Introduction to First-order Optimization Methods in Wasserstein Space

Clément Bonet

SGM
16/04/2026

Program SGM

(Wasserstein) Gradient Flows

1. Introduction optimization on $\mathcal{P}_2(\mathbb{R}^d)$
2. **Anna Korba**: A Computable Measure of Suboptimality for Entropy-Regularised Variational Objectives.
3. **Pierre-Cyril Aubin**: Gradient flows with general costs

Motivations

Let $\mathcal{P}_2(\mathbb{R}^d) = \{\mu \in \mathcal{P}(\mathbb{R}^d),\ \int \|x\|_2^2 \, \mathrm{d}\mu(x) < \infty\}$, $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$.

Goal:

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{F}(\mu)$$

Many applications in Machine Learning:

- Generative modeling
- Sampling from $\nu \propto e^{-V}$ (Wibisono, 2018)
- Learning neural networks (Mei et al., 2018; Chizat and Bach, 2018)
- Modeling dynamic of population of cells (Schiebinger et al., 2019)

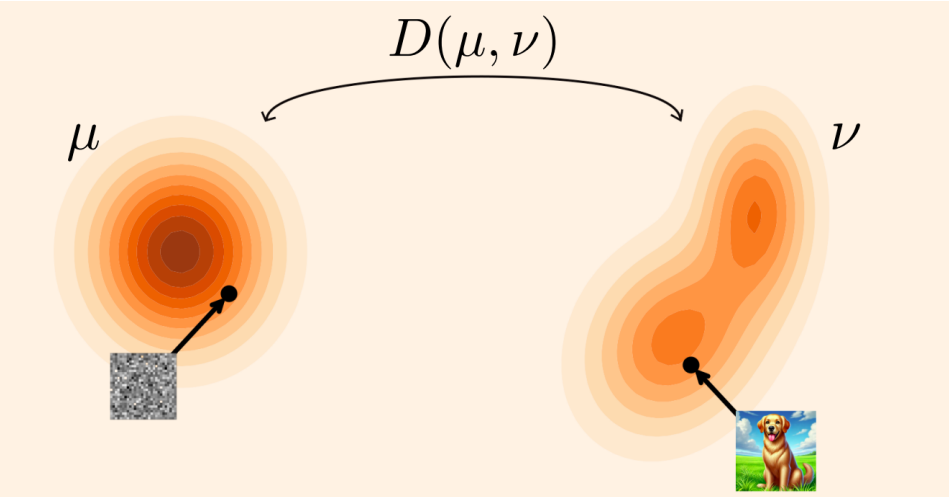
Generative Modeling

ν: unknown distribution, access to samples $y_1, \dots y_n \sim \nu$

$\rightarrow$ Minimize a distance $\mathcal{F}(\mu) = D(\mu, \nu)$

Example of divergences

- $\mathcal{F}(\mu) = \frac{1}{2}\mathrm{MMD}^2(\mu, \nu)$ (Arbel et al., 2019)
- $\mathcal{F}(\mu) = \mathrm{KL}(\nu||\mu)$
- $\mathcal{F}(\mu) = \mathrm{KL}(\varphi_\sigma * \mu||\varphi_\sigma * \nu)$ (Drifting (Deng et al., 2026; Cao et al., 2026; Turan and Ovsjanikov, 2026))



Sampling

$\nu \propto e^{-V}$ (*e.g.* in Bayesian inference)

Goal: provide samples from ν

$\rightarrow$ Minimize a distance $\mathcal{F}(\mu) = D(\mu, \nu)$ depending on V and μ

Sampling

$\nu \propto e^{-V}$ (*e.g.* in Bayesian inference)

Goal: provide samples from ν

$\rightarrow$ Minimize a distance $\mathcal{F}(\mu) = D(\mu, \nu)$ depending on V and μ

Example of divergence

$$\mathcal{F}(\mu) = \mathrm{KL}(\mu||\nu) = \int V \mathrm{d}\mu + \mathcal{H}(\mu) + \mathrm{cst},$$

where $\mathcal{H}(\mu) = \int \log\big(\mu(x)\big)\ \mathrm{d}\mu(x)$ for $\mu \ll \mathrm{Leb}$.

Methods:

- MCMC (Langevin...) (Wibisono, 2018)
- Variational Inference (Blei et al., 2017; Lambert et al., 2022)

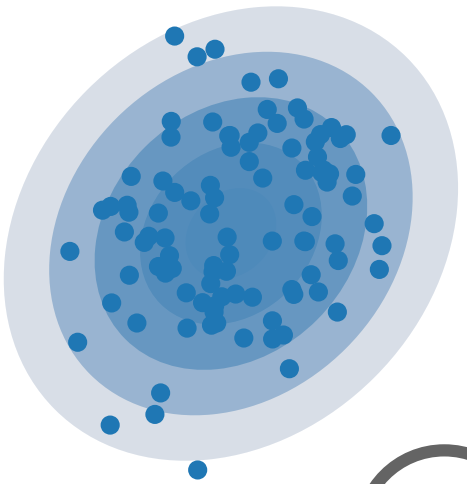


Table of Contents

Gradient Descent on $\mathbb{R}^d$

Let $f : \mathbb{R}^d \to \mathbb{R}$.

Goal: $\min_{x \in \mathbb{R}^d} \; f(x)$ via gradient flow

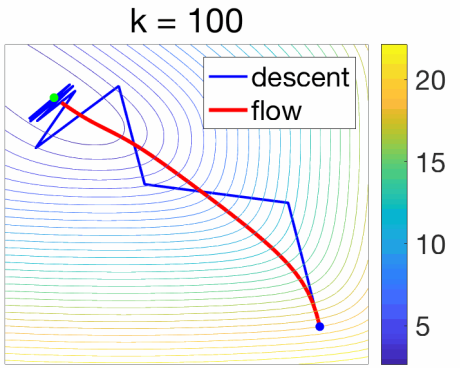
$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = -\nabla f(x_t), \quad x_0 = x_0$$

Gradient Descent on $\mathbb{R}^d$

Let $f : \mathbb{R}^d \to \mathbb{R}$.

Goal: $\min_{x \in \mathbb{R}^d} \; f(x)$ via gradient flow

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = -\nabla f(x_t), \quad x_0 = x_0$$



First algorithm: **Proximal Point**

$$\forall \tau > 0, \; \forall k \geq 0, \; x_{k+1} = x_k - \tau \nabla f(x_{k+1})$$
$$= \operatorname*{argmin}_{x \in \mathbb{R}^d} \; \frac{1}{2}\|x - x_k\|_2^2 + \tau f(x)$$

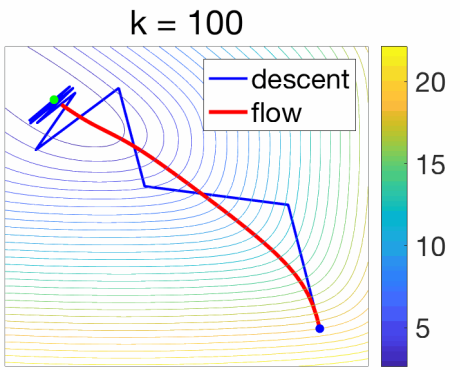
From (Bach, 2020)

Gradient Descent on $\mathbb{R}^d$

Let $f : \mathbb{R}^d \to \mathbb{R}$.

Goal: $\min_{x \in \mathbb{R}^d} \; f(x)$ via gradient flow

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = -\nabla f(x_t), \quad x_0 = x_0$$



Main algorithm: **Gradient Descent** (GD)

$$\forall \tau > 0, \; \forall k \geq 0, \; x_{k+1} = x_k - \tau \nabla f(x_k)$$

From (Bach, 2020)

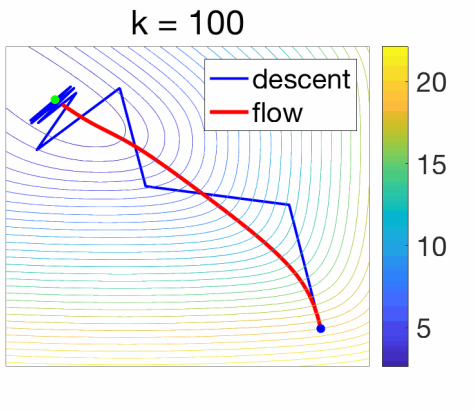
$$= \operatorname*{argmin}_{x \in \mathbb{R}^d} \; \frac{1}{2}\|x - x_k\|_2^2 + \tau \langle \nabla f(x_k), x - x_k \rangle$$

Gradient Descent on $\mathbb{R}^d$

Let $f : \mathbb{R}^d \to \mathbb{R}$.

Goal: $\min_{x \in \mathbb{R}^d} f(x)$ via gradient flow

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = -\nabla f(x_t), \quad x_0 = x_0$$



Main algorithm: **Gradient Descent** (GD)

$$\forall \tau > 0,\ \forall k \geq 0,\ x_{k+1} = x_k - \tau \nabla f(x_k)$$

From (Bach, 2020)

$$= \operatorname*{argmin}_{x \in \mathbb{R}^d} \frac{1}{2}\|x - x_k\|_2^2 + \tau \langle \nabla f(x_k), x - x_k \rangle$$

Convergence Analysis

- f β-smooth $\implies$ $f(x_{k+1}) \leq f(x_k) - \frac{1}{2\beta}\|\nabla f(x_k)\|_2^2 = f(x_k) - \frac{\beta}{2}\|x_{k+1} - x_k\|_2^2$
- f β-smooth and α-convex $\implies$ $f(x_k) - f(x^*) \leq \frac{\beta - \alpha}{2k}\|x_0 - x^*\|_2^2$

Reminder:

- f β-smooth $\iff$ $\forall x, y \in \mathbb{R}^d,\ f(x) - f(y) - \langle \nabla f(y), x - y \rangle \leq \frac{\beta}{2}\|x - y\|_2^2$
- f α-convex $\iff$ $f - \alpha \frac{\|\cdot\|_2^2}{2}$ convex

Mirror Descent on $\mathbb{R}^d$ (Beck and Teboulle, 2003)

If f not β-smooth: no guarantees for GD $\rightarrow$ change geometry

Mirror Descent on $\mathbb{R}^d$ (Beck and Teboulle, 2003)

