac{\gamma_*(\beta, t)^2}{\beta^4}\,. \tag{D.30}$$

Using this, together with the concentration property (D.27), we get, for all $0 < t_1 < t_2$,

$$\lim_{n\to\infty} \frac{1}{n^2} \int_{t_1}^{t_2} \Big(\mathbb{E}[\langle \boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2^0 \rangle] - \frac{\gamma_*(\beta, t)}{\beta^2}\Big)^2 \mathrm{d}t = 0\,.$$

Since $t \mapsto \mathbb{E}[\langle \boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2^0 \rangle] = \mathbb{E}[\|\mathbb{E}[\boldsymbol{\theta} | \boldsymbol{X}, \boldsymbol{y}(t)]\|^2]$ is non-decreasing (by Jensen), the last limit holds pointwise.

Namely, at all continuity points of $t \mapsto \gamma_*(\beta, t)$,

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{E}[\langle \boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2^0\rangle] = \frac{\gamma_*(\beta,t)}{\beta^2}\,. \tag{D.31}$$

Using this and (D.30), we get, on $t \in (0,\infty) \setminus \mathscr{D}(\beta)$

$$\lim_{n\to\infty} \frac{1}{n^2}\mathbb{E}[(\langle \boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2^0\rangle - \mathbb{E}[\langle \boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2^0\rangle])^2] = 0\,, \tag{D.32}$$

and therefore, using (D.28), we obtain the claim (D.25).

Finally, notice that the following is a consequence of Eq. (D.28):

$$\lim_{n\to\infty}\left(\frac{1}{n}\mathbb{E}[\langle \boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2^0\rangle] - \frac{1}{n}\mathbb{E}[\langle \boldsymbol{\theta}_1^+, \boldsymbol{\theta}_2^+\rangle]\right)^2 = 0.$$

Putting the last limit together with Eq. (D.31) implies Eq. (D.26).

Case II: $\boldsymbol{t} = 0$. We begin with establishing the following lemma.

Lemma D.4.3. *Let $\beta_*(\pi_\Theta)$ be as in the statement of Theorem 5.3.1. Then for any $\beta \ge \beta_*(\pi_\Theta)$ and any $t \ge 0$, we have*

$$\lim_{n\to\infty} \frac{1}{n^4}\mathbb{E}\Big[\Big(\langle \boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2^0\rangle^2 - \mathbb{E}[\langle \boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2^0\rangle^2]\Big)^2\Big] = 0\,. \tag{D.33}$$

Proof. Recall that Φ is defined in Eq. (5.15). We let $\gamma_*(\beta,t)$ be the first stationary point of $\gamma \mapsto \Phi(\gamma,\beta,t)$ on $(0,\infty)$. Following the notation of [125], we let

$$D_t := \{\beta > 0 : \gamma \mapsto \Phi(\gamma,\beta,t) \text{ has a unique minimizer}\}\,.$$

By the assumptions of Theorem 5.3.1, we know that $[\beta_*, \infty) \subseteq D_t$. Then, the claim of the lemma is a direct consequence of [125, Theorem 20].

□

By Lemma D.4.3, and repeating the argument of Eq. (D.28), we get

$$\begin{aligned}\mathbb{E}\left[(\langle \boldsymbol{\theta}_1^+, \boldsymbol{\theta}_2^+\rangle^2 - \mathbb{E}[\langle \boldsymbol{\theta}_1^+, \boldsymbol{\theta}_2^+\rangle^2])^2\right] \le& \mathbb{E}\left[(\langle \boldsymbol{\theta}_1^+, \boldsymbol{\theta}_2^+\rangle^2 - \mathbb{E}[\langle \boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2^0\rangle^2])^2\right] \\ =& \frac{\mathbb{E}\left[(\langle \boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2^0\rangle^2 - \mathbb{E}[\langle \boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2^0\rangle^2])^2 \mathbf{1}_{\langle \boldsymbol{v}_1, \boldsymbol{\theta}_0^1\rangle \ge 0, \langle \boldsymbol{v}_1, \boldsymbol{\theta}_2^0\rangle \ge 0)}\right]}{\mathbb{P}(\langle \boldsymbol{v}_1, \boldsymbol{\theta}_1^0\rangle \ge 0, \langle \boldsymbol{v}_1, \boldsymbol{\theta}_2^0\rangle \ge 0)} \\ \le& 4\mathbb{E}[(\langle \boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2^0\rangle^2 - \mathbb{E}[\langle \boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2^0\rangle^2])^2]\,.\end{aligned} \tag{D.34}$$

Therefore, by the last lemma,

$$\lim_{n\to\infty} \frac{1}{n^4}\mathbb{E}\Big[\Big(\langle \boldsymbol{\theta}_1^+, \boldsymbol{\theta}_2^+\rangle^2 - \mathbb{E}[\langle \boldsymbol{\theta}_1^+, \boldsymbol{\theta}_2^+\rangle^2]\Big)^2\Big] = 0\,. \tag{D.35}$$

Let $\boldsymbol{\nu}_0 := \sqrt{n}\boldsymbol{v}_1(\boldsymbol{X})/\|\boldsymbol{v}_1(\boldsymbol{X})\|_2$. By [28], we know that $\langle \boldsymbol{\theta}, \boldsymbol{\nu}_0\rangle^2/n^2 \overset{\mathrm{d}}{=} \langle \boldsymbol{\theta}_1^0, \boldsymbol{\nu}_0\rangle^2/n^2 \xrightarrow{P} 1 - \beta^{-2}$. Since

$\mathbb{P}_+$ is contiguous to $\mathbb{P}_0$, we obtain $\langle \boldsymbol{\theta}_1^+, \boldsymbol{\nu}_0\rangle/n \xrightarrow{P} \sqrt{1-\beta^{-2}}$. and therefore:

$$\operatorname*{p-lim}_{n\to\infty} \frac{1}{n}\|\boldsymbol{\theta}_1^+ - \boldsymbol{\nu}_0\|_2^2 = 2 - 2\sqrt{1-\beta^{-2}}. \tag{D.36}$$

Recall that $\gamma_*(\beta, t)$ is the first positive stationary point of $\gamma \mapsto \Phi(\gamma, \beta, t)$. From Eqs. (D.34) and (D.35), we see that $|\langle \boldsymbol{\theta}_1^+, \boldsymbol{\theta}_2^+\rangle|/n = \mathbb{E}[\langle \boldsymbol{\theta}, \boldsymbol{\theta}_1^0\rangle^2]^{1/2}/n + o_P(1)$. Further by[125, Theorem 2], we obtain $|\langle \boldsymbol{\theta}, \boldsymbol{\theta}_1^0\rangle|/n = \beta^{-2}\gamma_*(\beta, 0) + o_P(1)$, whence $|\langle \boldsymbol{\theta}_1^+, \boldsymbol{\theta}_2^+\rangle|/n = \beta^{-2}\gamma_*(\beta, 0) + o_P(1)$.

By Cauchy-Schwarz,

$$\frac{1}{n}\|\boldsymbol{\theta}_1^+ - \boldsymbol{\theta}_2^+\|_2^2 \le \frac{2}{n}\|\boldsymbol{\theta}_1^+ - \boldsymbol{\nu}_0\|_2^2 + \frac{2}{n}\|\boldsymbol{\theta}_2^+ - \boldsymbol{\nu}_0\|_2^2 = 4 - 4\sqrt{1-\beta^{-2}} + o_P(1),$$

hence $\langle \boldsymbol{\theta}_1^+, \boldsymbol{\theta}_2^+\rangle/n \ge 2\sqrt{1-\beta^{-2}} - 1 + o_P(1)$. Recall that $|\langle \boldsymbol{\theta}_1^+, \boldsymbol{\theta}_2^+\rangle|/n = \beta^{-2}\gamma_*(\beta, 0) + o_P(1)$, then for β_0 large enough and all $\beta > \beta_0$, it holds that $\langle \boldsymbol{\theta}_1^+, \boldsymbol{\theta}_2^+\rangle/n = \beta^{-2}\gamma_*(\beta, 0) + o_P(1)$. Applying bounded convergence (since $|\langle \boldsymbol{\theta}_1^+, \boldsymbol{\theta}_2^+\rangle/n| \le M_\Theta$), we see that $\mathbb{E}[(\langle \boldsymbol{\theta}_1^+, \boldsymbol{\theta}_2^+\rangle/n - \beta^{-2}\gamma_*(\beta, 0))^2] = o_n(1)$, thus concluding the proof of Lemma D.4.1 for $t = 0$.

□

Next, we will apply Lemma D.4.1 to prove Lemma 5.6.1. By the state evolution of the AMP algorithm, cf. Proposition 5.6.1, we see that

$$\frac{1}{n}\langle \boldsymbol{\theta}^+, \hat{\boldsymbol{m}}^k(\boldsymbol{y}(t), t)\rangle \xrightarrow{P} \mathbb{E}[\mathbb{E}[\Theta \mid \alpha_t^k\Theta + (\alpha_t^k)^{1/2}G]^2] = 1 - \mathsf{mmse}(\alpha_t^k),$$
$$\frac{1}{n}\|\hat{\boldsymbol{m}}^k(\boldsymbol{y}(t), t)\|_2^2 \xrightarrow{P} \mathbb{E}[\mathbb{E}[\Theta \mid \alpha_t^k\Theta + (\alpha_t^k)^{1/2}G]^2] = 1 - \mathsf{mmse}(\alpha_t^k).$$

By Proposition D.4.1, we see that as $k \to \infty$, α_t^k converges linearly to $\gamma_*(\beta, t) + t$, which further implies that $1 - \mathsf{mmse}(\alpha_t^k)$ converges linearly to $1 - \mathsf{mmse}(\gamma_*(\beta, t) + t) = 1 - \beta^{-2}\gamma_*(\beta, t)$. Furthermore, the convergence is uniform in $t \in [0, T]$. Therefore, for all $\varepsilon > 0$, there exists $K(\beta, T, \varepsilon) \in \mathbb{N}_{>0}$ depending only on $(\beta, T, \varepsilon, \pi_\Theta)$, such that for all $k \ge K(\beta, T, \varepsilon)$,

$$\left|\mathbb{E}[\mathbb{E}[\Theta \mid \alpha_t^k\Theta + (\alpha_t^k)^{1/2}G]^2] - \beta^{-2}\gamma_*(\beta, t)\right| \le \frac{\varepsilon}{2}.$$

and therefore, for all $k \ge K(\beta, T, \varepsilon)$, with high probability,

$$\left|\frac{1}{n}\langle \boldsymbol{\theta}^+, \hat{\boldsymbol{m}}^k(\boldsymbol{y}(t), t)\rangle - \beta^{-2}\gamma_*(\beta, t)\right| \le \varepsilon\,, \tag{D.37}$$
$$\left|\frac{1}{n}\|\hat{\boldsymbol{m}}^k(\boldsymbol{y}(t), t)\|_2^2 - \beta^{-2}\gamma_*(\beta, t)\right| \le \varepsilon\,. \tag{D.38}$$

By Lemma D.4.1, it holds that $\text{p-lim}_{n\to\infty}\langle \boldsymbol{\theta}_1^+, \boldsymbol{\theta}_2^+\rangle/n = \beta^{-2}\gamma_*(\beta, t)$. Since $\boldsymbol{\theta}_1^+$ and $\boldsymbol{\theta}_2^+$ are conditionally independent given $(\boldsymbol{X}, \boldsymbol{y}(t), \boldsymbol{v}_1(\boldsymbol{X}))$ this in particular implies:

$$\frac{1}{n}\|\boldsymbol{m}(\boldsymbol{y}(t), t)\|_2^2 = \beta^{-2}\gamma_*(\beta, t) + o_P(1)\,. \tag{D.39}$$

Further $\boldsymbol{\theta}^+$ and $\boldsymbol{m}(\boldsymbol{y}(t), t)$ are conditionally independent given $(\boldsymbol{X}, \boldsymbol{y}(t), \boldsymbol{v}_1(\boldsymbol{X}))$. Hence, from Eq. (D.37), it follows that with high probability

$$\left|\frac{1}{n}\langle \boldsymbol{m}(\boldsymbol{y}(t),t), \hat{\boldsymbol{m}}^{K(\beta,T,\varepsilon)}(\boldsymbol{y}(t),t)\rangle - -\beta^{-2}\gamma_*(\beta,t)\right| \le 2\varepsilon\,. \tag{D.40}$$

Putting together (D.38), (D.39), (D.40), we get

$$\frac{1}{n}\|\boldsymbol{m}(\boldsymbol{y}(t),t) - \hat{\boldsymbol{m}}^{K(\beta,t,\varepsilon)}(\boldsymbol{y}(t),t)\|_2^2 \le 10\,\varepsilon\,,$$

with high probability, thus completing the proof of Lemma 5.6.1.

