Journal of Machine Learning Research 24 (2023) 1-44 Submitted 4/23; Published 8/23

Mixed Regression via Approximate Message Passing

Nelvin Tan TCNT2@CAM.AC.UK
Department of Engineering, University of Cambridge
Cambridge, CB2 1PZ, United Kingdom

Ramji Venkataramanan RV285@CAM.AC.UK
Department of Engineering, University of Cambridge
Cambridge, CB2 1PZ, United Kingdom

Editor: David Sontag

Abstract

We study the problem of regression in a generalized linear model (GLM) with multiple signals and latent variables. This model, which we call a matrix GLM, covers many widely studied problems in statistical learning, including mixed linear regression, max-affine regression, and mixture-of-experts. The goal in all these problems is to estimate the signals, and possibly some of the latent variables, from the observations. We propose a novel approximate message passing (AMP) algorithm for estimation in a matrix GLM 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. The state evolution characterization can be used to tailor the AMP algorithm to take advantage of any structural information known about the signals. Using state evolution, we derive an optimal choice of AMP 'denoising' functions that minimizes the estimation error in each iteration.

The theoretical results are validated by numerical simulations for mixed linear regression, max-affine regression, and mixture-of-experts. For max-affine regression, we propose an algorithm that combines AMP with expectation-maximization to estimate the intercepts of the model along with the signals. The numerical results show that AMP significantly outperforms other estimators for mixed linear regression and max-affine regression in most parameter regimes.

Keywords: Approximate Message Passing, Mixed Linear Regression, Max-Affine Regression, Mixture-of-Experts, Expectation-Maximization

1. Introduction

We study the problem of regression in a generalized linear model with multiple signals (regressors) and latent variables. Specifically, consider L signal vectors $\beta^{(1)}, \dots, \beta^{(L)} \in \mathbb{R}^p$, and define the signal matrix $B := (\beta^{(1)}, \dots, \beta^{(L)}) \in \mathbb{R}^{p \times L}$. Then, the goal is to estimate B from an observed matrix $Y := [Y_1, \dots, Y_n]^\top \in \mathbb{R}^{n \times L_{\text{out}}}$, whose ith row $Y_i \in \mathbb{R}^{L_{\text{out}}}$ is generated as:

$$Y_i = q(B^\top X_i \,, \Psi_i), \quad i \in [n]. \tag{1}$$

Here $X_i \in \mathbb{R}^p$ is the ith feature vector, $\Psi_i \in \mathbb{R}^{L_\Psi}$ is a vector of unobserved auxiliary variables, and $q : \mathbb{R}^L \times \mathbb{R}^{L_\Psi} \to \mathbb{R}^{L_{\text{out}}}$ is a known function. We refer to the model (1) as

the matrix generalized linear model, or *matrix GLM*. As we show below, the matrix GLM covers many widely studied regression models including mixed linear regression, max-affine regression, mixed GLMs, and mixture-of-experts.

1.1 Mixed Linear Regression

In this model, we wish to estimate L signal vectors from *unlabeled* observations of each. Specifically, the components of the observed vector $Y := (Y_1, \dots, Y_n)^\top$ are generated as:

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

Here 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 and $[n] := \{1, \dots, n\}$. In words, each observation comes from exactly one of the L signal vectors, but we do not know which one. The mixed linear regression (MLR) model in (2) is a special case of the matrix GLM in (1), with the rows of the auxiliary matrix given by $\Psi_i = (c_{i,1}, \dots, c_{i,L}, \epsilon_i)$, for $i \in [n]$.

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 MLR model is more flexible as it allows for differences in regressors across unobserved sub-populations. MLR has been used for analyzing heterogeneous 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 (2), 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:

$$\widehat{\beta}^{(1)}, \dots, \widehat{\beta}^{(L)} = \underset{\substack{\beta^{(1)}, \dots, \beta^{(L)} \in \mathbb{R}^p \\ c_1, \dots, c_L \in \{0,1\}^n \\ \sum_{l=1}^{L} c_{il} = 1,\, i \in [n]}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - \sum_{l=1}^{L} \langle X_i, \beta^{(l)} \rangle \, c_{il} \right)^2. \tag{3}$$

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: Bayesian methods (Viele and Tong, 2002), spectral methods (Chaganty and Liang, 2013; Yi et al., 2014); expectation-maximization (Faria and Soromenho, 2010; Städler et al., 2010; Zhang et al., 2020); alternating minimization and its variants (Yi et al., 2014; Shen and Sanghavi, 2019; Ghosh and Kannan, 2020; Zilber and Nadler, 2023); 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 sub-optimal 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.

1.2 Max-Affine Regression

In the max-affine regression (MAR) model, we have

$$Y_i = \max\left\{\langle X_i, \beta^{(1)}\rangle + b_1, \ldots, \langle X_i, \beta^{(L)}\rangle + b_L\right\} + \epsilon_i, \quad i \in [n]. \tag{4}$$

Here $b_1, \ldots, b_L \in \mathbb{R}$ are the intercepts (typically unknown), and ϵ_i is a zero-mean noise variable that is independent of X_i. In words, each observation comes from the maximum of L affine functions, each defined via a different signal vector.

When $L = 1$ and $b_1 = 0$, the model (4) corresponds to standard linear regression. When $L = 2$ and $\beta^{(1)} = -\beta^{(2)} = \beta$ along with $b_1 = b_2 = 0$, then (4) reduces to $Y_i = |\langle X_i, \beta\rangle| + \epsilon$. This is the widely studied phase retrieval problem (Netrapalli et al., 2013; Candès et al., 2015), which arises in applications such as scientific imaging (Fogel et al., 2016). For general L, the function $x \mapsto \max_{l\in\{1,\ldots,L\}}\{\langle x, \beta_l\rangle + b_l\}$ is always convex and thus, estimation under model (4) can be used to fit convex functions to the observed data. Indeed, the MAR model serves as a parametric approximation to the non-parametric convex regression model

$$Y_i = \varphi(X_i) + \epsilon_i, \quad i \in [n] \tag{5}$$

where $\varphi : \mathbb{R}^p \to \mathbb{R}$ is an unknown convex function (Balázs et al., 2015; Ghosh et al., 2022). Unfortunately, convex regression suffers from the curse of dimensionality unless p is small (Guntuboyina and Sen, 2013). Since convex functions can be approximated to arbitrary accuracy by maxima of affine functions, it is reasonable to simplify the problem by considering only those convex functions that can be written as a maximum of a fixed number of affine functions. This assumption directly leads to the MAR model (4), which has been studied as a tractable alternative to the non-parametric convex regression model (5) in applications where p is large, such as data in economics, finance and operations research (Balázs, 2016). MAR can also be used as a tractable model for the problem of estimating convex sets from support function measurements (Soh and Chandrasekaran, 2021), which arises in tomography applications (Prince and Willsky, 1990; Gregor and Rannou, 2002).

To write the MAR model as an instance of the matrix GLM (1), let us concisely denote the unknown parameters by $\beta_{\mathrm{ma}}^{(l)} = \begin{bmatrix}\beta^{(l)} \\ b_l\end{bmatrix} \in \mathbb{R}^{p+1}$ for $l \in [L]$, and the observations by $\left(X_i^{(\mathrm{ma})}, y_i\right)$ for $i \in [n]$, where $X_i^{(\mathrm{ma})} = \begin{bmatrix}X_i \\ 1\end{bmatrix} \in \mathbb{R}^{p+1}$ are the augmented features. Under the augmented features and signals, the model (4) becomes

$$Y_i = \max_{l\in[L]} \left\{\langle X_i^{(\mathrm{ma})}, \beta_{\mathrm{ma}}^{(l)}\rangle\right\} + \epsilon_i, \quad i \in [n], \tag{6}$$

which is of the form in (1). A natural approach for estimating $\beta_{\mathrm{ma}}^{(1)}, \ldots, \beta_{\mathrm{ma}}^{(L)}$ is the least squares estimator, defined as

$$\widehat{\beta}_{\mathrm{ma}}^{(1)}, \ldots, \widehat{\beta}_{\mathrm{ma}}^{(L)} = \underset{\beta_{\mathrm{ma}}^{(1)}, \ldots, \beta_{\mathrm{ma}}^{(L)} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - \max_{l\in[L]} \left\{\langle X_i^{(\mathrm{ma})}, \beta_{\mathrm{ma}}^{(l)}\rangle\right\}\right)^2. \tag{7}$$

Ghosh et al. (2022, Lemma 1) showed that a global minimizer of the least-squares criterion above always exists but will not in general be unique, since any relabelling of the indices (of the signal vectors) of a minimizer will also be a minimizer. Furthermore, the optimization problem in (7) is non-convex, and for a worst-case choice of the design matrix $X = (X_1, \dots, X_n)^\top$, the problem is NP-hard (Ghosh et al., 2022).

1.3 Mixed Generalized Linear Models and Mixture-of-Experts

A mixed GLM is a generalization of the MLR model (2), where the output function is not necessarily linear. Specifically, for some known function $\breve{q} : \mathbb{R}^2 \to \mathbb{R}$, we have

$$Y_i = \breve{q}(\langle X_i, \beta^{(1)} \rangle c_{i1} + \cdots + \langle X_i, \beta^{(L)} \rangle c_{iL}\,,\, \epsilon_i), \quad i \in [n]. \tag{8}$$

As before, 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 case of $L = 1$ is the standard GLM which, with suitable choices for $\breve{q}$ and ϵ, covers a range of statistical learning problems including logistic regression, phase retrieval, and one-bit compressed sensing. In all these settings, the mixed GLM model (8) allows the flexibility to account for unlabeled data coming from multiple sub-populations (Khalili and Chen, 2007; Sedghi et al., 2016).

The mixture-of-experts model, introduced by Jacobs et al. (1991); Jordan and Jacobs (1994), is a generalization of the mixed GLM, where the probability of selecting each regressor can depend on the feature vector. In addition to the L regressors $\beta^{(1)}, \dots, \beta^{(L)} \in \mathbb{R}^p$, here we have L gating parameters $w^{(1)}, \dots, w^{(L)} \in \mathbb{R}^p$, using which the observations are generated as follows. For each $i \in [n]$:

$$Y_i = \tilde{q}(\langle X_i, \beta^{(l)} \rangle\,,\, \epsilon_i) \quad \text{with probability } \frac{\exp(\langle X_i, w^{(l)} \rangle)}{\sum_{l'=1}^{L} \exp(\langle X_i, w^{(l')} \rangle)} \quad \text{for } l \in [L]. \tag{9}$$

Here $\tilde{q} : \mathbb{R} \to \mathbb{R}$ is a known activation function and ϵ_i is a noise variable. Mixture-of-experts models and its variants have been widely studied in machine learning (Yuksel et al., 2012; Huang and Yao, 2012; Makkuva et al., 2019, 2020) and applications such as computer vision (Gross et al., 2017), natural language processing (Shazeer et al., 2017), and econometrics (Huang et al., 2013; Compiani and Kitamura, 2016).

To see that the mixture-of-experts model is a special case of the matrix GLM in (1), we take the signal matrix to be $B = [\beta^{(1)}, \dots, \beta^{(L)}, w^{(1)}, \dots, w^{(L)}]$ and the auxiliary matrix $\Psi \in \mathbb{R}^{n \times 2}$ with rows $\Psi_i = (\psi_i, \epsilon_i)$, where $\psi_i \sim_{\text{i.i.d.}} \text{Uniform}[0, 1]$ for $i \in [n]$ and independent of $\{\epsilon_i\}_{i \in [n]}$. Then the model (9) can be written as:

$$\begin{aligned} Y_i &= q(B^\top X_i, \Psi_i) \\ &= \sum_{l=1}^{L} \tilde{q}(\langle X_i, \beta^{(l)} \rangle,\, \epsilon_i) \mathbb{1}\Bigg\{ \sum_{l'=1}^{l-1} \frac{\exp(\langle X_i, w^{(l')} \rangle)}{\sum_{l^*=1}^{L} \exp(\langle X_i, w^{(l^*)} \rangle)} < \psi_i \le \sum_{l'=1}^{l} \frac{\exp(\langle X_i, w^{(l')} \rangle)}{\sum_{l^*=1}^{L} \exp(\langle X_i, w^{(l^*)} \rangle)} \Bigg\}, \end{aligned} \tag{10}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function.

1.4 Approximate Message Passing

The main contribution of this work is to design and analyze an Approximate message passing (AMP) algorithm for estimation in the matrix GLM model (1). We then apply the algorithm to mixed linear regression, max-affine regression, and mixture-of-experts.

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; Donoho et al., 2009; Bayati and Montanari, 2011a; Krzakala et al., 2012), 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; Li and Wei, 2023). 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.

1.5 Main Contributions

We propose an AMP algorithm for the matrix GLM (1), 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 (1).

In Section 4, we present numerical simulation results for mixed linear regression, max-affine regression, and mixture-of-experts. The case of max-affine regression requires special attention as the AMP derived for the matrix GLM cannot be directly applied. This is because the matrix GLM AMP and its state evolution analysis is derived assuming that the features are all i.i.d. Gaussian. However, to write MAR as an instance of the matrix GLM, recall from (6) that we use the augmented features $X_i^{(\mathrm{ma})} = \begin{bmatrix} X_i \\ 1 \end{bmatrix} \in \mathbb{R}^{p+1}$, $i \in [n]$, which are not i.i.d. Gaussian due to the last component being 1. We address this by

using the original formulation of MAR in (4), with the intercepts $b_1, \dots, b_L$ treated as unknown model parameters. We estimate these intercepts via an expectation-maximization (EM) algorithm that uses the AMP iterates to approximate certain intractable quantities. This leads to a combined EM-AMP algorithm which is described in Section 4.3. For both mixed linear regression and max-affine regression, the numerical results show that AMP significantly outperforms other popular estimators (such as alternating minimization) in most parameter regimes.

Though the algorithms and results in this paper focus on estimating the signals $\beta^{(1)}, \dots \beta^{(L)}$, they can be often be translated to estimating the latent variables as well. For example, in mixed linear regression, given signal estimates $\widehat{\beta}^{(1)}, \dots, \widehat{\beta}^{(L)}$, the labels can be estimated as $\hat{c}_i = \text{argmin}_{l \in [L]} \, (Y_i - \langle X_i, \widehat{\beta}^{(l)} \rangle)^2$, for $i \in [n]$.

A preliminary version of this paper was published in the proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS 2023) (Tan and Venkataramanan, 2023). The focus of the preliminary version was largely on mixed linear regression. In the current paper, in addition to MLR, we provide results for max-affine regression and mixture-of experts, including the novel EM-AMP algorithm for MAR.

Technical Ideas. 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.

Our combined EM-AMP algorithm for max-affine regression is inspired by the work of Vila and Schniter (2013), who used a similar approach for the problem of sparse linear regression with unknown parameters in the signal prior.

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 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).

1.6 Other Related Work

Mixtures of linear and generalized linear models. The special case of symmetric mixed linear regression where $\beta^{(1)} = -\beta^{(2)}$ has been studied in many recent works. We note that symmetric MLR is a version of the phase retrieval problem (Netrapalli et al., 2013; Candès et al., 2015; Fogel et al., 2016). 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 symmetric MLR. 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 case of MLR with sparse signals was studied by Krishnamurthy et al. (2019) and Pal et al. (2021), and the gap between statistical and computational performance limits for sparse MLR was recently characterized by Arpino and Venkataramanan (2023). Pal et al. (2022) studied the prediction error of MLR in the non-realizable setting, where no generative model is assumed for the data.

The convergence rate of maximum-likelihood estimation for the parameters of a mixed GLM was derived by Ho et al. (2022). Chandrasekher et al. (2023) analyzed the performance of a class of iterative algorithms (not including AMP) for mixed GLMs, providing 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. Spectral estimators for mixed GLMs were studied in the recent work of Zhang et al. (2022), which characterizes their asymptotic performance for Gaussian designs and independent signals.

Statistical and computational limits for a two-layers neural network, a model similar to the matrix GLM, were studied by Aubin et al. (2018). A Vector AMP algorithm for MAP and MMSE inference in a similar multi-layer model was proposed by Pandit et al. (2020). Despite the similarities with the matrix GLM, to the best of our knowledge these works do not investigate AMP for the settings of mixed and max-affine regression.

Max-affine regression. For the non-parametric convex regression model in (5), the least squares estimator is $\hat{\varphi}^{(\mathrm{ls})} \in \mathrm{argmin}_{\varphi} \sum_{i=1}^{n} (Y_i - \varphi(X_i))^2$, where the minimization is over all convex functions φ. This least-squares estimator can be computed by solving a quadratic program. Theoretical properties of this estimator and algorithms to compute it were studied by Seijo and Sen (2011), Lim and Glynn (2012) and Mazumder et al. (2019). For the MAR model (4), several approaches for signal estimation have been proposed, including alternating minimization (Magnani and Boyd, 2009; Ghosh et al., 2022), convex adaptive partitioning (Hannah and Dunson, 2013), and adaptive max-affine partitioning (Balázs, 2016). Among them, theoretical guarantees have been established only for alternating minimization; these guarantees are in the regime where n is at least of order $p\log(n/p)$ (Ghosh et al., 2022). In contrast, in this paper we consider the high-dimensional regime where n is proportional to p as $n \to \infty$.

2. Preliminaries

Notation. All vectors (even rows of matrices) are treated as 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 $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, so that $\|x\|_r = \left(\sum_{i=1}^{n} |x_i|^r\right)^{1/r}$. Given random variables U, V, we write $U \stackrel{d}{=} V$ 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 $\stackrel{c}{\to}$. 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

$X_n \xrightarrow{c} x$, 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 $Y \sim Q$.

2.1 Model Assumptions

In the model (1), 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 sensing vectors $X_1, \dots, X_n$, i.e., $X = [X_1, \dots, X_n]^\top$. Similarly, the auxiliary variable matrix $\Psi \in \mathbb{R}^{n \times L_\Psi}$ is defined as $\Psi = [\Psi_1, \dots, \Psi_n]^\top$. The design 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 $d_r(\nu_p(B), P_{\bar{B}}) \xrightarrow{c} 0$ and $d_r(\nu_n(\Psi), P_{\bar{\Psi}}) \xrightarrow{c} 0$.

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 (1).

