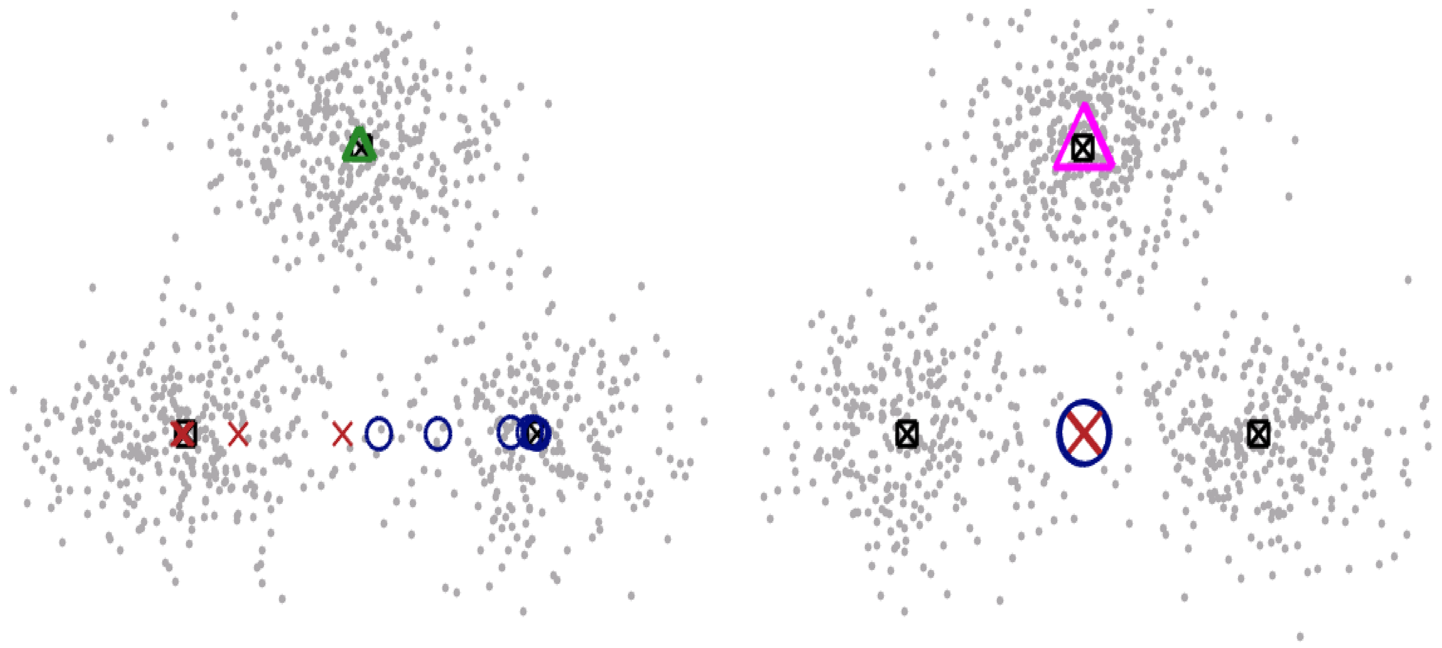
Convergence of Gradient EM for Multi-component Gaussian Mixture

Bowei Yan, Mingzhang Yin, Purnamrita Sarkar

Gaussian Mixture Models



- Data comes from M clusters in d dimensional space;
- Assume there exists a latent variable Z,

$$Z \sim \text{Multinomial}(\boldsymbol{\pi}); \qquad \boldsymbol{\pi} = (\pi_1, \cdots, \pi_M)$$
$$X|Z \sim \mathcal{N}(\boldsymbol{\mu}_Z, \Sigma); \qquad \boldsymbol{\mu} = (\boldsymbol{\mu}_1^T, \cdots, \boldsymbol{\mu}_M^T)^T \in \mathbb{R}^{Md}.$$

- Density of the mixture is $p(x|\boldsymbol{\mu}) = \sum_{i=1}^M \pi_i \phi(x|\boldsymbol{\mu}_i, \Sigma)$, where $\phi(x; \boldsymbol{\mu}, \Sigma)$ is the PDF of $N(\boldsymbol{\mu}, \Sigma)$.