D.5 Proof of Lemma 5.6.2

The subsequent proof is analogous to the one of [4, Lemma 4.9]. Recall that Φ is defined in Eq. (5.15), and (for $\beta > \beta_*$), $\gamma_*(\beta,t)$ is the unique global maximizer of $\gamma \mapsto \Phi(\gamma,\beta,t)$ over $\gamma \in (0,\infty)$. Taking the partial derivative of $\Phi(\gamma,\beta,t)$ with respect to γ, we obtain

$$\frac{\partial}{\partial\gamma}\Phi(\gamma,\beta,t) = \frac{\gamma}{2\beta^2} - \frac{1}{2} + \frac{1}{2}\mathsf{mmse}(\gamma+t),$$

where we recall that, for $(\Theta,G) \sim \pi_\Theta \otimes \mathsf{N}(0,1)$,

$$\mathsf{mmse}(\gamma) = \mathbb{E}\big[(\Theta - \mathbb{E}[\Theta \mid \gamma\Theta + \sqrt{\gamma}G])^2\big].$$

Therefore, $\gamma_*(\beta,t)$ is a solution to the following fixed point equation.

$$\gamma = \beta^2\mathbb{E}[\mathbb{E}[\Theta \mid (\gamma+t)\Theta + \sqrt{\gamma+t}G]^2]. \tag{D.41}$$

For any $t_1 < t_2$, we have

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\mathbb{E}\left[\|\boldsymbol{m}(\boldsymbol{y}(t_2),t_2)\right.&\left. . - \boldsymbol{m}(\boldsymbol{y}(t_1),t_1)\|_2^2\right] = \\ &= \lim_{n\to\infty}\frac{1}{n}\left\{\mathbb{E}\left[\|\boldsymbol{\theta} - \boldsymbol{m}(\boldsymbol{y}(t_1),t_1)\|_2^2\right] - \mathbb{E}\left[\|\boldsymbol{\theta} - \boldsymbol{m}(\boldsymbol{y}(t_2),t_2)\|_2^2\right]\right\} \\ &= \frac{\gamma_*(\beta,t_2) - \gamma_*(\beta,t_1)}{\beta^2}.\end{aligned} \tag{D.42}$$

By Lemma 5.6.1 we know that for all $t \ge 0$, with high probability $\|\boldsymbol{m}(\boldsymbol{y}(t),t) - \hat{\boldsymbol{m}}^k(\boldsymbol{y}(t),t)\|_2^2/n \le \varepsilon_k$, for some deterministic constants ε_k satisfying $\varepsilon_k \to 0^+$ as $k \to \infty$. Therefore, using the concentration of $\|\hat{\boldsymbol{m}}^k(\boldsymbol{y}(t_2),t_2) - \boldsymbol{m}^k(\boldsymbol{y}(t_1),t_1)\|_2^2/n$, we get

$$\operatorname*{p-lim}_{n\to\infty}\frac{1}{n}\|\boldsymbol{m}(\boldsymbol{y}(t_2),t_2) - \boldsymbol{m}(\boldsymbol{y}(t_1),t_1)\|_2^2 = \frac{\gamma_*(\beta,t_2) - \gamma_*(\beta,t_1)}{\beta^2}. \tag{D.43}$$

Note that $t \mapsto \boldsymbol{m}(\boldsymbol{y}(t),t)$ is a bounded martingale. Hence, for any fixed constant c, the process $Y_{n,t} := (M_{n,t} - c)_+$ is a positive bounded submartingale, where $M_{n,t} = \|\boldsymbol{m}(\boldsymbol{y}(t),t) - \boldsymbol{m}(\boldsymbol{y}(t_1),t_1)\|_2/\sqrt{n}$. By Doob's maximal inequality, we then see that

$$\mathbb{P}\left(\sup_{t\in[t_1,t_2]} Y_{n,t} \geq a\right) \leq \frac{1}{a}\mathbb{E}[Y_{n,t_2}] \leq \frac{1}{a}\mathbb{E}[Y_{n,t_2}^2]^{1/2}$$

for any $a > 0$. Setting $c = \sqrt{\gamma_*(\beta,t_2) - \gamma_*(\beta,t_1)}/\beta$, we have $\text{p-lim}_{n\to\infty} M_{n,t_2}^2 = c^2$ by Eq. (D.43). Since $M_{n,t}$ is bounded, then for any fixed $a > 0$, we obtain:

$$\begin{aligned}\limsup_{n\to\infty}\mathbb{P}\left(\sup_{t\in[t_1,t_2]} M_{n,t} \geq c + a\right) \leq& \limsup_{n\to\infty}\mathbb{P}\left(\sup_{t\in[t_1,t_2]} Y_{n,t} \geq a\right)\\ \leq& \frac{1}{a}\lim_{n\to\infty}\mathbb{E}\left[(M_{n,t_2} - c)^2\right]^{1/2} = 0.\end{aligned}$$

A lower bound can be derived analogously. Thus,

$$\operatorname*{p-lim}_{n\to\infty}\sup_{t\in[t_1,t_2]} M_{n,t}^2 = \frac{\gamma_*(\beta,t_2) - \gamma_*(\beta,t_1)}{\beta^2},$$

which yields

$$\begin{aligned}\operatorname*{p-lim}_{n\to\infty}\sup_{t\in[t_1,t_2]}\frac{1}{n}\|\boldsymbol{m}(\boldsymbol{y}(t),t) - \boldsymbol{m}(\boldsymbol{y}(t_1),t_1))\|_2^2 =& \operatorname*{p-lim}_{n\to\infty}\frac{1}{n}\|\boldsymbol{m}(\boldsymbol{y}(t_2),t_2) - \boldsymbol{m}(\boldsymbol{y}(t_1),t_1)\|_2^2\\ =& \frac{\gamma_*(\beta,t_2) - \gamma_*(\beta,t_1)}{\beta^2}.\end{aligned}$$

Therefore, in order to prove the lemma, it suffices to show the existence of $C_{\text{reg}} > 1$ depending uniquely on (β, π_Θ), such that

$$\frac{|\gamma_*(\beta,t_2) - \gamma_*(\beta,t_1)|}{\beta^2} \leq C_{\text{reg}}|t_1 - t_2|,$$

which follows from Proposition D.4.1. This concludes the proof of the lemma.

D.6 Proof of Lemma 5.6.3

Before proving Lemma 5.6.3, we establish a simple estimate on the conditional variance.

Lemma D.6.1. *There exists a constant $C_{\text{conv}} > 0$ depending only on π_Θ, such that*

$$\mathbb{E}\big[\operatorname{Var}(\Theta \mid \beta^2\Theta + \beta G)\big] \leq C_{\text{conv}}^{-1}\exp(-4C_{\text{conv}}\beta^2)/2. \tag{D.44}$$

Without loss, we can and will assume that $C_{\text{conv}} < 1$.

Proof. We denote by $\{x_1, x_2, \cdots, x_s\}$ the support of π_Θ and assume without loss of generality $x_1 < x_2 < \cdots < x_s$. Define $\hat{\theta} : \mathbb{R} \to \{x_1, x_2, \cdots, x_s\}$ by

$$\hat{\theta}(y) := \operatorname{argmin}\left(|x - y| : \ x \in \operatorname{supp}(\pi_\Theta)\right).$$

In case of ties, we choose the smallest value. We immediately see that $\hat{\theta}(y) = x_i$ if and only if $(x_{i-1}+x_i)/2 <$

$y \le (x_i + x_{i+1})/2$ (with the convention that $x_0 = -\infty$ and $x_{s+1} = +\infty$). Let $Y = \Theta + \beta^{-1}G$, then

$$\mathbb{E}[\mathrm{Var}[\Theta \mid \beta^2\Theta + \beta G]] \le \mathbb{E}[(\Theta - \hat{\theta}(Y))^2].$$

Let $\delta_\Theta = \min\{|x_i - x_{i+1}|/2 : i \in [s-1]\}$. We then have

$$\begin{aligned}
&\mathbb{E}[(\Theta - \hat{\theta}(Y))^2 \mid \Theta = x_i] \\
\le& 0 \times \mathbb{P}\left(Y \in \left((x_{i-1} + x_i)/2, (x_i + x_{i+1})/2\right] \mid \Theta = x_i\right) \\
&+ 4M_\Theta^2 \mathbb{P}\left(Y \in \left((x_{i-1} + x_i)/2, (x_i + x_{i+1})/2\right]^c \mid \Theta = x_i\right) \\
\le& \frac{16M_\Theta^2}{\delta_\Theta \beta \sqrt{2\pi}} e^{-\delta_\Theta^2\beta^2/8},
\end{aligned}$$

where to arrive at the last inequality we make use of Lemma D.1.3. Combining the above bounds, we obtain that $\mathbb{E}[(\Theta - \hat{\theta}(Y))^2] \le \frac{16M_\Theta^2}{\delta_\Theta \beta \sqrt{2\pi}} e^{-\delta_\Theta^2\beta^2/8}$. Using this fact, we conclude that there exists a constant $C_{\text{conv}} > 0$ that is a function of π_Θ only, such that Eq. (D.44) holds.

□

Proof.[Proof of Lemma 5.6.3] By the state evolution of Bayes AMP, Proposition 5.6.1, we have

$$\frac{1}{n}\|\boldsymbol{D}_{\alpha_t^k}(\hat{\boldsymbol{m}}^k(\boldsymbol{y}(t), t))\|_F^2 \xrightarrow{P} \mathbb{E}[\mathrm{Var}[\Theta \mid \alpha_t^k\Theta + (\alpha_t^k)^{1/2}G]]. \tag{D.45}$$

Since $\alpha_t^0 = \beta^2 - 1$ and $\alpha_t^k \ge \alpha_t^0$ (this follows from instance by the fact that $\gamma \mapsto \mathsf{mmse}(\gamma)$ is non-increasing, see the discussion at the beginning of Section D.4.1), we conclude that for $\beta > 2$,

$$\mathbb{E}[\mathrm{Var}[\Theta \mid \alpha_t^k\Theta + (\alpha_t^k)^{1/2}G]] \le \mathbb{E}[\mathrm{Var}[\Theta \mid \beta^2\Theta/4 + \beta G/2]]. \tag{D.46}$$

By Lemma D.6.1 below, we obtain that there exists a constant $C_{\text{conv}} > 0$ depending uniquely on π_Θ, such that

$$\mathbb{E}[\mathrm{Var}[\Theta \mid \beta^2\Theta/4 + \beta G/2]] \le C_{\text{conv}}^{-1}\exp(-C_{\text{conv}}\beta^2)/2. \tag{D.47}$$

Equation (5.46) follow from Eqs. (D.45) to (D.47).

