Explicit Optimization Methods in (Wasserstein over) Wasserstein Space

Clément Bonet[1]

Joint works with Théo Uscidda[2], Adam David[3],
Pierre-Cyril Aubin-Frankowski[4], Christophe Vauthier[5], Anna Korba[2]

[1]Ecole Polytechnique, CMAP, Institut Polytechnique de Paris
[2]ENSAE, CREST, Institut Polytechnique de Paris
[3]TU Berlin
[4]CERMICS, ENPC, Institut Polytechnique de Paris
[5]Université Paris-Saclay, Laboratoire de Mathématique d'Orsay

Team Mind Seminar
03/03/2026

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:

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

Applications:

- 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)

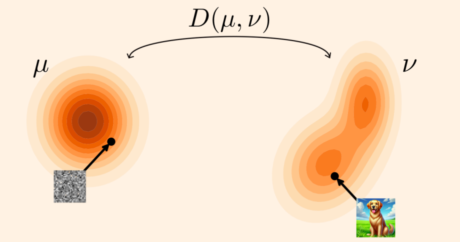
Generative Modeling

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

- $\mathcal{F}(\mu) = \frac{1}{2}\mathrm{MMD}^2(\mu, \nu)$
- $\mathcal{F}(\mu) = \mathrm{KL}(\nu||\mu)$
- $\mathcal{F}(\mu) = \frac{1}{2}\mathrm{W}_2^2(\mu, \nu)$



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)

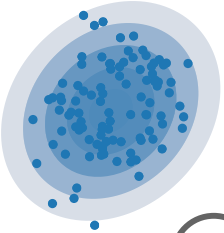


Table of Contents

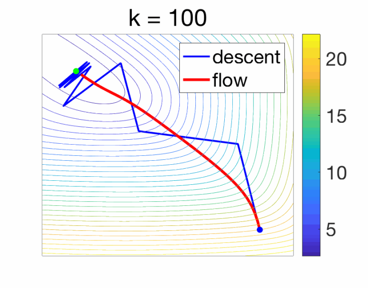
Gradient Descent on $\mathbb{R}^d$

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

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

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

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



From (Bach, 2020)

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

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

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

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

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

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

  $\iff$ $\forall x, y \in \mathbb{R}^d,\ f(x) - f(y) - \langle \nabla f(y), x - y\rangle \geq \dfrac{\alpha}{2}\|x - y\|_2^2$

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

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

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 $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ 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)

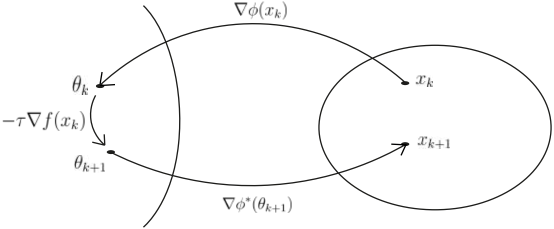
Let $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be convex, then the Bregman divergence is defined as

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

Mirror Descent (MD) algorithm:

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

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 $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ 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
- **Brenier's theorem**: If $\mu \ll \mathrm{Leb}$, then there exists a unique T_μ^ν such that

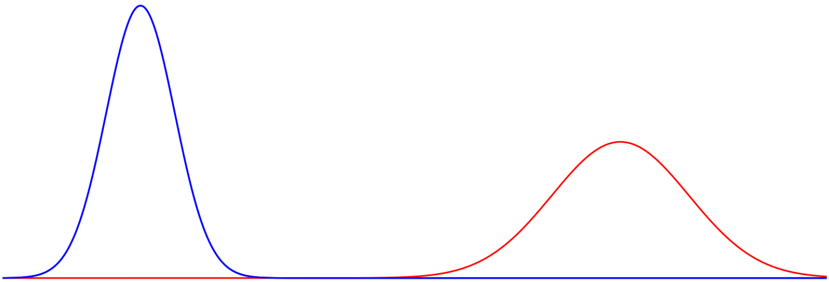
1. $(\mathrm{T}_\mu^\nu)_\# \mu = \nu$, *i.e.* if $X \sim \mu$, $\mathrm{T}_\mu^\nu(X) \sim \nu$
2. $\mathrm{W}_2^2(\mu, \nu) = \int \|x - \mathrm{T}_\mu^\nu(x)\|_2^2 \, \mathrm{d}\mu(x) = \|\mathrm{Id} - \mathrm{T}_\mu^\nu\|_{L^2(\mu)}^2$

