Homeomorphic-Invariance of EM: Non-Asymptotic Convergence KL Divergence for Exponential Families via Mirror Descent (Extended Abstract)*

Frederik Kunstner[1†], **Raunak Kumar**[2] and **Mark Schmidt**[1,3]
[1]University of British Columbia
[2]Cornell University
[3]Canada CIFAR AI Chair (Amii)
kunstner@cs.ubc.ca, raunak@cs.cornell.edu, schmidtm@cs.ubc.ca

Abstract

Expectation maximization (EM) is the default algorithm for fitting probabilistic models with missing or latent variables, yet we lack a full understanding of its non-asymptotic convergence properties. Previous works show results along the lines of "EM converges at least as fast as gradient descent" by assuming the conditions for the convergence of gradient descent apply. This approach is not only loose, in that it does not capture that EM can make more progress than a gradient step, but the assumptions fail to hold for textbook examples of EM like Gaussian mixtures. In this work, we show that for the common setting of exponential family distributions, viewing EM as a mirror descent algorithm leads to convergence rates in Kullback-Leibler (KL) divergence and how the KL divergence is related to first-order stationarity via Bregman divergences. In contrast to previous works, the analysis is invariant to the choice of parametrization and holds with minimal assumptions. We also show applications of these ideas to local linear (and superlinear) convergence rates, generalized EM, and non-exponential family distributions.

1 Introduction

Expectation maximization (EM) is the most common approach to fitting probabilistic models with missing data or latent variables. EM was formalized by Dempster *et al.*, who discussed a wide variety of earlier works that independently discovered the algorithm and domains where EM is used. They already listed multivariate sampling, normal linear models, finite mixtures, variance components, hyperparameter estimation, iteratively reweighted least squares, and factor analysis. To this day, EM continues to be used for these applications and others, like semi-supervised learning [Ghahramani and Jordan, 1994], hidden Markov models [Rabiner, 1989], continuous mixtures [Caron and Doucet, 2008], mixture of experts [Jordan and Xu, 1995], and image reconstruction [Figueiredo and Nowak, 2003]. The many applications of EM have made the work of Dempster *et al.* one of the most influential in the field.

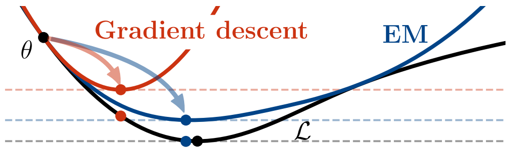

Figure 1: The surrogate optimized by EM is a tighter bound on the objective $\mathcal{L}$ than the quadratic bound implied by smoothness, optimized by gradient descent.

Since the development of EM and subsequent clarifications on the necessary conditions for convergence [Boyles, 1983; Wu, 1983], a large number of works have shown convergence results for EM and its many extensions, leading to a variety of insights about the algorithm, such as the effect of the ratio of missing information [Xu and Jordan, 1996] and the sample size [Wang *et al.*, 2015; Daskalakis *et al.*, 2017; Balakrishnan *et al.*, 2017]. However, existing results on the global, non-asymptotic convergence of EM rely on proof techniques developed for gradient descent on smooth functions, which rely on quadratic upper-bounds on the objective.[1] Informally, this approach argues that the maximization step of the surrogate constructed by EM does at least as well as gradient descent on a quadratic surrogate with a constant step-size, as illustrated in Figure 1.

The use of smoothness as a starting point leads to results that imply that EM behaves as a gradient method with a constant step-size. If true, there would be no difference between EM and its gradient-based variants [Lange *et al.*, 2000]. This does not hold, however, and the resulting convergence rates are inevitably loose; EM makes more progress than this worst-case bound even on simple problems, as shown in Figure 2.

Another issue is that, similarly to how Newton's method is invariant to affine reparametrizations, EM is invariant to any homeomorphism [Varadhan and Roland, 2004]; the steps taken are the same for any continuous, invertible reparametrization. This is not reflected by current analyses because the parametrization influences the smoothness of the function and

[*] This is an extended abstract of a paper that won best award at AISTATS 2021.
[†] Contact Author.

