SIAM J. MATH. DATA SCI.
Vol. 8, No. 2, pp. 411–439

Precise Asymptotics for Spectral Methods in Mixed Generalized Linear Models*

Yihan Zhang†, Marco Mondelli‡, and Ramji Venkataramanan§

Abstract. In a mixed generalized linear model, the goal is to learn multiple signals from unlabeled observations: each sample comes from exactly one signal, but it is not known which one. We consider the prototypical problem of estimating two statistically independent signals in a mixed generalized linear model with Gaussian covariates. Spectral methods are a popular class of estimators which output the top two eigenvectors of a suitable data-dependent matrix. However, despite the wide applicability, their design is still obtained via heuristic considerations, and the number of samples n needed to guarantee recovery is superlinear in the signal dimension d. In this paper, we develop exact asymptotics on spectral methods in the challenging proportional regime in which n, d grow large and their ratio converges to a finite constant. This allows us optimize the design of the spectral method, and combine it with a simple linear estimator, to minimize the estimation error. Our characterization exploits a mix of tools from random matrices, free probability, and the theory of approximate message passing algorithms. Numerical simulations for mixed linear regression and phase retrieval demonstrate the advantage enabled by our analysis over existing designs of spectral methods.

Key words. spectral estimator, generalized linear models, mixed regression, high-dimensional asymptotics, random matrix theory, approximate message passing (AMP)

MSC codes. 62E20, 62J05, 62J12

DOI. 10.1137/24M1702854

1. Introduction.

We consider the problem of learning multiple d-dimensional vectors from n unlabeled observations coming from a *mixed* generalized linear model (GLM):

$$y_i = q\left(\left\langle a_i, x^*_{\upsilon_i}\right\rangle, \varepsilon_i\right), \qquad i \in [n] = \{1, \dots, n\}. \tag{1.1}$$

Here, $x^*_1, \dots, x^*_\ell \in \mathbb{R}^d$ are the ℓ signals (regression vectors) to be recovered from the observation vector $y = (y_1, \dots, y_n) \in \mathbb{R}^n$ and the known design matrix $A = [a_1, \dots, a_n]^\top \in \mathbb{R}^{n \times d}$. For $i \in [n]$, ε_i is a noise variable, and υ_i is an $[\ell]$-valued latent variable, i.e., it indicates which signal each observation comes from, and is unknown to the statistician. The notation $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product, and $q : \mathbb{R}^2 \to \mathbb{R}$ is a known link function. For $\ell = 1$, (1.1) reduces to a generalized linear model [56], which covers many widely studied problems in statistical estimation including linear regression, logistic regression, phase retrieval [74, 33],

*Received by the editors October 18, 2024; accepted for publication (in revised form) December 8, 2025; published electronically June 15, 2026.
https://doi.org/10.1137/24M1702854
Funding: The first and second authors were partially supported by the 2019 Lopez-Loreta prize.
†School of Mathematics, University of Bristol, Bristol, UK (yihan.zhang@bristol.ac.uk).
‡Institute of Science and Technology Austria, Klosterneuburg, Austria (marco.mondelli@ist.ac.at).
§Department of Engineering, University of Cambridge, Cambridge, UK (rv285@cam.ac.uk).

and 1-bit compressed sensing [10]. The regression model with $\ell = 1$ implicitly assumes a homogeneous population, in which a single regression vector suffices to capture the features of the entire sample. In practice, it is often the case that the observations may come from multiple subpopulations. Mixed GLMs offer a flexible solution in settings with unlabeled heterogeneous data and have found applications in a variety of fields including biology, physics, and economics [57, 38, 49, 23]. When $q(g,\varepsilon) = g + \varepsilon$, (1.1) reduces to the widely studied mixture of linear regressions [83, 34, 76, 15, 88, 90, 75, 89, 37].

A natural approach to estimate the vectors $x_1^*, \dots, x_\ell^*$ from y and A is via the maximum-likelihood estimator (assuming a statistical model for $(\varepsilon_i)_{i\in[n]}$ is available). However, the corresponding optimization problem is nonconvex and NP-hard [88]. Thus, various low-complexity alternatives—mostly focusing on mixed linear regression—have been proposed: examples include expectation-maximization (EM) [41, 34, 76], alternating minimization [88, 75, 37], convex relaxation [20], moment descent methods [50, 17], and the use of tractable nonconvex objectives [90, 5]. Many of these methods are iterative in nature and require a "warm start" with an initial guess correlated with the ground truth. Spectral methods are a popular way to provide such initialization [88]. A variety of estimators based on the spectral decomposition of data-dependent matrices or tensors have been proposed for mixed GLMs [15, 88, 73]. In this paper, we focus on a spectral method that estimates the ℓ signals via the top-ℓ principal eigenvectors of the following data-dependent matrix:

$$D = \frac{1}{n}\sum_{i=1}^{n} \mathcal{T}(y_i) a_i a_i^\top \in \mathbb{R}^{d\times d}, \tag{1.2}$$

where $\mathcal{T}: \mathbb{R} \to \mathbb{R}$ is a suitably chosen preprocessing function. This spectral estimator with the preprocessing function $\mathcal{T}(y) = y^2$ was studied for mixed linear regression by Yi, Caramanis, and Sanghavi et al. [88], who showed that the signals can be accurately recovered when the number of observations n is of order $d \log d$. Furthermore, existing theoretical results for all estimators (including spectral, alternating minimization, and EM) require n to be of order at least $d \log d$ to guarantee accurate recovery [15, 88, 73, 50, 17]. This leads to the following natural questions:

What is the optimal sample complexity of a spectral estimator based on (1.2)?
Can we carry out a principled optimization of the preprocessing function $\mathcal{T}$?

A simpler alternative to obtain an initial estimate is to use the linear estimator

$$\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(y_i) a_i \ \in \mathbb{R}^d, \tag{1.3}$$

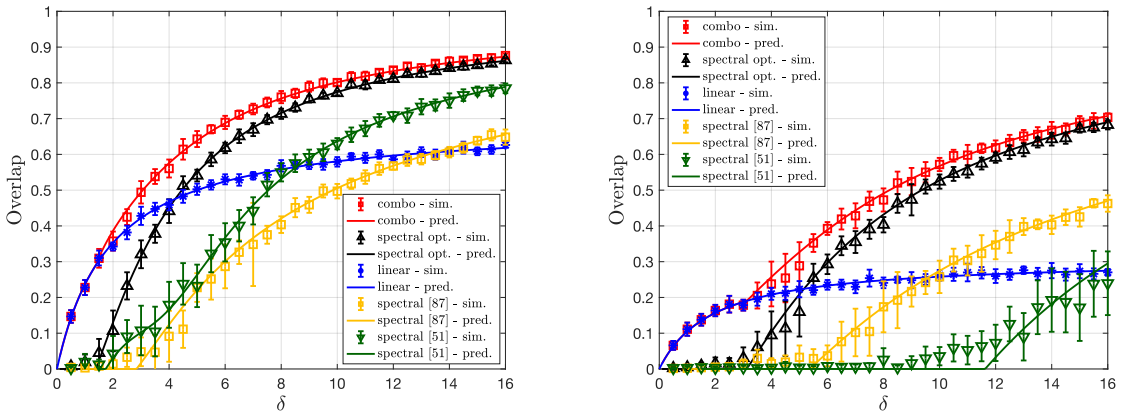
where $\mathcal{L}: \mathbb{R} \to \mathbb{R}$ is a suitable preprocessing function. The performance analysis of this linear estimator for the mixed GLM can be carried out similarly to that for the nonmixed case ($\ell = 1$); the analysis for the latter is given in [67, Proposition 1] and in [59, Lemma 2.1]. Thus, another natural question is the following:

What is the optimal way to combine a spectral estimator based on (1.2) *and the linear estimator in* (1.3)?

1.1. Main contributions. In this paper, we resolve the questions above for the recovery of two independent signals x_1^*, x_2^* with a Gaussian design matrix A. This is achieved by characterizing the high-dimensional limit of the joint empirical distribution of (i) the signals x_1^*, x_2^*, (ii) the linear estimator in (1.3), and (iii) spectral estimators based on the matrix in (1.2). Our analysis holds in the proportional setting where $n, d \to \infty$ with $n/d \to \delta \in (0, \infty)$. That is, we consider the regime where the ratio between sample size and signal dimension tends to a constant, as opposed to most analyses of mixed GLMs in the literature which assume $n = \Omega(d \log d)$. Our major findings are summarized as follows.

- Our master theorem (Theorem 3.1) characterizes the joint distribution of the linear estimator, the spectral estimator, and the signals in the high-dimensional limit. This joint distribution characterization holds for arbitrary preprocessing functions $\mathcal{L}, \mathcal{T} : \mathbb{R} \to \mathbb{R}$ in (1.2) and (1.3) (subject to mild regularity conditions). The limiting joint distribution is expressed as the law of a set of jointly Gaussian random variables whose covariance structure is explicitly derived in terms of the model and the preprocessing functions.
- As an immediate consequence of the distributional characterization, we derive the normalized correlations (or "overlaps") between the linear/spectral estimator and the signals (Corollary 3.8/Corollary 3.12). The linear estimator achieves a strictly positive overlap with each signal for any $\delta > 0$, provided a strictly positive overlap can be attained for some $\delta > 0$. In contrast, for the spectral estimator, we identify a threshold (depending on the preprocessing function $\mathcal{T}$) such that strictly positive overlap is attained as soon as δ exceeds this threshold. In general, there is no clear winner between the spectral and the linear estimator, and which one performs better depends on the setting.
- In fact, it is best to combine the linear and spectral estimators: our master theorem also allows us to compute the limiting overlap of a class of such combinations. In particular, the Bayes-optimal combination can be derived, which turns out to be linear in the two estimators due to the Gaussianity of their high-dimensional limits (Corollary 3.7).
- We determine the optimal preprocessing functions $\mathcal{L}^*, \mathcal{T}_1^*, \mathcal{T}_2^* : \mathbb{R} \to \mathbb{R}$ for the linear and spectral estimators that maximize the overlap between the estimator and each signal (Propositions 3.10 and 3.14). The optimal overlaps of linear and spectral estimators reveal intriguing behaviors of mixed models. In particular, there is a *single* function $\mathcal{L}^*$ that simultaneously maximizes the overlap between the linear estimator and each signal. In contrast, for the spectral method, one needs to employ two *different* functions $\mathcal{T}_1^*, \mathcal{T}_2^*$ in order to achieve the maximal overlaps with x_1^*, x_2^*, respectively. Furthermore, the optimal overlap of the spectral estimator with each signal approaches 1—the best possible value—as the aspect ratio δ grows. We remark that the same is not true for the linear estimator: the optimal overlap with each signal remains strictly less than 1 even as $\delta \to \infty$, as long as there is a strictly positive fraction of observations corresponding to each signal.

Our precise asymptotic analysis leads to a significant improvement over previous designs of spectral methods, as showcased in Figure 1 for noiseless mixed linear regression. The continuous lines correspond to our theoretical predictions ("pred."), which closely match the points coming from the simulations ("sim."). The following methods are compared: (i) optimal



(a) Recovery of x_1^* (b) Recovery of x_2^*

Figure 1. *Noiseless mixed linear regression with mixing parameter (i.e., probability that a sample corresponds to x_1^*) $\alpha = 0.6$. Overlaps with the first signal x_1^* (left) and the second signal x_2^* (right), computed via simulation ("sim.") and the theoretical prediction ("pred."), are plotted as a function of the aspect ratio $\delta = n/d$. The signal dimension is $d = 2000$. We note that our optimal spectral estimator enables weak recovery of both signals at a smaller δ than existing spectral estimators designed for nonmixed data. For example, in the right panel, our optimal spectral estimator starts weakly recovering x_2^* when $\delta > 3.1$, while other spectral estimators require at least $\delta > 5.3$.*

spectral method (black), obtained from Proposition 3.14; (ii) optimal linear method (blue), obtained from Proposition 3.10; (iii) combined estimator ("combo") (red), obtained from Corollary 3.7; (iv) spectral estimator for mixed linear regression proposed in [88] (yellow); (v) spectral estimator which optimizes the overlap in the nonmixed setting (green), proposed in [52]. The spectral methods resulting from our sharp analysis (red, black) significantly outperform existing methods (green, yellow), especially for low values of δ. More details on the experimental setup and additional simulation results can be found in section 4.

Proof techniques. We exploit a combination of tools from free probability, random matrices, and the theory of approximate message passing (AMP). Generalized approximate message passing (GAMP) refers to a family of iterative algorithms [68] with the following key feature: the joint distribution of the iterates is accurately tracked by a simple deterministic recursion, called *state evolution*. Our strategy to obtain the joint distribution of the linear/spectral estimators and the signals in the master theorem (Theorem 3.1) is to design a GAMP that (i) outputs the linear estimator as the first iterate, and (ii) then implements a power method, so that its fixed point corresponds to the spectral estimator. One challenge in the implementation of this strategy is that the state evolution of GAMP, in its original form for vanilla (nonmixed) GLMs, only records the correlation of its iterates with a single signal. To circumvent this issue, we equip GAMP with a state evolution recursion involving both signals and run a *pair* of GAMP iterations converging to the first and second top eigenvector of the spectral matrix D in (1.2), respectively. A second, even more fundamental, challenge is that, for the power method to converge to the desired eigenvector, a spectral gap between the corresponding eigenvalue and the rest of the spectrum is required. For nonmixed GLMs, the spectral analysis was carried out in earlier work [51, 58], which characterized the

limiting eigenvalues of D as well as the overlaps using tools from random matrix theory. Here, the difficulty comes from the mixed effect of the model, leading to additional matrix terms which appear challenging to control. Our approach is to decompose D into the sum of two matrices, D_1 and D_2, each consisting of components only pertaining to the first and second signal, respectively. Now, D_1, D_2 can be individually viewed as generated from a nonmixed GLM; hence their limiting spectra are well understood. The key observation is then that, by assuming both signals to be independent and uniformly distributed on the sphere, D_1 and D_2 become *asymptotically free.*[1] Thus, we are able to characterize the sum of these two spiked matrices by using techniques from free probability.

1.2. Related work. Mixtures of GLMs have been studied in machine learning as "hierarchical mixtures of experts" [40]. Bayesian methods for this problem were investigated in [66, 86, 83]. Khalili and Chen [41] proposed a penalized likelihood approach for variable selection in mixed GLMs, showing consistency in the low-dimensional setting (where the dimension d is fixed as n grows). Städler, Bühlmann, and van de Geer [76] analyzed ℓ_1-penalized estimators for high-dimensional mixed linear regression (MLR). Zhang et al. [89] considered MLR with two sparse components, when the mixing proportion and the covariance structure of the covariates are unknown. The works [41, 76, 89] all use variants of the EM algorithm for optimizing a suitable penalized likelihood function. Balakrishnan, Wainwright, and Yu [3] and Klusowski, Yang, and Brinda [42] obtained statistical guarantees on the performance of EM for a class of problems, including symmetric MLR with $x_1^* = -x_2^*$. Variants of EM for symmetric MLR were also analyzed in [85, 87, 92]. Minimax lower bounds were obtained in [31], and statistical-computational gaps were recently studied in [2]. Kong et al. [43] studied MLR as a canonical example of meta-learning: in the setting where the number of signals (ℓ) is large, they derived conditions under which a large number of signals with a few observations can compensate for the lack of signals with abundantly many observations. The prediction error of MLR in the nonrealizable setting, where no generative model is assumed for the data, was studied in [65]. Chandrasekher, Pananjady, and Thrampoulidis [16] analyzed the performance of iterative algorithms (not including AMP) for mixtures of GLMs. They provide a sharp characterization of the per-iteration error with sample-splitting in the regime $n \asymp d\,\mathrm{polylog}(d)$, assuming a Gaussian design and a random initialization. An AMP estimator for mixed GLMs was recently studied in [79]. We emphasize that the focus of the current paper is not on using the AMP algorithm as an estimator for mixed GLMs. Rather, we use AMP as a proof technique to obtain a precise distributional characterization of the spectral estimator, and we use this characterization to optimize its accuracy.

