Statistical guarantees for the EM algorithm:
From population to sample-based analysis

Sivaraman Balakrishnan[†]   Martin J. Wainwright[†,*]   Bin Yu[†,*]

Department of Statistics[†]   Department of Electrical Engineering and Computer Sciences[*]
University of California, Berkeley
Berkeley, CA 94720

{sbalakri,wainwrig,binyu}@berkeley.edu

August 12, 2014

Abstract

We develop a general framework for proving rigorous guarantees on the performance of the EM algorithm and a variant known as gradient EM. Our analysis is divided into two parts: a treatment of these algorithms at the population level (in the limit of infinite data), followed by results that apply to updates based on a finite set of samples. First, we characterize the domain of attraction of any global maximizer of the population likelihood. This characterization is based on a novel view of the EM updates as a perturbed form of likelihood ascent, or in parallel, of the gradient EM updates as a perturbed form of standard gradient ascent. Leveraging this characterization, we then provide non-asymptotic guarantees on the EM and gradient EM algorithms when applied to a finite set of samples. We develop consequences of our general theory for three canonical examples of incomplete-data problems: mixture of Gaussians, mixture of regressions, and linear regression with covariates missing completely at random. In each case, our theory guarantees that with a suitable initialization, a relatively small number of EM (or gradient EM) steps will yield (with high probability) an estimate that is within statistical error of the MLE. We provide simulations to confirm this theoretically predicted behavior.

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., [10, 11, 16, 25, 27, 30, 31, 32, 42, 46, 48]). 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. The goal of this paper is to address this potential gap between statistical and computational guarantees in application of the EM algorithm.

The EM algorithm has a lengthy and rich history. Various algorithms of the EM-type were analyzed in early work (e.g.,[4, 5, 17, 18, 37, 40, 41]), before Dempster et al. [16] introduced the EM algorithm in its modern general form. Among other results, they established its well-known monotonicity properties. The subsequent work of Wu [49] established some of the most general convergence results known for the EM algorithm; see also the more recent papers [14, 43]. Together with other results, Wu [49] showed that if the likelihood is unimodal

1

and certain regularity conditions hold, then the EM algorithm converges to the unique global optimum. However, in most interesting cases of the EM algorithm, the likelihood function is multi-modal, in which case the behavior of the EM algorithm remains a little more mysterious. Indeed, despite its popularity and widespread practical effectiveness, the EM algorithm is often considered a "sensible heuristic" with little or no theoretical backing.

One interesting observation with the EM algorithm is given a "suitable" initialization, it often converges to a statistically useful estimate. For instance, in application to a mixture of regressions problem (see Section 2.2.2 for more details), Chaganty and Liang [12] empirically demonstrate good performance for a two-stage estimator, in which the method of methods is used as an initialization, and then the EM algorithm is applied to refine this initial estimator. Although encouraging, this type of behavior is not well understood in a quantitative sense, especially how EM fixed points reached by this type of two-stage estimator are related to the global maximizers of the population likelihood. The goal of this paper is to address this question, and to develop some general tools for characterizing fixed points of the suitably initialized sample-based EM algorithm, and their relation to maximum likelihood estimates.

Some two-stage estimators have recently been analyzed in work on alternating minimization algorithms (see e.g. [21, 22, 35, 52]) which show that at least in certain special cases optimization methods can be locally effective despite non-convexity. Most directly related to our work is the paper of Yi et al. [52] which considers a special (degenerate) noiseless case of the EM algorithm for the mixtures of regressions problem. Results for the noisy mixtures of regressions problem follow from our general treatment of the EM algorithm (see Section 2.2.2).

In some settings, performing an exact M-step is computationally burdensome, in which case a natural alternative is some form of generalized EM updates. In such an algorithm, instead of performing an exact maximization, we simply choose a parameter value that does not decrease the likelihood. In addition to the standard EM updates, we also analyze a particular case of such an algorithm, known as gradient EM, based on a taking a single gradient step per iteration.

Our main results concern the population EM and gradient EM algorithms and their finite-sample counterparts. Our first set of results (Theorems 1 and 3) give conditions under which the population algorithms are contractive to the MLE, in a ball around the MLE. These results are completely deterministic. This population-level analysis is based on viewing these algorithms as perturbed versions of certain "oracle" algorithms which are known to be contractive around the MLE. Our second set of results (Theorem 2, Theorem 4 and Theorem 5) concern the sample-based EM and gradient EM algorithms which approximate the population-based algorithms using a subset of samples at each step. We give conditions under which these sample operators converge to an ε-ball around the population MLE. These results involve probabilistic bounds on the deviations between the iterates of the population and sample-based algorithms.

The remainder of this paper is organized as follows. Section 2 provides an introduction to the EM and gradient EM algorithms, as well as a description of the three examples treated in detail in this paper—namely, Gaussian mixture models (Section 2.2.1), mixture of regressions (Section 2.2.2), and regression with missing covariates (Section 2.2.3). Section 3 is devoted to our general convergence results on both the EM and gradient EM algorithms. In Section 4, we revisit the three model classes previously introduced, and illustrate the use of our general theory by deriving some concrete corollaries. In concrete examples our theory gives a characterization of the quality of initialization needed and the rate of convergence of the EM and gradient EM algorithms. We complement these theoretical results with simulations that confirm various aspects of the theoretical predictions. In order to promote readability, we defer the more technical aspects of proofs to the appendices.

2 Background and model examples

We begin with basic background on the EM algorithm and its variants, along with a number of specific models that we revisit later in the paper.

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 $\{f_\theta \mid \theta \in \Omega\}$, for a non-empty compact convex set Ω. Rather than observing the complete data (Y, Z), we observe only component Y. Thus, the component Z corresponds to the missing or latent structure in the data.

Our goal is to obtain an estimate of the unknown parameter θ^* via maximum likelihood—namely, to compute some $\widehat{\theta} \in \Omega$ maximizing the function $\theta \mapsto g_\theta(y)$, where

$$g_\theta(y) = \int_{\mathcal{Z}} f_\theta(y, z) dz \tag{1}$$

is the density function of the observed variable Y. Throughout this paper, we assume that θ^* is a maximizer of the population likelihood, but not that θ^* is a *unique* maximizer. Uniqueness is often violated in mixture models for which parameters are typically only identifiable up to permutation. In the examples that we consider, this non-identifiability will be resolved by appropriate initialization conditions.

In many settings, it can be difficult or computationally expensive to evaluate the log likelihood of the observed data, but relatively easy to compute the log likelihood $\log f_\theta(y, z)$ of both the latent and observed variables. The EM algorithm is well-suited to such settings. For each $\theta \in \Omega$, let $k_\theta(z \mid y)$ denote the conditional density of z given y. A straightforward application of Jensen's inequality then shows that the log likelihood at $\theta' \in \Omega$ can be lower bounded as

$$\log g_{\theta'}(y) \geq \underbrace{\int_{\mathcal{Z}} k_\theta(z \mid y) \log f_{\theta'}(y, z) dz}_{Q(\theta'|\theta)} - \int_{\mathcal{Z}} k_\theta(z \mid y) \log k_\theta(z \mid y) dz, \tag{2}$$

with equality holding when $\theta = \theta'$. Thus, we have a family of lower bounds on the log likelihood, and the EM algorithm successively maximizes this lower bound (M-step), and then reevaluates the lower bound at the new parameter value (E-step).

Standard EM updates: With this notation, it is easy to specify the EM iterations. The update $\theta^t \to \theta^{t+1}$ consists of the following two steps.

- E-step: Evaluate the expectation in equation (2) to compute $Q(\cdot|\theta^t)$.

- M-step: Compute the maximizer $\theta^{t+1} = \arg\max_{\theta' \in \Omega} Q(\theta'|\theta^t)$.

For future use, it is convenient to introduce the mapping $M : \Omega \to \Omega$ given by

$$M(\theta) := \arg\max_{\theta' \in \Omega} Q(\theta'|\theta). \tag{3}$$

With this choice, the M-step corresponds to the update $\theta^{t+1} = M(\theta^t)$.

Generalized EM updates: In a *generalized EM* algorithm, the requirements of the M-step are relaxed: instead of finding the exact optimum, the algorithm is required only to find a value $\theta^{t+1} \in \Omega$ such that

$$Q(\theta^{t+1}|\theta^t) \geq Q(\theta^t|\theta^t). \tag{4}$$

Depending on how θ^{t+1} is chosen, this requirement actually defines a family of algorithms.

Gradient EM updates: A closely related variant of the generalized EM updates is what we refer to as the *gradient EM updates*, applicable in the case when the function $Q(\cdot|\theta^t)$ is differentiable at each iteration t. Given a step size $\alpha > 0$, these updates take the form

$$\theta^{t+1} = \theta^t + \alpha \nabla Q(\theta^t|\theta^t), \tag{5}$$

where the gradient is taken in the first argument of Q. For ease of notation, we define the mapping $G \colon \Omega \to \Omega$ by

$$G(\theta) = \theta + \alpha \nabla Q(\theta|\theta). \tag{6}$$

An iteration of gradient EM can now be written compactly as $\theta^{t+1} = G(\theta^t)$.

There is a natural extension that includes a constraint arising from the parameter space Ω, in which the update is projected back onto the constraint set [1]. For simplicity, we focus on unconstrained problems in this paper, but all of our results extend in a straightforward way to constrained examples by incorporating the additional Euclidean projection. For appropriate choices of the step size parameter α, the gradient EM updates guarantee the ascent condition (4), so that it is a particular case of a generalized EM algorithm.

Population versus sample updates: Let us now make an important distinction, namely, that between the population and sample-based versions of the EM updates. Up to this point, we have suppressed dependence on the number of observed samples n. The population form of the (gradient) EM updates are an "oracle version", in which we effectively observe an infinite number of samples, and consequently, the function $Q(\cdot|\theta)$ takes the form

$$Q(\theta'|\theta) = \int_{\mathcal{Y}} \left(\int_{\mathcal{Z}} k_\theta(z \mid y) \log f_{\theta'}(y, z) dz \right) g_{\theta^*}(y) dy. \tag{7}$$

From here onwards, we use the notation M and G for the EM and gradient EM operators, respectively, both defined at the population level.

In the classical statistical settings, we observe only n i.i.d. samples $\{y_i\}_{i=1}^n$ of the Y component. Under the i.i.d. assumption, we define the function

$$Q_n(\theta'|\theta) = \frac{1}{n} \sum_{i=1}^n \left(\int_{\mathcal{Z}} k_\theta(z \mid y_i) \log f_{\theta'}(y_i, z) dz \right), \tag{8}$$

so that the expectation over Y in equation (7) is replaced by the empirical expectation defined by the samples. The function Q_n defines an analog of the population EM operator (3), namely

$$M_n(\theta) = \arg \max_{\theta' \in \Omega} Q_n(\theta'|\theta). \tag{9}$$

---

[1]To avoid pathologies additionally assume that the constraint set is closed.

In an analogous fashion, we define the sample-based analog of the gradient EM operator (6), namely

$$G_n(\theta) := \theta + \alpha \nabla Q_n(\theta|\theta), \tag{10}$$

where $\alpha > 0$ is an appropriately chosen step size parameter.

2.2 Illustrative examples

The EM algorithm is popular and a variety of examples can be found in the literature. In this section, we review three specific models analyzed in this paper, and derive the form of the population and sample-based updates, both for the usual EM algorithm and the gradient EM algorithm.

2.2.1 Gaussian mixture models

An isotropic, balanced two-component Gaussian mixture model can be specified by a density of the form

$$f_\theta(y) = \frac{1}{2}\phi(y; \theta^*, \sigma^2 I_d) + \frac{1}{2}\phi(y; -\theta^*, \sigma^2 I_d), \tag{11}$$

where $\phi(\,\cdot\,; \mu, \Sigma)$ denotes the density of a $\mathcal{N}(\mu, \Sigma)$ random vector in $\mathbb{R}^d$. Here we have assumed that the components are equally weighted; with the variance σ^2 known, the 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

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

Suppose that we are given n i.i.d. samples $\{y_i\}_{i=1}^n$ drawn from the mixture density (11). The complete data $\{(y_i, z_i)\}_{i=1}^n$ corresponds to the original samples along with the component indicator variables $z_i \in \{0, 1\}$. The sample-based function Q_n takes the form

$$Q_n(\theta'|\theta) = -\frac{1}{2n}\sum_{i=1}^n \left[w_\theta(y_i)\|y_i - \theta'\|_2^2 + (1 - w_\theta(y_i))\|y_i + \theta'\|_2^2 \right], \tag{12}$$

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

EM updates: This example is especially simple in that the EM operator $M_n : \mathbb{R}^d \to \mathbb{R}$ has a closed form solution, given by

$$M_n(\theta) := \arg\max_{\theta' \in \mathbb{R}^d} Q_n(\theta'|\theta) = \frac{2}{n}\sum_{i=1}^n w_\theta(y_i)y_i - \frac{1}{n}\sum_{i=1}^n y_i. \tag{13a}$$

The population EM operator $M : \mathbb{R}^d \to \mathbb{R}^d$ is defined analogously

$$M(\theta) = 2\mathbb{E}\left[w_\theta(Y)Y \right], \tag{13b}$$

where the empirical expectation has been replaced by expectation under the mixture distribution (11).

5

Gradient EM updates: On the other hand, the sample-based and population gradient EM operators with step size $\alpha > 0$ are given by

$$G_n(\theta) = \theta + \alpha \left\{ \frac{1}{n} \sum_{i=1}^{n} (2w_\theta(y_i) - 1)y_i - \theta \right\}, \quad \text{and} \quad G(\theta) = \theta + \alpha \big[2\mathbb{E}\big[w_\theta(Y)Y\big] - \theta \big]. \quad (14)$$

We return to analyze the EM updates for the Gaussian mixture model in Section 4.1.

2.2.2 Mixture of regressions

We now consider the mixture of regressions model, as has been analyzed in some recent work [12, 13, 52]. 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

$$y_i = \langle x_i, \, \theta^* \rangle + v_i, \quad (15)$$

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 (15) 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 (15), and $z_i = 0$ otherwise. In this symmetric form, the mixture of regressions model is closely related to models for phase retrieval, albeit over $\mathbb{R}^d$, as considered in a line of recent work (e.g., [3, 9, 35]).

EM updates: Define the weight function

$$w_\theta(x, y) = \frac{\exp\left(\frac{-(y - \langle x, \theta \rangle)^2}{2\sigma^2}\right)}{\exp\left(\frac{-(y - \langle x, \theta \rangle)^2}{2\sigma^2}\right) + \exp\left(\frac{-(y + \langle x, \theta \rangle)^2}{2\sigma^2}\right)}. \quad (16a)$$

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

$$Q(\theta'|\theta) = -\frac{1}{2n} \sum_{i=1}^{n} \left(w_\theta(x_i, y_i)(y_i - \langle x_i, \theta' \rangle)^2 + (1 - w_\theta(x_i, y_i))(y_i + \langle x_i, \theta' \rangle)^2 \right). \quad (16b)$$

Again, there is a closed form solution to this maximization problem: more precisely, the sample EM operator $M_n : \mathbb{R}^d \to \mathbb{R}^d$ takes the form

$$M_n(\theta) = \left(\sum_{i=1}^{n} x_i x_i^T \right)^{-1} \left(\sum_{i=1}^{n} (2w_\theta(x_i, y_i) - 1)y_i x_i \right). \quad (17a)$$

Similarly, by an easy calculation, we find that the population EM operator $M : \mathbb{R}^d \to \mathbb{R}^d$ has the form

$$M(\theta) = 2\mathbb{E}\big[w_\theta(X, Y)YX\big], \quad (17b)$$

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

6

Gradient EM updates: On the other hand, the gradient EM operators are given by

$$G_n(\theta) = \theta + \alpha \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[(2w_\theta(x_i, y_i) - 1)y_i x_i - x_i x_i^T \theta \right] \right\}, \quad \text{and} \tag{18a}$$

$$G(\theta) = \theta + \alpha \, 2\mathbb{E}\left[w_\theta(X, Y) \, YX - \theta \right], \tag{18b}$$

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

We return to analyze the EM updates for the mixture of regressions model in Section 4.2.

2.2.3 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.

