Submitted to the Annals of Statistics
arXiv: arXiv:0000.0000

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

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 gradient 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, i.e. 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 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, 46, 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

MSC 2010 subject classifications: Primary 62F10, 60K35; secondary 90C30
Keywords and phrases: EM algorithm, first-order EM algorithm, non-convex optimization, maximum likelihood estimation

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, for instance [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 for 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 MLE (i.e. a maximizer of the sample 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 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 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 regression 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.

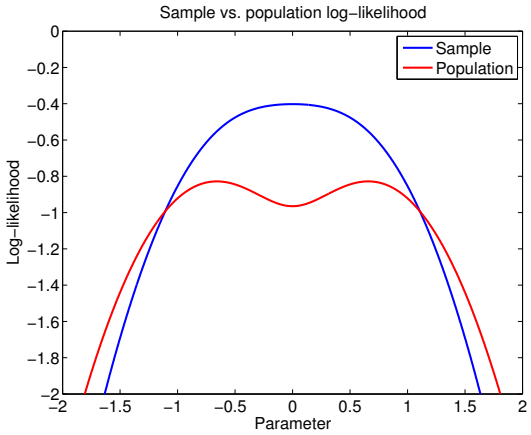

Fig 1. An illustration of the inadequacy of purely sample-based theory guaranteeing linear convergence to any fixed point of the sample-based likelihood. The figure illustrates the population and sample-based likelihoods for samples $y \sim \frac{1}{2}\mathcal{N}(-\theta^*, 1) + \frac{1}{2}\mathcal{N}(\theta^*, 1)$ with $\theta^* = 0.7$. There are two global optima for the population-likelihood corresponding to θ^* 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 θ^* (of radius roughly $\|\theta^*\|_2$), all maxima of the sample-based likelihood are extremely close to θ^*, with an equivalent statement for a neighborhood of $-\theta^*$.

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 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 (Theorem 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 supplement.

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_{θ^*} that belongs to some parameterized family $\{f_\theta \mid \theta \in \Omega\}$ where Ω is some non-empty 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 $k_\theta(z \mid y)$ denote the conditional density of z given y.

Our goal is to obtain an estimate of the unknown parameter θ^* via maximizing the log-likelihood. Throughout this paper we assume that the generative model is correctly specified, with an unknown true parameter θ^*. 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,

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

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).

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

The quantity $Q_n(\theta|\theta')$ provides a lower bound on the log-likelihood $\ell_n(\theta)$ for any θ, 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[1]. 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.[2] There is also a natural extension of the first-order EM iterates that includes a constraint arising from the parameter space Ω, in which the update is projected back using a Euclidean projection onto the constraint set Ω.

It is important to note that in typical examples, several of which are considered in detail in this paper, the likelihood function ℓ_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 θ^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.

2.2. *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

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

where θ^* denotes the true, unknown parameter and g_{θ^*} 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).

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

We can then consider the population analogs of the standard EM and first-order EM updates, obtained by replacing ∇Q_n with ∇Q in equations (2.2) and (2.3), respectively. Our main

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

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

goal is to understand the region of the parameter space over which these iterative schemes, are convergent to θ^*. For the remainder of this section, let us focus exclusively on the population first-order EM updates, given by

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

The concepts developed here are also useful in understanding the EM algorithm; we provide a full treatment of it in Appendix 5 of the supplementary material.

2.3. *Oracle auxiliary function and iterates.* Our key insight is that in a local neighborhood of θ^*, 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 θ^*. 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 θ^* is a maximizer of the population log-likelihood, a classical result [29] guarantees that it must then satisfy the condition

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

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 θ^* as the solution of a concave maximization problem—namely one involving the auxiliary function $q : \Omega \to \mathbb{R}$ given by:

Definition 3 (Oracle auxiliary function).

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

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 α, 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 θ^*. Of course, even in the idealized population setting, the statistician cannot compute the oracle function q, since it presumes knowledge of the unknown parameter θ^*. 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 stability condition* which ensures the closeness of these quantities. Particularly, we consider a condition of the form

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

where $\mathbb{B}_2(r; \theta^*)$ denotes a Euclidean ball[3] of radius r around the fixed point θ^*, and γ is a smoothness parameter. Our first main result (Theorem 1) shows that when the gradient smoothness condition (2.10) holds for appropriate values of γ, then for any initial point $\theta^0 \in \mathbb{B}_2(r; \theta^*)$, the first-order EM iterates converge at a geometric rate to θ^*. In this way, we have

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

a method for explicitly characterizing the region of the parameter space Ω over which the first-order EM iterates converge to θ^*.

Of course, there is no *a priori* reason to suspect that that gradient stability condition (2.10) holds for any non-trivial values of the radius r 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 stability condition (2.10) does hold for reasonable choices of the parameters r and γ, ones which yield accurate predictions of behavior of EM in practice.