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 $k \geq 0$ we compute:

$$\begin{aligned} \Theta^k &= X\widehat{B}^k - \widehat{R}^{k-1}(F^k)^\top, \quad \widehat{R}^k = g_k(\Theta^k, Y), \\ B^{k+1} &= X^\top \widehat{R}^k - \widehat{B}^k(C^k)^\top, \quad \widehat{B}^{k+1} = f_{k+1}(B^{k+1}). \end{aligned} \tag{11}$$

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

$$C^k = \frac{1}{n} \sum_{i=1}^n g_k'(\Theta_i^k, Y_i), \quad F^{k+1} = \frac{1}{n} \sum_{j=1}^p f_{k+1}'(B_j^{k+1}), \tag{12}$$

where $g_k', f_{k+1}' \in \mathbb{R}^{L \times L}$ 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 (11) is $\mathcal{O}(npL)$.

State evolution. The "memory" terms $-\widehat{R}^{k-1}(F^k)^\top$ and $-\widehat{B}^k(C^k)^\top$ in (11) 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 $\mathrm{M}_B^k, \mathrm{T}_B^k, \mathrm{M}_\Theta^k, \mathrm{T}_\Theta^k \in \mathbb{R}^{L \times L}$. Recalling that the observation Y is generated via the function q according to (1), it is convenient to rewrite g_k in (11) 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:

$$h_k(z, u, v) := g_k(u,\, q(z, v)). \tag{13}$$

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

$$\mathrm{M}_B^{k+1} = \mathbb{E}[\partial_Z h_k(Z, Z^k, \bar{\Psi})], \tag{14}$$

$$\mathrm{T}_B^{k+1} = \mathbb{E}[h_k(Z, Z^k, \bar{\Psi}) h_k(Z, Z^k, \bar{\Psi})^\top], \tag{15}$$

$$\Sigma^{k+1} = \begin{bmatrix} \Sigma_{(11)}^{k+1} & \Sigma_{(12)}^{k+1} \\ \Sigma_{(21)}^{k+1} & \Sigma_{(22)}^{k+1} \end{bmatrix}, \tag{16}$$

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

$$\begin{aligned} \Sigma_{(11)}^{k+1} &= \frac{1}{\delta}\mathbb{E}\big[\bar{B}\bar{B}^\top\big], \\ \Sigma_{(12)}^{k+1} &= \Big(\Sigma_{(21)}^{k+1}\Big)^\top = \frac{1}{\delta}\mathbb{E}\big[\bar{B} f_{k+1}(\mathrm{M}_B^{k+1}\bar{B} + G_B^{k+1})^\top\big], \\ \Sigma_{(22)}^{k+1} &= \frac{1}{\delta}\mathbb{E}\big[f_{k+1}(\mathrm{M}_B^{k+1}\bar{B} + G_B^{k+1}) f_{k+1}(\mathrm{M}_B^{k+1}\bar{B} + G_B^{k+1})^\top\big]. \end{aligned} \tag{17}$$

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 (14)-(16) is initialized with $\Sigma^0 \in \mathbb{R}^{2L \times 2L}$ defined below in (21).

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

$$(Z, Z^k, \bar{\Psi}) \stackrel{d}{=} (Z, \mathrm{M}_\Theta^k Z + G_\Theta^k, \bar{\Psi}), \tag{18}$$

where $G_\Theta^k \sim \mathcal{N}(0, \mathrm{T}_\Theta^k)$,

$$\mathrm{M}_\Theta^k = \Sigma_{(21)}^k \big(\Sigma_{(11)}^k\big)^{-1}, \tag{19}$$

$$\mathrm{T}_\Theta^k = \Sigma_{(22)}^k \,-\, \Sigma_{(21)}^k \big(\Sigma_{(11)}^k\big)^{-1} \Sigma_{(12)}^k. \tag{20}$$

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 (11).

(A1) There exists $\Sigma^0 \in \mathbb{R}^{2L\times 2L}$ and $c_0 \in \mathbb{R}$ such that as $n, p \to \infty$ (with $n/p \to \delta$), we have

$$\frac{1}{n}\begin{bmatrix} B^\top B & B^\top \widehat{B}^0 \\ (\widehat{B}^0)^\top B & (\widehat{B}^0)^\top \widehat{B}^0 \end{bmatrix} \xrightarrow{c} \Sigma^0, \tag{21}$$

$$\frac{1}{p}\sum_{j=1}^{p}\sum_{l=1}^{L} |\widehat{B}^0_{jl}|^r \xrightarrow{c} c_0. \tag{22}$$

Here $r \in [2, \infty)$ is the same as that used for the assumptions on the signal matrix at the end of Section 2.1. Furthermore, there exists a Lipschitz $F_0 : \mathbb{R}^L \to \mathbb{R}^L$ such that $\frac{1}{p}(\widehat{B}^0)^\top \phi(B) \xrightarrow{c} \mathbb{E}[F_0(\bar{B})\phi(\bar{B})^\top]$ 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 (13) 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-matrix) 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* (11) *for the matrix GLM model in* (1)*. Suppose that the model assumptions in Section 2.1 as well as* ***(A1)*** *and* ***(A2)*** *are satisfied, and that* T^1_B *is positive definite. Then for each* $k \geq 0$*, we have*

$$\sup_{\phi\in \mathrm{PL}_{2L}(r,1)} \Big|\frac{1}{p}\sum_{j=1}^{p} \phi(B^{k+1}_j, B_j) - \mathbb{E}[\phi(\mathrm{M}^{k+1}_B \bar{B} + G^{k+1}_B, \bar{B})]\Big| \xrightarrow{c} 0, \tag{23}$$

$$\sup_{\phi\in \mathrm{PL}_{2L+L_\Psi}(r,1)} \Big|\frac{1}{n}\sum_{i=1}^{n} \phi(\Theta^k_i, \Theta_i, \Psi_i) - \mathbb{E}[\phi(\mathrm{M}^k_\Theta Z + G^k_\Theta, Z, \bar{\Psi})]\Big| \xrightarrow{c} 0, \tag{24}$$

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,* $G^{k+1}_B \sim \mathcal{N}(0, \mathrm{T}^{k+1}_B)$ *is independent of* $\bar{B}$*, and* $G^k_\Theta \sim \mathcal{N}(0, \mathrm{T}^k_\Theta)$ *is independent of* $(Z, \bar{\Psi})$*.*

The proof of the theorem is given in Section 5.1. The result (23) 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 $(\mathrm{M}^{k+1}_B \bar{B} + G^{k+1}_B, \bar{B})$; see (Feng et al., 2022, Corollary 7.21). An analogous statement holds for (24).

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 $\widehat{B}^k = \left[\hat{\beta}^{(1),k}, \ldots, \hat{\beta}^{(L),k}\right]$. Note that $\hat{\beta}^{(\ell),k}$ is the estimate of the signal $\beta^{(\ell)}$ after k iterations, and 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:

$$\frac{\langle \hat{\beta}^{(\ell),k}, \beta^{(\ell)} \rangle^2}{\|\hat{\beta}^{(\ell),k}\|_2^2 \|\beta^{(\ell)}\|_2^2} \xrightarrow{c} \frac{(\mathbb{E}[f_{k,\ell}(\bar{B}^k)\bar{B}_\ell])^2}{\mathbb{E}[f_{k,\ell}(\bar{B}^k)^2]\mathbb{E}[\bar{B}_\ell^2]}, \quad \ell \in [L]. \tag{25}$$

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:

$$\frac{\|\beta^{(\ell)} - \hat{\beta}^{(\ell),k}\|_2^2}{p} \xrightarrow{c} \mathbb{E}\Big[\big(\bar{B}_\ell - f_{k,\ell}(\bar{B}^k)\big)^2\Big], \quad \ell \in [L]. \tag{26}$$

3.1 Choosing the Functions of AMP

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:

$$\begin{aligned} \widetilde{Z}^k &:= Z + \left(\mathrm{M}_\Theta^k\right)^{-1} G_\Theta^k, \\ \widetilde{B}^{k+1} &:= \bar{B} + \left(\mathrm{M}_B^{k+1}\right)^{-1} G_B^{k+1}. \end{aligned} \tag{27}$$

(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:

$$\begin{aligned} \mathrm{Cov}(\widetilde{Z}^k - Z) &= \left(\mathrm{M}_\Theta^k\right)^{-1} \mathrm{T}_\Theta^k \left(\left(\mathrm{M}_\Theta^k\right)^{-1}\right)^\top =: N_\Theta^k, \\ \mathrm{Cov}(\widetilde{B}^{k+1} - \bar{B}) &= \left(\mathrm{M}_B^{k+1}\right)^{-1} \mathrm{T}_B^{k+1} \left(\left(\mathrm{M}_B^{k+1}\right)^{-1}\right)^\top =: N_B^{k+1}. \end{aligned} \tag{28}$$

From (20), 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 (17)). Similarly, from (14) and (15), M_B^{k+1} and 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 (28). 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 $\widetilde{Z}^k, \widetilde{B}^{k+1}$, 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*

$$f_k^*(s) = \mathbb{E}[\bar{B} \mid \mathrm{M}_B^k \bar{B} + G_B^k = s], \tag{29}$$

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*

$$g_k^*(u,\, y) = \mathrm{Cov}[Z \mid Z^k = u]^{-1}\big(\mathbb{E}[Z \mid Z^k = u, \bar{Y} = y] - \mathbb{E}[Z \mid Z^k = u]\big). \tag{30}$$

Here $\begin{bmatrix} Z \\ Z^k \end{bmatrix} \sim \mathcal{N}(0, \Sigma^k)$ *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

In this section, we present numerical results for mixed linear regression (Eq. (2)), max-affine regression (Eq. (4)), and mixture-of experts (Eq. (9)). For MLR and MAR, we compare the performance of AMP with other popular estimators.

4.1 Mixed Linear Regression (Two Signals)

Consider the MLR model (2) with two signals, where

$$Y_i = \langle X_i, \beta^{(1)} \rangle c_i \,+\, \langle X_i, \beta^{(2)} \rangle (1 - c_i) \,+\, \epsilon_i, \quad i \in [n]. \tag{31}$$

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

The AMP algorithm in (11) is implemented with $g_k = g_k^*$, the optimal choice given by (30). For the function f_k, we use the Bayes-optimal f_k^* in (29) unless stated otherwise. In Appendix A, we provide the implementation details, and show how the functions f_k, g_k and their derivatives can approximated from the data.

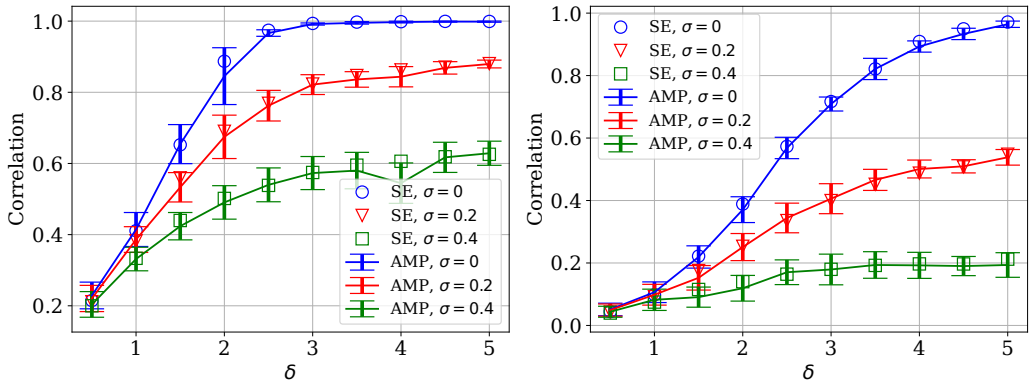
The performance in all the plots is measured via the normalized squared correlation between the AMP estimate and the signal (see (25)). Each point on the plots is obtained from 10 independent runs, where in each run, AMP is executed for 10 iterations. We report the average and error bars at 1 standard deviation of the final iteration.

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

$$(\beta_j^{(1)}, \beta_j^{(2)}) \sim_{\text{i.i.d.}} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right), \quad j \in [p]. \tag{32}$$

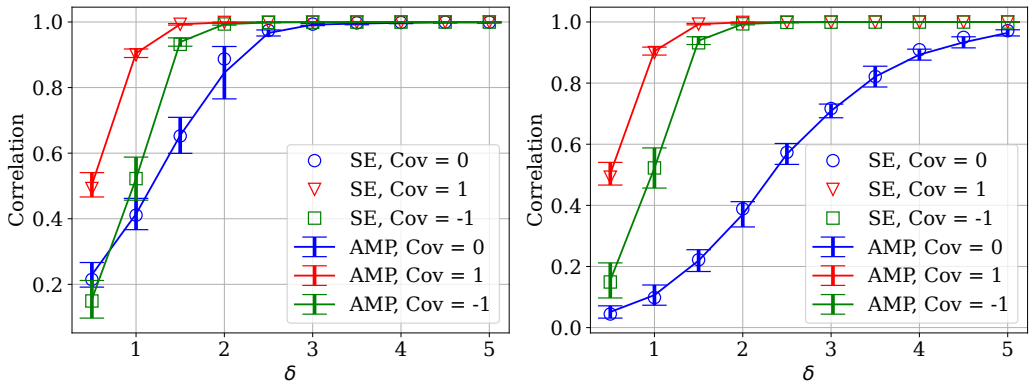
The initializer $\widehat{B}^0 \in \mathbb{R}^{p \times 2}$ is chosen randomly according to the same distribution, independently of the signal.

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 $\delta = n/p$, for different noise levels σ. The state evolution predictions closely match the performance of AMP for practical values of n, p, validating the result of Theorem 1. As expected, the correlation improves



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

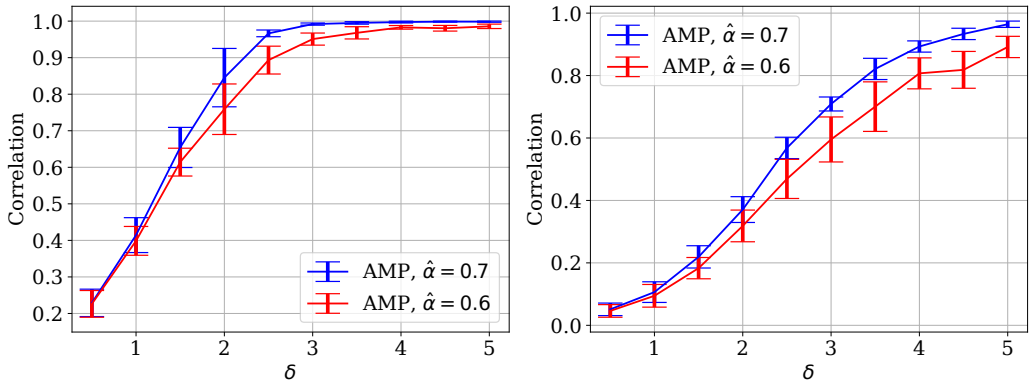
Figure 1: MLR, Gaussian prior with $\rho = 0$: normalized squared correlation vs. δ for various noise levels σ, with $\alpha = 0.7$.



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

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

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 $\beta^{(1)} = -\beta^{(2)} = \beta$, 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 β or $-\beta$ (with probabilities 0.7 and 0.3, respectively). We note that the case of $\rho = -1$ and $\alpha = 0.5$ is the phase retrieval problem, for which AMP algorithms have been studied in a number of works, e.g., (Schniter and Rangan, 2014; Ma et al., 2019). AMP needs to be initialized carefully in this setting since a random initialization independent of



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

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

the signal leads to state evolution predicting zero correlation between the signal and the AMP iterates (Ma et al., 2018; Mondelli and Venkataramanan, 2021).

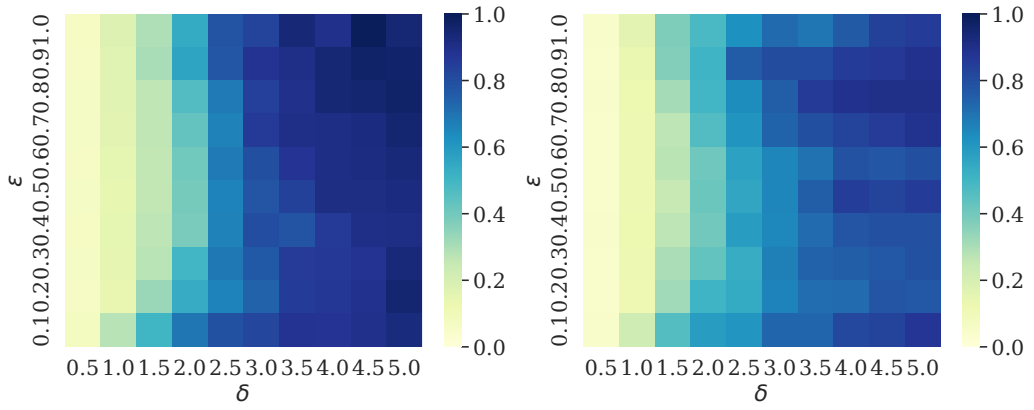
In practical applications, we may not know the exact proportion of observations that come that come from the first signal. Figure 3 shows the performance when AMP is run assuming a proportion parameter $\hat{\alpha} = 0.6$ which is different from the true value $\alpha = 0.7$. The functions f_k^*, g_k^* 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.

Sparse prior. We next consider a sparse prior for each of the two signals, with their entries generated as

$$\beta_j^{(1)}, \beta_j^{(2)} \sim_{\text{i.i.d.}} \frac{\varepsilon}{2}\delta_{+1} + (1-\varepsilon)\delta_0 + \frac{\varepsilon}{2}\delta_{-1}, \quad j \in [p]. \tag{33}$$

Here $\delta_{(\cdot)}$ denotes the Dirac delta function, and we note that the two signals are independent. The initializer is generated randomly from the same prior, independently of the signals. We investigate the performance of AMP with two choices of denoising function: the Bayes-optimal denoising function (defined in (29)) and the soft-thresholding denoising function (defined in (34)-(36) below). For the case of standard linear regression, the soft-thresholding function is a popular choice of denoiser for AMP when the signal is known to be sparse, but the exact sparsity level and the distribution of the non-zero coefficients are not known (Montanari, 2012). AMP with soft-thresholding denoising is also closely related to LASSO, which is widely used for sparse linear regression (Bayati and Montanari, 2011b).