As introduced in Section 2.2.2, 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 (15). 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

$$\tilde{x}_{ij} = \begin{cases} x_{ij} & \text{with probability } 1 - \rho \\ * & \text{with probability } \rho, \end{cases} \tag{19}$$

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

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

$$\mu_\theta(x_{\text{obs}}, y) := \begin{bmatrix} \mathbb{E}(x_{\text{mis}} | x_{\text{obs}}, y, \theta) \\ x_{\text{obs}} \end{bmatrix} = \begin{bmatrix} U_\theta z_{\text{obs}} \\ x_{\text{obs}} \end{bmatrix}, \tag{20a}$$

where

$$U_\theta = \frac{1}{\|\theta_{\text{mis}}\|_2^2 + \sigma^2} \begin{bmatrix} -\theta_{\text{mis}}\, \theta_{\text{obs}}^T & \theta_{\text{mis}} \end{bmatrix} \quad \text{and} \quad z_{\text{obs}} := \begin{bmatrix} x_{\text{obs}} \\ y \end{bmatrix} \in \mathbb{R}^{|x_{\text{obs}}|+1}. \tag{20b}$$

Similarly, the conditional second moment matrix takes the form

$$\Sigma_\theta(x_{\text{obs}}, y) := \mathbb{E}\left[XX^T \mid x_{\text{obs}}, y, \theta \right] = \begin{bmatrix} I & U_\theta z_{\text{obs}} x_{\text{obs}}^T \\ x_{\text{obs}} z_{\text{obs}}^T U_\theta^T & x_{\text{obs}} x_{\text{obs}}^T \end{bmatrix}. \tag{20c}$$

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

We now have the necessary notation in place to describe the EM and gradient EM updates. For a given parameter θ, the EM update is based on maximizing

$$Q_n(\theta'|\theta) := -\frac{1}{2n}\sum_{i=1}^{n}\langle\theta',\, \Sigma_\theta(x_{\text{obs},i}, y_i)\theta'\rangle + \frac{1}{n}\sum_{i=1}^{n} y_i\langle\mu_\theta(x_{\text{obs},i}, y_i),\, \theta'\rangle. \tag{21}$$

Again, this optimization problem has an explicit solution, so that the sample-based EM operator is given by

$$M_n(\theta) := \Big[\sum_{i=1}^{n}\Sigma_\theta(x_{\text{obs},i}, y_i)\Big]^{-1}\Big[\sum_{i=1}^{n} y_i\mu_\theta(x_{\text{obs},i}, y_i)\Big], \tag{22a}$$

accompanied by its population counterpart

$$M(\theta) := \big\{\mathbb{E}\big[\Sigma_\theta(X_{\text{obs}}, Y)\big]\big\}^{-1}\,\mathbb{E}\big[Y\mu_\theta(X_{\text{obs}}, Y)\big]. \tag{22b}$$

On the other hand, the gradient EM algorithm with step size α takes the form

$$G_n(\theta) = \theta + \alpha\Big\{\frac{1}{n}\sum_{i=1}^{n}\big[y_i\mu_\theta(x_{\text{obs},i}, y_i) - \Sigma_\theta(x_{\text{obs},i}, y_i)\theta\big]\Big\}, \tag{23a}$$

along with the population counterpart

$$G(\theta) = \theta + \alpha\mathbb{E}\Big[Y\mu_\theta(X_{\text{obs}}, Y) - \Sigma_\theta(X_{\text{obs}}, Y)\theta\Big], \tag{23b}$$

We return to analyze the gradient EM updates for this model in Section 4.3.

3 General convergence results

We now turn to analysis of the EM algorithm and gradient EM algorithms. In both cases, we let θ^* denote a maximizer of the population likelihood. In this section, we give general sufficient conditions under which the population algorithms converge to θ^* and under which the sample-based algorithms converge to an ε-ball around θ^*. Our analysis of each algorithm is organized as follows:

Our first result in Sections 3.1 and 3.2 concern the population EM and gradient EM operators respectively. Theorems 1 and 3 give conditions under which the population operators are contractive on a ball containing the fixed point θ^*, say $\mathbb{B}_2(r;\theta^*) = \{\theta\in\Omega \mid \|\theta-\theta^*\|_2 \leq r\}$ for some radius r. This population-level analysis is developed by viewing the population operators as perturbed versions of *oracle* operators which are known to be contractive around θ^*. Our conditions which relate the population EM and gradient EM operators to the oracle operators are then verified in concrete examples in Section 4. The analysis here is entirely deterministic.

Our second result in Sections 3.1 and 3.2 concern the sample-based EM and gradient EM operators. These sample-based operators approximate the population-based update using a subset of samples at each step. Theorem 2 and Theorem 4 for sample-based EM and gradient EM, respectively give conditions under which the sample-based operator is guaranteed to converge to an ε-ball around the fixed point θ^*. These results involve probabilistic bounds on the deviations between the population-based and sample-based operators. In addition, for gradient EM, we also analyze a stochastic update that uses a single sample per update in the flavor of stochastic approximation algorithms (see Theorem 5 in Section 3.2.3).

3.1 Analysis of EM algorithm

Let us begin with analysis of the standard EM updates, starting with the population version before turning to a sample-based version.

3.1.1 Guarantees for population-level EM

Recall that we always assume that the vector θ^* maximizes the population likelihood. It is a classical fact [29] that it must then satisfy the condition

$$\theta^* = \arg\max_{\theta \in \Omega} Q(\theta | \theta^*), \tag{24}$$

a property known as *self-consistency*. For this reason, the function $q(\cdot) := Q(\cdot | \theta^*)$ plays an important role in our analysis.

We assume throughout this section that the function q is λ-*strongly concave*, meaning that

$$q(\theta_1) - q(\theta_2) - \langle \nabla q(\theta_2),\, \theta_1 - \theta_2 \rangle \leq -\frac{\lambda}{2}\|\theta_1 - \theta_2\|_2^2, \tag{25}$$

for all pairs (θ_1, θ_2) in a neighborhood of θ^*. As we will illustrate, this condition holds in most concrete instantiations of EM, including the three model classes introduced in the previous section.

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 update $M(\theta)$ respectively. Indeed, by virtue of the self-consistency property (24) and the convexity of Ω, the fixed point satisfies the first-order optimality condition

$$\langle \nabla Q(\theta^* | \theta^*),\, \theta' - \theta^* \rangle \leq 0 \qquad \text{for all } \theta' \in \Omega. \tag{26}$$

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

$$\langle \nabla Q(M(\theta) | \theta),\, \theta' - M(\theta) \rangle \leq 0 \qquad \text{for all } \theta' \in \Omega. \tag{27}$$

Equations (26) and (27) are sets of inequalities that *characterize* the points $M(\theta)$ and θ^*. Thus, at an intuitive level, in order to establish that $M(\theta)$ and θ^* are close, it suffices to verify that these two characterizations are close in a suitable sense. We also note that inequalities similar to the condition (27) 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 (24) that characterizes θ^*.

With this intuition in mind, we introduce the following regularity condition in order to relate conditions (27) and (24): The condition involves a Euclidean ball of radius r around the fixed point θ^*, given by

$$\mathbb{B}_2(r; \theta^*) := \big\{ \theta \in \Omega \mid \|\theta - \theta^*\|_2 \leq r \big\}. \tag{28}$$

Definition 1 (First-order Stability (FOS)). The functions $\{Q(\cdot|\theta), \theta \in \Omega\}$ satisfy condition FOS (γ) over $\mathbb{B}_2(r; \theta^*)$ if

$$\|\nabla Q(M(\theta) | \theta^*) - \nabla Q(M(\theta) | \theta)\|_2 \leq \gamma \|\theta - \theta^*\|_2 \qquad \text{for all } \theta \in \mathbb{B}_2(r; \theta^*). \tag{29}$$

To provide some high-level intuition, observe the condition (29) 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 a concrete example, recall the Gaussian mixture model first introduced in Section 2.2.1. For this model, the condition (29) is equivalent to

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

where w_θ was previously defined following equation (12). Given that the function $\theta \mapsto w_\theta(y)$ is smooth in θ, provided that γ is not too small, it is reasonable to expect that this condition will hold in a neighborhood of θ^*, and we confirm this intuition in Corollary 1 to follow.

Under the conditions we have introduced, the following result guarantees that the population EM operator is locally contractive:

Theorem 1. *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 (25), and that the FOS(γ) condition (29) 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*

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

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.

$$\|\theta^t - \theta^*\|_2 \leq \left(\frac{\gamma}{\lambda}\right)^t \|\theta^0 - \theta^*\|_2 \qquad \text{for all } t = 1, 2, \ldots. \tag{30}$$

Proof. Since both $M(\theta)$ and θ^* are in Ω, we may apply condition (26) with $\theta' = M(\theta)$ and condition (27) with $\theta' = \theta^*$. Doing so, adding the resulting inequalities and then performing some algebra yields the condition

$$\langle \nabla Q(M(\theta)|\theta^*) - \nabla Q(\theta^*|\theta^*), \theta^* - M(\theta)\rangle \leq \langle \nabla Q(M(\theta)|\theta^*) - \nabla Q(M(\theta)|\theta), \theta^* - M(\theta)\rangle. \tag{31}$$

Now the λ-strong concavity condition (25) implies that the left-hand side is lower bounded as

$$\langle \nabla Q(M(\theta)|\theta^*) - \nabla Q(\theta^*|\theta^*), \theta^* - M(\theta)\rangle \geq \lambda \|\theta^* - M(\theta)\|_2^2. \tag{32a}$$

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

$$\langle \nabla Q(M(\theta)|\theta^*) - \nabla Q(M(\theta)|\theta), \theta^* - M(\theta)\rangle \leq \gamma \|\theta^* - M(\theta)\|_2 \|\theta - \theta^*\|_2, \tag{32b}$$

Combining inequalities (32a) and (32b) with the original bound (31) yields

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

and canceling terms completes the proof. $\qquad\square$

3.1.2 Guarantees for sample-based EM

We now turn to theoretical results on sample-based versions of the EM algorithm. More specifically, we consider two forms of the EM algorithm, the first being the standard form in which the operator $M_n : \Omega \to \Omega$, as previously defined (9), is applied repeatedly, thereby generating the sequence $\theta^{t+1} = M_n(\theta^t)$. We also analyze a sample-splitting [2] 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

$$\|M_n(\theta) - M(\theta)\|_2 \leq \varepsilon_M(n,\delta) \tag{33}$$

with probability at least $1-\delta$. This tolerance parameter (33) 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 (33) holds uniformly over the ball $\mathbb{B}_2(r;\theta^*)$. Accordingly, we let $\varepsilon_M^{\mathrm{unif}}(n,\delta)$ be the smallest scalar for which

$$\sup_{\theta \in \mathbb{B}_2(r;\theta^*)} \|M_n(\theta) - M(\theta)\|_2 \leq \varepsilon_M^{\mathrm{unif}}(n,\delta) \tag{34}$$

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

Theorem 2. *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

$$\varepsilon_M^{\mathrm{unif}}(n,\delta) \leq (1-\kappa)r, \tag{35a}$$

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

$$\|\theta^t - \theta^*\|_2 \leq \kappa^t \|\theta^0 - \theta^*\|_2 + \frac{1}{1-\kappa}\, \varepsilon_M^{\mathrm{unif}}(n,\delta) \tag{35b}$$

with probability at least $1 - \delta$.

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

$$\varepsilon_M\left(\frac{n}{T}, \frac{\delta}{T}\right) \leq (1-\kappa)r. \tag{36a}$$

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

$$\|\theta^t - \theta^*\|_2 \leq \kappa^t \|\theta^0 - \theta^*\|_2 + \frac{1}{1-\kappa}\, \varepsilon_M\left(\frac{n}{T}, \frac{\delta}{T}\right). \tag{36b}$$

---

[2]From a practical point, a potential advantage of sample splitting is that each iteration may be cheaper, since it is based on a smaller sample size. In contrast, a disadvantage is that it can be difficult to correctly specify the number of iterations in advance.

Figure 1 provides an illustration of the behavior predicted by Theorem 2: both algorithms are expected to show geometric convergence to the target parameter θ^*, up to some tolerance. For the bound (35b) note that the first term is decreasing in t, whereas the second term is independent of t. Thus, for a fixed sample size n, the bounds in Theorem 2 suggests a reasonable choice of the number of iterations. In particular, focusing on the standard EM algorithm, consider any positive integer[3] T such that

$$T \geq \log_{1/\kappa} \frac{(1-\kappa)\,\|\theta^0 - \theta^*\|_2}{\varepsilon_M^{\mathrm{unif}}(n,\delta)}. \tag{37}$$

This choice ensures that the first term in the bound (35b) is dominated by the second term, and hence that

$$\|\theta^T - \theta^*\|_2 \leq \frac{2}{1-\kappa}\,\varepsilon_M^{\mathrm{unif}}(n,\delta), \tag{38}$$

with probability at least $1 - \delta$. For the sample-splitting update in (36b) the first term is decreasing in t, whereas the second term is increasing in t. In this case, a similar conclusion holds when T is chosen to be the smallest positive integer such that

$$T \geq \log_{1/\kappa} \frac{(1-\kappa)\,\|\theta^0 - \theta^*\|_2}{\varepsilon_M\left(\frac{n}{T}, \frac{\delta}{T}\right)}. \tag{39}$$

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 $\varepsilon_M^{\mathrm{unif}}$.
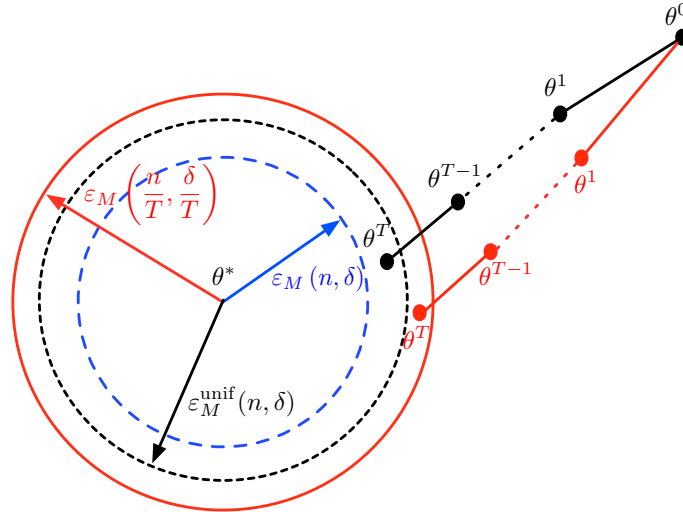

Figure 1. An illustration of Theorem 2. The first part of the theorem describes the geometric convergence of iterates of the EM algorithm to the ball of radius $\mathcal{O}(\varepsilon_M^{\mathrm{unif}}(n,\delta))$ (in black). The second part describes the geometric convergence of the sample-splitting EM algorithm to the ball of radius $\mathcal{O}(\varepsilon_M(n/T, \delta/T))$ (in red). In typical examples the ball to which sample-splitting EM converges is only a logarithmic factor larger than the ball $\mathcal{O}(\varepsilon_M(n,\delta))$ (in blue).

Let us now turn to the proof of the theorem.

---

[3] As will be clarified in the sequel, such a choice of T exists in various concrete models considered here.

Proof. We give a detailed proof of the claim (36b), from which it will be clear that the claim (35b) follows by a nearly identical argument. For any iteration $s \in \{1, 2, \ldots, T\}$, we have

$$\|M_{n/T}(\theta^s) - M(\theta^s)\|_2 \leq \varepsilon_M\left(\frac{n}{T}, \frac{\delta}{T}\right) \tag{40}$$

with probability at least $1 - \frac{\delta}{T}$. Consequently, by a union bound over all T indices, the bound (40) holds uniformly with probability at least $1 - \delta$. We perform the remainder of our analysis under this event.

It suffices to show that

$$\|\theta^{s+1} - \theta^*\|_2 \leq \kappa\|\theta^s - \theta^*\|_2 + \varepsilon_M\left(\frac{n}{T}, \frac{\delta}{T}\right) \qquad \text{for each iteration } s \in \{1, 2, \ldots, T-1\}. \tag{41}$$

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

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

where the final step follows by summing the geometric series.

It remains to prove the claim (41), and we do so via induction on the iteration number. Beginning with $s = 1$, we have

$$\begin{aligned}
\|\theta^1 - \theta^*\|_2 &= \|M_{n/T}(\theta^0) - \theta^*\|_2 \overset{(i)}{\leq} \|M(\theta^0) - \theta^*\|_2 + \|M_{n/T}(\theta^0) - M(\theta^0)\|_2 \\
&\overset{(ii)}{\leq} \kappa\|\theta^0 - \theta^*\|_2 + \varepsilon_M\left(\frac{n}{T}, \frac{\delta}{T}\right),
\end{aligned}$$

where step (i) follows by triangle inequality, whereas step (ii) follows from the bound (40), and the contractivity of the population operator applied to $\theta^0 \in \mathbb{B}_2(r; \theta^*)$. By our initialization condition and the bound (36a), 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 (41) holds at iteration s. The same argument then implies that the bound (41) also holds for iteration $s + 1$, and that $\|\theta^{s+1} - \theta^*\|_2 \leq r$, thus completing the proof. $\square$

3.2 Analysis of gradient EM algorithm

We now turn to analysis of the gradient EM algorithm. As before, we separate our analysis into two parts, the first (Theorem 3) addressing the behavior of the population-level operator, and the second (Theorems 4 and 5) providing guarantees for sample-based updates.

