Converging Fast and Slow: Different Avatars of EM

Raaz Dwivedi
Department of EECS, UC Berkeley
Mixture models: Usefulness

- Heterogenous sub-populations in various datasets
- Topic modeling, Financial returns
- Image annotation, classification, segmentation

Source: Blei et al. 2003

Source: Rotem et al. 2007
Mixture models: Formulation

- Distribution of observed variable $X$ in a latent variable model with labels $Z$

$$Z \sim \text{multinomial}(w_1, \ldots, w_K)$$

$$[X | Z = k] \sim \mathcal{P}_k$$

$$X \sim \sum_{k=1}^{K} w_k \mathcal{P}_k$$

- $\mathcal{P}_k = \mathcal{N}(\mu_k, \Sigma_k)$ results in Gaussian mixture model, arguably the most popular in practice

- Given $X_1, \ldots, X_n$, how do we estimate the parameters? Lack of $Z$ makes the problem non-convex
Mixture models:
Parameter estimation

Mixture models:
Parameter estimation

• Method of choice: Expectation-Maximization
  (Dempster-Laird-Rubin, Sundberg, Martin-Löf, Jeff Wu 1970-80)
Mixture models: Parameter estimation

Theoretical Guarantees for EM: Asymptotic and non-asymptotic analysis


• Several recent works on the non-asymptotic behavior of EM in $\mathbb{R}^d$ with $n$ samples
Theoretical Guarantees for EM: Well-specified 2-Gaussian Mixtures

True Model: \[ \frac{1}{2} \mathcal{N}(-\theta^*, \mathbb{I}_d) + \frac{1}{2} \mathcal{N}(\theta^*, \mathbb{I}_d) \]

Fitted model: \[ \frac{1}{2} \mathcal{N}(-\theta, \mathbb{I}_d) + \frac{1}{2} \mathcal{N}(\theta, \mathbb{I}_d) \]
Theoretical Guarantees for EM: Well-specified 2-Gaussian Mixtures

**True Model:** \[ \frac{1}{2} \mathcal{N}(-\theta^*, \mathbb{I}_d) + \frac{1}{2} \mathcal{N}(\theta^*, \mathbb{I}_d) \]

**Fitted model:** \[ \frac{1}{2} \mathcal{N}(-\theta, \mathbb{I}_d) + \frac{1}{2} \mathcal{N}(\theta, \mathbb{I}_d) \]

(Balakrishnan, Wainwright, Yu ’17)

EM with good initialization + Strong Signal \( \|\theta^*\| > C \):

\[ \|\theta^t_n - \theta^*\|_2 \approx \sqrt{\frac{d}{n}} \quad \text{for} \quad t \approx \log \left( \frac{n}{d} \right) \quad \text{and} \quad n \approx d \]
Well-initialized EM on well-specified well-separated mixtures:

\[ \sqrt{d \over n} \text{ error in } \log {n \over d} \text{ steps} \]

Cai, Ma and Zhang, 2019
General, well-separated 2-mixtures
Fitted with 2-mixtures

Yan, Yin and Sarkar, 2017
Spherical, well-separated k-mixtures
Fitted with k spherical mixtures

Other works: Wang+ 2015, Daskalakis+ 2017, Hao+ 2018, …
“But what happens when the components are too close to each other?”

“Or, when the number of components is over-specified in the fitted model?”

EM slows down… some old works? but can we quantify it?
We consider the simplest over-specified case: True model has **one** component and we fit **two** components

**True Model:** \( \mathcal{N}(0, \mathbb{I}_d) \)

\[
= \frac{1}{2} \mathcal{N}(\theta^*, \mathbb{I}_d) + \frac{1}{2} \mathcal{N}(-\theta^*, \mathbb{I}_d) \quad \text{with} \quad \theta^* = 0
\]

**Fitted model:** \( \frac{1}{2} \mathcal{N}(\theta, \mathbb{I}_d) + \frac{1}{2} \mathcal{N}(-\theta, \mathbb{I}_d) \)
Converging fast and slow: Statistical rates for EM estimates vs SNR

\[ \mu_n = \begin{cases} 0, & \text{No signal}: \theta^* = 0. \\ 1, & \text{Strong signal}: \theta^* = 1 \end{cases} \]

slope = -0.25

slope = -0.50
Our main result:
Convergence of sample EM with weak signal

In the case of no signal $\theta^* = 0$, for arbitrary initialization, the sample EM iterates satisfy

$$\|\theta^t_n - \theta^*\|_2 \lesssim \left( \frac{d}{n} \right)^{1/4} \quad \text{for} \quad t \gtrsim \left( \frac{n}{d} \right)^{1/2} \quad \text{and} \quad n \gtrsim d,$$
Converging fast and slow

In the case of no signal $\theta^* = 0$, for arbitrary initialization, the sample EM iterates satisfy

\[
\|\theta_n^t - \theta^*\|_2 \lesssim \left(\frac{d}{n}\right)^{1/4} \quad \text{for} \quad t \gtrsim \left(\frac{n}{d}\right)^{1/2} \quad \text{and} \quad n \gtrsim d,
\]

