Learning Gaussian Mixtures with Wasserstein-Fisher-Rao Gradient Flow

Yuling Yan (MIT)

joint work with K. Wang and P. Rigollet

IDSS MIT INSTITUTE FOR DATA, SYSTEMS, AND SOCIETY

LIDS

Overview

A story that connects statistics, machine learning, and mathematics

- Gaussian mixture model
- overparameterization
- nonparametric MLE
- gradient flow over the space of probability measures
- interacting particle systems

"Learning Gaussian Mixtures Using Wasserstein-Fisher-Rao Gradient Flow," Y. Yan, K. Wang, P. Rigollet, accepted to Annals of Statistics, 2024.

Part 1: Gaussian mixture model

Finite component GMM

K-component Gaussian mixture model:

$$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N \overset{\text{iid}}{\sim} \sum_{k=1}^{K} \omega_k^\star \mathcal{N}(\boldsymbol{\mu}_k^\star, \boldsymbol{I}_d)$$

- assume the weights $\{\omega_k^\star\}_{k=1}^K$ are known
- goal: estimate the centers $\{\boldsymbol{\mu}_k^\star\}_{k=1}^K$ from the data
- maximum likelihood estimation:

$$\underset{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K \in \mathbb{R}^d}{\text{minimize}} \;\; \ell(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K) := -\frac{1}{N} \sum_{i=1}^{N} \log \left[\sum_{j=1}^{K} \omega_j^\star \phi(\boldsymbol{X}_i - \boldsymbol{\mu}_j) \right]$$

- solve by EM or gradient descent (GD)?

Instability of GD and EM

Solving MLE is notoriously hard...

- When $K = 2$, everything is fine and easy
  - EM with random init converges to MLE (Wu and Zhou, 2019)
  - even a simple spectral method achieves optimality
- When $K \geq 3$, things become a little scary
  - $\exists$ hard instances s.t. with infinite samples, EM and any first-order algorithm fails with constant probability (Jin et al. 2016)
  - $\#$local minimizers is exponential in K (Chen and Xi, 2020)

Instability of GD and EM

A hard instance when $K = 3$ and $N = 1500$:

$$\mu_1^\star = -1, \quad \mu_2^\star = 1, \quad \mu_3^\star = 10, \quad \omega_1^\star = \omega_2^\star = \omega_3^\star = \frac{1}{3}$$

Run GD/EM with random initialization from data 100 times

- global minima: $\mu_1 \approx -1$, $\mu_2 \approx 1$, $\mu_3 \approx 10$
- bad local minima: $\mu_1 \approx 0$, $\mu_2 \approx \mu_3 \approx 10$
- GD (resp. EM) converges to bad local min 32 (resp. 28) times

Instability of GD and EM

A hard instance when $K = 3$ and $N = 1500$:

$$\mu_1^\star = -1, \quad \mu_2^\star = 1, \quad \mu_3^\star = 10, \quad \omega_1^\star = \omega_2^\star = \omega_3^\star = \frac{1}{3}$$

Run GD/EM with random initialization from data 100 times

- global minima: $\mu_1 \approx -1$, $\mu_2 \approx 1$, $\mu_3 \approx 10$
- bad local minima: $\mu_1 \approx 0$, $\mu_2 \approx \mu_3 \approx 10$
- GD (resp. EM) converges to bad local min 32 (resp. 28) times

Is there a way to avoid such difficulty?

Overparameterization

Choose $m \gg K$, imagine that GMM has m components, and solve the overparameterized MLE

$$\underset{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_m \in \mathbb{R}^d}{\text{minimize}} \quad \ell(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_m) := -\frac{1}{N} \sum_{i=1}^{N} \log \left[\frac{1}{m} \sum_{j=1}^{m} \phi(\boldsymbol{X}_i - \boldsymbol{\mu}_j) \right]$$

using GD with random initialization from the data.

- Motivation:
  - success of overparameterization in deep learning
  - stable, accurate numerical performance
  - in practice, K and weights are usually unknown
- Where do we expect it will converge to?

Lifting to $\mathcal{P}(\mathbb{R}^d)$

$$\underset{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K\in\mathbb{R}^d}{\text{minimize}} \sum_{i=1}^{N} \log\left[\sum_{j=1}^{K} \omega_j^\star \phi(\boldsymbol{X}_i - \boldsymbol{\mu}_j)\right] \tag{MLE}$$

Lifting to $\mathcal{P}(\mathbb{R}^d)$

$$\underset{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K \in \mathbb{R}^d}{\text{minimize}} \ \sum_{i=1}^{N} \log\left[\sum_{j=1}^{K} \omega_j^\star \phi(\boldsymbol{X}_i - \boldsymbol{\mu}_j)\right] \qquad \text{(MLE)}$$

$$\Updownarrow$$

$$\underset{\rho \in \mathbb{P}(\mathbb{R}^d)}{\text{minimize}} \ \ell(\rho) \coloneqq -\frac{1}{N}\sum_{i=1}^{N} \log\Big[(\rho * \phi)(\boldsymbol{X}_i)\Big] \quad \text{s.t.} \quad \underbrace{\rho = \sum_{j=1}^{K} \omega_j^\star \delta_{\boldsymbol{\mu}_j}}_{\rho \text{ is } K\text{-atomic}}$$

Lifting to $\mathcal{P}(\mathbb{R}^d)$

$$\underset{\boldsymbol{\mu}_1,\dots,\boldsymbol{\mu}_K \in \mathbb{R}^d}{\text{minimize}} \; \sum_{i=1}^{N} \log \left[\sum_{j=1}^{K} \omega_j^\star \phi(\boldsymbol{X}_i - \boldsymbol{\mu}_j) \right] \qquad \text{(MLE)}$$

$$\Updownarrow$$

$$\underset{\rho \in \mathbb{P}(\mathbb{R}^d)}{\text{minimize}} \; \ell(\rho) := -\frac{1}{N} \sum_{i=1}^{N} \log \Big[(\rho * \phi)(\boldsymbol{X}_i) \Big] \quad \text{s.t.} \quad \underbrace{\rho = \sum_{j=1}^{K} \omega_j^\star \delta_{\boldsymbol{\mu}_j}}_{\rho \text{ is } K\text{-atomic}}$$

$$\Downarrow$$

$$\underset{\rho \in \mathbb{P}(\mathbb{R}^d)}{\text{minimize}} \; \ell(\rho) \qquad \text{(nonparametric MLE)}$$