We evaluate the two denoisers in Figures 4 and 5, respectively, by plotting heatmaps showing the performance for various values of the pair (δ, ε). For each point in the heatmap, we take the minimum of the mean normalized squared correlation of the two estimates with the respective signals. This is obtained by executing 10 runs of AMP using the desired f_k function with 10 iterations per run (i.e., $k = 1, \ldots, 10$), and taking the average of the 10 correlations (from the 10 runs) at the final iteration.



(a) $\alpha = 0.6$ (b) $\alpha = 0.7$

Figure 4: MLR: Heatmap of minimum normalized correltion for Bayes-optimal f_k, with $p = 500$, $\sigma = 0$.

For the Bayes-optimal denoiser f_k, the heatmaps are shown in Figure 4. We have the following observations:

- Performance is better for $\alpha = 0.6$ compared to $\alpha = 0.7$. This is because we are taking the minimum between the two correlations, and when α is larger ($\alpha = 0.7$), there is less data available for the group with fewer observations (for a given (ε, δ)).

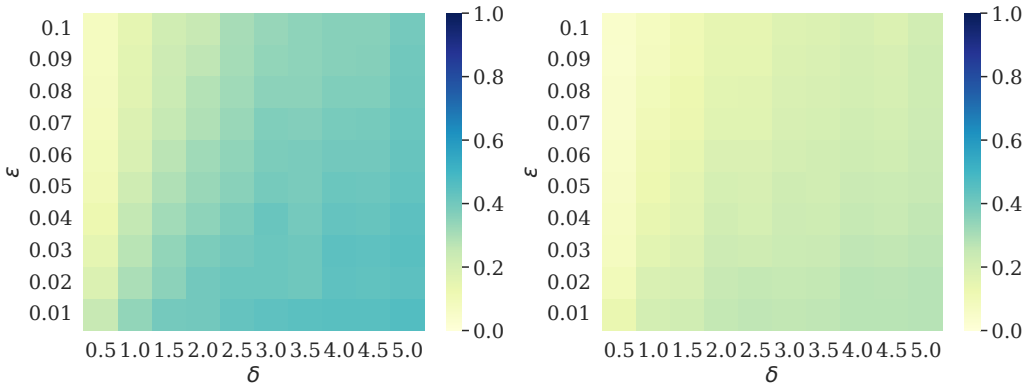
- For a given δ, performance is generally better at ε closer to 0 or 1. This is because at $\varepsilon = 0.1$, most of the signal entries are 0, and at $\varepsilon = 1$, all the values are either 1 or -1. Around $\varepsilon = 0.5$, we have a significant proportion of all three values, causing estimation to be harder.

The soft-thresholding function with threshold θ, denoted by $\mathrm{ST}(\cdot\,;\theta) : \mathbb{R} \to \mathbb{R}$, is defined as

$$\mathrm{ST}(x;\theta) = \begin{cases} x - \theta & \text{if } x > \theta \\ 0 & \text{if } -\theta \le x \le \theta \\ x + \theta & \text{if } x \le -\theta. \end{cases} \tag{34}$$

To set the threshold for the soft-thresholding denoiser f_k, we recall from Theorem 1 that the empirical distribution of $\{B_j^k\}$ converges to the distribution of the random vector $\mathrm{M}_B^k \bar{B} + G_B^k$. Therefore, the empirical distribution of $\{(\mathrm{M}_B^k)^{-1} B_j^k\}_{j\in[p]}$ converges to the distribution of $\bar{B} + (\mathrm{M}_B^k)^{-1} G_B^k$, where $(\mathrm{M}_B^k)^{-1} G_B^k \sim \mathcal{N}(0, N_B^k)$, where

$$N_B^k = \mathrm{Cov}\Big((\mathrm{M}_B^k)^{-1} G_B^k\Big) = (\mathrm{M}_B^k)^{-1} T_B^k (\mathrm{M}_B^k)^{-1}. \tag{35}$$



(a) $\alpha = 0.5$ (b) $\alpha = 0.6$

Figure 5: MLR: Heatmap of minimum normalized correlation for soft-thresholding f_k, with $p = 1000$, $\sigma = 0$. Soft-thresholding tuning parameter $\zeta = 1.1402$.

Letting $\zeta > 0$ be a tuning parameter, we set the soft-thresholding denoiser f_k to be:

$$f_k(B_j^k) = \begin{bmatrix} \mathrm{ST}\Big(\big\{(\mathrm{M}_B^k)^{-1}B_j^k\big\}_1; \zeta\sqrt{\big\{N_B^k\big\}_{11}}\Big) \\ \mathrm{ST}\Big(\big\{(\mathrm{M}_B^k)^{-1}B_j^k\big\}_2; \zeta\sqrt{\big\{N_B^k\big\}_{22}}\Big) \end{bmatrix}, \tag{36}$$

This implies that

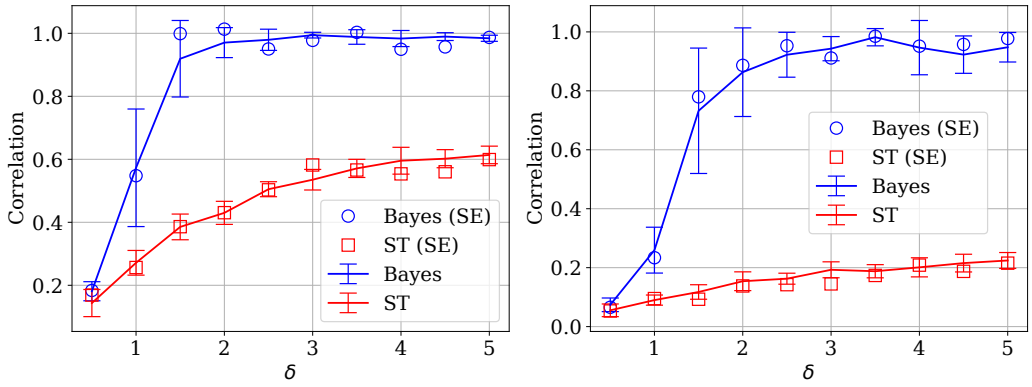
$$\nabla f_k(B_j^k) = \begin{bmatrix} \frac{\partial\{f_k\}_1}{\partial\{B_j^k\}_1} & \frac{\partial\{f_k\}_1}{\partial\{B_j^k\}_2} \\ \frac{\partial\{f_k\}_2}{\partial\{B_j^k\}_1} & \frac{\partial\{f_k\}_2}{\partial\{B_j^k\}_2} \end{bmatrix}, \tag{37}$$

where for $i_1, i_2 \in \{1, 2\}$,

$$\frac{\partial\{f_k\}_{i_1}}{\partial\{B_j^k\}_{i_2}} = \begin{cases} \{(M_B^k)^{-1}\}_{i_1 i_2} & \text{if } \big\{(\mathrm{M}_B^k)^{-1}B_j^k\big\}_{i_1} > \zeta\sqrt{\big\{N_B^k\big\}_{i_1 i_1}} \\ 0 & \text{if } \big|\big\{(\mathrm{M}_B^k)^{-1}B_j^k\big\}_{i_1}\big| \le \zeta\sqrt{\big\{N_B^k\big\}_{i_1 i_1}} \\ \{(M_B^k)^{-1}\}_{i_1 i_2} & \text{if } \big\{(\mathrm{M}_B^k)^{-1}B_j^k\big\}_{i_1} < -\zeta\sqrt{\big\{N_B^k\big\}_{i_1 i_1}}. \end{cases} \tag{38}$$

Here the notation $\{\cdot\}_{i_1}$ denotes the i_1-th entry of the vector and $\{\cdot\}_{i_1 i_2}$ the i_1, i_2-th entry of the matrix inside the parentheses.

Figure 5 shows the heatmaps for the soft-thresholding, with the tuning parameter ζ set to 1.1402. This value of ζ attains the minimax MSE of the soft-thresholding denoiser over the class of sparse signal priors which assign a probability mass at least 0.9 to the value 0 (Montanari, 2012). We observe that the performance for $\alpha = 0.5$ is stronger as the samples are more evenly spread out between the two signals. As expected the correlation improves as the signal becomes sparser (i.e., ε decreases), and as δ increases. Figure 6 compares the Bayes-optimal function with the soft-thresholding function for fixed values of sparsity level $\varepsilon = 0.1$ and mixture parameter $\alpha = 0.6$. The significantly better performance of the



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

Figure 6: MLR: Comparison of minimum normalized correlation for Bayes-optimal f_k vs. soft-thresholding f_k, with $p = 1000$, $\alpha = 0.6$, $\sigma = 0$, $\zeta = 1.1402$, and $\varepsilon = 0.1$.

the Bayes-optimal denoiser compared to soft-thresholding is because the former optimally utilizes knowledge of the signal prior, whereas soft-thresholding only uses an estimate of the proportion of zeros in the signals.

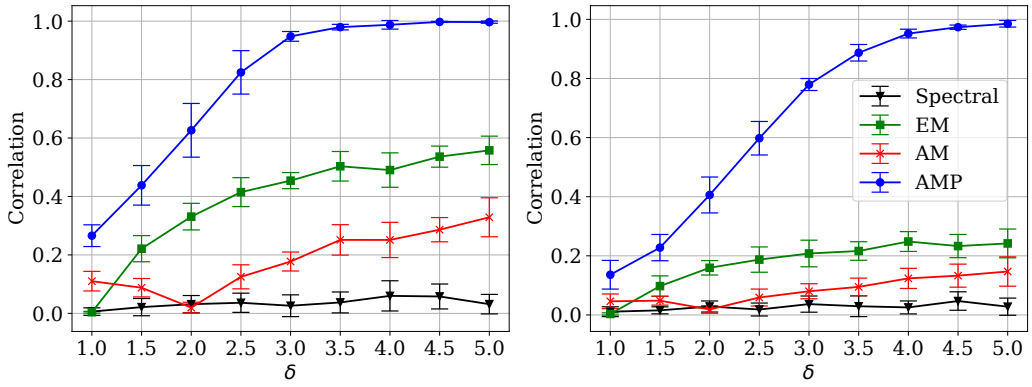
Comparison with other estimators. Figure 7 compares the performance of AMP with other widely studied estimators for mixed linear regression, for the Gaussian signal prior in (32) with independent signals ($\rho = 0$). The other estimators are: the spectral estimator proposed in (Yi et al., 2014, Algorithm 2); alternating minimization (AM) (Yi et al., 2014, Algorithm 1); and expectation maximization (EM) (Faria and Soromenho, 2010, Section 2.1). Figure 8 compares the performance of AMP with these estimators for the sparse signal prior in (33) with $\varepsilon = 0.1$. 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. (2010), but it did not 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.

4.2 Mixed Linear Regression (Three Signals)

To illustrate AMP's ability to tackle MLR with more than two signals, we now consider the model (2) with three signals:

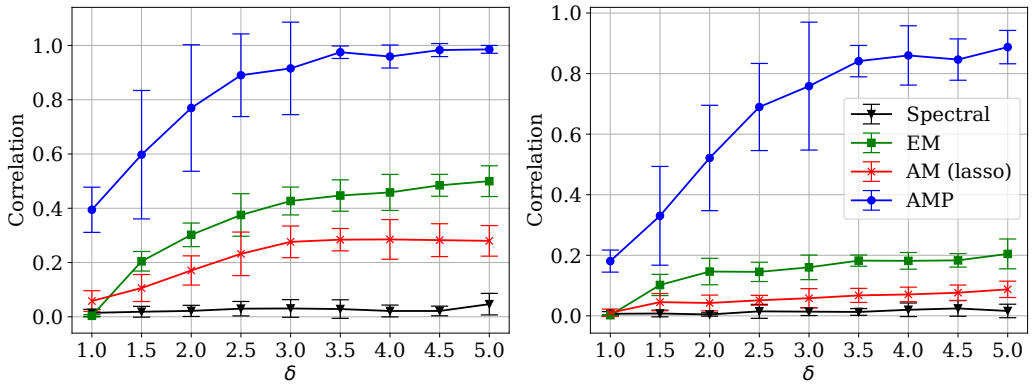
$$Y_i = \langle X_i, \beta^{(1)} \rangle c_{i1} + \langle X_i, \beta^{(2)} \rangle c_{i2} + \langle X_i, \beta^{(3)} \rangle c_{i3} + \epsilon_i, \qquad i \in [n]. \tag{39}$$

We take $[c_{i1}, c_{i2}, c_{i3}]^\top$ to be a one-hot vector, and denote the position of the one in the one-hot vector by $c_i \sim_{\text{i.i.d.}} \text{Categorical}(\{\alpha_1, \alpha_2, \alpha_3\})$. As before, $\epsilon_i \sim_{\text{i.i.d.}} \mathcal{N}(0, \sigma^2)$, 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 (11) is implemented with $g_k = g_k^*$ and $f_k = f_k^*$ (i.e., the optimal choices given in Proposition 2).



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

Figure 7: MLR, 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) $\beta^{(2)}$

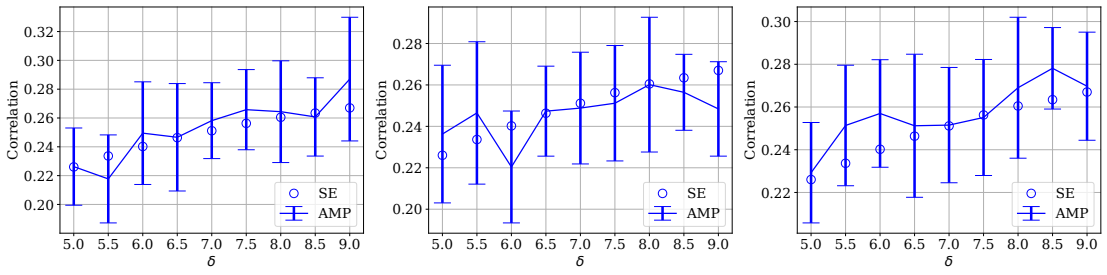
Figure 8: MLR, comparison of different estimators for sparse prior: Normalized squared correlation vs. δ, with $\alpha = 0.6$, $\sigma = 0.1$.

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

$$\begin{aligned}(\beta_j^{(1)}, \beta_j^{(2)}, \beta_j^{(3)}) &\sim_{\text{i.i.d.}} \mathcal{N}(\mathbb{E}[\bar{B}], I_3), \quad j \in [p] \\ c_i &\sim_{\text{i.i.d.}} \text{Categorical}(\{\alpha_1, \alpha_2, \alpha_3\}), \quad i \in [n].\end{aligned} \tag{40}$$

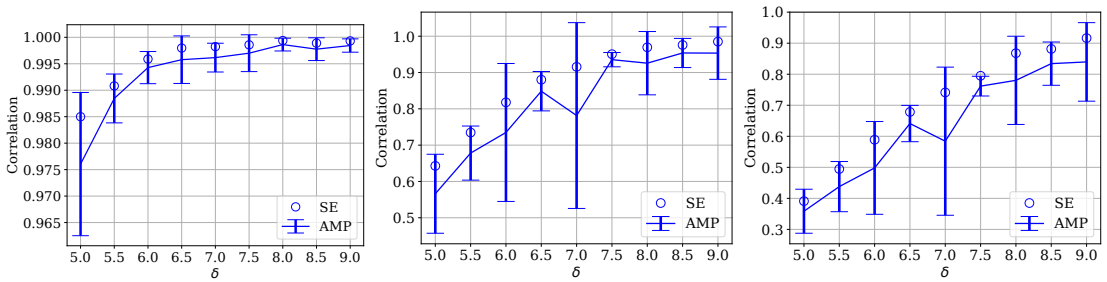
The initializer $\widehat{B}^0 \in \mathbb{R}^{p \times 3}$ is chosen randomly according to the same distribution, independent of the signal. We consider the following three scenarios, where $\sigma = 0$ (noiseless):

- **Signals with same mean and same proportions.** Figure 9 shows the performance with $\mathbb{E}[\bar{B}] = [0, 0, 0]^\top$ and $(\alpha_1, \alpha_2, \alpha_3) = (1/3, 1/3, 1/3)$. We observe that the



(a) $\beta^{(1)}$ (b) $\beta^{(2)}$ (c) $\beta^{(3)}$

Figure 9: MLR with three signals: signals with same mean and same proportions.



(a) $\beta^{(1)}$ (b) $\beta^{(2)}$ (c) $\beta^{(3)}$

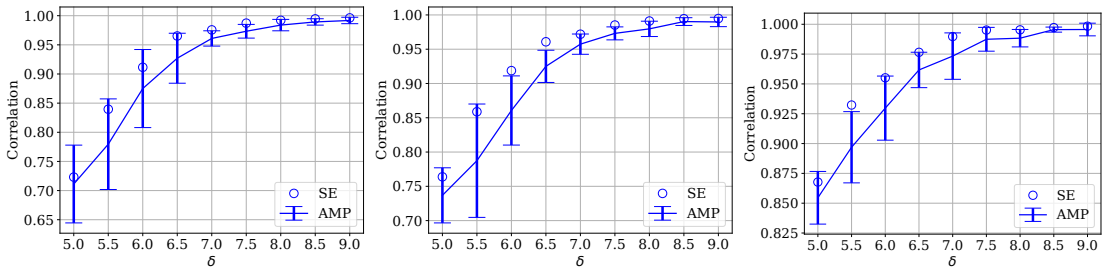
Figure 10: MLR with three signals: signals with same mean and different proportions.

performance does not improve much with increasing δ as the algorithm finds it challenging to distinguish the signals when they all have the same prior and correspond to the same proportion of observations.

- **Signals with same mean and different proportions.** Figure 10 shows the performance with $\mathbb{E}[\bar{B}] = [0, 0, 0]^\top$ and $(\alpha_1, \alpha_2, \alpha_3) = (0.5, 0.3, 0.2)$. The performance here is significantly better than the previous case where signals have the same mean and proportions. As expected, the correlation for β_1 is significantly better than that for β_2 and β_3 since β_1 has the highest proportion of observations.

