Supplementary Materials for "Learning Gaussian Mixtures Using the Wasserstein-Fisher-Rao Gradient Flow"

April 21, 2023

A Preliminaries

In the main text, we focused on the Gaussian mixture model where $\phi(x) = (2\pi)^{-d/2}\exp(-\|x\|_2^2/2)$. In fact, the algorithms and theorems in the current paper are also valid for a more general class of probability density ϕ. In the appendices, we only assume that ϕ satisfies the following regularity assumption.

Assumption 1 (Regularity)**.** *Assume that the density $\phi(x) > 0$ for any $x \in \mathbb{R}^d$. Furthermore, $\phi \in C^{\max\{d,2\}}(\mathbb{R}^d)$, $\lim_{\|x\|_2\to\infty}\phi(x) = 0$, $\sup_{x\in\mathbb{R}^d}\|\nabla\phi(x)\|_2 < \infty$ and $\sup_{x\in\mathbb{R}^d}\|\nabla^2\phi(x)\|_2 < \infty$.*

It is clear that the Gaussian kernel $\phi(x) = (2\pi)^{-d/2}\exp(-\|x\|_2^2/2)$ satisfies Assumption 1. The following lemma shows that NPMLE is compactly supported. The proof can be found in Appendix B.2.

Lemma 1 (Compact support of NPMLE)**.** *Let Assumption 1 hold and $\widehat{\rho}$ be any optimal solution to (1.1). Define $R_1 = \inf\{r \geq 0 :\ \bar{\phi}(r) \leq \underline{\phi}[\mathrm{diam}(\Omega)]/2\}$ and*

$$R = \inf\left\{r \geq 0 :\ \bar{\phi}(r) \leq \frac{\bar{\phi}(R_1)\underline{\phi}(R_1 + \mathrm{diam}(\Omega))}{8\bar{\phi}(0)}\right\}.$$

Then, we have $\mathsf{supp}(\widehat{\rho}) \subseteq \Omega_R$.

B Proof of structual results for NPMLE

In this section, we prove the two structural results for NPMLE, namely Theorem 1 and Lemma 1 under Assumption 1.

B.1 Proof of Theorem 1

Part 1: existence of NPMLE. Note that the loss function ℓ_N is lower bounded

$$\ell_N(\rho) = -\frac{1}{N}\sum_{i=1}^N \log\left[(\rho * \phi)(X_i)\right] \geq -\log\|\phi\|_\infty, \tag{B.1}$$

where the last inequality follows from $\rho * \phi(x) = \int_{\mathbb{R}^d}\phi(y-x)\,\rho(\mathrm{d}y) \leq \|\phi\|_\infty$ for any $x \in \mathbb{R}^d$. Therefore there exists a sequence of probability distribution $\{\rho_n\}$ such that

$$\ell_N(\rho_n) \leq \inf_{\rho\in\mathcal{P}(\mathbb{R}^d)}\ell_N(\rho) + \frac{1}{n}. \tag{B.2}$$

Now we argue that $\{\rho_n\}$ is tight. To that end, we show that there exists $r > 0$ such that for any $\varepsilon > 0$, it holds $\rho_n(\Omega_r) \geq 1 - \varepsilon$ for n large enough.

For any n and $r > 0$, define

$$\rho_{n,r} \coloneqq \rho_n(\Omega_r)\cdot\rho_n|_{\Omega_r} + \rho_n(\Omega_r^{\mathrm{c}})\cdot\mathsf{Unif}(\Omega),$$

where $\rho_n|_{\Omega_r}(\cdot) = \rho_n(\cdot|\Omega_r)$ is the conditional distribution of ρ_n given Ω_r, and $\mathsf{Unif}(\Omega)$ is the uniform distribution on Ω. We have

$$\ell_N(\rho_n) - \ell_N(\rho_{n,r}) = -\frac{1}{N}\sum_{i=1}^N \log\left[(\rho_n * \phi)(X_i)\right] + \frac{1}{N}\sum_{i=1}^N \log\left[(\rho_{n,r} * \phi)(X_i)\right] = \frac{1}{N}\sum_{i=1}^N \log\left[\frac{(\rho_{n,r} * \phi)(X_i)}{(\rho_n * \phi)(X_i)}\right].$$

Note that for each $i \in [N]$

$$\begin{aligned}
\log\left[\frac{(\rho_{n,r} * \phi)(X_i)}{(\rho_n * \phi)(X_i)}\right] &= \log\left[\frac{\int_{\Omega_r}\phi(X_i - y)\,\rho_n(\mathrm{d}y) + \rho_n(\Omega_r^{\mathrm{c}})\int_\Omega \phi(X_i - y)\,\mathrm{d}y}{\int_{\Omega_r}\phi(X_i - y)\,\rho_n(\mathrm{d}y) + \int_{\Omega_r^{\mathrm{c}}}\phi(X_i - y)\,\rho_n(\mathrm{d}y)}\right] \\
&= \log\left[1 + \frac{\rho_n(\Omega_r^{\mathrm{c}})\int_\Omega \phi(X_i - y)\,\mathrm{d}y - \int_{\Omega_r^{\mathrm{c}}}\phi(X_i - y)\,\rho_n(\mathrm{d}y)}{(\rho_n * \phi)(X_i)}\right] \\
&\geq \log\left[1 + \frac{\rho_n(\Omega_r^{\mathrm{c}})\left[\underline{\phi}(\mathrm{diam}(\Omega)) - \overline{\phi}(r)\right]}{\|\phi\|_\infty}\right].
\end{aligned}$$

We can choose $r > 0$ to be sufficiently large so that $\overline{\phi}(r) \leq \underline{\phi}(\mathrm{diam}(\Omega))/2$, and therefore for each $i \in [N]$

$$\ell_N(\rho_n) - \ell_N(\rho_{n,r}) \geq \log\left[1 + \frac{\rho_n(\Omega_r^{\mathrm{c}})\,\underline{\phi}(\mathrm{diam}(\Omega))}{2\|\phi\|_\infty}\right].$$

In view of (B.2), we know that

$$\ell_N(\rho_n) - \ell_N(\rho_{n,r}) \leq \frac{1}{n}.$$

Taking the above two inequalities collectively give

$$\rho_n(\Omega_r^{\mathrm{c}}) \leq \frac{2\|\phi\|_\infty\left[\exp\left(\frac{1}{n}\right) - 1\right]}{\underline{\phi}(\mathrm{diam}(\Omega))}.$$

Therefore we have

$$\rho_n(\Omega_r^{\mathrm{c}}) \leq \varepsilon \qquad \text{as long as} \qquad n \geq n_\varepsilon := \left\lceil 1/\log\left[1 + \frac{\varepsilon\underline{\phi}(\mathrm{diam}(\Omega))}{2\|\phi\|_\infty}\right]\right\rceil,$$

which implies that $\{\rho_n\}$ is tight. We conclude using Prokhorov's theorem: there exists a subsequence $\{\rho_{n_k}\}$ and $\widehat{\rho} \in \mathcal{P}(\mathbb{R}^d)$ such that ρ_{n_k} converges weakly to $\widehat{\rho}$ which must be a minimizer of (1.1).

Part 2: optimality condition. First of all, it is straightforward to check that for any $\rho \in \mathcal{M}(\mathbb{R}^d)$

$$\int_{\mathbb{R}^d} \delta\ell_N(\rho)(x)\,\rho(\mathrm{d}x) = -\frac{1}{N}\sum_{i=1}^N \frac{\int \phi(x - X_i)\,\rho(\mathrm{d}x)}{(\rho * \phi)(X_i)} = -\frac{1}{N}\sum_{i=1}^N \frac{(\rho * \phi)(X_i)}{(\rho * \phi)(X_i)} = -1.$$

If $\widehat{\rho} \in \mathcal{M}(\mathbb{R}^d)$ is the optimal solution to (1.1), then for any $x \in \mathbb{R}^d$ and any $\varepsilon \in [0, 1]$ we have

$$\ell_N(\widehat{\rho}) \leq \ell_N\left((1 - \varepsilon)\,\widehat{\rho} + \varepsilon\delta_x\right) = \ell_N\left(\rho + \varepsilon\left(\delta_x - \widehat{\rho}\right)\right).$$

As a result we have

$$\delta\ell_N(\widehat{\rho})(x) + 1 = \int_{\mathbb{R}^d} \delta\ell_N(\widehat{\rho})\,\mathrm{d}\left(\delta_x - \widehat{\rho}\right) = \lim_{\varepsilon \to 0}\frac{1}{\varepsilon}\left[\ell_N\left(\widehat{\rho} + \varepsilon\left(\delta_x - \widehat{\rho}\right)\right) - \ell_N(\widehat{\rho})\right] \geq 0.$$

Since x is arbitrary, this implies that $\delta\ell_N(\widehat{\rho})(x) \geq -1$ for any $x \in \mathbb{R}^d$. Combine this with $\int_{\mathbb{R}^d}\delta\ell_N(\widehat{\rho})\mathrm{d}\widehat{\rho} = -1$ readily gives $\delta\ell_N(\widehat{\rho})(x) = -1$ for $\widehat{\rho}$-a.e. x.

Conversely, if $\widehat{\rho} \in \mathcal{M}(\mathbb{R}^d)$ satisfies $\delta\ell_N(\widehat{\rho})(x) \geq -1$ for all $x \in \mathbb{R}^d$, then for any $\rho \in \mathcal{M}(\mathbb{R}^d)$, it holds

$$0 \leq \int_{\mathbb{R}^d} \delta\ell_N(\widehat{\rho})\,\mathrm{d}\rho + 1 = \int_{\mathbb{R}^d} \delta\ell_N(\widehat{\rho})\,\mathrm{d}\left(\rho - \widehat{\rho}\right) = \lim_{\varepsilon \to 0}\frac{1}{\varepsilon}\left[\ell_N\left(\widehat{\rho} + \varepsilon\left(\rho - \widehat{\rho}\right)\right) - \ell_N(\widehat{\rho})\right] \leq \ell_N(\rho) - \ell_N(\widehat{\rho})$$

where the last inequality follows from convexity of the functional $\rho \mapsto \ell_N(\rho)$. The above display yields that that $\widehat{\rho}$ is a global minimizer of (1.1).

B.2 Proof of Lemma 1

By Theorem 1, $\delta\ell_N(\widehat{\rho}) = -1$ over $\mathsf{supp}(\widehat{\rho})$. We will show that $|\delta\ell_N(\widehat{\rho})(y)| < 1/2$ when y is too far away from Ω, and then conclude that $\mathsf{supp}(\widehat{\rho})$ must stay close to Ω. To that end, we present some useful estimates in the following lemma that will also be useful later.

Lemma 2. *Let Assumption 1 hold. For any $\rho \in \mathcal{P}(\mathbb{R}^d)$, $x \in \Omega$ and $r \geq 0$, we have*

$$\rho(\Omega_r)\underline{\phi}(r + \mathrm{diam}(\Omega)) \leq (\rho * \phi)(x) \leq \rho(\Omega_r^{\mathrm{c}})\bar{\phi}(r) + \rho(\Omega_r)\bar{\phi}(0) \leq \bar{\phi}(r) + \rho(\Omega_r)\bar{\phi}(0).$$

As a result,

$$-\log\Big(\bar{\phi}(r) + \rho(\Omega_r)\bar{\phi}(0)\Big) \leq \ell_N(\rho) \leq -\log\Big(\rho(\Omega_r)\underline{\phi}(r + \mathrm{diam}(\Omega))\Big).$$

Take any $R \geq 0$ such that $\bar{\phi}(R) \leq e^{-\ell_N(\rho)}/2$. For any $\mu \in \mathcal{P}(\mathbb{R}^d)$ with $\ell_N(\mu) \leq \ell_N(\rho)$, we have

$$\begin{aligned}
&\mu(\Omega_R) \geq e^{-\ell_N(\rho)}/[2\bar{\phi}(0)],\\
&\inf_{x\in\Omega}(\mu * \phi)(x) \geq e^{-\ell_N(\rho)}\underline{\phi}(R + \mathrm{diam}(\Omega))/[2\bar{\phi}(0)],\\
&\sup_{y\in\Omega_r^{\mathrm{c}}}|\delta\ell_N(\mu)(y)| \leq \frac{2e^{\ell_N(\rho)}\bar{\phi}(0)}{\underline{\phi}(R + \mathrm{diam}(\Omega))} \cdot \bar{\phi}(r), \qquad \forall r \geq 0.
\end{aligned}$$

The proof of Lemma 2 is deferred to the end of this section. We come back to proving Lemma 1. Choose any $\rho_0 \in \mathcal{P}(\mathbb{R}^d)$ supported on Ω. By Lemma 2, we have

$$\ell_N(\rho_0) \leq -\log\Big(\underline{\phi}[\mathrm{diam}(\Omega)]\Big), \qquad \forall r \geq 0.$$

Take $R_1 = \inf\{r \geq 0 :\ \bar{\phi}(r) \leq \underline{\phi}[\mathrm{diam}(\Omega)]/2\}$. By the continuity of $\bar{\phi}$,

$$\bar{\phi}(R_1) \leq \underline{\phi}[\mathrm{diam}(\Omega)]/2 \leq e^{-\ell_N(\rho_0)}/2.$$

Lemma 2 implies that

$$\sup_{y\in\Omega_r^{\mathrm{c}}}|\delta\ell_N(\widehat{\rho})(y)| \leq \frac{2e^{\ell_N(\rho_0)}\bar{\phi}(0)}{\underline{\phi}(R_1 + \mathrm{diam}(\Omega))} \cdot \bar{\phi}(r) \leq \frac{4\bar{\phi}(0)}{\bar{\phi}(R_1)\underline{\phi}(R_1 + \mathrm{diam}(\Omega))} \cdot \bar{\phi}(r), \qquad \forall r \geq 0.$$

Let

$$R = \inf\left\{r \geq 0 :\ \bar{\phi}(r) \leq \frac{\bar{\phi}(R_1)\underline{\phi}(R_1 + \mathrm{diam}(\Omega))}{8\bar{\phi}(0)}\right\}.$$

The continuity of $\bar{\phi}$ leads to $\bar{\phi}(R) \leq \bar{\phi}(R_1)\underline{\phi}(R_1 + \mathrm{diam}(\Omega))/[8\bar{\phi}(0)]$ and thus $\sup_{y\in\Omega_R^{\mathrm{c}}}|\delta\ell_N(\widehat{\rho})(y)| \leq 1/2 < 1$. As a result, $\mathsf{supp}(\widehat{\rho}) \cap \Omega_R^{\mathrm{c}} = \varnothing$ and $\mathsf{supp}(\widehat{\rho}) \subseteq \Omega_R$.

Proof of Lemma 2. Note that $\phi(x - y) \leq \bar{\phi}(0)$ for any $x, y \in \mathbb{R}^d$. If $x \in \Omega$ and $y \in \Omega_r^{\mathrm{c}}$, then $\|x - y\|_2 \geq r$ and thus $\phi(x - y) \leq \bar{\phi}(r)$. Hence,

$$(\rho * \phi)(x) \leq \int_{\Omega_r}\phi(x - y)\rho(\mathrm{d}y) + \int_{\Omega_r^{\mathrm{c}}}\phi(x - y)\rho(\mathrm{d}y) \leq \bar{\phi}(0)\rho(\Omega_r) + \bar{\phi}(r)\rho(\Omega_r^{\mathrm{c}}).$$

If $x \in \Omega$ and $y \in \Omega_r$, then $\|x - y\|_2 \geq r + \mathrm{diam}(\Omega)$ and thus $\phi(x - y) \geq \underline{\phi}(r + \mathrm{diam}(\Omega))$. Therefore,

$$(\rho * \phi)(x) \geq \int_{\Omega_r}\phi(x - y)\rho(\mathrm{d}y) \geq \underline{\phi}(r + \mathrm{diam}(\Omega))\rho(\Omega_r).$$

The desired bounds on $\ell_N(\cdot)$ become obvious. If $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\ell_N(\mu) \leq \ell_N(\rho)$, then our estimates of ℓ implies that

$$-\log\Big(\bar{\phi}(r) + \mu(\Omega_r)\bar{\phi}(0)\Big) \leq \ell_N(\mu) \leq \ell_N(\rho), \qquad \forall r \geq 0.$$

Hence, $\bar{\phi}(r) + \mu(\Omega_r)\bar{\phi}(0) \geq e^{-\ell_N(\rho)}$. By Assumption 1, $\lim_{r\to\infty} \bar{\phi}(r) = 0$. Take any $R \geq 0$ such that $\bar{\phi}(R) \leq e^{-\ell_N(\rho)}/2$. Then,

$$\mu(\Omega_R) \geq [e^{-\ell_N(\rho)} - \bar{\phi}(r)]/\bar{\phi}(0) \geq e^{-\ell_N(\rho)}/[2\bar{\phi}(0)],$$
$$\inf_{x\in\Omega}(\mu * \phi)(x) \geq \mu(\Omega_R)\underline{\phi}(R + \mathrm{diam}(\Omega)) \geq e^{-\ell_N(\rho)}\underline{\phi}(R + \mathrm{diam}(\Omega))/[2\bar{\phi}(0)].$$

For any $r \geq 0$, we have $\|X - y\|_2 \geq r$ whenever $X \in \mathsf{supp}(\nu)$ and $y \in \Omega_r^{\mathrm{c}}$. Then,

$$|\delta\ell_N(\mu)(y)| = \frac{1}{N}\sum_{i=1}^{N} \frac{\phi(X_i - y)}{(\mu * \phi)(X_i)} \leq \frac{\bar{\phi}(r)}{\inf_{x\in\Omega}(\mu * \phi)(x)}, \qquad \forall r \geq 0, \quad y \in \Omega_r^{\mathrm{c}}.$$

The proof is finished by combining this and the lower bound on $\inf_{x\in\Omega}(\mu * \phi)(x)$ we have established above. □

C Derivation of gradient flows over $\mathcal{P}_2(\mathbb{R}^d)$

C.1 First variation

Recall that the population and finite-sample loss functions are

$$\ell_\infty(\rho) = \mathbb{E}_{X\sim(\rho^\star * \phi)}\left\{\log\left[(\rho * \phi)(X)\right]\right\}, \qquad \ell_N(\rho) = -\frac{1}{N}\sum_{i=1}^{N}\log\left[(\rho * \phi)(X_i)\right].$$

The first variation of ℓ_N is defined as nay measurable function $\delta\ell(\rho) : \mathbb{R}^d \to \mathbb{R}$ satisfying

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

for all signed measures $\mathcal{X}$ satisfying $\int \mathrm{d}\mathcal{X} = 0$. In particular, it is easy to see that the first variation is defined up to an additive constant.