Lifting to $\mathcal{P}(\mathbb{R}^d)$

$$\underset{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K\in\mathbb{R}^d}{\text{minimize}} \quad \sum_{i=1}^N \log\left[\sum_{j=1}^K \omega_j^\star \phi(\boldsymbol{X}_i - \boldsymbol{\mu}_j)\right] \qquad \text{(MLE)}$$

$$\Updownarrow$$

$$\underset{\rho\in\mathbb{P}(\mathbb{R}^d)}{\text{minimize}} \;\; \ell(\rho) := -\frac{1}{N}\sum_{i=1}^N \log\Big[(\rho * \phi)(\boldsymbol{X}_i)\Big] \quad \text{s.t.} \quad \underbrace{\rho = \sum_{j=1}^K \omega_j^\star \delta_{\boldsymbol{\mu}_j}}_{\rho \text{ is } K\text{-atomic}}$$

$$\Downarrow$$

$$\underset{\rho\in\mathbb{P}(\mathbb{R}^d)}{\text{minimize}} \;\; \ell(\rho) \qquad \text{(nonparametric MLE)}$$

$$\Uparrow$$

$$\underset{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_m\in\mathbb{R}^d}{\text{minimize}} \;\; \ell(\rho) \quad \text{s.t.} \quad \rho = \frac{1}{m}\sum_{j=1}^m \delta_{\boldsymbol{\mu}_j} \qquad \text{(overparameterized MLE)}$$

Lifting to $\mathcal{P}(\mathbb{R}^d)$

$$\underset{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K \in \mathbb{R}^d}{\text{minimize}} \ \sum_{i=1}^{N} \log\left[\sum_{j=1}^{K} \omega_j^\star \phi(\boldsymbol{X}_i - \boldsymbol{\mu}_j)\right] \qquad \text{(MLE)}$$

$$\Updownarrow$$

$$\underset{\rho \in \mathbb{P}(\mathbb{R}^d)}{\text{minimize}} \ \ell(\rho) := -\frac{1}{N}\sum_{i=1}^{N} \log\left[(\rho * \phi)(\boldsymbol{X}_i)\right] \quad \text{s.t.} \quad \underbrace{\rho = \sum_{j=1}^{K} \omega_j^\star \delta_{\boldsymbol{\mu}_j}}_{\rho \text{ is } K\text{-atomic}}$$

$$\Downarrow$$

$$\underset{\rho \in \mathbb{P}(\mathbb{R}^d)}{\text{minimize}} \ \ell(\rho) \qquad \text{(nonparametric MLE)}$$

$$\Uparrow$$

$$\underset{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_m \in \mathbb{R}^d}{\text{minimize}} \ \ell(\rho) \quad \text{s.t.} \quad \rho = \frac{1}{m}\sum_{j=1}^{m} \delta_{\boldsymbol{\mu}_j} \qquad \text{(overparameterized MLE)}$$

Does GD for overparameterized MLE converge to NPMLE?

Part 2: Nonparametric MLE (NPMLE)

A general formulation

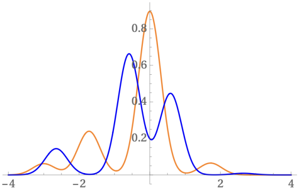
- **Observations**: $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \rho^\star * \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$
  - $\rho^\star \in \mathcal{P}(\mathbb{R}^d)$: unknown mixing distribution over $\mathbb{R}^d$

A general formulation

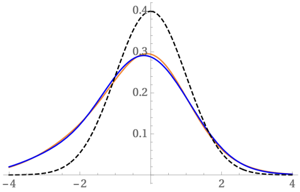
- **Observations**: $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \rho^\star * \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$
  - $\rho^\star \in \mathcal{P}(\mathbb{R}^d)$: unknown mixing distribution over $\mathbb{R}^d$
- **Goal**: learn the mixture distribution $\rho^\star * \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$

A general formulation

- **Observations**: $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \rho^\star * \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$
  - $\rho^\star \in \mathcal{P}(\mathbb{R}^d)$: unknown mixing distribution over $\mathbb{R}^d$
- **Goal**: learn the mixture distribution $\rho^\star * \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$
- **Approach**:
  - construct an estimate $\widehat{\rho}$ of the mixing distribution $\rho^\star$
  - estimate the mixture distribution with $\widehat{\rho} * \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$



(a) Mixing distributions



(b) Mixture distributions

figure credit: Y. Wu and P. Yang

Nonparametric MLE (NPMLE)

— Kiefer, Wolfowitz '56, Jewell '82, Lindsay '83, Polyanskiy, Wu '20

$$\widehat{\rho} \triangleq \operatorname*{arg\,min}_{\rho \in \mathcal{P}(\mathbb{R}^d)} \ell_N(\rho) = -\frac{1}{N} \sum_{i=1}^{N} \log \left[\rho * \phi(\boldsymbol{X}_i)\right]$$

- existence ✓
- uniqueness: $d = 1$ ✓, $d \geq 2$ ×
- structure: $d = 1$ ✓ $\widehat{\rho}$ is discrete, $O(\log N)$-atomic
  $d \geq 2$ ***?***
- minimax optimality ✓
- optimality condition: $\widehat{\rho}$ is NPMLE if and only if

$$\underbrace{\delta \ell_N(\widehat{\rho})(\boldsymbol{x})}_{\text{first variation}} \triangleq -\frac{1}{N} \sum_{i=1}^{N} \frac{\phi(\boldsymbol{x} - \boldsymbol{X}_i)}{(\widehat{\rho} * \phi)(\boldsymbol{X}_i)} \geq -1, \qquad \forall \boldsymbol{x} \in \mathbb{R}^d.$$