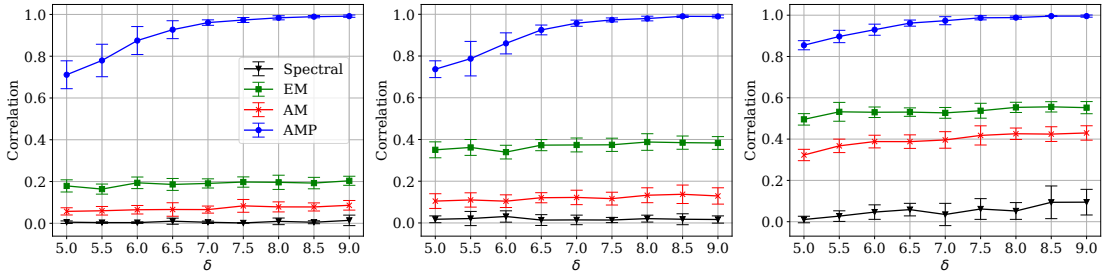
- **Signals with different means and same proportions.** Figure 11 shows the performance with $\mathbb{E}[\bar{B}] = [0, 0.5, 1]^\top$ and $(\alpha_1, \alpha_2, \alpha_3) = (1/3, 1/3, 1/3)$. This is the case with the best estimation performance. This is because the distinct means help distinguish the signals from one another and the equal proportions ensure that all three have sufficient number of observations for large enough δ.

Finally, Figure 12 compares the performances of AMP with other widely studied estimators for MLR, for the Gaussian signal prior in (40) with $\mathbb{E}[\bar{B}] = [0, 0.5, 1]^\top$ and $(\alpha_1, \alpha_2, \alpha_3) = (1/3, 1/3, 1/3)$ (this is the case where signals have different prior distributions but appear in the same proportion of observations). 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,



(a) $\beta^{(1)}$ (b) $\beta^{(2)}$ (c) $\beta^{(3)}$

Figure 11: MLR with three signals: signals with different means and same proportions.



(a) $\beta^{(1)}$ (b) $\beta^{(2)}$ (c) $\beta^{(3)}$

Figure 12: MLR with three signals: Comparison with different estimators. The signals have different means and occur in the same proportions.

2010, Section 2.1). 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 (to account for the fact 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 (Álvaro González, 2010) to achieve this approximately and efficiently in our 3D case. As in the case of two-signal MLR, 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.

4.3 Max-Affine Regression

We consider on the MAR model (4) with two signals, which is given by:

$$Y_i = \max\left\{\langle X_i, \beta^{(1)}\rangle + b_1,\, \langle X_i, \beta^{(2)}\rangle + b_2\right\} + \epsilon_i, \quad i \in [n], \tag{41}$$

where $\epsilon_i \sim_{\text{i.i.d.}} \mathcal{N}(0, \sigma^2)$, $X_i \sim_{\text{i.i.d.}} \mathcal{N}(0, I_p/n)$ for $i \in [n]$. Recall from (6) that MAR can be written as an instance of the matrix GLM using the augmented features $X_i^{(\text{ma})} =$

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.

Algorithm 1 Expectation-maximization approximate message passing (EM-AMP)

1: Initialize the intercepts $b^0 := (b_1^0, b_2^0)$, and $\widehat{B}^{k_{\max},0}$.
2: **for** iteration $m := 1, \ldots, m_{\max}$ **do**
3: Compute $\mathbb{E}\big[Z|\bar{Y}; b^m\big]$, and run AMP with intercept estimates b^m as part of the model for $k_{\max}$ iterations to produce $\widehat{\Theta}^m = X\widehat{B}^{k_{\max},m}$. Let $\widehat{\Theta}^{(1)}$, $\widehat{\Theta}^{(2)}$ be the two columns of $\widehat{\Theta}^m$.
4: Compute $b_1^{m+1} := \frac{1}{|\{i:\widehat{\Theta}_i^{(1)}+b_1^m > \widehat{\Theta}_i^{(2)}+b_2^m\}|} \sum_{i:\widehat{\Theta}_i^{(1)}+b_1^m > \widehat{\Theta}_i^{(2)}+b_2^m} Y_i - \mathbb{E}\big[Z_1|\bar{Y}; b^m\big]$.
5: Compute $b_2^{m+1} := \frac{1}{|\{i:\widehat{\Theta}_i^{(1)}+b_1^m \le \widehat{\Theta}_i^{(2)}+b_2^m\}|} \sum_{i:\widehat{\Theta}_i^{(1)}+b_1^m \le \widehat{\Theta}_i^{(2)}+b_2^m} Y_i - \mathbb{E}\big[Z_2|\bar{Y}; b^m\big]$.
6: Output $b^{m_{\max}}$ and $\widehat{B}^{k_{\max},m_{\max}}$.

$\begin{bmatrix} X_i \\ 1 \end{bmatrix} \in \mathbb{R}^{p+1}$, $i \in [n]$. Since the augmented features are not i.i.d. Gaussian (due to the last component being 1), we use the original formulation of MAR in (4) and consider the intercepts b_1 and b_2 to be unknown parameters of the output function $q(\cdot, \cdot)$.

Our solution is to estimate the unknown intercepts using an expectation-maximization (EM) algorithm. The EM algorithm iteratively produces intercept estimates, denoted by $b^m \equiv (b_1^m, b_2^m)$, for $m \ge 1$. However, the expectation step of the EM algorithm requires an estimate of $\Theta = XB$, which we approximate via AMP. This leads to a combined expectation-maximization approximate message passing (EM-AMP) algorithm, which is described in Algorithm 1. The idea of combining AMP with the EM algorithm was introduced by Vila and Schniter (2013), for sparse linear regression with unknown parameters in the signal prior.

The AMP stage in step 3 of Algorithm 1 is implemented with $g_k = g_k^*$ and $f_k = f_k^*$ (i.e., the optimal choices), computed using the current intercept estimates. The details of computing the g_k^* and the conditional expectation $\mathbb{E}\big[Z|\bar{Y}; b^m\big]$ in this step are given in Appendix B. The derivation of the EM updates in steps 4 and 5 of Algorithm 1 is given in Section 5.3.

We set the signal dimension $p = 500$ and vary the value of n in our experiments. We consider different choices for the intercepts $b := (b_1, b_2)$ and use a Gaussian prior for $B = \big(\beta^{(1)}, \beta^{(2)}\big)$, where we generate

$$\big(\beta_j^{(1)}, \beta_j^{(2)}\big) \sim_{\text{i.i.d.}} \mathcal{N}\big(\mathbb{E}[\bar{B}], I_2\big), \quad j \in [p]. \tag{42}$$

The initializer $\widehat{B}^0 \in \mathbb{R}^{p\times 2}$ is chosen according to the same distribution, independently of the signal. The EM initialization b^0 is taken to be $(0, 0)$.

Figures 13-15 show the performance of EM-AMP for max-affine regression with different choices of prior, intercepts, and noise level. The performance in all the plots is measured via the normalized squared correlation between the full signals $\beta_{\text{ma}}^{(1)} = ((\beta^{(1)})^\top, b_1)^\top$ and $\beta_{\text{ma}}^{(2)} = ((\beta^{(2)})^\top, b_2)^\top$ and their respective estimates $\hat{\beta}_{\text{ma}}^{(1)} = ((\hat{\beta}^{(1)})^\top, \hat{b}_1)^\top$ and $\hat{\beta}_{\text{ma}}^{(2)} = ((\hat{\beta}^{(2)})^\top, \hat{b}_2)^\top$, i.e.,

$$\frac{\langle \beta_{\text{ma}}^{(l)}, \hat{\beta}_{\text{ma}}^{(l)} \rangle^2}{\|\hat{\beta}_{\text{ma}}^{(l)}\|_2^2 \|\beta_{\text{ma}}^{(l)}\|_2^2}, \quad \text{where } l \in \{1, 2\}. \tag{43}$$

Each point on the plots is obtained from 5 independent runs, where in each run, we execute EM-AMP with $m_{\max} = 5$ and $k_{\max} = 5$. We report the average and error bars at 1 standard deviation of the final iteration. EM-AMP is compared with: (i) the alternating minimization algorithm (Ghosh et al., 2022) (the only known algorithm for MAR with theoretical guarantees), and (ii) the Oracle AMP (OR-AMP) where we assume that the true intercepts b are known and are part of the matrix GLM model. Though the intercepts are not known in practice, OR-AMP provides an upper bound on the best correlation achievable by AMP since it uses the optimal denoising functions and the correct intercepts. Hence, it is reasonable to expect that OR-AMP would provide the best performance.

We study the performance of our algorithms for the following three scenarios:

- **Same intercept, signals with different means.** Figure 13 shows the results for the setting $b = (1, 1)$, $\mathbb{E}[\bar{B}] = [0, 1]^\top$, $\sigma = 0.1$. When the intercepts are the same, the proportion of observations from each signal is roughly the same for large enough δ. This is because $Y_i = \max_{l\in\{1,2\}} \left(\langle X_i, \beta^{(l)}\rangle + b_l\right)$, where $\langle X_i, \beta^{(l)}\rangle$ is a zero-mean Gaussian regardless of the mean of $\beta^{(l)}$. (The variance of $\langle X_i, \beta^{(l)}\rangle$ depends on the mean of $\beta^{(l)}$.) In this setting, AM performs poorly for smaller δ values, but matches or slightly exceeds the performance of EM-AMP for large δ.

- **Different intercepts, signals with different means.** Figure 14 shows the results for the setting $b = (1, 0)$, $\mathbb{E}[\bar{B}] = [0, 1]^\top$, $\sigma = 0.1$. As mentioned above, the signal mean does not affect the proportion of observations from each sample. Hence, the signal with a larger intercept will have more observations. EM-AMP outperforms AM for the signal with fewer observations for all values δ, while for the other signal, AM is slightly better for larger values of δ.

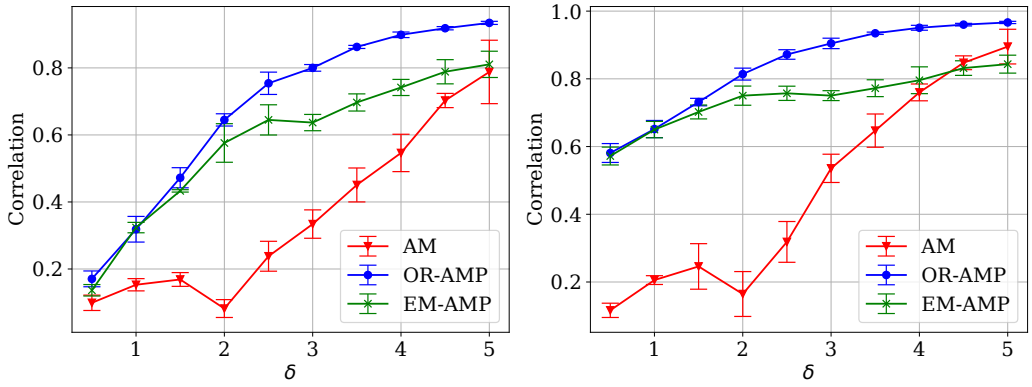
- **Same intercept, signals with different means, higher noise level.** Figure 15 shows the results for $b = (1, 1)$, $\mathbb{E}[\bar{B}] = [0, 1]^\top$, $\sigma = 0.4$. The plots show that EM-AMP significantly outperforms AM for all values of δ. AM is quite sensitive to the presence of noise unlike EM-AMP, which is more robust and nearly matches the performance OR-AMP.

Hard case. When entries of both $\beta^{(1)}$ and $\beta^{(2)}$ are generated from the same distribution, and the intercepts b_1 and b_2 are the same, the estimation problem becomes very challenging. In this case, AM, EM-AMP, and AMP all struggle to give an accurate estimate.

4.4 Mixture-of-Experts

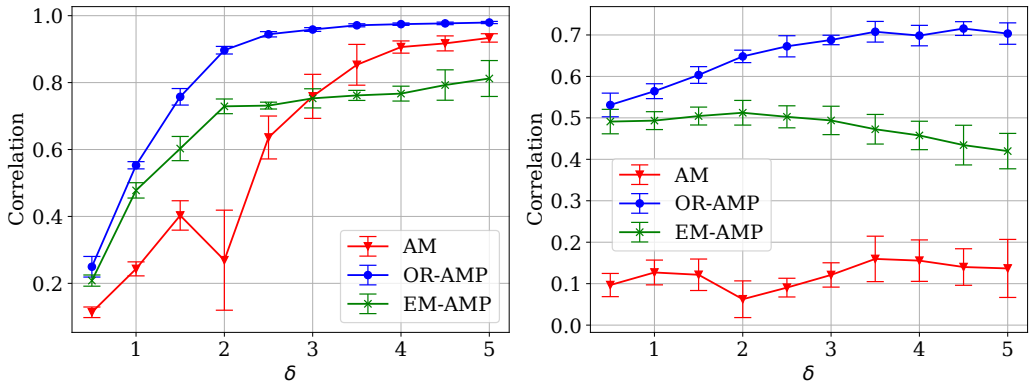
We consider the MOE model (9) with two regressors, two gating parameters, and the identity activation function $\tilde{q}(x) = x$. The model is given by

$$Y_i = \langle X_i, \beta^{(1)}\rangle \mathbb{1}\left\{\psi_i \le \frac{\exp(\langle X_i, w^{(1)}\rangle)}{\exp(\langle X_i, w^{(1)}\rangle) + \exp(\langle X_i, w^{(2)}\rangle)}\right\} + \langle X_i, \beta^{(2)}\rangle \mathbb{1}\left\{\psi_i > \frac{\exp(\langle X_i, w^{(1)}\rangle)}{\exp(\langle X_i, w^{(1)}\rangle) + \exp(\langle X_i, w^{(2)}\rangle)}\right\} + \epsilon_i, \tag{44}$$



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

Figure 13: MAR: Same intercepts, signals with different means, noise level $\sigma = 0.1$.



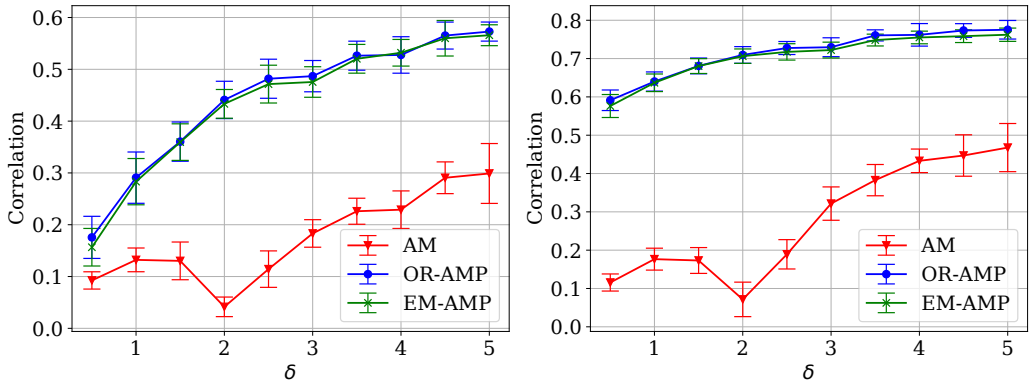
(a) $\beta^{(1)}$ (b) $\beta^{(2)}$

Figure 14: MAR: Different intercepts, signals with different means, noise level $\sigma = 0.1$.

where $\psi_i \sim_{\text{i.i.d.}} \text{Uniform}[0,1]$, $\epsilon_i \sim_{\text{i.i.d.}} \mathcal{N}(0,\sigma^2)$, $X_i \sim_{\text{i.i.d.}} \mathcal{N}(0, I_p/n)$ for $i \in [n]$. The signal dimension is set to $p = 500$ and the value of n is varied in our experiments. We use a Gaussian prior for $B = (\beta^{(1)}, \beta^{(2)}, w^{(1)}, w^{(2)})$, where we generate

$$(\beta_j^{(1)}, \beta_j^{(2)}, w_j^{(1)}, w_j^{(2)}) \sim_{\text{i.i.d.}} \mathcal{N}([1,2,3,4]^\top, I_4), \quad j \in [p]. \tag{45}$$

We run the AMP algorithm in (11) with $g_k = g_k^*$ and $f_k = f_k^*$ (i.e., the optimal choices). The initializer $B^0 \in \mathbb{R}^{p\times 4}$ is chosen according to the same distribution, independently of the signal. The details of the implementation are given in Appendix C. Figure 16 shows the performance of AMP for MOE. The performance in the plots is measured via the normalized squared correlation given in (25). Each point on the plots is obtained from 5 independent runs, where in each run, we execute AMP with $k = 5$. We report the average and error bars at 1 standard deviation of the final iteration. Figure 16 indicates a good match between the empirical performance of AMP and the theoretical state evolution predictions. The



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

Figure 15: MAR: Same intercept, signals with different means, noise level $\sigma = 0.4$.

improvement across increasing δ's is more significant for the regressors than the gating parameters since the latter are easier to estimate because of their larger means.

5. Proofs and Derivations

In this section, we provide detailed proofs and derivations of our results.

5.1 Proof of Theorem 1

To prove the theorem, we use a change of variables to rewrite (11) as a new matrix-valued AMP iteration. The new iteration is a special case of an abstract AMP iteration for 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 (23)-(24).