2.4. *From population to sample-based analysis.* Recall that 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 ∇Q_n instead of the population gradient ∇Q, a central object in our analysis is the empirical process given by

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

Let $\varepsilon_Q^{\text{unif}}$ be a high probability upper bound on the supremum of this empirical process. 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 θ^* is at most a constant multiple of $\varepsilon_Q^{\text{unif}}$. Figure 2 provides an illustration of the convergence guarantee provided by Theorem 2.

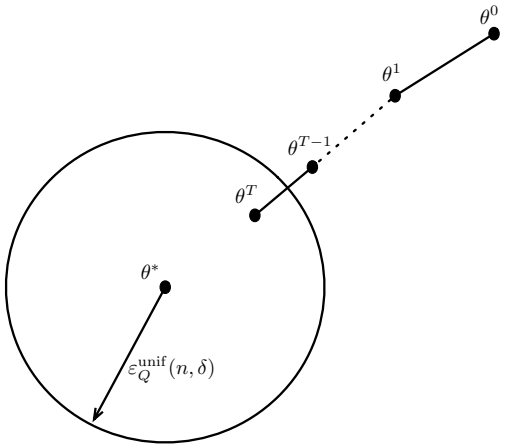

Fig 2. An illustration of Theorem 2. The theorem describes the geometric convergence of iterates of the first-order EM algorithm to the ball of radius $\mathcal{O}(\varepsilon_Q^{\text{unif}}(n,\delta))$.

Of course, this type of approximate convergence to θ^* is only useful if the bound $\varepsilon_Q^{\text{unif}}$ is small enough—ideally, of the same or lower order than the statistical precision, as measured by the Euclidean distance from the MLE to θ^*. 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 non-asymptotic, 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 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 for a sequence of T 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 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 θ^* denote a given global maximum of the population likelihood, our first main result (Theorem 1), as discussed in Section 3.1, characterizes a Euclidean ball around θ^* 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 θ^*. 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 θ^*.

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

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 γ, 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. **λ-strong concavity:** There is some $\lambda > 0$ such that

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

CONDITION 3. **μ-smoothness:** There is some $\mu > 0$ such that

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

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 θ^* maximizes the function q, so that in the unconstrained case, we are guaranteed that $\nabla q(\theta^*) = 0$. Now suppose that the λ-strong concavity and γ-stability conditions hold for some $\gamma < \lambda$. Under these conditions, it is easy to show (see Appendix A.4) that

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

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 non-concave log-likelihood, we see condition (3.4) ensures that the first-order EM algorithm makes progress towards θ^*. Our first main theorem makes this intuition precise, and in fact guarantees a geometric rate of convergence towards θ^*.

THEOREM 1. *For some radius $r > 0$, and a triplet (γ, λ, μ) 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 $\left(1 - \frac{2\lambda-\gamma}{\mu+\lambda}\right) < 1$, the bound (3.5) ensures that at the population level, the first-order EM iterates converge geometrically to θ^*.

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 stability (Condition 1) is more challenging. Intuitively, the gradient stability 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

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

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 σ^2 is known, so that our goal is to estimate the unknown mean vector θ^*. In this example, the hidden variable $Z \in \{0, 1\}$ is an indicator variable for the underlying mixture component—that is

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

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 η 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 η, 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*

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

REMARKS. • The above corollary guarantees that when the SNR is sufficiently large, the population-level first-order EM algorithm converges to θ^* when initialized at any point in a ball of radius $\|\theta^*\|_2/4$ around θ^*. 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 θ^* and $-\theta^*$, along with a hyperplane of bad local fixed-points, i.e. any point that is equi-distant from θ^* and $-\theta^*$ is a (bad) fixed-point of the population EM algorithm. Observing that, for instance, 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 θ^* 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 θ^*. 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 stability, λ-strong concavity, and μ-smoothness (Conditions 1 through 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. *Mixtures 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, $x_i \sim \mathcal{N}(0, I)$ 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 θ^* 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 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 θ^* 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 η, 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*

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

REMARKS. • As with the Gaussian mixture model, the population likelihood has global maxima at θ^* and $-\theta^*$, and a local minimum at 0. Consequently, the largest Euclidean ball over which the iterates could converge to θ^* would have radius $\|\theta^*\|_2$. Thus, we see that our framework gives an order-optimal characterization of the region of convergence.[4]

- Our analysis shows that the rate of convergence is again a decreasing function of the SNR parameter η. 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 function q for the MOR model satisfies the required gradient stability, 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).