- **Riemannian structure**

Riemannian Structure of the Wasserstein Space

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

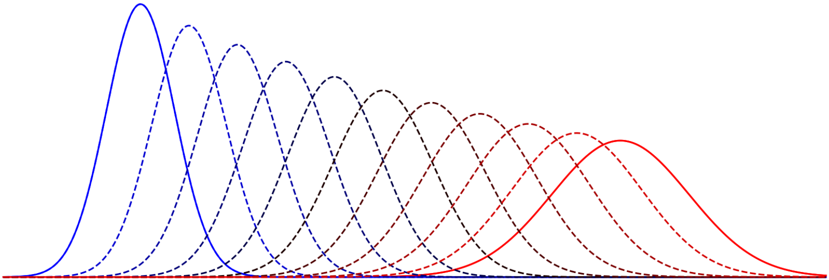
$$\forall t \in [0,1],\ \mu_t = \big((1-t)\mathrm{Id} + t\mathrm{T}_\mu^\nu\big)_\#\mu$$



Riemannian Structure of the Wasserstein Space

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

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



Riemannian Structure of the Wasserstein Space

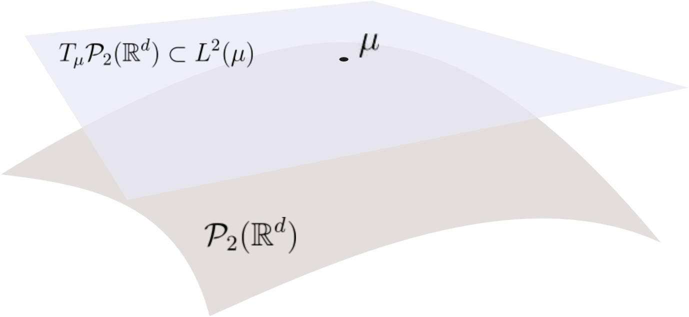
- Geodesics between $\mu \ll \mathrm{Leb}$ and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$:

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

- 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

Definition (Wasserstein gradient (Bonnet, 2019))

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

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

If such a gradient exists, then we say that $\mathcal{F}$ is W_2-differentiable at μ.

Wasserstein Gradient

Definition (Wasserstein gradient (Bonnet, 2019))

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

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

If such a gradient exists, then we say that $\mathcal{F}$ is W_2-differentiable at μ.

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

Definition (Wasserstein gradient (Bonnet, 2019))

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

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

If such a gradient exists, then we say that $\mathcal{F}$ is W_2-differentiable at μ.

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,

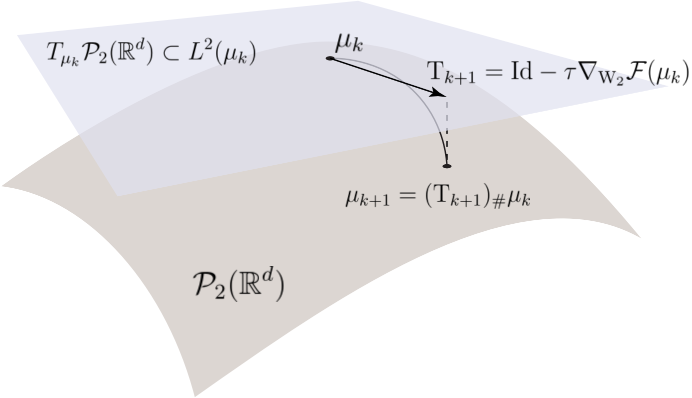
$$\nabla_{\mathrm{W}_2}\mathcal{W}(\mu) = \nabla W \star \mu$$