Existing algorithm: fixed-location EM

— *Jiang, Zhang '09, Koenker, Mizera '14*

- **Initialize:** fixed grid $\{\boldsymbol{\mu}_j\}_{1\le j\le m} \subset \mathbb{R}^d$, $\omega_0^{(1)} = \cdots = \omega_0^{(m)} = \frac{1}{m}$.
- **Iterative update:** for $t = 0, 1, \ldots, t_0 - 1$

$$\omega_{t+1}^{(j)} = \frac{1}{N} \sum_{i=1}^{N} \frac{\phi\left(\boldsymbol{X}_i - \boldsymbol{\mu}_j\right) \omega_t^{(j)}}{\sum_{l=1}^{m} \phi\left(\boldsymbol{X}_i - \boldsymbol{\mu}_l\right) \omega_t^{(l)}} \qquad \forall\, 1 \le j \le m$$

- **Output:**

$$\rho := \sum_{j=1}^{m} \omega_{t_0}^{(j)} \delta_{\boldsymbol{\mu}_j}$$

Existing algorithm: fixed-location EM

— Jiang, Zhang '09, Koenker, Mizera '14

- **Initialize:** fixed grid $\{\boldsymbol{\mu}_j\}_{1 \le j \le m} \subset \mathbb{R}^d$, $\omega_0^{(1)} = \cdots = \omega_0^{(m)} = \frac{1}{m}$.
- **Iterative update:** for $t = 0, 1, \ldots, t_0 - 1$

$$\omega_{t+1}^{(j)} = \frac{1}{N} \sum_{i=1}^{N} \frac{\phi\left(\boldsymbol{X}_i - \boldsymbol{\mu}_j\right) \omega_t^{(j)}}{\sum_{l=1}^{m} \phi\left(\boldsymbol{X}_i - \boldsymbol{\mu}_l\right) \omega_t^{(l)}} \qquad \forall\, 1 \le j \le m$$

- **Output:**

$$\rho := \sum_{j=1}^{m} \omega_{t_0}^{(j)} \delta_{\boldsymbol{\mu}_j}$$

This algorithm is basically solving NPMLE subject to ρ being supported on $\{\boldsymbol{\mu}_j\}_{1 \le j \le m}$.

Existing algorithm: fixed-location EM

— Jiang, Zhang '09, Koenker, Mizera '14

- **Initialize:** fixed grid $\{\boldsymbol{\mu}_j\}_{1\le j\le m} \subset \mathbb{R}^d$, $\omega_0^{(1)} = \cdots = \omega_0^{(m)} = \frac{1}{m}$.
- **Iterative update:** for $t = 0, 1, \ldots, t_0 - 1$

$$\omega_{t+1}^{(j)} = \frac{1}{N} \sum_{i=1}^{N} \frac{\phi\left(\boldsymbol{X}_i - \boldsymbol{\mu}_j\right) \omega_t^{(j)}}{\sum_{l=1}^{m} \phi\left(\boldsymbol{X}_i - \boldsymbol{\mu}_l\right) \omega_t^{(l)}} \qquad \forall\, 1 \le j \le m$$

- **Output:**

$$\rho := \sum_{j=1}^{m} \omega_{t_0}^{(j)} \delta_{\boldsymbol{\mu}_j}$$

Disadvantage: induces systematic approximation error.

Existing algorithm: fixed-location EM

— Jiang, Zhang '09, Koenker, Mizera '14

- **Initialize:** fixed grid $\{\boldsymbol{\mu}_j\}_{1\le j\le m} \subset \mathbb{R}^d$, $\omega_0^{(1)} = \cdots = \omega_0^{(m)} = \frac{1}{m}$.
- **Iterative update:** for $t = 0, 1, \ldots, t_0 - 1$

$$\omega_{t+1}^{(j)} = \frac{1}{N}\sum_{i=1}^{N} \frac{\phi\left(\boldsymbol{X}_i - \boldsymbol{\mu}_j\right)\omega_t^{(j)}}{\sum_{l=1}^{m}\phi\left(\boldsymbol{X}_i - \boldsymbol{\mu}_l\right)\omega_t^{(l)}} \qquad \forall\, 1 \le j \le m$$

- **Output:**

$$\rho := \sum_{j=1}^{m} \omega_{t_0}^{(j)} \delta_{\boldsymbol{\mu}_j}$$

Can we design an efficient algorithm that is capable of solving NPMLE exactly?

Part 3: gradient flow over the space of probability measures

Recap: gradient flow in Euclidean space

In $\mathbb{R}^d$, the gradient flow for a function $f : \mathbb{R}^d \to \mathbb{R}$ is defined as a curve $\boldsymbol{x}_t : [0, \infty) \to \mathbb{R}^d$ such that

$$\dot{\boldsymbol{x}}_t = -\nabla f(\boldsymbol{x}_t).$$

An equivalent definition can be given by

$$\dot{\boldsymbol{x}}_t = \lim_{\eta \to 0} \frac{\boldsymbol{x}_t^\eta - \boldsymbol{x}_t}{\eta}$$

where

$$\boldsymbol{x}_t^\eta = \arg\min_{\boldsymbol{x} \in \mathbb{R}^d} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x} - \boldsymbol{x}_t \rangle + \frac{1}{2\eta} \|\boldsymbol{x} - \boldsymbol{x}_t\|_2^2.$$

Gradient flow in the space of probability measures

- The first variation of $f : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ at some $\rho \in \mathcal{P}(\mathbb{R}^d)$ is defined to be any function $\delta f(\rho) : \mathbb{R}^d \to \mathbb{R}$ such that

$$\lim_{\varepsilon \to 0} \frac{f(\rho + \varepsilon \mathcal{X}) - f(\rho)}{\varepsilon} = \int \delta f(\rho) \mathrm{d}\mathcal{X}$$

  for any signed measure $\mathcal{X}$ over $\mathbb{R}^d$ satisfying $\int \mathrm{d}\mathcal{X} = 0$.

