Homeomorphic-Invariance of EM: Non-Asymptotic Convergence in KL Divergence for Exponential Families via Mirror Descent

Frederik Kunstner
University of British Columbia

Raunak Kumar
Cornell University

Mark Schmidt
University of British Columbia
Canada CIFAR AI Chair (Amii)

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 to EM. 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 first 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. Then, we show 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. (1977), 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), image reconstruction (Figueiredo and Nowak, 2003), and graphical models (Lauritzen, 1995). The many applications of EM have made the work of Dempster et al. one of the most influential in the field.

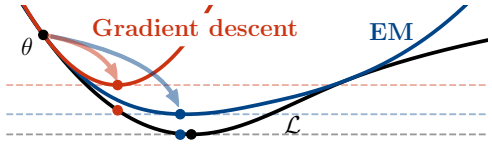

Figure 1: The surrogate optimized by EM is a tighter bound on the objective $\mathcal{L}$ than the the quadratic bound implied by smoothness, optimized by gradient descent.

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; Ma et al., 2000) and the sample size (Wang et al., 2015; Yi and Caramanis, 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.

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

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

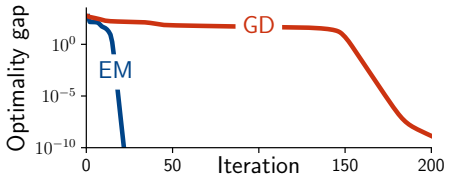

Figure 2: Performance of EM and gradient descent (GD) with constant step-size, selected by grid-search, for a Gaussian mixture model on the Old Faithful dataset. The large gap between the two methods suggests that existing theory for gradient descent is insufficient to explain the performance of EM.

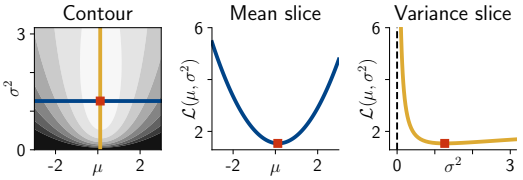

Figure 3: An exponential family distribution that cannot be smooth; fitting a Gaussian $\mathcal{N}(\mu, \sigma^2)$, including its variance. As the loss diverges to ∞, the objective cannot be upper-bounded by a quadratic function.

The use of smoothness as a starting assumption 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 (e.g. 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 by EM are the same for any continuous, invertible reparametrization. This is not reflected by current analyses because the parametrization of the problem influences the smoothness of the function and the resulting convergence rate. For these reasons, the general frameworks proposed in the optimization literature (Xu and Yin, 2013; Mairal, 2013; Razaviyayn, 2014; Paquette et al., 2018) where EM is 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 standard textbook illustration of EM: Gaussian mixtures with learned covariance matrices (Bishop, 2007; Murphy, 2012). This is shown in Figure 3. The smoothness assumption might be 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 the application 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 extension of EM also rely on the form of exponential families to efficiently summarize information about past data (Neal and Hinton, 1998; Sato, 1999; Delyon et al., 1999; Cappé and Moulines, 2009).

The main tool we use for the analysis is the Kullback-Leibler (KL) divergence to describe distances between parameters. This approach was initially used to derive asymptotic convergence results (Csiszár and Tusnády, 1984; Chrétien and Hero, 2000; Tseng, 2004) but has, to the best of our knowledge, not yet been applied to non-asymptotic analyses. By characterizing distances in KL divergence between the distributions induced by the parameters, rather than their Euclidean distance, 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 the EM objective $\mathcal{L}$ is non-convex, an informal summary of the main difference between previous analyses using smoothness and our results is that, after T iterations,

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

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

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 covariances (Fig. 3), $L = \infty$ and the bound is vacuous, whereas bounds in KL divergence do not depend on problem-specific constants. We show how the KL divergence relates to stationarity conditions for non-degenerate problems in Section 5.

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 in this setting, EM can be interpreted as a mirror descent update, where each iteration minimizes the linearization of the objective and a KL divergence penalization term (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. (1977) (Blight, 1970), 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. (2017) and Lu et al. (2018).

Our main results are that we:

- Show that EM for the exponential family is a mirror descent algorithm, 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 is related to stationary points 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 EXPECTATION-MAXIMIZATION AND EXPONENTIAL FAMILIES

Before stating our results, we introduce the EM algorithm and necessary background on exponential families. For completeness, we provide additional details in Appendix A and refer the reader to Wainwright and Jordan (2008) for a full treatment of the subject.

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

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

where $p(x, z \mid \theta)$ is the complete-data likelihood. The integral here 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,

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

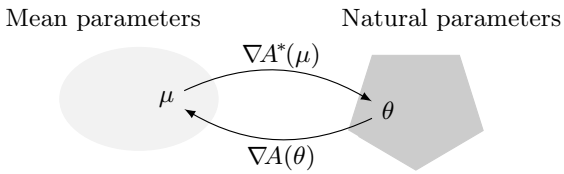

Figure 4: The gradient of the log-partition function and its dual, $(\nabla A, \nabla A^*)$, form a bijection between the natural and mean parameters θ, μ.

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.

A useful decomposition of the surrogate, shown by Dempster et al. (1977), is the equality

$$\begin{aligned} &Q_\theta(\phi) = \mathcal{L}(\phi) + H_\theta(\phi), \\ &\text{where } H_\theta(\phi) = -\int \log p(z \mid x, \phi)\, p(z \mid x, \theta)\, \mathrm{d}z \end{aligned} \qquad (2)$$

is an entropy-like term minimized at $\phi = \theta$. That is,

$$0 \le H_\theta(\theta) \le H_\theta(\phi) \qquad \text{and} \qquad \nabla H_\theta(\theta) = 0.$$

This gives two fundamental results about EM. As $H_\theta(\cdot) \ge 0$, 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, $\nabla Q_\theta(\theta) = \nabla \mathcal{L}(\theta)$.