By definition $b_t^k = \beta^2\mathbb{E}[\mathrm{Var}[\Theta \mid \alpha_t^k\Theta + (\alpha_t^k)^{1/2}G]]$, which by Eqs. (D.46) and (D.47) is no larger than $C_{\text{conv}}^{-1}\exp(-C_{\text{conv}}\beta^2)/2$, thus completing the proof of Eq. (5.47).

As for Eqs. (5.48) and (5.49), we will in fact show a stronger result and prove that these two inequalities hold for all $k \le k_0(\beta) + 1$, via induction over k. We already observed that, with probability $1 - o_n(1)$ we have $\|\boldsymbol{X}\|_{\text{op}} \le \beta + \|\boldsymbol{W}\|_{\text{op}} \le \beta + 2$ [7], and will work on this high-probability event.

For the base case $k = 0$, the claim directly follows as $\hat{\boldsymbol{m}}^0(\boldsymbol{y}_1, t) = \hat{\boldsymbol{m}}^0(\boldsymbol{y}_2, t) = \mathbb{E}[\Theta \mid \alpha_t^0\Theta + (\alpha_t^0)^{1/2}G = \boldsymbol{\nu}]$ and hence $\hat{\boldsymbol{p}}^0(\boldsymbol{y}_1, t) = \hat{\boldsymbol{p}}^0(\boldsymbol{y}_2, t)$. Now suppose for all $k \le k_1$ and all $\boldsymbol{y}_1$, $\boldsymbol{y}_2$, we have

$$\begin{aligned}
\frac{1}{\sqrt{n}}\|\hat{\boldsymbol{m}}^k(\boldsymbol{y}_1, t) - \hat{\boldsymbol{m}}^k(\boldsymbol{y}_2, t)\|_2 &\le \frac{\mathrm{Lip}_0(\beta, k)}{\sqrt{n}}\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2, \\
\frac{1}{\sqrt{n}}\|\hat{\boldsymbol{p}}^k(\boldsymbol{y}_1, t) - \hat{\boldsymbol{p}}^k(\boldsymbol{y}_2, t)\|_2 &\le \frac{\mathrm{Lip}_0(\beta, k)}{\sqrt{n}}\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2,
\end{aligned}$$

where $\mathrm{Lip}_0(\beta, k) > 0$ is a function of (β, k) only. We then prove that the above statement also holds for $k = k_1 + 1$.

Recalling the definition $\mathsf{F}(x; \alpha) := \mathbb{E}[\Theta \mid \alpha\Theta + \alpha^{1/2}G = z]$, a computation of the derivatives shows that the mappings

$$z \mapsto \mathsf{F}(z; \alpha_t^{k+1}), \qquad z \mapsto \Phi^{-1}_{\alpha_t^{k+1}}\left(\mathsf{F}(z; \alpha_t^{k+1})\right),$$

are M_Θ^2-Lipschitz and M_Θ-Lipschitz, respectively, where $\mathrm{supp}(\pi_\Theta) \subseteq [-M_\Theta, M_\Theta]$. As a result, on the event $\|\boldsymbol{X}\|_{\mathrm{op}} \le \beta + 2$, we have

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\|\hat{\boldsymbol{m}}^{k_1+1}(\boldsymbol{y}_1, t_1) - \hat{\boldsymbol{m}}^{k_1+1}(\boldsymbol{y}_2, t_2)\|_2 \\
\le &\frac{M_\Theta^2}{\sqrt{n}}\|\beta\boldsymbol{X}(\hat{\boldsymbol{m}}^{k_1}(\boldsymbol{y}_1, t) - \hat{\boldsymbol{m}}^{k_1}(\boldsymbol{y}_2, t)) + \boldsymbol{y}_1 - \boldsymbol{y}_2 - b_t^{k_1}(\hat{\boldsymbol{m}}^{k_1-1}(\boldsymbol{y}_1, t) - \hat{\boldsymbol{m}}^{k_1-1}(\boldsymbol{y}_2, t))\|_2 \\
\le &\frac{M_\Theta^2(\beta^2 + \beta + 1)\mathrm{Lip}_0(\beta, k_1)}{\sqrt{n}}\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2 + \frac{M_\Theta^2}{\sqrt{n}}\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2 + \frac{C_{\mathrm{conv}}^{-1}M_\Theta^2\mathrm{Lip}_0(\beta, k_1 - 1)}{\sqrt{n}}\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2.
\end{aligned}$$

Similarly, under the $\boldsymbol{p}$-parameterization we have

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\|\hat{\boldsymbol{p}}^{k_1+1}(\boldsymbol{y}_1, t) - \hat{\boldsymbol{p}}^{k_1+1}(\boldsymbol{y}_2, t)\|_2 \\
\le &\frac{M_\Theta}{\sqrt{n}}\|\beta\boldsymbol{X}(\hat{\boldsymbol{m}}^{k_1}(\boldsymbol{y}_1, t) - \hat{\boldsymbol{m}}^{k_1}(\boldsymbol{y}_2, t)) + \boldsymbol{y}_1 - \boldsymbol{y}_2 - b_t^{k_1}(\hat{\boldsymbol{m}}^{k_1-1}(\boldsymbol{y}_1, t) - \hat{\boldsymbol{m}}^{k_1-1}(\boldsymbol{y}_2, t))\|_2 \\
\le &\frac{M_\Theta(\beta^2 + \beta + 1)\mathrm{Lip}_0(\beta, k_1)}{\sqrt{n}}\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2 + \frac{M_\Theta}{\sqrt{n}}\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2 + \frac{C_{\mathrm{conv}}^{-1}M_\Theta\mathrm{Lip}_0(\beta, k_1 - 1)}{\sqrt{n}}\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2.
\end{aligned}$$

As a result, we see that setting $\mathrm{Lip}_0(\beta, k_1 + 1) = M_\Theta^2((\beta^2 + \beta + 1)\mathrm{Lip}_0(\beta, k_1) + 1 + C_{\mathrm{conv}}^{-1}\mathrm{Lip}_0(\beta, k_1 - 1))$ concludes the proof of the induction step. This further completes the proof of Eq. (5.48) and Eq. (5.49), thus finishing the proof of the lemma.

□

D.7 Proof of Lemma 5.6.5

We first state a simplified version of Lemma 5.6.4. More precisely, for $q \in (0, 1)$, note that $\sqrt{q\log(e/q)} \le 3q^{1/4}$. If we substitute this result into Eq. (5.51), normalize $\boldsymbol{t}_1, \boldsymbol{t}_2$, and set $\xi = \Delta^{1/6}$, $q = \Delta^{2/3}$, then we obtain the next corollary.

Corollary D.7.1. *Under the conditions of Lemma 5.6.4, for all $\Delta > 0$ and $M > 0$, we have*

$$\mathbb{P}\left(\sup_{\substack{\boldsymbol{t}_1, \boldsymbol{t}_2 \in [0, M]^n, \\ \|\boldsymbol{t}_1\|_2^2/n \le \Delta, \|\boldsymbol{t}_2\|_2^2/n \le \Delta}} \|\mathrm{diag}(\boldsymbol{t}_1)\boldsymbol{W}\,\mathrm{diag}(\boldsymbol{t}_2)\|_{\mathrm{op}} \ge 4C'M^{5/3}\Delta^{1/6}\right) \le Ce^{-cn\Delta^{2/3}M^{-4/3}}. \tag{D.48}$$

We denote by $\mathscr{E}_{\beta,n}^{(4)}$ the event depicted by Eq. (D.48), with $\Delta = \Delta(\beta)$. More precisely, let

$$\mathscr{E}_{\beta,n}^{(4)} := \left\{ \sup_{\boldsymbol{t}_1,\boldsymbol{t}_2 \in [0,M_\Theta]^n, \|\boldsymbol{t}_1\|_2^2/n \leq \Delta(\beta), \|\boldsymbol{t}_2\|_2^2/n \leq \Delta(\beta)} \|\operatorname{diag}(\boldsymbol{t}_1)\boldsymbol{W}\operatorname{diag}(\boldsymbol{t}_2)\|_{\mathrm{op}} \geq 4C' M_\Theta^{5/3}\Delta(\beta)^{1/6} \right\}. \tag{D.49}$$

Throughout the proof, we will make use of the following functions:

$$\begin{aligned}
&\Delta(\beta) := C_{\mathrm{conv}}^{-1} e^{-C_{\mathrm{conv}}\beta^2} + M_\Theta^2 \cdot (2\beta^2 + 2\beta + 4 + 2C_{\mathrm{conv}}^{-1} e^{-C_{\mathrm{conv}}\beta^2}) e^{-C_{\mathrm{conv}}\beta^2/4}, \\
&\rho(\beta) := \beta^2 M_\Theta^2 \Delta(\beta) + 4\beta C' M_\Theta^{5/3}\Delta(\beta)^{1/6}, \\
&F(\beta) := \rho(\beta) + M_\Theta^2 \cdot C_{\mathrm{conv}}^{-1} e^{-C_{\mathrm{conv}}\beta^2}.
\end{aligned} \tag{D.50}$$

We can and will choose β_0 large enough such that $F(\beta) \leq 1/2$ holds for all $\beta \geq \beta_0$. To simplify notations, we define

$$\begin{aligned}
\boldsymbol{m}^k &= \hat{\boldsymbol{m}}^k(\boldsymbol{y}_1, t), \\
\tilde{\boldsymbol{m}}^k &= \hat{\boldsymbol{m}}^k(\boldsymbol{y}_2, t), \\
\boldsymbol{p}^k &= \Psi_{\alpha_t^k}^{-1}(\boldsymbol{m}^k), \\
\tilde{\boldsymbol{p}}^k &= \Psi_{\alpha_t^k}^{-1}(\tilde{\boldsymbol{m}}^k).
\end{aligned}$$

We will choose $r(\beta)$ (depending uniquely on π_Θ, β) small enough so that $2r(\beta) \cdot (\mathrm{Lip}_0(\beta) + 1) \cdot (M_\Theta^2 + 1) \leq 2e^{-C_{\mathrm{conv}}\beta^2/4}$. Notice that indeed the choice of $r(\beta)$ can only depend on (β, π_Θ). By Lemma 5.6.3, for all $\boldsymbol{y}_1, \boldsymbol{y}_2 \in B^n(\boldsymbol{y}(t), r(\beta))$, we know that

$$\frac{1}{\sqrt{n}}\|\boldsymbol{p}^{k_*} - \tilde{\boldsymbol{p}}^{k_*}\|_2 \leq \frac{\mathrm{Lip}_0(\beta)}{\sqrt{n}}\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2 \leq 2\exp\left(-\frac{1}{4}C_{\mathrm{conv}}\beta^2\right) \tag{D.51}$$

for $k_* \in \{k_0(\beta), k_0(\beta) \pm 1\}$. Without loss, we can and will assume that $\mathrm{Lip}_0(\beta) \geq 2M_\Theta$.