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

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, i.e. the radius of the region around θ^* within which the population EM algorithm is globally convergent. 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 ρ satisfying the bound (3.13b), the first-order EM iterates with stepsize 1, satisfy the bound*

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

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

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 7 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 smoothness examples is relatively straightforward. In contrast, verifying the gradient stability (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 missing-ness—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

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

where $\alpha > 0$ is an appropriately chosen stepsize. As mentioned previously, we need to control the deviations of the sample gradient ∇Q_n from the population version ∇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

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

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 (γ, λ, μ) with $0 \leq \gamma < \lambda \leq \mu$ such that the gradient stability, strong-concavity and smoothness conditions hold (Conditions 1 through 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

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

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 3 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 θ^*. 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 integer T such that

$$T \geq \log_{1/\kappa} \frac{(1-\kappa)(\mu+\lambda)\,\|\theta^0 - \theta^*\|_2}{2\varepsilon_Q^{\text{unif}}(n, \delta)}. \tag{4.5}$$

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

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

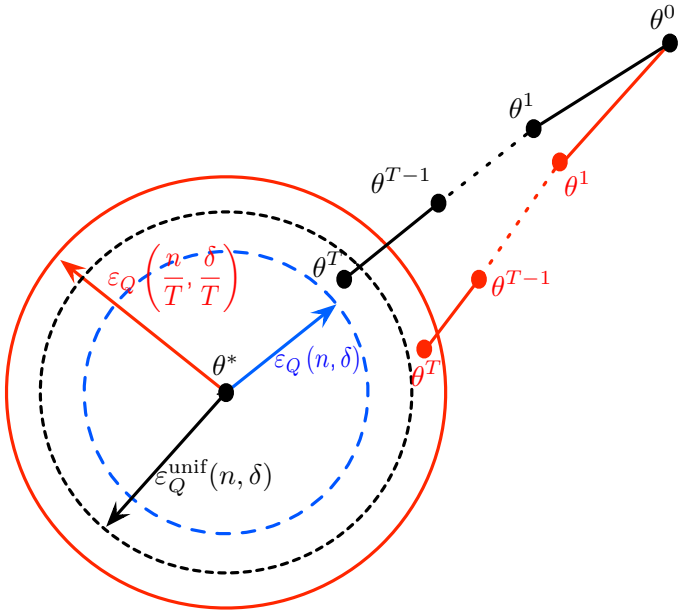

Fig 3. An illustration of Theorems 2 and 3. The first part of the theorem describes the geometric convergence of iterates of the EM algorithm to the ball of radius $\mathcal{O}(\varepsilon_Q^{\text{unif}}(n, \delta))$ (in black). The second part describes the geometric convergence of the sample-splitting 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).

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 ε_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 θ. 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 (γ, λ, μ) such that $0 \leq \gamma < \lambda \leq \mu$, the gradient stability, concavity and smoothness properties (Conditions 1 through 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

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

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 ε_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 $\varphi(\sigma; \|\theta^*\|_2) := \|\theta^*\|_2\big(1 + \frac{\|\theta^*\|_2^2}{\sigma^2}\big)$, 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 $n \geq c_1 d \log(1/\delta)$. 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 first-order EM iterates $\{\theta^t\}_{t=0}^{\infty}$ satisfy the bound*

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

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 supplement ([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 σ, 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

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

With this choice, we are guaranteed that the iterate θ^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 θ^* and the contraction coefficient κ. 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:

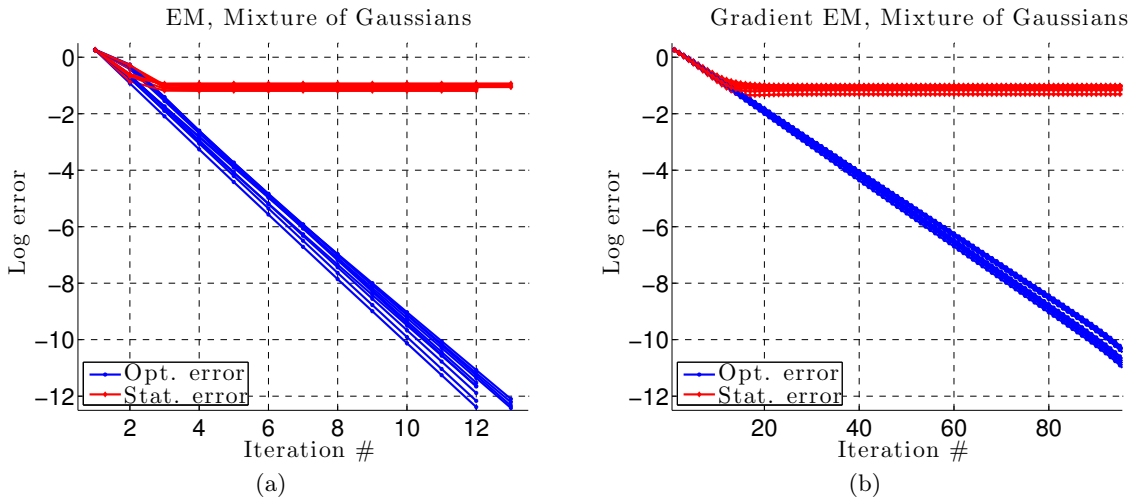

Fig 4. Plots of the iteration number versus log optimization error $\log(\|\theta^t - \widehat{\theta}\|_2)$ and log statistical error $\log(\|\theta^t - \theta^*\|_2)$. (a) Results for the EM algorithm[6]. (b) Results for the first-order EM algorithm. Each plot shows 10 different problem instances with 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.