Gradient EM

- E-step: $Q(\boldsymbol{\mu}|\boldsymbol{\mu}^t) = \mathbb{E}_X\left[\sum_{i=1}^M p(Z=i|X; \boldsymbol{\mu}^t) \log \phi(X; \boldsymbol{\mu}_i, \Sigma)\right]$;
- M-step: $\boldsymbol{\mu}_i^{t+1} = \boldsymbol{\mu}_i^t + s[\nabla Q(\boldsymbol{\mu}^t|\boldsymbol{\mu}^t)]_i = \boldsymbol{\mu}_i^t + s\mathbb{E}_X\left[\pi_i w_i(X; \boldsymbol{\mu}^t)(X - \boldsymbol{\mu}_i^t)\right]$.

Gradient Stability Condition

The Gradient Stability (GS) condition [1], denoted by $\text{GS}(\gamma, a)$, is satisfied if there exists $\gamma > 0$, such that for $\boldsymbol{\mu}_i^t \in \mathbb{B}(\boldsymbol{\mu}_i^*, a)$ with some $a > 0$, for $\forall i \in [M]$.

$$\|\nabla Q(\boldsymbol{\mu}^t|\boldsymbol{\mu}^*) - \nabla Q(\boldsymbol{\mu}^t|\boldsymbol{\mu}^t)\| \leq \gamma \|\boldsymbol{\mu}^t - \boldsymbol{\mu}^*\|$$

Theorem 1: Main Result for Population EM

Define $d_0 = \min\{d, M\}$, $\kappa = \frac{\pi_{\max}}{\pi_{\min}}$, $R_{\min} = \min_{i \neq j} \|\boldsymbol{\mu}_i^* - \boldsymbol{\mu}_j^*\|$. If $R_{\min} = \tilde{\Omega}(\sqrt{d_0})$, with initialization $\boldsymbol{\mu}^0$ satisfying, $\|\boldsymbol{\mu}_i^0 - \boldsymbol{\mu}_i^*\| \leq a, \forall i \in [M]$, where

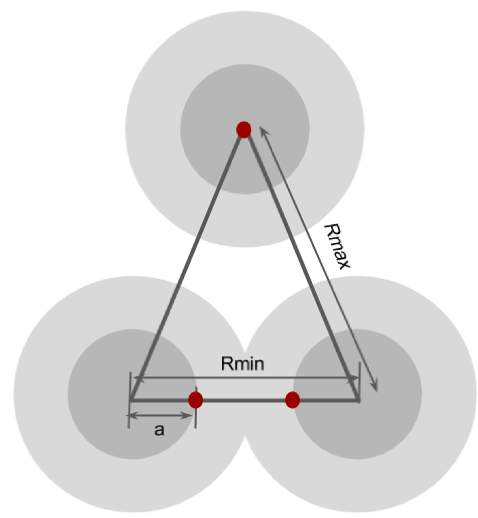
$$a \leq \frac{R_{\min}}{2} - \tilde{O}(\log(R_{\min})).$$

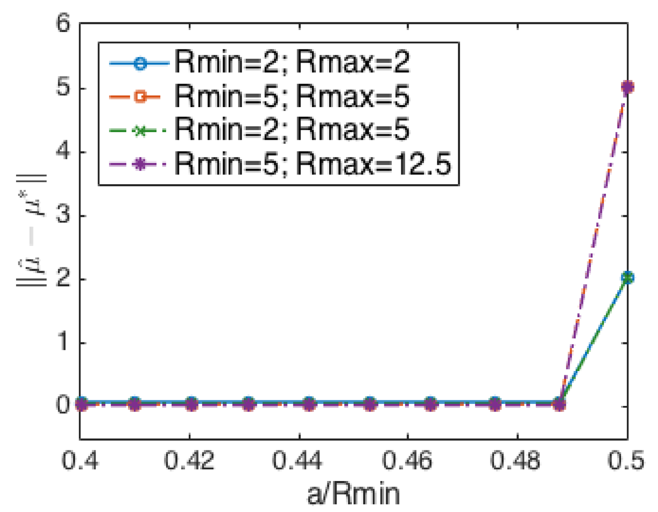
Then the Population EM converges with rate ζ to the center

