Proceedings of Machine Learning Research vol 125:1–64, 2020 33rd Annual Conference on Learning Theory

The estimation error of general first order methods

Michael Celentano MCELEN@STANFORD.EDU
Department of Statistics, Stanford University

Andrea Montanari MONTANARI@STANFORD.EDU
Department of Statistics and Department of Electrical Engineering, Stanford University

Yuchen Wu WUYC14@STANFORD.EDU
Department of Statistics, Stanford University

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

Modern large-scale statistical models require the estimation of thousands to millions of parameters. This is often accomplished by iterative algorithms such as gradient descent, projected gradient descent or their accelerated versions. What are the fundamental limits of these approaches? This question is well understood from an optimization viewpoint when the underlying objective is convex. Work in this area characterizes the gap to global optimality as a function of the number of iterations. However, these results have only indirect implications on the gap to *statistical* optimality.

Here we consider two families of high-dimensional estimation problems: high-dimensional regression and low-rank matrix estimation, and introduce a class of 'general first order methods' that aim at efficiently estimating the underlying parameters. This class of algorithms is broad enough to include classical first order optimization (for convex and non-convex objectives), but also other types of algorithms. Under a random design assumption, we derive lower bounds on the estimation error that hold in the high-dimensional asymptotics in which both the number of observations and the number of parameters diverge. These lower bounds are optimal in the sense that there exist algorithms in this class whose estimation error matches the lower bounds up to asymptotically negligible terms. We illustrate our general results through applications to sparse phase retrieval and sparse principal component analysis.

1. Introduction

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 $\boldsymbol{\theta}^1,\dots,\boldsymbol{\theta}^t$. 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 the 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., Candés

and Tao (2007); Bickel et al. (2009). The seminal work of Nemirovsky and Yudin (1983) 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 after t iterations. 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 (Keshavan et al., 2010; Loh and Wainwright, 2011; Chen and Candés, 2015). 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 paper 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.

Our results are novel even in the case of a convex cost function $\mathcal{L}(\boldsymbol{\vartheta})$, for two reasons. First, classical theory (Nesterov, 2018) lower bounds the objective value $\mathcal{L}(\boldsymbol{\theta}^t) - \min_{\boldsymbol{\vartheta}} \mathcal{L}(\boldsymbol{\vartheta})$ after t iterations. This has only indirect implications on estimation error, e.g., $\|\boldsymbol{\theta}^t - \boldsymbol{\theta}\|_2$ where $\boldsymbol{\theta}$ is the true value of the parameters (not the minimizer of the cost $\mathcal{L}(\boldsymbol{\vartheta})$). Second, the classical lower bounds on the objective value are worst case with respect to the function $\mathcal{L}(\boldsymbol{\vartheta})$ and do not take into account the data distribution.

Concretely, we consider two families of estimation problems:

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 $\boldsymbol{X}$.

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 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.

1. In particular, maximum likelihood need not be statistically optimal in high dimension (Bean et al., 2013).

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 $\boldsymbol{X}^\top$. 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.

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 $t \geq 1$. 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 $(\boldsymbol{v}^t)_{t\geq 1}$, $\boldsymbol{v}^t \in \mathbb{R}^{p\times r}$,

$$\begin{aligned} \boldsymbol{v}^{t+1} &= \boldsymbol{X}^\top F_t^{(1)}(\boldsymbol{u}^1, \dots, \boldsymbol{u}^t; \boldsymbol{y}, \boldsymbol{u}) + F_t^{(2)}(\boldsymbol{v}^1, \dots, \boldsymbol{v}^t; \boldsymbol{v}), \\ \boldsymbol{u}^t &= \boldsymbol{X} G_t^{(1)}(\boldsymbol{v}^1, \dots, \boldsymbol{v}^t; \boldsymbol{v}) + G_t^{(2)}(\boldsymbol{u}^1, \dots, \boldsymbol{u}^{t-1}; \boldsymbol{y}, \boldsymbol{u}), \end{aligned} \tag{1}$$

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 $(\boldsymbol{u}_i^s)^\mathsf{T}$ 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 $\boldsymbol{v}^1 = \boldsymbol{X}^\mathsf{T} F_0^{(1)}(\boldsymbol{y}, \boldsymbol{u}) + F_0^{(2)}(\boldsymbol{v})$. The unknown matrices (or vectors) $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ are estimated after t_* iterations by $\hat{\boldsymbol{\theta}}^{t_*} = G_*(\boldsymbol{v}^1, \cdots, \boldsymbol{v}^{t_*}; \boldsymbol{v})$ 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.

The paper is organized as follows: Section 2 illustrates the setting introduced above in two examples; Section 3 contains the statement of our general lower bounds; Section 4 applies these lower bounds to the two examples; Section 5 presents an outline of the proof, deferring technical details to appendices.

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

$$\mathcal{L}_n(\boldsymbol{\vartheta}) := \sum_{i=1}^n \ell(y_i; \langle \boldsymbol{x}_i, \boldsymbol{\vartheta}\rangle) + \Omega_n(\boldsymbol{\vartheta}) = \hat{\ell}_n(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{\vartheta}) + \Omega_n(\boldsymbol{\vartheta}). \tag{2}$$

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 (Parikh and Boyd, 2013):

$$\begin{aligned} \boldsymbol{\theta}^{t+1} &= \mathsf{Prox}_{\gamma_t \Omega_1}\big(\boldsymbol{\theta}^t - \gamma_t \nabla_{\boldsymbol{\vartheta}} \hat{\ell}_n(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{\theta}^t)\big), \\ \mathsf{Prox}_{\gamma\Omega_1}(y) &:= \arg\min_{\theta\in\mathbb{R}} \left\{ \frac{1}{2}(y-\theta)^2 + \gamma\Omega_1(\theta) \right\}. \end{aligned} \tag{3}$$

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

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

Therefore proximal gradient for the cost function (2) is an example of a GFOM: see Appendix H for the explicit change of variables. 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 focus on the sparse phase retrieval problem. We want to reconstruct a sparse signal $\boldsymbol{\theta}$ from noisy measurements of the modulus $|\langle \boldsymbol{\theta}, \boldsymbol{x}_i\rangle|$; that is, we lose the 'phase' of these projections.[2] Concretely, we assume $\|\boldsymbol{\theta}\|_0 \le s_0$. Information theoretically, accurate reconstruction is possible if $n \ge C s_0 \log(p/s_0)$, with C a sufficiently large constant (Eldar and Mendelson, 2012). Several groups have investigated practical reconstruction algorithms including semidefinite programming relaxations (Li and Voroninski, 2013) or first order methods (Schniter and Rangan, 2014; Candés et al., 2015; Chen and Candés, 2015; Cai et al., 2016; Sanghavi et al., 2017; Chen et al., 2019; Ma et al., 2020). A standard approach is to apply proximal gradient to the cost function (2) with $\Omega_n(\boldsymbol{\vartheta}) = \lambda\|\boldsymbol{\vartheta}\|_1$. However, all existing global convergence guarantees for these methods require $n \ge C s_0^2 \log p$. Soltanolkotabi (2019) proved that a first order method can accurately reconstruct the signal with $n \ge C s_0 \log(p/s_0)$ if it is initialized close enough to $\boldsymbol{\theta}$, but does not provide gaurantees for weaker initializations. Is the dependence on s_0^2 due to a fundamental computational barrier or an artifact of the theoretical analysis?

Example #2: Sparse PCA

A simple model for sparse principal component analysis (PCA) involves taking $r = 1$, $(\lambda_i)_{i\le n} \overset{\text{iid}}{\sim} \mathsf{N}(0,1)$, and $\boldsymbol{\theta} \in \mathbb{R}^p$ is a sparse vector with $\|\boldsymbol{\theta}\|_0 \le s_0 \ll p$ in the low-rank matrix estimation model above. Given data $\boldsymbol{X}$, we would like to reconstruct the signal $\boldsymbol{\theta}$. Information-theoretically, accurate reconstruction of $\boldsymbol{\theta}$ is possible if $n \ge C s_0 \log(p/s_0)$, with C a sufficiently large constant (Amini and Wainwright, 2008). A number of polynomial-time algorithms have been studied, ranging from simple thresholding algorithms (Johnstone and Lu, 2009; Deshpande and Montanari, 2016) to sophisticated convex relaxations (Amini and Wainwright, 2008; Ma and Wigderson, 2015). One natural idea is to modify the power iteration algorithm of standard PCA by computing

$$\boldsymbol{\theta}^{t+1} = c_t\, \boldsymbol{X}^{\mathsf{T}} \boldsymbol{X} \eta(\boldsymbol{\theta}^t; \gamma_t). \tag{4}$$

Here $(c_t)_{t\ge 0}$ is a deterministic normalization, and $\eta(\,\cdot\,;\gamma)$ is a thresholding function at level γ, e.g., soft thresholding $\eta(x;\gamma) = \mathrm{sign}(x)(|x| - \gamma)_+$. This algorithm is a GFOM: see Appendix H for the explicit change of variables. More elaborate versions of non-linear power iteration were developed, for example, by Journée et al. (2010); Ma et al. (2013), and are typically equivalent to suitable GFOMs.

2. We consider the real-valued case, but the generalization to the complex case should be immediate.

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 Berthet et al. (2013); Brennan et al. (2018) via reduction from the planted clique problem. Our analysis provides new evidence towards the same conclusion.

3. Main results

In this section we state formally our general results about high-dimensional regression and low-rank matrix estimation. 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 $t \geq 1$.
- A2. The covariates matrix $\boldsymbol{X}$ (for high-dimensional regression) or the noise matrix $\boldsymbol{Z}$ (for low-rank estimation) have entries $x_{ij} \overset{\text{iid}}{\sim} \mathsf{N}(0, 1/n)$, $z_{ij} \overset{\text{iid}}{\sim} \mathsf{N}(0, 1/n)$.

We denote by $\mathscr{P}_q(\mathbb{R}^k)$ and $\mathscr{P}_{\mathrm{c}}(\mathbb{R}^k)$ the set of probability distributions with finite q-th moment and compact support on $\mathbb{R}^k$, respectively. A function $f : \mathbb{R}^k \to \mathbb{R}$ is *pseudo-Lipschitz of order 2* if there exists C such that $|f(\boldsymbol{x}) - f(\boldsymbol{x}')| \leq C(1 + \|\boldsymbol{x}\| + \|\boldsymbol{x}'\|)\|\boldsymbol{x} - \boldsymbol{x}'\|$ for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^k$. A non-negative function $\ell : (\mathbb{R}^k)^2 \to \mathbb{R}$ is a *quadratically-bounded loss* if it is pseudo-Lipschitz of order 2 and there exists C such that for all $\boldsymbol{x}, \boldsymbol{x}', \boldsymbol{d} \in \mathbb{R}^k$, $|\ell(\boldsymbol{x}, \boldsymbol{d}) - \ell(\boldsymbol{x}', \boldsymbol{d})| \leq C(1 + \sqrt{\ell(\boldsymbol{x}, \boldsymbol{d})} + \sqrt{\ell(\boldsymbol{x}', \boldsymbol{d})})\|\boldsymbol{x} - \boldsymbol{x}'\|$. Whenever we take $n, p \to \infty$, we do so in such a way that $n/p \to \delta$.

3.1. High-dimensional regression

We make the following additional assumptions:

- 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.

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. Recall the variables u_i, v_i model side information available to the statistician which may, for example, take the form of an informative initialization. In many cases of interest, like phase retrieval, existing guarantees for first order methods require an informative initialization (Candés et al., 2015; Cai et al., 2016). Including side information in our analysis allows us to study its importance in achieving good estimation performance.

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

$$\begin{aligned} \tilde{\tau}_s^2 = \frac{1}{\delta}\,\mathsf{mmse}_{\Theta,V}(\tau_s^2), \qquad \sigma_s^2 = \frac{1}{\delta}(\tau_\Theta^2 - \mathsf{mmse}_{\Theta,V}(\tau_s^2)), \\ \frac{1}{\tau_{s+1}^2} = \frac{1}{\tilde{\tau}_s^2}\mathbb{E}\left[\mathbb{E}[G_1|Y, G_0, U]^2\right], \end{aligned} \tag{5}$$

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 1 *Assume A1, A2, R1, and R2. Let $\hat{\boldsymbol{\theta}}^t$ be output of any GFOM after t iterations. Then*

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

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

$$\lim_{n,p\to\infty} \frac{1}{p}\sum_{j=1}^{p} \ell(\theta_j, \hat{\theta}_j^t) \geq \inf_{\hat{\theta}(\,\cdot\,)} \mathbb{E}\big\{\ell(\Theta, \hat{\theta}(\Theta + \tau_t G, V))\big\}, \tag{6}$$

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. For all $\epsilon > 0$, there exist GFOMs with limiting risk within ϵ of the right-hand side of (6).

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

$$\boldsymbol{Q}_{t+1} = \boldsymbol{V}_{\boldsymbol{\Lambda},\boldsymbol{U}}(\hat{\boldsymbol{Q}}_t), \qquad \hat{\boldsymbol{Q}}_t = \frac{1}{\delta}\boldsymbol{V}_{\boldsymbol{\Theta},\boldsymbol{V}}(\boldsymbol{Q}_t), \tag{7}$$

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. (7) is a simple scalar recursion.

Theorem 2 *Assume A1, A2, and M1. Let $\hat{\boldsymbol{\theta}}^t$ be output of any GFOM after t iterations. 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 $\ell : \mathbb{R}^{2r} \to \mathbb{R}_{\geq 0}$,

$$\lim_{n\to\infty} \frac{1}{p}\sum_{j=1}^{p} \ell(\boldsymbol{\theta}_j, \hat{\boldsymbol{\theta}}_j^t) \geq \inf_{\hat{\boldsymbol{\theta}}(\,\cdot\,)} \mathbb{E}\big\{\ell(\boldsymbol{\Theta}, \hat{\boldsymbol{\theta}}(\boldsymbol{Q}_t^{1/2}\boldsymbol{\Theta} + \boldsymbol{G}, \boldsymbol{V}))\big\}, \tag{8}$$

where $(\boldsymbol{\Theta}, \boldsymbol{V}) \sim \mu_{\boldsymbol{\Theta},\boldsymbol{V}}$ independent of $\boldsymbol{G} \sim \mathsf{N}(0, \boldsymbol{I}_r)$, and the infimum on the right-hand side is over measurable functions $\hat{\boldsymbol{\theta}} : \mathbb{R}^r \to \mathbb{R}^r$. The limits are in probability and to a constant. As above, for all $\epsilon > 0$ there exist GFOMs with limiting risk within ϵ of the right-hand side of (8).

3.3. Discussion

Our motivations are similar to the ones for statistical query (SQ) lower bounds (Feldman et al., 2017a,b): 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 methods of Section 2 cannot be framed as SQ algorithms.

The lower bounds of Theorems 1 and 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 (Koller and Friedman, 2009), or an iterative implementation of the TAP equations from spin glass theory (Mézard et al., 1987).

Our proof uses the asymptotically exact analysis of AMP algorithms developed in Bolthausen (2014); Bayati and Montanari (2011); Javanmard and Montanari (2018); Berthier et al. (2019). However we need to overcome three technical obstacles: (1) Show that any GFOM can be reduced (in a suitable sense) to a certain AMP algorithm, 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 (Sly, 2009; Mossel and Xu, 2016), the current setting is technically more challenging because the underlying variables θ_i are continuous.

While the line of argument above is –in hindsight– very natural, the conclusion is broadly useful. For instance, Antenucci et al. (2019) study a class of of message passing algorithms inspired by replica symmetry breaking and survey propagation (Mézard et al., 2002), and observe that they do not perform better than Bayes AMP. These algorithms are within the scope of our Theorem 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 (Lelarge and Miolane, 2019; Barbier et al., 2019). It was conjectured that, in this context, no polynomial-time algorithm with access to an arbitrarily small amount of side informationcan can outperform Bayes AMP.[3] Theorems 1 and 2 establish this result within the restricted class of GFOMs.

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.

3. 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:

$$\delta_{\text{sp}} := \left(\int_{\mathbb{R}} \frac{\mathbb{E}_G[p(y|G)(G^2-1)]^2}{\mathbb{E}_G[p(y|G)]}\,\mathrm{d}y\right)^{-1}. \tag{9}$$

Here expectation is with respect to $G \sim \mathsf{N}(0,1)$. It was proved in Mondelli and Montanari (2019) that, if $\|\boldsymbol{\theta}\|_2 = \sqrt{p}$ and $n > (\delta_{\text{sp}} + \eta)p$, for some η bounded away from zero, then there exists a simple spectral estimator $\hat{\boldsymbol{\theta}}_{\text{sp}}$ that achieves weak recovery, i.e., a positive correlation with the true signal. Namely, $\frac{|\langle\hat{\boldsymbol{\theta}}_{\text{sp}},\boldsymbol{\theta}\rangle|}{\|\hat{\boldsymbol{\theta}}_{\text{sp}}\|_2\|\boldsymbol{\theta}\|_2}$ 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 δ_{sp} is information-theoretically optimal: for $n < (\delta_{\text{sp}} - \eta)p$ no estimator achieves a correlation that is bounded away from 0 (Mondelli and Montanari, 2019). 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 (Li and Voroninski, 2013; Eldar and Mendelson, 2012).

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 we are given

$$\overline{\boldsymbol{v}} = \sqrt{\alpha}\,\boldsymbol{\theta}/\|\boldsymbol{\theta}\|_2 + \sqrt{1-\alpha}\tilde{\boldsymbol{g}}, \qquad (\tilde{g}_i)_{i\le p} \overset{\text{iid}}{\sim} \mathsf{N}(0,1/p). \tag{10}$$

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 *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 $p(\,\cdot\,|x)$ satisfies the conditions of Theorem 1 and is twice differentiable with respect to x.*

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

$$\sup_{t\ge 0}\lim_{n,p\to\infty}\inf_{\boldsymbol{\theta}\in\mathscr{T}(\varepsilon,\sqrt{p})}\mathbb{E}\frac{|\langle\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^t\rangle|}{\|\boldsymbol{\theta}\|_2\|\hat{\boldsymbol{\theta}}^t\|_2} \le C_*\sqrt{\alpha}. \tag{11}$$

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 we have access to an initialization $\overline{\boldsymbol{v}}$ correlated with $\overline{\boldsymbol{\theta}}$, as per Eq. (10). In order to apply Theorem 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 2.

Lemma 4 *Assume the sparse PCA model with the distribution of $\overline{\boldsymbol{\theta}}$ given above. Define $(q_t)_{t\ge 0}$ by*

$$q_{t+1} = \frac{V_\pm(q_t + \tilde{\alpha})}{1 + V_\pm(q_t + \tilde{\alpha})}, \qquad q_0 = 0\,, \tag{12}$$

$$V_\pm(q) := e^{-\delta q\mu^2}\mu^2\varepsilon^2\mathbb{E}\left\{\frac{\sinh(\mu\sqrt{\delta q}G)^2}{1-\varepsilon+\varepsilon e^{-\delta q\mu^2/2}\cosh(\mu\sqrt{\delta q}G)}\right\}, \tag{13}$$

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

$$\lim_{n,p\to\infty}\frac{\langle\overline{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}^t\rangle}{\|\overline{\boldsymbol{\theta}}\|_2\|\hat{\boldsymbol{\theta}}^t\|_2} \le \sqrt{\frac{V_\pm(q_t+\tilde{\alpha})}{\mu^2\varepsilon}}. \tag{14}$$

The bound in Lemma 4, which holds for random vectors $\overline{\boldsymbol{\theta}}$ with i.i.d. entries from the three-points distribution, implies a minimax bound for non-random vectors $\overline{\boldsymbol{\theta}}$ with given ℓ_2-norm and sparsity. given in the corollary below.

Corollary 5 *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

$$\sup_{t\ge 0}\lim_{n,p\to\infty}\inf_{\overline{\boldsymbol{\theta}}\in\mathscr{T}(\varepsilon,R)}\mathbb{E}\frac{\langle\overline{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}^t\rangle}{\|\overline{\boldsymbol{\theta}}\|_2\|\hat{\boldsymbol{\theta}}^t\|_2} \le C_*\sqrt{\alpha}. \tag{15}$$

(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 if simple PCA fails, then every GFOM fails. Vice versa, if simple PCA succeeds, 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.

5. Proof of main results

In this section, we prove Theorems 1 and 2 under stronger assumptions than in their statements. These assumptions amount to stronger regularity requirements on the data generating distributions. Because they do not clarify the conceptual structure of our argument, we defer the precise statements of these assumptions to Appendix A. We label these assumptions R3 and R4 in the high-dimensional regression model and assumption M2 in the low-rank matrix estimation model. In Appendix E, we show that Theorem 1 (resp. Theorem 2) under assumptions R3 and R4 (resp. M2) implies the theorem under the weaker assumptions R1 and R2 (resp. M1).

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* (Bayati and Montanari, 2011). 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

$$\begin{aligned} \boldsymbol{a}^{t+1} &= \boldsymbol{X}^{\mathsf{T}} f_t(\boldsymbol{b}^1, \dots, \boldsymbol{b}^t; \boldsymbol{y}, \boldsymbol{u}) + \text{Onsager correction}, \\ \boldsymbol{b}^t &= \boldsymbol{X} g_t(\boldsymbol{a}^1, \dots, \boldsymbol{a}^t; \boldsymbol{v}) + \text{Onsager correction}. \end{aligned} \tag{16}$$

with initialization $\boldsymbol{a}^1 = \boldsymbol{X}^{\mathsf{T}} f_0(\boldsymbol{y}, \boldsymbol{u})$. Here the "Onsager correction" is a term determined in a specific way by the functions $(f_t)_{t\geq 0}$, $(g_t)_{t\geq 1}$ and the properties of the model in which AMP is applied. We specify the Onsager correction explicitly in Appendix B. Importantly, state evolution characterizes the iterates $\boldsymbol{a}^t$, $\boldsymbol{b}^t$ –and the iteration is referred to as *AMP*– only if the Onsager correction is chosen in the specific way detailed there. The next lemma, proved in Appendix B, describes the state evolution of the resulting AMP algorithm. Its characterization of AMP via state evolution uses existing results (Javanmard and Montanari, 2013). Its contribution is in showing that, in addition, all GFOMs are equivalent under a change a variables to an appropriately chosen AMP algorithm.

Lemma 6 *Under assumptions A1, A2, R3, R4 (for high-dimensional regression) or assumptions A1, A2, M2 (for low-rank matrix estimation), for any GFOM 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. Define $\{\boldsymbol{a}^s, \boldsymbol{b}^s\}_{s\geq 0}$ via the AMP algorithm (16). Then we have*

$$\begin{aligned} \boldsymbol{v}^t &= \varphi_t(\boldsymbol{a}^1, \dots, \boldsymbol{a}^t; \boldsymbol{v}), \quad t \geq 1, \\ \boldsymbol{u}^t &= \phi_t(\boldsymbol{b}^1, \dots, \boldsymbol{b}^t; \boldsymbol{y}, \boldsymbol{u}), \quad t \geq 1, \end{aligned}$$

where $\boldsymbol{v}^t, \boldsymbol{u}^t$ are as in Eq. (1). Further, state evolution determines two collections 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,

$$\frac{1}{p}\sum_{j=1}^{p} \psi(\boldsymbol{a}_j^1, \dots, \boldsymbol{a}_j^t, \boldsymbol{v}_j, \boldsymbol{\theta}_j) \xrightarrow{\text{p}} \mathbb{E}[\psi(\boldsymbol{\alpha}_1\boldsymbol{\Theta} + \boldsymbol{Z}^1, \dots, \boldsymbol{\alpha}_t\boldsymbol{\Theta} + \boldsymbol{Z}^t, \boldsymbol{V}, \boldsymbol{\Theta})], \tag{17}$$

where $(\boldsymbol{\Theta}, \boldsymbol{V}) \sim \mu_{\boldsymbol{\Theta},\boldsymbol{V}}$ independent of $(\boldsymbol{Z}^1, \dots, \boldsymbol{Z}^t) \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{T}_{[1:t]})$. Here $\boldsymbol{T}_{[1:t]} \in \mathbb{R}^{tr\times tr}$ is a positive semi-definite block matrix with block (s, s') given by $\boldsymbol{T}_{s,s'}$.[4]

Lemma 6 implies that the estimator $\hat{\boldsymbol{\theta}}^t$ in Theorem 1 and 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 6 implies that the limits in Theorems 1 and 2 exist and have lower bound

$$\inf R_\ell(g_*, (\boldsymbol{\alpha}_s), (\boldsymbol{T}_{s,s'})) := \inf \mathbb{E}[\ell(\boldsymbol{\Theta}, g_*(\boldsymbol{\alpha}_1\boldsymbol{\Theta} + \boldsymbol{Z}^1, \dots, \boldsymbol{\alpha}_t\boldsymbol{\Theta} + \boldsymbol{Z}^t, \boldsymbol{V}))], \tag{18}$$

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.

