Mixed Linear Regression via Approximate Message Passing

Nelvin Tan
University of Cambridge

Ramji Venkataramanan
University of Cambridge

Abstract

In mixed linear regression, each observation comes from one of L regression vectors (signals), but we do not know which one. The goal is to estimate the signals from the unlabeled observations. We propose a novel approximate message passing (AMP) algorithm for estimation and rigorously characterize its performance in the high-dimensional limit. This characterization is in terms of a state evolution recursion, which allows us to precisely compute performance measures such as the asymptotic mean-squared error. This can be used to tailor the AMP algorithm to take advantage of any known structural information about the signals. Using state evolution, we derive an optimal choice of AMP 'denoising' functions that minimizes the estimation error in each iteration. Numerical simulations are provided to validate the theoretical results, and show that AMP significantly outperforms other estimators including spectral methods, expectation maximization, and alternating minimization. Though our numerical results focus on mixed linear regression, the proposed AMP algorithm can be applied to a broader class of models including mixtures of generalized linear models and max-affine regression.

1 INTRODUCTION

We consider the mixed linear regression problem where we wish to estimate L signal vectors $\beta^{(1)}, \dots, \beta^{(L)} \in \mathbb{R}^p$ from *unlabeled* observations of each. Specifically, the components of the observed vector $Y := (Y_1, \dots, Y_n)$ are generated as:

$$Y_i = \langle X_i, \beta^{(1)} \rangle c_{i1} + \cdots + \langle X_i, \beta^{(L)} \rangle c_{iL} + \epsilon_i, \;\; i \in [n]. \tag{1}$$

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

Here $X_i \in \mathbb{R}^p$ is the ith feature vector, ϵ_i is a noise variable, and $c_{i1}, \dots, c_{iL} \in \{0, 1\}$ are binary-valued latent variables such that $\sum_{l=1}^{L} c_{il} = 1$, for $i \in [n]$. The notation $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product. In words, each observation comes from exactly one of the L signal vectors (regressors), but we do not know which one.

The case of $L = 1$ is standard linear regression, which implicitly assumes a homogeneous population, i.e., a single regression vector captures the population characteristics of the entire sample. However, this assumption may not be realistic in some situations as the sample may contain several sub-populations. Standard linear regression may provide biased estimates in such situations when the population heterogeneity is unobserved. The mixed linear regression (MLR) model is more flexible as it allows for differences in regressors across unobserved sub-populations. MLR has been used for analyzing heterogenous data in a variety of fields including biology, physics, and economics (McLachlan and Peel, 2004; Grün and Leisch, 2007; Li et al., 2019; Devijver et al., 2020).

In the MLR model (1), a natural approach for estimating $\beta^{(1)}, \dots, \beta^{(L)}$ from $\{X_i, Y_i\}_{i=1}^n$ is via the global least-squares estimator given by:

$$\widehat{\beta}^{(1)}, \dots, \widehat{\beta}^{(L)} = \underset{\substack{\beta^{(1)}, \dots, \beta^{(L)} \in \mathbb{R}^p \\ c_1, \dots, c_L \in \{0,1\}^n \\ \sum_{l=1}^{L} c_{il} = 1, \, i \in [n]}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - \sum_{l=1}^{L} \langle X_i, \beta^{(l)} \rangle \, c_{il} \right)^2. \tag{2}$$

However, this optimization problem is non-convex, and computing the global minimum is known to be NP-hard (Yi et al., 2014). A range of alternative approaches has been proposed including estimators based on: spectral methods (Chaganty and Liang, 2013; Yi et al., 2014); expectation-maximization (EM) (Faria and Soromenho, 2010; Städler et al., 2010; Zhang et al., 2020); alternating minimization (Yi et al., 2014; Shen and Sanghavi, 2019; Ghosh and Kannan, 2020); convex relaxation (Chen et al., 2014); moment descent methods (Li and Liang, 2018; Chen et al., 2020); and tractable non-convex objective functions (Zhong et al., 2016; Barik and Honorio, 2022). Most of these techniques are generic, and while some can incorporate certain constraints like sparsity, they are not well-equipped to exploit

specific structural information about $\beta^{(1)}, \dots, \beta^{(L)}$, such as a known prior on the signals. Moreover, these methods are suboptimal with respect to sample complexity: for accurate recovery they require the number of observations n to be at least of order $p \log p$ (Yi et al., 2014; Li and Liang, 2018; Chen et al., 2020). In contrast, here we consider the high-dimensional regime where n is proportional to p and provide *exact* asymptotics for the performance of the proposed estimator.

Approximate message passing (AMP) is a family of iterative algorithms which can be tailored to take advantage of structural information about the signals and the model, e.g., a known prior on the signal vector or on the proportion of observations that come from each signal. AMP algorithms were first proposed for the standard linear model (Kabashima, 2003; Bayati and Montanari, 2011; Donoho et al., 2009; Krzakala et al., 2012; Maleki et al., 2013), but have since been applied to a range of statistical problems, including estimation in generalized linear models (Rangan, 2011; Schniter and Rangan, 2014; Barbier et al., 2019; Ma et al., 2019; Sur and Candès, 2019; Maillard et al., 2020; Mondelli and Venkataramanan, 2021) and low-rank matrix estimation (Deshpande and Montanari, 2014; Fletcher and Rangan, 2018; Kabashima et al., 2016; Lesieur et al., 2017; Montanari and Venkataramanan, 2021; Barbier et al., 2020). In all these settings, under suitable model assumptions the performance of AMP in the high-dimensional limit is characterized by a succinct deterministic recursion called *state evolution*. The state evolution characterization has been used to show that AMP achieves Bayes-optimal performance for some models (Deshpande and Montanari, 2014; Donoho et al., 2013; Montanari and Venkataramanan, 2021; Barbier et al., 2019), and a conjecture from statistical physics states that AMP is optimal among polynomial-time algorithms for a wide range of statistical estimation problems.

In this work, we design and analyze an AMP algorithm for a general regression model, of which mixed linear regression is a special case. Define the signal matrix $B := (\beta^{(1)}, \dots, \beta^{(L)}) \in \mathbb{R}^{p \times L}$, and consider the problem of estimating B from an observed matrix $Y := (Y_1, \dots, Y_n)^\top \in \mathbb{R}^{n \times L_{\text{out}}}$, whose ith row Y_i is generated as:

$$Y_i = q(B^\top X_i \,,\, \Psi_i) \in \mathbb{R}^{L_{\text{out}}} \quad i \in [n]. \tag{3}$$

Here $\Psi \in \mathbb{R}^{n \times L_\Psi}$ is a matrix of unobserved auxiliary variables (with Ψ_i its ith row), and $q : \mathbb{R}^L \times \mathbb{R}^{L_\Psi} \to \mathbb{R}^{L_{\text{out}}}$ is a known function. The model (3) can be viewed as a generalized linear model with matrix-valued signals and auxiliary variables, so we call it the *matrix GLM*.

The MLR model (1) is a special case of the matrix GLM, obtained by taking $\Psi_i = (c_{i,1}, \dots, c_{i,L}, \epsilon_i)$. In addition, (3) includes other popular latent variable models including mixtures of generalized linear models (Khalili and Chen, 2007; Sedghi et al., 2016) and max-affine regression (Ghosh et al., 2022). The AMP and the theoretical results in Section 3 are presented for the matrix GLM, but we focus on mixed linear regression for the numerical simulations. This allows us to easily compare with other approaches such as spectral methods and the EM algorithm, and highlight the interesting features of the proposed AMP.

Main contributions. We propose an AMP algorithm for the matrix GLM (3), under the assumption that the features $\{X_i\}_{i \in [n]}$ are i.i.d. Gaussian. Our first technical contribution is a state evolution result for the AMP algorithm (Theorem 1), which gives a rigorous characterization of its performance in the high-dimensional limit as $n, p \to \infty$ with a fixed ratio $\delta = n/p$, for a constant $\delta > 0$. This allows us to compute exact asymptotic formulas for performance measures such as the mean-squared error (MSE) and the normalized correlation between the signals and their estimates. The AMP algorithm uses a pair of 'denoising' functions to produce updated signal estimates in each iteration. The accuracy of these estimates can be tracked using a signal-to-noise ratio defined in terms of the state evolution parameters. Our second contribution (Proposition 2) is to derive an optimal choice of denoising functions that maximizes this signal-to-noise ratio. The optimal choice for one of the these functions depends on the prior on the signals, while the other depends only on the output function $q(\cdot, \cdot)$ in (3). We present numerical simulation results for the mixed linear regression setting, and show that AMP significantly outperforms other estimators, including those based on spectral methods, alternating minimization, and expectation maximization.

The state evolution performance characterization in Theorem 1 is proved using a change of variables that maps the proposed algorithm to an abstract AMP recursion with matrix-valued iterates. A state evolution characterization for this abstract AMP was established by Javanmard and Montanari (2013); this result is translated via the change of variables to obtain the state evolution characterization for the proposed AMP. Though our AMP algorithm and its analysis assume i.i.d. Gaussian features, we expect that they can be extended to a much broader class of i.i.d. designs using the recent universality results of Wang et al. (2022). Another exciting direction for future work is to generalize the AMP algorithm and its state evolution to the mixed regression models with rotationally invariant design matrices. This can be done via a reduction to an abstract AMP recursion for rotationally invariant matrices, similar to the ones studied in (Fan, 2022) and Zhong et al. (2021).

Other related work. Mixtures of generalized linear models have been studied in machine learning under the name 'hierarchical mixtures of experts', see e.g., (Jordan and Jacobs, 1994). Bayesian methods for inference in this model were investigated by Peng et al. (1996) and Waterhouse et al. (1995), and Bayesian inference for the special

case of MLR was analyzed by Viele and Tong (2002). Balakrishnan et al. (2017) and Klusowski et al. (2019) obtained statistical guarantees on the performance of the EM algorithm for a class of problems, including the special case of symmetric mixed linear regression where $\beta^{(1)} = -\beta^{(2)}$. Variants of the EM algorithm for symmetric MLR in the high-dimensional setting (with sparse signals) were analyzed by Wang et al. (2015),Yi and Caramanis (2015), and Zhu et al. (2017). Fan et al. (2018) obtained minimax lower bounds for a class of computationally feasible algorithms for symmetric MLR.

Kong et al. (2020) studied MLR as a canonical example of meta-learning. They consider the setting where the number of signals (L) is large, and derive conditions under which a large number of signals with a few observations can compensate for the lack of signals with abundantly many observations. The special case of MLR with sparse signals was studied by Krishnamurthy et al. (2019) and Pal et al. (2021). Pal et al. (2022) analyzed the prediction error of MLR in the non-realizable setting, where no generative model is assumed for the data. Chandrasekher et al. (2021) recently analyzed the performance of a class of iterative algorithms (not including AMP) for mixtures of GLMs. They provide a sharp characterization of the per-iteration error with sample-splitting in the regime $n \sim p\,\mathrm{polylog}(p)$, assuming a Gaussian design and a random initialization.

2 PRELIMINARIES

Notation. We write $[n]$ for the set $\{1, \dots, n\}$. All vectors (even rows of matrices) are assumed to be column vectors unless otherwise stated. Matrices are denoted by upper case letters, and given a matrix A, we write A_i for its ith row. The notation $M \succeq 0$ denotes that the square matrix M is positive semidefinite. We write I_p for the $p \times p$ identity matrix. For $r \in [1, \infty)$, we write $\|x\|_r$ for the ℓ_r-norm of $x = (x_1, \dots, x_n) \in \mathbb{R}^n$, so that $\|x\|_r = \big(\sum_{i=1}^n |x_i|^r\big)^{1/r}$. Given random variables U, V, we write $U \overset{d}{=} V$ to denote equality in distribution.

Complete convergence. The asymptotic results in this paper are stated in terms of *complete convergence* (Hsu and Robbins, 1947), (Feng et al., 2022, Sec. 1.1). This is a stronger mode of stochastic convergence than almost sure convergence, and is denoted using the symbol $\overset{c}{\to}$. Let $\{X_n\}$ be a sequence of random elements taking values in a Euclidean space E. We say that X_n converges completely to a deterministic limit $x \in E$, and write $X_n \overset{c}{\to} x$, if $Y_n \to x$ almost surely for any sequence of E-valued random elements $\{Y_n\}$ with $Y_n \overset{d}{=} X_n$ for all n.