By direct computation, we have

$$\begin{aligned}
\lim_{\varepsilon\to 0}\frac{\ell_N(\rho + \varepsilon\mathcal{X}) - \ell_N(\rho)}{\varepsilon} &= -\frac{1}{N}\sum_{i=1}^{N}\lim_{\varepsilon\to 0}\frac{\log\left[[(\rho + \varepsilon\mathcal{X}) * \phi](X_i)\right] - \log\left[(\rho * \phi)(X_i)\right]}{\varepsilon} \\
&= -\frac{1}{N}\sum_{i=1}^{N}\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\log\left[1 + \varepsilon\frac{(\mathcal{X} * \phi)(X_i)}{(\rho * \phi)(X_i)}\right] = -\frac{1}{N}\sum_{i=1}^{N}\frac{(\mathcal{X} * \phi)(X_i)}{(\rho * \phi)(X_i)} \\
&= -\frac{1}{N}\sum_{i=1}^{N}\int\frac{\phi(x - X_i)}{(\rho * \phi)(X_i)}\mathcal{X}(\mathrm{d}x),
\end{aligned}$$

As a result, we have

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

Similarly, we can also compute

$$\begin{aligned}
\lim_{\varepsilon\to 0}\frac{\ell_\infty(\rho + \varepsilon\mathcal{X}) - \ell_\infty(\rho)}{\varepsilon} &= \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\left[-\int\log\left[\frac{(\rho + \varepsilon\mathcal{X}) * \phi(x)}{(\rho * \phi)(x)}\right](\rho^\star * \phi)(x)\,\mathrm{d}x\right] \\
&= -\int\frac{(\mathcal{X} * \phi)(x)}{(\rho * \phi)(x)}(\rho^\star * \phi)(x)\,\mathrm{d}x = -\int\frac{(\rho^\star * \phi)(x)}{(\rho * \phi)(x)}\left[\int\phi(x - y)\,\mathcal{X}(\mathrm{d}y)\right]\mathrm{d}x \\
&= -\int\left[\int\frac{(\rho^\star * \phi)(x)}{(\rho * \phi)(x)}\phi(x - y)\,\mathrm{d}x\right]\mathcal{X}(\mathrm{d}y).
\end{aligned}$$

which gives

$$\delta\ell_\infty(\rho) : x \to -\int\frac{(\rho^\star * \phi)(y)}{(\rho * \phi)(y)}\phi(x - y)\,\mathrm{d}y.$$

C.2 Fisher-Rao gradient flow

C.2.1 A formal derivation of gradient flow using Riemannian geometry

We first introduce a Riemannian structure over $\mathcal{P}_2(\mathbb{R}^d)$ underlying the Fisher-Rao metric. Define the tangent space at $\rho \in \mathcal{P}_2(\mathbb{R}^d)$ as

$$\mathrm{Tan}_\rho^{\mathsf{FR}}\mathcal{P}_2(\mathbb{R}^d) \coloneqq \left\{\zeta : \zeta = \rho\left(\alpha - \int \alpha \mathrm{d}\rho\right) \text{ for some } \alpha \text{ satisfying } \int \alpha^2 \mathrm{d}\rho < \infty\right\}.$$

We equip the tangent space $T_\rho^{\mathsf{FR}}\mathcal{P}_2(\mathbb{R}^d)$ with the following Riemannian metric tensor $g_\rho^{\mathsf{FR}}(\cdot,\cdot) : \mathrm{Tan}_\rho^{\mathsf{FR}}\mathcal{P}_2(\mathbb{R}^d) \times \mathrm{Tan}_\rho^{\mathsf{FR}}\mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ as

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

for any $\zeta_1 = \rho(\alpha_1 - \int \alpha_1 \mathrm{d}\rho)$ and $\zeta_2 = \rho(\alpha_2 - \int \alpha_2 \mathrm{d}\rho)$. The metric induced by this Riemannian structure, namely the Fisher-Rao metric $d_{\mathsf{FR}}(\cdot,\cdot)$, satisfies the following property:

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

Then we follow Gallouët and Monsaingeon (2017); Lu et al. (2019) to derive the Fisher-Rao gradient flow with respect to the functional ℓ_N. Let $(\rho_t)_{t\geq 0}$ be a C^1 curve satisfying $\rho_0 = \rho$ with initial velocity

$$\partial_t\rho_t|_{t=0} = \zeta = \rho\left(\alpha - \int \alpha \mathrm{d}\rho\right).$$

The Fisher-Rao gradient of ℓ_N at ρ is defined as the function $\mathrm{grad}_{\mathsf{FR}}\ell_N(\rho) \in L^2(\rho)$ such that

$$\frac{\mathrm{d}}{\mathrm{d}t}\ell_N(\rho_t)\Big|_{t=0} = g_\rho^{\mathsf{FR}}(\mathrm{grad}_{\mathsf{FR}}\ell_N(\rho),\zeta).$$

To compute it, observe that the right-hand side of the above identity is given by

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\ell_N(\rho_t)\Big|_{t=0} &= \int \delta\ell_N(\rho)\cdot\partial_t\rho_t\Big|_{t=0} \\ &= \int \delta\ell_N(\rho),\zeta = \int \delta\ell_N(\rho)\left(\alpha - \int \alpha \mathrm{d}\rho\right)\mathrm{d}\rho \\ &= \int\left(\delta\ell_N(\rho) - \int \delta\ell_N(\rho)\,\mathrm{d}\rho\right)\left(\alpha - \int \alpha \mathrm{d}\rho\right)\mathrm{d}\rho \\ &= g_\rho^{\mathsf{FR}}\left(\rho\left(\delta\ell_N(\rho) - \int \delta\ell_N(\rho)\,\mathrm{d}\rho\right),\zeta\right). \end{aligned}$$

Therefore

$$g_\rho^{\mathsf{FR}}(\mathrm{grad}_{\mathsf{FR}}\ell_N(\rho)\cdot\zeta) = g_\rho^{\mathsf{FR}}\left(\rho\left(\delta\ell_N(\rho) - \int \delta\ell_N(\rho)\,\mathrm{d}\rho\right),\zeta\right)$$

holds for any $\zeta \in \mathrm{Tan}_\rho^{\mathsf{FR}}\mathcal{P}_2(\mathbb{R}^d)$, and as a result

$$\mathrm{grad}_{\mathsf{FR}}\ell_N(\rho) = \rho\left[\delta\ell_N(\rho) - \int \delta\ell_N(\rho)\,\mathrm{d}\rho\right] = \rho\left[\delta\ell_N(\rho) + 1\right],$$

where we have used the fact that $\int \delta\ell_N(\rho)\mathrm{d}\rho = -1$. Hence, the gradient flow of ℓ_N with respect to the Fisher-Rao metric d_{FR} is given by

$$\partial_t\rho_t = -\mathrm{grad}_{\mathsf{FR}}\ell_N(\rho_t) = -\rho_t\left[\delta\ell_N(\rho_t) + 1\right].$$

C.2.2 Other perspectives of Fisher-Rao gradient flow

In this section, we formally illustrate the connection between Fisher-Rao gradient flow (4.1) with proximal gradient descent and mirror descent. For simplicity, we focus on the case when ρ_t is continuous; the case when ρ_t is discrete is similar. We also show the connection between the particle Fisher-Rao gradient descent (2) and the EM algorithm.

Fisher-Rao gradient flow as proximal gradient flow. Consider the proximal gradient update in (4.3). Recall that for $\mu, \nu \in \mathcal{P}_{\mathsf{ac}}(\mathbb{R}^d)$, the Fisher-Rao distance can be expressed as

$$d_{\mathsf{FR}}^2(\mu, \nu) = 4 \int \left| \sqrt{\mu(x)} - \sqrt{\nu(x)} \right|^2 \mathrm{d}x.$$

Note that for the purpose of defining a gradient flow, the metric only matters up to its second-order local expansion

$$d_{\mathsf{FR}}^2(\mu, \nu) = \int \frac{\Delta^2(x)}{\nu(x)} \mathrm{d}x + \mathsf{higher\text{-}order\ terms},$$

where $\Delta = \mu - \nu$. Therefore we obtain an asymptotically (as $\eta \to 0$) equivalent problem

$$\rho_t^\eta := \operatorname*{arg\,min}_{\rho \in \mathcal{P}_{\mathsf{ac}}(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \delta \ell_N(\rho_t) \, \mathrm{d}(\rho - \rho_t) + \frac{1}{2\eta} \int \frac{[\rho(x) - \rho_t(x)]^2}{\rho_t(x)} \mathrm{d}x \right\}.$$

The first-order optimality condition is

$$\delta \ell_N(\rho_t)(x) + \frac{1}{\eta} \cdot \frac{\rho(x) - \rho_t(x)}{\rho_t(x)} = c$$

for some constant $c \in \mathbb{R}$. This gives

$$\rho_t^\eta(x) = \rho_t(x) \left[1 + x\eta - \eta \delta \ell_N(\rho_t)(x) \right].$$

Since ρ_t^η is a probability density, we have

$$1 = \int_{\mathbb{R}^d} \rho_t^\eta(x) \, \mathrm{d}x = \int_{\mathbb{R}^d} \rho_t(x) \left[1 + c\eta - \eta \delta \ell_N(\rho_t)(x) \right] \mathrm{d}x = 1 + c\eta - \eta,$$

where we use the fact that $\int_{\mathbb{R}^d} \delta \ell_N(\rho) \mathrm{d}\rho = -1$ for any $\rho \in \mathcal{P}_2(\mathbb{R}^d)$. This gives $c = -1$. Therefore

$$\rho_t^\eta(x) = \rho_t(x) \left[1 - \eta - \eta \delta \ell_N(\rho_t)(x) \right],$$

and as a result,

$$\partial_t \rho_t = \lim_{\eta \to 0+} \frac{\rho_t^\eta - \rho_t}{\eta} = -\left[1 + \delta \ell_N(\rho_t) \right],$$

which recovers the Fisher-Rao gradient flow (4.1).

Fisher-Rao gradient flow as mirror flow. Recall that the mirror descent update is defined as

$$\rho_t^\eta := \operatorname*{arg\,min}_{\rho \in \mathcal{P}_{\mathsf{ac}}(\mathbb{R}^d)} \int_{\mathbb{R}^d} \delta \ell_N(\rho_t) \, \mathrm{d}(\rho - \rho_t) + \frac{1}{\eta} \mathsf{KL}(\rho \,\|\, \rho_t).$$

The first variation of $f(\cdot) := \mathsf{KL}(\cdot \,\|\, \rho_t)$ is given by

$$\delta f(\rho)(x) = \log \left[\frac{\rho(x)}{\rho_t(x)} \right],$$

therefore the first-order optimality condition reads

$$\delta \ell_N(\rho_t)(x) + \frac{1}{\eta} \log \left[\frac{\rho(x)}{\rho_t(x)} \right] = c$$

for some constant $c > 0$. This gives

$$\frac{\rho(x)}{\rho_t(x)} = \exp\{\eta[c - \delta\ell_N(\rho_t)(x)]\} \propto \exp[-\eta\delta\ell_N(\rho_t)(x)].$$

Since $\int_{\mathbb{R}^d} \rho_t^\eta(x)\mathrm{d}x = 1$, we know that the closed-form solution is given by

$$\rho_t^\eta(x) = \frac{\rho_t(x)\exp[-\eta\delta\ell_N(\rho_t)(x)]}{\int \rho_t(y)\exp[-\eta\delta\ell_N(\rho_t)(y)]\,\mathrm{d}y}. \tag{C.1}$$

Then as $\eta \to 0$, we can compute

$$\begin{aligned}\rho_t^\eta(x) &= \frac{\rho_t(x)\left[1 - \eta\delta\ell_N(\rho_t)(x) + O(\eta^2)\right]}{\int \rho_t(y)\left[1 - \eta\delta\ell_N(\rho_t)(y) + O(\eta^2)\right]\mathrm{d}y} = \frac{\rho_t(x)\left[1 - \eta\delta\ell_N(\rho_t)(x) + O(\eta^2)\right]}{1 + \eta + O(\eta^2)} \\ &= \rho_t(x)\left\{1 - \eta\left[1 + \delta\ell_N(\rho_t)(x)\right] + O(\eta^2)\right\},\end{aligned}$$

where we use the fact that $\int \delta\ell_N(\rho)\mathrm{d}\rho = -1$ for any $\rho \in \mathcal{P}_2(\mathbb{R}^d)$. Therefore the continuous-time limit of mirror descent is

$$\partial_t\rho_t = \lim_{\eta\to 0+} \frac{\rho_t^\eta(x) - \rho_t(x)}{\eta} = -\left[1 + \delta\ell_N(\rho_t)(x)\right],$$

which recovers the Fisher-Rao gradient flow (4.1).

Fisher-Rao gradient descent as EM algorithm. Now we consider fitting a m-component Gaussian mixture model with unknown weights $\{\omega^{(j)}\}_{1\le j\le m}$, known location parameters $\{\mu_j\}_{1\le j\le m}$ and isotropic covariance. Given the data $\{X_i\}_{1\le i\le N}$, the MLE is given by

$$\arg\max_{\omega\in\Delta^{m-1}} \ell(\omega) = \frac{1}{N}\sum_{i=1}^N \log\left[\sum_{j=1}^m \omega^{(j)}\phi(X_i - \mu_j)\right].$$

The Expectation-Maximization algorithm for solving the above MLE is given as follows. We first introduce the latent i.i.d. random variables $\{J_i\}_{1\le i\le N}$ distributed $\mathbb{P}(J_i = j) = \omega^{(j)}$ for $1 \le j \le m$, then the distribution of the observed samples is $X_i|J_i = j \sim \mathcal{N}(\mu_j, I_d)$. The joint distribution of (X_i, J_i) is given by

$$p_\omega(x, j) = \phi(X_i - \mu_j)\,\omega^{(j)},$$

and conditional on $X_i = x$, the conditional distribution of J_i is given by

$$p_\omega(j|x) = \frac{p_\omega(x,j)}{\sum_{l=1}^m p_\omega(x,l)} = \frac{\phi(X_i - \mu_j)\,\omega^{(j)}}{\sum_{l=1}^m \phi(X_i - \mu_l)\,\omega^{(l)}}.$$

Given the current estimate ω_t, the E-step amounts to computing

$$\begin{aligned}Q(\omega|\omega_t) &= \frac{1}{N}\sum_{i=1}^N\sum_{j=1}^m p_{\omega_t}(j|X_i)\log p_\omega(X_i, j) \\ &= \frac{1}{N}\sum_{i=1}^N\sum_{j=1}^m \frac{\phi(X_i - \mu_j)\,\omega_t^{(j)}}{\sum_{l=1}^m \phi(X_i - \mu_l)\,\omega_t^{(l)}}\log\left[\phi(X_i - \mu_j)\,\omega^{(j)}\right].\end{aligned}$$

The M-step is to update

$$\omega_{t+1} \coloneqq \arg\max_{\omega\in\Delta^{m-1}} Q(\omega|\omega_t),$$

which is given by

$$\omega_{t+1}^{(j)} = \frac{1}{N}\sum_{i=1}^N \frac{\phi(X_i - \mu_j)\,\omega_t^{(j)}}{\sum_{l=1}^m \phi(X_i - \mu_l)\,\omega_t^{(l)}} \qquad \forall\, 1 \le j \le m.$$

This is equivalent to Algorithm 2 with step size $\eta = 1$.

