Refined Convergence Rates for Maximum Likelihood Estimation under Finite Mixture Models

Tudor Manole

Department of Statistics and Data Science
Carnegie Mellon University

Joint work with: Nhat Ho (University of Texas, Austin)

International Conference on Machine Learning, July 2022

Finite Mixture Models

- Let $\mathcal{F} = \{f(\cdot;\theta) : \theta \in \Theta\}$ be a parameteric density family with parameter space Θ.

Finite Mixture Models

- Let $\mathcal{F} = \{f(\cdot;\theta) : \theta \in \Theta\}$ be a parameteric density family with parameter space Θ.
  - e.g. $f(\cdot;\theta)$ could be the $N(\mu, \Sigma)$ density with parameter $\theta = (\mu, \Sigma) \in \Theta \subseteq \mathbb{R}^d \times \mathbb{S}^d_{++}$.

Finite Mixture Models

- Let $\mathcal{F} = \{f(\cdot;\theta) : \theta \in \Theta\}$ be a parameteric density family with parameter space Θ.
  - e.g. $f(\cdot;\theta)$ could be the $N(\mu,\Sigma)$ density with parameter $\theta = (\mu,\Sigma) \in \Theta \subseteq \mathbb{R}^d \times \mathbb{S}^d_{++}$.
- Given an integer $K \geq 1$, assume

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p_{G_0}(x) = \sum_{k=1}^{K} \pi_k^0 f(x;\theta_k^0)$$

Finite Mixture Models

- Let $\mathcal{F} = \{f(\cdot;\theta) : \theta \in \Theta\}$ be a parameteric density family with parameter space Θ.
  - e.g. $f(\cdot;\theta)$ could be the $N(\mu,\Sigma)$ density with parameter $\theta = (\mu,\Sigma) \in \Theta \subseteq \mathbb{R}^d \times \mathbb{S}^d_{++}$.
- Given an integer $K \geq 1$, assume

$$X_1,\ldots,X_n \overset{\text{i.i.d.}}{\sim} p_{G_0}(x) = \sum_{k=1}^{K} \pi_k^0 f(x;\theta_k^0)$$

  - **Mixing Proportions**: $0 < \pi_k^0 \leq 1$, $\sum_{k=1}^{K} \pi_k^0 = 1$.

Finite Mixture Models

- Let $\mathcal{F} = \{f(\cdot;\theta) : \theta \in \Theta\}$ be a parameteric density family with parameter space Θ.
  - e.g. $f(\cdot;\theta)$ could be the $N(\mu,\Sigma)$ density with parameter $\theta = (\mu,\Sigma) \in \Theta \subseteq \mathbb{R}^d \times \mathbb{S}^d_{++}$.
- Given an integer $K \geq 1$, assume

$$X_1,\dots,X_n \overset{\text{i.i.d.}}{\sim} p_{G_0}(x) = \sum_{k=1}^{K} \pi_k^0 f(x;\theta_k^0)$$

  - **Mixing Proportions**: $0 < \pi_k^0 \leq 1$, $\sum_{k=1}^{K} \pi_k^0 = 1$.
  - **Atoms**: $\theta_k^0 \in \Theta$, possibly overlapping.

Finite Mixture Models

- Let $\mathcal{F} = \{f(\cdot;\theta) : \theta \in \Theta\}$ be a parameteric density family with parameter space Θ.
  - e.g. $f(\cdot;\theta)$ could be the $N(\mu, \Sigma)$ density with parameter $\theta = (\mu, \Sigma) \in \Theta \subseteq \mathbb{R}^d \times \mathbb{S}^d_{++}$.
- Given an integer $K \geq 1$, assume

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p_{G_0}(x) = \sum_{k=1}^{K} \pi_k^0 f(x;\theta_k^0) = \int_\Theta f(x;\theta) dG_0(\theta)$$

  - **Mixing Proportions**: $0 < \pi_k^0 \leq 1$, $\sum_{k=1}^{K} \pi_k^0 = 1$.
  - **Atoms**: $\theta_k^0 \in \Theta$, possibly overlapping.
  - **Mixing Measure**:

$$G_0 = \sum_{k=1}^{K} \pi_k^0 \delta_{\theta_k^0}$$

Maximum Likelihood Estimation (MLE) in Finite Mixtures

Let $\mathcal{O}_K$ denote the set of mixing measures with at most K components, and define

$$\widehat{G}_n = \sum_{j=1}^{K} \hat{\pi}_j \delta_{\hat{\theta}_j} = \operatorname*{argmax}_{G \in \mathcal{O}_K} \sum_{i=1}^{n} \log p_G(X_i).$$