Wasserstein distances. For $D \in \mathbb{N}$, let $\mathcal{P}_D(r)$ be the set of all Borel probability measures on $\mathbb{R}^D$ with finite rth-moment. That is, any $P \in \mathcal{P}_D(r)$ satisfies $\int_{\mathbb{R}^D} \|x\|_2^r dP(x) < \infty$. For $P, Q \in \mathcal{P}_D(r)$, the *r-Wasserstein distance* between P and Q is defined by $d_r(P, Q) = \inf_{(X,Y)} \mathbb{E}[\|X - Y\|_2^r]^{1/r}$, where the infimum is taken over all pairs of random vectors (X, Y) defined on a common probability space with $X \sim P$ and $Y \sim Q$.

Model assumptions. In the model (3), each feature vector $X_i \in \mathbb{R}^p$ is assumed to have independent Gaussian entries with zero mean and variance $1/n$, i.e., $X_i \sim_{\text{i.i.d.}} \mathcal{N}(0, I_p/n)$. The $n \times p$ design matrix X is formed by stacking the feature vectors $X_1, \dots, X_n$, i.e., $X = [X_1, \dots, X_n]^\top$. The matrix X is independent of both the signal matrix $B = (\beta^{(1)}, \dots, \beta^{(L)}) \in \mathbb{R}^{p \times L}$ and the auxiliary variable matrix $\Psi \in \mathbb{R}^{n \times L_\Psi}$.

As $p \to \infty$, we assume that $n/p = \delta$, for some constant $\delta > 0$. As $p \to \infty$, the empirical distributions of the rows of the signal matrix and the auxiliary variable matrix are assumed to converge in Wasserstein distance to well-defined limits. More precisely, for some $r \in [2, \infty)$, there exist random variables $\bar{B} \sim P_{\bar{B}}$ (where $\bar{B} \in \mathbb{R}^L$) and $\bar{\Psi} \sim P_{\bar{\Psi}}$ (where $\bar{\Psi} \in \mathbb{R}^{L_\Psi}$) with $\mathbb{E}[\bar{B}^\top \bar{B}] > 0$ and $\mathbb{E}\big[\sum_{l=1}^L |\bar{B}_l|^r\big], \mathbb{E}\big[\sum_{l=1}^{L_\Psi} |\bar{\Psi}_l|^r\big] < \infty$, such that writing $\nu_p(B)$ and $\nu_n(\Psi)$ for the empirical distributions of the rows of B and Ψ respectively, we have $d_r(\nu_p(B), P_{\bar{B}}) \overset{c}{\to} 0$ and $d_r(\nu_n(\Psi), P_{\bar{\Psi}}) \overset{c}{\to} 0$.

3 AMP FOR THE MATRIX-GLM

Consider the task of estimating the signal matrix B given $\{X_i, Y_i\}_{i \in [n]}$, generated according to (3).

Algorithm. In each iteration $k \geq 1$, the AMP algorithm iteratively produces estimates $\widehat{B}^k$ and Θ^k of $B \in \mathbb{R}^{p \times L}$ and $\Theta := XB \in \mathbb{R}^{n \times L}$, respectively. Starting with an initializer $\widehat{B}^0 \in \mathbb{R}^{p \times L}$ and defining $\widehat{R}^{-1} := 0 \in \mathbb{R}^{n \times L}$, for $k \geq 0$ the algorithm computes:

$$\Theta^k = X\widehat{B}^k - \widehat{R}^{k-1}(F^k)^\top, \quad \widehat{R}^k = g_k(\Theta^k, Y), \qquad (4)$$
$$B^{k+1} = X^\top \widehat{R}^k - \widehat{B}^k(C^k)^\top, \quad \widehat{B}^{k+1} = f_{k+1}(B^{k+1}).$$

Here the functions $g_k : \mathbb{R}^L \times \mathbb{R}^{L_{\text{out}}} \to \mathbb{R}^L$ and $f_{k+1} : \mathbb{R}^L \to \mathbb{R}^L$ act row-wise on their matrix inputs, and the matrices $C^k, F^{k+1} \in \mathbb{R}^{L \times L}$ are defined as

$$C^k = \frac{1}{n} \sum_{i=1}^n g_k'(\Theta_i^k, Y_i), \quad F^{k+1} = \frac{1}{n} \sum_{j=1}^p f_{k+1}'(B_j^{k+1}),$$

where g_k', f_{k+1}' denote the Jacobians of g_k, f_{k+1}, respectively, with respect to their first arguments. We note that the time complexity of each iteration of (4) is $\mathcal{O}(npL)$.

State evolution. The "memory" terms $-\widehat{R}^{k-1}(F^k)^\top$ and $-\widehat{B}^k(C^k)^\top$ in (4) play a crucial role in debiasing the iterates Θ^k and B^{k+1}, ensuring that their joint empirical

distributions are accurately captured by state evolution in the high-dimensional limit. Theorem 1 below shows that for each $k \geq 1$, the empirical distribution of the rows of B^k converges to the distribution of $\mathrm{M}_B^k \bar{B} + G_B^k \in \mathbb{R}^L$, where $G_B^k \sim \mathcal{N}(0, \mathrm{T}_B^k)$ is independent of $\bar{B}$, the random variable representing the limiting distribution of the rows of the signal matrix B. The deterministic matrices $\mathrm{M}_B^k, \mathrm{T}_B^k \in \mathbb{R}^{L \times L}$ are recursively defined below. The result implies that the empirical distribution of the rows of $\widehat{B}^k$ converges to the distribution of $f_k(\mathrm{M}_B^k \bar{B} + G_B^k)$. Thus f_k can be viewed as a denoising function that can be tailored to take advantage of the prior on $\bar{B}$. Theorem 1 also shows that the empirical distribution of the rows of Θ^k converges to the distribution of $\mathrm{M}_\Theta^k Z + G_\Theta^k \in \mathbb{R}^L$, where $Z \sim \mathcal{N}(0, \frac{1}{\delta}\mathbb{E}[\bar{B}\bar{B}^\top])$ and $G_\Theta^k \sim \mathcal{N}(0, \mathrm{T}_\Theta^k)$ are independent.

We now describe the state evolution recursion defining the matrices $\mathrm{M}_\Theta^k, \mathrm{T}_\Theta^k, \mathrm{M}_B^k, \mathrm{T}_B^k \in \mathbb{R}^{L \times L}$. Recalling that the observation Y is generated via the function q according to (3), it is convenient to rewrite g_k in (4) in terms of another function $h_k : \mathbb{R}^L \times \mathbb{R}^L \times \mathbb{R}^{L_\Psi} \to \mathbb{R}^L$ defined as:

$$h_k(z, u, v) := g_k(u,\, q(z, v)). \tag{5}$$

Then, for $k \geq 0$, given $\Sigma^k \in \mathbb{R}^{2L \times 2L}$, take $\begin{pmatrix} Z \\ Z^k \end{pmatrix} \sim \mathcal{N}(0, \Sigma^k)$ to be independent of $\bar{\Psi} \sim P_{\bar{\Psi}}$ and compute:

$$\mathrm{M}_B^{k+1} = \mathbb{E}[\partial_Z h_k(Z, Z^k, \bar{\Psi})], \tag{6}$$

$$\mathrm{T}_B^{k+1} = \mathbb{E}[h_k(Z, Z^k, \bar{\Psi}) h_k(Z, Z^k, \bar{\Psi})^\top], \tag{7}$$

$$\Sigma^{k+1} = \begin{pmatrix} \Sigma_{(11)}^{k+1} & \Sigma_{(12)}^{k+1} \\ \Sigma_{(21)}^{k+1} & \Sigma_{(22)}^{k+1} \end{pmatrix}, \tag{8}$$

where the four $L \times L$ matrices constituting $\Sigma^{k+1} \in \mathbb{R}^{2L \times 2L}$ are given by:

$$\Sigma_{(11)}^{k+1} = \frac{1}{\delta}\mathbb{E}[\bar{B}\bar{B}^\top], \tag{9}$$

$$\Sigma_{(12)}^{k+1} = \left(\Sigma_{(21)}^{k+1}\right)^\top = \frac{1}{\delta}\mathbb{E}[\bar{B} f_{k+1}(\mathrm{M}_B^{k+1}\bar{B} + G_B^{k+1})^\top],$$

$$\Sigma_{(22)}^{k+1} = \frac{1}{\delta}\mathbb{E}[f_{k+1}(\mathrm{M}_B^{k+1}\bar{B} + G_B^{k+1}) f_{k+1}(\mathrm{M}_B^{k+1}\bar{B} + G_B^{k+1})^\top].$$

Here we take $G_B^{k+1} \sim \mathcal{N}(0, \mathrm{T}_B^{k+1})$ to be independent of $\bar{B} \sim P_{\bar{B}}$. Note that $\partial_Z h_k$ denotes the partial derivative (Jacobian) of h_k with respect to its first argument $Z \in \mathbb{R}^L$, so it is an $L \times L$ matrix. The state evolution recursion (6)-(8) is initialized with $\Sigma^0 \in \mathbb{R}^{2L \times 2L}$ defined below in (13).

For $\begin{pmatrix} Z \\ Z^k \end{pmatrix} \sim \mathcal{N}(0, \Sigma^k)$, using standard properties of Gaussian random vectors, we have

$$(Z, Z^k, \bar{\Psi}) \overset{d}{=} (Z, \mathrm{M}_\Theta^k Z + G_\Theta^k, \bar{\Psi}), \tag{10}$$

where $G_\Theta^k \sim \mathcal{N}(0, \mathrm{T}_\Theta^k)$ is independent of Z, with

$$\mathrm{M}_\Theta^k = \Sigma_{(21)}^k \left(\Sigma_{(11)}^k\right)^{-1}, \tag{11}$$

$$\mathrm{T}_\Theta^k = \Sigma_{(22)}^k - \Sigma_{(21)}^k \left(\Sigma_{(11)}^k\right)^{-1} \Sigma_{(12)}^k. \tag{12}$$

Main result. We begin with two assumptions required for the main result. The first is on the AMP initializer $\widehat{B}^0 \in \mathbb{R}^{p \times L}$, and the second is on the functions g_k, f_{k+1} used to define the AMP in (4).

(A1) There exists $\Sigma^0 \in \mathbb{R}^{L \times L}$ and $c_0 \in \mathbb{R}$ such that as $n, p \to \infty$ (with $n/p \to \delta$), we have

$$\frac{1}{n}\begin{pmatrix} B^\top B & B^\top \widehat{B}^0 \\ (\widehat{B}^0)^\top B & (\widehat{B}^0)^\top \widehat{B}^0 \end{pmatrix} \overset{c}{\to} \Sigma^0, \tag{13}$$

$$\frac{1}{p}\sum_{j=1}^{p}\sum_{l=1}^{L} |\widehat{B}_{jl}^0|^r \overset{c}{\to} c_0.$$

Here $r \in [2, \infty)$ is the same as that used for the assumptions on the signal matrix at the end of Section 2. Furthermore, there exists a Lipschitz $F_0 : \mathbb{R}^L \to \mathbb{R}^L$ such that $\frac{1}{p}(\widehat{B}^0)^\top \phi(B) \overset{c}{\to} \mathbb{E}[F_0(\bar{B})\phi(\bar{B})^\top]$ and $\Sigma_0^{(22)} - \mathbb{E}[F_0(\bar{B})F_0(\bar{B})^\top]$ is positive semi-definite for all Lipschitz $\phi : \mathbb{R}^L \to \mathbb{R}^L$.

(A2) For $k \geq 0$, the function f_{k+1} is non-constant and Lipschitz on $\mathbb{R}^L$, and h_k defined in (5) is Lipschitz on $\mathbb{R}^{2L+L_\Psi}$ with $P_{\bar{\Psi}}(\{v : (z, u) \to h_k(z, u, v) \text{ is a non-constant}\}) > 0$. Furthermore, f'_{k+1} is continuous Lebesgue almost everywhere, and writing $\mathcal{D}_k \subseteq \mathbb{R}^{L+L_\Psi}$ for the set of discontinuities of g'_k, we have $\mathbb{P}[(Z^k, \bar{Y}) \in \mathcal{D}_k] = 0$.

Assumptions **(A1)** and **(A2)** are similar to those required for AMP initialization in (non-mixed) generalized linear models (Feng et al., 2022, Section 4). Moreover, **(A1)** is implied by the assumptions on the signal matrix if an initialization $\widehat{B}^0$ is chosen to be a scaled version of the all ones matrix.