C.3 Wasserstein gradient flow

We introduce the Riemannian structure over $\mathcal{P}_2(\mathbb{R}^d)$ underlying the quadratic Wasserstein distance. We define the tangent space at $\rho \in \mathcal{P}_2(\mathbb{R}^d)$ to be

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

We equip this tangent space with the $L^2(\rho)$ metric, namely we define the Riemannian metric tensor $g_\rho^{\mathsf{W}}(\cdot,\cdot) : \mathrm{Tan}_\rho^{\mathsf{W}}\mathcal{P}_2(\mathbb{R}^d) \times \mathrm{Tan}_\rho^{\mathsf{W}}\mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ as

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

for any $\zeta_1 = -\mathsf{div}(\rho\nabla u_1)$ and $\zeta_2 = -\mathsf{div}(\rho\nabla u_2)$. The metric induced by this Riemannian structure recovers the quadratic Wasserstein distance, namely

$$\begin{aligned} d_{\mathsf{W}}^2\left(\rho_0, \rho_1\right) &= \inf\left\{\int_0^1 \int \|v_t\|_2^2 \,\mathrm{d}\rho_t \mathrm{d}t : \left(\rho_t, v_t\right)_{t\in[0,1]} \text{ solves } \partial_t\rho_t + \mathsf{div}(\rho_t v_t) = 0\right\} \\ &= \inf_{\pi\in\Pi(\rho_0,\rho_1)} \int \|x-y\|_2^2 \,\pi\left(\mathrm{d}x, \mathrm{d}y\right), \end{aligned}$$

where $\Pi(\rho_0, \rho_1)$ is the set of couplings of ρ_0 and ρ_1. This is known as the Benamou-Brenier formula for the Wasserstein distance. Then we derive the Wasserstein gradient flow with respect to the functional ℓ_N. Interested readers are referred to Ambrosio et al. (2008) for detailed introduction to Wasserstein gradient flow. Let $(\rho_t)_{t\geq 0}$ be a C^1 curve satisfying $\rho_0 = \rho$ with initial velocity

$$\partial_t\rho_t|_{t=0} = \zeta = -\mathsf{div}\left(\rho\nabla u\right).$$

Then it should hold that

$$\frac{\mathrm{d}}{\mathrm{d}t}\ell_N\left(\rho_t\right)\Big|_{t=0} = g_\rho^{\mathsf{W}}\left(\mathrm{grad}_{\mathsf{W}}\ell_N\left(\rho\right), \zeta\right).$$

The left hand side of the above equation equals to

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\ell_N\left(\rho_t\right)\Big|_{t=0} &= \int \delta\ell_N\left(\rho\right)\zeta \mathrm{d}x = -\int \delta\ell_N\left(\rho\right)\mathsf{div}\left(\rho\nabla u\right)\mathrm{d}x \\ &= -\int \langle \nabla\delta\ell_N\left(\rho\right), \nabla u\rangle \,\mathrm{d}\rho \\ &= g_\rho^{\mathsf{W}}\left(-\mathsf{div}\left(\nabla\delta\ell_N\left(\rho\right)\rho\right), \zeta\right). \end{aligned}$$

Therefore

$$g_\rho^{\mathsf{W}}\left(\mathrm{grad}_{\mathsf{W}}\ell_N\left(\rho\right), \zeta\right) = g_\rho^{\mathsf{W}}\left(-\mathsf{div}\left(\nabla\delta\ell_N\left(\rho\right)\rho\right), \zeta\right)$$

holds for any $\zeta \in \mathrm{Tan}_\rho^{\mathsf{W}}\mathcal{P}_2(\mathbb{R}^d)$, and as a result

$$\mathrm{grad}_{\mathsf{W}}\ell_N\left(\rho\right) = -\mathsf{div}\left(\nabla\delta\ell_N\left(\rho\right)\rho\right).$$

This shows that the gradient flow of ℓ_N with respect to the quadratic Wasserstein distance d_{W} is given by

$$\partial_t\rho_t = -\mathrm{grad}_{\mathsf{W}}\ell_N\left(\rho_t\right) = \mathsf{div}\left(\nabla\delta\ell_N\left(\rho_t\right)\rho_t\right).$$

C.4 Wasserstein-Fisher-Rao gradient flow

We introduce the Riemannian structure over $\mathcal{P}_2(\mathbb{R}^d)$ underlying the Wasserstein-Fisher-Rao metric. Define the tangent space at $\rho \in \mathcal{P}_2(\mathbb{R}^d)$ to be

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

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

We equip this tangent space with the Riemannian metric tensor $g_\rho^{\mathsf{WFR}}(\cdot,\cdot):\mathrm{Tan}_\rho^{\mathsf{WFR}}\mathcal{P}_2(\mathbb{R}^d)\times\mathrm{Tan}_\rho^{\mathsf{WFR}}\mathcal{P}_2(\mathbb{R}^d)\to\mathbb{R}$ defined as

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

for any $\zeta_1=-\mathsf{div}(\rho\nabla u_1)+\rho(\alpha_1-\int\alpha_1\mathrm{d}\rho)$ and $\zeta_2=-\mathsf{div}(\rho\nabla u_2)+\rho(\alpha_2-\int\alpha_2\mathrm{d}\rho)$. The metric induced by the above Riemannian structure, namely the Wasserstein-Fisher-Rao metric $d_{\mathsf{WFR}}(\cdot,\cdot)$, is defined as

$$\begin{aligned}d_{\mathsf{WFR}}^2(\rho_0,\rho_1)=\inf\Bigg\{\int_0^1\int\Big[\|v_t\|^2+\Big(\alpha_t-\int\alpha_t\mathrm{d}\rho_t\Big)^2\Big]\mathrm{d}\rho_t\mathrm{d}t:(\rho_t,v_t,\alpha_t)_{0\le t\le1}\\\text{solves }\partial_t\rho_t=-\mathsf{div}(\rho_t v_t)+\rho_t\alpha_t\Bigg\}.\end{aligned}$$

Then we follow Gallouët and Monsaingeon (2017); Lu et al. (2019) to derive the Wasserstein-Fisher-Rao gradient flow with respect to the functional ℓ_N. Let $(\rho_t)_{t\ge0}$ be a C^1 curve satisfying $\rho_0=\rho$ with initial velocity

$$\partial_t\rho_t|_{t=0}=\zeta=-\mathsf{div}\,(\rho\nabla u)+\rho\left(\alpha-\int\alpha\mathrm{d}\rho\right).$$

Then it should hold that

$$\frac{\mathrm{d}}{\mathrm{d}t}\ell_N(\rho_t)\Big|_{t=0}=g_\rho^{\mathsf{WFR}}\left(\mathrm{grad}_{\mathsf{WFR}}\ell_N(\rho),\zeta\right).$$

The left hand side of the above equation equals to

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\ell_N(\rho_t)\Big|_{t=0}&=\int\delta\ell_N(\rho)\,\zeta\mathrm{d}x=\int\delta\ell_N(\rho)\left[-\mathsf{div}\,(\rho\nabla u)+\rho\left(\alpha-\int\alpha\mathrm{d}\rho\right)\right]\mathrm{d}x\\&=-\int\langle\nabla\delta\ell_N(\rho),\nabla u\rangle\,\mathrm{d}\rho+\int\left(\delta\ell_N(\rho)-\int\delta\ell_N(\rho)\,\mathrm{d}\rho\right)\left(\alpha-\int\alpha\mathrm{d}\rho\right)\mathrm{d}\rho\\&=g_\rho^{\mathsf{WFR}}\left(-\mathsf{div}\,(\nabla\delta\ell_N(\rho)\,\rho)+\rho\left(\delta\ell_N(\rho)-\int\delta\ell_N(\rho)\,\mathrm{d}\rho\right),\zeta\right).\end{aligned}$$

Therefore

$$g_\rho^{\mathsf{WFR}}\left(\mathrm{grad}_{\mathsf{W}}\ell_N(\rho),\zeta\right)=g_\rho^{\mathsf{WFR}}\left(-\mathsf{div}\,(\nabla\delta\ell_N(\rho)\,\rho)+\rho\left(\delta\ell_N(\rho)-\int\delta\ell_N(\rho)\,\mathrm{d}\rho\right),\zeta\right)$$

holds for any $\zeta\in\mathrm{Tan}_\rho^{\mathsf{WFR}}\mathcal{P}_2(\mathbb{R}^d)$, and as a result

$$\begin{aligned}\mathrm{grad}_{\mathsf{WFR}}\ell_N(\rho)&=-\mathsf{div}\,(\nabla\delta\ell_N(\rho)\,\rho)+\rho\left(\delta\ell_N(\rho)-\int\delta\ell_N(\rho)\,\mathrm{d}\rho\right)\\&=-\mathsf{div}\,(\nabla\delta\ell_N(\rho)\,\rho)+\rho\left[1+\delta\ell_N(\rho)\right],\end{aligned}$$

where we have used the fact that $\int\delta\ell_N(\rho)\mathrm{d}\rho=-1$. This shows that the gradient flow of ℓ_N with respect to the Wasserstein-Fisher-Rao metric d_{WFR} is given by

$$\partial_t\rho_t=-\mathrm{grad}_{\mathsf{WFR}}\ell_N(\rho_t)=\mathsf{div}\,(\nabla\delta\ell_N(\rho_t)\,\rho_t)-\rho_t\left[1+\delta\ell_N(\rho_t)\right].$$

D Proofs for the convergence theory

D.1 Proof of Lemma 1

We invoke a descent lemma Wasserstein gradient descent, whose proof is deferred to Appendix D.2. Such a lemma is standard in convex optimization optimization (see, e.g., Bubeck, 2015, eq. (3.5)). It has appeared for optimization over the Wasserstein space in Salim et al. (2020) under for functionals that are convex along generalized geodesics, an assumption that does not hold for the negative log-likelihood ℓ_N.

Lemma 3 (A descent lemma)**.** *Let Assumption 1 hold. Choose any $\rho \in \mathcal{P}(\mathbb{R}^d)$. Define $c = \inf_{x\in\Omega}(\rho * \phi)(x)$, $G = \sup_{x\in\mathbb{R}^d} \|\nabla\phi(x)\|_2$ and $H = \sup_{x\in\mathbb{R}^d} \|\nabla^2\phi(x)\|_2$.*

- *We have*

$$\ell_N(\rho^{\mathrm{W},\eta}) - \ell_N(\rho) \le -\eta\left[1 - \frac{\eta}{2c}\left(H + \frac{G^2}{c}\right)\right]\mathbb{E}_{Y\sim\rho}\|\nabla\delta\ell_N(\rho)(Y)\|_2^2.$$

  In addition, we have $\sup_{x\in\mathbb{R}^d} \|\nabla^2\delta\ell_N(\rho)(x)\|_2 \le H/c$.

- *If $0 \le \eta < c/H$ and $\mathsf{supp}(\rho) = \mathbb{R}^d$, then $\mathsf{supp}(\rho^{\mathrm{W},\eta}) = \mathbb{R}^d$.*

We now come back to Lemma 1. Let $R = \inf\{r \ge 0 : \ \bar{\phi}(r) \le e^{-\ell_N(\rho_0)}/2\}$. Lemma 2 implies that for any $\mu \in \mathcal{P}(\mathbb{R}^d)$ with $\ell_N(\mu) \le \ell_N(\rho_0)$, we have

$$\inf_{x\in\Omega}(\mu * \phi)(x) \ge e^{-\ell_N(\rho_0)}\underline{\phi}(R + \mathrm{diam}(\Omega))/[2\bar{\phi}(0)] = c_0.$$

In particular, $\inf_{x\in\Omega}(\rho_0 * \phi)(x) \ge c_0$. When $\eta \le c_0/(H + G^2/c_0)$, Lemma 3 and the definition $\rho_1 = (\mathrm{id} - \eta\nabla\delta\ell_N(\rho_0))_{\#}\rho_0$ together yield

$$\ell_N(\rho_1) - \ell_N(\rho_0) \le -\frac{\eta}{2}\mathbb{E}_{Y\sim\rho_0}\|\nabla\delta\ell_N(\rho_0)(Y)\|_2^2 \le -\frac{1}{2\eta}W_2^2(\rho_1,\rho_0) \le 0.$$

Also, if $\mathsf{supp}(\rho_0) = \mathbb{R}^d$, then $\mathsf{supp}(\rho_1) = \mathbb{R}^d$. From $\ell_N(\rho_1) \le \ell_N(\rho_0)$ we obtain that $\inf_{x\in\Omega}(\rho_1 * \phi)(x) \ge c_0$. Then, the proof is completed by induction.

D.2 Proof of Lemma 3

We prove the two results in Lemma 3 in sequence.

Part 1. Let $\nu = N^{-1}\sum_{i=1}^N \delta_{X_i}$ be the empirical data distribution, and let $h(x) = -\log x$ for $x > 0$. Then we can wirte

$$\ell_N(\rho) = -\frac{1}{N}\sum_{i=1}^N \log\left[\rho * \phi\left(X_i\right)\right] = \mathbb{E}_{X\sim\nu}\left[h\left((\rho * \phi)(X)\right)\right] = \mathbb{E}_{X\sim\nu}\left[h\left(\mathbb{E}_{Y\sim\rho}\left[\phi(X - Y)\right]\right)\right]$$

as well as

$$\ell_N(\rho^{\mathrm{W},\eta}) = \mathbb{E}_{X\sim\nu}\left[h\left(\mathbb{E}_{Y\sim\rho^{\mathrm{W},\eta}}\left[\phi(X - Y)\right]\right)\right] = \mathbb{E}_{X\sim\nu}\left[h\left(\mathbb{E}_{Y\sim\rho}\left[\phi\left(X - Y + \eta\nabla\delta\ell_N(\rho)(Y)\right)\right]\right)\right].$$

Define $a(x) = \mathbb{E}_{Y\sim\rho}[\phi(x - Y + \eta\nabla\delta\ell_N(\rho)(Y))]$ and $b(x) = \mathbb{E}_{Y\sim\rho}[\phi(x - Y)] = (\rho * \phi)(x)$ for any $x \in \mathbb{R}^d$. Then we can write

$$\ell_N(\rho) = \mathbb{E}_{X\sim\nu}\left[h\left(b(X)\right)\right], \qquad \ell_N(\rho^{\mathrm{W},\eta}) = \mathbb{E}_{X\sim\nu}\left[h\left(a(X)\right)\right].$$

Note that $h'(x) = -x^{-1} < 0$, $h''(x) = x^{-2} > 0$ and $h'''(x) = -2x^{-3} < 0$. For any $a, b > 0$, by Taylor's theorem

$$h(a) \le h(b) + h'(b)(a - b) + \frac{h''(b)}{2}(a - b)^2 = h(b) - \frac{a - b}{b} + \frac{(a - b)^2}{2b^2}.$$

Taking the above two equations collectively gives

$$\begin{aligned}\ell_N(\rho^{\mathrm{W},\eta}) - \ell_N(\rho) &= \mathbb{E}_{X\sim\nu}\left[h\left(a(X)\right) - h\left(b(X)\right)\right] \\ &\le \underbrace{-\mathbb{E}_{X\sim\nu}\left[\frac{a(X)-b(X)}{b(X)}\right]}_{=:\alpha_1} + \underbrace{\frac{1}{2}\mathbb{E}_{X\sim\nu}\left[\frac{[a(X)-b(X)]^2}{b(X)^2}\right]}_{=:\alpha_2}. \end{aligned} \tag{D.1}$$

Now derive upper bounds for α_1 and α_2 respectively.

- To control α_1, let $G = \sup_{x\in\mathbb{R}^d}\|\nabla\phi(x)\|_2$ and observe that

  $$\left|\phi\left(x - y + \eta\nabla\delta\ell_N(\rho)(y)\right) - \phi(x-y)\right| \le G\eta\left\|\nabla\delta\ell_N(\rho)(y)\right\|_2$$

  for all $x, y \in \mathbb{R}^d$, therefore

  $$\left|a\left(x\right) - b\left(x\right)\right| \le G\eta\mathbb{E}_{Y\sim\rho}\left[\left\|\nabla\delta\ell_N\left(\rho\right)\left(Y\right)\right\|_2\right] \tag{D.2}$$

  holds for all $x \in \mathbb{R}^d$. Then we have

  $$\alpha_1 \overset{\text{(i)}}{\le} \frac{G^2\eta^2\mathbb{E}^2_{Y\sim\rho}\left[\left\|\nabla\delta\ell_N\left(\rho\right)\left(Y\right)\right\|\right]}{c^2} \overset{\text{(ii)}}{\le} \frac{G^2\eta^2}{c^2}\mathbb{E}_{Y\sim\rho}\left[\left\|\nabla\delta\ell_N\left(\rho\right)\left(Y\right)\right\|_2^2\right], \tag{D.3}$$

  where (i) follows from (D.2) and the fact

  $$c = \inf_{x\in\Omega}(\rho * \phi)(x) = \inf_{x\in\Omega} b(x) \tag{D.4}$$

  and (ii) follows from Jensen's inequality.

- Regarding α_2, let $H = \sup_{x\in\mathbb{R}^d}\|\nabla^2\phi(x)\|_2$ and we have

  $$\phi\left(x - y + \eta\nabla\delta\ell_N(\rho)(y)\right) - \phi(x-y) \ge \langle\nabla\phi(x-y), \eta\nabla\delta\ell_N(\rho)(y)\rangle - \frac{H}{2}\eta^2\|\nabla\delta\ell_N(\rho)(y)\|_2^2,$$

  for any $x, y \in \mathbb{R}^d$, and therefore

  $$a(x) - b(x) \ge \mathbb{E}_{Y\sim\rho}\left[\langle\nabla\phi(x-Y), \eta\nabla\delta\ell_N(\rho)(Y)\rangle\right] - \frac{H}{2}\eta^2\mathbb{E}_{Y\sim\rho}\left[\left\|\nabla\delta\ell_N(\rho)(Y)\right\|_2^2\right] \tag{D.5}$$

  for any $x \in \mathbb{R}^d$. Since $b(x) > 0$, we have

  $$\begin{aligned}\alpha_2 &\overset{\text{(i)}}{\le} -\mathbb{E}_{X\sim\nu}\left[\frac{\mathbb{E}_{Y\sim\rho}\left[\langle\nabla\phi(X-Y), \eta\nabla\delta\ell_N(\rho)(Y)\rangle\right] - \frac{H}{2}\eta^2\mathbb{E}_{Y\sim\rho}\left[\left\|\nabla\delta\ell_N(\rho)(Y)\right\|_2^2\right]}{b\left(X\right)}\right] \\ &= -\eta\mathbb{E}_{Y\sim\rho}\left[\left\langle\mathbb{E}_{X\sim\nu}\left(\frac{\nabla\phi(X-Y)}{b(X)}\right), \nabla\delta\ell_N(\rho)(Y)\right\rangle\right] + \frac{H\eta^2}{2}\mathbb{E}_{X\sim\nu}\left[\frac{1}{b(X)}\right]\mathbb{E}_{Y\sim\rho}\left[\left\|\nabla\delta\ell_N(\rho)(Y)\right\|_2^2\right] \\ &\overset{\text{(ii)}}{=} -\eta\mathbb{E}_{Y\sim\rho}\left[\left\|\nabla\delta\ell_N(\rho)(Y)\right\|_2^2\right] + \frac{H\eta^2}{2}\mathbb{E}_{X\sim\nu}\left[\frac{1}{b(X)}\right]\mathbb{E}_{Y\sim\rho}\left[\left\|\nabla\delta\ell_N(\rho)(Y)\right\|_2^2\right] \\ &\overset{\text{(iii)}}{\le} \left(-\eta + \frac{H\eta^2}{2c}\right)\mathbb{E}_{Y\sim\rho}\left[\left\|\nabla\delta\ell_N(\rho)(Y)\right\|_2^2\right]. \end{aligned} \tag{D.6}$$

  Here (i) utilizes (D.5); (ii) holds since

  $$\delta\ell_N(\rho)(y) = -\mathbb{E}_{X\sim\nu}\left[\frac{\phi(X-y)}{b(X)}\right], \qquad \nabla\delta\ell_N(\rho)(y) = \mathbb{E}_{X\sim\nu}\left[\frac{\nabla\phi(X-y)}{b(X)}\right], \tag{D.7}$$

  and therefore

  $$\mathbb{E}_{Y\sim\rho}\left[\left\langle\mathbb{E}_{X\sim\nu}\left(\frac{\nabla\phi(X-Y)}{b(X)}\right), \nabla\delta\ell_N(\rho)(Y)\right\rangle\right] = \mathbb{E}_{Y\sim\rho}\left[\left\|\nabla\delta\ell_N(\rho)(Y)\right\|_2^2\right];$$

  and (iii) follows from (D.4).