2.1 Exponential families

Many canonical applications of EM, including mixture of Gaussians, are special cases where the complete-data distribution, given a value for the latent variable z, is an exponential family distribution;

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

where S, θ, and A are the sufficient statistics, natural parameters, and log-partition function of the distribution. Exponential family models are an important special case as the M-step has a closed form solution, and the update depends on the data only through the sufficient statistics. The solution for the maximum likelihood estimate (MLE) given x and z can be found from the stationary point of the complete log-likelihood,

$$\nabla \log p(x, z \mid \theta) = S(x, z) - \nabla A(\theta) = 0. \qquad (4)$$

The gradient ∇A yields the expected sufficient statistics, $\nabla A(\theta) = \mathbb{E}_{p(x,z \mid \theta)}[S(x, z)]$, also called mean parameters and denoted by μ. The log-partition function defines a bijection between the natural and mean parameters. Its inverse is given by ∇A^*, the gradient of the convex conjugate of A, $A^*(\mu) = \sup_\theta \{\langle \theta, \mu \rangle - A(\theta)\}$, such that $\mu = \nabla A(\theta)$ and $\theta = \nabla A^*(\mu)$, as illustrated in Figure 4. The solution to Equation (4), given by

$$\nabla A(\theta) = S(x, z) \qquad \Longrightarrow \qquad \theta = \nabla A^*(S(x, z)),$$

is called moment matching as its finds the parameter that, in expectation, generates the observed statistics.

To connect EM and mirror descent, we use the Bregman divergence induced by a convex function h; the difference between the function and its linearization,

$$D_h(\phi, \theta) = h(\phi) - h(\theta) - \langle \nabla h(\theta), \phi - \theta \rangle. \quad (5)$$

For exponential families, the Bregman divergence induced by the log-partition A is the KL divergence

$$D_A(\phi, \theta) = \mathrm{KL}[p(x, z \,|\, \theta) \| p(x, z \,|\, \phi)].$$

The Bregman divergences induced by A and its conjugate A^* have the following relation (note the ordering)

$$D_A(\phi, \theta) = D_{A^*}(\nabla A(\theta), \nabla A(\phi)). \quad (6)$$

Both expressions are the same KL divergence, but differ in the parametrization used to express the distributions.

3 EM AND MIRROR DESCENT

Although EM iterations strictly decrease in the objective function if such decrease is possible locally, this does not directly imply convergence to stationary points, even asymptotically (Boyles, 1983; Wu, 1983). The progress at each step could decrease faster than the objective. 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^*,

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

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

$$\|\nabla Q_\cdot(\theta) - \nabla Q_\cdot(\phi)\| \leq L \|\theta - \phi\|.$$

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

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

While the local analyses assumptions are reasonable, 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 without additional assumptions.

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

$$\mathcal{L}(\phi) \leq \mathcal{L}(\theta) + \langle \nabla \mathcal{L}(\theta), \phi - \theta \rangle + D_A(\phi, \theta), \quad (7)$$

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

While the upper bound is still expressed in a specific parametrization to describe 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; Beck and Teboulle, 2003), the minimization of a first-order Taylor expansion and Bregman divergence as in Equation (7), 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 (e.g. Lu et al., 2018, Theorem 3.1) then directly imply the following local result, up to non-degeneracy assumptions **A1**–**A3** discussed in the next section.

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

$$\mathcal{L}(\theta_T) - \mathcal{L}(\theta^*) \leq \frac{1}{T} \mathrm{KL}[p(x, z \,|\, \theta_1) \| p(x, z \,|\, \theta^*)]. \quad (8)$$

To the best of our knowledge, this is the first non-asymptotic convergence rate for EM that does not depend on problem-specific constants.

Proof of Proposition 1. Recall the decomposition of the surrogate in terms of the objective and entropy term, $Q_\theta(\phi) = \mathcal{L}(\phi) + H_\theta(\phi)$ in Equation (2). It gives

$$\mathcal{L}(\phi) - \mathcal{L}(\theta) = Q_\theta(\phi) - Q_\theta(\theta) + H_\theta(\theta) - H_\theta(\phi),$$

where $H_\theta(\theta) - H_\theta(\phi) \leq 0$ as $H_\theta(\phi)$ is minimized at $\phi = \theta$. We will show that for exponential families,

$$Q_\theta(\phi) - Q_\theta(\theta) = \langle \nabla \mathcal{L}(\theta), \phi - \theta \rangle + D_A(\phi, \theta),$$

which implies the upper-bound in Equation (7) and that its minima matches that of $Q_\theta(\phi)$.