Spectral methods based on (1.2) were introduced in [48] for standard GLMs (nonmixed, with $\ell = 1$). For the special case of phase retrieval, a series of works has provided increasingly refined bounds on the number of samples needed to guarantee signal recovery via the spectral method [63, 13, 19]. This type of analysis is based on matrix concentration inequalities, a technique that typically does not return exact values for the overlap between the signal and the estimate. More recently, an exact high-dimensional analysis for generalized linear models was carried out in [51, 58]. These works focus on the regime of interest in this paper: n and

[1] Asymptotic freeness can be thought of as the random matrix analogue of *independence* of random variables.

d growing at a proportional rate δ. This sharp analysis allows for the optimization of the preprocessing function: the choice of $\mathcal{T}$ minimizing the value of δ (and, hence, the amount of data) needed to achieve a strictly positive overlap was provided in [58]; furthermore, the choice of $\mathcal{T}$ maximizing the overlap was provided in [52]. Going beyond the proportional regime in which n is linear in d, bounds on the sample complexity required for moment methods (including spectral) to achieve nonvanishing overlap were recently obtained in [21]. The aforementioned analyses assume a Gaussian design matrix. Beyond this assumption, [27] provides precise asymptotics for design matrices sampled from the Haar distribution, and [54] studies rotationally invariant designs. Moving to the mixed regression setting ($\ell > 1$), Yi, Caramanis, and Sanghavi [88] proposed a spectral estimator based on (1.2) with $\mathcal{T}(y) = y^2$. The analysis is based on concentration inequalities and requires the number of samples n to be of order $d \log d$ for accurate recovery. Estimators based on spectral decomposition of data-dependent tensors were proposed for MLR in [15] and for mixed GLMs in [73]. However, these methods require n to be of order at least d^3 for accurate recovery. Our work is the *first* to establish exact asymptotics for a mixed GLM in the linear sample-size regime: $n, d \to \infty$ with $n/d \to \delta \in (0, \infty)$. To achieve this goal, our strategy differs from analyses of spectral methods in the nonmixed setting [51, 58] which reduce the study of the spectrum of D to that of a rank-1 perturbation. In contrast, our analysis is based on a combination of techniques from free probability and approximate message passing (AMP).

AMP is a family of iterative algorithms that has been applied to several problems in high-dimensional statistics, including estimation in linear models [25, 7, 45], generalized linear models [68, 72, 77], and low-rank matrix estimation [22, 69, 46]; see also the review [36]. A key feature of AMP algorithms is that under suitable model assumptions, the empirical joint distribution of their iterates can be exactly characterized in the high-dimensional limit, in terms of a simple scalar recursion called *state evolution.* By taking advantage of this characterization, AMP methods have been used to derive exact high-dimensional asymptotics for convex penalized estimators such as LASSO [8], M-estimators [24], logistic regression [77], and SLOPE [12]. AMP algorithms have been initialized via spectral methods in the context of low-rank matrix estimation [62] and GLMs [60]. Furthermore, they have been used—in a nonmixed setting—to combine linear and spectral estimators [59]. A finite-sample analysis which allows the number of iterations to grow roughly as $\log n$ (n being the ambient dimension) was put forward in [71], and the recent paper [47] improves this guarantee to a linear (in n) number of iterations. This could potentially allow one to study settings in which $\delta = n/d$ approaches the spectral threshold. The works on AMP discussed above all assume i.i.d. Gaussian matrices. A number of recent papers have proposed generalizations of AMP for the much broader class of rotationally invariant matrices, e.g., [64, 53, 70, 78, 91, 32, 61, 82].

Finally, we mention the recent paper [44] that derived precise asymptotics of spectral estimators for multi-index models by generalizing the techniques in [51, 58]. However, [44] did not derive the *joint* distribution of spectral and linear estimators, or the (optimal) combination of the two. To the best of our knowledge, such results do not follow immediately from the pure random matrix theory-type results in [44], but require additional work to handle the correlation between spectral and linear estimators. These are achieved in the present work using a mix of tools from free probability and AMP.

2. Preliminaries. The ith element in $a \in \mathbb{R}^p$ is denoted by a_i. If a vector has multiple subscripts, the component index is the last one. For a symmetric $M \in \mathbb{R}^{p\times p}$, we denote by μ_M its empirical spectral distribution. The (real) eigenvalues of M are $\lambda_1(M) \geq \lambda_2(M) \geq \cdots \geq \lambda_p(M)$, and the corresponding eigenvectors are $v_1(M), v_2(M), \cdots, v_p(M)$. The (i,j)th entry of M is denoted by $M_{i,j}$. For a random variable X, $\mathrm{supp}(X)$ denotes the support of its density function. The orthogonal group in dimension p is $\mathbb{O}(p) := \{O \in \mathbb{R}^{p\times p} : OO^\top = O^\top O = I_p t\}$. The unit sphere in dimension p is $\mathbb{S}^{p-1} := \{x \in \mathbb{R}^p : \|x\|_2 = 1\}$. For two distributions P and Q, $P \otimes Q$ is their product distribution, and $P^{\otimes k}$ is the k-fold product distribution of P.

Model. We consider a two-component mixed GLM with signal vectors $x_1^*, x_2^* \in \mathbb{S}^{d-1}$, covariate vectors $a_1, \cdots, a_n \in \mathbb{R}^d$, and a known link function $q\colon \mathbb{R}^2 \to \mathbb{R}$. Let P_ε be a noise distribution over $\mathbb{R}$. The n observations $y_1, \cdots, y_n \in \mathbb{R}$ are generated as

$$y_i = q\left(\langle a_i, \eta_i x_1^* + (1-\eta_i)x_2^*\rangle, \varepsilon_i\right), \qquad i \in [n]. \tag{2.1}$$

Here, the vector of latent variables $\underline{\eta} := (\eta_1, \cdots, \eta_n) \sim \mathrm{Bern}(\alpha)^{\otimes n}$ indicates which signal is selected by each observation, and is *unobserved.* The latent variable vector $\underline{\eta}$, the signals x_1^*, x_2^*, the covariate vectors $a_1, \ldots, a_n$, and the noise vector $\underline{\varepsilon} := (\varepsilon_1, \cdots, \varepsilon_n) \sim P_\varepsilon^{\otimes n}$ are mutually independent. Then, (2.1) is equivalent to

$$y_i \,|\, \langle a_i, \eta_i x_1^* + (1-\eta_i)x_2^*\rangle \sim p(\cdot \,|\, \langle a_i, \eta_i x_1^* + (1-\eta_i)x_2^*\rangle), \tag{2.2}$$

where $p(\cdot|g)$ denotes the distribution of $q(g,\varepsilon)$ for a fixed $g \in \mathbb{R}$ and $\varepsilon \sim P_\varepsilon$ independent of g. The design matrix is $A = [a_1^\top, \ldots, a_n^\top]^\top \in \mathbb{R}^{n\times d}$. Given A, upon observing $y = (y_1, \cdots, y_n) \in \mathbb{R}^n$, our goal is to estimate x_1^* and x_2^*. Given a pair of estimators $\widehat{x}_1 = \widehat{x}_1(y, A)$, $\widehat{x}_2 = \widehat{x}_2(y, A) \in \mathbb{R}^d$, we measure the performance via their *overlap* with the respective signals:

$$\lim_{d\to\infty} \frac{|\langle \widehat{x}_1, x_1^*\rangle|}{\|\widehat{x}_1\|_2 \|x_1^*\|_2}, \qquad \lim_{d\to\infty} \frac{|\langle \widehat{x}_2, x_2^*\rangle|}{\|\widehat{x}_2\|_2 \|x_2^*\|_2}.$$

Throughout the paper, the following assumptions are imposed:

(A1) x_1^*, x_2^* are independent and uniform on the unit sphere, $(x_1^*, x_2^*) \sim \mathrm{Unif}(\mathbb{S}^{d-1})^{\otimes 2}$.
(A2) $\alpha \in (1/2, 1)$.
(A3) The noise sequence $\underline{\varepsilon} \in \mathbb{R}^n$ is i.i.d. according to $\underline{\varepsilon} \sim P_\varepsilon^{\otimes n}$, and P_ε has finite second moment.
(A4) $a_1, \cdots, a_n \in \mathbb{R}^d$ are i.i.d., each distributed according to $a_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0_d, I_d)$.
(A5) We consider the proportional regime where $n, d \to \infty$ and $n/d \to \delta$ for some constant $\delta > 0$ which we call *aspect ratio.*

As for assumption (A1), choosing signals uniform on the sphere corresponds to having no structural information about them. This requirement is natural, since spectral methods are typically unable to exploit prior information about the signal. We expect that all results of the present paper hold for the more relaxed setting where the signals x_1^*, x_2^* are independent of the design matrix A and the noise vector $\underline{\varepsilon}$, and satisfy

$$\lim_{d\to\infty} \|x_1^*\|_2 = \lim_{d\to\infty} \|x_2^*\|_2 = 1, \qquad \lim_{d\to\infty} \langle x_1^*, x_2^*\rangle = 0.$$

In particular, we believe that the assumption that the signals are uniformly distributed on the sphere is not required. In that case, (5.4) in the argument of the reduction to the free

sum of two random matrices is no longer an exact equality (in distribution). However, we still expect asymptotic freeness to hold in the proportional limit. We leave its formal justification to future work. Understanding the effect of correlation on the performance of spectral estimators and the design of the optimal preprocessing function is an exciting future direction. Assumption (A2) is without loss of generality: if $0 < \alpha < 1/2$, one can simply interchange the roles of x_1^* and x_2^*. When $\alpha = 1/2$, the top two eigenvectors given by the spectral method correspond to the same limiting eigenvalue as $n \to \infty$. These eigenvectors provide an estimate on the space spanned by x_1^*, x_2^*, and in order to estimate the individual signals, an additional one-dimensional grid search is required. Provided this extra step is carried out, our results still apply; see Remark 3.6 in the current paper and Remarks SM3.3 and SM4.4 in the accompanying supplementary material file (supplement.pdf [local/web 690KB]). Assumption (A4) is common in the related literature [88, 58, 51, 52], and the potential universality beyond Gaussian design matrices is discussed in section 6.

Linear estimator. Given the preprocessing function $\mathcal{L}\colon \mathbb{R} \to \mathbb{R}$, the *linear estimator* is

$$\widehat{x}^{\text{lin}} := \frac{1}{n} A^\top \mathcal{L}(y) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i) a_i \in \mathbb{R}^d, \tag{2.3}$$

where $\mathcal{L}$ is applied componentwise, i.e., $\mathcal{L}(y) = (\mathcal{L}(y_1), \cdots, \mathcal{L}(y_n))$. Let Y be defined as

$$Y = q(G, \varepsilon), \quad \text{where } (G, \varepsilon) \sim \mathcal{N}(0,1) \otimes P_\varepsilon. \tag{2.4}$$

We make the following assumption on $\mathcal{L}$.

(A6) $\mathcal{L}\colon \mathbb{R} \to \mathbb{R}$ is Lipschitz and satisfies

$$\mathbb{E}\left[G\mathcal{L}(Y)\right] \neq 0, \quad \mathbb{E}\left[|G\mathcal{L}(Y)|\right] < \infty.$$

The first condition guarantees that the linear method w.r.t. $\mathcal{L}$ attains positive overlaps with both signals, and the second condition is rather mild and purely technical.

Spectral estimator. Let $\mathcal{T}\colon \mathbb{R} \to \mathbb{R}$ be a preprocessing function, and consider

$$T := \operatorname{diag}(\mathcal{T}(y)) \in \mathbb{R}^{n \times n}, \quad D := \frac{1}{n} A^\top T A = \frac{1}{n} \sum_{i=1}^{n} \mathcal{T}(y_i) a_i a_i^\top \in \mathbb{R}^{d \times d}, \tag{2.5}$$

where $\mathcal{T}(y) = (\mathcal{T}(y_1), \cdots, \mathcal{T}(y_n))$. Then, the spectral method computes the top two eigenvectors $v_1(D), v_2(D)$ of D as estimates of x_1^*, x_2^*. We make the following assumption on $\mathcal{T}$.

(A7) Let Y be defined in (2.4). Then, $\mathcal{T}(Y)$ is not almost surely zero, i.e., $\Pr[\mathcal{T}(Y) = 0] < 1$, $\mathcal{T}$ is Lipschitz and satisfies

$$\inf_{y \in \operatorname{supp}(Y)} \mathcal{T}(y) > -\infty \quad \text{and} \quad 0 < \sup_{y \in \operatorname{supp}(Y)} \mathcal{T}(y) < \infty.$$

In words, we require $\mathcal{T}$ to be bounded, with strictly positive upper edge of its range. A bounded preprocessing function is also required in the nonmixed setting [58, 51]. The requirement on the sup to be strictly positive is purely technical, and it simply rules out the trivial cases in which the spectral matrix D is all-zero with high probability. Assumption (A7) is satisfied by the preprocessing function that maximizes the overlap (cf. Proposition 3.14).

3. Main results. We start by defining a few auxiliary quantities. Let $\delta_1 = \alpha\delta, \delta_2 = (1-\alpha)\delta$, and $Z = \mathcal{T}(Y)$, with Y defined as in (2.4). Define $\varphi\colon (\sup \operatorname{supp}(Z), \infty) \to \mathbb{R}$ and $\psi\colon (\sup \operatorname{supp}(Z), \infty) \times (0, \infty) \to \mathbb{R}$ as

$$\varphi(\lambda) := \lambda \mathbb{E}\left[\frac{ZG^2}{\lambda - Z}\right], \tag{3.1}$$

$$\psi(\lambda; \Delta) := \lambda\left(\frac{1}{\Delta} + \mathbb{E}\left[\frac{Z}{\lambda - Z}\right]\right). \tag{3.2}$$

In what follows, we will set the second argument Δ of ψ to δ, δ_1, and δ_2. For $\Delta \in \{\delta, \delta_1, \delta_2\}$, let $\overline{\lambda}(\Delta) > \sup \operatorname{supp}(Z)$ be the minimum point of $\psi(\cdot; \Delta)$, i.e.,

$$\overline{\lambda}(\Delta) := \operatorname*{argmin}_{\lambda > \sup \operatorname{supp}(Z)} \psi(\lambda; \Delta). \tag{3.3}$$

Since ψ is convex in its first argument (see Lemma SM10.1), this minimum point is obtained by setting the derivative to 0. Furthermore, define $\zeta\colon (\sup \operatorname{supp}(Z), \infty) \times (0, \infty) \to \mathbb{R}$ as

$$\zeta(\lambda; \Delta) := \psi(\max\{\lambda, \overline{\lambda}(\Delta)\}; \Delta). \tag{3.4}$$

Finally, for $i \in \{1, 2\}$, by [58, Lemma 2], the equation $\zeta(\lambda; \delta_i) = \varphi(\lambda)$ admits a unique solution in $\lambda \in (\sup \operatorname{supp}(Z), \infty)$ which we call $\lambda^*(\delta_i)$. The functions $\psi(\lambda; \Delta), \varphi(\lambda), \zeta(\lambda; \Delta)$ together with the parameters $\lambda^*(\Delta), \overline{\lambda}(\Delta)$ are plotted in Figure 2 for $\Delta \in \{\delta, \delta_1, \delta_2\}$. Some convexity and monotonicity properties of these functions can be found in Lemma SM10.1.