Taking (D.1), (D.3) and (D.6) collectively gives

$$\ell_N(\rho^{\mathrm{W},\eta}) - \ell_N(\rho) \le \alpha_1 + \alpha_2 \le \left(-\eta + \frac{H\eta^2}{2c} + \frac{G^2\eta^2}{2c^2} \right) \mathbb{E}_{Y\sim\rho}\left[\|\nabla \delta\ell_N(\rho)(Y)\|_2^2 \right]$$
$$= -\eta\left[1 - \frac{\eta}{2c}\left(H + \frac{G^2}{c} \right) \right] \mathbb{E}_{Y\sim\rho}\left[\|\nabla \delta\ell_N(\rho)(Y)\|_2^2 \right].$$

Finally, we learn from (D.7) and (D.4) that

$$\left\| \nabla^2 \delta\ell_N(\rho)(x) \right\|_2 = \left\| \mathbb{E}_{X\sim\nu}\left[\frac{\nabla^2\phi(X-x)}{b(X)} \right] \right\|_2 \le \frac{\sup_{x\in\mathbb{R}^d} \|\nabla^2\phi(x)\|_2}{\inf_{x\in\mathbb{R}^d} b(x)} = \frac{H}{c}.$$

Part 2. When $\eta < c/H$, the mapping $x \mapsto \eta\nabla\delta\ell_N(\rho)(x)$ is a contraction. Let $\varphi(x) = x - \eta\nabla\delta\ell_N(\rho)(x)$. By Lemma 3, $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ is a bijection and φ^{-1} is Lipschitz. The second-order differentiability of $\delta\ell_N(\rho)$ implies the differentiability of φ and thus φ^{-1}. If $\mathsf{supp}(\rho) = \mathbb{R}^d$, then $\mathsf{supp}(\rho^{\mathrm{W},\eta}) = \mathsf{supp}(\varphi_{\#}\rho) = \mathbb{R}^d$.

D.3 Proof of Lemma 2

Let $\nu = N^{-1}\sum_{i=1}^N \delta_{X_i}$ be the empirical data distribution. For any $\mu \ll \rho$, we have

$$\ell_N(\mu) = -\mathbb{E}_{X\sim\nu}\left[\log\left((\mu * \phi)(X) \right) \right] = -\mathbb{E}_{X\sim\nu}\left[\log\left(\mathbb{E}_{Y\sim\mu}\left[\phi(X-Y) \right] \right) \right]$$
$$= -\mathbb{E}_{X\sim\nu}\left[\log\left(\mathbb{E}_{Y\sim\rho}\left[\frac{\mathrm{d}\mu}{\mathrm{d}\rho}(Y)\cdot\phi(X-Y) \right] \right) \right]$$
$$= -\mathbb{E}_{X\sim\nu}\left[\log\left(\mathbb{E}_{Y\sim\rho}\left[\frac{\mathrm{d}\mu}{\mathrm{d}\rho}(Y)\cdot(\rho*\phi)(X)\cdot\frac{\phi(X-Y)}{(\rho*\phi)(X)} \right] \right) \right].$$

For any $x \in \mathbb{R}^d$, we can define a new probability measure $\rho_x^{\mathrm{FR},1} \in \mathcal{P}(\mathbb{R}^d)$ through

$$\frac{\mathrm{d}\rho_x^{\mathrm{FR},1}}{\mathrm{d}\rho}(\cdot) = \frac{\phi(x-\cdot)}{(\rho*\phi)(x)}, \tag{D.8}$$

and we can check that $\rho_x^{\mathrm{FR},1}$ is indeed a probability measure since

$$\int_{\mathbb{R}^d} \mathrm{d}\rho_x^{\mathrm{FR},1} = \int_{\mathbb{R}^d} \frac{\phi(x-y)}{(\rho*\phi)(x)}\rho(\mathrm{d}x) = 1.$$

Then we have, by the convexity of $t \mapsto -\log t$ and Jensen's inequality,

$$\ell_N(\mu) \overset{\text{(i)}}{=} -\mathbb{E}_{X\sim\nu}\left[\log\left(\mathbb{E}_{Y\sim\rho_X^{\mathrm{FR},1}}\left[\frac{\mathrm{d}\mu}{\mathrm{d}\rho}(Y)\cdot(\rho*\phi)(X) \right] \right) \right]$$
$$\overset{\text{(ii)}}{\le} -\mathbb{E}_{X\sim\nu}\left[\mathbb{E}_{Y\sim\rho_X^{\mathrm{FR},1}}\left[\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\rho}(Y)\cdot(\rho*\phi)(X) \right) \right] \right]$$
$$\overset{\text{(iii)}}{=} -\mathbb{E}_{X\sim\nu}\left[\mathbb{E}_{Y\sim\rho}\left[\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\rho}(Y)\cdot(\rho*\phi)(X) \right)\cdot\frac{\phi(X-Y)}{(\rho*\phi)(X)} \right] \right]$$
$$= \underbrace{-\mathbb{E}_{X\sim\nu,Y\sim\rho}\left[\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\rho}(Y) \right)\cdot\frac{\phi(X-Y)}{(\rho*\phi)(X)} \right]}_{=:\beta_1} \underbrace{-\mathbb{E}_{X\sim\nu,Y\sim\rho}\left[\log\left((\rho*\phi)(X) \right)\cdot\frac{\phi(X-Y)}{(\rho*\phi)(X)} \right]}_{=:\beta_2}. \tag{D.9}$$

Here (i) and (iii) utilizes (D.8), and (ii) follows from Jensen's inequality and the convexity of $t \mapsto -\log t$. Then we study the two terms β_1 and β_2 respectively. Regarding β_1, we have

$$\beta_1 = \mathbb{E}_{Y\sim\rho}\left[\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\rho}(Y) \right) \mathbb{E}_{X\sim\nu}\left[\frac{\phi(X-Y)}{(\rho*\phi)(X)} \right] \right]$$
$$= \mathbb{E}_{Y\sim\rho}\left[\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\rho}(Y) \right)\cdot\frac{\mathrm{d}\rho^{\mathrm{FR},1}}{\mathrm{d}\rho}(Y) \right] = \mathbb{E}_{Y\sim\rho^{\mathrm{FR},1}}\left[\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\rho}(Y) \right) \right]$$

$$= \mathbb{E}_{Y\sim\rho^{\mathrm{FR},1}}\left[\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\rho^{\mathrm{FR},1}}(Y)\right)\right] + \mathbb{E}_{Y\sim\rho^{\mathrm{FR},1}}\left[\log\left(\frac{\mathrm{d}\rho^{\mathrm{FR},1}}{\mathrm{d}\rho}(Y)\right)\right]$$
$$= -\mathsf{KL}\left(\rho^{\mathrm{FR},1}\,\|\,\mu\right) + \mathsf{KL}\left(\rho^{\mathrm{FR},1}\,\|\,\rho\right). \tag{D.10}$$

Regarding β_2, we have

$$\beta_2 = \mathbb{E}_{X\sim\nu}\left[\log\left((\rho*\phi)(X)\right)\mathbb{E}_{Y\sim\rho}\left[\frac{\phi(X-Y)}{(\rho*\phi)(X)}\right]\right] = \mathbb{E}_{X\sim\nu}\left[\log\left((\rho*\phi)(X)\right)\right] = -\ell_N(\rho). \tag{D.11}$$

Taking (D.9), (D.10) and (D.11) collectively yields

$$\ell_N(\mu) \le \ell_N(\rho) + \mathsf{KL}\left(\rho^{\mathrm{FR},1}\,\|\,\mu\right) - \mathsf{KL}\left(\rho^{\mathrm{FR},1}\,\|\,\rho\right), \qquad \forall\,\mu \ll \rho.$$

By taking $\mu = \rho^{\mathrm{FR},1}$, we get

$$\ell_N\left(\rho^{\mathrm{FR},1}\right) \le \ell_N(\rho) - \mathsf{KL}\left(\rho^{\mathrm{FR},1}\,\|\,\rho\right). \tag{D.12}$$

Recall that for any $\gamma \in (0,1)$ we have $\rho^{\mathrm{FR},\gamma} = (1-\gamma)\rho + \gamma\rho^{\mathrm{FR},1}$. Therefore

$$\ell_N\left(\rho^{\mathrm{FR},\gamma}\right) \overset{\text{(i)}}{\le} (1-\gamma)\ell_N(\rho) + \gamma\ell_N(\rho^{\mathrm{FR},1}) \overset{\text{(ii)}}{\le} \ell_N(\rho) - \gamma\mathsf{KL}\left(\rho^{\mathrm{FR},1}\,\|\,\rho\right), \tag{D.13}$$

where (i) holds since $\ell_N(\rho)$ is ℓ_2-convex in ρ, and (ii) follows from (D.12). By the ℓ_2-convexity of $\mathsf{KL}(\cdot\,\|\,\rho)$, we have

$$\mathsf{KL}\left(\rho^{\mathrm{FR},\gamma}\,\|\,\rho\right) \le (1-\gamma)\mathsf{KL}\left(\rho\,\|\,\rho\right) + \gamma\mathsf{KL}\left(\rho^{\mathrm{FR},1}\,\|\,\rho\right) = \gamma\mathsf{KL}\left(\rho^{\mathrm{FR},1}\,\|\,\rho\right). \tag{D.14}$$

Combine (D.13) and (D.14) to achieve

$$\ell_N\left(\rho^{\mathrm{FR},\gamma}\right) \le \ell_N(\rho) - \mathsf{KL}\left(\rho^{\mathrm{FR},\gamma}\,\|\,\rho\right).$$

E Well-posedness of particle gradient flows

E.1 Fisher-Rao gradient flow (Proof of Theorem 5)

In this section, we will show that for the ODE system (4.5)

$$\dot{\omega}_t^{(j)} = -\omega_t^{(j)}\left[1 - \frac{1}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i-\mu^{(j)}\right)}{\sum_{l=1}^{m}\omega_t^{(l)}\phi\left(X_i-\mu^{(l)}\right)}\right], \quad \forall\, t\ge 0,\, j\in[m]$$

with initial value $\omega_0 \in \Delta^{m-1}$, the solution exists, is unique, and $(\rho_t)_{t\ge0}$ where $\rho_t \coloneqq \sum_{l=1}^{m}\omega_t^{(l)}\delta_{\mu^{(l)}}$ is a Fisher-Rao gradient flow in the sense of (4.1).

First of all, we will use Picard-Lindelöf theorem to prove existence and uniqueness of the solution. Define a function $f:\mathbb{R}^m\to\mathbb{R}^m$ as

$$f(y) = [f_j(y)]_{1\le j\le m}, \qquad f_j(y) = -y_j\left[1 - \frac{1}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i-\mu^{(j)}\right)}{\sum_{l=1}^{m}y_l\phi\left(X_i-\mu^{(l)}\right)}\right].$$

Then we can rewrite the ODE system as $\dot{\omega}_t = f(\omega_t)$. For any $y\in\mathbb{R}^m$ such that $\|y-\omega_0\|_2 \le \varepsilon$ for some $\delta>0$ to be specified later, we know that

$$\begin{aligned}
\sum_{l=1}^{m} y_l\phi\left(X_i-\mu^{(l)}\right) &\ge \sum_{l=1}^{m}\omega_0^{(l)}\phi\left(X_i-\mu^{(l)}\right) - \sum_{l=1}^{m}\left(\omega_0^{(l)}-y_l\right)\phi\left(X_i-\mu^{(l)}\right)\\
&\overset{\text{(i)}}{\ge} \min_{i\in[N],l\in[m]}\phi\left(X_i-\mu^{(l)}\right) - \|y-\omega_0\|_2\sum_{l=1}^{m}\phi^2\left(X_i-\mu^{(l)}\right)\\
&\overset{\text{(ii)}}{\ge} \min_{i\in[N],l\in[m]}\phi\left(X_i-\mu^{(l)}\right) - \frac{\varepsilon m}{(2\pi)^d}
\end{aligned}$$

for any $i \in [N]$, where (i) follows from $\omega_0^{(l)} \in \Delta^{m-1}$ and Cauchy-Schwarz inequality, while (ii) holds since $\|\phi\|_\infty \le (2\pi)^{-d/2}$. Therefore by taking

$$\varepsilon := \frac{(2\pi)^d}{2m} \min_{i\in[N], l\in[m]} \phi\left(X_i - \mu^{(l)}\right),$$

we know that for any $\|y - \omega_0\|_2 \le \varepsilon$ it holds that

$$\sum_{l=1}^m y_l \phi\left(X_i - \mu^{(l)}\right) \ge \frac{1}{2} \min_{i\in[N], l\in[m]} \phi\left(X_i - \mu^{(l)}\right) \triangleq \delta. \tag{E.1}$$

Therefore we can check that

$$\begin{aligned}\left|\left[\nabla f_j(y)\right]_j\right| &= \left| -\left[1 - \frac{1}{N}\sum_{i=1}^N \frac{\phi\left(X_i - \mu^{(j)}\right)}{\sum_{l=1}^m y_l \phi\left(X_i - \mu^{(l)}\right)}\right] - \left[\frac{1}{N}\sum_{i=1}^N \frac{y_j \phi^2\left(X_i - \mu^{(j)}\right)}{\left[\sum_{l=1}^m y_l \phi\left(X_i - \mu^{(l)}\right)\right]^2}\right]\right| \\ &\le 1 + \frac{\|\phi\|_\infty}{\delta} + \frac{(1+\varepsilon)\|\phi\|_\infty^2}{\delta^2} = 1 + \frac{1}{(2\pi)^{d/2}\delta} + \frac{1+\varepsilon}{(2\pi)^d \delta^2},\end{aligned}$$

and for $l \ne j$

$$\left|\left[\nabla f_j(y)\right]_l\right| = \left| -y_j \left[\frac{1}{N}\sum_{i=1}^N \frac{\phi\left(X_i - \mu^{(j)}\right)\phi\left(X_i - \mu^{(l)}\right)}{\left[\sum_{l=1}^m y_l \phi\left(X_i - \mu^{(l)}\right)\right]^2}\right]\right| \le \frac{(1+\varepsilon)\|\phi\|_\infty^2}{\delta^2} = \frac{1+\varepsilon}{(2\pi)^d \delta^2}.$$

As a result, we know that for any

$$\max_{y:\|y-\omega_0\|_2\le\varepsilon} \|\nabla f_j(y)\|_2 \le \sqrt{m}\left(1 + \frac{1}{(2\pi)^{d/2}\delta} + \frac{1+\varepsilon}{(2\pi)^d\delta^2}\right) \triangleq C_{\mathsf{Lip}}, \tag{E.2}$$

and hence $f(y)$ is C_{Lip}-Lipschitz continuous in $\{y : \|y - \omega_0\|_2 \le \varepsilon\}$ where $C_{\mathsf{Lip}} := \sqrt{m} C_{\mathsf{lip}}$. In addition, it is easy to show that

$$\max_{y:\|y-\omega_0\|_2\le\varepsilon} \|f_j(y)\|_2 \le -y_j + \frac{1}{N}\sum_{i=1}^N \frac{y_j \phi\left(X_i - \mu^{(j)}\right)}{\sum_{l=1}^m y_l \phi\left(X_i - \mu^{(l)}\right)} \le 1 + \varepsilon \triangleq M.$$

By Picard-Lindelöf theorem, there exists $t_0 > 0$ such that the ODE has a unique solution on the time interval $[0, t_0]$. We first check that $\omega_t^{(j)} > 0$ for any $j \in [m]$ and $t \in [0, t_0]$. If this is not true, define

$$t_\star := \min\left\{t \in [0, t_0] : \exists\, j \in [m] \text{ s.t. } \omega_t^{(j)} \le 0\right\}$$

and suppose $\omega_{t_\star}^{(j_\star)} \le 0$ for $j_\star \in [m]$. Then we know that $\omega_t^{(j)} > 0$ for any $j \in [m]$ and $0 \le t \le t_\star$, and hence $\dot\omega_t^{(j)} \ge -\omega_t^{(j)}$ for all $t \in [0, t_\star]$. Then we can use Grönwall's lemma to achieve $\omega_t^{(j)} \ge \omega_0^{(j)} e^{-t}$ for all $t \in [0, t_\star]$, and as a result $\omega_{t_\star}^{(j_\star)} > 0$, which is a contradiction. In addition, we can also check that

$$\frac{\mathrm{d}}{\mathrm{d}t}\sum_{j=1}^m \omega_t^{(j)} = \sum_{j=1}^m \dot\omega_t^{(j)} = -\sum_{j=1}^m \omega_t^{(j)}\left[1 - \frac{1}{N}\sum_{i=1}^N \frac{\phi\left(X_i - \mu^{(j)}\right)}{\sum_{l=1}^m \omega_t^{(l)}\phi\left(X_i - \mu^{(l)}\right)}\right] = 0$$

for all $t \in [0, t_0]$. As as a result $\omega_t \in \Delta^{m-1}$ for any $0 \le t \le t_0$. By repeating the same procedure as above (notice that the above proof only depends on $\omega_0 \in \Delta^{m-1}$, and t_0 only depends on universal constants C_{Lip} and M and does not depend on ω_0), we can show that the ODE has a unique solution on $[t_0, 2t_0]$, $[2t_0, 3t_0]$, and so on. This shows the existence and uniqueness of the solution to the ODE system (4.5).