[1] As EM is a maximization algorithm, we should say "gradient ascent" and "lower-bound". We use the language of minimization to make the connections with the optimization literature more explicit.

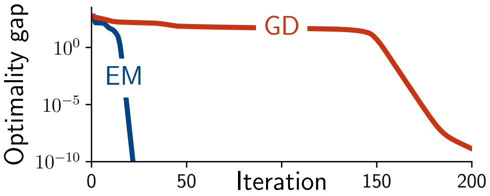

Figure 2: Performance of EM and gradient descent (GD) with constant step-size, selected by grid-search, for a Gaussian mixture model on the Old Faithful dataset. The large gap between the two methods suggests that existing theory for gradient descent is insufficient to explain the performance of EM.

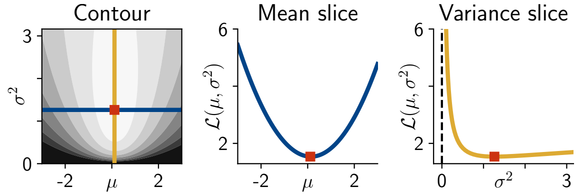

Figure 3: An exponential family distribution that cannot be smooth in Euclidean norm; fitting a Gaussian $\mathcal{N}(\mu, \sigma^2)$, including its variance. As the loss diverges to ∞ as $\sigma \to 0$, the objective cannot be upper-bounded by a quadratic function.

the resulting convergence rate. For these reasons, the general frameworks proposed in the optimization literature [Razaviyayn, 2014; Mairal, 2015] which view EM as a special case, do not reflect that EM is faster than typical members of these frameworks and yield loose analyses.

Most importantly, the assumption that the objective function is bounded by a quadratic does not hold in general. Results relying on smoothness do not apply, for example, to the textbook illustration of EM: Gaussian mixtures with learned covariance matrices [Bishop, 2007; Murphy, 2012]. This is shown in Figure 3. The smoothness assumption is a reasonable simplification for local analyses, as it only needs to hold over a small subspace of the parameter space. In this setting, it does not detract from the main contribution of works investigating statistical properties or large-sample behavior. It does not hold, however, for global convergence analyses with arbitrary initializations. Our focus in this work is analyzing the classic EM algorithm when run for a finite number of iterations on a finite dataset, the setting in which people have been using EM for over 40 years and continue to use today.

We focus on applications of EM to exponential family models, of which Gaussian mixtures are a special case. Exponential families are by far the most common setting and an important special case as the M-step has a closed form solution. Modern stochastic and online extensions of EM also rely on the form of exponential families to efficiently summarize past data [Neal and Hinton, 1998; Sato, 1999; Cappé and Moulines, 2009].

The main tool for the analysis is the Kullback-Leibler (KL) divergence to describe change across iterations. This approach was used for asymptotic convergence [Csiszár and Tusnády, 1984; Chrétien and Hero, 2000; Tseng, 2004] and describe extensions of EM or EM-like algorithms [Banerjee *et al.* 2005, Amid and Warmuth 2020]. But it has not yet been applied to non-asymptotic convergence. Using the KL divergence between distributions rather than the Euclidean distance between their parameters, the results do not rely on invalid smoothness assumptions and are invariant to the choice of parametrization.

Focusing on convergence to a stationary point, as EM objective $\mathcal{L}$ is usually non-convex, an informal summary of the main difference between previous analyses using Euclidean smoothness and our results is that, after T iterations,

$$\textbf{Smoothness: } \min_{t \leq T} \|\nabla \mathcal{L}(\theta_t)\|^2 \leq L \tfrac{\mathcal{L}(\theta_1) - \mathcal{L}^*}{T},$$

$$\textbf{KL divergence: } \min_{t \leq T} \mathrm{KL}[\theta_{t+1} \| \theta_t] \leq \tfrac{\mathcal{L}(\theta_1) - \mathcal{L}^*}{T},$$

where $\mathcal{L}^*$ is the optimal value of the objective, $\mathcal{L}(\theta_1) - \mathcal{L}^*$ is the initial optimality gap and L is the smoothness constant. For non-smooth models, such as Gaussians with learned variances (Fig. 3), $L = \infty$ and the bound is vacuous, whereas bounds in KL divergence do not depend on problem-specific constants.