$$\|\boldsymbol{\mu}^t - \boldsymbol{\mu}^*\| \leq \zeta^t \|\boldsymbol{\mu}_0 - \boldsymbol{\mu}^*\|, \quad \zeta = \frac{\pi_{\max} - \pi_{\min} + 2\gamma}{\pi_{\max} + \pi_{\min}} < 1$$

where

$$\gamma = M^2(2\kappa + 4)\left(2R_{\max} + d_0\right)^2 \exp\left(-\left(\frac{R_{\min}}{2} - a\right)^2 \frac{\sqrt{d_0}}{8}\right) < \pi_{\min}.$$





Theorem 2: Main Result for Sample-based EM

Let ζ be the contraction parameter in the main theorem, and

$$\epsilon^{\text{unif}}(n) = \tilde{O}\left(\frac{1}{\sqrt{n}} \max\{M^3(1+R_{\max})^3\sqrt{d}\max\{1, \log(\kappa)\}, (1+R_{\max})d\}\right).$$

If $\epsilon^{\text{unif}}(n) \leq (1-\zeta)a$, then sample-based gradient EM satisfies

$$\left\|\hat{\boldsymbol{\mu}}_i^t - \boldsymbol{\mu}_i^*\right\| \leq \zeta^t \left\|\boldsymbol{\mu}^0 - \boldsymbol{\mu}^*\right\|_2 + \frac{1}{1-\zeta}\epsilon^{\text{unif}}(n); \quad \forall i \in [M]$$

with probability at least $1 - n^{-cd}$, where c is positive constant.

Proof based on Rademacher complexity

For any unit vector u and cluster i, define the function class of gradient operator

$$\mathcal{F}_i^u = \{f^i : \mathcal{X} \to \mathbb{R} | f^i(X; \boldsymbol{\mu}, u) = w_i(X; \boldsymbol{\mu})\langle X - \boldsymbol{\mu}_i, u\rangle\}$$

And the target function

$$g_i^u(X) = \sup_{\boldsymbol{\mu} \in \mathbb{A}} \frac{1}{n}\sum_{i=1}^n w_1(X_i; \boldsymbol{\mu})\langle X_i - \boldsymbol{\mu}_1, u\rangle - \mathbb{E}w_1(X; \boldsymbol{\mu})\langle X - \boldsymbol{\mu}_1, u\rangle. \tag{1}$$

The proof consists of two steps:

1. First we show that $g(X)$ is close to its expectation by using concentration results.
   - Typically one uses McDiarmid's inequality for this step, which requires bounded differences. In our case we have differences which are bounded with high probability. We use a result from [2] to go around this problem. This gives a suboptimal $d/\sqrt{n}$ rate of convergence.
   - We are now working on a result which establishes the optimal $\sqrt{d/n}$ rate using similar arguments as in [3]. For details see the arxiv version.
2. Second we upper bound $\mathbb{E}g(X)$ by the Rademacher complexity of $\mathcal{F}_i^u$ by the symmetrization lemma.
   - In order to establish the Radamacher complexity we need the following vector-contraction result.