- Gradient flow for f under a geodesic distance $d(\cdot, \cdot)$:

$$\partial_t \rho_t = \lim_{\eta \to 0} \frac{\rho_t^\eta - \rho_t}{\eta},$$

  where

$$\rho_t^\eta \coloneqq \arg\min_{\rho \in \mathcal{P}(\mathbb{R}^d)} \int_{\mathbb{R}^d} \delta f(\rho_t) \mathrm{d}(\rho - \rho_t) + \frac{1}{2\eta} \mathrm{d}^2(\rho, \rho_t).$$

Computing first variation of ℓ_N

Recall that ℓ_N is negative log-likelihood

$$\ell_N(\rho) = -\frac{1}{N}\sum_{i=1}^{N}\log\left[\rho * \phi(\boldsymbol{X}_i)\right].$$

For any signed measure $\mathcal{X}$ satisfying $\int \mathrm{d}\mathcal{X} = 0$,

$$\lim_{\varepsilon\to 0}\frac{\ell_N(\rho+\varepsilon\mathcal{X})-\ell_N(\rho)}{\varepsilon} = -\frac{1}{N}\sum_{i=1}^{N}\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\log\left[1+\varepsilon\frac{(\mathcal{X}*\phi)(X_i)}{(\rho*\phi)(X_i)}\right]$$

$$= -\frac{1}{N}\sum_{i=1}^{N}\frac{(\mathcal{X}*\phi)(X_i)}{(\rho*\phi)(X_i)} = -\frac{1}{N}\sum_{i=1}^{N}\int\frac{\phi(x-X_i)}{(\rho*\phi)(X_i)}\mathcal{X}(\mathrm{d}x),$$

Therefore the first variation is given by

$$\delta\ell_N(\rho) : \boldsymbol{x} \to -\frac{1}{N}\sum_{i=1}^{N}\frac{\phi(\boldsymbol{x}-\boldsymbol{X}_i)}{(\rho*\phi)(\boldsymbol{X}_i)}$$

Fisher-Rao geometry

— *Bauer et al., 2016*

- The tangent space at $\rho \in \mathcal{P}(\mathbb{R}^d)$ is

$$\mathrm{Tan}_\rho^{\mathsf{FR}} \mathcal{P}(\mathbb{R}^d) \coloneqq \left\{ \zeta : \zeta = \rho \left(\alpha - \int \alpha \mathrm{d}\rho \right) \text{ for some } \alpha \right.$$

$$\left. \text{satisfying } \int \alpha^2 \mathrm{d}\rho < \infty \right\}.$$

- We equip the above tangent space with Riemannian metric tensor

$$\begin{aligned} g_\rho^{\mathsf{FR}}(\zeta_1, \zeta_2) &\coloneqq \int \frac{\zeta_1 \cdot \zeta_2}{\rho^2} \mathrm{d}\rho \\ &= \int_{\mathbb{R}^d} \alpha_1(x) \alpha_2(x) \rho(\mathrm{d}x) - \left(\int_{\mathbb{R}^d} \alpha_1 \mathrm{d}\rho \right) \left(\int_{\mathbb{R}^d} \alpha_2 \mathrm{d}\rho \right) \end{aligned}$$

for any $\zeta_1 = \rho(\alpha_1 - \int \alpha_1 \mathrm{d}\rho)$ and $\zeta_2 = \rho(\alpha_2 - \int \alpha_2 \mathrm{d}\rho)$.

Fisher-Rao gradient flow/descent

- Fisher-Rao distance between probability measures:

$$\mathrm{d}^2_{\mathsf{FR}}\left(\rho_0, \rho_1\right) = \inf \left\{ \int_0^1 \int \left[\left(\alpha_t - \int \alpha_t \mathrm{d}\rho_t \right)^2 \right] \mathrm{d}\rho_t \mathrm{d}t : \left(\rho_t, \alpha_t\right)_{t\in[0,1]} \right.$$
$$\left. \text{solves } \partial_t \rho_t = \rho_t \left(\alpha_t - \int \alpha_t \mathrm{d}\rho_t \right) \right\}$$

Fisher-Rao gradient flow/descent

- Fisher-Rao distance between probability measures:

$$\mathrm{d}^2_{\mathsf{FR}}(\rho_0, \rho_1) = \inf \left\{ \int_0^1 \int \left[\left(\alpha_t - \int \alpha_t \mathrm{d}\rho_t \right)^2 \right] \mathrm{d}\rho_t \mathrm{d}t : (\rho_t, \alpha_t)_{t \in [0,1]} \right.$$
$$\left. \text{solves } \partial_t \rho_t = \rho_t \left(\alpha_t - \int \alpha_t \mathrm{d}\rho_t \right) \right\}$$

  - when ρ_0 and ρ_1 are continuous, this is the squared Hellinger distance

Fisher-Rao gradient flow/descent

- Fisher-Rao distance between probability measures:

$$\mathrm{d}^2_{\mathsf{FR}}\left(\rho_0,\rho_1\right)=\inf\left\{\int_0^1\int\left[\left(\alpha_t-\int\alpha_t\mathrm{d}\rho_t\right)^2\right]\mathrm{d}\rho_t\mathrm{d}t:\left(\rho_t,\alpha_t\right)_{t\in[0,1]}\right.$$
$$\left.\text{solves }\ \partial_t\rho_t=\rho_t\left(\alpha_t-\int\alpha_t\mathrm{d}\rho_t\right)\right\}$$