If the complete-data distribution is in the exponential family, the surrogate in natural parameters is

$$\begin{aligned} Q_\theta(\phi) &= -\int \log p(x, z \,|\, \phi)\, p(z \,|\, x, \theta)\, \mathrm{d}z, \\ &= -\int [\langle S(x, z), \phi \rangle - A(\phi)]\, p(z \,|\, x, \theta)\, \mathrm{d}z, \\ &= -\big\langle \mathbb{E}_{p(z \,|\, x, \theta)}[S(x, z)], \phi \big\rangle + A(\phi). \end{aligned} \quad (9)$$

Using $s(\theta) = \mathbb{E}_{p(z \,|\, x, \theta)}[S(x, z)]$ for the expected sufficient statistics[2] and expanding $Q_\theta(\phi) - Q_\theta(\theta)$ yields

$$\begin{aligned} Q_\theta(\phi) - Q_\theta(\theta) &= -\langle s(\theta), \phi - \theta \rangle + A(\phi) - A(\theta), \\ &\overset{(1)}{=} -\langle s(\theta) - \nabla A(\theta), \phi - \theta \rangle + D_A(\phi, \theta), \\ &\overset{(2)}{=} \langle \nabla \mathcal{L}(\theta), \phi - \theta \rangle + D_A(\phi, \theta), \end{aligned}$$

where (1) adds and subtracts $\langle \nabla A(\theta), \phi - \theta \rangle$ to com-

[2]The sufficient statistics $s(\theta)$ also depend on x. We do not write $s(\theta, x)$ as x is fixed and the same at each iteration.

plete the Bregman divergence and (2) uses that the gradient of the surrogate and the objective match at θ,

$$\nabla\mathcal{L}(\theta) = \nabla Q_\theta(\theta) = \nabla A(\theta) - s(\theta). \qquad \square$$

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 incremental, stochastic and online EM (Neal and Hinton, 1998; Sato, 1999; Delyon et al., 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 $1/t$. For brevity, we leave the derivation to Appendix B.

4 ASSUMPTIONS AND OPEN CONSTRAINTS

Before diving into convergence results, we discuss the assumptions needed for the method to be well defined.

A1 The complete-data distribution $p(x, z \mid \theta)$ is a minimal exponential family distribution; no two parameters lead to the same distribution.

A1 implies the continuity and differentiability of $\mathcal{L}$, the convexity of the surrogate and that natural and mean parameters are well defined. The minimality assumption ensures the log-partition function A is strictly convex, which is the common assumption in the mirror descent literature that the mirror map is strictly convex. It implies the mappings $\nabla A, \nabla A^*$ are unique and that the surrogate has a unique solution. **A1** is not strictly necessary, as similar results can be derived with regularization, but greatly simplifies the presentation.

The next assumptions deal with a further subtle issue that arises when we attempt to apply results from the optimization literature to EM, like the generic frameworks of Xu and Yin (2013), Mairal (2013) or Razaviyayn (2014). The parameters of the distributions optimized by EM are typically constrained to a subset $\theta \in \Omega$, like that probabilities sum to one and that covariance matrices are positive-definite. To handle constraints, those analyses assume access to a projection onto the constraint set Ω. However, this does not hold for common settings of EM like mixtures of Gaussians. When the boundaries of the constraint set are open, the projection operator does not exist (there is no "closest positive-definite matrix" to a matrix that is not positive-definite). An additional complication is related to the existence of a lower-bound on the objective. For example, in Gaussian mixtures, we can drive the objective to $-\infty$ by centering a Gaussian on a single data point and shrinking the variance towards zero. The existence of such degenerate solutions is challenging for non-asymptotic convergence rates, as results typically depend on the optimality gap $\mathcal{L}(\theta) - \mathcal{L}^*$ and are vacuous if it is unbounded. To avoid those degenerate cases, we make the following assumptions.

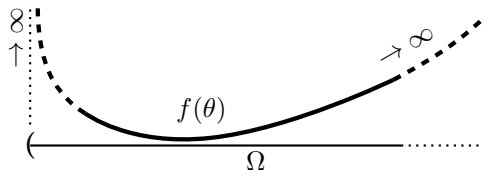

Figure 5: Example of a barrier function with compact sub-level sets on an open set Ω, satisfying **A2** and **A3**. Even if Ω is open, as f goes to ∞ at the boundary and is convex, the minimum is guaranteed to be in Ω.

A2 The objective function is lower-bounded by some $\mathcal{L}^* > -\infty$ on the constraint set Ω.

A3 The sub-level sets $\Omega_\theta = \{\phi \in \Omega : Q_\theta(\phi) \leq Q_\theta(\theta)\}$ are compact (closed and bounded).

One approach to ensure the EM updates are well-defined is to add regularization, in the form of a proper conjugate prior. If the parameters approach the boundary (or diverge in an unbounded direction), the prior acts as a barrier and diverges to ∞ rather than $-\infty$. The minimum of the surrogate is then finite and in Ω at every iteration, without the need for projections. This is illustrated in Figure 5. For simplicity of presentation, we assume **A2** and **A3** hold and discuss maximum a posteriori (MAP) estimation in Appendix C.

5 CONVERGENCE OF EM FOR EXPONENTIAL FAMILIES

We now give the main results for the convergence of EM to stationary points for exponential families. This analysis takes advantage of existing tools for the analysis of mirror descent, but in the less-common non-convex setting. Detailed proofs are deferred to Appendix D.

PROPOSITION 2. *Under assumptions **A1**–**A3**, EM for exponential family distributions converges at the rate*

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

While this result implies the distribution fit by EM stops changing, it does not—in itself—guarantee progress toward a stationary point as it is also satisfied by an algorithm that does not move, $\theta_{t+1} = \theta_t$. In the standard 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^*}.