The empirical distribution of a vector $u \in \mathbb{R}^d$ is given by $\frac{1}{d}\sum_{i=1}^d \delta_{u_i}$, where δ_{u_i} denotes a Dirac delta mass on u_i. Similarly, the joint empirical distribution of the rows of a matrix $(u^1, u^2, \dots, u^t) \in \mathbb{R}^{d \times t}$ is $\frac{1}{d}\sum_{i=1}^d \delta_{(u_i^1, \dots, u_i^t)}$. Our master theorem is an exact characterization in the high-dimensional limit of the joint empirical distribution of the rows of the signals, the linear estimator, and the spectral estimators. In particular, we show that this joint empirical distribution converges to the law of a Gaussian random vector with a specified covariance matrix. The result is stated in terms of the following parameters: the asymptotic correlations $\rho_1^{\text{lin}}, \rho_2^{\text{lin}}$ between the linear estimator and the two signals; the asymptotic normalized Euclidean

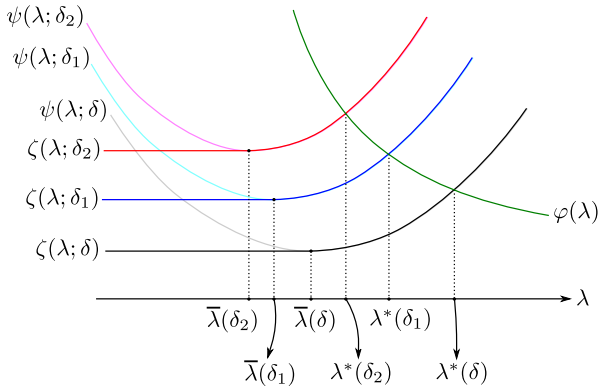


Figure 2. *Plot of $\psi(\lambda; \Delta), \varphi(\lambda), \zeta(\lambda; \Delta)$ as functions of λ with $\Delta \in \{\delta, \delta_1, \delta_2\}$.*

norm n^{lin} of the linear estimator; and the asymptotic correlations $\rho_1^{\text{spec}}, \rho_2^{\text{spec}}$ between the spectral estimators and the two signals. The formulas for these quantities are

$$n^{\text{lin}} := \left((\alpha^2 + (1-\alpha)^2)\mathbb{E}\left[G\mathcal{L}(Y)\right]^2 + \frac{\mathbb{E}\left[\mathcal{L}(Y)^2\right]}{\delta} \right)^{\frac{1}{2}}, \tag{3.5a}$$

$$\rho_1^{\text{lin}} := \frac{\alpha \mathbb{E}\left[G\mathcal{L}(Y)\right]}{n^{\text{lin}}}, \quad \rho_2^{\text{lin}} := \frac{(1-\alpha)\mathbb{E}\left[G\mathcal{L}(Y)\right]}{n^{\text{lin}}}, \tag{3.5b}$$

$$\rho_1^{\text{spec}} := \left(\frac{\frac{1}{\delta} - \mathbb{E}\left[\left(\frac{Z}{\lambda^*(\delta_1) - Z} \right)^2 \right]}{\frac{1}{\delta} + \alpha \mathbb{E}\left[\left(\frac{Z}{\lambda^*(\delta_1) - Z} \right)^2 (G^2 - 1) \right]} \right)^{\frac{1}{2}}, \tag{3.5c}$$

$$\rho_2^{\text{spec}} := \left(\frac{\frac{1}{\delta} - \mathbb{E}\left[\left(\frac{Z}{\lambda^*(\delta_2) - Z} \right)^2 \right]}{\frac{1}{\delta} + (1-\alpha) \mathbb{E}\left[\left(\frac{Z}{\lambda^*(\delta_2) - Z} \right)^2 (G^2 - 1) \right]} \right)^{\frac{1}{2}}.$$

Theorem 3.1 is stated in terms of *pseudo-Lipschitz* test functions. A function $\Psi : \mathbb{R}^m \to \mathbb{R}$ is pseudo-Lipschitz of order $k \geq 1$, denoted $\Psi \in \text{PL}(k)$, if there is a constant $C > 0$ such that

$$\|\Psi(x) - \Psi(y)\|_2 \leq C(1 + \|x\|_2^{k-1} + \|y\|_2^{k-1}) \|x - y\|_2 \tag{3.6}$$

for all $x, y \in \mathbb{R}^m$. Examples of pseudo-Lipschitz functions of order two are $\Psi(u) = u^2$ and $\Psi(u, v) = |uv|$ for $u, v \in \mathbb{R}$. We consider pseudo-Lipschitz test functions of order two, as those suffice to compute the asymptotic overlaps between the signals and the various estimators. One could extend Theorem 3.1 to test functions in $\text{PL}(k)$ for $k > 2$, at the cost of a more involved argument and an additional assumption on the finiteness of the moments of P_ε.

Theorem 3.1 (master theorem on joint distribution). *Consider the setting of section 2, and let assumptions (A1)–(A7) hold. Define the following rescaled vectors of Euclidean norm $\sqrt{d}$: $x^{\text{lin}} = \sqrt{d}\widehat{x}^{\text{lin}}/\|\widehat{x}^{\text{lin}}\|_2$, and for $i \in \{1, 2\}$, $\overline{x}_i^* = \sqrt{d}x_i^*$, $x_i^{\text{spec}} = s_i\sqrt{d}\, v_i(D)$, where the sign $s_i \in \{-1, 1\}$ is chosen such that $\langle s_i v_i(D), x_i^* \rangle \geq 0$. Then, the following holds almost surely for any* $\text{PL}(2)$ *function $\Psi : \mathbb{R}^3 \to \mathbb{R}$. If $\lambda^*(\delta_1) > \overline{\lambda}(\delta)$, then*

$$\lim_{d\to\infty} \frac{1}{d} \sum_{i=1}^{d} \Psi\left(\overline{x}_{1,i}^*, x_i^{\text{lin}}, x_{1,i}^{\text{spec}} \right) = \mathbb{E}\left[\Psi(X_1, \rho_1^{\text{lin}} X_1 + \rho_2^{\text{lin}} X_2 + W^{\text{lin}}, \rho_1^{\text{spec}} X_1 + W_1^{\text{spec}}) \right]. \tag{3.7}$$

Similarly, if $\lambda^*(\delta_2) > \overline{\lambda}(\delta)$, then

$$\lim_{d\to\infty} \frac{1}{d} \sum_{i=1}^{d} \Psi\left(\overline{x}_{2,i}^*, x_i^{\text{lin}}, x_{2,i}^{\text{spec}} \right) = \mathbb{E}\left[\Psi(X_2, \rho_1^{\text{lin}} X_1 + \rho_2^{\text{lin}} X_2 + W^{\text{lin}}, \rho_2^{\text{spec}} X_2 + W_2^{\text{spec}}) \right]. \tag{3.8}$$

Here $(X_1, X_2) \sim \mathcal{N}(0, 1)^{\otimes 2}$, the pairs $(W^{\text{lin}}, W_1^{\text{spec}})$ and $(W^{\text{lin}}, W_2^{\text{spec}})$ are independent of (X_1, X_2), and each pair is jointly Gaussian with zero mean and covariance given by

$$\mathbb{E}[(W^{\mathrm{lin}})^2] = 1 - (\rho_1^{\mathrm{lin}})^2 - (\rho_2^{\mathrm{lin}})^2, \quad \mathbb{E}[(W_1^{\mathrm{spec}})^2] = 1 - (\rho_1^{\mathrm{spec}})^2, \quad \mathbb{E}[(W_2^{\mathrm{spec}})^2] = 1 - (\rho_2^{\mathrm{spec}})^2,$$
$$\mathbb{E}[W^{\mathrm{lin}} W_1^{\mathrm{spec}}] = \frac{\alpha \rho_1^{\mathrm{spec}}}{n^{\mathrm{lin}}} \mathbb{E}\left[\frac{G\mathcal{L}(Y)Z}{\lambda^*(\delta_1) - Z}\right], \quad \mathbb{E}[W^{\mathrm{lin}} W_2^{\mathrm{spec}}] = \frac{(1-\alpha)\rho_2^{\mathrm{spec}}}{n^{\mathrm{lin}}} \mathbb{E}\left[\frac{G\mathcal{L}(Y)Z}{\lambda^*(\delta_2) - Z}\right].$$

The outline of the argument is presented in section 5. The full proof is given in section SM4, and it relies on the characterization of the eigenvalues of D carried out in Theorem SM3.1, which is stated and proved in section SM3.

Remark 3.2 (equivalence to convergence of empirical distribution). Equation (3.7) is equivalent to the statement that the joint empirical distribution of $(\overline{x}_1^*, x^{\mathrm{lin}}, x_1^{\mathrm{spec}})$ converges in Wasserstein-2 distance to the joint law of $(X_2, \rho_1^{\mathrm{lin}} X_1 + \rho_2^{\mathrm{lin}} X_2 + W^{\mathrm{lin}}, \rho_1^{\mathrm{spec}} X_1 + W_1^{\mathrm{spec}})$. The equivalence between convergence of empirical distributions in Wasserstein distance and convergence of empirical averages of pseudo-Lipschitz functions is proved in [36, Corollary 7.21].

Remark 3.3 (what if either the linear or spectral estimator is ineffective). The validity of the description of the joint law of the first signal and the linear/spectral estimators in (3.7) relies on two assumptions: $\mathbb{E}[G\mathcal{L}(Y)] \neq 0$ for the linear estimator and $\lambda^*(\delta_1) > \overline{\lambda}(\delta)$ for the spectral one. They guarantee that both estimators achieve nonzero asymptotic overlaps with x_1^*, i.e., $\rho_1^{\mathrm{lin}} \neq 0$ and $\rho_1^{\mathrm{spec}} > 0$. If either condition is not satisfied, a conclusion similar to (3.7) still holds with $\Psi\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ only taking x_1^* and the nontrivial estimator as inputs. Specifically, if only the linear estimator is effective, then we terminate GAMP in (5.8) after one step $t = 0$ and obtain the distributional characterization; if only the spectral estimator is effective, then the initializer in (5.13) ensures that the same proof goes through without modifications, again leading to the desired conclusion. An analogous argument holds for the second signal.

Remark 3.4 ($v_i(D)$ estimates x_i^*). We have been using $v_i(D)$ to estimate x_i^* for $i \in \{1, 2\}$. In fact, $v_1(D)$ is asymptotically uncorrelated with x_2^* (and $v_2(D)$ is asymptotically uncorrelated with x_1^*), according to the characterization of the asymptotic distribution of the top two eigenvectors in subsection SM4.1. Intuitively, this phenomenon arises due to the orthogonality between the two signals and fails to hold otherwise. For instance, if $\lim_{d\to\infty} \langle x_1^*, x_2^* \rangle = \rho \neq 0$, then both the first and second eigenvectors are asymptotically correlated with both signals. This can be formally justified by specializing [44, Theorem 4.2] to mixtures of GLMs and is numerically corroborated in [44, Figure 2].

Remark 3.5 (sign calibration of spectral estimator). As the eigenvectors of a matrix are insensitive to sign flip, the spectral estimators $x_1^{\mathrm{spec}}, x_2^{\mathrm{spec}}$ are defined up to a change of sign. In Theorem 3.1, we pick the signs so that the resulting overlaps $\rho_1^{\mathrm{spec}}, \rho_2^{\mathrm{spec}}$ are positive. In practice, there is a simple way to resolve the sign ambiguity: one can match the sign of $\mathbb{E}[(\rho_1^{\mathrm{lin}} X_1 + \rho_2^{\mathrm{lin}} X_2 + W^{\mathrm{lin}})(\rho_i^{\mathrm{spec}} X_i + W_i^{\mathrm{spec}})]$ with that of the scalar product $\langle x^{\mathrm{lin}}, x_i^{\mathrm{spec}} \rangle$, as the latter can be computed empirically (without knowing x_1^*, x_2^*).

Remark 3.6 (master theorem for $\alpha = 1/2$). Even though we assume $\alpha \in (1/2, 1)$ (see assumption (A2)), the conclusion of Theorem 3.1 still holds for $\alpha = 1/2$ with a slight modification in the definition of the spectral estimators. In this case, as $n \to \infty$ the top two eigenvectors given by the spectral method correspond to the same limiting eigenvalue. These eigenvectors, $v_1(D)$ and $v_2(D)$, estimate the subspace spanned by x_1^*, x_2^*. To estimate each

individual signal, we search for a vector in $\mathrm{span}\{v_1(D), v_2(D)\}$ whose correlation with x^{lin} is closest to the theoretical prediction from Theorem 3.1. Indeed, let $x_1^{\mathrm{spec}}, x_2^{\mathrm{spec}}$ be defined as

(3.9)
$$x_i^{\mathrm{spec}} := \underset{v \in \mathrm{span}\{v_1(D), v_2(D)\} \cap \sqrt{d}\mathbb{S}^{d-1}}{\mathrm{argmin}} \left| \frac{\langle v, x^{\mathrm{lin}} \rangle}{\sqrt{d}} - \left(\rho_i^{\mathrm{lin}} \rho_i^{\mathrm{spec}} + \mathbb{E}\left[W^{\mathrm{lin}} W_i^{\mathrm{spec}} \right] \right) \right| \quad \text{for } i \in \{1, 2\}.$$

Then, (3.7) and (3.8) hold, provided $\mathbb{E}[G\mathcal{L}(Y)] \neq 0$ (which guarantees that the linear estimator attains nonzero overlaps; see assumption (A6) and (3.5b)). We stress that (3.9) is computable in practice since it only involves x^{lin} and theoretical predictions. If x^{lin} is ineffective (which is the case, for example, in mixed phase retrieval, as mentioned in section SM1), a similar grid search can still be performed if the statistician is given as side information a vector with known correlation with a signal. The reader is referred to Remarks SM3.3 and SM3.4 for the adaptation of our proofs to the case $\alpha = 1/2$.

Equipped with Theorem 3.1, we can combine the linear and spectral estimators to improve the performance in the recovery of x_1^* and x_2^*. Formally, consider the (rescaled) linear and spectral estimators $x^{\mathrm{lin}} \in \sqrt{d}\mathbb{S}^{d-1}$ and $x_1^{\mathrm{spec}}, x_2^{\mathrm{spec}} \in \sqrt{d}\mathbb{S}^{d-1}$. Define

(3.10)
$$X^{\mathrm{lin}} := \rho_1^{\mathrm{lin}} X_1 + \rho_2^{\mathrm{lin}} X_2 + W^{\mathrm{lin}}, \quad X_1^{\mathrm{spec}} := \rho_1^{\mathrm{spec}} X_1 + W_1^{\mathrm{spec}}, \quad X_2^{\mathrm{spec}} := \rho_2^{\mathrm{spec}} X_2 + W_2^{\mathrm{spec}}.$$

Theorem 3.1 states that the joint empirical distribution of the estimators $(x^{\mathrm{lin}}, x_1^{\mathrm{spec}}, x_2^{\mathrm{spec}})$ converges to the law of $(X^{\mathrm{lin}}, X_1^{\mathrm{spec}}, X_2^{\mathrm{spec}})$. For $i \in \{1, 2\}$, define the set of functions

$$\mathcal{C}_i := \left\{ C_i \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R} \text{ s.t. } \mathbb{E}\left[C_i(X^{\mathrm{lin}}, X_i^{\mathrm{spec}})^2 \right] \in (0, \infty) \right\}. \tag{3.11}$$

Then, for any $C_i \in \mathcal{C}_i$, the *combined estimator* x_i^{comb} is defined as

$$x_i^{\mathrm{comb}} := C_i(x^{\mathrm{lin}}, x_i^{\mathrm{spec}}), \tag{3.12}$$

where C_i acts on its inputs componentwise, i.e., $x_{i,j}^{\mathrm{comb}} = C_i(x_j^{\mathrm{lin}}, x_{i,j}^{\mathrm{spec}})$ for any $j \in [d]$. Now, (3.7) reduces the vector problem of estimating x_i^* given $(x^{\mathrm{lin}}, x_i^{\mathrm{spec}})$ to the scalar problem of estimating X_i from X^{lin} and X_i^{spec}. The Bayes-optimal combined estimator that minimizes the expected squared error for this scalar problem is $\mathbb{E}[X_i \,|\, X^{\mathrm{lin}}, X_i^{\mathrm{spec}}]$. Recalling from Theorem 3.1 that $(X_i, X^{\mathrm{lin}}, X_i^{\mathrm{spec}})$ are jointly Gaussian, the Bayes-optimal combined estimator is a linear combination of $(X^{\mathrm{lin}}, X_i^{\mathrm{spec}})$. The performance of this combined estimator is formalized in the following corollary, whose proof is given in section SM5.

