*The Annals of Statistics*
2017, Vol. 45, No. 1, 77–120
DOI: 10.1214/16-AOS1435
© Institute of Mathematical Statistics, 2017

# STATISTICAL GUARANTEES FOR THE EM ALGORITHM: FROM POPULATION TO SAMPLE-BASED ANALYSIS[1]

BY SIVARAMAN BALAKRISHNAN[*,†], MARTIN J. WAINWRIGHT[†] AND BIN YU[†]

*University of California, Berkeley[*] and Carnegie Mellon University[†]*

The EM algorithm is a widely used tool in maximum-likelihood estimation in incomplete data problems. Existing theoretical work has focused on conditions under which the iterates or likelihood values converge, and the associated rates of convergence. Such guarantees do not distinguish whether the ultimate fixed point is a near global optimum or a bad local optimum of the sample likelihood, nor do they relate the obtained fixed point to the global optima of the idealized population likelihood (obtained in the limit of infinite data). This paper develops a theoretical framework for quantifying when and how quickly EM-type iterates converge to a small neighborhood of a given global optimum of the population likelihood. For correctly specified models, such a characterization yields rigorous guarantees on the performance of certain two-stage estimators in which a suitable initial pilot estimator is refined with iterations of the EM algorithm. Our analysis is divided into two parts: a treatment of the EM and first-order EM algorithms at the population level, followed by results that apply to these algorithms on a finite set of samples. Our conditions allow for a characterization of the region of convergence of EM-type iterates to a given population fixed point, that is, the region of the parameter space over which convergence is guaranteed to a point within a small neighborhood of the specified population fixed point. We verify our conditions and give tight characterizations of the region of convergence for three canonical problems of interest: symmetric mixture of two Gaussians, symmetric mixture of two regressions and linear regression with covariates missing completely at random.

**1. Introduction.** Data problems with missing values, corruptions and latent variables are common in practice. From a computational standpoint, computing the maximum likelihood estimate (MLE) in such incomplete data problems can be quite complex. To a certain extent, these concerns have been assuaged by the development of the expectation-maximization (EM) algorithm, along with growth

Received September 2015; revised January 2016.

[1]Supported in part by ONR-MURI Grant DOD 002888 and NSF Grant CIF-31712-23800, as well by NSF Grants DMS-11-07000, CDS&E-MSS 1228246, ARO Grant W911NF-11-1-0114, the Center for Science of Information (CSoI) and NSF Science and Technology Center, under Grant agreement CCF-0939370.

*MSC2010 subject classifications.* Primary 62F10, 60K35; secondary 90C30.

*Key words and phrases.* EM algorithm, first-order EM algorithm, nonconvex optimization, maximum likelihood estimation.

in computational resources. The EM algorithm is widely applied to incomplete data problems, and there is now a very rich literature on its behavior (e.g., [11, 12, 17, 25, 27, 30, 32, 34, 42, 44, 49]). However, a major issue is that in most models, although the MLE is known to have good statistical properties, the EM algorithm is only guaranteed to return a local optimum of the sample likelihood function. The goal of this paper is to address this gap between statistical and computational guarantees, in particular by developing an understanding of conditions under which the EM algorithm is guaranteed to converge to a local optimum that matches the performance of maximum likelihood estimate up to constant factors.

The EM algorithm has a lengthy and rich history. Various algorithms of the EM-type were analyzed in early work (e.g., [5, 6, 18, 19, 37, 40, 41]), before the EM algorithm in its modern general form was introduced by Dempster, Laird and Rubin [17]. Among other results, these authors established its well-known monotonicity properties. Wu [50] established some of the most general convergence results known for the EM algorithm; see also the more recent papers [15, 43]. Among the results in the paper, [50] is a guarantee for the EM algorithm to converge to the unique global optimum when the likelihood is unimodal and certain regularity conditions hold. However, in most interesting cases of the EM algorithm, the likelihood function is multi-modal, in which case the best that can be guaranteed is convergence to some local optimum of the likelihood at an asymptotically geometric rate (see, e.g., [20, 29, 31, 33]). A guarantee of this type does not preclude that the EM algorithm converges to a "poor" local optimum—meaning one that is far away from any global optimum of the likelihood. For this reason, despite its popularity and widespread practical effectiveness, the EM algorithm is in need of further theoretical backing.

The goal of this paper is to take the next step in closing this gap between the practical use of EM and its theoretical understanding. At a high level, our main contribution is to provide a quantitative characterization of a basin of attraction around the population global optimum with the following property: if the EM algorithm is initialized within this basin, then it is guaranteed to converge to an EM fixed point that is within *statistical precision* of a global optimum. The statistical precision is a measure of the error in the maximum likelihood estimate, or any other minimax optimal method; we define it more precisely in the sequel. Thus, in sharp contrast with the classical theory [20, 29, 31, 33]—which guarantees asymptotic convergence to an *arbitrary EM fixed point*—our theory guarantees geometric convergence to a "*good*" *EM fixed point*.

In more detail, we make advances over the classical results in the following specific directions:

- Existing results on the rate of convergence of the EM algorithm guarantee that there is some neighborhood of a fixed point over which the algorithm converges to this fixed point, but do not quantify its size. In contrast, we formulate conditions on the auxiliary $Q$-function underlying the EM algorithm, which allow us

to give a quantitative characterization of the region of attraction around the population global optimum. As shown by our analysis for specific statistical models, its size is determined by readily interpretable problem-dependent quantities, such as the signal-to-noise ratio (SNR) in mixture models, or the probability of missing-ness in models with missing data. As a consequence, we can provide concrete guarantees on the initializations of EM that lead to good fixed points. For example, for Gaussian mixture models with a suitably large mean separation, we show that a relatively poor initialization suffices for the EM algorithm to converge to a near-globally optimal solution.

- Classical results on the EM algorithm are all sample-based, in particular applying to any fixed point of the sample likelihood. However, given the nonconvexity of the likelihood, there is a priori no reason to believe that any fixed points of the sample likelihood are close to the population MLE (i.e., a maximizer of the population likelihood), or equivalently (for a well-specified model) close to the underlying true parameter. Indeed, it is easy to find cases in which the likelihood function has spurious local maxima; see Figure 1 for one simple example. In our approach, we first study the EM algorithm in the idealized limit of infinite samples, referred to as the population level. For specific models, we provide conditions under which there are in fact no spurious fixed points for two algorithms of interest (the EM and first-order EM algorithms) at the population

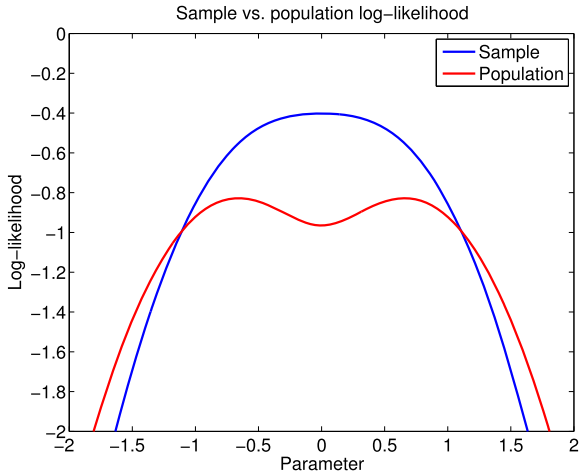


FIG. 1. *An illustration of the inadequacy of purely sample-based theory guaranteeing linear convergence to any fixed point of the sample-based likelihood. The figure illustrates the population and sample-based likelihoods for samples* $y \sim \frac{1}{2}\mathcal{N}(-\theta^*, 1) + \frac{1}{2}\mathcal{N}(\theta^*, 1)$ *with* $\theta^* = 0.7$. *There are two global optima for the population-likelihood corresponding to* $\theta^*$ *and* $-\theta^*$, *while the sample-based likelihood, for a small sample size, can have a single spurious maximum near* 0. *Our theory guarantees that for a sufficiently large sample size this phenomenon is unlikely, and that in a large region around* $\theta^*$ (*of radius roughly* $\|\theta^*\|_2$), *all maxima of the sample-based likelihood are extremely close to* $\theta^*$, *with an equivalent statement for a neighborhood of* $-\theta^*$.

level. We then give a precise lower bound on the sample size that suffices to ensure that, with high probability, the sample likelihood does not have spurious fixed points far away from the population MLE. These results show that the behavior shown in Figure 1 is unlikely given a sufficiently large sample.

- In simulations, it is frequently observed that if the EM algorithm is given a "suitable" initialization, then it converges to a statistically consistent estimate. For instance, in application to a mixture of regressions problem, Chaganty and Liang [13] empirically demonstrate good performance for a two-stage estimator, in which the method of moments is used as an initialization, and then the EM algorithm is applied to refine this initial estimator. Our theory allows us to give a precise characterization of what type of initialization is suitable for these types of two-stage methods. When the pilot estimator is consistent but does not achieve the minimax-optimal rate (as is often the case for various moment-based estimators in high dimensions), then these two-stage approaches are often much better than the initial pilot estimator alone. Our theoretical results help explain this behavior, and can further be used to characterize the refinement stage in new examples.

In well-specified statistical models, our results provide sufficient conditions on initializations that ensure that the EM algorithm converges geometrically to a fixed point that is within statistical precision of the unknown true parameter. Such a characterization is useful for a variety of reasons. First, there are many settings (including mixture modeling) in which the statistician has the ability to collect a few labeled samples in addition to many unlabeled ones, and understanding the size of the region of convergence of EM can be used to guide the efforts of the statistician, by characterizing the number of labeled samples that suffice to (with high-probability) provide an initialization from which she can leverage the unlabeled samples. In this setting, the typically small set of labeled samples are used to construct an initial estimator which is then refined by the EM algorithm applied on the larger pool of unlabeled samples. Second, in practice, the EM algorithm is run with numerous random initializations. Although we do not directly attempt to address this in this paper, we note that a tight characterization of the region of attraction can be used in a straightforward way to answer the question: how many random initializations (from a specified distribution) suffice (with high-probability) to find a near-globally optimal solution?

*Roadmap*. Our main results concern the population EM and first-order EM algorithms and their finite-sample counterparts. We give conditions under which the population algorithms are contractive to the MLE, when initialized in a ball around the MLE. These conditions allow us to establish the region of attraction of the population MLE. A bulk of our technical effort is in the treatment of three examples—namely, a symmetric mixture of two Gaussians, a symmetric mixture of two regressions and regression with missing covariates—for which we show

that our conditions hold in a large region around the MLE, and that the size of this region is determined by interpretable problem-dependent quantities.

The remainder of this paper is organized as follows. Section 2 provides an Introduction to the EM and first-order EM algorithms, and develops some intuition for the theoretical treatment of the first-order EM algorithm. Section 3 is devoted to the analysis of the first-order EM at the population level: in particular, Theorem 1 specifies concrete conditions that ensure geometric convergence, and Corollaries 1, 2 and 3 show that these conditions hold for three specific classes of statistical models: Gaussian mixtures, mixture of regressions and regression with missing covariates. We follow with analysis of the sample-based form of the first-order EM updates in Section 4, again stating two general theorems (Theorems 2 and 3), and developing their consequences for our three specific models in Corollaries 4, 5 and 6. We also provide an analogous set of population and sample results for the standard EM updates. The main results appear in Section 5. Due to space constraints, we defer detailed proofs as well as a treatment of concrete examples to the Supplementary Material [3]. In addition, Appendix C contains additional analysis of stochastic online forms of the first-order EM updates. Section 6 is devoted to the proofs of our results on the first-order EM updates, with some more technical aspects again deferred to appendices in the Supplementary Material.

**2. Background and intuition.** We begin with basic background on the standard EM algorithm as well as the first-order EM algorithm as they are applied at the sample level. We follow this background by introducing the population-level perspective that underlies the analysis of this paper, including the notion of the oracle iterates at the population level and the gradient smoothness condition, as well as discussing the techniques required to translate from population based results to finite-sample based results.

2.1. *EM algorithm and its relatives.* Let $Y$ and $Z$ be random variables taking values in the sample spaces $\mathcal{Y}$ and $\mathcal{Z}$, respectively. Suppose that the pair $(Y, Z)$ has a joint density function $f_{\theta^*}$ that belongs to some parameterized family $\{f_\theta | \theta \in \Omega\}$ where $\Omega$ is some nonempty convex set of parameters. Suppose that rather than observing the complete data $(Y, Z)$, we observe only component $Y$. The component $Z$ corresponds to the missing or latent structure in the data. For each $\theta \in \Omega$, we let $k_\theta(z|y)$ denote the conditional density of $z$ given $y$.

Our goal is to obtain an estimate of the unknown parameter $\theta^*$ via maximizing the log-likelihood. Throughout this paper, we assume that the generative model is correctly specified, with an unknown true parameter $\theta^*$. In the classical statistical setting, we observe $n$ i.i.d. samples $\{y_i\}_{i=1}^n$ of the $Y$ component. Formally, under the i.i.d. assumption, we are interested in computing some $\widehat{\theta} \in \Omega$ maximizing the log-likelihood function $\theta \mapsto \ell_n(\theta)$ where

$$\ell_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log\Big[\int_{\mathcal{Z}} f_\theta(y_i, z_i)\, dz_i\Big].$$

Rather than attempting to maximize the likelihood directly, the EM framework is based on using an auxiliary function to lower bound the log likelihood. More precisely, we define a bivariate function $Q_n : \Omega \times \Omega \to \mathbb{R}$ as follows.

DEFINITION 1 (Finite-sample $Q$-function).

$$Q_n(\theta|\theta') = \frac{1}{n}\sum_{i=1}^{n}\Big(\int_{\mathcal{Z}} k_{\theta'}(z|y_i)\log f_\theta(y_i, z)\,dz\Big). \tag{2.1}$$

The quantity $Q_n(\theta|\theta')$ provides a lower bound on the log-likelihood $\ell_n(\theta)$ for any $\theta$, with equality holding when $\theta = \theta'$—that is, $\ell_n(\theta') = Q_n(\theta'|\theta')$.

The standard EM algorithm operates by maximizing this auxiliary function, whereas the first-order EM algorithm operates by taking a gradient step.[2] In more detail:

- Given some initialization $\theta^0 \in \Omega$, the standard EM algorithm performs the updates

$$\theta^{t+1} = \arg\max_{\theta\in\Omega} Q_n(\theta|\theta^t), \qquad t = 0, 1, \ldots. \tag{2.2}$$

- Given some initialization $\theta^0 \in \Omega$ and an appropriately chosen step-size $\alpha \geq 0$, the first-order EM algorithm performs the updates:

$$\theta^{t+1} = \theta^t + \alpha\nabla Q_n(\theta|\theta^t)|_{\theta=\theta^t} \qquad \text{for } t = 0, 1, \ldots, \tag{2.3}$$

  where the gradient is taken with respect to the first argument of the $Q$-function.[3] There is also a natural extension of the first-order EM iterates that includes a constraint arising from the parameter space $\Omega$, in which the update is projected back using a Euclidean projection onto the constraint set $\Omega$.

It is important to note that in typical examples, several of which are considered in detail in this paper, the likelihood function $\ell_n$ is not concave, which makes direct computation of a maximizer challenging. On the other hand, there are many cases in which, for each fixed $\theta' \in \Omega$, the functions $Q_n(\cdot|\theta')$ are concave, thereby rendering the EM updates tractable. In this paper, as is often the case in examples, we focus on cases when the functions $Q_n(\cdot|\theta')$ are concave.

It is easy to verify that the gradient $\nabla Q_n(\theta|\theta^t)$, when evaluated at the specific point $\theta = \theta^t$, is actually equal to the gradient $\nabla\ell_n(\theta^t)$ of the log-likelihood at $\theta^t$. Thus, the first-order EM algorithm is *actually* gradient ascent on the marginal log-likelihood function. However, the description given in equation (2.3) emphasizes the role of the $Q$-function, which plays a key role in our theoretical development, and allows us to prove guarantees even when the log likelihood is not concave.

[2]We assume throughout that $Q_n$ and $Q$ are differentiable in their first argument.

[3]Throughout this paper, we always consider the derivative of the $Q$-function with respect to its first argument.

2.2. *Population-level perspective.* The core of our analysis is based on analyzing the log likelihood and the $Q$-functions at the population level, corresponding to the idealized limit of an infinite sample size. The population counterpart of the log likelihood is the function $\theta \mapsto \ell(\theta)$ given by

$$\ell(\theta) = \int_{\mathcal{Y}} \log\Big[\int_{\mathcal{Z}} f_\theta(y,z)\,dz\Big] g_{\theta^*}(y)\,dy, \tag{2.4}$$

where $\theta^*$ denotes the true, unknown parameter and $g_{\theta^*}$ is the marginal density of the observed data. A closely related object is the population analog of the $Q$-function, defined as follows.

DEFINITION 2 (Population $Q$-function).

$$Q(\theta|\theta') = \int_{\mathcal{Y}} \Big(\int_{\mathcal{Z}} k_{\theta'}(z|y) \log f_\theta(y,z)\,dz\Big) g_{\theta^*}(y)\,dy. \tag{2.5}$$

We can then consider the population analogs of the standard EM and first-order EM updates, obtained by replacing $Q_n$ and $\nabla Q_n$ with $Q$ and $\nabla Q$ in equations (2.2) and (2.3), respectively. Our main goal is to understand the region of the parameter space over which these iterative schemes, are convergent to $\theta^*$. For the remainder of this section, let us focus exclusively on the population first-order EM updates, given by

$$\theta^{t+1} = \theta^t + \alpha \nabla Q\big(\theta|\theta^t\big)|_{\theta=\theta^t}, \qquad \text{for } t = 0, 1, 2, \ldots. \tag{2.6}$$

The concepts developed here are also useful in understanding the EM algorithm; we provide a brief treatment of the EM algorithm in Section 5 and a full treatment of it in Appendix B of the Supplementary Material.

2.3. *Oracle auxiliary function and iterates.* Our key insight is that in a local neighborhood of $\theta^*$, the first-order EM iterates (2.6) can be viewed as perturbations of an alternate *oracle iterative scheme*, one that is guaranteed to converge to $\theta^*$. This leads us to a natural condition, relating the perturbed and oracle iterative schemes, which gives an explicit way to characterize the region of convergence of the first-order EM algorithm.

Since the vector $\theta^*$ is a maximizer of the population log-likelihood, a classical result [29] guarantees that it must then satisfy the condition

$$\theta^* = \arg\max_{\theta\in\Omega} Q\big(\theta|\theta^*\big), \tag{2.7}$$

a property known as *self-consistency*. Whenever the function $Q$ is concave in its first argument, this property allows us to express the fixed-point of interest $\theta^*$ as the solution of a concave maximization problem—namely one involving the auxiliary function $q : \Omega \to \mathbb{R}$ given by the following.

DEFINITION 3 (Oracle auxiliary function).

$$q(\theta) := Q(\theta|\theta^*) = \int_{\mathcal{Y}} \left( \int_{\mathcal{Z}} k_{\theta^*}(z|y) \log f_\theta(y,z)\, dz \right) g_{\theta^*}(y)\, dy. \tag{2.8}$$

Why is this oracle function useful? Assuming that it satisfies some standard regularity conditions—namely, strong concavity and smoothness—classical theory on gradient methods yields that, with an appropriately chosen stepsize $\alpha$, the iterates

$$\tilde{\theta}^{t+1} = \tilde{\theta}^t + \alpha \nabla q(\tilde{\theta}^t) \qquad \text{for } t = 0, 1, 2, \ldots \tag{2.9}$$

converge at a geometric rate to $\theta^*$. Of course, even in the idealized population setting, the statistician cannot compute the oracle function $q$, since it presumes knowledge of the unknown parameter $\theta^*$. However, with this perspective in mind, the first-order EM iterates (2.3) can be viewed as a perturbation of the idealized oracle iterates (2.9).

By comparing these two iterative schemes, we see that the only difference is the replacement of $\nabla q(\theta^t) = \nabla Q(\theta^t|\theta^*)$ with the quantity $\nabla Q(\theta^t|\theta^t)$. Thus, we are naturally led to consider a *gradient smoothness condition* which ensures the closeness of these quantities. Particularly, we consider a condition of the form

$$\|\nabla q(\theta) - \nabla Q(\theta|\theta)\|_2 \le \gamma \|\theta - \theta^*\|_2 \qquad \text{for all } \theta \in \mathbb{B}_2(r;\theta^*), \tag{2.10}$$

where $\mathbb{B}_2(r;\theta^*)$ denotes a Euclidean ball[4] of radius $r$ around the fixed point $\theta^*$, and $\gamma$ is a smoothness parameter. Our first main result (Theorem 1) shows that when the gradient smoothness condition (2.10) holds for appropriate values of $\gamma$, then for any initial point $\theta^0 \in \mathbb{B}_2(r;\theta^*)$, the first-order EM iterates converge at a geometric rate to $\theta^*$. In this way, we have a method for explicitly characterizing the region of the parameter space $\Omega$ over which the first-order EM iterates converge to $\theta^*$.

Of course, there is no a priori reason to suspect that gradient smoothness condition (2.10) holds for any nontrivial values of the radius $r$ and with a sufficiently small $\gamma$ in concrete examples. Indeed, much of the technical work in our paper is devoted to studying important and canonical examples of the EM algorithm, and showing that the smoothness condition (2.10) does hold for reasonable choices of the parameters $r$ and $\gamma$, ones which yield accurate predictions of the behavior of EM in practice.

2.4. *From population to sample-based analysis.* Our ultimate interest is in the behavior of the finite-sample first-order EM algorithm. Since the finite-sample updates (2.3) are based on the sample gradient $\nabla Q_n$ instead of the population gradient $\nabla Q$, a central object in our analysis is the empirical process given by

$$\{\nabla Q(\theta|\theta) - \nabla Q_n(\theta|\theta), \theta \in \mathbb{B}_2(r;\theta^*)\}. \tag{2.11}$$

[4]Our choice of a Euclidean ball is for concreteness; as the analysis in the sequel clarifies, other convex local neighborhoods of $\theta^*$ could also be used.

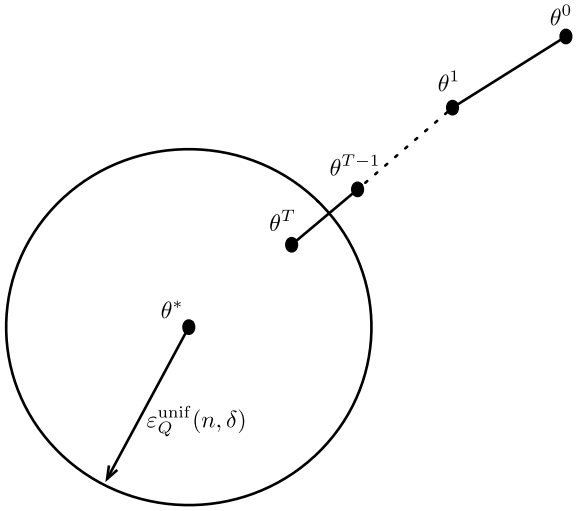


FIG. 2. *An illustration of Theorem 2. The theorem describes the geometric convergence of iterates of the first-order EM algorithm to the ball of radius* $\mathcal{O}(\varepsilon_Q^{\mathrm{unif}}(n,\delta))$.

Let $\varepsilon_Q^{\mathrm{unif}}(n,\delta)$ be an upper bound on the supremum of this empirical process that holds with probability at least $1-\delta$. With this notation, our second main result (Theorem 2) shows that under our previous conditions at the population level, the sample first-order EM iterates converge geometrically to a near-optimal solution—namely, a point whose distance from $\theta^*$ is at most a constant multiple of $\varepsilon_Q^{\mathrm{unif}}(n,\delta)$. Figure 2 provides an illustration of the convergence guarantee provided by Theorem 2.

Of course, this type of approximate convergence to $\theta^*$ is only useful if the bound $\varepsilon_Q^{\mathrm{unif}}(n,\delta)$ is small enough—ideally, of the same or lower order than the statistical precision, as measured by the Euclidean distance from the MLE to $\theta^*$. Consequently, a large part of our technical effort is devoted to establishing such bounds on the empirical process (2.11), making use of several techniques such as symmetrization, contraction and concentration inequalities. All of our finite-sample results are nonasymptotic, and allow for the problem dimension $d$ to scale with the sample-size $n$. Our finite-sample bounds are minimax-optimal up to logarithmic factors, and in typical cases are only sensible for scalings of $d$ and $n$ for which $d \ll n$. This is the best one can hope for without additional structural assumptions. We also note that after the initial posting of this work, the paper of Wang et al. [48] utilized our population-level analysis in the analysis of a truncated EM algorithm which under the structural assumption of sparsity of the unknown true parameter achieves near minimax-optimal rates in the regime when $d \gg n$.

The empirical process in equation (2.11) is tailored for analyzing the batch version of sample EM, in which the entire data set is used in each update. In other settings, it can also be useful to consider sample-splitting EM variants, in which

each iteration uses a fresh batch of samples. The key benefit from a theoretical standpoint of the sample-splitting variant is that at the price of a typically logarithmic overhead in sample size, analysis of the sample-splitting variant requires much weaker control on the empirical process: instead of controlling the supremum of the empirical process in equation (2.11), we only require a point-wise bound that needs to hold at the sequence of iterates. Our third main result (Theorem 3) provides analogous guarantees on such a sample-splitting form of the EM updates. Finally, in Appendix C, we analyze using a different technique, based on stochastic approximation, the most extreme form of sample-splitting, in which each iterate is based on a single fresh sample, corresponding to a form of stochastic EM. This form of extreme sample-splitting leads to an estimator that can be computed in an online/streaming fashion on an extremely large data-set which is an important consideration in modern statistical practice.

**3. Population-level analysis of the first-order EM algorithm.** This section is devoted to a detailed analysis of the first-order EM algorithm at the population level. Letting $\theta^*$ denote a given global maximum of the population likelihood, our first main result (Theorem 1) characterizes a Euclidean ball around $\theta^*$ over which the population update is contractive. Thus, for any initial point falling in this ball, we are guaranteed that the first-order EM updates converge to $\theta^*$. In Section 3.2, we derive some corollaries of this general theorem for three specific statistical models: mixtures of Gaussians, mixtures of regressions and regression with missing data.

3.1. *A general population-level guarantee.* Recall that the population-level first-order EM algorithm is based on the recursion $\theta^{t+1} = \theta^t + \alpha \nabla Q(\theta|\theta^t)|_{\theta=\theta^t}$, where $\alpha > 0$ is a step size parameter to be chosen. The main contribution of this section is to specify a set of conditions, *defined on a Euclidean ball* $\mathbb{B}_2(r; \theta^*)$ *of radius r around this point*, that ensure that any such sequence, when initialized in this ball, converges geometrically $\theta^*$.

Our first requirement is the gradient smoothness condition previously discussed in Section 2.3. Formally, we require the following.

CONDITION 1 (Gradient smoothness). For an appropriately small parameter $\gamma \geq 0$, we have that

$$\|\nabla q(\theta) - \nabla Q(\theta|\theta)\|_2 \leq \gamma \|\theta - \theta^*\|_2 \qquad \text{for all } \theta \in \mathbb{B}_2(r; \theta^*). \tag{3.1}$$

As specified more clearly in the sequel, a key requirement in the above condition is that the parameter $\gamma$, be sufficiently small. Our remaining two requirements apply to the oracle auxiliary function $q(\theta) := Q(\theta|\theta^*)$, as previously introduced in Definition 3. We require the following.

CONDITION 2 ($\lambda$-strong concavity). There is some $\lambda > 0$ such that

$$q(\theta_1) - q(\theta_2) - \langle \nabla q(\theta_2), \theta_1 - \theta_2 \rangle \leq -\frac{\lambda}{2}\|\theta_1 - \theta_2\|_2^2 \quad \text{for all pairs } \theta_1, \theta_2 \in \mathbb{B}_2(r; \theta^*). \tag{3.2}$$

When we require this condition to hold for all pairs $\theta_1, \theta_2 \in \Omega$ we refer to this as *global* $\lambda$-strong concavity.

CONDITION 3 ($\mu$-smoothness). There is some $\mu > 0$ such that

$$q(\theta_1) - q(\theta_2) - \langle \nabla q(\theta_2), \theta_1 - \theta_2 \rangle \geq -\frac{\mu}{2}\|\theta_1 - \theta_2\|_2^2 \quad \text{for all } \theta_1, \theta_2 \in \mathbb{B}_2(r; \theta^*). \tag{3.3}$$

As we illustrate, these conditions hold in many concrete instantiations of EM, including the three model classes we study in the next section.

Before stating our first main result, let us provide some intuition as to why these conditions ensure good behavior of the first-order EM iterates. As noted in Section 2.3, the point $\theta^*$ maximizes the function $q$, so that in the unconstrained case, we are guaranteed that $\nabla q(\theta^*) = 0$. Now suppose that the $\lambda$-strong concavity and $\gamma$-smoothness conditions hold for some $\gamma < \lambda$. Under these conditions, it is easy to show (see Appendix A.4) that

$$\langle \nabla Q(\theta^t|\theta^t), \nabla q(\theta^t) \rangle > 0 \qquad \text{for any } \theta^t \in \mathbb{B}_2(r; \theta^*) \setminus \{\theta^*\}. \tag{3.4}$$

This condition guarantees that for any $\theta^t \neq \theta^*$, the direction $\nabla Q(\theta^t|\theta^t)$ taken by the first-order EM algorithm at iteration $t$ always makes a positive angle with $\nabla q(\theta^t)$, which is an ascent direction for the function $q$. Given our perspective of $q$ as a concave surrogate function for the nonconcave log-likelihood, we see condition (3.4) ensures that the first-order EM algorithm makes progress toward $\theta^*$. Our first main theorem makes this intuition precise, and in fact guarantees a geometric rate of convergence toward $\theta^*$.

THEOREM 1. *For some radius $r > 0$, and a triplet $(\gamma, \lambda, \mu)$ such that $0 \leq \gamma < \lambda \leq \mu$, suppose that Conditions 1, 2 and 3 hold, and suppose that the stepsize is chosen as $\alpha = \frac{2}{\mu+\lambda}$. Then given any initialization $\theta^0 \in \mathbb{B}_2(r; \theta^*)$, the population first-order EM iterates satisfy the bound*

$$\|\theta^t - \theta^*\|_2 \leq \left(1 - \frac{2\lambda - \gamma}{\mu + \lambda}\right)^t \|\theta^0 - \theta^*\|_2 \qquad \textit{for all } t = 1, 2, \ldots. \tag{3.5}$$

Since $(1 - \frac{2\lambda-\gamma}{\mu+\lambda}) < 1$, the bound (3.5) ensures that at the population level, the first-order EM iterates converge geometrically to $\theta^*$.

Although its proof (see Section 6.1) is relatively straightforward, applying Theorem 1 to concrete examples requires some technical work in order to certify that

Conditions 1 through 3 hold over the ball $\mathbb{B}_2(r; \theta^*)$ for a reasonably large choice of the radius $r$. In the examples considered in this paper, the strong concavity and smoothness conditions are usually relatively straightforward, whereas establishing gradient smoothness (Condition 1) is more challenging. Intuitively, the gradient smoothness condition is a smoothness condition on the $Q$-function with respect to its second argument. Establishing that the gradient condition holds over (nearly) optimally-sized regions involves carefully leveraging properties of the generative model as well as smoothness properties of the log-likelihood function.

3.2. *Population-level consequences for specific models.* In this section, we derive some concrete consequences of Theorem 1 in application to three classes of statistical models for which the EM algorithm is frequently applied: Gaussian mixture models in Section 3.2.1, mixtures of regressions in Section 3.2.2 and regression with missing covariates in Section 3.2.3. We refer the reader to Appendix A for derivations of the exact form of the EM and first-order EM updates for these three models, thereby leaving this section to focus on the consequences on the theory.

3.2.1. *Gaussian mixture models.* Consider the two-component Gaussian mixture model with balanced weights and isotropic covariances. It can be specified by a density of the form

$$f_\theta(y) = \tfrac{1}{2}\phi\big(y; \theta^*, \sigma^2 I_d\big) + \tfrac{1}{2}\phi\big(y; -\theta^*, \sigma^2 I_d\big), \tag{3.6}$$

where $\phi(\cdot; \mu, \Sigma)$ denotes the density of a $\mathcal{N}(\mu, \Sigma)$ random vector in $\mathbb{R}^d$, and we have assumed that the two components are equally weighted. Suppose that the variance $\sigma^2$ is known, so that our goal is to estimate the unknown mean vector $\theta^*$. In this example, the hidden variable $Z \in \{0, 1\}$ is an indicator variable for the underlying mixture component, that is,

$$(Y|Z=0) \sim \mathcal{N}\big(-\theta^*, \sigma^2 I_d\big) \quad \text{and} \quad (Y|Z=1) \sim \mathcal{N}\big(\theta^*, \sigma^2 I_d\big).$$

The difficulty of estimating such a mixture model can be characterized by the signal-to-noise ratio $\frac{\|\theta^*\|_2}{\sigma}$, and our analysis requires the SNR to be lower bounded as

$$\frac{\|\theta^*\|_2}{\sigma} > \eta, \tag{3.7}$$

for a sufficiently large constant $\eta > 0$. Past work by Redner and Walker [39] provides empirical evidence for the necessity of this assumption: for Gaussian mixtures with low SNR, they show that the ML solution has large variance, and furthermore verify empirically that the convergence of the EM algorithm can be quite slow. Other researchers [28, 51] also provide theoretical justification for the slow convergence of EM on poorly separated Gaussian mixtures.