If f not β-smooth: no guarantees for GD $\rightarrow$ change geometry

Definition (Bregman Divergence)

Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be convex, then the Bregman divergence is defined as

$$\forall x, y \in \mathbb{R}^d,\ \mathrm{d}_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y \rangle.$$

Properties:

- ϕ convex $\implies$ $\mathrm{d}_\phi(x, y) \geq 0$ for all $x, y \in \mathbb{R}^d$
- ϕ strictly convex $\implies$ "$\mathrm{d}_\phi(x, y) = 0 \iff x = y$"
- For $\phi(x) = \frac{1}{2}\|x\|_2^2$, $\mathrm{d}_\phi(x, y) = \frac{1}{2}\|x - y\|_2^2$

Mirror Descent on $\mathbb{R}^d$ (Beck and Teboulle, 2003)

If f not β-smooth: no guarantees for GD $\rightarrow$ change geometry

Definition (Bregman Divergence)

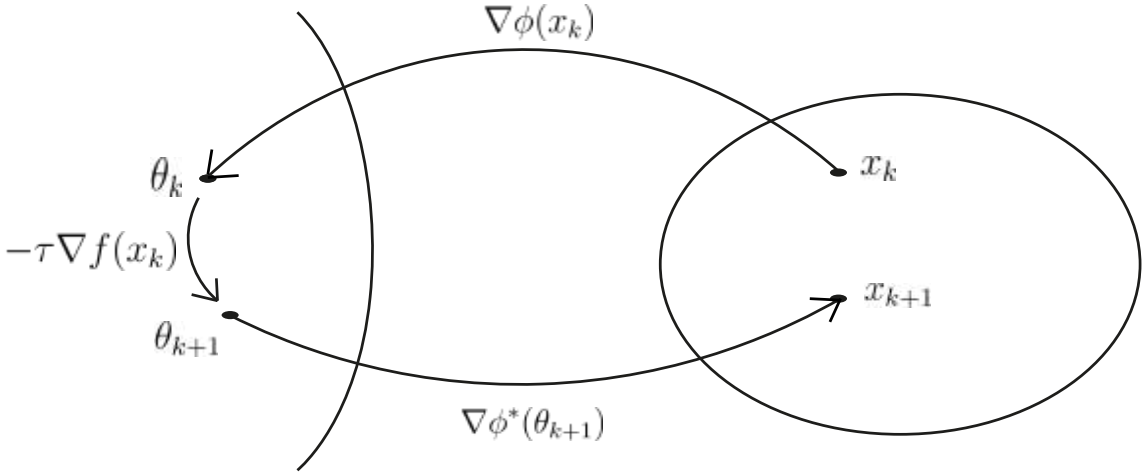
Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be convex, then the Bregman divergence is defined as

$$\forall x, y \in \mathbb{R}^d,\ \mathrm{d}_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y \rangle.$$

Mirror Descent (MD) algorithm:

$$\begin{aligned} \forall k \geq 0,\ x_{k+1} &= \operatorname*{argmin}_{x \in \mathbb{R}^d}\ \mathrm{d}_\phi(x, x_k) + \tau \langle \nabla f(x_k), x - x_k \rangle \\ &= \nabla\phi^*\big(\nabla\phi(x_k) - \tau\nabla f(x_k)\big) \end{aligned}$$

with $\phi^*(y) = \sup_x\ \langle x, y \rangle - \phi(x)$, $\nabla\phi^* = (\nabla\phi)^{-1}$.



Mirror Descent on $\mathbb{R}^d$ (Beck and Teboulle, 2003)

If f not β-smooth: no guarantees for GD $\rightarrow$ change geometry

Definition (Bregman Divergence)

Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be convex, then the Bregman divergence is defined as

$$\forall x, y \in \mathbb{R}^d,\ \mathrm{d}_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y\rangle.$$

Mirror Descent (MD) algorithm:

$$\begin{aligned}\forall k \geq 0,\ x_{k+1} &= \operatorname*{argmin}_{x\in\mathbb{R}^d}\ \mathrm{d}_\phi(x, x_k) + \tau\langle \nabla f(x_k), x - x_k\rangle \\ &= \nabla\phi^*\big(\nabla\phi(x_k) - \tau\nabla f(x_k)\big)\end{aligned}$$

Convergence analysis (Lu et al., 2018)

- f β-smooth relative to ϕ, *i.e.* $\mathrm{d}_f(x, y) \leq \beta\mathrm{d}_\phi(x, y)$ (equivalently $\beta\phi - f$ convex) $\implies$ $f(x_{k+1}) \leq f(x_k) - \beta\mathrm{d}_\phi(x_k, x_{k+1})$
- f β-smooth and α-convex relative to ϕ, *i.e.* $\alpha\mathrm{d}_\phi(x, y) \leq \mathrm{d}_f(x, y)$ (equivalently $f - \alpha\phi$ convex) $\implies$ $f(x_k) - f(x^*) \leq \frac{\beta-\alpha}{k}\mathrm{d}_\phi(x^*, x_0)$

Table of Contents

Wasserstein Geometry (Ambrosio et al., 2005)

Definition (Wasserstein distance)

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and denote by $\Pi(\mu, \nu)$ the set of coupling between μ, ν. Then, the Wasserstein distance is

$$\mathrm{W}_2^2(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int \|x - y\|_2^2 \, \mathrm{d}\gamma(x, y).$$

Wasserstein Geometry (Ambrosio et al., 2005)

Definition (Wasserstein distance)

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and denote by $\Pi(\mu, \nu)$ the set of coupling between μ, ν. Then, the Wasserstein distance is

$$\mathrm{W}_2^2(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int \|x - y\|_2^2 \, \mathrm{d}\gamma(x, y).$$

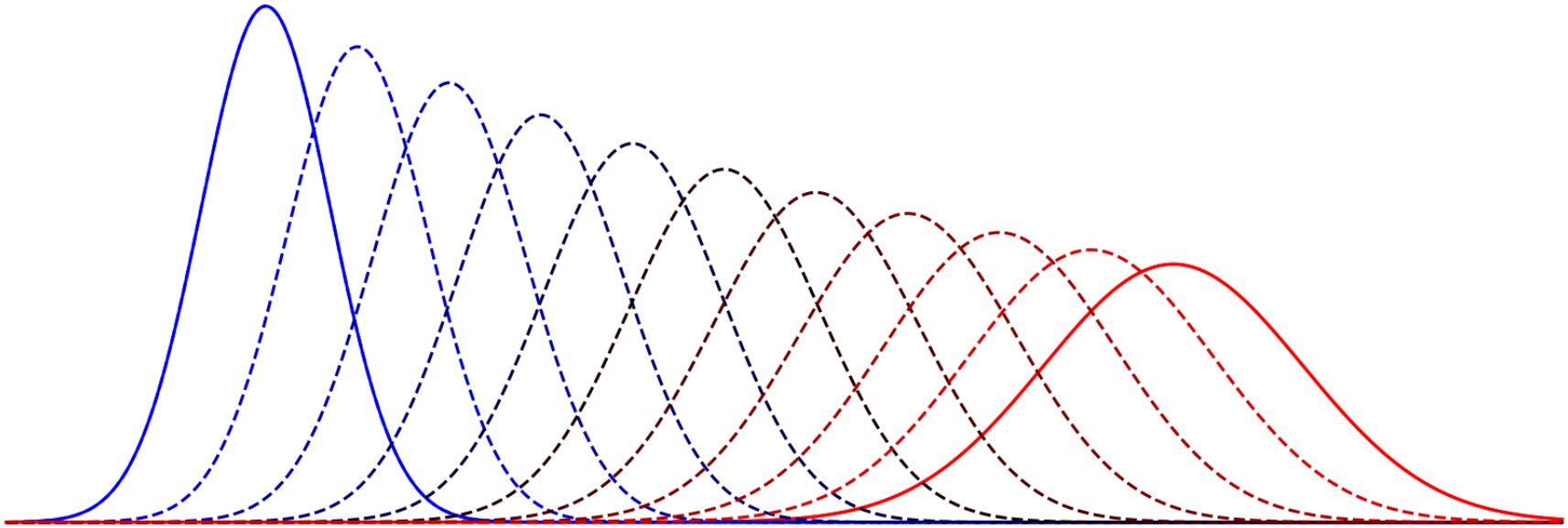
Properties:

- W_2 distance, $(\mathcal{P}_2(\mathbb{R}^d), \mathrm{W}_2)$: Wasserstein space
- $\mathrm{W}_2(\delta_x, \delta_y) = \|x - y\|_2$
- **Brenier's theorem**: If $\mu \ll \mathrm{Leb}$, then there exists a unique T_μ^ν such that
  1. $(\mathrm{T}_\mu^\nu)_{\#}\mu = \nu$, *i.e.* if $X \sim \mu$, $\mathrm{T}_\mu^\nu(X) \sim \nu$
  2. $\mathrm{W}_2^2(\mu, \nu) = \int \|x - \mathrm{T}_\mu^\nu(x)\|_2^2 \, \mathrm{d}\mu(x) = \|\mathrm{Id} - \mathrm{T}_\mu^\nu\|_{L^2(\mu)}^2$
- For $\eta \ll \mathrm{Leb}$, $\mathrm{W}_2^2(\mu, \nu) \leq \|\mathrm{T}_\eta^\mu - \mathrm{T}_\eta^\nu\|_{L^2(\eta)}^2$
- **Riemannian structure**

Riemannian Structure of the Wasserstein Space

- Geodesics between $\mu \ll \mathrm{Leb}$ and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$:

$$\forall t \in [0,1],\ \mu_t = \left((1-t)\mathrm{Id} + t\mathrm{T}_\mu^\nu\right)_\# \mu$$