Corollary 3.7 (Bayes-optimal linear-spectral combination). *Consider the setting of Theorem 3.1. For $i \in \{1, 2\}$, define $C_i^* \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ as follows:*

$$C_i^*(X^{\mathrm{lin}}, X_i^{\mathrm{spec}}) := \mathbb{E}\left[X_i \,|\, X^{\mathrm{lin}}, X_i^{\mathrm{spec}} \right] = \frac{1}{1 - \nu_i^2} \left(\xi_i X^{\mathrm{lin}} + \zeta_i X_i^{\mathrm{spec}} \right), \tag{3.13}$$

where

$$\nu_i := \rho_i^{\mathrm{lin}} \rho_i^{\mathrm{spec}} + \mathbb{E}\left[W^{\mathrm{lin}} W_i^{\mathrm{spec}} \right], \quad \xi_i := \rho_i^{\mathrm{lin}} - \rho_i^{\mathrm{spec}} \nu_i, \quad \zeta_i := \rho_i^{\mathrm{spec}} - \rho_i^{\mathrm{lin}} \nu_i.$$

For $i \in \{1,2\}$, let x_i^{comb} be the combined estimators defined in (3.12) w.r.t. C_i^*, respectively. Then, almost surely we have

$$\lim_{d\to\infty} \frac{\left|\langle x_i^{\mathrm{comb}}, x_i^*\rangle\right|}{\left\|x_i^{\mathrm{comb}}\right\|_2 \|x_i^*\|_2} = \frac{1}{1-\nu_i^2}\left(\xi_i^2+\zeta_i^2+2\xi_i\zeta_i\left(\rho_i^{\mathrm{lin}}\rho_i^{\mathrm{spec}}+\mathbb{E}\left[W^{\mathrm{lin}}W_i^{\mathrm{spec}}\right]\right)\right)^{1/2} =: \mathsf{OL}_i^{\mathrm{comb}}.$$

Furthermore, for any $(C_1,C_2)\in\mathcal{C}_1\times\mathcal{C}_2$, the corresponding combined estimators $\widetilde{x}_1^{\mathrm{comb}},\widetilde{x}_2^{\mathrm{comb}}$ defined w.r.t. C_1,C_2 through (3.12) satisfy

$$\lim_{d\to\infty} \frac{\left|\langle \widetilde{x}_i^{\mathrm{comb}}, x_i^*\rangle\right|}{\left\|\widetilde{x}_i^{\mathrm{comb}}\right\|_2 \|x_i^*\|_2} = \frac{\left|\mathbb{E}\left[X_i C_i(X^{\mathrm{lin}}, X_i^{\mathrm{spec}})\right]\right|}{\sqrt{\mathbb{E}\left[C_i(X^{\mathrm{lin}}, X_i^{\mathrm{spec}})^2\right]}} \le \mathsf{OL}_i^{\mathrm{comb}}, \qquad i\in\{1,2\}.$$

3.1. Linear estimator. Theorem 3.1 allows us to derive the asymptotic overlap of each signal with the linear estimator in (2.3).

Corollary 3.8 (overlaps, linear). *Consider the setting of section 2, and let assumptions (A1)–(A6) hold. Then, almost surely,*

$$\lim_{d\to\infty} \frac{\langle \widehat{x}^{\mathrm{lin}}, x_i^*\rangle}{\|\widehat{x}^{\mathrm{lin}}\|_2 \|x_i^*\|_2} = \rho_i^{\mathrm{lin}}, \quad i\in\{1,2\}. \tag{3.14}$$

Proof. Choose $\Psi(a,b,c)=ab$, and note that $\Psi\in \mathrm{PL}(2)$. Then, as $\|\widehat{x}^{\mathrm{lin}}\|_2 = \|\overline{x}_i^*\|_2 = \sqrt{d}$, the left-hand sides of (3.7) and (3.8) recover the overlaps in (3.14) for $i=1,2$, and the right-hand sides of (3.7) and (3.8) become $\rho_1^{\mathrm{lin}},\rho_2^{\mathrm{lin}}$ (defined in (3.5b)). ∎

Remark 3.9 (overlap of linear estimator does not approach 1). From (3.14) and the definitions of $\rho_1^{\mathrm{lin}},\rho_2^{\mathrm{lin}}$ in (3.5b), we have that the linear estimator achieves positive overlap with each signal for *any* positive δ, as long as $\mathbb{E}[G\mathcal{L}(Y)]>0$. As $\delta\to\infty$, the limiting overlaps approach $\sqrt{\frac{\alpha^2}{\alpha^2+(1-\alpha)^2}}$ and $\sqrt{\frac{(1-\alpha)^2}{\alpha^2+(1-\alpha)^2}}$, and they are strictly less than 1 for any $\alpha\in(1/2,1)$. In contrast, the overlap of the spectral estimator becomes positive only when δ exceeds a certain threshold (see Remark 3.16). However, once this threshold is exceeded, the spectral estimator yields overlaps approaching 1 as δ grows (see Remark 3.17). We also note that beyond the spectral threshold, the Bayes-optimal combination of the linear and spectral estimators has a larger overlap than either of the individual estimators (see Figure 1).

Using the limiting overlap of a linear estimator in Corollary 3.8, we can optimize the performance over the choice of $\mathcal{L}$ (subject to assumption (A6)). Let

$$\mathcal{I} := \{\mathcal{L}\colon \mathbb{R}\to\mathbb{R} \text{ Lipschitz s.t. } \mathbb{E}\left[G\mathcal{L}(Y)\right]\neq 0,\ \mathbb{E}\left[|G\mathcal{L}(Y)|\right]<\infty\} \tag{3.15}$$

be the set of functions $\mathcal{L}$ satisfying assumption (A6). For $i\in\{1,2\}$ and $\delta\in(0,\infty)$, define the *optimal overlaps* among linear estimators as

$$\mathsf{OL}_i^{\mathrm{lin}} := \sup_{\mathcal{L}\in\mathcal{I}} \rho_i^{\mathrm{lin}}.$$

Furthermore, if $\mathcal{I}=\emptyset$, we set $\mathsf{OL}_1^{\mathrm{lin}}=\mathsf{OL}_2^{\mathrm{lin}}=0$. In words, $\mathsf{OL}_i^{\mathrm{lin}}$ ($i\in\{1,2\}$) is the largest overlap with the ith signal that can be achieved by a linear estimator. Then, we have the following characterization of the optimal overlaps. The proof is contained in section SM6.

Proposition 3.10 (optimal linear estimator). *Consider the setting of section 2, and let assumptions (A1)– (A5) hold. Assume further that*

$$\int_{\mathrm{supp}(Y)} \frac{\mathbb{E}\left[Gp(y|G)\right]^2}{\mathbb{E}\left[p(y|G)\right]}\mathrm{d}y \in (0,\infty), \tag{3.16}$$

where $p(y|g)$ is the conditional law in (2.2) and the expectation is taken w.r.t. $G \sim \mathcal{N}(0,1)$. Then, for any $\delta \in (0,\infty)$, writing $\alpha_1 := \alpha$ and $\alpha_2 := (1-\alpha)$, we have

$$\mathsf{OL}_i^{\mathrm{lin}} = \left(\frac{\alpha_1^2+\alpha_2^2}{\alpha_i^2} + \frac{1}{\alpha_i^2\delta}\cdot\frac{1}{\int_{\mathrm{supp}(Y)}\frac{\mathbb{E}[Gp(y|G)]^2}{\mathbb{E}[p(y|G)]}\mathrm{d}y}\right)^{-1/2}, \qquad i\in\{1,2\}. \tag{3.17}$$

Moreover, define $\mathcal{L}^*\colon \mathbb{R}\to\mathbb{R}$ as

$$\mathcal{L}^*(y) = \frac{\mathbb{E}\left[Gp(y|G)\right]}{\mathbb{E}\left[p(y|G)\right]}.$$

Then, $\mathcal{L}^* \in \mathcal{I}$ and for any $\delta\in(0,\infty)$, both $\mathsf{OL}_1^{\mathrm{lin}}, \mathsf{OL}_2^{\mathrm{lin}}$ are simultaneously achieved by $\mathcal{L}^*$.

Remark 3.11 (when linear estimator is ineffective). Equation (3.16) ensures that the linear estimator asymptotically achieves strictly positive overlap with the signals. In fact, if

$$\int_{\mathrm{supp}(Y)} \frac{\mathbb{E}\left[Gp(y|G)\right]^2}{\mathbb{E}\left[p(y|G)\right]}\mathrm{d}y = 0,$$

then, from the right-hand side of (3.17), we obtain that $\mathsf{OL}_1^{\mathrm{lin}} = \mathsf{OL}_2^{\mathrm{lin}} = 0$ for any $\delta\in(0,\infty)$. This is the case for mixed phase retrieval, as mentioned in section SM1. We note that the condition in (3.16) also appears in the nonmixed setting (see Appendix C.1 of [59]).

3.2. Spectral estimator. The limiting value of the overlaps for the spectral estimator can be obtained similarly to Corollary 3.8.

Corollary 3.12 (overlaps, spectral). *Consider the setting of section 2, and let assumptions (A1)–(A5) and (A7) hold. Then, for $i\in\{1,2\}$, if $\lambda^*(\delta_i) > \overline{\lambda}(\delta)$, we have that, almost surely,*

$$\lim_{d\to\infty} \frac{|\langle v_i(D), x_i^*\rangle|}{\|v_i(D)\|_2\,\|x_i^*\|_2} = \rho_i^{\mathrm{spec}}. \tag{3.18}$$

Remark 3.13 (condition for vanishing overlap). We focus here on the recovery of the first signal, and an analogous discussion is valid for the second one. As $\lambda^*(\delta_1)$ approaches $\overline{\lambda}(\delta)$ from above, the right-hand side of (3.18) tends to 0. Indeed, as $\lambda^*(\delta_1)\searrow\overline{\lambda}(\delta)$, one can readily verify that $\mathbb{E}[(\frac{Z}{\lambda^*(\delta_1)-Z})^2]\nearrow\frac{1}{\delta}$ and consequently the numerator of ρ_1^{spec} (cf. (3.5c)) decreases to 0. Furthermore, in the nonmixed setting ($\alpha=1$), the analysis of [51, 58] gives that, when $\lambda^*(\delta) < \overline{\lambda}(\delta)$, the corresponding overlap vanishes. While we do not formally prove that the condition $\lambda^*(\delta_1) > \overline{\lambda}(\delta)$ is *necessary* for the spectral method to have nonvanishing overlap, these two observations point strongly in that direction. A third piece of supporting evidence is provided in Remark SM3.2.

Equipped with Corollary 3.12, we can optimize both (i) the spectral threshold, namely, the minimum value of δ needed to satisfy the condition $\lambda^*(\delta_1) > \overline{\lambda}(\delta)$ which gives a strictly positive overlap, and (ii) the limiting overlap given by the right-hand side of (3.18). Formally, for $i \in \{1,2\}$ and $\delta \in (0,\infty)$, let

$$\mathcal{H}_i := \left\{ \mathcal{T}\colon \mathbb{R} \to \mathbb{R} \text{ Lipschitz s.t. } \begin{array}{c} \inf_{y\in\mathrm{supp}(Y)} \mathcal{T}(y) > -\infty,\ 0 < \sup_{y\in\mathrm{supp}(Y)} \mathcal{T}(y) < \infty, \\ \Pr\left[\mathcal{T}(Y) = 0\right] < 1,\ \lambda^*(\delta_i) > \overline{\lambda}(\delta) \end{array} \right\} \tag{3.19}$$

be the set of functions $\mathcal{T}$ satisfying assumption (A7) such that $\lambda^*(\delta_i) > \overline{\lambda}(\delta)$ holds. We recall that $\delta_1 = \alpha\delta, \delta_2 = (1-\alpha)\delta$ and $\lambda^*(\cdot), \overline{\lambda}(\cdot)$ depend on the choice of the preprocessing function. Noting that $\mathcal{H}_i$ depends on δ, we can define the *spectral threshold* for the ith signal as

$$\delta_i^{\mathrm{spec}} := \inf\left\{\delta \in (0,\infty) : \mathcal{H}_i \neq \emptyset\right\}, \quad i \in \{1,2\}.$$

In words, this is the smallest δ such that there exists a preprocessing function satisfying $\lambda^*(\delta_i) > \overline{\lambda}(\delta)$ (and, hence, leading to nonvanishing limiting overlap). Furthermore, for $i \in \{1,2\}$ and $\delta > \delta_i^{\mathrm{spec}}$, define the *optimal overlap* as

$$\mathsf{OL}_i^{\mathrm{spec}} := \sup_{\mathcal{T}\in\mathcal{H}_i} \rho_i^{\mathrm{spec}}.$$

In words, for a given $\delta > \delta_i^{\mathrm{spec}}$, $\mathsf{OL}_i^{\mathrm{spec}}$ is the largest overlap with preprocessing functions that satisfy $\lambda^*(\delta_i) > \overline{\lambda}(\delta)$. We note that the supremum is guaranteed to be over a nonempty set as $\delta > \delta_i^{\mathrm{spec}}$. At this point, we can state the following result whose proof is given in section SM7.