Next, we show that $(\rho_t)_{t\ge0}$ defined as $\rho_t := \sum_{l=1}^m \omega_t^{(l)} \delta_{\mu^{(l)}}$ solves (4.1). Note that ρ_t is a probability measure since we have shown that $\omega_t \in \Delta^{m-1}$ for any $t \ge 0$. For any test function $\varphi(x) \in C_{\mathrm{c}}^\infty$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d} \varphi(x)\,\rho_t(\mathrm{d}x) = \frac{\mathrm{d}}{\mathrm{d}t}\left[\sum_{j=1}^m \omega_t^{(j)} \varphi\left(\mu_t^{(j)}\right)\right] = \sum_{j=1}^m \dot\omega_t^{(j)} \varphi\left(\mu_t^{(j)}\right)$$

$$= -\sum_{j=1}^{m}\left[1+\delta\ell_N\left(\rho_t\right)\left(\mu_t^{(j)}\right)\right]\omega_t^{(j)}\varphi\left(\mu_t^{(j)}\right)$$
$$= -\int_{\mathbb{R}^d}\left[1+\delta\ell_N\left(\rho_t\right)(x)\right]\varphi(x)\,\rho_t(\mathrm{d}x).$$

This proves that

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

holds in the sense of distributions.

E.2 Wasserstein gradient flow (Proof of Theorem 6)

In this section we will show that the ODE system (4.9)

$$\dot{\mu}_t^{(j)} = \frac{1}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i-\mu_t^{(j)}\right)}{m^{-1}\sum_{l=1}^{m}\phi\left(X_i-\mu_t^{(l)}\right)}\left(X_i-\mu_t^{(j)}\right),\quad \forall\, t\geq 0,\, j\in[m]$$

has unique solution, and $(\rho_t)_{t\geq 0}$ where $\rho_t \coloneqq m^{-1}\sum_{l=1}^{m}\delta_{\mu_t^{(l)}}$ is a Wasserstein gradient flow in the sense of (4.8).

We will use Picard-Lindelöf theorem to prove existence and uniqueness of the solution. Define a sufficiently large constant

$$R \coloneqq \max_{1\leq i\leq N}\|X_i\|_2.$$

For each $j\in[m]$, define a function $f^{(j)}:\mathbb{R}^{md}\to\mathbb{R}^d$ as

$$f^{(j)}(z) = \frac{1}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i-z_j\right)}{m^{-1}\sum_{l=1}^{m}\phi\left(X_i-z_l\right)}\left(X_i-z_j\right),$$

where $z=[z_j]_{1\leq j\leq m}\in\mathbb{R}^{md}$ and $z_1,\ldots,z_m\in\mathbb{R}^d$, and let $f(z)=[f^{(j)}(z)]_{1\leq j\leq m}$. Then we can write the ODE system as $\dot{\mu}_t=f(\mu_t)$ where $\mu_t=[\mu_t^{(j)}]_{1\leq j\leq m}$. Denote by $f^{(j)}(z)=[f_k^{(j)}(z)]_{1\leq k\leq d}$. For any $z\in\mathbb{R}^{md}$ satisfying $\max_{j\in[m]}\|z_j\|_2\leq 2R$, we have

$$\min_{1\leq i\leq N}\frac{1}{m}\sum_{l=1}^{m}\phi\left(X_i-z_l\right)\geq\frac{1}{(2\pi)^{d/2}}\exp\left(-\frac{9}{2}R^2\right)\triangleq\delta. \tag{E.3}$$

Then we can compute for $l\neq j$

$$\begin{aligned}\left\|\nabla_{z_l}f_k^{(j)}(z)\right\|_2 &= \left\|\nabla_{z_l}\frac{1}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i-z_j\right)}{m^{-1}\sum_{l=1}^{m}\phi\left(X_i-z_l\right)}e_k^\top\left(X_i-z_j\right)\right\|_2\\ &= \left\|-\frac{1}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i-z_j\right)m^{-1}\phi\left(X_i-z_l\right)}{\left[m^{-1}\sum_{l=1}^{m}\phi\left(X_i-z_l\right)\right]^2}e_k^\top\left(X_i-z_j\right)\left(X_i-z_l\right)\right\|_2\\ &\overset{\text{(i)}}{\leq}\frac{m^{-1}\|\phi\|_\infty^2}{\delta^2}\frac{1}{N}\sum_{i=1}^{N}\left|e_k^\top\left(X_i-z_j\right)\right|\|X_i-z_l\|_2\overset{\text{(ii)}}{\leq}\frac{9m^{-1}}{\delta^2(2\pi)^d}R^2,\end{aligned}$$

where (i) utilizes (E.3) and (ii) follows from $\max_{i\in[N]}\|X_i\|_2\leq R$ and $\max_{j\in[m]}\|z_j\|_2\leq 2R$. Similarly we have

$$\begin{aligned}\left\|\nabla_{z_j}f_k^{(j)}(z)\right\|_2 &= \left\|\nabla_{z_j}\frac{1}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i-z_j\right)}{m^{-1}\sum_{l=1}^{m}\phi\left(X_i-z_l\right)}e_k^\top\left(X_i-z_j\right)\right\|_2\\ &\leq\left\|-\frac{1}{N}\sum_{i=1}^{N}\frac{\phi^2\left(X_i-z_j\right)m^{-1}}{\left[m^{-1}\sum_{l=1}^{m}\phi\left(X_i-z_l\right)\right]^2}e_k^\top\left(X_i-z_j\right)\left(X_i-z_j\right)\right\|_2\end{aligned}$$

$$+\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i-z_j\right)}{m^{-1}\sum_{l=1}^{m}\phi\left(X_i-z_l\right)}\left[e_k^\top\left(X_i-z_j\right)\left(X_i-z_j\right)-e_k\right]\right\|_2$$
$$\leq 9\left(\frac{m^{-1}}{\delta^2\left(2\pi\right)^d}+\frac{1}{\delta\left(2\pi\right)^{d/2}}\right)R^2+\frac{\|\phi\|_\infty}{\delta}.$$

As a result, we have

$$\left\|\nabla_z f_k^{(j)}\left(z\right)\right\|_2=\sqrt{\sum_{l=1}^{m}\left\|\nabla_{z_l} f_k^{(j)}\left(z\right)\right\|_2^2}\leq\sqrt{\left(m+1\right)\left(\frac{9m^{-1}}{\delta^2\left(2\pi\right)^d}R^2\right)^2+2\frac{1}{\delta^2\left(2\pi\right)^d}}\triangleq C_{\mathsf{lip}}. \tag{E.4}$$

Therefore $f^{(j)}(z)$ is $\sqrt{d}C_{\mathsf{Lip}}$-Lipschitz continuous in $\{z:\max_{j\in[m]}\|z_j\|_2\leq 2R\}$, and hence $f(z)$ is C_{Lip}-Lipschitz continous where $C_{\mathsf{Lip}}\coloneqq C_{\mathsf{lip}}\sqrt{md}$. In addition, it is straightforward to show that for any $z\in\mathbb{R}^{md}$ satisfying $\max_{j\in[m]}\|z_j\|_2\leq 2R$,

$$\left\|f^{(j)}\left(z\right)\right\|_2\leq\frac{\|\phi\|_\infty}{\delta}3R\triangleq M$$

holds for all $1\leq j\leq m$.

Recall that $\mu_0^{(1)},\ldots,\mu_0^{(m)}$ are i.i.d. sampled from $\mathsf{Uniform}(\{X_i\}_{1\leq i\leq N})$, therefore $\max_{j\in[m]}\|\mu_0^{(j)}\|_2\leq R$. Therefore we have

$$\left\{z:\|z-\mu_0\|_2\leq R\right\}\subseteq\left\{z:\max_{j\in[m]}\|z_j\|_2\leq 2R\right\},$$

where $\mu_0=[\mu_0^{(j)}]_{1\leq j\leq m}$. Hence $f(z)$ is $\sqrt{md}C_{\mathsf{Lip}}$-Lipschitz continous in $\{z:\|z-\mu_0\|_2\leq R\}$. Then we can use Picard-Lindelöf theorem to show that, there exists $t_0>0$ such that the ODE has a unique solution on the time interval $[0,t_0]$. For any $t\in[0,t_0]$ and $j\in[m]$, we can compute

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|\mu_t^{(j)}\|_2^2=2\langle\mu_t^{(j)},\dot{\mu}_t^{(j)}\rangle&=\frac{2}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i-\mu_t^{(j)}\right)}{m^{-1}\sum_{l=1}^{m}\phi\left(X_i-\mu_t^{(l)}\right)}\mu_t^{(j)\top}\left(X_i-\mu_t^{(j)}\right)\\&\leq\frac{2}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i-\mu_t^{(j)}\right)}{m^{-1}\sum_{l=1}^{m}\phi\left(X_i-\mu_t^{(l)}\right)}\left(\|X_i\|_2\|\mu_t^{(j)}\|_2-\|\mu_t^{(j)}\|_2^2\right)\\&\leq\frac{2}{N}\sum_{i=1}^{N}\frac{\phi\left(X_i-\mu_t^{(j)}\right)}{m^{-1}\sum_{l=1}^{m}\phi\left(X_i-\mu_t^{(l)}\right)}\left(R-\|\mu_t^{(j)}\|_2\right)\|\mu_t^{(j)}\|_2,\end{aligned}$$

where we use Cauchy-Schwarz inequality in the penultimate step. This shows that $\frac{\mathrm{d}}{\mathrm{d}t}\|\mu_t^{(j)}\|_2^2<0$ as long as $\|\mu_{t_0}^{(j)}\|_2>R$, and as a result $\max_{j\in[m]}\|\mu_{t_0}^{(j)}\|_2\leq R$. Then we can repeat the same analysis as above (notice that the above proof only requires $\max_{j\in[m]}\|\mu_0^{(j)}\|_2\leq R$, and t_0 only depends on universal constants C_{Lip} and M) to show that the ODE has a unique solution on $[t_0,2t_0]$, $[2t_0,3t_0]$, and so on. This shows the existence and uniqueness of the solution to the ODE system (4.5).

Finally we check that $(\rho_t)_{t\geq 0}$ defined as $\rho_t\coloneqq\sum_{l=1}^{m}m^{-1}\delta_{\mu_t^{(l)}}$ solves (4.1). For any test function $\varphi(x)\in C_{\mathrm{c}}^\infty$, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d}\varphi\left(x\right)\rho_t\left(\mathrm{d}x\right)=\frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{1}{m}\sum_{j=1}^{m}\varphi\left(\mu_t^{(j)}\right)\right]&=\frac{1}{m}\sum_{j=1}^{m}\left\langle\nabla\varphi\left(\mu_t^{(j)}\right),\dot{\mu}_t^{(j)}\right\rangle\\&=\frac{1}{m}\sum_{j=1}^{m}\left\langle\nabla\varphi\left(\mu_t^{(j)}\right),-\nabla\delta\ell_N\left(\rho_t\right)\left(\mu_t^{(j)}\right)\right\rangle\\&=-\int_{\mathbb{R}^d}\left\langle\nabla\varphi\left(x\right),\nabla\delta\ell_N\left(\rho_t\right)\right\rangle\rho_t\left(\mathrm{d}x\right).\end{aligned}$$

This proves that

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

holds in the sense of distributions.

E.3 Wasserstein-Fisher-Rao gradient flow (Proof of Theorem 3)

In this section we will show that the ODE system (3.8)

$$\begin{aligned}
\dot{\mu}_t^{(j)} &= \frac{1}{N}\sum_{i=1}^N \frac{\phi\left(X_i - \mu_t^{(j)}\right)}{m^{-1}\sum_{l=1}^m \phi\left(X_i - \mu_t^{(l)}\right)}\left(X_i - \mu_t^{(j)}\right), \\
\dot{\omega}_t^{(j)} &= \left[\frac{1}{N}\sum_{i=1}^N \frac{\phi\left(X_i - \mu_t^{(j)}\right)}{\sum_{l=1}^m \omega_t^{(j)}\phi\left(X_i - \mu_t^{(l)}\right)} - 1\right]\omega_t^{(j)},
\end{aligned}$$

has unique solution, and $(\rho_t)_{t\geq 0}$ where $\rho_t \coloneqq \sum_{l=1}^m \omega_t^{(l)}\delta_{\mu_t^{(l)}}$ is a Wasserstein-Fisher-Rao gradient flow in the sense of (3.6). We will integrate the proof techniques used in the previous two sections.

We will again use Picard-Lindelöf theorem to prove existence and uniqueness of the solution. For each $j \in [m]$, define two functions $f^{(j)} : \Delta^{m-1} \times \mathbb{R}^{md} \to \mathbb{R}^d$ and $g^{(j)} : \Delta^{m-1} \times \mathbb{R}^{md} \to \mathbb{R}$ as

$$\begin{aligned}
f^{(j)}\left(y, z\right) &= \frac{1}{N}\sum_{i=1}^N \frac{\phi\left(X_i - z_j\right)}{\sum_{l=1}^m y_l \phi\left(X_i - z_l\right)}\left(X_i - z_j\right), \\
g^{(j)}\left(y, z\right) &= -\left[1 - \frac{1}{N}\sum_{i=1}^N \frac{\phi\left(X_i - z_j\right)}{\sum_{l=1}^m y_l \phi\left(X_i - z_l\right)}\right] y_j,
\end{aligned}$$

where $y = [y_j]_{1\leq j\leq m} \in \Delta^{m-1}$, $z = [z_j]_{1\leq j\leq m} \in \mathbb{R}^{md}$ with $z_1, \ldots, z_m \in \mathbb{R}^d$. Let

$$f\left(y, z\right) \coloneqq \begin{bmatrix} f^{(1)}\left(y, z\right) \\ \vdots \\ f^{(m)}\left(y, z\right) \end{bmatrix}, \quad g\left(y, z\right) \coloneqq \begin{bmatrix} g^{(1)}\left(y, z\right) \\ \vdots \\ g^{(m)}\left(y, z\right) \end{bmatrix}, \quad \text{and} \quad h\left(y, z\right) = \begin{bmatrix} f\left(y, z\right) \\ g\left(y, z\right) \end{bmatrix}.$$

Then we can write the ODE system as

$$\begin{bmatrix} \dot{\mu}_t \\ \dot{\omega}_t \end{bmatrix} = h\left(\begin{bmatrix} \mu_t \\ \omega_t \end{bmatrix}\right),$$

where $\mu_t = [\mu_t^{(j)}]_{1\leq j\leq m}$ and $\omega_t = [\omega_t^{(j)}]_{1\leq j\leq m}$. Denote by $f^{(j)}(z) = [f_k^{(j)}(z)]_{1\leq k\leq d}$.

For any $z \in \mathbb{R}^{md}$ satisfying $\max_{j\in[m]} \|z_j\|_2 \leq 2R$ and any $y \in \mathbb{R}^m$ satisfying $\|y - \omega_0\|_2 \leq \varepsilon$ where

$$R \coloneqq \max_{1\leq i\leq N} \|X_i\|_2, \qquad \varepsilon \coloneqq \frac{(2\pi)^{d/2}}{2m}\exp\left(-\frac{9}{2}R^2\right),$$

we have for any $i \in [N]$

$$\begin{aligned}
\sum_{l=1}^m y_l \phi\left(X_i - z_l\right) &\geq \sum_{l=1}^m \omega_0^{(l)}\phi\left(X_i - z_l\right) - \sum_{l=1}^m \left(\omega_0^{(l)} - y_l\right)\phi\left(X_i - z_l\right) \\
&\overset{\text{(i)}}{\geq} \min_{i\in[N], l\in[m]} \phi\left(X_i - z_l\right) - \|y - \omega_0\|_2 \sum_{l=1}^m \phi^2\left(X_i - \mu^{(l)}\right) \\
&\overset{\text{(ii)}}{\geq} \frac{1}{(2\pi)^{d/2}}\exp\left(-\frac{9}{2}R^2\right) - \frac{\varepsilon m}{(2\pi)^d} = \frac{1}{2\,(2\pi)^{d/2}}\exp\left(-\frac{9}{2}R^2\right) \triangleq \delta.
\end{aligned} \tag{E.5}$$

Here (i) follows from $\omega_0 \in \Delta^{m-1}$ and the Cauchy-Schwarz inequality, while (ii) and (iii) holds since $\|X_i - z_l\|_2 \leq 3R$ for any $i \in [N]$ and $l \in [m]$. For any $j \in [m]$, denote by $f^{(j)}(z) = [f_k^{(j)}(z)]_{1\leq k\leq d}$. Similar to the proof of (E.4), we can use (E.5) to show that

$$\left\|\nabla_z f_k^{(j)}(y,z)\right\|_2 \leq \sqrt{(m+1)\left(\frac{9m^{-1}}{\delta^2(2\pi)^d}R^2\right)^2 + 2\frac{1}{\delta^2(2\pi)^d}}.$$

We also have

$$\begin{aligned}\nabla_y f_k^{(j)}(y,z) &= \nabla_y \frac{1}{N}\sum_{i=1}^N \frac{\phi(X_i - z_j)}{\sum_{l=1}^m y_l \phi(X_i - z_l)} e_k^\top (X_i - z_j) \\ &= -\frac{1}{N}\sum_{i=1}^N \frac{\phi(X_i - z_j)}{\left[\sum_{l=1}^m y_l \phi(X_i - z_l)\right]^2} e_k^\top (X_i - z_j) \begin{bmatrix} \phi(X_i - z_1) \\ \vdots \\ \phi(X_i - z_m) \end{bmatrix},\end{aligned}$$

and as a result

$$\left\|\nabla_y f_k^{(j)}(y,z)\right\|_2 \leq \frac{\sqrt{m}\|\phi\|_\infty^2}{\delta^2} \max_{i\in[N], j\in[m]} \|X_i - z_j\|_2 \leq \frac{3\sqrt{m}R}{\delta^2(2\pi)^d}.$$