The result is stated in terms of *pseudo-Lipschitz* test functions. Let $\mathrm{PL}_m(r, C)$ be the set of functions $\phi : \mathbb{R}^m \to \mathbb{R}$ such that $|\phi(x) - \phi(y)| \leq C(1 + \|x\|_2^{r-1} + \|y\|_2^{r-1})\|x - y\|_2$ for all $x, y \in \mathbb{R}^m$. A function $\phi \in \mathrm{PL}_m(r, C)$ is called pseudo-Lipschitz of order r.

Theorem 1. *Consider the AMP in* (4) *for the general mixed model in* (3)*. Suppose that the model assumptions in Section 2 as well as* ***(A1)*** *and* ***(A2)*** *are satisfied, and that* T_B^1

is positive definite. Then for each $k \geq 0$, we have

$$\sup_{\phi \in \mathrm{PL}_{2L}(r,1)} \Big| \frac{1}{p} \sum_{j=1}^{p} \phi(B_j^{k+1}, B_j) - \mathbb{E}[\phi(\mathrm{M}_B^{k+1} \bar{B} + G_B^{k+1}, \bar{B})] \Big| \xrightarrow{c} 0, \quad (14)$$

$$\sup_{\phi \in \mathrm{PL}_{2L+L_\Psi}(r,1)} \Big| \frac{1}{n} \sum_{i=1}^{n} \phi(\Theta_i^k, \Theta_i, \Psi_i) - \mathbb{E}[\phi(\mathrm{M}_\Theta^k Z + G_\Theta^k, Z, \bar{\Psi})] \Big| \xrightarrow{c} 0, \quad (15)$$

as $n, p \to \infty$ with $n/p \to \delta$, where $\Theta_i = B^\top X_i$ for $1 \leq i \leq n$. In the above, $G_B^{k+1} \sim \mathcal{N}(0, \mathrm{T}_B^{k+1})$ is independent of $\bar{B}$, and $G_\Theta^k \sim \mathcal{N}(0, \mathrm{T}_\Theta^k)$ is independent of $(Z, \bar{\Psi})$.

The proof of the theorem is given in Section 5.1.

The result (14) is equivalent to the statement that the joint empirical distributions of the rows of $(B^{k+1},\, B)$ converges completely in r-Wasserstein distance to the joint distribution of $(\mathrm{M}_B^{k+1}\bar{B} + G_B^{k+1},\, \bar{B})$; see (Feng et al., 2022, Corollary 7.21). An analogous statement holds for (15).

Performance measures. Theorem 1 allows us to compute the limiting values of performance measures such as the mean-squared error (MSE), and the normalized correlation between each signal and its AMP estimate. For $k \geq 1$, writing $\hat{\beta}^{(\ell),k}$ for the ℓth column of the AMP iterate $\widehat{B}^k$, we have $\widehat{B}^k = \big(\hat{\beta}^{(1),k}, \ldots, \hat{\beta}^{(L),k}\big)$. Note that $\hat{\beta}^{(\ell),k}$ is the estimate of the signal $\beta^{(\ell)}$ after k iterations. We also define the shorthand $\bar{B}^k := \mathrm{M}_B^k \bar{B} + G_k^B$. Then Theorem 1 implies that the normalized squared correlation between each signal and its AMP estimate after k iterations converges as:

$$\frac{\langle \hat{\beta}^{(\ell),k}, \beta^{(\ell)} \rangle^2}{\|\hat{\beta}^{(\ell),k}\|_2^2 \|\beta^{(\ell)}\|_2^2} \xrightarrow{c} \frac{(\mathbb{E}[f_{k,\ell}(\bar{B}^k)\bar{B}_\ell])^2}{\mathbb{E}[f_{k,\ell}(\bar{B}^k)^2]\mathbb{E}[\bar{B}_\ell^2]}, \quad \ell \in [L]. \quad (16)$$

Here $f_{k,\ell}$ is the ℓth component of the function $f_k : \mathbb{R}^L \to \mathbb{R}^L$, and $\bar{B}_\ell$ is the ℓth component of $\bar{B} \in \mathbb{R}^L$. Similarly, the MSE of the AMP estimate after k iterations converges as:

$$\frac{\|\beta^{(\ell)} - \hat{\beta}^{(\ell),k}\|_2^2}{p} \xrightarrow{c} \mathbb{E}\Big[\big(\bar{B}_\ell - f_{k,\ell}(\bar{B}^k)\big)^2\Big], \quad \ell \in [L]. \quad (17)$$

3.1 Choosing the Functions f_k and g_k

Recalling that the empirical distributions of the rows of Θ^k and B^{k+1} converge to the laws of $\mathrm{M}_\Theta^k Z + G_\Theta^k$ and $\mathrm{M}_B^{k+1}\bar{B} + G_B^{k+1}$, respectively, we define the random vectors:

$$\widetilde{Z}^k := Z + \big(\mathrm{M}_\Theta^k\big)^{-1} G_\Theta^k, \qquad \widetilde{B}^{k+1} := \bar{B} + \big(\mathrm{M}_B^{k+1}\big)^{-1} G_B^{k+1}. \quad (18)$$

(If the inverse doesn't exist we premultiply by the pseudoinverse.) Since $G_B^{k+1} \sim \mathcal{N}(0, \mathrm{T}_B^{k+1})$ and $G_\Theta^k \sim \mathcal{N}(0, \mathrm{T}_\Theta^k)$, the effective noise covariance matrices are:

$$\mathrm{cov}(\widetilde{Z}^k - Z) = \big(\mathrm{M}_\Theta^k\big)^{-1} \mathrm{T}_\Theta^k \Big(\big(\mathrm{M}_\Theta^k\big)^{-1}\Big)^\top =: N_\Theta^k,$$
$$\mathrm{cov}(\widetilde{B}^{k+1} - \bar{B}) = \big(\mathrm{M}_B^{k+1}\big)^{-1} \mathrm{T}_B^{k+1} \Big(\big(\mathrm{M}_B^{k+1}\big)^{-1}\Big)^\top =: N_B^{k+1}. \quad (19)$$

From (12), we observe that $\mathrm{M}_\Theta^k, \mathrm{T}_\Theta^k$ are both determined by Σ^k, which in turn is determined by the choice of f_k (from (9)). Similarly, from (5)-(6), M_B^{k+1}, T_B^{k+1} are determined by g_k. A natural objective is to choose f_k and g_k to minimize the trace of the effective noise covariance matrices N_Θ^k and N_B^{k+1} in (19). We can interpret the quantities $\mathrm{Tr}(N_\Theta^k)$ and $\mathrm{Tr}(N_B^{k+1})$ as the effective noise variances for estimating $Z, \bar{B}$ from $\tilde{Z}^k, \tilde{B}^{k+1}$, respectively. In the special case where there is only one signal, minimizing these effective noise variances is equivalent to maximizing the scalar signal-to-noise ratios $(\mathrm{M}_\Theta^k)^2/\mathrm{T}_\Theta^k$ and $(\mathrm{M}_B^{k+1})^2/\mathrm{T}_B^{k+1}$, respectively, which is achieved by the Bayes-optimal AMP for generalized linear models (Rangan, 2011; Feng et al., 2022).

Assuming that the signal prior $P_{\bar{B}}$ and the distribution of auxiliary variables P_Ψ are known, the following proposition gives optimal choices for f_k, g_k.

Proposition 2. *Let $k \geq 1$. Then:*

1) Given M_B^k, T_B^k, the quantity $\mathrm{Tr}(N_\Theta^k)$ is minimized when $f_k = f_k^*$, where

$$f_k^*(s) = \mathbb{E}[\bar{B} \mid \mathrm{M}_B^k \bar{B} + G_B^k = s], \quad (20)$$

where $G_B^k \sim \mathcal{N}(0, \mathrm{T}_B^k)$ and $\bar{B} \sim P_{\bar{B}}$ are independent.

2) Given M_Θ^k, T_Θ^k, the quantity $\mathrm{Tr}(N_B^{k+1})$ is minimized when $g_k = g_k^*$, where

$$g_k^*(u,\, y) = \mathrm{Cov}[Z \mid Z^k = u]^{-1}\big(\mathbb{E}[Z \mid Z^k = u, \bar{Y} = y] - \mathbb{E}[Z \mid Z^k = u]\big). \quad (21)$$

Here $\begin{pmatrix} Z \\ Z^k \end{pmatrix} \sim \mathcal{N}(0, \Sigma^k)$ and $\bar{Y} = q(Z, \bar{\Psi})$, with $\bar{\Psi} \sim P_{\bar{\Psi}}$ independent of Z.

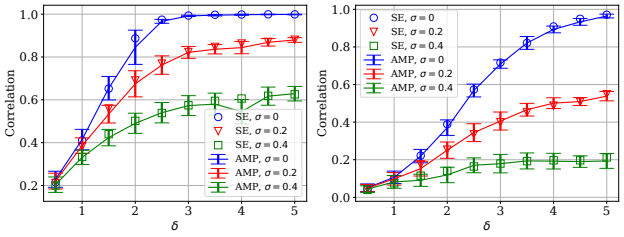
The proof is given in Section 5.2.

4 NUMERICAL SIMULATIONS

We first focus on the MLR model (1) with two signals, where for $i \in [n]$,

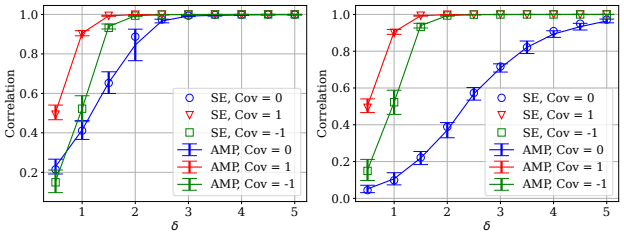
$$Y_i = \langle X_i, \beta^{(1)} \rangle c_i + \langle X_i, \beta^{(2)} \rangle (1 - c_i) + \epsilon_i. \quad (22)$$

We take $c_i \sim_{\text{i.i.d.}} \text{Bernoulli}(\alpha)$ for $\alpha \in (0,1)$, $\epsilon_i \sim_{\text{i.i.d.}} \mathcal{N}(0, \sigma^2)$, and $X_i \sim_{\text{i.i.d.}} \mathcal{N}(0, I_p/n)$, for $i \in [n]$. We set



(a) $\beta^{(1)}$ (b) $\beta^{(2)}$

Figure 1: Gaussian prior with $\rho = 0$: normalized squared correlation vs. δ for various noise levels σ, with $\alpha = 0.7$.



(a) $\beta^{(1)}$ (b) $\beta^{(2)}$

Figure 2: Gaussian prior with different values of signal covariance ρ: Normalized squared correlation vs. δ, with $\alpha = 0.7$, $\sigma = 0$.

the signal dimension $p = 500$ and vary the value of n in our experiments.

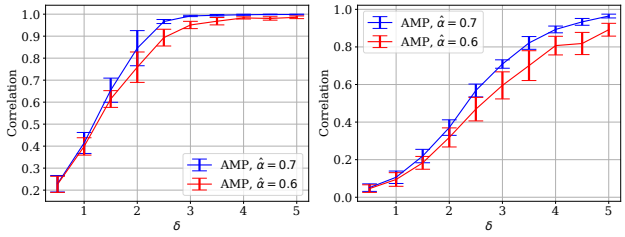
The AMP algorithm in (4) is implemented with $g_k = g_k^*$, the optimal choice given by (21). For the function f_k, we use the Bayes-optimal f_k^* in (20) unless stated otherwise. The performance in all the plots is measured via the normalized squared correlation between the AMP estimate and the signal (see (16)). Each point on the plots is obtained from 10 independent runs, where in each run, AMP is executed for 10 iterations. We report the average and error bars at 1 standard deviation of the final iteration. Additional implementation details are given in Appendix A.

Gaussian prior. In Figures 1, 2, and 3, we set the Bernoulli parameter $\alpha = 0.7$ and choose the two signals to be jointly Gaussian, with their entries generated as

$$(\beta_i^{(1)}, \beta_i^{(2)}) \sim_{\text{i.i.d.}} \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right), \quad i \in [p]. \quad (23)$$

The initializer $\widehat{B}^0 \in \mathbb{R}^{p\times 2}$ is chosen randomly according to the same distribution, independently of the signal.