Given the iteration (11), for $k \geq 0$ define

$$\check{B}^{k+1} := B^{k+1} - B(\mathrm{M}_B^{k+1})^\top, \quad \check{\Theta}^k := (\Theta, \Theta^k), \tag{46}$$

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

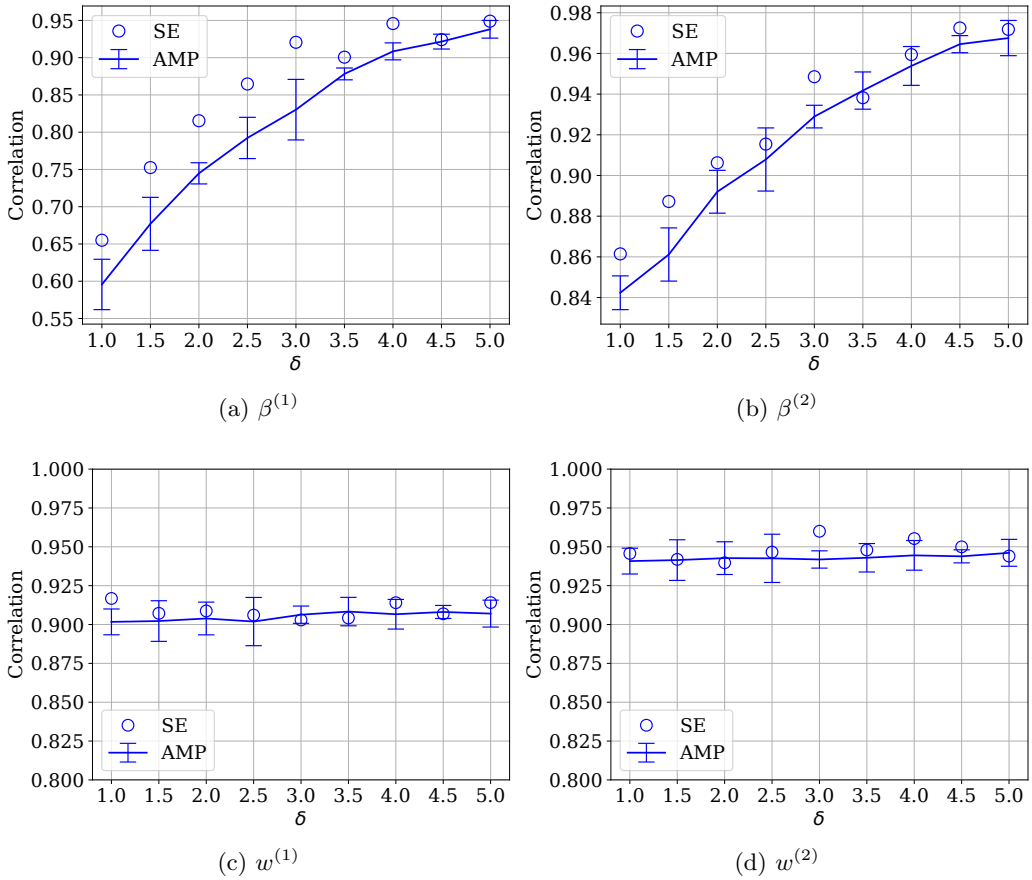
$$\check{f}_k(\check{B}^k, B) = (B,\ f_k(\check{B}^k + B(\mathrm{M}_B^k)^\top)). \tag{47}$$

Then, we claim that the original AMP iteration (11) is equivalent to the following one:

$$\begin{aligned} \check{\Theta}^k &= X\check{f}_k(\check{B}^k, B) \ - \ h_{k-1}(\check{\Theta}^{k-1}, \Psi)(\check{F}^k)^\top \\ \check{B}^{k+1} &= X^\top h_k(\check{\Theta}^k, \Psi) \ - \ \check{f}_k(\check{B}^k, B)(\check{C}^k)^\top, \end{aligned} \tag{48}$$

where h_k is defined in (13), 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:

$$\begin{aligned} \check{C}^k &= \left[\mathbb{E}[\partial_Z h_k(Z, Z^k, \bar{\Psi})],\ \ \tfrac{1}{n}\textstyle\sum_{i=1}^n \partial_{\Theta_i^k} h_k(\Theta_i, \Theta_i^k, \Psi_i)\right] \\ \check{F}^{k+1} &= \begin{bmatrix} 0_{L\times L} \\ \frac{1}{n}\sum_{j=1}^p f'_{k+1}(\check{B}_j^k + B_j(\mathrm{M}_B^k)^\top) \end{bmatrix}. \end{aligned} \tag{49}$$



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

(c) $w^{(1)}$ (d) $w^{(2)}$

Figure 16: Mixture-of-experts: noise level $\sigma = 0.1$.

The iteration (48) 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 (48) and the original AMP in (11) can be seen by substituting the definitions (46) and (47) into (48), and recalling from (14) that $\mathrm{M}_B^{k+1} = \mathbb{E}[\partial_Z h_k(Z, Z^k, \bar{\Psi})]$.

A key feature of the new iteration in (48) is that, 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 (11) where the input Y to the function g_k is not independent of X. The recursion in (48) 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 state evolution for the abstract AMP can also be obtained using the general AMP framework in Gerbelot and Berthier (2021). The standard form of the abstract AMP recursion uses empirical estimates (instead of expected values) for the first L columns of $\check{C}^k$ in (49). However, the state evolution result still remains valid for the recursion (48) (see Remark 4.3 of Feng et al. (2022)). This result, stated in Proposition 3 below, guarantees 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, where 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 $k \geq 0$:

$$\check{\mathrm{T}}^{k+1} = \mathbb{E}\big[h_k(G_\sigma^k, \bar{\Psi})h_k(G_\sigma^k, \bar{\Psi})^\top\big] \tag{50}$$

$$\check{\Sigma}^{k+1} = \delta^{-1}\mathbb{E}\big[\check{f}_{k+1}(G_\tau^{k+1}, \bar{B})\check{f}_{k+1}(G_\tau^{k+1}, \bar{B})^\top\big] = \begin{bmatrix} \delta^{-1}\mathbb{E}[\bar{B}\bar{B}^\top] & \check{\Sigma}_{(12)}^{k+1} \\ \big(\check{\Sigma}_{(12)}^{k+1}\big)^\top & \check{\Sigma}_{(22)}^{k+1} \end{bmatrix}, \tag{51}$$

where

$$\check{\Sigma}_{(12)}^{k+1} = \Big(\check{\Sigma}_{(21)}^{k+1}\Big)^\top = \delta^{-1}\mathbb{E}\big[\bar{B}f_{k+1}(G_\tau^{k+1} + \mathrm{M}_{k+1}^B\bar{B})^\top\big] \tag{52}$$

$$\check{\Sigma}_{(22)}^{k+1} = \delta^{-1}\mathbb{E}\big[f_{k+1}(G_\tau^{k+1} + \mathrm{M}_{k+1}^B\bar{B}) \cdot f_{k+1}(G_\tau^{k+1} + \mathrm{M}_{k+1}^B\bar{B})^\top\big]. \tag{53}$$

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 (15)-(17) and of $(\check{\mathrm{T}}^{k+1}, \check{\Sigma}^{k+1})$ in (50)-(51), 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* (48)*, for $k \geq 0$ we have:*

$$\sup_{\eta\in \mathrm{PL}_{2L}(r,1)} \Big|\frac{1}{p}\sum_{j=1}^p \eta(\check{B}_j^{k+1},\, B_j) - \mathbb{E}[\eta(G_\tau^{k+1},\, \bar{B})]\Big| \xrightarrow{c} 0, \tag{54}$$

$$\sup_{\eta\in \mathrm{PL}_{2L+L_\Psi}(r,1)} \Big|\frac{1}{n}\sum_{i=1}^n \eta(\check{\Theta}_i^k, \Psi_i) - \mathbb{E}[\eta(G_\sigma^k, \bar{\Psi})]\Big| \xrightarrow{c} 0, \tag{55}$$

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

To obtain the result (23), we recall the definition of $\check{B}^{k+1}$ from (46), and in (54) 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 (24), we recall the definition of $\check{\Theta}^k$ from (46), and in (55) take $\eta(\check{\Theta}^k, \Psi) = \phi(\Theta^k, \Theta, \Psi)$. Since $\check{\Sigma}^k = \Sigma^k$, we have:

$$(G_\sigma^k, \bar{\Psi}) \overset{d}{=} (Z, Z^k, \bar{\Psi}) \overset{d}{=} (Z,\, \mathrm{M}_\Theta^k Z + G_\Theta^k, \bar{\Psi}), \tag{56}$$

where the last equality follows from (18). 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*

$$\mathbb{E}[UU^\top] - \mathbb{E}[UV^\top]\big(\mathbb{E}[VV^\top]\big)^{-1}\mathbb{E}[VU^\top] \succeq 0. \tag{57}$$

Proof of part 1. Using the law of total expectation, $\Sigma^k_{(12)}$ in (17) can be written as:

$$\delta\Sigma^k_{(12)} = \mathbb{E}[\bar{B} f_k(\mathrm{M}^k_B \bar{B} + G^k_B)^\top] = \mathbb{E}\big[\mathbb{E}[\bar{B} f_k(\mathrm{M}^k_B \bar{B} + G^k_B)^\top \mid \mathrm{M}^k_B \bar{B} + G^k_B]\big] = \mathbb{E}[f^*_k f_k^\top], \quad (58)$$

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

$$\begin{aligned} &\mathbb{E}[f^*_k (f^*_k)^\top] - \mathbb{E}[f^*_k f_k^\top]\mathbb{E}[f_k f_k^\top]^{-1}\mathbb{E}[f_k (f^*_k)^\top] \succeq 0 \\ &\implies \delta^{-1}\mathbb{E}[f^*_k (f^*_k)^\top] - \Sigma^k_{(12)}(\Sigma^k_{(22)})^{-1}\Sigma^k_{(21)} \succeq 0, \end{aligned} \quad (59)$$

where we have used (58) and (17) for the second line. Adding and subtracting T^k_Θ in (59) we obtain

$$\mathrm{T}^k_\Theta - \underbrace{\big(\mathrm{T}^k_\Theta - \delta^{-1}\mathbb{E}[f^*_k (f^*_k)^\top] + \Sigma^k_{(12)}(\Sigma^k_{(22)})^{-1}\Sigma^k_{(21)}\big)}_{:=\Gamma^k_\Theta} \succeq 0. \quad (60)$$

Multiplying the matrix $(\mathrm{T}^k_\Theta - \Gamma^k_\Theta)$ in (60) by $\big(\mathrm{M}^k_\Theta\big)^{-1}$ on the left and $\Big(\big(\mathrm{M}^k_\Theta\big)^{-1}\Big)^\top$ on the right maintains positive definiteness. This yields

$$N^k_\Theta - \Big(\mathrm{M}^k_\Theta\Big)^{-1} \Gamma^k_\Theta \Big(\big(\mathrm{M}^k_\Theta\big)^{-1}\Big)^\top \succeq 0, \quad (61)$$

where we have used the formula for N^k_Θ from (28). Eq. (61) implies

$$\mathrm{Tr}(N^k_\Theta) \geq \mathrm{Tr}\left(\Big(\mathrm{M}^k_\Theta\Big)^{-1} \Gamma^k_\Theta \Big(\big(\mathrm{M}^k_\Theta\big)^{-1}\Big)^\top\right). \quad (62)$$

Now, using the formula for T^k_Θ in (20) it can be verified that when $f_k = f^*_k$, we have

$$\mathrm{T}^k_\Theta = \Gamma^k_\Theta = \frac{1}{\delta}\Big(\mathbb{E}\big[f^*_k (f^*_k)^\top\big] - \mathbb{E}\big[f^*_k (f^*_k)^\top\big]\big(\mathbb{E}\big[\bar{B}\bar{B}^\top\big]\big)^{-1}\mathbb{E}\big[f^*_k (f^*_k)^\top\big]\Big). \quad (63)$$

Therefore (60)-(62) are satisfied with equality when $f_k = f^*_k$, which proves the first part of the proposition.

Proof of part 2. We begin by introducing the multivariate Stein's 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*

$$\mathbb{E}[\nabla g(X)] = \Big(\Sigma^{-1}\mathbb{E}\big[(X - \mu) g(X)^\top\big]\Big)^\top. \quad (64)$$

Proof We have

$$\begin{aligned}\mathbb{E}[(X-\mu)g(X)^\top] &= \big(\mathbb{E}[(X-\mu)g_1(X)],\dots,\mathbb{E}[(X-\mu)g_L(X)]\big)\\ &\overset{(a)}{=} (\Sigma\mathbb{E}[\nabla g_1(X)],\dots,\Sigma\mathbb{E}[\nabla g_L(X)])\\ &= \Sigma\mathbb{E}[(\nabla g_1(X),\dots,\nabla g_L(X))]\\ &\overset{(b)}{=} \Sigma\mathbb{E}[\nabla g(X)]^\top,\end{aligned} \tag{65}$$

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 $l = 1,\dots,L$, and (b) uses the definition of $\nabla g(x)$. Finally, rearranging the above equation and taking the transpose gives the result. ■

Next, we use Lemma 5 to show that

$$\mathrm{M}_B^{k+1} = \mathbb{E}\big[g_k(Z^k,\bar{Y})g_k^*(Z^k,\bar{Y})^\top\big], \tag{66}$$

where g_k^* is defined in (30). Indeed, using the law of total expectation we have

$$\begin{aligned}\mathrm{M}_B^{k+1} &= \mathbb{E}\Big[\mathbb{E}[\partial_Z h_k(Z,Z^k,\bar{\Psi})|Z^k]\Big]\\ &\overset{(a)}{=} \mathbb{E}\Big[\mathrm{Cov}[Z|Z^k]^{-1}\mathbb{E}\big[(Z-\mathbb{E}[Z|Z^k])h_k(Z,Z^k,\bar{\Psi})^\top\big|\,Z^k\big]\Big]^\top\\ &= \mathbb{E}[\mathrm{Cov}[Z|Z^k]^{-1}(Z-\mathbb{E}[Z|Z^k])h_k(Z,Z^k,\bar{\Psi})^\top]^\top\\ &= \mathbb{E}\Big[\mathbb{E}\big[\mathrm{Cov}[Z|Z^k]^{-1}(Z-\mathbb{E}[Z|Z^k])h_k(Z,Z^k,\bar{\Psi})^\top\big|\,Z^k,\bar{Y}\big]\Big]^\top\\ &\overset{(b)}{=} \mathbb{E}\Big[g_k^*(Z^k,\bar{Y})h_k(Z,Z^k,\bar{\Psi})^\top\Big]^\top\\ &\overset{(c)}{=} \mathbb{E}\big[g_k(Z^k,\bar{Y})g_k^*(Z^k,\bar{Y})^\top\big].\end{aligned} \tag{67}$$

Here (a) applies Lemma 5, (b) follows from the definition of g_k^* in (30), and (c) from (13). 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:

$$\mathbb{E}[g_k^*(g_k^*)^\top] - \mathbb{E}[g_k^*g_k^\top]\Big(\mathbb{E}[g_kg_k^\top]\Big)^{-1}\mathbb{E}[g_k(g_k^*)^\top] \succeq 0 \Leftrightarrow \mathbb{E}[g_k^*(g_k^*)^\top] - \Big(N_B^{k+1}\Big)^{-1} \succeq 0 \tag{68}$$

$$\Leftrightarrow \Big(\mathbb{E}[g_k^*(g_k^*)^\top]\Big)^{-1} - N_B^{k+1} \preceq 0, \tag{69}$$

where (68) is obtained by recalling from (28) that $(N_B^{k+1})^{-1} = \mathrm{M}_B^{k+1}\left(\mathrm{T}_B^{k+1}\right)^{-1}\left(\mathrm{M}_B^{k+1}\right)^\top$, and using the expressions for M_B^{k+1} and T_B^{k+1} in (66) and (15). Eq. (69) 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 (69), we have that

$$\mathrm{Tr}(N_B^{k+1}) \le \mathrm{Tr}\left(\left(\mathbb{E}[g_k^*(g_k^*)^\top]\right)^{-1}\right), \tag{70}$$

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

5.3 Derivation of the EM Step for Max-Affine Regression

In this section, we derive steps 4 and 5 of EM-AMP for max-affine regression (Algorithm 1). We follow an approach similar to the one in Vila and Schniter (2013), where EM was combined with AMP for compressed sensing with unknown parameters in the signal prior. We adapt their derivation to the MAR model. Recall that

$$\begin{aligned} Y_i &= \max\left\{\left\langle X_i, \beta^{(1)}\right\rangle + b_1, \left\langle X_i, \beta^{(2)}\right\rangle + b_2\right\} + \epsilon_i \\ &= \max\left\{\Theta_i^{(1)} + b_1, \Theta_i^{(2)} + b_2\right\} + \epsilon_i, \qquad i \in [n]. \end{aligned} \tag{71}$$

Here $\Theta^{(1)}$ and $\Theta^{(2)}$ are the first and second columns of $\Theta \in \mathbb{R}^{n\times 2}$ respectively, and $\Theta_i := \left(\Theta_i^{(1)}, \Theta_i^{(2)}\right) \in \mathbb{R}^2$.

The parameter that we would like to estimate using EM is $b = (b_1, b_2)$. Before providing the derivation, we briefly review the main idea behind the EM algorithm. The EM algorithm iteratively produces estimates $b^m \equiv (b_1^m, b_2^m)$ for $m \geq 1$, with the goal of increasing the likelihood $p(Y; b)$ at each iteration, where $Y = [Y_1, \ldots, Y_n]^\top$. It achieves this by iteratively increasing a lower bound on $p(Y; b)$, thus guaranteeing that the likelihood converges to a local maximum or at least a saddle point (Wu, 1983). In our case, for an arbitrary distribution $\hat{p}$ on $(\Theta^{(1)}, \Theta^{(2)})$ we have

$$\begin{aligned} \log p(Y;b) &= \int_{\Theta^{(1)}}\int_{\Theta^{(2)}} \hat{p}(\Theta^{(1)},\Theta^{(2)}) d\Theta^{(1)} d\Theta^{(2)} \log p(Y;b) \\ &= \int_{\Theta^{(1)}}\int_{\Theta^{(2)}} \hat{p}(\Theta^{(1)},\Theta^{(2)}) d\Theta^{(1)} d\Theta^{(2)} \log\left(\frac{p(\Theta^{(1)},\Theta^{(2)},Y;b)}{\hat{p}(\Theta^{(1)},\Theta^{(2)})} \cdot \frac{\hat{p}(\Theta^{(1)},\Theta^{(2)})}{p(\Theta^{(1)},\Theta^{(2)}|Y;b)}\right) \\ &\overset{(a)}{=} \mathbb{E}_{\Theta^{(1)},\Theta^{(2)}\sim\hat{p}}\big[\log p(\Theta^{(1)},\Theta^{(2)},Y;b)\big] + H(\hat{p}) + D(\hat{p}\,\|\,p(\cdot|Y;b)) \\ &\overset{(b)}{\geq} \mathbb{E}_{\Theta^{(1)},\Theta^{(2)}\sim\hat{p}}\big[\log p(\Theta^{(1)},\Theta^{(2)},Y;b)\big] + H(\hat{p}) := \mathcal{L}(Y;b), \end{aligned} \tag{72}$$

where in (a), $H(\cdot)$ is the Shannon entropy and $D(\cdot\|\cdot)$ the Kullback-Leibler (KL) divergence, and the inequality (b) follows from the non-negativity of the KL divergence. The EM-algorithm at step $m+1$ iterates over two steps:

- In the E-step, we choose $\hat{p}$ to maximize the lower bound $\mathcal{L}(Y;b)$ for fixed $b = b^m$,
- In the M-step, we choose b to maximize the lower bound $\mathcal{L}(Y;b)$ for fixed $\hat{p} = \hat{p}^m$.