  - when ρ_0 and ρ_1 are continuous, this is the squared Hellinger distance

- Fisher-Rao gradient flow of $\ell_N(\rho)$:

$$\partial_t\rho_t=-\left[1+\delta\ell_N\left(\rho_t\right)\right]\rho_t$$

Fisher-Rao gradient flow/descent

- Fisher-Rao distance between probability measures:

$$\mathrm{d}^2_{\mathsf{FR}}\left(\rho_0, \rho_1\right) = \inf \left\{ \int_0^1 \int \left[\left(\alpha_t - \int \alpha_t \mathrm{d}\rho_t \right)^2 \right] \mathrm{d}\rho_t \mathrm{d}t : \left(\rho_t, \alpha_t\right)_{t\in[0,1]} \right.$$
$$\left. \text{solves } \partial_t \rho_t = \rho_t \left(\alpha_t - \int \alpha_t \mathrm{d}\rho_t \right) \right\}$$

  ◦ when ρ_0 and ρ_1 are continuous, this is the squared Hellinger distance

- Fisher-Rao gradient flow of $\ell_N(\rho)$:

$$\partial_t \rho_t = -\left[1 + \delta \ell_N\left(\rho_t\right)\right] \rho_t$$

- Fisher-Rao gradient descent of $\ell_N(\rho)$:

$$\frac{\mathrm{d}\rho_{t+1}}{\mathrm{d}\rho_t} = 1 - \gamma\left[1 + \delta \ell_N\left(\rho_t\right)\right]$$

Fixed-location EM as Fisher-Rao GD

- Fixed-location EM:

$$\rho_t = \sum_{j=1}^{m} \omega_t^{(j)} \delta_{\boldsymbol{\mu}_j} \quad \text{where} \quad \omega_{t+1}^{(j)} = \frac{1}{N} \sum_{i=1}^{N} \frac{\phi\left(\boldsymbol{X}_i - \boldsymbol{\mu}_j\right) \omega_t^{(j)}}{\sum_{l=1}^{m} \phi\left(\boldsymbol{X}_i - \boldsymbol{\mu}_l\right) \omega_t^{(l)}}$$

- Fisher-Rao gradient descent:

$$\frac{\mathrm{d}\rho_{t+1}}{\mathrm{d}\rho_t} = 1 - \gamma\left[1 + \delta\ell_N\left(\rho_t\right)\right]$$

Fixed-location EM as Fisher-Rao GD

- Fixed-location EM:

$$\rho_t = \sum_{j=1}^{m} \omega_t^{(j)} \delta_{\boldsymbol{\mu}_j} \quad \text{where} \quad \omega_{t+1}^{(j)} = \frac{1}{N} \sum_{i=1}^{N} \frac{\phi\left(\boldsymbol{X}_i - \boldsymbol{\mu}_j\right) \omega_t^{(j)}}{\sum_{l=1}^{m} \phi\left(\boldsymbol{X}_i - \boldsymbol{\mu}_l\right) \omega_t^{(l)}}$$

- Fisher-Rao gradient descent:

$$\frac{\mathrm{d}\rho_{t+1}}{\mathrm{d}\rho_t} = 1 - \gamma \left[1 + \delta \ell_N \left(\rho_t\right)\right]$$

Theorem 1

Fixed-location EM algorithm is Fisher-Rao gradient descent with step size $\gamma = 1$.

A geometric perspective of fixed-location EM

Fixed-location EM can be viewed as

- an interacting particle system in $\mathbb{R}^d$
- each particle has two attributes: location, weight
- the locations are fixed
- the weights evolve according to the gradient descent in the space of probability measures endowed with Fisher-Rao geometry

A geometric perspective of fixed-location EM

Fixed-location EM can be viewed as

- an interacting particle system in $\mathbb{R}^d$
- each particle has two attributes: location, weight
- the locations are fixed
- the weights evolve according to the gradient descent in the space of probability measures endowed with Fisher-Rao geometry

Issue: fixed-location EM has approximation error due to fixed grid

A geometric perspective of fixed-location EM

Fixed-location EM can be viewed as

- an interacting particle system in $\mathbb{R}^d$
- each particle has two attributes: location, weight
- the locations are fixed
- the weights evolve according to the gradient descent in the space of probability measures endowed with Fisher-Rao geometry

Reason: the location of particles doesn't change

Discussion: Fisher-Rao GD

$$\frac{\mathrm{d}\rho_{t+1}}{\mathrm{d}\rho_t} = 1 - \gamma \left[1 + \delta\ell_N\left(\rho_t\right)\right]$$

- Pros: we can establish convergence theory in the mean field limit (infinite number of particles), i.e. when ρ_0 is continuous
  - the optimality condition is $\delta\ell_N(\rho)(\boldsymbol{x}) \geq -1$ for all $\boldsymbol{x} \in \mathbb{R}^d$
- Cons: computationally inefficient
  - impossible to implement continuous dynamic
  - Fisher-Rao geometry is not able to move particles
  - incur approximation error that is exponential in d

Discussion: Fisher-Rao GD

$$\frac{\mathrm{d}\rho_{t+1}}{\mathrm{d}\rho_t} = 1 - \gamma\left[1 + \delta\ell_N\left(\rho_t\right)\right]$$

- Pros: we can establish convergence theory in the mean field limit (infinite number of particles), i.e. when ρ_0 is continuous
  - the optimality condition is $\delta\ell_N(\rho)(\boldsymbol{x}) \geq -1$ for all $\boldsymbol{x} \in \mathbb{R}^d$
- Cons: computationally inefficient
  - impossible to implement continuous dynamic
  - Fisher-Rao geometry is not able to move particles
  - incur approximation error that is exponential in d

Is it possible to find another geometry that leads to better algorithm?