4. We emphasize that the construction of all relevant functions and matrices depend on the model. We describe these constructions and prove Lemma 6 in Appendix B.

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 (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*. See Figure 1 in Appendix C for a diagram of 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 $(\boldsymbol{u}_f)_{f\in\mathcal{F}}$.

When ambiguity will result, we will refer to the models of Section 3 as high-dimensional regression and low-rank matrix estimation *on the graph*.[5] 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

$$\begin{aligned} \boldsymbol{a}^{t+1}_{v\to f} &= \sum_{f'\in\partial v\setminus f} x_{f'v}\boldsymbol{r}^t_{f'\to v}, \qquad \boldsymbol{r}^t_{f\to v} = f_t(\boldsymbol{b}^1_{f\to v}, \dots, \boldsymbol{b}^t_{f\to v}; y_f, \boldsymbol{u}_f), \\ \boldsymbol{b}^t_{f\to v} &= \sum_{v'\in\partial f\setminus v} x_{fv'}\boldsymbol{q}^t_{v'\to f}, \qquad \boldsymbol{q}^t_{v\to f} = g_t(\boldsymbol{a}^1_{v\to f}, \dots, \boldsymbol{a}^t_{v\to f}; \boldsymbol{v}_v), \end{aligned} \tag{19}$$

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

$$\boldsymbol{a}^{t+1}_v = \sum_{f\in\partial v} x_{fv}\boldsymbol{r}^t_{f\to v}, \qquad \boldsymbol{b}^t_f = \sum_{v\in\partial f} x_{fv}\boldsymbol{q}^t_{v\to f}. \tag{20}$$

5. This terminology is motivated by viewing the models of Section 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.

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

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 7 *In the high-dimensional regression and low-rank matrix estimation models on the computation 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 fixed t and node v, the message passing algorithm (19) generates beliefs at v satisfying in the proportional asymptotics*

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

where $(\boldsymbol{\Theta}, \boldsymbol{V}) \sim \mu_{\boldsymbol{\Theta},\boldsymbol{V}}$ independent of $(\boldsymbol{Z}^1, \dots, \boldsymbol{Z}^t) \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{T}_{[1:t]})$, and $\overset{\mathrm{W}}{\to}$ denotes convergence in the Wasserstein metric of order 2 (see Appendix A). Moreover, the matrices $(\boldsymbol{T}_{s,t})_{s,t\geq 1}, (\boldsymbol{\alpha}_t)_{t\geq 1}$ agree with those in Lemma 6 when the functions $(f_t)_{t\geq 0}$, $(g_t)_{t\geq 1}$ also agree.

We prove Lemma 7 in Appendix C. Lemma 7 and the properties of convergence in the Wasserstein metric of order 2 (see Lemma 10, Appendix A) imply that for any message passing estimator $\hat{\boldsymbol{\theta}}_v^t$ and loss ℓ, the risk $\mathbb{E}[\ell(\boldsymbol{\theta}_v, \hat{\boldsymbol{\theta}}_v^t)] = \mathbb{E}[\ell(\boldsymbol{\theta}_v, g_*(\boldsymbol{a}_v^1, \dots, \boldsymbol{a}_v^t; \boldsymbol{v}_v)]$ 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 $\hat{\boldsymbol{\theta}}_v^t$ 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 $\hat{\boldsymbol{\theta}}_v^t$ by the optimal risk of any measurable 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 8 *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 in the proportional asymptotics*

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

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. (6) (resp. Eq. (8)).

We prove Lemma 8 in Appendix D. Combining Lemma 8 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 (6) and (8) now follow. Moreover, as we show in Appendix F, the bounds (6) and (8) are achieved by a certain AMP algorithm. The proof is complete.

Acknowledgements

MC was supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE 1656518. AM was partially supported by NSF grants CCF-1714305, IIS-1741162 and by the ONR grant N00014-18-1-2729.

References

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.

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.

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.

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.

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, 2 2011.

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.

Quentin Berthet, Philippe Rigollet, et al. Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):1780–1815, 2013.

Raphael Berthier, Andrea Montanari, and Phan-Minh Nguyen. State evolution for approximate message passing with non-separable functions. *Information and Inference*, 1 2019.

Peter J Bickel, Yaacov Ritov, Alexandre B Tsybakov, et al. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Patrick Billingsley. *Probability and Measure*. John Wiley & Sons, Inc., Hoboken, New Jersey, anniversary edition, 2012.

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.

Matthew Brennan, Guy Bresler, and Wasim Huleihel. Reducibility and computational lower bounds for problems with planted sparse structure. In *Proceedings of the 31st Conference On Learning Theory*, volume 75, pages 48–166, 7 2018.

T Tony Cai, Xiaodong Li, Zongming Ma, et al. Optimal rates of convergence for noisy sparse phase retrieval via thresholded wirtinger flow. *The Annals of Statistics*, 44(5):2221–2251, 2016.

Emmanuel Candés and Terence Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35:2313–2351, 2007.

Emmanuel J Candés, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

Sourav Chatterjee. A generalization of the lindeberg principle. *Ann. Probab.*, 34(6):2061–2076, 11 2006.

Yuxin Chen and Emmanuel Candés. 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.

Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Gradient descent with random initialization: fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176(1):5–37, 7 2019.

Yash Deshpande and Andrea Montanari. Sparse pca via covariance thresholding. *The Journal of Machine Learning Research*, 17(1):4913–4953, 2016.

Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, New York, NY, fourth edition, 2010.

Yonina Eldar and Shahar Mendelson. Phase retrieval: Stability and recovery guarantees. *Applied and Computational Harmonic Analysis*, 36, 11 2012.

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.

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, 2017a.

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, 2017b.

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.

Adel Javanmard and Andrea Montanari. Debiasing the lasso: Optimal sample size for gaussian designs. *Ann. Statist.*, 46(6A):2593–2622, 12 2018.

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.

Michel Journée, Yurii Nesterov, Peter Richtárik, and Rodolphe Sepulchre. Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11: 517–553, 2 2010.

Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11:2057–2078, 7 2010.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

E.L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer Science+Business Media, Inc., New York, NY, third edition, 2005.

Marc Lelarge and Léo Miolane. Fundamental limits of symmetric low-rank matrix estimation. *Probability Theory and Related Fields*, 173(3-4):859–929, 2019.

Xiaodong Li and Vladislav Voroninski. Sparse signal recovery from quadratic measurements via convex programming. *SIAM Journal on Mathematical Analysis*, 45(5):3019–3033, 2013.

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.

Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations of Computational Mathematics*, 20(3):451–632, 8 2020.

Tengyu Ma and Avi Wigderson. Sum-of-squares lower bounds for sparse pca. In *Advances in Neural Information Processing Systems*, pages 1612–1620, 2015.

Zongming Ma et al. Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 41(2):772–801, 2013.

Marc Mézard, Giorgio Parisi, and Miguel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

Marc Mézard, Giorgio Parisi, and Riccardo Zecchina. Analytic and algorithmic solution of random satisfiability problems. *Science*, 297(5582):812–815, 2002.

Marco Mondelli and Andrea Montanari. Fundamental limits of weak recovery with applications to phase retrieval. *Found Comput Math*, 19:703–773, 06 2019.

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.

Arkadi S. Nemirovsky and D. B. Yudin. Problem complexity and method efficiency in optimization; john wile. 1983.

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

Neal Parikh and Stephen Boyd. Proximal Algorithms. *Foundations and Trends in Optimization*, 1 (3):123–231, 2013.

Sujay Sanghavi, Rachel Ward, and Chris D. White. The local convexity of solving systems of quadratic equations. *Results in Mathematics*, 71(3):569–608, 06 2017.

Philip Schniter and Sundeep Rangan. Compressive phase retrieval via generalized approximate message passing. *IEEE Transactions on Signal Processing*, 63(4):1043–1055, 2014.

Allan Sly. Reconstruction for the potts model. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 581–590, 2009.

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.

Charles M. Stein. Estimation of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics*, 9(6):1135–1151, 11 1981.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

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.

Cèdric Villani. *Optimal Transport, old and new*. Springer-Verlag Berlin Heidelberg, New York, NY, 2010.

Appendix A. Strengthened assumptions, technical definitions, and lemmas

A.1. Strengthened assumptions

As mentioned in the main body, we will first prove Theorems 1 and 2, and in particular, Lemmas 6, 7, and 8, 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 $\{(\theta_i, v_i)\}_{i \leq p} \overset{\rm iid}{\sim} \mu_{\Theta,V}$, $\{(\boldsymbol{w}_i, u_i)\}_{i \leq n} \overset{\rm iid}{\sim} \mu_{\boldsymbol{W},U}$.

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 $\{(\boldsymbol{\lambda}_i, \boldsymbol{u}_i)\}_{i \leq n} \overset{\rm iid}{\sim} \mu_{\boldsymbol{\Lambda},\boldsymbol{U}}$, $\{(\boldsymbol{\theta}_j, \boldsymbol{v}_j)\}_{j \leq p} \overset{\rm iid}{\sim} \mu_{\boldsymbol{\Theta},\boldsymbol{V}}$.

We relax these assumptions in Appendix E. In Appendix E, we show that Theorem 1 (resp. Theorem 2) under assumptions R3 and R4 (resp. M2) implies the theorem under the weaker assumptions R1 and R2 (resp. M1).

A.2. 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, $\Pi \in \mathscr{P}_2(\mathbb{R}^k \times \mathbb{R}^k)$ with first and second marginals μ and μ', respectively (where a marginal here involves a block of k coordinates). It is well known that $W_2(\mu, \mu')$ is a metric on $\mathscr{P}_2(\mathbb{R}^k)$ (Villani, 2010, pg. 94). When a sequence of probability distributions μ_n converges to μ in the Wasserstein metric of order 2, we write $\mu_n \overset{\rm W}{\to} \mu$. We also write $\boldsymbol{A}_n \overset{\rm W}{\to} \boldsymbol{A}$ when $\boldsymbol{A}_n \sim \mu_n$, $\boldsymbol{A} \sim \mu$ for such a sequence.

Second, we generalize the definition of *pseudo-Lipschitz of order 2*. A function $f : \mathbb{R}^r \to \mathbb{R}$ is *pseudo-Lipschitz of order k* if there exists C such that $|f(\boldsymbol{x}) - f(\boldsymbol{x}')| \leq C(1 + \|\boldsymbol{x}\|^{k-1} + \|\boldsymbol{x}'\|^{k-1})\|\boldsymbol{x} - \boldsymbol{x}'\|$ for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^k$. To be pseudo-Lipshitz of order 1 is equivalent to being Lipschitz.

Finally, we provide several technical lemmas which we will need in our proofs. Some standard lemmas are stated without proof.

Lemma 9 *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 10 *If a sequence of random vectors $\boldsymbol{X}_n \overset{\rm W}{\to} \boldsymbol{X}$, 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 11 *Consider a sequence of random variables $(A_n, \boldsymbol{B}_n) \xrightarrow{\mathrm{d}} (A, \boldsymbol{B})$ with values in $\mathbb{R} \times \mathbb{R}^k$ such that $(A_n, \boldsymbol{B}_n) \xrightarrow{\mathrm{d}} (A, \boldsymbol{B})$ 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 $\phi(A_n, \boldsymbol{B}_n) \xrightarrow{\mathrm{d}} \phi(A, \boldsymbol{B})$.

Proof [Lemma 11] 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 closed rectangle whose boundary has probability 0 under $\boldsymbol{B}$ that

$$\mu_{A_n, \boldsymbol{B}_n}(S \times I) \to \mu_{A, \boldsymbol{B}}(S \times I). \tag{21}$$

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 separated from K by distance ϵ. Similarly define $\phi_I^\epsilon : \mathbb{R}^k \to \mathbb{R}$. Then

$$\begin{aligned}\mathbb{E}[\phi_K^\epsilon(A_n)\phi_I^\epsilon(\boldsymbol{B}_n)] &\geq \mu_{A_n, \boldsymbol{B}_n}(K \times I) \\ &\geq \mathbb{E}[\phi_K^\epsilon(A_n)\phi_I^\epsilon(\boldsymbol{B}_n)] - \mu_A(\mathsf{spt}(\phi_K^\epsilon) \setminus K) - \mu_{\boldsymbol{B}_n}(\mathsf{spt}(\phi_I^\epsilon) \setminus I),\end{aligned}$$

where we have used that $A_n \overset{\mathrm{d}}{=} A$. Because the boundary of I has measure 0 under $\mu_{\boldsymbol{B}}$, we have $\lim_{\epsilon \to 0} \limsup_{n \to \infty} \mu_{\boldsymbol{B}_n}(\mathsf{spt}(\phi_I^\epsilon) \setminus I) = 0$. Moreover, $\lim_{\epsilon \to 0} \mu_A(\mathsf{spt}(\phi_K^\epsilon) \setminus K) = 0$. 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 $\epsilon \to 0$ 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_G^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 Billingsley (2012), 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

$$\begin{aligned}\mu_{A, \boldsymbol{B}}(S \times I) - \epsilon \leq \mu_{A, \boldsymbol{B}}(K \times I) = \lim_{n \to \infty} \mu_{A_n, \boldsymbol{B}_n}(K \times I) \leq \liminf_{n \to \infty} \mu_{A_n, \boldsymbol{B}_n}(S \times I) \\ \leq \limsup_{n \to \infty} \mu_{A_n, \boldsymbol{B}_n}(S \times I) \leq \lim_{n \to \infty} \mu_{A_n, \boldsymbol{B}_n}(G \times I) = \mu_{A, \boldsymbol{B}}(G \times I) \leq \mu_{A, \boldsymbol{B}}(S \times I) + \epsilon.\end{aligned}$$

Taking $\epsilon \to 0$, we conclude (21).

We now show (21) 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 it is 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 $f_+(a, \boldsymbol{b}) := \sum_\iota \mathbf{1}\{\boldsymbol{b} \in I_\iota\} \sup_{\boldsymbol{b}' \in I_\iota} f(a, \boldsymbol{b}')$, where $\{I_\iota\}_\iota$ is the partition. Note that on $\{a \mid \delta(a, \epsilon) < \delta^*\} \times [-M, M]^k$, 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 $\{\delta(a, \epsilon) < \delta^*\} \times [-M, M]^k$

$$\begin{aligned} \mathbb{E}[f_-(A_n, \boldsymbol{B}_n)] - 2\epsilon < \mathbb{E}[f(A_n, \boldsymbol{B}_n)] < \mathbb{E}[f_+(A_n, \boldsymbol{B}_n)] + 2\epsilon, \\ \mathbb{E}[f_-(A, \boldsymbol{B})] - 2\epsilon < \mathbb{E}[f(A, \boldsymbol{B})] < \mathbb{E}[f_+(A, \boldsymbol{B})] + 2\epsilon. \end{aligned} \tag{22}$$

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 (21), 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 (22).

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 (5) defining the lower bound in the high-dimensional regression model.

Lemma 12 *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 $Y|G_0, G_1, U$ has 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 (5).

Proof [Lemma 12] 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 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 ℓ is quadratically-bounded if it is non-negative, pseudo-Lipschitz of order 2, and also satisfies

$$|\ell(\boldsymbol{\vartheta}, \boldsymbol{d}) - \ell(\boldsymbol{\vartheta}', \boldsymbol{d})| \leq C \left(1 + \sqrt{\ell(\boldsymbol{\vartheta}, \boldsymbol{d})} + \sqrt{\ell(\boldsymbol{\vartheta}', \boldsymbol{d})}\right) \|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}'\|. \tag{23}$$

Consider $(\boldsymbol{\Theta}, \boldsymbol{V}) \sim \mu_{\boldsymbol{\Theta},\boldsymbol{V}} \in \mathscr{P}_2(\mathbb{R}^k \times \mathbb{R}^k)$, $\boldsymbol{Z} \sim \mathsf{N}(0, \boldsymbol{I}_k)$ 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

$$R(\boldsymbol{y}, \tau, \boldsymbol{v}, K, M) := \inf_{\|\boldsymbol{d}\|_\infty \leq M} \int \frac{1}{Z} \ell(\boldsymbol{\vartheta}, \boldsymbol{d}) e^{-\frac{1}{2\tau^2}\|\boldsymbol{y}-\boldsymbol{\vartheta}\|^2} \mu_{\boldsymbol{\Theta}^{(K)}|\boldsymbol{V}}(\boldsymbol{v}, \mathrm{d}\boldsymbol{\vartheta}), \tag{24}$$

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 13 *The following properties hold for the Bayes risk with respect to quadratically bounded losses ℓ.*

(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,

$$\inf_{\hat{\boldsymbol{\theta}}(\cdot)} \mathbb{E}[\ell(\boldsymbol{\Theta}^{(K)}, \hat{\boldsymbol{\theta}}(\boldsymbol{\Theta}^{(K)} + \tau \boldsymbol{z}, \boldsymbol{V})] = \mathbb{E}[R(\boldsymbol{Y}^{(K)}, \tau, \boldsymbol{V}, K, M)], \tag{25}$$

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,

$$\mathbb{E}[R(\boldsymbol{\Theta}^{(K)} + \tau \boldsymbol{Z}, \tau, \boldsymbol{V}, K, \infty)] = \lim_{M \to \infty} \mathbb{E}[R(\boldsymbol{\Theta}^{(K)} + \tau \boldsymbol{Z}, \tau, \boldsymbol{V}, K, M)]. \tag{26}$$

(b) For a fixed $K < \infty$, the posterior Bayes risk is bounded: $R(\boldsymbol{y}, \tau, \boldsymbol{v}, K, M) \leq \bar{R}(K)$ 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 the truncation level K and noise variance τ. This is true also at $K = \infty$:

$$\mathbb{E}[R(\boldsymbol{\Theta}^{(K)} + \tau \boldsymbol{Z}, \tau, \boldsymbol{V}, K, \infty)] = \lim_{\substack{K \to \infty \\ \tau' \to \tau}} \mathbb{E}[R(\boldsymbol{\Theta}^{(K)} + \tau' \boldsymbol{Z}, \tau', \boldsymbol{V}, K, \infty)], \tag{27}$$

where the limit holds for any way of taking K, τ' to their limits (ie., sequentially or simultaneously).

Proof [Lemma 13(a)] For any measurable $\hat{\boldsymbol{\theta}} : \mathbb{R}^k \times \mathbb{R}^k \to [-M, M]^k$,

$$\begin{aligned}\mathbb{E}[\ell(\boldsymbol{\Theta}^{(K)}, \hat{\boldsymbol{\theta}}(\boldsymbol{\Theta}^{(K)} + \tau\boldsymbol{Z}, \boldsymbol{V}))] &= \mathbb{E}[\mathbb{E}[\ell(\boldsymbol{\Theta}^{(K)}, \hat{\boldsymbol{\theta}}(\boldsymbol{\Theta}^{(K)} + \tau\boldsymbol{Z}, \boldsymbol{V}))|\boldsymbol{\Theta}^{(K)} + \tau\boldsymbol{Z}, \boldsymbol{V}]] \\ &\geq \mathbb{E}[R(\boldsymbol{\Theta}^{(K)} + \tau\boldsymbol{Z}, \tau, \boldsymbol{V}, K, M)]. \end{aligned} \tag{28}$$

For $M < \infty$, equality obtains. Indeed, we may define

$$\hat{\boldsymbol{\theta}}^{(M)}(\boldsymbol{y}, \boldsymbol{v}; \tau) = \arg\min_{\|\boldsymbol{d}\|_\infty \leq M} \int \frac{1}{Z}\ell(\boldsymbol{\vartheta}, \boldsymbol{d}) e^{-\frac{1}{2\tau^2}\|\boldsymbol{y}-\boldsymbol{\vartheta}\|^2} \mu_{\boldsymbol{\Theta}^{(K)}|\boldsymbol{V}}(\boldsymbol{v}, \mathrm{d}\boldsymbol{\vartheta}), \tag{29}$$

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, by Eq. (28), $\mathbb{E}[R(\boldsymbol{\Theta}^{(K)} + \tau\boldsymbol{Z}, \tau, \boldsymbol{V}, K, \infty)]$ is a lower bound on the Bayes risk at $M = \infty$ and we may achieve risk arbitrarily close to this lower bound by taking $M \to \infty$ in (29), we conclude (25) at $M = \infty$ as well. ■

Proof [Lemma 13(b)] The quantity $R(\boldsymbol{y}, \tau, \boldsymbol{v}, K, M)$ is non-negative. Define

$$\bar{R}(K) = \max_{\|\boldsymbol{\vartheta}\|_\infty \leq K} \ell(\boldsymbol{\vartheta}, \boldsymbol{0}).$$

Observe that $R(\boldsymbol{y}, \tau, \boldsymbol{v}, K, M) \leq \bar{R}(K)$ 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

$$\begin{aligned}&\left\|\nabla_{\boldsymbol{y}} \int \ell(\boldsymbol{\vartheta}, \boldsymbol{d}) p^*(\boldsymbol{\vartheta}|\boldsymbol{y}, \tau, \boldsymbol{v}, K) \mu_{\boldsymbol{\Theta}^{(K)}|\boldsymbol{V}}(\boldsymbol{v}, \mathrm{d}\boldsymbol{\vartheta})\right\| \\ &\qquad \leq \int \ell(\boldsymbol{\vartheta}, \boldsymbol{d}) p^*(\boldsymbol{\vartheta}|\boldsymbol{y}, \tau, v) \left\|\nabla_{\boldsymbol{y}} \log p^*(\boldsymbol{\vartheta}|\boldsymbol{y}, \tau, \boldsymbol{v})\right\| \mu_{\boldsymbol{\Theta}^{(K)}|\boldsymbol{V}}(\boldsymbol{v}, \mathrm{d}\boldsymbol{\vartheta}) \\ &\qquad \leq \frac{2K\sqrt{k}}{\tau^2} \int \ell(\boldsymbol{\vartheta}, \boldsymbol{d}) p^*(\boldsymbol{\vartheta}|\boldsymbol{y}, \tau, \boldsymbol{v}) \mu_{\boldsymbol{\Theta}^{(K)}|\boldsymbol{V}}(\boldsymbol{v}, \mathrm{d}\boldsymbol{\vartheta}),\end{aligned}$$

where we have used that $\|\nabla_{\boldsymbol{y}} \log p^*(\boldsymbol{\vartheta}|\boldsymbol{y}, \tau, \boldsymbol{v})\| = \frac{1}{\tau^2}(\boldsymbol{\vartheta} - \mathbb{E}_{\boldsymbol{\Theta}^{(K)}}[\boldsymbol{\Theta}^{(K)}]) \leq 2K\sqrt{k}/\tau^2$, 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 $\int \ell(\boldsymbol{\vartheta}, \boldsymbol{d}) p^*(\boldsymbol{\vartheta}|\boldsymbol{y}, \tau, \boldsymbol{v}) \mu_{\Theta|V}(\boldsymbol{v}, \mathrm{d}\boldsymbol{\vartheta}) \leq \bar{R}$, 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 13(b) has been shown. ■

Proof [Lemma 13(c)] Finally, we prove (27). For any $K > 0$, we may write[6]

$$\mu_{\boldsymbol{\Theta}^{(K)}|\boldsymbol{V}}(\boldsymbol{v}, \cdot) = \mu_{\boldsymbol{\Theta}|\boldsymbol{V}}(\boldsymbol{v}, \cdot)|_{[-K,K]^k} + \mu_{\boldsymbol{\Theta}|\boldsymbol{V}}(\boldsymbol{v}, ([-K, K]^k)^c)\delta_{\boldsymbol{0}}(\cdot). \tag{30}$$

6. 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.