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.

- 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 5 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.

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

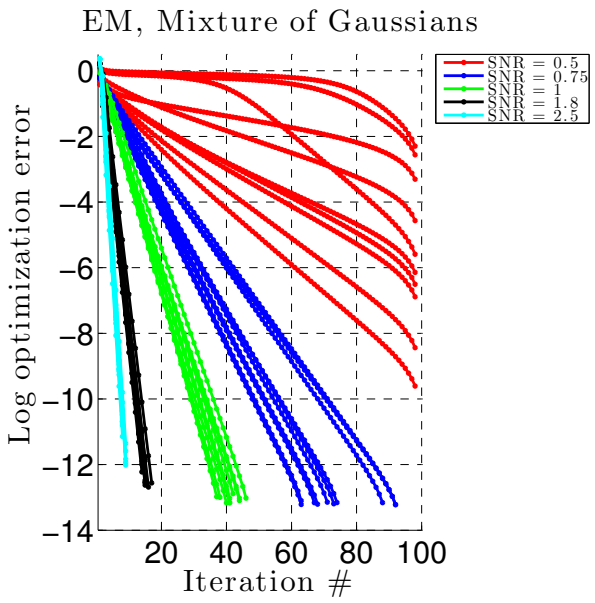

Fig 5. Plot of the iteration number versus the (log) optimization error $\log(\|\theta^t - \widehat{\theta}\|_2)$[9] for different values of the SNR $\frac{\|\theta^*\|_2}{\sigma}$. For each SNR, we performed 10 independent trials of a Gaussian mixture model with dimension $d = 10$ and sample size $n = 1000$. 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

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

with probability at least $1 - \delta$. Apart from the logarithmic penalty $\log^2\big(\frac{n}{d\varphi^2(\sigma;\|\theta^*\|_2)}\big)$, 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 6 shows the results of simulations designed to test this prediction: see the caption for the details.

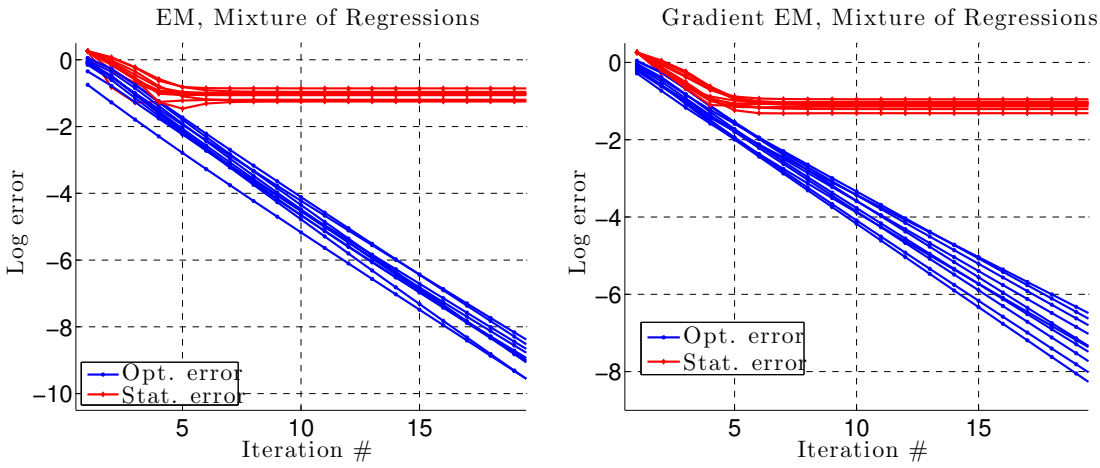

Fig 6. Plots of the iteration number versus log optimization error $\log(\|\theta^t - \widehat{\theta}\|_2)$ and log statistical error $\log(\|\theta^t - \theta^*\|_2)$ for mixture of regressions. (a) Results for the EM algorithm. (b) Results for the first-order EM algorithm. Each plot shows 10 independent 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.

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*

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

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 (ξ_1, ξ_2), but does not otherwise depend on n, d, σ^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 $1 - \delta$. Modulo the logarithmic penalty in n, incurred due to the sample-splitting, this estimate achieves the optimal $\sqrt{\frac{d}{n}}$ scaling of the ℓ_2-error.