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

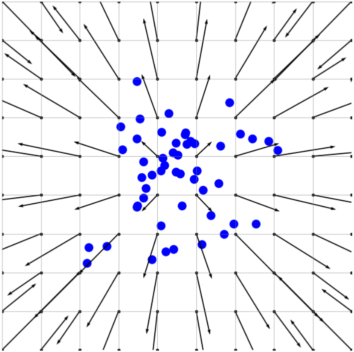


Wasserstein Gradient Descent in Practice

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

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: $\forall i \in \{1,\dots,n\},\ x_i^{k+1} = \mathrm{T}_{k+1}(x_i^k)$

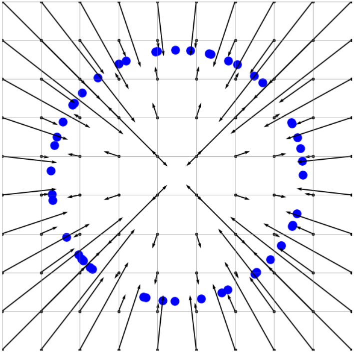


Table of Contents

Mirror Descent (Bonet et al., 2024)

Study schemes of the form

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

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

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

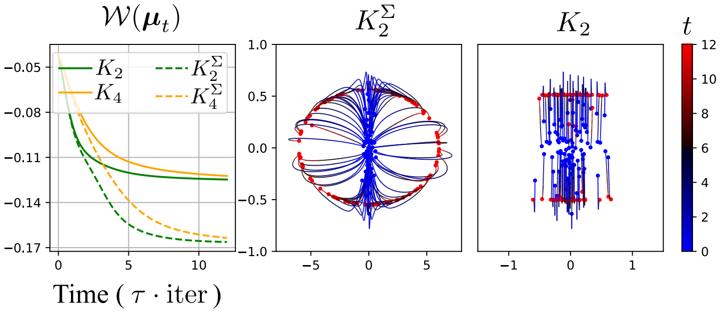
Mirror Descent on Interaction Energy

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

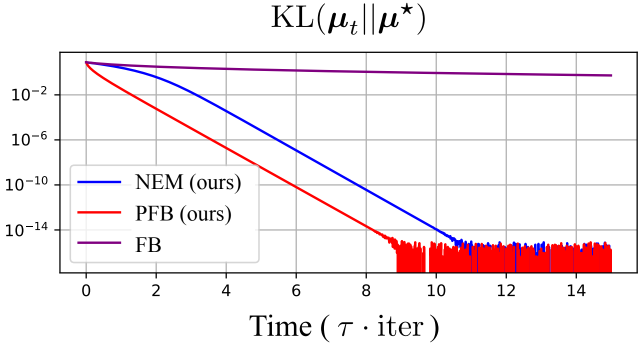
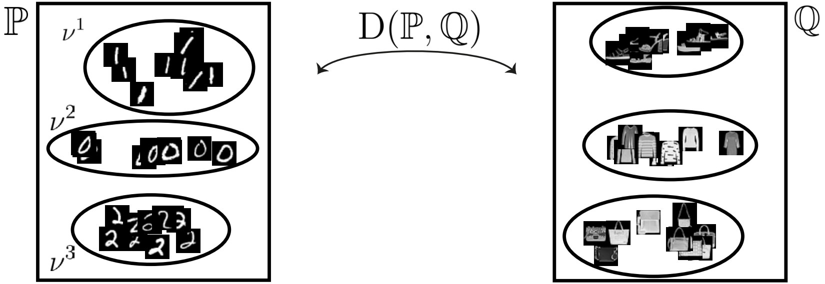


Table of Contents

Contributions

- Model datasets as $\mathbb{P} = \frac{1}{C}\sum_{c=1}^{C} \delta_{\nu^c} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$ where $\nu^c = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i^c}$
- Flow a dataset $\mathbb{P}$ towards $\mathbb{Q}$ by minimizing a discrepancy D on $\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$



Wasserstein over Wasserstein Distance (WoW)

Definition (WoW distance)