Choose $\bar{K}, \epsilon' > 0$ such that $|\tau' - \tau| < \epsilon'$ implies

$$\int_{[-\bar{K},\bar{K}]^k} \frac{1}{Z(\boldsymbol{y}, \tau', \boldsymbol{v}, \infty)} e^{-\frac{1}{2\tau'^2}\|\boldsymbol{y}-\boldsymbol{\vartheta}\|^2} \mu_{\boldsymbol{\Theta}|\boldsymbol{V}}(\boldsymbol{v}, \mathrm{d}\boldsymbol{\vartheta}) \geq \frac{1}{2} \int \frac{1}{Z(\boldsymbol{y}, \tau, \boldsymbol{v}, \infty)} e^{-\frac{1}{2\tau^2}\|\boldsymbol{y}-\boldsymbol{\vartheta}\|^2} \mu_{\boldsymbol{\Theta}|\boldsymbol{V}}(\boldsymbol{v}, \mathrm{d}\boldsymbol{\vartheta}).$$

Fix $\epsilon > 0$ and $K' > K > 0$ with K' possibly equal to infinity. By (24), we may choose $\boldsymbol{d}^*$ such that

$$\int \frac{1}{Z(\boldsymbol{y}, \tau, \boldsymbol{v}, K)} \ell(\boldsymbol{\vartheta}, \boldsymbol{d}^*) e^{-\frac{1}{2\tau^2}\|\boldsymbol{y}-\boldsymbol{\vartheta}\|^2} \mu_{\boldsymbol{\Theta}^{(K)}|\boldsymbol{V}}(\boldsymbol{v}, \mathrm{d}\boldsymbol{\vartheta}) \leq (1+\epsilon) R(\boldsymbol{y}, \tau, \boldsymbol{v}, K, \infty). \tag{31}$$

By the definition of $\bar{K}$, there exists $\boldsymbol{\vartheta}^* \in [-\bar{K}, \bar{K}]^k$ such that

$$\ell(\boldsymbol{\vartheta}^*, \boldsymbol{d}^*) \leq 2(1+\epsilon) R(\boldsymbol{y}, \tau, \boldsymbol{v}, K, \infty).$$

By (23), we conclude that

$$\ell(\boldsymbol{\vartheta}, \boldsymbol{d}^*) \leq C \left(1 + \sqrt{2(1+\epsilon) R(\boldsymbol{y}, \tau, \boldsymbol{v}, K, \infty)} + \sqrt{\ell(\boldsymbol{\vartheta}, \boldsymbol{d}^*)}\right) \|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}^*\|,$$

whence

$$\ell(\boldsymbol{\vartheta}, \boldsymbol{d}^*) \leq \left(1 + \sqrt{2(1+\epsilon) R(\boldsymbol{y}, \tau, \boldsymbol{v}, K, \infty)} + 3C\|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}^*\|\right)^2. \tag{32}$$

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| \\
&\leq \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 + \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| \\
&\leq \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) \\
&\quad + \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| \\
&\leq \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 (32). Then, by (31),

$$\begin{aligned}
&Z(\boldsymbol{y}, \tau', \boldsymbol{v}, K') R(\boldsymbol{y}, \tau', \boldsymbol{v}, K', \infty) \\
&\qquad \leq \int \ell(\boldsymbol{\vartheta}, \boldsymbol{d}^*) e^{-\frac{1}{2\tau'^2}\|\boldsymbol{y}-\boldsymbol{\vartheta}\|^2} \mu_{\boldsymbol{\Theta}^{(K')}|\boldsymbol{V}}(\boldsymbol{v}, \mathrm{d}\boldsymbol{\vartheta}) \\
&\qquad \leq (1+\epsilon) Z(\boldsymbol{y}, \tau, \boldsymbol{v}, K) R(\boldsymbol{y}, \tau, \boldsymbol{v}, K, \infty) + \xi(K, \tau')(1 + R(\boldsymbol{y}, \tau, \boldsymbol{v}, K, \infty)).
\end{aligned}$$

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, $\bar{R}(K) = \max_{\|\boldsymbol{\vartheta}\|_\infty \leq K} \ell(\boldsymbol{\vartheta}, \boldsymbol{0})$ 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

$$\lim_{\substack{K\to\infty \\ \tau'\to\tau}} R(\boldsymbol{y}, \tau', \boldsymbol{v}, K, \infty) = R(\boldsymbol{y}, \tau, \boldsymbol{v}, \infty, \infty), \tag{33}$$

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. (33) and uniform integrability imply Eq. (27). ∎

Appendix B. Proof for reduction from GFOMs to AMP (Lemma 6)

In this section, we prove Lemma 6.

B.1. A general change of variables

For any GFOM (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 change of variables.

The change of variables is determined by a collection of $r \times r$ matrices $(\boldsymbol{\xi}_{t,s})_{t\ge 1, 1\le s\le t}$, $(\boldsymbol{\zeta}_{t,s})_{t\ge 1, 0\le s<t}$. We will often omit subscripts outside of the parentheses. Define recursively the functions $(f_t)_{t\ge 0}$, $(\phi_t)_{t\ge 1}$

$$\begin{aligned} f_t(\boldsymbol{b}^1, \dots, \boldsymbol{b}^t; y, \boldsymbol{u}) &= F_t^{(1)}(\phi_1(\boldsymbol{b}^1; y, \boldsymbol{u}), \dots, \phi_t(\boldsymbol{b}^1, \dots, \boldsymbol{b}^t; y, \boldsymbol{u}); y, \boldsymbol{u}) \\ \phi_t(\boldsymbol{b}^1, \dots, \boldsymbol{b}^t; y, \boldsymbol{u}) &= \boldsymbol{b}^t + \sum_{s=0}^{t-1} f_s(\boldsymbol{b}^1, \dots, \boldsymbol{b}^s; y, \boldsymbol{u}) \boldsymbol{\zeta}_{t,s}^{\mathsf{T}} \\ &\quad + G_t^{(2)}(\phi_1(\boldsymbol{b}^1; y, \boldsymbol{u}), \dots, \phi_{t-1}(\boldsymbol{b}^1, \dots, \boldsymbol{b}^{t-1}; y, \boldsymbol{u}); y, \boldsymbol{u}), \end{aligned} \tag{34a}$$

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 $(g_t)_{t\geq 1}$, $(\varphi_t)_{t\geq 1}$

$$\begin{aligned}\varphi_{t+1}(\boldsymbol{a}^1, \dots, \boldsymbol{a}^{t+1}; \boldsymbol{v}) &= \boldsymbol{a}^{t+1} + \sum_{s=1}^{t} g_s(\boldsymbol{a}^1, \dots, \boldsymbol{a}^{t+1}; \boldsymbol{v})\boldsymbol{\xi}_{t,s}^{\mathsf{T}} \\ &\quad + F_t^{(2)}(\phi_1(\boldsymbol{a}^1; \boldsymbol{v}), \dots, \phi_t(\boldsymbol{a}^1, \dots, \boldsymbol{a}^t; \boldsymbol{v}); \boldsymbol{v}), \\ g_{t+1}(\boldsymbol{a}^1, \dots, \boldsymbol{a}^t; \boldsymbol{v}) &= G_{t+1}^{(1)}(\varphi_1(\boldsymbol{a}^1; \boldsymbol{v}), \dots, \varphi_{t+1}(\boldsymbol{a}^1, \dots, \boldsymbol{a}^{t+1}; \boldsymbol{v}); \boldsymbol{v}),\end{aligned} \tag{34b}$$

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

$$\begin{aligned}\boldsymbol{a}^{t+1} &= \boldsymbol{X}^{\mathsf{T}} f_t(\boldsymbol{b}^1, \dots, \boldsymbol{b}^t; y, \boldsymbol{u}) - \sum_{s=1}^{t} g_s(\boldsymbol{a}^1, \dots, \boldsymbol{a}^s; \boldsymbol{v})\boldsymbol{\xi}_{t,s}^{\mathsf{T}}, \\ \boldsymbol{b}^t &= \boldsymbol{X} g_t(\boldsymbol{a}^1, \dots, \boldsymbol{a}^t; \boldsymbol{v}) - \sum_{s=0}^{t-1} f_s(\boldsymbol{b}^1, \dots, \boldsymbol{b}^s; y, \boldsymbol{u})\boldsymbol{\zeta}_{t,s}^{\mathsf{T}}\end{aligned} \tag{35}$$

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. Approximate message passing and state evolution

We call the iteration (35) 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. When this recursion is satisfied, the sums on the right-hand sides are the "Onsager corrections" which we left unspecified in Eq. (16). For future reference, the AMP iteration with the Onsager correction terms explicitly specified is

$$\begin{aligned}\boldsymbol{a}^{t+1} &= \boldsymbol{X}^{\mathsf{T}} f_t(\boldsymbol{b}^1, \dots, \boldsymbol{b}^t; \boldsymbol{y}, \boldsymbol{u}) - \sum_{s=1}^{t} g_s(\boldsymbol{a}^1, \dots, \boldsymbol{a}^s; \boldsymbol{v})\boldsymbol{\xi}_{t,s}^{\mathsf{T}}, \\ \boldsymbol{b}^t &= \boldsymbol{X} g_t(\boldsymbol{a}^1, \dots, \boldsymbol{a}^t; \boldsymbol{v}) - \sum_{s=0}^{t-1} f_s(\boldsymbol{b}^1, \dots, \boldsymbol{b}^s; \boldsymbol{y}, \boldsymbol{u})\boldsymbol{\zeta}_{t,s}^{\mathsf{T}},\end{aligned} \tag{36}$$

where $\boldsymbol{\xi}_{t,s}$, $\boldsymbol{\zeta}_{t,s}$ satisfy an AMP- and model-specific recursion.

The state evolution characterization of the iterates (see Eq. (17)) holds whenever the matrices $\boldsymbol{\xi}_{t,s}$, $\boldsymbol{\zeta}_{t,s}$ satisfy this recursion. In this section, we specify this recursion. We simultaneously specify the parameters $(\boldsymbol{\alpha}_s), (\boldsymbol{T}_{s,s'})$ appearing in Lemma 6 in both the high-dimensional regression and low-rank matrix estimation models.

B.2.1. 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 $(\Sigma_{s,t})_{s,t\ge 0}$ which did not appear in the statement of Lemma 6. The recursion, whose lines are implemented in the order in which they appear, is

$$\begin{aligned}
\xi_{t,s} &= \mathbb{E}[\partial_{B^s} f_t(B^1,\dots,B^t; h(B^0,W),U)], \quad 1\le s\le t,\\
\alpha_{t+1} &= \mathbb{E}[\partial_{B^0} f_t(B^1,\dots,B^t; h(B^0,W),U)],\\
T_{s+1,t+1} &= \mathbb{E}[f_s(B^1,\dots,B^s; h(B^0,W),U) f_t(B^1,\dots,B^t; h(B^0,W),U)], \quad 0\le s\le t,\\
\zeta_{t,s} &= \frac{1}{\delta}\mathbb{E}[\partial_{Z^{s+1}} g_t(\alpha_1\Theta + Z^1,\dots,\alpha_t\Theta+Z^t;V)], \quad 0\le s\le t-1,\\
\Sigma_{0,t} &= \frac{1}{\delta}\mathbb{E}[\Theta g_t(\alpha_1\Theta + Z^1,\dots,\alpha_t\Theta+Z^t;V)],\\
\Sigma_{s,t} &= \frac{1}{\delta}\mathbb{E}[g_s(\alpha_1\Theta + Z^1,\dots,\alpha_t\Theta+Z^s;V) g_t(\alpha_1\Theta + Z^1,\dots,\alpha_t\Theta+Z^t;V)], \quad 1\le s\le t,
\end{aligned} \tag{37}$$

where $\Theta\sim\mu_\Theta$, $U\sim\mu_U$, $V\sim\mu_V$, $W\sim\mu_W$, $(B^0,\dots,B^t)\sim \mathsf{N}(\mathbf{0},\boldsymbol{\Sigma}_{[0:t]})$, $(Z^1,\dots,Z^t)\sim \mathsf{N}(\mathbf{0},\boldsymbol{T}_{[1:t]})$, all independent. We initialize just before the second line with $\Sigma_{0,0}=\mathbb{E}[\Theta^2]$.

Eq. (17) for $(\alpha_s),(T_{s,s'})$ defined in this way is a special case of Proposition 5 of Javanmard and Montanari (2013), as we now explain. First, we fix iteration t. Then, we design an interation which agrees, after a change of variables, with iteration (36) up to iteration t and to which we can apply the results of Javanmard and Montanari (2013). After this change of variables, the state evolution up to iteration t follows from Proposition 5 of Javanmard and Montanari (2013). Because t was arbitrary and we take $n,p\to\infty$ before $t\to\infty$, this establishes the result. This section simply provides explicitly the appropriate change of variables, but it does not provide substantial explanation for why the state evolution holds. For such an exposition, we refer the reader to Javanmard and Montanari (2013). Other works establishing state evolution for other versions of AMP include Bayati and Montanari (2011); Berthier et al. (2019). These contain many of the same ideas, but do not consider an AMP iteration which translates as easily into our setting, in which the AMP updates are allowed to depend upon the full past.

Javanmard and Montanari (2013) study an interation in which the iterates are matrices $u^s\in\mathbb{R}^{n\times(t+1)}$ and $v^s\in\mathbb{R}^{p\times(t+1)}$. The AMP iteration (36) up to time t, in which each update depends upon the full past, is an instance of the AMP of Javanmard and Montanari (2018) in which the columns of the matrices v^s and u^s contain functions of the iterates $\boldsymbol{a}^s$, $\boldsymbol{b}^s$, and parameters of the high-dimensional regression model. Specifically,

$$u^s \leftarrow \begin{pmatrix} | & | & & | & | & & | \\ \boldsymbol{X\theta} & \boldsymbol{b}^1 & \cdots & \boldsymbol{b}^s & \mathbf{0} & \cdots & \mathbf{0} \\ | & | & & | & | & & | \end{pmatrix},$$

$$v^s \leftarrow \begin{pmatrix} | & | & & | & | & & | \\ \mathbf{0} & \boldsymbol{a}^1-\alpha_1\boldsymbol{\theta} & \cdots & \boldsymbol{a}^s-\alpha_s\boldsymbol{\theta} & \mathbf{0} & \cdots & \mathbf{0} \\ | & | & & | & | & & | \end{pmatrix},$$

The following change of variables transforms (36) into equations (28) and (29) of Proposition 5 in Javanmard and Montanari (2013). Our notation is on the right and is separated from the notation of Javanmard and Montanari (2013) by the symbol "$\leftarrow$".

$$\tilde{A}\leftarrow \boldsymbol{X}, \quad m\leftarrow n, \quad n\leftarrow p, \quad q\leftarrow t+1,$$

$$u^s(i) \leftarrow \begin{cases} \boldsymbol{X}\boldsymbol{\theta} & i = 1, \\ \boldsymbol{b}^{i-1} & 2 \le i \le s+1, \\ \mathbf{0} & \text{otherwise}, \end{cases} \qquad \text{and} \qquad v^s(i) \leftarrow \begin{cases} \boldsymbol{a}^{i-1} - \alpha_{i-1}\boldsymbol{\theta} & 2 \le i \le s+1, \\ \mathbf{0} & \text{otherwise}, \end{cases}$$

$$y(i) \leftarrow \begin{cases} \boldsymbol{v} & i = 1, \\ \boldsymbol{\theta} & i = 2, \\ \mathbf{0} & \text{otherwise}, \end{cases} \qquad \text{and} \qquad w(i) \leftarrow \begin{cases} \boldsymbol{u} & i = 1, \\ \boldsymbol{w} & i = 2, \\ \mathbf{0} & \text{otherwise}, \end{cases}$$

The "(i)" notation indexes columns of a matrix and u^s, v^s, y, and w are matrices with $t+1$ columns. In matrix form

$$y \leftarrow \begin{pmatrix} | & | & | & & | \\ \boldsymbol{v} & \boldsymbol{\theta} & \mathbf{0} & \cdots & \mathbf{0} \\ | & | & | & & | \end{pmatrix}, \qquad w \leftarrow \begin{pmatrix} | & | & | & & | \\ \boldsymbol{u} & \boldsymbol{w} & \mathbf{0} & \cdots & \mathbf{0} \\ | & | & | & & | \end{pmatrix}.$$

The update functions of Javanmard and Montanari (2013) are

$$\widehat{e}(v, y; s)(i) \leftarrow \begin{cases} y(2) & i = 1, \\ g_{i-1}(v(2) + \alpha_1 y(2), \ldots, v(i) + \alpha_{i-1} y(2); y(1)) & 2 \le i \le s+1, \\ \mathbf{0} & \text{otherwise}, \end{cases}$$

$$\widehat{h}(u, w; s)(i) \leftarrow \begin{cases} f_{i-2}(u(2), \ldots, u(i-1); h(u(1), w(2)), w(1)), & 2 \le i \le s+2, \\ \mathbf{0} & \text{otherwise}. \end{cases}$$

Indeed, then in matrix form

$$\widehat{e}(v^s, y; s)(i) \leftarrow \begin{pmatrix} | & | & & | & | & & | \\ \boldsymbol{\theta} & g_1(\boldsymbol{a}^1; \boldsymbol{v}) & \cdots & g_s(\boldsymbol{a}^1, \ldots, \boldsymbol{a}^s; \boldsymbol{v}) & \mathbf{0} & \cdots & \mathbf{0} \\ | & | & & | & | & & | \end{pmatrix},$$

and

$$\widehat{h}(u^s, w; s)(i) \leftarrow \begin{pmatrix} | & | & & & | & | & | \\ \mathbf{0} & f_0(\boldsymbol{y}, \boldsymbol{u}) & \cdots & f_s(\boldsymbol{b}^1, \ldots, \boldsymbol{b}^s; \boldsymbol{y}, \boldsymbol{u}) & \mathbf{0} & \cdots & \mathbf{0} \\ | & | & & | & | & & | \end{pmatrix}.$$

The Onsager correction coefficients $(\xi_{t,s})$ and $(\zeta_{t,s})$ are related to the operators B_s and D_s in Javanmard and Montanari (2013). The operator B_s is a linear operator $\mathbb{R}^{n\times(t+1)} \to \mathbb{R}^{n\times(t+1)}$ which in our context consists of the row-wise application of the same linear transformation, which we call $\overline{\mathsf{B}}_s \in \mathbb{R}^{(t+1)\times(t+1)}$. The entries of this linear transformation are

$$\begin{aligned} (\overline{\mathsf{B}}_s)_{i,j} &= \frac{1}{\delta}\mathbb{E}[\partial_{V^j}\widehat{e}(V, Y; s)(i)] \\ &\leftarrow \begin{cases} 0 & i = 1 \text{ or } j = 1, \\ \frac{1}{\delta}\mathbb{E}[\partial_{Z^{j-1}} g_{i-1}(\alpha_1\Theta + Z^1, \ldots, \alpha_{i-1}\Theta + Z^{i-1}; V)] & 2 \le j \le i \le s+1, \\ 0 & \text{otherwise}. \end{cases} \end{aligned}$$

$$(\overline{\mathsf{D}}_s)_{i,j} = \mathbb{E}[\partial_{U^j} \widehat{h}(U, W; s)(i)]$$
$$\leftarrow \begin{cases} \mathbb{E}[\partial_{B^{j-1}} f_{i-1}(B^1, \dots, B^i; h(B^0, W), U)] & 1 \le j \le i-1 \le s+1, \\ 0 & \text{otherwise.} \end{cases}$$

The Onsager coefficients and state evolution coefficients are arrived at through the change of variables:

$$(\overline{\mathsf{B}}_s)_{s+1,s'+2} \leftarrow \zeta_{s,s'}, \quad (\overline{\mathsf{D}}_s)_{s+2,s'+1} \leftarrow \xi_{s,s'} \quad (\overline{\mathsf{D}}_s)_{s+1,1} \leftarrow \alpha_s,$$

over the relevant ranges of s, s'. In matrix form,

$$\overline{\mathsf{B}}_s \leftarrow \begin{pmatrix} 0 & 0 & \cdots & & & 0 & \cdots & 0 \\ 0 & \zeta_{1,0} & 0 & \cdots & & 0 & \cdots & 0 \\ 0 & \zeta_{2,0} & \zeta_{2,1} & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \ddots & \vdots & & \vdots \\ 0 & \zeta_{s,0} & \cdots & & \zeta_{s,s-1} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & & 0 & 0 & \cdots & 0 \\ \vdots & & & & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & & 0 & 0 & \cdots & 0 \end{pmatrix},$$

and

$$\overline{\mathsf{D}}_s \leftarrow \begin{pmatrix} 0 & 0 & \cdots & & & 0 & \cdots & 0 \\ \alpha_1 & 0 & \cdots & & & 0 & \cdots & 0 \\ \alpha_2 & \xi_{1,1} & 0 & \cdots & & 0 & \cdots & 0 \\ \alpha_3 & \xi_{2,1} & \xi_{2,2} & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \ddots & \vdots & & \vdots \\ \alpha_{s+1} & \xi_{s,1} & \cdots & & \xi_{s,s} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & & 0 & 0 & \cdots & 0 \\ \vdots & & & & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

We remark that in Javanmard and Montanari (2013) the quantities $(\overline{\mathsf{B}}_s)_{s+1,s'+2}$, $(\overline{\mathsf{D}}_s)_{s+1,s'+1}$, and $(\overline{\mathsf{D}}_s)_{s+1,1}$ are empirical averages rather than population averages. Because empirical averages concentrate well on their expecation, 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 Berthier et al. (2019). The state evolution matrices now correspond to

$$\mathbb{E}[V^{s+1}(s+2)V^{s+1}(s'+2)] = \mathbb{E}[\widehat{h}(U, W; s)(s+2)\widehat{h}(U, W; s)(s'+2)]$$
$$\leftarrow \mathbb{E}[f_s(B^1, \dots, B^s; h(B^0, W), U) f_{s'}(B^1, \dots, B^{s'}; h(B^0, W), U)]$$
$$= T_{s+1,s'+1}, \quad \text{for } s \ge s' \ge 0,$$

and

$$\begin{aligned}\mathbb{E}[U^s(s+1)U^s(s'+1)] &= \frac{1}{\delta}\mathbb{E}[\widehat{e}(V,Y;s)(s+1)\widehat{e}(V,Y;s)(s'+1)] \\ &\leftarrow \frac{1}{\delta}\mathbb{E}[g_s(\alpha_1\Theta + Z^1,\dots,\alpha_s\Theta+Z^s;V)g_{s'}(\alpha_1\Theta+Z^1,\dots,\alpha_{s'}\Theta+Z^{s'};V)] \\ &= \Sigma_{s,s'}, \quad \text{for } s \geq s' \geq 1,\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[U^s(s+1)U^s(1)] &= \frac{1}{\delta}\mathbb{E}[\widehat{e}(V,Y;s)(s+1)\widehat{e}(V,Y;s)(1)] \\ &\leftarrow \frac{1}{\delta}\mathbb{E}[g_s(\alpha_1\Theta + Z^1,\dots,\alpha_s\Theta+Z^s;V)\Theta] \\ &= \Sigma_{s,0}, \quad \text{for } s \geq 1.\end{aligned}$$

From this change of variables, Eq. (17) holds in the high-dimensional regression model from Theorem 1 and Proposition 5 of Javanmard and Montanari (2013).