In addition, we can compute

$$\begin{aligned}\nabla_y g^{(j)}(y,z) &= -\nabla_y \left[\left(1 - \frac{1}{N}\sum_{i=1}^N \frac{\phi(X_i - z_j)}{\sum_{l=1}^m y_l \phi(X_i - z_l)}\right) y_j\right] \\ &= -\left[1 - \frac{1}{N}\sum_{i=1}^N \frac{\phi(X_i - z_j)}{\sum_{l=1}^m y_l \phi(X_i - z_l)}\right] e_j + \frac{1}{N}\sum_{i=1}^N \frac{y_j \phi(X_i - z_j)}{\left[\sum_{l=1}^m y_l \phi(X_i - z_l)\right]^2} \begin{bmatrix} \phi(X_i - z_1) \\ \vdots \\ \phi(X_i - z_m) \end{bmatrix}\end{aligned}$$

and for each $l \in [m]$

$$\begin{aligned}\nabla_{z_l} g^{(j)}(y,z) &= -\nabla_{z_l}\left[\left(1 - \frac{1}{N}\sum_{i=1}^N \frac{\phi(X_i - z_j)}{\sum_{l=1}^m y_l \phi(X_i - z_l)}\right) y_j\right] \\ &= y_j \frac{1}{N}\sum_{i=1}^N \left[\frac{\nabla_{z_l}\phi(X_i - z_j)}{\sum_{l=1}^m y_l \phi(X_i - z_l)} - \frac{\phi(X_i - z_j) y_l \nabla_{z_l}\phi(X_i - z_l)}{\left[\sum_{l=1}^m y_l \phi(X_i - z_l)\right]^2}\right] \\ &= \frac{1}{N}\sum_{i=1}^N \left[\mathbb{1}\{l = j\} \frac{y_j \phi(X_i - z_j)}{\sum_{l=1}^m y_l \phi(X_i - z_l)}(X_i - z_j) - \frac{y_j \phi(X_i - z_j) y_l \phi(X_i - z_l)}{\left[\sum_{l=1}^m y_l \phi(X_i - z_l)\right]^2}(X_i - z_l)\right].\end{aligned}$$

As a result, we have

$$\begin{aligned}\left\|\nabla_y g^{(j)}(y,z)\right\|_2 &\leq \left|1 - \frac{1}{N}\sum_{i=1}^N \frac{\phi(X_i - z_j)}{\sum_{l=1}^m y_l \phi(X_i - z_l)}\right| + \frac{1}{N}\sum_{i=1}^N \left|\frac{y_j \phi(X_i - z_j)}{\left[\sum_{l=1}^m y_l \phi(X_i - z_l)\right]^2}\right| \sqrt{m}\|\phi\|_\infty \\ &\leq 1 + \frac{\|\phi\|_\infty}{\delta} + \frac{(1+\varepsilon)\sqrt{m}}{\delta^2}\|\phi\|_\infty^2 = 1 + \frac{1}{(2\pi)^{d/2}\delta} + \frac{(1+\varepsilon)\sqrt{m}}{(2\pi)^d \delta^2}\end{aligned}$$

and

$$\begin{aligned}\left\|\nabla_{z_l} g^{(j)}(y,z)\right\|_2 &\leq \frac{1}{N}\sum_{i=1}^N \frac{(1+\varepsilon)\|\phi\|_\infty}{\delta}\|X_i - z_j\|_2 + \frac{1}{N}\sum_{i=1}^N \frac{(1+\varepsilon)^2\|\phi\|_\infty^2}{\delta^2}\|X_i - z_l\|_2 \\ &\leq \frac{3(1+\varepsilon)R}{(2\pi)^{d/2}\delta} + \frac{3(1+\varepsilon)^2 R\|\phi\|_\infty^2}{(2\pi)^d \delta^2}.\end{aligned}$$

Therefore for any $k \in [d]$, $j \in [m]$, $z \in \mathbb{R}^{md}$ satisfying $\max_{j\in[m]} \|z_j\|_2 \le 2R$ and $y \in \mathbb{R}^m$ such that $\|y-\omega_0\|_2 \le \varepsilon$, we have

$$\begin{aligned}\left\|\nabla f_k^{(j)}(y,z)\right\|_2 &= \sqrt{\left\|\nabla_z f_k^{(j)}(y,z)\right\|_2^2 + \left\|\nabla_y f_k^{(j)}(y,z)\right\|_2^2} \\ &\le \sqrt{(m+1)\left(\frac{9m^{-1}}{\delta^2(2\pi)^d}R^2\right)^2 + \frac{2}{\delta^2(2\pi)^d} + \left(\frac{3\sqrt{m}R}{\delta^2(2\pi)^d}\right)^2} \triangleq C_{\mathsf{lip},f},\end{aligned}$$

which suggests that $f^{(j)}(y,z)$ is $\sqrt{d}C_{\mathsf{lip},f}$-Lipschitz continuous, and

$$\begin{aligned}\left\|\nabla g^{(j)}(y,z)\right\|_2 &= \sqrt{\left\|\nabla_y g^{(j)}(y,z)\right\|_2^2 + \sum_{l=1}^m \left\|\nabla_{z_l} g^{(j)}(y,z)\right\|_2^2} \\ &\le \sqrt{\left(1+\frac{1}{(2\pi)^{d/2}\delta} + \frac{(1+\varepsilon)\sqrt{m}}{(2\pi)^d\delta^2}\right)^2 + m\left(\frac{3(1+\varepsilon)R}{(2\pi)^{d/2}\delta} + \frac{3(1+\varepsilon)^2 R\|\phi\|_\infty^2}{(2\pi)^d\delta^2}\right)^2} \triangleq C_{\mathsf{lip},g}.\end{aligned}$$

which suggests that $g^{(j)}(y,z)$ is $C_{\mathsf{lip},g}$-Lipschitz continuous. This allows us to conclude that $h(y,z)$ is C_{Lip}-continuous in $\{(y,z) : \|y-\omega_0\|_2 \le \varepsilon, \max_{j\in[m]}\|z_j\|_2 \le 2R\}$, where $C_{\mathsf{Lip}} = \sqrt{mC_{\mathsf{lip},g}^2 + mdC_{\mathsf{lip},f}^2}$. In addition, it is easy to check that for any $z \in \mathbb{R}^{md}$ satisfying $\max_{j\in[m]}\|z_j\|_2 \le 2R$ and any $y \in \mathbb{R}^m$ satisfying $\|y-\omega_0\|_2 \le \varepsilon$,

$$\max_{j\in[m]}\left\|f^{(j)}(y,z)\right\|_2 \le \frac{3R}{\delta(2\pi)^{d/2}}, \quad \max_{j\in[m]}\left|g^{(j)}(y,z)\right| \le \left(1+\frac{1}{\delta(2\pi)^{d/2}}\right)(1+\varepsilon)$$

and therefore

$$\|h(y,z)\|_2 \le \sqrt{m\left(\frac{3R}{\delta(2\pi)^{d/2}}\right)^2 + m\left[\left(1+\frac{1}{\delta(2\pi)^{d/2}}\right)(1+\varepsilon)\right]^2} \triangleq M.$$

Recall that $\mu_0^{(1)}, \ldots, \mu_0^{(m)}$ are i.i.d. sampled from $\mathsf{Uniform}(\{X_i\}_{1\le i\le N})$, therefore

$$(\omega_0,\mu_0) \in (y,z) : \left\{(y,z) : \|y-\omega_0\|_2 \le \varepsilon, \max_{j\in[m]}\|z_j\|_2 \le 2R\right\}$$

where $\mu_0 = [\mu_0^{(j)}]_{1\le j\le m}$. We are ready to apply Picard-Lindelöf theorem to show that there exists $t_0 > 0$, only depending on C_{Lip} and M, such that the ODE has a unique solution on the time interval $[0,t_0]$. We can use the same argument in the proof of Theorem 5 in Appendix E.1 to show that $\omega_t \in \Delta^{m-1}$ for all $t \in [0,t_0]$, and can use the same argument in the proof of Theorem 6 in Appendix E.2 to show that $\max_{j\in[m]}\|\mu_t^{(j)}\|_2 \le R$ for all $t \in [0,t_0]$. Then we can repeat the same analysis as above (notice that the above proof only requires $\omega_0 \in \Delta^{m-1}$ and $\max_{j\in[m]}\|\mu_0^{(j)}\|_2 \le R$, and t_0 only depends on universal constants C_{Lip} and M) to show that the ODE has a unique solution on $[t_0,2t_0]$, $[2t_0,3t_0]$, and so on. This shows the existence and uniqueness of the solution to the ODE system (3.8).

Finally we check that $(\rho_t)_{t\ge0}$ defined as $\rho_t \coloneqq \sum_{l=1}^m \omega_t^{(l)}\delta_{\mu_t^{(l)}}$ solves (4.1). Note that ρ_t is a probability measure since we have shown that $\omega_t \in \Delta^{m-1}$ for any $t \ge 0$. For any test function $\varphi(x) \in C_c^\infty$, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d}\varphi(x)\,\rho_t(\mathrm{d}x) &= \frac{\mathrm{d}}{\mathrm{d}t}\left[\sum_{j=1}^m \omega_t^{(j)}\varphi\left(\mu_t^{(j)}\right)\right] = \sum_{j=1}^m\left[\dot\omega_t^{(j)}\varphi\left(\mu_t^{(j)}\right) + \omega_t^{(j)}\left\langle\nabla\varphi\left(\mu_t^{(j)}\right), \dot\mu_t^{(j)}\right\rangle\right] \\ &= -\sum_{j=1}^m\left[1+\delta\ell_N(\rho_t)\left(\mu_t^{(j)}\right)\right]\omega_t^{(j)}\varphi\left(\mu_t^{(j)}\right) + \sum_{j=1}^m \omega_t^{(j)}\left\langle\nabla\varphi\left(\mu_t^{(j)}\right), -\nabla\delta\ell_N(\rho_t)\left(\mu_t^{(j)}\right)\right\rangle\end{aligned}$$

$$= -\int_{\mathbb{R}^d} \left[1 + \delta \ell_N\left(\rho_t\right)(x)\right] \varphi(x)\, \rho_t(\mathrm{d}x) - \int_{\mathbb{R}^d} \left\langle \nabla \varphi(x), \nabla \delta \ell_N\left(\rho_t\right)\right\rangle \rho_t(\mathrm{d}x).$$

This proves that

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

holds in the sense of distributions.

F Properties of Wasserstein gradient flow

In this section, we present some preliminary results on Wasserstein gradient flow for learning Gaussian mixtures. We also discuss the implications of these results, as well as the technical difficulty of obtaining more general results.

We first establish the connection between the Wasserstein gradient flow and the classical gradient flow in the Euclidean space. Suppose we fit the data $\{X_i\}_{1\le i\le N}$ using a m-component Gaussian mixture model

$$\frac{1}{m}\sum_{j=1}^{m} \mathcal{N}\left(\mu^{(j)}, I_d\right),$$

where $\{\mu^{(j)}\}_{1\le j\le m}$ is the location of the m Gaussian components. The negative likelihood function is

$$\ell_{N,m}\left(\mu^{(1)}, \ldots, \mu^{(m)}\right) \coloneqq -\frac{1}{N}\sum_{i=1}^{N} \log\left[\frac{1}{m}\sum_{j=1}^{m} \phi\left(X_i - \mu^{(j)}\right)\right]. \tag{F.1}$$

The gradient flow for minimizing (F.1), denoted by $(\mu_t)_{t\ge 0}$ where $\mu_t = [\mu_t^{(j)}]_{1\le j\le m}$, is given by the following ODE system

$$\dot{\mu}_t^{(j)} = -\nabla_{\mu^{(j)}} \ell_{N,m}\left(\mu_t^{(1)}, \ldots, \mu_t^{(m)}\right) \tag{F.2}$$

with initialization $\mu_0^{(1)}, \ldots, \mu_0^{(m)} \overset{i.i.d.}{\sim} \mathsf{Uniform}(\{X_i\}_{1\le i\le N})$. The following theorem shows that the gradient flow (F.2) captures the evolution of the location of particles in the Wasserstein gradient flow (4.8) initialized from a discrete distribution $\frac{1}{m}\sum_{l=1}^{m} \delta_{\mu_0^{(l)}}$. The proof is deferred to Appendix F.1.

Theorem 1. *Consider the Euclidean gradient flow $(\mu_t)_{t\ge 0}$ in (F.2). Then the flow $(\rho_t)_{t\ge 0}$ defined as*

$$\rho_t \coloneqq \frac{1}{m}\sum_{l=1}^{m} \delta_{\mu_t^{(l)}} \tag{F.3}$$

is the Wasserstein gradient flow, i.e. (F.3) is a distributional solution to the PDE (4.8).

Similar connection can also be established for the gradient descent algorithm for minimizing (F.1) and the particle Wasserstein gradient descent (cf. Algorithm 3), which is omitted for brevity.

Then we focus on the infinite sample limit of Wasserstein gradient flow and analyze its convergence property. The population level loss function is

$$\ell_\infty(\rho) = -\mathbb{E}_{X\sim \rho^\star * \phi}\left\{\log\left[\rho * \phi(X)\right]\right\} = \mathsf{KL}\left(\rho^\star * \phi \,\|\, \rho * \phi\right) + \mathsf{const}. \tag{F.4}$$

In Appendix C.1 we have computed that

$$\delta \ell_\infty(\rho) = -\int_{\mathbb{R}^d} \frac{\rho^\star * \phi(y)}{\rho * \phi(y)} \phi(x - y)\, \mathrm{d}y. \tag{F.5}$$

We know that the Wasserstein gradient flow $(\rho_t)_{t\ge 0}$ with respect to $\ell_\infty(\rho)$ is described by the following PDE:

$$\partial_t \rho_t = \mathsf{div}\left(\rho_t \nabla \delta \ell_\infty\left(\rho_t\right)\right) \tag{F.6}$$

with $\rho_0 = \rho^\star * \mathcal{N}(0, I_d)$, which is the data distribution when we have infinite samples. This Wasserstein gradient flow has the following particle interpretation: suppose at time $t = 0$ we initialize a particle $x_0 \sim \rho_0$ in the vector field $(v_t)_{t\geq 0}$ where $v_t = -\nabla \delta \ell_\infty(\rho_t)$, namely

$$\dot{x}_t = v_t\left(x_t\right),$$

then $x_t \sim \rho_t$, namely the marginal distribution of $(x_t)_{t\geq 0}$ evolves according to the Wasserstein gradient flow.

The following theorem shows that, when the true mixing distribution $\rho^\star$ is a singleton (we assume without loss of generality that $\rho^\star = \delta_0$), Wasserstein gradient flow converges to $\rho^\star$. The proof can be found in Appendix F.2.

Theorem 2. *Consider the Wasserstein gradient flow in (F.6) with $\rho^\star = \delta_0$. For any $\varepsilon < 1$, we have $\int_{\mathbb{R}^d} \|x\|_2^2 \rho_t(\mathrm{d}x) = O(\varepsilon)$ as long as*

$$t \geq \exp\left(2d\right) \varepsilon^{-1-\max\left\{8, \sqrt{8d}\right\}}.$$

Although Theorem 2 only focuses on the case when $\rho^\star$ is a singleton, the convergence result already provides some intuition about the behavior of Wasserstein gradient flow in more general setting. Consider a well-separated Gaussian mixture model with K components. Assume that the mixing distribution is $\rho^\star = \sum_{j=1}^K \omega_j^\star \delta_{\mu_j^\star}$, and the location of each Gaussian components, $\{\mu_j^\star\}_{1\leq j\leq K}$, are well-separated. Since the push-forward mapping $v_t = -\nabla \delta \ell_\infty(\rho_t)$ is localized (see F.5), there exists some $T > 0$ such that the Wasserstein gradient flow (F.6) initialized from $\rho^\star * \mathcal{N}(0, I_d)$ can be approximated, up to time T, by

$$\rho_t \approx \sum_{j=1}^K \omega_j^\star \rho_t^{(j)} \qquad \forall\, t \in [0, T],$$

where for each $j \in [K]$, $\rho_t^{(j)}$ is the Wasserstein gradient flow $\partial_t \rho_t^{(j)} = \mathsf{div}(\rho_t^{(j)} \nabla \delta \ell_\infty(\rho_t^{(j)}))$ with initialization $\rho_0^{(j)} = \mathcal{N}(\mu_j^\star, I_d)$. This suggests that Wasserstein gradient flow approximately converges to $\rho^\star$ since, by Theorem 2, each $\rho_t^{(j)}$ converges to $\delta_{\mu_j^\star}$. However this observation also suggests that Wasserstein gradient flow is not robust to weight mismatch. Consider initializing the Wasserstein gradient flow (F.6) with $\rho_0 = \widetilde{\rho} * \mathcal{N}(0, 1)$, where $\widetilde{\rho} = \sum_{j=1}^K \widetilde{\omega}_j \delta_{\mu_j^\star}$ is a mixing distribution with correct support $\{\mu_j^\star\}_{1\leq j\leq K}$ but wrong weights $\{\widetilde{\omega}_j\}_{1\leq j\leq K} \neq \{\omega_j^\star\}_{1\leq j\leq K}$. Then we also have

$$\rho_t \approx \sum_{j=1}^K \widetilde{\omega}_j \rho_t^{(j)} \qquad \forall\, t \in [0, T],$$

which shows that ρ_t approximately converges to $\widetilde{\rho}$ instead of $\rho^\star$ when $0 \leq t \leq T$. Note that the time length T that such approximations are valid can be arbitrarily large as long as the separation $\min_{i\neq j} \|\mu_i^\star - \mu_j^\star\|_2 \to \infty$. The above discussion suggests that using the correct initial weights are important for Wasserstein gradient flow to converge to the true mixing distribution.