We define $\mathscr{E}_{\beta,L,\delta,\varepsilon,n} = \mathscr{E}_{L,\delta,\varepsilon,n}^{(1)} \cap \mathscr{E}_{\beta,L,\delta,n}^{(2)} \cap \mathscr{E}_{\beta,L,\delta,\varepsilon,n}^{(3)}$. The subsequent proof will be based on the following lemma:

Lemma D.7.1. *On the set $\mathscr{E}_{\beta,L,\delta,\varepsilon,n}$, if in addition we have*

$$\frac{1}{\sqrt{n}}\|\boldsymbol{p}^{k_*} - \tilde{\boldsymbol{p}}^{k_*}\|_2 \leq \frac{\mathrm{Lip}_0(\beta)}{\sqrt{n}}\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2 \leq 2\exp\left(-\frac{1}{4}C_{\mathrm{conv}}\beta^2\right) \tag{D.52}$$

holds for all $k_* \in \{k, k+1, k+2\}$ with $k_0(\beta) - 1 \leq k \leq K(\beta, T, \varepsilon) - 3$, then it also holds for $k_* = k+3$. Furthermore, the following inequality holds for all $k_0(\beta) - 1 \leq k \leq K(\beta, T, \varepsilon) - 3$:

$$\frac{1}{\sqrt{n}}\|\boldsymbol{p}^{k+3} - \tilde{\boldsymbol{p}}^{k+3}\|_2 \leq \frac{\rho(\beta)}{\sqrt{n}}\|\boldsymbol{p}^{k+2} - \tilde{\boldsymbol{p}}^{k+2}\|_2 + \frac{\rho(\beta)}{\sqrt{n}}\|\boldsymbol{p}^{k+1} - \tilde{\boldsymbol{p}}^{k+1}\|_2 + \frac{1}{\sqrt{n}}\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2, \tag{D.53}$$

where we recall that ρ is defined in Eq. (D.50).

Lemma D.7.1 and Eq. (D.51) imply the following upper bound via induction argument:

$$\frac{1}{\sqrt{n}}\|\boldsymbol{p}^{K(\beta,T,\varepsilon)} - \tilde{\boldsymbol{p}}^{K(\beta,T,\varepsilon)}\|_2 \leq \frac{1 + 2\mathrm{Lip}_0(\beta)}{1 - 2\rho(\beta)} \times \frac{1}{\sqrt{n}}\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2. \tag{D.54}$$

Define

$$\mathrm{Lip}_*(\beta) := \frac{M_\Theta \cdot (1 + 2\mathrm{Lip}_0(\beta))}{1 - 2\rho(\beta)}. \tag{D.55}$$

The claim of the lemma follows from Eq. (D.54) using the fact that $\Psi_{\alpha_t^{K(\beta,T,\varepsilon)}}$ has Lipschitz constant M_Θ.

The remainder of this section is dedicated to proving Lemma D.7.1.

Proof.[Proof of Lemma D.7.1] The condition of Lemma D.7.1 assumes that for all $k_* \in \{k, k+1, k+2\}$,

$$\frac{1}{\sqrt{n}}\|\boldsymbol{p}^{k_*} - \tilde{\boldsymbol{p}}^{k_*}\|_2 \le 2\exp\left(-\frac{1}{4}C_{\text{conv}}\beta^2\right). \tag{D.56}$$

Next, we will make use of the Jacobian matrices given in Eq. (5.45) to provide a preliminary upper bound for $\|\boldsymbol{p}^{k+3} - \tilde{\boldsymbol{p}}^{k+3}\|_2$.

On the set $\mathscr{E}_{\beta,L,\delta,\varepsilon,n}$, for all $\boldsymbol{m}, \boldsymbol{m'} \in [a_\Theta, b_\Theta]^n$ and $k_0(\beta) \le k \le K(\beta, T, \varepsilon) - 3$, it holds that

$$\begin{aligned}
&\|\beta \boldsymbol{D}(\boldsymbol{m})\boldsymbol{X}\boldsymbol{D}(\boldsymbol{m'})\|_{\text{op}} \le M_\Theta^2 \cdot (\beta^2 + 2\beta), \qquad \|b_t^k \boldsymbol{D}(\boldsymbol{m})\boldsymbol{D}(\boldsymbol{m'})\|_{\text{op}} \le M_\Theta^2 \cdot C_{\text{conv}}^{-1}\exp(-C_{\text{conv}}\beta^2),\\
&\|\boldsymbol{D}(\boldsymbol{m})\|_{\text{op}} \le M_\Theta,
\end{aligned}$$

where we used Eq. (5.47). Combining these upper bounds and Eq. (5.44), we obtain a crude upper bound for $\|\boldsymbol{p}^{k+3} - \tilde{\boldsymbol{p}}^{k+3}\|_2$:

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\|\tilde{\boldsymbol{p}}^{k+3} - \boldsymbol{p}^{k+3}\|_2\\
\le&\frac{M_\Theta^2 \cdot (\beta^2 + \beta + 1)}{\sqrt{n}}\|\tilde{\boldsymbol{p}}^{k+2} - \boldsymbol{p}^{k+2}\|_2 + \frac{M_\Theta^2 \cdot C_{\text{conv}}^{-1}\exp(-C_{\text{conv}}\beta^2)}{\sqrt{n}}\|\tilde{\boldsymbol{p}}^{k+1} - \boldsymbol{p}^{k+1}\|_2\\
&+ \frac{M_\Theta^2}{\sqrt{n}}\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2\\
\overset{(i)}{\le}&M_\Theta^2 \cdot (2\beta^2 + 2\beta + 4 + 2C_{\text{conv}}^{-1}\exp(-C_{\text{conv}}\beta^2))\exp\left(-\frac{1}{4}C_{\text{conv}}\beta^2\right),
\end{aligned}$$

where to obtain *(i)*, we use the following facts: (1) $\|\tilde{\boldsymbol{p}}^{k+i} - \boldsymbol{p}^{k+i}\|_2/\sqrt{n} \le 2e^{-C_{\text{conv}}\beta^2/4}$ for all $i \in \{1, 2\}$; (2) $\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2/\sqrt{n} \le 2r(\beta) \le 2e^{-C_{\text{conv}}\beta^2/4}$. Since $\boldsymbol{y}(t) \in B^n(\boldsymbol{y}(t), r(\beta))$, we can also control the difference between $\boldsymbol{p}^{k+3}$, $\tilde{\boldsymbol{p}}^{k+3}$ and $\hat{\boldsymbol{p}}^{k+3}(\boldsymbol{y}(t), t)$ following exactly the same manner, and produce exactly the same upper bound.

Before completing the proof, it is useful to establish the following lemma.

Lemma D.7.2. *For any π_Θ such that* $\mathrm{supp}(\pi_\Theta) \subseteq [-M_\Theta, M_\Theta]$ *and any $\gamma > 0$, the mapping*

$$Q(p) := \mathrm{Var}[\Theta \mid \gamma\Theta + \sqrt{\gamma}G = \Gamma_\gamma^{-1}(p)]$$

is $3M_\Theta^2$-Lipschitz continuous.

Proof.[Proof of Lemma D.7.2] Let $h = \Gamma_\gamma^{-1}(p)$. Taking the derivative of $Q(\cdot)$, we obtain via chain rule

$$\begin{aligned}
&\frac{\mathrm{d}Q}{\mathrm{d}p} = \frac{\mathrm{d}Q}{\mathrm{d}h}\cdot\frac{\mathrm{d}h}{\mathrm{d}p}\\
&= \big(\mathbb{E}[\Theta^2(\Theta - \mathbb{E}[\Theta \mid \gamma\Theta + \sqrt{\gamma}G = h]) \mid \gamma\Theta + \sqrt{\gamma}G = h] -
\end{aligned}$$

$$2\mathbb{E}[\Theta \mid \gamma\Theta + \sqrt{\gamma}G = h] \operatorname{Var}[\Theta \mid \gamma\Theta + \sqrt{\gamma}G = h]) \cdot \operatorname{Var}[\Theta \mid \gamma\Theta + \sqrt{\gamma}G = h]^{-1/2}.$$

Applying Cauchy–Schwarz inequality and the bounded support assumption,

$$\begin{aligned}
&\left|\mathbb{E}[\Theta^2(\Theta - \mathbb{E}[\Theta \mid \gamma\Theta + \sqrt{\gamma}G = h]) \mid \gamma\Theta + \sqrt{\gamma}G = h]\right| \leq M_\Theta^2 \cdot \operatorname{Var}[\Theta \mid \gamma\Theta + \sqrt{\gamma}G = h]^{1/2}, \\
&|\mathbb{E}[\Theta \mid \gamma\Theta + \sqrt{\gamma}G = h] \operatorname{Var}[\Theta \mid \gamma\Theta + \sqrt{\gamma}G = h]| \leq M_\Theta^2 \cdot \operatorname{Var}[\Theta \mid \gamma\Theta + \sqrt{\gamma}G = h]^{1/2}.
\end{aligned}$$

Putting together the above analysis, we conclude that $\|\frac{\mathrm{d}Q}{\mathrm{d}p}\|_\infty \leq 3M_\Theta^2$, thus completing the proof of the lemma.

□

By Eq. (5.50), on the set $\mathscr{E}_{\beta,L,\delta,\varepsilon,n}$, it holds that $\|\boldsymbol{D}_{\alpha_t^k}(\hat{\boldsymbol{m}}^k(\boldsymbol{y}(t),t))\|_F^2/n \leq C_{\mathrm{conv}}^{-1}\exp(-C_{\mathrm{conv}}\beta^2)$. Invoking this result, triangle inequality, Lemma D.7.2, and Cauchy-Schwartz inequality, we can conclude that for all $\boldsymbol{\zeta}^{k+3}$ that lies on the line segment connecting $\boldsymbol{p}^{k+3}$ and $\tilde{\boldsymbol{p}}^{k+3}$, it holds that

$$\begin{aligned}
&\frac{1}{n}\|\boldsymbol{D}_{\alpha_t^{k+3}}(\Psi_{\alpha_t^{k+3}}(\boldsymbol{\zeta}^{k+3}))\|_F^2 \\
\leq&\frac{1}{n}\|\boldsymbol{D}_{\alpha_t^{k+3}}(\Psi_{\alpha_t^{k+3}}(\hat{\boldsymbol{p}}^{k+3}(\boldsymbol{y}(t),t)))\|_F^2 + \frac{3M_\Theta^2}{n}\|\hat{\boldsymbol{p}}^{k+3}(\boldsymbol{y}(t),t) - \boldsymbol{\zeta}^{k+3}\|_1 \\
\leq&C_{\mathrm{conv}}^{-1}\exp(-C_{\mathrm{conv}}\beta^2) + M_\Theta^2 \cdot (2\beta^2 + 2\beta + 4 + 2C_{\mathrm{conv}}^{-1}\exp(-C_{\mathrm{conv}}\beta^2))\exp\left(-\frac{1}{4}C_{\mathrm{conv}}\beta^2\right) \\
=&\Delta(\beta)\,,
\end{aligned}$$

where we recall that $\Delta(\beta)$ is defined in Eq. (D.50). Similarly, we can derive that for all $\boldsymbol{\zeta}^{k+2}$ that is on the line segment connecting $\boldsymbol{p}^{k+2}$ and $\tilde{\boldsymbol{p}}^{k+2}$,

$$\frac{1}{n}\|\boldsymbol{D}_{\alpha_t^{k+2}}(\Psi_{\alpha_t^{k+2}}(\boldsymbol{\zeta}^{k+2}))\|_F^2 \leq \Delta(\beta).$$

In addition, note that for all $\boldsymbol{\zeta}^{k+2}, \boldsymbol{\zeta}^{k+3}$ as above, it holds that

$$\max\left\{\|\boldsymbol{D}_{\alpha_t^{k+2}}(\Psi_{\alpha_t^{k+2}}(\boldsymbol{\zeta}^{k+2}))\|_\infty, \|\boldsymbol{D}_{\alpha_t^{k+3}}(\Psi_{\alpha_t^{k+2}}(\boldsymbol{\zeta}^{k+2}))\|_\infty\right\} \leq M_\Theta.$$