Let $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$ and denote by $\Pi(\mathbb{P}, \mathbb{Q})$ the set of coupling between $\mathbb{P}, \mathbb{Q}$. Then, the WoW distance is

$$\mathrm{W}_{\mathrm{W}_2}^2(\mathbb{P}, \mathbb{Q}) = \inf_{\Gamma \in \Pi(\mathbb{P}, \mathbb{Q})} \int \mathrm{W}_2^2(\mu, \nu)\ \mathrm{d}\Gamma(\mu, \nu).$$

Wasserstein over Wasserstein Distance (WoW)

Definition (WoW distance)

Let $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$ and denote by $\Pi(\mathbb{P}, \mathbb{Q})$ the set of coupling between $\mathbb{P}, \mathbb{Q}$. Then, the WoW distance is

$$\mathrm{W}^2_{\mathrm{W}_2}(\mathbb{P}, \mathbb{Q}) = \inf_{\Gamma \in \Pi(\mathbb{P}, \mathbb{Q})} \int \mathrm{W}_2^2(\mu, \nu)\ \mathrm{d}\Gamma(\mu, \nu).$$

Properties:

- $\mathrm{W}_{\mathrm{W}_2}$ distance, $\big(\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big), \mathrm{W}_{\mathrm{W}_2}\big)$: WoW space
- **Riemannian structure**
- Can be rewritten (Bonet et al., 2025; Pinzi and Savaré, 2025):

$$\mathrm{W}^2_{\mathrm{W}_2}(\mathbb{P}, \mathbb{Q}) = \inf_{\mathbb{\Gamma} \in \Lambda(\mathbb{P}, \mathbb{Q})} \iint \|y - x\|_2^2\ \mathrm{d}\gamma(x, y)\mathrm{d}\mathbb{\Gamma}(\gamma),$$

where $\Lambda(\mathbb{P}, \mathbb{Q}) = \{\mathbb{\Gamma} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)\big),\ \phi^1_\#\mathbb{\Gamma} = \mathbb{P},\ \phi^2_\#\mathbb{\Gamma} = \mathbb{Q},\ \iint \|x - y\|_2^2\ \mathrm{d}\gamma(x, y)\mathrm{d}\mathbb{\Gamma}(\gamma) = \mathrm{W}^2_{\mathrm{W}_2}(\mathbb{P}, \mathbb{Q})\}$, $\phi^1(\gamma) = \pi^1_\#\gamma$ and $\phi^2(\gamma) = \pi^2_\#\gamma$

WoW Gradient

Let $\mathbb{F} : \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big) \to \mathbb{R}$.

Definition (WoW gradient)

Let $\mathbb{P} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$. $\nabla_{\mathrm{W}_{\mathrm{W}_2}}\mathbb{F}(\mathbb{P}) \in L^2(\mathbb{P}, T\mathcal{P}_2(\mathbb{R}^d))$ is a WoW gradient of $\mathbb{F}$ at $\mathbb{P}$ if for any $\mathbb{Q} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$ and any $\mathbb{\Gamma} \in \Lambda(\mathbb{P}, \mathbb{Q})$,

$$\mathbb{F}(\mathbb{Q}) = \mathbb{F}(\mathbb{P}) + \iint \langle \nabla_{\mathrm{W}_2}\mathbb{F}(\mathbb{P})(\pi^1_\#\gamma)(x), y - x\rangle \,\mathrm{d}\gamma(x,y)\mathrm{d}\mathbb{\Gamma}(\gamma) + o\big(\mathrm{W}_{\mathrm{W}_2}(\mathbb{P}, \mathbb{Q})\big).$$

If such a gradient exists, then we say that $\mathbb{F}$ is $\mathrm{W}_{\mathrm{W}_2}$-differentiable at $\mathbb{P}$.

WoW Gradient

Let $\mathbb{F} : \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big) \to \mathbb{R}$.