Wasserstein geometry

— *Otto, 2001; Ambrosio et al., 2008*

- The tangent space at $\rho \in \mathcal{P}(\mathbb{R}^d)$ is

$$\mathrm{Tan}_\rho^{\mathsf{W}}\mathcal{P}(\mathbb{R}^d) := \Big\{ \zeta : \zeta = -\mathsf{div}\,(\rho \nabla u) \ \text{ for some } u$$

$$\text{satisfying } \int \|\nabla u\|_2^2 \mathrm{d}\rho < \infty \Big\}.$$

- We equip the above tangent space with Riemannian metric tensor

$$g_\rho^{\mathsf{W}}\left(\zeta_1, \zeta_2\right) := \int_{\mathbb{R}^d} \left\langle \nabla u_1, \nabla u_2 \right\rangle \rho\left(\mathrm{d}x\right)$$

for any $\zeta_1 = -\mathsf{div}(\rho \nabla u_1)$ and $\zeta_2 = -\mathsf{div}(\rho \nabla u_2)$.

Wasserstein gradient flow/descent

- The (quadratic) Wasserstein distance:

$$d_{\mathsf{W}}^2(\rho_0, \rho_1) = \inf \left\{ \int_0^1 \int \|v_t\|^2 \, \mathrm{d}\rho_t \mathrm{d}t : (\rho_t, v_t)_{t \in [0,1]} \text{ solves } \partial_t \rho_t = -\mathsf{div}(\rho_t v_t) \right\}$$

Wasserstein gradient flow/descent

- The (quadratic) Wasserstein distance:

$$d_{\mathsf{W}}^2(\rho_0,\rho_1) = \inf\left\{ \int_0^1 \int \|v_t\|^2 \,\mathrm{d}\rho_t \mathrm{d}t : (\rho_t, v_t)_{t\in[0,1]} \text{ solves } \partial_t \rho_t = -\mathsf{div}\left(\rho_t v_t\right) \right\}$$

- Wasserstein gradient flow of $\ell_N(\rho)$:

$$\partial_t \rho_t = \mathsf{div}\left(\nabla \delta \ell_N\left(\rho_t\right)\rho_t\right).$$

Wasserstein gradient flow/descent

- The (quadratic) Wasserstein distance:

$$d_{\mathsf{W}}^2(\rho_0, \rho_1) = \inf \left\{ \int_0^1 \int \|v_t\|^2 \, \mathrm{d}\rho_t \mathrm{d}t : (\rho_t, v_t)_{t\in[0,1]} \text{ solves } \partial_t \rho_t = -\mathsf{div}(\rho_t v_t) \right\}$$

- Wasserstein gradient flow of $\ell_N(\rho)$:

$$\partial_t \rho_t = \mathsf{div}\left(\nabla \delta \ell_N(\rho_t)\, \rho_t\right).$$

- Wasserstein gradient descent of $\ell_N(\rho)$:

$$\rho_{t+1} = \left[\mathrm{Id} - \eta \nabla \delta \ell_N(\rho_t)\right]_{\#} \rho_t$$

  - Here $T_{\#}\rho(A) = \rho(T^{-1}(A))$ for any Borel set A.

Euclidean GD is Wasserstein GD

Euclidean gradient flow for

$$\underset{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_m \in \mathbb{R}^d}{\text{minimize}} \; \ell(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_m) := -\frac{1}{N}\sum_{i=1}^{N} \log\left[\frac{1}{m}\sum_{j=1}^{m} \phi(\boldsymbol{X}_i - \boldsymbol{\mu}_j)\right]$$

is given by

$$\dot{\boldsymbol{\mu}}_t^{(j)} = \frac{1}{N}\sum_{i=1}^{N} \frac{\phi\left(\boldsymbol{X}_i - \boldsymbol{\mu}_t^{(j)}\right)}{\sum_{l=1}^{m} \omega_t^{(j)} \phi\left(\boldsymbol{X}_i - \boldsymbol{\mu}_t^{(l)}\right)} \left(\boldsymbol{X}_i - \boldsymbol{\mu}_t^{(j)}\right)$$

Euclidean GD is Wasserstein GD

Euclidean gradient flow for

$$\underset{\boldsymbol{\mu}_1,\dots,\boldsymbol{\mu}_m\in\mathbb{R}^d}{\text{minimize}} \quad \ell(\boldsymbol{\mu}_1,\dots,\boldsymbol{\mu}_m) \coloneqq -\frac{1}{N}\sum_{i=1}^{N}\log\left[\frac{1}{m}\sum_{j=1}^{m}\phi(\boldsymbol{X}_i-\boldsymbol{\mu}_j)\right]$$

is given by

$$\dot{\boldsymbol{\mu}}_t^{(j)} = \frac{1}{N}\sum_{i=1}^{N}\frac{\phi\left(\boldsymbol{X}_i-\boldsymbol{\mu}_t^{(j)}\right)}{\sum_{l=1}^{m}\omega_t^{(j)}\phi\left(\boldsymbol{X}_i-\boldsymbol{\mu}_t^{(l)}\right)}\left(\boldsymbol{X}_i-\boldsymbol{\mu}_t^{(j)}\right)$$

Theorem 2

The flow $(\rho_t)_{t\geq 0}$ given by $\rho_t = \frac{1}{m}\sum_{j=1}^{m}\delta_{\boldsymbol{\mu}_t^{(j)}}$ is a Wasserstein gradient flow. This connection is also true for gradient descent.

A geometric perspective of gradient descent

(Overparameterized) gradient descent can be viewed as

- an interacting particle system in $\mathbb{R}^d$
- each particle has two attributes: location, weight
- the locations evolve according to the gradient descent in the space of probability measures endowed with Wasserstein geometry
- the weights are fixed

A geometric perspective of gradient descent

(Overparameterized) gradient descent can be viewed as

- an interacting particle system in $\mathbb{R}^d$
- each particle has two attributes: location, weight
- the locations evolve according to the gradient descent in the space of probability measures endowed with Wasserstein geometry
- the weights are fixed

Question: is Wasserstein geometry the right one to use?