Proposition 3.14 (optimal spectral estimator). *Consider the setting of section 2, and let assumptions (A1)–(A5) hold. Let $\alpha_1 := \alpha$ and $\alpha_2 := (1-\alpha)$. Then, for $i \in \{1,2\}$ we have*

$$\delta_i^{\mathrm{spec}} = \frac{1}{\alpha_i^2 \int_{\mathrm{supp}(Y)} \frac{\mathbb{E}[p(y|G)(G^2-1)]^2}{\mathbb{E}[p(y|G)]} \mathrm{d}y}, \tag{3.20}$$

and for $\delta > \delta_i^{\mathrm{spec}}$,

$$\mathsf{OL}_i^{\mathrm{spec}} = \frac{1}{\sqrt{\beta_i^*(\delta,\alpha) + \alpha_i}}, \tag{3.21}$$

where $\beta_i^*(\delta,\alpha) \in (1-\alpha_i, \infty)$ are the unique solutions to the following pair of fixed point equations:

$$(\beta_i^*(\delta,\alpha) - (1-\alpha_i)) \int_{\mathrm{supp}(Y)} \frac{\mathbb{E}\left[p(y|G)(G^2-1)\right]^2}{\alpha_i \mathbb{E}\left[p(y|G)G^2\right] + \beta_i^*(\delta,\alpha)\mathbb{E}\left[p(y|G)\right]} \mathrm{d}y = \frac{1}{\alpha_i^2 \delta}, \quad i \in \{1,2\}. \tag{3.22}$$

Finally, for $i \in \{1,2\}$, define $\mathcal{T}_i^*\colon \mathbb{R} \to \mathbb{R}$ as

$$\mathcal{T}_i^*(y) = 1 - \frac{1}{\alpha_i \cdot \frac{\mathbb{E}[p(y|G)G^2]}{\mathbb{E}[p(y|G)]} + (1-\alpha_i)}, \quad \text{where } G \sim \mathcal{N}(0,1). \tag{3.23}$$

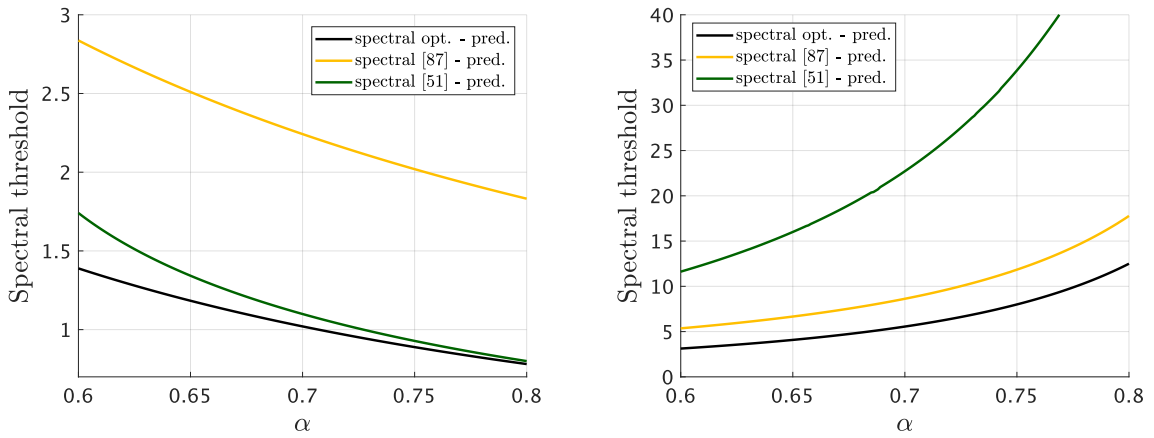
Then, for $\delta > \delta_i^{\mathrm{spec}}$, we have (i) *$\mathcal{T}_i^* \in \mathcal{H}_i$, and* (ii) *the value of $\mathsf{OL}_i^{\mathrm{spec}}$ is achieved by $\mathcal{T}_i^*$.*

Remark 3.15 (jointly optimal $\mathcal{T}, \mathcal{L}$ for C_i^*). Finding the spectral and linear estimators that jointly maximize the overlap between the optimal combination of the two and the ith signal (where $i \in \{1, 2\}$) amounts to solving the following constrained optimization problem over a pair of functions $\mathcal{T}, \mathcal{L}$:

$$\sup_{(\mathcal{T},\mathcal{L})\in\mathcal{H}_i\times\mathcal{I}} \mathsf{OL}_i^{\mathrm{comb}}. \tag{3.24}$$

In the above display, $\mathcal{H}_i$ (defined in (3.19)) is the set of spectral preprocessing functions that satisfy assumption (A7) and are effective for estimating x_i^* (i.e., $\lambda^*(\delta_i) > \overline{\lambda}(\delta)$); $\mathcal{I}$ (defined in (3.15)) is the set of linear preprocessing functions that satisfy assumption (A6) (and therefore are effective for estimating both signals); $\mathsf{OL}_i^{\mathrm{comb}}$ (defined in Corollary 3.7) is the asymptotic overlap between the optimally combined estimator (with respect to fixed $\mathcal{T}, \mathcal{L}$) and x_i^*. Equation (3.24) is an explicit yet challenging functional optimization problem that remains open. Note that in the special cases of $\alpha = 0$ or $\alpha = 1$, (3.24) reduces to an analogous optimization problem for (nonmixed) GLMs whose resolution was left open in [59, section C.4].

Remark 3.16 (universal lower bounds on spectral thresholds). In section SM8, we show that the spectral thresholds δ_1^{spec} and δ_2^{spec} are always at least $\delta_1^* := \frac{1}{2\alpha^2}$ and $\delta_2^* := \frac{1}{2(1-\alpha)^2}$ for any conditional law $p(\cdot \,|\, g)$ in (2.2) (i.e., regardless of the model). These lower bounds coincide with the spectral thresholds for both noiseless linear regression and noiseless phase retrieval; see Remark SM1.4. Thus, unlike the linear estimator, the spectral estimator (even the optimal one) does not achieve weak recovery for all $\delta > 0$; it gives positive overlaps only when the aspect ratio δ exceeds a certain value. We highlight that the threshold associated to our proposed optimal spectral estimator is significantly lower than that corresponding to spectral estimators proposed earlier in the literature [88, 52]; see Figure 3.



(a) Recovery of x_1^* (b) Recovery of x_2^*

Figure 3. *Smallest δ required by different spectral estimators to weakly recover signals for noiseless mixed phase retrieval. The spectral threshold is plotted as a function of a varying mixing parameter $\alpha \in [0.6, 0.8]$. Our optimal spectral estimator always attains the lowest threshold. We note that these thresholds remain the same for noiseless mixed linear regression, due to the design of the corresponding estimators.*

Remark 3.17 (overlap of spectral estimator approaches 1). The optimal limiting overlaps in (3.21) approach 1 as $\delta \to \infty$ provided

$$\int_{\mathrm{supp}(Y)} \frac{\mathbb{E}\left[p(y|G)(G^2-1)\right]^2}{\alpha \mathbb{E}\left[p(y|G)G^2\right] + (1-\alpha)\mathbb{E}\left[p(y|G)\right]} \mathrm{d}y \in (0,\infty). \tag{3.25}$$

To show this, consider the optimal limiting overlap between the spectral estimator and the first signal, which by (3.21) equals $\frac{1}{\sqrt{\beta_1^*(\delta,\alpha)+\alpha}}$. To show the claim, it suffices to show $\beta_1^*(\delta,\alpha) \xrightarrow{\delta\to\infty} 1-\alpha$. From (3.22), the fixed point equation defining $\beta_1^*(\infty,\alpha)$ becomes

$$(\beta_1^*(\infty,\alpha) - (1-\alpha)) \int_{\mathrm{supp}(Y)} \frac{\mathbb{E}[p(y|G)(G^2-1)]^2}{\alpha \mathbb{E}[p(y|G)G^2] + \beta_1^*(\infty,\alpha)\mathbb{E}[p(y|G)]} \mathrm{d}y = 0 \tag{3.26}$$

as $\delta \to \infty$. Since (3.25) holds, the unique solution to (3.26) has to be $\beta_1^*(\infty,\alpha) = 1-\alpha$. This proves the claim. We note that the condition (3.25) is satisfied by the mixed linear regression model.

4. Numerical experiments. The experimental results in Figures 1, 4, and 5 show that the performance of the various estimators (linear, spectral, and combined) closely match the asymptotic predictions in various settings. Furthermore, Figure 6 shows that our estimators exhibit improvements over existing spectral estimators designed for nonmixed data even when the signals have mild correlation. In all plots, the signal dimension is $d = 2000$, and the vertical and horizontal axes represent the overlap and the aspect ratio δ. The solid curves correspond to the theoretical predictions whose analytic expressions are in section SM1. Discrete points (little squares, triangles, asterisks, etc.) are computed using synthetic data. Each of these points is the mean of 10 i.i.d. trials together with error bars at 1 standard deviation. Additional comments on experimental setup and results are deferred to section SM2.

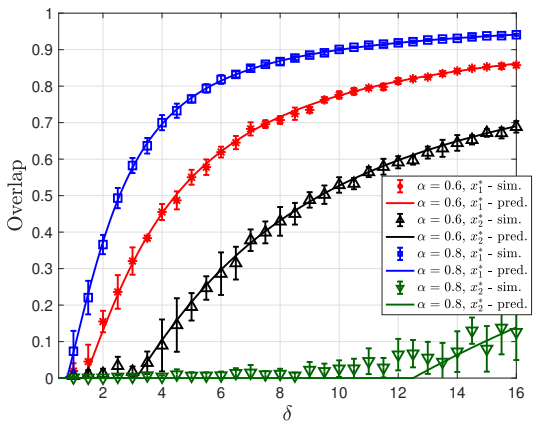
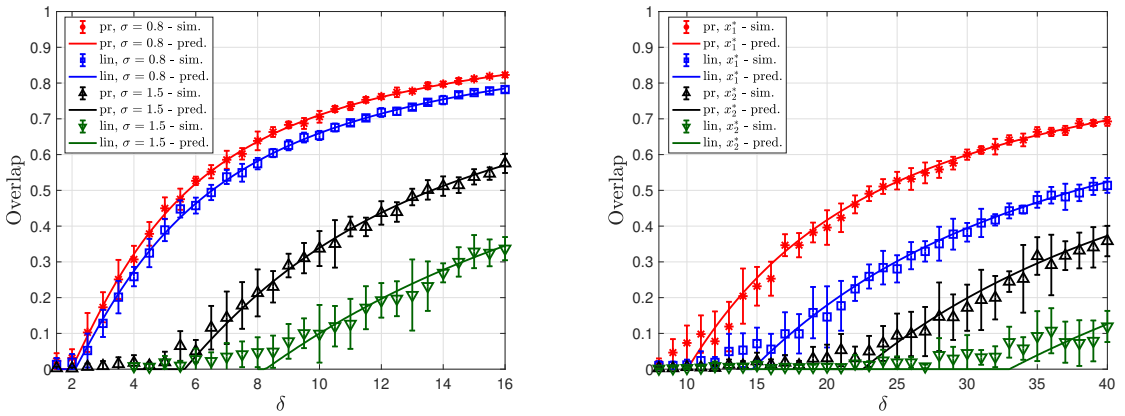
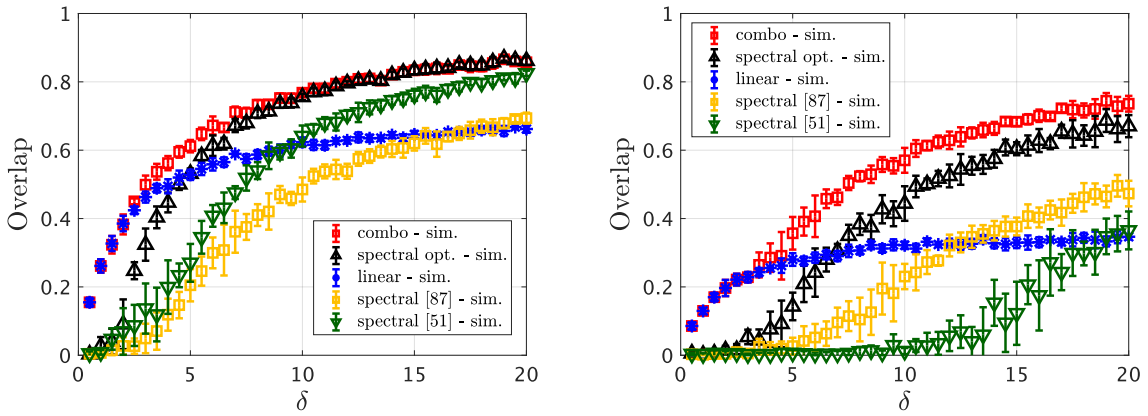


Figure 4. *Spectral estimators for noiseless mixed linear regression, with mixing parameter* $\alpha \in \{0.6, 0.8\}$. *Optimal spectral estimators given by* (SM1.7) *are used. Overlaps with both signals* x_1^*, x_2^*, *computed from simulation ("sim.") and prediction ("pred."), are plotted as a function of the aspect ratio* δ. *The same numerics apply to noiseless phase retrieval (see Remark* SM1.4*).*



(a) $\alpha = 0.8$ and $\sigma \in \{0.8, 1.5\}$. (b) $\alpha = 0.6$ and $\sigma = 1.5$.

Figure 5. *Spectral estimators for mixed linear regression and mixed phase retrieval. Optimal spectral estimators ((SM1.4) and (SM1.5)) are used. Overlaps with the first signal x_1^* (left plot) and with both signals x_1^*, x_2^* (right plot), computed from simulation ("sim.") and prediction ("pred."), are plotted as a function of the aspect ratio δ.*



(a) Recovery of x_1^* (b) Recovery of x_2^*

Figure 6. *Performance comparison for correlated signals. The setting is noiseless mixed linear regression with mixing parameter $\alpha = 0.6$ and signal correlation $\langle x_1^*, x_2^* \rangle = \rho$ where $\rho = 0.1$. Overlaps with x_1^* (left) and x_2^* (right) are plotted as a function of the aspect ratio δ. The signal dimension is $d = 2000$.*

5. Proof outline. The proof of Theorem 3.1 combines AMP with random matrix theory (RMT) tools. We now outline the high-level ideas in the analysis.

Eigenvalues via random matrix theory. The first step is to understand the spectrum of D, in particular, the right edge of the bulk and the outlier(s). This involves the following challenges:

- The matrix D in (2.5) can be thought of as an instance of *spiked matrix model*. Its structure is, however, more sophisticated than the canonical "signal plus noise" model. Indeed, the potential spikes of D result from two signals through the composition of the link function q and the spectral preprocessing function $\mathcal{T}$.

- The analysis of the limiting spectrum for nonmixed GLMs is provided in [51, 58]. In our mixed setting, applying the strategy of [51, 58] to analyze the spectrum of D results in additional matrix terms which are hard to bound.

The key idea is to decompose D into the sum of two asymptotically free random matrices, consisting of the observations corresponding to the first and the second signal. To be more specific, let us condition on $\eta_1, \cdots, \eta_n$ and assume for notational convenience that $\eta_i = 1$ for $1 \le i \le n_1$ and $\eta_i = 0$ for $n_1 + 1 \le i \le n$, for some $0 \le n_1 \le n$. Let $n_2 = n - n_1$. Note that, almost surely, $n_1/d \to \delta_1, n_2/d \to \delta_2$. Now, we can write the matrices of interest in block form:

$$A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}, \quad T = \begin{bmatrix} T_1 & 0_{n_1 \times n_2} \\ 0_{n_2 \times n_1} & T_2 \end{bmatrix}, \tag{5.1}$$

where $A_1 \in \mathbb{R}^{n_1 \times d}, A_2 \in \mathbb{R}^{n_2 \times d}$ and $T_1 \in \mathbb{R}^{n_1 \times n_1}, T_2 \in \mathbb{R}^{n_2 \times n_2}$. We also let $\underline{\varepsilon}_1 = (\varepsilon_1, \cdots, \varepsilon_{n_1})$ and $\underline{\varepsilon}_2 = (\varepsilon_{n_1+1}, \cdots, \varepsilon_n)$. Then,

$$A^\top T A = \begin{bmatrix} A_1^\top & A_2^\top \end{bmatrix} \begin{bmatrix} T_1 & 0_{n_1 \times n_2} \\ 0_{n_2 \times n_1} & T_2 \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = A_1^\top T_1 A_1 + A_2^\top T_2 A_2. \tag{5.2}$$

Note that, for $i \in \{1, 2\}$, $A_i^\top T_i A_i = A_i^\top \operatorname{diag}(\mathcal{T}(q(A_i x_i^*, \underline{\varepsilon}_i)))A_i$. Since $A_1, x_1^*, \underline{\varepsilon}_1$ and $A_2, x_2^*, \underline{\varepsilon}_2$ are mutually independent, $A_1^\top T_1 A_1$ is independent of $A_2^\top T_2 A_2$. However, A_1 and T_1 are *not* independent; neither are A_2 and T_2. When considered in isolation, $A_1^\top T_1 A_1$ and $A_2^\top T_2 A_2$ are obtained from a nonmixed GLM with aspect ratio discounted by α and $1 - \alpha$, respectively. Thanks to [51, 58], their limiting spectra are well understood. Now, the crucial observation is that $A_1^\top T_1 A_1$ and $A_2^\top T_2 A_2$ are *asymptotically free*. Indeed, let $O \sim \operatorname{Haar}(\mathbb{O}(d))$ be a matrix sampled uniformly from the orthogonal group $\mathbb{O}(d)$ and independent of everything else. Then,

$$\begin{aligned} A_1^\top T_1 A_1 + A_2^\top T_2 A_2 &= A_1^\top \operatorname{diag}(\mathcal{T}(q(A_1 x_1^*, \underline{\varepsilon}_1)))A_1 + A_2^\top \operatorname{diag}(\mathcal{T}(q(A_2 x_2^*, \underline{\varepsilon}_2)))A_2 \\ &\overset{\mathrm{d}}{=} A_1^\top \operatorname{diag}(\mathcal{T}(q(A_1 x_1^*, \underline{\varepsilon}_1)))A_1 + (A_2 O)^\top \operatorname{diag}(\mathcal{T}(q((A_2 O) x_2^*, \underline{\varepsilon}_2)))(A_2 O) \qquad (5.3) \\ &= A_1^\top \operatorname{diag}(\mathcal{T}(q(A_1 x_1^*, \underline{\varepsilon}_1)))A_1 + O^\top A_2^\top \operatorname{diag}(\mathcal{T}(q(A_2 (O x_2^*), \underline{\varepsilon}_2)))A_2 O \\ &\overset{\mathrm{d}}{=} A_1^\top \operatorname{diag}(\mathcal{T}(q(A_1 x_1^*, \underline{\varepsilon}_1)))A_1 + O^\top A_2^\top \operatorname{diag}(\mathcal{T}(q(A_2 x_2^*, \underline{\varepsilon}_2)))A_2 O \qquad (5.4) \\ &= A_1^\top T_1 A_1 + O^\top A_2^\top T_2 A_2 O. \qquad (5.5) \end{aligned}$$

Equation (5.3) follows from the independence of A_1, A_2 and from the rotational invariance of isotropic Gaussians. Equation (5.4) follows since O and $O x_2^*$ are independent if $O \sim \operatorname{Haar}(\mathbb{O}(d))$ and $x_2^* \sim \operatorname{Unif}(\mathbb{S}^{d-1})$. In this step, we crucially use the assumption that x_1^* and x_2^* are independent and each uniformly distributed over $\mathbb{S}^{d-1}$.