Definition (WoW gradient)

Let $\mathbb{P} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$. $\nabla_{\mathrm{W}_{\mathrm{W}_2}}\mathbb{F}(\mathbb{P}) \in L^2(\mathbb{P}, T\mathcal{P}_2(\mathbb{R}^d))$ is a WoW gradient of $\mathbb{F}$ at $\mathbb{P}$ if for any $\mathbb{Q} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$ and any $\mathbb{\Gamma} \in \Lambda(\mathbb{P}, \mathbb{Q})$,

$$\mathbb{F}(\mathbb{Q}) = \mathbb{F}(\mathbb{P}) + \iint \langle \nabla_{\mathrm{W}_2}\mathbb{F}(\mathbb{P})(\pi^1_\#\gamma)(x), y - x\rangle \,\mathrm{d}\gamma(x,y)\mathrm{d}\mathbb{\Gamma}(\gamma) + o\big(\mathrm{W}_{\mathrm{W}_2}(\mathbb{P}, \mathbb{Q})\big).$$

If such a gradient exists, then we say that $\mathbb{F}$ is $\mathrm{W}_{\mathrm{W}_2}$-differentiable at $\mathbb{P}$.

Properties:

- Study of existence and uniqueness of $\nabla_{\mathrm{W}_{\mathrm{W}_2}}\mathbb{F}(\mathbb{P})$ in $\partial\mathbb{F}(\mathbb{P}) \cap T_{\mathbb{P}}\mathcal{P}_2\big(\mathcal{P}_2(\mathcal{M})\big)$
- Strong differentials (*i.e.* we can choose non optimal $\mathbb{\Gamma}$)

WoW Gradient

Let $\mathbb{F} : \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big) \to \mathbb{R}$.

Definition (WoW gradient)

Let $\mathbb{P} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$. $\nabla_{\mathrm{W}_{\mathrm{W}_2}}\mathbb{F}(\mathbb{P}) \in L^2(\mathbb{P}, T\mathcal{P}_2(\mathbb{R}^d))$ is a WoW gradient of $\mathbb{F}$ at $\mathbb{P}$ if for any $\mathbb{Q} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$ and any $\mathbb{\Gamma} \in \Lambda(\mathbb{P}, \mathbb{Q})$,

$$\mathbb{F}(\mathbb{Q}) = \mathbb{F}(\mathbb{P}) + \iint \langle \nabla_{\mathrm{W}_2}\mathbb{F}(\mathbb{P})(\pi^1_{\#}\gamma)(x), y - x\rangle \, \mathrm{d}\gamma(x,y)\mathrm{d}\mathbb{\Gamma}(\gamma) + o\big(\mathrm{W}_{\mathrm{W}_2}(\mathbb{P},\mathbb{Q})\big).$$

If such a gradient exists, then we say that $\mathbb{F}$ is $\mathrm{W}_{\mathrm{W}_2}$-differentiable at $\mathbb{P}$.

Example of functionals

- Potential energies $\mathbb{V}(\mathbb{P}) = \int \mathcal{F}(\mu)\mathrm{d}\mathbb{P}(\mu)$: For $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ differentiable and smooth,

$$\nabla_{\mathrm{W}_{\mathrm{W}_2}}\mathbb{V}(\mathbb{P}) = \nabla_{\mathrm{W}_2}\mathcal{F}$$

- Interaction energies $\mathbb{W}(\mathbb{P}) = \iint \mathrm{W}(\mu,\nu)\, \mathrm{d}\mathbb{P}(\mu)\mathrm{d}\mathbb{P}(\nu)$: For W differentiable and smooth,

$$\nabla_{\mathrm{W}_{\mathrm{W}_2}}\mathbb{W}(\mathbb{P})(\mu) = \int \big(\nabla_1 \mathrm{W}(\mu,\cdot) + \nabla_2 \mathrm{W}(\cdot,\mu)\big)\, \mathrm{d}\mathbb{P}$$

WoW Gradient Descent

Forward scheme:

$$\forall k \geq 0,\ \mathbb{P}_{k+1} = \exp_{\mathbb{P}_k}\big(-\tau \nabla_{\mathrm{W}_{\mathrm{W}_2}} \mathbb{F}(\mathbb{P}_k)\big)$$

