Proceedings of Machine Learning Research vol 65:1–7, 2017

Ten Steps of EM Suffice for Mixtures of Two Gaussians

Constantinos Daskalakis COSTIS@MIT.EDU
Christos Tzamos CTZAMOS@GMAIL.COM
Manolis Zampetakis MZAMPET@MIT.EDU
EECS and CSAIL, MIT

Extended Abstract[1]

The *Expectation-Maximization (EM) algorithm* Dempster et al. (1977); Wu (1983); Redner and Walker (1984) is one of the most widely used heuristics for maximizing likelihood in statistical models with latent variables. Consider a probability distribution $p_{\boldsymbol{\lambda}}$ sampling $(\boldsymbol{X}, \boldsymbol{Z})$, where $\boldsymbol{X}$ is a vector of observable random variables, $\boldsymbol{Z}$ a vector of non-observable random variables and $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ a vector of parameters. Given independent samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of the observed random variables, the goal of maximum likelihood estimation is to select $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ maximizing the log-likelihood of the samples, namely $\sum_i \log p_{\boldsymbol{\lambda}}(\boldsymbol{x}_i)$. Unfortunately, computing $p_{\boldsymbol{\lambda}}(\boldsymbol{x}_i)$ involves summing $p_{\boldsymbol{\lambda}}(\boldsymbol{x}_i, \boldsymbol{z}_i)$ over all possible values of $\boldsymbol{z}_i$, which commonly results in a log-likelihood function that is non-convex with respect to $\boldsymbol{\lambda}$ and therefore hard to optimize. In this context, the EM algorithm proposes the following heuristic:

- Start with an initial guess $\boldsymbol{\lambda}^{(0)}$ of the parameters.
- For all $t \geq 0$, until convergence:
  - (E-Step) For each sample i, compute the posterior $Q_i^{(t)}(\boldsymbol{z}) := p_{\boldsymbol{\lambda}^{(t)}}(\boldsymbol{Z} = \boldsymbol{z} | \boldsymbol{X} = \boldsymbol{x}_i)$.
  - (M-Step) Set $\boldsymbol{\lambda}^{(t+1)} := \arg\max_{\boldsymbol{\lambda}} \sum_i \sum_{\boldsymbol{z}} Q_i^{(t)}(\boldsymbol{z}) \log \frac{p_{\boldsymbol{\lambda}}(\boldsymbol{x}_i, \boldsymbol{z})}{Q_i^{(t)}(\boldsymbol{z})}$.

Intuitively, the E-step of the algorithm uses the current guess of the parameters, $\boldsymbol{\lambda}^{(t)}$, to form beliefs, $Q_i^{(t)}$, about the state of the (non-observable) $\boldsymbol{Z}$ variables for each sample i. Then the M-step uses the new beliefs about the state of $\boldsymbol{Z}$ for each sample to maximize with respect to $\boldsymbol{\lambda}$ a lower bound on $\sum_i \log p_{\boldsymbol{\lambda}}(\boldsymbol{x}_i)$. Indeed, by the concavity of the $\log$ function, the objective function used in the M-step of the algorithm is a lower bound on the true log-likelihood for all values of $\boldsymbol{\lambda}$, and it equals the true log-likelihood for $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(t)}$. From these observations, it follows that the above alternating procedure improves the true log-likelihood until convergence.

Despite its wide use and practical significance, little is known about whether and under what conditions EM converges to the true maximum likelihood estimator. A few works establish local convergence of the algorithm to stationary points of the log-likelihood function Wu (1983); Tseng (2004); Chrétien and Hero (2008), and even fewer local convergence to the MLE Redner and Walker (1984); Balakrishnan et al. (2017). Besides local convergence, it is also known that badly initialized

1. Full version appears as arXiv preprint arXiv:1609.00368, v5 Daskalakis et al. (2016).

EM may settle far from the MLE both in parameter and in likelihood distance Wu (1983). The lack of theoretical understanding of the convergence properties of EM is intimately related to the non-convex nature of the optimization it performs.