Therefore, if we view the diagonal elements of matrices $\boldsymbol{D}_{\alpha_t^{k+2}}(\Psi_{\alpha_t^{k+2}}(\boldsymbol{\zeta}^{k+2}))$ and $\boldsymbol{D}_{\alpha_t^{k+3}}(\Psi_{\alpha_t^{k+3}}(\boldsymbol{\zeta}^{k+2}))$ as vectors, then they belong to the set $\{\boldsymbol{x} \in [0, M_\Theta]^n : \|\boldsymbol{x}\|_2^2/n \leq \Delta(\beta)\}$. Hence, recalling the definition of event $\mathscr{E}_{\beta,n}^{(4)}$ in Eq. Eq. (D.49), we see that for all $\boldsymbol{\zeta}^{k+2}$ and $\boldsymbol{\zeta}^{k+3}$, the following inequalities hold on $\mathscr{E}_{\beta,L,\delta,\varepsilon,n} \cap \mathscr{E}_{\beta,n}^{(4)}$:

$$\begin{aligned}
\|\beta\boldsymbol{D}(\boldsymbol{\zeta}^{k+3})\boldsymbol{X}\boldsymbol{D}(\boldsymbol{\zeta}^{k+2})\|_{\mathrm{op}} &\leq \frac{\beta^2}{n}\|\boldsymbol{D}(\boldsymbol{\zeta}^{k+3})\boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T}\boldsymbol{D}(\boldsymbol{\zeta}^{k+2})\|_{\mathrm{op}} + \beta\|\boldsymbol{D}(\boldsymbol{\zeta}^{k+3})\boldsymbol{W}\boldsymbol{D}(\boldsymbol{\zeta}^{k+2})\|_{\mathrm{op}} \\
&\leq \beta^2 M_\Theta^2\Delta(\beta) + 4\beta C' M_\Theta^{5/3}\Delta(\beta)^{1/6}, \\
\|b_t^k\boldsymbol{D}(\boldsymbol{\zeta}^{k+3})\boldsymbol{D}(\boldsymbol{\zeta}^{k+1})\|_{\mathrm{op}} &\leq M_\Theta^2 \cdot C_{\mathrm{conv}}^{-1}\exp(-C_{\mathrm{conv}}\beta^2).
\end{aligned}$$

Putting together the above inequalities and Eq. (5.45), we obtain that

$$\frac{1}{\sqrt{n}}\|\boldsymbol{p}^{k+3} - \tilde{\boldsymbol{p}}^{k+3}\|_2$$

$$
\begin{aligned}
&\le \frac{\beta^2 M_\Theta^2 \Delta(\beta) + 4\beta C' M_\Theta^{5/3} \Delta(\beta)^{1/6}}{\sqrt{n}} \|\boldsymbol{p}^{k+2} - \tilde{\boldsymbol{p}}^{k+2}\|_2 + \\
&\quad \frac{M_\Theta^2 \cdot C_{\text{conv}}^{-1} \exp(-C_{\text{conv}}\beta^2)}{\sqrt{n}} \|\boldsymbol{p}^{k+1} - \tilde{\boldsymbol{p}}^{k+1}\|_2 + \frac{M_\Theta}{\sqrt{n}} \|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2 \\
&\le \frac{\rho(\beta)}{\sqrt{n}} \|\boldsymbol{p}^{k+2} - \tilde{\boldsymbol{p}}^{k+2}\|_2 + \frac{M_\Theta^2 \cdot C_{\text{conv}}^{-1} \exp(-C_{\text{conv}}\beta^2)}{\sqrt{n}} \|\boldsymbol{p}^{k+1} - \tilde{\boldsymbol{p}}^{k+1}\|_2 + \frac{M_\Theta}{\sqrt{n}} \|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2 \qquad \text{(D.57)} \\
&\overset{(d)}{\le} \frac{\text{Lip}_0(\beta) \cdot \rho(\beta) + \text{Lip}_0(\beta) \cdot M_\Theta^2 \cdot C_{\text{conv}}^{-1} e^{-C_{\text{conv}}\beta^2} + M_\Theta}{\sqrt{n}} \cdot \|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2 \\
&\le \frac{\text{Lip}_0(\beta) \cdot F(\beta) + M_\Theta}{\sqrt{n}} \cdot \|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2 \\
&\overset{(e)}{\le} \frac{\text{Lip}_0(\beta)}{\sqrt{n}} \cdot \|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2,
\end{aligned}
$$

where in step *(d)* we used Eq. (D.52) with $k_* \in \{k+1, k+2\}$, and in step *(e)* we make use of the following facts: (1) $\text{Lip}_0(\beta) \ge 2M_\Theta$; (2) $F(\beta) \le 1/2$ for all $\beta \ge \beta_0$.

Recall that $2r(\beta) \cdot (\text{Lip}_0(\beta) + 1) \cdot (M_\Theta^2 + 1) \le 2e^{-C_{\text{conv}}\beta^2/4}$. Therefore, we can conclude that Eq. (D.52) holds for $k_* = k + 3$. In addition, for β_0 large enough clearly we have $M_\Theta^2 \cdot C_{\text{conv}}^{-1} \exp(-C_{\text{conv}}\beta^2) < \rho(\beta)$ holds for all $\beta \ge \beta_0$. As a result, we can deduce from Eq. (D.57) that Eq. (D.53) holds for all desired k, thus completing the proof of Lemma D.7.1.

□

D.7.1 Proof of Lemma D.4.2

Throughout this proof, we work with $\boldsymbol{X}, \boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2^0$ with distribution $\mathbb{P}_0$ defined in Section D.4.2. We will lighten notations by writing $\boldsymbol{\theta}_i := \boldsymbol{\theta}_i^0$.

We write the posterior distribution as

$$
\mu_t(\mathrm{d}\boldsymbol{\theta}) = \frac{1}{Z(t)} e^{H_t(\boldsymbol{\theta})} \, \pi_\Theta^{\otimes n}(\mathrm{d}\boldsymbol{\theta}) \, , \tag{D.58}
$$

$$
H_t(\boldsymbol{\theta}) := \frac{\beta}{2} \langle \boldsymbol{\theta}, \boldsymbol{X}\boldsymbol{\theta} \rangle - \frac{\beta^2}{4n} \|\boldsymbol{\theta}\|_2^4 + \langle \boldsymbol{y}(t), \boldsymbol{\theta} \rangle - \frac{t}{2} \|\boldsymbol{\theta}\|^2 \, . \tag{D.59}
$$

In this proof, we will never consider the joint distribution of these objects at two distinct values of t. Hence, we can carry out derivations with

$$
\boldsymbol{y}(t) = t\,\boldsymbol{\theta}_* + \sqrt{t}\,\boldsymbol{z} \, , \tag{D.60}
$$

for a fixed $\boldsymbol{z} \sim \mathsf{N}(0, \boldsymbol{I}_n)$. Further, we will write $\mu_t(F(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)) := \int F(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \, \mu_t^{\otimes 2}(\mathrm{d}\boldsymbol{\theta})$. Finally, we define

$$
\begin{aligned}
U_t(\boldsymbol{\theta}) &= \frac{2}{n} \frac{\partial}{\partial t} H_t(\boldsymbol{\theta}) & \text{(D.61)} \\
&= \frac{1}{n} \Big\{ 2\langle \boldsymbol{\theta}, \boldsymbol{\theta}_* \rangle + \frac{1}{\sqrt{t}} \langle \boldsymbol{z}, \boldsymbol{\theta} \rangle - \|\boldsymbol{\theta}\|_2^2 \Big\} \, . & \text{(D.62)}
\end{aligned}
$$

Using Gaussian integration by parts and Remark D.1.1, we have

$$
\mathbb{E}\left[\mu_t\big(U_t(\boldsymbol{\theta})\big)\right] = \frac{1}{n} \mathbb{E}\left[\mu_t\big(\langle \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \rangle\big)\right] \, , \tag{D.63}
$$

$$\mathbb{E}\left[\mu_t\Big(U_t(\boldsymbol{\theta}_1)\langle\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\rangle\Big)\right] = \frac{1}{n}\mathbb{E}\left[\mu_t\big(\langle\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\rangle^2\big)\right]. \tag{D.64}$$

Recall that $\mathrm{supp}(\pi_\Theta) \subseteq [-M_\Theta, M_\Theta]$, and therefore $|\langle\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\rangle/n| \le M_\Theta^2$, which further yields

$$\left|\mathbb{E}\Big[U_t(\boldsymbol{\theta}_1)\frac{1}{n}\langle\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\rangle\Big] - \mathbb{E}\Big[U_t(\boldsymbol{\theta})\Big]\mathbb{E}\Big[\frac{1}{n}\langle\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\rangle\Big]\right| \le M_\Theta^2 \cdot \mathbb{E}\left[\Big|U_t(\boldsymbol{\theta}) - \mathbb{E}\big[U_t(\boldsymbol{\theta})\big]\Big|\right]. \tag{D.65}$$

Combining Eqs. (D.63) to (D.65) gives

$$\frac{1}{n^2}\mathbb{E}\left[(\langle\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\rangle - \mathbb{E}[\langle\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\rangle])^2\right] \le M_\Theta^2 \cdot \mathbb{E}\left[\Big|U_t(\boldsymbol{\theta}) - \mathbb{E}\big[U_t(\boldsymbol{\theta})\big]\Big|\right]. \tag{D.66}$$

Next, we will prove that the right hand side of Eq. (D.66) is $o_n(1)$ for all $t > 0$, thus completing the proof of the lemma.

We define the free energy density:

$$\phi(t) := \frac{1}{n}\log\Big\{\int e^{H_t(\boldsymbol{\theta})}\,\pi_\Theta^{\otimes n}(\mathrm{d}\boldsymbol{\theta})\Big\}.$$

We can compute the first and second derivatives of ϕ:

$$\begin{aligned}
&\frac{\partial\phi}{\partial\sqrt{t}}(t) = \sqrt{t}\mu_t\big(U_t(\boldsymbol{\theta})\big), &\text{(D.67)}\\
&\frac{\partial^2\phi}{\partial(\sqrt{t})^2}(t) = nt\,\mathrm{Var}_{\mu_t}\big(U_t(\boldsymbol{\theta})\big) + \frac{1}{n}\mu_t\Big(2\langle\boldsymbol{\theta}_*,\boldsymbol{\theta}\rangle - \|\boldsymbol{\theta}\|_2^2\Big).
\end{aligned}$$

Therefore, defining $\psi(r) := \phi(r^2)$, we obtain that $r \mapsto \bar{\psi}(r) := \psi(r) + 3M_\Theta^2 r^2/2$ is convex for $r \in (0,\infty)$. Applying Lemma D.1.1 to the functions $\bar{\psi}(r)$ and $\mathbb{E}\bar{\psi}(r)$, for $0 < \varepsilon < r/2$ we have

$$\mathbb{E}\left[|\psi'(r) - \mathbb{E}[\psi'(r)]|\right] \le \mathbb{E}[\psi'(r+\varepsilon) - \psi'(r-\varepsilon)] + \frac{3}{\varepsilon}\sup_{|r'-r|\le\varepsilon}\mathbb{E}\left[|\psi(r') - \mathbb{E}[\psi(r')]|\right] + 6M_\Theta^2\varepsilon. \tag{D.68}$$

The next lemma proves that $\sup_{|r'-r|\le\varepsilon}\mathbb{E}\left[|\psi(r') - \mathbb{E}[\psi(r')]|\right]$ is small.