Figure 1 shows the performance of AMP for independent signals ($\rho = 0$). The normalized squared correlation is plotted as a function of the sampling ratio $\delta = \frac{n}{p}$, for different noise levels σ. The state evolution predictions closely match the performance of AMP for practical values of n, p,



(a) $\beta^{(1)}$ (b) $\beta^{(2)}$

Figure 3: Gaussian prior with $\rho = 0$ and different values of estimated proportion $\hat{\alpha}$: Normalized squared correlation vs. δ, with true $\alpha = 0.7$, $\sigma = 0$.

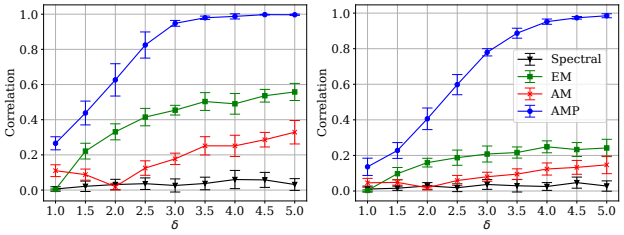
validating the result of Theorem 1. As expected, the correlation improves with increasing δ and degrades with increasing σ. The performance for $\beta^{(1)}$ is better than for $\beta^{(2)}$ as 70% of the observations come from $\beta^{(1)}$. Figure 2 plots the performance as a function of δ for signal correlation $\rho \in \{0, 1, -1\}$, with $\sigma = 0$ (noiseless). When $\rho = 1$, both signals are identical and the problem reduces to standard linear regression. When $\rho = -1$, we have $\beta^{(2)} = -\beta^{(1)}$, so there is still effectively only one signal vector. However, the $\rho = -1$ case is harder than $\rho = 1$ since each measurement is unlabelled and could come from either $\beta^{(1)}$ or $-\beta^{(1)}$ (with probabilities 0.7 and 0.3, respectively).

In practical applications, we may not know the exact proportion of observations that come that come from the first signal. Figure 3 shows the performance when AMP is run assuming a proportion parameter $\hat{\alpha} = 0.6$ which is different from the true value $\alpha = 0.7$. The functions f_k^*, g_k^* defining the AMP depend on α, hence replacing α with $\hat{\alpha}$ in these functions is effectively running AMP with a different (sub-optimal) choice of denoising functions.

Comparison with other estimators. Figure 4 compares the performance of AMP with other widely studied estimators for mixed linear regression, for the Gaussian signal prior in (23) with independent signals ($\rho = 0$). The other estimators are: the spectral estimator proposed in (Yi et al., 2014, Algorithm 2); alternating minimization (AM) (Yi et al., 2014, Algorithm 1); and expectation maximization (EM) (Faria and Soromenho, 2010, Section 2.1). Figure 5 compares the performance of AMP with these estimators for a sparse signal prior given by:

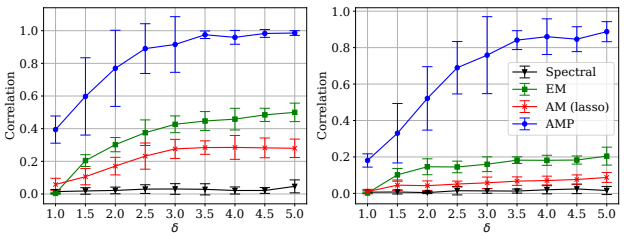
$$(\beta_i^{(1)}, \beta_i^{(2)}) \sim_{\text{i.i.d.}} (0.9)\,\delta_0 + (0.05)\,\delta_{+1} + (0.05)\,\delta_{-1}, \quad i \in [p]. \quad (24)$$

For this prior, we modified the least squares step of the AM algorithm in (Yi et al., 2014, Algorithm 2) to use Lasso instead of standard least squares – this gives better performance as it takes advantage of the signal sparsity. We also tried using the lasso-type EM algorithm (Städler et al.,



(a) $\beta^{(1)}$ (b) $\beta^{(2)}$

Figure 4: Comparison of different estimators for Gaussian prior with $\rho = 0$: Normalized squared correlation vs. δ, with $\alpha = 0.6$, $\sigma = 0$.



(a) $\beta^{(1)}$ (b) $\beta^{(2)}$

Figure 5: Comparison of different estimators for sparse prior: Normalized squared correlation vs. δ, with $\alpha = 0.6$, $\sigma = 0.1$.

2010); however, it was not found to give a noticeable improvement in performance. In both setups, AMP significantly outperforms the other estimators as it is tailored to take advantage of the signal prior via the choice of the denoising function f_k.

Three signals. To illustrate AMP's ability to tackle MLR with more than two signals, we now consider the model (1) with three signals:

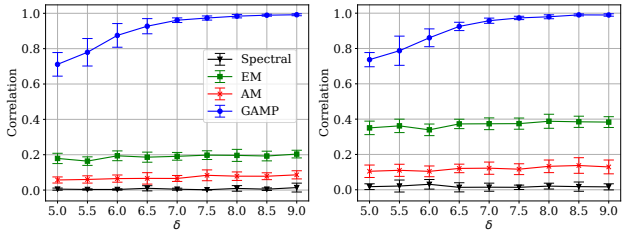
$$Y_i = \langle X_i, \beta^{(1)} \rangle c_{i1} + \langle X_i, \beta^{(2)} \rangle c_{i2} + \langle X_i, \beta^{(3)} \rangle c_{i3} + \epsilon_i, \quad i \in [n]. \tag{25}$$

We take $[c_{i1}, c_{i2}, c_{i3}]^\top$ to be a one-hot vector, where we denote the position of the one in the one-hot vector to be $c_i \sim_{\text{i.i.d.}}$ Categorical$(\{\alpha_1, \alpha_2, \alpha_3\})$, $\epsilon_i \sim_{\text{i.i.d.}} \mathcal{N}(0, \sigma^2)$, and $X_i \sim_{\text{i.i.d.}} \mathcal{N}(0, I_p/n)$, for $i \in [n]$. We set the signal dimension $p = 500$ and vary the value of n in our experiments. The AMP algorithm in (4) is implemented with $g_k = g_k^*$ and $f_k = f_k^*$ (i.e., the optimal choices).

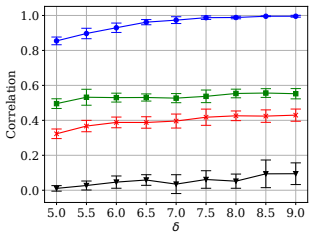
We use independent Gaussian priors for the three signals. Specifically, we generate:

$$(\beta_j^{(1)}, \beta_j^{(2)}, \beta_j^{(3)}) \sim_{\text{i.i.d.}} \mathcal{N}(\mathbb{E}[\bar{B}], I_3), \quad j \in [p] \tag{26}$$

$$c_i \sim_{\text{i.i.d.}} \text{Categorical}(\{\alpha_1, \alpha_2, \alpha_3\}), \quad i \in [n]. \tag{27}$$



(a) $\beta^{(1)}$ (b) $\beta^{(2)}$



(c) $\beta^{(3)}$

Figure 6: MLR with three signals. Comparison of different estimators for Gaussian prior: Normalized squared correlation vs. δ, with $\sigma = 0$.

The initializer $\widehat{B}^0 \in \mathbb{R}^{p \times 3}$ is chosen randomly according to the same distribution, independent of the signal. Figure 6 compares the performance of AMP with other MLR estimators, for the noiseless case, i.e., $\sigma = 0$. The parameters for the Gaussian signal prior in (26)–(27) were chosen to be $\mathbb{E}[\bar{B}] = [0, 0.5, 1]^\top$ and $(\alpha_1, \alpha_2, \alpha_3) = (1/3, 1/3, 1/3)$. (This is the case where signals have different means but appear in the same proportion of observations.) We modified the grid search[1] step of the spectral estimator in (Yi et al., 2014, Algorithm 2) to sample evenly across a sphere instead of a circle (the reason being that we now have three signals instead of two). Since this step cannot be done exactly like in the 2D case, we used the Fibonacci sphere algorithm by Álvaro González (2010) to achieve this approximately and efficiently in our 3D case. As in the case of two-signal MLR, AMP is shown to significantly outperform the other estimators since it is tailored to take advantage of the signal prior via the choice of the denoising function f_k.

5 PROOFS

5.1 Proof of Theorem 1

To prove the theorem, we use a change of variables to rewrite (4) as a new matrix-valued AMP iteration. The new iteration is a special case of an abstract AMP iteration for

[1] In the two signal case, grid search was used to iterate over all possible combinations of the top two eigenvectors of $\frac{1}{n}\sum_{i=1}^n Y_i X_i X_i^\top$ to get the best combination for each signal.

which a state evolution result has been established by Javanmard and Montanari (2013). This state evolution result is then translated to obtain the results in (14)-(15).

Given the iteration (4), for $k \geq 0$ define

$$\check{B}^{k+1} := B^{k+1} - B(\mathrm{M}_B^{k+1})^\top, \quad \check{\Theta}^k := (\Theta, \Theta^k), \quad (28)$$

where we recall that $\Theta = XB$. For $k \geq 0$, we also define the function $\check{f}_k : \mathbb{R}^{2L} \to \mathbb{R}^{2L}$:

$$\check{f}_k(\check{B}^k, B) = (B,\ f_k(\check{B}^k + B(\mathrm{M}_B^k)^\top)). \quad (29)$$

Then, we claim that the original AMP iteration (4) is equivalent to the following one:

$$\begin{aligned} \check{\Theta}^k &= X\check{f}_k(\check{B}^k, B) \ - \ h_{k-1}(\check{\Theta}^{k-1}, \Psi)(\check{F}^k)^\top \\ \check{B}^{k+1} &= X^\top h_k(\check{\Theta}^k, \Psi) \ - \ \check{f}_k(\check{B}^k, B)(\check{C}^k)^\top, \end{aligned} \quad (30)$$

where h_k is defined in (5), and the matrices $\check{C}^k \in \mathbb{R}^{L\times 2L}$, $\check{F}^{k+1} \in \mathbb{R}^{2L\times L}$ are defined as:

$$\begin{aligned} \check{C}^k &= \big(\mathbb{E}[\partial_Z h_k(Z, Z^k, \bar{\Psi})] \ \tfrac{1}{n}\textstyle\sum_{i=1}^n \partial_{\Theta_i^k} h_k(\Theta_i, \Theta_i^k, \Psi_i)\big) \\ \check{F}^{k+1} &= \begin{pmatrix} 0_{L\times L} \\ \frac{1}{n}\sum_{j=1}^p f'_{k+1}(\check{B}_j^k + B_j(\mathrm{M}_B^k)^\top) \end{pmatrix}. \end{aligned} \quad (31)$$

The iteration (30) is initialized with $\check{\Theta}^0 = (\Theta,\ X\widehat{B}^0)$, where $\widehat{B}^0$ is the initializer of the original AMP. The equivalence between the iteration in (30) and the original AMP in (4) can be seen by substituting the definitions (28) and (29) into (30), and recalling from (6) that $\mathrm{M}_B^{k+1} = \mathbb{E}[\partial_Z h_k(Z, Z^k, \bar{\Psi})]$.