Our paper aims to illuminate why EM works well in practice and develop techniques for understanding its behavior. We do so by analyzing one of the most basic and natural, yet still challenging, statistical models EM may be applied to, namely balanced mixtures of two multi-dimensional Gaussians with equal and known covariance matrices. In particular, we study the convergence of EM when applied to the following family of parametrized density functions:

$$p_{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2}(\boldsymbol{x}) = 0.5 \cdot \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_1, \Sigma) + 0.5 \cdot \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_2, \Sigma),$$

where Σ is a known covariance matrix, $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ are unknown (vector) parameters, and $\mathcal{N}(\boldsymbol{\mu}, \Sigma; \boldsymbol{x})$ represents the Gaussian density with mean $\boldsymbol{\mu}$ and covariance matrix Σ, i.e.

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{2\pi \det \Sigma}} \exp\left(-0.5(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right).$$

Our main contribution is to provide global convergence guarantees for EM applied to the above family of distributions. We establish our result for both the "population version" of the algorithm, and the finite-sample version, as described below.

Analysis of Population EM for Mixtures of Two Gaussians. To elucidate the optimization features of the algorithm and avoid analytical distractions arising due to sampling error, it has been standard practice in the literature of theoretical analyses of EM to consider the "population version" of the algorithm, where the EM iterations are performed assuming access to infinitely many samples from a distribution $p_{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2}$ as above. With infinitely many samples, we can identify the mean, $\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2}$, of $p_{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2}$, and re-parametrize the density around the mean as follows:

$$p_{\boldsymbol{\mu}}(\boldsymbol{x}) = 0.5 \cdot \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) + 0.5 \cdot \mathcal{N}(\boldsymbol{x}; -\boldsymbol{\mu}, \Sigma). \tag{1}$$

We first study the convergence of EM when we perform iterations with respect to the parameter $\boldsymbol{\mu}$ of $p_{\boldsymbol{\mu}}(\boldsymbol{x})$ in (1). Starting with an initial guess $\boldsymbol{\lambda}^{(0)}$ for the unknown mean vector $\boldsymbol{\mu}$, the t-th iteration of EM amounts to the following update:

$$\boldsymbol{\lambda}^{(t+1)} = M(\boldsymbol{\lambda}^{(t)}, \boldsymbol{\mu}) \triangleq \frac{\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\mu}}}\left[\frac{0.5\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda}^{(t)},\Sigma)}{p_{\boldsymbol{\lambda}^{(t)}}(\boldsymbol{x})}\boldsymbol{x}\right]}{\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{\mu}}}\left[\frac{0.5\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda}^{(t)},\Sigma)}{p_{\boldsymbol{\lambda}^{(t)}}(\boldsymbol{x})}\right]}, \tag{2}$$

where we have compacted both the E- and M-step of EM into one update.

The intuition behind the EM update formula is as follows. First, we take expectations with respect to $\boldsymbol{x} \sim p_{\boldsymbol{\mu}}$ because we are studying the population version of EM, hence we assume access to infinitely many samples from $p_{\boldsymbol{\mu}}$. For each sample $\boldsymbol{x}$, the ratio $\frac{0.5\mathcal{N}(\boldsymbol{x};\boldsymbol{\lambda}^{(t)},\Sigma)}{p_{\boldsymbol{\lambda}^{(t)}}(\boldsymbol{x})}$ is our belief, at step t, that $\boldsymbol{x}$ was sampled from the first Gaussian component of $p_{\boldsymbol{\mu}}$, namely the one for which our current estimate of its mean vector is $\boldsymbol{\lambda}^{(t)}$. (The complementary probability is our present belief that $\boldsymbol{x}$ was sampled from the other Gaussian component.) Given these beliefs for all vectors $\boldsymbol{x}$, the update (2) is the result of the M-step of EM. Intuitively, our next guess $\boldsymbol{\lambda}^{(t+1)}$ for the mean vector of the first Gaussian component is a weighted combination over all samples $\boldsymbol{x} \sim p_{\boldsymbol{\mu}}$ where the weight of every $\boldsymbol{x}$ is our belief that it came from the first Gaussian component.