The key observation for exponential families is that M-step iterations match the moments of the model to the sufficient statistics of the data. We show that EM can be interpreted as a mirror descent update, where each iteration minimizes the linearization of the objective and a KL divergence penalization (rather than the gradient descent update which uses the Euclidean distance between parameters instead). While the connection between EM and exponential families is far from new, as it predates the codification of EM by Dempster *et al.*, the further connection to mirror descent to describe its behavior is, to the best of our knowledge, not acknowledged in the literature. More closely related to general optimization, our work can be seen as an application of the recent perspective of mirror descent as defining smoothness relative to a reference function, as presented by Bauschke *et al.* and Lu *et al.*. Our main results are:

- Show that EM for the exponential family is equivalent to mirror descent, and that the EM objective is relatively smooth in KL divergence.
- Show the first homeomorphic-invariant non-asymptotic EM convergence rate, and how the KL divergence between iterates relates to stationarity and the natural gradient.
- Show how the ratio of missing information affects the non-asymptotic linear (or superlinear) convergence rate of EM around minimizers.
- Extend the results to generalized EM, where the M-step is only solved approximately.
- Discuss how to handle cases where the M-step is not in the exponential family (and might be non-differentiable) by analyzing the E-step.

2 EM and Exponential Families

EM applies when we want to maximize the likelihood $p(x \,|\, \theta)$ of data x given parameters θ, but the likelihood depends on unobserved variables z. By marginalizing over z, we obtain the negative log-likelihood (NLL), that we want to minimize,

$$\mathcal{L}(\theta) = -\log p(x \,|\, \theta) = -\log \int p(x, z \,|\, \theta)\, \mathrm{d}z, \qquad (1)$$

where $p(x, z \mid \theta)$ is the complete-data likelihood. The integral is multi-dimensional if z is, and a summation for discrete values, but we write all cases as a single integral for simplicity. EM is most useful when the complete-data NLL, $-\log p(x, z \mid \theta)$, is a convex function of θ and solvable in closed form if z were known. EM defines the surrogate $Q_\theta(\phi)$, which estimates $\mathcal{L}(\phi)$ using the expected values for the latent variables at θ,

$$\mathcal{L}(\phi) \le Q_\theta(\phi) = -\int \log p(x, z \mid \phi)\, p(z \mid x, \theta)\, \mathrm{d}z, \quad (2)$$

and iteratively updates $\theta_{t+1} \in \arg\min_\phi Q_{\theta_t}(\phi)$. The computation of the surrogate $Q_\theta(\cdot)$ and its minimization are typically referred to as the E-step and M-step. Two fundamental results about EM are that, up to a constant, the surrogate is an upper-bound on the objective and improvement on Q_θ translates to improvement on $\mathcal{L}$, and the gradients of the loss and the surrogate match at the point it is formed, $\nabla Q_\theta(\theta) = \nabla \mathcal{L}(\theta)$.

Many canonical applications of EM, including mixture of Gaussians, are special cases where the complete-data distribution $p(x, z \mid \theta)$ is an exponential family distribution;

$$p(x, z \mid \theta) \propto \exp(\langle S(x, z), \theta\rangle - A(\theta)), \quad (3)$$

where S, θ, and A are the sufficient statistics, natural parameters, and log-partition function. Exponential families are an important special case as the M-step has a closed-form solution. The update depends on the data only through the sufficient statistics, and the minimization of the surrogate reduces to

$$\nabla A(\theta_{t+1}) = \mathbb{E}_{p(z \mid x, \theta_t)}[S(x, z)]. \quad (4)$$

The E-step computes the expected sufficient statistics the M-step finds the parameters θ that satisfy Equation (4).