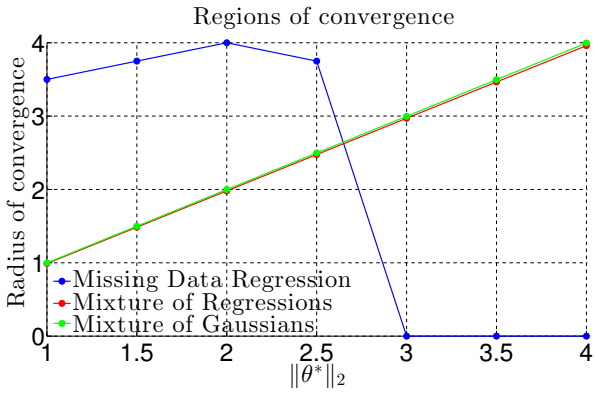

Fig 7. Simulations of the radius of convergence for problems of dimension $d = 10$, sample size $n = 1000$, and variance $\sigma^2 = 1$. Radius of convergence is defined as the maximum value of $\|\theta^0 - \theta^*\|_2$ for which initialization at θ^0 leads to convergence to an optimum near θ^*. Consistent 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.

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 non-asymptotic finite-sample guarantees.

5.1. *Analysis of the EM algorithm at the population level.* We assume throughout this section that the function q is λ-*strongly concave* (but not necessarily smooth). For any fixed θ, in order to relate the population EM updates to the fixed point θ^*, 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 θ^* and the EM update. In order to compactly represent the EM update, we define the operator $M : \Omega \to \Omega$,

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

Using this notation, the EM algorithm given some initialization θ^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 Ω, the fixed point satisfies the first-order optimality condition

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

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

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

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 θ^*. Thus, at an intuitive level, in order to establish that θ^{t+1} and θ^* are close, it suffices to verify

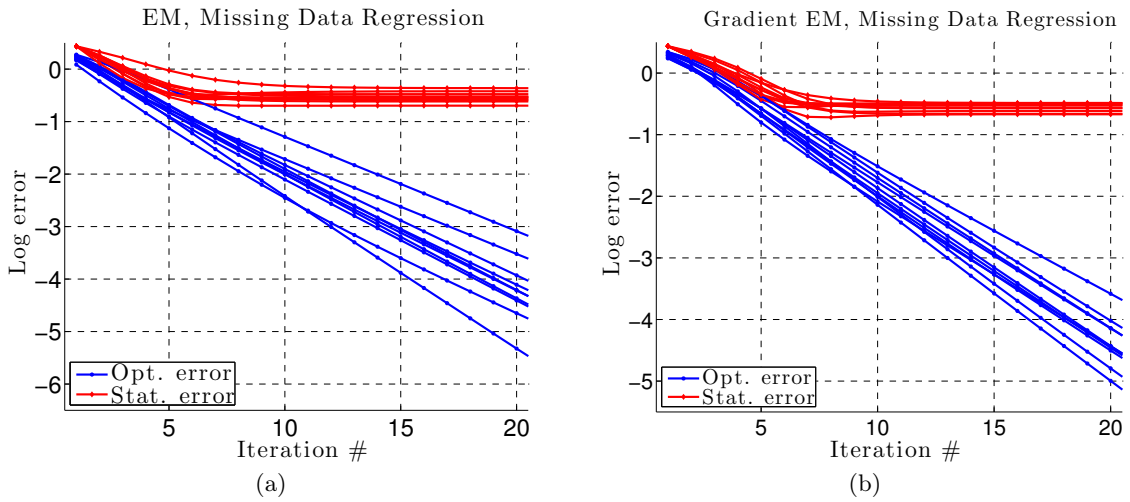

Fig 8. Plots of the iteration number versus log optimization error $\log(\|\theta^t - \widehat{\theta}\|_2)$ and log statistical error $\log(\|\theta^t - \theta^*\|_2)$ for regression with missing covariates. (a) Results for the EM algorithm. (b) Results for the first-order EM algorithm. Each plot shows 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.

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 [44]). In the analysis of EM, we obtain additional leverage from the self-consistency condition (2.7) that characterizes θ^*.

With this intuition in mind, we introduce the following regularity condition in order to relate conditions (5.3) and (2.7): The condition involves a Euclidean ball of radius r around the fixed point θ^*, given by

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

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

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

To provide some high-level intuition, observe the condition (5.5) is always satisfied at the fixed point θ^*, in particular with parameter $\gamma = 0$. Intuitively then, by allowing for a strictly positive parameter γ, one might expect that this condition would hold in a local neighborhood $\mathbb{B}_2(r; \theta^*)$ of the fixed point θ^*, 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 θ^* for which the FOS condition holds (with an appropriate γ) is in the region of attraction of θ^*—i.e. the population EM update produces an iterate closer to θ^* that 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 (γ, λ) such that $0 \leq \gamma < \lambda$, suppose that the function $Q(\cdot|\theta^*)$ is λ-strongly concave (3.2), and that the FOS(γ) 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

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

The proof is a consequence of the KKT conditions from equations (5.2) and (5.3), along with consequences of the strong convexity 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—viz.