Recall that the M-step finds a stationary point of the upper-bound in Equation (7). Setting its derivative to 0, using $\nabla_\phi D_A(\phi, \theta_t) = \nabla A(\phi) - \nabla A(\theta_t)$ yields

$$\nabla \mathcal{L}(\theta_t) - \nabla A(\theta_t) + \nabla A(\theta_{t+1}) = 0.$$

Using the expansion of the gradient in terms of the observed statistics $s(\theta_t)$ and the mean parametrization $\mu_t = \nabla A(\theta_t)$, we obtain the moment matching update; finding the mean parameters μ_{t+1} that generate the observed sufficient statistics $s(\theta_t)$ in expectation;

$$\mu_{t+1} = s(\theta_t) = \mu_t - \nabla \mathcal{L}(\theta_t).$$

Expressing the KL divergence as the dual Bregman divergence $D_{A^*}(\mu_{t+1}, \mu_t)$ (Equation 6) then gives

$$\mathrm{KL}[\theta_{t+1} \| \theta_t] = D_{A^*}(\mu_{t+1}, \mu_t) = D_{A^*}(s(\theta_t), \mu_t).$$

This adds a measure of stationarity to Proposition 2;

Corollary 2. *Under assumptions **A1**–**A3**,*

$$\min_{t \le T} D_{A^*}(s(\theta_t), \mu_t) \le \frac{\mathcal{L}(\theta_1) - \mathcal{L}^*}{T}.$$

The observed sufficient statistics $s(\theta_t)$ and mean parameters μ_t are the two parts of the gradient, $\nabla \mathcal{L}(\theta_t) = \mu_t - s(\theta_t)$, and $D_{A^*}(s(\theta_t), \mu_t) = 0$ implies $\nabla \mathcal{L}(\theta_t) = 0$.

Corollary 2 is the Bregman divergence analog of the standard result for steepest descent in an arbitrary norm $\|\cdot\|$, giving convergence in the dual norm $\|\nabla \mathcal{L}\|_*$. If the smoothness assumption is satisfied with constant L, we recover existing results in Euclidean norm,

$$\min_{t \le T} \|\nabla \mathcal{L}(\theta_t)\|^2 \le \frac{L}{T}(\mathcal{L}(\theta_1) - \mathcal{L}^*),$$

as the L-smoothness of A implies the $1/L$-strong convexity of A^* and $D_{A^*}(\mu_{t+1}, \mu_t) \ge 1/L \, \|\nabla \mathcal{L}(\theta_t)\|^2$.

The convergence in KL divergence, however, does not depend on additional smoothness assumptions and is a stronger guarantee as it implies the probabilistic models being optimized stop changing. This can not be directly guaranteed by small gradient norms, as differences in distributions do not only depend on the difference between parameters. For example, how much a Gaussian distribution changes when changing the mean depends on its variance; if the variance is small, the change will be big, but if the variance is large, the change will be comparatively smaller. This is illustrated in Figure 6, and is not captured by gradient norms.

5.1 Connection to the Natural Gradient

A useful simplification to interpret the divergence is to consider the norm it is locally equivalent to. By a second-order Taylor expansion, we have that

$$D_{A^*}(\mu + \delta, \mu) \approx \|\delta\|^2_{\nabla^2 A^*(\mu)},$$

where $\|\delta\|^2_{\nabla^2 A^*(\mu)} = \langle \delta, \nabla^2 A^*(\mu) \delta \rangle$. For exponential families, $\nabla^2 A^*$ is the inverse of the Fisher information matrix of the complete-data distribution $p(x, z \mid \theta)$,

$$\nabla^2 A^*(\mu) = I_{x,z}(\theta)^{-1} = \mathbb{E}_{x,z \sim p(x,z \mid \theta)}\left[\nabla^2 \log p(x, z \mid \theta)\right].$$

The left side of Corollary 2 is then, locally,

$$D_{A^*}(\mu_t - \nabla \mathcal{L}(\theta_t), \mu_t) \approx \|\nabla \mathcal{L}(\theta_t)\|^2_{I_{x,z}(\theta_t)^{-1}}. \tag{10}$$

This quantity is the analog of the Newton decrement,

$$\|\nabla \mathcal{L}(\theta_t)\|^2_{\nabla^2 \mathcal{L}(\theta_t)^{-1}},$$

used in the affine-invariant analysis of Newton's method (Nesterov and Nemirovski, 1994). But Equation (10) is for the natural gradient in information geometry, $I_{x,z}(\theta_t)^{-1} \nabla \mathcal{L}(\theta_t)$ (Amari and Nagaoka, 2000). While the Newton decrement is invariant to affine reparametrizations, this "natural decrement" is also invariant to any homeomorphism.

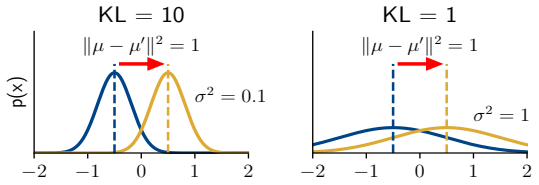

Figure 6: The similarity between two Gaussians depends on their variance, even if it is fixed. The Euclidean distance between parameters, and by extension gradient norms, is a poor measure of stationarity as it ignores unchanged parameters.

5.2 Invariant Local Linear Rates

It was already established by Dempster et al. (1977) that, asymptotically, the EM algorithm converges r-linearly, meaning that if it is in a convex region,[3]

$$\mathcal{L}(\theta_{t+1}) - \mathcal{L}^* \le r(\mathcal{L}(\theta_t) - \mathcal{L}^*) \qquad \text{for } r < 1, \tag{11}$$

near a strict minima θ^*, where the rate r is determined by the amount of "missing information". In this section, we strengthen Corollary 1 to show this result extends to local but non-asymptotic rates.