Our main result for population-EM is the following:

Informal Theorem 1 (Population EM Analysis) *Whenever the initial guess $\boldsymbol{\lambda}^{(0)}$ is not equidistant to $\boldsymbol{\mu}$ and $-\boldsymbol{\mu}$, EM converges geometrically to either $\boldsymbol{\mu}$ or $-\boldsymbol{\mu}$, with convergence rate that improves as $t \to \infty$. We provide a simple, closed form expression of the convergence rate as a function of $\boldsymbol{\lambda}^{(t)}$ and $\boldsymbol{\mu}$. If the initial guess $\boldsymbol{\lambda}^{(0)}$ is equidistant to $\boldsymbol{\mu}$ and $-\boldsymbol{\mu}$, EM converges to the unstable fixed point* $\mathbf{0}$.

A formal statement of Informal Theorem 1 is provided in the full version of the paper. As a simple illustration of our result, we show that, in one dimension, when our original guess $\lambda^{(0)} = +\infty$ and the signal-to-noise ratio $\mu/\sigma = 1$, 10 steps of the EM algorithm result in 1% error.

Despite the simplicity of the case we consider, no global convergence results were known prior to our work, even for the population EM. Balakrishnan et al. (2017) studied the same setting proving only local convergence, i.e. convergence only when the initial guess is close to the true parameters. They argue that the population EM update is contracting close to the true parameters. Unfortunately, the EM update is non-contracting outside a small neighborhood of the true parameters so this argument cannot be used for a global convergence guarantee.

In this work, we study the problem under arbitrary starting points and completely characterize the fixed points of EM. We show that other than a measure-zero subset of the space (namely points that are equidistant from the centers of the two Gaussians), any initialization of the EM algorithm converges to the true centers of the Gaussians, providing explicit bounds for the convergence rate. To achieve this, we follow an orthogonal approach to Balakrishnan et al. (2017): Instead of trying to directly compute the number of steps required to reach convergence *for a specific instance of the problem*, we study *the sensitivity of the EM iteration as the instance varies*. The intuition is that if the EM update is sensitive to updating the instance, then changing the instance should also attract the update towards the changing instance; see Figure 1. We can use this, in turn, to argue that keeping the instance fixed, one EM update makes progress towards the true parameters. In particular, we gain a handle on the convergence rate of EM on all instances at once.

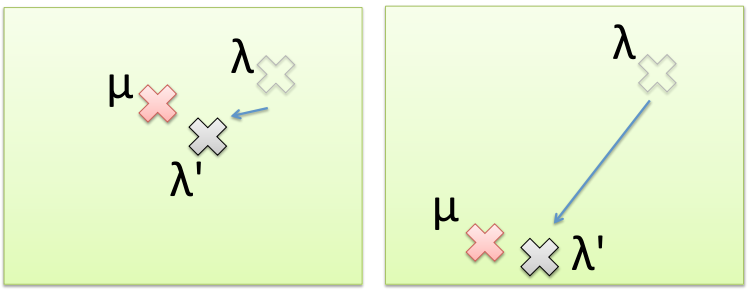

Figure 1: Sensitivity of the EM update when changing the true parameters. Large sensitivity implies large progress towards the true parameters.

Analysis of Finite-Sample EM for Mixtures of Two Gaussians. The finite sample analysis proceeds in three steps. First, in the finite sample regime we do not know the average of the two mean vectors, $(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$, exactly. We show that, with $\tilde{O}(d/\epsilon^2)$ samples, we can approximate the average to within Mahalanobis distance ϵ. We then chain two coupling arguments. The first compares the progress towards the true mean made by the correctly centered population EM update to that

of the incorrectly centered population EM update. The second compares the progress towards the true mean made by the incorrectly centered population EM update with the progress made by the incorrectly centered finite sample EM update. See Figure 2. Given the error incurred in the approximation of the center $(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$, we propose to stabilize the sample-based EM iteration by including in the sample for each sampled point $\boldsymbol{x}_i$ its symmetric point $-\boldsymbol{x}_i$. This is the sample based version that we analyze, although our analysis goes through without this stabilization. Our result is the following, formally given in the full version.