3.2.1 Guarantees for population-level gradient EM

Recall that the gradient EM algorithm generates a sequence of iterates $\{\theta^t\}_{t=0}^{\infty}$ via the recursion $\theta^{t+1} = G(\theta^t)$, where

$$G(\theta) := \theta + \alpha\nabla Q(\theta|\theta). \tag{42}$$

Here $\alpha > 0$ is a step size parameter to be chosen. For analyzing gradient EM, we also require an additional condition on the function $q(\theta) = Q(\theta|\theta^*)$, previously defined in Section 3.1. In addition to the λ-strong concavity assumption (25), we also assume that q is μ-smooth, meaning that

$$q(\theta_1) - q(\theta_2) - \langle \nabla q(\theta_2),\, \theta_1 - \theta_2 \rangle \geq -\frac{\mu}{2}\|\theta_1 - \theta_2\|_2^2, \tag{43}$$

for all pairs (θ_1, θ_2).

In order to gain intuition into the gradient EM algorithm, it is instructive to compare its iterates with those of standard gradient ascent on the function q. Gradient ascent on q performs the updates $\tilde{\theta}^{t+1} = T(\tilde{\theta}^t)$, where

$$T(\theta) := \theta + \alpha \nabla q(\theta). \tag{44}$$

Under the stated strong concavity and smoothness assumptions, it is a standard result from optimization theory [6, 7, 34] that the gradient operator $T : \Omega \to \Omega$ with step size choice $\alpha = \frac{2}{\mu+\lambda}$ is contractive, in particular with

$$\|T(\theta) - \theta^*\|_2 \leq \left(\frac{\mu - \lambda}{\mu + \lambda}\right)\|\theta - \theta^*\|_2 \qquad \text{for all } \theta \in \mathbb{B}_2(r; \theta^*). \tag{45}$$

Intuitively, then, if the function $Q(\cdot|\theta)$ is "close enough" to the function $q(\cdot) = Q(\cdot|\theta^*)$, then the gradient EM operator might be expected to satisfy a similar contractivity condition. The closeness requirement is formalized in the following condition:

Definition 2 (Gradient Stability (GS))**.** The functions $\{Q(\cdot|\theta), \theta \in \Omega\}$ satisfy condition GS (γ) over $\mathbb{B}_2(r; \theta^*)$ if

$$\|\nabla Q(\theta|\theta^*) - \nabla Q(\theta|\theta)\|_2 \leq \gamma\|\theta - \theta^*\|_2 \qquad \text{for all } \theta \in \mathbb{B}_2(r; \theta^*). \tag{46}$$

See Figure 2 for an illustration of this condition. We give concrete examples of this condition and its verification in Section 4. As with the FOS condition observe that the GS condition is always satisfied at the fixed point θ^*, i.e. for $r = 0$ with $\gamma = 0$. Allowing for strictly positive γ, if the functions $Q(\cdot|\theta)$ are sufficiently regular we expect the condition to hold in a region around θ^*. Observe that this condition involves the gradient of the functions $Q(\cdot|\theta)$ and $Q(\cdot|\theta^*)$ at θ, as opposed to $M(\theta)$ in the case of the FOS condition. For this reason, it can be easier to verify for specific models.

Under this condition, the following result guarantees local contractivity of the gradient EM operator (42):

Theorem 3. *For some radius $r > 0$, and a triplet (γ, λ, μ) such that $0 \leq \gamma < \lambda \leq \mu$, suppose that the function $q(\theta) = Q(\theta|\theta^*)$ is λ-strongly concave (25), μ-smooth (43), and that the $GS(\gamma)$ condition (46) holds on the ball $\mathbb{B}_2(r; \theta^*)$. Then the population gradient EM operator G with step size $\alpha = \frac{2}{\mu+\lambda}$ is contractive over $\mathbb{B}_2(r; \theta^*)$, in particular with*

$$\|G(\theta) - \theta^*\|_2 \leq \left(1 - \frac{2\lambda - 2\gamma}{\mu + \lambda}\right)\|\theta - \theta^*\|_2 \qquad \text{for all } \theta \in \mathbb{B}_2(r; \theta^*). \tag{47}$$
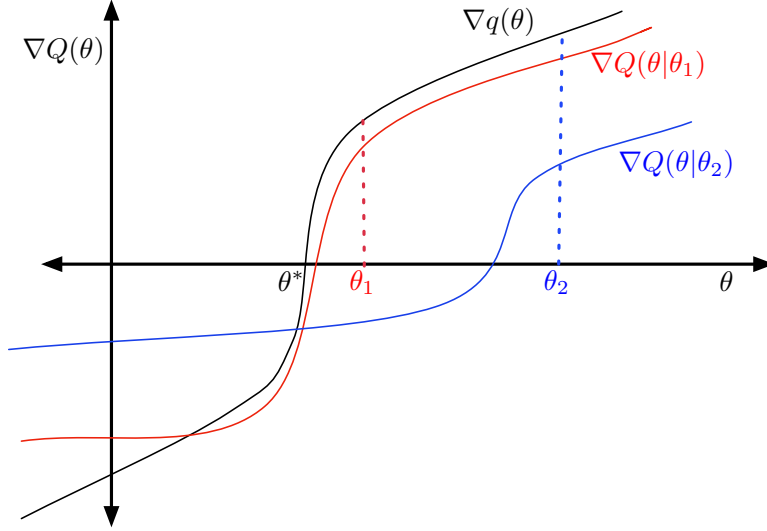
14

Figure 2. Illustration of the gradient stability condition (46): for a point θ_1 close to the population optimum θ^*, the gradients $\nabla Q(\theta_1|\theta_1)$ and $\nabla q(\theta_1)$ must be close, whereas for a point θ_2 distant from θ^*, the gradients $\nabla Q(\theta_2|\theta_2)$ and $\nabla q(\theta_2)$ can be quite different.

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

$$\|\theta^t - \theta^*\|_2 \leq \left(1 - \frac{2\lambda - 2\gamma}{\mu + \lambda}\right)^t \|\theta^0 - \theta^*\|_2 \qquad \text{for all } t = 1, 2, \ldots. \tag{48}$$

Proof. By definition of the gradient EM update (42), we have

$$\|G(\theta) - \theta^*\|_2 = \|\theta + \alpha \nabla Q(\theta|\theta) - \theta^*\|_2$$

$$\overset{(i)}{\leq} \underbrace{\|\theta + \alpha \nabla Q(\theta|\theta^*) - \theta^*\|_2}_{\|T(\theta) - \theta^*\|_2} + \alpha \|\nabla Q(\theta|\theta) - \nabla Q(\theta|\theta^*)\|_2$$

$$\overset{(ii)}{\leq} \left(\frac{\mu - \lambda}{\mu + \lambda}\right) \|\theta - \theta^*\|_2 + \alpha \gamma \|\theta - \theta^*\|_2.$$

where step (i) follows from the triangle inequality, and step (ii) uses the contractivity of T from equation (45), and condition GS. Substituting $\alpha = \frac{2}{\mu + \lambda}$ and performing some algebra yields the claim. $\qquad\square$

3.2.2 Guarantees for sample-based gradient EM

In this section, in parallel with our earlier analysis of sample-based version of the EM algorithm, we analyze two sample-based variants of the gradient EM algorithm, the first when the update operator G_n is computed using all n samples and applied repeatedly, and the second based on sample-splitting.

We begin by introducing quantities that measure the deviations of the sample operator G_n from the population version G. For a given sample size n and tolerance parameter $\delta \in (0, 1)$,

15

we let $\varepsilon_G(n, \delta)$ be the smallest scalar such that, for any fixed vector $\theta \in \mathbb{B}_2(r; \theta^*)$,

$$\|G_n(\theta) - G(\theta)\|_2 \leq \varepsilon_G(n, \delta) \tag{49}$$

with probability at least $1 - \delta$. The uniform analogue of this deviation is defined similarly: the quantity $\varepsilon_G^{\text{unif}}(n, \delta)$ is the smallest scalar for which

$$\sup_{\theta \in \mathbb{B}_2(r; \theta^*)} \|G_n(\theta) - G(\theta)\|_2 \leq \varepsilon_G^{\text{unif}}(n, \delta) \tag{50}$$

with probability at least $1 - \delta$.

Theorem 4. *Suppose that the population gradient EM operator $G : \Omega \to \Omega$ 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

$$\varepsilon_G^{\text{unif}}(n, \delta) \leq (1 - \kappa)r, \tag{51a}$$

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

$$\|\theta^t - \theta^*\|_2 \leq \kappa^t \|\theta^0 - \theta^*\|_2 + \frac{1}{1 - \kappa} \, \varepsilon_G^{\text{unif}}(n, \delta) \tag{51b}$$

with probability at least $1 - \delta$.

(b) If the sample size n is large enough to ensure that

$$\varepsilon_G\left(\frac{n}{T}, \frac{\delta}{T}\right) \leq (1 - \kappa)r, \tag{52a}$$

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

$$\|\theta^t - \theta^*\|_2 \leq \kappa^t \|\theta^0 - \theta^*\|_2 + \frac{1}{1 - \kappa} \, \varepsilon_G\left(\frac{n}{T}, \frac{\delta}{T}\right) \tag{52b}$$

with probability at least $1 - \delta$.

Note that the guarantees (51b) and (52b) are identical to the earlier bounds (35b) and (36b) from Theorem 2, modulo the replacements of $(\varepsilon_M, \varepsilon_M^{\text{unif}})$ by $(\varepsilon_G, \varepsilon_G^{\text{unif}})$. We omit the proofs, since they follow from essentially the same argument as Theorem 2. Thus, in order to obtain interpretable bounds for gradient EM applied to specific models, it only remains to establish the κ-contractivity of the population operator, and to compute the functions ε_G or $\varepsilon_G^{\text{unif}}$.

3.2.3 Stochastic version of gradient EM

In this section, we analyze a sample-based variant of gradient EM that is inspired by stochastic approximation. It can be viewed as an extreme form of sample-splitting, in which we use only a single sample per iteration, but compensate for the noisiness using a decaying step size. Throughout this section we assume that (a lower bound on) the radius of convergence r of the population operator is known to the algorithm[4].

---

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

In particular, given a sequence of positive step sizes $\{\alpha^t\}_{t=0}^\infty$, we analyze the recursion

$$\theta^{t+1} = \Pi\Big(\theta^t + \alpha^t \nabla Q_1(\theta^t|\theta^t)\Big), \tag{53}$$

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

Theorem 5. *For a triplet (γ, λ, μ) such that $0 \leq \gamma < \lambda \leq \mu$, suppose that the population function q is λ-strongly concave* (25), *μ-smooth* (43), *and satisfies the GS(γ) condition* (46) *over the ball $\mathbb{B}_2(r;\theta^*)$. Then given an initialization $\theta^0 \in \mathbb{B}_2(\frac{r}{2};\theta^*)$, the stochastic EM gradient updates* (53) *with step size $\alpha^t := \frac{3}{2\xi\,(t+2)}$ satisfy the bound*

$$\mathbb{E}[\|\theta^t - \theta^*\|_2^2] \leq \frac{9\sigma_G^2}{\xi^2}\frac{1}{(t+2)} + \Big(\frac{2}{t+2}\Big)^{3/2}\|\theta^0 - \theta^*\|_2^2 \qquad \text{for iterations } t = 1, 2, \ldots . \tag{54}$$

While the stated claim (54) provides bounds in expectation, it is also possible to obtain high-probability results.[5]

Proof. In order to prove this theorem we first establish a recursion on the expected mean-squared error. As with Theorem 3 this result is established by relating the population gradient EM operator to the gradient ascent operator on the function $q(\cdot)$. This key recursion along with some algebra will yield the theorem.

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

$$\mathbb{E}[\|\Delta^{t+1}\|_2^2] \leq \Big\{1 - \alpha^t\xi\Big\}\mathbb{E}[\|\Delta^t\|_2^2] + (\alpha^t)^2\sigma_G^2, \tag{55}$$

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

We prove this lemma in Appendix A.

Using this result, we can now complete the proof of the bound (54). With the step size choice $\alpha^t := \frac{a}{\xi\,(t+2)}$ where $a = \frac{3}{2}$, unwrapping the recursion (55) yields

$$\mathbb{E}[\|\Delta^{t+1}\|_2^2] \leq \frac{a^2\sigma_G^2}{\xi^2}\sum_{\tau=2}^{t+1}\Big\{\frac{1}{\tau^2}\prod_{\ell=\tau+1}^{t+2}\Big(1 - \frac{a}{\ell}\Big)\Big\} + \frac{a^2\sigma_G^2}{\xi^2(t+2)^2} + \prod_{\ell=2}^{t+2}\Big(1 - \frac{a}{\ell}\Big)\mathbb{E}[\|\Delta^0\|_2^2]. \tag{56}$$

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

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

---

[5]Although we do not consider this extension here, stronger exponential concentration results follow from controlling the moment generating function of the random variable $\sup_{\theta \in \mathbb{B}_2(r;\theta^*)}\|\nabla Q_1(\theta|\theta)\|_2^2$. For instance, see Nemirovski et al. [33] for such results in the context of stochastic optimization.

See Noorshams and Wainright [36] for a proof. Using this fact in Equation (56) yields

$$\mathbb{E}[\|\Delta^{t+1}\|_2^2] \leq \frac{a^2\sigma_{\text{G}}^2}{\xi^2\,(t+3)^a} \sum_{\tau=2}^{t+2} \frac{(\tau+1)^a}{\tau^2} + \left(\frac{2}{t+3}\right)^a \mathbb{E}[\|\Delta^0\|_2^2]$$

$$\leq \frac{2a^2\sigma_{\text{G}}^2}{\xi^2\,(t+3)^a} \sum_{\tau=2}^{t+2} \frac{1}{\tau^{2-a}} + \left(\frac{2}{t+3}\right)^a \mathbb{E}[\|\Delta^0\|_2^2].$$

Finally, applying the integral upper bound $\sum_{\tau=2}^{t+2} \frac{1}{\tau^{2-a}} \leq \int_1^{t+2} \frac{1}{x^{2-a}}dx \leq 2(t+3)^{a-1}$ yields the claim (54). □

In order to obtain guarantees for stochastic gradient EM applied to specific models, it only remains to prove the concavity and smoothness properties of the population function q, and to bound the uniform variance σ_{G}.

A summary: For the convenience of the reader, let us now summarize the theorems given in this section, including the assumptions on which they rely and the results that they provide.

Condition	Result	Thm.
Strong concavity of q and FOS	Pop. contractivity of EM (R1)	Thm. 1
Bound on $\varepsilon_M^{\text{unif}}$ and (R1)	Fin.-sample bound for EM	Thm. 2
Bound on ε_M and (R1)	Fin.-sample bound for sample splitting EM	Thm. 2
Strong concavity, smoothness of q and GS	Pop. contractivity of grad. EM (R2)	Thm. 3
Bound on $\varepsilon_G^{\text{unif}}$ and (R2)	Fin.-sample bound for grad. EM	Thm. 4
Bound on ε_G and (R2)	Fin.-sample bound for sample splitting grad. EM	Thm. 4
Bound on σ_{G} and (R2)	Fin.-sample bound for stochastic gradient EM	Thm. 5

4   Consequences for specific models

In the previous section, we provided a number of general theorems on the behavior of the EM algorithm as well as the gradient EM algorithm, at both the population and sample levels. In this section, we develop some concrete consequences of this general theory for the three specific model classes previously introduced in Section 2.2.

4.1   Gaussian mixture models

We begin by analyzing the EM updates for the Gaussian mixture model previously introduced in Section 2.2.1. Our first result (Corollary 1) establishes contractivity for the population operator (13b), whereas our second result (Corollary 2) provides bounds for the sample-based EM updates.

Recall that our mixture model consists of two equally weighted components, with distributions $\mathcal{N}(\theta^*, \sigma^2 I)$ and $\mathcal{N}(-\theta^*, \sigma^2 I)$ respectively. The difficulty of estimating this mixture model can be characterized by the signal-to-noise ratio $\frac{\|\theta^*\|_2}{\sigma}$, and our analysis requires a lower bound of the form

$$\frac{\|\theta^*\|_2}{\sigma} > \eta, \tag{57}$$

for a sufficiently large constant $\eta > 0$. Past work by Redner and Walker [39] provides evidence for the necessity of this assumption: for Gaussian mixtures with low signal-to-noise ratio, 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, 50] 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 contractivity for Gaussian mixtures)**.** *Consider a Gaussian mixture model for which the SNR condition* (57) *holds for a sufficiently large η. Then there is a universal constant $c > 0$ such that the population EM operator* (13b) *is κ-contractive over the ball $\mathbb{B}_2(r; \theta^*)$ with*

$$r = \frac{\|\theta^*\|_2}{4}, \quad and \quad \kappa(\eta) \leq e^{-c\eta^2}. \tag{58}$$

This corollary guarantees that when the SNR is sufficiently large, then the MLE θ^* has a basin of attraction that is at least a constant fraction of the signal strength. Moreover, the convergence rate of the population updates is geometric, with the contraction factor κ decreasing exponentially in the signal-to-noise ratio. The proof of Corollary 1 involves establishing that for a sufficiently large SNR, the strong concavity and FOS (γ) conditions hold for a Gaussian mixture model, so that Theorem 1 can be applied. Although the proof structure is conceptually straightforward, the details are quite technical, so that we defer it to Appendix B.1.