Riemannian Structure of the Wasserstein Space

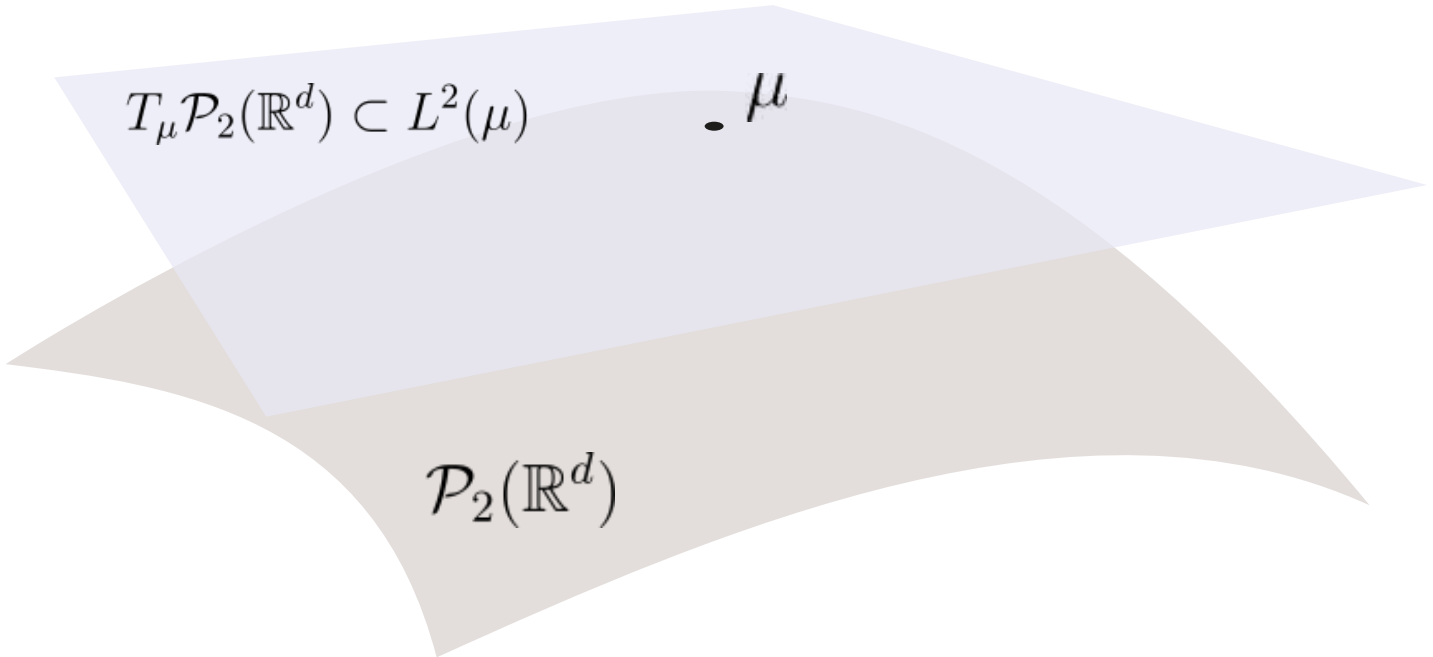
- Geodesics between $\mu \ll \mathrm{Leb}$ and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$:

$$\forall t \in [0,1],\ \mu_t = \left((1-t)\mathrm{Id} + t\mathrm{T}_\mu^\nu\right)_\# \mu$$

- Tangent space at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ (Ambrosio et al., 2005):

$$\mathcal{T}_\mu \mathcal{P}_2(\mathbb{R}^d) = \overline{\{\nabla\psi,\ \psi \in C_c^\infty(\mathbb{R}^d)\}} \subset L^2(\mu),$$

where $L^2(\mu) = \{f \in \mathbb{R}^d \to \mathbb{R}^d,\ \int \|f(x)\|_2^2 \,\mathrm{d}\mu(x) < \infty\}$.



Wasserstein Gradient (Ambrosio et al., 2005)

Definition (Wasserstein gradient (Bonnet, 2019))

Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. $\nabla_{\mathrm{W}_2}\mathcal{F}(\mu) : \mathbb{R}^d \to \mathbb{R}^d \in L^2(\mu)$ is a Wasserstein gradient of $\mathcal{F}$ at μ if for any $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ and any optimal coupling $\gamma \in \Pi_o(\mu, \nu)$,

$$\mathcal{F}(\nu) = \mathcal{F}(\mu) + \int \langle \nabla_{\mathrm{W}_2}\mathcal{F}(\mu)(x), y - x \rangle \, \mathrm{d}\gamma(x, y) + o\big(\mathrm{W}_2(\mu, \nu)\big).$$

Wasserstein Gradient (Ambrosio et al., 2005)

Definition (Wasserstein gradient (Bonnet, 2019))

Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. $\nabla_{\mathrm{W}_2}\mathcal{F}(\mu) : \mathbb{R}^d \to \mathbb{R}^d \in L^2(\mu)$ is a Wasserstein gradient of $\mathcal{F}$ at μ if for any $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ and any optimal coupling $\gamma \in \Pi_o(\mu, \nu)$,

$$\mathcal{F}(\nu) = \mathcal{F}(\mu) + \int \langle \nabla_{\mathrm{W}_2}\mathcal{F}(\mu)(x), y - x \rangle \, \mathrm{d}\gamma(x, y) + o\big(\mathrm{W}_2(\mu, \nu)\big).$$

Properties:

- There is a unique gradient in $\mathcal{T}_\mu\mathcal{P}_2(\mathbb{R}^d)$ (Lanzetti et al., 2022, Proposition 2.5)
- Differential are strong (Lanzetti et al., 2022, Proposition 2.6), *i.e.* for any $\gamma \in \Pi(\mu, \nu)$,

$$\mathcal{F}(\nu) = \mathcal{F}(\mu) + \int \langle \nabla_{\mathrm{W}_2}\mathcal{F}(\mu)(x), y - x \rangle \, \mathrm{d}\gamma(x, y) + o\left(\sqrt{\int \|x - y\|_2^2 \, \mathrm{d}\gamma(x, y)}\right).$$

In particular, for $\gamma = (\mathrm{Id}, \mathrm{T})_{\#}\mu$,

$$\mathcal{F}(\mathrm{T}_{\#}\mu) = \mathcal{F}(\mu) + \langle \nabla_{\mathrm{W}_2}\mathcal{F}(\mu), \mathrm{T} - \mathrm{Id} \rangle_{L^2(\mu)} + o(\|\mathrm{T} - \mathrm{Id}\|_{L^2(\mu)})$$

Wasserstein Gradient

Example of functionals

- Potential energies $\mathcal{V}(\mu) = \int V \mathrm{d}\mu$: For V differentiable and L-smooth,

$$\nabla_{\mathrm{W}_2} \mathcal{V}(\mu) = \nabla V$$

- Interaction energies $\mathcal{W}(\mu) = \frac{1}{2} \iint W(x-y)\ \mathrm{d}\mu(x)\mathrm{d}\mu(y)$: For W even, differentiable and L-smooth,

$$\nabla_{\mathrm{W}_2} \mathcal{W}(\mu) = \nabla W \star \mu$$

Negative entropy

$\mathcal{H}(\mu) = \int \log\left(p_\mu(x)\right)\ \mathrm{d}\mu(x)$ for $\mu \ll \mathrm{Leb}$ not W_2-differentiable but can consider subgradients under regularity assumptions:

$$\forall x \in \mathbb{R}^d,\ \nabla_{\mathrm{W}_2} \mathcal{H}(\mu)(x) = \nabla \log p_\mu(x)$$

JKO Scheme (Jordan et al., 1998)

The implicit scheme:

$$\mu_{k+1} = \underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\operatorname{argmin}} \; \frac{1}{2\tau} \mathrm{W}_2^2(\mu, \mu_k) + \mathcal{F}(\mu) = J(\mu)$$

$\rightarrow$ used in generative modeling (Fan et al., 2022; Choi et al., 2024)

$\rightarrow$ Hard to compute

JKO Scheme (Jordan et al., 1998)

The implicit scheme:

$$\mu_{k+1} = \underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\operatorname{argmin}} \ \frac{1}{2\tau} \mathrm{W}_2^2(\mu, \mu_k) + \mathcal{F}(\mu) = J(\mu)$$

$\rightarrow$ used in generative modeling (Fan et al., 2022; Choi et al., 2024)

$\rightarrow$ Hard to compute

Remark

- *If $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, equivalent to (by Brenier's theorem)*

$$\begin{cases} \mathrm{T}_{k+1} = \operatorname{argmin}_{\mathrm{T} \in L^2(\mu_k)} \ \frac{1}{2\tau} \|\mathrm{T} - \mathrm{Id}\|^2_{L^2(\mu_k)} + \mathcal{F}(\mathrm{T}_{\#}\mu_k) \\ \mu_{k+1} = (\mathrm{T}_{k+1})_{\#}\mu_k \end{cases}$$

- *For any $\mathrm{T} \in L^2(\mu_k)$,*

$$J(\mu) \leq \frac{1}{2\tau} \|\mathrm{T} - \mathrm{Id}\|^2_{L^2(\mu_k)} + \mathcal{F}(\mathrm{T}_{\#}\mu_k) = J_{\mu_k}(\mathrm{T})$$

- *If $\mathcal{F}(\mu) = D(\mu||\nu)$: source-fixed Unbalanced OT problem*

JKO Scheme (Jordan et al., 1998)

The implicit scheme:

$$\mu_{k+1} = \underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\operatorname{argmin}} \ \frac{1}{2\tau} \mathrm{W}_2^2(\mu, \mu_k) + \mathcal{F}(\mu) = J(\mu)$$

$\rightarrow$ used in generative modeling (Fan et al., 2022; Choi et al., 2024)

$\rightarrow$ Hard to compute

Remark

- *If $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, equivalent to (by Brenier's theorem)*

$$\begin{cases} \mathrm{T}_{k+1} = \operatorname{argmin}_{\mathrm{T} \in L^2(\mu_k)} \ \frac{1}{2\tau} \|\mathrm{T} - \mathrm{Id}\|^2_{L^2(\mu_k)} + \mathcal{F}(\mathrm{T}_{\#}\mu_k) \\ \mu_{k+1} = (\mathrm{T}_{k+1})_{\#}\mu_k \end{cases}$$

- *For any $\mathrm{T} \in L^2(\mu_k)$,*

$$J(\mu) \leq \frac{1}{2\tau} \|\mathrm{T} - \mathrm{Id}\|^2_{L^2(\mu_k)} + \mathcal{F}(\mathrm{T}_{\#}\mu_k) = J_{\mu_k}(\mathrm{T})$$