The improvement ratio r is determined by the eigenvalues of the missing information matrix M at θ^*, defined as (Orchard and Woodbury, 1972)

$$M(\theta^*) = I_{x,z}(\theta^*)^{-1} I_{z \mid x}(\theta^*), \tag{12}$$

where $I_{x,z}(\theta)$ and $I_{z \mid x}(\theta)$ are the Fisher information matrices of the complete-data distribution $p(x, z \mid \theta)$ and conditional missing-data distribution $p(z \mid x, \theta)$. Intuitively this matrix measures how much information is missing and how much easier the problem would be if we had access to the true values of the latent variables.

[3] The result of Dempster et al. (1977) concerned the convergence of the distance to the optimum, $\|\theta_t - \theta^*\|$, but we use function values for simplicity, as it also applies.

If $I_{z\,|\,x}(\theta)$ is small, there is little information to be gained from observing the latent variables as most of this information is already contained in x. But $I_{z\,|\,x}(\theta)$ is high if the known values of x do not constrain the possible values of z and the problem is more difficult.

The matrix $M(\theta)$ is not a fixed quantity and evolves with the parameters θ. In regions where we have a good model of the data, for example if we found well-separated clusters fit with Gaussian mixtures, there is little uncertainty about the latent variables (the cluster membership) and $M(\theta)$ will be small. However, $M(\theta)$ is often large at the start of the optimization procedure. The linear rate r in Equation (11) is then determined by the maximum eigenvalue of the missing information, $r = \lambda_{\max}(M(\theta^*))$. Linear convergence occurs if the missing information at θ^* is small, $M(\theta^*) \prec 1$, otherwise r can be larger than 1.

This result, however, is only asymptotic and existing non-asymptotic linear rates rely on strong-convexity assumptions instead (e.g. Balakrishnan et al., 2017). A twice-differentiable function f is α-strongly convex if

$$\nabla^2 f(\theta) \succeq \alpha I \qquad \text{for } \alpha > 0,$$

which means the eigenvalues of $\nabla^2 f(\theta)$ are bounded below by α. If f is also L-smooth, as defined in Section 3, a gradient-EM type of analysis gives that

$$f(\theta_{t+1}) - f^* \leq \Big(1 - \frac{\alpha}{L}\Big)(f(\theta_t) - f^*).$$

This implies EM converges linearly if it enters a smooth and strongly-convex region. However, in these works, the connection to the ratio of missing information is lost and the rate is not invariant to reparametrization.

We showed in Section 3 that, instead of measuring smoothness in Euclidean norms, the EM objective is 1-smooth relative to its log-partition function A. Likewise, we can characterize strong convexity relative to a reference function h (Lu et al., 2018), requiring that

$$\nabla^2 \mathcal{L}(\theta) \succeq \alpha \nabla^2 h(\theta) \qquad \text{for } \alpha > 0. \tag{13}$$

For EM, where we care about strong convexity relative to the log-partition A, *the relative strong-convexity parameter α is directly related to the missing information;*

Proposition 3. *For exponential families, the EM objective is α-strongly convex relative to A on a region Θ iff the missing information M (Equation 12) satisfies*

$$\lambda_{\max}(M(\theta)) \leq (1 - \alpha) \qquad \textit{for all } \theta \in \Theta.$$

We provide a detailed proof in Appendix D and give here the main intuition. For exponential families, the Hessian of the surrogate $Q_\theta(\theta)$ coincides with the Hessian of $A(\theta)$ (Equation 9), which is the Fisher information matrix of the complete-data distribution, $I_{x,z}(\theta)$. Using the decomposition of Equation (2), the Hessian of the objective can be shown to be equal to

$$\nabla^2 \mathcal{L}(\theta) = I_{x,z}(\theta) - I_{z\,|\,x}(\theta).$$

The definition of α-strong convexity relative to A (Equation 13) for EM is then equivalent to

$$I_{x,z}(\theta) - I_{z\,|\,x}(\theta) \succeq \alpha I_{x,z}(\theta).$$

Multiplying by $I_{x,z}(\theta)^{-1}$ and rearranging terms yields

$$M(\theta) = I_{x,z}(\theta)^{-1} I_{z\,|\,x}(\theta) \preceq (1 - \alpha) I. \qquad \square$$

Convergence results for mirror descent on relatively 1-smooth and α-strongly convex functions (Lu et al., 2018) then directly give the following local linear rate.

Corollary 3. *Under **A1**–**A3**, if EM is initialized in a locally convex region Θ with minimum $\mathcal{L}^*$ and the ratio of missing information is bounded, $\lambda_{\max}(M(\theta)) \leq r$,*

$$\mathcal{L}(\theta_{t+1}) - \mathcal{L}^* \leq r(\mathcal{L}(\theta_t) - \mathcal{L}^*).$$

If the ratio of missing information goes to zero, as in the case of well-separated clusters for Gaussian mixtures with suitable initialization, then EM converges superlinearly in the neighborhood of a solution (Salakhutdinov et al., 2003; Xu and Jordan, 1996; Ma et al., 2000).

5.3 Generalized EM

We now consider generalized EM schemes, which do not optimize the surrogate exactly in the M-step but output an approximate (possibly randomized) update. Given θ_t, we assume we can solve the surrogate problem with some expected guarantee on the optimality gap, $\mathbb{E}\big[Q_{\theta_t}(\theta_{t+1}) - Q^*_{\theta_t}\big]$, where $Q^*_{\theta_t}$ is the minimum value of the surrogate. The M-step achieve $Q_{\theta_t}(\theta_{t+1}) = Q^*_{\theta_t}$, but it might be more efficient to solve the problem only partially, or the M-step might be intractable. We consider two types of guarantees for those two cases, multiplicative and additive errors.