Based on the population-level contractivity guaranteed by Corollary 1, we can also establish guarantees for the standard EM sequence $\theta^{t+1} = M_n(\theta^t)$, where the sample-based operator M_n was previously defined in equation (13a). This guarantee involves the function $\varphi(\sigma; \|\theta^*\|_2) := \|\theta^*\|_2 \sqrt{\|\theta^*\|_2^2 + \sigma^2}$, as well as positive universal constants (c, c_1, c_2).

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

$$\|\theta^t - \theta^*\|_2 \leq \kappa^t \|\theta^0 - \theta^*\|_2 + \frac{c_2}{1-\kappa} \varphi(\sigma; \|\theta^*\|_2) \sqrt{\frac{d}{n} \log(1/\delta)} \tag{59}$$

with probability at least $1 - \delta$.

See Appendix B.2 for the proof of this result. In Appendix B.3, we also give guarantees for EM with sample-splitting which achieves better dependence on $\|\theta^*\|_2$ and σ with an easier proof at the cost of additional logarithmic factors in sample complexity.

A related result of Dasgupta and Schulman [15] shows that when the SNR is sufficiently high 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 (59) provides a rough guide of how many iterations are required: consider the smallest positive integer such that

$$T \geq \log_{1/\kappa} \left(\frac{\|\theta^0 - \theta^*\|_2 (1-\kappa)}{\varphi(\sigma; \|\theta^*\|_2)} \sqrt{\frac{n}{d} \frac{1}{\log(1/\delta)}} \right). \tag{60a}$$
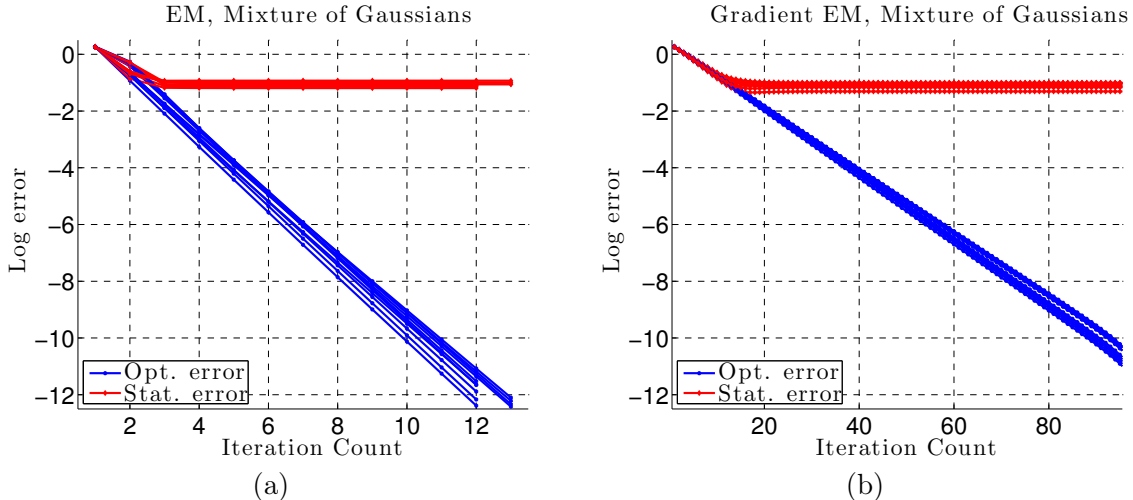
19

Figure 3. Plots of the iteration count 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[7]. (b) Results for the gradient 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.

With this choice, we are guaranteed that the iterate θ^T satisfies the bound

$$\|\theta^T - \theta^*\|_2 \le \frac{(1 + c_2)\varphi(\sigma; \|\theta^*\|_2)}{1 - \kappa} \sqrt{\frac{d}{n} \, \log(1/\delta)} \tag{60b}$$

with probability at least $1 - \delta$. Treating σ and $\|\theta^*\|_2$ as fixed there is no point in performing additional iterations, since by standard minimax results, any estimator of θ^* based on n samples must have ℓ_2-error of the order $\sqrt{\frac{d}{n}}$. Of course, the iteration choice (60a) 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 suggests that the iteration complexity should grow logarithmically in the ratio n/d.

Corollary 2 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 4 shows the results of simulations designed to test this prediction: for dimension $d = 10$ and sample size $n = 1000$, we performed 10 trials with the standard EM updates applied to Gaussian mixture models with SNR $\frac{\|\theta^*\|_2}{\sigma} = 2$. In panel (a), the red curves plot the log statistical error versus the iteration number, whereas the blue curves show the log optimization error versus iteration. As can be seen by the red curves, the statistical error decreases geometrically before leveling off at a plateau. On the other hand, the optimization error decreases geometrically to numerical tolerance. Panel (b) shows that the gradient 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 2 also predicts that the convergence rate should increase as the signal-to-noise ratio $\frac{\|\theta^*\|_2}{\sigma}$ is increased. Figure 4 shows the results of simulations designed to test this prediction: again, for mixture models with dimension $d = 10$ and sample

---

[7]In this and subsequent figures we show simulations for the standard (i.e. not sample-splitting) versions of the EM and gradient EM algorithms.
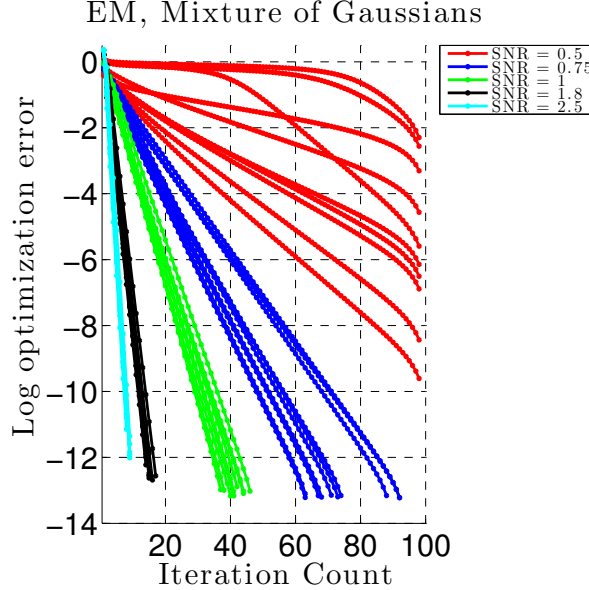
Figure 4. Plot of the iteration count versus the (log) optimization error $\log(\|\theta^t - \widehat{\theta}\|_2)$ for different values of the SNR $\frac{\|\theta^*\|_2}{\sigma}$. For each SNR, we performed 10 independent trials of a Gaussian mixture model with dimension $d = 10$ and sample size $n = 1000$. Larger values of SNR lead to faster convergence rates, consistent with Corollary 2.

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[8].

4.2 Mixtures of regressions

In this section, we analyze the EM and gradient EM algorithms for the mixture of regressions (MOR) model, previously introduced in Section 2.2.2. 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

$$\frac{\|\theta^*\|_2}{\sigma} > \eta \tag{61}$$

Under a suitable lower bound on this quantity, our first result guarantees that the population level operators (17b) and (18a) are locally contractive.

Corollary 3 (Population contractivity for MOR). *Consider any mixture of regressions model satisfying the SNR condition (61) for a sufficiently large constant η. Then the population EM operator M from equation (17b) and the population gradient EM operator G from equation (18a) are κ-contractive over the ball $\mathbb{B}_2(r; \theta^*)$ with*

$$r = \frac{\|\theta^*\|_2}{32}, \quad and \quad \kappa \leq \frac{1}{2}. \tag{62}$$

---

[8]To be clear, Corollary 2 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 2 can also be used to show geometric convergence of the optimization error.

As shown in the proof, the contraction coefficient κ is again a decreasing function of the SNR parameter η. However, its functional form is not as explicit as in the Gaussian mixture case. The proof of Corollary 3 involves verifying that the function q for the MOR model satisfies the required concavity, smoothness, GS(γ) and FOS(γ) conditions. It is quite technically involved, so that we defer it to Appendix C.1.

Let us now provide guarantees for a sample-splitting version of the EM updates. Recall that sample-based EM operator was previously defined in equation (17a). For a given sample size n and iteration number T, suppose that we split[9] our full data set into T subsets, each of size n/T. We then generate the sequence $\theta^{t+1} = M_{n/T}(\theta^t)$, where we use a fresh subset at each iteration. In the following result, we use $\varphi(\sigma; \|\theta^*\|_2) = \sqrt{\sigma^2 + \|\theta^*\|_2^2}$, along with positive universal constants (c_1, c_2).

Corollary 4 (Sample-splitting EM guarantees for MOR). *In addition to the conditions of Corollary 3, suppose that the sample size is lower bounded as $n \geq c_1 d \log(T/\delta)$. Then there is a contraction coefficient $\kappa \leq 1/2$ such that, for any initial vector $\theta^0 \in \mathbb{B}_2(\frac{\|\theta^*\|_2}{32}; \theta^*)$, the sample-splitting EM iterates $\{\theta^t\}_{t=1}^T$ based on n/T samples per step satisfy the bound*

$$\|\theta^t - \theta^*\|_2 \leq \kappa^t \|\theta^0 - \theta^*\|_2 + c_2 \varphi(\sigma; \|\theta^*\|_2) \sqrt{\frac{d}{n} \, T \log(T/\delta)} \tag{63}$$

with probability at least $1 - \delta$.

We prove this corollary in Appendix C.2. Note the bound (63) again provides guidance on the number of iterations to perform. For a given sample size n, suppose we perform $T = c \log(n/d\varphi^2(\sigma; \|\theta^*\|_2))$ iterations for a constant c. The bound (63) then implies that

$$\|\theta^T - \theta^*\|_2 \leq c_3 \varphi(\sigma; \|\theta^*\|_2) \sqrt{\frac{d}{n} \, \log^2\left(\frac{n}{d\varphi^2(\sigma; \|\theta^*\|_2)}\right) \, \log(1/\delta)} \tag{64}$$

with probability at least $1 - \delta$. Apart from the logarithmic penalty $\log^2\left(\frac{n}{d\varphi^2(\sigma; \|\theta^*\|_2)}\right)$, 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 EM updates, as we did for the Gaussian mixture model.

We conclude our discussion of the MOR model by stating a result for the stochastic form of gradient EM analyzed in Theorem 5. In particular, given a data set of size n, we run the algorithm for n iterations, with a step size $\alpha^t := \frac{3}{t+2}$ for iterations $t = 1, \ldots, n$. Once again our result is terms of $\varphi(\sigma; \|\theta^*\|_2) = \sqrt{\sigma^2 + \|\theta^*\|_2^2}$ and positive universal constants (c_1, c_2).

Corollary 5 (Stochastic gradient EM guarantees for MOR). *In addition to the conditions of Corollary 3, suppose that the sample size is lower bounded as $n \geq c_1 d \log(1/\delta)$. Then given any initialization $\theta^0 \in \mathbb{B}_2(\frac{\|\theta^*\|_2}{32}; \theta^*)$, performing n iterations of the stochastic gradient EM gradient updates (53) yields an estimate $\widehat{\theta} = \theta^n$ such that*

$$\mathbb{E}[\|\widehat{\theta} - \theta^*\|_2^2] \leq c_2 \, \varphi^2(\sigma; \|\theta^*\|_2) \, \frac{d}{n}. \tag{65}$$

We prove this corollary in Appendix C.3. Figure 5 illustrates this corollary showing the error as a function of iteration number (sample size) for the stochastic gradient EM algorithm.

---

[9]To simplify exposition, assume that n/T is an integer.
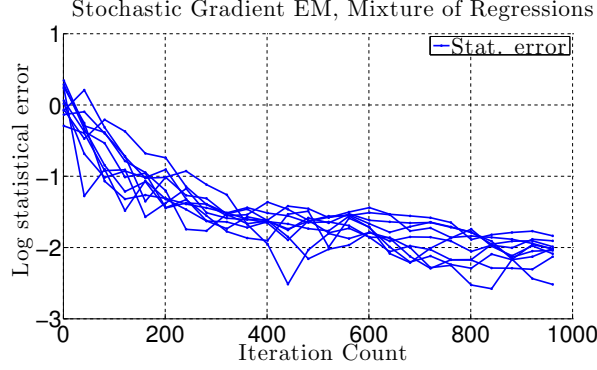
Figure 5. A plot of the (log) statistical error for the stochastic gradient EM algorithm as a function of iteration number (sample size) for the mixture of regressions example. The plot shows 10 different problem instances with $d = 10$, $\frac{\|\theta^*\|_2}{\sigma} = 2$ and $\frac{\|\theta^0 - \theta^*\|_2}{\sigma} = 1$. The statistical error decays at the sub-linear rate $\mathcal{O}(1/\sqrt{t})$ as a function of the iteration number t. An iteration of stochastic gradient EM is however typically much faster and uses only a single sample.

4.3   Linear regression with missing covariates

This section is devoted to analysis of the gradient EM algorithm for the problem of linear regression with missing covariates, as previously introduced in Section 2.2.3. Here the central parameter is the probability ρ 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. Define ξ_1 and ξ_2 to be such that the following bounds hold,

$$\frac{\|\theta^*\|_2}{\sigma} \leq \xi_1, \quad \text{and} \quad \|\theta - \theta^*\|_2 \leq r := \xi_2 \sigma. \tag{66}$$

For any given choice of (ξ_1, ξ_2) define $\xi := (\xi_1 + \xi_2)^2$. Our guarantees apply whenever the missing probability is bounded as

$$\rho < \frac{1}{1 + 2\xi(1 + \xi)}. \tag{67}$$

Corollary 6 (Population contractivity for missing covariates). *Given any missing covariate regression model with missing probability ρ satisfying the bound* (67), *the gradient EM operator* (22b) *is κ-contractive over the ball $\mathbb{B}_2(r; \theta^*)$ with*

$$r = \xi_2 \sigma, \quad \text{and} \quad \kappa = \frac{\xi + \rho(1 + 2\xi(1 + \xi))}{1 + \xi} < 1. \tag{68}$$

See Appendix D.1 for the proof of Corollary 6. Relative to our previous results, this corollary is somewhat unusual, in that we require an *upper bound* on the 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 not an artifact of our analysis nor of the gradient EM algorithm. Roughly these lower bounds formalize the intuition that as the norm $\|\theta^*\|_2$ increases, 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.
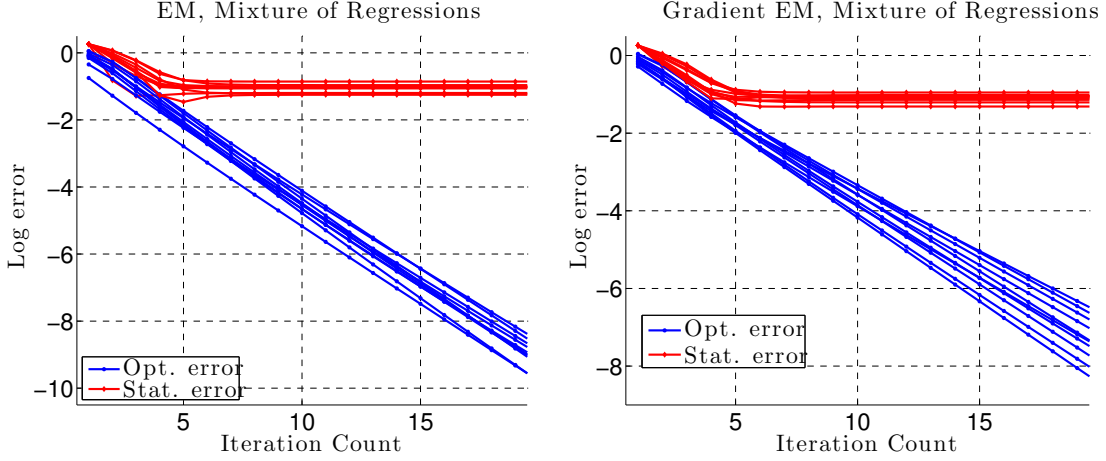
23

Figure 6. Plots of the iteration count 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 gradient EM algorithm. Each plot shows 10 different problem instances of dimension $d = 10$, 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.

Let us now provide guarantees for a sample-splitting version of the EM updates, based on the sample-based EM operator in equation (22a). As usual, for a given sample size n and iteration number T, suppose that we split our full data set into T subsets, each of size n/T. We then generate the sequence $\theta^{t+1} = M_{n/T}(\theta^t)$, where we use a fresh subset at each iteration.

Corollary 7 (Sample-splitting EM guarantees for missing covariates). *In addition to the conditions of Corollary 6, 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 EM iterates $\{\theta^t\}_{t=1}^T$ based on n/T samples per iteration satisfy the bound*

$$\|\theta^t - \theta^*\|_2 \leq \kappa^t \|\theta^0 - \theta^*\|_2 + \frac{c_2\sqrt{1+\sigma^2}}{1-\kappa}\sqrt{\frac{d}{n}\,T\log(T/\delta)} \tag{69}$$

with probability at least $1 - \delta$.