Maximum Likelihood Estimation (MLE) in Finite Mixtures

Let $\mathcal{O}_K$ denote the set of mixing measures with at most K components, and define

$$\widehat{G}_n = \sum_{j=1}^{K} \hat{\pi}_j \delta_{\hat{\theta}_j} = \operatorname*{argmax}_{G \in \mathcal{O}_K} \sum_{i=1}^{n} \log p_G(X_i).$$

$\boxed{\textbf{What is the risk of } \widehat{G}_n\text{?}}$

The Wasserstein Distances

- To quantify the risk of $\widehat{G}_n$, we require a loss function on $\mathcal{O}_K$.

The Wasserstein Distances

- To quantify the risk of $\widehat{G}_n$, we require a loss function on $\mathcal{O}_K$.
- Nguyen'13 proposed to use the r-Wasserstein distances ($r \geq 1$):

$$W_r^r(G, G') = \inf_{\substack{\theta, \theta' \\ \theta \sim G \\ \theta' \sim G'}} \mathbb{E}\Big[\|\theta - \theta'\|^r\Big], \quad G, G' \in \mathcal{O}_K,$$

State of the Art

- Pointwise convergence rate for "strongly identifiable" families $\mathcal{F}$:

$$\mathbb{E}\Big[W_2(\widehat{G}_n, G_0)\Big] \lesssim_{G_0} n^{-\frac{1}{4}}$$

(Chen'95, Ho & Nguyen'16)

State of the Art

- Pointwise convergence rate for "strongly identifiable" families $\mathcal{F}$:

$$\mathbb{E}\Big[W_2(\widehat{G}_n, G_0)\Big] \lesssim_{G_0} n^{-\frac{1}{4}} \qquad \text{(Chen'95, Ho \& Nguyen'16)}$$

- Slower pointwise rates hold for location-scale Gaussian mixtures (Ho & Nguyen'16).

State of the Art

- Pointwise convergence rate for "strongly identifiable" families $\mathcal{F}$:

$$\mathbb{E}\Big[W_2(\widehat{G}_n, G_0)\Big] \lesssim_{G_0} n^{-\frac{1}{4}} \qquad \text{(Chen'95, Ho \& Nguyen'16)}$$

- Slower pointwise rates hold for location-scale Gaussian mixtures (Ho & Nguyen'16).
- Uniform rate for strongly identifiable families $\mathcal{F}$:

$$\sup_{G_0 \in \mathcal{O}_K} \mathbb{E}\Big[W_1(\widehat{G}_n, G_0)\Big] \lesssim n^{-\frac{1}{4K-2}} \qquad \text{(Heinrich \& Kahn'18)}$$

State of the Art

- Pointwise convergence rate for "strongly identifiable" families $\mathcal{F}$:

$$\mathbb{E}\Big[W_2(\widehat{G}_n, G_0)\Big] \lesssim_{G_0} n^{-\frac{1}{4}} \qquad \text{(Chen'95, Ho \& Nguyen'16)}$$

- Slower pointwise rates hold for location-scale Gaussian mixtures (Ho & Nguyen'16).
- Uniform rate for strongly identifiable families $\mathcal{F}$:

$$\sup_{G_0 \in \mathcal{O}_K} \mathbb{E}\Big[W_1(\widehat{G}_n, G_0)\Big] \lesssim n^{-\frac{1}{4K-2}} \qquad \text{(Heinrich \& Kahn'18)}$$

Our Contribution: In each of these settings, the Wasserstein distance can be replaced by a stronger loss function which implies faster convergence rates for the atoms of $\widehat{G}_n$.

State of the Art

- Pointwise convergence rate for "strongly identifiable" families $\mathcal{F}$:

$$\mathbb{E}\left[W_2(\widehat{G}_n, G_0)\right] \lesssim_{G_0} n^{-\frac{1}{4}}$$ (Chen'95, Ho & Nguyen'16)

- Slower pointwise rates hold for location-scale Gaussian mixtures (Ho & Nguyen'16).
- Uniform rate for strongly identifiable families $\mathcal{F}$:

$$\sup_{G_0 \in \mathcal{O}_K} \mathbb{E}\left[W_2(\widehat{G}_n, G_0)\right] \lesssim n^{-\frac{1}{4K-2}}$$ (Heinrich & Kahn'18)

Our Contribution: In each of these settings, the Wasserstein distance can be replaced by a stronger loss function which implies faster convergence rates for the atoms of $\widehat{G}_n$.