Lemma D.7.3. *There exists a constant $C(t,\beta,\pi_\Theta) > 0$ which is a function of (t,β,π_Θ) only, and is bounded compact intervals $[t_1,t_2] \subseteq (0,\infty)$ such that*

$$\mathbb{E}\left[|\phi(t) - \mathbb{E}[\phi(t)]|\right] \le C(t,\beta,\pi_\Theta)n^{-1/2}.$$

Proof. Letting $\boldsymbol{X} = \beta\boldsymbol{\theta}_*\boldsymbol{\theta}^\mathsf{T}/n + \boldsymbol{W}$, consider the mapping

$$f : (\boldsymbol{W}, \boldsymbol{z}) \mapsto \phi(t).$$

We denote by W_{ij} the (i,j)-th entry of $\boldsymbol{W}$. The following upper bounds on the partial derivatives are straightforward:

$$\begin{aligned}
&\left|\frac{\partial}{\partial\sqrt{n}W_{ij}}f(\boldsymbol{W},\boldsymbol{z})\right| \le \beta M_\Theta^2 n^{-3/2},\\
&\left|\frac{\partial}{\partial z_i}f(\boldsymbol{W},\boldsymbol{z})\right| \le 2rM_\Theta n^{-1}.
\end{aligned} \tag{D.69}$$

Hence by Gaussian concentration, we obtain

$$\mathbb{E}_{\boldsymbol{W},\boldsymbol{z}}\left[(\phi(t) - \mathbb{E}_{\boldsymbol{W},\boldsymbol{z}}[\phi(t)])\right] \le C_1 n^{-1}. \tag{D.70}$$

Here and below, we denote by C_i constants that depend only on β, M_Θ, r and and are bounded over compacts.

Finally, we show that $\mathbb{E}_{\boldsymbol{W},\boldsymbol{z}}[\phi(t)]$, as a function of $\boldsymbol{\theta}$, concentrates around its expectation. This follows from the estimate:

$$\left|\frac{\partial}{\partial\theta_i}\mathbb{E}_{\boldsymbol{W},\boldsymbol{z}}[\phi(t)]\right| \le C_2 n^{-1}.$$

Using Efron Stein's inequality, we get

$$\mathbb{E}_{\boldsymbol{\theta}}[(\mathbb{E}_{\boldsymbol{W},\boldsymbol{z}}[\phi(t)] - \mathbb{E}_{\boldsymbol{W},\boldsymbol{z},\boldsymbol{\theta}}[\phi(t)])^2] \le C_3 n^{-1}. \tag{D.71}$$

The proof of Lemma D.7.3 follows from Eq. (D.70) and Eq. (D.71).

□

We now conclude the proof of Lemma D.4.2. We have

$$\psi'(r) = \frac{\partial\phi}{\partial\sqrt{t}}(t)\bigg|_{t=r^2} = \sqrt{t}\mu_t\big(U_t(\boldsymbol{\theta})\big)\Big|_{t=r^2}. \tag{D.72}$$

Therefore letting $t_\pm := \sqrt{r\pm\varepsilon}$, Eq. (D.63) and Eq. (D.68) imply

$$\mathbb{E}\left[|\mu_t\big(U_t(\boldsymbol{\theta})\big) - \mathbb{E}[\mu_t\big(U_t(\boldsymbol{\theta})\big)]|\right] \le \frac{1}{n}\mathbb{E}[\mu_{t_+}(\langle\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\rangle) - \mu_{t_-}(\langle\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\rangle)] + \frac{C_4}{\varepsilon}n^{-1/2} + 6M_\Theta^2\varepsilon. \tag{D.73}$$

Proposition D.4.1 implies that the mapping $t \mapsto \gamma_*(\beta,t)$ is locally Lipschitz continuous on $(0,\infty)$. Since $\lim_{n\to\infty}\mathbb{E}\left[\mu_t(\langle\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\rangle/n)\right] = \gamma_*(\beta,t)$, this yields

$$\lim_{n\to\infty}\frac{1}{n}\mathbb{E}[\mu_{t_+}(\langle\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\rangle) - \mu_{t_-}(\langle\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\rangle)] \le \delta(\varepsilon)\,, \tag{D.74}$$

for some $\delta(\varepsilon)\downarrow 0$ as $\varepsilon\to 0$.

Since ε is arbitrary, we obtain

$$\lim_{n\to\infty}\mathbb{E}\left[|\mu_t\big(U_t(\boldsymbol{\theta})\big) - \mathbb{E}[\mu_t\big(U_t(\boldsymbol{\theta})\big)]|\right] = 0\,. \tag{D.75}$$

The proof is completed by showing that, for all $0 < t_1 < t_2$

$$\lim_{n\to\infty}\int_{t_1}^{t_2}\mathbb{E}\Big[\mathrm{Var}_{\mu_t}(U_t(\boldsymbol{\theta}))\Big]\,\mathrm{d}t = 0\,. \tag{D.76}$$

We notice that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mu_t(U_t(\boldsymbol{\theta})) = \frac{n}{2}\,\mathrm{Var}_{\mu_t}(U_t(\boldsymbol{\theta})) - \frac{1}{2nt^{3/2}}\mu_t(\langle\boldsymbol{z},\boldsymbol{\theta}\rangle)\,, \tag{D.77}$$

and therefore

$$\int_{t_1}^{t_2} \mathbb{E}\Big[\operatorname{Var}_{\mu_t}(U_t(\boldsymbol{\theta}))\Big]\,\mathrm{d}t = \frac{2}{n}\big\{\mathbb{E}\mu(U_{t_2}(\boldsymbol{\theta})) - \mathbb{E}\mu(U_{t_1}(\boldsymbol{\theta}))\big\} + \frac{1}{n^2}\int_{t_1}^{t_2} \frac{1}{t^{3/2}}\mathbb{E}\mu_t(\langle \boldsymbol{z}, \boldsymbol{\theta}\rangle)\,\mathrm{d}t\,. \tag{D.78}$$

On the other hand, using Eq. (D.63) and $\|\boldsymbol{\theta}\|_\infty \le M_\Theta$, we get

$$\Big|\mathbb{E}\mu_t\big(U_t(\boldsymbol{\theta})\big)\Big| \le M_\Theta^2\,, \tag{D.79}$$

$$\Big|\mathbb{E}\mu_t(\langle \boldsymbol{z}, \boldsymbol{\theta}\rangle)\Big| \le M_\Theta\sqrt{n}\mathbb{E}\|\boldsymbol{z}\|_2 \le 2M_\Theta n\,. \tag{D.80}$$

Substituting above, we get

$$\int_{t_1}^{t_2} \mathbb{E}\Big[\operatorname{Var}_{\mu_t}(U_t(\boldsymbol{\theta}))\Big]\,\mathrm{d}t \le \frac{4}{n}M_\Theta^2 + \frac{2}{n^2}M_\Theta\frac{t_2 - t_1}{t_1^{3/2}}\,, \tag{D.81}$$

which implies the claim (D.76).

Bibliography

[1] Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

[2] Emmanuel Abbe, Laurent Massoulie, Andrea Montanari, Allan Sly, and Nikhil Srivastava. Group synchronization on grids. *Mathematical Statistics and Learning*, 1(3):227–256, 2018.

[3] Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *International Conference on Computational Learning Theory*, pages 458–469. Springer, 2005.

[4] Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Sampling from the sherrington-kirkpatrick gibbs measure via algorithmic stochastic localization. *arXiv preprint arXiv:2203.05093*, 2022.

[5] Arash A Amini and Martin J Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. In *2008 IEEE International Symposium on Information Theory*, pages 2454–2458. IEEE, 2008.

[6] Nima Anari, Vishesh Jain, Frederic Koehler, Huy Tuan Pham, and Thuy-Duong Vuong. Entropic independence i: modified log-sobolev inequalities for fractionally log-concave distributions and high-temperature ising models. *arXiv e-prints*, pages arXiv–2106, 2021.

[7] Greg W. Anderson, Alice Guionnet, and Ofer Zeitouni. *An introduction to random matrices*. Cambridge University Press, 2009.

[8] Fabrizio Antenucci, Silvio Franz, Pierfrancesco Urbani, and Lenka Zdeborová. Glassy nature of the hard phase in inference problems. *Physical Review X*, 9(1):011020, 2019.

[9] Pranjal Awasthi, Afonso S Bandeira, Moses Charikar, Ravishankar Krishnaswamy, Soledad Villar, and Rachel Ward. Relax, no need to round: Integrality of clustering formulations. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pages 191–200, 2015.

[10] Pranjal Awasthi and Or Sheffet. Improved spectral-norm bounds for clustering. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 37–49. Springer, 2012.

[11] Zhi-Dong Bai and Yong-Qua Yin. Necessary and sufficient conditions for almost sure convergence of the largest eigenvalue of a wigner matrix. *The Annals of Probability*, pages 1729–1741, 1988.

[12] Zhi-Dong Bai and Yong-Qua Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. In *Advances In Statistics*, pages 108–127. World Scientific, 2008.

[13] Zhidong D Bai and Yong Q Yin. Convergence to the semicircle law. *The Annals of Probability*, pages 863–875, 1988.

[14] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

[15] Jinho Baik and Jack W Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of multivariate analysis*, 97(6):1382–1408, 2006.

[16] Sivaraman Balakrishnan, Martin J Wainwright, Bin Yu, et al. Statistical guarantees for the em algorithm: From population to sample-based analysis. *Annals of Statistics*, 45(1):77–120, 2017.

[17] Afonso S Bandeira, Ramon Van Handel, et al. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *Annals of Probability*, 44(4):2479–2506, 2016.

[18] Jess Banks, Cristopher Moore, Roman Vershynin, Nicolas Verzelen, and Jiaming Xu. Information-theoretic bounds and phase transitions in clustering, sparse pca, and submatrix localization. *IEEE Transactions on Information Theory*, 64(7):4872–4894, 2018.

[19] Jean Barbier, Mohamad Dia, Nicolas Macris, Florent Krzakala, Thibault Lesieur, and Lenka Zdeborová. Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula. In *Advances in Neural Information Processing Systems*, pages 424–432, 2016.

[20] Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

[21] Jean Barbier and Nicolas Macris. The adaptive interpolation method: a simple scheme to prove replica formulas in bayesian inference. *Probability theory and related fields*, 174(3):1133–1185, 2019.

[22] Roland Bauerschmidt and Thierry Bodineau. A very simple proof of the lsi for high temperature spin systems. *Journal of Functional Analysis*, 276(8):2582–2588, 2019.

[23] Mohsen Bayati, Marc Lelarge, and Andrea Montanari. Universality in polytope phase transitions and message passing algorithms. *The Annals of Applied Probability*, 25(2):753–822, 2015.

[24] Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011.

[25] Derek Bean, Peter J Bickel, Noureddine El Karoui, and Bin Yu. Optimal M-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences of the United States of America*, 110(36):14563–8, 9 2013.

[26] Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. *SIAM Journal on Computing*, 44(4):889–911, 2015.

[27] Gérard Ben Arous and Aukosh Jagannath. Spectral gap estimates in mean field spin glasses. *Communications in Mathematical Physics*, 361(1):1–52, 2018.

[28] Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low-rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.

[29] Florent Benaych-Georges and Raj Rao Nadakuditi. The singular values and vectors of low-rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135, 2012.

[30] Quentin Berthet and Philippe Rigollet. Computational lower bounds for sparse pca. *arXiv preprint arXiv:1304.0828*, 2013.

[31] Quentin Berthet and Philippe Rigollet. Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):1780–1815, 2013.

[32] Raphael Berthier, Andrea Montanari, and Phan-Minh Nguyen. State evolution for approximate message passing with non-separable functions. *Information and Inference*, 01 2019.