Discussion: Wasserstein GD

$$\rho_{t+1} = \left[\mathrm{Id} - \eta \nabla \delta \ell_N \left(\rho_t\right)\right]_{\#} \rho_t$$

- Pros: good empirical performance when m is large and locations are randomly initialized from data
- Cons: (most theoretically)
  - difficult to establish convergence results even in the mean-field limit (due to geodesic non-convexity)
  - not able to teleport mass like Fisher-Rao GD (requires exponential time to converge for imperfect initialization)
  - approximation error of order $O(1/m)$

Discussion: Wasserstein GD

$$\rho_{t+1} = \left[\mathrm{Id} - \eta \nabla \delta \ell_N \left(\rho_t\right)\right]_{\#} \rho_t$$

- Pros: good empirical performance when m is large and locations are randomly initialized from data
- Cons: (most theoretically)
  - difficult to establish convergence results even in the mean-field limit (due to geodesic non-convexity)
  - not able to teleport mass like Fisher-Rao GD (requires exponential time to converge for imperfect initialization)
  - approximation error of order $O(1/m)$

Is there an even better solution with provable theoretical guarantees?

Geodesic convexity

- The correct notion of "convexity" in general metric space is geodesic convexity.
- Suppose $\rho_0, \rho_1 \in \mathcal{P}(\mathbb{R}^d)$, and let $(\rho_t)_{0 \le t \le 1}$ be the geodesic, i.e., "shortest, constant speed" curve that connects ρ_0 and ρ_1.
- $f(\rho)$ is geodesically convex if $g(t) = f(\rho_t)$ is convex on $[0, 1]$.
- Example: when equipped with Wasserstein-2 distance, the geodesic between ρ_0 and ρ_1 is given by

$$\rho_t = \left[(1-t)\mathrm{id} + tT\right]_{\#} \rho_0$$

  where T is the optimal transport map between ρ_0 and ρ_1.
- $\ell_N(\rho)$ is not geodesically convex...

Wasserstein-Fisher-Rao geometry

— *Chizat et al., 2018; Gallouet et al., 2017; Kondratyev et al., 2016; Liero et al., 2018*

- Key idea: incorporate both mass transportation and teleportation
- WFR geometry $\leftarrow$ coupling Wasserstein and Fisher-Rao geometry
- The tangent space at $\rho \in \mathcal{P}(\mathbb{R}^d)$ is

$$\mathrm{Tan}_{\rho}^{\mathsf{WFR}}\mathcal{P}(\mathbb{R}^d) := \left\{\zeta : \zeta = -\mathsf{div}\left(\rho \nabla u\right) + \rho\left(\alpha - \int \alpha \mathrm{d}\rho\right) \text{ for some } u, \alpha \text{ satisfying } \int (\alpha^2 + \|\nabla u\|_2^2)\mathrm{d}\rho < \infty\right\}.$$

- We equip the above tangent space with Riemannian metric tensor

$$g_{\rho}^{\mathsf{WFR}}\left(\zeta_1, \zeta_2\right) := \int_{\mathbb{R}^d} \left\langle \nabla u_1, \nabla u_2 \right\rangle \rho\left(\mathrm{d}x\right) + \int_{\mathbb{R}^d} \alpha_1\left(x\right)\alpha_2\left(x\right)\rho\left(\mathrm{d}x\right) - \left(\int_{\mathbb{R}^d} \alpha_1 \mathrm{d}\rho\right)\left(\int_{\mathbb{R}^d} \alpha_2 \mathrm{d}\rho\right)$$

for any $\zeta_i = -\mathsf{div}(\rho\nabla u_i) + \rho(\alpha_i - \int \alpha_i \mathrm{d}\rho)$ where $i = 1, 2$.

WFR gradient flow/descent

- WFR distance between probability measures:

$$d^2_{\mathsf{WFR}}(\rho_0,\rho_1) = \inf\left\{ \int_0^1 \int \left[\|v_t\|^2 + \left(\alpha_t - \int \alpha_t \mathrm{d}\rho_t\right)^2 \right] \mathrm{d}\rho_t \mathrm{d}t : \right.$$

$$\left. (\rho_t, v_t, \alpha_t)_{t\in[0,1]} \text{ solves } \partial_t \rho_t = -\mathsf{div}(\rho_t v_t) + \rho_t \left(\alpha_t - \int \alpha_t \mathrm{d}\rho_t\right) \right\}$$

WFR gradient flow/descent

- WFR distance between probability measures:

$$d^2_{\mathsf{WFR}}\left(\rho_0, \rho_1\right) = \inf\left\{ \int_0^1 \int \Big[\left\|v_t\right\|^2 + \Big(\alpha_t - \int \alpha_t \mathrm{d}\rho_t\Big)^2 \Big] \mathrm{d}\rho_t \mathrm{d}t : \right.$$

$$\left. \left(\rho_t, v_t, \alpha_t\right)_{t\in[0,1]} \text{ solves } \partial_t \rho_t = -\mathsf{div}\left(\rho_t v_t\right) + \rho_t \Big(\alpha_t - \int \alpha_t \mathrm{d}\rho_t\Big) \right\}$$

- Wasserstein-Fisher-Rao gradient flow of $\ell_N(\rho)$:

$$\partial_t \rho_t = \mathsf{div}\left(\rho_t \nabla \delta \ell_N\left(\rho_t\right)\right) - \left[1 + \delta \ell_N\left(\rho_t\right)\right] \rho_t$$

WFR gradient flow/descent

- WFR distance between probability measures:

$$d^2_{\mathsf{WFR}}(\rho_0, \rho_1) = \inf \Bigg\{ \int_0^1 \int \Big[\|v_t\|^2 + \Big(\alpha_t - \int \alpha_t \mathrm{d}\rho_t\Big)^2 \Big] \mathrm{d}\rho_t \mathrm{d}t :$$

$$(\rho_t, v_t, \alpha_t)_{t\in[0,1]} \text{ solves } \partial_t \rho_t = -\mathsf{div}(\rho_t v_t) + \rho_t \Big(\alpha_t - \int \alpha_t \mathrm{d}\rho_t\Big) \Bigg\}$$