Vector-valued contraction

To get the Rademacher complexity, we build upon the recent vector-contraction result from [4]. Define $\eta_j(\boldsymbol{\mu}) : \mathbb{R}^{Md} \to \mathbb{R}^M$ as a vector valued function with the k-th coordinate

$$[\eta_j(\boldsymbol{\mu})]_k = \frac{\|\boldsymbol{\mu}_1\|^2}{2} - \frac{\|\boldsymbol{\mu}_k\|^2}{2} + \langle X_j, \boldsymbol{\mu}_k - \boldsymbol{\mu}_1\rangle + \log\left(\frac{\pi_k}{\pi_1}\right)$$

It can be shown that

$$|w_1(X_j; \boldsymbol{\mu}) - w_1(X_j; \boldsymbol{\mu}')| \leq \frac{\sqrt{M}}{4}\|\eta_j(\boldsymbol{\mu}) - \eta_j(\boldsymbol{\mu}')\|$$

Applying the vector-valued contraction lemma,

$$\mathbb{E}\left[\sup_{\boldsymbol{\mu} \in \mathbb{A}} \frac{1}{n}\sum_{j=1}^n \epsilon_j w_i(X_j; \boldsymbol{\mu})\langle X_j, u\rangle\right] \leq \mathbb{E}\left[\frac{\sqrt{2}\sqrt{M}}{4n}\sup_{\boldsymbol{\mu} \in \mathbb{A}} \sum_{j=1}^n \sum_{k=1}^M \epsilon_{jk}[\eta_j(\boldsymbol{\mu})]_k\right]$$

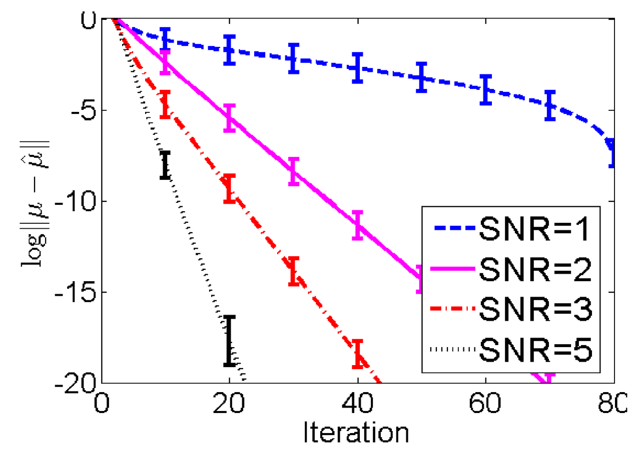
Bounding the right hand side, we have

$$R_n(\mathcal{F}) \leq \frac{cM^{3/2}(1+R_{\max})^3\sqrt{d}\max\{1, \log(\kappa)\}}{\sqrt{n}}$$

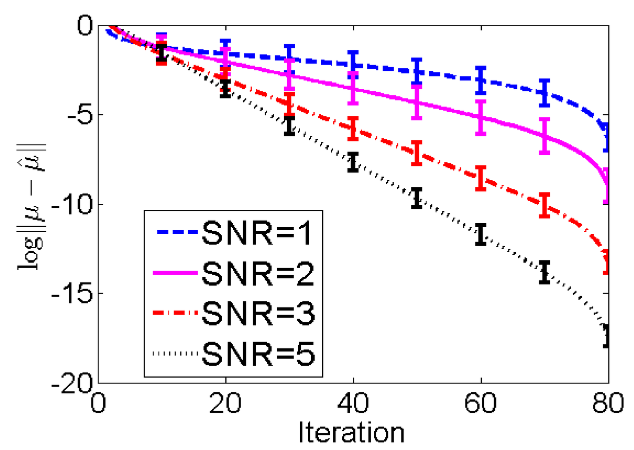
Simulation

All settings indicate the linear convergence rate as shown in the analysis;

Increasing imbalance of cluster weights slows down the local convergence rate.



$\boldsymbol{\pi} = (1/3, 1/3, 1/3)$



$\boldsymbol{\pi} = (0.6, 0.3, 0.1)$

References

[1] Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *Ann. Statist.*, 45(1):77–120, 02 2017.

[2] Richard Combes. An extension of mcdiarmid's inequality. *arXiv preprint arXiv:1511.05240*, 2015.

[3] Aryeh Kontorovich. Concentration in unbounded metric spaces and algorithmic stability. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 28–36, 2014.

[4] Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pages 3–17. Springer, 2016.

Extended version at `https://arxiv.org/abs/1705.08530`