The key difference between the new iteration in (30) and the original AMP (4) is that in (30), in addition to the previous iterate, the inputs to the functions $\check{f}_k$ and h_k are auxiliary variables (B, Ψ, respectively) that are independent of X. This is in contrast to the AMP in (4) where the input Y to the function g_k is not independent of X. The recursion in (30) is a special case of an abstract AMP recursion with matrix-valued iterates for which a state evolution result has been established by Javanmard and Montanari (2013). (We will use a version of the result described in (Feng et al., 2022, Sec. 6.7)).) The standard form of the abstract AMP recursion uses empirical estimates (instead of expected values) for the first two entries of $\check{C}^k$ in (31). However, the state evolution result remains valid for the recursion (30) (see Remark 4.3 of Feng et al. (2022)). This result states that the empirical distributions of the rows of $\check{\Theta}^k$ and $\check{B}^{k+1}$ converge to the Gaussian distributions $\mathcal{N}(0, \check{\Sigma}^k)$ and $\mathcal{N}(0, \check{\mathrm{T}}^{k+1})$, respectively. The deterministic covariance matrices $\check{\Sigma}^k \in \mathbb{R}^{2L\times 2L}$, $\check{\mathrm{T}}^{k+1} \in \mathbb{R}^{L\times L}$ are defined by the following state evolution recursion. Let $\check{\Sigma}^0 = \Sigma^0$ (defined in Assumption **(A1)**), and for $k \geq 0$:

$$\check{\mathrm{T}}^{k+1} = \mathbb{E}\big[h_k(G_\sigma^k, \bar{\Psi})h_k(G_\sigma^k, \bar{\Psi})^\top\big], \quad (32)$$

$$\begin{aligned} \check{\Sigma}^{k+1} &= \delta^{-1}\mathbb{E}\big[\check{f}_{k+1}(G_\tau^{k+1}, \bar{B})\check{f}_{k+1}(G_\tau^{k+1}, \bar{B})^\top\big], \\ &= \begin{pmatrix} \delta^{-1}\mathbb{E}[\bar{B}\bar{B}^\top] & \check{\Sigma}_{(12)}^{k+1} \\ \big(\check{\Sigma}_{(12)}^{k+1}\big)^\top & \check{\Sigma}_{(22)}^{k+1} \end{pmatrix}, \end{aligned} \quad (33)$$

where

$$\begin{aligned} \check{\Sigma}_{(12)}^{k+1} &= \Big(\check{\Sigma}_{(21)}^{k+1}\Big)^\top = \delta^{-1}\mathbb{E}\big[\bar{B}f_{k+1}(G_\tau^{k+1} + \mathrm{M}_{k+1}^B\bar{B})^\top\big] \\ \check{\Sigma}_{(22)}^{k+1} &= \delta^{-1}\mathbb{E}\big[f_{k+1}(G_\tau^{k+1} + \mathrm{M}_{k+1}^B\bar{B}) \\ &\qquad \cdot f_{k+1}(G_\tau^{k+1} + \mathrm{M}_{k+1}^B\bar{B})^\top\big]. \end{aligned} \quad (34)$$

Here we take $G_\sigma^k \sim N(0, \check{\Sigma}^k)$ independent of $\bar{\Psi} \sim P_{\bar{\Psi}}$, and $G_\tau^{k+1} \sim N(0, \check{\mathrm{T}}^{k+1})$ independent of $\bar{B} \sim P_{\bar{B}}$. Comparing the recursive definitions of $(\mathrm{T}_B^{k+1}, \Sigma^{k+1})$ in (7)-(9) and of $(\check{\mathrm{T}}^{k+1}, \check{\Sigma}^{k+1})$ in (32)-(33), and noting that they are both initialized with Σ^0, we have that $\check{\mathrm{T}}^{k+1} = \mathrm{T}_B^{k+1}$ and $\check{\Sigma}^{k+1} = \Sigma^{k+1}$ for $k \geq 0$.

The following proposition follows from the state evolution result (Feng et al., 2022, Sec. 6.7) for an abstract AMP recursion with matrix-valued iterates.

Proposition 3. *Assume the setting of Theorem 1. For the abstract AMP in* (30)*, for $k \geq 0$ we have:*

$$\sup_{\eta\in \mathrm{PL}_{2L}(r,1)} \Big|\frac{1}{p}\sum_{j=1}^p \eta(\check{B}_j^{k+1},\ B_j) - \mathbb{E}[\eta(G_\tau^{k+1},\ \bar{B})]\Big| \xrightarrow{c} 0, \quad (35)$$

$$\sup_{\eta\in \mathrm{PL}_{2L+L_\Psi}(r,1)} \Big|\frac{1}{n}\sum_{i=1}^n \eta(\check{\Theta}_i^k, \Psi_i) - \mathbb{E}[\eta(G_\sigma^k, \bar{\Psi})]\Big| \xrightarrow{c} 0, \quad (36)$$

as $n, p \to \infty$ with $n/p \to \delta$.

To obtain the result (14), we recall the definition of $\check{B}^{k+1}$ from (28), and in (35) we take $\eta(\check{B}^{k+1}, B) = c_{k,r}\phi(\check{B}^{k+1} + B(\mathrm{M}_B^{k+1})^\top, B)$ for a suitably small constant $c_{k,r} > 0$, and recall that $G_\tau^{k+1} \sim \mathcal{N}(0, \mathrm{T}_B^{k+1})$. To obtain (15), we recall the definition of $\check{\Theta}^k$ from (28), and in (36) take $\eta(\check{\Theta}^k, \Psi) = \phi(\Theta^k, \Theta, \Psi)$. Since $\check{\Sigma}^k = \Sigma^k$, we have:

$$(G_\sigma^k, \bar{\Psi}) \overset{d}{=} (Z, Z^k, \bar{\Psi}) \overset{d}{=} (Z,\ \mathrm{M}_\Theta^k Z + G_\Theta^k, \bar{\Psi}), \quad (37)$$

where the last equality follows from (10). This completes the proof of the theorem. □

5.2 Proof of Proposition 2

The proof relies on the following generalized Cauchy-Schwarz inequality for covariance matrices.

Lemma 4. (Lavergne, 2008, Lemma 1) *Let $U, V \in \mathbb{R}^L$ random vectors such that $\mathbb{E}[\|U\|_2^2] < \infty$, $\mathbb{E}[\|V\|_2^2] < \infty$, and $\mathbb{E}[VV^\top]$ is invertible. Then*

$$\mathbb{E}[UU^\top] - \mathbb{E}[UV^\top]\big(\mathbb{E}[VV^\top]\big)^{-1}\mathbb{E}[VU^\top] \succeq 0. \quad (38)$$

Proof of part 1. Using the law of total expectation, $\Sigma_{(12)}^k$ in (9) can be written as:

$$\begin{aligned} \delta\Sigma_{(12)}^k &= \mathbb{E}[\bar{B}f_k(\mathrm{M}_B^k\bar{B} + G_B^k)^\top] \quad (39) \\ &= \mathbb{E}\big[\mathbb{E}[\bar{B}f_k(\mathrm{M}_B^k\bar{B} + G_B^k)^\top \mid \mathrm{M}_B^k\bar{B} + G_B^k]\big] = \mathbb{E}[f_k^* f_k^\top], \end{aligned}$$

where we use the shorthand $f_k \equiv f_k(\mathrm{M}_B^k \bar{B} + G_B^k)$ and $f_k^* \equiv \mathbb{E}[\bar{B} \,|\, \mathrm{M}_B^k \bar{B} + G_B^k]$. Using Lemma 4 we have that

$$\begin{aligned} &\mathbb{E}[f_k^*(f_k^*)^\top] - \mathbb{E}[f_k^* f_k^\top]\mathbb{E}[f_k f_k^\top]^{-1}\mathbb{E}[f_k(f_k^*)^\top] \succeq 0 \\ &\implies \delta^{-1}\mathbb{E}[f_k^*(f_k^*)^\top] - \Sigma^k_{(12)}(\Sigma^k_{(22)})^{-1}\Sigma^k_{(21)} \succeq 0, \end{aligned} \tag{40}$$

where we have used (39) and (9) for the second line. Adding and subtracting T_Θ^k in (40) we obtain

$$\mathrm{T}_\Theta^k - \underbrace{\big(\mathrm{T}_\Theta^k - \delta^{-1}\mathbb{E}[f_k^*(f_k^*)^\top] + \Sigma^k_{(12)}(\Sigma^k_{(22)})^{-1}\Sigma^k_{(21)}\big)}_{:=\Gamma_\Theta^k} \succeq 0. \tag{41}$$

Multiplying the matrix $(\mathrm{T}_\Theta^k - \Gamma_\Theta^k)$ in (41) by $\big(\mathrm{M}_\Theta^k\big)^{-1}$ on the left and $\Big(\big(\mathrm{M}_\Theta^k\big)^{-1}\Big)^\top$ on the right maintains positive definiteness. This yields

$$N_\Theta^k - \big(\mathrm{M}_\Theta^k\big)^{-1}\Gamma_\Theta^k\Big(\big(\mathrm{M}_\Theta^k\big)^{-1}\Big)^\top \succeq 0, \tag{42}$$

where we have used the formula for N_Θ^k from (19). Eq. (42) implies

$$\mathrm{Tr}(N_\Theta^k) \geq \mathrm{Tr}\left(\big(\mathrm{M}_\Theta^k\big)^{-1}\Gamma_\Theta^k\Big(\big(\mathrm{M}_\Theta^k\big)^{-1}\Big)^\top\right). \tag{43}$$

Now, using the formula for T_Θ^k in (12) it can be verified that when $f_k = f_k^*$, we have

$$\mathrm{T}_\Theta^k = \Gamma_\Theta^k = \tag{44}$$
$$\frac{1}{\delta}\Big(\mathbb{E}[f_k^*(f_k^*)^\top] - \mathbb{E}[f_k^*(f_k^*)^\top]\big(\mathbb{E}[\bar{B}\bar{B}^\top]\big)^{-1}\mathbb{E}[f_k^*(f_k^*)^\top]\Big).$$

Therefore (41)-(43) are satisfied with equality when $f_k = f_k^*$, which proves the first part of the proposition.

Proof of part 2. We begin with the following lemma:

Lemma 5. *Let $x = (x_1, \ldots, x_L)$ and $g : \mathbb{R}^L \to \mathbb{R}^L$ be such that for $j = 1, \ldots, L$, the function $x_j \to g_l(x_1, \ldots, x_L)$ (where $g_l(x_1, \ldots, x_L)$ is the lth entry of $g(x_1, \ldots, x_L)$) is absolutely continuous for Lebesgue almost every $(x_i : i \neq j) \in \mathbb{R}^{L-1}$, with weak derivative $\partial_{x_j} g_l : \mathbb{R}^L \to \mathbb{R}$ satisfying $\mathbb{E}[|\partial_{x_j} g_l(x)|] < \infty$. Let $\nabla g(x) = (\nabla g_1(x), \ldots, \nabla g_L(x))^\top \in \mathbb{R}^{L\times L}$ where $\nabla g_l(x) = \big(\partial_{x_1} g_l(x), \ldots, \partial_{x_L} g_l(x)\big)^\top$ for $x \in \mathbb{R}^L$. If $X \sim \mathcal{N}(\mu, \Sigma)$ with Σ positive definite, then*

$$\mathbb{E}[\nabla g(X)] = \Big(\Sigma^{-1}\mathbb{E}\big[(X-\mu)g(X)^\top\big]\Big)^\top. \tag{45}$$

Proof. We have

$$\begin{aligned} &\mathbb{E}[(X-\mu)g(X)^\top] \\ &= \big(\mathbb{E}[(X-\mu)g_1(X)], \ldots, \mathbb{E}[(X-\mu)g_L(X)]\big) && (46)\\ &\overset{(a)}{=} (\Sigma\mathbb{E}[\nabla g_1(X)], \ldots, \Sigma\mathbb{E}[\nabla g_L(X)]) && (47)\\ &= \Sigma\mathbb{E}[(\nabla g_1(X), \ldots, \nabla g_L(X))] && (48)\\ &\overset{(b)}{=} \Sigma\mathbb{E}[\nabla g(X)]^\top, && (49) \end{aligned}$$

where (a) uses the multivariate Stein's Lemma from (Feng et al., 2022, Lemma 6.20) which states that under our conditions we have $\mathbb{E}[Xg_l(X)] = \Sigma\mathbb{E}[\nabla g_l(X)]$ for $l = 1, \ldots, L$, and (b) uses the definition of $\nabla g(x)$. Finally, rearranging the above equation and taking the transpose gives the result. □

Next, we use our multivariate generalization of Stein's Lemma (Lemma 5) to show that

$$\mathrm{M}_B^{k+1} = \mathbb{E}\big[g_k(Z^k, \bar{Y})g_k^*(Z^k, \bar{Y})^\top\big], \tag{50}$$

where g_k^* is defined in (21). Indeed, using the law of total expectation we have

$$\begin{aligned} \mathrm{M}_B^{k+1} &= \mathbb{E}\Big[\mathbb{E}[\partial_Z h_k(Z, Z^k, \bar{\Psi})|Z^k]\Big] \\ &\overset{(a)}{=} \mathbb{E}\Big[\mathrm{Cov}[Z|Z^k]^{-1}\mathbb{E}\big[(Z - \mathbb{E}[Z|Z^k])h_k(Z, Z^k, \bar{\Psi})^\top \\ &\qquad\qquad \big| \, Z^k\big]\Big]^\top \\ &= \mathbb{E}[\mathrm{Cov}[Z|Z^k]^{-1}(Z - \mathbb{E}[Z|Z^k])h_k(Z, Z^k, \bar{\Psi})^\top]^\top \\ &= \mathbb{E}\Big[\mathbb{E}\big[\mathrm{Cov}[Z|Z^k]^{-1}(Z - \mathbb{E}[Z|Z^k])h_k(Z, Z^k, \bar{\Psi})^\top \\ &\qquad\qquad \big| \, Z^k, \bar{Y}\big]\Big]^\top \\ &\overset{(b)}{=} \mathbb{E}\Big[g_k^*(Z^k, \bar{Y})h_k(Z, Z^k, \bar{\Psi})^\top\Big]^\top \\ &\overset{(c)}{=} \mathbb{E}\big[g_k(Z^k, \bar{Y})g_k^*(Z^k, \bar{Y})^\top\big]. \end{aligned} \tag{51}$$