We prove this corollary in Appendix D.2. 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.

As with Corollary 4, 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 (69) 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.

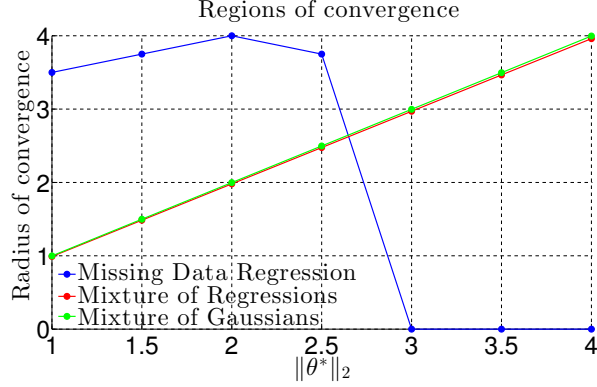We conclude our discussion of the missing covariates model by stating a result for the

24

Figure 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 6.

stochastic form of gradient EM analyzed in Theorem 5. In particular, given a data set of size n, we run the algorithm for n iterations, with a step size $\alpha^t := \frac{3}{t+2}$ for iterations $t = 1, \ldots, n$.

Corollary 8 (Stochastic gradient EM guarantees for missing covariates). *In addition to the conditions of Corollary 6, suppose that the sample size is lower bounded as $n \geq c_1 d \log(1/\delta)$. Then given any initialization $\theta^0 \in \mathbb{B}_2(\xi_2 \sigma; \theta^*)$, performing n iterations of the stochastic EM gradient updates (53) with step sizes $\alpha^t = \frac{3}{2(1-\kappa)(t+2)}$ yields an estimate $\widehat{\theta} = \theta^n$ such that*

$$\mathbb{E}[\|\widehat{\theta} - \theta^*\|_2^2] \leq c_2(1 + \sigma^2) \frac{d}{n}. \tag{70}$$

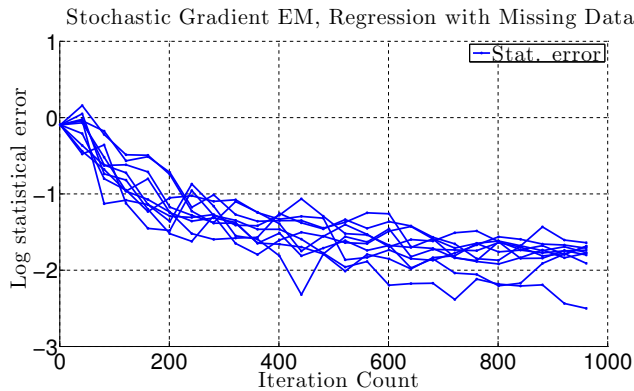We prove this corollary in Appendix D.3. Figure 8 illustrates this.

Figure 8. A plot of the (log) statistical error for the stochastic gradient EM algorithm as a function of iteration number (sample size) for the problem of linear regression with missing covariates. The plot shows 10 different problem instances with $d = 10$, $\frac{\|\theta^*\|_2}{\sigma} = 2$ and $\frac{\|\theta^0 - \theta^*\|_2}{\sigma} = 1$. The statistical error decays at the sub-linear rate $\mathcal{O}(1/\sqrt{t})$ as a function of the iteration number t.
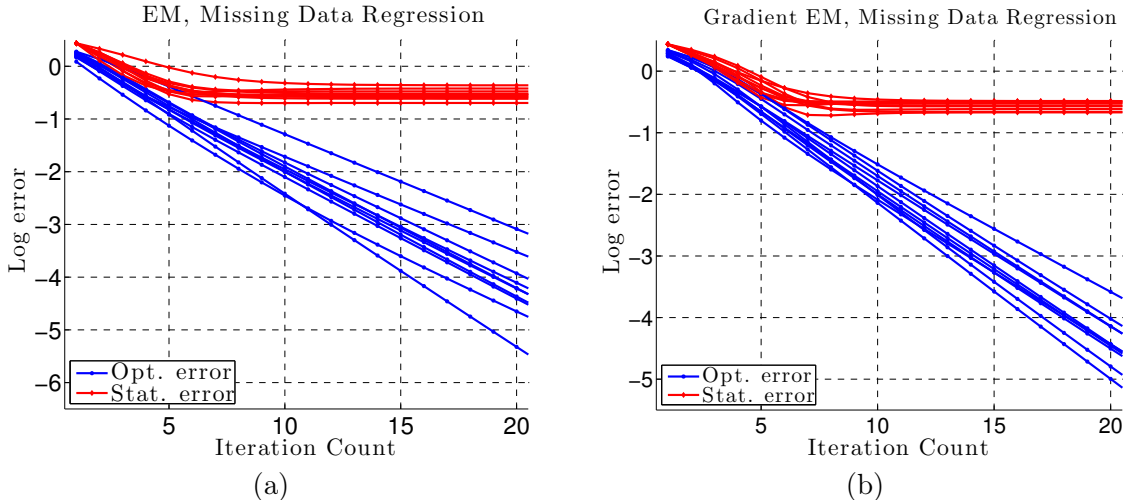
25

Figure 9. Plots of the iteration count 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 gradient 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.

5 Discussion

In this paper, we have provided some general techniques for studying the EM and gradient 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 gradient 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., [52]), and the plug-in principle for regression with missing data (e.g., [20, 51]). 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, 12, 19]) 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

A   Proofs for stochastic gradient EM

In this section we provide proofs of results related to Theorem 5 from Section 3.2.3. It only remains to prove Lemma 1.

In order to establish Lemma 1 we require an analogue of Theorem 3 that allows for a wider range of step sizes. Recall the classical gradient ascent operator on the function $q(\theta) = Q(\theta|\theta^*)$. For step size $\alpha > 0$, it takes the form $T(\theta) = \theta + \alpha \nabla q(\theta)$. Under the stated λ-concavity and μ-smoothness conditions, for any step size $0 < \alpha \leq \frac{2}{\lambda+\mu}$, the classical gradient operator T is contractive with parameter

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

This follows from the classical analysis of gradient descent (e.g., [6, 7, 34]). Using this fact, we can prove the following about the population gradient EM operator:

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

$$\xi := \frac{2\mu\lambda}{\lambda+\mu} - \gamma. \tag{71}$$

We omit the proof, since it follows from a similar argument to that of Theorem 3. With this preliminary in place we can now begin the proof of Lemma 1.

A.1   Proof of Lemma 1

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

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

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

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

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

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

Consequently, by iterated expectations, we have

$$\mathbb{E}[\|\Delta^{t+1}\|_2^2] \leq \mathbb{E}[\|\Delta^t\|_2^2] + (\alpha^t)^2 \mathbb{E}\|\widehat{W}(\theta^t)\|_2^2 + 2\alpha^t \mathbb{E}\Big[\langle W(\theta^t), \Delta^t \rangle\Big]. \tag{72}$$

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

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

Combining with our earlier inequality (72) yields

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

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

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

where step (i) uses the contractivity of G^t established in Lemma 2 and step (ii) uses the definition of ξ from equation (71). Putting together the pieces yields the claim (55).

B Proofs for Gaussian mixture models

In this section, we provide proofs of results related to the Gaussian mixture model, as presented in Section 4.1. More specifically, we first prove Corollary 1 on the population level behavior, followed by the proof of Corollary 2 on the behavior of the standard sample-based EM updates.

B.1 Proof of Corollary 1

In order to apply Theorem 1, we need to verify the λ-concavity condition (25), and the FOS(γ) condition (29) over the ball $\mathbb{B}_2(r; \theta^*)$. The population EM operator for the Gaussian mixture model was previously defined in equation (13b). The update $\theta \mapsto M(\theta)$ is based on maximizing the function

$$Q(\theta'|\theta) = -\frac{1}{2}\mathbb{E}\big[w_\theta(Y)\|Y - \theta'\|_2^2 + (1 - w_\theta(Y))\|Y + \theta'\|_2^2\big] \qquad \text{over } \theta' \in \mathbb{R}^d.$$

Here the weighting function takes the form

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

By inspection, the function $q(\theta') = Q(\theta'|\theta^*)$ is strongly concave on $\mathbb{R}^d$ with $\lambda = 1$.

It remains to verify the FOS(γ) condition (29). The following auxiliary lemma is central to the proof:

Lemma 3. *Under the conditions of Corollary 1, there is a constant $\gamma \in (0, 1)$ with $\gamma \leq \exp(-c_2\eta^2)$ such that*

$$\|\mathbb{E}\big[2\Delta_w(Y)Y\big]\|_2 \leq \gamma \|\theta - \theta^*\|_2, \tag{73}$$

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

Taking this result as given for the moment, let us now verify the FOS condition (29). By symmetry, we have $\mathbb{E}[w_\theta(Y)] = 1 - \mathbb{E}[w_\theta(Y)] = \frac{1}{2}$ for any $\theta \in \Omega$. Using this fact, it suffices to show that

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

This follows immediately from Lemma 3. Thus, the FOS 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.

Proof of Lemma 3: We now prove Lemma 3. Our proof makes use of the following elementary facts:

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

$$\sup_{t\in[0,\infty]} f(t) = \frac{4}{(e\,\mu)^2}, \qquad \text{achieved at } t^* = \frac{2}{\mu} \text{ and} \tag{74a}$$

$$\sup_{t\in[t^*,\infty]} f(t) = f(t^*), \qquad \text{for } t^* \geq \frac{2}{\mu}. \tag{74b}$$

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

$$g(t) \leq \frac{1}{4} \quad \text{for all } t \in \mathbb{R}, \text{ and} \tag{75a}$$

$$\sup_{t\in[\mu,\infty]} g(t) \leq \frac{1}{(\exp(\mu)+\exp(-\mu))^2} \leq \exp(-2\mu), \qquad \text{valid for any } \mu \geq 0. \tag{75b}$$

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

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

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

$$\Big\|\mathbb{E}\Big[\big(w_\theta(Y) - w_{\theta^*}(Y)\big)Y\Big]\Big\|_2 \leq 2\sup_{u\in[0,1]}\|\mathbb{E}[\Gamma_u(\widetilde{Y})]\|_{\mathrm{op}}\|\Delta\|_2. \tag{76}$$

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

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

29

By construction, the matrix D is diagonal, so that it suffices to bound the diagonal terms. Beginning with the first diagonal entry, we have

$$D_{11} = \mathbb{E}\Big[\frac{V_1^2}{\sigma^2\big(\exp\big(\frac{\|\theta_u\|_2 V_1}{\sigma^2}\big) + \exp\big(-\frac{\|\theta_u\|_2 V_1}{\sigma^2}\big)\big)^2}\Big] \leq \mathbb{E}\Big[\frac{V_1^2/\sigma^2}{\exp\big(\frac{2\|\theta_u\|_2 V_1}{\sigma^2}\big)}\Big].$$

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

$$D_{11} \leq \mathbb{E}\Big[\frac{V_1^2/\sigma^2}{\exp\big(\frac{2\|\theta_u\|_2 V_1}{\sigma^2}\big)} \mid \mathcal{E}\Big]\mathbb{P}[\mathcal{E}] + \mathbb{E}\Big[\frac{V_1^2/\sigma^2}{\exp\big(\frac{2\|\theta_u\|_2 V_1}{\sigma^2}\big)} \mid \mathcal{E}^c\Big].$$

Conditioned on $\mathcal{E}$ and $\mathcal{E}^c$, respectively, we then apply the bounds (74a) and (74b) to obtain

$$D_{11} \leq \frac{\sigma^2}{e^2\|\theta_u\|_2^2}\mathbb{P}[\mathcal{E}] + \frac{\|\theta^*\|_2^2}{16\sigma^2\exp\big(\frac{\|\theta_u\|_2\|\theta^*\|_2}{2\sigma^2}\big)},$$

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

$$\|\theta_u\|_2 = \|\theta^* + u(\theta - \theta^*)\|_2 \geq \|\theta^*\|_2 - \frac{1}{4}\|\theta^*\|_2 = \frac{3}{4}\|\theta^*\|_2, \tag{77}$$

we obtain the bound $D_{11} \leq \frac{16\sigma^2}{9e^2\|\theta^*\|_2^2}\mathbb{P}(\mathcal{E}) + \frac{\|\theta^*\|_2^2\exp\big(-\frac{3\|\theta^*\|_2^2}{8\sigma^2}\big)}{16\sigma^2}$, whenever $\|\theta^*\|_2^2 \geq 16\sigma^2/3$.

Note that the mean of V_1 is lower bounded as

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

where step (i) follows from the lower bound (77). Consequently, by standard Gaussian tail bounds, we have

$$\mathbb{P}[\mathcal{E}] \leq \exp\Big(\frac{-\|\theta^*\|_2^2}{32\sigma^2}\Big). \tag{78}$$

Combining the pieces yields

$$D_{11} \leq \frac{16\sigma^2}{9e^2\|\theta^*\|_2^2}e^{-\frac{\|\theta^*\|_2^2}{32\sigma^2}} + \frac{\|\theta^*\|_2^2}{16\sigma^2}e^{-\frac{3\|\theta^*\|_2^2}{8\sigma^2}} \qquad \text{whenever } \|\theta^*\|_2^2 \geq 16\sigma^2/3.$$

On the other hand, for any index $j \neq 1$, we have

$$D_{jj} = \mathbb{E}\Big[\frac{1}{\big(\exp\big(\frac{\|\theta_u\|_2 Y_1}{\sigma^2}\big) + \exp\big(-\frac{\|\theta_u\|_2 Y_1}{\sigma^2}\big)\big)^2}\Big] = \mathbb{E}\Big[g\Big(\frac{\|\theta_u\|_2 V_1}{\sigma^2}\Big)\Big],$$

where the reader should recall the function g from equation (75a). Once again, conditioning on the event $\mathcal{E} = \{V_1 \leq \frac{\|\theta^*\|_2}{4}\}$ and its complement yields

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

30

where step (i) follows by applying bound (75a) to the first term, and the bound (75b) with $\mu = \frac{\|\theta^*\|_2 \|\theta_u\|_2}{4\sigma^2}$ to the second term; and step (ii) follows from the bound (77). Applying the bound (78) on $\mathbb{P}[\mathcal{E}]$ yields

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

Returning to equation (76), we have shown that

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

whenever $\frac{\|\theta^*\|_2^2}{\sigma^2} \geq \eta^2 \geq 16/3$. On this basis, the bound (73) holds as long as the signal-to-noise ratio is sufficiently large,

B.2   Proof of Corollary 2

In order to prove this corollary, it suffices to bound the function $\varepsilon_M^{\text{unif}}(n, \delta)$, as previously defined (34). Defining the set $\mathbb{A} := \{\theta \in \mathbb{R}^d \mid \|\theta - \theta^*\|_2 \leq \|\theta^*\|_2/4\}$, our goal is to control the random variable $Z := \sup_{\theta \in \mathbb{A}} \|M(\theta) - M_n(\theta)\|_2$. For each unit-norm vector $u \in \mathbb{R}^d$, define the random variable

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

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

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

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

$$Z = \sup_{v \in \mathbb{S}^d} Z_v \leq 2\max_{j \in [M]} Z_{u^j}. \tag{79}$$

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

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

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

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

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

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

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

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

where $\|\cdot\|_{\text{op}}$ 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

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

Putting together the pieces, we conclude that

$$\mathbb{E}\left[e^{\lambda Z_u}\right] \leq \mathbb{E}\left[\exp\left(10\lambda\|\theta^*\|_2 \max_{j \in [M]}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\langle y_i, u^j\rangle^2\right)\right] \leq \sum_{j=1}^{M}\mathbb{E}\left[\exp\left(10\lambda\|\theta^*\|_2\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\langle y_i, u^j\rangle^2\right)\right].$$
(80)

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 $v \sim \mathcal{N}(0, \sigma^2 I)$. Consequently, for any $u \in \mathbb{R}^d$, we have

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

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 $\mathbb{E}[e^{t(\langle y_i, u\rangle)^2}] \leq e^{\frac{\gamma^2 t^2}{2}}$ for all $t > 0$ sufficiently small. Combined with our earlier inequality (80), we conclude that

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

for all λ sufficiently small. Combined with our first discretization (79), we have thus shown that

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

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

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

with probability at least $1 - \delta$.

B.3   Guarantees for EM with sample-splitting

In this section, we state and prove a result for the EM algorithm with sample-splitting for the mixture of Gaussians.