B.2.2. LOW-RANK MATRIX ESTIMATION AMP

In the low-ank matrix estimation model, the recursion defining $(\boldsymbol{x}_{t,x})$, $(\boldsymbol{\zeta}_{t,s})$ also defines $(\boldsymbol{\alpha}_t)$, $(\boldsymbol{T}_{s,t})_{s,t\geq 1}$ as well as collections of $r \times r$ matrices $(\boldsymbol{\gamma}_t)_{t\geq 1}$, $(\boldsymbol{\Sigma}_{s,t})_{s,t\geq 0}$ which did not appear in Lemma 6. The recursion, whose lines are implemented in the order in which they appear, is

$$\begin{aligned}\boldsymbol{\xi}_{t,s} &= \mathbb{E}[\nabla_{\tilde{\boldsymbol{Z}}^s} f_t(\boldsymbol{\gamma}_1\boldsymbol{\Lambda}+\tilde{\boldsymbol{Z}}^1,\dots,\boldsymbol{\gamma}_t\boldsymbol{\Lambda}+\tilde{\boldsymbol{Z}}^t;0,\boldsymbol{U})], \quad 1\leq s\leq t, \\ \boldsymbol{\alpha}_{t+1} &= \mathbb{E}[f_t(\boldsymbol{\gamma}_1\boldsymbol{\Lambda}+\tilde{\boldsymbol{Z}}^1,\dots,\boldsymbol{\gamma}_t\boldsymbol{\Lambda}+\tilde{\boldsymbol{Z}}^t;0,\boldsymbol{U})\boldsymbol{\Lambda}^{\mathsf{T}}], \\ \boldsymbol{T}_{s+1,t+1} &= \mathbb{E}[f_s(\boldsymbol{\gamma}_1\boldsymbol{\Lambda}+\tilde{\boldsymbol{Z}}^1,\dots,\boldsymbol{\gamma}_t\boldsymbol{\Lambda}+\tilde{\boldsymbol{Z}}^s;0,\boldsymbol{U})f_t(\boldsymbol{\gamma}_1\boldsymbol{\Lambda}+\tilde{\boldsymbol{Z}}^1,\dots,\boldsymbol{\gamma}_t\boldsymbol{\Lambda}+\tilde{\boldsymbol{Z}}^t;0,\boldsymbol{U})^{\mathsf{T}}], \ s\leq t, \\ \boldsymbol{\zeta}_{t,s} &= \frac{1}{\delta}\mathbb{E}[\nabla_{\boldsymbol{Z}^{s+1}} g_t(\boldsymbol{\alpha}_1\boldsymbol{\Theta}+Z^1,\dots,\boldsymbol{\alpha}_t\boldsymbol{\Theta}+\boldsymbol{Z}^t;\boldsymbol{V})], \quad 0\leq s\leq t-1, \\ \boldsymbol{\gamma}_t &= \frac{1}{\delta}\mathbb{E}[g_t(\boldsymbol{\alpha}_1\boldsymbol{\Theta}+\boldsymbol{Z}^1,\dots,\boldsymbol{\alpha}_t\boldsymbol{\Theta}+\boldsymbol{Z}^t;\boldsymbol{V})\boldsymbol{\Theta}^{\mathsf{T}}], \\ \boldsymbol{\Sigma}_{s,t} &= \frac{1}{\delta}\mathbb{E}[g_s(\boldsymbol{\alpha}_1\boldsymbol{\Theta}+\boldsymbol{Z}^1,\dots,\boldsymbol{\alpha}_t\boldsymbol{\Theta}+\boldsymbol{Z}^s;\boldsymbol{V})g_t(\boldsymbol{\alpha}_1\boldsymbol{\Theta}+\boldsymbol{Z}^1,\dots,\boldsymbol{\alpha}_t\boldsymbol{\Theta}+\boldsymbol{Z}^t;\boldsymbol{V})^{\mathsf{T}}], \\ &\qquad\qquad 1\leq s\leq t,\end{aligned} \tag{38}$$

where $\boldsymbol{\Lambda}\sim\mu_{\boldsymbol{\Lambda}}$ $\boldsymbol{U}\sim\mu_{\boldsymbol{U}}$, $\boldsymbol{\Theta}\sim\mu_{\boldsymbol{\Theta}}$, $\boldsymbol{V}\sim\mu_{\boldsymbol{V}}$, $(\tilde{\boldsymbol{Z}}^1,\dots,\tilde{\boldsymbol{Z}}^t)\sim\mathsf{N}(\boldsymbol{0},\boldsymbol{\Sigma}_{[1:t]})$, and $(\boldsymbol{Z}^1,\dots,\boldsymbol{Z}^t)\sim\mathsf{N}(\boldsymbol{0},\boldsymbol{T}_{[1:t]})$, all independent. Here ∇ denotes the Jacobian with respect to the subscripted (vectorial) argument, which exists almost everywhere because the functions involved are Lipschitz and the random variables have density with respect to Lebesgue measure (Evans and Gariepy, 2015, 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 (17), we will show

$$\frac{1}{n}\sum_{i=1}^n \psi(\boldsymbol{b}_i^1,\dots,\boldsymbol{b}_i^t;\boldsymbol{u}_i,\boldsymbol{\lambda}_i) \xrightarrow{\text{p}} \mathbb{E}[\psi(\boldsymbol{\gamma}_1\boldsymbol{\Lambda}+\tilde{\boldsymbol{Z}}^1,\dots,\boldsymbol{\gamma}_t\boldsymbol{\Lambda}+\tilde{\boldsymbol{Z}}^t;\boldsymbol{U},\boldsymbol{\Lambda})],$$

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. (17) for $(\alpha_s), (T_{s,s'})$ defined in this way. We consider the $r = 1$ case, as $r > 1$ is similar but requires more notational overhead. Recall $\boldsymbol{X} = \frac{1}{n}\boldsymbol{\lambda}\boldsymbol{\theta}^\mathsf{T} + \boldsymbol{Z}$. Because $\boldsymbol{X}$ is not a Gaussian matrix due to its low-rank component, the results of Javanmard and Montanari (2013) do not directly apply. To apply them, we use a technique common in the AMP literature: we design an AMP iteration which, via a certain set of change of variables, closely tracks the iterates of (36) and to which the results of Javanmard and Montanari (2013) apply. Then, the state evolution for the new AMP iteration will transfer to a state evolution for the iteration (36) via an approximation argument.

First, note

$$\boldsymbol{a}^{t+1} - \frac{1}{n}\langle \boldsymbol{\lambda}, f_t(\boldsymbol{b}^1, \dots, \boldsymbol{b}^t; 0, \boldsymbol{u})\rangle \boldsymbol{\theta} = \boldsymbol{Z}^\mathsf{T} f_t(\boldsymbol{b}^1, \dots, \boldsymbol{b}^t; 0, \boldsymbol{u}) - \sum_{s=1}^{t} \xi_{t,s} g_s(\boldsymbol{a}^1, \dots, \boldsymbol{a}^s; \boldsymbol{v}),$$
$$\boldsymbol{b}^t - \frac{1}{n}\langle \boldsymbol{\theta}, g_t(\boldsymbol{a}^1, \dots, \boldsymbol{a}^t; \boldsymbol{v})\rangle \boldsymbol{\lambda} = \boldsymbol{Z} g_t(\boldsymbol{a}^1, \dots, \boldsymbol{a}^t; \boldsymbol{v}) - \sum_{s=0}^{t-1} \zeta_{t,s} f_s(\boldsymbol{b}^1, \dots, \boldsymbol{b}^s; \boldsymbol{y}, \boldsymbol{u}).$$

We introduce a change of variables:

$$\hat{f}_t(d^1, \dots, d^t; u, \lambda) := f_t(d^1 + \gamma_1\lambda, \dots, d^t + \gamma_t\lambda; 0, u), \qquad \boldsymbol{d}^t = \boldsymbol{b}^t - \gamma_t\boldsymbol{\lambda} \in \mathbb{R}^n,$$
$$\hat{g}_t(c^1, \dots, c^t; v, \theta) := g_t(c^1 + \alpha_1\theta, \dots, c^t + \alpha_t\theta; v), \qquad \boldsymbol{c}^t = \boldsymbol{a}^t - \alpha_t\boldsymbol{\theta} \in \mathbb{R}^p.$$

Because f_t, g_t are Lipschitz continuous, so too are $\hat{f}_t$, $\hat{g}_t$. We have

$$\boldsymbol{a}^{t+1} - \frac{1}{n}\langle \boldsymbol{\lambda}, \hat{f}_t(\boldsymbol{d}^1, \dots, \boldsymbol{d}^t; \boldsymbol{u}, \boldsymbol{\lambda})\rangle \boldsymbol{\theta} = \boldsymbol{Z}^\mathsf{T} \hat{f}_t(\boldsymbol{d}^1, \dots, \boldsymbol{d}^t; \boldsymbol{u}, \boldsymbol{\lambda}) - \sum_{s=1}^{t} \xi_{t,s} \hat{g}_s(\boldsymbol{c}^1, \dots, \boldsymbol{c}^s; \boldsymbol{v}, \boldsymbol{\theta}),$$
$$\boldsymbol{b}^t - \frac{1}{n}\langle \boldsymbol{\theta}, \hat{g}_t(\boldsymbol{c}^1, \dots, \boldsymbol{c}^t; \boldsymbol{v}, \boldsymbol{\theta})\rangle \boldsymbol{\lambda} = \boldsymbol{Z} \hat{g}_t(\boldsymbol{c}^1, \dots, \boldsymbol{c}^t; \boldsymbol{v}, \boldsymbol{\theta}) - \sum_{s=0}^{t-1} \zeta_{t,s} \hat{f}_s(\boldsymbol{b}^1, \dots, \boldsymbol{b}^s; \boldsymbol{u}, \boldsymbol{\lambda}).$$

Define

$$\hat{\boldsymbol{c}}^{t+1} = \boldsymbol{Z}^\mathsf{T} \hat{f}_t(\hat{\boldsymbol{d}}^1, \dots, \hat{\boldsymbol{d}}^t; \boldsymbol{u}, \boldsymbol{\lambda}) - \sum_{s=1}^{t} \xi_{t,s} \hat{g}_s(\hat{\boldsymbol{c}}^1, \dots, \hat{\boldsymbol{c}}^t; \boldsymbol{v}, \boldsymbol{\theta}),$$
$$\hat{\boldsymbol{d}}^t = \boldsymbol{Z} \hat{g}_t(\hat{\boldsymbol{c}}^1, \dots, \hat{\boldsymbol{c}}^t; \boldsymbol{v}, \boldsymbol{\theta}) - \sum_{s=0}^{t-1} \zeta_{t,s} \hat{f}_s(\hat{\boldsymbol{d}}^1, \dots, \hat{\boldsymbol{d}}^t; \boldsymbol{u}, \boldsymbol{\lambda}).$$

Because this iteration involves only multiplication by a Gaussian matrix, state evolution holds for it by Theorem 1 and Proposition 5 of Javanmard and Montanari (2013) via a change of variables as in high-dimensional regression AMP (see previous section). In this case, the state evolution states that for any pseudo-Lipschitz function $\psi : \mathbb{R}^{t+2} \to \mathbb{R}$ of order 2, we have

$$\begin{aligned} \frac{1}{p}\sum_{j=1}^{p} \psi(\hat{c}_j^1, \dots, \hat{c}_j^t, v_j, \theta_j) &\xrightarrow{\text{p}} \mathbb{E}[\psi(Z^1, \dots, Z^t, V, \Theta)], \\ \frac{1}{n}\sum_{i=1}^{n} \psi(\hat{d}_i^1, \dots, \hat{d}_i^t, u_i, \lambda_i) &\xrightarrow{\text{p}} \mathbb{E}[\psi(\tilde{Z}^1, \dots, \tilde{Z}^t, U, \Lambda)], \end{aligned} \tag{39}$$

where $Z^s, \tilde{Z}^s$ are as in the state evolution Eq. (38).

Because ψ is pseudo-Lipschitz of order 2, to establish (17), it suffices to show

$$\frac{1}{n}\|\hat{\boldsymbol{c}}^t - \boldsymbol{c}^t\|_2^2 \xrightarrow{\text{p}} 0, \qquad \frac{1}{n}\|\hat{\boldsymbol{d}}^t - \boldsymbol{d}^t\|_2^2 \xrightarrow{\text{p}} 0. \tag{40}$$

We proceed by induction. By the weak law of large numbers, we have that $\frac{1}{n}\langle \boldsymbol{\lambda}, \hat{f}_0(\boldsymbol{u}, \boldsymbol{\lambda})\rangle = \frac{1}{n}\langle \boldsymbol{\lambda}, f_0(0, \boldsymbol{u})\rangle \xrightarrow{\text{p}} \alpha_1$. Therefore, $\boldsymbol{c}^1 = \boldsymbol{Z}^\mathsf{T}\hat{f}_0(\boldsymbol{u}, \boldsymbol{\lambda}) + o_p(1)\boldsymbol{\theta} = \hat{\boldsymbol{c}}^1 + o_p(1)\boldsymbol{\theta}$. Since $\frac{1}{p}\|\boldsymbol{\theta}\|_2^2 \xrightarrow{\text{p}} \mathbb{E}[\Theta^2]$, we have that $\frac{1}{n}\|\boldsymbol{c}^1 - \hat{\boldsymbol{c}}^1\|_2^2 \xrightarrow{\text{p}} 0$.

Because $\hat{g}_1$ is Lipschitz and $\frac{1}{p}\|\boldsymbol{\theta}\|^2 = O_p(1)$, we have $|\frac{1}{n}\langle \boldsymbol{\theta}, \hat{g}_1(\boldsymbol{c}^1; \boldsymbol{v}, \boldsymbol{\theta})\rangle - \frac{1}{n}\langle \boldsymbol{\theta}, \hat{g}_1(\hat{\boldsymbol{c}}^1; \boldsymbol{v}, \boldsymbol{\theta})\rangle| \xrightarrow{\text{p}} 0$. By (39), we have that $\frac{1}{n}\langle \boldsymbol{\theta}, \hat{g}_1(\hat{\boldsymbol{c}}^1; \boldsymbol{v}, \boldsymbol{\theta})\rangle \xrightarrow{\text{p}} \gamma_1$. We have

$$\frac{1}{n}\|\hat{g}_1(\boldsymbol{c}^1; \boldsymbol{v}, \boldsymbol{\theta}) - \hat{g}_1(\hat{\boldsymbol{c}}^1; \boldsymbol{v}, \boldsymbol{\theta})\|_2^2 \le \frac{1}{n}L^2\|\boldsymbol{c}^1 - \hat{\boldsymbol{c}}^1\|_2^2 \xrightarrow{\text{p}} 0,$$

where L is a Lipschitz constant for $\hat{g}_1$. By Bai and Yin (2008), the maximal singular value of $\boldsymbol{Z}^T\boldsymbol{Z}$ is $O_p(1)$. Therefore, $\frac{1}{n}\|\boldsymbol{Z}\hat{g}_1(\boldsymbol{c}^1; \boldsymbol{v}, \boldsymbol{\theta}) - \boldsymbol{Z}\hat{g}_1(\hat{\boldsymbol{c}}^1; \boldsymbol{v}, \boldsymbol{\theta})\|_2^2 \xrightarrow{\text{p}} 0$. As a result, and using that $\frac{1}{n}\|\boldsymbol{\lambda}\|_2^2$ converges almost surely to a constant,

$$\frac{1}{n}\|\hat{\boldsymbol{d}}^1 - \boldsymbol{d}^1\|_2^2 = \frac{1}{n}\|\boldsymbol{Z}\hat{g}_t(\hat{\boldsymbol{c}}^1; \boldsymbol{v}, \boldsymbol{\theta}) - \boldsymbol{Z}\hat{g}_t(\boldsymbol{c}^1; \boldsymbol{v}, \boldsymbol{\theta}) + (\frac{1}{n}\langle \boldsymbol{\theta}, \hat{g}_1(\boldsymbol{c}^1; \boldsymbol{v}, \boldsymbol{\theta})\rangle - \gamma_1)\boldsymbol{\lambda}\|_2^2 \xrightarrow{\text{p}} 0.$$

Now assume that (40) holds for $1, 2, \ldots, t$. For the $(t+1)$-th iteration, we have

$$|\frac{1}{n}\langle \boldsymbol{\lambda}, \hat{f}_t(\boldsymbol{d}^1, \ldots, \boldsymbol{d}^t; \boldsymbol{u}, \boldsymbol{\lambda})\rangle - \frac{1}{n}\langle \boldsymbol{\lambda}, \hat{f}_t(\hat{\boldsymbol{d}}^1, \ldots, \hat{\boldsymbol{d}}^t; \boldsymbol{u}, \boldsymbol{\lambda})\rangle| \le \frac{L}{n}\|\boldsymbol{\lambda}\|_2 \sum_{s=1}^t \|\boldsymbol{d}^s - \hat{\boldsymbol{d}}^s\|_2 \xrightarrow{\text{p}} 0,$$

where L is a Lipschitz constant for $\hat{f}$. By (39), we have $\frac{1}{n}\langle \boldsymbol{\lambda}, \hat{f}_t(\hat{\boldsymbol{d}}^1, \ldots, \hat{\boldsymbol{d}}^t; \boldsymbol{u}, \boldsymbol{\lambda})\rangle \xrightarrow{\text{p}} \alpha_{t+1}$. As a result, we have $\frac{1}{n}\langle \boldsymbol{\lambda}, \hat{f}_t(\boldsymbol{d}^1, \ldots, \boldsymbol{d}^t; \boldsymbol{u}, \boldsymbol{\lambda})\rangle \xrightarrow{\text{p}} \alpha_{t+1}$. Furthermore, for any $1 \le s \le t$, we have

$$\frac{1}{n}\|\hat{f}_s(\boldsymbol{d}^1, \ldots, \boldsymbol{d}^s; \boldsymbol{u}, \boldsymbol{\lambda}) - \hat{f}_s(\hat{\boldsymbol{d}}^1, \ldots, \hat{\boldsymbol{d}}^s; \boldsymbol{u}, \boldsymbol{\lambda})\|_2^2 \le \frac{\hat{L}_t^2}{n}\sum_{i=1}^s \|\boldsymbol{d}^i - \hat{\boldsymbol{d}}^i\|_2^2 \xrightarrow{\text{p}} 0,$$

$$\frac{1}{n}\|\hat{g}_s(\boldsymbol{c}^1, \ldots, \boldsymbol{c}^s; \boldsymbol{v}, \boldsymbol{\theta}) - \hat{g}_s(\hat{\boldsymbol{c}}^1, \ldots, \hat{\boldsymbol{c}}^s; \boldsymbol{v}, \boldsymbol{\theta})\|_2^2 \le \frac{\hat{L}_t^2}{n}\sum_{i=1}^s \|\boldsymbol{c}^i - \hat{\boldsymbol{c}}^i\|_2^2 \xrightarrow{\text{p}} 0.$$

Again using that the maximal singular value of $\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}$ is $O_p(1)$, we have

$$\frac{1}{n}\|\boldsymbol{Z}^\mathsf{T}\hat{f}_t(\hat{\boldsymbol{d}}^1, \ldots, \hat{\boldsymbol{d}}^t; \boldsymbol{u}, \boldsymbol{\lambda}) - \boldsymbol{Z}^\mathsf{T}\hat{f}_t(\boldsymbol{d}^1, \ldots, \boldsymbol{d}^t; \boldsymbol{u}, \boldsymbol{\lambda})\|_2^2 \xrightarrow{\text{p}} 0.$$

As a result, we have

$$\begin{aligned}
&\frac{1}{n}\|\hat{\boldsymbol{c}}^{t+1} - \boldsymbol{c}^{t+1}\|_2^2 \\
= \;& \frac{1}{n}\|(\frac{1}{n}\langle \boldsymbol{\lambda}, \hat{f}_t(\boldsymbol{d}^1, \ldots, \boldsymbol{d}^t; \boldsymbol{u}, \boldsymbol{\lambda}) - \alpha_{t+1})\boldsymbol{\theta} + \boldsymbol{Z}^\mathsf{T}(\hat{f}_t(\hat{\boldsymbol{d}}^1, \ldots, \hat{\boldsymbol{d}}^t; \boldsymbol{u}, \boldsymbol{\lambda}) - \hat{f}_t(\boldsymbol{d}^1, \ldots, \boldsymbol{d}^t; \boldsymbol{u}, \boldsymbol{\lambda})) - \\
& \sum_{s=1}^t \xi_{t,s}(\hat{g}_s(\hat{\boldsymbol{c}}^1, \ldots, \hat{\boldsymbol{c}}^s; \boldsymbol{v}, \boldsymbol{\theta}) - \hat{g}_s(\boldsymbol{c}^1, \ldots, \boldsymbol{c}^s; \boldsymbol{v}, \boldsymbol{\theta}))\|_2^2 \xrightarrow{\text{p}} 0.
\end{aligned}$$

Similarly, we have

$$|\frac{1}{n}\langle \boldsymbol{\theta}, \hat{g}_{t+1}(\hat{\boldsymbol{c}}^1, \ldots, \hat{\boldsymbol{c}}^{t+1}; \boldsymbol{v}, \boldsymbol{\theta})\rangle - \frac{1}{n}\langle \boldsymbol{\theta}, \hat{g}_{t+1}(\boldsymbol{c}^1, \ldots, \boldsymbol{c}^{t+1}; \boldsymbol{v}, \boldsymbol{\theta})\rangle| \leq \frac{L}{n}\|\boldsymbol{\theta}\|_2 \sum_{s=1}^{t+1} \|\hat{\boldsymbol{c}}^{t+1} - \boldsymbol{c}^{t+1}\|_2 \xrightarrow{\text{p}} 0,$$

where L is a Lipschitz constant for $\hat{g}_{t+1}$. By (39), we have that $\frac{1}{n}\langle \boldsymbol{\theta}, \hat{g}_{t+1}(\hat{\boldsymbol{c}}^1, \ldots, \hat{\boldsymbol{c}}^{t+1}; \boldsymbol{v}, \boldsymbol{\theta})\rangle \xrightarrow{\text{p}} \gamma_{t+1}$. As a result, we have that $\frac{1}{n}\langle \boldsymbol{\theta}, \hat{g}_{t+1}(\boldsymbol{c}^1, \ldots, \boldsymbol{c}^{t+1}; \boldsymbol{v}, \boldsymbol{\theta})\rangle \xrightarrow{\text{p}} \gamma_{t+1}$. Furthermore, for any $1 \leq s \leq t$, we have

$$\frac{1}{n}\|\hat{f}_s(\boldsymbol{d}^1, \ldots, \boldsymbol{d}^s; \boldsymbol{u}, \boldsymbol{\lambda}) - \hat{f}_s(\hat{\boldsymbol{d}}^1, \ldots, \hat{\boldsymbol{d}}^s; \boldsymbol{u}, \boldsymbol{\lambda})\|_2^2 \leq \frac{L^2}{n}\sum_{i=1}^{s} \|\boldsymbol{d}^i - \hat{\boldsymbol{d}}^i\|_2^2 \xrightarrow{\text{p}} 0.$$

Also, for any $1 \leq s \leq t+1$, we have

$$\frac{1}{n}\|\hat{g}_s(\boldsymbol{c}^1, \ldots, \boldsymbol{c}^s; \boldsymbol{v}, \boldsymbol{\theta}) - \hat{g}_s(\hat{\boldsymbol{c}}^1, \ldots, \hat{\boldsymbol{c}}^s; \boldsymbol{v}, \boldsymbol{\theta})\|_2^2 \leq \frac{L^2}{n}\sum_{i=1}^{s} \|\boldsymbol{c}^i - \hat{\boldsymbol{c}}^i\|_2^2 \xrightarrow{\text{p}} 0.$$