- Wasserstein-Fisher-Rao gradient flow of $\ell_N(\rho)$:

$$\partial_t \rho_t = \mathsf{div}(\rho_t \nabla \delta \ell_N(\rho_t)) - [1 + \delta \ell_N(\rho_t)]\, \rho_t$$

- Wasserstein-Fisher-Rao gradient descent of $\ell_N(\rho)$:

$$\frac{\mathrm{d}\rho_{t+0.5}}{\mathrm{d}\rho_t} = 1 - \eta\,[1 + \delta \ell_N(\rho_t)] \qquad \text{(Fisher-Rao GD)}$$

$$\rho_{t+1} = [\mathrm{Id} - \eta \nabla \delta \ell_N(\rho_{t+0.5})]_{\#}\, \rho_{t+0.5} \qquad \text{(Wasserstein GD)}$$

WFR gradient descent

- **Initialize:** number of particles m, $\omega_0^{(1)} = \cdots = \omega_0^{(m)} = \frac{1}{m}$,
  $\mu_0^{(1)}, \ldots, \mu_0^{(m)} \overset{\text{i.i.d.}}{\sim} \text{Uniform}(\{X_i\}_{1 \le i \le m})$.
- **Iterative update:** for $t = 0, 1, \ldots, t_0 - 1$

$$\mu_{t+1}^{(j)} = \mu_t^{(j)} + \eta \frac{1}{N} \sum_{i=1}^{N} \frac{\phi\left(X_i - \mu_t^{(j)}\right)}{\sum_{l=1}^{m} \omega_t^{(j)} \phi\left(X_i - \mu_t^{(l)}\right)} \left(X_i - \mu_t^{(j)}\right),$$

$$\omega_{t+1}^{(j)} = \omega_t^{(j)} + \eta \left[\frac{1}{N} \sum_{i=1}^{N} \frac{\phi\left(X_i - \mu_t^{(j)}\right)}{\sum_{l=1}^{m} \omega_t^{(j)} \phi\left(X_i - \mu_t^{(l)}\right)} - 1\right] \omega_t^{(j)},$$

- **Output:**

$$\rho := \sum_{j=1}^{m} \omega_{t_0}^{(j)} \delta_{\mu_{t_0}^{(j)}}$$

A geometric perspective of WFR gradient descent

Wasserstein-Fisher-Rao gradient descent can be viewed as

- an interacting particle system in $\mathbb{R}^d$
- each particle has two attributes: location, weight
- the locations and weights evolve according to the gradient descent in the space of probability measures endowed with Wasserstein-Fisher-Rao geometry

A geometric perspective of WFR gradient descent

Wasserstein-Fisher-Rao gradient descent can be viewed as

- an interacting particle system in $\mathbb{R}^d$
- each particle has two attributes: location, weight
- the locations and weights evolve according to the gradient descent in the space of probability measures endowed with Wasserstein-Fisher-Rao geometry

WFR GD change the location and weight of particles simultaneously: no systematic approximation error!

Convergence theory

Theorem 3

Suppose that $\mathrm{supp}(\rho_0) = \mathbb{R}^d$. *There exists* $\eta_0 > 0$ *determined by the samples* $\{X_i\}_{1 \le i \le N}$, *such that if* $0 < \eta \le \eta_0$, *then*

1. $\ell_N(\rho_t)$ *is decreasing*
2. *if* $\rho_t \xrightarrow{\mathrm{w}} \widehat{\rho}$ *when* $t \to \infty$, *then* $\widehat{\rho}$ *is the NPMLE.*

Convergence theory

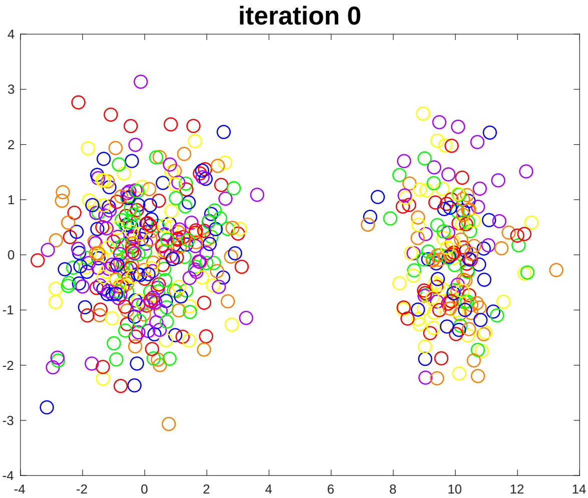
Theorem 3

Suppose that supp$(\rho_0) = \mathbb{R}^d$. *There exists* $\eta_0 > 0$ *determined by the samples* $\{X_i\}_{1 \le i \le N}$, *such that if* $0 < \eta \le \eta_0$, *then*

1. $\ell_N(\rho_t)$ *is decreasing*
2. *if* $\rho_t \xrightarrow{\mathrm{w}} \widehat{\rho}$ *when* $t \to \infty$, *then* $\widehat{\rho}$ *is the NPMLE.*

- also holds for WFR gradient flow
- conditional convergence (similar to Chizat and Bach, 2018)
- only works in the mean-field regime (infinite particle limit)
- suggests overparameterization (using a large m)

Implementation



$$N = 1500,\ d = 2,\ m = 500,\ \rho^\star = \frac{1}{3}\delta_{(-1,0)} + \frac{1}{3}\delta_{(1,0)} + \frac{1}{3}\delta_{(10,0)}$$

Conclusion

- We identify prior algorithm for solving NPMLE for Gaussian mixture model as Fisher-Rao gradient descent
- We design an efficient algorithm (WFR gradient descent) that is capable of computing NPMLE exactly.
- Our work also demonstrates the role of overparameterization in learning Gaussian mixtures.

Future directions

- Is it possible to prove "unconditional" convergence theory?
- How to establish convergence guarantees for WFR gradient descent beyond mean-field regime (finite particle)? If so, how large should m be?
- Does all this holds for Wasserstein gradient descent (gradient descent for overparameterized MLE)? Is weight update really necessary?

Thank you!