For the E-step, since $\mathcal{L}(Y;b^m) = \log p(Y;b^m) - D(\hat{p}\,\|\,p(\cdot|Y;b^m))$ (via rearranging (72)), the maximizing probability density function (pdf) would be $\hat{p}^m = p(\Theta^{(1)},\Theta^{(2)}|Y;b^m)$. Then, for the M-step, from the definition of $\mathcal{L}(Y;b)$, the maximizing b is:

$$b^{m+1} = \operatorname*{argmax}_{b\in\mathbb{R}^2} \mathbb{E}_{\Theta^{(1)},\Theta^{(2)}\sim p(\Theta^{(1)},\Theta^{(2)}|Y;b^m)}\big[\log p\big(\Theta^{(1)},\Theta^{(2)},Y;b\big)\big]. \tag{73}$$

We can further expand the $p(\cdot)$ above as

$$p\big(\Theta^{(1)},\Theta^{(2)},Y;b\big) = p\big(\Theta^{(1)},\Theta^{(2)}\big)p\big(Y|\Theta^{(1)},\Theta^{(2)};b\big) = p\big(\Theta^{(1)},\Theta^{(2)}\big)\prod_{i=1}^n p(Y_i|\Theta_i;b), \tag{74}$$

where $p\big(\Theta^{(1)}, \Theta^{(2)}\big) = \prod_{i=1}^n p\big(\Theta_i^{(1)}, \Theta_i^{(2)}\big)$ does not depend on b (recall that $\big(\Theta^{(1)}, \Theta^{(2)}\big) = XB$). It is challenging to jointly optimize over $b = (b_1, b_2)$ in (73), so we update one component of b at a time while holding the other fixed. This is a well known "incremental" variant of EM (Neal and Hinton, 1998). Using (74) and (73), the updated b_1 estimate is

$$\begin{aligned} b_1^{m+1} &= \operatorname*{argmax}_{b_1 \in \mathbb{R}} \mathbb{E}\Big[\log \prod_{i=1}^n p(Y_i|\Theta_i; b_1, b_2^m) \mid Y; b^m \Big] \\ &= \operatorname*{argmax}_{b_1 \in \mathbb{R}} \sum_{i=1}^n \mathbb{E}\Big[\log p(Y_i|\Theta_i; b_1, b_2^m) \mid Y; b^m \Big] \\ &= \operatorname*{argmax}_{b_1 \in \mathbb{R}} \sum_{i=1}^n \int_{\Theta_i} p(\Theta_i|Y; b^m) \log p(Y_i|\Theta_i; b_1, b_2^m) d\Theta_i \,, \end{aligned} \tag{75}$$

where we recall that $b^m = (b_1^m, b_2^m)$. At this point, note that

$$p(Y_i|\Theta_i; b_1, b_2^m) \sim \mathcal{N}\Big(\max\big\{\Theta_i^{(1)} + b_1, \Theta_i^{(2)} + b_2^m\big\}, \sigma^2 \Big). \tag{76}$$

To get b_1^{m+1}, we need to solve

$$\begin{aligned} &\frac{\partial}{\partial b_1} \sum_{i=1}^n \int_{\Theta_i} p(\Theta_i|Y; b^m) \log p(Y_i|\Theta_i; b_1, b_2^m) d\Theta_i = 0 \\ \iff &\sum_{i=1}^n \int_{\Theta_i} p(\Theta_i|Y; b^m) \frac{\partial}{\partial b_1} \log p(Y_i|\Theta_i; b_1, b_2^m) d\Theta_i = 0, \end{aligned} \tag{77}$$

where we used Leibniz's integral rule to exchange differentiation and integration. Using (76), the derivative in (77) is

$$\begin{aligned} \frac{\partial}{\partial b_1} \log p(Y_i|\Theta_i; b_1, b_2^m) &= \frac{\partial}{\partial b_1} \log \left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{Y_i - \max\{\Theta_i^{(1)} + b_1, \Theta_i^{(2)} + b_2^m\}}{\sigma} \right)^2 \right) \right) \\ &= \begin{cases} \frac{1}{\sigma^2}(Y_i - \Theta_i^{(1)} - b_1) & \text{if } \Theta_i^{(1)} + b_1 > \Theta_i^{(2)} + b_2^m \\ 0 & \text{if } \Theta_i^{(1)} + b_1 \le \Theta_i^{(2)} + b_2^m \end{cases}. \end{aligned} \tag{78}$$

Substituting the above back into (77), we get

$$\begin{aligned} &\sum_{i: \Theta_i^{(1)} + b_1 > \Theta_i^{(2)} + b_2^m} \int_{\Theta_i} p_{Z|Y}(\Theta_i|Y; b^m) \frac{1}{\sigma^2}(Y_i - \Theta_i^{(1)} - b_1) d\Theta_i = 0 \\ \iff &\sum_{i: \Theta_i^{(1)} + b_1 > \Theta_i^{(2)} + b_2^m} \Bigg(Y_i \int_{\Theta_i} p_{Z|Y}(\Theta_i|Y; b^m) d\Theta_i - \int_{\Theta_i} p_{Z|Y}(\Theta_i|Y; b^m) \Theta_i^{(1)} d\Theta_i \\ &\qquad - b_1 \int_{\Theta_i} p_{Z|Y}(\Theta_i|Y; b^m) d\Theta_i \Bigg) = 0 \\ \iff &\sum_{i: \Theta_i^{(1)} + b_1 > \Theta_i^{(2)} + b_2^m} \Big(Y_i - \mathbb{E}[\Theta_i^{(1)}|Y; b^m] - b_1 \Big) = 0. \end{aligned} \tag{79}$$

Rearranging gives

$$b_1 = \frac{1}{|\{i : \Theta_i^{(1)} + b_1 > \Theta_i^{(2)} + b_2^m\}|} \sum_{i:\Theta_i^{(1)}+b_1>\Theta_i^{(2)}+b_2^m} \Big(Y_i - \mathbb{E}\big[\Theta_i^{(1)}|Y; b^m\big]\Big)$$
$$\approx \frac{1}{|\{i : \Theta_i^{(1)} + b_1 > \Theta_i^{(2)} + b_2^m\}|} \sum_{i:\Theta_i^{(1)}+b_1>\Theta_i^{(2)}+b_2^m} Y_i - \mathbb{E}\big[Z_1|\bar{Y}; b^m\big], \tag{80}$$

where we approximate $\mathbb{E}\big[\Theta_i^{(1)}|Y; b^m\big]$ by $\mathbb{E}\big[Z_1|\bar{Y}; b^m\big]$ because computing $\mathbb{E}\big[\Theta_i^{(1)}|Y; b^m\big]$ is intractable. The computation of $\mathbb{E}\big[Z_1|\bar{Y}; b^m\big]$ is detailed in Appendix B.

Note that in (80) it is not possible to compute b_1 on the LHS while using it in the RHS. An easy (but admittedly non-principled) fix is to just use b_1^m on the RHS. This gives the update:

$$b_1^{m+1} := \frac{1}{|\{i : \Theta_i^{(1)} + b_1^m > \Theta_i^{(2)} + b_2^m\}|} \sum_{i:\Theta_i^{(1)}+b_1^m>\Theta_i^{(2)}+b_2^m} Y_i - \mathbb{E}\big[Z_1|\bar{Y}; b^m\big], \tag{81}$$

where Θ_i and $\mathbb{E}\big[Z_1|Y; b^m\big]$ can be obtained from AMP in the previous iteration. Similarly, the update for the other intercept is:

$$b_2^{m+1} := \frac{1}{|\{i : \Theta_i^{(1)} + b_1^m \leq \Theta_i^{(2)} + b_2^m\}|} \sum_{i:\Theta_i^{(1)}+b_1^m\leq\Theta_i^{(2)}+b_2^m} Y_i - \mathbb{E}\big[Z_2|\bar{Y}; b^m\big]. \tag{82}$$

Since $\Theta^{(1)}, \Theta^{(2)}$ are unknown, we approximate them using AMP iterates. Specifically, the two columns of $\widehat{\Theta}^m = X\widehat{B}^{k_{\max},m}$ provide estimates of $\Theta^{(1)}, \Theta^{(2)}$, respectively. Using these in (81) and (82) yields Steps 4 and 5 of the EM-AMP in Algorithm 1.

Acknowledgments

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

Appendix A. Implementation Details for MLR

In this appendix, we consider MLR with two signals 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 MLR with 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 writing the matrix-AMP algorithm in (11)-(12) with more details:

- Initialize $\widehat{R}^{-1} = 0 \in \mathbb{R}^{n\times 2}$, $F_0 = I_2$. Next, we initialize the rows of B^0 and $\widehat{B}^0$ independently using the jointly Gaussian prior. Letting $\bar{B} = (\bar{\beta}^{(1)}, \bar{\beta}^{(2)}) \in \mathbb{R}^2$ be a random variable distributed according to the jointly Gaussian prior, we initialize:

$$B_j^0, \widehat{B}_j^0 \sim_{\text{i.i.d.}} \mathcal{N}\left(\begin{bmatrix}\mathbb{E}[\bar{\beta}^{(1)}]\\ \mathbb{E}[\bar{\beta}^{(2)}]\end{bmatrix}, \begin{bmatrix}\text{Var}[\bar{\beta}^{(1)}] & \text{Cov}[\bar{\beta}^{(1)}, \bar{\beta}^{(2)}]\\ \text{Cov}[\bar{\beta}^{(2)}, \bar{\beta}^{(1)}] & \text{Var}[\bar{\beta}^{(2)}]\end{bmatrix}\right), \quad j \in [p], \tag{83}$$

  and

$$\Sigma^0 = \frac{p}{n}\begin{bmatrix}\mathbb{E}[(\bar{\beta}^{(1)})^2] & \mathbb{E}[\bar{\beta}^{(1)}\bar{\beta}^{(2)}] & (\mathbb{E}[\bar{\beta}^{(1)}])^2 & \mathbb{E}[\bar{\beta}^{(1)}]\mathbb{E}[\bar{\beta}^{(2)}]\\ \mathbb{E}[\bar{\beta}^{(1)}\bar{\beta}^{(2)}] & \mathbb{E}[(\bar{\beta}^{(2)})^2] & \mathbb{E}[\bar{\beta}^{(1)}]\mathbb{E}[\bar{\beta}^{(2)}] & (\mathbb{E}[\bar{\beta}^{(2)}])^2\\ (\mathbb{E}[\bar{\beta}^{(1)}])^2 & \mathbb{E}[\bar{\beta}^{(1)}]\mathbb{E}[\bar{\beta}^{(2)}] & \mathbb{E}[(\bar{\beta}^{(1)})^2] & \mathbb{E}[\bar{\beta}^{(1)}\bar{\beta}^{(2)}]\\ \mathbb{E}[\bar{\beta}^{(1)}]\mathbb{E}[\bar{\beta}^{(2)}] & (\mathbb{E}[\bar{\beta}^{(2)}])^2 & \mathbb{E}[\bar{\beta}^{(1)}\bar{\beta}^{(2)}] & \mathbb{E}[(\bar{\beta}^{(2)})^2]\end{bmatrix}. \tag{84}$$

- 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. Compute $\widehat{R}^k := g_k(\Theta^k, Y)$
  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}

The quantities in steps 1, 2, and 4 can be directly computed. The other steps require some form of numerical approximation (based on limiting properties of the iterates) to make the computation tractable. We now explain how the quantities in steps 2, 3, 5, 6, and 7 can be computed or approximated.

Step 2: We assume that Σ^k has been approximated in the previous iteration. From Proposition 2, for MLR the function $g_k : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$ is given by

$$g_k(Z^k, \bar{Y}) = \text{Cov}[Z|Z^k]^{-1}(\mathbb{E}[Z|Z^k, \bar{Y}] - \mathbb{E}[Z|Z^k]). \tag{85}$$

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

$$\mathrm{Cov}[Z|Z^k] = \Sigma^k_{(11)} - \Sigma^k_{(12)}(\Sigma^k_{(22)})^{-1}\Sigma^k_{(21)}, \qquad \mathbb{E}[Z|Z^k] = \Sigma^k_{(12)}(\Sigma^k_{(22)})^{-1}Z^k. \tag{86}$$

To compute $\mathbb{E}[Z|Z^k, \bar{Y}]$, we recall that $Z = (Z_1, Z_2)^\top$ and let $\bar{\Psi} = (\bar{c}, \bar{\epsilon})$, with $\bar{c} \sim$ Bernoulli(α) and $\bar{\epsilon} \sim \mathcal{N}(0, \sigma^2)$ independent. Then,

$$Y = q(Z, \bar{\Psi}) = Z_1\bar{c} + Z_2(1 - \bar{c}) + \bar{\epsilon}, \tag{87}$$

using which we have that

$$\mathbb{E}[Z|Z^k, \bar{Y}] = \mathbb{E}[Z|Z^k, \bar{Y}, \bar{c} = 1]\mathbb{P}[\bar{c} = 1|Z^k, \bar{Y}] \; + \; \mathbb{E}[Z|Z^k, \bar{Y}, \bar{c} = 0]\mathbb{P}[\bar{c} = 0|Z^k, \bar{Y}]. \tag{88}$$

We now show how each of the four terms in (88) can be computed.

We first find the joint distribution of $(Z, Z^k, \bar{Y}|\bar{c} = 1)$, which from (87) is jointly Gaussian. We denote this distribution by $\mathcal{N}(\mathbf{0}, \Sigma^{k,1}_Y)$, and proceed to derive $\Sigma^{k,1}_Y \in \mathbb{R}^{5\times 5}$. Given a matrix $M \in \mathbb{R}^{n_1 \times n_2}$, will use the notation $M_{[a],[b]}$ to denote the submatrix consisting of the first a rows and first b columns of M, and $M_{[a^+],[b^+]}$ to denote the submatrix with rows $\{a, \ldots, n_1\}$ and columns $\{b, \ldots, n_2\}$.

We know from the joint distribution of (Z, Z^k) that $(\Sigma^{k,1}_Y)_{[4],[4]} = \Sigma^k$. Hence, we only need to determine the remaining entries:

$$\begin{aligned}
(\Sigma^{k,1}_Y)_{5,5} &= \mathrm{Var}[\bar{Y} \mid \bar{c} = 1] = \mathrm{Var}[Z_1 + \bar{\epsilon}] = \Sigma^k_{11} + \sigma^2, \\
(\Sigma^{k,1}_Y)_{1,5} &= (\Sigma^{k,1}_Y)_{5,1} = \mathrm{Cov}[\bar{Y}, Z_1|\bar{c} = 1] = \mathrm{Cov}[Z_1 + \bar{\epsilon}, Z_1] = \Sigma^k_{11}, \\
(\Sigma^{k,1}_Y)_{2,5} &= (\Sigma^{k,1}_Y)_{5,2} = \mathrm{Cov}[\bar{Y}, Z_2|\bar{c} = 1] = \mathrm{Cov}[Z_1 + \bar{\epsilon}, Z_2] = \Sigma^k_{12}, \\
(\Sigma^{k,1}_Y)_{1,3} &= (\Sigma^{k,1}_Y)_{3,1} = \mathrm{Cov}[\bar{Y}, Z^k_1|\bar{c} = 1] = \mathrm{Cov}[Z_1 + \bar{\epsilon}, Z^k_1] = \Sigma^k_{13}, \\
(\Sigma^{k,1}_Y)_{1,4} &= (\Sigma^{k,1}_Y)_{4,1} = \mathrm{Cov}[\bar{Y}, Z^k_2|\bar{c} = 1] = \mathrm{Cov}[Z_1 + \bar{\epsilon}, Z^k_2] = \Sigma^k_{14},
\end{aligned} \tag{89}$$

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

$$\Sigma^{k,1}_Y = \begin{bmatrix}
\Sigma^k_{11} & \Sigma^k_{12} & \Sigma^k_{13} & \Sigma^k_{14} & \Sigma^k_{11} \\
\Sigma^k_{21} & \Sigma^k_{22} & \Sigma^k_{23} & \Sigma^k_{24} & \Sigma^k_{21} \\
\Sigma^k_{31} & \Sigma^k_{32} & \Sigma^k_{33} & \Sigma^k_{34} & \Sigma^k_{31} \\
\Sigma^k_{41} & \Sigma^k_{42} & \Sigma^k_{43} & \Sigma^k_{44} & \Sigma^k_{41} \\
\Sigma^k_{11} & \Sigma^k_{12} & \Sigma^k_{13} & \Sigma^k_{14} & \Sigma^k_{11} + \sigma^2
\end{bmatrix}. \tag{90}$$

From the joint distribution, we can compute

$$\mathbb{E}[Z|Z^k, \bar{Y}, \bar{c} = 1] = (\Sigma^{k,1}_Y)_{[2],[3^+]}(\Sigma^{k,1}_Y)^{-1}_{[3^+],[3^+]} \begin{bmatrix} Z^k \\ \bar{Y} \end{bmatrix}, \tag{91}$$

where $[3^+] := \{3, 4, 5\}$. Using the same approach, we can determine the joint distribution of $(Z, Z^k, \bar{Y}|\bar{c} = 0)$ as $\mathcal{N}(\mathbf{0}, \Sigma^{k,0}_Y)$, where

$$\Sigma^{k,0}_Y = \begin{bmatrix}
\Sigma^k_{11} & \Sigma^k_{12} & \Sigma^k_{13} & \Sigma^k_{14} & \Sigma^k_{12} \\
\Sigma^k_{21} & \Sigma^k_{22} & \Sigma^k_{23} & \Sigma^k_{24} & \Sigma^k_{22} \\
\Sigma^k_{31} & \Sigma^k_{32} & \Sigma^k_{33} & \Sigma^k_{34} & \Sigma^k_{32} \\
\Sigma^k_{41} & \Sigma^k_{42} & \Sigma^k_{43} & \Sigma^k_{44} & \Sigma^k_{42} \\
\Sigma^k_{21} & \Sigma^k_{22} & \Sigma^k_{23} & \Sigma^k_{24} & \Sigma^k_{22} + \sigma^2
\end{bmatrix}. \tag{92}$$

From this joint distribution, we can compute

$$\mathbb{E}[Z|Z^k, \bar{Y}, \bar{c}=0] = (\Sigma_Y^{k,0})_{[2],[3^+]}(\Sigma_Y^{k,0})^{-1}_{[3^+],[3^+]}\begin{bmatrix} Z^k \\ \bar{Y}\end{bmatrix}. \tag{93}$$

The first conditional probability term in (88) can be computed as:

$$\begin{aligned}\mathbb{P}[\bar{c}=1|Z^k,\bar{Y}] &= \frac{\mathbb{P}[\bar{c}=1]\mathbb{P}[Z^k,\bar{Y}|\bar{c}=1]}{\mathbb{P}[\bar{c}=1]\mathbb{P}[Z^k,\bar{Y}|\bar{c}=1]+\mathbb{P}[\bar{c}=0]\mathbb{P}[Z^k,\bar{Y}|\bar{c}=0]} \\ &= \frac{\alpha\mathbb{P}[Z^k,\bar{Y}|\bar{c}=1]}{\alpha\mathbb{P}[Z^k,\bar{Y}|\bar{c}=1]+(1-\alpha)\mathbb{P}[Z^k,\bar{Y}|\bar{c}=0]},\end{aligned} \tag{94}$$

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

$$\mathbb{P}[\bar{c}=0|Z^k,\bar{Y}] = \frac{(1-\alpha)\mathbb{P}[Z^k,\bar{Y}|\bar{c}=0]}{\alpha\mathbb{P}[Z^k,\bar{Y}|\bar{c}=1]+(1-\alpha)\mathbb{P}[Z^k,\bar{Y}|\bar{c}=0]}, \tag{95}$$

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

Using (91)-(95), we can compute (88), which together with the quantities in (86) allows us to compute $g_k(Z^k,\bar{Y})$ in (85). Finally, compute $\widehat{R}^k$ by applying g_k row wise to Θ^k and Y (i.e., compute $g_k(\Theta_i^k, Y_i)$).