WoW Gradient Descent

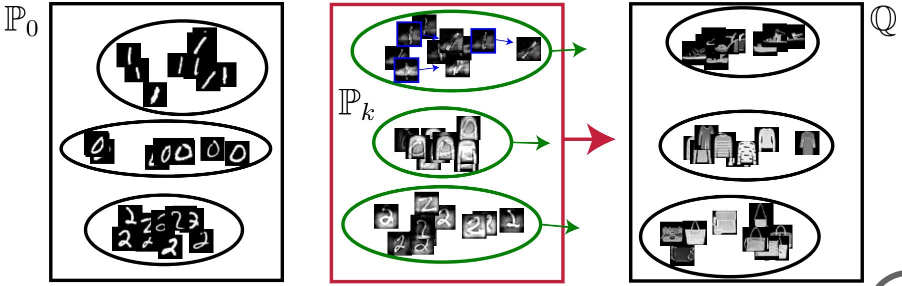
Forward scheme:

$$\forall k \geq 0,\ \mathbb{P}_{k+1} = \exp_{\mathbb{P}_k}\big(-\tau \nabla_{\mathrm{W_{W_2}}} \mathbb{F}(\mathbb{P}_k)\big)$$

In practice: For $\mathbb{P}_k = \frac{1}{C}\sum_{c=1}^{C} \delta_{\mu_k^{c,n}}$ with $\mu_k^{c,n} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_{i,k}^c} \in \mathcal{P}_2(\mathbb{R}^d)$:

$$\forall k \geq 0,\ \text{particle (image) } i,\ \text{class } c,\ x_{i,k+1}^c = x_{i,k}^c - \tau \nabla_{\mathrm{W_{W_2}}} \mathbb{F}(\mathbb{P}_k)(\mu_k^{c,n})(x_{i,k}^c).$$

$\mathbb{P}_k$: inter-class interaction, $\mu_k^{c,n}$: intra-class interaction, $x_{i,k}^c$ image



Synthetic Data

$$\mathbb{F}(\mathbb{P}) = \frac{1}{2}\mathrm{MMD}_K^2(\mathbb{P},\mathbb{Q}) = \mathbb{V}(\mathbb{P}) + \mathbb{W}(\mathbb{P}) + \mathrm{cst},$$

$$\text{where} \quad \begin{cases} \mathbb{V}(\mathbb{P}) = \int \mathcal{V}(\mu)\ \mathrm{d}\mathbb{P}(\mu), \quad \mathcal{V}(\mu) = -\int K(\mu,\nu)\ \mathrm{d}\mathbb{Q}(\nu) \\ \mathbb{W}(\mathbb{P}) = \frac{1}{2}\iint K(\mu,\nu)\ \mathrm{d}\mathbb{P}(\mu)\mathrm{d}\mathbb{P}(\nu) \end{cases}$$

- WoW gradient computed in **closed-form** or using **auto-differentiation**
- Kernel K based on the **Sliced-Wasserstein** distance
- **Complexity**: $O(C^2 Ln\log n)$, $\mathbb{P} = \frac{1}{C}\sum_{c=1}^{C}\delta_{\mu^{c,n}}$, $\mu^{c,n} = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$

Sliced-Wasserstein distance (Rabin et al., 2011; Bonneel et al., 2015):

$$\mathrm{SW}_2^2(\mu,\nu) = \int_{S^{d-1}} \mathrm{W}_2^2(P_\#^\theta\mu, P_\#^\theta\nu)\mathrm{d}\sigma(\theta) \approx \frac{1}{L}\sum_{\ell=1}^{L}\mathrm{W}_2^2(P_\#^{\theta_\ell}\mu, P_\#^{\theta_\ell}\nu),$$

with $S^{d-1} = \{\theta\in\mathbb{R}^d,\ \|\theta\|_2 = 1\}$, $P^\theta(x) = \langle x,\theta\rangle$, $\theta_1,\dots,\theta_L \sim \sigma = \mathrm{Unif}(S^{d-1})$.