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

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

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

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 θ^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

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

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

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

REMARKS. In order to obtain readily interpretable bounds for specific models, it only remains to establish the κ-contractivity of the population operator, and to compute either the function ε_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 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 λ-concavity and μ-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

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

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.* For any $\theta^s \in \mathbb{B}_2(r;\theta^*)$, we have that

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

with probability at least $1-\delta$. We perform the remainder of our analysis under this event.

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

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

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

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

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

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

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 $\theta^0 \in \mathbb{B}_2(r;\theta^*)$, i.e. 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 through 3 on the behavior of first-order EM at the population level for concrete models.

6.3.1. *Proof of Corollary 1.* In order to apply Theorem 1, we need to verify the λ-concavity (3.2) and μ-smoothness (3.3) conditions, and the GS(γ) 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

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

where the weighting function is given by

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

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

It remains to verify the GS(γ) 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 (θ^* 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 θ and θ^* 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

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

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 λ-concavity (3.2) and μ-smoothness (3.3) conditions, and the GS(γ) condition (3.1) over the ball $\mathbb{B}_2(r;\theta^*)$. In this example, the q-function takes the form:

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

where $w_\theta(x,y) := \frac{\exp\left(\frac{-(y-\langle x,\,\theta\rangle)^2}{2\sigma^2}\right)}{\exp\left(\frac{-(y-\langle x,\,\theta\rangle)^2}{2\sigma^2}\right)+\exp\left(\frac{-(y+\langle x,\,\theta\rangle)^2}{2\sigma^2}\right)}$. Observe that function $Q(\cdot|\theta^*)$ is λ-strongly concave and μ-smooth with λ and μ equal to the smallest and largest (respectively) 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

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

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

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

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

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

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 $\langle \mathbb{E}\big[\Delta_w(X,Y)YX\big], \widetilde{\Delta}\rangle \leq \gamma \|\Delta\|_2 \|\widetilde{\Delta}\|_2$ 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 μ-smooth, λ-strongly concave, and that the GS condition is satisfied. In this case, q is a quadratic of the form

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

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

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

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

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

$$V := \Sigma_{\theta^*}(X_{\text{obs}}, Y)\theta - Y\mu_{\theta^*}(X_{\text{obs}}, Y) - \Sigma_{\theta}(X_{\text{obs}}, Y)\theta + Y\mu_{\theta}(X_{\text{obs}}, Y). \tag{6.7}$$

For a fixed pattern of missingness, we can compute the expectation over (X_{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

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

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 ρ, the expectation of the i^{th} coordinate is at most $\big|\mathbb{E}[V]|_i \leq \big|\rho|\theta_i - \theta^*_i| + \rho\pi_1|\theta_i| + (1-\rho)\pi_2|\theta_i - \theta^*_i|\big|$. Thus, defining $\eta := (1-\rho)\delta + \rho < 1$, we have

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

where we have used our upper bound (6.9) on π_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 γ^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 $\|\theta_{\mathrm{mis}}\|_2 - \|\theta^*_{\mathrm{mis}}\|_2 \leq \|\theta_{\mathrm{mis}} - \theta^*_{\mathrm{mis}}\|_2$, and moreover

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

As consequence, we have

$$\|\theta^*_{\mathrm{mis}}\|_2^2 - \|\theta_{\mathrm{mis}}\|_2^2 = (\|\theta_{\mathrm{mis}}\|_2 - \|\theta^*_{\mathrm{mis}}\|_2)(\|\theta_{\mathrm{mis}}\|_2 + \|\theta^*_{\mathrm{mis}}\|_2) \leq (2\xi_1 + \xi_2)\sigma\|\theta_{\mathrm{mis}} - \theta^*_{\mathrm{mis}}\|_2$$

Since $\|\theta_{\mathrm{obs}} - \theta^*_{\mathrm{obs}}\|_2^2 \leq \xi_2\sigma\|\theta_{\mathrm{obs}} - \theta^*_{\mathrm{obs}}\|_2$, the stated bound on π_1 follows.

On the other hand, we have

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

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 $\mathbb{A} := \{\theta \in \mathbb{R}^d \mid \|\theta - \theta^*\|_2 \leq \|\theta^*\|_2/4\}$, we need to control the random variable variable

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

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

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

with probability at least $1 - \delta$.