A4 Multiplicative error: The approximate solution θ_{t+1} satisfies the guarantee that, for some $c \in (0, 1]$,
$\mathbb{E}\big[Q_{\theta_t}(\theta_{t+1}) - Q^*_{\theta_t}\big] \leq (1 - c)\big(Q_{\theta_t}(\theta_t) - Q^*_{\theta_t}\big)$.

If $c = 1$, the algorithm is exact and θ_{t+1} minimizes the surrogate, while for $c = 0$ there is no guarantee of progress. An example of an algorithm satisfying this condition for mixture models would be the exact optimization of only one of the mixture components, chosen at random, like the ECM algorithm of Meng and Rubin (1993). As the surrogate problem is separable among components, the guarantee is satisfied with $c = 1/k$, where k is the number of clusters.

Theorem 1. *Under assumptions **A1**–**A3**, if the M-steps are solved up to c-multiplicative error (**A4**),*

$$\min_{t \leq T} \mathbb{E}[D_{A^*}(s(\theta_t), \mu_t)] \leq \frac{1}{c} \frac{\mathcal{L}(\theta_1) - \mathcal{L}^*}{T}.$$

This can give speedup in overall time if some iterations can be made more than than c times faster by leveraging the structure of the problem.

Multiplicative error, however, is a strong assumption if the closed-form solution is intractable. Instead, additive error is almost always satisfied. For example, 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.

A5 Additive error: The algorithm returns a solution θ_{t+1} with the guarantee that, in expectation,

$$\mathbb{E}\big[Q_{\theta_t}(\theta_{t+1}) - Q^*_{\theta_t}\big] \leq \epsilon_t.$$

If $\epsilon_t = 0$, the optimization is exact. Otherwise, the algorithm might not guarantee progress and the sequence ϵ_t needs to converge to 0 for the iterations to converge.

THEOREM 2. *Under assumptions **A1**–**A3**, if the M-step at step t is solved up to ϵ_t-additive error (**A5**),*

$$\min_{t \leq T} \mathbb{E}[D_{A^*}(s(\theta_t), \mu_t)] \leq \frac{\mathcal{L}(\theta_1) - \mathcal{L}^*}{T} + \frac{1}{T}\sum_{t=1}^{T} \epsilon_t.$$

For example, for $\epsilon_t = \mathcal{O}(1/t)$, the rate reduces to $\mathcal{O}(\log(T)/T)$, but we recover a $\mathcal{O}(1/T)$ rate if the errors decrease faster, $\epsilon_t = \mathcal{O}(1/t^2)$. As in Section 5.2, the results can be extended to give convergence in function values in a locally convex region. The proofs of Theorems 1 and 2 are deferred to Appendix D.

5.4 EM for General Models

While the exponential family covers many applications of EM, some are not smooth, in Euclidean distance or otherwise. For example, in a mixture of Laplace distributions the gradient of the surrogate is discontinuous (the Laplace distribution is not an exponential family). In this case, the progress of the M-step need not be related to the gradient and results similar to Corollary 2 do not hold. To the best of our knowledge, there are no general non-asymptotic convergence results for the general non-differentiable, non-convex setting and all we can guarantee is asymptotic convergence, as in the works of Chrétien and Hero (2000) and Tseng (2004).

The tools presented here can still obtain partial answers for the Laplace mixture and similar examples. The analyses in previous sections considered the progress of the M-step, as is common in non-asymptotic literature. We can instead view the E-step as the primary driver of progress, as is more common in the asymptotic literature. Assuming relative smoothness on the conditional distribution $p(z \mid x, \theta)$ only, we derive in Appendix E an analog of Corollary 2 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, as for the Laplace mixture.

6 DISCUSSION

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 showed that EM for exponential families always satisfies a notion of smoothness relative to a Bregman divergence. In this setting, EM and its stochastic variants are equivalent to mirror descent updates. This perspective leads to convergence rates that hold without additional assumptions and that are invariant to reparametrization. We also showed how the ratio of missing information can be integrated in non-asymptotic convergence rates, and analyzed the use of approximate M-steps. Although we focused on the MLE, Appendix C discusses MAP estimation. We show that results similar to Proposition 2 on the convergence to stationary points in KL divergence still hold, with minor changes to incorporate the prior. Viewing EM as a mirror descent procedure also highlights that it is a first-order method. It is thus susceptible to similar issues as classical first-order methods, such as slow progress in "flat" regions. However, flatness is measured in a different geometry (KL divergence) rather than the Euclidean geometry of gradient descent.

Beyond non-asymptotic convergence, smoothness relative to a KL divergence could be applied to extend statistical results, such as that of Daskalakis et al. (2017) to settings other than well-separated mixtures of Gaussians. In addition to the EM algorithm, our results could be extended to variational methods, such as the works of Hoffman et al. (2013) and Khan et al. (2016), due to the similarity between the EM surrogate and the evidence lower-bound.

Stochastic variants of EM are becoming increasingly relevant as they allow the algorithm to scale to large datasets, and recent recent work by Chen et al. (2018) and Karimi et al. (2019) combined stochastic EM updates with variance reduction methods like SAG, SVRG, and MISO (Le Roux et al., 2012; Johnson and Zhang, 2013; Mairal, 2015). Those works have taken the perspective that EM can be expressed as a preconditioned gradient step, and the resulting worst-case analysis not only depends on the smoothness constant, but the prescribed step-size is proportional to $1/L$. For Gaussian mixtures with arbitrary initialization, this implies using a step-size of 0. Our results highlights the gap between EM and gradient-EM methods, using a combination of classic and modern tools from a variety of fields, and we hope that the tools developed here may help to fix this and similar practical issues.