With the signal-to-noise ratio lower bound $\eta$ defined above, we have the following guarantee.

COROLLARY 1 (Population result for the first-order EM algorithm for Gaussian mixtures). *Consider a Gaussian mixture model for which the SNR condition* (3.7) *holds for a sufficiently large* $\eta$, *and define the radius* $r = \frac{\|\theta^*\|_2}{4}$. *Then there is a contraction coefficient* $\kappa(\eta) \leq e^{-c\eta^2}$ *where* $c$ *is a universal constant such that for any initialization* $\theta^0 \in \mathbb{B}_2(r; \theta^*)$, *the population first-order EM iterates with stepsize* 1, *satisfy the bound*

$$\|\theta^t - \theta^*\|_2 \leq \kappa^t \|\theta^0 - \theta^*\|_2 \qquad \text{for all } t = 1, 2, \ldots. \tag{3.8}$$

REMARKS.

- The above corollary guarantees that when the SNR is sufficiently large, the population-level first-order EM algorithm converges to $\theta^*$ when initialized at any point in a ball of radius $\|\theta^*\|_2/4$ around $\theta^*$. Of course, an identical statement is true for the other global maximum at $-\theta^*$. At the population-level the log-likelihood function is not concave: it has two global maxima at $\theta^*$ and $-\theta^*$, a local minimum at 0 and a hyperplane of points that are attracted toward 0, that is, any point that is equi-distant from $\theta^*$ and $-\theta^*$ is a point of the population EM algorithm that is not attracted toward a global maximum. Observing that the all-zeroes vectors is also a fixed point of the (population) first-order EM algorithm—albeit a bad one—our corollary gives a characterization of the basin of attraction that is optimal up to the factor of $1/4$.
- In addition, the result shows that the first-order EM algorithm has two appealing properties: (a) as the mean separation grows, the initialization can be further away $\theta^*$ while retaining the global convergence guarantee; and (b) as the SNR grows, the first-order EM algorithm converges more rapidly. In particular, in a high SNR problem a few iterations of first-order EM suffice to obtain a solution that is very close to $\theta^*$. Both of these effects have been observed empirically in the work of Redner and Walker [39], and we give further evidence in our later simulations in Section 4. To the best of our knowledge, Corollary 1 provides the first rigorous theoretical characterization of this behavior.
- The proof of Corollary 1 involves establishing that for a sufficiently large SNR, the Gaussian mixture model satisfies the gradient smoothness, $\lambda$-strong concavity and $\mu$-smoothness (Conditions 1–3). We provide the body of the proof in Section 6.3.1, with the more technical details deferred to the Supplementary Material ([3], Appendix D).

3.2.2. *Mixture of regressions*. We now consider the mixture of regressions model, which is a latent variable extension of the usual regression model. In the standard linear regression model, we observe i.i.d. samples of the pair $(Y, X) \in \mathbb{R} \times \mathbb{R}^d$ linked via the equation

$$y_i = \langle x_i, \theta^* \rangle + v_i, \tag{3.9}$$

where $v_i \sim \mathcal{N}(0, \sigma^2)$ is the observation noise assumed to be independent of $x_i$. We assume a random design setting where $x_i \sim \mathcal{N}(0, I)$ are the design vectors and $\theta^* \in \mathbb{R}^d$ is the unknown regression vector to be estimated. In the mixture of regressions problem, there are two underlying choices of regression vector—say $\theta^*$ and $-\theta^*$—and we observe a pair $(y_i, x_i)$ drawn from the model (3.9) with probability $\frac{1}{2}$, and otherwise generated according to the alternative regression model $y_i = \langle x_i, -\theta^* \rangle + v_i$. Here, the hidden variables $\{z_i\}_{i=1}^n$ correspond to labels of the underlying regression model: say $z_i = 1$ when the data is generated according to the model (3.9), and $z_i = 0$ otherwise. Some recent work [13, 14, 53] has analyzed different methods for estimating a mixture of regressions. The work [14] analyzes a convex relaxation approach while the work [13] analyzes an estimator based on the method-of-moments. The work [53] focuses on the noiseless mixture of regressions problem (where $v_i = 0$), and provides analysis for an iterative algorithm in this context. In the symmetric form we consider, the mixture of regressions problem is also closely related to models for phase retrieval, albeit over $\mathbb{R}^d$, as considered in another line of recent work (e.g., [4, 10, 36]).

As in our analysis of the Gaussian mixture model, our theory applies when the signal-to-noise ratio is sufficiently large, as enforced by a condition of the form

$$\frac{\|\theta^*\|_2}{\sigma} > \eta, \tag{3.10}$$

for a sufficiently large constant $\eta > 0$. Under a suitable lower bound on this quantity, our first result guarantees that the first-order EM algorithm is locally convergent to the global optimum $\theta^*$ and provides a quantification of the local region of convergence.

COROLLARY 2 (Population result for the first-order EM algorithm for MOR). *Consider any mixture of regressions model satisfying the SNR condition* (3.10) *for a sufficiently large constant* $\eta$, *and define the radius* $r := \frac{\|\theta^*\|_2}{32}$. *Then for any* $\theta^0 \in \mathbb{B}_2(r; \theta^*)$, *the population first-order EM iterates with stepsize* 1, *satisfy the bound*

$$\|\theta^t - \theta^*\|_2 \leq \left(\tfrac{1}{2}\right)^t \|\theta^0 - \theta^*\|_2 \qquad \text{for } t = 1, 2, \ldots. \tag{3.11}$$

REMARKS.

- As with the Gaussian mixture model, the population likelihood has global maxima at $\theta^*$ and $-\theta^*$, and a local minimum at 0. Consequently, the largest Euclidean ball over which the iterates could converge to $\theta^*$ would have radius $\|\theta^*\|_2$. Thus, we see that our framework gives an order-optimal characterization of the region of convergence.[5]

[5]Possibly the factor $1/32$ could be sharpened with a more detailed analysis.

- Our analysis shows that the rate of convergence is again a decreasing function of the SNR parameter $\eta$. However, its functional form is not as explicit as in the Gaussian mixture case, so to simplify the statement, we used the fact that it is upper bounded by $1/2$. The proof of Corollary 2 involves verifying that the family of Q functions for the MOR model satisfies the required gradient smoothness, concavity and smoothness properties (Conditions 1 through 3). We provide the body of the argument in Section 6.3.2, with more technical aspects deferred to the Supplementary Material ([3], Appendix E).

3.2.3. *Linear regression with missing covariates*. Our first two examples involved mixture models in which the class membership variable was hidden. Another canonical use of the EM algorithm is in cases with corrupted or missing data. In this section, we consider a particular instantiation of such a problem, namely that of linear regression with the covariates missing completely at random.

In standard linear regression, we observe response–covariate pairs $(y_i, x_i) \in \mathbb{R} \times \mathbb{R}^d$ generated according to the linear model (3.9). In the missing data extension of this problem, instead of observing the covariate vector $x_i \in \mathbb{R}^d$ directly, we observe the corrupted version $\tilde{x}_i \in \mathbb{R}^d$ with components

$$\tilde{x}_{ij} = \begin{cases} x_{ij}, & \text{with probability } 1-\rho, \\ *, & \text{with probability } \rho, \end{cases} \tag{3.12}$$

where $\rho \in [0, 1)$ is the probability of missingness.

For this model, the key parameter is the probability $\rho \in [0, 1)$ that any given coordinate of the covariate vector is missing, and our analysis links this quantity to the signal-to-noise ratio and the radius of contractivity $r$, that is, the radius of the region around $\theta^*$ within which the population EM algorithm is convergent to a global optimum. Define

$$\xi_1 := \frac{\|\theta^*\|_2}{\sigma} \quad \text{and} \quad \xi_2 := \frac{r}{\sigma}. \tag{3.13a}$$

With this notation, our theory applies whenever the missing probability satisfies the bound

$$\rho < \frac{1}{1 + 2\xi(1+\xi)} \qquad \text{where } \xi := (\xi_1 + \xi_2)^2. \tag{3.13b}$$

COROLLARY 3 (Population contractivity for missing covariates). *Given any missing covariate regression model with missing probability $\rho$ satisfying the bound* (3.13b), *the first-order EM iterates with stepsize* 1, *satisfy the bound*

$$\left\|\theta^t - \theta^*\right\|_2 \leq \kappa^t \left\|\theta^0 - \theta^*\right\|_2 \qquad \text{for } t = 1, 2, \ldots, \tag{3.14}$$

*where* $\kappa \equiv \kappa(\xi, \rho) := (\frac{\xi + \rho(1+2\xi(1+\xi))}{1+\xi})$.

REMARKS.

- When the inequality (3.13b) holds, it can be verified that $\kappa(\xi, \rho)$ is strictly less than 1, which guarantees that the iterates converge at a geometric rate.
- Relative to our previous results, this corollary is somewhat unusual, in that we require an *upper bound* on the signal-to-noise ratio $\frac{\|\theta^*\|_2}{\sigma}$. Although this requirement might seem counter-intuitive at first sight, known minimax lower bounds on regression with missing covariates [26] show that it is unavoidable, that is, it is neither an artifact of our analysis nor a deficiency of the first-order EM algorithm. Intuitively, such a bound is required because as the norm $\|\theta^*\|_2$ increases, unlike in the mixture models considered previously, the amount of missing information increases in proportion to the amount of observed information. Figure 3 provides the results of simulations that confirm this behavior, in particular showing that for regression with missing data, the radius of convergence eventually decreases as $\|\theta^*\|_2$ grows.
- We provide the proof of this corollary in Section 6.3.3. Understanding the tightness of the above result remains an open problem. In particular, unlike in the mixture model examples, we do not know of a natural way to upper bound the radius of the region of convergence.

In conclusion, we have derived consequences of our main population-level result (Theorem 1) for three specific concrete models. In each of these examples, the auxiliary function $q$ is quadratic, so that verifying the strong concavity and

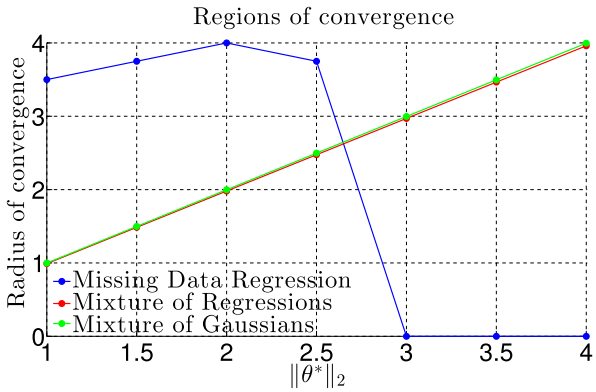


FIG. 3. *Simulations of the radius of convergence for the first-order EM algorithm for problems of dimension* $d = 10$, *sample size* $n = 1000$ *and variance* $\sigma^2 = 1$. *Radius of convergence is defined as the maximum value of* $\|\theta^0 - \theta^*\|_2$ *for which initialization at* $\theta^0$ *leads to convergence to an optimum near* $\theta^*$. *Consistent with the theory, for both the Gaussian mixture and mixture of regression models, the radius of convergence grows with* $\|\theta^*\|_2$. *In contrast, in the missing data case* (*here with* $\rho = 0.2$), *increasing* $\|\theta^*\|_2$ *can cause the EM algorithm to converge to bad local optima, which is consistent with the prediction of Corollary* 3.

smoothness examples is relatively straightforward. In contrast, verifying the gradient smoothness (GS) bound in Condition 1 requires substantially more effort. We believe that the GS condition is a canonical concept in the understanding of EM-type iterations, as evidenced by its role in highlighting critical problem dependent quantities—such as signal-to-noise ratio and probability of missingness—that determine the region of attraction for global maxima of the population likelihood.

**4. Analysis of sample-based first-order EM updates.** Up to this point, we have analyzed the first-order EM updates at the population level (2.6), whereas in practice, the algorithm is applied with a finite set of samples. Accordingly, we now turn to theoretical guarantees for the sample-based first-order EM updates (2.3). As discussed in Section 2.4, the main challenge here is in controlling the empirical process defined by the difference between the sample-based and population-level updates.

4.1. *Standard form of sample-based first-order EM.* Recalling the definition (2.1) of the sample based $Q$-function, we are interested in the behavior of the recursion

$$\theta^{t+1} = \theta^t + \alpha \nabla Q_n\big(\theta|\theta^t\big)|_{\theta=\theta^t}, \tag{4.1}$$

where $\alpha > 0$ is an appropriately chosen stepsize. As mentioned previously, we need to control the deviations of the sample gradient $\nabla Q_n$ from the population version $\nabla Q$. Accordingly, for a given sample size $n$ and tolerance parameter $\delta \in (0, 1)$, we let $\varepsilon_Q^{\text{unif}}(n, \delta)$ be the smallest scalar such that

$$\sup_{\theta \in \mathbb{B}_2(r;\theta^*)} \big\|\nabla Q_n(\theta|\theta) - \nabla Q(\theta|\theta)\big\|_2 \leq \varepsilon_Q^{\text{unif}}(n, \delta) \tag{4.2}$$

with probability at least $1 - \delta$.

Our first main result on the performance of the sample-based first-order EM algorithm depends on the same assumptions as Theorem 1: namely, that there exists a radius $r > 0$ and a triplet $(\gamma, \lambda, \mu)$ with $0 \leq \gamma < \lambda \leq \mu$ such that the gradient smoothness, strong-concavity and smoothness conditions hold (Conditions 1–3), and that we implement the algorithm with stepsize $\alpha = \frac{2}{\mu+\lambda}$.

THEOREM 2. *Suppose that, in addition to the conditions of Theorem 1, the sample size $n$ is large enough to ensure that*

$$\varepsilon_Q^{\text{unif}}(n, \delta) \leq (\lambda - \gamma) r. \tag{4.3}$$

*Then with probability at least $1 - \delta$, given any initial vector $\theta^0 \in \mathbb{B}_2(r; \theta^*)$, the finite-sample first-order EM iterates $\{\theta^t\}_{t=0}^{\infty}$ satisfy the bound*

$$\big\|\theta^t - \theta^*\big\|_2 \leq \left(1 - \frac{2\lambda - 2\gamma}{\mu + \lambda}\right)^t \big\|\theta^0 - \theta^*\big\|_2 + \frac{\varepsilon_Q^{\text{unif}}(n, \delta)}{\lambda - \gamma} \tag{4.4}$$

*for all $t = 1, 2, \ldots$.*

REMARKS.

- This result leverages the population-level result in Theorem 1. It is particularly crucial that we have linear convergence at the population level, since this ensures that errors made at each iteration, which are bounded by $\varepsilon_Q^{\text{unif}}(n, \delta)$ with probability at least $1 - \delta$, do not accumulate too fast. The bound in equation (4.3) ensures that the iterates of the finite-sample first-order EM algorithm remain in $\mathbb{B}_2(r; \theta^*)$ with the same probability.
- Note that the bound (4.4) involves two terms, the first of which decreases geometrically in the iteration number $t$, whereas the second is independent of $t$. Thus, we are guaranteed that the iterates converge geometrically to a ball of radius $\mathcal{O}(\varepsilon_Q^{\text{unif}}(n, \delta))$. See Figure 4 for an illustration of this guarantee. In typical examples, we show that $\varepsilon_Q^{\text{unif}}(n, \delta)$ is on the order of the minimax rate for estimating $\theta^*$. For the $d$-dimensional parametric problems considered in this paper, the minimax rate typically scales as $\mathcal{O}(\sqrt{d/n})$. In these cases, Theorem 2 guarantees that the first-order EM algorithm, when initialized in $\mathbb{B}_2(r; \theta^*)$, converges rapidly to a point that is within the minimax distance of the unknown true parameter.
- For a fixed sample size $n$, the bound (4.4) suggests a reasonable choice of the number of iterations. In particular, letting $\kappa = 1 - \frac{2\lambda - 2\gamma}{\mu + \lambda}$, consider any positive

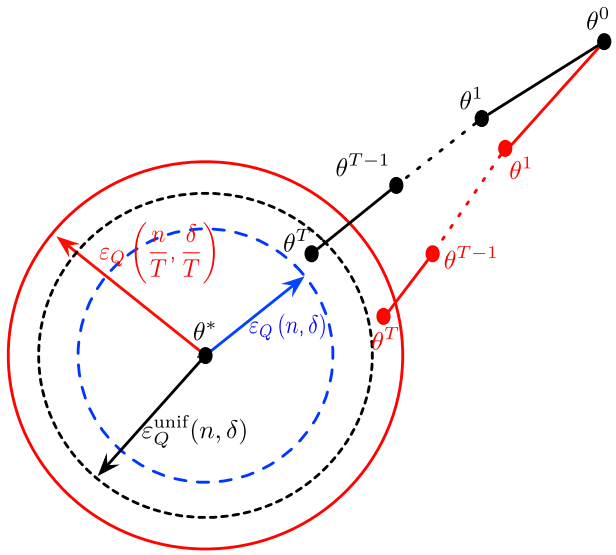


FIG. 4. *An illustration of Theorems 2 and 3. The first part of the theorem describes the geometric convergence of iterates of the EM algorithm to the ball of radius* $\mathcal{O}(\varepsilon_Q^{\text{unif}}(n, \delta))$ *(in black). The second part describes the geometric convergence of the sample-splitting EM algorithm to the ball of radius* $\mathcal{O}(\varepsilon_Q(n/T, \delta/T))$ *(in red). In typical examples, the ball to which sample-splitting EM converges is only a logarithmic factor larger than the ball* $\mathcal{O}(\varepsilon_Q(n, \delta))$ *(in blue).*

integer $T$ such that

$$T \geq \log_{1/\kappa} \frac{(\lambda - \gamma)\|\theta^0 - \theta^*\|_2}{\varepsilon_Q^{\text{unif}}(n, \delta)}. \tag{4.5}$$

As will be clarified in the sequel, such a choice of $T$ exists in various concrete models considered here. This choice ensures that the first term in the bound (4.4) is dominated by the second term, and hence that

$$\|\theta^T - \theta^*\|_2 \leq \frac{2\varepsilon_Q^{\text{unif}}(n, \delta)}{\lambda - \gamma} \qquad \text{with probability at least } 1 - \delta. \tag{4.6}$$

4.2. *Sample-splitting in first-order EM.* In this section, we consider the finite-sample performance of a variant of the first-order EM algorithm that uses a *fresh batch* of samples for each iteration. Although we introduce the sample-splitting variant primarily for theoretical convenience, there are also some potential practical advantages, such as computational savings from having a smaller data set per update. A disadvantage is that it can be difficult to correctly specify the number of iterations in advance, and the first-order EM algorithm that uses sample-splitting is likely to be less efficient from a statistical standpoint. Indeed, in our theory, the statistical guarantees are typically weaker by a logarithmic factor in the total sample size $n$.

Formally, given a total of $n$ samples and $T$ iterations, suppose that we divide the full data set into $T$ subsets of size $\lfloor n/T \rfloor$, and then perform the updates

$$\theta^{t+1} = \theta^t + \alpha \nabla Q_{\lfloor n/T \rfloor}(\theta|\theta^t)|_{\theta=\theta^t}, \tag{4.7}$$

where $\nabla Q_{\lfloor n/T \rfloor}$ denotes the $Q$-function computed using a fresh subset of $\lfloor n/T \rfloor$ samples at each iteration. For a given sample size $n$ and tolerance parameter $\delta \in (0, 1)$, we let $\varepsilon_Q(n, \delta)$ be the smallest scalar such that, for any *fixed* $\theta \in \mathbb{B}_2(r; \theta^*)$,

$$\mathbb{P}\big[\|\nabla Q_n(\theta|\theta^t)|_{\theta=\theta^t} - \nabla Q(\theta|\theta^t)|_{\theta=\theta^t}\|_2 > \varepsilon_Q(n, \delta)\big] \leq 1 - \delta. \tag{4.8}$$

The quantity $\varepsilon_Q$ provides a bound that needs only to hold pointwise for each $\theta \in \mathbb{B}_2(r; \theta^*)$, as opposed to the quantity $\varepsilon_Q^{\text{unif}}$ for which the bound (4.2) must hold uniformly over all $\theta$. Due to this difference, establishing bounds on $\varepsilon_Q(n, \delta)$ can be significantly easier than bounding $\varepsilon_Q^{\text{unif}}(n, \delta)$.

Our theory for the iterations (4.7) applies under the same conditions as Theorem 1: namely, for some radius $r > 0$, and a triplet $(\gamma, \lambda, \mu)$ such that $0 \leq \gamma < \lambda \leq \mu$, the gradient smoothness, concavity and smoothness properties (Conditions 1–3) hold, and the stepsize is chosen as $\alpha = \frac{2}{\mu + \lambda}$.

THEOREM 3. *Suppose that, in addition to the conditions of Theorem 1, the sample size n is large enough to ensure that*

$$\varepsilon_Q\left(\frac{n}{T}, \frac{\delta}{T}\right) \leq (\lambda - \gamma)r. \tag{4.9a}$$

*Then with probability at least* $1-\delta$, *given any initial vector* $\theta^0 \in \mathbb{B}_2(r;\theta^*)$, *the sample-splitting first-order EM iterates satisfy the bound*

$$\|\theta^t - \theta^*\|_2 \le \left(1 - \frac{2\lambda - 2\gamma}{\mu + \lambda}\right)^t \|\theta^0 - \theta^*\|_2 + \frac{\varepsilon_Q(n/T, \delta/T)}{\lambda - \gamma}. \tag{4.9b}$$

See Appendix B.2 for the proof of this result. It has similar flavor to the guarantee of Theorem 2, but requires a number of iterations $T$ to be specified beforehand. The optimal choice of $T$ balances the two terms in the bound. As will be clearer in the sequel, in typical cases the optimal choice of $T$ will depend logarithmically in $\varepsilon_Q$. Each iteration uses roughly $n/\log n$ samples, and the iterates converge to a ball of correspondingly larger radius.

4.3. *Finite-sample consequences for specific models*. We now state some consequences of Theorems 2 and 3 for the three models previously considered at the population-level in Section 3.2.

4.3.1. *Mixture of gaussians*. We begin by analyzing the sample-based first-order EM updates (4.1) for the Gaussian mixture model, as previously introduced in Section 3.2.1, where we showed in Corollary 1 that the population iterates converge geometrically given a lower bound on the signal-to-noise ratio $\frac{\|\theta^*\|_2}{\sigma}$. In this section, we provide an analogous guarantee for the sample-based updates, again with a stepsize $\alpha = 1$. See Appendix A for derivation of the specific form of the first-order EM updates for this model.

Our guarantee involves the function $\varphi(\sigma; \|\theta^*\|_2) := \|\theta^*\|_2(1 + \frac{\|\theta^*\|_2^2}{\sigma^2})$, as well as positive universal constants $(c, c_1, c_2)$.

COROLLARY 4 (Sample-based first-order EM guarantees for Gaussian mixture). *In addition to the conditions of Corollary* 1, *suppose that the sample size is lower bounded as* $n \ge c_1 d \log(1/\delta)$. *Then given any initialization* $\theta^0 \in \mathbb{B}_2(\frac{\|\theta^*\|_2}{4}; \theta^*)$, *there is a contraction coefficient* $\kappa(\eta) \le e^{-c\eta^2}$ *such that the first-order EM iterates* $\{\theta^t\}_{t=0}^{\infty}$ *satisfy the bound*

$$\|\theta^t - \theta^*\|_2 \le \kappa^t \|\theta^0 - \theta^*\|_2 + \frac{c_2}{1-\kappa}\varphi(\sigma; \|\theta^*\|_2)\sqrt{\frac{d}{n}\log(1/\delta)} \tag{4.10}$$

*with probability at least* $1-\delta$.

REMARKS.

- We provide the proof of this result in Section 6.4.1, with some of the more technical aspects deferred to the Supplementary Material ([3], Appendix D). In the

supplement ([3], Corollary 8) we also give guarantees for the EM updates with sample-splitting, as described in equation (4.7) for the first-order EM algorithm. These results have better dependence on $\|\theta^*\|_2$ and $\sigma$, but the sample size requirement is greater by a logarithmic factor.

- It is worth comparing with a related result of Dasgupta and Schulman [16] on estimating Gaussian mixture models. They show that when the SNR is sufficiently high—scaling roughly as $d^{1/4}$—then a modified EM algorithm, with an intermediate pruning step, reaches a near-optimal solution in two iterations. On one hand, the SNR condition in our corollary is significantly weaker, requiring only that it is larger than a fixed constant independent of dimension (as opposed to scaling with $d$), but their theory is developed for more general $k$-mixtures.
- The bound (4.10) provides a rough guide of how many iterations are required in order to achieve an estimation error of order $\sqrt{d/n}$, corresponding to the minimax rate. In particular, consider the smallest positive integer such that

$$T \geq \log_{1/\kappa}\left(\frac{\|\theta^0 - \theta^*\|_2(1-\kappa)}{\varphi(\sigma; \|\theta^*\|_2)}\sqrt{\frac{n}{d}\frac{1}{\log(1/\delta)}}\right). \tag{4.11a}$$

With this choice, we are guaranteed that the iterate $\theta^T$ satisfies the bound

$$\|\theta^T - \theta^*\|_2 \leq \frac{(1+c_2)\varphi(\sigma; \|\theta^*\|_2)}{1-\kappa}\sqrt{\frac{d}{n}\log(1/\delta)} \tag{4.11b}$$

with probability at least $1-\delta$. To be fair, the iteration choice (4.11a) is not computable based only on data, since it depends on unknown quantities such as $\theta^*$ and the contraction coefficient $\kappa$. However, as a rough guideline, it shows that the number of iterations to be performed should grow logarithmically in the ratio $n/d$.

- Corollary 4 makes a number of qualitative predictions that can be tested. To begin, it predicts that the statistical error $\|\theta^t - \theta^*\|_2$ should decrease geometrically, and then level off at a plateau. Figure 5 shows the results of simulations designed to test this prediction: for dimension $d = 10$ and sample size $n = 1000$, we performed 10 trials with the standard EM updates applied to Gaussian mixture models with SNR $\frac{\|\theta^*\|_2}{\sigma} = 2$. In panel (a), the red curves plot the log statistical error versus the iteration number, whereas the blue curves show the log optimization error versus iteration. As can be seen by the red curves, the statistical error decreases geometrically before leveling off at a plateau. On the other hand, the optimization error decreases geometrically to numerical tolerance. Panel (b) shows that the first-order EM updates have a qualitatively similar behavior for this model, although the overall convergence rate appears to be slower.

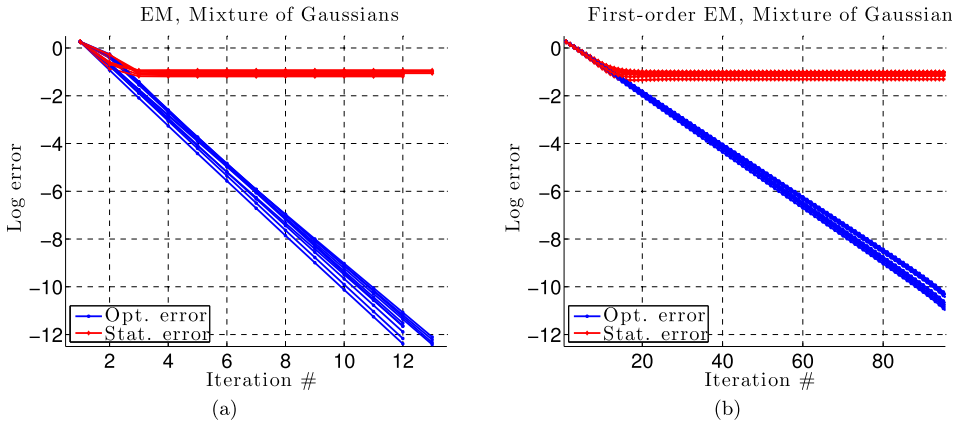


FIG. 5. *Plots of the iteration number versus log optimization error* $\log(\|\theta^t - \widehat{\theta}\|_2)$ *and log statistical error* $\log(\|\theta^t - \theta^*\|_2)$. (a) *Results for the EM algorithm*[6]. (b) *Results for the first-order EM algorithm. Each plot shows* 10 *different problem instances with dimension* $d = 10$, *sample size* $n = 1000$ *and signal-to-noise ratio* $\frac{\|\theta^*\|_2}{\sigma} = 2$. *The optimization error decays geometrically up to numerical precision, whereas the statistical error decays geometrically before leveling off.*

- In conjunction with Corollary 1, Corollary 4 also predicts that the convergence rate should increase as the signal-to-noise ratio $\frac{\|\theta^*\|_2}{\sigma}$ is increased. Figure 6 shows the results of simulations designed to test this prediction: again, for mixture models with dimension $d = 10$ and sample size $n = 1000$, we applied the standard EM updates to Gaussian mixture models with varying SNR $\frac{\|\theta^*\|_2}{\sigma}$. For each choice of SNR, we performed 10 trials, and plotted the log optimization error $\log \|\theta^t - \widehat{\theta}\|_2$ versus the iteration number. As expected, the convergence rate is geometric (linear on this logarithmic scale), and the rate of convergence increases as the SNR grows.[7]

4.3.2. *Mixture of regressions.* Recall the mixture of regressions (MOR) model previously introduced in Section 3.2.2. In this section, we analyze the sample-splitting first-order EM updates (4.7) for the MOR model. See Appendix A for a derivation of the specific form of the updates for this model. Our result involves the quantity $\varphi(\sigma; \|\theta^*\|_2) = \sqrt{\sigma^2 + \|\theta^*\|_2^2}$, along with positive universal constants $(c_1, c_2)$.

[6]In this and subsequent figures, we show simulations for the standard (i.e., not sample-splitting) versions of the EM and first-order EM algorithms.

[7]To be clear, Corollary 4 predicts geometric convergence of the statistical error $\|\theta^t - \theta^*\|_2$, whereas these plots show the optimization error $\|\theta^t - \widehat{\theta}\|_2$. However, the analysis underlying Corollary 4 can also be used to show geometric convergence of the optimization error.

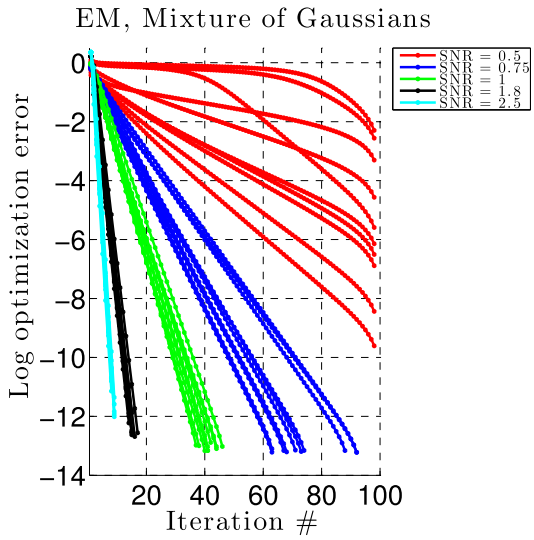


FIG. 6. *Plot of the iteration number versus the (log) optimization error* $\log(\|\theta^t - \widehat{\theta}\|_2)$[8] *for different values of the SNR* $\frac{\|\theta^*\|_2}{\sigma}$. *For each SNR, we performed* 10 *independent trials of a Gaussian mixture model with dimension* $d = 10$ *and sample size* $n = 1000$. *Larger values of SNR lead to faster convergence rates, consistent with Corollaries* 4 *and* 7.