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

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

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 $\{u^1, \ldots, u^M\}$ 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

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

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

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

using a standard symmetrization result for empirical processes (e.g., [23, 24]). Now observe that for any triplet of d-vectors y, θ and θ', 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 [23, 24], 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

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

where $\|\cdot\|_{\mathrm{op}}$ denotes the ℓ_2-operator norm of a matrix (maximum singular value). Repeating the same discretization argument over $\{u^1, \ldots, u^M\}$, we find that

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

Putting together the pieces, we conclude that

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

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 η 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 $\mathbb{E}[e^{t(\langle y_i, u\rangle)^2}] \leq e^{\frac{\gamma^4 t^2}{2}}$ for all $t > 0$ sufficiently small. Combined with our earlier inequality (6.12), we conclude that

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

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

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

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 $1 - \delta$ 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

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

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

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

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

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

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

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

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

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

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 [45], 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 [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 $(y_i \mid x_i)$ for each index i, we find that

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

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

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

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 [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 θ^* 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,

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

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

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

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 Σ_θ and μ_θ when the arguments are understood. We use the notation $\odot$ to denote the element-wise product.

Controlling T_1:. Define the matrices $\bar{\Sigma} = \mathbb{E}[\Sigma_\theta(x_{\text{obs}}, y)]$ and $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\Sigma_\theta(x_{\text{obs},i}, y_i)$. With this notation, we have $T_1 \leq |||\bar{\Sigma} - \widehat{\Sigma}|||_{\text{op}}\, \|\theta\|_2 \leq |||\bar{\Sigma} - \widehat{\Sigma}|||_{\text{op}}\, (\xi_1 + \xi_2)\,\sigma$, where the second step follows since any vector $\theta \in \mathbb{B}_2(r; \theta^*)$ has ℓ_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 $|||\bar{\Sigma} - \widehat{\Sigma}|||_{\text{op}} \leq c\sqrt{\frac{d}{n}\log(1/\delta)}$ 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

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

where I_{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 μ_θ 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 $\mu_\theta(x_{\text{obs}}, y)$ is sub-Gaussian with a constant parameter.*

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

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

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

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

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 ω are independent. Moreover, with ω fixed, the variable $\langle x, \omega \rangle$ is sub-Gaussian with parameter $\|\omega\|_2^2$, whence

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

where the final inequality uses the fact that $\langle \omega, u \rangle^2 \leq \|\omega\|_2^2$. We have thus shown that B_2 is sub-Gaussian with parameter one. Since $\|\theta - \theta^*\|_2 \leq \xi_2 \sigma$, the same argument shows that B_3 is sub-Gaussian with parameter at most ξ_2. Since v is sub-Gaussian with parameter σ and independent of ω, 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 $\mathbb{P}[T_2 \geq t] \leq 2M \exp\left(- c \min\left\{ \frac{nt}{\sqrt{1+\sigma^2}}, \frac{nt^2}{1+\sigma^2} \right\} \right)$. 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 $\varepsilon_Q(n, \delta) \leq c\sqrt{1+\sigma^2} \sqrt{\frac{d}{n} \log(1/\delta)}$ 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 non-convex 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 particular 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. This research was partially supported by ONR-MURI grant DOD 002888. and NSF grant CIF-31712-23800, as well by US NSF grants DMS-1107000, CDS&E-MSS 1228246, ARO grant W911NF-11-1-0114, the Center for Science of Information (CSoI), and US NSF Science and Technology Center, under grant agreement CCF-0939370. The authors would also 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: xxx; .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. *CoRR* **abs/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.
[3] Balakrishnan, S., Wainwright, M. J. and Yu, B. (2014). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *XXX*.
[4] Balan, R., Casazza, P. and Edidin, D. (2006). On signal reconstruction without phase. *Applied and Computational Harmonic Analysis* **20** 345 - 356.
[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. *The Annals of Mathematical Statistics* **41** 164–171.
[6] Beale, E. M. L. and Little, R. J. A. (1975). Missing Values in Multivariate Analysis. *Journal of the Royal Statistical Society. Series B (Methodological)* **37** pp. 129-145.
[7] Bertsekas, D. P. (1995). *Nonlinear Programming*. Athena Scientific.
[8] Bubeck, S. (2014). Theory of Convex Optimization for Machine Learning.
[9] Buldygin, V. V. and Kozachenko, Y. V. (2000). *Metric characterization of random variables and random processes*. American Mathematical Society, Providence, RI.
[10] Candès, E. J., Strohmer, T. and Voroninski, V. (2013). PhaseLift: Exact and Stable Signal Recovery from Magnitude Measurements via Convex Programming. *Communications on Pure and Applied Mathematics* **66** 1241–1274.
[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. Stat. Data Anal.* **14** 315–332.
[13] Chaganty, A. T. and Liang, P. (2013). Spectral Experts for Estimating Mixtures of Linear Regressions.
[14] Chen, Y., Yi, X. and Caramanis, C. (2013). A Convex Formulation for Mixed Regression: Near Optimal Rates in the Face of Noise.
[15] Chrétien, S. and Hero, A. O. (2008). On EM algorithms and their proximal generalizations. *ESAIM: Probability and Statistics* **12** 308–326.
[16] Dasgupta, S. and Schulman, L. J. (2007). A Probabilistic Analysis of EM for Mixtures of Separated, Spherical Gaussians. *Journal of Machine Learning Research* **8** 203-226.
[17] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39** 1–38.