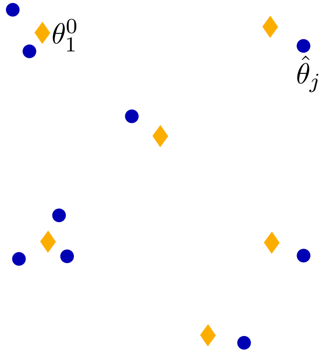
Interpreting Convergence in Wasserstein Distance

$$\boxed{\mathbb{E}\Big[W_2(\widehat{G}_n, G_0)\Big] \lesssim_{G_0} n^{-\frac{1}{4}}}$$

• Atoms $\hat{\theta}_j$ of $\widehat{G}_n$ ♦Atoms θ_k^0 of G_0.

Interpreting Convergence in Wasserstein Distance

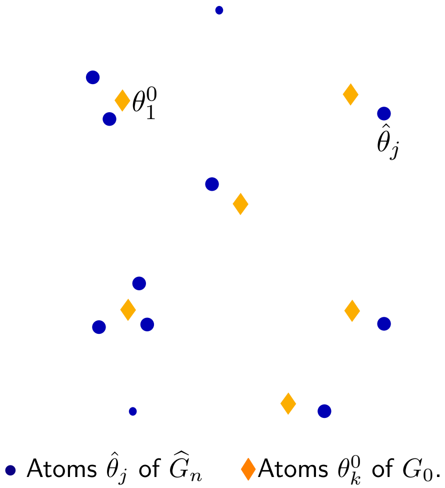
$$\boxed{\mathbb{E}\Big[W_2(\widehat{G}_n, G_0)\Big] \lesssim_{G_0} n^{-\frac{1}{4}}}$$



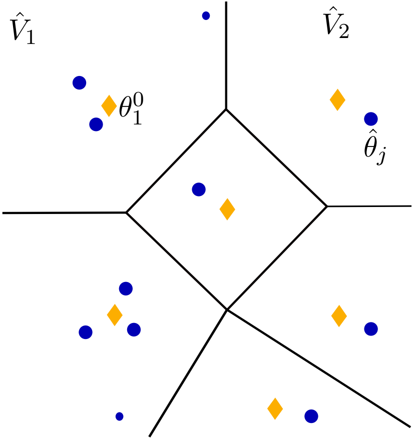
● Atoms $\hat{\theta}_j$ of $\widehat{G}_n$ ◆ Atoms θ_k^0 of G_0.

Interpreting Convergence in Wasserstein Distance

$$\boxed{\mathbb{E}\left[W_2(\widehat{G}_n, G_0)\right] \lesssim_{G_0} n^{-\frac{1}{4}}}$$



Interpreting Convergence in Wasserstein Distance



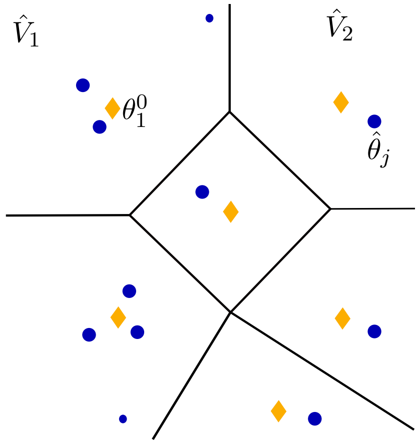
• Atoms $\hat{\theta}_j$ of $\widehat{G}_n$ ◆Atoms θ_k^0 of G_0.

$$\boxed{\mathbb{E}\Big[W_2(\widehat{G}_n, G_0)\Big] \lesssim_{G_0} n^{-\frac{1}{4}}}$$

- Voronoi Cells:

$$\hat{V}_k = \Big\{ j : \|\hat{\theta}_j - \theta_k^0\| < \|\hat{\theta}_j - \theta_l^0\|, \forall l \neq k \Big\}$$

Interpreting Convergence in Wasserstein Distance



• Atoms $\hat{\theta}_j$ of $\widehat{G}_n$ ◆Atoms θ_k^0 of G_0.

$$\boxed{\mathbb{E}\Big[W_2(\widehat{G}_n, G_0)\Big] \lesssim_{G_0} n^{-\frac{1}{4}}}$$