Informal Theorem 2 (Finite Sample EM Analysis) *Whenever $\epsilon < SNR$, $\tilde{O}(d/\epsilon^2 \cdot \mathrm{poly}(1/SNR))$ samples suffice to approximate $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ to within Mahalanobis distance ϵ using the EM algorithm. In particular, the error rate of the EM based estimator is $\tilde{O}\left(\sqrt{\frac{d}{n}}\right)$ where n is the number of samples, which is optimal up to logarithmic factors.*[2]

Bootstrapping EM for Faster Convergence. We note that, in multiple dimensions, care must be taken in initializing the EM algorithm, even in the infinite sample regime, as the convergence guarantee depends on the angle between the current iterate and the true mean vector. While a randomly chosen unit vector will have projection of $\Theta(1/\sqrt{d})$ in the direction of $\boldsymbol{\mu}$, we argue that we can boostrap EM to turn this projection larger than a constant. This allows us to work with similar convergence rates as in the single-dimensional case, namely only SNR (and not dimension) dependent.

Informal Theorem 3 (EM Initialization) *EM can be boostrapped so that the number of iterations required to approximate $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ to within Mahalanobis distance ϵ depends logarithmically in the dimension.*

Related Work on Learning Mixtures of Gaussians. We have already outlined the literature on the Expectation-Maximization algorithm. Several results study its local convergence properties and there are known cases where badly initialized EM fails to converge. See above.

There is also a large body of literature on learning mixtures of Gaussians. A long line of work initiated by Dasgupta Dasgupta (1999); Arora and Kannan (2001); Vempala and Wang (2004); Achlioptas and McSherry (2005); Kannan et al. (2005); Dasgupta and Schulman (2007); Chaudhuri and Rao (2008); Brubaker and Vempala (2008); Chaudhuri et al. (2009) provides rigorous guarantees on recovering the parameters of Gaussians in a mixture under separability assumptions, while later work Kalai et al. (2010); Moitra and Valiant (2010); Belkin and Sinha (2010) has established guarantees under minimal information theoretic assumptions. More recent work Hardt and Price (2015) provides tight bounds on the number of samples necessary to recover the parameters of the Gaussians as well as improved algorithms, while another strand of the literature studies proper learning with improved running times and sample sizes Suresh et al. (2014); Daskalakis and Kamath (2014). Finally, there has been work on methods exploiting general position assumptions or performing smoothed analysis Hsu and Kakade (2013); Ge et al. (2015).

In practice, the most common algorithm for learning mixtures of Gaussians is the Expectation-Maximization algorithm, with the practical experience that it performs well in a broad range of

2. Note that even if SNR is arbitrarily large (so that the two Gaussian components are "perfectly separated") the problem degenerates to finding the mean of one Gaussian whose optimal rate is $\Omega\left(\sqrt{\frac{d}{n}}\right)$.

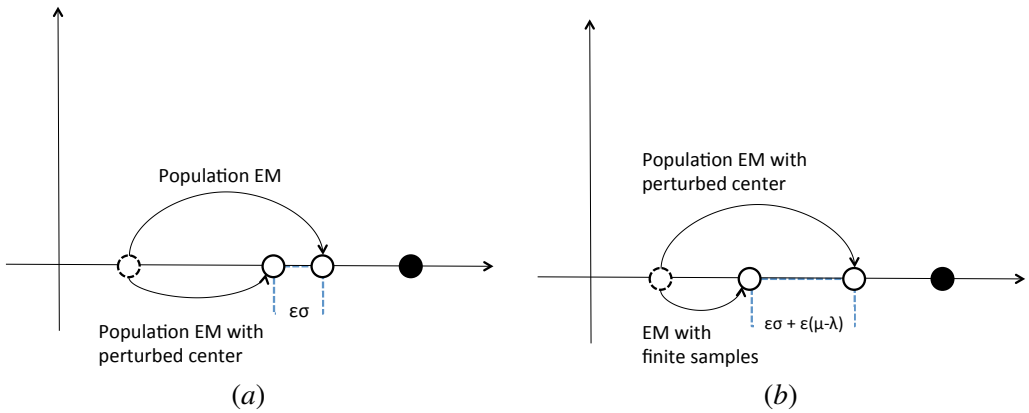

Figure 2: (a) Coupling correctly and incorrectly centered population EM updates. We show that, starting from the same iterate, the correctly and incorrectly centered population EM updates will land to close-by points. (b) Coupling incorrectly centered population EM and finite sample EM updates. We show that, starting from the same iterate, the incorrectly centered population EM update and the finite sample update land to close-by points.

scenarios despite the lack of theoretical guarantees. Recently, Balakrishnan et al. (2017) studied the convergence of EM in the case of an equal-weight mixture of two Gaussians with the same and known covariance matrix, showing local convergence guarantees. In particular, they show that when EM is initialized close enough to the actual parameters, then it converges. In this work, we revisit the same setting considered by Balakrishnan et al. (2017) but establish *global convergence guarantees*. We show that, for any initialization of the parameters, the EM algorithm converges geometrically to the true parameters. We also provide a simple and explicit formula for the rate of convergence.