For strong signal $\|\theta^*\| > C$, sample EM iterates satisfy

\[
\|\theta_n^t - \theta^*\|_2 \lesssim \left(\frac{d}{n}\right)^{1/2} \quad \text{for} \quad t \gtrsim \log \left(\frac{n}{d}\right) \quad \text{and} \quad n \gtrsim d.
\]

Balakrishnan+ 2017
Converging fast and slow

In the case of no signal $\theta^* = 0$, for arbitrary initialization, the sample EM iterates satisfy

$$\|\theta_n^t - \theta^*\|_2 \lesssim \left(\frac{d}{n}\right)^{1/4} \quad \text{for} \quad t \gtrsim \left(\frac{n}{d}\right)^{1/2} \quad \text{and} \quad n \gtrsim d,$$

For strong signal $\|\theta^*\| > C$, sample EM iterates satisfy

$$\|\theta_n^t - \theta^*\|_2 \lesssim \left(\frac{d}{n}\right)^{1/2} \quad \text{for} \quad t \gtrsim \log \left(\frac{n}{d}\right) \quad \text{and} \quad n \gtrsim d.$$

Balakrishnan+ 2017
Zero SNR:
Statistical rates for non-special fits

Statistical error vs \( n \)

All cases show scaling \( n^{-1/4} \)

Number of samples \( n \)

- 2-mixture-fit: \( 0.5\mathcal{N}(\theta_1, 1) + 0.5\mathcal{N}(\theta_2, 1) \)
- 2-mixture-fit: \( w\mathcal{N}(\theta_1, 1) + (1 - w)\mathcal{N}(\theta_2, 1) \)
- 3-mixture-fit: \( \sum_{k=1}^{3} w_k\mathcal{N}(\theta_k, 1) \)
- slope = -0.26
Zero signal
= Degenerate Fisher matrix
= Flatter log-likelihood

Flatness: EM takes more iterations to converge
Zero signal
= Degenerate Fisher matrix
= Slow rate for MLE [Chen 1995, Rousseau 2011, Nguyen 2013, Ho+ 2018]

MLE farther from $\theta^*$: slower statistical rate for EM estimates
Proving the slow rates
Closed form updates for EM

Fitted model:
\[ \frac{1}{2} \mathcal{N}(\theta, 1) + \frac{1}{2} \mathcal{N}(-\theta, 1) \]

Population EM iteration:
\[ \theta^{t+1} = \mathbb{E}[X \tanh(X^\top \theta^t)] =: M(\theta^t) \]

Sample EM iteration:
\[ \theta^{t+1}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \tanh(X_i^\top \theta^t_n) =: M_n(\theta^t_n) \]

Can study the updates via the operators \( M \) and \( M_n \)
Proof strategy:
From population to sample analysis

\[ \| \theta_{n}^{t+1} - \theta^* \| = \| M_n(\theta_{n}^t) - \theta^* \| \]

\[ \leq \| M(\theta_{n}^t) - \theta^* \| + \| M_n(\theta_{n}^t) - M(\theta_{n}^t) \| \]

- **Population**-level behavior
- **Deterministic** analysis
- Characterizes the “algorithmic” rate of convergence

- **Finite sample** perturbation error
- **Probabilistic** analysis
- Characterizes the “statistical” rate of convergence
Proof strategy:
From population to sample analysis

\[ \| \theta_{n+1}^t - \theta^* \| = \| M_n(\theta_n^t) - \theta^* \| \]
\[ \leq \| M(\theta_n^t) - \theta^* \| + \| M_n(\theta_n^t) - M(\theta_n^t) \| \]

Balakrishnan+ 2017: For strong signal

\[ \| M(\theta) - \theta^* \| \leq \kappa \| \theta - \theta^* \| \]
\[ (\kappa < 1 - c) \]

Our work: for no signal

\[ \| M(\theta) - \theta^* \| \asymp (1 - c \| \theta - \theta^* \|^2) \cdot \| \theta - \theta^* \| \]
\[ \kappa(\theta) \to 1 \text{ as } \theta \to \theta^* \]
Proof strategy:
From population to sample analysis

\[ ||M(\theta^t_n) - \theta^*|| \]

Population EM sequence \[ \theta^{t+1} = M(\theta^t) \]

\[ e^{-t} \sim \epsilon = n^{-1/2} \Rightarrow t = O(\log n) \]
Proof strategy:
From population to sample analysis

\[ \| M(\theta_n^t) - \theta^* \| \]

Population EM sequence \( \theta^{t+1} = M(\theta^t) \)

- \( e^{-t} \sim \epsilon = n^{-1/4} \Rightarrow t = O(\sqrt{n}) \)
- \( e^{-t} \sim \epsilon = n^{-1/2} \Rightarrow t = O(\log n) \)