Step 3: Recalling that $g_k(Z^k,\bar{Y}) = h_k(Z,Z^k,\bar{\Psi})$, we approximate $C_k = \frac{1}{n}\sum_{i=1}^b g_k'(\Theta_i^k, Y_i)$ by calculating $\mathbb{E}[g_k'(Z^k,\bar{Y})] = \mathbb{E}[\nabla_{Z^k} h_k(Z,Z^k,\bar{\Psi})]$. Here $\nabla_{Z^k}h_k$ denotes the Jacobian with respect to Z^k. We compute the latter expectation by applying the generalized Stein's lemma (see Lemma 5) to $(Z,Z^k)^\top \sim \mathcal{N}(0,\Sigma^k)$ and $h_k(Z,Z^k,\bar{\Psi})$. This gives:

$$\mathbb{E}\left[\begin{bmatrix} Z \\ Z^k\end{bmatrix} h(Z,Z^k,\bar{\Psi})^\top\right] = \Sigma^k \mathbb{E}\Big[\nabla_{(Z,Z^k)} h_k(Z,Z^k,\bar{\Psi})\Big]^\top \in \mathbb{R}^{4\times 2}. \tag{96}$$

Writing the above explicitly, we have

$$\begin{aligned}\begin{bmatrix}\mathbb{E}[Zh_k(Z,Z^k,\bar{\Psi})^\top] \\ \mathbb{E}[Z^k h_k(Z,Z^k,\bar{\Psi})^\top]\end{bmatrix} &= \begin{bmatrix}\Sigma^k_{(11)} & \Sigma^k_{(12)} \\ \Sigma^k_{(21)} & \Sigma^k_{(22)}\end{bmatrix}\begin{bmatrix}\mathbb{E}[\nabla_Z h_k(Z,Z^k,\bar{\Psi})]^\top \\ \mathbb{E}[\nabla_{Z^k} h_k(Z,Z^k,\bar{\Psi})]^\top\end{bmatrix} \\ &= \begin{bmatrix}\Sigma^k_{(11)}\mathbb{E}[\nabla_Z h_k(Z,Z^k,\bar{\Psi})]^\top + \Sigma^k_{(12)}\mathbb{E}[\nabla_{Z^k} h_k(Z,Z^k,\bar{\Psi})]^\top \\ \Sigma^k_{(21)}\mathbb{E}[\nabla_Z h_k(Z,Z^k,\bar{\Psi})]^\top + \Sigma^k_{(22)}\mathbb{E}[\nabla_{Z^k} h_k(Z,Z^k,\bar{\Psi})]^\top\end{bmatrix},\end{aligned} \tag{97}$$

where the matrices $\Sigma^k_{(11)}, \Sigma^k_{(12}, \Sigma^k_{(21)}, \Sigma^k_{(22)} \in \mathbb{R}^{2\times 2}$ are as defined in (17). Looking at just the second row above and rearranging, we obtain:

$$\mathbb{E}[\nabla_{Z^k} h_k(Z,Z^k,\bar{\Psi})] = \Big\{(\Sigma^k_{(22)})^{-1}\Big(\mathbb{E}[Z^k h_k(Z,Z^k,\bar{\Psi})^\top] - \Sigma^k_{(21)}\mathbb{E}[\nabla_Z h_k(Z,Z^k,\bar{\Psi})]^\top\Big)\Big\}^\top. \tag{98}$$

Here $\mathbb{E}[Z^k h_k(Z,Z^k,\bar{\Psi})^\top] = \mathbb{E}[Z^k g_k(Z^k,\bar{Y})]$ 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 $\frac{1}{n} g_k(\Theta^k,Y)^\top g_k(\Theta^k,Y)$ (this follows from the equivalent expressions for M_B^{k+1} in (14) and (66), noting that we have used $g_k = g_k^*$).

Step 5: Since $\bar{B}$ is independent of $G_B^{k+1} \sim \mathcal{N}(0, \mathrm{T}_B^{k+1})$, we have that

$$\begin{bmatrix} \bar{B} \\ \mathrm{M}_B^{k+1}\bar{B} + G_B^{k+1} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbb{E}[\bar{B}] \\ \mathrm{M}_B^{k+1}\mathbb{E}[\bar{B}] \end{bmatrix}, \begin{bmatrix} \mathrm{Cov}[\bar{B}] & \mathrm{Cov}[\bar{B}](\mathrm{M}_B^{k+1})^\top \\ \mathrm{M}_B^{k+1}\mathrm{Cov}[\bar{B}] & \mathrm{M}_B^{k+1}\mathrm{Cov}[\bar{B}](\mathrm{M}_B^{k+1})^\top + \mathrm{T}_B^{k+1} \end{bmatrix} \right) \tag{99}$$

This implies that

$$\begin{aligned} &f_{k+1}(\mathrm{M}_B^{k+1}\bar{B} + G_B^{k+1} =: s) = \mathbb{E}[\bar{B}|s] \\ &= \mathbb{E}[\bar{B}] + \mathrm{Cov}[\bar{B}](\mathrm{M}_B^{k+1})^\top \Big(\mathrm{M}_B^{k+1}\mathrm{Cov}[\bar{B}](\mathrm{M}_B^{k+1})^\top + \mathrm{T}_B^{k+1} \Big)^{-1} \Big(s - \mathrm{M}_B^{k+1}\mathbb{E}[\bar{B}] \Big). \end{aligned} \tag{100}$$

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 choices, by (15) and (66)). Using (66), this can be calculated using

$$\mathrm{T}_B^{k+1} = \mathrm{M}_B^{k+1} \approx \frac{1}{n} g_k(\Theta^k, Y)^\top g_k(\Theta^k, Y). \tag{101}$$

Step 6: The expression for this can be obtained by taking the derivative of (100) w.r.t. s, which gives

$$f'_{k+1}(s) = \Big(\mathrm{M}_B^{k+1}\mathrm{Cov}[\bar{B}](\mathrm{M}_B^{k+1})^\top + \mathrm{T}_B^{k+1} \Big)^{-1} \mathrm{M}_B^{k+1}\mathrm{Cov}[\bar{B}], \tag{102}$$

where M_B^{k+1} and T_B^{k+1} can be approximated using (101).

Step 7: Using the formulas in (16)-(17) for Σ^{k+1} and noting that f_{k+1} is a conditional expectation, the covariance Σ^{k+1} can be approximated as

$$\Sigma^{k+1} \approx \frac{p}{n} \begin{bmatrix} \Sigma_{(11)}^k & \frac{1}{p} f_{k+1}(B^{k+1})^\top f_{k+1}(B^{k+1}) \\ \frac{1}{p} f_{k+1}(B^{k+1})^\top f_{k+1}(B^{k+1}) & \frac{1}{p} f_{k+1}(B^{k+1})^\top f_{k+1}(B^{k+1}) \end{bmatrix}. \tag{103}$$

A.2 Sparse Discrete Prior

As described in Appendix A.1, there are seven main steps in the AMP algorithm. A change in the prior requires us to make changes to our denoiser f_k which affects steps 5, 6, and 7; the other steps remain unchanged. The changes are as follows, for the Bayes-optimal and soft-thresholding choices for the denoiser f_k.

Bayes-optimal f_k:

Step 5: We have

$$f_{k+1}(\mathrm{M}_B^{k+1}\bar{B} + G_B^{k+1} =: s) = \mathbb{E}[\bar{B}|s] = \frac{\sum_{\bar{b}} \bar{b}\, \mathbb{P}[\bar{B} = \bar{b}] \mathbb{P}[s|\bar{B} = \bar{b}]}{\sum_{\bar{b}} \mathbb{P}[\bar{B} = \bar{b}] \mathbb{P}[s|\bar{B} = \bar{b}]}, \tag{104}$$

where $(s|\bar{B} = \bar{b}) \sim \mathcal{N}(\mathrm{M}_B^{k+1}\bar{b}, \mathrm{T}_B^{k+1})$, i.e., $\mathbb{P}[s|\bar{B} = \bar{b}]$ is the bivariate Gaussian pdf with mean vector $\mathrm{M}_B^{k+1}\bar{b}$ and covariance matrix T_B^{k+1}.

Step 6: In the following, for brevity we write $f \equiv f_{k+1}$, with f_1, f_2 denoting its two components. By the definition of a Jacobian, we have

$$\nabla_s f(\mathrm{M}_B^{k+1}\bar{B} + G_B^{k+1} = s) = \begin{bmatrix} \frac{\partial f_1}{\partial s_1} & \frac{\partial f_1}{\partial s_2} \\ \frac{\partial f_2}{\partial s_1} & \frac{\partial f_2}{\partial s_2} \end{bmatrix} = \begin{bmatrix} (\nabla_s f_1)^\top \\ (\nabla_s f_2)^\top \end{bmatrix} \tag{105}$$

To compute $\nabla_s f_1(s)$, letting $\bar{b} = [\bar{b}^{(1)}, \bar{b}^{(2)}]^\top$, we write

$$f_1(s) = \frac{\sum_{\bar{b}} \bar{b}^{(1)}\, \mathbb{P}[\bar{B} = \bar{b}]\mathbb{P}[s|\bar{B} = \bar{b}]}{\sum_{\bar{b}} \mathbb{P}[\bar{B} = \bar{b}]\mathbb{P}[s|\bar{B} = \bar{b}]} =: \frac{\mathrm{num}_1}{\mathrm{denom}_1}. \tag{106}$$

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

$$\nabla_s f_1(s) = \frac{(\nabla_s \mathrm{num}_1)(\mathrm{denom}_1) - (\mathrm{num}_1)(\nabla_s \mathrm{denom}_1)}{\mathrm{denom}_1^2} \tag{107}$$

Since $\mathbb{P}[s|\bar{B} = \bar{b}]$ is the bivariate Gaussian pdf with mean $\mathrm{M}_B^{k+1}\bar{b}$ and covariance matrix T_B^{k+1}, we have that

$$\begin{aligned} \nabla_s \mathbb{P}[s|\bar{B} = \bar{b}] &= \nabla_s \left(\frac{\exp\{-\frac{1}{2}(s - \mathrm{M}_B^{k+1}\bar{b})^\top (\mathrm{T}_B^{k+1})^{-1}(s - \mathrm{M}_B^{k+1}\bar{b})\}}{\sqrt{\det(2\pi \mathrm{T}_B^{k+1})}} \right) \\ &= (\mathrm{T}_B^{k+1})^{-1}\big(\mathrm{M}_B^{k+1}\bar{b} - s\big)\mathbb{P}[s|\bar{B} = \bar{b}] \end{aligned} \tag{108}$$

Using the above equation, we get

$$\begin{aligned} \nabla_s \mathrm{num}_1 &= \sum_{\bar{b}} \bar{b}^{(1)} (\mathrm{T}_B^{k+1})^{-1}\big(\mathrm{M}_B^{k+1}\bar{b} - s\big)\mathbb{P}[\bar{B} = \bar{b}]\mathbb{P}[s|\bar{B} = \bar{b}], \\ \nabla_s \mathrm{denom}_1 &= \sum_{\bar{b}} (\mathrm{T}_B^{k+1})^{-1}\big(\mathrm{M}_B^{k+1}\bar{b} - s\big)\mathbb{P}[\bar{B} = \bar{b}]\mathbb{P}[s|\bar{B} = \bar{b}], \end{aligned} \tag{109}$$

using which $\nabla_s f_1(s)$ can be computed using (106). The Jacobian $\nabla_s f_2(s)$ can be similarly computed.

Soft-Thresholding f_k:

Step 5: We can directly compute the function in (36).

Step 6: We can directly compute the Jacobian as shown in (37).

Step 7: We observe that the unlike the Bayes-optimal case, we no longer have the equality $\mathbb{E}[\bar{B} f_{k+1}(\mathrm{M}_B^{k+1}\bar{B} + G_B^{k+1})^\top] = \mathbb{E}[f_{k+1}(\mathrm{M}_B^{k+1}\bar{B} + G_B^{k+1}) f_{k+1}(\mathrm{M}_B^{k+1}\bar{B} + G_B^{k+1})^\top]$. Hence, we need to compute $\mathbb{E}[\bar{B} f_{k+1}(\mathrm{M}_B^{k+1}\bar{B} + G_B^{k+1})^\top]$ separately. To do so, we evaluate each entry of $\mathbb{E}[\bar{B} f_{k+1}(\mathrm{M}_B^{k+1}\bar{B} + G_B^{k+1})^\top]$ separately. We start by substituting the definitions of $\bar{B}$ and f_{k+1}, with $\bar{B} = (\bar{\beta}^{(1)}, \bar{\beta}^{(2)})$. This gives:

$$\begin{aligned} &\{\bar{B} f_{k+1}^\top\}_{11} = \bar{\beta}^{(1)} \mathrm{ST}\Big(\{\bar{B} + (\mathrm{M}_B^{k+1})^{-1} G_B^{k+1}\}_1; \alpha\sqrt{\{N_B^{k+1}\}_{11}}\Big) \\ &= \bar{\beta}^{(1)} \mathrm{ST}\Big(\bar{\beta}^{(1)} + \{(\mathrm{M}_B^{k+1})^{-1}\}_{11}\{G_B^{k+1}\}_1 + \{(\mathrm{M}_B^{k+1})^{-1}\}_{12}\{G_B^{k+1}\}_2; \alpha\sqrt{\{N_B^{k+1}\}_{11}}\Big). \end{aligned} \tag{110}$$

Expanding the function out and taking expectations over $\bar{\beta}^{(1)}$ and G_B^{k+1}, we get

$$\mathbb{E}[\{\bar{B} f_{k+1}^\top\}_{11}] = \begin{cases} \varepsilon & \text{if } |\{\bar{B} + (\mathrm{M}_B^{k+1})^{-1} G_B^{k+1}\}_1| > \alpha \sqrt{\{(\mathrm{M}_B^{k+1})^{-1} \mathrm{T}_B^{k+1} (\mathrm{M}_B^{k+1})^{-1}\}_{11}} \\ 0 & \text{otherwise} \end{cases}. \tag{111}$$

Following a similar set of steps for the other entries, we have

$$\mathbb{E}[\{\bar{B} f_{k+1}^\top\}_{22}] = \begin{cases} \varepsilon & \text{if } |\{\bar{B} + (\mathrm{M}_B^{k+1})^{-1} G_B^{k+1}\}_2| > \alpha \sqrt{\{(\mathrm{M}_B^{k+1})^{-1} \mathrm{T}_B^{k+1} (\mathrm{M}_B^{k+1})^{-1}\}_{22}} \\ 0 & \text{otherwise} \end{cases}, \tag{112}$$

and

$$\mathbb{E}[\{\bar{B} f_{k+1}^\top\}_{12}] = \mathbb{E}[\{\bar{B} f_{k+1}^\top\}_{21}] = 0. \tag{113}$$

In our algorithm, we do not have access to $\bar{B} + (\mathrm{M}_B^{k+1})^{-1} G_B^{k+1}$, but have B^{k+1} whose empirical distribution (of rows) converges to $\bar{B} + (\mathrm{M}_B^{k+1})^{-1} G_B^{k+1}$. We can therefore estimate the expectations in (111) and (112) by evaluating the right side for each row B_j^{k+1} and taking the average. For example, we compute $\mathbb{E}[\{\bar{B} f_{k+1}^\top\}_{11}]$ by evaluating (111) for each of the p rows of B^{k+1} and taking the average.