COROLLARY 5 (Sample-splitting first-order EM guarantees for MOR). *In addition to the conditions of Corollary* 2, *suppose that the sample size is lower bounded as* $n \geq c_1 d \log(T/\delta)$. *Then there is a contraction coefficient* $\kappa \leq 1/2$ *such that, for any initial vector* $\theta^0 \in \mathbb{B}_2(\frac{\|\theta^*\|_2}{32}; \theta^*)$, *the sample-splitting first-order EM iterates* (4.7) *with stepsize* 1, *based on* $n/T$ *samples per step satisfy the bound*

$$\|\theta^t - \theta^*\|_2 \leq \kappa^t \|\theta^0 - \theta^*\|_2 + c_2 \varphi(\sigma; \|\theta^*\|_2) \sqrt{\frac{d}{n} T \log(T/\delta)} \tag{4.12}$$

*with probability at least* $1 - \delta$.

REMARKS.

- See Section 6.4.3 for the proof of this claim. As with Corollary 4, the bound (4.12) again provides guidance on the number of iterations to perform: in particular, for a given sample size $n$, suppose we perform $T = \lceil \log(n/d\varphi^2(\sigma; \|\theta^*\|_2)) \rceil$ iterations. The bound (4.12) then implies that

$$\|\theta^T - \theta^*\|_2 \leq c_3 \varphi(\sigma; \|\theta^*\|_2) \sqrt{\frac{d}{n} \log^2\left(\frac{n}{d\varphi^2(\sigma; \|\theta^*\|_2)}\right) \log(1/\delta)} \tag{4.13}$$

[8] The fixed point $\widehat{\theta}$ is determined by running the algorithm to convergence up to machine precision.

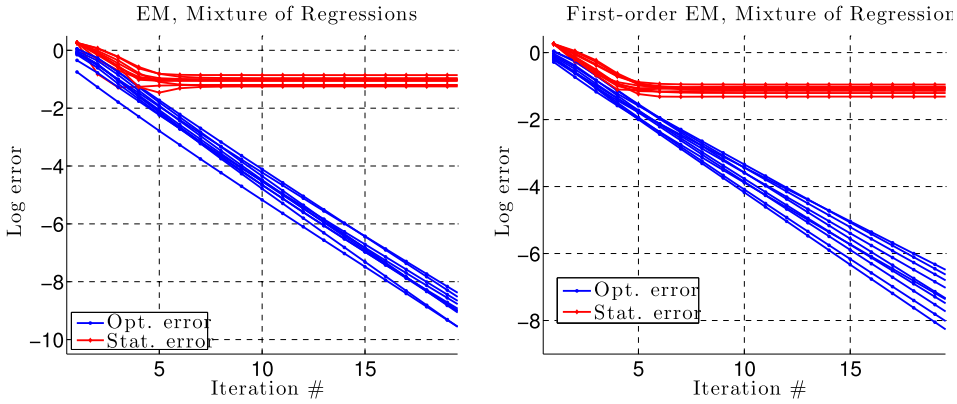


FIG. 7. *Plots of the iteration number versus log optimization error* $\log(\|\theta^t - \widehat{\theta}\|_2)$ *and log statistical error* $\log(\|\theta^t - \theta^*\|_2)$ *for mixture of regressions.* (a) *Results for the EM algorithm.* (b) *Results for the first-order EM algorithm. Each plot shows* 10 *independent trials with* $d = 10$, *sample size* $n = 1000$, *and signal-to-noise ratio* $\frac{\|\theta^*\|_2}{\sigma} = 2$. *In both plots, the optimization error decays geometrically while the statistical error decays geometrically before leveling off.*

with probability at least $1 - \delta$. Apart from the logarithmic penalty $\log^2(\frac{n}{d\varphi^2(\sigma;\|\theta^*\|_2)})$, this guarantee matches the minimax rate for estimation of a $d$-dimensional regression vector. We note that the logarithmic penalty can be removed by instead analyzing the standard form of the first-order EM updates, as we did for the Gaussian mixture model.

- As with Corollary 4, this corollary predicts that the statistical error $\|\theta^t - \theta^*\|_2$ should decrease geometrically, and then level off at a plateau. Figure 7 shows the results of simulations designed to test this prediction: see the caption for the details.

4.3.3. *Linear regression with missing covariates.* Recall the problem of linear regression with missing covariates, as previously described in Section 3.2.3. In this section, we analyze the sample-splitting version (4.7) version of the first-order EM updates. See Appendix A for the derivation of the concrete form of these updates for this specific model.

COROLLARY 6 (Sample-splitting first-order EM guarantees for missing covariates). *In addition to the conditions of Corollary 3, suppose that the sample size is lower bounded as* $n \geq c_1 d \log(1/\delta)$. *Then there is a contraction coefficient* $\kappa < 1$ *such that, for any initial vector* $\theta^0 \in \mathbb{B}_2(\xi_2\sigma; \theta^*)$, *the sample-splitting first-order EM iterates* (4.7) *with stepsize* 1, *based on* $n/T$ *samples per iteration satisfy the bound*

$$\|\theta^t - \theta^*\|_2 \leq \kappa^t \|\theta^0 - \theta^*\|_2 + \frac{c_2\sqrt{1+\sigma^2}}{1-\kappa}\sqrt{\frac{d}{n}T\log(T/\delta)} \tag{4.14}$$

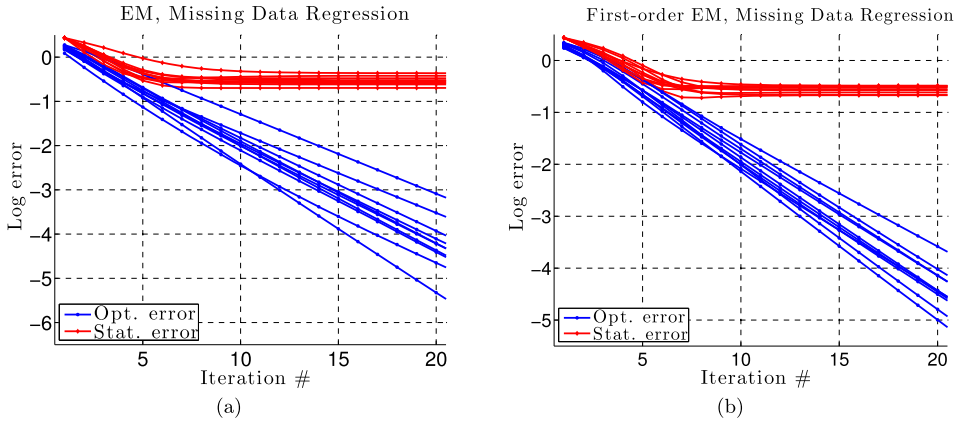


FIG. 8. *Plots of the iteration number versus log optimization error* $\log(\|\theta^t - \widehat{\theta}\|_2)$ *and log statistical error* $\log(\|\theta^t - \theta^*\|_2)$ *for regression with missing covariates.* (a) *Results for the EM algorithm.* (b) *Results for the first-order EM algorithm. Each plot shows* 10 *different problem instances of dimension* $d = 10$, *sample size* $n = 1000$, *signal-to-noise ratio* $\frac{\|\theta^*\|_2}{\sigma} = 2$, *and missing probability* $\rho = 0.2$. *In both plots, the optimization error decays geometrically while the statistical error decays geometrically before leveling off.*

*with probability at least* $1 - \delta$.

We prove this corollary in the Supplementary Material ([3], Appendix 6.4.3). We note that the constant $c_2$ is a monotonic function of the parameters $(\xi_1, \xi_2)$, but does not otherwise depend on $n$, $d$, $\sigma^2$ or other problem-dependent parameters.

REMARK. As with Corollary 9, this result provides guidance on the appropriate number of iterations to perform: in particular, if we set $T = c \log n$ for a sufficiently large constant $c$, then the bound (4.14) implies that

$$\|\theta^T - \theta^*\|_2 \leq c' \sqrt{1 + \sigma^2} \sqrt{\frac{d}{n} \log^2(n/\delta)}$$

with probability at least $1 - \delta$. This is illustrated in Figure 8. Modulo the logarithmic penalty in $n$, incurred due to the sample-splitting, this estimate achieves the optimal $\sqrt{\frac{d}{n}}$ scaling of the $\ell_2$-error.

**5. Extension of results to the EM algorithm.** In this section, we develop unified population and finite-sample results for the EM algorithm. Particularly, at the population-level we show in Theorem 4 that a closely related condition to the GS condition can be used to give a bound on the region and rate of convergence of the EM algorithm. Our next main result shows how to leverage this population-level result along with control on an appropriate empirical process in order to provide nonasymptotic finite-sample guarantees.

5.1. *Analysis of the EM algorithm at the population level.* We assume throughout this section that the function $q$ is $\lambda$-*strongly concave* (but not necessarily smooth). For any fixed $\theta$, in order to relate the population EM updates to the fixed point $\theta^*$, we require control on the two gradient mappings $\nabla q(\cdot) = \nabla Q(\cdot|\theta^*)$ and $\nabla Q(\cdot|\theta)$. These mappings are central in characterizing the fixed point $\theta^*$ and the EM update. In order to compactly represent the EM update, we define the operator $M : \Omega \to \Omega$,

$$M(\theta) = \arg\max_{\theta' \in \Omega} Q\big(\theta'|\theta\big). \tag{5.1}$$

Using this notation, the EM algorithm given some initialization $\theta^0$, produces a sequence of iterates $\{\theta^t\}_{t=0}^{\infty}$, where $\theta^{t+1} = M(\theta^t)$.

By virtue of the self-consistency property (2.7) and the convexity of $\Omega$, the fixed point satisfies the first-order optimality (KKT) condition

$$\big\langle \nabla Q\big(\theta^*|\theta^*\big), \theta' - \theta^* \big\rangle \leq 0 \qquad \text{for all } \theta' \in \Omega. \tag{5.2}$$

Similarly, for any $\theta \in \Omega$, since $M(\theta)$ maximizes the function $\theta' \mapsto Q(\theta'|\theta)$ over $\Omega$, we have

$$\big\langle \nabla Q\big(M(\theta)|\theta\big), \theta' - \theta \big\rangle \leq 0 \qquad \text{for all } \theta' \in \Omega. \tag{5.3}$$

We note that for unconstrained problems, the terms $\nabla Q(\theta^*|\theta^*)$ and $\nabla Q(M(\theta)|\theta)$ will be equal to zero, but we retain the forms of equations (5.2) and (5.3) to make the analogy with the GS condition clearer.

Equations (5.2) and (5.3) are sets of inequalities that *characterize* the points $M(\theta)$ and $\theta^*$. Thus, at an intuitive level, in order to establish that $\theta^{t+1}$ and $\theta^*$ are close, it suffices to verify that these two characterizations are close in a suitable sense. We also note that inequalities similar to the condition (5.3) are often used as a starting point in the classical analysis of M-estimators (e.g., see van de Geer [45]). In the analysis of EM, we obtain additional leverage from the condition (5.2) that characterizes $\theta^*$.

With this intuition in mind, we introduce the following regularity condition in order to relate conditions (5.3) and (5.2): The condition involves a Euclidean ball of radius $r$ around the fixed point $\theta^*$, given by

$$\mathbb{B}_2\big(r; \theta^*\big) := \big\{\theta \in \Omega \big| \big\|\theta - \theta^*\big\|_2 \leq r\big\}. \tag{5.4}$$

DEFINITION 4 [First-order stability (FOS)]. The functions $\{Q(\cdot|\theta), \theta \in \Omega\}$ satisfy condition FOS($\gamma$) over $\mathbb{B}_2(r; \theta^*)$ if

$$\big\|\nabla Q\big(M(\theta)|\theta^*\big) - \nabla Q\big(M(\theta)|\theta\big)\big\|_2 \leq \gamma \big\|\theta - \theta^*\big\|_2 \quad \text{for all } \theta \in \mathbb{B}_2\big(r; \theta^*\big). \tag{5.5}$$

To provide some high-level intuition, observe the condition (5.5) is always satisfied at the fixed point $\theta^*$, in particular with parameter $\gamma = 0$. Intuitively then, by allowing for a strictly positive parameter $\gamma$, one might expect that this condition would hold in a local neighborhood $\mathbb{B}_2(r;\theta^*)$ of the fixed point $\theta^*$, as long as the functions $Q(\cdot|\theta)$ and the map $M$ are sufficiently regular. As before with the GS condition, we show in the sequel that every point around $\theta^*$ for which the FOS condition holds (with an appropriate $\gamma$) is in the region of attraction of $\theta^*$—the population EM update produces an iterate closer to $\theta^*$ than the original point.

Formally, under the conditions we have introduced, the following result guarantees that the population EM operator is locally contractive.

THEOREM 4. *For some radius $r > 0$ and pair $(\gamma, \lambda)$ such that $0 \le \gamma < \lambda$, suppose that the function $Q(\cdot|\theta^*)$ is* globally $\lambda$*-strongly concave* (3.2)*, and that the FOS$(\gamma)$ condition* (5.5) *holds on the ball $\mathbb{B}_2(r;\theta^*)$. Then the population EM operator $M$ is contractive over $\mathbb{B}_2(r;\theta^*)$, in particular with*

$$\|M(\theta) - \theta^*\|_2 \le \frac{\gamma}{\lambda}\|\theta - \theta^*\|_2 \qquad \text{for all } \theta \in \mathbb{B}_2(r;\theta^*).$$

The proof is a consequence of the KKT conditions from equations (5.2) and (5.3), along with consequences of the strong concavity of $Q(\cdot|\theta^*)$. We defer a detailed proof to Appendix B.1.

REMARKS. As an immediate consequence, under the conditions of the theorem, for any initial point $\theta^0 \in \mathbb{B}_2(r;\theta^*)$, the population EM sequence $\{\theta^t\}_{t=0}^{\infty}$ exhibits linear convergence, namely

$$\|\theta^t - \theta^*\|_2 \le \left(\frac{\gamma}{\lambda}\right)^t \|\theta^0 - \theta^*\|_2 \qquad \text{for all } t = 1, 2, \ldots. \tag{5.6}$$

5.2. *Finite-sample analysis for the EM algorithm.* We now turn to theoretical results on the sample-based version of the EM algorithm. More specifically, we define the sample-based operator $M_n : \Omega \to \Omega$,

$$M_n(\theta) = \arg\max_{\theta' \in \Omega} Q_n(\theta'|\theta), \tag{5.7}$$

where the sample-based $Q$-function was defined previously in equation (2.1). Analogous to the situation with the first-order EM algorithm we also consider a sample-splitting version of the EM algorithm, in which given a total of $n$ samples and $T$ iterations, we divide the full data set into $T$ subsets of size $\lfloor n/T \rfloor$, and then perform the updates $\theta^{t+1} = M_{n/T}(\theta^t)$, using a fresh subset of samples at each iteration.

For a given sample size $n$ and tolerance parameter $\delta \in (0, 1)$, we let $\varepsilon_M(n, \delta)$ be the smallest scalar such that, for any fixed $\theta \in \mathbb{B}_2(r;\theta^*)$, we have

$$\|M_n(\theta) - M(\theta)\|_2 \le \varepsilon_M(n, \delta) \tag{5.8}$$

with probability at least $1 - \delta$. This tolerance parameter (5.8) enters our analysis of the sample-splitting form of EM. On the other hand, in order to analyze the standard sample-based form of EM, we require a stronger condition, namely one in which the bound (5.8) holds uniformly over the ball $\mathbb{B}_2(r; \theta^*)$. Accordingly, we let $\varepsilon_M^{\text{unif}}(n, \delta)$ be the smallest scalar for which

$$\sup_{\theta \in \mathbb{B}_2(r;\theta^*)} \|M_n(\theta) - M(\theta)\|_2 \leq \varepsilon_M^{\text{unif}}(n, \delta) \tag{5.9}$$

with probability at least $1 - \delta$. With these definitions, we have the following guarantees.

THEOREM 5. *Suppose that the population EM operator $M : \Omega \to \Omega$ is contractive with parameter $\kappa \in (0, 1)$ on the ball $\mathbb{B}_2(r; \theta^*)$, and the initial vector $\theta^0$ belongs to $\mathbb{B}_2(r; \theta^*)$.*

(a) *If the sample size $n$ is large enough to ensure that*

$$\varepsilon_M^{\text{unif}}(n, \delta) \leq (1 - \kappa)r, \tag{5.10a}$$

*then the EM iterates $\{\theta^t\}_{t=0}^{\infty}$ satisfy the bound*

$$\|\theta^t - \theta^*\|_2 \leq \kappa^t \|\theta^0 - \theta^*\|_2 + \frac{1}{1-\kappa} \varepsilon_M^{\text{unif}}(n, \delta) \tag{5.10b}$$

*with probability at least $1 - \delta$.*

(b) *For a given iteration number $T$, suppose the sample size $n$ is large enough to ensure that*

$$\varepsilon_M\left(\frac{n}{T}, \frac{\delta}{T}\right) \leq (1 - \kappa)r. \tag{5.11a}$$

*Then the sample-splitting EM iterates $\{\theta^t\}_{t=0}^{T}$ based on $\frac{n}{T}$ samples per round satisfy the bound*

$$\|\theta^t - \theta^*\|_2 \leq \kappa^t \|\theta^0 - \theta^*\|_2 + \frac{1}{1-\kappa} \varepsilon_M\left(\frac{n}{T}, \frac{\delta}{T}\right). \tag{5.11b}$$

We provide a detailed proof of this Theorem in Appendix B.

REMARKS. In order to obtain readily interpretable bounds for specific models, it only remains to establish the $\kappa$-contractivity of the population operator, and to compute either the function $\varepsilon_M$ or the function $\varepsilon_M^{\text{unif}}$. In the Supplementary Material, we revisit each of the three examples considered in this paper, and provide population and finite-sample guarantees for the EM algorithm.

**6. Proofs.** In this section, we provide proofs of some of our previously stated results, beginning with Theorems 1 and 2, followed by the proofs of Corollaries 1 through 3.

6.1. *Proof of Theorem* 1. This proof relies on a classical result that ensures linear convergence of gradient ascent when applied to a smooth and strongly concave function (see, e.g., [7, 8, 35]).

LEMMA 1. *For a function $q$ with the $\lambda$-strong concavity and $\mu$-smoothness properties* (*Conditions* 2 *and* 3), *the oracle iterates* (2.9) *with stepsize $\alpha = \frac{2}{\mu+\lambda}$ are linearly convergent*:

$$\|\theta^t + \alpha \nabla q(\theta)|_{\theta=\theta^t} - \theta^*\|_2 \leq \left(\frac{\mu-\lambda}{\mu+\lambda}\right)\|\theta^t - \theta^*\|_2. \tag{6.1}$$

Taking this result as given, we can now prove the theorem. By definition of the first-order EM update (2.6), we have

$$\begin{aligned}
&\|\theta^t + \alpha \nabla Q(\theta|\theta^t)|_{\theta=\theta^t} - \theta^*\|_2 \\
&\quad = \|\theta^t + \alpha \nabla q(\theta)|_{\theta=\theta^t} - \alpha \nabla q(\theta)|_{\theta=\theta^t} + \alpha \nabla Q(\theta|\theta^t)|_{\theta=\theta^t} - \theta^*\|_2 \\
&\quad \overset{\text{(i)}}{\leq} \|\theta^t + \alpha \nabla q(\theta)|_{\theta=\theta^t} - \theta^*\|_2 + \alpha\|\nabla q(\theta)|_{\theta=\theta^t} - \nabla Q(\theta|\theta^t)|_{\theta=\theta^t}\|_2 \\
&\quad \overset{\text{(ii)}}{\leq} \left(\frac{\mu-\lambda}{\mu+\lambda}\right)\|\theta^t - \theta^*\|_2 + \alpha\gamma\|\theta^t - \theta^*\|_2,
\end{aligned}$$

where step (i) follows from the triangle inequality, and step (ii) uses Lemma 1 and condition GS. Substituting $\alpha = \frac{2}{\mu+\lambda}$ and performing some algebra yields the claim.

6.2. *Proof of Theorem* 2. With probability at least $1-\delta$ we have that for any $\theta^s \in \mathbb{B}_2(r;\theta^*)$,

$$\|\nabla Q_n(\theta|\theta^s)|_{\theta=\theta^s} - \nabla Q(\theta|\theta^s)|_{\theta=\theta^s}\|_2 \leq \varepsilon_Q^{\text{unif}}(n,\delta). \tag{6.2}$$

We perform the remainder of our analysis under this event.

Defining $\kappa = (1 - \frac{2\lambda-2\gamma}{\lambda+\mu})$, it suffices to show that

$$\|\theta^{s+1} - \theta^*\|_2 \leq \kappa\|\theta^s - \theta^*\|_2 + \alpha\varepsilon_Q^{\text{unif}}(n,\delta), \quad \text{for each iteration } s \in \{0,1,2,\ldots\}. \tag{6.3}$$

Indeed, when this bound holds, we may iterate it to show that

$$\begin{aligned}
\|\theta^t - \theta^*\|_2 &\leq \kappa\|\theta^{t-1} - \theta^*\|_2 + \alpha\varepsilon_Q^{\text{unif}}(n,\delta) \\
&\leq \kappa\{\kappa\|\theta^{t-2} - \theta^*\|_2 + \alpha\varepsilon_Q^{\text{unif}}(n,\delta)\} + \alpha\varepsilon_Q^{\text{unif}}(n,\delta) \\
&\leq \kappa^t\|\theta^0 - \theta^*\|_2 + \left\{\sum_{s=0}^{t-1}\kappa^s\right\}\alpha\varepsilon_Q^{\text{unif}}(n,\delta) \\
&\leq \kappa^t\|\theta^0 - \theta^*\|_2 + \frac{\alpha}{1-\kappa}\varepsilon_Q^{\text{unif}}(n,\delta),
\end{aligned}$$

where the final step follows by summing the geometric series.

It remains to prove the claim (6.3), and we do so via induction on the iteration number. Beginning with $s = 0$, we have

$$\begin{aligned}\|\theta^1 - \theta^*\|_2 &= \|\theta^0 + \alpha \nabla Q_n(\theta|\theta^0)|_{\theta=\theta^0} - \theta^*\|_2 \\ &\overset{\text{(i)}}{\leq} \|\theta^0 + \alpha \nabla Q(\theta|\theta^0)|_{\theta=\theta^0} - \theta^*\|_2 \\ &\quad + \alpha \|\nabla Q(\theta|\theta^0)|_{\theta=\theta^0} - \nabla Q_n(\theta|\theta^0)|_{\theta=\theta^0}\| \\ &\overset{\text{(ii)}}{\leq} \kappa \|\theta^0 - \theta^*\|_2 + \alpha \varepsilon_Q^{\text{unif}}(n, \delta),\end{aligned}$$

where step (i) follows by triangle inequality, whereas step (ii) follows from the bound (6.2), and the contractivity of the population operator applied to $\theta^0 \in \mathbb{B}_2(r; \theta^*)$, that is, Theorem 1. By our initialization condition and the assumed bound (4.3), note that we are guaranteed that $\|\theta^1 - \theta^*\|_2 \leq r$.

In the induction from $s \mapsto s + 1$, suppose that $\|\theta^s - \theta^*\|_2 \leq r$, and the bound (6.3) holds at iteration $s$. The same argument then implies that the bound (6.3) also holds for iteration $s + 1$, and that $\|\theta^{s+1} - \theta^*\|_2 \leq r$, thus completing the proof.

6.3. *Proofs of population-based corollaries for first-order EM.* In this section, we prove Corollaries 1–3 on the behavior of first-order EM at the population level for concrete models.

6.3.1. *Proof of Corollary 1.* We note at this point, and for subsequent examples that scaling the family of $Q$ functions by a fixed constant does not affect any of our conditions and their consequences. Particularly, in various examples, we will re-scale $Q$ functions by constants such as $\sigma^2$. In order to apply Theorem 1, we need to verify the $\lambda$-concavity (3.2) and $\mu$-smoothness (3.3) conditions, and the GS($\gamma$) condition (3.1) over the ball $\mathbb{B}_2(r; \theta^*)$. The first-order EM update is given in Appendix A. In this example, the $q$-function takes the form

$$q(\theta) = Q(\theta|\theta^*) = -\tfrac{1}{2}\mathbb{E}\big[w_{\theta^*}(Y)\|Y - \theta\|_2^2 + \big(1 - w_{\theta^*}(Y)\big)\|Y + \theta\|_2^2\big],$$

where the weighting function is given by

$$w_\theta(y) := \frac{\exp(-\|\theta - y\|_2^2/(2\sigma^2))}{\exp(-\|\theta - y\|_2^2/(2\sigma^2)) + \exp(-\|\theta + y\|_2^2/(2\sigma^2))}.$$

The $q$-function is smooth and strongly-concave with parameters 1.

It remains to verify the GS($\gamma$) condition (3.1). The main technical effort, deferred to the appendices, is in showing the following central lemma.

LEMMA 2. *Under the conditions of Corollary 1, there is a constant $\gamma \in (0, 1)$ with $\gamma \leq \exp(-c_2\eta^2)$ such that*

$$\|\mathbb{E}\big[2\Delta_w(Y)Y\big]\|_2 \leq \gamma \|\theta - \theta^*\|_2, \tag{6.4}$$

*where $\Delta_w(y) := w_\theta(y) - w_{\theta^*}(y)$.*

The proof of this result crucially exploits the generative model, as well as the smoothness of the weighting function, in order to establish that the GS condition holds over a relatively large region around the population global optima ($\theta^*$ and $-\theta^*$). Intuitively, the generative model allows us to argue that with large probability the weighting function $w_\theta(y)$ and the weighting function $w_{\theta^*}(y)$ are quite close, even when $\theta$ and $\theta^*$ are relatively far, so that in expectation the GS condition is satisfied.

Taking this result as given for the moment, let us now verify the GS condition (3.1). An inspection of the updates in equation (A.3), along with the claimed smoothness and strong-concavity parameters lead to the conclusion that it suffices to show that

$$\|\mathbb{E}[2\Delta_w(Y)Y]\|_2 < \|\theta - \theta^*\|_2.$$

This follows immediately from Lemma 2. Thus, the GS condition holds when $\gamma < 1$. The bound on the contraction parameter follows from the fact that $\gamma \leq \exp(-c_2\eta^2)$ and applying Theorem 1 yields Corollary 1.

6.3.2. *Proof of Corollary 2.* Once again we need to verify the $\lambda$-strong concavity (3.2) and $\mu$-smoothness (3.3) conditions, and the GS($\gamma$) condition (3.1) over the ball $\mathbb{B}_2(r;\theta^*)$. In this example, the $q$-function takes the form:

$$\begin{aligned} q(\theta) &= Q(\theta|\theta^*) \\ &:= -\tfrac{1}{2}\mathbb{E}\big[w_{\theta^*}(X,Y)\big(Y - \langle X,\theta\rangle\big)^2 + \big(1 - w_{\theta^*}(X,Y)\big)\big(Y + \langle X,\theta\rangle\big)^2\big], \end{aligned}$$

where $w_\theta(x,y) := \frac{\exp(-(y-\langle x,\theta\rangle)^2/(2\sigma^2))}{\exp(-(y-\langle x,\theta\rangle)^2/(2\sigma^2))+\exp(-(y+\langle x,\theta\rangle)^2/(2\sigma^2))}$. Observe that function $Q(\cdot|\theta^*)$ is $\lambda$-strongly concave and $\mu$-smooth with $\lambda$ and $\mu$ equal to the smallest and largest (resp.) eigenvalue of the matrix $\mathbb{E}[XX^T]$. Since $\mathbb{E}[XX^T] = I$ by assumption, we see that strong concavity and smoothness hold with $\lambda = \mu = 1$.

It remains to verify condition GS. Define the difference function $\Delta_w(X,Y) := w_\theta(X,Y) - w_{\theta^*}(X,Y)$, and the difference vector $\Delta = \theta - \theta^*$. Using the updates given in Appendix A in equation (A.6a), we need to show that

$$\|2\mathbb{E}[\Delta_w(X,Y)YX]\|_2 < \|\Delta\|_2.$$

Fix any $\widetilde{\Delta} \in \mathbb{R}^d \setminus \{0\}$. It suffices for us to show that

$$\langle 2\mathbb{E}[\Delta_w(X,Y)YX], \widetilde{\Delta}\rangle < \|\Delta\|_2\|\widetilde{\Delta}\|_2.$$

Note that we can write $Y \stackrel{d}{=} (2Z-1)\langle X,\theta^*\rangle + v$, where $Z \sim \mathrm{Ber}(1/2)$ is a Bernoulli variable, and $v \sim \mathcal{N}(0,1)$. Using this notation, it is sufficient to show

$$\begin{aligned} &\mathbb{E}[\Delta_w(X,Y)(2Z-1)\langle X,\theta^*\rangle\langle X,\widetilde{\Delta}\rangle] + \mathbb{E}[\Delta_w(X,Y)v\langle X,\widetilde{\Delta}\rangle] \\ &\qquad \leq \gamma\|\Delta\|_2\|\widetilde{\Delta}\|_2 \end{aligned} \tag{6.5}$$

for $\gamma \in [0, 1/2)$ in order to establish contractivity. In order to prove the theorem with the desired upper bound on the coefficient of contraction we need to show (6.5) with $\gamma \in [0, 1/4)$. Once again, the main technical effort is in establishing the following lemma which provides control on the two terms.

LEMMA 3. *Under the conditions of Corollary 2, there is a constant $\gamma < 1/4$ such that for any fixed vector $\widetilde{\Delta}$ we have*

$$\big|\mathbb{E}\big[\Delta_w(X, Y)(2Z - 1)\langle X, \theta^*\rangle\langle X, \widetilde{\Delta}\rangle\big]\big| \leq \frac{\gamma}{2}\|\Delta\|_2\|\widetilde{\Delta}\|_2 \quad \textit{and} \tag{6.6a}$$

$$\big|\mathbb{E}\big[\Delta_w(X, Y)v\langle X, \widetilde{\Delta}\rangle\big]\big| \leq \frac{\gamma}{2}\|\Delta\|_2\|\widetilde{\Delta}\|_2. \tag{6.6b}$$

In conjunction, these bounds imply that $\langle \mathbb{E}[\Delta_w(X, Y)YX], \widetilde{\Delta}\rangle \leq \gamma\|\Delta\|_2\|\widetilde{\Delta}\|_2$ with $\gamma \in [0, 1/4)$, as claimed.

6.3.3. *Proof of Corollary 3.* We need to verify the conditions of Theorem 1, namely that the function $q$ is $\mu$-smooth, $\lambda$-strongly concave and that the GS condition is satisfied. In this case, $q$ is a quadratic of the form

$$q(\theta) = \tfrac{1}{2}\langle\theta, \mathbb{E}\big[\Sigma_{\theta^*}(X_{\mathrm{obs}}, Y)\big]\theta\rangle - \big\langle\mathbb{E}\big[Y\mu_{\theta^*}(X_{\mathrm{obs}}, Y)\big], \theta\big\rangle,$$

where the vector $\mu_{\theta^*} \in \mathbb{R}^d$ and matrix $\Sigma_{\theta^*}$ are defined formally in the Appendix [see equations (A.7a) and (A.7c), resp.]. Here, the expectation is over both the patterns of missingness and the random $(X_{\mathrm{obs}}, Y)$.

*Smoothness and strong concavity.* Note that $q$ is a quadratic function with Hessian $\nabla^2 q(\theta) = \mathbb{E}[\Sigma_{\theta^*}(X_{\mathrm{obs}}, Y)]$. Let us fix a pattern of missingness, and then average over $(X_{\mathrm{obs}}, Y)$. Recalling the matrix $U_{\theta^*}$ from equation (A.7b), we find that a simple calculation yields

$$\mathbb{E}\big[\Sigma_{\theta^*}(X_{\mathrm{obs}}, Y)\big] = \begin{bmatrix} I & U_{\theta^*}\begin{bmatrix} I \\ \theta^{*T}_{\mathrm{obs}} \end{bmatrix} \\ \begin{bmatrix} I & \theta^*_{\mathrm{obs}} \end{bmatrix} U^T_{\theta^*} & I \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix},$$

showing that the expectation does not depend on the pattern of missingness. Consequently, the quadratic function $q$ has an identity Hessian, showing that smoothness and strong concavity hold with $\mu = \lambda = 1$.