- *If $\mathcal{F}(\mu) = D(\mu||\nu)$: source-fixed Unbalanced OT problem*

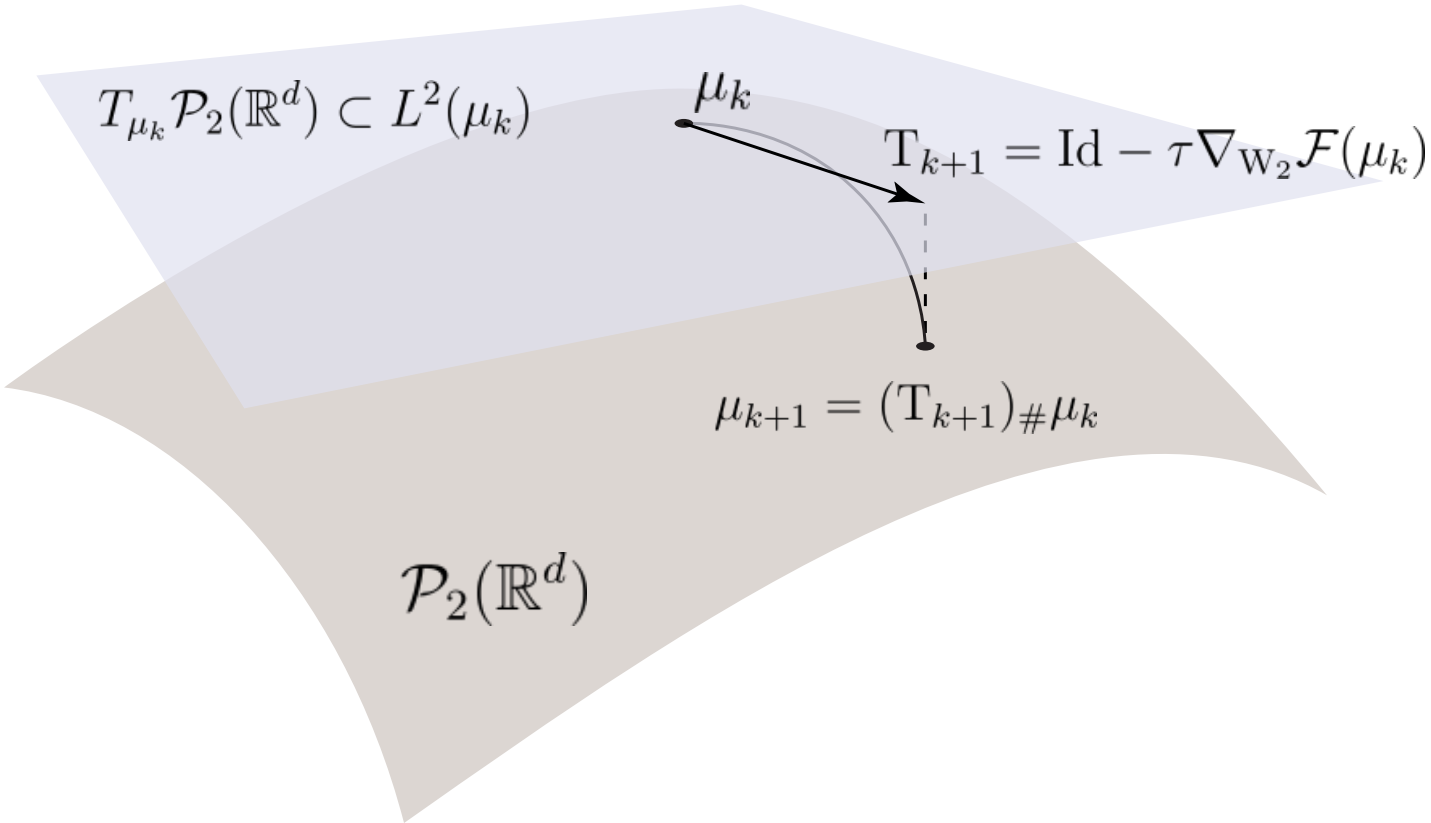
WGD: Linearize $\mathcal{F}$ around μ_k with coupling $\gamma = (\mathrm{Id}, \mathrm{T})_{\#}\mu_k$

Wasserstein Gradient Descent

Wasserstein Gradient Descent:

$$\begin{cases} \mathrm{T}_{k+1} = \mathrm{argmin}_{\mathrm{T} \in L^2(\mu_k)} \ \frac{1}{2}\|\mathrm{T} - \mathrm{Id}\|^2_{L^2(\mu_k)} + \tau \langle \nabla_{\mathrm{W}_2} \mathcal{F}(\mu_k), \mathrm{T} - \mathrm{Id} \rangle_{L^2(\mu_k)} \\ \mu_{k+1} = (\mathrm{T}_{k+1})_{\#} \mu_k \end{cases}$$

Taking the FOC: $\mathrm{T}_{k+1} = \mathrm{Id} - \tau \nabla_{\mathrm{W}_2} \mathcal{F}(\mu_k)$

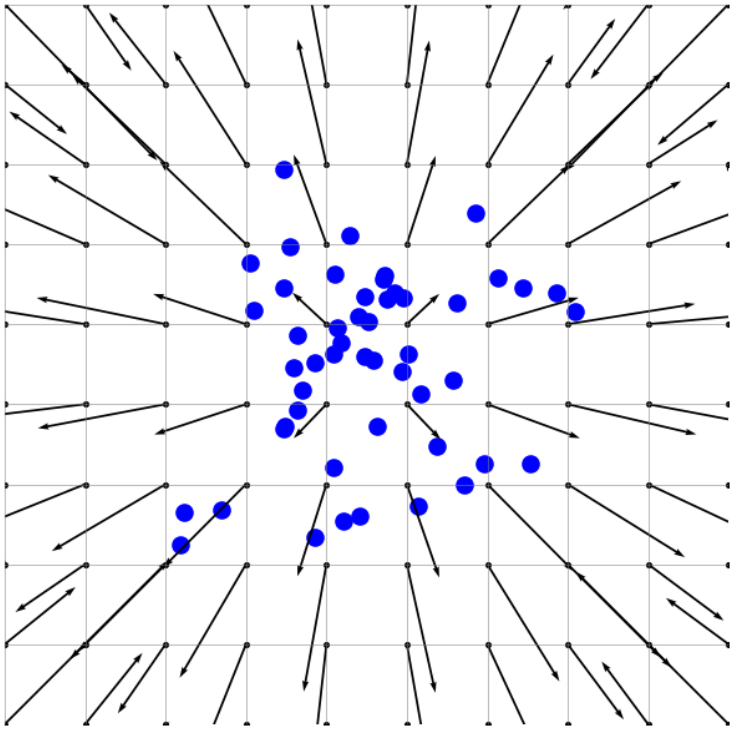


Wasserstein Gradient Descent in Practice

$$\mathcal{F}(\mu) = \frac{1}{2}\iint W(x-y)\ \mathrm{d}\mu(x)\mathrm{d}\mu(y), \qquad W(z) = \frac{\|z\|_2^4}{4} - \frac{\|z\|_2^2}{2}$$

Particle approximation:

- $\hat{\mu}_0 = \frac{1}{n}\sum_{i=1}^n \delta_{x_i^0}$ where $x_i^0 \sim \mu_0$
- At each iteration k, $\hat{\mu}_k^n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i^k}$
- Approximate $\mathrm{T}_{k+1} = \mathrm{Id} - \tau\nabla_{\mathrm{W}_2}\mathcal{F}(\hat{\mu}_k^n) = \mathrm{Id} - \int \nabla W(\cdot - y)\ \mathrm{d}\hat{\mu}_k^n(y)$
- Update particles: $\forall i \in \{1,\dots,n\},\ x_i^{k+1} = \mathrm{T}_{k+1}(x_i^k)$



Wasserstein Gradient Descent in Practice

$$\mathcal{F}(\mu) = \frac{1}{2}\iint W(x-y)\ \mathrm{d}\mu(x)\mathrm{d}\mu(y), \qquad W(z) = \frac{\|z\|_2^4}{4} - \frac{\|z\|_2^2}{2}$$

Particle approximation:

- $\hat{\mu}_0 = \frac{1}{n}\sum_{i=1}^n \delta_{x_i^0}$ where $x_i^0 \sim \mu_0$
- At each iteration k, $\hat{\mu}_k^n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i^k}$
- Approximate $\mathrm{T}_{k+1} = \mathrm{Id} - \tau\nabla_{\mathrm{W}_2}\mathcal{F}(\hat{\mu}_k^n) = \mathrm{Id} - \int \nabla W(\cdot - y)\ \mathrm{d}\hat{\mu}_k^n(y)$
- Update particles: $\forall i \in \{1,\dots,n\},\ x_i^{k+1} = \mathrm{T}_{k+1}(x_i^k)$

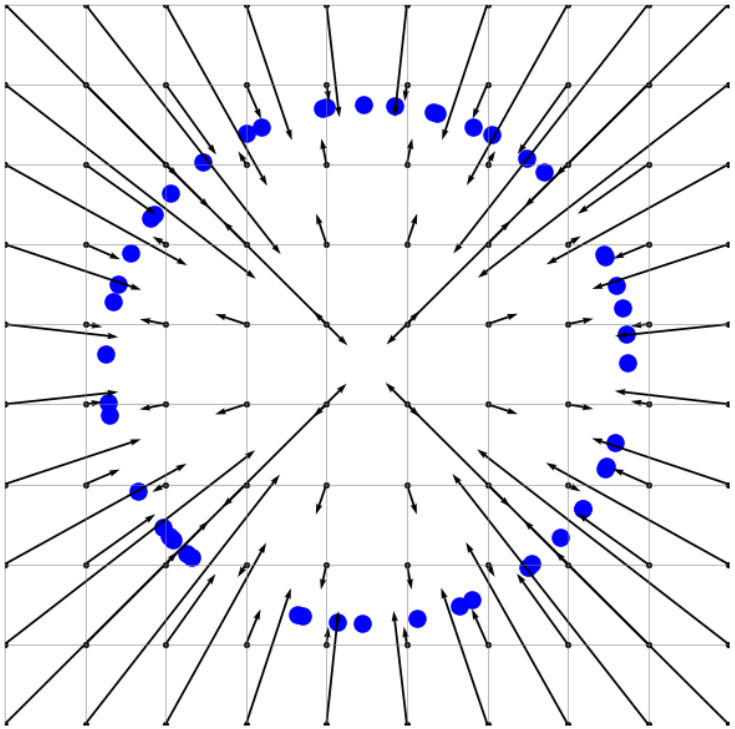


Table of Contents

Mirror Descent (Bonet et al., 2024)

Study schemes of the form

$$\begin{cases} \mathrm{T}_{k+1} = \operatorname{argmin}_{\mathrm{T} \in L^2(\mu_k)} \ \mathrm{d}(\mathrm{T}, \mathrm{Id}) + \tau \langle \nabla_{\mathrm{W}_2} \mathcal{F}(\mu_k), \mathrm{T} - \mathrm{Id} \rangle_{L^2(\mu_k)} \\ \mu_{k+1} = (\mathrm{T}_{k+1})_{\#} \mu_k, \end{cases}$$

and provide **convergence conditions**.