[33] Raphael Berthier, Andrea Montanari, and Phan-Minh Nguyen. State evolution for approximate message passing with non-separable functions. *Information and Inference: A Journal of the IMA*, 9(1):33–79, 2020.

[34] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[35] Patrick Billingsley. *Probability and Measure.* John Wiley & Sons, Inc., Hoboken, New Jersey, anniversar edition, 2012.

[36] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[37] Erwin Bolthausen. An iterative construction of solutions of the TAP equations for the Sherrington–Kirkpatrick model. *Communications in Mathematical Physics*, 325(1):333–366, 2014.

[38] Matthew Brennan, Guy Bresler, and Wasim Huleihel. Reducibility and computational lower bounds for problems with planted sparse structure. *arXiv:1806.07508*, 2018.

[39] S Charles Brubaker and Santosh S Vempala. Isotropic pca and affine-invariant clustering. In *Building Bridges*, pages 241–281. Springer, 2008.

[40] Sébastien Bubeck, Jian Ding, Ronen Eldan, and Miklós Z Rácz. Testing for high-dimensional geometry in random graphs. *Random Structures & Algorithms*, 49(3):503–532, 2016.

[41] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications.* Springer Science & Business Media, 2011.

[42] T Tony Cai, Xiaodong Li, and Zongming Ma. Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow. *The Annals of Statistics*, 44(5):2221–2251, 2016.

[43] T Tony Cai, Tengyuan Liang, and Alexander Rakhlin. Computational and statistical boundaries for submatrix localization in a large noisy matrix. *The Annals of Statistics*, 45(4):1403–1430, 2017.

[44] T Tony Cai, Jing Ma, Linjun Zhang, et al. Chime: Clustering of high-dimensional gaussian mixtures with em algorithm and its optimality. *The Annals of Statistics*, 47(3):1234–1267, 2019.

[45] Emmanuel Candés and Terence Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35:2313–2351, 2007.

[46] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[47] Emmanuel J Candès and Michael B Wakin. An introduction to compressive sampling. *IEEE signal processing magazine*, 25(2):21–30, 2008.

[48] Michael Celentano. Sudakov-fernique post-amp, and a new proof of the local convexity of the tap free energy. *arXiv:2208.09550*, 2022.

[49] Michael Celentano, Zhou Fan, and Song Mei. Local convexity of the tap free energy and amp convergence for z2-synchronization. *arXiv preprint arXiv:2106.11428*, 2021.

[50] Michael Celentano, Andrea Montanari, and Yuchen Wu. The estimation error of general first order methods. In *Conference on Learning Theory*, pages 1078–1141. PMLR, 2020.

[51] Sourav Chatterjee. A generalization of the lindeberg principle. *Ann. Probab.*, 34(6):2061–2076, 11 2006.

[52] Hong-Bin Chen and Jiaming Xia. Hamilton–jacobi equations for inference of matrix tensor products. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 58, pages 755–793. Institut Henri Poincaré, 2022.

[53] Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. *arXiv:2211.01916*, 2022.

[54] Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv:2209.11215*, 2022.

[55] Wei-Kuo Chen and Wai-Kit Lam. Universality of approximate message passing algorithms. *Electronic Journal of Probability*, 26(none):1 – 44, 2021.

[56] Yuansi Chen and Ronen Eldan. Localization schemes: A framework for proving mixing bounds for markov chains. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 110–122. IEEE, 2022.

[57] Yuxin Chen and Emmanuel Candes. Solving random quadratic systems of equations is nearly as easy as solving linear systems. In *Advances in Neural Information Processing Systems*, pages 739–747, 2015.

[58] Yuxin Chen and Emmanuel J Candès. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Communications on pure and applied mathematics*, 70(5):822–883, 2017.

[59] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

[60] Katalin Csilléry, Michael GB Blum, Oscar E Gaggiotti, and Olivier François. Approximate bayesian computation (abc) in practice. *Trends in ecology & evolution*, 25(7):410–418, 2010.

[61] Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 651–676, 2017.

[62] Sanjoy Dasgupta. Learning mixtures of gaussians. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*, pages 634–644. IEEE, 1999.

[63] Sanjoy Dasgupta and Leonard J Schulman. A probabilistic analysis of em for mixtures of separated, spherical gaussians. *Journal of Machine Learning Research*, 8:203–226, 2007.

[64] Yash Deshpande, Emmanuel Abbe, and Andrea Montanari. Asymptotic mutual information for the balanced binary stochastic block model. *Information and Inference: A Journal of the IMA*, 6(2):125–170, 2017.

[65] Yash Deshpande and Andrea Montanari. Information-theoretically optimal sparse pca. In *2014 IEEE International Symposium on Information Theory*, pages 2197–2201. IEEE, 2014.

[66] Yash Deshpande and Andrea Montanari. Sparse pca via covariance thresholding. *Advances in Neural Information Processing Systems*, 27, 2014.

[67] Yash Deshpande and Andrea Montanari. Sparse pca via covariance thresholding. *The Journal of Machine Learning Research*, 17(1):4913–4953, 2016.

[68] Mohamad Dia, Nicolas Macris, Florent Krzakala, Thibault Lesieur, Lenka Zdeborová, et al. Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula. *Advances in Neural Information Processing Systems*, 29, 2016.

[69] PL Dobrushin. The description of a random field by means of conditional probabilities and conditions of its regularity. *Theory of Probability & Its Applications*, 13(2):197–224, 1968.

[70] Tomas Dominguez and Jean-Christophe Mourrat. Mutual information for the sparse stochastic block model. *arXiv preprint arXiv:2209.04513*, 2022.

[71] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[72] David L Donoho and Michael J Feldman. Optimal eigenvalue shrinkage in the semicircle limit. *arXiv:2210.04488*, 2022.

[73] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[74] John C Duchi and Feng Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *Information and Inference: A Journal of the IMA*, 8(3):471–529, 2019.

[75] Alain Durmus and Eric Moulines. High-dimensional bayesian inference via the unadjusted langevin algorithm. 2019.

[76] Rick Durrett. *Probability: Theory and Examples.* Cambridge University Press, New York, NY, fourth edition, 2010.

[77] Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-hastings algorithms are fast! In *Conference on learning theory*, pages 793–797. PMLR, 2018.

[78] Ahmed El Alaoui and Florent Krzakala. Estimation in the spiked wigner model: a short proof of the replica formula. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 1874–1878. IEEE, 2018.

[79] Ronen Eldan. Thin shell implies spectral gap up to polylog via a stochastic localization scheme. *Geometric and Functional Analysis*, 23(2):532–569, 2013.

[80] Ronen Eldan. Taming correlations through entropy-efficient measure decompositions with applications to mean-field approximation. *Probability Theory and Related Fields*, 176(3):737–755, 2020.

[81] Ronen Eldan. Analysis of high-dimensional distributions using pathwise methods. *Proceedings of ICM, to appear*, 2022.

[82] Ronen Eldan, Frederic Koehler, and Ofer Zeitouni. A spectral condition for spectral gap: fast mixing in high-temperature ising models. *Probability Theory and Related Fields*, 182(3):1035–1051, 2022.

[83] Lawrence C. Evans and Ronald F. Gariepy. *Measure Theory and Fine Properties of Functions.* CRC Press, Taylor & Francis Group, Boca Raton, FL, revised edition, 2015.

[84] Zhou Fan, Song Mei, and Andrea Montanari. Tap free energy, spin glasses and variational inference. *The Annals of Probability*, 49(1):1–45, 2021.

[85] Albert Fannjiang and Thomas Strohmer. The numerics of phase retrieval. *Acta Numerica*, 29:125–228, 2020.

[86] Yingjie Fei and Yudong Chen. Hidden integrality of sdp relaxations for sub-gaussian mixture models. In *Conference On Learning Theory*, pages 1931–1965. PMLR, 2018.

[87] Jon Feldman, Rocco A Servedio, and Ryan O'Donnell. Pac learning axis-aligned mixtures of gaussians with no separation assumption. In *International Conference on Computational Learning Theory*, pages 20–34. Springer, 2006.

[88] Michael J Feldman. Spiked singular values and vectors under extreme aspect ratios. *arXiv:2104.15127*, 2021.

[89] Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh S Vempala, and Ying Xiao. Statistical algorithms and a lower bound for detecting planted cliques. *Journal of the ACM (JACM)*, 64(2):1–37, 2017.

[90] Vitaly Feldman, Cristobal Guzman, and Santosh Vempala. Statistical query algorithms for mean vector estimation and stochastic convex optimization. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1265–1277. SIAM, 2017.

[91] Andreas Galanis, Daniel Štefankovič, and Eric Vigoda. Inapproximability of the partition function for the antiferromagnetic ising and hard-core models. *Combinatorics, Probability and Computing*, 25(4):500–559, 2016.

[92] Dani Gamerman and Hedibert F Lopes. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference.* CRC press, 2006.

[93] Antoine Gerschenfeld and Andrea Montanari. Reconstruction for models on random graphs. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 194–204. IEEE, 2007.

[94] Behrooz Ghorbani, Hamid Javadi, and Andrea Montanari. An instability in variational inference for topic models. In *International Conference on Machine Learning (ICML)*, pages 2221–2231. PMLR, 2019.

[95] Walter R Gilks, Sylvia Richardson, and David Spiegelhalter. *Markov chain Monte Carlo in practice.* CRC press, 1995.

[96] Christophe Giraud and Nicolas Verzelen. Partial recovery bounds for clustering with the relaxed k-means. *Mathematical Statistics and Learning*, 1(3):317–374, 2019.

[97] Dongning Guo, Shlomo Shamai, and Sergio Verdú. Mutual information and minimum mean-square error in gaussian channels. *IEEE transactions on information theory*, 51(4):1261–1282, 2005.

[98] Bruce Hajek, Yihong Wu, and Jiaming Xu. Submatrix localization via message passing. *The Journal of Machine Learning Research*, 18(1):6817–6868, 2017.

[99] Moritz Hardt and Eric Price. Tight bounds for learning a mixture of two gaussians. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 753–760, 2015.

[100] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[101] Daniel Hsu and Sham M Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 11–20, 2013.

[102] Takayuki Iguchi, Dustin G Mixon, Jesse Peterson, and Soledad Villar. On the tightness of an sdp relaxation of k-means. *arXiv preprint arXiv:1505.04778*, 2015.

[103] Ernst Ising. *Beitrag zur theorie des ferro-und paramagnetismus.* PhD thesis, Grefe & Tiedemann, 1924.

[104] Adel Javanmard and Andrea Montanari. State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference: A Journal of the IMA*, 2(2):115–144, 2013.

[105] Adel Javanmard and Andrea Montanari. Debiasing the lasso: Optimal sample size for gaussian designs. *Ann. Statist.*, 46(6A):2593–2622, 12 2018.

[106] Chi Jin, Yuchen Zhang, Sivaraman Balakrishnan, Martin J Wainwright, and Michael I Jordan. Local maxima in the likelihood of gaussian mixture models: Structural results and algorithmic consequences. *Advances in neural information processing systems*, 29, 2016.

[107] Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, pages 295–327, 2001.

[108] Iain M Johnstone. High dimensional statistical inference and random matrices. *arXiv preprint math/0611589*, 2006.

[109] Iain M Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009.

[110] Iain M Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009.

[111] Iain M Johnstone and Alexei Onatski. Testing in high-dimensional spiked models. *The Annals of Statistics*, 48(3):1231–1254, 2020.

[112] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[113] Michel Journée, Yurii Nesterov, Peter Richtárik, and Rodolphe Sepulchre. Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11(2), 2010.

[114] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two gaussians. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 553–562, 2010.

[115] Noureddine El Karoui. On the largest eigenvalue of wishart matrices with identity covariance when n, p and p/n tend to infinity. *arXiv preprint math/0309355*, 2003.