[18] Hartley, H. O. (1958). Maximum Likelihood Estimation from Incomplete Data. *Biometrics* **14** pp. 174-194.
[19] Healy, M. and Westmacott, M. (1956). Missing Values in Experiments Analysed on Automatic Computers. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **5** pp. 203-206.
[20] Hero, A. O. and Fessler, J. A. (1995). Convergence in Norm for Alternating Expectation-Maximization (EM) Type Algorithms. *Statistica Sinica* **5** 41–54.
[21] Hsu, D. and Kakade, S. M. (2012). Learning Gaussian Mixture Models: Moment Methods and Spectral Decompositions. *CoRR* **abs/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. *Journal of the Royal Statistical Society Series B - Statistical Methodology* **61** 547–561.
[23] 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.
[24] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York, NY.
[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.
[26] Loh, P.-L. and Wainwright, M. J. (2012). Corrupted and missing predictors: Minimax bounds for high-dimensional linear regression. In *ISIT* 2601-2605.
[27] Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society: Series B* **44** 226-233.
[28] Ma, J. and Xu, L. (2005). Asymptotic Convergence Properties of the EM Algorithm with Respect to the Overlap in the Mixture. *Neurocomputing* **68**.
[29] McLachlan, G. and Krishnan, T. (2007). *The EM Algorithm and Extensions. Wiley Series in Probability and Statistics*. Wiley.
[30] Meilijson, I. (1989). A fast improvement of the EM algorithm on its own terms. *Journal of the Royal Statistical Society: Series B* **51** 127-138.
[31] Meng, X.-L. (1994). On the Rate of Convergence of the ECM Algorithm. *The Annals of Statistics* **22** 326–339.
[32] Meng, X. L. and Rubin, D. B. (1993). Maximum likelihood via the ECM algorithm: a general framework. *Biometrika* **80** 267-278.
[33] Meng, X.-L. and Rubin, D. B. (1994). On the global and componentwise rates of convergence of the EM algorithm. *Linear Algebra and its Applications* **199, Supplement 1** 413 - 425. Special Issue Honoring Ingram Olkin.
[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, USA.
[35] Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization: A Basic Course. Applied Optimization*. Springer.
[36] Netrapalli, P., Jain, P. and Sanghavi, S. (2013). Phase Retrieval using Alternating Minimization. In *NIPS* 2796-2804.
[37] Orchard, T. and Woodbury, M. A. (1972). A missing information principle: theory and applications. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of Statistics* 697–715.
[38] Pearson, K. (1894). *Contributions to the Mathematical Theory of Evolution*. Harrison and Sons.
[39] Redner, R. A. and Walker, H. F. (1984). Mixture Densities, Maximum Likelihood and the EM Algorithm. *SIAM Review* **26** 195–239.
[40] Rubin, D. B. (1974). Characterizing the Estimation of Parameters in Incomplete-Data Problems. *Journal of the American Statistical Association* **69** pp. 467-474.
[41] Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. *Scand. J. Statist* **1** 49-58.
[42] Tanner, M. A. and Wong, W. H. (1987). The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association* **82** 528-550.
[43] Tseng, P. (2004). An Analysis of the EM Algorithm and Entropy-Like Proximal Point Methods. *Mathematics of Operations Research* **29** pp. 27-44.
[44] van de Geer, S. (2000). *Empirical Processes in M-Estimation*. Cambridge University Press.
[45] van der Vaart, A. W. and Wellner, J. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, NY.
[46] 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 31st Symposium on*

the Interface 230-239. Berk and M. Pourahmadi.

[47] 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.

[48] Wang, Z., Gu, Q., Ning, Y. and Liu, H. (2014). High Dimensional Expectation-Maximization Algorithm: Statistical Optimization and Asymptotic Normality.

[49] Wei, G. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *Journal of the American Statistical Association* **85** 699-704.

[50] Wu, C. F. J. (1983). On the Convergence Properties of the EM Algorithm. *The Annals of Statistics* **11** 95–103.

[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. *Communications in Statistics - Theory and Methods* **36** 375-386.

[53] Yi, X., Caramanis, C. and Sanghavi, S. (2013). Alternating Minimization for Mixed Linear Regression.

Department of Statistics
University of California, Berkeley
Berkeley, CA 94720
E-mail: sbalakri@berkeley.edu
wainwrig@berkeley.edu
binyu@berkeley.edu

Department of Electrical Engineering and Computer Science
University of California, Berkeley
Berkeley, CA 94720