Appendix B. Implementation Details for MAR

Changing the matrix GLM model $q(\cdot, \cdot)$ only affects the denoising function g_k in AMP. Hence, in the seven-step implementation described in Appendix A.1, the only change is in the computation of g_k in steps 2 and 3. (The Jacobian g_k' in step 3 is approximated using g_k, as described on p. 34.) Recall that

$$g_k(Z^k, \bar{Y}) = \mathrm{Cov}[Z|Z^k]^{-1} \big(\mathbb{E}[Z|Z^k, \bar{Y}] - \mathbb{E}[Z|Z^k] \big). \tag{114}$$

Note that $\mathrm{Cov}[Z|Z^k]$ and $\mathbb{E}[Z|Z^k]$ are the same as that for mixed linear regression (with the formulas given in (86)), so we only need to evaluate $\mathbb{E}[Z|Z^k, \bar{Y}] \in \mathbb{R}^2$. This will be approximated using a Monte Carlo approach which we now describe. We have

$$\begin{aligned} \mathbb{E}[Z|Z^k = u, \bar{Y} = y] &= \frac{\int z \, p_{Z^k}(u) p_{Z|Z^k}(z|u) p_{\bar{Y}|Z,Z^k}(y|z,u) dz}{\int p_{Z^k}(u) p_{Z|Z^k}(z|u) p_{\bar{Y}|Z,Z^k}(y|z,u) dz} \\ &= \frac{\mathbb{E}_{Z|Z^k=u}[Z \, p_{Z^k}(u) p_{\bar{Y}|Z,Z^k}(y|Z,u)]}{\mathbb{E}_{Z|Z^k=u}[p_{Z^k}(u) p_{\bar{Y}|Z,Z^k}(y|Z,u)]}, \end{aligned} \tag{115}$$

where given $Z^k = u$ and $\bar{Y} = y$, the probability density functions inside the expectations can be easily evaluated since $Z^k \sim \mathcal{N}(0, \Sigma_{(22)}^k)$, and for $\sigma > 0$ and $z = (z_1, z_2)$,

$$\begin{aligned} p_{\bar{Y}|Z,Z^k}(y|z,u) &= p_{\bar{Y}|Z}(y|z) \\ &= \begin{cases} \phi_{\mathcal{N}}\left(\frac{y - z_1 - b_1}{\sigma}\right), & \text{if } z_1 + b_1 > z_2 + b_2 \\ \phi_{\mathcal{N}}\left(\frac{y - z_2 - b_2}{\sigma}\right), & \text{otherwise} \end{cases}, \end{aligned} \tag{116}$$

where $\phi_{\mathcal{N}}$ is the standard Gaussian density. With the above density functions, we can approximate the numerator and denominator of (115) by sampling z's from

$$(Z|Z^k = u) \sim \mathcal{N}\Big(\Sigma^k_{(12)}\big(\Sigma^k_{(22)}\big)^{-1}u,\, \Sigma^k_{(11)} - \Sigma^k_{(12)}\big(\Sigma^k_{(22)}\big)^{-1}\Sigma^k_{(21)}\Big), \tag{117}$$

and then taking the averages of the functions inside the expectations.

Additionally, the EM part of the EM-AMP algorithm requires $\mathbb{E}[Z|\bar{Y}]$. This is computed using Monte Carlo in a similar fashion to $\mathbb{E}[Z|Z^k, \bar{Y}]$:

$$\mathbb{E}[Z|\bar{Y} = y] = \frac{\int z\, p_Z(z) p_{\bar{Y}|Z}(y|z)dz}{\int p_Z(z) p_{\bar{Y}|Z}(y|zdz} = \frac{\mathbb{E}_Z[Z\, p_{\bar{Y}|Z}(y|Z)]}{\mathbb{E}_Z[p_{\bar{Y}|Z}(y|Z)]}, \tag{118}$$

where $p_{\bar{Y}|Z}(y|z)$ is given in (116). We can approximate the numerator and denominator of (118) by sampling z's from $Z \sim \mathcal{N}(0, \Sigma^k_{(11)})$ and then taking the averages of the functions inside the expectations.

In the mth iteration of EM-AMP, the expectations in (115) and (118) are computed by using the current intercept estimates (b_1^m, b_2^m) in the formula for $p_{\bar{Y}|Z}$ in (116).

Appendix C. Implementation Details for MOE

The changes required here are similar to those of MAR stated in Appendix B. In the seven-step implementation described in Appendix A.1, the only change is in the computation of g_k in steps 2 and 3. Recall that

$$g_k(Z^k, \bar{Y}) = \mathrm{Cov}[Z|Z^k]^{-1}\big(\mathbb{E}[Z|Z^k, \bar{Y}] - \mathbb{E}[Z|Z^k]\big). \tag{119}$$

Note that $\mathrm{Cov}[Z|Z^k]$ and $\mathbb{E}[Z|Z^k]$ are the same as that for mixed linear regression (with the formulas given in (86)), so we only need to evaluate $\mathbb{E}[Z|Z^k, \bar{Y}] \in \mathbb{R}^4$. This will be approximated using the same Monte Carlo approach as MAR, by writing

$$\mathbb{E}[Z|Z^k = u, \bar{Y} = y] = \frac{\mathbb{E}_{Z|Z^k=u}[Z\, p_{Z^k}(u) p_{\bar{Y}|Z,Z^k}(y|Z,u)]}{\mathbb{E}_{Z|Z^k=u}[p_{Z^k}(u) p_{\bar{Y}|Z,Z^k}(y|Z,u)]}, \tag{120}$$

where given $Z^k = u$ and $\bar{Y} = y$, the probability density functions inside the expectations can be easily evaluated since $Z^k \sim \mathcal{N}(0, \Sigma^k_{(22)})$, and for $\sigma > 0$ and $z = (z_1, z_2, z_3, z_4)$,

$$\begin{aligned}
p_{\bar{Y}|Z,Z^k}(y|z,u) = p_{\bar{Y}|Z}(y|z) &= \int_0^1 p_{\bar{Y},\bar{\psi}|Z}(y, v|z)dv \overset{(a)}{=} \int_0^1 p_{\bar{\psi}}(v) p_{\bar{Y}|\bar{\psi},Z}(y|v,z)dv \\
&\overset{(b)}{=} \int_0^{\frac{\exp(z_3)}{\exp(z_3)+\exp(z_4)}} p_{\bar{\psi}}(v)\phi_{\mathcal{N}}\Big(\frac{y-z_1}{\sigma}\Big)dv + \int_{\frac{\exp(z_3)}{\exp(z_3)+\exp(z_4)}}^1 p_{\bar{\psi}}(v)\phi_{\mathcal{N}}\Big(\frac{y-z_2}{\sigma}\Big)dv \\
&\overset{(c)}{=} \frac{\exp(z_3)}{\exp(z_3)+\exp(z_4)}\phi_{\mathcal{N}}\Big(\frac{y-z_1}{\sigma}\Big) + \Big(1 - \frac{\exp(z_3)}{\exp(z_3)+\exp(z_4)}\Big)\phi_{\mathcal{N}}\Big(\frac{y-z_2}{\sigma}\Big),
\end{aligned}$$

where $\phi_{\mathcal{N}}$ is the standard Gaussian density. Here (a) uses the independence between $\bar{\psi}$ and Z (see assumption in sentence below (13)), (b) uses the fact that $\bar{Y} = Z_1 + \bar{\epsilon}$ when

$\bar{\psi} \leq \frac{\exp(Z_3)}{\exp(Z_3)+\exp(Z_4)}$ and $\bar{Y} = Z_2 + \bar{\epsilon}$ when $\bar{\psi} > \frac{\exp(Z_3)}{\exp(Z_3)+\exp(Z_4)}$, (c) uses $p_{\bar{\psi}}(v) = 1$ for all $v \in [0,1]$ since $\bar{\psi} \sim \text{Uniform}[0,1]$. With the above density functions, we can approximate the numerator and denominator of (120) by sampling z's from

$$(Z|Z^k = u) \sim \mathcal{N}\Big(\Sigma^k_{(12)}\big(\Sigma^k_{(22)}\big)^{-1}u,\, \Sigma^k_{(11)} - \Sigma^k_{(12)}\big(\Sigma^k_{(22)}\big)^{-1}\Sigma^k_{(21)}\Big), \tag{121}$$

and then taking the averages of the functions inside the expectations.

References

Gabriel Arpino and Ramji Venkataramanan. Statistical-computational tradeoffs in mixed sparse linear regression, 2023. `https://arxiv.org/abs/2303.02118`.

Benjamin Aubin, Antoine Maillard, Florent Krzakala, Nicolas Macris, Lenka Zdeborová, et al. The committee machine: Computational to statistical gaps in learning a two-layers neural network. In *Advances in Neural Information Processing Systems*, 2018.

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

Gábor Balázs. *Convex Regression: Theory, Practice, and Applications*. PhD thesis, University of Alberta, 2016.

Gábor Balázs, András György, and Csaba Szepesvari. Near-optimal max-affine estimators for convex regression. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 56–64, 2015.

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.

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

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

Mohsen Bayati and Andrea Montanari. The lasso risk for Gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2011b.

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

Arun Tejasvi Chaganty and Percy Liang. Spectral experts for estimating mixtures of linear regressions. *International Conference on Machine Learning*, page 1040–1048, 2013.

Kabir Aladin Chandrasekher, Ashwin Pananjady, and Christos Thrampoulidis. Sharp global convergence guarantees for iterative nonconvex optimization with random data. *The Annals of Statistics*, 51(1):179 – 210, 2023.

Sitan Chen, Jerry Li, and Zhao Song. Learning mixtures of linear regressions in subexponential time via Fourier moments. In *ACM Symposium on Theory of Computing (STOC)*, 2020.

Yudong Chen, Xinyang Yi, and Constantine Caramanis. A convex formulation for mixed regression with two components: Minimax optimal rates. *Conference on Learning Theory*, pages 560–604, 2014.

Giovanni Compiani and Yuichi Kitamura. Using mixtures in econometric models: a brief review and some new results. *The Econometrics Journal*, 19(3):C95–C127, 2016.

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

Emilie Devijver, Yannig Goude, and Jean-Michel Poggi. Clustering electricity consumers using high-dimensional regression mixture models. *Applied Stochastic Models in Business and Industry*, pages 159–177, 2020.

David L. Donoho, Arian Maleki, and Andrea Montanari. Message passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106:18914–18919, 2009.

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

Jianqing Fan, Han Liu, Zhaoran Wang, and Zhuoran Yang. Curse of heterogeneity: Computational barriers in sparse mixture models and phase retrieval, 2018. `https://arxiv.org/abs/1808.06996`.

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

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

Oliver Y. Feng, Ramji Venkataramanan, Cynthia Rush, and Richard J. Samworth. A unifying tutorial on approximate message passing. *Foundations and Trends in Machine Learning*, 2022.

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

Fajwel Fogel, Irène Waldspurger, and Alexandre d'Aspremont. Phase retrieval for imaging problems. *Math. Prog. Comp.*, 8:311–335, 2016.

Cédric Gerbelot and Raphaël Berthier. Graph-based approximate message passing iterations, 2021. `https://arxiv.org/abs/2109.11905`.

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

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

Jens Gregor and Fernando R Rannou. Three-dimensional support function estimation and application for projection magnetic resonance imaging. *International journal of imaging systems and technology*, 12(1):43–50, 2002.

Sam Gross, Marc'Aurelio Ranzato, and Arthur Szlam. Hard mixtures of experts for large scale weakly supervised vision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6865–6873, 2017.

Bettina Grün and Friedrich Leisch. Applications of finite mixtures of regression models, 2007. `https://tinyurl.com/3sfyrwbs`.

Adityanand Guntuboyina and Bodhisattva Sen. Covering numbers for convex functions. *IEEE Transactions on Information Theory*, 59(4):1957–1965, 2013.

Lauren A. Hannah and David B. Dunson. Multivariate convex regression with adaptive partitioning. *Journal of Machine Learning Research*, 14(3):3261–3294, 2013.

Nhat Ho, Chiao-Yu Yang, and Michael I Jordan. Convergence rates for Gaussian mixtures of experts. *The Journal of Machine Learning Research*, 23(1):14523–14603, 2022.

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

Mian Huang and Weixin Yao. Mixture of regression models with varying mixing proportions: a semiparametric approach. *Journal of the American Statistical Association*, 107(498):711–724, 2012.

Mian Huang, Runze Li, and Shaoli Wang. Nonparametric mixture of regression models. *Journal of the American Statistical Association*, 108(503):929–941, 2013.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

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

Michael I. Jordan and Robert A. Jacobs. Hierarchical Mixtures of Experts and the EM Algorithm. *Neural Computation*, 6(2):181–214, 03 1994.

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

Yoshiyuki Kabashima, Florent Krzakala, Marc Mézard, Ayaka Sakata, and Lenka Zdeborová. Phase transitions and sample complexity in Bayes-optimal matrix factorization. *IEEE Transactions on Information Theory*, 62(7):4228–4265, 2016.

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

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

Weihao Kong, Raghav Somani, Zhao Song, Sham Kakade, and Sewoong Oh. Meta-learning for mixed linear regression. *International Conference on Machine Learning*, 119:5394–5404, 2020.

Akshay Krishnamurthy, Arya Mazumdar, Andrew McGregor, and Soumyabrata Pal. Sample Complexity of Learning Mixture of Sparse Linear Regressions. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Florent Krzakala, Marc Mézard, François Sausset, Yifan Sun, and Lenka Zdeborová. Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(8), 2012.

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

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

Gen Li and Yuting Wei. A non-asymptotic framework for approximate message passing in spiked models, 2023. `https://arxiv.org/abs/2208.03313`.

Qianyun Li, Runmin Shi, and Faming Liang. Drug sensitivity prediction with high-dimensional mixture regression. *PloS One*, pages 1–18, 2019.

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

Eunji Lim and Peter W. Glynn. Consistency of multidimensional convex regression. *Operations Research*, 60(1):196–208, 2012.

Junjie Ma, Ji Xu, and Arian Maleki. Approximate message passing for amplitude based optimization. In *International Conference on Machine Learning (ICML)*, pages 3371–3380, 2018.

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

Alessandro Magnani and Stephen P. Boyd. Convex piecewise-linear fitting. *Optimization and Engineering*, 10(1):1–17, 2009.

Antoine Maillard, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase retrieval in high dimensions: Statistical and computational phase transitions. In *Neural Information Processing Systems (NeurIPS)*, 2020.

Ashok Makkuva, Pramod Viswanath, Sreeram Kannan, and Sewoong Oh. Breaking the gridlock in mixture-of-experts: Consistent and efficient algorithms. *International Conference on Machine Learning (ICML)*, pages 4304–4313, 2019.

Ashok Makkuva, Sewoong Oh, Sreeram Kannan, and Pramod Viswanath. Learning in gated neural networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 3338–3348, 2020.

Rahul Mazumder, Arkopal Choudhury, Garud Iyengar, and Bodhisattva Sen. A computational framework for multivariate convex regression and its variants. *Journal of the American Statistical Association*, 114(525):318–331, 2019.

Geoffrey J McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2004.

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

Andrea Montanari. Graphical models concepts in compressed sensing. In *Compressed Sensing: Theory and Applications*, pages 394–438. Oxford University Press, 2012.

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

Radford M. Neal and Geoffrey E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. *Learning in graphical models*, pages 355–368, 1998.

Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. Phase retrieval using alternating minimization. *Advances in Neural Information Processing Systems*, pages 2796–2804, 2013.

Soumyabrata Pal, Arya Mazumdar, and Venkata Gandikota. Support recovery of sparse signals from a mixture of linear measurements. In *Advances in Neural Information Processing Systems*, 2021.

Soumyabrata Pal, Arya Mazumdar, Rajat Sen, and Avishek Ghosh. On learning mixture of linear regressions in the non-realizable setting. In *International Conference on Machine Learning*, pages 17202–17220, 2022.

Parthe Pandit, Mojtaba Sahraee Ardakan, Sundeep Rangan, Philip Schniter, and Alyson K Fletcher. Matrix inference and estimation in multi-layer models. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Jerry Ladd Prince and Alan S Willsky. Reconstructing convex sets from support line measurements. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(4):377–389, 1990.

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

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

Hanie Sedghi, Majid Janzamin, and Anima Anandkumar. Provable tensor methods for learning mixtures of generalized linear models. *International Conference on Artificial Intelligence and Statistics*, 51:1223–1231, 2016.

Emilio Seijo and Bodhisattva Sen. Nonparametric least squares estimation of a multivariate convex regression function. *Annals of Statistics*, 39(3):1633–1657, 2011.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

Yanyao Shen and Sujay Sanghavi. Iterative least trimmed squares for mixed linear regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Yong Sheng Soh and Venkat Chandrasekaran. Fitting tractable convex sets to support function evaluations. *Discrete & Computational Geometry*, 66(2):510–551, 2021.

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

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

Nelvin Tan and Ramji Venkataramanan. Mixed linear regression via approximate message passing. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

Kert Viele and Barbara Tong. Modeling with mixtures of linear regressions. *Statistics and Computing*, 12:315–330, 2002.

Jeremy P Vila and Philip Schniter. Expectation-maximization Gaussian-mixture approximate message passing. *IEEE Transactions on Signal Processing*, 61(19):4658–4672, 2013.

Tianhao Wang, Xinyi Zhong, and Zhou Fan. Universality of approximate message passing algorithms and tensor networks, 2022. `https://arxiv.org/abs/2206.13037`.

Zhaoran Wang, Quanquan Gu, Yang Ning, and Han Liu. High dimensional EM algorithm: Statistical optimization and asymptotic normality. *Advances in neural information processing systems*, pages 2521–2529, 2015.

CF Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, pages 95–103, 1983.

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

Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Alternating minimization for mixed linear regression. *International Conference on Machine Learning*, pages 613–621, 2014.

Seniha Esen Yuksel, Joseph N. Wilson, and Paul D. Gader. Twenty years of mixture of experts. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1177–1193, 2012.

Linjun Zhang, Rong Ma, T. Tony Cai, and Hongzhe Li. Estimation, confidence intervals, and large-scale hypotheses testing for high-dimensional mixed linear regression, 2020. `https://arxiv.org/abs/2011.03598`.

Yihan Zhang, Marco Mondelli, and Ramji Venkataramanan. Precise asymptotics for spectral methods in mixed generalized linear models, 2022. `https://arxiv.org/abs/2211.11368`.

Kai Zhong, Prateek Jain, and Inderjit S. Dhillon. Mixed linear regression with multiple components. *Advances in Neural Information Processing Systems*, pages 2190–2198, 2016.

Xinyi Zhong, Chang Su, and Zhou Fan. Approximate Message Passing for orthogonally invariant ensembles: Multivariate non-linearities and spectral initialization, 2021. `https://arxiv.org/abs/2110.02318`.

Rongda Zhu, Lingxiao Wang, Chengxiang Zhai, and Quanquan Gu. High-dimensional variance-reduced stochastic gradient expectation-maximization algorithm. In *International Conference on Machine Learning (ICML)*, pages 4180–4188, 2017.

Pini Zilber and Boaz Nadler. Imbalanced mixed linear regression, 2023. `https://arxiv.org/abs/2301.12559`.

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