*Condition GS.* We need to prove the existence of a scalar $\gamma \in [0, 1)$ such that $\|\mathbb{E}[V]\|_2 \leq \gamma\|\theta - \theta^*\|_2$, where the vector $V = V(\theta, \theta^*)$ is given by

$$\begin{aligned} V := {} & \Sigma_{\theta^*}(X_{\mathrm{obs}}, Y)\theta - Y\mu_{\theta^*}(X_{\mathrm{obs}}, Y) - \Sigma_{\theta}(X_{\mathrm{obs}}, Y)\theta \\ & + Y\mu_{\theta}(X_{\mathrm{obs}}, Y). \end{aligned} \tag{6.7}$$

For a fixed pattern of missingness, we can compute the expectation over $(X_{\text{obs}}, Y)$ in closed form. Supposing that the first block is missing, we have

$$\mathbb{E}_{X_{\text{obs}},Y}[V] = \begin{bmatrix} (\theta_{\text{mis}} - \theta^*_{\text{mis}}) + \pi_1 \theta_{\text{mis}} \\ \pi_2(\theta_{\text{obs}} - \theta^*_{\text{obs}}) \end{bmatrix}, \tag{6.8}$$

where $\pi_1 := \frac{\|\theta^*_{\text{mis}}\|_2^2 - \|\theta_{\text{mis}}\|_2^2 + \|\theta_{\text{obs}} - \theta^*_{\text{obs}}\|_2^2}{\|\theta_{\text{mis}}\|_2^2 + \sigma^2}$ and $\pi_2 := \frac{\|\theta_{\text{mis}}\|_2^2}{\|\theta_{\text{mis}}\|_2^2 + \sigma^2}$. We claim that these scalars can be bounded, independently of the missingness pattern, as

$$\pi_1 \leq 2(\xi_1 + \xi_2)\frac{\|\theta - \theta^*\|_2}{\sigma} \quad \text{and} \quad \pi_2 \leq \delta := \frac{1}{1 + (1/(\xi_1 + \xi_2))^2} < 1. \tag{6.9}$$

Taking these bounds (6.9) as given for the moment, we can then average over the missing pattern. Since each coordinate is missing independently with probability $\rho$, the expectation of the $i$th coordinate is at most $|\mathbb{E}[V]|_i \leq |\rho|\theta_i - \theta^*_i| + \rho\pi_1|\theta_i| + (1-\rho)\pi_2|\theta_i - \theta^*_i||$. Thus, defining $\eta := (1-\rho)\delta + \rho < 1$, we have

$$\begin{aligned} \|\mathbb{E}[V]\|_2^2 &\leq \eta^2\|\theta - \theta^*\|_2^2 + \rho^2\pi_1^2\|\theta\|_2^2 + 2\pi_1\eta\rho|\langle\theta, \theta - \theta^*\rangle| \\ &\leq \underbrace{\left\{\eta^2 + \rho^2\|\theta\|_2^2\frac{4(\xi_1+\xi_2)^2}{\sigma^2} + \frac{4\eta\rho\|\theta\|_2(\xi_1+\xi_2)}{\sigma}\right\}}_{\gamma^2}\|\theta - \theta^*\|_2^2, \end{aligned}$$

where we have used our upper bound (6.9) on $\pi_1$. We need to ensure that $\gamma < 1$. By assumption, we have $\|\theta^*\|_2 \leq \xi_1\sigma$ and $\|\theta - \theta^*\|_2 \leq \xi_2\sigma$, and hence $\|\theta\|_2 \leq (\xi_1 + \xi_2)\sigma$. Thus, the coefficient $\gamma^2$ is upper bounded as

$$\gamma^2 \leq \eta^2 + 4\rho^2(\xi_1+\xi_2)^4 + 4\eta\rho(\xi_1+\xi_2)^2.$$

Under the stated conditions of the corollary, we have $\gamma < 1$, thereby completing the proof.

It remains to prove the bounds (6.9). By our assumptions, we have $\|\theta_{\text{mis}}\|_2 - \|\theta^*_{\text{mis}}\|_2 \leq \|\theta_{\text{mis}} - \theta^*_{\text{mis}}\|_2$, and moreover

$$\|\theta_{\text{mis}}\|_2 \leq \|\theta^*_{\text{mis}}\|_2 + \xi_2\sigma \leq (\xi_1 + \xi_2)\sigma. \tag{6.10}$$

As consequence, we have

$$\begin{aligned} \|\theta^*_{\text{mis}}\|_2^2 - \|\theta_{\text{mis}}\|_2^2 &= (\|\theta_{\text{mis}}\|_2 - \|\theta^*_{\text{mis}}\|_2)(\|\theta_{\text{mis}}\|_2 + \|\theta^*_{\text{mis}}\|_2) \\ &\leq (2\xi_1 + \xi_2)\sigma\|\theta_{\text{mis}} - \theta^*_{\text{mis}}\|_2. \end{aligned}$$

Since $\|\theta_{\text{obs}} - \theta^*_{\text{obs}}\|_2^2 \leq \xi_2\sigma\|\theta_{\text{obs}} - \theta^*_{\text{obs}}\|_2$, the stated bound on $\pi_1$ follows.

On the other hand, we have

$$\pi_2 = \frac{\|\theta_{\text{mis}}\|_2^2}{\|\theta_{\text{mis}}\|_2^2 + \sigma^2} = \frac{1}{1 + \sigma^2/\|\theta_{\text{mis}}\|_2^2} \overset{\text{(i)}}{\leq} \underbrace{\frac{1}{1 + (1/(\xi_1+\xi_2))^2}}_{\delta} < 1,$$

where step (i) follows from (6.10).

6.4. *Proofs of sample-based corollaries for first-order EM.* This section is devoted to proofs of Corollaries 4 through 6 on the behavior of the first-order EM algorithm in the finite sample setting.

6.4.1. *Proof of Corollary 4.* In order to prove this result, it suffices to bound the quantity $\varepsilon_Q^{\mathrm{unif}}(n,\delta)$ defined in equation (4.2). Utilizing the updates defined in equation (A.3), and defining the set $\mathbb{A} := \{\theta \in \mathbb{R}^d \mid \|\theta - \theta^*\|_2 \le \|\theta^*\|_2/4\}$, we need to control the random variable

$$Z := \sup_{\theta\in\mathbb{A}} \left\| \alpha\left\{ \frac{1}{n}\sum_{i=1}^n \bigl(2w_\theta(y_i)-1\bigr)y_i - \theta \right\} - \alpha\bigl[2\mathbb{E}\bigl[w_\theta(Y)Y\bigr]-\theta\bigr] \right\|_2.$$

In order to establish the Corollary it suffices to show that for sufficiently large universal constants $c_1, c_2$ we have that, for $n \ge c_1 d\log(1/\delta)$

$$Z \le \frac{c_2\|\theta^*\|_2(\|\theta^*\|_2^2+\sigma^2)}{\sigma^2}\sqrt{\frac{d\log(1/\delta)}{n}}$$

with probability at least $1-\delta$.

For each unit-norm vector $u \in \mathbb{R}^d$, define the random variable

$$Z_u := \sup_{\theta\in\mathbb{A}} \left\{ \frac{1}{n}\sum_{i=1}^n \bigl(2w_\theta(y_i)-1\bigr)\langle y_i, u\rangle - \mathbb{E}\bigl(2w_\theta(Y)-1\bigr)\langle Y, u\rangle \right\}.$$

Recalling that we choose $\alpha = 1$, we note that $Z = \sup_{u\in\mathbb{S}^d} Z_u$. We begin by reducing our problem to a finite maximum over the sphere $\mathbb{S}^d$. Let $\{u^1,\dots,u^M\}$ denote a $1/2$-covering of the sphere $\mathbb{S}^d = \{v \in \mathbb{R}^d \mid \|v\|_2 = 1\}$. For any $v \in \mathbb{S}^d$, there is some index $j \in [M]$ such that $\|v - u^j\|_2 \le 1/2$, and hence we can write

$$Z_v \le Z_{u^j} + |Z_v - Z_{u^j}| \le \max_{j\in[M]} Z_{u^j} + Z\|v-u^j\|_2,$$

where the final step uses the fact that $|Z_u - Z_v| \le Z\|u-v\|_2$ for any pair $(u,v)$. Putting together the pieces, we conclude that

$$Z = \sup_{v\in\mathbb{S}^d} Z_v \le 2\max_{j\in[M]} Z_{u^j}. \tag{6.11}$$

Consequently, it suffices to bound the random variable $Z_u$ for a fixed $u \in \mathbb{S}^d$. Letting $\{\varepsilon_i\}_{i=1}^n$ denote an i.i.d. sequence of Rademacher variables, for any $\lambda > 0$, we have

$$\mathbb{E}\bigl[e^{\lambda Z_u}\bigr] \le \mathbb{E}\left[\exp\left(\frac{2}{n}\sup_{\theta\in\mathbb{A}}\sum_{i=1}^n \varepsilon_i\bigl(2w_\theta(y_i)-1\bigr)\langle y_i,u\rangle\right)\right],$$

using a standard symmetrization result for empirical processes (e.g., [23, 24]). Now observe that for any triplet of $d$-vectors $y$, $\theta$ and $\theta'$, we have the Lipschitz property

$$\bigl|2w_\theta(y) - 2w_{\theta'}(y)\bigr| \le \frac{1}{\sigma^2}\bigl|\langle\theta, y\rangle - \langle\theta', y\rangle\bigr|.$$

Consequently, by the Ledoux–Talagrand contraction for Rademacher processes [23, 24], we have

$$\mathbb{E}\left[\exp\left(\frac{2}{n}\sup_{\theta\in\mathbb{A}}\sum_{i=1}^{n}\varepsilon_i\big(2w_\theta(y_i)-1\big)\langle y_i,u\rangle\right)\right] \le \mathbb{E}\left[\exp\left(\frac{4}{n\sigma^2}\sup_{\theta\in\mathbb{A}}\sum_{i=1}^{n}\varepsilon_i\langle\theta,y_i\rangle\langle y_i,u\rangle\right)\right].$$

Since any $\theta\in\mathbb{A}$ satisfies $\|\theta\|_2\le\frac{5}{4}\|\theta^*\|_2$, we have

$$\sup_{\theta\in\mathbb{A}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\langle\theta,y_i\rangle\langle y_i,u\rangle \le \frac{5}{4}\|\theta^*\|_2\left|\!\left|\!\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i y_i y_i^T\right|\!\right|\!\right|_{\mathrm{op}},$$

where $|||\cdot|||_{\mathrm{op}}$ denotes the $\ell_2$-operator norm of a matrix (maximum singular value). Repeating the same discretization argument over $\{u^1,\dots,u^M\}$, we find that

$$\left|\!\left|\!\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i y_i y_i^T\right|\!\right|\!\right|_{\mathrm{op}} \le 2\max_{j\in[M]}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\big\langle y_i,u^j\big\rangle^2.$$

Putting together the pieces, we conclude that

$$\begin{aligned}\mathbb{E}\big[e^{\lambda Z_u}\big] &\le \mathbb{E}\left[\exp\left(\frac{10\lambda\|\theta^*\|_2}{\sigma^2}\max_{j\in[M]}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\big\langle y_i,u^j\big\rangle^2\right)\right]\\ &\le \sum_{j=1}^{M}\mathbb{E}\left[\exp\left(\frac{10\lambda\|\theta^*\|_2}{\sigma^2}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\big\langle y_i,u^j\big\rangle^2\right)\right].\end{aligned} \tag{6.12}$$

Now by assumption, the random vectors $\{y_i\}_{i=1}^n$ are generated i.i.d. according to the model $y=\eta\theta^*+w$, where $\eta$ is a Rademacher sign variable, and $w\sim\mathcal{N}(0,\sigma^2 I)$. Consequently, for any $u\in\mathbb{R}^d$, we have

$$\mathbb{E}\big[e^{\langle u,y\rangle}\big]=\mathbb{E}\big[e^{\eta\langle u,\theta^*\rangle}\big]\mathbb{E}\big[e^{\langle u,w\rangle}\big]\le e^{(\|\theta^*\|_2^2+\sigma^2)/2},$$

showing that the vectors $y_i$ are sub-Gaussian with parameter at most $\gamma=\sqrt{\|\theta^*\|_2^2+\sigma^2}$. Therefore, $\varepsilon_i\langle y_i,u\rangle^2$ is zero mean sub-exponential, and has moment generating function bounded as $\mathbb{E}[e^{t\varepsilon_i\langle y_i,u\rangle^2}]\le e^{\gamma^4t^2/2}$ for all $t>0$ sufficiently small. Combined with our earlier inequality (6.12), we conclude that

$$\mathbb{E}\big[e^{\lambda Z_u}\big]\le Me^{c\frac{\lambda^2\|\theta^*\|_2^2\gamma^4}{n\sigma^4}}\le e^{c\frac{\lambda^2\|\theta^*\|_2^2\gamma^4}{n\sigma^4}+2d}$$

for all $\lambda$ sufficiently small. Combined with our first discretization (6.11), we have thus shown that

$$\mathbb{E}\big[e^{\frac{\lambda}{2}Z}\big]\le Me^{c\frac{\lambda^2\|\theta^*\|_2^2\gamma^4}{n\sigma^4}+2d}\le e^{c\frac{\lambda^2\|\theta^*\|_2^2\gamma^4}{n\sigma^4}+4d}.$$

Combined with the Chernoff approach, this bound on the MGF implies that, as long as $n \geq c_1 d \log(1/\delta)$ for a sufficiently large constant $c_1$, we have

$$Z \leq \frac{c_2 \|\theta^*\|_2 \gamma^2}{\sigma^2} \sqrt{\frac{d \log(1/\delta)}{n}}$$

with probability at least $1 - \delta$ as desired.

6.4.2. *Proof of Corollary 5.* As before, it suffices to find a suitable upper bound on the $\varepsilon_Q(n, \delta)$ from equation (4.8). Based on the specific form of the first-order EM updates for this model [see equation (A.6a) in Appendix A], we need to control the random variable

$$Z := \left\| \alpha \left\{ \frac{1}{n} \sum_{i=1}^{n} \big(2 w_\theta(y_i) - 1\big) y_i - \theta \right\} - \alpha \big[ 2 \mathbb{E}\big[ w_\theta(Y) Y \big] - \theta \big] \right\|_2.$$

We claim that there are universal constants $(c_1, c_2)$ such that given a sample size $n \geq c_1 d \log(1/\delta)$, we have

$$\mathbb{P}\left[ Z > \frac{c_2 \|\theta^*\|_2 (\|\theta^*\|_2^2 + \sigma^2)}{\sigma^2} \sqrt{\frac{d \log(1/\delta)}{n}} \right] \leq \delta.$$

Given our choice of stepsize $\alpha = 1$, we have

$$\begin{aligned} Z \leq & \left\| \frac{1}{n} \sum_{i=1}^{n} \big(2 w_\theta(x_i, y_i) - 1\big) y_i x_i - \mathbb{E}\big(2 w_\theta(X, Y) - 1\big) Y X \right\|_2 \\ & + \left\| I - \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \right\|_{\mathrm{op}} \|\theta\|_2. \end{aligned}$$

Now define the matrices $\widehat{\Sigma} := \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T$ and $\Sigma = \mathbb{E}[X X^T] = I$, as well as the vector

$$\widehat{v} := \frac{1}{n} \sum_{i=1}^{n} \big[ \mu_\theta(x_i, y_i) y_i x_i \big] \quad \text{and} \quad v := \mathbb{E}\big[ \mu_\theta(X, Y) Y X \big],$$

where $\mu_\theta(x, y) := 2 w_\theta(x, y) - 1$. Noting that $\mathbb{E}[Y X] = 0$, we have the bound

$$Z \leq \underbrace{\|\widehat{v} - v\|_2}_{T_1} + \underbrace{\big\|\widehat{\Sigma} - \Sigma\big\|_{\mathrm{op}} \|\theta\|_2}_{T_2}. \tag{6.13}$$

We bound each of the terms $T_1$ and $T_2$ in turn.

*Bounding $T_1$.* Let us write $\|\widehat{v} - v\|_2 = \sup_{u \in \mathbb{S}^d} Z(u)$, where

$$Z(u) := \frac{1}{n} \sum_{i=1}^{n} \mu_\theta(x_i, y_i) y_i \langle x, u \rangle - \mathbb{E}\big[\mu_\theta(X, Y) Y \langle X, u \rangle\big].$$

By a discretization argument over a $1/2$-cover of the sphere $\mathbb{S}^d$—say $\{u^1, \ldots, u^M\}$, we have the upper bound $\|\widehat{v} - v\|_2 \leq 2 \max_{j \in [M]} Z(u^j)$. Thus, it suffices to control the random variable $Z(u)$ for a fixed $u \in \mathbb{S}^d$. By a standard symmetrization argument [46], we have

$$\mathbb{P}\big[Z(u) \geq t\big] \leq 2\mathbb{P}\left[\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \mu_\theta(x_i, y_i) y_i \langle x_i, u \rangle \geq t/2\right],$$

where $\{\varepsilon_i\}_{i=1}^n$ are an i.i.d. sequence of Rademacher variables. Let us now define the event $\mathcal{E}\{\frac{1}{n}\sum_{i=1}^n \langle x_i, u\rangle^2 \leq 2\}$. Since each variable $\langle x_i, u\rangle$ is sub-Gaussian with parameter one, standard tail bounds imply that $\mathbb{P}[\mathcal{E}^c] \leq e^{-n/32}$. Therefore, we can write

$$\mathbb{P}\big[Z(u) \geq t\big] \leq 2\mathbb{P}\left[\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \mu_\theta(x_i, y_i) y_i \langle x_i, u \rangle \geq t/2 \,\middle|\, \mathcal{E}\right] + 2e^{-n/32}.$$

As for the remaining term, we have

$$\mathbb{E}\left[\exp\left(\frac{\lambda}{n} \sum_{i=1}^{n} \varepsilon_i \mu_\theta(x_i, y_i) y_i \langle x_i, u \rangle\right) \,\middle|\, \mathcal{E}\right] \leq \mathbb{E}\left[\exp\left(\frac{2\lambda}{n} \sum_{i=1}^{n} \varepsilon_i y_i \langle x_i, u \rangle\right) \,\middle|\, \mathcal{E}\right],$$

where we have applied the Ledoux–Talagrand contraction for Rademacher processes [23, 24], using the fact that $|\mu_\theta(x, y)| \leq 1$ for all pairs $(x, y)$. Now conditioned on $x_i$, the random variable $y_i$ is zero-mean and sub-Gaussian with parameter at most $\sqrt{\|\theta^*\|_2^2 + \sigma^2}$. Consequently, taking expectations over the distribution $(y_i | x_i)$ for each index $i$, we find that

$$\begin{aligned}\mathbb{E}\left[\exp\left(\frac{2\lambda}{n} \sum_{i=1}^{n} \varepsilon_i y_i \langle x_i, u \rangle\right) \,\middle|\, \mathcal{E}\right] &\leq \left[\exp\left(\frac{4\lambda^2}{n^2}(\|\theta^*\|_2^2 + \sigma^2) \sum_{i=1}^{n} \langle x_i, u\rangle^2\right) \,\middle|\, \mathcal{E}\right] \\ &\leq \exp\left(\frac{8\lambda^2}{n}(\|\theta^*\|_2^2 + \sigma^2)\right),\end{aligned}$$

where the final inequality uses the definition of $\mathcal{E}$. Using this bound on the moment-generating function, we find that

$$\mathbb{P}\left[\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \mu_\theta(x_i, y_i) y_i \langle x_i, u \rangle \geq t/2 \,\middle|\, \mathcal{E}\right] \leq \exp\left(-\frac{nt^2}{256(\|\theta^*\|_2^2 + \sigma^2)}\right).$$

Since the $1/2$-cover of the unit sphere $\mathbb{S}^d$ has at most $2^d$ elements, we conclude that there is a universal constant $c$ such that $T_1 \leq c\sqrt{\|\theta^*\|_2^2 + \sigma^2}\sqrt{\frac{d}{n}\log(1/\delta)}$ with probability at least $1 - \delta$.

*Bounding* $T_2$. Since $n > d$ by assumption, standard results in random matrix theory [47] imply that $|||\widehat{\Sigma} - \Sigma|||_{\mathrm{op}} \leq c\sqrt{\frac{d}{n}\log(1/\delta)}$ with probability at least $1 - \delta$. On the other hand, observe that $\|\theta\|_2 \leq 2\|\theta^*\|_2$, since with the chosen stepsize, each iteration decreases the distance to $\theta^*$ and our initial iterate satisfies $\|\theta\|_2 \leq 2\|\theta^*\|_2$. Combining the pieces, we see that $T_2 \leq c\|\theta^*\|_2\sqrt{\frac{d}{n}\log(1/\delta)}$ with probability at least $1 - \delta$.

Finally, substituting our bounds on $T_1$ and $T_2$ into the decomposition (6.13) yields the claim.

6.4.3. *Proof of Corollary* 6. We need to upper bound the deviation function $\varepsilon_Q(n, \delta)$ previously defined (4.8). For any fixed $\theta \in \mathbb{B}_2(r; \theta^*) = \{\theta \in \mathbb{R}^d \mid \|\theta - \theta^*\|_2 \leq \xi_2\sigma\}$, we need to upper bound the random variable,

$$Z = \left\| \frac{1}{n}\sum_{i=1}^{n}\big[y_i\mu_\theta(x_{\mathrm{obs},i}, y_i) - \Sigma_\theta(x_{\mathrm{obs},i}, y_i)\theta\big] - \mathbb{E}\big[Y\mu_\theta(X_{\mathrm{obs}}, Y) - \Sigma_\theta(X_{\mathrm{obs}}, Y)\theta\big] \right\|_2,$$

with high probability. We define: $T_1 := \|[\mathbb{E}\Sigma_\theta(x_{\mathrm{obs}}, y)\theta - \frac{1}{n}\sum_{i=1}^{n}\Sigma_\theta(x_{\mathrm{obs},i}, y_i)\theta]\|_2$, and

$$T_2 := \left\| \left[ \mathbb{E}\big(y\mu_\theta(x_{\mathrm{obs}}, y)\big) - \frac{1}{n}\sum_{i=1}^{n} y_i\mu_\theta(x_{\mathrm{obs},i}, y_i) \right] \right\|_2.$$

For convenience, we let $z_i \in \mathbb{R}^d$ be a $\{0, 1\}$-valued indicator vector, with ones in the positions of observed covariates. For ease of notation, we frequently use the abbreviations $\Sigma_\theta$ and $\mu_\theta$ when the arguments are understood. We use the notation $\odot$ to denote the element-wise product.

*Controlling* $T_1$. Define the matrices

$$\bar{\Sigma} = \mathbb{E}\big[\Sigma_\theta(x_{\mathrm{obs}}, y)\big] \quad \text{and} \quad \widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\Sigma_\theta(x_{\mathrm{obs},i}, y_i).$$

With this notation, we have $T_1 \leq |||\bar{\Sigma} - \widehat{\Sigma}|||_{\mathrm{op}}\|\theta\|_2 \leq |||\bar{\Sigma} - \widehat{\Sigma}|||_{\mathrm{op}}(\xi_1 + \xi_2)\sigma$, where the second step follows since any vector $\theta \in \mathbb{B}_2(r; \theta^*)$ has $\ell_2$-norm bounded as $\|\theta\|_2 \leq (\xi_1 + \xi_2)\sigma$. We claim that for any fixed vector $u \in \mathbb{S}^d$, the random variable $\langle u, (\bar{\Sigma} - \widehat{\Sigma})u\rangle$ is zero-mean and sub-exponential. When this tail condition holds and $n > d$, standard arguments in random matrix theory [47] ensure that $|||\bar{\Sigma} - \widehat{\Sigma}|||_{\mathrm{op}} \leq c\sqrt{\frac{d}{n}\log(1/\delta)}$ with probability at least $1 - \delta$.

It is clear that $\langle u, (\bar{\Sigma} - \widehat{\Sigma})u\rangle$ has zero mean. It remains to prove that $\langle u, (\bar{\Sigma} - \widehat{\Sigma})u\rangle$ is sub-exponential. Note that $\widehat{\Sigma}$ is a rescaled sum of rank one matrices, each of the form

$$\Sigma_\theta(x_{\mathrm{obs}}, y) = I_{\mathrm{mis}} + \mu_\theta \mu_\theta^T - \big((1-z)\odot\mu_\theta\big)\big((1-z)\odot\mu_\theta\big)^T,$$

where $I_{\mathrm{mis}}$ denotes the identity matrix on the diagonal sub-block corresponding to the missing entries. The square of any sub-Gaussian random variable has sub-exponential tails. Thus, it suffices to show that each of the random variables $\langle \mu_\theta, u\rangle$, and $\langle (1-z)\odot\mu_\theta, u\rangle$ are sub-Gaussian. The random vector $z \odot x$ has i.i.d. sub-Gaussian components with parameter at most 1 and $\|u\|_2 = 1$, so that $\langle z\odot x, u\rangle$ is sub-Gaussian with parameter at most 1. It remains to verify that $\mu_\theta$ is sub-Gaussian, a fact that we state for future reference as a lemma.

LEMMA 4. *Under the conditions of Corollary 3, the random vector $\mu_\theta(x_{\mathrm{obs}}, y)$ is sub-Gaussian with a constant parameter.*

PROOF. Introducing the shorthand $\omega = (1-z)\odot\theta$, we have

$$\mu_\theta(x_{\mathrm{obs}}, y) = z\odot x + \frac{1}{\sigma^2 + \|\omega\|_2^2}\big[y - \langle z\odot\theta, z\odot x\rangle\big]\omega.$$

Moreover, since $y = \langle x, \theta^*\rangle + v$, we have

$$\big\langle \mu_\theta(x_{\mathrm{obs}}, y), u\big\rangle = \underbrace{\langle z\odot x, u\rangle}_{B_1} + \underbrace{\frac{\langle x,\omega\rangle\langle\omega,u\rangle}{\sigma^2+\|\omega\|_2^2}}_{B_2} + \underbrace{\frac{\langle x,\theta^*-\theta\rangle\langle\omega,u\rangle}{\sigma^2+\|\omega\|_2^2}}_{B_3} + \underbrace{\frac{v\langle\omega,u\rangle}{\sigma^2+\|\omega\|_2^2}}_{B_4}.$$

It suffices to show that each of the variables $\{B_j\}_{j=1}^4$ is sub-Gaussian with a constant parameter. As discussed previously, the variable $B_1$ is sub-Gaussian with parameter at most one. On the other hand, note that $x$ and $\omega$ are independent. Moreover, with $\omega$ fixed, the variable $\langle x, \omega\rangle$ is sub-Gaussian with parameter $\|\omega\|_2^2$, whence

$$\mathbb{E}\big[e^{\lambda B_2}\big] \le \exp\left(\lambda^2 \frac{\|\omega\|_2^2\langle\omega,u\rangle^2}{2(\sigma^2+\|\omega\|_2^2)^2}\right) \le e^{\lambda^2/2},$$

where the final inequality uses the fact that $\langle\omega, u\rangle^2 \le \|\omega\|_2^2$. We have thus shown that $B_2$ is sub-Gaussian with parameter one. Since $\|\theta - \theta^*\|_2 \le \xi_2\sigma$, the same argument shows that $B_3$ is sub-Gaussian with parameter at most $\xi_2$. Since $v$ is sub-Gaussian with parameter $\sigma$ and independent of $\omega$, the same argument shows that $B_4$ is sub-Gaussian with parameter at most one, thereby completing the proof of the lemma. □

*Controlling* $T_2$. We now turn to the second term. Note the variational representation

$$T_2 = \sup_{\|u\|_2=1} \left| \mathbb{E}\big[y\langle \mu_\theta(x_{\text{obs}}, y), u\rangle\big] - \frac{1}{n}\sum_{i=1}^{n} y_i \langle \mu_\theta(x_{\text{obs},i}, y_i), u\rangle \right|.$$

By a discretization argument—say with a $1/2$ cover $\{u^1, \ldots, u^M\}$ of the sphere with $M \leq 2^d$ elements, we obtain

$$T_2 \leq 2 \max_{j\in[M]} \left| \mathbb{E}\big[y\langle \mu_\theta(x_{\text{obs}}, y), u^j\rangle\big] - \frac{1}{n}\sum_{i=1}^{n} y_i \langle \mu_\theta(x_{\text{obs},i}, y_i), u^j\rangle \right|.$$

Each term in this maximum is the product of two zero-mean variables, namely $y$ and $\langle \mu_\theta, u\rangle$. On one hand, the variable $y$ is sub-Gaussian with parameter at most $\sqrt{\|\theta^*\|_2^2 + \sigma^2} \leq c\sigma$; on the other hand, Lemma 4 guarantees that $\langle \mu_\theta, u\rangle$ is sub-Gaussian with constant parameter. The product of any two sub-Gaussian variables is sub-exponential, and thus, by standard sub-exponential tail bounds [9], we have $\mathbb{P}[T_2 \geq t] \leq 2M \exp(-c \min\{\frac{nt}{\sqrt{1+\sigma^2}}, \frac{nt^2}{1+\sigma^2}\})$. Since $M \leq 2^d$ and $n > c_1 d$, we conclude that $T_2 \leq c\sqrt{1+\sigma^2}\sqrt{\frac{d}{n}\log(1/\delta)}$ with probability at least $1-\delta$.

Combining our bounds on $T_1$ and $T_2$, we conclude that $\varepsilon_Q(n, \delta) \leq c\sqrt{1+\sigma^2} \times \sqrt{\frac{d}{n}\log(1/\delta)}$ with probability at least $1-\delta$. Thus, we see that Corollary 6 follows from Theorem 2.

**7. Discussion.** In this paper, we have provided some general techniques for studying the EM and first-order EM algorithms, at both the population and finite-sample levels. Although this paper focuses on these specific algorithms, we expect that the techniques could be useful in understanding the convergence behavior of other algorithms for potentially nonconvex problems.

The analysis of this paper can be extended in various directions. For instance, in the three concrete models that we treated, we assumed that the model was correctly specified, and that the samples were drawn in an i.i.d. manner, both conditions that may be violated in statistical practice. Maximum likelihood estimation is known to have various robustness properties under model mis-specification. Developing an understanding of the EM algorithm in this setting is an important open problem.

Finally, we note that in concrete examples our analysis guarantees good behavior of the EM and first-order EM algorithms when they are given suitable initialization. For the three model classes treated in this paper, simple pilot estimators can be used to obtain such initializations—in particular using PCA for Gaussian mixtures and mixtures of regressions (e.g., [53]), and the plug-in principle for regression with missing data (e.g., [22, 52]). These estimators can be seen as par-

ticular instantiations of the method of moments [38]. Although still an active area of research, a line of recent work (e.g., [1, 2, 13, 21]) has demonstrated the utility of moment-based estimators or initializations for other types of latent variable models, and it would be interesting to analyze the behavior of EM for such models.

**Acknowledgments.** The authors would like to thank Andrew Barron, Jason Klusowski and John Duchi for helpful discussions, as well as Amirreza Shaban, Fanny Yang and Xinyang Yi for various corrections to the original manuscript. They are also grateful to the Editor, Associate Editor and anonymous reviewers for their helpful suggestions and improvements to the paper.

## SUPPLEMENTARY MATERIAL

**Supplement to "Statistical guarantees for the EM algorithm: From population to sample-based analysis"** (DOI: 10.1214/16-AOS1435SUPP; .pdf). The supplement [3] contains all remaining technical proofs omitted from the main text due to space constraints.

## REFERENCES

[1] ANANDKUMAR, A., GE, R., HSU, D., KAKADE, S. M. and TELGARSKY, M. (2012). Tensor decompositions for learning latent variable models. Preprint. Available at arXiv:1210.7559.

[2] ANANDKUMAR, A., JAIN, P., NETRAPALLI, P. and TANDON, R. (2013). Learning sparsely used overcomplete dictionaries via alternating minimization. Technical report, Microsoft Research, Redmond, OR.

[3] BALAKRISHNAN, S., WAINWRIGHT, M. J. and YU, B. (2014). Supplement to "Statistical guarantees for the EM algorithm: From population to sample-based analysis." DOI:10.1214/16-AOS1435SUPP.

[4] BALAN, R., CASAZZA, P. and EDIDIN, D. (2006). On signal reconstruction without phase. *Appl. Comput. Harmon. Anal.* **20** 345–356. MR2224902

[5] BAUM, L. E., PETRIE, T., SOULES, G. and WEISS, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.* **41** 164–171. MR0287613

[6] BEALE, E. M. L. and LITTLE, R. J. A. (1975). Missing values in multivariate analysis. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **37** 129–145. MR0373113

[7] BERTSEKAS, D. P. (1995). *Nonlinear Programming*. Athena Scientific, Belmont, CA.

[8] BUBECK, S. (2014). Theory of convex optimization for machine learning. Unpublished manuscript.

[9] BULDYGIN, V. V. and KOZACHENKO, YU. V. (2000). *Metric Characterization of Random Variables and Random Processes*. *Translations of Mathematical Monographs* **188**. Amer. Math. Soc., Providence, RI. Translated from the 1998 Russian original by V. Zaiats. MR1743716

[10] CANDÈS, E. J., STROHMER, T. and VORONINSKI, V. (2013). PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming. *Comm. Pure Appl. Math.* **66** 1241–1274. MR3069958

[11] CELEUX, G., CHAUVEAU, D. and DIEBOLT, J. (1995). On stochastic versions of the EM algorithm. Technical Report, No. 2514, INRIA.

