Are there local maxima in the infinite-sample likelihood of Gaussian mixture estimation?

Nathan Srebro

Toyota Technological Institute-Chicago IL, USA
nati@uchicago.edu

Consider the problem of estimating the centers $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$ of a uniform mixture of unit-variance spherical Gaussians in $\mathbb{R}^d$,

$$f_{(\mu_1,\mu_2,\ldots,\mu_k)}(x) = \sum_{i=1}^{k} \frac{1}{k} \frac{1}{(2\pi)^{d/2}} e^{|x-\mu_i|^2/2}, \tag{1}$$

from i.i.d. samples $x_1, \ldots, x_m$ drawn from this distribution. This can be done by maximizing the (average log) likelihood $L_{(x_1,\ldots,x_m)}(\boldsymbol{\mu}) = \frac{1}{m}\sum_i \log f_{\boldsymbol{\mu}}(x_i)$. Maximizing the likelihood is guaranteed to recover the correct centers, in the large-sample limit, for any mixture model of the form (1). Unfortunately, maximizing the likelihood is hard in the worst case, and we usually revert to local search heuristics such as Expectation Maximization (EM) which can get trapped in the many local minima the likelihood function might have.

Despite this, a string of results establishes that the centers *can* be tractably recovered, given enough data sampled from a well-separated mixture, using projection-based methods [1–3], and even using EM [4].

These results require a large separation between centers. In practice, even with a much smaller separation, given enough data and proper initialization, EM converges to the global ML solution and allows recovery of the centers [5]. It seems that when data is plentiful, the local minima disappear.

Although the likelihood function for finite data sets may admit many local minima, the conjecture proposed here is that in the infinite sample limit, for data sampled from a distribution of the form (1), with *any* true centers $\boldsymbol{\mu}^0 = (\mu_1^0, \ldots, \mu_k^0)$, the only local maxima are the global maxima, given by permutations of the true centers $\mu_1^0, \ldots, \mu_k^0$.

At the infinite sample limit, the likelihood is given by the KL-divergence between mixture models: $L(\boldsymbol{\mu}) \overset{m\to\infty}{\longrightarrow} \mathbf{E}_{X\sim f_{\boldsymbol{\mu}^0}}[\log f_{\boldsymbol{\mu}}(X)] = -D\left(\boldsymbol{\mu}^0 \| \boldsymbol{\mu}\right) - H(\boldsymbol{\mu}^0)$, where the entropy $H(\boldsymbol{\mu}^0)$ of $f_{\boldsymbol{\mu}^0}$ is constant. Maxima of the infinite-sample likelihood are thus exactly minima of the KL-divergence $D\left(\boldsymbol{\mu}^0 \| \boldsymbol{\mu}\right) = \mathbf{E}_{f_{\boldsymbol{\mu}^0}}\left[\log \frac{f_{\boldsymbol{\mu}^0}}{f_{\boldsymbol{\mu}}}\right]$. The KL-divergence is non-negative and zero only when $f_{\boldsymbol{\mu}} = f_{\boldsymbol{\mu}^0}$. This happens iff $\mu_1, \ldots, \mu_k$ are a permutation of $\mu_1^0, \ldots, \mu_k^0$, and so these are the only global minima of the KL-divergence. Our conjecture can therefore be equivalently stated as: **for any set of centers $\boldsymbol{\mu}^0$, the only local minima of $D\left(\boldsymbol{\mu}^0 \| \boldsymbol{\mu}\right)$, with respect to $\boldsymbol{\mu}$, are the global minima obtained at permutations of $\boldsymbol{\mu}^0$.**

The KL-divergence has many stable points which are not local minima, but rather saddle points. For example, such a saddle point arises when two centers

coincide in $\boldsymbol{\mu}$ (but not in $\boldsymbol{\mu}^0$). There are also several different basins, one for each permutation of the centers, separated by non-convex ridges and near-plateaus. When only a finite data set is considered, local minima easily arise in these near-plateaus. Even in the infinite-sample limit, EM, or other local-search methods, might take a very large number of steps to traverse these near-plateaus and converge. For this reason the conjecture does not directly imply tractability.

The conjecture *does* imply that no minimum separation is required in order to establish convergence to the global minimum at the infinite sample limit—if it is true, what remains is to study the relationship between the speed of convergence, the sample size and the separation. Moreover, the conjecture implies that local search (e.g. EM) will converge to the correct model *regardless of initialization* (except for a measure zero set of "ridge manifolds" between the attraction basins of different permutations of the correct centers). Empirical simulations with "infinite sample" EM (working directly on the KL-divergence) on three centers in two dimensions confirm this by showing eventual convergence to the global likelihood, even when initialized with two nearby centers. Current large-separation results require careful initialization ensuring at least one initial center from the vicinity of each true center [4, 6].

Of course, the real quantity of interest is the probability P_m, under some specific random initialization scheme, of being in the basin of attraction of the global maximum of the likelihood given a random sample of finite size m. In fact, our interest in the problem stemmed from study of the probability P_m for some reasonable initialization schemes. The conjecture can be equivalently stated as $P_m \to 1$ for any initialization scheme, and can thus be seen as a prerequisite to understanding P_m.

Clarification: the KL-divergence $D(p\|\boldsymbol{\mu})$ between a fixed arbitrary distribution p and mixture models (1), may have non-global local minima. The conjecture only applies when p itself is a mixture model of the form (1). In particular, if p is a mixture of more than k Gaussians, and we are trying to fit it with a mixture of only k Gaussians, non-global local minima can arise.

References

1. Dasgupta, S.: Learning mixtures of gaussians. In: Proc. of the 40th Ann. Symp. on Foundations of Computer Science. (1999)
2. Arora, S., Kannan, R.: Learning mixtures of arbitrary gaussians. In: Proceedings of the thirty-third annual ACM symposium on Theory of computing. (2001)
3. Vempala, S., Wang, G.: A spectral algorithm for learning mixture models. J. Comput. Syst. Sci. **68** (2004) 841–860
4. Dasgupta, S., Schulman, L.: A two-round variant of em for gaussian mixtures. In: Proc. of the 16th Ann. Conf. on Uncertainty in Artificial Intelligence. (2000)
5. Srebro, N., Shakhnarovich, G., Roweis, S.: An investigation of computational and informational limits in gaussian mixture clustering. In: Proceedings of the 23rd international conference on Machine learning (ICML). (2006)
6. Ostrovsky, R., Rabani, Y., Schulman, L.J., Swamy, C.: The effectiveness of lloyd-type methods for the k-means problem. In: Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06). (2006) 165–176