3 Main Result: EM as Mirror Descent

Although EM iterations strictly decrease the objective function, this does not directly imply convergence to stationary points, even asymptotically [Boyles, 1983; Wu, 1983]. Characterizing the progress to ensure convergence requires additional assumptions. Local analyses typically assume that the EM update contracts the distance to a local minima θ^*,

$$\|\theta_{t+1} - \theta^*\| \le c\|\theta_t - \theta^*\|,$$

for some $c < 1$. On the other hand, global analyses typically assume the surrogate is *smooth* in Euclidean norm,

$$\|\nabla Q.(\theta) - \nabla Q.(\phi)\| \le L\|\theta - \phi\|,$$

for all θ and ϕ, and some fixed constant L. This is equivalent to assuming the following upper bound holds,

$$\mathcal{L}(\phi) \le \mathcal{L}(\theta) + \langle \nabla \mathcal{L}(\theta), \phi - \theta\rangle + L\frac{1}{2}\|\theta - \phi\|^2.$$

Such assumptions are reasonable for local analyses, but the worst-case value of L for global results can be infinite, as in the simple example of Figure 3. Instead, we show that the following upper-bound in KL divergence holds.

PROPOSITION 1. *For exponential family distributions, the* M*-step update in Expectation-Maximization is equivalent to the minimization of the following upper-bound;*

$$\mathcal{L}(\phi) \le \mathcal{L}(\theta) + \langle \nabla \mathcal{L}(\theta), \phi - \theta\rangle + D_A(\phi, \theta), \quad (5)$$

where A is the log-partition of the complete-data distribution, and $D_A(\phi, \theta) = \mathrm{KL}[p(x, z \mid \theta) \| p(x, z \mid \phi)]$ is the Bregman divergence induced by A;

$$D_A(\phi, \theta) = A(\phi) - A(\theta) - \langle \nabla A(\theta), \phi - \theta\rangle.$$

While the upper bound is expressed in a specific parametrization to specify the distributions, the KL divergence is a property of the distributions, independent of their representation. As this upper-bound is the one minimized by the M-step, it is a direct description of the algorithm rather than an additional surrogate used for convenience, as was illustrated in Figure 1.

This gives an interpretation of EM in terms of the mirror descent algorithm [Nemirovski and Yudin, 1983], the minimization of a first-order Taylor expansion and Bregman divergence as in Equation (5), with step-size $\alpha = 1$. In the recent perspective of mirror descent framed as relative smoothness [Bauschke *et al.*, 2017; Lu *et al.*, 2018], the objective function is 1-smooth, relative to A. Existing results [Lu *et al.* 2018, Theorem 3.1] then directly imply the following local result, when the algorithm enters a convex region (up to non-degeneracy assumptions).

COROLLARY 1. *For exponential families, if* EM *is initialized in a locally-convex region with minimum θ^*,*

$$\mathcal{L}(\theta_T) - \mathcal{L}(\theta^*) \le \frac{1}{T}\mathrm{KL}[p(x, z \mid \theta_1) \| p(x, z \mid \theta^*)]. \quad (6)$$

And adapting the proofs to the non-convex setting yields the first global, non-asymptotic convergence rate for EM to stationary points that does not depend on problem-specific constants.

PROPOSITION 2. EM *for exponential family distributions converges at the rate*

$$\min_{t \le T} \mathrm{KL}[p(x, z \mid \theta_{t+1}) \| p(x, z \mid \theta_t)] \le \frac{\mathcal{L}(\theta_1) - \mathcal{L}^*}{T}.$$

This result implies that the distribution fit by EM stops changing, but it does not—in itself—guarantee progress toward a stationary point, as it would also be satisfied by an algorithm that does not move, if $\theta_{t+1} = \theta_t$. In the Euclidean setting of gradient descent with constant step-size, Proposition 2 is the equivalent of the statement that the distance between iterates $\|\theta_{t+1} - \theta_t\|$ converges. As $\|\theta_{t+1} - \theta_t\| \propto \|\nabla \mathcal{L}(\theta_t)\|$, it also implies that the gradient norm converges. A similar result holds for EM, where measuring distances between iterates with D_A leads to stationarity in the dual divergence D_{A^*}.