Corollary 9 (Sample-splitting EM guarantees for Gaussian mixtures). *Consider a Gaussian mixture model satisfying the SNR(η) condition* (57), *and any initialization θ^0 such that $\|\theta^0 - \theta^*\|_2 \leq \frac{\|\theta^*\|_2}{4}$. Given a sample size $n \geq 16T \log(6T/\delta)$, then with probability at least $1 - \delta$, the sample-splitting EM iterates $\{\theta^t\}_{t=0}^T$ satisfy the bound*

$$\|\theta^t - \theta^*\|_2 \leq \kappa^t \|\theta^0 - \theta^*\|_2 + \frac{c}{1-\kappa}\left(\sigma\sqrt{\frac{dT\log(T/\delta)}{n}} + \sqrt{\frac{T\log(T/\delta)}{n}}\|\theta^*\|_2\right). \quad (81)$$

It is worth comparing the result here to the result established earlier in Corollary 2. The sample-splitting EM algorithm is more sensitive to the number of iterations which determines the batch size and needs to be chosen in advance. Supposing that the number of iterations were chosen optimally however the result has better dependence on $\|\theta^*\|_2$ and σ at the cost of a logarithmic factor in n.

Proof. The proof follows by establishing a bound on the function $\varepsilon_M(n, \delta)$. Define $\mathcal{S} = \{\theta : \|\theta - \theta^*\|_2 \leq \frac{\|\theta^*\|_2}{4}\}$. Recalling the updates in (13a) and (13b), note that

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

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

$$\max\{T_1, T_2\} \leq \sqrt{\frac{\log(8/\delta)}{2n}}\|\theta^*\|_2 + c\sigma\sqrt{\frac{d\log(1/\delta)}{n}}, \quad (82)$$

with probability at least $1 - \delta$.

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

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

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

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

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

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

with probability at least $1 - \delta/4$. Combining the pieces yields the claimed bound (82) on T_1.

33

Control of T_2: By triangle inequality, we have

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

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

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

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

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

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

with probability at least $1 - \delta/4$. Putting together the pieces yields the claimed bound (82) on T_2, thereby completing the proof of the corollary. $\qquad\square$

C   Proofs for mixtures of regressions

In this appendix, we provide proofs of results related to the mixture of regressions model, as presented in Section 4.2. More specifically, we first prove Corollary 3 on the population level behavior, followed by the proof of Corollaries 4 and 5 on the behavior of sample-splitting EM updates and stochastic gradient EM updates, respectively.

C.1   Proof of Corollary 3

We begin by proving part (a) of the corollary on the population EM update, which is based on maximizing the function

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

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

It remains to verify condition FOS. Define the difference function $\Delta_w(X, Y) := w_\theta(X, Y) - w_{\theta^*}(X, Y)$, and the difference vectors $\Delta = \theta - \theta^*$. Using this notation, for this model, 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. Using this notation, it is equivalent to show

$$\mathbb{E}\big[\Delta_w(X,Y)(2Z-1)\langle X, \theta^* \rangle \langle X, \widetilde{\Delta} \rangle\big] + \mathbb{E}\big[\Delta_w(X,Y)v\langle X, \widetilde{\Delta} \rangle\big] \leq \gamma \|\Delta\|_2 \,\|\widetilde{\Delta}\|_2 \qquad (83)$$

for $\gamma \in [0, 1/2)$ in order to establish contractivity. In order to prove the theorem with the desired upper bound on κ we need to show (83) with $\gamma \in [0, 1/4)$. The following lemma provides control on the two terms:

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

$$\big|\mathbb{E}\big[\Delta_w(X,Y)(2Z-1)\langle X, \theta^* \rangle \langle X, \widetilde{\Delta} \rangle\big]\big| \leq \frac{\gamma}{2} \|\Delta\|_2 \|\widetilde{\Delta}\|_2, \quad and \qquad (84a)$$

$$\big|\mathbb{E}\big[\Delta_w(X,Y)v\langle X, \widetilde{\Delta} \rangle\big]\big| \leq \frac{\gamma}{2} \|\Delta\|_2 \,\|\widetilde{\Delta}\|_2. \qquad (84b)$$

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.

Part (b) of the corollary is nearly immediate given part (a). Our first task is to verify smoothness of the objective $Q(\cdot|\theta^*)$. The smoothness parameter is given by the largest eigenvalue of the Hessian of $Q(\cdot|\theta^*)$ which is $\mathbb{E}[XX^T]$. Since $\mathbb{E}[XX^T] = I$ by assumption, we see that smoothness holds with $\mu = 1$. Finally, we need to verify the condition GS with the desired contraction coefficient. Some algebra shows that it suffices to show that under the stated assumptions of the corollary we have

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

for $\kappa \in [0, \frac{1}{2})$. This is an immediate consequence of Lemma 4.

It remains to prove Lemma 4. Since the standard deviation σ is known, a simple rescaling argument allows us to take $\sigma = 1$, and replace the weight function in (16a) with

$$w_\theta(x,y) = \frac{\exp\big(\frac{-(y-\langle x, \theta \rangle)^2}{2}\big)}{\exp\big(\frac{-(y-\langle x, \theta \rangle)^2}{2}\big) + \exp\big(\frac{-(y+\langle x, \theta \rangle)^2}{2}\big)}. \qquad (85)$$

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

Lemma 5. *Given a Gaussian random vector $X \sim \mathcal{N}(0, I)$ and any fixed vectors $u, v \in \mathbb{R}^d$, we have*

$$\mathbb{E}[\langle X, u \rangle^2 \langle X, v \rangle^2] \leq 3\|u\|_2^2 \|v\|_2^2 \quad with \ equality \ when \ u = v, \ and \qquad (86a)$$
$$\mathbb{E}[\langle X, u \rangle^4 \langle X, v \rangle^2] \leq 15\|u\|_2^4 \|v\|_2^2. \qquad (86b)$$

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

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

$$\leq 3\|u\|_2^2 \|z\|_2^2 \ = 3\|u\|_2^2 \|v\|_2^2.$$

A similar argument yields the second claim. $\qquad \square$

With these preliminaries in place, we can now begin the proof of Lemma 4. Recall that $\Delta = \theta - \theta^*$ and that $\widetilde{\Delta}$ is any fixed vector in $\mathbb{R}^d$. Define $\theta_u = \theta^* + u\Delta$ for a scalar $u \in [0, 1]$. Recall that by our assumptions guarantee that

$$\|\Delta\|_2 \leq \frac{\|\theta^*\|_2}{32}, \qquad \text{and} \qquad \|\theta^*\|_2 \geq \eta. \tag{87a}$$

For future reference, we observe that

$$\|\theta_u\|_2 \geq \|\theta^*\|_2 - \|\Delta\|_2 \geq \frac{\|\theta^*\|_2}{2}. \tag{87b}$$

Noting that Lemma 4 consists of two separate inequalities (84a) and (84b), we treat these cases separately.

C.1.1 Proof of inequality (84a)

We split the proof of this bound into two separate cases: namely, $\|\Delta\|_2 \leq 1$ and $\|\Delta\|_2 > 1$.

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

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

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

$$\Delta_w(X, Y) = \int_0^1 \frac{2Y\langle X, \Delta\rangle}{(\exp(Z_u) + \exp(-Z_u))^2} \, du, \tag{88}$$

where $Z_u := Y\langle X, \theta^* + u\Delta\rangle$. Substituting for $\Delta_w(X, Y)$ in inequality (84a), we see that it suffices to show

$$\int_0^1 \underbrace{\mathbb{E}\Big[\frac{2Y\langle X, \theta^*\rangle}{(\exp(Z_u) + \exp(-Z_u))^2}(2Z - 1)\langle X, \Delta\rangle\langle X, \widetilde{\Delta}\rangle\Big]}_{A_u} \, du \leq \frac{\gamma}{2}\|\Delta\|_2\|\widetilde{\Delta}\|_2. \tag{89}$$

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

Lemma 6. *There is a $\gamma \in [0, 1/4)$ such that for each $u \in [0, 1]$, we have*

$$\sqrt{\mathbb{E}\Big[\frac{Y^2\langle X, \theta_u\rangle^2}{(\exp(Z_u) + \exp(-Z_u))^4}\Big]} \leq \frac{\gamma}{14}, \quad and \tag{90a}$$

$$\sqrt{\mathbb{E}\Big[\frac{Y^2}{(\exp(Z_u) + \exp(-Z_u))^4}\Big]} \leq \frac{\gamma}{32} \qquad whenever \ \|\Delta\|_2 \leq 1. \tag{90b}$$

See Section C.1.5 for the proof of this lemma.

Using Lemma 6, let us bound the quantity A_u from equation (89). Since $\theta^* = \theta_u - u\Delta$, we have $A_u = B_1 + B_2$, where

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

$$B_2 := -\mathbb{E}\Big[\frac{2Yu\langle X, \Delta\rangle}{(\exp(Z_u) + \exp(-Z_u))^2}(2Z - 1)\langle X, \Delta\rangle\langle X, \widetilde{\Delta}\rangle\Big].$$

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

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

$$B_1 \le \sqrt{\mathbb{E}\Big[\frac{y^2\langle X, \theta_u\rangle^2}{(\exp(Z_u) + \exp(-Z_u))^4}\Big]}\sqrt{\mathbb{E}\big[4(2Z-1)^2\langle X, \Delta\rangle^2\langle X, \widetilde{\Delta}\rangle^2\big]}$$

$$\le \frac{\gamma}{14}\sqrt{\mathbb{E}\big[4\langle X, \Delta\rangle^2\langle X, \widetilde{\Delta}\rangle^2\big]},$$

where the second step follows from the bound (90a), and the fact that $(2Z-1)^2 = 1$. Next we observe that $\mathbb{E}\big[4\langle X, \Delta\rangle^2\langle X, \widetilde{\Delta}\rangle^2\big] \le 12\|\Delta\|_2^2\|\widetilde{\Delta}\|_2^2$, where we have used the bound (86a) from Lemma 5. Combined with our earlier bound, we conclude that $B_1 \le \frac{\gamma}{4}\|\Delta\|_2\|\widetilde{\Delta}\|_2$, as claimed.

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

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

$$\le \frac{\gamma}{32}\sqrt{\mathbb{E}\big[4u^2\langle X, \Delta\rangle^4\langle X, \widetilde{\Delta}\rangle^2\big]},$$

where the second step follows from the bound (90b), and the fact that $(2Z-1)^2 = 1$. In this case, we have

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

where step (i) uses the bound (86b) from Lemma 5, and step (ii) that $\|\Delta\|_2 \le 1$. Combining the pieces, we conclude that $B_2 \le \frac{\gamma}{4}\|\Delta\|_2\|\widetilde{\Delta}\|_2$, which completes the proof of inequality (84a) in the case $\|\Delta\|_2 \le 1$.

Case $\|\Delta\|_2 > 1$: We now turn to the second case of the bound (84a). Our argument (here and in later sections) makes use of various probability bounds on different events, which we state here for future reference. These events involve the scalar $\tau := C_\tau\sqrt{\log\|\theta^*\|_2}$ for a constant C_τ, as well as the vectors

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

Lemma 7 (Event bounds)**.**

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

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

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

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

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

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

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

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

Lemma 8 (Conditional covariance bounds). *Conditioned on any event $\mathcal{E} \in \{\mathcal{E}_1 \cap \mathcal{E}_2, \mathcal{E}_1^c, \mathcal{E}_5^c, \mathcal{E}_6^c\}$, we have $\|\mathbb{E}[XX^T \mid \mathcal{E}]\|_{op} \leq 2$.*

See Section C.1.7 for the proof of this result.

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

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

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

$$T \leq \Psi(\mathcal{E}_1 \cap \mathcal{E}_2) + \Psi(\mathcal{E}_1^c) + \Psi(\mathcal{E}_4^c) + \Psi(\mathcal{E}_5^c) + \Psi(\mathcal{E}_6^c). \tag{91}$$

We bound each of these five terms in turn.

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

$$\Psi(\mathcal{E}_1 \cap \mathcal{E}_2) \leq \sqrt{\mathbb{E}\left[\Delta_w(X,Y)^2 \langle X, \widetilde{\Delta} \rangle^2 | \mathcal{E}_1 \cap \mathcal{E}_2 \right]} \sqrt{\mathbb{E}\left[\langle X, \theta^* \rangle^2 | \mathcal{E}_1 \cap \mathcal{E}_2 \right]}. \tag{92}$$

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

$$\text{sign}(Y\langle X, \theta^* \rangle) = \text{sign}(Y\langle X, \theta_u \rangle). \tag{93a}$$

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

$$\min\left\{ |Y\langle X, \theta^* \rangle|, |Y\langle X, \theta \rangle| \right\} \geq \frac{\tau^2}{2}. \tag{93b}$$

Recalling the weight function (85), we claim that when conditions (93a) and (93b) hold, then

$$|\Delta_w(X,Y)| = |w_{\theta_u}(X,Y) - w_{\theta^*}(X,Y)| \overset{(i)}{\leq} \frac{\exp(-\tau^2/2)}{\exp(-\tau^2/2) + \exp(\tau^2/2)} \leq \exp(-\tau^2). \tag{94}$$

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

Combined with our earlier bound (92), we have shown

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

Applying Lemma 8 with $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2$ yields $\Psi(\mathcal{E}_1 \cap \mathcal{E}_2) \leq 2\|\widetilde{\Delta}\|_2 \|\theta^*\|_2 e^{-\tau^2}$.

Bounding $\Psi(\mathcal{E}_1^c)$: Combining the Cauchy-Schwarz inequality with Lemma 7(i), we have

$$\Psi(\mathcal{E}_1^c) \leq \sqrt{\mathbb{E}\big[\langle X, \widetilde{\Delta}\rangle^2|\mathcal{E}_1^c\big]}\sqrt{\mathbb{E}\big[\langle X, \theta^*\rangle^2|\mathcal{E}_1^c\big]}\frac{\|\Delta\|_2}{\|\theta^*\|_2}. \tag{95}$$

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

$$\langle X, \theta^*\rangle^2 = \frac{1}{4}\langle X, (\theta^* - \theta_u) + (\theta_u + \theta^*)\rangle^2 \leq \frac{1}{2}\langle X, \theta^* - \theta_u\rangle^2 + \frac{1}{2}\langle X, \theta_u + \theta^*\rangle^2$$

$$\overset{(i)}{\leq} \langle X, \theta^* - \theta_u\rangle^2$$

$$\overset{(ii)}{\leq} \langle X, \Delta\rangle^2$$

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

Returning to equation (95), we have

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

where step (i) follows from the conditional covariance bound of Lemma 8.

Bounding $\Psi(\mathcal{E}_4^c)$: Combining the Cauchy-Schwarz inequality with Lemma 7(iv) yields

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

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

Bounding $\Psi(\mathcal{E}_5^c)$: Combining the Cauchy-Schwarz inequality with Lemma 7(v) yields $\Psi(\mathcal{E}_5^c) \leq \frac{\tau}{\|\theta_u\|_2}\sqrt{\mathbb{E}\big[\langle X, \widetilde{\Delta}\rangle^2|\mathcal{E}_5^c\big]}\sqrt{\mathbb{E}\big[\langle X, \theta^*\rangle^2|\mathcal{E}_5^c\big]}$. Conditioned on the event $\mathcal{E}_5^c$, we have

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

Together with Lemma 8, we obtain the bound

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

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

Bounding $\Psi(\mathcal{E}_6^c)$: Combining the Cauchy-Schwarz inequality with Lemma 7(vi) yields $\Psi(\mathcal{E}_6^c) \leq \frac{\tau}{\|\theta^*\|_2}\sqrt{\mathbb{E}\big[\langle X, \widetilde{\Delta}\rangle^2|\mathcal{E}_6^c\big]}\sqrt{\mathbb{E}\big[\langle X, \theta^*\rangle^2|\mathcal{E}_6^c\big]}$. Conditioned on the event $\mathcal{E}_6^c$, we have $\langle X, \theta^*\rangle^2 \leq \tau^2$,

and so applying Lemma 8 with $\mathcal{E} = \mathcal{E}_6^c$ yields $\Psi(\mathcal{E}_6^c) \leq \frac{\sqrt{2}\tau^2\|\widetilde{\Delta}\|_2}{\|\theta^*\|_2}$.

We have thus obtained bounds on all five terms in the decomposition (91). We combine these bounds with the with lower bound $\|\theta_u\|_2 \geq \frac{\|\theta^*\|_2}{2}$ from equation (87b), and then perform some algebra to obtain

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

where c is a universal constant. In particular, selecting $\tau = c_\tau \sqrt{\log\|\theta^*\|_2}$ for a sufficient large constant c_τ, selecting the constant η in (87a) sufficiently large yields the claim (84a).

C.1.2   Proof of inequality (84b)

As in Section C.1.1, we treat the cases $\|\Delta\|_2 \leq 1$ and $\|\Delta\|_2 \geq 1$ separately.