[116] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11(Jul):2057–2078, 2010.

[117] Jason M Klusowski and WD Brinda. Statistical guarantees for estimating the centers of a two-component gaussian mixture by em. *arXiv preprint arXiv:1608.02280*, 2016.

[118] Frederic Koehler, Holden Lee, and Andrej Risteski. Sampling approximately low-rank ising models: Mcmc meets variational methods. *arXiv preprint arXiv:2202.08907*, 2022.

[119] Mladen Kolar, Sivaraman Balakrishnan, Alessandro Rinaldo, and Aarti Singh. Minimax localization of structural information in large noisy matrices. *Advances in Neural Information Processing Systems*, 24, 2011.

[120] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[121] Dirk P Kroese, Thomas Taimre, and Zdravko I Botev. *Handbook of Monte Carlo methods*. John Wiley & Sons, 2013.

[122] Amit Kumar and Ravindran Kannan. Clustering with spectral norm and the k-means algorithm. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 299–308. IEEE, 2010.

[123] Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. *arXiv:2206.06227*, 2022.

[124] E.L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses.* Springer Science+Business Media, Inc., New York, NY, third edition, 2005.

[125] Marc Lelarge and Léo Miolane. Fundamental limits of symmetric low-rank matrix estimation. *Probability Theory and Related Fields*, 173(3):859–929, 2019.

[126] Thibault Lesieur, Caterina De Bacco, Jess Banks, Florent Krzakala, Cris Moore, and Lenka Zdeborová. Phase transitions and optimal algorithms in high-dimensional gaussian mixture clustering. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 601–608. IEEE, 2016.

[127] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. Mmse of probabilistic low-rank matrix estimation: Universality with respect to the output channel. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 680–687. IEEE, 2015.

[128] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

[129] Xiaodong Li, Yang Li, Shuyang Ling, Thomas Strohmer, and Ke Wei. When do birds of a feather flock together? k-means, proximity, and conic programming. *Mathematical Programming*, 179(1):295–341, 2020.

[130] Xiaodong Li and Vladislav Voroninski. Sparse signal recovery from quadratic measurements via convex programming. *SIAM Journal on Mathematical Analysis*, 45(5):3019–3033, 2013.

[131] Robert Shevilevich Liptser and Al'bert Nikolaevich Shiriaev. *Statistics of random processes: General theory*, volume 394. Springer, 1977.

[132] Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *Advances in Neural Information Processing Systems*, pages 2726–2734, 2011.

[133] Yu Lu and Harrison H Zhou. Statistical and computational guarantees of lloyd's algorithm and its variants. *arXiv preprint arXiv:1612.02099*, 2016.

[134] Junjie Ma, Ji Xu, and Arian Maleki. Optimization-Based AMP for Phase Retrieval: The Impact of Initialization and ℓ_2 Regularization. *IEEE Transactions on Information Theory*, 65(6):3600–3629, 2019.

[135] Tengyu Ma and Avi Wigderson. Sum-of-squares lower bounds for sparse PCA. In *Advances in Neural Information Processing Systems*, pages 1612–1620, 2015.

[136] Zongming Ma. Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 41(2):772–801, 2013.

[137] Nicolas Macris, Cynthia Rush, et al. All-or-nothing statistical and computational phase transitions in sparse spiked matrix estimation. *Advances in Neural Information Processing Systems*, 33:14915–14926, 2020.

[138] Antoine Maillard, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase retrieval in high dimensions: Statistical and computational phase transitions. *arXiv:2006.05228*, 2020.

[139] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[140] Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond*, volume 9. World Scientific Publishing Company, 1987.

[141] Marc Mézard, Giorgio Parisi, and Riccardo Zecchina. Analytic and algorithmic solution of random satisfiability problems. *Science*, 297(5582):812–815, 2002.

[142] Léo Miolane. Fundamental limits of low-rank matrix estimation: the non-symmetric case. *arXiv preprint arXiv:1702.00473*, 2017.

[143] Dustin G Mixon, Soledad Villar, and Rachel Ward. Clustering subgaussian mixtures by semidefinite programming. *Information and Inference: A Journal of the IMA*, 6(4):389–415, 2017.

[144] Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 93–102. IEEE, 2010.

[145] Marco Mondelli and Andrea Montanari. Fundamental limits of weak recovery with applications to phase retrieval. In *Conference On Learning Theory*, pages 1445–1450. PMLR, 2018.

[146] Marco Mondelli and Andrea Montanari. Fundamental limits of weak recovery with applications to phase retrieval. *Found Comput Math*, 19:703–773, 06 2019.

[147] Marco Mondelli and Ramji Venkataramanan. Approximate message passing with spectral initialization for generalized linear models. In *International Conference on Artificial Intelligence and Statistics*, pages 397–405. PMLR, 2021.

[148] Marco Mondelli and Ramji Venkataramanan. Pca initialization for approximate message passing in rotationally invariant models. *Advances in Neural Information Processing Systems*, 34:29616–29629, 2021.

[149] Andrea Montanari. Optimization of the Sherrington-Kirkpatrick Hamiltonian. In *IEEE Symposium on the Foundations of Computer Science, FOCS*, November 2019.

[150] Andrea Montanari, Feng Ruan, and Jun Yan. Adapting to unknown noise distribution in matrix denoising. *arXiv:1810.02954*, 2018.

[151] Andrea Montanari and Subhabrata Sen. A short tutorial on mean-field spin glass techniques for non-physicists. *arXiv:2204.02909*, 2022.

[152] Andrea Montanari and Ramji Venkataramanan. Estimation of low-rank matrices via approximate message passing. *The Annals of Statistics*, 49(1):321–345, 2021.

[153] Andrea Montanari and Alexander S Wein. Equivalence of approximate message passing and low-degree polynomials in rank-one matrix estimation. *arXiv:2212.06996*, 2022.

[154] Andrea Montanari and Yuchen Wu. Fundamental limits of low-rank matrix estimation with diverging aspect ratios. *arXiv preprint arXiv:2211.00488*, 2022.

[155] Andrea Montanari and Yuchen Wu. Statistically optimal first order algorithms: A proof via orthogonalization. *arXiv preprint arXiv:2201.05101*, 2022.

[156] Andrea Montanari and Yuchen Wu. Posterior sampling from the spiked models via diffusion processes. *arXiv preprint arXiv:2304.11449*, 2023.

[157] Elchanan Mossel and Eric Vigoda. Limitations of markov chain monte carlo algorithms for bayesian inference of phylogeny. 2006.

[158] Elchanan Mossel and Jiaming Xu. Local algorithms for block models with side information. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 71–80, 2016.

[159] Mohamed Ndaoud. Sharp optimal recovery in the two-component gaussian mixture model. *arXiv preprint arXiv:1812.08078*, 2018.

[160] Mohamed Ndaoud. Sharp optimal recovery in the two component gaussian mixture model. *The Annals of Statistics*, 50(4):2096–2126, 2022.

[161] Arkadii Semenovich Nemirovsky and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

[162] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer, 2003.

[163] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, et al. Genes mirror geography within europe. *Nature*, 456(7218):98–101, 2008.

[164] John Novembre and Matthew Stephens. Interpreting principal component analyses of spatial population genetic variation. *Nature genetics*, 40(5):646–649, 2008.

[165] Dmitry Panchenko. *The Sherrington-Kirkpatrick model.* Springer Science & Business Media, 2013.

[166] Neal Parikh and Stephen Boyd. Proximal Algorithms. *Foundations and Trends in Optimization*, 1(3):123–231, 2013.

[167] Debashis Paul. Asymptotic distribution of the smallest eigenvalue of wishart (N, n) when $N, n \to \infty$ such that $N/n \to 0$. In *Nonparametric Statistical Methods and Related Topics: A Festschrift in Honor of Professor PK Bhattacharya on the Occasion of His 80th Birthday*, pages 423–458. World Scientific, 2012.

[168] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

[169] Jiming Peng and Yu Wei. Approximating k-means-type clustering via semidefinite programming. *SIAM journal on optimization*, 18(1):186–205, 2007.

[170] Amelia Perry, Alexander S Wein, and Afonso S Bandeira. Statistical limits of spiked tensor models. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 56, pages 230–264. Institut Henri Poincaré, 2020.

[171] Amelia Perry, Alexander S Wein, Afonso S Bandeira, and Ankur Moitra. Optimality and sub-optimality of pca for spiked random matrices and synchronization. *arXiv preprint arXiv:1609.05573*, 2016.

[172] Jiaze Qiu and Subhabrata Sen. The tap free energy for high-dimensional linear regression. *arXiv:2203.07539*, 2022.

[173] Oded Regev and Aravindan Vijayaraghavan. On learning mixtures of well-separated gaussians. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 85–96. IEEE, 2017.

[174] Martin Royer. Adaptive clustering through semidefinite programming. *Advances in Neural Information Processing Systems*, 30, 2017.

[175] Philip Schniter and Sundeep Rangan. Compressive phase retrieval via generalized approximate message passing. *IEEE Transactions on Signal Processing*, 63(4):1043–1055, 2014.

[176] Amit Singer. Angular synchronization by eigenvectors and semidefinite programming. *Applied and computational harmonic analysis*, 30(1):20–36, 2011.

[177] Allan Sly. Reconstruction for the potts model. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 581–590, 2009.

[178] Allan Sly and Nike Sun. The computational hardness of counting in two-spin models on d-regular graphs. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pages 361–369. IEEE, 2012.

[179] Mahdi Soltanolkotabi. Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization. *IEEE Transactions on Information Theory*, 65(4):2374–2400, 2019.

[180] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[181] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

[182] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[183] Aart J Stam. Some inequalities satisfied by the quantities of information of fisher and shannon. *Information and Control*, 2(2):101–112, 1959.

[184] Charles M. Stein. Estimation of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics*, 9(6):1135–1151, 11 1981.

[185] Andrew M Stuart. Inverse problems: a bayesian perspective. *Acta numerica*, 19:451–559, 2010.

[186] David J Thouless, Philip W Anderson, and Robert G Palmer. Solution of 'solvable model of a spin glass'. *Philosophical Magazine*, 35(3):593–601, 1977.

[187] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[188] Aad W. van der Vaart. *Asymptotic Statistics.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[189] Ramon Van Handel. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2014.

[190] Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.

[191] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok, editors, *Compressed Sensing, Theory and Applications*, volume 23, chapter 5, pages 210–268. Cambridge University Press, 2012.

[192] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[193] Cèdric Villani. *Optimal Transport, old and new.* Springer-Verlag Berlin Heidelberg, New York, NY, 2010.

[194] Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

[195] Irene Waldspurger. Phase retrieval with random gaussian sensing vectors by alternating projections. *IEEE Transactions on Information Theory*, 64(5):3301–3312, 2018.

[196] Gang Wang, Georgios B Giannakis, and Yonina C Eldar. Solving systems of random quadratic equations via truncated amplitude flow. *IEEE Transactions on Information Theory*, 64(2):773–794, 2017.

[197] Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

[198] Dawn B Woodard and Jeffrey S Rosenthal. Convergence rate of markov chain methods for genomic motif discovery. 2013.

[199] Yun Yang, Martin J Wainwright, and Michael I Jordan. On the computational complexity of high-dimensional bayesian variable selection. 2016.

[200] Xinyi Zhong, Chang Su, and Zhou Fan. Empirical bayes pca in high dimensions. *arXiv preprint arXiv:2012.11676*, 2020.

[201] Xinyi Zhong, Chang Su, and Zhou Fan. Empirical bayes pca in high dimensions. *Journal of the Royal Statistical Society Series B*, pages 853–878, 2022.