Considered divergences:

- For $\mathrm{d}(\mathrm{T}, \mathrm{Id}) = \frac{1}{2} \|\mathrm{T} - \mathrm{Id}\|^2_{L^2(\mu)}$: **Wasserstein gradient descent**
- For $\mathrm{d}_{\phi_\mu}(\mathrm{T}, \mathrm{Id}) = \phi_\mu(\mathrm{T}) - \phi_\mu(\mathrm{Id}) - \langle \nabla \phi_\mu(\mathrm{Id}), \mathrm{T} - \mathrm{Id} \rangle_{L^2(\mu)}$ (**Bregman divergence** on $L^2(\mu)$): extends **Mirror Descent** (Beck and Teboulle, 2003) to $\mathcal{P}_2(\mathbb{R}^d)$.
- For $\mathrm{d}(\mathrm{T}, \mathrm{Id}) = \int h\big(\mathrm{T}(x) - x\big) \, \mathrm{d}\mu(x)$: extends **Preconditioned Gradient Descent** (Maddison et al., 2021) to $\mathcal{P}_2(\mathbb{R}^d)$.

Background on $L^2(\mu)$

Definition (Bregman Divergence (Frigyik et al., 2008))

Let $\phi_\mu : L^2(\mu) \to \mathbb{R}$ be convex. The Bregman divergence is defined for all $\mathrm{T}, \mathrm{S} \in L^2(\mu)$ as

$$\mathrm{d}_{\phi_\mu}(\mathrm{T}, \mathrm{S}) = \phi_\mu(\mathrm{T}) - \phi_\mu(\mathrm{S}) - \langle \nabla\phi_\mu(\mathrm{S}), \mathrm{T} - \mathrm{S} \rangle_{L^2(\mu)}.$$

- If $\phi_\mu(\mathrm{T}) = \frac{1}{2}\|\mathrm{T}\|^2_{L^2(\mu)}$, $\mathrm{d}_{\phi_\mu}(\mathrm{T}, \mathrm{S}) = \frac{1}{2}\|\mathrm{T} - \mathrm{S}\|^2_{L^2(\mu)}$
- We call ϕ_μ **pushforward compatible** if there exists $\phi : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ such that

$$\forall \mu \in \mathcal{P}_2(\mathbb{R}^d),\ \forall \mathrm{T} \in L^2(\mu),\ \phi_\mu(\mathrm{T}) = \phi(\mathrm{T}_{\#}\mu).$$

In this case,

$$\nabla\phi_\mu(\mathrm{T}) = \nabla_{\mathrm{W}_2}\phi(\mathrm{T}_{\#}\mu) \circ \mathrm{T}$$

Mirror Descent on the Wasserstein Space

Let $\phi_\mu : L^2(\mu) \to \mathbb{R}$ be strictly convex, proper and differentiable.

Mirror Descent scheme:

$$\begin{cases} \mathrm{T}_{k+1} = \mathrm{argmin}_{\mathrm{T} \in L^2(\mu_k)} \ \mathrm{d}_{\phi_{\mu_k}}(\mathrm{T}, \mathrm{Id}) + \tau \langle \nabla_{\mathrm{W}_2} \mathcal{F}(\mu_k), \mathrm{T} - \mathrm{Id} \rangle_{L^2(\mu_k)} \\ \mu_{k+1} = (\mathrm{T}_{k+1})_{\#} \mu_k. \end{cases}$$

By FOC: $\nabla \phi_{\mu_k}(\mathrm{T}_{k+1}) = \nabla \phi_{\mu_k}(\mathrm{Id}) - \tau \nabla_{\mathrm{W}_2} \mathcal{F}(\mu_k)$

Mirror Descent on the Wasserstein Space

Let $\phi_\mu : L^2(\mu) \to \mathbb{R}$ be strictly convex, proper and differentiable.

Mirror Descent scheme:

$$\begin{cases} \mathrm{T}_{k+1} = \mathrm{argmin}_{\mathrm{T} \in L^2(\mu_k)} \ \mathrm{d}_{\phi_{\mu_k}}(\mathrm{T}, \mathrm{Id}) + \tau \langle \nabla_{\mathrm{W}_2} \mathcal{F}(\mu_k), \mathrm{T} - \mathrm{Id} \rangle_{L^2(\mu_k)} \\ \mu_{k+1} = (\mathrm{T}_{k+1})_{\#} \mu_k. \end{cases}$$

By FOC: $\nabla \phi_{\mu_k}(\mathrm{T}_{k+1}) = \nabla \phi_{\mu_k}(\mathrm{Id}) - \tau \nabla_{\mathrm{W}_2} \mathcal{F}(\mu_k)$

Computing the scheme:

- For $\phi_\mu(\mathrm{T}) = \int V \circ \mathrm{T} \, \mathrm{d}\mu$, $\mathrm{T}_{k+1} = \nabla V^* \circ \big(\nabla V - \tau \nabla_{\mathrm{W}_2} \mathcal{F}(\mu_k)\big)$
- For ϕ_μ pushforward compatible (*i.e.* $\phi_\mu(\mathrm{T}) = \phi(\mathrm{T}_{\#}\mu)$ with $\phi : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$):

$$\nabla_{\mathrm{W}_2} \phi(\mu_{k+1}) \circ \mathrm{T}_{k+1} = \nabla_{\mathrm{W}_2} \phi(\mu_k) - \tau \nabla_{\mathrm{W}_2} \mathcal{F}(\mu_k)$$

In general: implicit in $\mathrm{T}_{k+1} \to$ Newton method

Relative Convexity and Smoothness

Let $\phi_\mu, \psi_\mu : L^2(\mu) \to \mathbb{R}$ convex, $\mathcal{F}, \mathcal{G} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$.

Relative smoothness/convexity on $L^2(\mu)$

- ϕ_μ is β-smooth relative to ψ_μ if for all $\mathrm{T}, \mathrm{S} \in L^2(\mu)$, $\mathrm{d}_{\phi_\mu}(\mathrm{T}, \mathrm{S}) \leq \beta \mathrm{d}_{\psi_\mu}(\mathrm{T}, \mathrm{S})$.
- ϕ_μ is α-convex relative to ψ_μ if for all $\mathrm{T}, \mathrm{S} \in L^2(\mu)$, $\mathrm{d}_{\phi_\mu}(\mathrm{T}, \mathrm{S}) \geq \alpha \mathrm{d}_{\psi_\mu}(\mathrm{T}, \mathrm{S})$.

Relative Convexity and Smoothness

Let $\phi_\mu, \psi_\mu : L^2(\mu) \to \mathbb{R}$ convex, $\mathcal{F}, \mathcal{G} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$.

Relative smoothness/convexity on $L^2(\mu)$

- ϕ_μ is β-smooth relative to ψ_μ if for all $\mathrm{T}, \mathrm{S} \in L^2(\mu)$, $\mathrm{d}_{\phi_\mu}(\mathrm{T}, \mathrm{S}) \leq \beta \mathrm{d}_{\psi_\mu}(\mathrm{T}, \mathrm{S})$.
- ϕ_μ is α-convex relative to ψ_μ if for all $\mathrm{T}, \mathrm{S} \in L^2(\mu)$, $\mathrm{d}_{\phi_\mu}(\mathrm{T}, \mathrm{S}) \geq \alpha \mathrm{d}_{\psi_\mu}(\mathrm{T}, \mathrm{S})$.

Define $\tilde{\mathcal{F}}_\mu(\mathrm{T}) = \mathcal{F}(\mathrm{T}_\# \mu)$, $\tilde{\mathcal{G}}_\mu(\mathrm{T}) = \mathcal{G}(\mathrm{T}_\# \mu)$.

Relative smoothness/convexity on $\mathcal{P}_2(\mathbb{R}^d)$

Relative smoothness/convexity along a curve $\mu_t = (\mathrm{T}_t)_\# \mu$ with $\mathrm{T}_t = (1-t)\mathrm{S} + t\mathrm{T}$ for all $t \in [0,1]$, $\mathrm{T}, \mathrm{S} \in L^2(\mu)$.

- $\mathcal{F}$ β-smooth relative to $\mathcal{G}$ along $t \mapsto \mu_t$ if $\forall s, t \in [0,1]$,

$$\mathrm{d}_{\tilde{\mathcal{F}}_\mu}(\mathrm{T}_s, \mathrm{T}_t) \leq \beta \mathrm{d}_{\tilde{\mathcal{G}}_\mu}(\mathrm{T}_s, \mathrm{T}_t)$$

- $\mathcal{F}$ α-convex relative to $\mathcal{G}$ along $t \mapsto \mu_t$ if $\forall s, t \in [0,1]$,

$$\mathrm{d}_{\tilde{\mathcal{F}}_\mu}(\mathrm{T}_s, \mathrm{T}_t) \geq \alpha \mathrm{d}_{\tilde{\mathcal{G}}_\mu}(\mathrm{T}_s, \mathrm{T}_t)$$

Descent Lemma

Let $\phi_\mu : L^2(\mu) \to \mathbb{R}$ be strictly convex, proper and differentiable.