C.1.3   Case $\|\Delta\|_2 \leq 1$:

As before, by a Taylor expansion of the function $\theta \mapsto \Delta_w(X, Y)$, it suffices to show that

$$\int_0^1 \mathbb{E}\Big[\frac{2Yv}{(\exp(Z_u) + \exp(-Z_u))^2}\langle X, \Delta\rangle\langle X, \widetilde{\Delta}\rangle\Big]du \leq \frac{\gamma}{2}\|\Delta\|_2\|\widetilde{\Delta}\|_2.$$

For any fixed $u \in [0, 1]$, the Cauchy-Schwarz inequality implies that

$$\mathbb{E}\Big[\frac{2Yv\,\langle X, \Delta\rangle\langle X, \widetilde{\Delta}\rangle}{(\exp(Z_u) + \exp(-Z_u))^2}\Big] \leq \sqrt{\mathbb{E}\Big[\frac{4Y^2}{(\exp(Z_u) + \exp(-Z_u))^4}\Big]} \sqrt{\mathbb{E}\big[v^2\langle X, \Delta\rangle^2\langle X, \widetilde{\Delta}\rangle^2\big]}$$

$$\overset{(i)}{\leq} \sqrt{\mathbb{E}\Big[\frac{4Y^2}{(\exp(Z_u) + \exp(-Z_u))^4}\Big]} \sqrt{3\,\|\Delta\|_2^2\,\|\widetilde{\Delta}\|_2^2}$$

$$\overset{(ii)}{\leq} \frac{\sqrt{3}\gamma}{16}\|\Delta\|_2\,\|\widetilde{\Delta}\|_2,$$

where step (i) follows from inequality (86a) in Lemma 5, the independence of v and X, and the fact that $\mathbb{E}[v^2] = 1$; and step (ii) follows from the bound (90b) in Lemma 6.

C.1.4   Case $\|\Delta\|_2 > 1$:

After applying the Cauchy-Schwarz inequality, it suffices show that $\sqrt{\mathbb{E}\big[\Delta_w^2(X, Y)\big]} \leq \frac{\gamma}{2}$. The remainder of this section is devoted to the proof of this claim.

Recall the scalar $\tau := C_\tau\sqrt{\log\|\theta^*\|_2}$, as well as the events $\mathcal{E}_1$ and $\mathcal{E}_2$ from Lemma 7. For any measurable event $\mathcal{E}$, define the function $\Psi(\mathcal{E}) = \mathbb{E}\big[\Delta_w^2(X, Y) \mid \mathcal{E}\big]\mathbb{P}[\mathcal{E}]$. With this notation, by successive conditioning, we have the upper bound

$$\mathbb{E}\big[\Delta_w^2(X, Y)\big] \leq \Psi(\mathcal{E}_1^c) + \Psi(\mathcal{E}_1 \cap \mathcal{E}_2^c) + \Psi(\mathcal{E}_1 \cap \mathcal{E}_2). \tag{96}$$

We control each of these terms in turn.

Controlling term $\Psi(\mathcal{E}_1^c)$: Noting that $\sup_{x,y} |\Delta_w(x,y)| \leq 2$ and applying Lemma 7(i), we have $\Psi(\mathcal{E}_1^c) \leq 4\mathbb{P}[\mathcal{E}_1^c] \leq 4\frac{\|\Delta\|_2}{\|\theta^*\|_2}$.

Controlling term $\Psi(\mathcal{E}_1 \cap \mathcal{E}_2^c)$: Similarly, Lemma 7(ii) implies that

$$\Psi(\mathcal{E}_1 \cap \mathcal{E}_2^c) \leq 4\mathbb{P}[\mathcal{E}_2^c] \leq 4\Big\{ \frac{\tau}{\|\theta^*\|_2} + \frac{\tau}{\|\theta_u\|_2} + 2e^{-\frac{\tau^2}{2}} \Big\}.$$

Controlling term $\Psi(\mathcal{E}_1 \cap \mathcal{E}_2)$: Conditioned on the event $\mathcal{E}_1 \cap \mathcal{E}_2$, the bound (94) implies that $|\Delta_w(X,Y)| \leq \exp(-\tau^2)$, and hence $\Psi(\mathcal{E}_1 \cap \mathcal{E}_2) \leq e^{-2\tau^2}$.

Thus, we have derived bounds on each of the three terms in the decomposition (96): putting them together yields

$$\sqrt{\mathbb{E}\big[\Delta_w^2(X,Y)\big]} \leq \sqrt{4\frac{\|\Delta\|_2}{\|\theta^*\|_2} + 4\Big\{ \frac{\tau}{\|\theta^*\|_2} + \frac{\tau}{\|\theta_u\|_2} + 2e^{-\frac{\tau^2}{2}} \Big\} + e^{-2\tau^2}}$$

By choosing C_τ sufficiently large in the definition of τ, selecting the signal-to-noise constant η in condition (87a) sufficiently large, the claim follows.

C.1.5   Proof of Lemma 6

The lemma statement consists of two inequalities, and we divide our proof accordingly.

Proof of inequality (90a): For any measurable event $\mathcal{E}$, let us introduce the function $\Psi(\mathcal{E}) := \mathbb{E}\Big[\frac{Y^2\langle X, \theta_u\rangle^2}{(\exp(Z_u)+\exp(-Z_u))^4} \mid \mathcal{E} \Big]\mathbb{P}[\mathcal{E}]$. With this notation, successive conditioning yields the decomposition

$$\mathbb{E}\Big[\frac{Y^2\langle X, \theta_u\rangle^2}{(\exp(Z_u)+\exp(-Z_u))^4} \Big] = \Psi(\mathcal{E}_4^c) + \Psi(\mathcal{E}_4 \cap \mathcal{E}_3^c) + \Psi(\mathcal{E}_2), \tag{97}$$

and we bound each of these terms in turn. The reader should recall the constant $\tau := C_\tau \sqrt{\log \|\theta^*\|_2}$, as well as the events $\mathcal{E}_3$ and $\mathcal{E}_4$ from Lemma 7.

Bounding $\Psi(\mathcal{E}_4^c)$: Observe that

$$\frac{Y^2\langle X, \theta_u\rangle^2}{(\exp(Z_u)+\exp(-Z_u))^4} \leq \sup_{t\geq 0} \frac{t^2}{\exp(4t)} \leq \frac{1}{4e^2}, \tag{98}$$

where the final step follows from inequality (74a). Combined with Lemma 7(iv), we conclude that $\Psi(\mathcal{E}_4^c) \leq \frac{1}{2e^2}e^{-\frac{\tau^2}{2}}$.

Bounding $\Psi(\mathcal{E}_4 \cap \mathcal{E}_3^c)$: In this case, we have

$$\Psi(\mathcal{E}_4 \cap \mathcal{E}_3^c) \overset{(i)}{\leq} \frac{1}{4e^2}\mathbb{P}[\mathcal{E}_3^c] \overset{(ii)}{\leq} \frac{1}{4e^2}\Big\{ \frac{\tau}{\|\theta^*\|_2} + \frac{\tau}{\|\theta_u\|_2} \Big\},$$

where step (i) follows from inequality (98), and step (ii) follows from Lemma 7(iii).

Bounding $\Psi(\mathcal{E}_2)$: Conditioned on the event $\mathcal{E}_2$, we have $Y^2\langle X, \theta_u\rangle^2 \geq \frac{\tau^2}{2}$, where we have used the lower bound (93b). Introducing the shorthand $t^* = \tau^2/2$, this lower bound implies that

$$\Psi(\mathcal{E}_2) \leq \sup_{t \geq t^*} \frac{t^2}{e^{4t}} \overset{(i)}{\leq} \frac{(t^*)^2}{e^{4t^*}} = \frac{\tau^4}{4e^{2\tau^2}},$$

where inequality (i) is valid as long as $t^* = \frac{\tau^2}{2} \geq \frac{1}{2}$, or equivalently $\tau^2 \geq 1$.

Substituting our upper bounds on three components in the decomposition (97) yields

$$\mathbb{E}\Big[\frac{Y^2\langle X, \theta_u\rangle^2}{(\exp(Z_u) + \exp(-Z_u))^4}\Big] \leq \frac{1}{2e^2} e^{-\frac{\tau^2}{2}} + \frac{1}{4e^2}\Big(\frac{\tau}{\|\theta^*\|_2} + \frac{\tau}{\|\theta_u\|_2}\Big) + \frac{\tau^4}{4}e^{-2\tau^2}.$$

Setting C_τ sufficiently large in the definition of τ and choosing sufficiently large values of the signal-to-noise constant η in the condition (87a) yields the claim.

Proof of inequality (90b): For any measurable event $\mathcal{E}$, let us introduce the function $\Psi(\mathcal{E}) = \mathbb{E}\big[\frac{Y^2}{(\exp(Z_u) + \exp(-Z_u))^4} \mid \mathcal{E}\big]\mathbb{P}[\mathcal{E}]$. Recalling the event $\mathcal{E}_5$ from Lemma 7, successive conditioning yields the decomposition

$$\mathbb{E}\Big[\frac{Y^2}{(\exp(Z_u) + \exp(-Z_u))^4}\Big] = \Psi(\mathcal{E}_5^c) + \Psi(\mathcal{E}_5). \tag{99}$$

We bound each of these terms in turn.

Bounding $\Psi(\mathcal{E}_5^c)$: Simple algebra combined with Lemma 7(v) yields the upper bound $\Psi(\mathcal{E}_5^c) \leq \frac{\tau}{16\|\theta_u\|_2}\mathbb{E}[Y^2]$. Conditioned on $\mathcal{E}_5$, we have the upper bound $|\langle X, \theta_u\rangle| \leq \tau$, whence

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

Combining Lemma 8 with the bound $\|\Delta\|_2 \leq 1$, we find that $\langle X, \theta^*\rangle^2 \leq 2\tau^2 + 4$. Since $Y \overset{d}{=} (2Z - 1)\langle X, \theta^*\rangle + v$, we have

$$\mathbb{E}[Y^2 \mid \mathcal{E}_5^c] \leq \mathbb{E}[2\langle X, \theta^*\rangle^2 + 2v^2 \mid \mathcal{E}_5^c] \overset{(i)}{\leq} 4\tau^2 + 10.$$

Putting together the pieces, we conclude that $\Psi(\mathcal{E}_5^c) \leq \frac{4\tau^3 + 10\tau}{16\|\theta_u\|_2}$.

Bounding $\Psi(\mathcal{E}_5)$: Recall that $Z_u = Y\langle X, \theta_u\rangle$, so we have that

$$\Psi(\mathcal{E}_5) \leq \mathbb{E}\Big[\frac{Y^2}{(e^{Y\langle X, \theta_u\rangle} + e^{-Y\langle X, \theta_u\rangle})^4} \mid \mathcal{E}_5\Big] \overset{(i)}{\leq} \frac{4}{(e\,\tau)^2},$$

where step (i) follows from the bound (74a) and the observation that $|\langle X, \theta_u\rangle| \geq \tau$ conditioned on the event $\mathcal{E}_5$.

Substituting our bounds on the two terms into the decomposition (97) yields

$$\mathbb{E}\Big[\frac{Y^2}{(e^{Z_u} + e^{-Z_u})^4}\Big] \leq \frac{4\tau^3 + 10\tau}{16\|\theta_u\|_2} + \frac{4}{(e\,\tau)^2} \leq \frac{8\tau^3 + 20\tau}{16\|\theta^*\|_2} + \frac{4}{(e\,\tau)^2}.$$

Once again, sufficiently large choices of the constant c_τ and the signal-to-noise constant η in equation (87a) yields the claim.

C.1.6   Proof of Lemma 7

In this section, we prove the probability bounds on events $\mathcal{E}_1$ through $\mathcal{E}_6$ stated in Lemma 7. In doing so, we make use of the following auxiliary result, due to Yi et al. [52] (see Lemma 1 in their paper):

Lemma 9. *Given vectors $v, z \in \mathbb{R}^d$ and a Gaussian random vector $X \sim \mathcal{N}(0, I)$, the matrix $\Sigma = \mathbb{E}[XX^T \mid \langle X, v \rangle^2 > \langle X, z \rangle^2]$ has singular values*

$$\left(1 + \frac{\sin \alpha}{\alpha}, 1 - \frac{\sin \alpha}{\alpha}, 1, \ldots, 1\right), \qquad \text{where } \alpha = \cos^{-1} \frac{\langle z-v, z+v \rangle}{\|z+v\|_2 \|z-v\|_2}. \tag{100a}$$

Moreover, whenever $\|v\|_2 \le \|z\|_2$, we have

$$\mathbb{P}[\langle X, v \rangle^2 > \langle X, z \rangle^2] \le \frac{\|v\|_2}{\|z\|_2}. \tag{100b}$$

Proof of Lemma 7(i):   Note that the event $\mathcal{E}_1^c$ holds if and only if $\langle X, \theta^* \rangle \langle X, \theta_u \rangle < 0$, or equivalently, if and only if

$$4\langle X, \theta^* \rangle \langle X, \theta_u \rangle = \langle X, \theta^* + \theta_u \rangle^2 - \langle X, \theta^* - \theta_u \rangle^2 < 0.$$

Now observe that

$$\|\theta^* - \theta_u\|_2 \le u\|\Delta\|_2 \le \|\Delta\|_2, \quad \text{and} \quad \|\theta^* + \theta_u\|_2 \quad \ge 2\|\theta^*\|_2 - \|\Delta\|_2 \ge \|\theta^*\|_2 \ge \|\Delta\|_2.$$

Consequently, we may apply the bound (100b) from Lemma 9 with $v = \theta^* + \theta_u$ and $z = \theta^* - \theta_u$ to obtain $\mathbb{P}[\mathcal{E}_1^c] \le \frac{\|\theta^* - \theta_u\|_2}{\|\theta^* + \theta_u\|_2} \le \frac{\|\Delta\|_2}{\|\theta^*\|_2}$, as claimed.

Proof of Lemma 7(iv):   For $X \sim \mathcal{N}(0, \sigma^2)$, we have $\mathbb{P}[|X| \le \tau] \le 2 \exp e^{-\frac{\tau^2}{2\sigma^2}}$ for any $\tau \ge 0$, from which the claim follows.

Proof of Lemma 7(v):   For $X \sim \mathcal{N}(0, \sigma^2)$, we have

$$\mathbb{P}[|X| \le \tau] \le \sqrt{\frac{2}{\pi}} \frac{\tau}{\sigma} \qquad \text{for any } \tau \ge 0 \tag{101}$$

from which the claim follows.

Proof of Lemma 7(vi):   Similarly, this inequality follows from the tail bound (101).

Proof of Lemma 7(iii):   This claim follows from parts (v) and (vi) of Lemma 7, combined with the union bound.

Proof of Lemma 7(ii):   This bound follows from parts (iii) and (iv) of Lemma 7, combined with the union bound.

C.1.7   Proof of Lemma 8

For an event $\mathcal{E}$, define the matrix $\Gamma(\mathcal{E}) = \mathbb{E}[XX^T \mid \mathcal{E}]$. The lemma concerns the operator norm of this matrix for different choices of the event $\mathcal{E}$.

Conditioned on $\mathcal{E}_1 \cap \mathcal{E}_2$:   In this case, we write

$$\mathbb{E}[XX^T] = \Gamma(\mathcal{E}_1 \cap \mathcal{E}_2)\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2] + \Gamma((\mathcal{E}_1 \cap \mathcal{E}_2)^c)\mathbb{P}[(\mathcal{E}_1 \cap \mathcal{E}_2)^c] \;\succeq\; \Gamma(\mathcal{E}_1 \cap \mathcal{E}_2)\,\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2].$$

Since $\mathbb{E}[XX^T] = I$, we conclude that $\|\Gamma(\mathcal{E}_1 \cap \mathcal{E}_2)\|_{\mathrm{op}} \le \frac{1}{\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2]}$, and hence it suffices show that $\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2] \ge \frac{1}{2}$. Parts (i) and (ii) of Lemma 7 imply that

$$\mathbb{P}[\mathcal{E}_1 \cap \mathcal{E}_2] \ge 1 - \frac{\|\Delta\|_2}{\|\theta^*\|_2} - \frac{\tau}{\|\theta^*\|_2} - \frac{\tau}{\|\theta_u\|_2} - 2e^{-\frac{\tau^2}{2}}.$$

For appropriate choices of c_τ and the constant η in the signal-to-noise condition (87a), the claim follows.

Conditioned on $\mathcal{E}_1^c$:   As before, note that the event $\mathcal{E}_1^c$ holds if and only if the inequality $|\langle X,\, \theta^* + \theta_u \rangle| < |\langle X,\, \theta^* - \theta_u \rangle|$ holds. Consequently, Lemma 9 implies that $\|\Gamma(\mathcal{E}_1^c)\|_{\mathrm{op}} \le 2$.

Conditioned on $\mathcal{E}_5^c$:   We make note of an elementary fact about Gaussians: for any scalar $\alpha > 0$ and unit norm vector $\|v\|_2 = 1$, for $X \sim \mathcal{N}(0, I_d)$, we have

$$\|\mathbb{E}[XX^T \mid |\langle X,\, v\rangle| \le \alpha]\|_{\mathrm{op}} \le \max\left(1, \alpha^2\right). \tag{102}$$

In particular, when $\alpha \le 1$, then the operator norm is at most 1. This claim follows easily from the rotation invariance of the Gaussian, which allows us to assume that $v = e_1$ without loss of generality. It is thus equivalent to bound the largest eigenvalue of the matrix

$$D := \mathbb{E}[XX^T \mid |X_1| \le \alpha],$$

which is a diagonal matrix by independence of the entries of X. Noting that $D_{11} \le \alpha^2$ and $D_{jj} = 1$ for $j \neq 1$ completes the proof of the bound (102).