Then $\frac{1}{n}\|\boldsymbol{Z}\hat{g}_{t+1}(\hat{\boldsymbol{c}}^1, \ldots, \hat{\boldsymbol{c}}^{t+1}; \boldsymbol{v}, \boldsymbol{\theta}) - \boldsymbol{Z}\hat{g}_{t+1}(\boldsymbol{c}^1, \ldots, \boldsymbol{c}^{t+1}; \boldsymbol{v}, \boldsymbol{\theta})\|_2^2 \xrightarrow{\text{p}} 0$. As a result, we have

$$\begin{aligned}
&\frac{1}{n}\|\hat{\boldsymbol{d}}^{t+1} - \boldsymbol{d}^{t+1}\|_2^2 \\
&\quad = \frac{1}{n}\|(\frac{1}{n}\langle \boldsymbol{\theta}, \hat{g}_{t+1}(\boldsymbol{c}^1, \ldots, \boldsymbol{c}^{t+1}; \boldsymbol{v}, \boldsymbol{\theta})\rangle - \gamma_{t+1})\boldsymbol{\lambda} \\
&\quad + \boldsymbol{Z}(\hat{g}_{t+1}(\hat{\boldsymbol{c}}^1, \ldots, \hat{\boldsymbol{c}}^{t+1}; \boldsymbol{v}, \boldsymbol{\theta}) - \hat{g}_{t+1}(\boldsymbol{c}^1, \ldots, \boldsymbol{c}^{t+1}; \boldsymbol{v}, \boldsymbol{\theta})) \\
&\quad - \sum_{s=0}^{t} \zeta_{t,s}(\hat{f}_s(\hat{\boldsymbol{d}}^1, \ldots, \hat{\boldsymbol{d}}^s; \boldsymbol{u}, \boldsymbol{\lambda}) - \hat{f}_s(\boldsymbol{d}^1, \ldots, \boldsymbol{d}^s; \boldsymbol{u}, \boldsymbol{\lambda}))\|_2^2 \xrightarrow{\text{p}} 0.
\end{aligned}$$

Thus, we have proved (40). Therefore, for all pseudo-Lipschitz functions ψ of order 2, we have that there exists a numerical constant C such that

$$\begin{aligned}
&\left|\frac{1}{p}\sum_{j=1}^{p} \psi(c_j^1 + \alpha_1\theta_j, \ldots, c_j^t + \alpha_t\theta_j, v_j, \theta_j) - \frac{1}{p}\sum_{j=1}^{p} \psi(\hat{c}_j^1 + \alpha_1\theta_j, \ldots, \hat{c}_j^t + \alpha_t\theta_j, v_j, \theta_j)\right| \\
&\qquad \leq L_\psi(1 + \sum_{s=1}^{t}\|\boldsymbol{a}^s\|_2 + \|\boldsymbol{\theta}\|_2 + \|\boldsymbol{v}\|_2)\sum_{s=1}^{t}\|\hat{\boldsymbol{c}}^s - \boldsymbol{c}^s\|_2 \xrightarrow{\text{p}} 0.
\end{aligned}$$

By (39),

$$\frac{1}{p}\sum_{j=1}^{p} \psi(\hat{c}_j^1 + \alpha_1\theta_j, \ldots, \hat{c}_j^t + \alpha_t\theta_j, v_j, \theta_j) \xrightarrow{\text{p}} \mathbb{E}[\psi(\boldsymbol{\alpha}_1\boldsymbol{\Theta} + \boldsymbol{Z}^1, \ldots, \boldsymbol{\alpha}_t\boldsymbol{\Theta} + \boldsymbol{Z}^t, \boldsymbol{V}, \boldsymbol{\Theta})].$$

Therefore, $\frac{1}{p}\sum_{j=1}^{p} \psi(a_j^1, \ldots, a_j^t, v_j, \theta_j) \xrightarrow{\text{p}} \mathbb{E}[\psi(\boldsymbol{\alpha}_1\boldsymbol{\Theta} + \boldsymbol{Z}^1, \ldots, \boldsymbol{\alpha}_t\boldsymbol{\Theta} + \boldsymbol{Z}^t, \boldsymbol{V}, \boldsymbol{\Theta})]$. Similarly, we can show that $\frac{1}{n}\sum_{i=1}^{n} \psi(\boldsymbol{b}_i^1, \ldots, \boldsymbol{b}_i^t, \boldsymbol{u}_i, \boldsymbol{\lambda}_i) \xrightarrow{\text{p}} \mathbb{E}[\psi(\boldsymbol{\gamma}_1\boldsymbol{\Lambda} + \tilde{\boldsymbol{Z}}^1, \ldots, \boldsymbol{\gamma}_t\boldsymbol{\Lambda} + \tilde{\boldsymbol{Z}}^t, \boldsymbol{U}, \boldsymbol{\Lambda})]$. Thus, we have finished the proof.

B.3. The AMP change of variables

To prove Lemma 6, all that remains is to show that for any GFOM (1), there exist matrices $(\boldsymbol{\xi}_{t,s})$, $(\boldsymbol{\zeta}_{t,s})$ such that the change of variables in Eqs. (34) generates an iteration (35) which is an AMP iteration. That is, in addition to satisfying Eq. (34), the matrices $(\boldsymbol{\xi}_{t,s})$, $(\boldsymbol{\zeta}_{t,s})$ and functions (f_t), (g_t) satisfy Eqs. (37) and (38) 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 (34) with either (37) or (38). Specifically, in the high-dimensional regression model, we place (34a) before the first line of (37) and (34b) before the fourth line of (37). To illustrate: we first define $f_0(y,u) = F_0(y;u)$. Then we get $\alpha_1, T_{1,1}, T_{1,2}, T_{2,2}$ from (37). Next we get $\varphi_1(a,v) = a + F_0^{(2)}(v)$, and $g_1(a^1;v) = G_1^{(1)}(\varphi_1(a^1;v);v)$ from (34b). We then get $\zeta_{1,0}$, $\Sigma_{0,1}$, $\Sigma_{1,1}$ from (37). Then we get f_1, ϕ_1 from (34a). We can then get $\xi_{1,1}$, α_2, and $T_{s,2}$ for $1 \le s \le 2$. This recursion continues. 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 (34) and (37). 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 (34a) before the first line of (38) and (34b) before the fourth line of (38).

The proof of Lemma 6 is complete.

Appendix C. Proof of state evolution for message passing (Lemma 7)

In this section, we prove Lemma 7. 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.

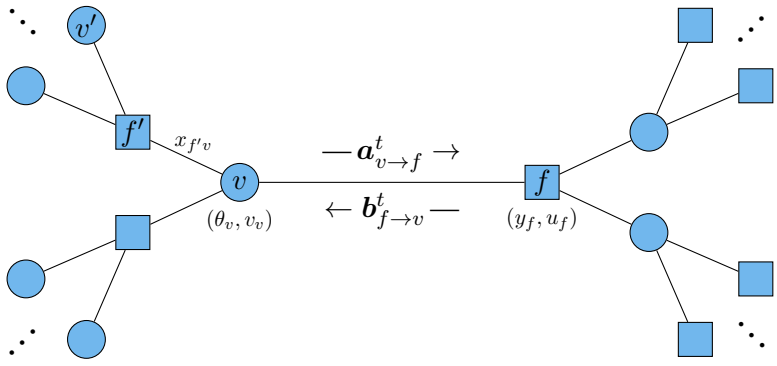

Figure 1: The computation tree $\mathcal{T}$.

The computation tree with $n = p = 3$ is shown in Figure 1. We have labeled some of the nodes or edges with some of the random variables or messages associated to them. 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 $f' \in \partial v \backslash f$. We define $\mathcal{T}_{f\to v}, \mathcal{V}_{f\to v}, \mathcal{F}_{f\to v}, \mathcal{E}_{f\to v}$

similarly. For example, Figure 2 displays $\mathcal{T}_{v\to f}$ for the computation tree shown in Figure 1. 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.

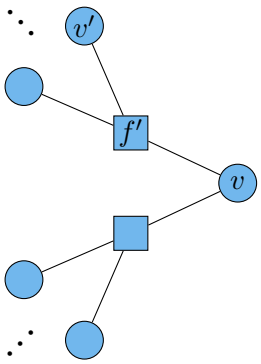

Figure 2: The sub-tree $\mathcal{T}_{v\to f}$.

C.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 (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 $\mu_{\Theta,V}, \mu_{W,U} \in \mathscr{P}_4(\mathbb{R}^2)$.[7] 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 v by f (including, by convention, f).

Gaussian message passing algorithms are defined by sequences of Lipschitz functions $(\tilde{f}_t : \mathbb{R}^{t+3} \to \mathbb{R})_{t\ge 0}$, $(\tilde{g}_t : \mathbb{R}^{t+2} \to \mathbb{R})_{t\ge 0}$. We initialize the indexing differently with Gaussian message passing algorithms than with the message passing algorithms in Section 5 in anticipation of notational simplifications that will occur later. For every pair of neighboring nodes v, f, we generate sequences of messages $(\tilde{a}^t_{v\to f})_{t\ge 1}$, $(\tilde{q}^t_{v\to f})_{t\ge 0}$, $(\tilde{b}^t_{f\to v})_{t\ge 0}$, $(\tilde{r}^t_{f\to v})_{t\ge 0}$ according to the iteration

$$\tilde{a}^{t+1}_{v\to f} = \sum_{f'\in\partial v\setminus f} z_{f'v}\tilde{r}^t_{f'\to v}, \qquad \tilde{r}^t_{f\to v} = \tilde{f}_t(\tilde{b}^0_{f\to v}, \dots, \tilde{b}^t_{f\to v}; w_f, u_f), \tag{41a}$$

$$\tilde{b}^t_{f\to v} = \sum_{v'\in\partial f\setminus v} z_{fv'}\tilde{q}^t_{v'\to f}, \qquad \tilde{q}^t_{v\to f} = \tilde{g}_t(\tilde{a}^1_{v\to f}, \dots, \tilde{a}^t_{v\to f}; \theta_v, v_v), \tag{41b}$$

7. 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.

with initialization $\tilde{q}^0_{v\to f} = g_0(\theta_v, v_v)$. For $t \geq 0$, define the node beliefs

$$\tilde{a}_v^{t+1} = \sum_{f\in\partial v} z_{fv}\tilde{r}^t_{f\to v}, \qquad \tilde{b}_f^t = \sum_{v\in\partial f} z_{fv}\tilde{q}^t_{v\to f}. \tag{42}$$

To compactify notation, denote $\tilde{\boldsymbol{a}}_v^t = (\tilde{a}_v^1, \dots, \tilde{a}_v^t)^\mathsf{T}$, and likewise for $\tilde{\boldsymbol{a}}^t_{v\to f}$, $\tilde{\boldsymbol{q}}^t_{v\to f}$, $\tilde{\boldsymbol{b}}_f^t$, $\tilde{\boldsymbol{b}}^t_{f\to v}$, $\tilde{\boldsymbol{r}}^t_{f\to v}$ (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 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

$$\Sigma_{s,s'} = \mathbb{E}[\tilde{g}_s(\tilde{\boldsymbol{A}}^s; \Theta, V)\tilde{g}_{s'}(\tilde{\boldsymbol{A}}^{s'}; \Theta, V)], \quad T_{s+1,s'+1} = \mathbb{E}[\tilde{f}_s(\tilde{\boldsymbol{B}}^s; W, U)\tilde{f}_{s'}(\tilde{\boldsymbol{B}}^{s'}; W, U)], \tag{43}$$

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 14 *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*

$$(\tilde{\boldsymbol{a}}_v^t, \theta_v, v_v) \overset{\mathrm{W}}{\to} \mathsf{N}(\boldsymbol{0}_t, \boldsymbol{T}_{[1:t]}) \otimes \mu_{\Theta,V} \;\; \textit{and} \;\; (\tilde{\boldsymbol{a}}^t_{v\to f}, \theta_v, v_v) \overset{\mathrm{W}}{\to} \mathsf{N}(\boldsymbol{0}_t, \boldsymbol{T}_{[1:t]}) \otimes \mu_{\Theta,V}, \tag{44a}$$

$$(\tilde{\boldsymbol{b}}_f^t, w_f, u_f) \overset{\mathrm{W}}{\to} \mathsf{N}(\boldsymbol{0}_{t+1}, \boldsymbol{\Sigma}_{[0:t]}) \otimes \mu_{W,U} \;\; \textit{and} \;\; (\tilde{\boldsymbol{b}}^t_{f\to v}, w_f, u_f) \overset{\mathrm{W}}{\to} \mathsf{N}(\boldsymbol{0}_{t+1}, \boldsymbol{\Sigma}_{[0:t]}) \otimes \mu_{W,U}. \tag{44b}$$

Further, all the random variables in the preceding displays have bounded fourth moments and $\mathbb{E}[\|\tilde{\boldsymbol{a}}_v^t - \tilde{\boldsymbol{a}}^t_{v\to f}\|^2] \to 0$ and $\mathbb{E}[\|\tilde{\boldsymbol{b}}_f^t - \tilde{\boldsymbol{b}}^t_{f\to v}\|^2] \to 0$.

The analysis of message passing on the tree is facilitated by the many independence relationships between messages, which follow from the following lemma.

Lemma 15 *For all $(f, v) \in \mathcal{E}$ and all t, the messages $\tilde{r}^t_{f\to v}, \tilde{b}^t_{f\to v}$ are $\mathcal{T}_{f\to v}$-measurable, and the messages $\tilde{q}^t_{v\to f}, \tilde{a}^t_{a\to f}$ is $\mathcal{T}_{v\to f}$-measurable.*

Proof [Lemma 15] The proof is by induction. The base case is that $\tilde{q}^0_{v\to f} = g_0(\theta_v, v_v)$ is $\mathcal{T}_{v\to f}$-measurable. Then, if $\tilde{q}^s_{v\to f}$ are $\mathcal{T}_{v\to f}$-measurable and $\tilde{b}^s_{f\to v}$ are $\mathcal{T}_{f\to v}$-measurable for $0 \leq s \leq t$ and all $(f, v) \in \mathcal{E}$, then $\tilde{b}^t_{f\to v}, \tilde{r}^t_{f\to v}$ are $\mathcal{T}_{f\to v}$-measurable by (41). Similarly, if $\tilde{r}^s_{f\to v}$ are $\mathcal{T}_{f\to v}$-measurable and $\tilde{a}^s_{v\to f}$ are $\mathcal{T}_{v\to r}$-measurable for $0 \leq s \leq t$ and all $(f, v) \in \mathcal{E}$, then $\tilde{a}^{t+1}_{f\to v}, \tilde{r}^{t+1}_{f\to v}$ are $\mathcal{T}_{v\to f}$-measurable by (41). The induction is complete. ∎

We now prove Lemma 14.

Proof [Lemma 14] The proof is by induction.

Base case: $(\theta_v, v_f) \overset{\mathrm{W}}{\to} \mu_{\Theta,V}$.

This is the exact distribution in finite samples by assumption.

Inductive step 1: Eq. (44a) *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.* (44b) *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 (42), 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,

$$\tilde{\boldsymbol{b}}_f^t \mid \sigma((\mathcal{T}_{v\to f})_{v\in\partial f}) \sim \mathsf{N}(\mathbf{0}_{t+1}, \widehat{\boldsymbol{\Sigma}}_{[0:t]}).$$

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 9, 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 10, and (43). The terms in the sum defining $\widehat{\boldsymbol{\Sigma}}_{[0:t]}$ are mutually independent by Lemma 15 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, $\widehat{\boldsymbol{\Sigma}}_{[0:t]} \overset{L_1}{\to} \boldsymbol{\Sigma}_{[0:t]}$, whence by Slutsky's theorem, $\tilde{\boldsymbol{b}}_f^t \overset{\mathrm{d}}{\to} \mathsf{N}(\mathbf{0}_{t+1}, \boldsymbol{\Sigma}_{[0:t]})$. 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 (Villani, 2010, Theorem 6.9), so $\tilde{\boldsymbol{b}}_f^t \overset{\mathrm{W}}{\to} \mathsf{N}(\mathbf{0}_{t+1}, \boldsymbol{\Sigma}_{[0:t]})$.

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. (44) *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.* (44) *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 14 follows. ∎

C.2. Message passing in the high-dimensional regression model

We prove Lemma 7 for the high-dimensional regression model by showing that the iteration (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 (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 $(\tilde{a}^t_{v\to f})_{t\geq 1}$, $(\tilde{a}^t_v)_{t\geq 1}$, $(\tilde{q}^t_{v\to f})_{t\geq 0}$, $(\tilde{b}^t_{f\to v})_{t\geq 0}$, $(\tilde{b}^t_f)_{t\geq 0}$, $(\tilde{r}^t_{f\to v})_{t\geq 0}$ via the Gaussian message passing algorithm (41) 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 (43) and (37) are equivalent. Thus, Lemma 14 holds for the iterates of this Gaussian message passing algorithm with the $\boldsymbol{T}_{[1:t]}$, $\boldsymbol{\Sigma}_{[0:t]}$ defined by (37).

We claim that for fixed $s \geq 1$, as $n \to \infty$ we have

$$\mathbb{E}[(\alpha_s\theta_v + \tilde{a}^s_{v\to f} - a^s_{v\to f})^2] \to 0 \text{ and } \mathbb{E}[(\tilde{b}^s_{f\to v} - b^s_{f\to v})^2] \to 0, \tag{45a}$$

and

$$\mathbb{E}[(a^s_{v\to f})^4] \text{ and } \mathbb{E}[(b^s_{f\to v})^4] \text{ are uniformly bounded with respect to } n, \tag{45b}$$

where (α_s) are defined by (37). These are the same coefficients appearing in the AMP state evolution (Lemma 6), as claimed. We show (45) by induction. There is no base case because the inductive steps work for $t = 0$ as written.

Inductive step 1: If (45) *holds for* $1 \leq s \leq t$, *then* (45a) *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

$$|\tilde{f}_t(\tilde{\boldsymbol{b}}^t_{f'\to v}; w_{f'}, u_{f'}) - \tilde{f}_t(\tilde{b}^0_{f'\to v}, \boldsymbol{b}^t_{f'\to v}; w_{f'}, u_{f'})| \leq L\sum_{s=1}^t |\tilde{b}^s_{f'\to v} - b^s_{f'\to v}|,$$

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,

$$\begin{aligned}
\mathbb{E}[\mathsf{I}^2] &= \frac{n-1}{n}\mathbb{E}[(\tilde{f}_t(\tilde{\boldsymbol{b}}^t_{f'\to v}; w_{f'}, u_{f'}) - \tilde{f}_t(\tilde{b}^0_{f'\to v}, \boldsymbol{b}^t_{f'\to v}; w_{f'}, u_{f'}))^2] \\
&\leq \frac{L^2(n-1)t}{n}\sum_{s=1}^t \mathbb{E}[(\tilde{b}^s_{f'\to v} - b^s_{f'\to v})^2] \to 0,
\end{aligned}$$

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) (Stein, 1981) to get

$$\begin{aligned}
&\mathbb{E}[\alpha_{t+1}\theta_v + \mathrm{II} \;\big|\; \theta_v, \sigma((\mathcal{T}_{v'' \to f'})_{v'' \in \partial f' \setminus v})] \\
&= \alpha_{t+1}\theta_v + (n-1)\mathbb{E}\big[z_{f'v}(\tilde{f}_t(\tilde{b}^0_{f' \to v}, \boldsymbol{b}^t_{f' \to v}; w_{f'}, u_{f'}) - \tilde{f}_t(\tilde{b}^0_{f'}, \boldsymbol{b}^t_{f' \to v}; w_{f'}, u_{f'})) \;\big|\; \theta_v\big] \\
&= \theta_v \left(\alpha_{t+1} - \frac{n-1}{n}\mathbb{E}[\partial_{\tilde{b}^0}\tilde{f}_t(\tilde{b}^0_{f'}, \boldsymbol{b}^t_{f' \to v}; w_{f'}, u_{f'}) \;\big|\; \theta_v]\right),
\end{aligned}$$

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 (Evans and Gariepy, 2015, 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,

$$(\tilde{b}^0_{f' \to v} + z_{f'v}\theta_v, \boldsymbol{b}^t_{f' \to v})|\theta_v, \sigma((\mathcal{T}_{v'' \to f'})_{v'' \in \partial f' \setminus v}) \stackrel{\mathrm{d}}{=} \mathsf{N}(\mathbf{0}, \widehat{\boldsymbol{\Sigma}}),$$

where we define $\widehat{\boldsymbol{\Sigma}} \in \mathbb{R}^{(t+1)\times(t+1)}$ by

$$\widehat{\Sigma}_{0,0} = \frac{1}{n}\sum_{v' \in \partial f'} \theta^2_{v'} \;\text{ and }\; \widehat{\Sigma}_{s,s'} = \frac{1}{n}\sum_{v' \in \partial f' \setminus v} q^s_{v' \to f'} q^{s'}_{v' \to f'} \;\text{ for } s \geq 1 \text{ or } s' \geq 1,$$

and for the purposes of the preceding display we set $q^0_{v' \to f'} = \theta_{v'}$. By the Lipschitz continuity of the functions (g_s), Lemmas 9 and 10, 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 (45b) 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 $\mathbb{E}[\partial_{\tilde{b}^0}\tilde{f}_t(\tilde{b}^0_{f'}, \boldsymbol{b}^t_{f' \to v}; w_{f'}, u_{f'})|\theta_v, (\mathcal{T}_{v'' \to f'})_{v'' \in \partial f' \setminus v}] = \mathbb{E}[\partial_{\tilde{b}^0}\tilde{f}_t(\widehat{\boldsymbol{\Sigma}}^{1/2}\boldsymbol{Z}; W, U)]$, where on the right-hand side the expectation is with respect to $(W, U) \sim \mu_{W,U}$ and $\boldsymbol{Z} \sim \mathsf{N}(\mathbf{0}_{t+1}, \boldsymbol{I}_{t+1})$ 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 (37), we conclude $\mathbb{E}[\partial_{\tilde{b}^0}\tilde{f}_t(\tilde{b}^0_{f'}, \boldsymbol{b}^t_{f' \to v}; w_{f'}, u_{f'})|\theta_v, (\mathcal{T}_{v'' \to f'})_{v'' \in \partial f' \setminus v}] \xrightarrow{\mathrm{p}} \alpha_{t+1}$. Then, by dominated convergence, $\mathbb{E}[\alpha_{t+1}\theta_v + \mathrm{II} \;\big|\; \theta_v] \xrightarrow{L_2} 0$. Moreover, because the terms in the sum defining II are mutually independent given θ_v

$$\begin{aligned}
&\mathrm{Var}(\alpha_{t+1}\theta_v + \mathrm{II} \mid \theta_v) \\
&\qquad \leq (n-1)\mathbb{E}\left[z^2_{f'v}(\tilde{f}_t(\tilde{b}^0_{f' \to v}, \boldsymbol{b}^t_{f' \to v}; w_{f'}, u_{f'}) - \tilde{f}_t(\tilde{b}^0_{f'}, \boldsymbol{b}^t_{f' \to v}; w_{f'}, u_{f'}))^2 \mid \theta_v\right] \\
&\qquad \leq L^2(n-1)\mathbb{E}[z^4_{f'v}\theta^2_v \mid \theta_v] \leq 3\theta^2_v/n,
\end{aligned}$$

where L is a Lipschitz constant of $\tilde{f}_t$. We conclude that $\mathbb{E}[\mathrm{Var}(\alpha_{t+1}\theta_v + \mathrm{II} \mid \theta_v)] \to 0$. Combined with $\mathbb{E}[\alpha_{t+1}\theta_v + \mathrm{II} \;\big|\; \theta_v] \xrightarrow{L_2} 0$, we get $\mathrm{Var}(\alpha_{t+1}\theta_v + \mathrm{II}) = \mathrm{Var}(\mathbb{E}[\alpha_{t+1}\theta_v + \mathrm{II}|\theta_v]) + \mathbb{E}[\mathrm{Var}(\alpha_{t+1}\theta_v +$

$\mathsf{II}|\theta_v)] \to 0$, so that $\alpha_{t+1}\theta_v + \mathsf{II} \overset{L_2}{\to} 0$. Combining $\mathsf{I} \overset{L_2}{\to} 0$ and $\alpha_{t+1}\theta_v + \mathsf{II} \overset{L_2}{\to} 0$ gives $\mathbb{E}[(\alpha_{t+1}\theta_v + \tilde{a}^{t+1}_{v\to f} - a^{t+1}_{v\to f})^2] \to 0$, as desired.