Acknowledgements

We thank the anonymous reviewers, whose comments helped improve the clarity of the manuscript.

We thank Si Yi (Cathy) Meng, Aaron Mishkin, and Victor Sanches Portella for providing comments on the manuscript and earlier versions of this work, and for suggesting references on related material. We are also grateful to Jason Hartford, Jonathan Wilder Lavington, and Yihan (Joey) Zhou for conversations that informed the ideas presented here.

This research was partially supported by the Canada CIFAR AI Chair Program, the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants RGPIN-2015-06068 and the NSERC Postgraduate Scholarships-Doctoral Fellowship 545847-2020.

References

Agarwal, Arvind and Hal Daumé III (2010). "A geometric view of conjugate priors". In: *Machine Learning* 81.1, pp. 99–113.

Amari, Shun-ichi and Hiroshi Nagaoka (2000). *Methods of Information Geometry*. Translations of Mathematical Monographs. Oxford University Press.

Balakrishnan, Sivaraman, Martin J. Wainwright, and Bin Yu (2017). "Statistical guarantees for the EM algorithm: From population to sample-based analysis". In: *Annals of Statistics* 45.1, pp. 77–120.

Banerjee, Arindam, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh (2005). "Clustering with Bregman Divergences". In: *Journal of Machine Learning Research* 6, pp. 1705–1749.

Bauschke, Heinz H., Jérôme Bolte, and Marc Teboulle (2017). "A descent Lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications". In: *Mathematics of Operations Research* 42.2, pp. 330–348.

Beck, Amir and Marc Teboulle (2003). "Mirror descent and nonlinear projected subgradient methods for convex optimization". In: *Operations Research Letters* 31.3, pp. 167–175.

Bishop, Christopher M. (2007). *Pattern recognition and machine learning, 5th Edition*. Information science and statistics. Springer.

Blight, B. J. N. (1970). "Estimation from a Censored Sample for the Exponential Family". In: *Biometrika* 57.2, pp. 389–395.

Boyles, Russell A. (1983). "On the Convergence of the EM Algorithm". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 45.1, pp. 47–50.

Cappé, Olivier and Eric Moulines (2009). "On-line expectation–maximization algorithm for latent data models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.3, pp. 593–613.

Caron, François and Arnaud Doucet (2008). "Sparse Bayesian nonparametric regression". In: *Proceedings of the International Conference on Machine Learning*, pp. 88–95.

Chen, Jianfei, Jun Zhu, Yee Whye Teh, and Tong Zhang (2018). "Stochastic expectation maximization with variance reduction". In: *Advances in Neural Information Processing Systems*, pp. 7978–7988.

Chrétien, Stéphane and Alfred O. Hero (2000). "Kullback proximal algorithms for maximum likelihood estimation". In: *IEEE Transactions on Information Theory* 46.5, pp. 1800–1810.

Csiszár, Imre and Gábor Tusnády (1984). "Information geometry and alternating minimization procedures". In: *Statistics & Decisions, Supplemental Issue* 1, pp. 205–237.

Daskalakis, Constantinos, Christos Tzamos, and Manolis Zampetakis (2017). "Ten steps of EM suffice for mixtures of two Gaussians". In: vol. 65. Proceedings of Machine Learning Research, pp. 704–710.

Delyon, Bernard, Marc Lavielle, and Eric Moulines (1999). "Convergence of a Stochastic Approximation Version of the EM Algorithm". In: *Annals of Statistics* 27.1, pp. 94–128.

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 39.1, pp. 1–38.

Diaconis, Persi and Donald Ylvisaker (1979). "Conjugate Priors for Exponential Families". In: *The Annals of Statistics* 7.2, pp. 269–281.

Figueiredo, Mário A. T. and Robert D. Nowak (2003). "An EM algorithm for wavelet-based image restoration". In: *IEEE Transactions on Image Processing* 12.8, pp. 906–916.

Ghahramani, Zoubin and Michael I. Jordan (1994). "Supervised learning from incomplete data via an EM approach". In: *Advances in Neural Information Processing Systems*, pp. 120–127.

Hoffman, Matthew D., David M. Blei, Chong Wang, and John W. Paisley (2013). "Stochastic variational inference". In: *J. Mach. Learn. Res.* 14.1, pp. 1303–1347.

Johnson, Rie and Tong Zhang (2013). "Accelerating stochastic gradient descent using predictive variance reduction". In: *Advances in Neural Information Processing Systems*, pp. 315–323.

Jordan, Michael I. and Lei Xu (1995). "Convergence results for the EM approach to mixtures of experts architectures". In: *Neural Networks* 8.9, pp. 1409–1431.

Karimi, Belhal, Hoi-To Wai, Eric Moulines, and Marc Lavielle (2019). "On the global convergence of (fast) incremental expectation maximization methods". In: *Advances in Neural Information Processing Systems*, pp. 2837–2847.

Khan, Mohammad Emtiyaz, Reza Babanezhad, Wu Lin, Mark Schmidt, and Masashi Sugiyama (2016). "Faster Stochastic Variational Inference using Proximal-Gradient Methods with General Divergence Functions". In: *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*. Ed. by Alexander T. Ihler and Dominik Janzing. AUAI Press.