Mirror Descent scheme:

$$\begin{cases} \mathrm{T}_{k+1} = \operatorname{argmin}_{\mathrm{T} \in L^2(\mu_k)} \ \mathrm{d}_{\phi_{\mu_k}}(\mathrm{T}, \mathrm{Id}) + \tau \langle \nabla_{\mathrm{W}_2} \mathcal{F}(\mu_k), \mathrm{T} - \mathrm{Id} \rangle_{L^2(\mu_k)} \\ \mu_{k+1} = (\mathrm{T}_{k+1})_{\#} \mu_k. \end{cases}$$

Proposition (Descent Lemma)

Assumptions:

- *For all $k \geq 0$, $\mathcal{F}$ is β-smooth relative to ϕ along $t \mapsto \big((1-t)\mathrm{Id} + t\mathrm{T}_{k+1}\big)_{\#}\mu_k$, i.e. $\mathrm{d}_{\tilde{\mathcal{F}}_{\mu_k}}(\mathrm{T}_{k+1}, \mathrm{Id}) \leq \beta \mathrm{d}_{\phi_{\mu_k}}(\mathrm{T}_{k+1}, \mathrm{Id})$ for $\tilde{\mathcal{F}}_\mu(\mathrm{T}) = \mathcal{F}(\mathrm{T}_{\#}\mu)$.*

Then, for all $k \geq 0$,

$$\mathcal{F}(\mu_{k+1}) \leq \mathcal{F}(\mu_k) - \beta \mathrm{d}_{\phi_{\mu_k}}(\mathrm{Id}, \mathrm{T}_{k+1}).$$

Descent Lemma

Let $\phi_\mu : L^2(\mu) \to \mathbb{R}$ be strictly convex, proper and differentiable.

Mirror Descent scheme:

$$\begin{cases} \mathrm{T}_{k+1} = \mathrm{argmin}_{\mathrm{T} \in L^2(\mu_k)} \ \mathrm{d}_{\phi_{\mu_k}}(\mathrm{T}, \mathrm{Id}) + \tau \langle \nabla_{\mathrm{W}_2} \mathcal{F}(\mu_k), \mathrm{T} - \mathrm{Id} \rangle_{L^2(\mu_k)} \\ \mu_{k+1} = (\mathrm{T}_{k+1})_{\#} \mu_k. \end{cases}$$

Proposition

Assumptions: Let $\beta > 0, \alpha \geq 0$ *and* $\mathrm{T}_{\phi_{\mu_k}}^{\mu_k, \mu^*} = \mathrm{argmin}_{\mathrm{T}_{\#}\mu_k = \mu^*} \ \mathrm{d}_{\phi_{\mu_k}}(\mathrm{T}, \mathrm{Id})$.

- $\mathcal{F}$ *β-smooth relative to ϕ along* $t \mapsto \big((1-t)\mathrm{Id} + t\mathrm{T}_{k+1}\big)_{\#}\mu_k$
- $\mathcal{F}$ *α-convex relative to ϕ along* $t \mapsto \big((1-t)\mathrm{Id} + t\mathrm{T}_{\phi_{\mu_k}}^{\mu_k, \mu^*}\big)_{\#}\mu_k$
- *Assume* $\mathrm{d}_{\phi_{\mu_k}}(\mathrm{T}_{\phi_{\mu_k}}^{\mu_k, \mu^*}, \mathrm{T}_{k+1}) \geq \mathrm{d}_{\phi_{\mu_{k+1}}}(\mathrm{T}_{\phi_{\mu_{k+1}}}^{\mu_{k+1}, \mu^*}, \mathrm{Id})$

Then, for all $k \geq 1$, $\mathcal{F}(\mu_k) - \mathcal{F}(\mu^*) \leq \frac{\beta - \alpha}{k} \mathrm{d}_{\phi_{\mu_0}}(\mathrm{T}_{\phi_{\mu_0}}^{\mu_0, \mu^*}, \mathrm{Id})$.

Showing Relative Smoothness and Convexity

Smoothness and convexity of $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ relative to $\phi : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$?
$\to$ In general: look at the hessian

Showing Relative Smoothness and Convexity

Smoothness and convexity of $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ relative to $\phi : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$?
$\to$ In general: look at the hessian

Particular cases where it is simpler: For $V, U, W, K : \mathbb{R}^d \to \mathbb{R}$,

- Let $\mathcal{F}(\mu) = \int V \mathrm{d}\mu$ and $\phi(\mu) = \int U \mathrm{d}\mu$:

$$V \ \beta\text{-smooth relative to } U \implies \mathcal{F} \ \beta\text{-smooth relative to } \phi$$
$$V \ \alpha\text{-convex relative to } U \implies \mathcal{F} \ \alpha\text{-convex relative to } \phi$$

- Let $\mathcal{F}(\mu) = \iint W(x-y) \ \mathrm{d}\mu(x)\mathrm{d}\mu(y)$ and $\phi(\mu) = \iint K(x-y) \ \mathrm{d}\mu(x)\mathrm{d}\mu(y)$:

$$W \ \beta\text{-smooth relative to } K \implies \mathcal{F} \ \beta\text{-smooth relative to } \phi$$
$$W \ \alpha\text{-convex relative to } K \implies \mathcal{F} \ \alpha\text{-convex relative to } \phi$$

- For $\mathcal{F} = \mathcal{G} + \mathcal{H}$, $\mathrm{d}_{\tilde{\mathcal{F}}_\mu} = \mathrm{d}_{\tilde{\mathcal{G}}_\mu} + \mathrm{d}_{\tilde{\mathcal{H}}_\mu}$ and $\mathcal{F}$ 1-convex relative to $\mathcal{G}$ and $\mathcal{H}$

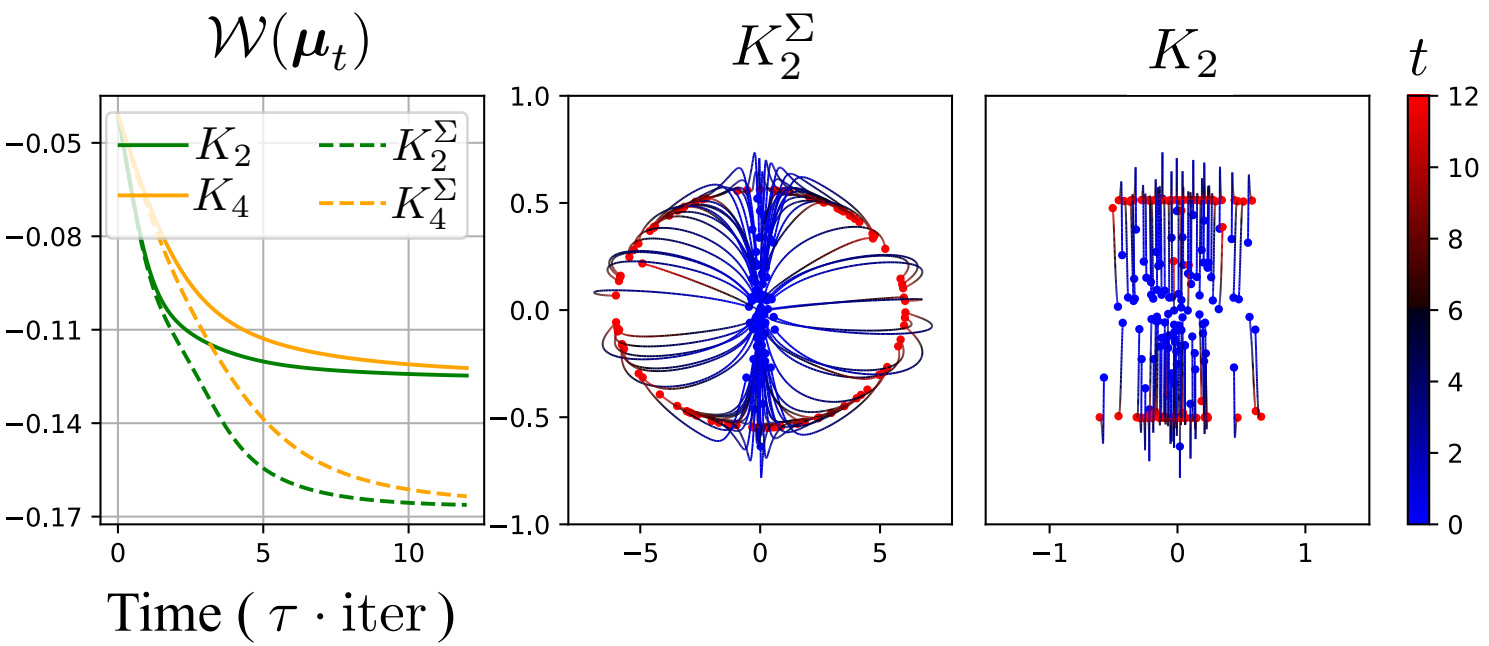
Mirror Descent on Interaction Energy

Goal: Let $\Sigma \in S_d^{++}(\mathbb{R})$ possibly ill-conditioned,

$$\min_{\mu} \ \mathcal{W}(\mu) = \iint W(x-y)\ \mathrm{d}\mu(x)\mathrm{d}\mu(y) \quad \text{with} \quad W(z) = \frac{1}{4}\|z\|_{\Sigma^{-1}}^4 - \frac{1}{2}\|z\|_{\Sigma^{-1}}^2$$

Bregman potential: $\phi_\mu(\mathrm{T}) = \iint K\big(\mathrm{T}(x) - \mathrm{T}(y)\big)\ \mathrm{d}\mu(x)\mathrm{d}\mu(y)$ with

$$K_2(z) = \frac{1}{2}\|z\|_2^2, \quad K_2^\Sigma(z) = \frac{1}{2}\|z\|_{\Sigma^{-1}}^2,$$
$$K_4(z) = \frac{1}{4}\|z\|_2^4 + \frac{1}{2}\|z\|_2^2, \quad K_4^\Sigma(z) = \frac{1}{4}\|z\|_{\Sigma^{-1}}^4 + \frac{1}{2}\|z\|_{\Sigma^{-1}}^2.$$