We now expand

$$\tilde{b}^{t+1}_{f\to v} - b^{t+1}_{f\to v} = \sum_{v'\in\partial f\setminus v} z_{fv'}(g_t(\boldsymbol{\alpha}_{t+1}\theta_{v'} + \tilde{\boldsymbol{a}}^{t+1}_{v'\to f}; v_{v'}) - g_t(\boldsymbol{a}^{t+1}_{v'\to f}; v_{v'})).$$

The terms in this sum are mutually independent, and $\tilde{\boldsymbol{a}}^{t+1}_{v'\to f}, \boldsymbol{a}^{t+1}_{v'\to f}, \theta_{v'}$ are independent of $z_{f'v}$. Thus,

$$\begin{aligned}\mathbb{E}[(\tilde{b}^{t+1}_{f\to v} - b^{t+1}_{f\to v})^2] &= \frac{p-1}{n}\mathbb{E}[(g_t(\boldsymbol{\alpha}_{t+1}\theta_{v'} + \tilde{\boldsymbol{a}}^{t+1}_{v'\to f}; v_{v'}) - g_t(\boldsymbol{a}^{t+1}_{v'\to f}; v_{v'}))^2] \\ &\le \frac{L^2(p-1)(t+1)}{n}\sum_{s=1}^{t+1}\mathbb{E}[(\alpha_s\theta_{v'} + \tilde{a}^s_{v'\to f} - a^s_{v'\to f})^2] \to 0.\end{aligned}$$

This completes the proof of (45a) at $s = t+1$.

Inductive step 2: If (45) *holds for* $1 \le s \le t$*, then* (45b) *holds for* $s = t+1$.

By Lipschitz continuity,

$$\left|a^{t+1}_{v\to f} - \sum_{f'\in\partial v\setminus f} z_{f'v}\tilde{f}_t(\tilde{b}^0_{f'\to v}, \boldsymbol{b}^t_{f'\to v}, u_{f'}, w_{f'})\right| \le L|\theta_v| \sum_{f'\in\partial v\setminus f}|z_{f'v}|,$$

where L is a Lipschitz constant for $\tilde{f}_t$. The right-hand side has bounded fourth moment, so we only need to show that the sum in the previous display has bounded fourth moment. The quantity $\tilde{f}_t(\tilde{b}^0_{f'\to v}, \boldsymbol{b}^t_{f'\to v}, u_{f'}, w_{f'})$ 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 (45b) at $s = t+1$.

The induction is complete, and (45a) holds for all $s \ge 1$. Lemma 7 follows by combining Lemma 14 and Eq. (45a).

C.3. Message passing in the low-rank matrix estimation model

Like in the preceding section, we prove Lemma 7 for the low-rank matrix estimation model by showing that the iteration (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 (19):

$$\begin{aligned}\tilde{f}_t(\tilde{b}^0, \cdots, \tilde{b}^t, w, u) &:= f_t(\tilde{b}^1 + \gamma_1 w, \cdots, \tilde{b}^t + \gamma_t w; 0, u), \\ \tilde{g}_t(\tilde{a}^1, \cdots, \tilde{a}^t; \theta, v) &:= g_t(\tilde{a}^1 + \alpha_1\theta, \cdots, \tilde{a}^t + \alpha_t\theta; v).\end{aligned}$$

Note that here $\tilde{f}_t$ does not depend on $\tilde{b}^0$, and we may define $\tilde{g}_0$ arbitrarily without affecting later iterates.[8] Define $(\tilde{a}^t_{v\to f})_{t\ge 1}$, $(\tilde{a}^t_v)_{t\ge 1}$, $(\tilde{q}^t_{v\to f})_{t\ge 0}$, $(\tilde{b}^t_{f\to v})_{t\ge 0}$, $(\tilde{b}^t_f)_{t\ge 0}$, $(\tilde{r}^t_{f\to v})_{t\ge 0}$ via the Gaussian message passing algorithm (41) 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 (43) and (38) are equivalent. Thus, Lemma 14 holds for the iterates of this Gaussian message passing algorithm with the $\boldsymbol{T}_{[1:t]}$, $\boldsymbol{\Sigma}_{[0:t]}$ defined by (38).

We claim that for fixed $s \ge 1$, as $n \to \infty$ we have

$$\mathbb{E}[(\alpha_s\theta_v + \tilde{a}^s_{v\to f} - a^s_{v\to f})^2] \to 0 \quad \text{and} \quad \mathbb{E}[(\gamma_s\lambda_f + \tilde{b}^s_{f\to v} - b^s_{f\to v})^2] \to 0, \tag{46a}$$

and

$$\mathbb{E}[\theta_v^2(a^s_{v\to f})^2] \quad \text{and} \quad \mathbb{E}[\lambda_f^2(b^s_{f\to v})^2] \quad \text{are bounded for fixed } s. \tag{46b}$$

We show this by induction. There is no base case because the inductive step works for $t = 0$ as written.

Inductive step: If (46) *holds for* $1 \le s \le t$*, then* (46) *holds for* $s = t + 1$.

We expand

$$\begin{aligned}\alpha_{t+1}\theta_v + \tilde{a}^{t+1}_{v\to f} - a^{t+1}_{v\to f} &= \alpha_{t+1}\theta_v + \sum_{f'\in\partial v\setminus f} z_{f'v}(f_t(\tilde{\boldsymbol{b}}^t_{f'\to v} + \boldsymbol{\gamma}_t\lambda_{f'}; 0, u_{f'}) - f_t(\boldsymbol{b}^t_{f'\to v}; 0, u_{f'})) \\ &\quad - \frac{1}{n}\theta_v \sum_{f'\in\partial v\setminus f} \lambda_{f'} f_t(\boldsymbol{b}^t_{f'\to v}; 0, u_{f'}) \\ &=: \alpha_{t+1}\theta_v + \mathsf{I} + \mathsf{II},\end{aligned}$$

where $\tilde{\boldsymbol{b}}^t_{f'\to v} = (\tilde{b}^1_{f'\to v}, \dots, \tilde{b}^t_{f'\to v})$ and $\boldsymbol{\gamma}_t = (\gamma_1, \dots, \gamma_t)$ (note that $\tilde{b}^0_{f'\to v}$ is excluded, which differs from the notation used in the proof of Lemma 7).

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,

$$\begin{aligned}\mathbb{E}[\mathsf{I}^2] &= \frac{n-1}{n}\mathbb{E}[(f_t(\tilde{\boldsymbol{b}}^t_{f'\to v} + \boldsymbol{\gamma}_t\lambda_{f'}; 0, u_{f'}) - f_t(\boldsymbol{b}^t_{f'\to v}; 0, u_{f'})^2] \\ &\le \frac{L^2(n-1)t}{n}\sum_{s=1}^t \mathbb{E}[(\tilde{b}^s_{f'\to v} + \gamma_s\lambda_{f'} - b^s_{f'\to v})^2] \to 0,\end{aligned}$$

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 14,

$$(\boldsymbol{b}^t_{f'\to v}, \lambda_{f'}, u_{f'}) \overset{\mathrm{W}}{\to} (\boldsymbol{\gamma}_t\Lambda + \tilde{B}^t, \Lambda, U),$$

where $(\Lambda, U) \sim \mu_{\Lambda,U}$ and $\tilde{B}^t \sim \mathsf{N}(\boldsymbol{0}_t, \boldsymbol{\Sigma}_{[1:t]})$ 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 9, we have $\mathbb{E}[\lambda_{f'} f_t(\boldsymbol{b}^t_{f'\to v}; 0, u_{f'})] \to \alpha_{t+1}$

8. The iterate $\tilde{b}^0$ only played a role in approximating the high-dimensional regression message passing algorithm by a Gaussian message passing algorithm.

by Lemma 10 and the state evolution recursion (38). Moreover, because f_t is Lipschitz, for some constant C

$$\begin{aligned}\mathbb{E}[\lambda_{f'}^2 f_t(\boldsymbol{b}_{f'\to v}^t; 0, u_{f'})^2] &\le C\mathbb{E}\left[\lambda_{f'}^2\left(1+\sum_{s=1}^t (b_{f'\to v}^s)^2 + u_{f'}^2\right)\right] \\ &= C\left(\mathbb{E}[\lambda_{f'}^2] + \sum_{s=1}^t \mathbb{E}[\lambda_{f'}^2 (b_{f'\to v}^s)^2] + \mathbb{E}[\lambda_{f'}^2 u_{f'}^2]\right),\end{aligned}$$

which is 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

$$\frac{1}{n}\sum_{f'\in\partial v\setminus f} \lambda_{f'} f_t(\boldsymbol{b}_{f'\to v}^t; 0, u_{f'}) \xrightarrow{L_2} \alpha_{t+1}.$$

Because θ_v is independent of this sum and has bounded second moment, we conclude that

$$\alpha_{t+1}\theta_v + \mathsf{II} = \theta_v\left(\alpha_{t+1} - \frac{1}{n}\sum_{f'\in\partial v\setminus f} \lambda_{f'} f_t(\boldsymbol{b}_{f'\to v}^t; 0, u_{f'})\right) \xrightarrow{L_2} 0.$$

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 $\mathbb{E}[(\alpha_{t+1}\theta_v + \tilde{a}_{v\to f}^{t+1} - a_{v\to f}^{t+1})^2] \to 0$ and $\mathbb{E}[\theta_v^2(\alpha_{t+1}\theta_v + \tilde{a}_{v\to f}^{t+1} - a_{v\to f}^{t+1})^2]$ is bounded. Because θ_v is independent of $\tilde{a}_{v\to f}^{t+1}$, the term $\mathbb{E}[\theta_v^2(\tilde{a}_{v\to f}^{t+1})^2]$ is bounded, so also $\mathbb{E}[\theta_v^2(a_{v\to f}^{t+1})^2]$ is bounded, as desired.

The argument establishing that $\mathbb{E}[(\gamma_{t+1}\lambda_f + \tilde{b}_{f\to v}^{t+1} - b_{f\to v}^{t+1})^2] \to 0$ and that $\mathbb{E}[\lambda_f^2 (b_{f\to v}^{t+1})^2]$ is bounded is equivalent. The induction is complete, and (46) holds for all s.

Lemma 7 follows by combining Lemma 14 and Eq. (46).

Appendix D. Proof of information-theoretic lower bounds on the computation tree (Lemma 8)

In this section, we prove Lemma 8 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 included. 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 model, it is the σ-algebra generated 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 additionally 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.

D.1. Information-theoretic lower bound in the high-dimensional regression model

In this section, we prove Lemma 8 in the high-dimensional regression model.

Note that the conditions on the conditional density in assumption R4 are equivalent to positivity, boundedness, and the existence of finite, non-negative constants q'_k such that $\frac{|\partial^k_x p(y|x)|}{p(y|x)}\le q'_k$ 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

$$\frac{|\partial^k_x\mathbb{E}[p(y|x+A)]|}{\mathbb{E}[p(y|x+A)]}\le\int\frac{|\partial^k_x p(y|x+a)|}{p(y|x+a)}\frac{p(y|x+a)}{\mathbb{E}[p(y|x+A)]}\mu_A(\mathrm{d}a)\le q'_k, \tag{47}$$

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 (5).

Lemma 16 *In the high-dimensional regression model, there exist $\mathcal{T}_{v,2t}$-measurable random variables $\tau_{v,t}$, $\chi_{v,t}$ such that*

$$p_v(\vartheta|\mathcal{T}_{v,2t})\propto\exp\left(-\frac{1}{2\tau^2_{v,t}}(\chi_{v,t}-\vartheta)^2+o_p(1)\right),$$

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 (5).

Proof [Lemma 16] 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 $(m^t_{v\to f}(\vartheta))_{t\ge 0}, (m^t_{f\to v}(\vartheta))_{t\ge 0}$. The iteration is

$$m^0_{v\to f}(\vartheta) = 1,$$
$$m^s_{f\to v}(\vartheta) \propto \int p(y_f|X_{fv}\vartheta + \sum_{v'\in\partial f\setminus v} X_{fv'}\vartheta_{v'}, u_f) \prod_{v'\in\partial f\setminus v} m^s_{v'\to f}(\vartheta_{v'}) \prod_{v'\in\partial f\setminus v} \mu_{\Theta|V}(v_{v'}, \mathrm{d}\vartheta_{v'}),$$
$$m^{s+1}_{v\to f}(\vartheta) \propto \prod_{f'\in\partial v\setminus f} m^s_{f'\to v}(\vartheta),$$

with normalization $\int m^t_{f\to v}(\vartheta)\mu_{\Theta|V}(v_v, \mathrm{d}\vartheta) = \int m^t_{v\to f}(\vartheta)\mu_{\Theta|V}(v_v, \mathrm{d}\vartheta) = 1$. For any variable node v,

$$p_v(\vartheta|\mathcal{T}_{v,2t}) \propto \prod_{f\in\partial v} m^{t-1}_{f\to v}(\vartheta). \tag{48}$$

This equation is exact.

We define several quantities related to the belief propagation iteration.

$$\mu^s_{v\to f} = \int \vartheta m^s_{v\to f}(\vartheta)\mu_{\Theta|V}(v_v, \mathrm{d}\vartheta), \quad (\tilde\tau^s_{v\to f})^2 = \int \vartheta^2 m^s_{v\to f}(\vartheta)\mu_{\Theta|V}(v_v, \mathrm{d}\vartheta) - (\mu^s_{v\to f})^2,$$
$$\mu^s_{f\to v} = \sum_{v'\in\partial f\setminus v} x_{fv'}\mu^s_{v'\to f}, \qquad (\tilde\tau^s_{f\to v})^2 = \sum_{v'\in\partial f\setminus v} x^2_{fv'}(\tilde\tau^s_{v'\to f})^2,$$
$$a^s_{f\to v} = \frac{1}{x_{fv}}\frac{\mathrm{d}}{\mathrm{d}\vartheta}\log m^s_{f\to v}(\vartheta)\Big|_{\vartheta=0}, \qquad b^s_{f\to v} = -\frac{1}{x^2_{fv}}\frac{\mathrm{d}^2}{\mathrm{d}\vartheta^2}\log m^s_{f\to v}(\vartheta)\Big|_{\vartheta=0},$$
$$a^s_{v\to f} = \frac{\mathrm{d}}{\mathrm{d}\vartheta}\log m^s_{v\to f}(\vartheta)\Big|_{\vartheta=0}, \qquad b^s_{v\to f} = -\frac{\mathrm{d}^2}{\mathrm{d}\vartheta^2}\log m^s_{v\to f}(\vartheta)\Big|_{\vartheta=0},$$
$$\chi^s_{v\to f} = a^s_{v\to f}/b^s_{v\to f}, \qquad (\tau^s_{v\to f})^2 = 1/b^s_{v\to f}.$$

Lemma 16 follows from the following asymptotic characterization of the quantities in the preceding display in the limit $n, p\to\infty$, $n/p\to\delta$:

$$\mathbb{E}[(\mu^s_{v\to f})^2] \to \delta\sigma^2_s, \qquad \mathbb{E}[(\tilde\tau^s_{v\to f})^2] \to \delta\tilde\tau^2_s,$$
$$(\mu^s_{f\to v}, u_f) \xrightarrow{\mathrm{d}} \mathsf{N}(0,\sigma^2_s)\otimes\mu_U, \qquad (\tilde\tau^s_{f\to v})^2 \xrightarrow{\mathrm{p}} \tilde\tau^2_s, \tag{49}$$
$$(\theta_v, v_v, a^s_{v\to f}/b^s_{v\to f}, b^s_{v\to f}) \xrightarrow{\mathrm{d}} (\Theta, V, \Theta + \tau_s G, 1/\tau^2_s),$$

where in the last line $\Theta\sim\mu_\Theta$, $G\sim\mathsf{N}(0,1)$ independent, and σ^2_s, τ^2_s are defined in (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: $\mathbb{E}[(\mu^0_{v\to f})^2] \to \delta\sigma^2_0$ *and* $\mathbb{E}[(\tilde\tau^0_{v\to f})^2] \to \delta\tilde\tau^2_0$.

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 $\mathbb{E}[(\mu^1_{v\to f})^2] = \mathbb{E}_{\Theta,V}[\mathbb{E}_{\Theta,V}[\Theta|V]^2] = \mathbb{E}[\Theta^2] - \mathsf{mmse}_{\Theta,V}(\infty) = \delta\sigma^2_1$. Similarly, $(\tilde\tau^1_{v\to f})^2 = \mathrm{Var}_{\Theta,V}(\Theta|V = v_v)$, so that $\mathbb{E}[(\tilde\tau^1_{v\to f})^2] = \mathsf{mmse}_{\Theta,V}(\infty) = \delta\tilde\tau^2_0$.

Inductive step 1: If $\mathbb{E}[(\mu^s_{v\to f})^2] \to \delta\sigma^2_s$, *then* $(\mu^s_{f\to v}, u_f) \xrightarrow{\mathrm{d}} \mathsf{N}(0,\sigma^2_s)\otimes\mu_U$.

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 $v' \in \partial f\backslash v$. Thus, $\mu^s_{f\to v}|\mathcal{T}^{2s+1,1}_{f\to v} \sim \mathsf{N}(0, \frac{1}{n}\sum_{v'\in\partial f\backslash v}(\mu^s_{v'\to f})^2)$. Note that $\mathbb{E}[\frac{1}{n}\sum_{v'\in\partial f\backslash v}(\mu^s_{v'\to f})^2] = (p-1)\mathbb{E}[(\mu^s_{v\to f})^2]/n \to \sigma_s^2$ 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, $\frac{1}{n}\sum_{v'\in\partial f\backslash v}(\mu^s_{v'\to f})^2 \xrightarrow{\mathrm{p}} \sigma_s^2$. 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,

$$\mathbb{E}[\sum_{v'\in\partial f\backslash v} x^2_{fv'}(\tilde\tau^s_{v'\to f})^2] = (p-1)\mathbb{E}[(\tilde\tau^s_{v\to f})^2]/n \to \tilde\tau_s^2.$$

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* $(\theta_v, v_v, a^{s+1}_{v\to f}/b^{s+1}_{v\to f}, b^{s+1}_{v\to f}) \xrightarrow{\mathrm{d}} (\Theta, V, \Theta + \tau_{s+1}G, 1/\tau^2_{s+1})$ *where* $G\sim\mathsf{N}(0,1)$ *independent of* $(\Theta,V)\sim\mu_{\Theta,V}$.

For all $(f,v)\in\mathcal{E}$ and $s\ge 1$, define

$$p^s_{f\to v}(y;x) = \int p(y|x + \sum_{v'\in\partial f\backslash v} x_{fv'}\vartheta_{v'}, u_f)\prod_{v'\in\partial f\backslash v} m^s_{v'\to f}(\vartheta_{v'})\prod_{v'\in\partial f\backslash v}\mu_{\Theta|V}(v_{v'},\mathrm{d}\vartheta_{v'}).$$

More compactly, we may write $p^s_{f\to v}(y;x,u_f) = \mathbb{E}_{\{\Theta_{v'}\}}[p(y|x+\sum_{v'\in\partial f\backslash v} x_{fv'}\Theta_{v'}, u_f)]$, 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 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

$$a^s_{f\to v} = a^s_{f\to v}(y_f) \qquad\text{and}\qquad a^{s+1}_{v\to f} = \sum_{f'\in\partial v\backslash f} x_{f'v}a^s_{f'\to v}(y_{f'}),$$
$$b^s_{f\to v} = b^s_{f\to v}(y_f) \qquad\text{and}\qquad b^{s+1}_{v\to f} = \sum_{f'\in\partial v\backslash f} x^2_{f'v}b^s_{f'\to v}(y_{f'}).$$

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. Denote by P^* the joint distribution of all random variables in the regression model. 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 that $(\theta_{v'}, v_{v'}) \overset{\text{iid}}{\sim} \mu_{\Theta,V}$ 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 D.1.1.

Lemma 17 *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 18 *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.

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 $(\mu^s_{f\to v}, \tilde\tau^s_{f\to v}) \xrightarrow[P_{v,0}]{\text{d}} \mathsf{N}(0, \sigma_s^2) \times \delta_{\tilde\tau_s}$.

By Lemma 17, the inductive hypothesis, and Lemma 11, we have for fixed y

$$\begin{pmatrix}
\mathbb{E}_{G_1}[p(y|\mu^s_{f\to v} + \tilde\tau^s_{f\to v}G_1, u_f)]\\
\mathbb{E}_{G_1}[\dot p(y|\mu^s_{f\to v} + \tilde\tau^s_{f\to v}G_1, u_f)]\\
\mathbb{E}_{G_1}[\ddot p(y|\mu^s_{f\to v} + \tilde\tau^s_{f\to v}G_1, u_f)]
\end{pmatrix} \xrightarrow[P_{v,0}]{\text{d}} \begin{pmatrix}
\mathbb{E}_{G_1}[p(y|\sigma_s G_0 + \tilde\tau_s G_1, U]\\
\mathbb{E}_{G_1}[\dot p(y|\sigma_s G_0 + \tilde\tau_s G_1, U)]\\
\mathbb{E}_{G_1}[\ddot p(y|\sigma_s G_0 + \tilde\tau_s G_1, U)]
\end{pmatrix},$$

where and $G_0, G_1 \sim \mathsf{N}(0,1)$ and $U \sim \mu_U$ independent. Applying Lemma 17 and Slutsky's Theorem, we have that

$$\begin{pmatrix}
p^s_{f\to v}(y;0)\\
\dot p^s_{f\to v}(y;0)\\
\ddot p^s_{f\to v}(y;0)
\end{pmatrix} \xrightarrow[P_{v,0}]{\text{d}} \begin{pmatrix}
\mathbb{E}_{G_1}[p(y|\sigma_s G_0 + \tilde\tau_s G_1, U)]\\
\mathbb{E}_{G_1}[\dot p(y|\sigma_s G_0 + \tilde\tau_s G_1, U)]\\
\mathbb{E}_{G_1}[\ddot p(y|\sigma_s G_0 + \tilde\tau_s G_1, U)]
\end{pmatrix}.$$

By the Continuous Mapping Theorem,

$$p^s_{f\to v}(y;0) \xrightarrow[P_{v,0}]{\mathrm{d}} \mathbb{E}_{G_1}[p(y|\sigma_s G_0 + \tilde{\tau}_s G_1, U)],$$
$$a^s_{f\to v}(y) \xrightarrow[P_{v,0}]{\mathrm{d}} \frac{\mathrm{d}}{\mathrm{d}x} \log \mathbb{E}_{G_1}[\dot{p}(y|\sigma_s G_0 + \tilde{\tau}_s G_1, U)]\Big|_{x=0},$$
$$b^s_{f\to v}(y) \xrightarrow[P_{v,0}]{\mathrm{d}} -\frac{\mathrm{d}^2}{\mathrm{d}x^2} \log \mathbb{E}_{G_1}[\dot{p}(y|\sigma_s G_0 + \tilde{\tau}_s G_1, U)]\Big|_{x=0}.$$

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 (47), we have

$$\mathbb{E}_{P_{v,0}}[p^s_{f\to v}(y|0)] \to \mathbb{E}_{G_0,G_1,U}[p(y|\sigma_s G_0 + \tilde{\tau}_s G_1, U)],$$
$$\mathbb{E}_{P_{v,0}}[a^s_{f\to v}(y)^2] \to \mathbb{E}_{G_0,U}\left[\left(\frac{\mathrm{d}}{\mathrm{d}x} \log \mathbb{E}_{G_1}[\dot{p}(y|\sigma_s G_0 + \tilde{\tau}_s G_1, U)]\Big|_{x=0}\right)^2\right],$$
$$\mathbb{E}_{P_{v,0}}[b^s_{f\to v}] \to -\mathbb{E}_{G_0,U}\left[\frac{\mathrm{d}^2}{\mathrm{d}x^2} \log \mathbb{E}_{G_1}[\dot{p}(y|\sigma_s G_0 + \tilde{\tau}_s G_1, U)]\Big|_{x=0}\right].$$

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 (47), 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

$$a^{s+1}_{v\to f} = \sum_{f'\in\partial v\setminus f} x_{f'v} a^s_{f'\to v}(y_{f'}) \xrightarrow[P_{v,0}]{\mathrm{d}} \mathsf{N}\left(0, 1/\tau^2_{s+1}\right). \tag{50}$$

We compute

$$\begin{aligned}
\mathbb{E}_{P_{v,0}}[(a^s_{f\to v}(y_f))^2] &= \mathbb{E}_{P_{v,0}}[\mathbb{E}_{P_{v,0}}[(a^s_{f\to v}(y_f))^2 | \sigma(\mathcal{T}^{2s+1,1}_{f\to v}, (x_{fv'})_{v'\in\partial f\setminus v}), u_f]] \\
&= \mathbb{E}_{P_{v,0}}\left[\int a^s_{f\to v}(y)^2 p^s_{f\to v}(y;0)\mathrm{d}y\right] \\
&= \int \mathbb{E}_{P_{v,0}}\left[a^s_{f\to v}(y)^2 p^s_{f\to v}(y;0)\right] \mathrm{d}y,
\end{aligned}$$

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 (5) from Lemma 12. We conclude (50).