The asymptotic freeness shown in (5.5) allows us to study the (free) sum of $A_1^\top T_1 A_1$ and $A_2^\top T_2 A_2$ using the tools developed in [9]. Indeed, the analysis carried out in subsections SM3.1 and SM3.2 implies the following characterization of the top three limiting eigenvalues of D (see Theorem SM3.1):

$$\lim_{d \to \infty} \lambda_1(D) = \zeta(\lambda^*(\delta_1); \delta), \ \lim_{d \to \infty} \lambda_2(D) = \zeta(\lambda^*(\delta_2); \delta), \ \lim_{d \to \infty} \lambda_3(D) = \zeta(\overline{\lambda}(\delta); \delta). \tag{5.6}$$

Here, it is helpful to recall the definitions of $\zeta(\cdot;\cdot)$ (see (3.4)), $\lambda^*(\cdot)$ (see the text following (3.4)), and $\overline{\lambda}(\cdot)$ (see (3.3)). Moreover, using the convexity of the function $\zeta(\cdot;\delta)$, it can be shown that for $i \in \{1,2\}$, $\lambda_i(D)$ is strictly larger than $\lambda_3(D)$ in the high-dimensional limit if $\lambda^*(\delta_i) > \overline{\lambda}(\delta)$, meaning that $\lambda_i(D)$ is detached from the bulk spectrum of D and becomes an outlier eigenvalue. Therefore, D exhibits a spectral gap between the ith eigenvalue and the right edge of the bulk. In that case, the limiting eigenvalues admit the more explicit expressions reported in Remark SM3.4. The existence of a spectral gap will be crucially used in proving the convergence of GAMP iterates to spectral estimators, as discussed below.

Joint distribution via GAMP. The convergence results in (3.7) and (3.8) are obtained using a generalized approximate message passing (GAMP) algorithm [68]. In a mixed GLM, since the observations $(y_i)_{i\in[n]}$ are unlabeled (i.e., it is unknown to the estimator whether each y_i is generated from the first or the second signal), estimating *both* signals is more challenging than estimating each one from an individual nonmixed GLM. However, the existing state evolution result for GAMP [68], [36, sect. 4] is derived for a nonmixed model, and only keeps track of the effect of a single signal. We generalize the GAMP state evolution result to mixed GLMs (see Proposition SM4.1), so that the state evolution recursion tracks the effect of both signals. For convenience, let us work with the following rescalings:

$$\overline{A} := \frac{1}{\sqrt{d}} A, \quad \overline{x}_1^* := \sqrt{d}\, x_1^*, \quad \overline{x}_2^* := \sqrt{d}\, x_2^*, \quad \overline{D} := \overline{A}^\top T \overline{A} = \frac{n}{d} A^\top T A. \tag{5.7}$$

Given two sequences of *denoising functions* $f_{t+1}\colon \mathbb{R}^3 \to \mathbb{R}, g_t\colon \mathbb{R}^2 \to \mathbb{R}$ (for each iteration $t \geq 0$), GAMP maintains a pair of iterates $u^t \in \mathbb{R}^n, v^{t+1} \in \mathbb{R}^d$ according to

$$\begin{aligned} u^t &= \frac{1}{\sqrt{\delta}} \overline{A} \tilde{v}^t - \mathsf{b}_t \tilde{u}^{t-1}, \quad \tilde{u}^t = g_t(u^t; y), \\ v^{t+1} &= \frac{1}{\sqrt{\delta}} \overline{A}^\top \tilde{u}^t - \mathsf{c}_t \tilde{v}^t, \quad \tilde{v}^{t+1} = f_{t+1}(v^{t+1}; \overline{x}_1^*, \overline{x}_2^*), \end{aligned} \tag{5.8}$$

where f_{t+1}, g_t are applied componentwise, i.e., $f_{t+1}(v^{t+1}; \overline{x}_1^*, \overline{x}_2^*) = (f_{t+1}(v_1^{t+1}; \overline{x}_{1,1}^*, \overline{x}_{2,1}^*), \ldots, f_{t+1}(v_d^{t+1}; \overline{x}_{1,d}^*, \overline{x}_{2,d}^*))$, $g_t(u^t; y) = (g_t(u_1^t; y_1), \ldots, g_t(u_n^t; y_n))$. The scalars $\mathsf{b}_t, \mathsf{c}_t$ are defined as

$$\mathsf{b}_t = \frac{1}{n} \sum_{i=1}^{d} f_t'(v_i^t; \overline{x}_{1,i}^*, \overline{x}_{2,i}^*), \qquad \mathsf{c}_t = \frac{1}{n} \sum_{i=1}^{n} g_t'(u_i^t; y_i),$$

where f_t' and g_t' each denote the derivative with respect to the first argument. The iteration is initialized with a given $\tilde{v}^0 \in \mathbb{R}^d$ and $\tilde{u}^{-1} = 0_n$. Under the assumption that the design matrix is Gaussian (indeed $\overline{A}_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1/d)$ according to assumption (A4)), the joint empirical distribution of u^t, v^{t+1} converges (as $n, d \to \infty$ with $n/d \to \delta$) to the law of a pair of jointly Gaussian random variables U_t, V_{t+1}:

$$U_t := \mu_{1,t} G_1 + \mu_{2,t} G_2 + W_{U,t}, \quad V_{t+1} := \chi_{1,t+1} X_1 + \chi_{2,t+1} X_2 + W_{V,t+1},$$

where $(G_1, G_2, W_{U,t}) \sim \mathcal{N}(0,1) \otimes \mathcal{N}(0,1) \otimes \mathcal{N}(0, \sigma_{U,t}^2)$ and $(X_1, X_2, W_{V,t+1}) \sim \mathcal{N}(0,1) \otimes \mathcal{N}(0,1) \otimes \mathcal{N}(0, \sigma_{V,t+1}^2)$. The covariance structure of these jointly Gaussian random variables is described by a set of recursions called *state evolution*:

$$\mu_{1,t} = \frac{1}{\sqrt{\delta}}\mathbb{E}[X_1 f_t(V_t; X_1, X_2)], \quad \mu_{2,t} = \frac{1}{\sqrt{\delta}}\mathbb{E}[X_2 f_t(V_t; X_1, X_2)],$$
$$\sigma^2_{U,t} = \frac{1}{\delta}\mathbb{E}[f_t(V_t; X_1, X_2)^2] - \mu^2_{1,t} - \mu^2_{2,t},$$
$$\chi_{1,t+1} = \sqrt{\delta}\left(\mathbb{E}[G_1 g_t(U_t; \tilde{Y})] - \mathbb{E}[g_t'(U_t; \tilde{Y})]\mu_{1,t}\right),$$
$$\chi_{2,t+1} = \sqrt{\delta}\left(\mathbb{E}[G_2 g_t(U_t; \tilde{Y})] - \mathbb{E}[g_t'(U_t; \tilde{Y})]\mu_{2,t}\right),$$
$$\sigma^2_{V,t+1} = \mathbb{E}[g_t(U_t; \tilde{Y})^2],$$

where the random variable $\tilde{Y}$ is given by

$$\tilde{Y} = q(\eta G_1 + (1-\eta)G_2, \varepsilon), \text{ with } (G_1, G_2, \eta, \varepsilon) \sim \mathcal{N}(0,1) \otimes \mathcal{N}(0,1) \otimes \mathrm{Bern}(\alpha) \otimes P_\varepsilon.$$

The recursion is initialized as

$$\mu_{1,0} = \frac{1}{\sqrt{\delta}} \lim_{d\to\infty} \frac{\langle \overline{x}_1^*, \tilde{v}^0\rangle}{d}, \quad \mu_{2,0} = \frac{1}{\sqrt{\delta}} \lim_{d\to\infty} \frac{\langle \overline{x}_2^*, \tilde{v}^0\rangle}{d}, \quad \sigma^2_{U,0} = \frac{1}{\delta} \lim_{d\to\infty} \frac{\|\tilde{v}^0\|_2^2}{d} - \mu^2_{1,0} - \mu^2_{2,0}.$$

The proof of convergence of the empirical distributions of u^t, v^{t+1} to the laws of U_t, V_{t+1}, given in section SM9, uses a reduction to an abstract AMP recursion with matrix-valued iterates for which a state evolution result was established in [39, 36]. For details, see the formal statements in Proposition SM4.1 which track the joint empirical distribution of *all* iterates.

At this point, the linear estimator is readily obtained via the iterate of GAMP run for one step ($t = 0$). For $t \geq 1$, we tailor the denoisers $(f_{t+1}, g_t)_{t\geq 1}$ so that the iterates of GAMP implement a power method, which for large enough t, gives the first and second eigenvectors of the spectral matrix D (defined in (2.5)). Specifically, consider the GAMP iteration in (5.8) with the initializer $\tilde{v}^0 = 0_d$, and the following choice of denoisers:

$$\begin{aligned} &g_0(u^0; y) = \sqrt{\delta}\mathcal{L}(y), \qquad f_1(v; \overline{x}_1^*, \overline{x}_2^*) = f(\overline{x}_1^*, \overline{x}_2^*), \\ &g_t(u; y) = \sqrt{\delta}\, u\, \mathcal{F}(y), \quad f_{t+1}(v; \overline{x}_1^*, \overline{x}_2^*) = \frac{v}{\beta_{t+1}}, \quad t \geq 1, \end{aligned} \tag{5.9}$$

where $\mathcal{F}: \mathbb{R} \to \mathbb{R}$ is bounded and Lipschitz, $f: \mathbb{R}^2 \to \mathbb{R}$ is Lipschitz, and

$$\beta_{t+1} := \sqrt{\chi^2_{1,t+1} + \chi^2_{2,t+1} + \sigma^2_{V,t+1}}.$$

To prove Theorem 3.1, we select two pairs of functions $(f, \mathcal{F})$, in terms of the spectral preprocessing function $\mathcal{T}$ (see (5.13) and (5.14)). With the above choice of $(f_{t+1})_{t\geq 1}, (g_t)_{t\geq 1}$, the GAMP iteration becomes

$$\begin{aligned} &u^0 = 0_n, \quad v^1 = \overline{A}^\top \mathcal{L}(y), \\ &u^t = \frac{1}{\sqrt{\delta}\,\beta_t}\left(\overline{A}v^t - F u^{t-1}\right), \quad v^{t+1} = \overline{A}^\top F u^t - \frac{\sqrt{\delta}}{\beta_t}\mathbb{E}[\mathcal{F}(\tilde{Y})]\, v^t, \qquad t \geq 2, \end{aligned} \tag{5.10}$$

where $F = \mathrm{diag}(\mathcal{F}(y_1), \ldots, \mathcal{F}(y_n))$. First, note that the iterate v^1 coincides with the linear estimator $\widehat{x}^{\mathrm{lin}}$ in (2.3). Furthermore, we show that in the high-dimensional limit, as $t \to \infty$, the iterate v^t is aligned with an eigenvector of the matrix

$$\overline{A}^\top F(\sqrt{\delta}\beta_\infty I_n + F)^{-1}\overline{A}, \tag{5.11}$$

where $\beta_\infty = \lim_{t\to\infty}\beta_t$. To justify the claim, assume the iterates u^t, v^{t+1} converge to the limits u^∞, v^∞ in the sense that $\lim_{t\to\infty}\lim_{d\to\infty}\frac{1}{d}\|u^t - u^\infty\|_2^2 = 0$ and $\lim_{t\to\infty}\lim_{d\to\infty}\frac{1}{d}\|v^t - v^\infty\|_2^2 = 0$. Then, from (5.10) we can derive

$$v^\infty\left(1 + \frac{\sqrt{\delta}}{\beta_\infty}\mathbb{E}[\mathcal{F}(\tilde{Y})]\right) = \overline{A}^\top F(\sqrt{\delta}\beta_\infty I_n + F)^{-1}\overline{A}v^\infty. \tag{5.12}$$

Therefore, v^∞ is an eigenvector of the matrix in (5.11), and the GAMP iteration of (5.10) is effectively a power method.

Recall that our goal is to obtain via GAMP the two leading eigenvectors of $\overline{A}^\top T\overline{A}$. Hence, we pick $\mathcal{F}$ so that $F(\sqrt{\delta}\beta_\infty I_n + F)^{-1} = cT$ for some constant c. To this end, we analyze the iteration in (5.10) with two choices for the function $\mathcal{F}(y)$ and initialization $\tilde{v}^0$:

$$\text{Choice 1}: \quad \mathcal{F}_1(y) := \frac{\mathcal{T}(y)}{\lambda^*(\delta_1) - \mathcal{T}(y)}, \quad f(\overline{x}_1^*, \overline{x}_2^*) = \overline{x}_1^*, \tag{5.13}$$

$$\text{Choice 2}: \quad \mathcal{F}_2(y) := \frac{\mathcal{T}(y)}{\lambda^*(\delta_2) - \mathcal{T}(y)}, \quad f(\overline{x}_1^*, \overline{x}_2^*) = \overline{x}_2^*, \tag{5.14}$$

where for $i \in \{1,2\}$, $\lambda^*(\delta_i)$ is the unique solution of $\zeta(\lambda;\delta_i) = \varphi(\lambda)$ (see the text following (3.4)). The above two choices are motivated by the characterization of the limiting eigenvalues in (5.6) and Remark SM3.4. As outlined below, choice 1 (resp., choice 2) ensures that (5.12) becomes an eigenequation for the first (resp., second) eigenvalue of $\overline{D}$.