Applying the bound (102), we find that $\|\Gamma(\mathcal{E}_5^c)\|_{\mathrm{op}} \le \max\left(1, \frac{\tau^2}{\|\theta_u\|_2^2}\right)$. Consequently, the claim follows by making sufficiently large choices of c_τ and the constant η in the signal-to-noise condition (87a).

Conditioned on $\mathcal{E}_6^c$:   The bound (102) implies that $\|\mathbb{E}[XX^T \mid \mathcal{E}_6^c]\|_{\mathrm{op}} \le \max\left\{1, \frac{\tau^2}{\|\theta^*\|_2^2}\right\}$. As in the previous case, choosing c_τ and η appropriately ensures that $\frac{\tau^2}{\|\theta^*\|_2^2} \le 1$.

C.2   Proof of Corollary 4

We need to compute an upper bound on the function $\varepsilon_M(n, \delta)$ previously defined in equation (33). For this particular model, we have

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

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

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

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

$$\|M(\theta) - M_n(\theta)\|_2 \leq \underbrace{\|\widehat{\Sigma}^{-1}\|_{\mathrm{op}} \|\widehat{v} - v\|_2}_{T_1} + \underbrace{\|\widehat{\Sigma}^{-1} - \Sigma^{-1}\|_{\mathrm{op}} \|v\|_2}_{T_2}. \tag{103}$$

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

Bounding T_1: Recall the assumed lower bound on the sample size—namely $n > c\,d \log(1/\delta)$ for a sufficiently large constant c. Under this condition, standard bounds in random matrix theory [47], guarantee that $\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}} \leq \frac{1}{2}$ with probability at least $1 - \delta$. When this bound holds, we have $\|\widehat{\Sigma}^{-1}\|_{\mathrm{op}} \geq 1/2$.

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

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

By a discretization argument over a $1/2$-cover of the sphere $\mathbb{S}^d$—say $\{u^1, \ldots, u^M\}$—we have the upper bound $\|\widehat{v} - v\|_2 \leq 2 \max_{j \in [M]} Z(u^j)$. Thus, it suffices to control the random variable $Z(u)$ for a fixed $u \in \mathbb{S}^d$. By a standard symmetrization argument [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 $\mathcal{E}\{\frac{1}{n} \sum_{i=1}^{n} \langle x_i, u \rangle^2 \leq 2\}$. Since each variable $\langle x_i, u \rangle$ is sub-Gaussian with parameter one, standard tail bounds imply that $\mathbb{P}[\mathcal{E}^c] \leq e^{-n/32}$. Therefore, we can write

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

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

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

$$\leq \exp\Big(\frac{8\lambda^2}{n}\big(\|\theta^*\|_2^2 + \sigma^2\big)\Big),$$

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

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

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

Bounding T_2: Since $n > d$ by assumption, standard results in random matrix theory [47] imply that $\|\widehat{\Sigma}^{-1} - \Sigma^{-1}\|_{\mathrm{op}} \leq c\sqrt{\frac{d}{n} \log(1/\delta)}$ with probability at least $1 - \delta$. On the other hand, observe that

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

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

Finally, substituting our bounds on T_1 and T_2 into the decomposition (103) yields the claim.

C.3   Proof of Corollary 5

We need to bound the uniform variance $\sigma_{\mathrm{G}}^2 = \sup_{\theta \in \mathbb{B}_2(r;\theta^*)} \mathbb{E}\|\nabla Q_1(\theta|\theta)\|_2^2$, where $r = \frac{\|\theta^*\|_2}{32}$. From the gradient update (18a), we have $\nabla Q_1(\theta \mid \theta) = (2w_\theta(x_1, y_1) - 1)y_1 x_1 - \langle x_1, \theta \rangle x_1$, and hence

$$\mathbb{E}\big[\|\nabla Q_1(\theta|\theta)\|_2^2\big] \leq 2 \underbrace{\mathbb{E}[y_1^2\|x_1\|_2^2]}_{T_1} + 2 \underbrace{\|\mathbb{E}[x_1 x_1^T \|x_1\|_2^2]\|_{\mathrm{op}}}_{T_2} \|\theta\|_2^2. \tag{104}$$

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

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

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

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

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

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

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

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

Putting together the pieces yields $\sigma_{\mathrm{G}}^2 \leq c(\sigma^2 + \|\theta^*\|_2^2)d$, so that Corollary 5 follows as a consequence of Theorem 5.

D   Proofs for missing covariates

In this appendix, we provide proofs of results related to regression with missing covariates, as presented in Section 4.3. More specifically, we first prove Corollary 6 on the population level behavior, followed by the proof of Corollaries 7 and 8 on the behavior of sample-splitting EM updates and stochastic gradient EM updates, respectively.

D.1   Proof of Corollary 6

We need to verify the conditions of Theorem 3, 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

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

where the vector $\mu_{\theta^*} \in \mathbb{R}^d$ and matrix Σ_{θ^*} were previously defined (see equations (20a) and (20c) 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 $\nabla^2 q(\theta) = \mathbb{E}\big[\Sigma_{\theta^*}(X_{\mathrm{obs}}, Y)\big]$. Let us fix a pattern of missingness, and then average over (X_{obs}, Y). Recalling the matrix U_{θ^*} from equation (20b), we find that yields

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

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

$$V := \Sigma_{\theta^*}(X_{\mathrm{obs}}, Y)\theta - Y\mu_{\theta^*}(X_{\mathrm{obs}}, Y) - \Sigma_\theta(X_{\mathrm{obs}}, Y)\theta + Y\mu_\theta(X_{\mathrm{obs}}, Y). \tag{105}$$

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

$$\mathbb{E}_{X_{\mathrm{obs}}, Y}[V] = \begin{bmatrix} (\theta_{\mathrm{mis}} - \theta_{\mathrm{mis}}^*) + \pi_1\theta_{\mathrm{mis}} \\ \pi_2(\theta_{\mathrm{obs}} - \theta_{\mathrm{obs}}^*) \end{bmatrix}. \tag{106}$$

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

$$\pi_1 \leq 2(\xi_1 + \xi_2)\frac{\|\theta - \theta^*\|_2}{\sigma}, \qquad \text{and} \quad \pi_2 \leq \delta := \frac{1}{1 + \left(\frac{1}{\xi_1 + \xi_2}\right)^2} < 1. \tag{107}$$

Taking these bounds (107) 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

47

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

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

where we have used our upper bound (107) 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 (107). By our assumptions, we have $\|\theta_{\text{mis}}\|_2 - \|\theta_{\text{mis}}^*\|_2 \leq \|\theta_{\text{mis}} - \theta_{\text{mis}}^*\|_2$, and moreover

$$\|\theta_{\text{mis}}\|_2 \leq \|\theta_{\text{mis}}^*\|_2 + \xi_2\sigma \leq (\xi_1 + \xi_2)\sigma. \tag{108}$$

As consequence, we have

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

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

On the other hand, we have

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

where step (i) follows from (108).

D.2 Proof of Corollary 7

We need to upper bound the deviation function $\varepsilon_G(n, \delta)$ previously defined (49). For any fixed $\theta \in \mathbb{B}_2(r; \theta^*) = \{\theta \in \mathbb{R}^d \mid \|\theta - \theta^*\|_2 \leq \xi_2\sigma\}$, we have the bound $\|G(\theta) - G_n(\theta)\|_2 \leq T_1 + T_2$, where

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

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

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 $\bar{\Sigma} = \mathbb{E}[\Sigma_\theta(x_{\mathrm{obs}}, y)]$ and $\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n \Sigma_\theta(x_{\mathrm{obs},i}, y_i)$. With this notation, we have

$$T_1 \leq \|\bar{\Sigma} - \widehat{\Sigma}\|_{\mathrm{op}} \, \|\theta\|_2 \; \leq \; \|\bar{\Sigma} - \widehat{\Sigma}\|_{\mathrm{op}} \, (\xi_1 + \xi_2)\, \sigma,$$

where the second step follows since any vector $\theta \in \mathbb{B}_2(r; \theta^*)$ has 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 $\|\bar{\Sigma} - \widehat{\Sigma}\|_{\mathrm{op}} \leq c\sqrt{\frac{d}{n} \, \log(1/\delta)}$ with probability at least $1 - \delta$.

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

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

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 10. *Under the conditions of Corollary 6, the random vector $\mu_\theta(x_{\mathrm{obs}}, y)$ is sub-Gaussian with a constant parameter.*

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

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

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

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

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

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

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. $\square$

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

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

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

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

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 10 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 [8], 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 $\varepsilon_G(n, \delta) \leq c\sqrt{1+\sigma^2} \sqrt{\frac{d}{n} \log(1/\delta)}$ with probability at least $1 - \delta$. Thus, we see that Corollary 7 follows from Theorem 2.

D.3 Proof of Corollary 8

Once again we focus on bounding the uniform variance σ_G^2. From the form of Q given in equation (21) (with $n = 1$), we have

$$\mathbb{E}\left[\|\nabla Q_1(\theta|\theta)\|_2^2 \right] \leq 2\Big\{ \underbrace{\mathbb{E}\left[\|\Sigma_\theta(x_{\mathrm{obs}}, y)\theta\|_2^2 \right]}_{T_1} + \underbrace{\mathbb{E}\big[y^2 \|\mu_\theta(x_{\mathrm{obs}}, y)\|_2^2 \big]}_{T_2} \Big\}. \tag{109}$$

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

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

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

By the Cauchy-Schwarz inequality, we have

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

From Lemma 10, the random vector μ_θ is sub-Gaussian with constant parameter, so that $\mathbb{E}[\|\mu_\theta\|_2^4] \leq c\,d^2$. Since $\|\theta\|_2 \leq c\|\theta^*\|_2$, the random variable $\langle \mu_\theta, \theta \rangle$ is sub-Gaussian with parameter $c\|\theta^*\|_2$, and hence $\mathbb{E}[\langle \mu_\theta, \theta \rangle^4] \leq c\|\theta^*\|_2^4$. Putting together the pieces, we see that $\mathbb{E}\big[\|\mu_\theta\|_2^2 \langle \mu_\theta, \theta \rangle^2 \big] \leq c\,d\|\theta^*\|_2^2$. A similar argument applies to other expectation, so that we conclude that $T_1 = \mathbb{E}\big[\|\Sigma_\theta \theta\|_2^2 \big] \leq c\|\theta^*\|_2^2 d$, a bound that holds uniformly for all $\theta \in \mathbb{B}_2(r; \theta^*)$.

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

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

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

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

Similarly, Lemma 10 implies that $\sqrt{\mathbb{E}[\|\mu_\theta(x_{\text{obs}}, y)\|_2^4]} \leq cd$, and hence $T_2 \leq c'(\|\theta^*\|_2^2 + \sigma^2)d$.

Substituting our upper bounds on T_1 and T_2 into the decomposition (109), we find that $\sigma_G^2 \leq c\left(\|\theta^*\|_2^2 + \sigma^2\right)d$. Thus, Corollary 8 follows from Theorem 5.

References

[1] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *CoRR*, abs/1210.7559, 2012.

[2] A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon. Learning sparsely used over-complete dictionaries via alternating minimization. Technical report, Microsoft Research, 2013.

[3] R. Balan, P. Casazza, and D. Edidin. On signal reconstruction without phase. *Applied and Computational Harmonic Analysis*, 20(3):345 – 356, 2006.

[4] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970.

[5] E. M. L. Beale and R. J. A. Little. Missing values in multivariate analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 37(1):pp. 129–145, 1975.

[6] D. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1995.

[7] S. Bubeck. Theory of convex optimization for machine learning. 2014. `arXiv:1405.4980`.

[8] V. V. Buldygin and Y. V. Kozachenko. *Metric characterization of random variables and random processes*. American Mathematical Society, Providence, RI, 2000.

[9] E. J. Candès, T. Strohmer, and V. Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[10] G. Celeux, D. Chauveau, and J. Diebolt. On stochastic versions of the EM algorithm. Technical Report 2514, INRIA, 1995.

[11] G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Comput. Stat. Data Anal.*, 14(3):315–332, 1992.

[12] A. T. Chaganty and P. Liang. Spectral experts for estimating mixtures of linear regressions. 2013. `arXiv:1306.3729`.

[13] Y. Chen, X. Yi, and C. Caramanis. A convex formulation for mixed regression: Near optimal rates in the face of noise. 2013. arXiv:1312.7006.

[14] S. Chrétien and A. O. Hero. On EM algorithms and their proximal generalizations. *ESAIM: Probability and Statistics*, 12:308–326, 2008.

[15] S. Dasgupta and L. J. Schulman. A probabilistic analysis of em for mixtures of separated, spherical gaussians. *Journal of Machine Learning Research*, 8:203–226, 2007.

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[17] H. O. Hartley. Maximum likelihood estimation from incomplete data. *Biometrics*, 14(2):pp. 174–194, 1958.

[18] M. Healy and M. Westmacott. Missing values in experiments analysed on automatic computers. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 5(3):pp. 203–206, 1956.

[19] D. Hsu and S. M. Kakade. Learning gaussian mixture models: Moment methods and spectral decompositions. *CoRR*, abs/1206.5766, 2012.

[20] S. J. Iturria, R. J. Carroll, and D. Firth. 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, 1999.

[21] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *STOC*, pages 665–674. 2013.

[22] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *Information Theory, IEEE Transactions on*, 56(6):2980–2998, 2010.

[23] V. Koltchinskii. *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, 2011.

[24] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York, NY, 1991.

[25] C. Liu and D. B. Rubin. The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. *Biometrika*, 81:633–648, 1994.

[26] P.-L. Loh and M. J. Wainwright. Corrupted and missing predictors: Minimax bounds for high-dimensional linear regression. In *ISIT*, pages 2601–2605. 2012.

[27] T. A. Louis. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 44:226–233, 1982.

[28] J. Ma and L. Xu. Asymptotic convergence properties of the EM algorithm with respect to the overlap in the mixture. *Neurocomputing*, 68, 2005.

[29] G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics. Wiley, 2007.

[30] I. Meilijson. A fast improvement of the EM algorithm on its own terms. *Journal of the Royal Statistical Society: Series B*, 51:127–138, 1989.

[31] X. L. Meng and D. B. Rubin. Maximum likelihood via the ECM algorithm: a general framework. *Biometrika*, 80:267–278, 1993.

[32] R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 355–368. MIT Press, Cambridge, MA, USA, 1999.

[33] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[34] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Springer, 2004.

[35] P. Netrapalli, P. Jain, and S. Sanghavi. Phase retrieval using alternating minimization. In *NIPS*, pages 2796–2804. 2013.

[36] N. Noorshams and M. J. Wainwright. Stochastic belief propagation: A low-complexity alternative to the sum-product algorithm. *IEEE Transactions on Information Theory*, 59(4):1981–2000, 2013.

[37] T. Orchard and M. A. Woodbury. A missing information principle: theory and applications. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of Statistics*, pages 697–715, 1972.

[38] K. Pearson. *Contributions to the Mathematical Theory of Evolution*. Harrison and Sons, 1894.

[39] R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, 1984.

[40] D. B. Rubin. Characterizing the estimation of parameters in incomplete-data problems. *Journal of the American Statistical Association*, 69(346):pp. 467–474, 1974.

[41] R. Sundberg. Maximum likelihood theory for incomplete data from an exponential family. *Scand. J. Statist*, 1:49–58, 1974.

[42] M. A. Tanner and W. H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:528–550, 1987.

[43] P. Tseng. An analysis of the em algorithm and entropy-like proximal point methods. *Mathematics of Operations Research*, 29(1):pp. 27–44, 2004.

[44] S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

[45] A. W. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, NY, 1996.

[46] D. A. van Dyk and X. L. Meng. Algorithms based on data augmentation: A graphical representation and comparison. In *Computing Science and Statistics: Proceedings of the 31st Symposium on the Interface*, pages 230–239. Berk and M. Pourahmadi, 2000.

[47] R. Vershynin. 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, 2010. arXiv:1011.3027.

[48] G. Wei and M. A. Tanner. 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, 1990.

[49] C. F. J. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

[50] L. Xu and M. I. Jordan. On convergence properties of the EM algorithm for gaussian mixtures. *Neural Comput.*, 8(1):129–151, 1996.

[51] Q. Xu and J. You. Covariate selection for linear errors-in-variables regression models. *Communications in Statistics - Theory and Methods*, 36(2):375–386, 2007.

[52] X. Yi, C. Caramanis, and S. Sanghavi. Alternating minimization for mixed linear regression. 2013. arXiv:1310.3745.