[12] CELEUX, G. and GOVAERT, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Comput. Statist. Data Anal.* **14** 315–332. MR1192205

[13] CHAGANTY, A. T. and LIANG, P. (2013). Spectral experts for estimating mixtures of linear regressions. Unpublished manuscript.

[14] CHEN, Y., YI, X. and CARAMANIS, C. (2013). A convex formulation for mixed regression: Near optimal rates in the face of noise. Unpublished manuscript.

[15] CHRÉTIEN, S. and HERO, A. O. (2008). On EM algorithms and their proximal generalizations. *ESAIM Probab. Stat.* **12** 308–326. MR2404033

[16] DASGUPTA, S. and SCHULMAN, L. (2007). A probabilistic analysis of EM for mixtures of separated, spherical Gaussians. *J. Mach. Learn. Res.* **8** 203–226. MR2320668

[17] DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **39** 1–38. MR0501537

[18] HARTLEY, H. O. (1958). Maximum likelihood estimation from incomplete data. *Biometrics* **14** 174–194.

[19] HEALY, M. and WESTMACOTT, M. (1956). Missing values in experiments analysed on automatic computers. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **5** 203–206.

[20] HERO, A. O. and FESSLER, J. A. (1995). Convergence in norm for alternating expectation-maximization (EM) type algorithms. *Statist. Sinica* **5** 41–54. MR1329288

[21] HSU, D. and KAKADE, S. M. (2012). Learning Gaussian mixture models: Moment methods and spectral decompositions. Preprint. Available at arXiv:1206.5766.

[22] ITURRIA, S. J., CARROLL, R. J. and FIRTH, D. (1999). Polynomial regression and estimating functions in the presence of multiplicative measurement error. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **61** 547–561. MR1707860

[23] KOLTCHINSKII, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems. Lecture Notes in Math.* **2033**. Springer, Heidelberg. MR2829871

[24] LEDOUX, M. and TALAGRAND, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, Berlin. MR1102015

[25] LIU, C. and RUBIN, D. B. (1994). The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. *Biometrika* **81** 633–648. MR1326414

[26] LOH, P. and WAINWRIGHT, M. J. (2012). Corrupted and missing predictors: Minimax bounds for high-dimensional linear regression. In *ISIT* 2601–2605. IEEE, Piscataway Township, NJ.

[27] LOUIS, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **44** 226–233. MR0676213

[28] MA, J. and XU, L. (2005). Asymptotic convergence properties of the EM algorithm with respect to the overlap in the mixture. *Neurocomputing* **68** 105–129.

[29] MCLACHLAN, G. and KRISHNAN, T. (2007). *The EM Algorithm and Extensions*. Wiley, New York.

[30] MEILIJSON, I. (1989). A fast improvement to the EM algorithm on its own terms. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **51** 127–138. MR0984999

[31] MENG, X. (1994). On the rate of convergence of the ECM algorithm. *Ann. Statist.* **22** 326–339. MR1272086

[32] MENG, X. and RUBIN, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80** 267–278. MR1243503

[33] MENG, X. and RUBIN, D. B. (1994). On the global and componentwise rates of convergence of the EM algorithm. *Linear Algebra Appl.* **199** 413–425. MR1274429

[34] NEAL, R. M. and HINTON, G. E. (1999). A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models* (M. I. Jordan, ed.) 355–368. MIT Press, Cambridge, MA.

[35] NESTEROV, Y. (2004). *Introductory Lectures on Convex Optimization*: *A Basic Course*. *Applied Optimization* **87**. Kluwer Academic, Boston, MA. MR2142598

[36] NETRAPALLI, P., JAIN, P. and SANGHAVI, S. (2013). Phase retrieval using alternating minimization. In *Neural Information Processing Systems* Curran Associates, Inc., Red Hook, New York 2796–2804.

[37] ORCHARD, T. and WOODBURY, M. A. (1972). A missing information principle: Theory and applications. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (*Univ. California*, *Berkeley*, *Calif.*, 1970/1971), *Theory of Statistics* **1** 697–715. Univ. California Press, Berkeley, CA. MR0400516

[38] PEARSON, K. (1894). *Contributions to the Mathematical Theory of Evolution*. Harrison and Sons, London.

[39] REDNER, R. A. and WALKER, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **26** 195–239. MR0738930

[40] RUBIN, D. B. (1974). Characterizing the estimation of parameters in incomplete-data problems. *J. Amer. Statist. Assoc.* **69** 467–474.

[41] SUNDBERG, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. *Scand. J. Stat.* **1** 49–58. MR0381110

[42] TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82** 528–550. MR0898357

[43] TSENG, P. (2004). An analysis of the EM algorithm and entropy-like proximal point methods. *Math. Oper. Res.* **29** 27–44. MR2065712

[44] VAN DYK, D. A. and MENG, X. L. (2000). Algorithms based on data augmentation: A graphical representation and comparison. In *Computing Science and Statistics*: *Proceedings of the* 31*st Symposium on the Interface* 230–239. Interface Foundation of North America, Fairfax Station, VA.

[45] VAN DE GEER, S. (2000). *Empirical Processes in M-Estimation*. Cambridge Univ. Press, Cambridge.

[46] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York. MR1385671

[47] VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge Univ. Press, Cambridge. MR2963170

[48] WANG, Z., GU, Q., NING, Y. and LIU, H. (2014). High dimensional expectation-maximization algorithm: Statistical optimization and asymptotic normality. Unpublished manuscript.

[49] WEI, G. and TANNER, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *J. Amer. Statist. Assoc.* **85** 699–704.

[50] WU, C.-F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11** 95–103. MR0684867

[51] XU, L. and JORDAN, M. I. (1996). On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Comput.* **8** 129–151.

[52] XU, Q. and YOU, J. (2007). Covariate selection for linear errors-in-variables regression models. *Comm. Statist. Theory Methods* **36** 375–386. MR2391878

[53] YI, X., CARAMANIS, C. and SANGHAVI, S. (2013). Alternating minimization for mixed linear regression. Unpublished manuscript.

S. BALAKRISHNAN
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720
USA
AND
DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15213
USA
E-MAIL: siva@stat.cmu.edu

M. J. WAINWRIGHT
B. YU
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: wainwrig@berkeley.edu
binyu@berkeley.edu

*Submitted to the Annals of Statistics*
arXiv: arXiv:0000.0000

# SUPPLEMENTARY MATERIAL FOR: STATISTICAL GUARANTEES FOR THE EM ALGORITHM: FROM POPULATION TO SAMPLE-BASED ANALYSIS

By Sivaraman Balakrishnan, Martin J. Wainwright and Bin Yu

*University of California, Berkeley*

## APPENDIX A: EM AND FIRST-ORDER EM UPDATES FOR EXAMPLES

In this appendix, we derive the precise forms of the EM and first-order EM updates at both the population and finite-sample level for the three examples we consider. In Section A.4, we prove the claim (3.4).

**A.1. Mixture of Gaussians.** Suppose that we are given $n$ i.i.d. samples $\{y_i\}_{i=1}^n$ drawn from the mixture density (3.6). The complete data $\{(y_i, z_i)\}_{i=1}^n$ corresponds to the original samples along with the component indicator variables $z_i \in \{0, 1\}$. The sample-based function $Q_n$ takes the form

$$Q_n(\theta'|\theta) = -\frac{1}{2n}\sum_{i=1}^{n}\big[w_\theta(y_i)\|y_i - \theta'\|_2^2 + (1 - w_\theta(y_i))\|y_i + \theta'\|_2^2\big], \tag{A.1}$$

where $w_\theta(y) := e^{-\frac{\|\theta - y\|_2^2}{2\sigma^2}}\Big[e^{-\frac{\|\theta-y\|_2^2}{2\sigma^2}} + e^{-\frac{\|\theta+y\|_2^2}{2\sigma^2}}\Big]^{-1}$.

*EM updates:.* This example is especially simple in that each iteration of the EM algorithm has a closed form solution, given by

$$\theta^{t+1} := \arg\max_{\theta' \in \mathbb{R}^d} Q_n(\theta'|\theta^t) \;=\; \frac{2}{n}\sum_{i=1}^{n} w_{\theta^t}(y_i)y_i - \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{A.2a}$$

Iterations of the population EM algorithm are specified analogously

$$\theta^{t+1} = 2\mathbb{E}\big[w_{\theta^t}(Y)Y\big], \tag{A.2b}$$

where the empirical expectation has been replaced by expectation under the mixture distribution (3.6).

*First-order EM updates:.* On the other hand, the sample-based and population first-order EM operators with step size $\alpha > 0$ are given by

$$\theta^{t+1} = \theta^t + \alpha\Big\{\frac{1}{n}\sum_{i=1}^{n}(2w_{\theta^t}(y_i) - 1)y_i - \theta^t\Big\}, \quad \text{and} \quad \theta^{t+1} = \theta^t + \alpha\big[\,2\mathbb{E}\big[w_{\theta^t}(Y)Y\big] - \theta^t\big], \tag{A.3}$$

respectively.

**A.2. Mixture of regressions.**

*EM updates:.* Define the weight function

$$w_\theta(x,y) = \frac{\exp\left(\frac{-(y-\langle x,\theta\rangle)^2}{2\sigma^2}\right)}{\exp\left(\frac{-(y-\langle x,\theta\rangle)^2}{2\sigma^2}\right) + \exp\left(\frac{-(y+\langle x,\theta\rangle)^2}{2\sigma^2}\right)}. \tag{A.4a}$$

In terms of this notation, the sample EM update is based on maximizing the function

$$Q_n(\theta'|\theta) = -\frac{1}{2n}\sum_{i=1}^n \Big(w_\theta(x_i,y_i)(y_i - \langle x_i,\, \theta'\rangle)^2 + (1 - w_\theta(x_i,y_i))(y_i + \langle x_i,\, \theta'\rangle)^2\Big). \tag{A.4b}$$

Again, there is a closed form solution to this maximization problem: more precisely,

$$\theta^{t+1} = \Big(\sum_{i=1}^n x_i x_i^T\Big)^{-1}\Big(\sum_{i=1}^n (2w_{\theta^t}(x_i,y_i) - 1)y_i x_i\Big). \tag{A.5a}$$

Similarly, by an easy calculation, we find that the population EM iterations have the form

$$\theta^{t+1} = 2\mathbb{E}\big[w_{\theta^t}(X,Y)YX\big], \tag{A.5b}$$

where the expectation is taken over the joint distribution of the pair $(Y,X) \in \mathbb{R}\times\mathbb{R}^d$.

*First-order EM updates:.* On the other hand, the first-order EM operators are given by

$$\theta^{t+1} = \theta^t + \alpha\Big\{\frac{1}{n}\sum_{i=1}^n \Big[(2w_{\theta^t}(x_i,y_i) - 1)y_i x_i - x_i x_i^T\theta^t\Big]\Big\}, \quad \text{and} \tag{A.6a}$$

$$\theta^{t+1} = \theta^t + \alpha\,\mathbb{E}\Big[2w_{\theta^t}(X,Y)\,YX - \theta^t\Big], \tag{A.6b}$$

where $\alpha > 0$ is a step size parameter.

**A.3. Linear regression with missing covariates.** In this example, the E-step involves imputing the mean and covariance of the jointly Gaussian distribution of covariate-response pairs. For a given sample $(x,y)$, let $x_{\text{obs}}$ denote the observed portion of $x$, and let $\theta_{\text{obs}}$ denote the corresponding sub-vector of $\theta$. Define the missing portions $x_{\text{mis}}$ and $\theta_{\text{mis}}$ in an analogous fashion. With this notation, the EM algorithm imputes the conditional mean and conditional covariance using the current parameter estimate $\theta$. Using properties of joint Gaussians, the conditional mean of $X$ given $(x_{\text{obs}},y)$ is found to be

$$\mu_\theta(x_{\text{obs}},y) := \begin{bmatrix}\mathbb{E}(x_{\text{mis}}|x_{\text{obs}},y,\theta)\\ x_{\text{obs}}\end{bmatrix} = \begin{bmatrix}U_\theta z_{\text{obs}}\\ x_{\text{obs}}\end{bmatrix}, \tag{A.7a}$$

where

$$U_\theta = \frac{1}{\|\theta_{\text{mis}}\|_2^2 + \sigma^2}\begin{bmatrix}-\theta_{\text{mis}}\,\theta_{\text{obs}}^T & \theta_{\text{mis}}\end{bmatrix} \quad \text{and} \quad z_{\text{obs}} := \begin{bmatrix}x_{\text{obs}}\\ y\end{bmatrix} \in \mathbb{R}^{|x_{\text{obs}}|+1}. \tag{A.7b}$$

Similarly, the conditional second moment matrix takes the form

$$\Sigma_\theta(x_{\text{obs}},y) := \mathbb{E}\Big[XX^T \mid x_{\text{obs}},y,\theta\Big] = \begin{bmatrix}I & U_\theta z_{\text{obs}} x_{\text{obs}}^T\\ x_{\text{obs}} z_{\text{obs}}^T U_\theta^T & x_{\text{obs}}x_{\text{obs}}^T\end{bmatrix}. \tag{A.7c}$$

In writing all these expressions, we have assumed that the coordinates are permuted so that the missing values are in the first block.

We now have the necessary notation in place to describe the EM and first-order EM updates. For a given parameter $\theta$, the EM update is based on maximizing

$$Q_n(\theta'|\theta) := -\frac{1}{2n}\sum_{i=1}^{n}\langle\theta',\, \Sigma_\theta(x_{\mathrm{obs},i}, y_i)\theta'\rangle + \frac{1}{n}\sum_{i=1}^{n} y_i\langle\mu_\theta(x_{\mathrm{obs},i}, y_i),\, \theta'\rangle. \tag{A.8}$$

The sample-based EM iterations are given as

$$\theta^{t+1} := \Big[\sum_{i=1}^{n}\Sigma_{\theta^t}(x_{\mathrm{obs},i}, y_i)\Big]^{-1}\Big[\sum_{i=1}^{n} y_i\mu_{\theta^t}(x_{\mathrm{obs},i}, y_i)\Big], \tag{A.9a}$$

accompanied by its population counterpart

$$\theta^{t+1} := \big\{\mathbb{E}\big[\Sigma_{\theta^t}(X_{\mathrm{obs}}, Y)\big]\big\}^{-1}\,\mathbb{E}\big[Y\mu_{\theta^t}(X_{\mathrm{obs}}, Y)\big]. \tag{A.9b}$$

On the other hand, the first-order EM algorithm with step size $\alpha$ performs the following iterations:

$$\theta^{t+1} = \theta^t + \alpha\Big\{\frac{1}{n}\sum_{i=1}^{n}\big[y_i\mu_{\theta^t}(x_{\mathrm{obs},i}, y_i) - \Sigma_{\theta^t}(x_{\mathrm{obs},i}, y_i)\theta^t\big]\Big\}, \tag{A.10a}$$

along with the population counterpart

$$\theta^{t+1} = \theta^t + \alpha\mathbb{E}\Big[Y\mu_{\theta^t}(X_{\mathrm{obs}}, Y) - \Sigma_{\theta^t}(X_{\mathrm{obs}}, Y)\theta^t\Big]. \tag{A.10b}$$

**A.4. Proof of the claim (3.4).** From the strong-concavity of $q$, it is straightforward to verify that

$$\|\nabla q(\theta)\|_2^2 \geq \lambda^2\|\theta - \theta^*\|_2^2.$$

This inequality together with the gradient smoothness condition (3.1) yields

$$\begin{aligned}\langle\nabla Q(\theta|\theta),\, \nabla q(\theta)\rangle &= \frac{1}{2}\Big(\|\nabla Q(\theta|\theta)\|_2^2 + \|\nabla q(\theta)\|_2^2 - \|\nabla Q(\theta|\theta) - \nabla q(\theta)\|_2^2\Big)\\ &\geq \frac{1}{2}\Big(\lambda^2\|\theta - \theta^*\|_2^2 - \gamma^2\|\theta - \theta^*\|_2^2\Big)\\ &\geq 0,\end{aligned}$$

where the final inequality is strict whenever $\theta \neq \theta^*$.

## APPENDIX B: RESULTS FOR THE EM ALGORITHM

In this section, we revisit the examples introduced previously, and develop guarantees for the EM algorithm applied to them.

COROLLARY 7 (Sample-based EM guarantees for Gaussian mixtures). *In addition to the conditions of Corollary 1, suppose that the sample size is lower bounded as $n \geq c_1 d\log(1/\delta)$. Then given any initialization $\theta^0 \in \mathbb{B}_2(\frac{\|\theta^*\|_2}{4}; \theta^*)$, there is a contraction coefficient $\kappa(\eta) \leq e^{-c\eta^2}$ such that the standard EM iterates $\{\theta^t\}_{t=0}^{\infty}$ satisfy the bound*

$$\|\theta^t - \theta^*\|_2 \leq \kappa^t\|\theta^0 - \theta^*\|_2 + \frac{c_2}{1-\kappa}\varphi(\sigma; \|\theta^*\|_2)\,\sqrt{\frac{d}{n}\,\log(1/\delta)} \tag{B.1}$$

*with probability at least $1 - \delta$.*

We also have an analogous result for the EM algorithm with sample-splitting for the mixture of Gaussians.

COROLLARY 8 (Sample-splitting EM guarantees for Gaussian mixtures). *Consider a Gaussian mixture model satisfying the SNR($\eta$) condition (3.7), and any initialization $\theta^0$ such that $\|\theta^0 - \theta^*\|_2 \leq \frac{\|\theta^*\|_2}{4}$.*

*Given a sample size $n \geq 16T\log(6T/\delta)$, then with probability at least $1-\delta$, the sample-splitting EM iterates $\{\theta^t\}_{t=0}^T$ satisfy the bound*

$$\|\theta^t - \theta^*\|_2 \leq \kappa^t \|\theta^0 - \theta^*\|_2 + \frac{c}{1-\kappa}\Big(\sigma\sqrt{\frac{dT\log(T/\delta)}{n}} + \sqrt{\frac{T\log(T/\delta)}{n}}\|\theta^*\|_2\Big). \tag{B.2}$$

It is worth comparing the result here to the result established earlier in Corollary 7. The sample-splitting EM algorithm is more sensitive to the number of iterations which determines the batch size and needs to be chosen in advance. Supposing that the number of iterations were chosen optimally however the result has better dependence on $\|\theta^*\|_2$ and $\sigma$ at the cost of a logarithmic factor in $n$.

Our next corollary gives a guarantee for the sample-splitting EM updates applied to the mixture of regressions example. Let us now provide guarantees for a sample-splitting version of the EM updates. For a given sample size $n$ and iteration number $T$, suppose that we split[1] our full data set into $T$ subsets, each of size $n/T$. We then generate the sequence $\theta^{t+1} = M_{n/T}(\theta^t)$, where we use a fresh subset at each iteration. In the following result, we use $\varphi(\sigma; \|\theta^*\|_2) = \sqrt{\sigma^2 + \|\theta^*\|_2^2}$, along with positive universal constants ($c_1$,$c_2$).

COROLLARY 9 (Sample-splitting EM guarantees for MOR). *In addition to the conditions of Corollary 2, suppose that the sample size is lower bounded as $n \geq c_1 d\log(T/\delta)$. Then there is a contraction coefficient $\kappa \leq 1/2$ such that, for any initial vector $\theta^0 \in \mathbb{B}_2(\frac{\|\theta^*\|_2}{32}; \theta^*)$, the sample-splitting EM iterates $\{\theta^t\}_{t=1}^T$ based on $n/T$ samples per step satisfy the bound*

$$\|\theta^t - \theta^*\|_2 \leq \kappa^t \|\theta^0 - \theta^*\|_2 + c_2\varphi(\sigma; \|\theta^*\|_2)\sqrt{\frac{d}{n}\, T\log(T/\delta)} \tag{B.3}$$

*with probability at least $1-\delta$.*

We devote the remaining technical sections of this Appendix section to the proofs of the various claims made in this Appendix and earlier in Section 5.

**B.1. Proof of Theorem 4.** Since both $M(\theta)$ and $\theta^*$ are in $\Omega$, we may apply condition (5.2) with $\theta' = M(\theta)$ and condition (5.3) with $\theta' = \theta^*$. Doing so, adding the resulting inequalities and then performing some algebra yields the condition

$$\langle \nabla Q(M(\theta)|\theta^*) - \nabla Q(\theta^*|\theta^*),\, \theta^* - M(\theta)\rangle \leq \langle \nabla Q(M(\theta)|\theta^*) - \nabla Q(M(\theta)|\theta),\, \theta^* - M(\theta)\rangle. \tag{B.4}$$

Now the $\lambda$-strong concavity condition (3.2) implies that the left-hand side is lower bounded as

$$\langle \nabla Q(M(\theta)|\theta^*) - \nabla Q(\theta^*|\theta^*),\, \theta^* - M(\theta)\rangle \geq \lambda\|\theta^* - M(\theta)\|_2^2. \tag{B.5a}$$

[1] To simplify exposition, assume that $n/T$ is an integer.

On the other hand, the FOS($\gamma$) condition together with the Cauchy-Schwarz inequality implies that the right-hand side is upper bounded as

$$\langle \nabla Q(M(\theta)|\theta^*) - \nabla Q(M(\theta)|\theta), \theta^* - M(\theta)\rangle \leq \gamma \|\theta^* - M(\theta)\|_2 \|\theta - \theta^*\|_2, \tag{B.5b}$$

Combining inequalities (B.5a) and (B.5b) with the original bound (B.4) yields

$$\lambda \|\theta^* - M(\theta)\|_2^2 \leq \gamma \|\theta^* - M(\theta)\|_2 \|\theta - \theta^*\|_2,$$

and canceling terms completes the proof.

**B.2. Proof of Theorems 3 and 5.** For concreteness we prove Theorem 3 in this section. The proof of Theorem 5 follows in a similar manner. The proof follows along similar lines to the proof of Theorem 2. For any iteration $s \in \{1, 2, \ldots, T\}$, we have

$$\|\nabla Q_{n/T}(\theta|\theta^s)|_{\theta=\theta^s} - \nabla Q(\theta|\theta^s)|_{\theta=\theta^s}\|_2 \leq \varepsilon_Q\Big(\frac{n}{T}, \frac{\delta}{T}\Big) \tag{B.6}$$

with probability at least $1 - \frac{\delta}{T}$. Consequently, by a union bound over all $T$ indices, the bound (B.6) holds uniformly with probability at least $1 - \delta$. We perform the remainder of our analysis under this event.

It suffices to show that

$$\|\theta^{s+1} - \theta^*\|_2 \leq \kappa \|\theta^s - \theta^*\|_2 + \alpha \varepsilon_Q\Big(\frac{n}{T}, \frac{\delta}{T}\Big) \qquad \text{for each iteration } s \in \{1, 2, \ldots, T-1\}. \tag{B.7}$$

Indeed, when this bound holds, we may iterate it to show that

$$\begin{aligned}
\|\theta^t - \theta^*\|_2 &\leq \kappa \|\theta^{t-1} - \theta^*\|_2 + \alpha \varepsilon_Q\Big(\frac{n}{T}, \frac{\delta}{T}\Big) \\
&\leq \kappa\Big\{\kappa \|\theta^{t-2} - \theta^*\|_2 + \alpha \varepsilon_Q\Big(\frac{n}{T}, \frac{\delta}{T}\Big)\Big\} + \alpha \varepsilon_Q\Big(\frac{n}{T}, \frac{\delta}{T}\Big) \\
&\leq \kappa^t \|\theta^0 - \theta^*\|_2 + \Big\{\sum_{s=0}^{t-1} \kappa^s\Big\} \alpha \varepsilon_Q\Big(\frac{n}{T}, \frac{\delta}{T}\Big) \\
&\leq \kappa^t \|\theta^0 - \theta^*\|_2 + \frac{1}{1-\kappa} \alpha \varepsilon_Q\Big(\frac{n}{T}, \frac{\delta}{T}\Big),
\end{aligned}$$

where the final step follows by summing the geometric series.

It remains to prove the claim (B.7), and we do so via induction on the iteration number. Beginning with $s = 1$, we have

$$\|\theta^1 - \theta^*\|_2 \;=\; \|\theta^0 + \alpha \nabla Q_{n/T}(\theta|\theta^0)|_{\theta=\theta^0} - \theta^*\|_2 \overset{\text{(i)}}{\leq} \kappa \|\theta^0 - \theta^*\|_2 + \alpha \varepsilon_Q\Big(\frac{n}{T}, \frac{\delta}{T}\Big),$$

where step (i) follows by triangle inequality, the bound (B.6), and the contractivity of the population operator applied to $\theta^0 \in \mathbb{B}_2(r; \theta^*)$. By our initialization condition and the bound (5.11a), note that we are guaranteed that $\|\theta^1 - \theta^*\|_2 \leq r$.

In the induction from $s \mapsto s+1$, suppose that $\|\theta^s - \theta^*\|_2 \leq r$, and the bound (6.3) holds at iteration $s$. The same argument then implies that the bound (6.3) also holds for iteration $s+1$, and that $\|\theta^{s+1} - \theta^*\|_2 \leq r$, thus completing the proof.

**B.3. Population contractivity of the EM operator.** Much of the work in establishing population level results for the first-order EM algorithm can be leveraged in establishing population level results for the EM algorithm.

In particular, we observe that the FOS condition differs from the GS condition only in that for the FOS condition we need to control the norm:

$$\|\nabla Q(M(\theta)|\theta) - \nabla Q(M(\theta)|\theta^*)\|_2$$

as opposed to the norm:

$$\|\nabla Q(\theta|\theta) - \nabla Q(\theta|\theta^*)\|_2.$$

It is a straightforward exercise to verify that in the two mixture model examples we consider in the paper (and more generally when the $Q$ function is a spherical quadratic function of its first argument) the gradient of the $Q$ function is independent of its first argument. Particularly, we note that the population level results for EM applied to the mixture of regressions and mixture of Gaussians follow directly from Corollaries 1 and 2, and the observation that for these examples the FOS and GS conditions are equivalent. It then remains to analyze the finite-sample performance of the EM algorithm in these examples, and we address this in the following sections.

**B.4. Proof of Corollary 8.** The proof follows by establishing a bound on the function $\varepsilon_M(n, \delta)$. Define $\mathcal{S} = \{\theta : \|\theta - \theta^*\|_2 \leq \frac{\|\theta^*\|_2}{4}\}$. Recalling the updates in (A.2a) and (A.2b), note that

$$\|M(\theta) - M_n(\theta)\|_2 \leq \underbrace{\|\frac{1}{n}\sum_{i=1}^n Y_i\|_2}_{T_1} + \underbrace{\|\frac{1}{n}\sum_{i=1}^n w_\theta(Y_i)Y_i - \mathbb{E}w_\theta(Y)Y\|_2}_{T_2}.$$

We bound each of these terms in turn, in particular showing that

$$\max\{T_1, T_2\} \leq \sqrt{\frac{\log(8/\delta)}{2n}}\|\theta^*\|_2 + c\sigma\sqrt{\frac{d\log(1/\delta)}{n}}, \tag{B.8}$$

with probability at least $1 - \delta$.

*Control of $T_1$:.* Observe that since $Y \sim (2Z - 1)\theta^* + v$ we have

$$T_1 = \|\frac{1}{n}\sum_{i=1}^n Y_i\|_2 \leq \|\frac{1}{n}\sum_{i=1}^n v_i\|_2 + \Big|\frac{1}{n}\sum_{i=1}^n (2Z_i - 1)\Big|\|\theta^*\|_2.$$

Since $Z_i$ are i.i.d Bernoulli variables, Hoeffding's inequality implies that

$$\Big|\frac{1}{n}\sum_{i=1}^n (2Z_i - 1)\Big| \leq \sqrt{\frac{\log(8/\delta)}{2n}}.$$

with probability at least $1 - \frac{\delta}{4}$. On the other hand, the vector $U_1 := \frac{1}{n}\sum_{i=1}^n v_i$ is zero-mean and sub-Gaussian with parameter $\sigma/\sqrt{n}$, whence the squared norm $\|U_1\|_2^2$ is sub-exponential. Using standard bounds for sub-exponential variates and the condition $n > \sigma d$, we obtain

$$\|U_1\|_2 \leq c_2\sigma\sqrt{\frac{d\log(1/\delta)}{n}}.$$

with probability at least $1 - \delta/4$. Combining the pieces yields the claimed bound (B.8) on $T_1$.

*Control of $T_2$:.* By triangle inequality, we have

$$T_2 \leq \big|\frac{1}{n}\sum_{i=1}^n w_\theta(Y_i)(2Z_i - 1) - \mathbb{E}w_\theta(Y)(2Z-1)\big|\|\theta^*\|_2 + \|\frac{1}{n}\sum_{i=1}^n w_\theta(Y_i)v_i - \mathbb{E}w_\theta(Y)v\|_2.$$

The random variable $w_\theta(Y)(2Z-1)$ lies in the interval $[-1,1]$, so that Hoeffding's inequality implies that

$$\big|\frac{1}{n}\sum_{i=1}^n w_\theta(Y_i)(2Z_i - 1) - \mathbb{E}w_\theta(Y)(2Z-1)\big|\|\theta^*\|_2 \leq \sqrt{\frac{\log(6/\delta)}{2n}}\|\theta^*\|_2.$$

with probability at least $1 - \delta/4$.

Next observe that the random vector $U_2 := \frac{1}{n}\sum_{i=1}^n w_{\theta^t}(X_i)v_i - \mathbb{E}w_{\theta^t}(X)v$ is zero mean and sub-Gaussian with parameter $\sigma/\sqrt{n}$. Consequently, as in our analysis of $T_1$,we conclude that

$$\|U_2\|_2 \leq c\sigma\sqrt{\frac{d\log(1/\delta)}{n}}.$$

with probability at least $1 - \delta/4$. Putting together the pieces yields the claimed bound (B.8) on $T_2$, thereby completing the proof of the corollary.