Mirror Descent on Gaussian

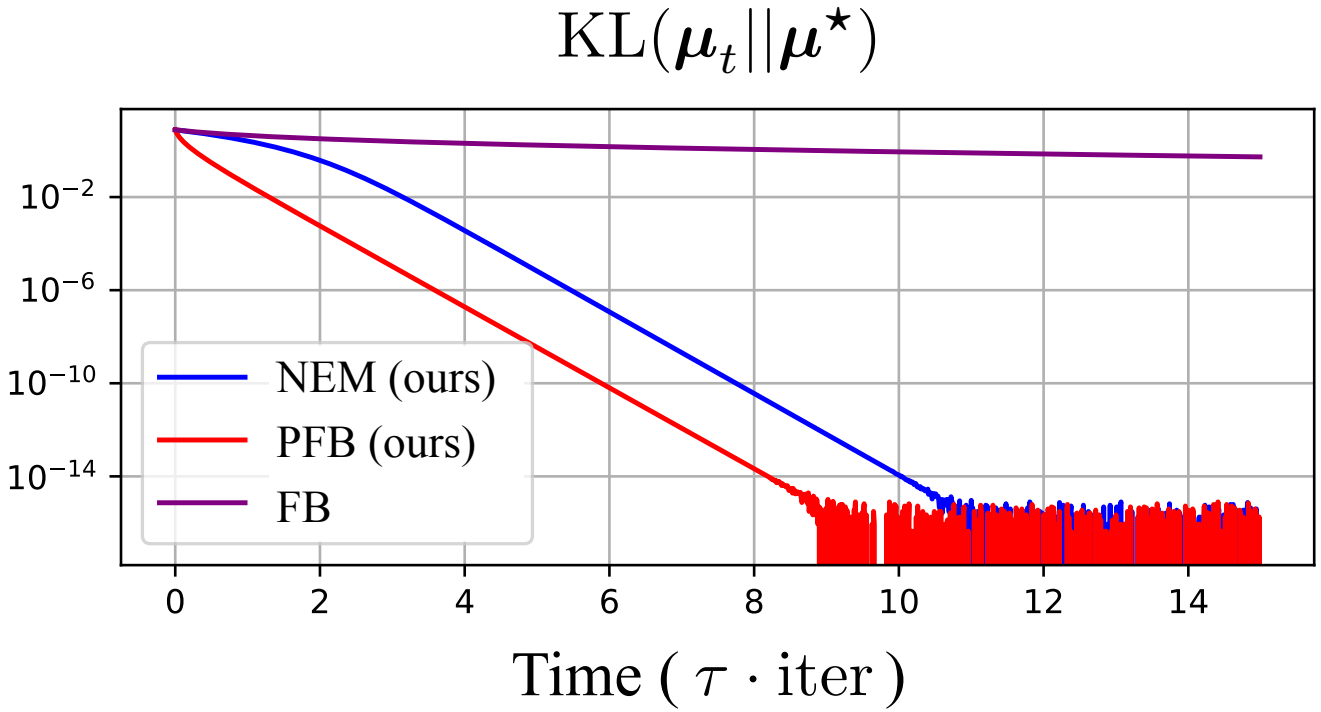
Goal:

$$\min_{\mu} \mathcal{F}(\mu) = \int V \mathrm{d}\mu + \mathcal{H}(\mu) \quad \text{with} \quad V(x) = \frac{1}{2} x^T \Sigma^{-1} x$$

$\rightarrow$ minimum $\mu^\star = \mathcal{N}(0, \Sigma)$.

Comparison between:

- Forward-Backward (FB) on the Bures-Wasserstein space (Diao et al., 2023)
- Preconditioned Forward-Backward (PFB) scheme with $\phi(\mu) = \int V \mathrm{d}\mu$
- NEM: MD with $\phi(\mu) = \mathcal{H}(\mu)$ and restriction to Gaussian



Convexity on the Wasserstein Space

Issue: Lot of functionals of interest are not convex on $\mathcal{P}_2(\mathbb{R}^d)$

Examples

- $\mathcal{F}(\mu) = \frac{1}{2}\mathrm{MMD}^2(\mu, \nu)$
- $\mathcal{F}(\mu) = \frac{1}{2}\mathrm{W}_2^2(\mu, \nu)$
- $\mathcal{F}(\mu) = \mathrm{KL}(\mu||\nu)$ for ν not log-concave

Given $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ a (positive definite) kernel,

$$\begin{aligned}\mathbb{F}(\mu) = \frac{1}{2}\mathrm{MMD}^2(\mu, \nu) &= \frac{1}{2}\iint k(x, y)\ \mathrm{d}(\mu - \nu)(x)\mathrm{d}(\mu - \nu)(y) \\ &= \frac{1}{2}\iint k(x, y)\ \mathrm{d}\mu(x)\mathrm{d}\mu(y) - \int V(x)\ \mathrm{d}\mu(x) + C,\end{aligned}$$

where $V(x) = \int k(x, y)\ \mathrm{d}\nu(y)$, $C = \frac{1}{2}\iint k(x, y)\ \mathrm{d}\nu(x)\mathrm{d}\nu(y)$.
$\to$ **not convex** (Arbel et al., 2019)

Difference-of-Convex Case

Goal: minimize

$$\mathbb{F}(\mu) = \mathcal{F}(\mu) - \mathcal{G}(\mu)$$

If $\mathcal{G}$ is convex along $t \mapsto \big((1-t)\mathrm{Id} + t\mathrm{T}\big)_{\#}\mu$,

$$\mathbb{F}(\mathrm{T}_{\#}\mu) \leq \mathcal{F}(\mathrm{T}_{\#}\mu) - \mathcal{G}(\mu) - \langle \nabla_{\mathrm{W}_2}\mathcal{G}(\mu), \mathrm{T} - \mathrm{Id}\rangle_{L^2(\mu)}$$

Wasserstein Convex Concave Procedure (CCCP)

$$\mathrm{T}_{k+1} = \operatorname*{argmin}_{\mathrm{T}\in L^2(\mu_k)} \mathcal{F}(\mathrm{T}_{\#}\mu_k) - \langle \nabla_{\mathrm{W}_2}\mathcal{G}(\mu_k), \mathrm{T} - \mathrm{Id}\rangle_{L^2(\mu_k)}, \quad \mu_{k+1} = (\mathrm{T}_{k+1})_{\#}\mu_k.$$

Difference-of-Convex Case

Goal: minimize

$$\mathbb{F}(\mu) = \mathcal{F}(\mu) - \mathcal{G}(\mu)$$

If $\mathcal{G}$ is convex along $t \mapsto \big((1-t)\mathrm{Id} + t\mathrm{T}\big)_{\#}\mu$,

$$\mathbb{F}(\mathrm{T}_{\#}\mu) \leq \mathcal{F}(\mathrm{T}_{\#}\mu) - \mathcal{G}(\mu) - \langle \nabla_{\mathrm{W}_2}\mathcal{G}(\mu), \mathrm{T} - \mathrm{Id}\rangle_{L^2(\mu)}$$

Wasserstein Convex Concave Procedure (CCCP)

$$\mathrm{T}_{k+1} = \operatorname*{argmin}_{\mathrm{T}\in L^2(\mu_k)} \mathcal{F}(\mathrm{T}_{\#}\mu_k) - \langle \nabla_{\mathrm{W}_2}\mathcal{G}(\mu_k), \mathrm{T} - \mathrm{Id}\rangle_{L^2(\mu_k)}, \quad \mu_{k+1} = (\mathrm{T}_{k+1})_{\#}\mu_k.$$

Equivalent to:

$$\begin{aligned}\mathrm{T}_{k+1} &= \operatorname*{argmin}_{\mathrm{T}\in L^2(\mu_k)} \mathrm{d}_{\tilde{\mathcal{G}}_{\mu_k}}(\mathrm{T}, \mathrm{Id}) + \mathbb{F}(\mathrm{T}_{\#}\mu_k) \\ &= \operatorname*{argmin}_{\mathrm{T}\in L^2(\mu_k)} \mathrm{d}_{\tilde{\mathcal{F}}_{\mu_k}}(\mathrm{T}, \mathrm{Id}) + \langle \nabla_{\mathrm{W}_2}\mathbb{F}(\mu_k), \mathrm{T} - \mathrm{Id}\rangle_{L^2(\mu_k)}\end{aligned}$$

DC Decomposition of MMD

DC decomposition for $k(x,y) = e^{-\|x-y\|_2^2/(2h)} = \psi(x-y)$:

$$\psi(z) = \psi_+(z) - \psi_-(z) = \cosh\big(\|z\|_2^2/(2h)\big) - \sinh\big(\|z\|_2^2/(2h)\big)$$

$$\mathbb{F}(\mu) = \mathcal{F}(\mu) - \mathcal{G}(\mu) + C,$$

$$\begin{cases} \mathcal{F}(\mu) = \frac{1}{2}\iint \psi_+(x-y)\ \mathrm{d}\mu(x)\mathrm{d}\mu(y) + \iint \psi_-(x-y)\ \mathrm{d}\nu(y)\mathrm{d}\mu(x) \\ \mathcal{G}(\mu) = \frac{1}{2}\iint \psi_-(x-y)\ \mathrm{d}\mu(x)\mathrm{d}\mu(y) + \iint \psi_+(x-y)\ \mathrm{d}\nu(y)\mathrm{d}\mu(x) \end{cases}$$