Lange, Kenneth, David R. Hunter, and Ilsoon Yang (2000). "Optimization transfer using surrogate objective functions". In: *Journal of Computational and Graphical Statistics* 9.1, pp. 1–20.

Lauritzen, Steffen L. (1995). "The EM algorithm for graphical association models with missing data". In: *Computational Statistics & Data Analysis* 19.2, pp. 191–201.

Le Roux, Nicolas, Mark Schmidt, and Francis Bach (2012). "A stochastic gradient method with an exponential convergence rate for finite training sets". In: *Advances in Neural Information Processing Systems*, pp. 2672–2680.

Lu, Haihao, Robert M. Freund, and Yurii Nesterov (2018). "Relatively smooth convex optimization by first-order methods, and applications". In: *SIAM Journal on Optimization* 28.1, pp. 333–354.

Ma, Jinwen, Lei Xu, and Michael I. Jordan (2000). "Asymptotic convergence rate of the EM algorithm for Gaussian mixtures". In: *Neural Computation* 12.12, pp. 2881–2907.

Mairal, Julien (2013). "Optimization with first-order surrogate functions". In: *Proceedings of the International Conference on Machine Learning*, pp. 783–791.

Mairal, Julien (2015). "Incremental majorization-minimization optimization with application to large-scale machine learning". In: *SIAM Journal on Optimization* 25.2, pp. 829–855.

McLachlan, Geoffrey and Thriyambakam Krishnan (2007). *The EM algorithm and extensions*. 2nd. Vol. 382. Wiley.

Meng, Xiao-Li and Donald B. Rubin (1993). "Maximum likelihood estimation via the ECM algorithm: A general framework". In: *Biometrika* 80.2, pp. 267–278.

Murphy, Kevin P. (2012). *Machine learning: A probabilistic perspective*. Adaptive computation and machine learning series. MIT Press.

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

Nemirovski, Arkadi, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro (2009). "Robust Stochastic Approximation Approach to Stochastic Programming". In: *SIAM Journal on Optimization* 19.4, pp. 1574–1609.

Nemirovski, Arkadi Semenovich and David Borisovich Yudin (1983). *Problem complexity and method efficiency in optimization*. translated by E.R. Dawson. Original title: Slozhnost' zadach i ėffektivnost' metodov optimizatsii. NY: Wiley.

Nesterov, Yurii (2013). *Introductory lectures on convex optimization: A basic course*. Vol. 87. Springer Science & Business Media.

Nesterov, Yurii and Arkadi Nemirovski (1994). *Interior-Point Polynomial Algorithms in Convex Programming*. Society for Industrial and Applied Mathematics.

Orchard, Terence and Max A. Woodbury (1972). "A missing information principle: theory and applications". In: *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of Statistics*. Berkeley, Califfornia: University of California Press, pp. 697–715.

Paquette, Courtney, Hongzhou Lin, Dmitriy Drusvyatskiy, Julien Mairal, and Zaid Harchaoui (2018). "Catalyst for gradient-based nonconvex optimization". In: *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 613–622.

Rabiner, Lawrence R. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition". In: *Proceedings of the IEEE* 77.2, pp. 257–286.

Raskutti, Garvesh and Sayan Mukherjee (2015). "The Information Geometry of Mirror Descent". In: *IEEE Transactions on Information Theory* 61.3, pp. 1451–1457.

Razaviyayn, Meisam (2014). "Successive convex approximation: Analysis and applications". PhD thesis. University of Minnesota.

Robbins, Herbert and Sutton Monro (1951). "A stochastic approximation method". In: *Annals of Mathematical Statistics* 22.3, pp. 400–407.

Salakhutdinov, Ruslan, Sam T. Roweis, and Zoubin Ghahramani (2003). "Optimization with EM and Expectation-Conjugate-Gradient". In: *Proceedings of the International Conference on Machine Learning*, pp. 672–679.

Sato, Masa-aki (1999). *Fast learning of on-line EM algorithm*. Tech. rep. ATR Human Information Processing Research Laboratories.

Tseng, Paul (2004). "An analysis of the EM algorithm and entropy-like proximal point methods". In: *Mathematics of Operations Research* 29.1, pp. 27–44.

Varadhan, Ravi and Christophe Roland (2004). *Squared Extrapolation Methods (SQUAREM): A New Class of Simple and Efficient Numerical Schemes for Accelerating the Convergence of the EM Algorithm*. Working Paper 63. Johns Hopkins University, Department of Biostatistics.

Wainwright, Martin J. and Michael I. Jordan (2008). "Graphical models, exponential families, and variational inference". In: *Foundations and Trends in Machine Learning* 1.1-2, pp. 1–305.

Wang, Zhaoran, Quanquan Gu, Yang Ning, and Han Liu (2015). "High dimensional EM algorithm: Statistical optimization and asymptotic normality". In: *Advances in Neural Information Processing Systems*, pp. 2521–2529.

Wu, C. F. Jeff (1983). "On the convergence properties of the EM algorithm". In: *Annals of statistics* 11.1, pp. 95–103.

Xu, Lei and Michael I. Jordan (1996). "On convergence properties of the EM algorithm for Gaussian mixtures". In: *Neural Computation* 8.1, pp. 129–151.

Xu, Yangyang and Wotao Yin (2013). "A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion". In: *SIAM Journal on Imaging Sciences* 6.3, pp. 1758–1789.

Yi, Xinyang and Constantine Caramanis (2015). "Regularized EM algorithms: A unified framework and statistical guarantees". In: *Advances in Neural Information Processing Systems*, pp. 1567–1575.