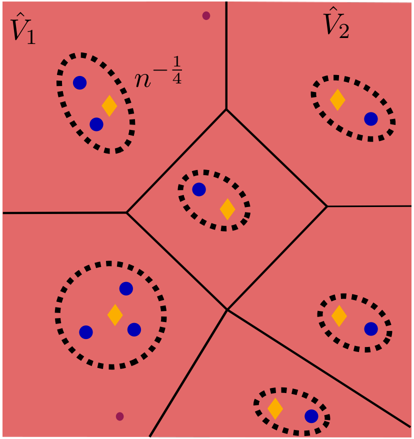
- Voronoi Cells:

$$\hat{V}_k = \Big\{ j : \|\hat{\theta}_j - \theta_k^0\| < \|\hat{\theta}_j - \theta_l^0\|, \forall l \neq k \Big\}$$

- Rate Interpretation: For all k, and $j \in \hat{V}_k$,

$$\mathbb{E}\|\hat{\theta}_j - \theta_k^0\| \lesssim n^{-\frac{1}{4}}, \quad \text{or} \quad \mathbb{E}[\hat{\pi}_j] \lesssim n^{-\frac{1}{2}}.$$

Interpreting Convergence in Wasserstein Distance



• Atoms $\hat{\theta}_j$ of $\widehat{G}_n$ ◆ Atoms θ_k^0 of G_0.

$$\boxed{\mathbb{E}\Big[W_2(\widehat{G}_n, G_0)\Big] \lesssim_{G_0} n^{-\frac{1}{4}}}$$

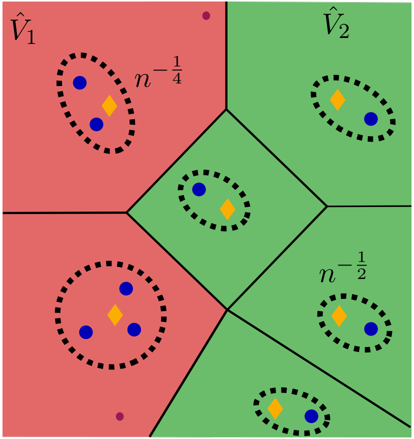
- Voronoi Cells:

$$\hat{V}_k = \Big\{ j : \|\hat{\theta}_j - \theta_k^0\| < \|\hat{\theta}_j - \theta_l^0\|, \forall l \neq k \Big\}$$

- Rate Interpretation: For all k, and $j \in \hat{V}_k$,

$$\mathbb{E}\|\hat{\theta}_j - \theta_k^0\| \lesssim n^{-\frac{1}{4}}, \quad \text{or} \quad \mathbb{E}[\hat{\pi}_j] \lesssim n^{-\frac{1}{2}}.$$

Interpreting Convergence in Wasserstein Distance



● Atoms $\hat{\theta}_j$ of $\widehat{G}_n$ ◆Atoms θ_k^0 of G_0.

$$\boxed{\mathbb{E}\Big[W_2(\widehat{G}_n, G_0)\Big] \lesssim_{G_0} n^{-\frac{1}{4}}}$$

- Voronoi Cells:

$$\hat{V}_k = \Big\{ j : \|\hat{\theta}_j - \theta_k^0\| < \|\hat{\theta}_j - \theta_l^0\|, \forall l \neq k \Big\}$$

- Rate Interpretation: For all k, and $j \in \hat{V}_k$,

$$\mathbb{E}\|\hat{\theta}_j - \theta_k^0\| \lesssim n^{-\frac{1}{4}}, \quad \text{or} \quad \mathbb{E}[\hat{\pi}_j] \lesssim n^{-\frac{1}{2}}.$$

- Key Observation: This rate is loose for all k such that $|\hat{V}_k| = 1$.

Main Result: Refined Convergence Rate of the MLE

Theorem (Informal). Assume strong identifiability and mild regularity conditions.

(1) (Fast Rate) For all k such that $|\hat{V}_k| = 1$, and $j \in \hat{V}_k$, it holds that

$$\mathbb{E}\|\hat{\theta}_j - \theta_k^0\| \lesssim_{G_0} n^{-\frac{1}{2}}.$$

Main Result: Refined Convergence Rate of the MLE

Theorem (Informal). Assume strong identifiability and mild regularity conditions.

(1) (Fast Rate) For all k such that $|\hat{V}_k| = 1$, and $j \in \hat{V}_k$, it holds that

$$\mathbb{E}\|\hat{\theta}_j - \theta_k^0\| \lesssim_{G_0} n^{-\frac{1}{2}}.$$

(2) (Slow Rate) For all k such that $|\hat{V}_k| > 1$, and $j \in \hat{V}_k$, it either holds that

$$\mathbb{E}\|\hat{\theta}_j - \theta_k^0\| \lesssim_{G_0} n^{-\frac{1}{4}}, \quad \text{or} \quad \mathbb{E}[\hat{\pi}_j] \lesssim_{G_0} n^{-\frac{1}{2}}.$$

Main Result: Refined Convergence Rate of the MLE

Theorem (Informal). Assume strong identifiability and mild regularity conditions.