We also would like to compare the convergence rate in Theorem 2 to a benchmark provided by the Bures-Wasserstein gradient flow. The Bures-Wasserstein gradient flow is defined on the space of non-degenerate Gaussian distributions on $\mathbb{R}^d$, denoted by $\mathsf{BW}(\mathbb{R}^d) = \mathbb{R}^d \times \mathbb{S}_{++}^d$ (where we identify a non-degenerate Gaussian distribution $\nu = \mathcal{N}(\mu, \Sigma)$ with $(\mu, \Sigma) \in \mathbb{R}^d \times \mathbb{S}_{++}^d$) equipped with the Wasserstein distance (3.3), which has the following closed form expression

$$d_{\mathsf{W}}^2\left(\nu_1, \nu_2\right) = \|\mu_1 - \mu_2\|_2^2 + \mathsf{tr}\left[\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2}\right)^{1/2}\right]$$

when $\nu_1 = \mathcal{N}(\mu_1, \Sigma_1)$ and $\nu_2 = \mathcal{N}(\mu_2, \Sigma_2)$ are both non-degenerate Gaussians. The Bures-Wasserstein gradient flow $(\nu_t)_{t\geq 0}$ can be viewed as the Wasserstein gradient flow $(\rho_t)_{t\geq 0}$ constrained to lie on $\mathsf{BW}(\mathbb{R}^d)$. We refer interested readers to Altschuler et al. (2021); Lambert et al. (2022) for more detailed discussion. We can see from the proof of Theorem 2 that the push-forward mapping $v_t(x)$ of Wasserstein gradient flow decays exponentially fast as $\|x\|_2 \to \infty$, this will make the Wasserstein gradient flow $(\rho_t)_{t\geq 0}$ becomes more and more heavy-tailed . However the push-forward mapping of Bures-Wasserstein gradient flow is always

linear, and the Bures-Wasserstein gradient flow $(\nu_t)_{t\geq 0}$ is always Gaussian. For example, in Appendix F.3 we can compute the push forward mapping explicitly for the two gradient flows at $t=0$:

$$v_0(x) = -\frac{1}{3}\left(\frac{4}{3}\right)^{d/2} \exp\left(-\frac{\|x\|_2^2}{6}\right) x \qquad \text{(Wasserstein)},$$

$$v_0(x) = \frac{x}{4} \qquad \text{(Bures-Wasserstein)}.$$

Therefore it is natural to expect that Bures-Wasserstein gradient flow $(\nu_t)_{t\geq 0}$ initialized from $\nu_0 = \mathcal{N}(0, I_d)$ converges faster than Wasserstein gradient flow $(\rho_t)_{t\geq 0}$ initialized from $\rho_0 = \delta_0$. In Appendix F.3 we show that the Bures-Wasserstein gradient flow $(\nu_t = \mathcal{N}(\mu_t, \Sigma_t))_{t\geq 0}$ is characterized by the following ODE:

$$\begin{aligned}\mu_t &= 0\\ \dot{\Sigma}_t &= -2\left(\Sigma_t + I_d\right)^{-1}\Sigma_t^2\left(\Sigma_t + I_d\right)^{-1}.\end{aligned}$$

We also show that Σ_t is sandwiched between

$$\frac{1}{1+2t} I \preceq \Sigma_t \preceq \frac{2}{2+t} I,$$

and as a result $\int_{\mathbb{R}^d} \|x\|_2^2 \nu_t(\mathrm{d}x) = O(d/t)$. Since Bures-Wasserstein gradient flow is not converging exponentially fast (we can see that the convergence rate is polynomial in t), we conjecture that Wasserstein gradient flow does not enjoy exponential convergence as well.

Lastly, we numerically show in Figure 1 that the loss function $\ell_\infty(\rho)$ (cf. (F.4)) is not geodesically convex (Ambrosio et al., 2008) even when $\rho^\star = \delta_0$. We can also check that Polyak-Łojasiewicz (PL) inequality

$$\forall\, \rho: \quad \|\nabla_{\mathsf{W}} \ell_\infty(\rho)\|_\rho^2 \geq C_{\mathsf{PL}}\left[\ell_\infty(\rho) - \ell_\infty(\rho^\star)\right] \quad \text{for some } C_{\mathsf{PL}} > 0$$

does not hold in general: consider $\rho^\star = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$ and $\rho = \delta_0$, then it is straightforward to check that $\nabla_{\mathsf{W}}\ell_\infty(\rho) = 0$ but $\ell_\infty(\rho) > \ell_\infty(\rho^\star)$. Therefore we cannot use standard proof technique (e.g. Ambrosio et al. (2008) when the loss function is geodesically convex, or Chewi et al. (2020) when there is a PL inequality) to show exponential convergence for the Wasserstein gradient flow (F.6).

F.1 Proof of Theorem 1

It is straightforward to compute the gradient of $\ell_{N,m}$. For any $j \in [m]$, we have

$$\begin{aligned}\nabla_{\mu^{(j)}} \ell_{N,m}\left(\mu^{(1)}, \ldots, \mu^{(m)}\right) &= -\frac{1}{N}\sum_{i=1}^N \frac{1}{\sum_{l=1}^m \phi\left(X_i - \mu^{(l)}\right)} \nabla_{\mu^{(j)}} \phi\left(X_i - \mu^{(j)}\right)\\ &= -\frac{1}{N}\sum_{i=1}^N \frac{\phi\left(X_i - \mu^{(l)}\right)}{\sum_{l=1}^m \phi\left(X_i - \mu^{(l)}\right)}\left(X_i - \mu^{(l)}\right).\end{aligned}$$

Therefore the Euclidean gradient flow (F.2) is given by

$$\dot{\mu}_t^{(j)} = -\nabla_{\mu^{(j)}} \ell_{N,m}\left(\mu_t^{(1)}, \ldots, \mu_t^{(m)}\right) = \frac{1}{N}\sum_{i=1}^N \frac{\phi\left(X_i - \mu_t^{(l)}\right)}{\sum_{l=1}^m \phi\left(X_i - \mu_t^{(l)}\right)}\left(X_i - \mu_t^{(l)}\right).$$

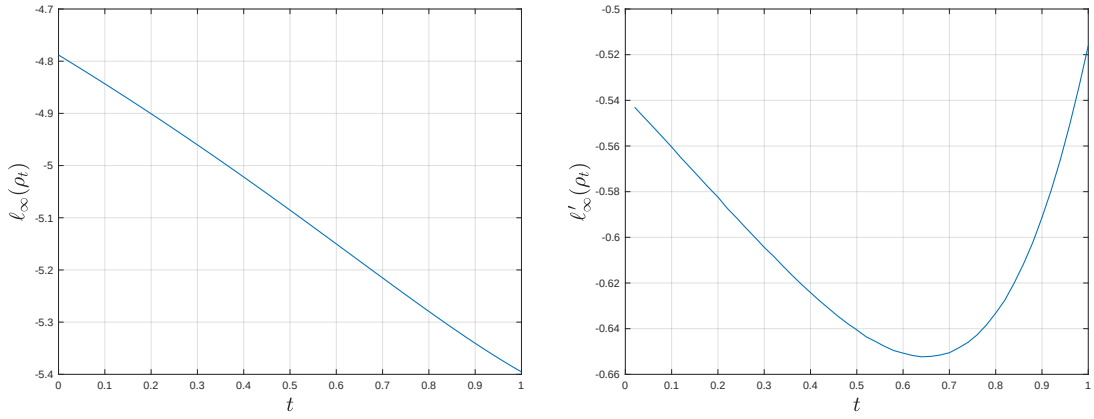
Then we can invoke Theorem 6 to finish the proof.

F.2 Proof of Theorem 2

Step 1: characterizing the push-forward mapping. First of all, it is straightforward to check that $(\rho_t)_{t\geq 0}$ is spherically symmetric for all $t \geq 0$, namely $\rho_t(\mathrm{d}x)$ only depends on $\|x\|_2$. The push-forward mapping $v_t(x): \mathbb{R}^d \to \mathbb{R}^d$ at time t is

$$v_t(x) = -\nabla \delta \ell_\infty(\rho_t)(x) = \nabla_x \int_{\mathbb{R}^d} \frac{\rho^\star * \phi(y)}{\rho_t * \phi(y)} \phi(x-y)\,\mathrm{d}y$$

(a) constant speed geodesic $(\rho_t)_{0\leq t\leq 1}$ joining $\rho_0 = \mathcal{N}(0,3)$ to $\rho_1 = \mathcal{N}(0,1)$



(b) constant speed geodesic $(\rho_t)_{0\leq t\leq 1}$ joining $\rho_0 = \mathcal{N}(0,1)$ to $\rho_1 = \delta_0$

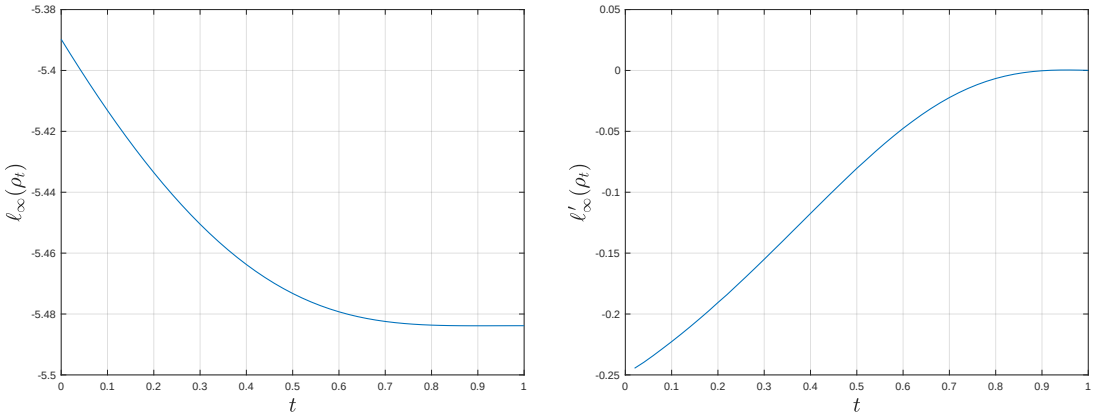


Figure 1: The loss function $\ell_\infty(\rho_t)$ or its derivative $\ell'_\infty(\rho_t)$ vs. t. In Figures (a), $(\rho_t)_{0\leq t\leq 1}$ is the constant speed geodesic joining $\rho_0 = \mathcal{N}(0,3)$ to $\rho_1 = \mathcal{N}(0,1)$. In Figures (b), $(\rho_t)_{0\leq t\leq 1}$ is the constant speed geodesic joining $\rho_0 = \mathcal{N}(0,1)$ to $\rho_1 = \delta_0$. This shows that $\ell_\infty(\rho)$ is not globally geodesically convex, but might be locally geodesically conex around $\rho^\star = \delta_0$.

$$
\begin{aligned}
&= -\int_{\mathbb{R}^d} \frac{\rho^\star * \phi(y)}{\rho_t * \phi(y)} (x-y)\, \phi(x-y)\, \mathrm{d}y \\
&= \int_{\mathbb{R}^d} \nabla_y \frac{\rho^\star * \phi(y)}{\rho_t * \phi(y)} \phi(y-x)\, \mathrm{d}y \\
&= \int_{\mathbb{R}^d} h_t(y)\, \phi(y-x)\, \mathrm{d}y, \qquad \text{(F.7)}
\end{aligned}
$$

where the penultimate line follows from Stein's lemma or Gaussian integration by parts, and $h_t(y)$ in the last line is defined as

$$
\begin{aligned}
h_t(y) &:= \nabla \frac{\phi(y)}{\rho_t * \phi(y)} = \nabla_y \frac{1}{\int \exp\left(-\frac{1}{2}\|z\|_2^2 + y^\top z\right) \rho_t(\mathrm{d}z)} \\
&= -\frac{\int \exp\left(-\frac{1}{2}\|z\|_2^2 + y^\top z\right) z \rho_t(\mathrm{d}z)}{\left[\int \exp\left(-\frac{1}{2}\|z\|_2^2 + y^\top z\right) \rho_t(\mathrm{d}z)\right]^2} = -\phi(y) \frac{\int \phi(y-z)\, z \rho_t(\mathrm{d}z)}{\left[\int \phi(y-z)\, \rho_t(\mathrm{d}z)\right]^2} \\
&= -\phi(y) \frac{\int \phi(y-z)\, z \rho_t(\mathrm{d}z)}{[\rho_t * \phi(y)]^2} = -\frac{\phi(y)}{\rho_t * \phi(y)} \cdot \frac{\int \phi(y-z)\, z \rho_t(\mathrm{d}z)}{\rho_t * \phi(y)}. \qquad \text{(F.8)}
\end{aligned}
$$

For any $y \in \mathbb{R}^d$, we can compute

$$\begin{aligned}
\int \phi(y-z)\, z \rho_t(\mathrm{d}z) &= \int_{y^\top z>0} \phi(y-z)\, z\rho_t(\mathrm{d}z) + \int_{y^\top z<0} \phi(y-z)\, z\rho_t(\mathrm{d}z) + \int_{y^\top z=0} \phi(y-z)\, z\rho_t(\mathrm{d}z) \\
&\overset{\text{(i)}}{=} \int_{y^\top z>0} [\phi(y-z) - \phi(y+z)]\, z\rho_t(\mathrm{d}z) \\
&= \phi(y) \int_{y^\top z>0} \left[\exp\left(y^\top z\right) - \exp\left(-y^\top z\right)\right] z \exp\left(-\frac{1}{2}\|z\|_2^2\right) \rho_t(\mathrm{d}z) \\
&\overset{\text{(ii)}}{=} \phi(y) \int_{y^\top z>0} \left[\exp\left(y^\top z\right) - \exp\left(-y^\top z\right)\right] \frac{y^\top z}{\|y\|_2^2} y \exp\left(-\frac{1}{2}\|z\|_2^2\right) \rho_t(\mathrm{d}z) \\
&= \underbrace{\int_{y^\top z>0} \left[\exp\left(y^\top z\right) - \exp\left(-y^\top z\right)\right] \frac{y^\top z}{\|y\|_2^2} \exp\left(-\frac{1}{2}\|z\|_2^2\right) \rho_t(\mathrm{d}z)}_{=:a_t} \cdot \phi(y)\, y. \qquad \text{(F.9)}
\end{aligned}$$

Here (i) and (ii) both follow from the spherical symmetry of ρ_t, and it is straightforward to check that the integral in the last line does not depend on y due to the spherical symmetry of ρ_t, therefore a_t is a universal constant that is independent of y. Note that when $y^\top z > 0$, we have $\exp(y^\top z) - \exp(-y^\top z) \geq 2y^\top z$, therefore

$$a_t \geq 2 \int_{y^\top z>0} \frac{\left(y^\top z\right)^2}{\|y\|_2^2} \exp\left(-\frac{1}{2}\|z\|_2^2\right) \rho_t(\mathrm{d}z) = \int_{\mathbb{R}^d} \frac{\left(y^\top z\right)^2}{\|y\|_2^2} \exp\left(-\frac{1}{2}\|z\|_2^2\right) \rho_t(\mathrm{d}z).$$

Since a_t does not depend on y, we take $y = e_i$ for $i \in [d]$ to achieve

$$a_t \geq \int_{\mathbb{R}^d} z_i^2 \exp\left(-\frac{1}{2}\|z\|_2^2\right) \rho_t(\mathrm{d}z), \qquad \forall\, i \in [d].$$

By taking average over d, we have

$$a_t \geq \frac{1}{d} \int_{\mathbb{R}^d} \|z\|_2^2 \exp\left(-\frac{1}{2}\|z\|_2^2\right) \rho_t(\mathrm{d}z) = \frac{m_t}{d}, \qquad \text{(F.10)}$$

where we define

$$m_t \coloneqq \int \|z\|_2^2 \exp\left(-\frac{1}{2}\|z\|_2^2\right) \rho_t(\mathrm{d}z).$$

Taking (F.8), (F.9) and (F.10) collectively gives

$$h_t(y) = -a_t y \left[\frac{\phi(y)}{\rho_t * \phi(y)}\right]^2, \qquad \text{where} \qquad a_t \geq \frac{m_t}{d}.$$

Then we use (F.7) to characterize the push-forward mapping:

$$\begin{aligned}
v_t(x) &= \int_{y^\top x>0} h(y)\, \phi(y-x)\, \mathrm{d}y + \int_{y^\top x<0} h(y)\, \phi(y-x)\, \mathrm{d}y + \int_{y^\top x=0} h(y)\, \phi(y-x)\, \mathrm{d}y \\
&\overset{\text{(i)}}{=} \int_{y^\top x>0} h(y) \left[\phi(y-x) - \phi(-y-x)\right] \mathrm{d}y \\
&= -\phi(x)\, a_t \int_{y^\top x>0} y \left[\frac{\phi(y)}{\rho_t * \phi(y)}\right]^2 \left[\exp\left(y^\top x\right) - \exp\left(-y^\top x\right)\right] \exp\left(-\frac{1}{2}\|y\|_2^2\right) \mathrm{d}y \\
&\overset{\text{(ii)}}{=} -\phi(x)\, a_t \int_{y^\top x>0} \frac{x^\top y}{\|x\|_2^2} x \left[\frac{\phi(y)}{\rho_t * \phi(y)}\right]^2 \left[\exp\left(y^\top x\right) - \exp\left(-y^\top x\right)\right] \exp\left(-\frac{1}{2}\|y\|_2^2\right) \mathrm{d}y \\
&= -a_t \underbrace{\int_{y^\top x>0} \frac{x^\top y}{\|x\|_2^2} \left[\frac{\phi(y)}{\rho_t * \phi(y)}\right]^2 \left[\exp\left(y^\top x\right) - \exp\left(-y^\top x\right)\right] \exp\left(-\frac{1}{2}\|y\|_2^2\right) \mathrm{d}y}_{=:b_t} \cdot \phi(x)\, x.
\end{aligned}$$