Now we compute the asymptotic behavior of $b_{v\to f}^{s+1}$ under $P_{v,0}$. Under $P_{v,0}$, $x_{f'v}$ is independent of $y_{f'}$, and $(x_{f'v}, b_{f'\to v}^s(y_{f'}))$ are mutually independent for $f' \in \partial v \backslash f$. Thus, $\mathbb{E}_{P_{v,0}}[x_{f'v}^2 b_{f'\to v}^s(y_{f'})] = \mathbb{E}_{P_{v,0}}[b_{f'\to v}^s(y_{f'})]/n$. Because $b_{f'\to v}^s(y_{f'})$ is bounded by (47), if we can show that $\mathbb{E}_{P_{v,0}}[b_{f'\to v}^s(y_{f'})] \to 1/\tau_{s+1}^2$, then $b_{v\to f}^{s+1} \xrightarrow[P_{v,0}]{\mathrm{p}} 1/\tau_{s+1}^2$ will follow by the weak law of large numbers. We compute

$$\begin{aligned}
\mathbb{E}_{P_{v,0}}[b_{f\to v}^s(y_f)] &= \mathbb{E}_{P_{v,0}}[\mathbb{E}_{P_{v,0}}[b_{f\to v}^s(y_f)|\sigma(\mathcal{T}_{f\to v}^{2s+1,1}, (x_{fv'})_{v'\in\partial f\backslash v}, u_f)]] \\
&= \mathbb{E}_{P_{v,0}}\left[\int b_{f\to v}^s(y) p_{f\to v}^s(y;0)\mathrm{d}y\right] \\
&= \int \mathbb{E}_{P_{v,0}}\left[b_{f\to v}^s(y) p_{f\to v}^s(y;0)\right]\mathrm{d}y,
\end{aligned}$$

where the last equation follows by Fubini's theorem (using that the integrand is bounded by the integrable function $q_2\mathbb{E}_{P_{v,0}}[p_{f\to v}^s(y;0)]$). The integrands converge point-wise, so that

$$\begin{aligned}
&\mathbb{E}_{P_{v,0}}[b_{f\to v}^s(y_f)] \\
&\quad\to \mathbb{E}_{G_0,U}\left[\int \frac{\mathbb{E}_{G_1}[\dot{p}(y|\sigma_s G_0 + \tilde{\tau}_s G_1, U)]^2}{\mathbb{E}_{G_1}[p(y|\sigma_s G_0 + \tilde{\tau}_s G_1, U)]}\mathrm{d}y\right] - \int \mathbb{E}_{G_0,G_1,U}[\ddot{p}(y|\sigma_s G_0 + \tilde{\tau}_s G_1, U)]\mathrm{d}y \\
&\quad= \frac{1}{\tau_{s+1}^2},
\end{aligned}$$

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 (47). Thus, we conclude that $b_{v\to f}^{s+1} \xrightarrow[P_{v,0}]{\mathrm{p}} 1/\tau_{s+1}^2$.

Now we compute the asymptotic distribution of $(a_{v\to f}^{s+1}, b_{v\to f}^{s+1})$ under $P_{v,\theta}$ for any $\theta \in [-M, M]$. The log-likelihood ratio between $P_{v,\theta}$ and $P_{v,0}$ is

$$\begin{aligned}
\sum_{f'\in\partial v} \log\frac{p_{f'\to v}^s(y_{f'}|x_{f'v}\theta)}{p_{f'\to v}^s(y_{f'}|0)} &= \log\frac{m_{v\to f}^{s+1}(\theta)}{m_{v\to f}^{s+1}(0)} + \log\frac{p_{f\to v}^s(y_f|x_{fv}\theta)}{p_{f\to v}^s(y_f|0)} \\
&= \theta a_{v\to f}^{s+1} - \frac{1}{2}\theta^2 b_{v\to f}^{s+1} + O_p(n^{-1/2}),
\end{aligned}$$

where we have used Lemma 18 and that $\left|\log\frac{p_{f\to v}^s(y_f|x_{fv}\theta)}{p_{f\to v}^s(y_f|0)}\right| \le M q_1|x_{fv}| = O_p(n^{-1/2})$. 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 (van der Vaart, 1998, 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 (van der Vaart, 1998, Theorem 2.3), we conclude $(a_{v\to f}^{s+1}/b_{v\to f}^{s+1}, b_{v\to f}^{s+1}) \xrightarrow[P_{v,\theta}]{\mathrm{d}} \mathsf{N}(\theta, \tau_{s+1}^2) \otimes \delta_{1/\tau_{s+1}^2}$.

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 of θ_v and the random vector $\boldsymbol{D} := \mathcal{T}_{v,2t} \setminus \{\theta_v, v_v\}$, 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., (Durrett, 2010, Example 5.1.5)). Because $\boldsymbol{D}$ has the same distribution under 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

$$\epsilon_{v\to f}^s = \sup_{\vartheta \in [-M,M]} \left| \log \frac{m_{v\to f}^s(\vartheta)}{m_{v\to f}^s(0)} - \left(\vartheta a_{v\to f}^s - \frac{1}{2}\vartheta^2 b_{v\to f}^s \right) \right|,$$

where because all the terms are continuous in ϑ, the random variable $\epsilon_{v\to f}^s$ is measurable and finite. We have that

$$\mu_{v\to f}^s \geq \frac{\int \vartheta \exp(\vartheta a_{v\to f}^s - \vartheta^2 b_{v\to f}^s/2 - \epsilon_{v\to f}^s)\mu_\Theta(v_v, \mathrm{d}\vartheta)}{\int \exp(\vartheta a_{v\to f}^s - \vartheta^2 b_{v\to f}^s/2 + \epsilon_{v\to f}^s)\mu_\Theta(v_v, \mathrm{d}\vartheta)} \geq e^{-2\epsilon_{v\to f}^s} \eta_{\Theta,V}(a_{v\to f}^s/b_{v\to f}^s, v_v; 1/b_{v\to f}^s),$$

where $\eta_{\Theta,V}(y, v; \tau^2) = \mathbb{E}_{\Theta,V,G}[\Theta | \Theta + \tau G = y; V = v]$ and $(\Theta, V) \sim \mu_{\Theta,V}$, $G \sim \mathsf{N}(0, 1)$ independent. Likewise,

$$\mu_{v\to f}^s \leq e^{2\epsilon_{v\to f}^s} \eta_{\Theta,V}(a_{v\to f}^s/b_{v\to f}^s, v_v; 1/b_{v\to f}^s).$$

Because $\eta_{\Theta,V}$ takes values in the bounded interval $[-M, M]$ and $\epsilon_{v\to f} = o_p(1)$ by Lemma 18, 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 $\mathbb{R}\times\mathbb{R}_{>0}$.[9] 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 11. 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

$$\begin{aligned} e^{-2\epsilon^s_{v\to f}} s^2_{\Theta,V}(a^s_{v\to f}/b^s_{v\to f}, v_v; 1/b^s_{v\to f}) &\le \int \vartheta^2 m^s_{v\to f}(\vartheta)\mu_\Theta(\mathrm{d}\vartheta) \\ &\le e^{2\epsilon^s_{v\to f}} s^2_{\Theta,V}(a^s_{v\to f}/b^s_{v\to f}, v_v; 1/b^s_{v\to f}), \end{aligned}$$

where $s^2_{\Theta,V}(y, v; \tau^2) = \mathbb{E}_{\Theta,V,G}[\Theta^2 | \Theta + \tau G = y, V = v]$ and $(\Theta, V) \sim \mu_{\Theta,V}$, $G \sim \mathsf{N}(0,1)$ 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 (49) is established.

To complete the proof of Lemma 16, first observe by (48) 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

$$\left|\log \frac{m^{t-1}_{f\to v}(\vartheta_v)}{m^{t-1}_{f\to v}(0)}\right| \le M|x_{fv}| \sup_{x\in\mathbb{R}} \left|\frac{\dot p^{t-1}_{f\to v}(y_f; x)}{p^{t-1}_{f\to v}(y_f; x)}\right| \le Mq_1|x_{fv}| = o_p(1).$$

By Lemma 18, 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 (49). ∎

We complete the proof of Lemma 8 for the high-dimensional regression model. Consider any estimator $\hat\theta : \mathcal{T}_{v,2t} \mapsto [-M, M]$ on the computation tree. We compute

$$\begin{aligned} &\mathbb{E}[\ell(\theta_v, \hat\theta(\mathcal{T}_{v,2t}))] = \mathbb{E}[\mathbb{E}[\ell(\theta_v, \hat\theta(\mathcal{T}_{v,2t}))|\mathcal{T}_{v,2t}]] \\ &\quad = \mathbb{E}\left[\int \ell(\vartheta, \hat\theta(\mathcal{T}_{v,2t})) \frac{1}{Z(\mathcal{T}_{v,2t})} \exp\left(-\frac{1}{2\tau^2_{v,t}}(\chi_{v,t} - \vartheta)^2 + o_p(1)\right) \mu_{\Theta|V}(v_v, \mathrm{d}\vartheta)\right] \\ &\quad \ge \mathbb{E}\left[\exp(-2\epsilon_v) \int \ell(\vartheta, \hat\theta(\mathcal{T}_{v,2t})) \frac{1}{Z(\chi_{v,t}, \tau_{v,t}, v_v)} \exp\left(-\frac{1}{2\tau^2_{v,t}}(\chi_{v,t} - \vartheta)^2\right) \mu_{\Theta|V}(v_v, \mathrm{d}\vartheta)\right] \\ &\quad \ge \mathbb{E}\left[\exp(-2\epsilon_v) R(\chi_{v,2t}, \tau_{v,2t}, v_v)\right], \end{aligned}$$

9. This commonly known fact holds, for example, by (Lehmann and Romano, 2005, 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.

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 13(b), $R(\chi,\tau,v)$ is continuous in (χ,τ) on $\mathbb{R}\times\mathbb{R}_{>0}$. By Lemma 16, $\epsilon_v = o_p(1)$. 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 11 and the equality holds by Lemma 13(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 8 in the high-dimensional regression model is complete.

D.1.1. TECHNICAL TOOLS

Proof [Lemma 17] By Lindeberg's principle (see, e.g., Chatterjee (2006)) and using that μ_Θ is supported on $[-M,M]$, we have

$$|p^s_{f\to v}(y;0) - \mathbb{E}_{G_1}[p(y|\mu^s_{f\to v}+\tilde{\tau}^s_{f\to v}G_1,u_f)]| \leq \frac{M^3\sup_{x\in\mathbb{R}}|\partial_x^3 p(y|x,u_f)|}{3}\sum_{v'\in\partial f\setminus v}|x_{fv'}|^3,$$

$$|\dot{p}^s_{f\to v}(y;0) - \mathbb{E}_{G_1}[\dot{p}(y|\mu^s_{f\to v}+\tilde{\tau}^s_{f\to v}G_1,u_f)]| \leq \frac{M^3\sup_{x\in\mathbb{R}}|\partial_x^4 p(y|x,u_f)|}{3}\sum_{v'\in\partial f\setminus v}|x_{fv'}|^3,$$

$$|\ddot{p}^s_{f\to v}(y;0) - \mathbb{E}_{G_1}[\ddot{p}(y|\mu^s_{f\to v}+\tilde{\tau}^s_{f\to v}G_1,u_f)]| \leq \frac{M^3\sup_{x\in\mathbb{R}}|\partial_x^5 p(y|x,u_f)|}{3}\sum_{v'\in\partial f\setminus v}|x_{fv'}|^3.$$

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 18] 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 the expectation is taken with respect to $\hat{G}_{f'} \stackrel{\mathrm{d}}{=} \sum_{v' \in \partial f' \setminus v} x_{f'v'} \Theta_{v' \to f'}$ and $x_{f'v'}$ 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. ■

D.2. Information-theoretic lower bound in the low-rank matrix estimation model

In this section, we prove Lemma 8 in the low-rank matrix estimation model.

Recall that conditions on the conditional density in assumption R4 are equivalent to positivity, boundedness, and the existence finite, non-negative constants q'_k such that $\frac{|\partial^k_x p(y|x)|}{p(y|x)} \le q'_k$ for $1 \le k \le 5$. In particular, we have (47) 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 products 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 19 *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, $(\theta_v, v_v, \chi_{v,t}, q_{v,t}) \stackrel{\mathrm{d}}{\to} (\Theta, V, q_t^{1/2}\Theta + G, q_t)$ where $(\Theta, V) \sim \mu_{\Theta,V}, G \sim \mathsf{N}(0,1)$ independent of Θ, V, and q_t is given by (7).

Proof [Lemma 19] As in the proof of Lemma 16, we compute the posterior density $p_v(\vartheta | \mathcal{T}_{v,2t-1})$ via belief propogation. The belief propagation iteration is

$$m^0_{f \to v}(\ell) = 1,$$

$$m^{s+1}_{v \to f}(\vartheta) \propto \int \prod_{f' \in \partial v \setminus f} \left(\exp\left(-\frac{n}{2}(x_{f'v} - \frac{1}{n}\ell_{f'}\vartheta)^2 \right) m^s_{f' \to v}(\ell_{f'}) \mu_{\Lambda|U}(u_{f'}, \mathrm{d}\ell_{f'}) \right),$$

$$m^s_{f \to v}(\ell) \propto \int \prod_{v' \in \partial f \setminus v} \left(\exp\left(-\frac{n}{2}(x_{fv'} - \frac{1}{n}\ell\vartheta_{v'})^2 \right) m^s_{v' \to f}(\vartheta_{v'}) \mu_{\Theta|V}(v_{v'}, \mathrm{d}\vartheta_{v'}) \right).$$

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 19 follows from the following asymptotic characterization of the quantities in the preceding display in the limit $n,p\to\infty$, $n/p\to\delta$:

$$\begin{aligned}
\mathbb{E}[\mu^s_{f\to v}\lambda_f] &\to q_{s+1}, & \mathbb{E}[(\mu^s_{f\to v})^2] &\to q_{s+1},\\
\alpha^{s+1}_{v\to f} &\xrightarrow{\mathrm{p}} q_{s+1}, & (\tau^{s+1}_{v\to f}) &\xrightarrow{\mathrm{p}} q_{s+1},\\
(\theta_v, v_v, a^s_{v\to f}, b^s_{v\to f}) &\xrightarrow{\mathrm{d}} (\Theta, V, q_s\Theta + q_s^{1/2}G_2, q_s),\\
\mathbb{E}[\mu^s_{v\to f}\theta_v] &\to \delta\hat q_s, & \mathbb{E}[(\mu^s_{v\to f})^2] &\to \delta\hat q_s,\\
\alpha^s_{f\to v} &\xrightarrow{\mathrm{p}} \hat q_s, & (\hat\tau^{s+1}_{f\to v})^2 &\xrightarrow{\mathrm{p}} \hat q_s,\\
(\lambda_f, u_f, a^s_{f\to v}, b^s_{f\to v}) &\xrightarrow{\mathrm{d}} (\Lambda, U, \hat q_s\Lambda + \hat q_s^{1/2}G, \hat q_s).
\end{aligned} \tag{51}$$

As in the proof of Lemma 16, 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

$$\frac{\mathrm{d}}{\mathrm{d}\vartheta}\mathbb{E}_{\Lambda_{f'}}\left[\exp\left(-\frac{1}{2n}\Lambda^2_{f'}\vartheta^2 + x_{f'v}\Lambda_{f'}\vartheta\right)\right]\Big|_{\vartheta=0} = \mathbb{E}_{\Lambda_{f'}}\left[x_{f'v}\Lambda_{f'}\right] = x_{f'v}\mu^s_{f'\to v},$$
$$\frac{\mathrm{d}^2}{\mathrm{d}\vartheta^2}\mathbb{E}_{\Lambda_{f'}}\left[\exp\left(-\frac{1}{2n}\Lambda^2_{f'}\vartheta^2 + x_{f'v}\Lambda_{f'}\vartheta\right)\right]\Big|_{\vartheta=0} = \mathbb{E}_{\Lambda_{f'}}\left[x^2_{f'v}\Lambda^2_{f'} - \frac{1}{n}\Lambda^2_{f'}\right] = \left(x^2_{f'v} - \frac{1}{n}\right)s^s_{f'\to v}.$$

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 of 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 (van der Vaart, 1998, 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 $B = \sum_{f'\in\partial v\setminus f}(x^2_{f'v} - 1/n)s^s_{f'\to v}$. We have

$$A = \frac{1}{n^2}\sum_{v'\in\partial f\setminus v}\lambda^2_{f'}\theta^2_v(\mu^s_{f'\to v})^2 + \frac{2}{n}\sum_{f'\in\partial v\setminus f}\lambda_{f'}\theta_v z_{f'v}(\mu^s_{f'\to v})^2 + \sum_{v'\in\partial f\setminus v} z^2_{f'v}(\mu^s_{f'\to v})^2.$$

Observe $\mathbb{E}[\lambda^2_{f'}\theta^2_v(\mu^s_{f'\to v})^2] \le M^6$, 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 $\mu^s_{f'\to v}$, $\mathbb{E}[|\lambda_{f'}\theta_v z_{f'v}(\mu^s_{f'\to v})^2|] \le M^4\sqrt{2/(\pi n)}$, 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 $\mu^s_{f'\to v}$ 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 $(z^2_{f'v}(\mu^s_{f'\to v})^2)_{f'\in\partial v\setminus f}$ are mutually independent and $\mathbb{E}[z^4_{f'v}(\mu^s_{f'\to v})^4] \le 3M^4/n^2$, so that by the weak law of large numbers we have that the last term converges to q_{s+1} in probability. Thus, $A \xrightarrow{\mathrm{p}} q_{s+1}$.

We have

$$B = \frac{1}{n^2}\sum_{v'\in\partial f\setminus v}\lambda_f^2\theta^2_{v'}s^s_{v'\to f} + \frac{2}{n}\sum_{v'\in\partial f\setminus v}\lambda_f\theta_{v'}s^s_{v'\to f} + \sum_{v'\in\partial f\setminus v}(z^2_{f'v}-1/n)s^s_{v'\to f}.$$

As in the analysis of the first two terms of A, we may use that $s^s_{v'\to f}\le M^2$ to argue that the first two terms of B converge to 0 in probability. Further, because $z_{f'v}$ is independent of $s^s_{v'\to f}$, the expectation of the last term is 0. Further, $\mathbb{E}[(z^2_{f'v}-1/n)^2(s^s_{v'\to f})^2] \le 2\mathbb{E}[(z^4_{f'v}+1/n^2)]\mathbb{E}[(s^s_{v'\to f})^2] \le 8M^4/n^2$, so that by the weak law of large numbers, the final term converges to 0 in probability. Thus, $B \xrightarrow{\mathrm{p}} 0$. Because, as we have shown, $A \xrightarrow{\mathrm{p}} q_{s+1}$, we conclude $b^{s+1}_{v\to f} \xrightarrow{\mathrm{p}} q_{s+1}$.

Combining with $(\theta_v, v_v, a^{s+1}_{v\to f}) \xrightarrow{\mathrm{d}} (\Theta, V, q_{s+1}\Theta + q^{1/2}_{s+1}G)$ and applying the Continuous Mapping Theorem (van der Vaart, 1998, Theorem 2.3), we have $(\theta_v, a^{s+1}_{v\to f}, b^{s+1}_{v\to f}) \xrightarrow{\mathrm{d}} (\Theta, q_{s+1}\Theta + q^{1/2}_{s+1}G, q_{s+1})$.

Inductive step 3: If $(\theta_v, v_v, a^s_{v\to f}, b^s_{v\to f}) \xrightarrow{\mathrm{d}} (\Theta, V, q_s\Theta + q^{1/2}_s G_1, q_s)$*, then* $\mathbb{E}[\mu^s_{v\to f}\theta_v] \to \delta\hat{q}_s$ *and* $\mathbb{E}[(\mu^s_{v\to f})^2] \to \delta\hat{q}_s$.

We will require the following lemma, whose proof is deferred to section D.2.1.

Lemma 20 *For any fixed* s*, we have for* $\vartheta, \ell \in [-M, M]$

$$\log\frac{m^s_{v\to f}(\vartheta)}{m^s_{v\to f}(0)} = \vartheta a^s_{v\to f} - \frac{1}{2}\vartheta^2 b^s_{v\to f} + O_p(n^{-1/2}),$$
$$\log\frac{m^s_{f\to v}(\ell)}{m^s_{f\to v}(0)} = \ell a^s_{f\to v} - \frac{1}{2}\ell^2 b^s_{f\to v} + O_p(n^{-1/2}),$$

where $O_p(n^{-1/2})$ *has no* ϑ *(or* ℓ*) dependence.*

Define

$$\epsilon^s_{f\to v} = \sup_{\vartheta\in[-M,M]}\left|\log\frac{m^s_{v\to f}(\vartheta)}{m^s_{v\to f}(0)} - \left(\vartheta a^s_{v\to f} - \frac{1}{2}\vartheta^2 b^s_{v\to f}\right)\right|.$$

By Lemma 20, we have $\epsilon^s_{v\to f} = o_p(1)$. Moreover, using the same argument as in inductive step 4 of the proof of Theorem 16, we have that

$$e^{-2\epsilon^s_{v\to f}}\eta_{\Theta,V}(a^s_{v\to f}(b^s_{v\to f})^{-1/2}, v_v; b^s_{v\to f}) \le \mu^s_{v\to f} \le e^{2\epsilon^s_{v\to f}}\eta_{\Theta,V}(a^s_{v\to f}(b^s_{v\to f})^{1/2}, v_v; b^s_{v\to f}),$$

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 20, we conclude that

$$\mu^s_{v\to f} = \eta_{\Theta,V}(a^s_{v\to f}/b^s_{v\to f}, v_v; b^s_{v\to f}) + o_p(1).$$

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 $|\theta_v|, |\mu^s_{v\to f}|, |\eta_{\Theta,V}(a^s_{v\to f}/b^s_{v\to f}, v_v; b^s_{v\to f})| \le M$, 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* $\alpha^s_{f\to v} \xrightarrow{\text{p}} \hat{q}_s$ *and* $(\hat{\tau}^{s+1}_{f\to v})^2 \xrightarrow{\text{p}} \hat{q}_s$.