A useful simplification to interpret the divergence in Proposition 2 is to consider a locally equivalent norm. By a second-order Taylor expansion, it can be shown that

$$\mathrm{KL}[p(x, z \mid \theta_{t+1}) \| p(x, z \mid \theta_t)] \approx \|\nabla \mathcal{L}(\theta_t)\|^2_{I_{x,z}(\theta_t)^{-1}},$$

where $I_{x,z}(\theta_t)$ is the Fisher information of the full data distribution, $p(x, z \mid \theta)$. This quantity is the analog of the Newton decrement used in the affine-invariant analysis of Newton's method [Nesterov and Nemirovski, 1994], but for the natural gradient direction [Amari and Nagaoka, 2000] rather than the Newton direction. While the Newton decrement is invariant to affine reparametrizations, this "natural decrement" is also invariant to any homeomorphism.

4 Additional Results

This section summarizes additional results presented in the full paper. Instead of assuming that the objective is smooth in Euclidean norm and applying the methodology for the convergence of gradient descent, which does not hold even for the standard Gaussian mixture examples found in textbooks, we show that EM for exponential families can be viewed as mirror descent and analyzed through smoothness relative to a KL divergence. This perspective leads to convergence rates that hold with minimal assumptions, extends known asymptotic result to the non-asymptotic regime, handles approximate solutions to the M-step, and can be extended to analyse the E-step for non-exponential families and describe stochastic variants.

4.1 Required Assumptions

The results presented here hold for regular exponential families, as long as the updates are well defined. A subtle issue is that the updates might lead to degenerate probability distributions. For example, in the case of a Gaussian mixture, we can drive the NLL to $-\infty$ by putting the mean of one cluster on a single point and driving its variance to 0. Such degenerate solutions are challenging for non-asymptotic rates, as results typically depend on the optimality gap $\mathcal{L}(\theta) - \mathcal{L}^*$ and are vacuous if it is unbounded. However, those issues can be dealt with by adding proper regularization, for example by using Maximum a Posteriori estimation under an appropriate prior.

4.2 Local Linear Rates

It was already established by Dempster *et al.* that, asymptotically, the EM algorithm converges r-linearly, meaning that near a strict minima θ^*, $\|\theta_{t+1} - \theta^*\| \leq r\|\theta_t - \theta^*\|$. The rate of convergence r is determined by the amount of "missing information" at θ^* that measures how much information is missing from not knowing z [Orchard and Woodbury, 1972]. For a mixture of Gaussians, r would be small (and convergence fast) if we find well-separated clusters, as there is little uncertainty about the latent variables (the cluster membership), and convergence would be fast.

This result, however, is only asymptotic. Existing non-asymptotic linear rates rely on strong-convexity assumptions [e.g. Balakrishnan *et al.* 2017], and show a linear rate of convergence in strongly-convex regions. But the results depend on the eigenvalues of the Hessian rather than the ratio of missing information. As in Proposition 1, which shows that the EM objective is 1-smooth relative to its log-partition function A rather than measure smoothness in Euclidean norms, we can characterize strong convexity relative to the log-partition function. Existing results on the convergence of mirror descent for relatively-smooth, relatively-strongly convex functions [Lu *et al.*, 2018] then directly give that, if in a relatively strongly-convex region, EM converges non-asymptotically at the rate $\mathcal{L}(\theta_{t+1}) - \mathcal{L}^* \leq r(\mathcal{L}(\theta_t) - \mathcal{L}^*)$, with the same missing information ratio. If the ratio of missing information diminishes with each iterations as we find clusters that better explain the data, the convergence rate improves and EM can converge superlinearly for well-separated clusters [Salakhutdinov *et al.*, 2003; Xu and Jordan, 1996].

4.3 Inexact Variants

The analyses extend to generalized EM schemes, which do not optimize the surrogate exactly in the M-step but output an approximate (possibly randomized) update.

We consider multiplicative error, where the algorithm is guaranteed to make at least some fraction of the progress of a full M-step, and additive error, where the algorithm can make arbitrary mistakes but has to eventually improve. An example of multiplicative error for mixture models is the exact optimization of only one of the mixture components, chosen at random, like the ECM algorithm of Meng and Rubin. For additive error, although suboptimal for the reasons mentioned earlier, running GD with a line-search on the surrogate guarantees additive error if the objective is (locally) smooth. The convergence rate of the algorithm is preserved, suffering a penalty depending on the error. For multiplicative error, the rate degrades by a multiplicative factor proportional to the fraction of progress made, while for additive error, the rate has an additional term depending on the average of all errors, which has to go down to 0 to ensure convergence.