Here (a) applies Lemma 5, (b) follows from the definition of g_k^* in (21), and (c) from (5). Using the shorthand $g_k \equiv g_k(Z^k, \bar{Y})$ and $g_k^* \equiv g_k^*(Z^k, \bar{Y})$, from Lemma 4 we have:

$$\begin{aligned} &\mathbb{E}[g_k^*(g_k^*)^\top] - \mathbb{E}[g_k^* g_k^\top]\Big(\mathbb{E}[g_k g_k^\top]\Big)^{-1}\mathbb{E}[g_k(g_k^*)^\top] \succeq 0 \\ &\Leftrightarrow \mathbb{E}[g_k^*(g_k^*)^\top] - \big(N_B^{k+1}\big)^{-1} \succeq 0 && (52)\\ &\Leftrightarrow \Big(\mathbb{E}[g_k^*(g_k^*)^\top]\Big)^{-1} - N_B^{k+1} \preceq 0, && (53) \end{aligned}$$

where (52) is obtained by recallling from (19) that $(N_B^{k+1})^{-1} = \mathrm{M}_B^{k+1}\big(\mathrm{T}_B^{k+1}\big)^{-1}\big(\mathrm{M}_B^{k+1}\big)^\top$, and using the expressions for M_B^{k+1} and T_B^{k+1} in (50) and (7). Eq. (53) follows from the fact that if P and Q are positive definite matrices such that $P - Q \succeq 0$, then $P^{-1} - Q^{-1} \preceq 0$. From (53), we have that

$$\mathrm{Tr}(N_B^{k+1}) \leq \mathrm{Tr}\left(\big(\mathbb{E}[g_k^*(g_k^*)^\top]\big)^{-1}\right), \tag{54}$$

with equality if $g_k = g_k^*$. This completes the proof of the second part of the proposition. □

Acknowledgements

N. Tan was supported by the Cambridge Trust and the Harding Distinguished Postgraduate Scholars Programme Leverage Scheme.

References

Balakrishnan, S., Wainwright, M. J., and Yu, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120.

Barbier, J., Krzakala, F., Macris, N., Miolane, L., and Zdeborová, L. (2019). Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460.

Barbier, J., Macris, N., and Rush, C. (2020). All-or-nothing statistical and computational phase transitions in sparse spiked matrix estimation. In *Neural Information Processing Systems (NeurIPS)*.

Barik, A. and Honorio, J. (2022). Sparse mixed linear regression with guarantees: Taming an intractable problem with invex relaxation. *International Conference on Machine Learning*, 162:1627–1646.

Bayati, M. and Montanari, A. (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57:764–785.

Chaganty, A. T. and Liang, P. (2013). Spectral experts for estimating mixtures of linear regressions. *International Conference on Machine Learning*, page 1040–1048.

Chandrasekher, K. A., Pananjady, A., and Thrampoulidis, C. (2021). Sharp global convergence guarantees for iterative nonconvex optimization: A Gaussian process perspective. arXiv:2109.09859.

Chen, S., Li, J., and Song, Z. (2020). Learning mixtures of linear regressions in subexponential time via Fourier moments. *Proceedings of the 52nd Annual ACM Symposium on Theory of Computing*, pages 587–600.

Chen, Y., Yi, X., and Caramanis, C. (2014). A convex formulation for mixed regression with two components: Minimax optimal rates. *Conference on Learning Theory*, pages 560–604.

Deshpande, Y. and Montanari, A. (2014). Information-theoretically optimal sparse PCA. In *IEEE International Symposium on Information Theory (ISIT)*, pages 2197–2201.

Devijver, E., Goude, Y., and Poggi, J.-M. (2020). Clustering electricity consumers using high-dimensional regression mixture models. *Applied Stochastic Models in Business and Industry*, pages 159–177.

Donoho, D. L., Javanmard, A., and Montanari, A. (2013). Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing. *IEEE Transactions on Information Theory*, 59(11):7434–7464.

Donoho, D. L., Maleki, A., and Montanari, A. (2009). Message passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106:18914–18919.

Fan, J., Liu, H., Wang, Z., and Yang, Z. (2018). Curse of heterogeneity: Computational barriers in sparse mixture models and phase retrieval. arXiv:1808.06996.

Fan, Z. (2022). Approximate message passing algorithms for rotationally invariant matrices. *Annals of Statistics*, 50(1):197–224.

Faria, S. and Soromenho, G. (2010). Fitting mixtures of linear regressions. *Journal of Statistical Computation and Simulation*, 80(2):201–225.

Feng, O. Y., Venkataramanan, R., Rush, C., and Samworth, R. J. (2022). A unifying tutorial on approximate message passing. *Foundations and Trends in Machine Learning*.

Fletcher, A. K. and Rangan, S. (2018). Iterative reconstruction of rank-one matrices in noise. *Information and Inference: A Journal of the IMA*, 7(3):531–562.

Ghosh, A. and Kannan, R. (2020). Alternating minimization converges super-linearly for mixed linear regression. In *International Conference on Artificial Intelligence and Statistics*, pages 1093–1103.

Ghosh, A., Pananjady, A., Guntuboyina, A., and Ramchandran, K. (2022). Max-affine regression: Parameter estimation for Gaussian designs. *IEEE Transactions on Information Theory*, 68(3):1851–1885.

Grün, B. and Leisch, F. (2007). Applications of finite mixtures of regression models. `https://tinyurl.com/3sfyrwbs`.

Hsu, P.-L. and Robbins, H. (1947). Complete convergence and the law of large numbers. *Proceedings of the national academy of sciences*, 33(2):25–31.

Javanmard, A. and Montanari, A. (2013). State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference*, 2(2):115–144.

Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical Mixtures of Experts and the EM Algorithm. *Neural Computation*, 6(2):181–214.

Kabashima, Y. (2003). A CDMA multiuser detection algorithm on the basis of belief propagation. *Journal of Physics A: Mathematical and General*, 36(43):11111–11121.

Kabashima, Y., Krzakala, F., Mézard, M., Sakata, A., and Zdeborová, L. (2016). Phase transitions and sample complexity in Bayes-optimal matrix factorization. *IEEE Transactions on Information Theory*, 62(7):4228–4265.

Khalili, A. and Chen, J. (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association*, 102(479):1025–1038.

Klusowski, J. M., Yang, D., and Brinda, W. D. (2019). Estimating the coefficients of a mixture of two linear regressions by expectation maximization. *IEEE Transactions on Information Theory*, 65:3515–3524.

Kong, W., Somani, R., Song, Z., Kakade, S., and Oh, S. (2020). Meta-learning for mixed linear regression. *International Conference on Machine Learning*, 119:5394–5404.

Krishnamurthy, A., Mazumdar, A., McGregor, A., and Pal, S. (2019). Sample Complexity of Learning Mixture of Sparse Linear Regressions. In *Advances in Neural Information Processing Systems*, volume 32.

Krzakala, F., Mézard, M., Sausset, F., Sun, Y., and Zdeborová, L. (2012). Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(8).

Lavergne, P. (2008). A Cauchy-Schwarz inequality for expectation of matrices. *Discussion Papers, Department of Economics, Simon Fraser University*.

Lesieur, T., Krzakala, F., and Zdeborová, L. (2017). Constrained low-rank matrix estimation: Phase transitions, approximate message passing and applications. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(7):073403.

Li, Q., Shi, R., and Liang, F. (2019). Drug sensitivity prediction with high-dimensional mixture regression. *PloS one*, pages 1–18.

Li, Y. and Liang, Y. (2018). Learning mixtures of linear regressions with nearly optimal complexity. *Conference On Learning Theory*, pages 1125–1144.

Ma, J., Xu, J., and Maleki, A. (2019). Optimization-based AMP for phase retrieval: The impact of initialization and ℓ_2 regularization. *IEEE Transactions on Information Theory*, 65(6):3600–3629.

Maillard, A., Loureiro, B., Krzakala, F., and Zdeborová, L. (2020). Phase retrieval in high dimensions: Statistical and computational phase transitions. In *Neural Information Processing Systems (NeurIPS)*.

Maleki, A., Anitori, L., Yang, Z., and Baraniuk, R. G. (2013). Asymptotic analysis of complex lasso via complex approximate message passing (CAMP). *IEEE Transactions on Information Theory*, 59(7):4290–4308.

McLachlan, G. J. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

Mondelli, M. and Venkataramanan, R. (2021). Approximate message passing with spectral initialization for generalized linear models. *International Conference on Artificial Intelligence and Statistics*, pages 397–405.

Montanari, A. and Venkataramanan, R. (2021). Estimation of low-rank matrices via approximate message passing. *Annals of Statistics*, 45(1):321–345.

Pal, S., Mazumdar, A., and Gandikota, V. (2021). Support recovery of sparse signals from a mixture of linear measurements. In *Advances in Neural Information Processing Systems*.

Pal, S., Mazumdar, A., Sen, R., and Ghosh, A. (2022). On learning mixture of linear regressions in the non-realizable setting. In *International Conference on Machine Learning*, pages 17202–17220.

Peng, F., Jacobs, R. A., and Tanner, M. A. (1996). Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. *Journal of the American Statistical Association*, 91(435):953–960.

Rangan, S. (2011). Generalized approximate message passing for estimation with random linear mixing. *IEEE International Symposium on Information Theory*.

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

Sedghi, H., Janzamin, M., and Anandkumar, A. (2016). Provable tensor methods for learning mixtures of generalized linear models. *International Conference on Artificial Intelligence and Statistics*, 51:1223–1231.

Shen, Y. and Sanghavi, S. (2019). Iterative least trimmed squares for mixed linear regression. *Advances in Neural Information Processing Systems*, 32.

Städler, N., Bühlmann, P., and van de Geer, S. (2010). ℓ_1-penalization for mixture regression models. *TEST*, 19(2):209–256.

Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525.

Viele, K. and Tong, B. (2002). Modeling with mixtures of linear regressions. *Statistics and Computing*, 12:315–330.

Wang, T., Zhong, X., and Fan, Z. (2022). Universality of approximate message passing algorithms and tensor networks. *arXiv:2206.13037*.

Wang, Z., Gu, Q., Ning, Y., and Liu, H. (2015). High dimensional EM algorithm: Statistical optimization and asymptotic normality. *Advances in neural information processing systems*, pages 2521–2529.

Waterhouse, S., MacKay, D., and Robinson, A. (1995). Bayesian methods for mixtures of experts. In *Advances in Neural Information Processing Systems*, volume 8.

Yi, X. and Caramanis, C. (2015). Regularized EM algorithms: A unified framework and statistical guarantees. *Advances in Neural Information Processing Systems*, pages 1567–1575.

Yi, X., Caramanis, C., and Sanghavi, S. (2014). Alternating minimization for mixed linear regression. *International Conference on Machine Learning*, pages 613–621.

Zhang, L., Ma, R., Cai, T. T., and Li, H. (2020). Estimation, confidence intervals, and large-scale hypotheses testing for high-dimensional mixed linear regression. arXiv:2011.03598.

Zhong, K., Jain, P., and Dhillon, I. S. (2016). Mixed linear regression with multiple components. *Advances in Neural Information Processing Systems*, pages 2190–2198.

Zhong, X., Su, C., and Fan, Z. (2021). Approximate Message Passing for orthogonally invariant ensembles: Multivariate non-linearities and spectral initialization. *arXiv:2110.02318*.

Zhu, R., Wang, L., Zhai, C., and Gu, Q. (2017). High-dimensional variance-reduced stochastic gradient expectation-maximization algorithm. *Proceedings of the 34th International Conference on Machine Learning*, 70:4180–4188.