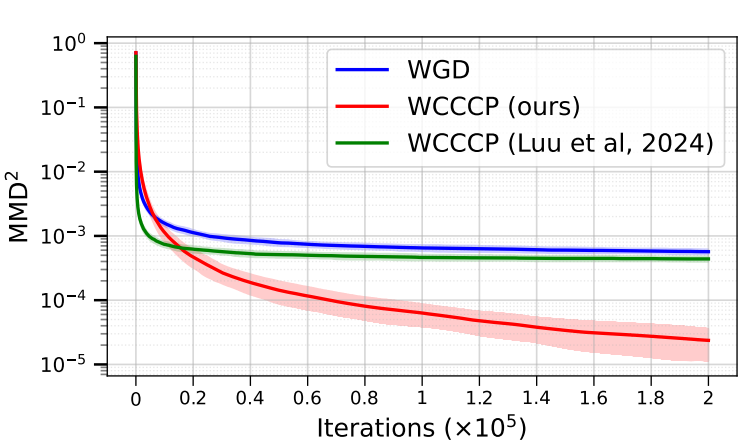
DC Decomposition of MMD

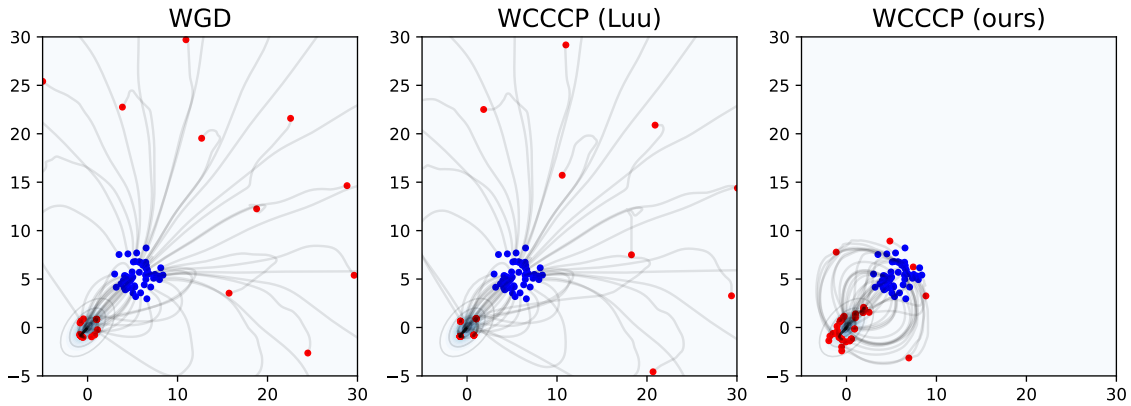
DC decomposition for $k(x,y) = e^{-\|x-y\|_2^2/(2h)} = \psi(x-y)$:

$$\psi(z) = \psi_+(z) - \psi_-(z) = \cosh\big(\|z\|_2^2/(2h)\big) - \sinh\big(\|z\|_2^2/(2h)\big)$$

$$\mathbb{F}(\mu) = \mathcal{F}(\mu) - \mathcal{G}(\mu) + C,$$

$$\begin{cases} \mathcal{F}(\mu) = \frac{1}{2}\iint \psi_+(x-y)\ \mathrm{d}\mu(x)\mathrm{d}\mu(y) + \iint \psi_-(x-y)\ \mathrm{d}\nu(y)\mathrm{d}\mu(x) \\ \mathcal{G}(\mu) = \frac{1}{2}\iint \psi_-(x-y)\ \mathrm{d}\mu(x)\mathrm{d}\mu(y) + \iint \psi_+(x-y)\ \mathrm{d}\nu(y)\mathrm{d}\mu(x) \end{cases}$$





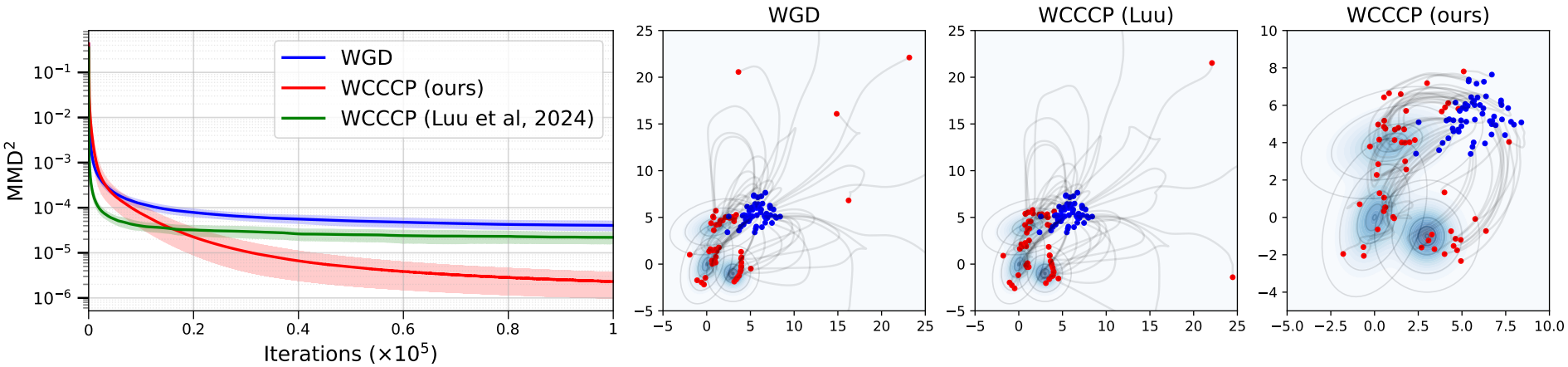
DC Decomposition of MMD

DC decomposition for $k(x,y) = e^{-\|x-y\|_2^2/(2h)} = \psi(x-y)$:

$$\psi(z) = \psi_+(z) - \psi_-(z) = \cosh\big(\|z\|_2^2/(2h)\big) - \sinh\big(\|z\|_2^2/(2h)\big)$$

$$\mathbb{F}(\mu) = \mathcal{F}(\mu) - \mathcal{G}(\mu) + C,$$

$$\begin{cases} \mathcal{F}(\mu) = \frac{1}{2}\iint \psi_+(x-y)\ \mathrm{d}\mu(x)\mathrm{d}\mu(y) + \iint \psi_-(x-y)\ \mathrm{d}\nu(y)\mathrm{d}\mu(x) \\ \mathcal{G}(\mu) = \frac{1}{2}\iint \psi_-(x-y)\ \mathrm{d}\mu(x)\mathrm{d}\mu(y) + \iint \psi_+(x-y)\ \mathrm{d}\nu(y)\mathrm{d}\mu(x) \end{cases}$$



Conclusion

Conclusion:

- Lifting optimization algorithms to $\mathcal{P}_2(\mathbb{R}^d)$
  $\rightarrow$ Mirror descent
  $\rightarrow$ Preconditioned Gradient Descent
  $\rightarrow$ Convex-Concave Procedure
- Convergence analysis of the discrete schemes

Perspectives:

- Analyze the Gaussian MD scheme (as Variational Inference)
- Adaptative DC Algorithms

Conclusion

Conclusion:

- Lifting optimization algorithms to $\mathcal{P}_2(\mathbb{R}^d)$
  $\rightarrow$ Mirror descent
  $\rightarrow$ Preconditioned Gradient Descent
  $\rightarrow$ Convex-Concave Procedure
- Convergence analysis of the discrete schemes

Perspectives:

- Analyze the Gaussian MD scheme (as Variational Inference)
- Adaptative DC Algorithms

Thank you!

References I

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: in Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2005.

Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow. *Advances in neural information processing systems*, 32, 2019.

Francis Bach. Effortless optimization through gradient flows, 2020. URL https://francisbach.com/gradient-flows/.

Amir Beck and Marc Teboulle. Mirror Descent and Nonlinear Projected Subgradient Methods for Convex Optimization. *Operations Research Letters*, 31 (3):167–175, 2003.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112 (518):859–877, 2017.

Clément Bonet, Théo Uscidda, Adam David, Pierre-Cyril Aubin-Frankowski, and Anna Korba. Mirror and Preconditioned Gradient Descent in Wasserstein Space. In *Thirty-eight Conference on Neural Information Processing Systems*, 2024.

References II

Benoît Bonnet. A Pontryagin Maximum Principle in Wasserstein Spaces for Constrained Optimal Control Problems. *ESAIM: Control, Optimisation and Calculus of Variations*, 25:52, 2019.

Jiarui Cao, Zixuan Wei, and Yuxin Liu. Gradient flow drifting: Generative modeling via wasserstein gradient flows of kde-approximated divergences. *arXiv preprint arXiv:2603.10592*, 2026.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

Jaemoo Choi, Jaewoong Choi, and Myungjoo Kang. Scalable wasserstein gradient flow for generative modeling through unbalanced optimal transport. *arXiv preprint arXiv:2402.05443*, 2024.

Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative modeling via drifting. *arXiv preprint arXiv:2602.04770*, 2026.

Michael Ziyang Diao, Krishna Balasubramanian, Sinho Chewi, and Adil Salim. Forward-backward Gaussian variational inference via JKO in the Bures-Wasserstein Space. In *International Conference on Machine Learning*, pages 7960–7991. PMLR, 2023.

References III

Jiaojiao Fan, Qinsheng Zhang, Amirhossein Taghvaei, and Yongxin Chen. Variational wasserstein gradient flow. In *International Conference on Machine Learning*, 2022.

Bela A Frigyik, Santosh Srivastava, and Maya R Gupta. Functional Bregman divergence. In *2008 IEEE International Symposium on Information Theory*, pages 1681–1685. IEEE, 2008.

Richard Jordan, David Kinderlehrer, and Felix Otto. The Variational Formulation of the Fokker–Planck Equation. *SIAM journal on mathematical analysis*, 29(1): 1–17, 1998.

Marc Lambert, Sinho Chewi, Francis Bach, Silvère Bonnabel, and Philippe Rigollet. Variational inference via wasserstein gradient flows. *Advances in Neural Information Processing Systems*, 35:14434–14447, 2022.

Nicolas Lanzetti, Saverio Bolognani, and Florian Dörfler. First-Order Conditions for Optimization in the Wasserstein Space. *arXiv preprint arXiv:2209.12197*, 2022.

Haihao Lu, Robert M Freund, and Yurii Nesterov. Relatively Smooth Convex Optimization by First-Order Methods, and Applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

References IV

Chris J Maddison, Daniel Paulin, Yee Whye Teh, and Arnaud Doucet. Dual Space Preconditioning for Gradient Descent. *SIAM Journal on Optimization*, 31(1): 991–1016, 2021.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

Erkan Turan and Maks Ovsjanikov. Generative drifting is secretly score matching: a spectral and variational perspective. *arXiv preprint arXiv:2603.09936*, 2026.

Andre Wibisono. Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem. In *Conference on Learning Theory*, pages 2093–3027. PMLR, 2018.