Concurrent and independent work by Xu, Hsu and Maleki Xu et al. (2016) has also provided global and geometric convergence guarantees for the same setting, as well as a slightly more general setting where the mean of the mixture is unknown, but they do not provide explicit convergence rates. They also do not provide an analysis of the finite-sample regime.

Acknowledgments

We thank Sham Kakade for suggesting the problem to us, and for initial discussions. The authors were supported by NSF Awards CCF-0953960 (CAREER), CCF-1551875, CCF-1617730, and CCF-1650733, ONR Grant N00014-12-1-0999, a Microsoft Faculty Fellowship, and a Simons Graduate Fellowship.

References

Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *International Conference on Computational Learning Theory*, pages 458–469. Springer, 2005.

Sanjeev Arora and Ravi Kannan. Learning mixtures of arbitrary gaussians. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 247–257. ACM, 2001.

Sivaraman Balakrishnan, Martin J Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017.

Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 103–112. IEEE, 2010.

S Charles Brubaker and Santosh S Vempala. Isotropic PCA and affine-invariant clustering. In *the 49th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2008.

Kamalika Chaudhuri and Satish Rao. Learning Mixtures of Product Distributions Using Correlations and Independence. In *the 21st International Conference on Computational Learning Theory (COLT)*, 2008.

Kamalika Chaudhuri, Sanjoy Dasgupta, and Andrea Vattani. Learning mixtures of gaussians using the k-means algorithm. *arXiv preprint arXiv:0912.0086*, 2009.

Stéphane Chrétien and Alfred O Hero. On EM algorithms and their proximal generalizations. *ESAIM: Probability and Statistics*, 12:308–326, 2008.

Sanjoy Dasgupta. Learning mixtures of gaussians. In *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pages 634–644. IEEE, 1999.

Sanjoy Dasgupta and Leonard Schulman. A probabilistic analysis of em for mixtures of separated, spherical gaussians. *Journal of Machine Learning Research*, 8(Feb):203–226, 2007.

Constantinos Daskalakis and Gautam Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of gaussians. In *Proceedings of The 27th Conference on Learning Theory*, pages 1183–1213, 2014.

Constantinos Daskalakis, Christos Tzamos, and Manolis Zampetakis. Ten Steps of EM Suffice for Mixtures of Two Gaussians. *arXiv*, 2016. http://arxiv.org/abs/1609.00368.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

Rong Ge, Qingqing Huang, and Sham M Kakade. Learning mixtures of gaussians in high dimensions. In *the 47th Annual ACM on Symposium on Theory of Computing (STOC)*, 2015.

Moritz Hardt and Eric Price. Tight bounds for learning a mixture of two gaussians. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pages 753–760. ACM, 2015.

Daniel Hsu and Sham M Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *the 4th conference on Innovations in Theoretical Computer Science (ITCS)*, 2013.

Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two gaussians. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 553–562. ACM, 2010.

Ravindran Kannan, Hadi Salmasian, and Santosh Vempala. The spectral method for general mixture models. In *the 18th International Conference on Computational Learning Theory (COLT)*, 2005.

Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 93–102. IEEE, 2010.

Richard A Redner and Homer F Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM review*, 26(2):195–239, 1984.

Ananda Theertha Suresh, Alon Orlitsky, Jayadev Acharya, and Ashkan Jafarpour. Near-optimal-sample estimators for spherical gaussian mixtures. In *Advances in Neural Information Processing Systems*, pages 1395–1403, 2014.

Paul Tseng. An analysis of the EM algorithm and entropy-like proximal point methods. *Mathematics of Operations Research*, 29(1):27–44, 2004.

Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.

CF Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of statistics*, pages 95–103, 1983.

Ji Xu, Daniel Hsu, and Arian Maleki. Global analysis of Expectation Maximization for mixtures of two Gaussians. In *the 30th Annual Conference on Neural Information Processing Systems (NIPS)*, 2016.