Lemma SM4.3 shows that the state evolution parameters $(\chi_{1,t}, \chi_{2,t}, \sigma_{V,t}^2)$ for choice 1 satisfy

$$\lim_{t\to\infty}\chi_{1,t} = \frac{\rho_1^{\text{spec}}}{\sqrt{\delta}}, \quad \lim_{t\to\infty}\sigma_{V,t}^2 = \frac{1 - (\rho_1^{\text{spec}})^2}{\delta}, \quad \text{and} \quad \chi_{2,t} = 0 \ \ \forall t \geq 2,$$

where the quantity ρ_1^{spec} was defined in (3.5c). Hence,

$$\beta_\infty = \lim_{t\to\infty}\sqrt{\chi_{1,t}^2 + \chi_{2,t}^2 + \sigma_{V,t}^2} = \frac{1}{\sqrt{\delta}},$$

and (5.12) becomes

$$v^\infty\left(1 + \delta\mathbb{E}\left[\frac{\mathcal{T}(Y)}{\lambda^*(\delta_1) - \mathcal{T}(Y)}\right]\right) = \frac{1}{\lambda^*(\delta_1)}\overline{A}^\top T\overline{A}v^\infty. \tag{5.15}$$

With choice 1, (5.15) gives that the GAMP iterate converges to an eigenvector of $\overline{D} = \overline{A}^\top T\overline{A}$ corresponding to the eigenvalue $\lambda^*(\delta_1)(1 + \delta\mathbb{E}[\frac{\mathcal{T}(Y)}{\lambda^*(\delta_1) - \mathcal{T}(Y)}])$. Similarly, with choice 2, the GAMP iterate converges to an eigenvector of $\overline{D}$ corresponding to the eigenvalue $\lambda^*(\delta_2)(1 + \delta\mathbb{E}[\frac{\mathcal{T}(Y)}{\lambda^*(\delta_2) - \mathcal{T}(Y)}])$. These claims match the rigorous eigenvalue characterization in (5.6) and Remark SM3.4. At this point, note that power methods (and therefore our GAMP iterations in (5.10)) crucially require a spectral gap to converge to the desired eigenvector. This spectral gap is guaranteed precisely by (5.6) provided $\lambda^*(\delta_1) > \overline{\lambda}(\delta)$ (resp., $\lambda^*(\delta_2) > \overline{\lambda}(\delta)$), which gives that $\lambda_1(\overline{D})$ (resp., $\lambda_2(\overline{D})$) is asymptotically an outlier in the spectrum of $\overline{D}$. As a consequence, we can rigorously prove the convergence of the GAMP iterates under choice 1 (resp., choice 2) to $v_1(\overline{D})$ (resp., $v_2(\overline{D})$).

To conclude, the iterate v^1 in the GAMP iteration in (5.10) equals the linear estimator, and v^{t+1} asymptotically aligns with the spectral estimator. Since the state evolution tracks the limiting joint distribution of all iterates, the characterization in (3.7) and (3.8) follows. We stress that GAMP in our argument is used only as a tool for analysis and is not part of the estimators. The actual estimators (spectral and linear) can be computed by a combination of the following simple operations: (i) applying a componentwise nonlinearity, (ii) matrix-vector/matrix-matrix multiplication, (iii) computation of eigenvectors.

Optimal linear and spectral estimators. The master theorem (Theorem 3.1) holds for arbitrary linear and spectral preprocessing functions $\mathcal{L}, \mathcal{T}$ satisfying the stated assumptions. Specializing Theorem 3.1 to linear and spectral estimators alone and using the explicit formulas for their limiting overlaps (given in Corollaries 3.8 and 3.12), we find the optimal preprocessing functions $\mathcal{L}^*, \mathcal{T}_1^*, \mathcal{T}_2^*$ that maximize the limiting overlaps. This is done in Propositions 3.10 and 3.14 by casting the optimization problem as a variational problem and solving it explicitly.

6. Discussion.

Universality beyond the Gaussian design matrix. A natural question is whether the predictions obtained under an i.i.d. Gaussian design are valid more generally. This topic has been investigated in the random matrix theory literature [80, 30, 81, 35, 1], and a recent line of research has focused on AMP [6, 18, 26, 84, 28, 29]. We note that none of these results is directly applicable to our setting, and the problem also remains open in the nonmixed (i.e., $\alpha = 1$) setup. However, the aforementioned body of work suggests that the Gaussian predictions may hold for much more general—even "almost deterministic"—design matrices.

Mixed GLM with multiple components. We focus on the mixed GLM with two components, but our approach is well suited to handle mixed GLMs with *multiple* components. We now briefly sketch how to generalize our main Theorem 3.1. The other results (overlaps for linear, spectral, and combined estimators, and their optimization) are generalized in a similar fashion.

Let $x_1^*, \cdots, x_\ell^* \in \mathbb{R}^d$ be ℓ signal vectors, and let the observation $y = (y_1, \cdots, y_n) \in \mathbb{R}^n$ be generated as $y_i = q(\langle a_i, x_{\upsilon_i}^* \rangle, \varepsilon_i)$. The latent vector $\underline{\upsilon} = (\upsilon_1, \cdots, \upsilon_n)$ is a sequence of i.i.d. mixing variables s.t. $\Pr[\upsilon_i = j] = \alpha_j$ for all $i \in [n]$ and $j \in [\ell]$. We assume that $x_1^*, \cdots, x_\ell^*$ are i.i.d. and uniform on the unit sphere, and $1 > \alpha_1 > \alpha_2 > \cdots > \alpha_\ell > 0$ (corresponding to assumptions (A1) and (A2)). We also impose our previous assumptions (A3)–(A5), and assume that ℓ is a constant (independent of n, d). For $i \in [\ell]$, let $\delta_i := \alpha_i \delta$ and

$$n^{\mathrm{lin}} := \left(\left(\sum_{k=1}^{\ell} \alpha_k^2 \right) \mathbb{E}\left[G\mathcal{L}(Y) \right]^2 + \frac{\mathbb{E}\left[\mathcal{L}(Y)^2 \right]}{\delta} \right)^{1/2},$$

$$\rho_i^{\mathrm{lin}} := \frac{\alpha_i \mathbb{E}\left[G\mathcal{L}(Y) \right]}{n^{\mathrm{lin}}}, \qquad \rho_i^{\mathrm{spec}} := \left(\frac{\frac{1}{\delta} - \mathbb{E}\left[\left(\frac{Z}{\lambda^*(\delta_i) - Z} \right)^2 \right]}{\frac{1}{\delta} + \alpha_i \mathbb{E}\left[\left(\frac{Z}{\lambda^*(\delta_i) - Z} \right)^2 (G^2 - 1) \right]} \right)^{1/2}.$$

Here $Z = \mathcal{T}(Y)$, $Y = q(G, \varepsilon)$, and $G \sim \mathcal{N}(0,1)$, as before. Then, under the same setting of Theorem 3.1 with $\overline{x}_i^*$ and x_i^{spec} defined similarly for $i \in [\ell]$, we have that, if $\lambda^*(\delta_i) > \overline{\lambda}(\delta)$,

$$\lim_{d\to\infty} \frac{1}{d}\sum_{j=1}^{d} \Psi\left(\overline{x}_{i,j}^*, x_j^{\text{lin}}, x_{i,j}^{\text{spec}}\right) = \mathbb{E}\left[\Psi\left(X_i, \sum_{k=1}^{\ell} \rho_k^{\text{lin}} X_k + W^{\text{lin}}, \rho_i^{\text{spec}} X_i + W_i^{\text{spec}}\right)\right],$$

where $(X_1, \cdots, X_\ell) \sim \mathcal{N}(0,1)^{\otimes \ell}$, $(W^{\text{lin}}, W_i^{\text{spec}})$ is independent of $(X_1, \cdots, X_\ell)$ and is jointly Gaussian with zero mean and covariance given by

$$\mathbb{E}\left[(W^{\text{lin}})^2\right] = 1 - \sum_{k=1}^{\ell} (\rho_k^{\text{lin}})^2, \ \mathbb{E}\left[(W_i^{\text{spec}})^2\right] = 1 - (\rho_i^{\text{spec}})^2,$$
$$\mathbb{E}\left[W^{\text{lin}} W_i^{\text{spec}}\right] = \frac{\alpha_i \rho_i^{\text{spec}}}{n^{\text{lin}}} \mathbb{E}\left[\frac{G\mathcal{L}(Y)Z}{\lambda^*(\delta_i) - Z}\right].$$

The result on the eigenvalues of the spectral matrix D can be obtained by following the strategy detailed in section SM3 (and sketched in section 5). Indeed, D can be decomposed into the (asymptotically) free sum of the ℓ components associated to each of the signals, and [9, Theorem 2.1] is well equipped to characterize its top $\ell+1$ eigenvalues. To derive the limiting joint empirical law of the ith signal and the linear and spectral estimators, we can then run a GAMP algorithm similar to (5.9) with denoisers tailored for the ith signal. The condition $\lambda^*(\delta_i) > \overline{\lambda}(\delta)$ guarantees the existence of a spectral gap between the ith largest eigenvalue of D and the rest of its spectrum, which in turn is leveraged to argue the convergence of GAMP to the desired eigenvector. This yields results analogous to Theorem 3.1 with α underlying (3.7) therein replaced with α_i, for $1 \le i \le \ell$.

Lower bounds for inference in mixed GLMs. In the nonmixed setting, [58] derives an information-theoretic threshold δ^{it} such that, for $\delta < \delta^{\text{it}}$, no estimation method gives a nontrivial estimate of the signal.[2] Furthermore, for noiseless phase retrieval, $\delta^{\text{it}} = 1/2$, which matches the threshold achieved by a spectral method. In the mixed setting, denoting by $\delta_1^{\text{it}}, \delta_2^{\text{it}}$ the information-theoretic thresholds corresponding to the two signals, we have

$$\delta_1^{\text{it}} \ge \frac{1}{\alpha}\delta^{\text{it}}, \qquad \delta_2^{\text{it}} \ge \frac{1}{1-\alpha}\delta^{\text{it}}. \tag{6.1}$$

To see this, note that, if a genie reveals the values of the mixing variables $(\eta_1, \cdots, \eta_n)$, then the estimation problem given mixed data with aspect ratio δ can be decoupled into two nonmixed ones with aspect ratios $\alpha\delta$ and $(1-\alpha)\delta$. We also remark that adapting the second moment method of [58] to our mixed setting does not improve the bound in (6.1) (hence, this derivation is omitted). Following the strategy of [4]—which establishes the exact asymptotics of the minimum mean squared error and, thus, gives a tight bound in the nonmixed case—requires additional ideas beyond the scope of this paper, so it is left for future research.

As a final remark, let us contrast (6.1) with the spectral bounds mentioned in Remark 3.16, which are universal in the sense that they hold for *any* mixed GLM. In particular, we note

[2] More formally, it is proved that the minimum mean squared error achieved by the Bayes-optimal estimator coincides with the error of a trivial estimator which always outputs the all-0 vector.

that the former scales as $(1/\alpha, 1/(1-\alpha))$, while the latter scales as $(1/\alpha^2, 1/(1-\alpha)^2)$, which suggests a gap between what is achievable information-theoretically and algorithmically. The possibility of a statistical-computational tradeoff is also suggested by the fact that, for $\alpha = 1/2$ and antipodal signals $(x_1^* = -x_2^*)$, mixed linear regression reduces to phase retrieval, which is widely believed to have such a gap; see, e.g., [55, 11, 14, 2]. Closing the gap or understanding its fundamental nature remains an intriguing open question for future investigation.

REFERENCES

[1] G. W. Anderson and B. Farrell, *Asymptotically liberating sequences of random unitary matrices*, Adv. Math., 255 (2014), pp. 381–413, https://doi.org/10.1016/j.aim.2013.12.026.

[2] G. Arpino and R. Venkataramanan, *Statistical-computational tradeoffs in mixed sparse linear regression*, in Proceedings of the 36th Annual Conference on Learning Theory, PMLR 195, 2023, pp. 921–986.

[3] S. Balakrishnan, M. J. Wainwright, and B. Yu, *Statistical guarantees for the EM algorithm: From population to sample-based analysis*, Ann. Statist., 45 (2017), pp. 77–120, https://doi.org/10.1214/16-AOS1435.

[4] J. Barbier, F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová, *Optimal errors and phase transitions in high-dimensional generalized linear models*, Proc. Natl. Acad. Sci. USA, 116 (2019), pp. 5451–5460, https://doi.org/10.1073/pnas.1802705116.

[5] A. Barik and J. Honorio, *Sparse mixed linear regression with guarantees: Taming an intractable problem with invex relaxation*, in Proceedings of the 39th International Conference on Machine Learning, PMLR 162, 2022, pp. 1627–1646.

[6] M. Bayati, M. Lelarge, and A. Montanari, *Universality in polytope phase transitions and message passing algorithms*, Ann. Appl. Probab., 25 (2015), pp. 753–822, https://doi.org/10.1214/14-AAP1010.

[7] M. Bayati and A. Montanari, *The dynamics of message passing on dense graphs, with applications to compressed sensing*, IEEE Trans. Inform. Theory, 57 (2011), pp. 764–785, https://doi.org/10.1109/TIT.2010.2094817.

[8] M. Bayati and A. Montanari, *The LASSO risk for Gaussian matrices*, IEEE Trans. Inform. Theory, 58 (2012), pp. 1997–2017, https://doi.org/10.1109/TIT.2011.2174612.

[9] S. T. Belinschi, H. Bercovici, M. Capitaine, and M. Février, *Outliers in the spectrum of large deformed unitarily invariant models*, Ann. Probab., 45 (2017), pp. 3571–3625, https://doi.org/10.1214/16-AOP1144.

[10] P. T. Boufounos and R. G. Baraniuk, 1-*bit compressive sensing*, in Conference on Information Sciences and Systems (CISS), 2008, pp. 16–21.

[11] M. Brennan and G. Bresler, *Reducibility and statistical-computational gaps from secret leakage*, in Proceedings of the 33rd Annual Conference on Learning Theory, PMLR 125, 2020, pp. 648–847.

[12] Z. Bu, J. M. Klusowski, C. Rush, and W. J. Su, *Algorithmic analysis and statistical estimation of SLOPE via approximate message passing*, IEEE Trans. Inform. Theory, 67 (2021), pp. 506–537, https://doi.org/10.1109/TIT.2020.3025272.

[13] E. J. Candès, T. Strohmer, and V. Voroninski, *PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming*, Commun. Pure Appl. Math., 66 (2013), pp. 1241–1274, https://doi.org/10.1002/cpa.21432.

[14] M. Celentano, A. Montanari, and Y. Wu, *The estimation error of general first order methods*, in Proceedings of the 33rd Annual Conference on Learning Theory, PMLR 125, 2020, pp. 1078–1141.

[15] A. T. Chaganty and P. Liang, *Spectral experts for estimating mixtures of linear regressions*, in Proceedings of the 30th International Conference on Machine Learning, PMLR 28, 2013, pp. 1040–1048.

[16] K. A. Chandrasekher, A. Pananjady, and C. Thrampoulidis, *Sharp global convergence guarantees for iterative nonconvex optimization with random data*, Ann. Statist., 51 (2023), pp. 179–210, https://doi.org/10.1214/22-AOS2246.