4.4 EM for General Models

While exponential families cover many applications, some are not smooth, in Euclidean distance or otherwise. For example, a mixture of Laplace distributions leads to surrogates with discontinuous gradients (the Laplace distribution is not an exponential family). In this case, the progress need not be related to the gradient and Proposition 2 does not hold.

The tools presented here can still obtain partial answers. The analyses in previous sections considered the progress of the M-step, but we can instead focus on the E-step as the primary driver of progress. Looking at the KL divergence between distributions on the latent variables only, $p(z \mid x, \theta)$, an analog of Proposition 2 holds for stationarity on the latent variables, rather than the complete-data distribution. This guarantee is weaker, but the assumption holds more generally. For example, it is satisfied by any finite mixture, even if the mixture components are non-differentiable.

4.5 Stochastic Variants

This perspective extends to stochastic approximation [Robbins and Monro, 1951] variants of EM, which are becoming increasingly relevant as they scale to large datasets. Algorithms such as stochastic and online EM [Sato, 1999; Cappé and Moulines, 2009] average the observed sufficient statistics to update the parameters. This can be cast as stochastic mirror descent [Nemirovski *et al.*, 2009] with step-sizes decreasing as $1/t$, while incremental EM [Neal and Hinton, 1998] and other variance-reduced variants [Chen *et al.*, 2018; Karimi *et al.*, 2019] can be viewed as applications of variance reduction methods like SAG or MISO [Le Roux *et al.*, 2012; Mairal, 2015] to mirror descent. However, analyses in those settings still view of EM as a preconditioned gradient step, using smoothness to handle the stochasticity. In some cases, those works prescribe a step-size a proportional to $1/L$ which, for Gaussian mixtures, implies using a step-size of 0. We hope the tools developed here may help to fix this and similar practical issues.

References

[Amari and Nagaoka, 2000] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*. Oxford University Press, 2000.

[Amid and Warmuth, 2020] Ehsan Amid and Manfred K. Warmuth. Divergence-based motivation for online EM and combining hidden variable models. In *UAI*, volume 124, pages 81–90, 2020.

[Balakrishnan *et al.*, 2017] Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Ann. Stat.*, 45(1):77–120, 2017.

[Banerjee *et al.*, 2005] Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *JMLR*, 6:1705–1749, 2005.

[Bauschke *et al.*, 2017] Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications. *Math. Oper. Res.*, 42(2):330–348, 2017.

[Bishop, 2007] Christopher M. Bishop. *Pattern recognition and machine learning, 5th Edition*. Springer, 2007.

[Boyles, 1983] Russell A. Boyles. On the convergence of the EM algorithm. *J. R. Stat. Soc. Ser. B Methodol.*, 45(1):47–50, 1983.

[Cappé and Moulines, 2009] Olivier Cappé and Eric Moulines. Online expectation–maximization algorithm for latent data models. *J. R. Stat. Soc. Ser. B Methodol.*, 71(3):593–613, 2009.

[Caron and Doucet, 2008] François Caron and Arnaud Doucet. Sparse Bayesian nonparametric regression. In *ICML*, pages 88–95, 2008.

[Chen *et al.*, 2018] Jianfei Chen, Jun Zhu, Yee Whye Teh, and Tong Zhang. Stochastic expectation maximization with variance reduction. In *NeurIPS*, pages 7978–7988, 2018.

[Chrétien and Hero, 2000] Stéphane Chrétien and Alfred O. Hero. Kullback proximal algorithms for maximum likelihood estimation. *IEEE Trans. Inf. Theory*, 46(5):1800–1810, 2000.

[Csiszár and Tusnády, 1984] Imre Csiszár and Gábor Tusnády. Information geometry and alternating minimization procedures. *Stat. decis.*, 1:205–237, 1984.

[Daskalakis *et al.*, 2017] Constantinos Daskalakis, Christos Tzamos, and Manolis Zampetakis. Ten steps of EM suffice for mixtures of two gaussians. In *PMLR*, volume 65, pages 704–710, 2017.