(1) (Fast Rate) For all k such that $|\hat{V}_k| = 1$, and $j \in \hat{V}_k$, it holds that

$$\mathbb{E}\|\hat{\theta}_j - \theta_k^0\| \lesssim_{G_0} n^{-\frac{1}{2}}.$$

(2) (Slow Rate) For all k such that $|\hat{V}_k| > 1$, and $j \in \hat{V}_k$, it either holds that

$$\mathbb{E}\|\hat{\theta}_j - \theta_k^0\| \lesssim_{G_0} n^{-\frac{1}{4}}, \quad \text{or} \quad \mathbb{E}[\hat{\pi}_j] \lesssim_{G_0} n^{-\frac{1}{2}}.$$

- In contrast, past work only implied option (2) **for all** k**.**

Main Result: Refined Convergence Rate of the MLE

Theorem (Informal). Assume strong identifiability and mild regularity conditions.

(1) (Fast Rate) For all k such that $|\hat{V}_k| = 1$, and $j \in \hat{V}_k$, it holds that

$$\mathbb{E}\|\hat{\theta}_j - \theta_k^0\| \lesssim_{G_0} n^{-\frac{1}{2}}.$$

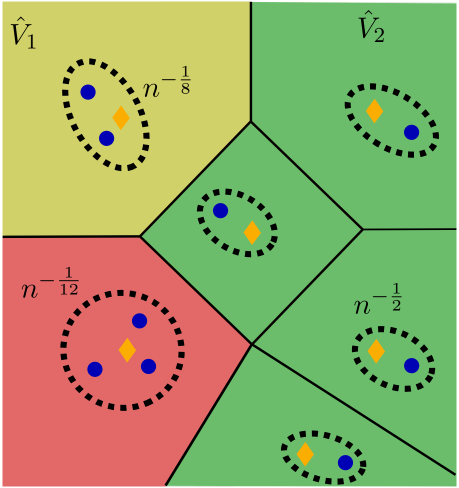
(2) (Slow Rate) For all k such that $|\hat{V}_k| > 1$, and $j \in \hat{V}_k$, it either holds that

$$\mathbb{E}\|\hat{\theta}_j - \theta_k^0\| \lesssim_{G_0} n^{-\frac{1}{4}}, \quad \text{or} \quad \mathbb{E}[\hat{\pi}_j] \lesssim_{G_0} n^{-\frac{1}{2}}.$$

- In contrast, past work only implied option (2) **for all** k**.**
- We prove this by introducing a new loss function $\mathcal{D}$ which satisfies $\mathcal{D} \gtrsim W_2$ and

$$\mathbb{E}\Big[\mathcal{D}(\widehat{G}_n, G_0)\Big] \lesssim_{G_0} n^{-\frac{1}{4}}.$$

A Peak at our Refinements for Location-Scale Gaussian Mixtures



Atoms $\hat{\theta}_j = \begin{pmatrix} \hat{\mu}_j \\ \hat{\sigma}_j \end{pmatrix}$ of $\widehat{G}_n$ Atoms $\theta_k^0 = \begin{pmatrix} \mu_k^0 \\ \sigma_k^0 \end{pmatrix}$ of G_0.

Summary and Discussion

- Past work painted an **overly pessimistic** view of parameter estimation in mixtures.

Summary and Discussion

- Past work painted an **overly pessimistic** view of parameter estimation in mixtures.
- W_r is only able to quantify the **worst-case** convergence rate among the atoms of $\widehat{G}_n$.

Summary and Discussion

- Past work painted an **overly pessimistic** view of parameter estimation in mixtures.
- W_r is only able to quantify the **worst-case** convergence rate among the atoms of $\widehat{G}_n$.
- Our divergences reveal the **heterogeneity** of convergence rates among these atoms.

Summary and Discussion

- Past work painted an **overly pessimistic** view of parameter estimation in mixtures.
- W_r is only able to quantify the **worst-case** convergence rate among the atoms of $\widehat{G}_n$.
- Our divergences reveal the **heterogeneity** of convergence rates among these atoms.
- Many open questions (EM algorithm, method of moments, etc.).

Thank You