**B.5. Proof of Corollary 7.** In order to prove this corollary, it suffices to bound the function $\varepsilon_M^{\text{unif}}(n,\delta)$, as previously defined (5.9). Defining the set $\mathbb{A} := \big\{\theta \in \mathbb{R}^d \mid \|\theta - \theta^*\|_2 \leq \|\theta^*\|_2/4\big\}$, our goal is to control the random variable $Z := \sup_{\theta\in\mathbb{A}}\|M(\theta) - M_n(\theta)\|_2$. For each unit-norm vector $u \in \mathbb{R}^d$, define the random variable

$$Z_u := \sup_{\theta\in\mathbb{A}}\big\{\frac{1}{n}\sum_{i=1}^n (2w_\theta(y_i) - 1)\langle y_i, u\rangle - \mathbb{E}(2w_\theta(Y) - 1)\langle Y, u\rangle\big\}.$$

Noting that $Z = \sup_{u\in\mathbb{S}^d} Z_u$, we begin by reducing our problem to a finite maximum over the sphere $\mathbb{S}^d$. Let $\{u^1, \ldots, u^M\}$ denote a $1/2$-covering of the sphere $\mathbb{S}^d = \{v \in \mathbb{R}^d \mid \|v\|_2 = 1\}$. For any $v \in \mathbb{S}^d$, there is some index $j \in [M]$ such that $\|v - u^j\|_2 \leq 1/2$, and hence we can write

$$Z_v \leq Z_{u^j} + |Z_v - Z_{u^j}| \leq \max_{j\in[M]} Z_{u^j} + Z\|v - u^j\|_2,$$

where the final step uses the fact that $|Z_u - Z_v| \leq Z\|u - v\|_2$ for any pair $(u,v)$. Putting together the pieces, we conclude that

$$Z = \sup_{v\in\mathbb{S}^d} Z_v \leq 2\max_{j\in[M]} Z_{u^j}. \tag{B.9}$$

Consequently, it suffices to bound the random variable $Z_u$ for a fixed $u \in \mathbb{S}^d$. Letting $\{\varepsilon_i\}_{i=1}^n$ denote an i.i.d. sequence of Rademacher variables, for any $\lambda > 0$, we have

$$\mathbb{E}\big[e^{\lambda Z_u}\big] \leq \mathbb{E}\Big[\exp\Big(\frac{2}{n}\sup_{\theta\in\mathbb{A}}\sum_{i=1}^n \varepsilon_i(2w_\theta(y_i) - 1)\langle y_i, u\rangle\Big)\Big],$$

using a standard symmetrization result for empirical processes (e.g., [4, 5]). Now observe that for any triplet of $d$-vectors $y$, $\theta$ and $\theta'$, we have the Lipschitz property

$$\big|2w_\theta(y) - 2w_{\theta'}(y)\big| \leq \frac{1}{\sigma^2}\big|\langle\theta, y\rangle - \langle\theta', y\rangle\big|.$$

Consequently, by the Ledoux-Talagrand contraction for Rademacher processes [4, 5], we have

$$\mathbb{E}\Big[\exp\Big(\frac{2}{n}\sup_{\theta\in\mathbb{A}}\sum_{i=1}^{n}\varepsilon_i(2w_\theta(y_i)-1)\langle y_i,\,u\rangle\Big)\Big] \leq \mathbb{E}\Big[\exp\Big(\frac{4}{n\sigma^2}\sup_{\theta\in\mathbb{A}}\sum_{i=1}^{n}\varepsilon_i\langle\theta,\,y_i\rangle\langle y_i,\,u\rangle\Big)\Big]$$

Since any $\theta\in\mathbb{A}$ satisfies $\|\theta\|_2\leq\frac{5}{4}\|\theta^*\|_2$, we have

$$\sup_{\theta\in\mathbb{A}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\langle\theta,\,y_i\rangle\langle y_i,\,u\rangle \leq \frac{5}{4}\|\theta^*\|_2\Big\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i y_i y_i^T\Big\|_{\rm op},$$

where $\|\cdot\|_{\rm op}$ denotes the $\ell_2$-operator norm of a matrix (maximum singular value). Repeating the same discretization argument over $\{u^1,\ldots,u^M\}$, we find that

$$\Big\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i y_i y_i^T\Big\|_{\rm op} \leq 2\max_{j\in[M]}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\langle y_i,\,u^j\rangle^2.$$

Putting together the pieces, we conclude that

(B.10)

$$\mathbb{E}\big[e^{\lambda Z_u}\big] \leq \mathbb{E}\Big[\exp\Big(\frac{10\lambda\|\theta^*\|_2}{\sigma^2}\max_{j\in[M]}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\langle y_i,\,u^j\rangle^2\Big)\Big] \;\leq\; \sum_{j=1}^{M}\mathbb{E}\Big[\exp\Big(\frac{10\lambda\|\theta^*\|_2}{\sigma^2}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\langle y_i,\,u^j\rangle^2\Big)\Big].$$

Now by assumption, the random vectors $\{y_i\}_{i=1}^n$ are generated i.i.d. according to the model $y=\eta\theta^*+w$, where $\eta$ is a Rademacher sign variable, and $w\sim\mathcal{N}(0,\sigma^2 I)$. Consequently, for any $u\in\mathbb{R}^d$, we have

$$\mathbb{E}[e^{\langle u,\,y\rangle}] = \mathbb{E}[e^{\eta\langle u,\,\theta^*\rangle}]\,\mathbb{E}[e^{\langle u,\,w\rangle}] \;\leq\; e^{\frac{\|\theta^*\|_2^2+\sigma^2}{2}},$$

showing that the vectors $\langle y_i,\,u\rangle$ are sub-Gaussian with parameter at most $\gamma=\sqrt{\|\theta^*\|_2^2+\sigma^2}$. Therefore, the vectors $\varepsilon_i\langle y_i,\,u\rangle^2$ are zero mean sub-exponential, and have moment generating function bounded as $\mathbb{E}[e^{t(\langle y_i,\,u\rangle)^2}]\leq e^{\frac{\gamma^4t^2}{2}}$ for all $t>0$ sufficiently small. Combined with our earlier inequality (B.10), we conclude that

$$\mathbb{E}\big[e^{\lambda Z_u}\big] \leq M\,e^{c\,\frac{\lambda^2\|\theta^*\|_2^2\gamma^4}{n\sigma^4}} \;\leq\; e^{c\,\frac{\lambda^2\|\theta^*\|_2^2\gamma^4}{n\sigma^4}+2d}$$

for all $\lambda$ sufficiently small. Combined with our first discretization (B.9), we have thus shown that

$$\mathbb{E}[e^{\frac{\lambda}{2}Z}] \leq Me^{c\,\frac{\lambda^2\|\theta^*\|_2^2\gamma^4}{n\sigma^4}+2d} \;\leq\; e^{c\,\frac{\lambda^2\|\theta^*\|_2^2\gamma^4}{n\sigma^4}+4d}.$$

Combined with the Chernoff approach, this bound on the MGF implies that, as long as $n\geq c_1 d\log(1/\delta)$ for a sufficiently large constant $c_1$, we have

$$Z \leq \frac{c_2\|\theta^*\|_2\gamma^2}{\sigma^2}\sqrt{\frac{d\log(1/\delta)}{n}}$$

with probability at least $1-\delta$.

B.5.1. *Proof of Corollary 9.* We need to compute an upper bound on the function $\varepsilon_M(n, \delta)$ previously defined in equation (5.8). For this particular model, we have

$$\|M(\theta) - M_n(\theta)\|_2 = \|\Big(\sum_{i=1}^n x_i x_i^T\Big)^{-1}\Big(\sum_{i=1}^n (2w_\theta(x_i, y_i) - 1) y_i x_i\Big) - 2\mathbb{E}[w_\theta(X, Y) YX]\|_2.$$

Define the matrices $\widehat{\Sigma} := \frac{1}{n}\sum_{i=1}^n x_i x_i^T$ and $\Sigma = \mathbb{E}[XX^T] = I$, as well as the vector

$$\widehat{v} := \frac{1}{n}\sum_{i=1}^n \big[\mu_\theta(x_i, y_i) y_i x_i\big], \quad \text{and} \quad v := \mathbb{E}\big[\mu_\theta(X, Y) YX\big],$$

where $\mu_\theta(x, y) := 2w_\theta(x, y) - 1$. Noting that $\mathbb{E}[YX] = 0$, some straightforward algebra then yields the bound

$$\|M(\theta) - M_n(\theta)\|_2 \leq \underbrace{|||\widehat{\Sigma}^{-1}|||_{\mathrm{op}} \|\widehat{v} - v\|_2}_{T_1} + \underbrace{|||\widehat{\Sigma}^{-1} - \Sigma^{-1}|||_{\mathrm{op}} \|v\|_2}_{T_2}. \tag{B.11}$$

We bound each of the terms $T_1$ and $T_2$ in turn.

*Bounding $T_1$:.* Recall the assumed lower bound on the sample size—namely $n > c\, d \log(1/\delta)$ for a sufficiently large constant $c$. Under this condition, standard bounds in random matrix theory [11], guarantee that $|||\widehat{\Sigma} - \Sigma|||_{\mathrm{op}} \leq \frac{1}{2}$ with probability at least $1 - \delta$. When this bound holds, we have $|||\widehat{\Sigma}^{-1}|||_{\mathrm{op}} \geq 1/2$.

As for the other part of $T_1$, let us write $\|\widehat{v} - v\|_2 = \sup_{u \in \mathbb{S}^d} Z(u)$, where

$$Z(u) := \frac{1}{n}\sum_{i=1}^n \mu_\theta(x_i, y_i) y_i \langle x, u\rangle - \mathbb{E}[\mu_\theta(X, Y) Y \langle X, u\rangle].$$

By a discretization argument over a 1/2-cover of the sphere $\mathbb{S}^d$—say $\{u^1, \ldots, u^M\}$—we have the upper bound $\|\widehat{v} - v\|_2 \leq 2 \max_{j \in [M]} Z(u^j)$. Thus, it suffices to control the random variable $Z(u)$ for a fixed $u \in \mathbb{S}^d$. By a standard symmetrization argument [10], we have

$$\mathbb{P}\big[Z(u) \geq t\big] \leq 2\mathbb{P}\Big[\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mu_\theta(x_i, y_i) y_i \langle x_i, u\rangle \geq t/2\Big],$$

where $\{\varepsilon_i\}_{i=1}^n$ are an i.i.d. sequence of Rademacher variables. Let us now define the event $\mathcal{E}\big\{\frac{1}{n}\sum_{i=1}^n \langle x_i, u\rangle^2 \leq 2\big\}$. Since each variable $\langle x_i, u\rangle$ is sub-Gaussian with parameter one, standard tail bounds imply that $\mathbb{P}[\mathcal{E}^c] \leq e^{-n/32}$. Therefore, we can write

$$\mathbb{P}\big[Z(u) \geq t\big] \leq 2\mathbb{P}\Big[\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mu_\theta(x_i, y_i) y_i \langle x_i, u\rangle \geq t/2 \mid \mathcal{E}\Big] + 2e^{-n/32}.$$

As for the remaining term, we have

$$\mathbb{E}\Big[\exp\Big(\frac{\lambda}{n}\sum_{i=1}^n \varepsilon_i \mu_\theta(x_i, y_i) y_i \langle x_i, u\rangle\Big) \mid \mathcal{E}\Big] \leq \mathbb{E}\Big[\exp\Big(\frac{2\lambda}{n}\sum_{i=1}^n \varepsilon_i y_i \langle x_i, u\rangle\Big) \mid \mathcal{E}\Big],$$

where we have applied the Ledoux-Talagrand contraction for Rademacher processes [4, 5], using the fact that $|\mu_\theta(x, y)| \leq 1$ for all pairs $(x, y)$. Now conditioned on $x_i$, the random variable $y_i$

is zero-mean and sub-Gaussian with parameter at most $\sqrt{\|\theta^*\|_2^2 + \sigma^2}$. Consequently, taking expectations over the distribution $(y_i \mid x_i)$ for each index $i$, we find that

$$\mathbb{E}\Big[\exp\Big(\frac{2\lambda}{n}\sum_{i=1}^n \varepsilon_i y_i \langle x_i,\, u\rangle\Big) \mid \mathcal{E}\Big] \leq \Big[\exp\Big(\frac{4\lambda^2}{n^2}\big(\|\theta^*\|_2^2 + \sigma^2\big)\sum_{i=1}^n \langle x_i,\, u\rangle^2\Big) \mid \mathcal{E}\Big]$$
$$\leq \exp\Big(\frac{8\lambda^2}{n}\big(\|\theta^*\|_2^2 + \sigma^2\big)\Big),$$

where the final inequality uses the definition of $\mathcal{E}$. Using this bound on the moment-generating function, we find that

$$\mathbb{P}\Big[\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mu_\theta(x_i, y_i) y_i \langle x_i,\, u\rangle \geq t/2 \mid \mathcal{E}\Big] \leq \exp\Big(-\frac{nt^2}{256(\|\theta^*\|_2^2 + \sigma^2)}\Big).$$

Since the $1/2$-cover of the unit sphere $\mathbb{S}^d$ has at most $2^d$ elements, we conclude that there is a universal constant $c$ such that $T_1 \leq c\sqrt{\|\theta^*\|_2^2 + \sigma^2}\sqrt{\frac{d}{n}\log(1/\delta)}$ with probability at least $1-\delta$.

*Bounding $T_2$:.* Since $n > d$ by assumption, standard results in random matrix theory [11] imply that $\|\!|\widehat{\Sigma}^{-1} - \Sigma^{-1}|\!\|_{\mathrm{op}} \leq c\sqrt{\frac{d}{n}\log(1/\delta)}$ with probability at least $1-\delta$. On the other hand, observe that

$$\|v\|_2 = \|M(\theta)\|_2 \leq 2\|\theta^*\|_2,$$

since the population operator $M$ is a contraction, and $\|\theta\|_2 \leq 2\|\theta^*\|_2$. Combining the pieces, we see that $T_2 \leq c\|\theta^*\|_2\sqrt{\frac{d}{n}\log(1/\delta)}$ with probability at least $1-\delta$.

Finally, substituting our bounds on $T_1$ and $T_2$ into the decomposition (B.11) yields the claim.

## APPENDIX C: A STOCHASTIC VERSION OF FIRST-ORDER EM

In this section, we analyze a sample-based variant of first-order EM that is inspired by stochastic approximation. Online variants of the EM algorithm have been studied by various authors (see for instance [3, 6]), who focus on the convergence and rate of convergence of these algorithms to any stationary point of the log-likelihood. On the contrary, our focus in particular applications is on the convergence of the algorithm to the MLE. The stochastic algorithm we study can be viewed as an extreme form of sample-splitting, in which we use only a single sample per iteration, but compensate for the noisiness using a decaying step size. Throughout this section we assume that (a lower bound on) the radius of convergence $r$ of the population operator is known to the algorithm[2].

**C.1. Analysis of stochastic first-order EM.** Given a sequence of positive step sizes $\{\alpha^t\}_{t=0}^\infty$, we analyze the recursion

$$\theta^{t+1} = \Pi\Big(\theta^t + \alpha^t \nabla Q_1(\theta^t|\theta^t)\Big), \tag{C.1}$$

where the gradient $\nabla Q_1(\theta^t|\theta^t)$ is computed using a single fresh sample at each iteration. Here $\Pi$ denotes the projection onto the Euclidean ball $\mathbb{B}_2(\frac{r}{2}; \theta^0)$ of radius $\frac{r}{2}$ centered at the initial iterate $\theta^0$. Thus, given any initial vector $\theta^0$ in the ball of radius $r/2$ centered at $\theta^*$, we are guaranteed that all iterates remain within an $r$-ball of $\theta^*$. The following result is stated in terms of the constant $\xi := \frac{2\mu\lambda}{\lambda+\mu} - \gamma > 0$, and the uniform variance $\sigma_{\mathrm{G}}^2 := \sup_{\theta \in \mathbb{B}_2(r;\theta^*)} \mathbb{E}\|\nabla Q_1(\theta|\theta)\|_2^2$.

[2]This assumption can be restrictive in practice. We believe the requirement can be eliminated by a more judicious choice of the step-size parameter in the first few iterations.

THEOREM 6. *For a triplet $(\gamma, \lambda, \mu)$ such that $0 \leq \gamma < \lambda \leq \mu$, suppose that the population function $q$ is $\lambda$-strongly concave (3.2), $\mu$-smooth (3.3), and satisfies the GS($\gamma$) condition (3.1) over the ball $\mathbb{B}_2(r; \theta^*)$. Then given an initialization $\theta^0 \in \mathbb{B}_2(\frac{r}{2}; \theta^*)$, the stochastic EM gradient updates (C.1) with step size $\alpha^t := \frac{3}{2\xi\,(t+2)}$ satisfy the bound*

$$\text{(C.2)} \quad \mathbb{E}[\|\theta^t - \theta^*\|_2^2] \leq \frac{9\sigma_{\mathrm{G}}^2}{\xi^2} \frac{1}{(t+2)} + \Big(\frac{2}{t+2}\Big)^{3/2} \|\theta^0 - \theta^*\|_2^2 \qquad \text{for iterations } t = 1, 2, \ldots.$$

While the stated claim (C.2) provides bounds in expectation, it is also possible to obtain high-probability results.[3]
We prove this result in Appendix C.4. In order to obtain guarantees for stochastic first-order EM applied to specific models, it only remains to prove the concavity and smoothness properties of the population function $q$, and to bound the uniform variance $\sigma_{\mathrm{G}}$.

**C.2. Stochastic updates for mixture of regressions.** In order to illustrate Theorem 6 we first revisit the MOR model considered in Section 3.2.2. In particular, given a data set of size $n$ from this model, we run the algorithm for $n$ iterations, with a step size $\alpha^t := \frac{3}{t+2}$ for iterations $t = 1, \ldots, n$. Once again our result is terms of $\varphi(\sigma; \|\theta^*\|_2) = \sqrt{\sigma^2 + \|\theta^*\|_2^2}$ and positive universal constants ($c_1$,$c_2$).

COROLLARY 10 (Stochastic first-order EM guarantees for MOR). *In addition to the conditions of Corollary 2, suppose that the sample size is lower bounded as $n \geq c_1 d \log(1/\delta)$. Then given any initialization $\theta^0 \in \mathbb{B}_2(\frac{\|\theta^*\|_2}{32}; \theta^*)$, performing $n$ iterations of the stochastic first-order EM gradient updates (C.1) yields an estimate $\widehat{\theta} = \theta^n$ such that*

$$\text{(C.3)} \qquad \mathbb{E}[\|\widehat{\theta} - \theta^*\|_2^2] \leq c_2\, \varphi^2(\sigma; \|\theta^*\|_2)\, \frac{d}{n}.$$

We prove this corollary in Appendix C.5. Figure 9 illustrates this corollary showing the error as a function of iteration number (sample size) for the stochastic first-order EM algorithm.

**C.3. Stochastic updates for missing data.** We conclude our discussion of the stochastic form of first-order EM by re-visiting the model with covariates missing completely at random considered in Section 3.2.3. In particular, given a data set of size $n$ from this model, we run the algorithm for $n$ iterations, with a step size $\alpha^t := \frac{3}{t+2}$ for iterations $t = 1, \ldots, n$.

COROLLARY 11 (Stochastic first-order EM guarantees for missing covariates). *In addition to the conditions of Corollary 3, suppose that the sample size is lower bounded as $n \geq c_1 d \log(1/\delta)$. Then given any initialization $\theta^0 \in \mathbb{B}_2(\xi_2 \sigma; \theta^*)$, performing $n$ iterations of the stochastic EM gradient updates (C.1) with step sizes $\alpha^t = \frac{3}{2\,(1-\kappa)(t+2)}$ yields an estimate $\widehat{\theta} = \theta^n$ such that*

$$\text{(C.4)} \qquad \mathbb{E}[\|\widehat{\theta} - \theta^*\|_2^2] \leq c_2(1 + \sigma^2)\, \frac{d}{n}.$$

We prove this corollary in Appendix C.6. Figure 10 provides an illustration of the performance in practice.

[3] Although we do not consider this extension here, stronger exponential concentration results follow from controlling the moment generating function of the random variable $\sup_{\theta \in \mathbb{B}_2(r;\theta^*)} \|\nabla Q_1(\theta|\theta)\|_2^2$. For instance, see Nemirovski et al. [7] for such results in the context of stochastic optimization.

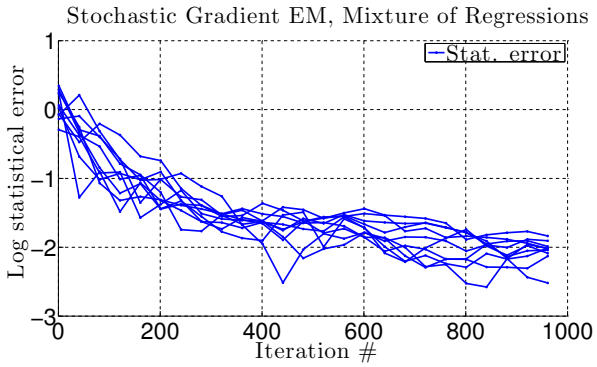


**Fig 9.** A plot of the (log) statistical error for the stochastic first-order EM algorithm as a function of iteration number (sample size) for the mixture of regressions example. The plot shows 10 different problem instances with $d = 10$, $\frac{\|\theta^*\|_2}{\sigma} = 2$ and $\frac{\|\theta^0 - \theta^*\|_2}{\sigma} = 1$. The statistical error decays at the sub-linear rate $\mathcal{O}(1/\sqrt{t})$ as a function of the iteration number $t$. An iteration of stochastic first-order EM is however typically much faster and uses only a single sample.

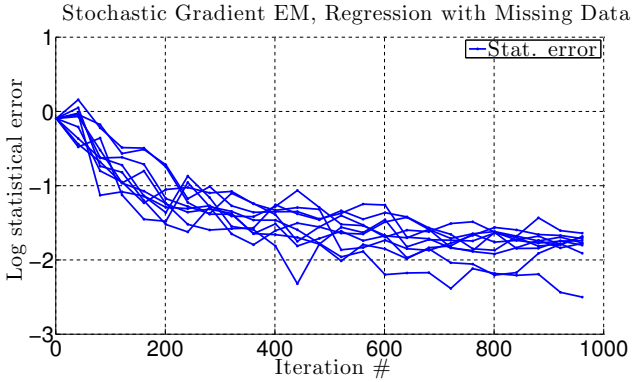


**Fig 10.** A plot of the (log) statistical error for the stochastic first-order EM algorithm as a function of iteration number (sample size) for the problem of linear regression with missing covariates. The plot shows 10 different problem instances with $d = 10$, $\frac{\|\theta^*\|_2}{\sigma} = 2$ and $\frac{\|\theta^0 - \theta^*\|_2}{\sigma} = 1$. The statistical error decays at the sub-linear rate $\mathcal{O}(1/\sqrt{t})$ as a function of the iteration number $t$.

**C.4. Proof of Theorem 6.** We first establish a recursion on the expected mean-squared error. As with Theorem 1 this result is established by relating the population first-order EM operator to the gradient ascent operator on the function $q(\cdot)$. This key recursion along with some algebra lead to the theorem.

LEMMA 5. *Given the stochastic EM gradient iterates with step sizes* $\{\alpha^t\}_{t=0}^{\infty}$, *the error* $\Delta^{t+1} := \theta^{t+1} - \theta^*$ *at iteration* $t+1$ *satisfies the recursion*

$$\mathbb{E}[\|\Delta^{t+1}\|_2^2] \leq \Big\{1 - \alpha^t \xi\Big\} \mathbb{E}[\|\Delta^t\|_2^2] + (\alpha^t)^2 \sigma_{\mathrm{G}}^2, \tag{C.5}$$

*where* $\sigma_{\mathrm{G}}^2 = \sup_{\theta \in \mathbb{B}_2(r;\theta^*)} \mathbb{E}[\|\nabla Q_1(\theta|\theta)\|_2^2]$.

We prove this lemma in the sequel.

Using this result, we can now complete the proof of the bound (C.2). With the step size

choice $\alpha^t := \frac{a}{\xi\,(t+2)}$ where $a = \frac{3}{2}$, unwrapping the recursion (C.5) yields

$$\text{(C.6)} \quad \mathbb{E}[\|\Delta^{t+1}\|_2^2] \leq \frac{a^2\sigma_{\mathrm{G}}^2}{\xi^2}\sum_{\tau=2}^{t+1}\Big\{\frac{1}{\tau^2}\prod_{\ell=\tau+1}^{t+2}\Big(1-\frac{a}{\ell}\Big)\Big\} + \frac{a^2\sigma_{\mathrm{G}}^2}{\xi^2(t+2)^2} + \prod_{\ell=2}^{t+2}\Big(1-\frac{a}{\ell}\Big)\,\mathbb{E}[\|\Delta^0\|_2^2].$$

In order to bound these terms we use the following fact: For any $a \in (1,2)$, we have

$$\prod_{\ell=\tau+1}^{t+2}\Big(1-\frac{a}{\ell}\Big) \leq \Big(\frac{\tau+1}{t+3}\Big)^a.$$

See Noorshams and Wainwright [9] for a proof. Applying this inequality in equation (C.6) yields

$$\begin{aligned}\mathbb{E}[\|\Delta^{t+1}\|_2^2] &\leq \frac{a^2\sigma_{\mathrm{G}}^2}{\xi^2\,(t+3)^a}\sum_{\tau=2}^{t+2}\frac{(\tau+1)^a}{\tau^2} + \Big(\frac{2}{t+3}\Big)^a\mathbb{E}[\|\Delta^0\|_2^2] \\ &\leq \frac{2a^2\sigma_{\mathrm{G}}^2}{\xi^2\,(t+3)^a}\sum_{\tau=2}^{t+2}\frac{1}{\tau^{2-a}} + \Big(\frac{2}{t+3}\Big)^a\mathbb{E}[\|\Delta^0\|_2^2].\end{aligned}$$

Finally, applying the integral upper bound $\sum_{\tau=2}^{t+2}\frac{1}{\tau^{2-a}} \leq \int_1^{t+2}\frac{1}{x^{2-a}}dx \;\leq 2(t+3)^{a-1}$ yields the claim (C.2).

It only remains to prove Lemma 5. In order to establish Lemma 5 we require an analogue of Theorem 1 that allows for a wider range of step sizes. Recall the classical gradient ascent operator on the function $q(\theta) = Q(\theta|\theta^*)$. For step size $\alpha > 0$, it takes the form $T(\theta) = \theta + \alpha\nabla q(\theta)$. Under the stated $\lambda$-concavity and $\mu$-smoothness conditions, for any step size $0 < \alpha \leq \frac{2}{\lambda+\mu}$, the classical gradient operator $T$ is contractive with parameter

$$\phi(\alpha) = 1 - \frac{2\alpha\mu\lambda}{\mu+\lambda}.$$

This follows from the classical analysis of gradient descent (e.g., [1, 2, 8]). Using this fact, we can prove the following about the population first-order EM operator:

Lemma 6. *For any step size $0 < \alpha \leq \frac{2}{\lambda+\mu}$, the population first-order EM operator $G : \Omega \to \Omega$ is contractive with parameter $\kappa(\alpha) = 1 - \alpha\xi$, where*

$$\text{(C.7)} \qquad \xi := \frac{2\mu\lambda}{\lambda+\mu} - \gamma.$$

We omit the proof, since it follows from a similar argument to that of Theorem 1. With this preliminary in place we can now begin the proof of Lemma 5.

C.4.1. *Proof of Lemma 5.* Let us write $\theta^{t+1} = \Pi(\widetilde{\theta}^{t+1})$, where $\tilde{\theta}^{t+1} := \theta^t + \alpha^t\nabla Q_1(\theta^t|\theta^t)$ is the update vector prior to projecting onto the ball $\mathbb{B}_2(\frac{r}{2};\theta^0)$. Defining the difference vectors $\Delta^{t+1} := \theta^{t+1} - \theta^*$ and $\widetilde{\Delta}^{t+1} := \widetilde{\theta}^{t+1} - \theta^*$, we have

$$\|\Delta^{t+1}\|_2^2 - \|\Delta^t\|_2^2 \leq \|\widetilde{\Delta}^{t+1}\|_2^2 - \|\Delta^t\|_2^2 = \langle\widetilde{\theta}^{t+1} - \theta^t,\, \widetilde{\theta}^{t+1} + \theta^t - 2\theta^*\rangle.$$

Introducing the shorthand $\widehat{W}(\theta) := \nabla Q_1(\theta|\theta)$, we have $\widetilde{\theta}^{t+1} - \theta^t = \alpha^t \widehat{W}(\theta)$, and hence

$$\begin{aligned}\|\Delta^{t+1}\|_2^2 - \|\Delta^t\|_2^2 &\leq \alpha^t \langle \widehat{W}(\theta^t),\, \alpha^t \widehat{W}(\theta^t) + 2(\theta^t - \theta^*)\rangle \\ &= (\alpha^t)^2 \|\widehat{W}(\theta^t)\|_2^2 + 2\alpha^t \langle \widehat{W}(\theta^t),\, \Delta^t\rangle.\end{aligned}$$

Letting $\mathcal{F}_t$ denote the $\sigma$-field of events up to the random variable $\theta^t$, note that

$$\mathbb{E}[\widehat{W}(\theta^t) \mid \mathcal{F}_t] = W(\theta^t) := \nabla Q(\theta^t|\theta^t).$$

Consequently, by iterated expectations, we have

$$\mathbb{E}[\|\Delta^{t+1}\|_2^2] \leq \mathbb{E}[\|\Delta^t\|_2^2] + (\alpha^t)^2 \mathbb{E}\|\widehat{W}(\theta^t)\|_2^2 + 2\alpha^t \mathbb{E}\Big[\langle W(\theta^t),\, \Delta^t\rangle\Big]. \tag{C.8}$$

Now since $\theta^*$ maximizes the function $q$ and $\theta^t$ belongs to $\mathbb{B}_2(\frac{r}{2}; \theta^0)$, we have

$$\langle W(\theta^*),\, \Delta^t\rangle = \langle \nabla q(\theta^*),\, \Delta^t\rangle \leq 0.$$

Combining with our earlier inequality (C.8) yields

$$\mathbb{E}[\|\Delta^{t+1}\|_2^2] \leq \mathbb{E}[\|\Delta^t\|_2^2] + (\alpha^t)^2 \mathbb{E}\|\widehat{W}(\theta^t)\|_2^2 + 2\alpha^t \mathbb{E}\Big[\langle W(\theta^t) - W(\theta^*),\, \Delta^t\rangle\Big].$$

Defining $G^t(\theta^t) := \theta^t + \alpha^t W(\theta^t)$, we see that

$$\begin{aligned}\alpha^t \langle W(\theta^t) - W(\theta^*),\, \Delta^t\rangle &= \langle G^t(\theta^t) - G^t(\theta^*) - (\theta^t - \theta^*),\, \theta^t - \theta^*\rangle \\ &= \langle G^t(\theta^t) - G^t(\theta^*),\, \theta^t - \theta^*\rangle - \|\theta^t - \theta^*\|_2^2 \\ &\overset{\text{(i)}}{\leq} (\kappa(\alpha^t) - 1)\|\theta^t - \theta^*\|_2^2 \\ &\overset{\text{(ii)}}{=} -\alpha^t \xi\, \|\Delta^t\|_2^2,\end{aligned}$$

where step (i) uses the contractivity of $G^t$ established in Lemma 6 and step (ii) uses the definition of $\xi$ from equation (C.7). Putting together the pieces yields the claim (C.5).

**C.5. Proof of Corollary 10.** We need to bound the uniform variance $\sigma_{\mathrm{G}}^2 = \sup_{\theta \in \mathbb{B}_2(r;\theta^*)} \mathbb{E}\|\nabla Q_1(\theta|\theta)\|_2^2$, where $r = \frac{\|\theta^*\|_2}{32}$. From the gradient update (A.6a), we have $\nabla Q_1(\theta \mid \theta) = (2w_\theta(x_1, y_1) - 1)y_1 x_1 - \langle x_1,\, \theta\rangle x_1$, and hence

$$\mathbb{E}\big[\|\nabla Q_1(\theta|\theta)\|_2^2\big] \leq 2\underbrace{\mathbb{E}[y_1^2 \|x_1\|_2^2]}_{T_1} + 2\underbrace{\|\mathbb{E}[x_1 x_1^T \|x_1\|_2^2]\|_{\mathrm{op}}}_{T_2} \|\theta\|_2^2. \tag{C.9}$$

First considering $T_1$, recall that $y_1 = z_1\langle x_1,\, \theta^*\rangle + v_1$, where $v \sim \mathcal{N}(0, \sigma^2)$ and $z_1$ is a random sign, independent of $(x_1, v_1)$. Consequently, we have

$$T_1 \leq 2\mathbb{E}[\langle x_1,\, \theta^*\rangle^2 \|x_1\|_2^2] + 2\mathbb{E}[v_1^2 \|x_1\|_2^2] \;\leq\; 2\sqrt{\mathbb{E}[\langle x_1,\, \theta^*\rangle^4]}\, \sqrt{\mathbb{E}[\|x_1\|_2^4]} + 2\sigma^2 d,$$

where we have applied the Cauchy-Schwarz inequality, and observed that $\mathbb{E}[\|x_1\|_2^2] = d$ and $\mathbb{E}[v_1^2] = \sigma^2$. Since the random variable $\langle x_1,\, \theta^*\rangle$ is sub-Gaussian with parameter at most $\|\theta^*\|_2$, we have $\mathbb{E}[\langle x_1,\, \theta^*\rangle^4] \leq 3\|\theta^*\|_2^4$. Moreover, since the random vector $x_1$ has i.i.d. components, we have

$$\mathbb{E}[\|x_1\|_2^4] = \sum_{j=1}^d \mathbb{E}[x_{1j}^4] + 2\sum_{i \neq j} \mathbb{E}[x_{1i}^2]\mathbb{E}[x_{1j}^2] = 3d + 2\binom{d}{2} \;\leq 4d^2.$$

Putting together the pieces, we conclude that $T_1 \leq 8\|\theta^*\|_2^2 d + 2\sigma^2 d$.

Turning to term $T_2$, by definition of the operator norm, there is a unit-norm vector $u \in \mathbb{R}^d$ such that

$$T_2 = |||\mathbb{E}[x_1 x_1^T \|x_1\|_2^2]|||_{\text{op}} = u^T \Big(\mathbb{E}[x_1 x_1^T \|x_1\|_2^2]\Big) u = \mathbb{E}[\langle x_1, u\rangle^2 \|x_1\|_2^2]$$
$$\overset{\text{(i)}}{\leq} \sqrt{\mathbb{E}[\langle x_1, u\rangle^4]}\, \sqrt{\mathbb{E}[\|x_1\|_2^4]}$$
$$\overset{\text{(ii)}}{\leq} \sqrt{3}\sqrt{4d^2} \leq 4d.$$

where step (i) applies the Cauchy-Schwarz inequality, and step (ii) uses the fact that $\langle x_1, u\rangle$ is sub-Gaussian with parameter 1, and our previous bound on $\mathbb{E}[\|x_1\|_2^4]$.