[Dempster *et al.*, 1977] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Methodol.*, 39(1):1–38, 1977.

[Figueiredo and Nowak, 2003] Mário A. T. Figueiredo and Robert D. Nowak. An EM algorithm for wavelet-based image restoration. *IEEE Trans. Image Process.*, 12(8):906–916, 2003.

[Ghahramani and Jordan, 1994] Zoubin Ghahramani and Michael I. Jordan. Supervised learning from incomplete data via an EM approach. In *NeurIPS*, pages 120–127, 1994.

[Jordan and Xu, 1995] Michael I. Jordan and Lei Xu. Convergence results for the EM approach to mixtures of experts architectures. *Neural Networks*, 8(9):1409–1431, 1995.

[Karimi *et al.*, 2019] Belhal Karimi, Hoi-To Wai, Eric Moulines, and Marc Lavielle. On the global convergence of (fast) incremental expectation maximization methods. In *NeurIPS*, pages 2837–2847, 2019.

[Lange *et al.*, 2000] Kenneth Lange, David R. Hunter, and Ilsoon Yang. Optimization transfer using surrogate objective functions. *J. Comput. Graph. Stat.*, 9(1):1–20, 2000.

[Le Roux *et al.*, 2012] Nicolas Le Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *NeurIPS*, pages 2672–2680, 2012.

[Lu *et al.*, 2018] Haihao Lu, Robert M. Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM J. Optim*, 28(1):333–354, 2018.

[Mairal, 2015] Julien Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM J. Optim*, 25(2):829–855, 2015.

[Meng and Rubin, 1993] Xiao-Li Meng and Donald B. Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–278, 1993.

[Murphy, 2012] Kevin P. Murphy. *Machine learning: A probabilistic perspective*. MIT Press, 2012.

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

[Nemirovski and Yudin, 1983] Arkadi Nemirovski and David B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley, NY, 1983. translated by E.R. Dawson.

[Nemirovski *et al.*, 2009] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4):1574–1609, 2009.

[Nesterov and Nemirovski, 1994] Yurii Nesterov and Arkadi Nemirovski. *Interior-Point Polynomial Algorithms in Convex Programming*. SIAM, 1994.

[Orchard and Woodbury, 1972] Terence Orchard and Max A. Woodbury. A missing information principle: theory and applications. In *Berkeley Symp. on Math. Stat. and Prob.*, pages 697–715, 1972.

[Rabiner, 1989] Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286, 1989.

[Razaviyayn, 2014] Meisam Razaviyayn. *Successive convex approximation: Analysis and applications*. PhD thesis, University of Minnesota, 2014.

[Robbins and Monro, 1951] Herbert Robbins and Sutton Monro. A stochastic approximation method. *Ann. Math. Stat.*, 22(3):400–407, 1951.

[Salakhutdinov *et al.*, 2003] Ruslan Salakhutdinov, Sam T. Roweis, and Zoubin Ghahramani. Optimization with EM and expectation-conjugate-gradient. In *ICML*, pages 672–679, 2003.

[Sato, 1999] Masa-aki Sato. Fast learning of on-line EM algorithm. Technical report, ATR Human Inf. Proc. Res. Lab., 1999.

[Tseng, 2004] Paul Tseng. An analysis of the EM algorithm and entropy-like proximal point methods. *Math. Oper. Res.*, 29(1):27–44, 2004.

[Varadhan and Roland, 2004] Ravi Varadhan and Christophe Roland. Squared extrapolation methods (SQUAREM): A new class of simple and efficient numerical schemes for accelerating the convergence of the EM algorithm. Technical Report 63, Johns Hopkins, 2004.

[Wang *et al.*, 2015] Zhaoran Wang, Quanquan Gu, Yang Ning, and Han Liu. High dimensional EM algorithm: Statistical optimization and asymptotic normality. In *NeurIPS*, pages 2521–2529, 2015.

[Wu, 1983] C. F. Jeff Wu. On the convergence properties of the EM algorithm. *Ann. Stat.*, 11(1):95–103, 1983.

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