Inductive step 5: If $\alpha^s_{f\to v} \xrightarrow{\text{p}} \hat{q}_s$ *and* $(\hat{\tau}^s_{f\to v})^2 \xrightarrow{\text{p}} \hat{q}_s$, *then* $(\lambda_f, u_f, a^s_{f\to v}, b^s_{f\to v}) \xrightarrow{\text{d}} (\Lambda, U, \hat{q}_s\Lambda + \hat{q}_s^{1/2}G, \hat{q}_s)$.

Inductive step 6: If $(\lambda_f, u_f, a^s_{f\to v}, b^s_{f\to v}) \xrightarrow{\text{d}} (\Lambda, U, \hat{q}_s\Lambda + \hat{q}_s^{1/2}G, \hat{q}_s)$, *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 (51).

To complete the proof of Lemma 19, 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

$$\left|\log \mathbb{E}_{\Lambda_f}[\exp(\vartheta x_{fv}\Lambda_f - \vartheta^2\Lambda_f^2/(2n))]\right| \le M^2|x_{fv}| + M^4/2n = o_p(1).$$

By Lemma 20, 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 (51) and Slutsky's theorem. ■

Lemma 8 in the low-rank matrix estimation model follows from Lemma 19 by exactly the same argument that derived Lemma 8 in the high-dimensional regression model from Lemma 16.

D.2.1. TECHNICAL TOOLS

Proof [Lemma 20] Fix any $\vartheta \in [-M, M]$. By Taylor's theorem, there exists $\vartheta_{f'} \in [-M, M]$ (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 (47) from R4 in the proof of Lemma 8(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. ∎

Appendix E. Weakening the assumptions

Section 5 and the preceding appendices establish under the assumptions A1, A2 and either R3, R4 or M2 all claims in Theorems 1 and 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.

E.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 21 *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 $\boldsymbol{y} = h(\boldsymbol{X}\boldsymbol{\theta}, \boldsymbol{w})$ 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

$$\mathbb{P}\left(\frac{1}{n}\|\boldsymbol{y}-\tilde{\boldsymbol{y}}\|^2>\epsilon\right)\to 0,\ \mathbb{P}\left(\frac{1}{p}\|\boldsymbol{v}-\tilde{\boldsymbol{v}}\|^2>\epsilon\right)\to 0,\ \mathbb{P}\left(\frac{1}{n}\|\boldsymbol{u}-\tilde{\boldsymbol{u}}\|^2>\epsilon\right)\to 0. \quad (52)$$

Note that because in any GFOM the functions $F_t^{(1)}, F_t^{(2)}, G_t^{(1)}, G_t^{(2)}, G_*$ are Lipschitz and $\|\boldsymbol{X}\|_{\mathsf{op}} \xrightarrow{\mathrm{p}} C_\delta < \infty$ as $n,p\to\infty, n/p\to 0$ (Vershynin, 2012, Theorem 5.31), the previous display and the iteration (1) imply

$$\mathbb{P}\left(\frac{1}{p}\|\hat{\boldsymbol{\theta}}^t-\tilde{\hat{\boldsymbol{\theta}}}^t\|^2>c(\epsilon,t)\right)\to 0, \quad (53)$$

for some $c(\epsilon,t)<\infty$ which goes to 0 as $\epsilon\to 0$ for fixed t.

3. We have

$$|\mathsf{mmse}_{\Theta,V}(\tau_s^2)-\mathsf{mmse}_{\tilde{\Theta},\tilde{V}}(\tau_s)^2|<\epsilon, \quad (54)$$

$$\left|\mathbb{E}\left[\mathbb{E}[G_1|h(G,W)+\epsilon^{1/2}Z,G_0]^2\right]-\mathbb{E}\left[\mathbb{E}[G_1|\tilde{h}(G,\tilde{\boldsymbol{W}}),G_0]^2\right]\right|<\tilde{\tau}_s^2\epsilon, \quad (55)$$

for all $s\le t$ where $G_0,G_1,Z\overset{\mathrm{iid}}{\sim}\mathsf{N}(0,1)$, $W\sim\mu_W$, 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 21, 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. (55) 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

$$\mathbb{E}[\Theta^2\mathbf{1}\{|\Theta|>K\}]<\delta\epsilon/L^2,\ \ \mathbb{E}[U^2\mathbf{1}\{|U|>K\}]<\epsilon/2,\ \ \mathbb{E}[V^2\mathbf{1}\{|V|>K\}]<\epsilon/2. \quad (56)$$

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 (54) 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 the original 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 (52) 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 (56). 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, 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

$$\begin{aligned}
&\mathbb{E}[(h(\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\theta}, w_i) - \hat{h}(\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\theta}, w_i))^2 | \boldsymbol{\theta}, w_i] \\
&\qquad = \mathbb{E}[(h(\tau_\Theta \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\theta}/\|\boldsymbol{\theta}\|, w_i) - \hat{h}(\tau_\Theta \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\theta}/\|\boldsymbol{\theta}\|, w_i))^2 | \boldsymbol{\theta}, w_i] + o_p(1).
\end{aligned}$$

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 (52) 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 $\|\boldsymbol{X}\|_{\mathsf{op}}^2 \xrightarrow{\mathrm{p}} C_\delta$ (Vershynin, 2012, 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

$$\mathbb{P}\left(\frac{1}{n}\|\boldsymbol{y}-\tilde{\boldsymbol{y}}\|^2 > 5(C_\delta+4)\epsilon\right) \to 0.$$

Because ϵ was arbitrary, we can in fact achieve (52) by considering a smaller ϵ (without affecting the validity of (54)).

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 (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. (54) and (55) 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 (Durrett, 2010, Theorem 5.5.7), we have that $\mathbb{E}[G_1|(h(G,W)+B_s)_{s\ge\epsilon},G_0]$ converges to $\mathbb{E}[G_1|(h(G,W)+B_s)_{s\ge0},G_0] = \mathbb{E}[G_1|h(G,W),G_0]$ 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 13(a) and (c), and specifically, Eqs. (25) and (27).

Because the lower bound of Theorem 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

$$\lim_{n\to\infty}\frac{1}{p}\sum_{j=1}^{p}\ell(\theta_j,\hat{\theta}_j^t) \ge \inf_{\hat{\theta}(\cdot)} \mathbb{E}[\ell(\tilde{\Theta},\hat{\theta}(\tilde{\Theta}+\tau_t^{(\epsilon)}G,V))].$$

Taking $\epsilon \to 0$ and applying (53), we conclude that (6) holds for $\hat{\boldsymbol{\theta}}^t$, as desired.

The reduction in the high-dimensional regression model is complete.

Proof [Lemma 21] It is enough to prove the result for $\tau = 1$. Note

$$\mathbb{E}[A|Y = y] = \frac{\int a e^{-(y-b)^2} \mu(\mathrm{d}a, \mathrm{d}b)}{\int e^{-(y-b)^2} \mu(\mathrm{d}a, \mathrm{d}b)}, \qquad \mathbb{E}[A|Y_n = y] = \frac{\int a e^{-(y-b)^2} \mu_n(\mathrm{d}a, \mathrm{d}b)}{\int e^{-(y-b)^2} \mu_n(\mathrm{d}a, \mathrm{d}b)}.$$

Because $\mu_n \overset{\mathrm{W}}{\to} \mu$, we have

$$\frac{\int a e^{-(y-b)^2} \mu_n(\mathrm{d}a, \mathrm{d}b)}{\int e^{-(y-b)^2} \mu_n(\mathrm{d}a, \mathrm{d}b)} \to \frac{\int a e^{-(y-b)^2} \mu(\mathrm{d}a, \mathrm{d}b)}{\int e^{-(y-b)^2} \mu(\mathrm{d}a, \mathrm{d}b)},$$

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. ■

E.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 $\tilde{\Lambda}_i = \Lambda_i \mathbf{1}\{|\Lambda_i| \le M\}$ for $1 \le i \le k$. 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 $\{(\boldsymbol{\lambda}_i, \boldsymbol{u}_i)\}_{i \le n} \overset{\mathrm{iid}}{\sim} \mu_{\boldsymbol{\Lambda},\boldsymbol{U}}$ and $\{(\boldsymbol{\theta}_j, \boldsymbol{v}_j)\}_{j \le p} \overset{\mathrm{iid}}{\sim} \mu_{\boldsymbol{\Theta},\boldsymbol{V}}$ and $\boldsymbol{Z} \in \mathbb{R}^{n \times p}$ independent with $z_{ij} \overset{\mathrm{iid}}{\sim} \mathsf{N}(0, 1/n)$. Construct $\tilde{\boldsymbol{\lambda}}_i, \tilde{\boldsymbol{u}}_i, \tilde{\boldsymbol{\theta}}_j, \tilde{\boldsymbol{v}}_j$ by truncating each coordinate at level M as above. Define $\boldsymbol{X}, \tilde{\boldsymbol{X}} \in \mathbb{R}^{n \times p}$ by $x_{ij} = \frac{1}{n} \boldsymbol{\lambda}_i^{\mathsf{T}} \boldsymbol{\theta}_j + z_{ij}$ and $\tilde{z}_{ij} = \frac{1}{n} \tilde{\boldsymbol{\lambda}}_i^{\mathsf{T}} \tilde{\boldsymbol{\theta}} + z_{ij}$. As in the previous section, we have for any $\epsilon > 0$ that

$$\mathbb{P}(\|\boldsymbol{X} - \tilde{\boldsymbol{X}}\|_{\mathsf{op}} > \epsilon) \to 0, \quad \mathbb{P}\left(\frac{1}{p}\|\boldsymbol{v} - \tilde{\boldsymbol{v}}\|^2 > \epsilon\right) \to 0, \quad \mathbb{P}\left(\frac{1}{p}\|\boldsymbol{u} - \tilde{\boldsymbol{u}}\|^2 > \epsilon\right) \to 0.$$

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 (7) after the truncation converges to the state evolution with no truncation.

The reduction now occurs exactly as in the previous section.

Appendix F. Achieving the bound

All that remains to prove Theorems 1 and 2 under assumptions $\mathsf{A1}$, $\mathsf{A2}$ and either $\mathsf{R1}$, $\mathsf{R2}$ or $\mathsf{M1}$, respectively, is to show that the lower bounds in Eqs. (6) and (8) can be achieved. In both cases, we can achieve the bound up to tolerance ϵ using a certain AMP algorithm.

F.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 (5). As we saw in Appendix D.1 and in particular, in Lemma 16, 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 16. Moreover, we saw in the same section that a consequence of Lemma 16 is that the asymptotic limiting Bayes risk with respect to loss ℓ for estimating 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 (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

$$f_t(b^t; y, u) = \mathbb{E}[B^0 - B^t | h(B^0, W) = y,\, B^t = b^t,\, U = u]/\tilde{\tau}_t,$$
$$g_t(a^t; v) = \mathbb{E}[\Theta | V = v,\, \alpha_t \Theta + Z^t = a^t],$$

where $(\Theta, V) \sim \mu_{\Theta,V}$), $(W, U) \sim \mu_{W,U}$, $(B^0, \ldots, B^t) \sim \mathsf{N}(\mathbf{0}, \boldsymbol{\Sigma}_{[0:t]})$, $(Z^1, \ldots, Z^t) \sim \mathsf{N}(\mathbf{0}, \boldsymbol{T}_{[1:t]})$, 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

$$\begin{aligned}
&\mathbb{E}[(B^0 - B^t) f_t(B^t; h(B^0, W), U)]/\tilde{\tau}_t^2 = \mathbb{E}[(\mathbb{E}[B^0 - B^t | Y, B^t, U]/\tilde{\tau}_t)^2]/\tilde{\tau}_t = \alpha_t, \\
&\mathbb{E}[f_s(B^s; h(B^0, W), U) f_t(B^t; h(B^0, W), U)] \\
&\qquad = \mathbb{E}[\mathbb{E}[B^0 - B^s | Y, B^s, U]\mathbb{E}[B^0 - B^t | Y, B^t, U]]/\tilde{\tau}_t^2 \\
&\qquad = \mathbb{E}[(B^0 - B^t)^2 | Y, B^t, U]/\tilde{\tau}_t^2 = T_{s,t}, \\
&\frac{1}{\delta}\mathbb{E}[\Theta g_t(\alpha_t \Theta + Z^t; V)] = \frac{1}{\delta}\mathbb{E}[\mathbb{E}[\Theta | \Theta + Z^t/\alpha_t, V]^2] = \sigma_t^2, \\
&\frac{1}{\delta}\mathbb{E}[g_s(\alpha_s \Theta + Z^s; V) g_t(\alpha_t \Theta + Z^t; V)] = \frac{1}{\delta}\mathbb{E}[\mathbb{E}[\Theta | \Theta + Z^t/\alpha_t, V]^2].
\end{aligned}$$

If f_t, g_t are Lipschitz, then, because h is also Lipschitz, Stein's lemma (Stein, 1981) 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 (37), and Lemma 6 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 (37) 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 evolution 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 1.

F.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

$$f_t(\boldsymbol{b}^t; \boldsymbol{u}) = \mathbb{E}[\boldsymbol{\Lambda}|\boldsymbol{\gamma}_t\boldsymbol{\Lambda} + \boldsymbol{\Sigma}_{t,t}^{1/2}\boldsymbol{G} = \boldsymbol{b}^t; \boldsymbol{U}],$$
$$g_t(\boldsymbol{a}^t; \boldsymbol{v}) = \mathbb{E}[\boldsymbol{\Theta}|\boldsymbol{\alpha}_t\boldsymbol{\Theta} + \boldsymbol{T}_{t,t}^{1/2}\boldsymbol{G} = \boldsymbol{a}^t; \boldsymbol{V}].$$

We check that the parameters so defined satisfy the AMP state evolution (38). Note that by (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 (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 (37), and Lemma 6 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 $\epsilon > 0$.

Appendix G. Proofs for sparse phase retrieval and sparse PCA

G.1. Proof of Lemma 4

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 $\boldsymbol{v} = \sqrt{\widetilde{\alpha}}\boldsymbol{\theta}_0 + \boldsymbol{g}$. We apply Theorem 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

$$q_{t+1} = \frac{\hat{q}_t}{1 + \hat{q}_t}, \quad \hat{q}_t = \frac{1}{\delta}\mathbb{E}\big\{\mathbb{E}[\sqrt{\delta}\Theta_0|(\delta q_t)^{1/2}\Theta_0 + G; V]^2\big\},$$

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 2. Also note that the second moment of the conditional expectation above is equal to $\mathbb{E}\{\mathbb{E}[\sqrt{\delta}\Theta_0|(\delta(q_t + \tilde{\alpha}))^{1/2}\Theta_0 + G]^2\}$ and a simple calculation yields

$$\hat{q}_{t+1} = V_\pm(q_t + \tilde{\alpha}), \quad q_t = \frac{\hat{q}_t}{1+\hat{q}_t},$$

which is equivalent to Eqs. (12), (13).

Let $Y = \sqrt{\delta(q_t + \tilde{\alpha})}\Theta_0 + G$, $G \sim \mathsf{N}(0,1)$. Theorem 2 then yields

$$\begin{aligned}\frac{1}{p}\|\hat{\boldsymbol{\theta}}^t - \boldsymbol{\theta}_0/\lambda\|_2^2 &\geq \inf_{\hat{\theta}(\,\cdot\,)} \mathbb{E}\big\{\big(\hat{\theta}(Y) - \Theta_0/\lambda\big)^2\big\} + o_p(1)\\ &= \frac{1}{\lambda^2}\mathbb{E}\big\{\big(\mathbb{E}(\Theta_0|Y) - \Theta_0\big)^2\big\} + o_p(1).\end{aligned}$$

In order to prove the upper bound (14), it is sufficient to consider $\|\hat{\boldsymbol{\theta}}^t\|_2^2 \leq p$. Then, for any $\lambda \geq 0$,

$$\begin{aligned}\frac{1}{p}\langle\hat{\boldsymbol{\theta}}^t, \boldsymbol{\theta}_0\rangle &\leq \frac{1}{p}\langle\hat{\boldsymbol{\theta}}^t, \boldsymbol{\theta}_0\rangle - \frac{\lambda}{2p}(\|\hat{\boldsymbol{\theta}}^t\|_2^2 - p)\\ &= \frac{\lambda}{2} + \frac{1}{2\lambda p}\|\boldsymbol{\theta}_0\|_2^2 - \frac{\lambda}{2p}\|\hat{\boldsymbol{\theta}}^t - \boldsymbol{\theta}_0/\lambda\|_2^2\\ &\leq \frac{\lambda}{2} + \frac{1}{2\lambda}\mathbb{E}\{\Theta_0^2\} - \frac{1}{2\lambda}\mathbb{E}\big\{\big(\mathbb{E}(\Theta_0|Y) - \Theta_0\big)^2\big\} + o(1)\\ &\leq \frac{\lambda}{2} + \frac{1}{2\lambda}V_\pm(q_t + \tilde{\alpha}) + o(1).\end{aligned}$$

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.

G.2. Proof of Corollary 5

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 4, we get

$$\lim_{n\to\infty} \inf_{\overline{\boldsymbol{\theta}}_0 \in \mathscr{T}(\varepsilon,R)} \mathbb{E}\left\{\frac{\langle\overline{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}^t\rangle}{\|\overline{\boldsymbol{\theta}}_0\|_2\|\hat{\boldsymbol{\theta}}^t\|_2}\right\} \leq \sqrt{\frac{V_\pm(q'_t + \tilde{\alpha}')}{(\mu')^2\varepsilon'}}, \tag{57}$$

where we used dominated convergence to pass from the limit in probability to the 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. (57) also holds with μ', ε', q'_t replaced by μ, ε, q_t:

$$\lim_{n\to\infty} \inf_{\overline{\boldsymbol{\theta}}_0 \in \mathscr{T}(\varepsilon,R)} \mathbb{E}\left\{\frac{\langle\overline{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}^t\rangle}{\|\overline{\boldsymbol{\theta}}_0\|_2\|\hat{\boldsymbol{\theta}}^t\|_2}\right\} \leq \sqrt{\frac{V_\pm(q_t + \tilde{\alpha})}{\mu^2\varepsilon}}. \tag{58}$$

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. (12), (13) imply that for any $\eta > 0$ there exists $q_* > 0$ such that, if $q_t + \tilde{\alpha} \le q_*$, then

$$q_{t+1} \le (\mu^4\varepsilon^2\delta + \eta)(q_t + \tilde{\alpha}).$$

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. (13) to get $q_t \le \overline{q}_t$ for all t, where $\overline{q}_0 = 0$ and

$$\overline{q}_{t+1} = F_0(\overline{q}_t + \tilde{\alpha}), \qquad F_0(q) := \frac{\mu^2\varepsilon^2}{1-\varepsilon}\sinh(\mu^2\delta q).$$

Further Eq. (58) implies

$$\lim_{n\to\infty}\inf_{\overline{\boldsymbol{\theta}}_0\in\mathscr{T}(\varepsilon,R)}\mathbb{E}\left\{\frac{\langle\overline{\boldsymbol{\theta}}_0,\hat{\boldsymbol{\theta}}^t\rangle}{\|\overline{\boldsymbol{\theta}}_0\|_2\|\hat{\boldsymbol{\theta}}^t\|_2}\right\} \le \sqrt{\frac{\overline{q}_{t+1}}{\mu^2\varepsilon}}. \tag{59}$$

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

$$x_{t+1} = a\sinh(x_t + b).$$

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 $b < 1/2$. Substituting this bound in Eq. (59), we obtain the desired claim.

G.3. Proof of Corollary 3

Consider first the case of a random vector $\boldsymbol{\theta}_0$ with i.i.d. entries $\theta_{0,i} \sim \mu_\theta$. Define, for $\Theta_0 \sim \mu_\theta$,

$$\begin{aligned}F_\varepsilon(q) &:= \mathbb{E}\{\mathbb{E}[\Theta_0|\sqrt{q}\Theta_0 + G]^2\}\\ &= e^{-q\mu^2}\mu^2\varepsilon^2\mathbb{E}\left\{\frac{\sinh(\mu\sqrt{q}G)^2}{1-\varepsilon+\varepsilon e^{-q\mu^2/2}\cosh(\mu\sqrt{q}G)}\right\}.\end{aligned}$$

Setting $q_t = \tau_t^{-2}$, $\hat{q}_t = \sigma_t^2$, and $\tilde{\alpha} = \alpha/(1-\alpha)$, and referring to Lemma 12, the state evolution recursion (5) takes the form

$$\begin{aligned}\hat{q}_t &= F_\varepsilon(q_t + \tilde{\alpha}), \quad q_{t+1} = \delta\, H(\hat{q}_t),\\ H(q) &:= \mathbb{E}_{G_0,Y}\left[\left(\frac{\mathbb{E}_{G_1}\partial_x p(Y|\sqrt{q}G_0 + \sqrt{1-q}G_1)}{\mathbb{E}_{G_1}p(Y|\sqrt{q}\,G_0 + \sqrt{1-q}G_1)}\right)^2\right].\end{aligned} \tag{60}$$

Notice the change in factors δ with respect to Eq. (5), which is due to the different normalization of the design matrix.

By the same argument used in the proof of Lemma 4, Theorem 1 implies that, for any GFOM with output $\hat{\boldsymbol{\theta}}_t$, we have

$$\lim_{n,p\to\infty}\mathbb{E}\frac{|\langle\overline{\boldsymbol{\theta}}_0,\hat{\boldsymbol{\theta}}^t\rangle|}{\|\overline{\boldsymbol{\theta}}_0\|_2\|\hat{\boldsymbol{\theta}}^t\|_2} \le \sqrt{\hat{q}_t}.$$

We next compute the first order Taylor-expansion of the iteration (60), 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 Mondelli and Montanari (2019)). As a consequence, for any $\eta > 0$, there exists α_0 such that, if $\tilde{\alpha} < \alpha_0$, $q_t < \alpha_0$, then

$$q_{t+1} \le (\frac{\delta}{\delta_{\rm sp}} + \eta)(q_t + \tilde{\alpha}).$$

The claim follows by taking $\eta = \eta(\delta) := (\delta_{\rm sp} - \delta)/(2\delta_{\rm sp})$, whence $q_t \le \tilde{\alpha}/\eta(\delta)$ for all t, provided $\tilde{\alpha} < \alpha_* := \alpha_0\eta(\delta)$. The deterministic argument follows in the same way as Corollary 5.

Appendix H. Proximal gradient and modified power iteration as GFOMs

The proximal gradient algorithm (3) is an instance of a general first order method (1) via the change of variables

$$\begin{aligned} \boldsymbol{v}^t &= \boldsymbol{\theta}^t, & \boldsymbol{u}^t &= \boldsymbol{X}\boldsymbol{\theta}^t, & F_t^{(1)}(u^t; y) &= -\gamma_t s(y, u), \\ F_t^{(2)}(v^t) &= v^t, & G_t^{(1)}(v^t) &= v^t, & G_t^{(2)} &= 0. \end{aligned}$$

The modified power iteration algorithm of standard PCA (4) is an instance of a general first order method (1) via the change of variables

$$\begin{aligned} \boldsymbol{v}^t &= \boldsymbol{\theta}^t, & \boldsymbol{u}^t &= \boldsymbol{X}\eta(\boldsymbol{\theta}^t; \gamma^t), & F_t^{(1)}(u^t) &= c_t u^t, \\ F_t^{(2)} &= 0, & G_t^{(1)}(v^t) &= \eta(v^t; \gamma_t), & G_t^{(2)} &= 0\,. \end{aligned}$$