- \( \theta^* = 1 \)
- \( \theta^* = 0 \)
- \( \theta^{t+1} = \theta^t / (1 + (\theta^t)^2) \)

computational slow-down
Proof strategy:
From population to sample analysis

\[
\|\theta_{t+1}^{n} - \theta^*\| = \|M_n(\theta_{t}^{n}) - \theta^*\|
\]

\[
\leq \|M(\theta_{t}^{n}) - \theta^*\| + \|M_n(\theta_{t}^{n}) - M(\theta_{t}^{n})\|
\]

\[
\leq \kappa \|\theta_{t}^{n} - \theta^*\| + C \sqrt{\frac{d}{n}}
\]

Strong Signal

statistical slow-down
Proof strategy:
From population to sample analysis

\[
\|\theta_{n}^{t+1} - \theta^*\| = \|M_n(\theta_n^t) - \theta^*\| \\
\leq \|M(\theta_n^t) - \theta^*\| + \|M_n(\theta_n^t) - M(\theta_n^t)\| \\
\leq \kappa \|\theta_n^t - \theta^*\| + C \sqrt{\frac{d}{n}} \\
\lesssim \sqrt{\frac{d}{n}} \cdot \frac{1}{1 - \kappa} \\
\text{for } t \gtrsim \log_{1/\kappa} \left( \frac{n}{d} \cdot \|\theta^0 - \theta^*\| \right) \\
\text{we are done since } 1 - \kappa > c > 0
\]
Proof strategy: From population to sample analysis

\[ \| \theta_{n}^{t+1} - \theta^* \| = \| M_n(\theta_n^t) - \theta^* \| \]
\[ \leq \| M(\theta_n^t) - \theta^* \| + \| M_n(\theta_n^t) - M(\theta_n^t) \| \]

Strong Signal

\[ \leq \kappa \| \theta_n^t - \theta^* \| + C \sqrt{\frac{d}{n}} \]
\[ \lesssim \sqrt{\frac{d}{n}} \cdot \frac{1}{1 - \kappa} \]

for \( t \geq \log_{1/\kappa} \left( \frac{n}{d} \cdot \| \theta^0 - \theta^* \| \right) \)

we are done since \( 1 - \kappa > c > 0 \)

Weak Signal

\[ 1 - \kappa(\theta) \approx \| \theta - \theta^* \|^2 \]
\[ \downarrow \text{(implicit equation)} \]
\[ \| \hat{\theta}_n - \theta^* \| \lesssim \sqrt{\frac{d}{n}} \cdot \frac{1}{\| \hat{\theta}_n - \theta^* \|^2} \]
\[ \downarrow \]
\[ \| \hat{\theta}_n - \theta^* \| \lesssim \left( \frac{d}{n} \right)^{1/6} \]

sub-optimal compared to \( n^{-1/4} \)
Sharpening the proof:
Localize the estimates in a ball

\[ \|\theta_n^t - \theta^*\| \leq n^{-b} \]

A standard technique in empirical process theory to derive sharp minimax rates

But \( \kappa \) gets too close to 1 if \( \theta_n^t \) is too close to \( \theta^* \)
Sharpening the proof:
Localize the estimates in an **annulus**

\[
n^{-a} \leq \|\theta_n^t - \theta^*\| \leq n^{-b}
\]
Sharpening the proof:
Localize the estimates in an **annulus**

\[
n^{-a} \leq \|\theta_t^n - \theta^*\| \leq n^{-b}
\]

Outer radius provides a control on the perturbation error

\[
\|M(\theta_t^n) - M_n(\theta_t^n)\| \leq \frac{n^{-b}}{\sqrt{n}}
\]

Inner radius helps to control the contraction

\[
1 - \kappa(\theta_t^n) \geq n^{-2a}
\]

Leads to a recursion between \(a\) and \(b\) with a unique fixed point \(1/4\)

\[
a = \frac{1}{3}(b + \frac{1}{2})
\]
Summary

Over-specification / weak signal is a double-edged sword

statistical slow-down
\[ n^{-\frac{1}{4}} \text{ vs } n^{-\frac{1}{2}} \]

computational slow-down
\[ n^{\frac{1}{2}} \text{ vs } \log n \]
Summary

Over-specification / weak signal is a double-edged sword

\[ n^{-1/4} \text{ vs } n^{-1/2} \]

\[ n^{1/2} \text{ vs } \log n \]

Blessing in disguise?
A future recipe for model selection: Look at EM iterations
Follow-up work

We assume zero signal:

Wu and Zhou [2019] generalize it to a minimax weak signal setting (under restrictive initialization conditions)

We assume known variance:

Our recent work shows that fitting an over-specified model with unknown variance may lead to further slow-down ($n^{-1/8}$)

Localization beyond EM:

We employ localization techniques to derive sharp rates beyond mixture models (draft in progress)
Thank you!

Over-specification / weak signal is a double-edged sword

- Statistical slow-down: $n^{-1/4}$ vs $n^{-1/2}$
- Computational slow-down: $n^{1/2}$ vs $\log n$