Putting together the pieces yields $\sigma_{\text{G}}^2 \leq c\,(\sigma^2 + \|\theta^*\|_2^2)d$, so that Corollary 10 follows as a consequence of Theorem 6.

**C.6. Proof of Corollary 11.** Once again we focus on bounding the uniform variance $\sigma_{\text{G}}^2$. From the form of $Q$ given in equation (A.8) (with $n = 1$), we have

$$\mathbb{E}\Big[\|\nabla Q_1(\theta|\theta)\|_2^2\Big] \leq 2\Big\{ \underbrace{\mathbb{E}\Big[\|\Sigma_\theta(x_{\text{obs}}, y)\theta\|_2^2\Big]}_{T_1} + \underbrace{\mathbb{E}[y^2\, \|\mu_\theta(x_{\text{obs}}, y)\|_2^2]}_{T_2} \Big\}. \tag{C.10}$$

We bound each of these terms in turn. To simplify notation, we omit the dependence of $\mu_\theta$ and $\Sigma_\theta$ on $(x_{\text{obs}}, y)$, but it should be implicitly understood.

*Bounding $T_1$:.* Letting $1 \in \mathbb{R}^d$ be the vector of all ones, and $z \in \mathbb{R}^d$ be an indicator of observed indices, we have $\Sigma_\theta = I_{\text{mis}} + \mu_\theta \mu_\theta^T - ((1-z)\odot\mu_\theta)((1-z)\odot\mu_\theta)^T$. Consequently,

$$\frac{1}{3}\mathbb{E}[\|\Sigma_\theta\theta\|_2^2] \leq \|\theta\|_2^2 + \mathbb{E}\big[\|\mu_\theta\|_2^2\, \langle\mu_\theta, \theta\rangle^2\big] + \mathbb{E}\big[\|(1-z)\odot\mu_\theta\|_2^2\, \langle(1-z)\odot\mu_\theta, \theta\rangle^2\big].$$

By the Cauchy-Schwarz inequality, we have

$$\mathbb{E}\big[\|\mu_\theta\|_2^2\, \langle\mu_\theta, \theta\rangle^2\big] \leq \sqrt{\mathbb{E}[\|\mu_\theta\|_2^4]}\, \sqrt{\mathbb{E}[\langle\mu_\theta, \theta\rangle^4]}.$$

From Lemma 4, the random vector $\mu_\theta$ is sub-Gaussian with constant parameter, so that $\mathbb{E}[\|\mu_\theta\|_2^4] \leq c\, d^2$. Since $\|\theta\|_2 \leq c\|\theta^*\|_2$, the random variable $\langle\mu_\theta, \theta\rangle$ is sub-Gaussian with parameter $c\|\theta^*\|_2$, and hence $\mathbb{E}[\langle\mu_\theta, \theta\rangle^4] \leq c\,\|\theta^*\|_2^4$. Putting together the pieces, we see that $\mathbb{E}\big[\|\mu_\theta\|_2^2\, \langle\mu_\theta, \theta\rangle^2\big] \leq c\, d\|\theta^*\|_2^2$. A similar argument applies to other expectation, so that we conclude that $T_1 = \mathbb{E}\big[\|\Sigma_\theta\theta\|_2^2\big] \leq c\,\|\theta^*\|_2^2 d$, a bound that holds uniformly for all $\theta \in \mathbb{B}_2(r; \theta^*)$.

*Bounding $T_2$:.* By the Cauchy-Schwarz inequality, we have

$$T_2 = \mathbb{E}[y^2 \|\mu_\theta(x_{\text{obs}}, y)\|_2^2] \leq \sqrt{\mathbb{E}[y^4]}\sqrt{\mathbb{E}[\|\mu_\theta(x_{\text{obs}}, y)\|_2^4]}.$$

Note that $y$ is sub-Gaussian with parameter at most $\sqrt{\|\theta^*\|_2^2 + \sigma^2}$, whence

$$\sqrt{\mathbb{E}[y^4]} \leq c\,(\|\theta^*\|_2^2 + \sigma^2).$$

Similarly, Lemma 4 implies that $\sqrt{\mathbb{E}[\|\mu_\theta(x_{\text{obs}}, y)\|_2^4]} \leq cd$, and hence $T_2 \leq c'\big(\|\theta^*\|_2^2 + \sigma^2\big)d$.

Substituting our upper bounds on $T_1$ and $T_2$ into the decomposition (C.10), we find that $\sigma_{\text{G}}^2 \leq c\,\big(\|\theta^*\|_2^2 + \sigma^2\big)d$. Thus, Corollary 11 follows from Theorem 6.

## APPENDIX D: TECHNICAL MATERIAL FOR GAUSSIAN MIXTURE MODELS

In this appendix, we provide proofs of technical results related to the mixture of Gaussians model.

**D.1. Some elementary properties.** We make frequent use of the following facts:

- For the function $f(t) = \frac{t^2}{\exp(\mu t)}$, we have

$$\sup_{t \in [0,\infty]} f(t) = \frac{4}{(e\,\mu)^2}, \qquad \text{achieved at } t^* = \tfrac{2}{\mu} \text{ and} \tag{D.1a}$$

$$\sup_{t \in [t^*,\infty]} f(t) = f(t^*), \qquad \text{for } t^* \geq \tfrac{2}{\mu}. \tag{D.1b}$$

- For the function $g(t) = \frac{1}{(\exp(t)+\exp(-t))^2}$, we have

$$g(t) \leq \frac{1}{4} \quad \text{for all } t \in \mathbb{R}, \text{ and} \tag{D.2a}$$

$$\sup_{t \in [\mu,\infty]} g(t) \leq \frac{1}{(\exp(\mu)+\exp(-\mu))^2} \leq \exp(-2\mu), \qquad \text{valid for any } \mu \geq 0. \tag{D.2b}$$

- Similarly, for the function $g^2(t) = \frac{1}{(\exp(t)+\exp(-t))^4}$, we have

$$g^2(t) \leq \frac{1}{16} \quad \text{for all } t \in \mathbb{R}, \text{ and} \tag{D.3a}$$

$$\sup_{t \in [\mu,\infty]} g^2(t) \leq \frac{1}{(\exp(\mu)+\exp(-\mu))^4} \leq \exp(-4\mu), \qquad \text{valid for any } \mu \geq 0. \tag{D.3b}$$

**D.2. Proof of Lemma 2.** With these preliminaries in place, we can now begin the proof. For each $u \in [0,1]$, define $\theta_u = \theta^* + u\Delta$, where $\Delta := \theta - \theta^*$. Taylor's theorem applied to the function $\theta \mapsto w_\theta(Y)$, followed by expectations, yields

$$\mathbb{E}\Big[Y\big(w_\theta(Y) - w_{\theta^*}(Y)\big)\Big] = 2\int_0^1 \mathbb{E}\Bigg[\underbrace{\frac{YY^T}{\sigma^2\big(\exp\big(-\frac{\langle\theta_u, Y\rangle}{\sigma^2}\big) + \exp\big(\frac{\langle\theta_u, Y\rangle}{\sigma^2}\big)\big)^2}}_{\Gamma_u(Y)}\Bigg]\Delta\, du.$$

For each choice of $u \in [0,1]$, the matrix-valued function $y \mapsto \Gamma_u(y)$ is symmetric—that is, $\Gamma_u(y) = \Gamma_u(-y)$. Since the distribution of $Y$ is symmetric around zero, we conclude that $\mathbb{E}[\Gamma_u(Y)] = \mathbb{E}[\Gamma_u(\widetilde{Y})]$, where $\widetilde{Y} \sim \mathcal{N}(\theta^*, \sigma^2 I)$, and hence that

$$\|\mathbb{E}\Big[\big(w_\theta(Y) - w_{\theta^*}(Y)\big)\,Y\Big]\|_2 \leq 2 \sup_{u\in[0,1]} |||\mathbb{E}[\Gamma_u(\widetilde{Y})]|||_{\mathrm{op}}\, \|\Delta\|_2. \tag{D.4}$$

The remainder of the proof is devoted to bounding $|||\mathbb{E}[\Gamma_u(\widetilde{Y})]|||_{\mathrm{op}}$ uniformly over $u \in [0,1]$. For an arbitrary fixed $u \in [0,1]$ let $R$ be an orthonormal matrix such that $R\theta_u = \|\theta_u\|_2 e_1$, where $e_1 \in \mathbb{R}^d$ denotes the first canonical basis vector. Define the rotated random vector $V = R\widetilde{Y}$, and note that $V \sim \mathcal{N}(R\theta^*, \sigma^2 I)$. Using this transformation, the operator norm of the matrix $\mathbb{E}[\Gamma_u(\widetilde{Y})]$ is equal to that of

$$D = \mathbb{E}\Big[\frac{VV^T}{\sigma^2\big(\exp\big(\frac{\langle V, \|\theta_u\|_2 e_1\rangle}{\sigma^2}\big) + \exp\big(-\frac{\langle V, \|\theta_u\|_2 e_1\rangle}{\sigma^2}\big)\big)^2}\Big].$$

In order to bound the operator norm of $D$, we need to introduce some intermediate quantities. We define

$$\alpha_1 := \mathbb{E}\left[\frac{V_1^2}{\sigma^2\big(\exp\big(\frac{\langle V, \|\theta_u\|_2 e_1\rangle}{\sigma^2}\big) + \exp\big(-\frac{\langle V, \|\theta_u\|_2 e_1\rangle}{\sigma^2}\big)\big)^2}\right],$$
$$\alpha_2 := \mathbb{E}\left[\frac{V_1}{\sigma^2\big(\exp\big(\frac{\langle V, \|\theta_u\|_2 e_1\rangle}{\sigma^2}\big) + \exp\big(-\frac{\langle V, \|\theta_u\|_2 e_1\rangle}{\sigma^2}\big)\big)^2}\right],$$
$$\alpha_3 := \mathbb{E}\left[\frac{1}{\sigma^2\big(\exp\big(\frac{\langle V, \|\theta_u\|_2 e_1\rangle}{\sigma^2}\big) + \exp\big(-\frac{\langle V, \|\theta_u\|_2 e_1\rangle}{\sigma^2}\big)\big)^2}\right].$$

Further denote,

$$\mu := R\theta^*,$$
$$\nu := [0, \mu_2, \mu_3, \ldots, \mu_d]^T.$$

In terms of these quantities observe that we can write,

$$D = \alpha_1 e_1 e_1^T + \alpha_2(\nu e_1^T + e_1 \nu^T) + \alpha_3 \nu\nu^T.$$

So we have that,

$$\text{(D.5)} \qquad \|\!|D|\!\|_{\text{op}} \le \|\!|D|\!\|_{\text{fro}} \le \alpha_1 + 2\alpha_2\|\nu\|_2 + \alpha_3\|\nu\|_2^2 \le \alpha_1 + 2\alpha_2\|\theta^*\|_2 + \alpha_3\|\theta^*\|_2^2.$$

In order to bound $\alpha_1, \alpha_2$ and $\alpha_3$ observe that,

$$\alpha_1 \le \mathbb{E}\Big[\frac{V_1^2/\sigma^2}{\exp\big(\frac{2\|\theta_u\|_2 V_1}{\sigma^2}\big)}\Big].$$

Defining the event $\mathcal{E} = \{V_1 \le \frac{\|\theta^*\|_2}{4}\}$, we condition on it and its complement to obtain

$$\alpha_1 \le \mathbb{E}\Big[\frac{V_1^2/\sigma^2}{\exp\big(\frac{2\|\theta_u\|_2 V_1}{\sigma^2}\big)} \mid \mathcal{E}\Big]\mathbb{P}[\mathcal{E}] + \mathbb{E}\Big[\frac{V_1^2/\sigma^2}{\exp\big(\frac{2\|\theta_u\|_2 V_1}{\sigma^2}\big)} \mid \mathcal{E}^c\Big].$$

Conditioned on $\mathcal{E}$ and $\mathcal{E}^c$, respectively, we then apply the bounds (D.1a) and (D.1b) to obtain

$$\alpha_1 \le \frac{\sigma^2}{e^2\|\theta_u\|_2^2}\mathbb{P}[\mathcal{E}] + \frac{\|\theta^*\|_2^2}{16\sigma^2\exp\big(\frac{\|\theta_u\|_2\|\theta^*\|_2}{2\sigma^2}\big)},$$

provided $\|\theta^*\|_2\|\theta_u\|_2 \ge 4\sigma^2$. Noting that

$$\text{(D.6)} \qquad \|\theta_u\|_2 = \|\theta^* + u(\theta - \theta^*)\|_2 \ge \|\theta^*\|_2 - \frac{1}{4}\|\theta^*\|_2 \ = \frac{3}{4}\|\theta^*\|_2,$$

we obtain the bound $\alpha_1 \le \frac{16\sigma^2}{9e^2\|\theta^*\|_2^2}\mathbb{P}(\mathcal{E}) + \frac{\|\theta^*\|_2^2\exp\big(-\frac{3\|\theta^*\|_2^2}{8\sigma^2}\big)}{16\sigma^2}$, whenever $\|\theta^*\|_2^2 \ge 16\sigma^2/3$.

Note that the mean of $V_1$ is lower bounded as

$$\mathbb{E}[V_1] = \langle R\theta^*, e_1\rangle = \langle R\theta_u, e_1\rangle + \langle R(\theta^* - \theta_u), e_1\rangle \ge \|\theta_u\|_2 - \|\theta^* - \theta_u\|_2 \overset{\text{(i)}}{\ge} \frac{\|\theta^*\|_2}{2},$$

where step (i) follows from the lower bound (D.6). Consequently, by standard Gaussian tail bounds, we have

$$\mathbb{P}[\mathcal{E}] \leq \exp\Big(\frac{-\|\theta^*\|_2^2}{32\sigma^2}\Big). \tag{D.7}$$

Combining the pieces yields

$$\alpha_1 \leq \frac{16\sigma^2}{9e^2\|\theta^*\|_2^2}\, e^{-\frac{\|\theta^*\|_2^2}{32\sigma^2}} + \frac{\|\theta^*\|_2^2}{16\sigma^2} e^{-\frac{3\|\theta^*\|_2^2}{8\sigma^2}} \qquad \text{whenever } \|\theta^*\|_2^2 \geq 16\sigma^2/3.$$

In a similar fashion we have that,

$$\begin{aligned}
\alpha_2 &= \mathbb{E}\Big[\frac{V_1}{\sigma^2\Big(\exp\Big(\frac{\|\theta_u\|_2 V_1}{\sigma^2}\Big) + \exp\Big(-\frac{\|\theta_u\|_2 V_1}{\sigma^2}\Big)\Big)^2}\Big] \\
&\leq \sqrt{\mathbb{E}\Big[\frac{V_1^2}{\sigma^2}\Big]}\sqrt{\mathbb{E}\Big[\frac{1}{\sigma^2\Big(\exp\Big(\frac{\|\theta_u\|_2 V_1}{\sigma^2}\Big) + \exp\Big(-\frac{\|\theta_u\|_2 V_1}{\sigma^2}\Big)\Big)^4}\Big]}.
\end{aligned}$$

Observe that, $\mathbb{E}\Big[\frac{V_1^2}{\sigma^2}\Big] \leq \frac{\|\theta^*\|_2^2}{\sigma^2}$, and it remains to bound the second term. We have that,

$$\mathbb{E}\Big[\frac{1}{\sigma^2\Big(\exp\Big(\frac{\|\theta_u\|_2 V_1}{\sigma^2}\Big) + \exp\Big(-\frac{\|\theta_u\|_2 V_1}{\sigma^2}\Big)\Big)^4}\Big] \;=\; \frac{1}{\sigma^2}\mathbb{E}\Big[g^2\Big(\frac{\|\theta_u\|_2 V_1}{\sigma^2}\Big)\Big],$$

where the reader should recall the function $g$ from equation (D.2a).

Conditioning on the event $\mathcal{E} = \{V_1 \leq \frac{\|\theta^*\|_2}{4}\}$ and its complement yields

$$\begin{aligned}
\frac{1}{\sigma^2}\mathbb{E}\Big[g^2\Big(\frac{\|\theta_u\|_2 V_1}{\sigma^2}\Big)\Big] &\leq \frac{1}{\sigma^2}\Big[\mathbb{E}\Big[g^2\Big(\frac{\|\theta_u\|_2 V_1}{\sigma^2}\Big) \mid \mathcal{E}\Big]\mathbb{P}[\mathcal{E}] + \mathbb{E}\Big[g^2\Big(\frac{\|\theta_u\|_2 V_1}{\sigma^2}\Big) \mid \mathcal{E}^c\Big]\Big] \\
&\overset{(i)}{\leq} \frac{1}{\sigma^2}\Big[\frac{1}{16}\mathbb{P}[\mathcal{E}] + \exp\Big(-\frac{\|\theta^*\|_2\|\theta_u\|_2}{\sigma^2}\Big)\Big] \\
&\overset{(ii)}{\leq} \frac{1}{\sigma^2}\Big[\frac{1}{16}\mathbb{P}[\mathcal{E}] + \exp\Big(-\frac{3\|\theta^*\|_2^2}{4\sigma^2}\Big)\Big],
\end{aligned}$$

where step (i) follows by applying bound (D.2a) to the first term, and the bound (D.2b) with $\mu = \frac{\|\theta^*\|_2\|\theta_u\|_2}{4\sigma^2}$ to the second term; and step (ii) follows from the bound (D.6). Applying the bound (D.7) on $\mathbb{P}[\mathcal{E}]$ yields

$$\frac{1}{\sigma^2}\mathbb{E}\Big[g^2\Big(\frac{\|\theta_u\|_2 V_1}{\sigma^2}\Big)\Big] \leq \frac{1}{\sigma^2}\Big[\frac{1}{16}\exp\big(-\frac{\|\theta^*\|_2^2}{32\sigma^2}\big) + \exp\big(-\frac{3\|\theta^*\|_2^2}{4\sigma^2}\big)\Big] \leq \frac{2}{\sigma^2}\exp\big(-\frac{\|\theta^*\|_2^2}{32\sigma^2}\big).$$

Putting this together we have that,

$$\alpha_2 \leq \frac{2\|\theta^*\|_2}{\sigma^2}\exp\Big(-\frac{\|\theta^*\|_2^2}{64\sigma^2}\Big).$$

In order to bound $\alpha_3$, we follow a similar argument. Once again, conditioning on the event $\mathcal{E} = \{V_1 \leq \frac{\|\theta^*\|_2}{4}\}$ and its complement yields

$$\begin{aligned}
\alpha_3 &\leq \frac{1}{\sigma^2}\Big[\mathbb{E}\Big[g\Big(\frac{\|\theta_u\|_2 V_1}{\sigma^2}\Big) \mid \mathcal{E}\Big]\mathbb{P}[\mathcal{E}] + \mathbb{E}\Big[g\Big(\frac{\|\theta_u\|_2 V_1}{\sigma^2}\Big) \mid \mathcal{E}^c\Big]\Big] \\
&\overset{(i)}{\leq} \frac{1}{\sigma^2}\Big[\frac{1}{4}\mathbb{P}[\mathcal{E}] + \exp\Big(-\frac{\|\theta^*\|_2\|\theta_u\|_2}{4\sigma^2}\Big)\Big] \\
&\overset{(ii)}{\leq} \frac{1}{\sigma^2}\Big[\frac{1}{4}\mathbb{P}[\mathcal{E}] + \exp\Big(-\frac{3\|\theta^*\|_2^2}{16\sigma^2}\Big)\Big],
\end{aligned}$$

where step (i) follows by applying bound (D.2a) to the first term, and the bound (D.2b) with $\mu = \frac{\|\theta^*\|_2\|\theta_u\|_2}{4\sigma^2}$ to the second term; and step (ii) follows from the bound (D.6). Applying the bound (D.7) on $\mathbb{P}[\mathcal{E}]$ yields

$$\alpha_3 \leq \frac{1}{\sigma^2}\Big[\frac{1}{4}\exp\big(-\frac{\|\theta^*\|_2^2}{32\sigma^2}\big) + \exp\big(-\frac{3\|\theta^*\|_2^2}{16\sigma^2}\big)\Big] \leq \frac{2}{\sigma^2}\exp\big(-\frac{\|\theta^*\|_2^2}{32\sigma^2}\big).$$

Returning to equations (D.5) and (D.4), we have shown that

$$\|2\mathbb{E}\Big[\big(w_\theta(Y) - w_{\theta^*}(Y)\big)\, Y\Big]\|_2 \leq c_1\big(1 + \frac{1}{\eta^2} + \eta^2\big)e^{-c_2\eta^2}\|\theta - \theta^*\|_2,$$

whenever $\frac{\|\theta^*\|_2^2}{\sigma^2} \geq \eta^2 \geq 16/3$. On this basis, the bound (6.4) holds as long as the signal-to-noise ratio is sufficiently large,

## APPENDIX E: TECHNICAL RESULTS FOR MIXTURE OF REGRESSIONS

In this appendix, we provide proofs of technical results related to the mixture of regressions model.

**E.1. Proof of Lemma 3.** Since the standard deviation $\sigma$ is known, a simple rescaling argument allows us to take $\sigma = 1$, and replace the weight function in (A.4a) with

$$w_\theta(x, y) = \frac{\exp\big(\frac{-(y-\langle x,\theta\rangle)^2}{2}\big)}{\exp\big(\frac{-(y-\langle x,\theta\rangle)^2}{2}\big) + \exp\big(\frac{-(y+\langle x,\theta\rangle)^2}{2}\big)}. \tag{E.1}$$

Our proof makes use of the following elementary result on Gaussian random vectors:

LEMMA 7. *Given a Gaussian random vector $X \sim \mathcal{N}(0, I)$ and any fixed vectors $u, v \in \mathbb{R}^d$, we have*

$$\mathbb{E}[\langle X, u\rangle^2\langle X, v\rangle^2] \leq 3\|u\|_2^2\|v\|_2^2 \quad \textit{with equality when } u = v, \textit{ and} \tag{E.2a}$$

$$\mathbb{E}[\langle X, u\rangle^4\langle X, v\rangle^2] \leq 15\|u\|_2^4\|v\|_2^2. \tag{E.2b}$$

PROOF. For any fixed orthonormal matrix $R \in \mathbb{R}^{d\times d}$, the transformed variable $R^T X$ also has a $\mathcal{N}(0, I)$ distribution, and hence $\mathbb{E}[\langle X, u\rangle^2\langle X, v\rangle^2] = \mathbb{E}[\langle X, Ru\rangle^2\langle X, Rv\rangle^2]$. Let us choose $R$ such that $Ru = \|u\|_2 e_1$. Introducing the shorthand $z = Rv$, we have

$$\begin{aligned}\mathbb{E}[\langle X, Ru\rangle^2\langle X, Rv\rangle^2] = \mathbb{E}[\|u\|_2^2 X_1^2 \sum_{i=1}^d\sum_{j=1}^d X_i X_j z_i z_j] &= \|u\|_2^2(3z_1^2 + (\|z\|_2^2 - z_1^2)) \\ &\leq 3\|u\|_2^2\|z\|_2^2 \;= 3\|u\|_2^2\|v\|_2^2.\end{aligned}$$

A similar argument yields the second claim. □

With these preliminaries in place, we can now begin the proof of Lemma 3. Recall that $\Delta = \theta - \theta^*$ and that $\widetilde{\Delta}$ is any fixed vector in $\mathbb{R}^d$. Define $\theta_u = \theta^* + u\Delta$ for a scalar $u \in [0, 1]$. Recall that by our assumptions guarantee that

$$\|\Delta\|_2 \leq \frac{\|\theta^*\|_2}{32}, \qquad \text{and} \qquad \|\theta^*\|_2 \geq \eta. \tag{E.3a}$$

For future reference, we observe that

$$\|\theta_u\|_2 \geq \|\theta^*\|_2 - \|\Delta\|_2 \geq \frac{\|\theta^*\|_2}{2}. \tag{E.3b}$$

Noting that Lemma 3 consists of two separate inequalities (6.6a) and (6.6b), we treat these cases separately.

**E.2. Proof of inequality (6.6a).** We split the proof of this bound into two separate cases: namely, $\|\Delta\|_2 \leq 1$ and $\|\Delta\|_2 > 1$.

*Case* $\|\Delta\|_2 \leq 1$*:.* We then have.

$$\frac{d}{d\theta} w_\theta(X, Y) = \frac{2YX}{\big(\exp\big(Y\langle X, \theta\rangle\big) + \exp\big(-Y\langle X, \theta\rangle\big)\big)^2}.$$

Thus, using a Taylor series with integral form remainder on the function $\theta \mapsto w_\theta(X, Y)$ yields

$$\Delta_w(X, Y) = \int_0^1 \frac{2Y\langle X, \Delta\rangle}{(\exp(Z_u) + \exp(-Z_u))^2} du, \tag{E.4}$$

where $Z_u := Y\langle X, \theta^* + u\Delta\rangle$. Substituting for $\Delta_w(X, Y)$ in inequality (6.6a), we see that it suffices to show

$$\int_0^1 \underbrace{\mathbb{E}\big[\frac{2Y\langle X, \theta^*\rangle}{(\exp(Z_u) + \exp(-Z_u))^2}(2Z - 1)\langle X, \Delta\rangle\langle X, \widetilde{\Delta}\rangle\big]}_{A_u} du \leq \frac{\gamma}{2}\|\Delta\|_2\|\widetilde{\Delta}\|_2. \tag{E.5}$$

for some $\gamma \in [0, 1/4)$. The following auxiliary result is central to establishing this claim:

LEMMA 8. *There is a* $\gamma \in [0, 1/4)$ *such that for each* $u \in [0, 1]$*, we have*

$$\sqrt{\mathbb{E}\big[\frac{Y^2\langle X, \theta_u\rangle^2}{(\exp(Z_u) + \exp(-Z_u))^4}\big]} \leq \frac{\gamma}{14}, \quad \textit{and} \tag{E.6a}$$

$$\sqrt{\mathbb{E}\big[\frac{Y^2}{(\exp(Z_u) + \exp(-Z_u))^4}\big]} \leq \frac{\gamma}{32} \qquad \textit{whenever } \|\Delta\|_2 \leq 1. \tag{E.6b}$$

See Section E.4 for the proof of this lemma.

Using Lemma 8, let us bound the quantity $A_u$ from equation (E.5). Since $\theta^* = \theta_u - u\Delta$, we have $A_u = B_1 + B_2$, where

$$B_1 := \mathbb{E}\big[\frac{2Y\langle X, \theta_u\rangle}{(\exp(Z_u) + \exp(-Z_u))^2}(2Z - 1)\langle X, \Delta\rangle\langle X, \widetilde{\Delta}\rangle\big], \quad \text{and}$$
$$B_2 := -\mathbb{E}\big[\frac{2Yu\langle X, \Delta\rangle}{(\exp(Z_u) + \exp(-Z_u))^2}(2Z - 1)\langle X, \Delta\rangle\langle X, \widetilde{\Delta}\rangle\big].$$

In order to show that $A_u \leq \frac{\gamma}{2}\|\Delta\|_2\,\|\widetilde{\Delta}\|_2$, it suffices to show that $\max\{B_1, B_2\} \leq \frac{\gamma}{4}\|\Delta\|_2\|\widetilde{\Delta}\|_2$.

*Bounding* $B_1$*:.* By the Cauchy-Schwarz inequality, we have

$$\begin{aligned} B_1 &\leq \sqrt{\mathbb{E}\big[\frac{y^2\langle X, \theta_u\rangle^2}{(\exp(Z_u) + \exp(-Z_u))^4}\big]}\sqrt{\mathbb{E}\big[4(2Z - 1)^2\langle X, \Delta\rangle^2\langle X, \widetilde{\Delta}\rangle^2\big]} \\ &\leq \frac{\gamma}{14}\sqrt{\mathbb{E}\big[4\langle X, \Delta\rangle^2\langle X, \widetilde{\Delta}\rangle^2\big]}, \end{aligned}$$

where the second step follows from the bound (E.6a), and the fact that $(2Z - 1)^2 = 1$. Next we observe that $\mathbb{E}\big[4\langle X, \Delta\rangle^2\langle X, \widetilde{\Delta}\rangle^2\big] \leq 12\|\Delta\|_2^2\|\widetilde{\Delta}\|_2^2$, where we have used the bound (E.2a) from Lemma 7. Combined with our earlier bound, we conclude that $B_1 \leq \frac{\gamma}{4}\|\Delta\|_2\,\|\widetilde{\Delta}\|_2$, as claimed.

*Bounding $B_2$:.* Similarly, another application of the Cauchy-Schwarz inequality yields

$$B_2 \leq \sqrt{\mathbb{E}\big[\frac{y^2}{(\exp(Z_u)+\exp(-Z_u))^4}\big]}\sqrt{\mathbb{E}\big[4u^2(2Z-1)^2\langle X,\,\Delta\rangle^4\langle X,\,\widetilde{\Delta}\rangle^2\big]}$$
$$\leq \frac{\gamma}{32}\sqrt{\mathbb{E}\big[4u^2\langle X,\,\Delta\rangle^4\langle X,\,\widetilde{\Delta}\rangle^2\big]},$$

where the second step follows from the bound (E.6b), and the fact that $(2Z-1)^2 = 1$. In this case, we have

$$\mathbb{E}\big[4u^2\langle X,\,\Delta\rangle^4\langle X,\,\widetilde{\Delta}\rangle^2\big] \overset{\text{(i)}}{\leq} 60\|\Delta\|_2^4\|\widetilde{\Delta}\|_2^2 \overset{\text{(ii)}}{\leq} 60\|\Delta\|_2^2\|\widetilde{\Delta}\|_2^2,$$

where step (i) uses the bound (E.2b) from Lemma 7, and step (ii) that $\|\Delta\|_2 \leq 1$. Combining the pieces, we conclude that $B_2 \leq \frac{\gamma}{4}\|\Delta\|_2\|\widetilde{\Delta}\|_2$, which completes the proof of inequality (6.6a) in the case $\|\Delta\|_2 \leq 1$.

*Case $\|\Delta\|_2 > 1$:.* We now turn to the second case of the bound (6.6a). Our argument (here and in later sections) makes use of various probability bounds on different events, which we state here for future reference. These events involve the scalar $\tau := C_\tau\sqrt{\log\|\theta^*\|_2}$ for a constant $C_\tau$, as well as the vectors

$$\Delta := \theta - \theta^*, \text{ and } \theta_u := \theta^* + u\,\Delta \quad \text{for some fixed } u \in [0,1].$$

LEMMA 9 (Event bounds).

*(i) For the event $\mathcal{E}_1 := \big\{\operatorname{sign}(\langle X,\,\theta^*\rangle) = \operatorname{sign}(\langle X,\,\theta_u\rangle)\big\}$, we have $\mathbb{P}[\mathcal{E}_1^c] \leq \frac{\|\Delta\|_2}{\|\theta^*\|_2}$.*

*(ii) For the event $\mathcal{E}_2 := \big\{|\langle X,\,\theta^*\rangle| > \tau\big\} \cap \big\{|\langle X,\,\theta_u\rangle| > \tau\big\} \cap \big\{|v| \leq \frac{\tau}{2}\big\}$, we have*

$$\mathbb{P}[\mathcal{E}_2^c] \leq \frac{\tau}{\|\theta^*\|_2} + \frac{\tau}{\|\theta_u\|_2} + 2\exp\Big(-\frac{\tau^2}{2}\Big).$$

*(iii) For the event $\mathcal{E}_3 := \big\{|\langle X,\,\theta^*\rangle| \geq \tau\big\} \bigcup \big\{|\langle X,\,\theta_u\rangle| \geq \tau\big\}$, we have $\mathbb{P}\big[\mathcal{E}_3^c\big] \leq \frac{\tau}{\|\theta^*\|_2} + \frac{\tau}{\|\theta_u\|_2}$.*

*(iv) For the event $\mathcal{E}_4 := \big\{|v| \leq \tau/2\big\}$, we have $\mathbb{P}[\mathcal{E}_4^c] \leq 2e^{-\frac{\tau^2}{2}}$.*

*(v) For the event $\mathcal{E}_5 := \big\{|\langle X,\,\theta_u\rangle| > \tau\big\}$, we have $\mathbb{P}[\mathcal{E}_5^c] \leq \frac{\tau}{\|\theta_u\|_2}$.*

*(vi) For the event $\mathcal{E}_6 := \big\{|\langle X,\,\theta^*\rangle| > \tau\big\}$, we have $\mathbb{P}[\mathcal{E}_6^c] \leq \frac{\tau}{\|\theta^*\|_2}$.*

Various stages of our proof involve controlling the second moment matrix $\mathbb{E}[XX^T]$ when conditioned on some of the events given above:

LEMMA 10 (Conditional covariance bounds). *Conditioned on any event $\mathcal{E} \in \{\mathcal{E}_1 \cap \mathcal{E}_2, \mathcal{E}_1^c, \mathcal{E}_5^c, \mathcal{E}_6^c\}$, we have $\|\!|\mathbb{E}[XX^T \mid \mathcal{E}]|\!\|_{op} \leq 2$.*

See Section E.4.3 for the proof of this result.