[17] S. Chen, J. Li, and Z. Song, *Learning mixtures of linear regressions in subexponential time via Fourier moments*, in Proceedings of the 52nd Annual ACM Symposium on Theory of Computing, 2020, pp. 587–600.
[18] W.-K. Chen and W.-K. Lam, *Universality of approximate message passing algorithms*, Electron. J. Probab., 26 (2021), pp. 1–44, https://doi.org/10.1214/21-EJP604.
[19] Y. Chen and E. J. Candès, *Solving random quadratic systems of equations is nearly as easy as solving linear systems*, in Advances in Neural Information Processing Systems (NIPS), 2015, pp. 739–747.
[20] Y. Chen, X. Yi, and C. Caramanis, *A convex formulation for mixed regression with two components: Minimax optimal rates*, in Proceedings of the 27th Conference on Learning Theory, PMLR 35, 2014, pp. 560–604.
[21] A. Damian, L. Pillaud-Vivien, J. Lee, and J. Bruna, *Computational-statistical gaps in Gaussian single-index models (extended abstract)*, in Proceedings of the 37th Annual Conference on Learning Theory, PMLR 247, 2024, pp. 1262–1262.
[22] Y. Deshpande and A. Montanari, *Information-theoretically optimal sparse PCA*, in IEEE International Symposium on Information Theory (ISIT), 2014, pp. 2197–2201.
[23] E. Devijver, Y. Goude, and J.-M. Poggi, *Clustering electricity consumers using high-dimensional regression mixture models*, Appl. Stoch. Model. Bus. Ind., 36 (2020), pp. 159–177, https://doi.org/10.1002/asmb.2453.
[24] D. Donoho and A. Montanari, *High dimensional robust M-estimation: Asymptotic variance via approximate message passing*, Probab. Theory Related Fields, 166 (2016), pp. 935–969, https://doi.org/10.1007/s00440-015-0675-z.
[25] D. L. Donoho, A. Maleki, and A. Montanari, *Message passing algorithms for compressed sensing*, Proc. Natl. Acad. Sci. USA, 106 (2009), pp. 18914–18919, https://doi.org/10.1073/pnas.0909892106.
[26] R. Dudeja and M. Bakhshizadeh, *Universality of linearized message passing for phase retrieval with structured sensing matrices*, IEEE Trans. Inform. Theory, 68 (2022), pp. 7545–7574, https://doi.org/10.1109/TIT.2022.3182018.
[27] R. Dudeja, M. Bakhshizadeh, J. Ma, and A. Maleki, *Analysis of spectral methods for phase retrieval with random orthogonal matrices*, IEEE Trans. Inform. Theory, 66 (2020), pp. 5182–5203, https://doi.org/10.1109/TIT.2020.2981910.
[28] R. Dudeja, Y. M. Lu, and S. Sen, *Universality of approximate message passing with semirandom matrices*, Ann. Probab., 51 (2023), pp. 1616–1683, https://doi.org/10.1214/23-AOP1628.
[29] R. Dudeja, S. Sen, and Y. M. Lu, *Spectral universality in regularized linear regression with nearly deterministic sensing matrices*, IEEE Trans. Inform. Theory, 70 (2024), pp. 7923–7951, https://doi.org/10.1109/TIT.2024.3458953.
[30] L. Erdős, H.-T. Yau, and J. Yin, *Bulk universality for generalized Wigner matrices*, Probab. Theory Related Fields, 154 (2012), pp. 341–407, https://doi.org/10.1007/s00440-011-0390-3.
[31] J. Fan, H. Liu, Z. Wang, and Z. Yang, *Curse of Heterogeneity: Computational Barriers in Sparse Mixture Models and Phase Retrieval*, preprint, arXiv:1808.06996, 2018.
[32] Z. Fan, *Approximate message passing algorithms for rotationally invariant matrices*, Ann. Statist., 50 (2022), pp. 197–224, https://doi.org/10.1214/21-AOS2101.
[33] A. Fannjiang and T. Strohmer, *The numerics of phase retrieval*, Acta Numer., 29 (2020), pp. 125–228, https://doi.org/10.1017/S0962492920000069.
[34] S. Faria and G. Soromenho, *Fitting mixtures of linear regressions*, J. Stat. Comput. Simul., 80 (2010), pp. 201–225, https://doi.org/10.1080/00949650802590261.
[35] B. Farrell, *Limiting empirical singular value distribution of restrictions of discrete Fourier transform matrices*, J. Fourier Anal. Appl., 17 (2011), pp. 733–753, https://doi.org/10.1007/s00041-010-9156-z.
[36] O. Y. Feng, R. Venkataramanan, C. Rush, and R. J. Samworth, *A unifying tutorial on approximate message passing*, Found. Trends Mach. Learn., 15 (2022), pp. 335–536, https://doi.org/10.1561/2200000092.
[37] A. Ghosh and R. Kannan, *Alternating minimization converges super-linearly for mixed linear regression*, in International Conference on Artificial Intelligence and Statistics, 2020, pp. 1093–1103.
[38] B. Grün and F. Leisch, *Applications of Finite Mixtures of Regression Models*, Technical report, 2007.

[39] A. Javanmard and A. Montanari, *State evolution for general approximate message passing algorithms, with applications to spatial coupling*, Inf. Inference, 2 (2013), pp. 115–144, https://doi.org/10.1093/imaiai/iat004.
[40] M. I. Jordan and R. A. Jacobs, *Hierarchical mixtures of experts and the EM algorithm*, Neural Comput., 6 (1994), pp. 181–214, https://doi.org/10.1162/neco.1994.6.2.181.
[41] A. Khalili and J. Chen, *Variable selection in finite mixture of regression models*, J. Amer. Statist. Assoc., 102 (2007), pp. 1025–1038, https://doi.org/10.1198/016214507000000590.
[42] J. M. Klusowski, D. Yang, and W. D. Brinda, *Estimating the coefficients of a mixture of two linear regressions by expectation maximization*, IEEE Trans. Inform. Theory, 65 (2019), pp. 3515–3524, https://doi.org/10.1109/TIT.2019.2891628.
[43] W. Kong, R. Somani, Z. Song, S. Kakade, and S. Oh, *Meta-learning for mixed linear regression*, in Proceedings of the 37th International Conference on Machine Learning, PMLR 119, 2020, pp. 5394–5404.
[44] F. Kovacevic, Y. Zhang, and M. Mondelli, *Spectral estimators for multi-index models: Precise asymptotics and optimal weak recovery*, in Proceedings of the 38th Annual Conference on Learning Theory (COLT), PMLR 291, 2025, pp. 3354–3404.
[45] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová, *Probabilistic reconstruction in compressed sensing: Algorithms, phase diagrams, and threshold achieving matrices*, J. Statist. Mech. Theory Exp., 2012 (2012), P08009, https://doi.org/10.1088/1742-5468/2012/08/P08009.
[46] T. Lesieur, F. Krzakala, and L. Zdeborová, *Constrained low-rank matrix estimation: Phase transitions, approximate message passing and applications*, J. Statist. Mech. Theory Exp., 2017 (2017), 073403, https://doi.org/10.1088/1742-5468/aa7284.
[47] G. Li and Y. Wei, *A Non-Asymptotic Framework for Approximate Message Passing in Spiked Models*, preprint, arXiv:2208.03313, 2022.
[48] K.-C. Li, *On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma*, J. Amer. Statist. Assoc., 87 (1992), pp. 1025–1039, https://doi.org/10.1080/01621459.1992.10476258.
[49] Q. Li, R. Shi, and F. Liang, *Drug sensitivity prediction with high-dimensional mixture regression*, PloS ONE, 14 (2019), e0212108, https://doi.org/10.1371/journal.pone.0212108.
[50] Y. Li and Y. Liang, *Learning mixtures of linear regressions with nearly optimal complexity*, in Proceedings of the 31st Annual Conference on Learning Theory, PMLR 75, 2018, pp. 1125–1144.
[51] Y. M. Lu and G. Li, *Phase transitions of spectral initialization for high-dimensional non-convex estimation*, Inf. Inference, 9 (2020), pp. 507–541, https://doi.org/10.1093/imaiai/iaz020.
[52] W. Luo, W. Alghamdi, and Y. M. Lu, *Optimal spectral initialization for signal recovery with applications to phase retrieval*, IEEE Trans. Signal Process., 67 (2019), pp. 2347–2356, https://doi.org/10.1109/TSP.2019.2904918.
[53] J. Ma and L. Ping, *Orthogonal AMP*, IEEE Access, 5 (2017), pp. 2020–2033, https://doi.org/10.1109/ACCESS.2017.2653119.
[54] A. Maillard, F. Krzakala, Y. M. Lu, and L. Zdeborová, *Construction of optimal spectral methods in phase retrieval*, in Proceedings of the 2nd Annual Conference on Mathematical and Scientific Machine Learning, PMLR 145, 2022, pp. 693–720.
[55] A. Maillard, B. Loureiro, F. Krzakala, and L. Zdeborová, *Phase retrieval in high dimensions: Statistical and computational phase transitions*, in Advances in Neural Information Processing Systems 33, 2020, pp. 11071–11082.
[56] P. McCullagh and J. Nedler, *Generalized Linear Models*, Chapman & Hall/CRC, 1989.
[57] G. J. McLachlan and D. Peel, *Finite Mixture Models*, John Wiley & Sons, 2004.
[58] M. Mondelli and A. Montanari, *Fundamental limits of weak recovery with applications to phase retrieval*, Found. Comput. Math., 19 (2019), pp. 703–773, https://doi.org/10.1007/s10208-018-9395-y.
[59] M. Mondelli, C. Thrampoulidis, and R. Venkataramanan, *Optimal combination of linear and spectral estimators for generalized linear models*, Found. Comput. Math., 22 (2022), pp. 1513–1566, https://doi.org/10.1007/s10208-021-09531-x.

[60] M. Mondelli and R. Venkataramanan, *Approximate message passing with spectral initialization for generalized linear models*, in Proceedings of the 24th International Conference on Artificial Intelligence and Statistics, PMLR 130, 2021, pp. 397–405.
[61] M. Mondelli and R. Venkataramanan, *PCA initialization for approximate message passing in rotationally invariant models*, in Advances in Neural Information Processing Systems 34, Curran Associates, 2021, pp. 29616–29629.
[62] A. Montanari and R. Venkataramanan, *Estimation of low-rank matrices via approximate message passing*, Ann. Statist., 49 (2021), pp. 321–345, https://doi.org/10.1214/20-AOS1958.
[63] P. Netrapalli, P. Jain, and S. Sanghavi, *Phase retrieval using alternating minimization*, in Advances in Neural Information Processing Systems 26), 2013, pp. 2796–2804.
[64] M. Opper, B. Cakmak, and O. Winther, *A theory of solving TAP equations for Ising models with general invariant random matrices*, J. Phys. A, 49 (2016), 114002, https://doi.org/10.1088/1751-8113/49/11/114002.
[65] S. Pal, A. Mazumdar, R. Sen, and A. Ghosh, *On learning mixture of linear regressions in the non-realizable setting*, in Proceedings of the 39th International Conference on Machine Learning, PMLR 162, 2022, pp. 17202–17220.
[66] F. Peng, R. A. Jacobs, and M. A. Tanner, *Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition*, J. Amer. Statist. Assoc., 91 (1996), pp. 953–960, https://doi.org/10.1080/01621459.1996.10476965.
[67] Y. Plan, R. Vershynin, and E. Yudovina, *High-dimensional estimation with geometric constraints*, Inf. Inference, 6 (2017), pp. 1–40, https://doi.org/10.1093/imaiai/iaw015.
[68] S. Rangan, *Generalized approximate message passing for estimation with random linear mixing*, in Proceedings of the 2011 IEEE International Symposium on Information Theory, 2011, pp. 2168–2172.
[69] S. Rangan, V. Goyal, and A. K. Fletcher, *Asymptotic analysis of MAP estimation via the replica method and compressed sensing*, in Advances in Neural Information Processing Systems 22, Curran Associates, 2009, pp. 1545–1553.
[70] S. Rangan, P. Schniter, and A. K. Fletcher, *Vector approximate message passing*, IEEE Trans. Inform. Theory, 65 (2019), pp. 6664–6684, https://doi.org/10.1109/TIT.2019.2916359.
[71] C. Rush and R. Venkataramanan, *Finite sample analysis of approximate message passing algorithms*, IEEE Trans. Inform. Theory, 64 (2018), pp. 7264–7286, https://doi.org/10.1109/TIT.2018.2816681.
[72] P. Schniter and S. Rangan, *Compressive phase retrieval via generalized approximate message passing*, IEEE Trans. Signal Process., 63 (2015), pp. 1043–1055, https://doi.org/10.1109/TSP.2014.2386294.
[73] H. Sedghi, M. Janzamin, and A. Anandkumar, *Provable tensor methods for learning mixtures of generalized linear models*, in Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, PMLR 51, 2016, pp. 1223–1231.
[74] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev, *Phase retrieval with application to optical imaging: A contemporary overview*, IEEE Signal Processing Mag., 32 (2015), pp. 87–109, https://doi.org/10.1109/MSP.2014.2352673.
[75] Y. Shen and S. Sanghavi, *Iterative least trimmed squares for mixed linear regression*, in Advances in Neural Information Processing Systems 32, Curran Associates, 2019.
[76] N. Städler, P. Bühlmann, and S. van de Geer, *ℓ_1-penalization for mixture regression models*, TEST, 19 (2010), pp. 209–256, https://doi.org/10.1007/s11749-010-0197-z.
[77] P. Sur and E. J. Candès, *A modern maximum-likelihood theory for high-dimensional logistic regression*, Proc. Natl. Acad. Sci. USA, 116 (2019), pp. 14516–14525, https://doi.org/10.1073/pnas.1810420116.
[78] K. Takeuchi, *Rigorous dynamics of expectation-propagation-based signal recovery from unitarily invariant measurements*, IEEE Trans. Inform. Theory, 66 (2020), pp. 368–386, https://doi.org/10.1109/TIT.2019.2947058.
[79] N. Tan and R. Venkataramanan, *Mixed regression via approximate message passing*, J. Mach. Learn. Res., 24 (2023), 317.
[80] T. Tao and V. Vu, *Random matrices: Localization of the eigenvalues and the necessity of four moments*, Acta Math. Vietnam., 36 (2011), pp. 431–449.
[81] A. M. Tulino, G. Caire, S. Shamai, and S. Verdu, *Capacity of channels with frequency-selective and time-selective fading*, IEEE Trans. Inform. Theory, 56 (2010), pp. 1187–1215, https://doi.org/10.1109/TIT.2009.2039041.

[82] R. Venkataramanan, K. Kögler, and M. Mondelli, *Estimation in rotationally invariant generalized linear models via approximate message passing*, in Proceedings of the 39th International Conference on Machine Learning, PMLR 162, 2022, pp. 22120–22144.
[83] K. Viele and B. Tong, *Modeling with mixtures of linear regressions*, Stat. Comput., 12 (2002), pp. 315–330, https://doi.org/10.1023/A:1020779827503.
[84] T. Wang, X. Zhong, and Z. Fan, *Universality of approximate message passing algorithms and tensor networks*, Ann. Appl. Probab., 34 (2024), pp. 3943–3994, https://doi.org/10.1214/24-AAP2056.
[85] Z. Wang, Q. Gu, Y. Ning, and H. Liu, *High dimensional EM algorithm: Statistical optimization and asymptotic normality*, in Advances in Neural Information Processing Systems 28, 2015, pp. 2521–2529.
[86] S. Waterhouse, D. MacKay, and A. Robinson, *Bayesian methods for mixtures of experts*, in Advances in Neural Information Processing Systems 8, MIT Press, 1995, pp. 351–357.
[87] X. Yi and C. Caramanis, *Regularized EM algorithms: A unified framework and statistical guarantees*, in Advances in Neural Information Processing Systems 28, 2015, pp. 1567–1575.
[88] X. Yi, C. Caramanis, and S. Sanghavi, *Alternating minimization for mixed linear regression*, in Proceedings of the 31st International Conference on Machine Learning, PMLR 32, 2014, pp. 613–621.
[89] L. Zhang, R. Ma, T. T. Cai, and H. Li, *Estimation, Confidence Intervals, and Large-Scale Hypotheses Testing for High-Dimensional Mixed Linear Regression*, arXiv:2011.03598, 2020.
[90] K. Zhong, P. Jain, and I. S. Dhillon, *Mixed linear regression with multiple components*, in Advances in Neural Information Processing Systems 29, 2016, pp. 2190–2198.
[91] X. Zhong, T. Wang, and Z. Fan, *Approximate message passing for orthogonally invariant ensembles: Multivariate non-linearities and spectral initialization*, Inf. Inference, 13 (2024), iaae024, https://doi.org/10.1093/imaiai/iaae024.
[92] R. Zhu, L. Wang, C. Zhai, and Q. Gu, *High-dimensional variance-reduced stochastic gradient expectation-maximization algorithm*, in Proceedings of the 34th International Conference on Machine Learning, PMLR 70, 2017, pp. 4180–4188.