Álvaro González (2010). Measurement of areas on a sphere using fibonacci and latitude–longitude lattices. *Mathematical Geosciences*, 42(1):49–64.

A IMPLEMENTATION DETAILS

In this appendix, we consider MLR with two signals (see (22)), and provide the implementation details of matrix-AMP with Bayes-optimal functions (see Proposition 2), for the Gaussian prior and the sparse discrete prior. While the implementation details stated here are for the case of two signals, it is straightforward to generalize them to the case of three signals which we have omitted.

A.1 Gaussian Prior

We start by rewriting the matrix-AMP algorithm with more details below:

- Initialize $\widehat{R}^{-1} = 0 \in \mathbb{R}^{n\times 2}$, $F_0 = I_2$. Next, we decide on a joint distribution of $B_j^0 = (\beta_j^{(1)}, \beta_j^{(2)})$ and $\widehat{B}_j^0 = (\widehat{\beta}_j^{0,(1)}, \widehat{\beta}_j^{0,(2)})$, and initialize

$$\Sigma_0 = \frac{p}{n}\begin{bmatrix} \mathbb{E}[(\beta_j^{(1)})^2] & \mathbb{E}[\beta_j^{(1)}\beta_j^{(2)}] & \mathbb{E}[\beta_j^{(1)}\widehat{\beta}_j^{0,(1)}] & \mathbb{E}[\beta_j^{(1)}\widehat{\beta}_j^{0,(2)}] \\ \mathbb{E}[\beta_j^{(1)}\beta_j^{(2)}] & \mathbb{E}[(\beta_j^{(2)})^2] & \mathbb{E}[\beta_j^{(2)}\widehat{\beta}_j^{0,(1)}] & \mathbb{E}[\beta_j^{(2)}\widehat{\beta}_j^{0,(2)}] \\ \mathbb{E}[\beta_j^{(1)}\widehat{\beta}_j^{0,(1)}] & \mathbb{E}[\beta_j^{(2)}\widehat{\beta}_j^{0,(1)}] & \mathbb{E}[(\widehat{\beta}_j^{0,(1)})^2] & \mathbb{E}[\widehat{\beta}_j^{0,(1)}\widehat{\beta}_j^{0,(2)}] \\ \mathbb{E}[\beta_j^{(1)}\widehat{\beta}_j^{0,(2)}] & \mathbb{E}[\beta_j^{(2)}\widehat{\beta}_j^{0,(2)}] & \mathbb{E}[\widehat{\beta}_j^{0,(1)}\widehat{\beta}_j^{0,(2)}] & \mathbb{E}[(\widehat{\beta}_j^{0,(2)})^2] \end{bmatrix}. \tag{55}$$

- For each iteration of matrix-AMP $k \in \mathbb{N}_0$, we have the following steps:
  1. Compute $\Theta^k := X\widehat{B}^k - \widehat{R}^{k-1}F_k^\top$
  2. Approximate $\widehat{R}^k := g_k(\Theta^k, Y)$
  3. Approximate $C^k := \frac{1}{n}\sum_{i=1}^n g_k'(\Theta_i^k, Y_i)$
  4. Compute $B^{k+1} := X^\top\widehat{R}^k - \widehat{B}^k C_k^\top$
  5. Approximate $\widehat{B}^{k+1} := f_{k+1}(B^{k+1})$
  6. Approximate $F^{k+1} := \frac{1}{n}\sum_{j=1}^p f_{k+1}'(B_j^{k+1})$
  7. Approximate Σ^{k+1}

Steps 1 and 4 are straightforward and thus, can be obtained through direct computation. The other steps are trickier and requires some form of approximation to make the computation tractable. We now proceed to explain in detail how approximation can be done for steps 2, 3, 5, 6, and 7.

Step 2: We approximate this by computing $g_k(\Theta_i^k, Y_i)$ first, which we denote by $g_k(Z^k, \bar{Y})$ (this is essentially $g_k(\Theta_i^k, Y_i)$ but written with random variables) for the purpose of our derivations. The computation consists of the followings steps:

- Compute

$$\text{Var}[Z|Z^k] = \Sigma_{(11)}^k - \Sigma_{(12)}^k(\Sigma_{(22)}^k)^{-1}\Sigma_{(21)}^k \tag{56}$$

$$\mathbb{E}[Z|Z^k] = \Sigma_{(12)}^k(\Sigma_{(22)}^k)^{-1}Z^k. \tag{57}$$

- Note that we have $(Z, Z^k) \sim N_4(0, \Sigma^k)$, $Z = (Z_1, Z_2)^\top$ and

$$Y = q(Z, \bar{\Psi}) = Z_1\bar{c} + Z_2(1-\bar{c}) + \bar{\epsilon}. \tag{58}$$

  Define $\bar{c}$ to be the random variable that the empirical distribution of c_i's converge to. Compute

$$\mathbb{E}[Z|Z^k, \bar{Y}] = \sum_{\bar{c}\in\{0,1\}} \mathbb{E}[Z|Z^k, \bar{Y}, \bar{c}]\mathbb{P}[\bar{c}|Z^k, \bar{Y}] \tag{59}$$

$$= \underbrace{\mathbb{E}[Z|Z^k, \bar{Y}, \bar{c}=1]}_{\text{Part 1}}\underbrace{\mathbb{P}[\bar{c}=1|Z^k, \bar{Y}]}_{\text{Part 2}} + \underbrace{\mathbb{E}[Z|Z^k, \bar{Y}, \bar{c}=0]}_{\text{Part 3}}\underbrace{\mathbb{P}[\bar{c}=0|Z^k, \bar{Y}]}_{\text{Part 4}} \tag{60}$$

Part 1. We first find the joint distribution of $(Z, Z^k, \bar{Y}|\bar{c}=1)$ which we write as $\mathcal{N}_5(\mathbf{0}, \Sigma_Y^{k,1})$. We now proceed to derive $\Sigma_Y^{k,1}$. We know from the joint distribution of (Z, Z^k) that $(\Sigma_Y^{k,1})_{[4],[4]} = \Sigma^k$. Hence, we only need to solve for the remaining entries:

$$(\Sigma_Y^{k,1})_{5,5} = \mathrm{Var}[\bar{Y}|\bar{c}=1] = \mathrm{Var}[Z_1 + \bar{\epsilon}] = \Sigma_{11}^k + \sigma^2 \tag{61}$$

$$(\Sigma_Y^{k,1})_{1,5} = (\Sigma_Y^{k,1})_{5,1} = \mathrm{Cov}[\bar{Y}, Z_1|\bar{c}=1] = \mathrm{Cov}[Z_1 + \bar{\epsilon}, Z_1] = \Sigma_{11}^k \tag{62}$$

$$(\Sigma_Y^{k,1})_{2,5} = (\Sigma_Y^{k,1})_{5,2} = \mathrm{Cov}[\bar{Y}, Z_2|\bar{c}=1] = \mathrm{Cov}[Z_1 + \bar{\epsilon}, Z_2] = \Sigma_{12}^k \tag{63}$$

$$(\Sigma_Y^{k,1})_{1,3} = (\Sigma_Y^{k,1})_{3,1} = \mathrm{Cov}[\bar{Y}, Z_1^k|\bar{c}=1] = \mathrm{Cov}[Z_1 + \bar{\epsilon}, Z_1^k] = \Sigma_{13}^k \tag{64}$$

$$(\Sigma_Y^{k,1})_{1,4} = (\Sigma_Y^{k,1})_{4,1} = \mathrm{Cov}[\bar{Y}, Z_2^k|\bar{c}=1] = \mathrm{Cov}[Z_1 + \bar{\epsilon}, Z_2^k] = \Sigma_{14}^k, \tag{65}$$

where we have used the fact that (Z, Z^k) and $\bar{\epsilon}$ are independent, and the notation Σ_{ij}^k refers to the (i,j)-th entry of the matrix Σ^k. This gives

$$\Sigma_Y^{k,1} = \begin{bmatrix} \Sigma_{11}^k & \Sigma_{12}^k & \Sigma_{13}^k & \Sigma_{14}^k & \Sigma_{11}^k \\ \Sigma_{21}^k & \Sigma_{22}^k & \Sigma_{23}^k & \Sigma_{24}^k & \Sigma_{21}^k \\ \Sigma_{31}^k & \Sigma_{32}^k & \Sigma_{33}^k & \Sigma_{34}^k & \Sigma_{31}^k \\ \Sigma_{41}^k & \Sigma_{42}^k & \Sigma_{43}^k & \Sigma_{44}^k & \Sigma_{41}^k \\ \Sigma_{11}^k & \Sigma_{12}^k & \Sigma_{13}^k & \Sigma_{14}^k & \Sigma_{11}^k + \sigma^2 \end{bmatrix}. \tag{66}$$

From the joint distribution, we can compute

$$\mathbb{E}[Z|Z^k, \bar{Y}, \bar{c}=1] = (\Sigma_Y^{k,1})_{[2],[3^+]}(\Sigma_Y^{k,1})^{-1}_{[3^+],[3^+]} \begin{bmatrix} Z^k \\ \bar{Y} \end{bmatrix}, \tag{67}$$

where $[3^+] := \{3, 4, 5\}$.

Part 3. Using the same approach as part 1, we can find the joint distribution of $(Z, Z^k, \bar{Y}|\bar{c}=0)$, which is

$$\begin{bmatrix} Z \\ Z^k \\ \bar{Y} \end{bmatrix} \sim \mathcal{N}_5 \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \Sigma_Y^{k,0} = \begin{bmatrix} \Sigma_{11}^k & \Sigma_{12}^k & \Sigma_{13}^k & \Sigma_{14}^k & \Sigma_{12}^k \\ \Sigma_{21}^k & \Sigma_{22}^k & \Sigma_{23}^k & \Sigma_{24}^k & \Sigma_{22}^k \\ \Sigma_{31}^k & \Sigma_{32}^k & \Sigma_{33}^k & \Sigma_{34}^k & \Sigma_{32}^k \\ \Sigma_{41}^k & \Sigma_{42}^k & \Sigma_{43}^k & \Sigma_{44}^k & \Sigma_{42}^k \\ \Sigma_{21}^k & \Sigma_{22}^k & \Sigma_{23}^k & \Sigma_{24}^k & \Sigma_{22}^k + \sigma^2 \end{bmatrix} \right). \tag{68}$$

From the joint distribution, we can compute

$$\mathbb{E}[Z|Z^k, \bar{Y}, \bar{c}=0] = (\Sigma_Y^{k,0})_{[2],[3^+]}(\Sigma_Y^{k,0})^{-1}_{[3^+],[3^+]} \begin{bmatrix} Z^k \\ \bar{Y} \end{bmatrix}. \tag{69}$$

Part 2. We compute

$$\mathbb{P}[\bar{c}=1|Z^k, \bar{Y}] = \frac{\mathbb{P}[\bar{c}=1]\mathbb{P}[Z^k, \bar{Y}|\bar{c}=1]}{\mathbb{P}[\bar{c}=1]\mathbb{P}[Z^k, \bar{Y}|\bar{c}=1] + \mathbb{P}[\bar{c}=0]\mathbb{P}[Z^k, \bar{Y}|\bar{c}=0]} \tag{70}$$

$$= \frac{\alpha\mathbb{P}[Z^k, \bar{Y}|\bar{c}=1]}{\alpha\mathbb{P}[Z^k, \bar{Y}|\bar{c}=1] + (1-\alpha)\mathbb{P}[Z^k, \bar{Y}|\bar{c}=0]}, \tag{71}$$

where given $\bar{c}=1$, we have $(Z^k, \bar{Y})^\top \sim \mathcal{N}_3\big(\mathbf{0}, (\Sigma_Y^{k,1})_{[3^+],[3^+]}\big)$, and given $\bar{c}=0$, we have $(Z^k, \bar{Y})^\top \sim \mathcal{N}_3\big(\mathbf{0}, (\Sigma_Y^{k,0})_{[3^+],[3^+]}\big)$.