Similar to (F.9), here (i) and (ii) both follow from the spherical symmetry of ρ_t, and the integral in the last line does not depend on x due to the spherical symmetry of ρ_t, as a result b_t is a universal constant that is independent of x. Note that when $y^\top x > 0$, we have $\exp(y^\top x) - \exp(-y^\top x) \geq 2y^\top x$, therefore

$$\begin{aligned} b_t &\geq 2\int_{y^\top x>0} \frac{x^\top y}{\|x\|_2^2}\left[\frac{\phi(y)}{\rho_t * \phi(y)}\right]^2 y^\top x \exp\left(-\frac{1}{2}\|y\|_2^2\right) \mathrm{d}y \\ &= \int_{\mathbb{R}^d} \frac{\left(x^\top y\right)^2}{\|x\|_2^2}\left[\frac{\phi(y)}{\rho_t * \phi(y)}\right]^2 \exp\left(-\frac{1}{2}\|y\|_2^2\right) \mathrm{d}y. \end{aligned}$$

Note that $\|\rho_t * \phi\|_\infty \leq \|\phi\|_\infty \leq (2\pi)^{-d/2}$, and as a result

$$b_t \geq \int \frac{\left(x^\top y\right)^2}{\|x\|_2^2} \exp\left(-\frac{3}{2}\|y\|_2^2\right) \mathrm{d}y = \frac{1}{3}\left(\frac{2\pi}{3}\right)^{d/2}$$

Therefore we have

$$v_t(x) = -a_t b_t \phi(x)\, x = -c_t \phi(x)\, x \tag{F.11}$$

where

$$c_t \coloneqq a_t b_t \geq \frac{1}{3d}\left(\frac{2\pi}{3}\right)^{d/2} m_t. \tag{F.12}$$

Step 2: showing the convergence of Wasserstein gradient flow. Recall the particle interpretation of Wasserstein gradient flow as follows: let $x_0 \sim \rho_0 = \mathcal{N}(0, I_d)$ and $\dot{x}_t = v_t(x_t)$, then for any $t \geq 0$ we have $x_t \sim \rho_t$. This allows us to compute

$$\begin{aligned} \partial_t \mathbb{E}\left[\|x_t\|_2^2\right] &= 2\mathbb{E}\left[\langle x_t, \dot{x}_t\rangle\right] = 2\mathbb{E}\left[\langle x_t, v_t(x_t)\rangle\right] \overset{\text{(i)}}{=} -2c_t \mathbb{E}\left[\|x_t\|_2^2 \phi(x_t)\right] \\ &\overset{\text{(ii)}}{\leq} -\frac{2}{3d}\left(\frac{2\pi}{3}\right)^{d/2} m_t \mathbb{E}\left[\|x_t\|_2^2 \phi(x_t)\right] \\ &\overset{\text{(iii)}}{=} -\frac{2}{3d}\left(\frac{4\pi^2}{3}\right)^{d/2} \mathbb{E}^2\left[\|x_t\|_2^2 \phi(x_t)\right], \end{aligned} \tag{F.13}$$

where (i) follows from (F.11), (ii) utilizes (F.12), and (iii) holds since

$$m_t = \int \|z\|_2^2 \exp\left(-\frac{1}{2}\|z\|_2^2\right) \rho_t(\mathrm{d}z) = (2\pi)^{d/2}\, \mathbb{E}\left[\|x_t\|_2^2 \phi(x_t)\right].$$

For any $\tau > 0$, by Cauchy-Schwarz inequality we have

$$\mathbb{E}\left[\|x_0\|_2^2 \mathbb{1}\left\{\|x_0\|_2^2 > d + \tau\right\}\right] \overset{\text{(i)}}{\leq} \left[\mathbb{E}\|x_0\|_2^4\right]^{1/2} \left[\mathbb{P}\left(\|x_0\|_2^2 > d + \tau\right)\right]^{1/2}.$$

Note that $\|x_0\|_2^2 \sim \chi^2(d)$, therefore $\mathbb{E}\|x_0\|_2^4 = \mathsf{var}(\|x_0\|_2^2) + (\mathbb{E}\|x_0\|_2^2)^2 = 2d + d^2$. In addition, by the tail probability bound for χ^2 random variables (e.g. Wainwright (2019, equation (2.18))), we have

$$\mathbb{P}\left(\|x_0\|_2^2 > d + \tau\right) \leq \exp\left(-\min\left\{\frac{\tau^2}{8d}, \frac{\tau}{8}\right\}\right).$$

Therefore we have

$$\begin{aligned} \mathbb{E}\left[\|x_0\|_2^2 \mathbb{1}\left\{\|x_0\|_2^2 > d + \tau\right\}\right] &\leq \sqrt{2d + d^2} \exp\left(-\min\left\{\frac{\tau^2}{8d}, \frac{\tau}{8}\right\}\right) \\ &\leq (d+1)\exp\left(-\min\left\{\frac{\tau^2}{8d}, \frac{\tau}{8}\right\}\right) \leq \varepsilon \end{aligned}$$

as long as we choose

$$\tau \triangleq \max\left\{8\log\frac{d+1}{\varepsilon}, \sqrt{8d\log\frac{d+1}{\varepsilon}}\right\}.$$

Since the push forward mapping $v_t(x)$ is always pointing towards zero (cf. (F.11) and (F.12)), we know that $\|x_t\|_2$ is non-increasing in t. Therefore we have

$$\begin{aligned}
\mathbb{E}\left[\|x_t\|_2^2 \phi(x_t)\right] &\geq \mathbb{E}\left[\|x_t\|_2^2 \phi(x_t) \mathbb{1}\left\{\|x_0\|_2^2 \leq d+\tau\right\}\right] \\
&\overset{\text{(i)}}{\geq} (2\pi)^{-d/2} \exp\left(-\frac{d+\tau}{2}\right) \mathbb{E}\left[\|x_t\|_2^2 \mathbb{1}\left\{\|x_0\|_2^2 \leq d+\tau\right\}\right] \\
&\geq (2\pi)^{-d/2} \exp\left(-\frac{d+\tau}{2}\right) \left(\mathbb{E}\left[\|x_t\|_2^2\right] - \mathbb{E}\left[\|x_t\|_2^2 \mathbb{1}\left\{\|x_0\|_2^2 > d+\tau\right\}\right]\right) \\
&\overset{\text{(ii)}}{\geq} (2\pi)^{-d/2} \exp\left(-\frac{d+\tau}{2}\right) \left(\mathbb{E}\left[\|x_t\|_2^2\right] - \mathbb{E}\left[\|x_0\|_2^2 \mathbb{1}\left\{\|x_0\|_2^2 > d+\tau\right\}\right]\right) \\
&\geq (2\pi)^{-d/2} \exp\left(-\frac{d+\tau}{2}\right) \left(\mathbb{E}\left[\|x_t\|_2^2\right] - \varepsilon\right), \qquad \text{(F.14)}
\end{aligned}$$

where both (i) and (ii) follows from the fact that $\|x_t\|_2$ is non-increasing. Taking (F.13) and (F.14) collectively gives

$$\begin{aligned}
\partial_t \mathbb{E}\left[\|x_t\|_2^2\right] &\leq -\frac{2}{3d}\left(\frac{4\pi^2}{3}\right)^{d/2} (2\pi)^{-d} \exp\left[-(d+\tau)\right] \left(\mathbb{E}\left[\|x_t\|_2^2\right] - \varepsilon\right)^2 \\
&= -\frac{2}{3d}\left(\frac{1}{3}\right)^{d/2} \exp\left[-(d+\tau)\right] \left(\mathbb{E}\left[\|x_t\|_2^2\right] - \varepsilon\right)^2.
\end{aligned}$$

Let $f(t) = \mathbb{E}[\|x_t\|_2^2]$, we know that $f(0) = d$ and

$$\frac{\mathrm{d}f}{\mathrm{d}t} \leq -\frac{2}{3d}\left(\frac{1}{3}\right)^{d/2} \exp\left[-(d+\tau)\right] (f-\varepsilon)^2.$$

Solving this ordinary differential inequality gives

$$\frac{1}{f(t)-\varepsilon} - \frac{1}{f(0)-\varepsilon} \geq \frac{2}{3d}\left(\frac{1}{3}\right)^{d/2} \exp\left[-(d+\tau)\right] t,$$

which is equivalent to

$$\mathbb{E}\left[\|x_t\|_2^2\right] \leq \varepsilon + \left\{\frac{2}{3d}\left(\frac{1}{3}\right)^{d/2} \exp\left(-d-\tau\right) t + \frac{1}{d-\varepsilon}\right\}^{-1}.$$

Then we immediately know that $\mathbb{E}[\|x_t\|_2^2] \leq O(\varepsilon)$ as long as

$$t \geq \exp(2d)\, \varepsilon^{-1-\max\left\{8, \sqrt{8d}\right\}}.$$

F.3 Calculation for Bures-Wasserstein gradient flow

Define $\ell(\mu, \Sigma) = \ell_\infty(\rho)$ where we parameterize $\rho = \mathcal{N}(\mu, \Sigma)$. Then we can compute

$$\begin{aligned}
\ell(\mu, \Sigma) &= -\int \log\left[(2\pi)^{-d/2} \left[\det(\Sigma + I_d)\right]^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^\top (\Sigma + I_d)^{-1} (x-\mu)\right)\right] \phi(x)\, \mathrm{d}x + \text{constant} \\
&= \frac{1}{2}\log\det(\Sigma + I_d) + \int \frac{1}{2}(x-\mu)^\top (\Sigma + I_d)^{-1} (x-\mu)\, \phi(x)\, \mathrm{d}x + \text{constant}
\end{aligned}$$

$$= \frac{1}{2}\log\det\left(\Sigma + I_d\right) + \frac{1}{2}\mathbb{E}_{x\sim\mathcal{N}(0,I)}\left[(x-\mu)^\top\left(\Sigma+I_d\right)^{-1}(x-\mu)\right] + \text{constant}$$
$$= \frac{1}{2}\log\det\left(\Sigma + I_d\right) + \frac{1}{2}\mathsf{tr}\left[\left(\Sigma+I_d\right)^{-1}\right] + \frac{1}{2}\mu^\top\left(\Sigma+I_d\right)^{-1}\mu + \text{constant}.$$

Then we can compute the Euclidean gradient of $\ell(\mu,\Sigma)$ as follows:

$$\begin{aligned}
\nabla_\mu \ell(\mu,\Sigma) &= \left(\Sigma+I_d\right)^{-1}\mu,\\
\nabla_\Sigma \ell(\mu,\Sigma) &= \frac{1}{2}\left(\Sigma+I_d\right)^{-1} - \frac{1}{2}\left(\Sigma+I_d\right)^{-2} - \frac{1}{2}\left(\Sigma+I_d\right)^{-1}\mu\mu^\top\left(\Sigma+I_d\right)^{-1}\\
&= \frac{1}{2}\left(\Sigma+I_d\right)^{-1}\left(\Sigma+I_d-I_d-\mu\mu^\top\right)\left(\Sigma+I_d\right)^{-1}\\
&= \frac{1}{2}\left(\Sigma+I_d\right)^{-1}\left(\Sigma-\mu\mu^\top\right)\left(\Sigma+I_d\right)^{-1}.
\end{aligned}$$

According to Lambert et al. (2022, Appendix B.3), when initialized from $(\mu_0,\Sigma_0) = (0, I_d)$, the Bures-Wasserstein gradient flow can be described using the following ODE:

$$\begin{aligned}
\dot\mu_t &= -\left(\Sigma_t+I_d\right)^{-1}\mu_t\\
\dot\Sigma_t &= -\Sigma_t\left(\Sigma_t+I_d\right)^{-1}\left[\Sigma_t-\mu\mu^\top\right]\left(\Sigma_t+I_d\right)^{-1} - \left(\Sigma_t+I_d\right)^{-1}\left[\Sigma_t-\mu\mu^\top\right]\left(\Sigma_t+I_d\right)^{-1}\Sigma_t
\end{aligned}$$

with initial condition $\mu_0 = 0$ and $\Sigma_0 = I_d$. It is straightforward to check that $\mu_t = 0$ for all $t \geq 0$, and the dynamic of Σ_t is governed by

$$\begin{aligned}
\dot\Sigma_t &= -\Sigma_t\left(\Sigma_t+I_d\right)^{-1}\Sigma_t\left(\Sigma_t+I_d\right)^{-1} - \left(\Sigma_t+I_d\right)^{-1}\Sigma_t\left(\Sigma_t+I_d\right)^{-1}\Sigma_t\\
&= -\Sigma_t\left(\Sigma_t+I_d\right)^{-1} + 2\left(\Sigma_t+I_d\right)^{-1}\Sigma_t\left(\Sigma_t+I_d\right)^{-1} - \left(\Sigma_t+I_d\right)^{-1}\Sigma_t\\
&= -2I_d + 2\left(\Sigma_t+I_d\right)^{-1} + 2\left(\Sigma_t+I_d\right)^{-1}\Sigma_t\left(\Sigma_t+I_d\right)^{-1}\\
&= -2\left(\Sigma_t+I_d\right)^{-1}\Sigma_t^2\left(\Sigma_t+I_d\right)^{-1}
\end{aligned}$$

with initial condition $\Sigma_0 = I_d$. We can check that the off-diagonal entries of Σ_t are always zero, and its diagonal entries are identical and evloves according to the following ODE

$$\dot\sigma_t = -2\frac{\sigma_t^2}{\left(\sigma_t+1\right)^2}$$

with initial condition $\sigma_0 = 1$. It is straightforward to check that σ_t is monotonically decreasing and is always non-negative, namely $0 \leq \sigma_t \leq 1$ always holds. Therefore we have

$$-2\sigma_t^2 \leq \dot\sigma_t \leq -\frac{1}{2}\sigma_t^2.$$

This gives

$$\frac{1}{1+2t} \leq \sigma_t \leq \frac{2}{2+t},$$

and therefore

$$\frac{1}{1+2t}I \preceq \Sigma_t \preceq \frac{2}{2+t}I,$$

which suggests that ρ_t converges to $\rho^\star$ at the speed of $O(d/t)$.

When $t = 0$, we can compute the push forward mapping of Wasserstein gradient flow explicitly, which intuitively explains why Wasserstein gradient flow does not converge exponentially fast. We first compute

$$\nabla\frac{\rho^\star * \phi(y)}{\rho_0 * \phi(y)} == \nabla\frac{\left(\det I\right)^{-d/2}\exp\left(-\frac{1}{2}\|y\|_2^2\right)}{\left(\det 2I\right)^{-d/2}\exp\left(-\frac{1}{4}\|y\|_2^2\right)} = 2^{d/2}\nabla\exp\left(-\frac{1}{4}\|y\|_2^2\right) = -2^{d/2-1}y\exp\left(-\frac{1}{4}\|y\|_2^2\right),$$

then the push forward mapping at $t = 0$ is given by $x \mapsto v_0(x)$ where

$$\begin{aligned}
v_0(x) &= \int \left(\nabla_y \frac{\rho^\star * \phi(y)}{\rho_0 * \phi(y)}\right) \phi(y-x)\,\mathrm{d}y = -2^{d/2-1} \int y \exp\left(-\frac{1}{4}\|y\|_2^2\right) \cdot \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\|x-y\|_2^2\right) \mathrm{d}y \\
&= -\frac{2^{d/2-1}}{(2\pi)^{d/2}} \int y \exp\left(-\frac{1}{2}\|x\|_2^2 + x^\top y - \frac{3}{4}\|y\|_2^2\right) \mathrm{d}y = -\frac{2^{d/2-1}}{(2\pi)^{d/2}} \int y \exp\left(-\frac{1}{6}\|x\|_2^2 - \frac{3}{4}\left\|y - \frac{2}{3}x\right\|_2^2\right) \mathrm{d}y \\
&= -2^{d/2-1}\left(\frac{2}{3}\right)^{d/2} \exp\left(-\frac{1}{6}\|x\|_2^2\right) \int \frac{1}{(2\pi)^{d/2}(2/3)^{d/2}} y \exp\left(-\frac{3}{4}\left\|y - \frac{2}{3}x\right\|_2^2\right) \mathrm{d}y \\
&= -\frac{1}{3}\left(\frac{4}{3}\right)^{d/2} \exp\left(-\frac{1}{6}\|x\|_2^2\right) x.
\end{aligned}$$

On the other hand, in view of Lambert et al. (2022, Appendix B.3), the Bures-Wasserstein gradient at time $t = 0$ is given by

$$\nabla_{\mathsf{BW}} \ell_\infty(\rho_0) = \begin{bmatrix} \nabla_\mu \ell(\mu_0, \Sigma_0) \\ 2\nabla_\Sigma \ell(\mu_0, \Sigma_0) \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{1}{4} I_d \end{bmatrix},$$

and therefore the push forward mapping at $t = 0$ is given by $x \mapsto x/4$.

References

Altschuler, J., Chewi, S., Gerber, P. R., and Stromme, A. (2021). Averaging on the bures-wasserstein manifold: dimension-free convergence of gradient descent. *Advances in Neural Information Processing Systems*, 34:22132–22145.

Ambrosio, L., Gigli, N., and Savaré, G. (2008). *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media.

Bubeck, S. (2015). *Convex optimization: algorithms and complexity*. Now Publishers Inc.

Chewi, S., Maunu, T., Rigollet, P., and Stromme, A. J. (2020). Gradient descent algorithms for bures-wasserstein barycenters. In *Conference on Learning Theory*, pages 1276–1304. PMLR.

Gallouët, T. O. and Monsaingeon, L. (2017). A jko splitting scheme for kantorovich–fisher–rao gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1100–1130.

Lambert, M., Chewi, S., Bach, F., Bonnabel, S., and Rigollet, P. (2022). Variational inference via wasserstein gradient flows. *arXiv preprint arXiv:2205.15902*.

Lu, Y., Lu, J., and Nolen, J. (2019). Accelerating langevin sampling with birth-death. *arXiv preprint arXiv:1905.09863*.

Salim, A., Korba, A., and Luise, G. (2020). The wasserstein proximal gradient algorithm. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12356–12366. Curran Associates, Inc.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.