With this set-up, our goal is to bound the quantity

$$T = \big|\mathbb{E}\big[\Delta_w(X,Y)(2Z-1)\langle X,\,\theta^*\rangle\langle X,\,\widetilde{\Delta}\rangle\big]\big| \leq \mathbb{E}\big[\big|\Delta_w(X,Y)(2Z-1)\langle X,\,\theta^*\rangle\langle X,\,\widetilde{\Delta}\rangle\big|\big].$$

For any measurable event $\mathcal{E}$, we define $\Psi(\mathcal{E}) := \mathbb{E}\big[\big|\Delta_w(X,Y)(2Z-1)\langle X,\,\theta^*\rangle\langle X,\,\widetilde{\Delta}\rangle\big| \mid \mathcal{E}\big]\,\mathbb{P}[\mathcal{E}]$, and note that by successive conditioning, we have

$$T \leq \Psi(\mathcal{E}_1 \cap \mathcal{E}_2) + \Psi(\mathcal{E}_1^c) + \Psi(\mathcal{E}_4^c) + \Psi(\mathcal{E}_5^c) + \Psi(\mathcal{E}_6^c). \tag{E.7}$$

We bound each of these five terms in turn.

*Bounding* $\Psi(\mathcal{E}_1 \cap \mathcal{E}_2)$*:* . Applying the Cauchy-Schwarz inequality and using the fact that $(2Z-1)^2 = 1$ yields

$$\Psi(\mathcal{E}_1 \cap \mathcal{E}_2) \leq \sqrt{\mathbb{E}\big[\Delta_w(X,Y)^2 \langle X, \widetilde{\Delta}\rangle^2 | \mathcal{E}_1 \cap \mathcal{E}_2\big]} \sqrt{\mathbb{E}\big[\langle X, \theta^*\rangle^2 | \mathcal{E}_1 \cap \mathcal{E}_2\big]}. \tag{E.8}$$

We now bound $\Delta_w(X,Y)$ conditioned on the event $\mathcal{E}_1 \cap \mathcal{E}_2$. Since $\operatorname{sign}(\langle X, \theta^*\rangle) = \operatorname{sign}(\langle X, \theta_u\rangle)$ on the event $\mathcal{E}_1$, we have

$$\operatorname{sign}(Y\langle X, \theta^*\rangle) = \operatorname{sign}(Y\langle X, \theta_u\rangle). \tag{E.9a}$$

Conditioned on the event $\mathcal{E}_2$, observe that $|Y| = |(2Z-1)\langle X, \theta^*\rangle + v| \geq |\langle X, \theta^*\rangle| - |v| \geq \frac{\tau}{2}$, which implies that

$$\min\big\{|Y\langle X, \theta^*\rangle|, |Y\langle X, \theta\rangle|\big\} \geq \frac{\tau^2}{2}. \tag{E.9b}$$

Recalling the weight function (E.1), we claim that when conditions (E.9a) and (E.9b) hold, then

$$|\Delta_w(X,Y)| = |w_{\theta_u}(X,Y) - w_{\theta^*}(X,Y)| \overset{\text{(i)}}{\leq} \frac{\exp(-\tau^2/2)}{\exp(-\tau^2/2) + \exp(\tau^2/2)} \leq \exp(-\tau^2). \tag{E.10}$$

We need to verify inequality (i): suppose first that $\operatorname{sign}(Y\langle X, \theta^*\rangle) = 1$. In this case, both $w_{\theta_u}(X,Y)$ and $w_{\theta^*}(X,Y)$ are at least $\frac{\exp(\tau^2/2)}{\exp(-\tau^2/2)+\exp(\tau^2/2)}$. Since each of these terms are upper bounded by 1, we obtain the claimed bound on $\Delta_w(X,Y)$. The case when $\operatorname{sign}(Y\langle X, \theta^*\rangle) = -1$ follows analogously.

Combined with our earlier bound (E.8), we have shown

$$\Psi(\mathcal{E}_1 \cap \mathcal{E}_2) \leq \exp(-\tau^2) \sqrt{\mathbb{E}\big[\langle X, \widetilde{\Delta}\rangle^2 | \mathcal{E}_1 \cap \mathcal{E}_2\big]} \sqrt{\mathbb{E}\big[\langle X, \theta^*\rangle^2 | \mathcal{E}_1 \cap \mathcal{E}_2\big]}.$$

Applying Lemma 10 with $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2$ yields $\Psi(\mathcal{E}_1 \cap \mathcal{E}_2) \leq 2\|\widetilde{\Delta}\|_2 \|\theta^*\|_2 e^{-\tau^2}$.

*Bounding* $\Psi(\mathcal{E}_1^c)$*:*. Combining the Cauchy-Schwarz inequality with Lemma 9(i), we have

$$\Psi(\mathcal{E}_1^c) \leq \sqrt{\mathbb{E}\big[\langle X, \widetilde{\Delta}\rangle^2 | \mathcal{E}_1^c\big]} \sqrt{\mathbb{E}\big[\langle X, \theta^*\rangle^2 | \mathcal{E}_1^c\big]} \frac{\|\Delta\|_2}{\|\theta^*\|_2}. \tag{E.11}$$

We first claim that $\mathbb{E}\big[\langle X, \theta^*\rangle^2 \mid \mathcal{E}_1^c\big] \leq \mathbb{E}\big[\langle X, \Delta\rangle^2 \mid \mathcal{E}_1^c\big]$. To establish this bound, it suffices to show that conditioned on $\mathcal{E}_1$, we have $\langle X, \theta^*\rangle^2 \leq \langle X, \Delta\rangle^2$. Note that event $\mathcal{E}_1$ implies that $\langle X, \theta^*\rangle \langle X, \theta_u\rangle \leq 0$. Consequently, conditioned on event $\mathcal{E}_1$, we have

$$\begin{aligned}
\langle X, \theta^*\rangle^2 = \frac{1}{4}\langle X, (\theta^* - \theta_u) + (\theta_u + \theta^*)\rangle^2 &\leq \frac{1}{2}\langle X, \theta^* - \theta_u\rangle^2 + \frac{1}{2}\langle X, \theta_u + \theta^*\rangle^2 \\
&\overset{(i)}{\leq} \langle X, \theta^* - \theta_u\rangle^2 \\
&\overset{(ii)}{\leq} \langle X, \Delta\rangle^2
\end{aligned}$$

where step (i) makes use of the bound $\langle X, \theta^*\rangle \langle X, \theta_u\rangle \leq 0$; and step (ii) follows since $\theta_u = \theta^* + u\Delta$, and $u \in [0,1]$.

Returning to equation (E.11), we have

$$\Psi(\mathcal{E}_1^c) \leq \sqrt{\mathbb{E}\big[\langle X, \widetilde{\Delta}\rangle^2 | \mathcal{E}_1^c\big]} \sqrt{\mathbb{E}\big[\langle X, \Delta\rangle^2 | \mathcal{E}_1^c\big]} \frac{\|\Delta\|_2}{\|\theta^*\|_2} \overset{\text{(i)}}{\leq} \frac{2\|\widetilde{\Delta}\|_2 \|\Delta\|_2^2}{\|\theta^*\|_2}$$

where step (i) follows from the conditional covariance bound of Lemma 10.

*Bounding* $\Psi(\mathcal{E}_4^c)$*:.* Combining the Cauchy-Schwarz inequality with Lemma 9(iv) yields

$$\Psi(\mathcal{E}_4^c) \leq 2\sqrt{\mathbb{E}\big[\langle X, \widetilde{\Delta}\rangle^2 | \mathcal{E}_4^c\big]}\sqrt{\mathbb{E}\big[\langle X, \theta^*\rangle^2 | \mathcal{E}_4^c\big]}\, e^{-\frac{\tau^2}{2}}.$$

Observe that by the independence of $v$ and $X$, conditioning on $\mathcal{E}_4^c$ has no effect on the second moment of $X$. Since $\mathbb{E}[XX^T] = I$, we conclude that $\Psi(\mathcal{E}_4^c) \leq 2\|\widetilde{\Delta}\|_2\|\theta^*\|_2\, e^{-\frac{\tau^2}{2}}$.

*Bounding* $\Psi(\mathcal{E}_5^c)$*: .* Combining the Cauchy-Schwarz inequality with Lemma 9(v) yields $\Psi(\mathcal{E}_5^c) \leq \frac{\tau}{\|\theta_u\|_2}\sqrt{\mathbb{E}\big[\langle X, \widetilde{\Delta}\rangle^2 | \mathcal{E}_5^c\big]}\sqrt{\mathbb{E}\big[\langle X, \theta^*\rangle^2 | \mathcal{E}_5^c\big]}$. Conditioned on the event $\mathcal{E}_5^c$, we have

$$\langle X, \theta^*\rangle^2 \leq 2\langle X, \theta_u\rangle^2 + 2\langle X, \Delta\rangle^2 \leq 2\tau^2 + 2\langle X, \Delta\rangle^2.$$

Together with Lemma 10, we obtain the bound

$$\Psi(\mathcal{E}_5^c) \leq \frac{2\tau\|\widetilde{\Delta}\|_2\sqrt{\tau^2 + 2\|\Delta\|_2^2}}{\|\theta_u\|_2} \overset{\text{(i)}}{\leq} \frac{2\tau\|\widetilde{\Delta}\|_2\|\Delta\|_2\sqrt{\tau^2+2}}{\|\theta_u\|_2},$$

where step (i) uses the fact that $\|\Delta\|_2 \geq 1$.

*Bounding* $\Psi(\mathcal{E}_6^c)$*: .* Combining the Cauchy-Schwarz inequality with Lemma 9(vi) yields $\Psi(\mathcal{E}_6^c) \leq \frac{\tau}{\|\theta^*\|_2}\sqrt{\mathbb{E}\big[\langle X, \widetilde{\Delta}\rangle^2 | \mathcal{E}_6^c\big]}\sqrt{\mathbb{E}\big[\langle X, \theta^*\rangle^2 | \mathcal{E}_6^c\big]}$. Conditioned on the event $\mathcal{E}_6^c$, we have $\langle X, \theta^*\rangle^2 \leq \tau^2$, and so applying Lemma 10 with $\mathcal{E} = \mathcal{E}_6^c$ yields $\Psi(\mathcal{E}_6^c) \leq \frac{\sqrt{2}\tau^2\|\widetilde{\Delta}\|_2}{\|\theta^*\|_2}$.

We have thus obtained bounds on all five terms in the decomposition (E.7). We combine these bounds with the with lower bound $\|\theta_u\|_2 \geq \frac{\|\theta^*\|_2}{2}$ from equation (E.3b), and then perform some algebra to obtain

$$T \leq c\,\|\Delta\|_2\|\widetilde{\Delta}\|_2\Big\{\frac{\tau^2}{\|\theta^*\|_2} + \|\theta^*\|_2 e^{-\tau^2/2}\Big\} + 2\|\widetilde{\Delta}\|_2\frac{\|\Delta\|_2^2}{\|\theta^*\|_2},$$

where $c$ is a universal constant. In particular, selecting $\tau = c_\tau\sqrt{\log\|\theta^*\|_2}$ for a sufficient large constant $c_\tau$, selecting the constant $\eta$ in (E.3a) sufficiently large yields the claim (6.6a).

**E.3. Proof of inequality (6.6b).** As in Section E.2, we treat the cases $\|\Delta\|_2 \leq 1$ and $\|\Delta\|_2 \geq 1$ separately.

E.3.1. *Case* $\|\Delta\|_2 \leq 1$*:.* As before, by a Taylor expansion of the function $\theta \mapsto \Delta_w(X, Y)$, it suffices to show that

$$\int_0^1 \mathbb{E}\big[\frac{2Yv}{(\exp(Z_u) + \exp(-Z_u))^2}\langle X, \Delta\rangle\langle X, \widetilde{\Delta}\rangle\big]du \leq \frac{\gamma}{2}\|\Delta\|_2\|\widetilde{\Delta}\|_2.$$

For any fixed $u \in [0, 1]$, the Cauchy-Schwarz inequality implies that

$$\begin{aligned}
\mathbb{E}\big[\frac{2Yv\,\langle X, \Delta\rangle\langle X, \widetilde{\Delta}\rangle}{(\exp(Z_u) + \exp(-Z_u))^2}\big] &\leq \sqrt{\mathbb{E}\big[\frac{4Y^2}{(\exp(Z_u) + \exp(-Z_u))^4}\big]}\,\sqrt{\mathbb{E}\big[v^2\langle X, \Delta\rangle^2\langle X, \widetilde{\Delta}\rangle^2\big]} \\
&\overset{\text{(i)}}{\leq} \sqrt{\mathbb{E}\big[\frac{4Y^2}{(\exp(Z_u) + \exp(-Z_u))^4}\big]}\,\sqrt{3\,\|\Delta\|_2^2\,\|\widetilde{\Delta}\|_2^2} \\
&\overset{\text{(ii)}}{\leq} \frac{\sqrt{3}\gamma}{16}\|\Delta\|_2\,\|\widetilde{\Delta}\|_2,
\end{aligned}$$

where step (i) follows from inequality (E.2a) in Lemma 7, the independence of $v$ and $X$, and the fact that $\mathbb{E}[v^2] = 1$; and step (ii) follows from the bound (E.6b) in Lemma 8.

E.3.2. *Case* $\|\Delta\|_2 > 1$*:.* After applying the Cauchy-Schwarz inequality, it suffices show that $\sqrt{\mathbb{E}\big[\Delta_w^2(X,Y)\big]} \leq \frac{\gamma}{2}$. The remainder of this section is devoted to the proof of this claim.

Recall the scalar $\tau := C_\tau \sqrt{\log \|\theta^*\|_2}$, as well as the events $\mathcal{E}_1$ and $\mathcal{E}_2$ from Lemma 9. For any measurable event $\mathcal{E}$, define the function $\Psi(\mathcal{E}) = \mathbb{E}\big[\Delta_w^2(X,Y) \mid \mathcal{E}\big]\mathbb{P}[\mathcal{E}]$. With this notation, by successive conditioning, we have the upper bound

$$\mathbb{E}\big[\Delta_w^2(X,Y)\big] \leq \Psi(\mathcal{E}_1^c) + \Psi(\mathcal{E}_1 \cap \mathcal{E}_2^c) + \Psi(\mathcal{E}_1 \cap \mathcal{E}_2). \tag{E.12}$$

We control each of these terms in turn.

*Controlling term* $\Psi(\mathcal{E}_1^c)$*:.* Noting that $\sup_{x,y} |\Delta_w(x,y)| \leq 2$ and applying Lemma 9(i), we have $\Psi(\mathcal{E}_1^c) \leq 4\mathbb{P}[\mathcal{E}_1^c] \leq 4\frac{\|\Delta\|_2}{\|\theta^*\|_2}$.

*Controlling term* $\Psi(\mathcal{E}_1 \cap \mathcal{E}_2^c)$*:.* Similarly, Lemma 9(ii) implies that

$$\Psi(\mathcal{E}_1 \cap \mathcal{E}_2^c) \leq 4\mathbb{P}[\mathcal{E}_2^c] \leq 4\Big\{\frac{\tau}{\|\theta^*\|_2} + \frac{\tau}{\|\theta_u\|_2} + 2e^{-\frac{\tau^2}{2}}\Big\}.$$

*Controlling term* $\Psi(\mathcal{E}_1 \cap \mathcal{E}_2)$*:.* Conditioned on the event $\mathcal{E}_1 \cap \mathcal{E}_2$, the bound (E.10) implies that $|\Delta_w(X,Y)| \leq \exp(-\tau^2)$, and hence $\Psi(\mathcal{E}_1 \cap \mathcal{E}_2) \leq e^{-2\tau^2}$.

Thus, we have derived bounds on each of the three terms in the decomposition (E.12): putting them together yields

$$\sqrt{\mathbb{E}\big[\Delta_w^2(X,Y)\big]} \leq \sqrt{4\frac{\|\Delta\|_2}{\|\theta^*\|_2} + 4\Big\{\frac{\tau}{\|\theta^*\|_2} + \frac{\tau}{\|\theta_u\|_2} + 2e^{-\frac{\tau^2}{2}}\Big\} + e^{-2\tau^2}}$$

By choosing $C_\tau$ sufficiently large in the definition of $\tau$, selecting the signal-to-noise constant $\eta$ in condition (E.3a) sufficiently large, the claim follows.

**E.4. Proof of Lemma 8.** The lemma statement consists of two inequalities, and we divide our proof accordingly.

E.4.1. *Proof of inequality* (E.6a). For any measurable event $\mathcal{E}$, let us introduce the function $\Psi(\mathcal{E}) := \mathbb{E}\Big[\frac{Y^2\langle X, \theta_u\rangle^2}{(\exp(Z_u)+\exp(-Z_u))^4} \mid \mathcal{E}\Big]\mathbb{P}[\mathcal{E}]$. With this notation, successive conditioning yields the decomposition

$$\mathbb{E}\big[\frac{Y^2\langle X,\, \theta_u\rangle^2}{(\exp(Z_u) + \exp(-Z_u))^4}\big] = \Psi(\mathcal{E}_4^c) + \Psi(\mathcal{E}_4 \cap \mathcal{E}_3^c) + \Psi(\mathcal{E}_2), \tag{E.13}$$

and we bound each of these terms in turn. The reader should recall the constant $\tau := C_\tau \sqrt{\log \|\theta^*\|_2}$, as well as the events $\mathcal{E}_3$ and $\mathcal{E}_4$ from Lemma 9.

*Bounding* $\Psi(\mathcal{E}_4^c)$*:.* Observe that

$$\frac{Y^2\langle X,\, \theta_u\rangle^2}{(\exp(Z_u) + \exp(-Z_u))^4} \leq \sup_{t \geq 0} \frac{t^2}{\exp(4t)} \leq \frac{1}{4e^2}, \tag{E.14}$$

where the final step follows from inequality (D.1a). Combined with Lemma 9(iv), we conclude that $\Psi(\mathcal{E}_4^c) \leq \frac{1}{2e^2} e^{-\frac{\tau^2}{2}}$.

*Bounding* $\Psi(\mathcal{E}_4 \cap \mathcal{E}_3^c)$*:.* In this case, we have

$$\Psi(\mathcal{E}_4 \cap \mathcal{E}_3^c) \stackrel{(i)}{\leq} \frac{1}{4e^2}\mathbb{P}[\mathcal{E}_3^c] \stackrel{(ii)}{\leq} \frac{1}{4e^2}\Big\{\frac{\tau}{\|\theta^*\|_2} + \frac{\tau}{\|\theta_u\|_2}\Big\},$$

where step (i) follows from inequality (E.14), and step (ii) follows from Lemma 9(iii).

*Bounding* $\Psi(\mathcal{E}_2)$*: .* Conditioned on the event $\mathcal{E}_2$, we have $Y^2\langle X, \theta_u\rangle^2 \geq \frac{\tau^2}{2}$, where we have used the lower bound (E.9b). Introducing the shorthand $t^* = \tau^2/2$, this lower bound implies that

$$\Psi(\mathcal{E}_2) \leq \sup_{t \geq t^*} \frac{t^2}{e^{4t}} \;\leq\; \frac{(t^*)^2}{e^{4t^*}} \;=\; \frac{\tau^4}{4e^{2\tau^2}},$$

where inequality (i) is valid as long as $t^* = \frac{\tau^2}{2} \geq \frac{1}{2}$, or equivalently $\tau^2 \geq 1$.

Substituting our upper bounds on three components in the decomposition (E.13) yields

$$\mathbb{E}\Big[\frac{Y^2\langle X, \theta_u\rangle^2}{(\exp(Z_u) + \exp(-Z_u))^4}\Big] \leq \frac{1}{2e^2}\, e^{-\frac{\tau^2}{2}} + \frac{1}{4e^2}\Big(\frac{\tau}{\|\theta^*\|_2} + \frac{\tau}{\|\theta_u\|_2}\Big) + \frac{\tau^4}{4} e^{-2\tau^2}.$$

Setting $C_\tau$ sufficiently large in the definition of $\tau$ and choosing sufficiently large values of the signal-to-noise constant $\eta$ in the condition (E.3a) yields the claim.

*Proof of inequality* (E.6b)*:.* For any measurable event $\mathcal{E}$, let us introduce the function

$$\Psi(\mathcal{E}) = \mathbb{E}\Big[\frac{Y^2}{(\exp(Z_u) + \exp(-Z_u))^4} \mid \mathcal{E}\Big]\mathbb{P}[\mathcal{E}].$$

Recalling the event $\mathcal{E}_5$ from Lemma 9, successive conditioning yields the decomposition

$$\mathbb{E}\Big[\frac{Y^2}{(\exp(Z_u) + \exp(-Z_u))^4}\Big] = \Psi(\mathcal{E}_5^c) + \Psi(\mathcal{E}_5). \tag{E.15}$$

We bound each of these terms in turn.

*Bounding* $\Psi(\mathcal{E}_5^c)$*: .* Simple algebra combined with Lemma 9(v) yields the upper bound $\Psi(\mathcal{E}_5^c) \leq \frac{\tau}{16\|\theta_u\|_2}\mathbb{E}[Y^2]$. Conditioned on $\mathcal{E}_5$, we have the upper bound $|\langle X, \theta_u\rangle| \leq \tau$, whence

$$\langle X, \theta^*\rangle^2 \leq 2\tau^2 + 2\langle X, \Delta\rangle^2.$$

Combining Lemma 10 with the bound $\|\Delta\|_2 \leq 1$, we find that $\langle X, \theta^*\rangle^2 \leq 2\tau^2 + 4$. Since $Y \stackrel{d}{=} (2Z-1)\langle X, \theta^*\rangle + v$, we have

$$\mathbb{E}[Y^2 \mid \mathcal{E}_5^c] \leq \mathbb{E}[2\langle X, \theta^*\rangle^2 + 2v^2 \mid \mathcal{E}_5^c] \stackrel{(i)}{\leq} 4\tau^2 + 10.$$

Putting together the pieces, we conclude that $\Psi(\mathcal{E}_5^c) \leq \frac{4\tau^3 + 10\tau}{16\|\theta_u\|_2}$.

*Bounding* $\Psi(\mathcal{E}_5)$*:.* Recall that $Z_u = Y\langle X, \theta_u\rangle$, so we have that

$$\Psi(\mathcal{E}_5) \leq \mathbb{E}\Big[\frac{Y^2}{(e^{Y\langle X, \theta_u\rangle} + e^{-Y\langle X, \theta_u\rangle})^4} \mid \mathcal{E}_5\Big] \stackrel{(i)}{\leq} \frac{4}{(e\,\tau)^2},$$

where step (i) follows from the bound (D.1a) and the observation that $|\langle X, \theta_u \rangle| \geq \tau$ conditioned on the event $\mathcal{E}_5$.

Substituting our bounds on the two terms into the decomposition (E.13) yields

$$\mathbb{E}\big[\frac{Y^2}{(e^{Z_u} + e^{-Z_u})^4}\big] \leq \frac{4\tau^3 + 10\tau}{16\|\theta_u\|_2} + \frac{4}{(e\,\tau)^2} \;\leq\; \frac{8\tau^3 + 20\tau}{16\|\theta^*\|_2} + \frac{4}{(e\,\tau)^2}.$$

Once again, sufficiently large choices of the constant $c_\tau$ and the signal-to-noise constant $\eta$ in equation (E.3a) yields the claim.

E.4.2. *Proof of Lemma 9.* In this section, we prove the probability bounds on events $\mathcal{E}_1$ through $\mathcal{E}_6$ stated in Lemma 9. In doing so, we make use of the following auxiliary result, due to Yi et al. [12] (see Lemma 1 in their paper):

LEMMA 11. *Given vectors $v, z \in \mathbb{R}^d$ and a Gaussian random vector $X \sim \mathcal{N}(0, I)$, the matrix $\Sigma = \mathbb{E}\big[XX^T \mid \langle X, v\rangle^2 > \langle X, z\rangle^2\big]$ has singular values*

$$\Big(1 + \frac{\sin\alpha}{\alpha}, 1 - \frac{\sin\alpha}{\alpha}, 1, \ldots, 1\Big), \qquad \text{where } \alpha = \cos^{-1}\tfrac{\langle z-v, z+v\rangle}{\|z+v\|_2\|z-v\|_2}. \tag{E.16a}$$

*Moreover, whenever $\|v\|_2 \leq \|z\|_2$, we have*

$$\mathbb{P}\big[\langle X, v\rangle^2 > \langle X, z\rangle^2\big] \leq \frac{\|v\|_2}{\|z\|_2}. \tag{E.16b}$$

*Proof of Lemma 9(i).* Note that the event $\mathcal{E}_1^c$ holds if and only if $\langle X, \theta^*\rangle\langle X, \theta_u\rangle < 0$, or equivalently, if and only if

$$4\langle X, \theta^*\rangle\langle X, \theta_u\rangle = \langle X, \theta^* + \theta_u\rangle^2 - \langle X, \theta^* - \theta_u\rangle^2 < 0.$$

Now observe that

$$\|\theta^* - \theta_u\|_2 \leq u\|\Delta\|_2 \leq \|\Delta\|_2, \quad \text{and} \quad \|\theta^* + \theta_u\|_2 \quad \geq 2\|\theta^*\|_2 - \|\Delta\|_2 \geq \|\theta^*\|_2 \geq \|\Delta\|_2.$$

Consequently, we may apply the bound (E.16b) from Lemma 11 with $v = \theta^* + \theta_u$ and $z = \theta^* - \theta_u$ to obtain $\mathbb{P}[\mathcal{E}_1^c] \leq \frac{\|\theta^* - \theta_u\|_2}{\|\theta^* + \theta_u\|_2} \leq \frac{\|\Delta\|_2}{\|\theta^*\|_2}$, as claimed.

*Proof of Lemma 9(iv):.* For $X \sim \mathcal{N}(0, \sigma^2)$, we have $\mathbb{P}\big[|X| \leq \tau\big] \leq 2\exp e^{-\frac{\tau^2}{2\sigma^2}}$ for any $\tau \geq 0$, from which the claim follows.

*Proof of Lemma 9(v):.* For $X \sim \mathcal{N}(0, \sigma^2)$, we have

$$\mathbb{P}\big[|X| \leq \tau\big] \leq \sqrt{\frac{2}{\pi}}\frac{\tau}{\sigma} \qquad \text{for any } \tau \geq 0 \tag{E.17}$$

from which the claim follows.

*Proof of Lemma 9(vi):.* Similarly, this inequality follows from the tail bound (E.17).

*Proof of Lemma 9(iii):.* This claim follows from parts (v) and (vi) of Lemma 9, combined with the union bound.

*Proof of Lemma 9(ii):.* This bound follows from parts (iii) and (iv) of Lemma 9, combined with the union bound.

E.4.3. *Proof of Lemma 10.* For an event $\mathcal{E}$, define the matrix $\Gamma(\mathcal{E}) = \mathbb{E}[XX^T \mid \mathcal{E}]$. The lemma concerns the operator norm of this matrix for different choices of the event $\mathcal{E}$.

*Conditioned on $\mathcal{E}_1 \cap \mathcal{E}_2$: .* In this case, we write

$$\mathbb{E}\big[XX^T\big] = \Gamma(\mathcal{E}_1 \cap \mathcal{E}_2)\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2] + \Gamma\big((\mathcal{E}_1 \cap \mathcal{E}_2)^c\big)\mathbb{P}[(\mathcal{E}_1 \cap \mathcal{E}_2)^c] \succeq \Gamma(\mathcal{E}_1 \cap \mathcal{E}_2)\, \mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2].$$

Since $\mathbb{E}[XX^T] = I$, we conclude that $|||\Gamma(\mathcal{E}_1 \cap \mathcal{E}_2)|||_{\mathrm{op}} \leq \frac{1}{\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2]}$, and hence it suffices show that $\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2] \geq \frac{1}{2}$. Parts (i) and (ii) of Lemma 9 imply that

$$\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2] \geq 1 - \frac{\|\Delta\|_2}{\|\theta^*\|_2} - \frac{\tau}{\|\theta^*\|_2} - \frac{\tau}{\|\theta_u\|_2} - 2e^{-\frac{\tau^2}{2}}.$$

For appropriate choices of $c_\tau$ and the constant $\eta$ in the signal-to-noise condition (E.3a), the claim follows.

*Conditioned on $\mathcal{E}_1^c$: .* As before, note that the event $\mathcal{E}_1^c$ holds if and only if the inequality $|\langle X, \theta^* + \theta_u\rangle| < |\langle X, \theta^* - \theta_u\rangle|$ holds. Consequently, Lemma 11 implies that $|||\Gamma(\mathcal{E}_1^c)|||_{\mathrm{op}} \leq 2$.

*Conditioned on $\mathcal{E}_5^c$: .* We make note of an elementary fact about Gaussians: for any scalar $\alpha > 0$ and unit norm vector $\|v\|_2 = 1$, for $X \sim \mathcal{N}(0, I_d)$, we have

$$|||\mathbb{E}\big[XX^T \mid |\langle X, v\rangle| \leq \alpha\big]|||_{\mathrm{op}} \leq \max\big(1, \alpha^2\big). \tag{E.18}$$

In particular, when $\alpha \leq 1$, then the operator norm is at most 1. This claim follows easily from the rotation invariance of the Gaussian, which allows us to assume that $v = e_1$ without loss of generality. It is thus equivalent to bound the largest eigenvalue of the matrix

$$D := \mathbb{E}\big[XX^T \mid |X_1| \leq \alpha\big],$$

which is a diagonal matrix by independence of the entries of $X$. Noting that $D_{11} \leq \alpha^2$ and $D_{jj} = 1$ for $j \neq 1$ completes the proof of the bound (E.18).

Applying the bound (E.18), we find that $|||\Gamma(\mathcal{E}_5^c)|||_{\mathrm{op}} \leq \max\Big(1, \frac{\tau^2}{\|\theta_u\|_2^2}\Big)$. Consequently, the claim follows by making sufficiently large choices of $c_\tau$ and the constant $\eta$ in the signal-to-noise condition (E.3a).

*Conditioned on $\mathcal{E}_6^c$: .* The bound (E.18) implies that $|||\mathbb{E}\big[XX^T \mid \mathcal{E}_6^c\big]|||_{\mathrm{op}} \leq \max\big\{1, \frac{\tau^2}{\|\theta^*\|_2^2}\big\}$. As in the previous case, choosing $c_\tau$ and $\eta$ appropriately ensures that $\frac{\tau^2}{\|\theta^*\|_2^2} \leq 1$.

## REFERENCES


[1] Bertsekas, D. P. (1995). *Nonlinear Programming.* Athena Scientific.

[2] Bubeck, S. (2014). Theory of Convex Optimization for Machine Learning.

[3] Cappé, O. and Moulines, E. (2009). On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71** 593–613.

[4] Koltchinskii, V. (2011). *Oracle inequalities in empirical risk minimization and sparse recovery problems École d'été de probabilités de Saint-Flour XXXVIII-2008.* Springer Verlag, Berlin Heidelberg New York.

[5] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes.* Springer-Verlag, New York, NY.

[6] Neal, R. M. and Hinton, G. E. (1999). A View of the EM Algorithm That Justifies Incremental, Sparse, and Other Variants. In *Learning in Graphical Models* (M. I. Jordan, ed.) 355–368. MIT Press, Cambridge, MA, USA.

[7] Nemirovski, A., Juditsky, A., Lan, G. and Shapiro, A. (2009). Robust Stochastic Approximation Approach to Stochastic Programming. *SIAM Journal on Optimization* **19** 1574–1609.

[8] Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization: A Basic Course. Applied Optimization.* Springer.

[9] Noorshams, N. and Wainwright, M. J. (2013). Stochastic belief propagation: A low-complexity alternative to the sum-product algorithm. *IEEE Transactions on Information Theory* **59** 1981–2000.
[10] van der Vaart, A. W. and Wellner, J. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, NY.
[11] Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. Chapter 5 of: Compressed Sensing, Theory and Applications. Edited by Y. Eldar and G. Kutyniok. Cambridge University Press, 2012.
[12] Yi, X., Caramanis, C. and Sanghavi, S. (2013). Alternating Minimization for Mixed Linear Regression.

Department of Statistics
University of California, Berkeley
Berkeley, California 94720
E-mail: sbalakri@berkeley.edu
wainwrig@berkeley.edu
binyu@berkeley.edu

Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
Berkeley, California 94720