Part 4. Similar to part 2, we compute

$$\mathbb{P}[\bar{c}=0|Z^k, \bar{Y}] = \frac{(1-\alpha)\mathbb{P}[Z^k, \bar{Y}|\bar{c}=0]}{\alpha\mathbb{P}[Z^k, \bar{Y}|\bar{c}=1] + (1-\alpha)\mathbb{P}[Z^k, \bar{Y}|\bar{c}=0]}, \tag{72}$$

where given $\bar{c}=0$, we have $(Z^k, \bar{Y})^\top \sim \mathcal{N}_3\big(\mathbf{0}, (\Sigma_Y^{k,0})_{[3^+],[3^+]}\big)$.

- Finally, compute

$$g_k(Z^k, \bar{Y}) = \mathrm{Var}[Z|Z^k]^{-1}(\mathbb{E}[Z|Z^k, \bar{Y}] - \mathbb{E}[Z|Z^k]).$$

Now that we know how to compute $g_k(Z^k, \bar{Y})$, we can compute $\hat{R}^k$ by applying g_k row wise to Θ^k and Y (i.e., compute $g_k(\Theta_i^k, Y_i)$).

Step 3: We approximate $C_k = \frac{1}{n}\sum_{i=1}^b g_k'(\Theta_i^k, Y_i)$ by calculating $\mathbb{E}[g_k'(Z^k, \bar{Y})]$. We do this by applying the generalized Stein's lemma (see Lemma 5) to $(Z, Z^k)^\top \sim \mathcal{N}_4(0, \Sigma^k)$ and $h_k(Z, Z^k, \bar{\Psi}) = g_k(Z^k, \bar{Y})$. This gives

$$\mathbb{E}\left[\begin{bmatrix} Z \\ Z^k \end{bmatrix} h(Z, Z^k, \bar{\Psi})^\top\right] = \Sigma^k \mathbb{E}\Big[\nabla_{(Z,Z^k)} h_k(Z, Z^k, \bar{\Psi})\Big]^\top. \tag{73}$$

Writing the above more explicitly gives

$$\begin{bmatrix} \mathbb{E}[Z h_k(Z, Z^k, \bar{\Psi})^\top] \\ \mathbb{E}[Z^k h_k(Z, Z^k, \bar{\Psi})^\top] \end{bmatrix} = \begin{bmatrix} \Sigma^k_{(11)} & \Sigma_{(12)} \\ \Sigma^k_{(21)} & \Sigma_{(22)} \end{bmatrix} \mathbb{E}\begin{bmatrix} \nabla_Z\{h_k(Z, Z^k, \bar{\Psi})\}_1 & \nabla_Z\{h_k(Z, Z^k, \bar{\Psi})\}_2 \\ \nabla_{Z^k}\{h_k(Z, Z^k, \bar{\Psi})\}_1 & \nabla_{Z^k}\{h_k(Z, Z^k, \bar{\Psi})\}_2 \end{bmatrix} \tag{74}$$

$$= \begin{bmatrix} \Sigma^k_{(11)} & \Sigma^k_{(12)} \\ \Sigma^k_{(21)} & \Sigma^k_{(22)} \end{bmatrix} \begin{bmatrix} \mathbb{E}[\nabla_Z h_k(Z, Z^k, \bar{\Psi})]^\top \\ \mathbb{E}[\nabla_{Z^k} h_k(Z, Z^k, \bar{\Psi})]^\top \end{bmatrix} \tag{75}$$

$$= \begin{bmatrix} \Sigma^k_{(11)}\mathbb{E}[\nabla_Z h_k(Z, Z^k, \bar{\Psi})]^\top + \Sigma^k_{(12)}\mathbb{E}[\nabla_{Z^k} h_k(Z, Z^k, \bar{\Psi})]^\top \\ \Sigma^k_{(21)}\mathbb{E}[\nabla_Z h_k(Z, Z^k, \bar{\Psi})]^\top + \Sigma^k_{(22)}\mathbb{E}[\nabla_{Z^k} h_k(Z, Z^k, \bar{\Psi})]^\top \end{bmatrix}, \tag{76}$$

where $\{h_k(Z, Z^k, \bar{\Psi})\}_i$ refers to the ith output of $h_k \in \mathbb{R}^2$. Looking at just the second row above and rearranging, we get

$$\mathbb{E}[\nabla_{Z^k} h_k(Z, Z^k, \bar{\Psi})] = \Big\{(\Sigma^k_{(22)})^{-1}\Big(\mathbb{E}[Z^k h_k(Z, Z^k, \bar{\Psi})^\top] - \Sigma^k_{(21)}\mathbb{E}[\nabla_Z h_k(Z, Z^k, \bar{\Psi})]^\top\Big)\Big\}^\top, \tag{77}$$

where $\mathbb{E}[Z^k h_k(Z, Z^k, \bar{\Psi})^\top]$ can be approximated by $\frac{1}{n}\langle \Theta^k, g_k(\Theta^k, Y)\rangle$ and $\mathbb{E}[\nabla_Z h_k(Z, Z^k, \bar{\Psi})]$ can be approximated by $\frac{1}{n} g_k(\Theta^k, Y)^\top g_k(\Theta^k, Y)$ (see (51) to see why this makes sense).

Step 5: We approximate this by first noting that from our state evolution $G_B^{k+1} \sim \mathcal{N}_2(0, \mathrm{T}_B^{k+1})$ is independent of $\bar{B}$. This implies that

$$\begin{bmatrix} \bar{B} \\ \mathrm{M}_B^{k+1}\bar{B} + G_B^{k+1} \end{bmatrix} \sim \mathcal{N}_4\left(\begin{bmatrix} \mathbb{E}[\bar{B}] \\ \mathrm{M}_B^{k+1}\mathbb{E}[\bar{B}] \end{bmatrix}, \begin{bmatrix} \mathrm{Cov}[\bar{B}] & \mathrm{Cov}[\bar{B}](\mathrm{M}_B^{k+1})^\top \\ \mathrm{M}_B^{k+1}\mathrm{Cov}[\bar{B}] & \mathrm{M}_B^{k+1}\mathrm{Cov}[\bar{B}](\mathrm{M}_B^{k+1})^\top + \mathrm{T}_B^{k+1} \end{bmatrix}\right) \tag{78}$$

This implies that

$$f_{k+1}(\mathrm{M}_B^{k+1}\bar{B} + G_B^{k+1} =: s) = \mathbb{E}[\bar{B}|s] \tag{79}$$

$$= \mathbb{E}[\bar{B}] + \mathrm{Cov}[\bar{B}](\mathrm{M}_B^{k+1})^\top\Big(\mathrm{M}_B^{k+1}\mathrm{Cov}[\bar{B}](\mathrm{M}_B^{k+1})^\top + \mathrm{T}_B^{k+1}\Big)^{-1}\Big(s - \mathrm{M}_B^{k+1}\mathbb{E}[\bar{B}]\Big). \tag{80}$$

We can use the above function to compute $f_{k+1}(B_j^{k+1})$ if we can approximate M_B^{k+1} and T_B^{k+1} (which is the same as M_B^{k+1} under the Bayes-optimal setting). This can be calculated using

$$\mathrm{M}_B^{k+1} \approx \frac{1}{n} g_k(\Theta^k, Y)^\top g_k(\Theta^k, Y). \tag{81}$$

Step 6: The expression for this can be obtained by taking the derivative of (80) w.r.t. s, which gives

$$f_{k+1}'(s) = \Big(\mathrm{M}_B^{k+1}\mathrm{Cov}[\bar{B}](\mathrm{M}_B^{k+1})^\top + \mathrm{T}_B^{k+1}\Big)^{-1}\mathrm{M}_B^{k+1}\mathrm{Cov}[\bar{B}], \tag{82}$$

where M_B^{k+1} and T_B^{k+1} can be approximated using (81).

Step 7: We have

$$\Sigma_{k+1} \approx \frac{p}{n}\begin{bmatrix} \mathbb{E}[\bar{B}\bar{B}^\top] & \frac{1}{p} f_{k+1}(B^{k+1})^\top f_{k+1}(B^{k+1}) \\ \frac{1}{p} f_{k+1}(B^{k+1})^\top f_{k+1}(B^{k+1}) & \frac{1}{p} f_{k+1}(B^{k+1})^\top f_{k+1}(B^{k+1}) \end{bmatrix}. \tag{83}$$

A.2 Sparse Discrete Prior

As presented in Appendix A.1, there are seven main steps in the AMP algorithm. A change in prior requires us to make changes to our denoiser f_k which affects steps 5, and 6 – the other steps remain unchanged. The changes are as follows:

Step 5: We have

$$f_{k+1}(\mathrm{M}_B^{k+1}\bar{B} + G_B^{k+1} =: s) = \mathbb{E}[\bar{B}|s] \tag{84}$$

$$= \sum_{\bar{b}} \bar{b}\mathbb{P}[\bar{B} = \bar{b}|s] \tag{85}$$

$$= \sum_{\bar{b}} \bar{b}\frac{\mathbb{P}[\bar{B} = \bar{b}]\mathbb{P}[s|\bar{B} = \bar{b}]}{\mathbb{P}[s]} \tag{86}$$

$$= \frac{\sum_{\bar{b}} \bar{b}\mathbb{P}[\bar{B} = \bar{b}]\mathbb{P}[s|\bar{B} = \bar{b}]}{\sum_{\bar{b}} \mathbb{P}[\bar{B} = \bar{b}]\mathbb{P}[s|\bar{B} = \bar{b}]}, \tag{87}$$

where $(s|\bar{B} = \bar{b}) = (\mathrm{M}_B^{k+1}\bar{B} + G_B^{k+1}|\bar{B} = \bar{b}) \sim \mathcal{N}(\mathrm{M}_B^{k+1}\bar{b}, \mathrm{T}_B^{k+1})$.

Step 6: By the definition of a Jacobian, we have

$$\nabla_s f(\mathrm{M}_B^{k+1}\bar{B} + G_B^{k+1} =: s) = \begin{bmatrix} \frac{\partial f_1}{\partial s_1} & \frac{\partial f_1}{\partial s_2} \\ \frac{\partial f_2}{\partial s_1} & \frac{\partial f_2}{\partial s_2} \end{bmatrix} = \begin{bmatrix} (\nabla_s f_1)^\top \\ (\nabla_s f_2)^\top \end{bmatrix} \tag{88}$$

Let us look at just

$$f_1(s) = \frac{\sum_{\bar{b}} \beta_j^{(1)}\mathbb{P}[\bar{B} = \bar{b}]\mathbb{P}[s|\bar{B} = \bar{b}]}{\sum_{\bar{b}} \mathbb{P}[\bar{B} = \bar{b}]\mathbb{P}[s|\bar{B} = \bar{b}]} =: \frac{\mathrm{num}_1}{\mathrm{denom}_1}. \tag{89}$$

By the quotient rule for functions with a vector input and an ouput in $\mathbb{R}$, we have

$$\nabla_s f_1(s) = \frac{(\nabla_s \mathrm{num}_1)(\mathrm{denom}_1) - (\mathrm{num}_1)(\nabla_s \mathrm{denom}_1)}{\mathrm{denom}_1^2} \tag{90}$$

Using vector calculus, we see that

$$\nabla_s \mathbb{P}[s|\bar{B} = \bar{b}] = \nabla_s \left(\frac{\exp\{-\frac{1}{2}(s - \mathrm{M}_B^{k+1}\bar{b})^\top (\mathrm{T}_B^{k+1})^{-1}(s - \mathrm{M}_B^{k+1}\bar{b})\}}{\sqrt{\det(2\pi \mathrm{T}_B^{k+1})}} \right) \tag{91}$$

$$= (\mathrm{T}_B^{k+1})^{-1}\big(\mathrm{M}_B^{k+1}\bar{b} - s\big)\mathbb{P}[s|\bar{B} = \bar{b}] \tag{92}$$

Using the above equation, we get

$$\nabla_s \mathrm{num}_1 = \sum_{\bar{b}} \beta_j^{(1)} (\mathrm{T}_B^{k+1})^{-1}\big(\mathrm{M}_B^{k+1}\bar{b} - s\big)\mathbb{P}[\bar{B} = \bar{b}]\mathbb{P}[s|\bar{B} = \bar{b}] \tag{93}$$

$$\nabla_s \mathrm{denom}_1 = \sum_{\bar{b}} (\mathrm{T}_B^{k+1})^{-1}\big(\mathrm{M}_B^{k+1}\bar{b} - s\big)\mathbb{P}[\bar{B} = \bar{b}]\mathbb{P}[s|\bar{B} = \bar{b}]. \tag{94}$$

We can perform the same steps to get $\nabla_s f_2(s)$.