- Gaussian SW kernel: $K(\mu,\nu) = e^{-\mathrm{SW}_2^2(\mu,\nu)/h}$ (Kolouri et al., 2016)
- Riesz SW kernel: $K(\mu,\nu) = -\mathrm{SW}_2(\mu,\nu)$

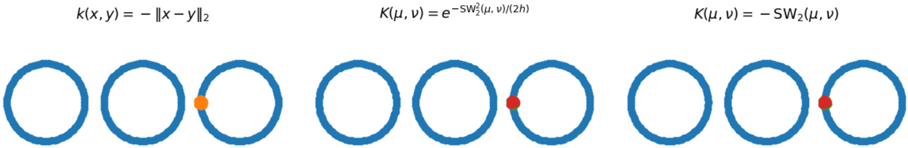
Synthetic Data

$$\mathbb{F}(\mathbb{P}) = \frac{1}{2}\mathrm{MMD}_K^2(\mathbb{P},\mathbb{Q}) = \mathbb{V}(\mathbb{P}) + \mathbb{W}(\mathbb{P}) + \mathrm{cst},$$

$$\text{where} \quad \begin{cases} \mathbb{V}(\mathbb{P}) = \int \mathcal{V}(\mu)\ \mathrm{d}\mathbb{P}(\mu), \quad \mathcal{V}(\mu) = -\int K(\mu,\nu)\ \mathrm{d}\mathbb{Q}(\nu) \\ \mathbb{W}(\mathbb{P}) = \frac{1}{2}\iint K(\mu,\nu)\ \mathrm{d}\mathbb{P}(\mu)\mathrm{d}\mathbb{P}(\nu) \end{cases}$$

- WoW gradient computed in **closed-form** or using **auto-differentiation**
- Kernel K based on the **Sliced-Wasserstein** distance
- **Complexity**: $O(C^2 L n \log n)$, $\mathbb{P} = \frac{1}{C}\sum_{c=1}^{C} \delta_{\mu^{c,n}}$, $\mu^{c,n} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$

Goal: $\min_{\mathbb{P}}\ \mathbb{F}(\mathbb{P}) = \frac{1}{2}\mathrm{MMD}_K^2(\mathbb{P},\mathbb{Q})$, where $\mathbb{Q} = \frac{1}{3}\sum_{c=1}^{3} \delta_{\nu^{c,n}}$, $\nu^{c,n}$ ring.



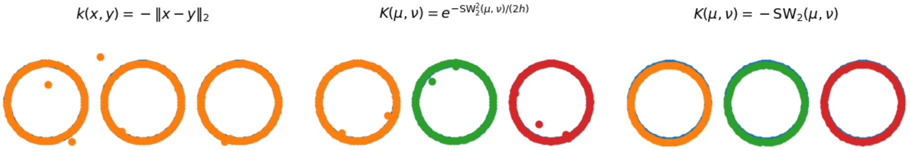
Synthetic Data

$$\mathbb{F}(\mathbb{P}) = \frac{1}{2}\mathrm{MMD}_K^2(\mathbb{P}, \mathbb{Q}) = \mathbb{V}(\mathbb{P}) + \mathbb{W}(\mathbb{P}) + \mathrm{cst},$$

where $\begin{cases} \mathbb{V}(\mathbb{P}) = \int \mathcal{V}(\mu)\ \mathrm{d}\mathbb{P}(\mu), \quad \mathcal{V}(\mu) = -\int K(\mu,\nu)\ \mathrm{d}\mathbb{Q}(\nu) \\ \mathbb{W}(\mathbb{P}) = \frac{1}{2}\iint K(\mu,\nu)\ \mathrm{d}\mathbb{P}(\mu)\mathrm{d}\mathbb{P}(\nu) \end{cases}$

- WoW gradient computed in **closed-form** or using **auto-differentiation**
- Kernel K based on the **Sliced-Wasserstein** distance
- **Complexity**: $O(C^2 L n \log n)$, $\mathbb{P} = \frac{1}{C}\sum_{c=1}^{C} \delta_{\mu^{c,n}}$, $\mu^{c,n} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$

Goal: $\min_{\mathbb{P}}\ \mathbb{F}(\mathbb{P}) = \frac{1}{2}\mathrm{MMD}_K^2(\mathbb{P}, \mathbb{Q})$, where $\mathbb{Q} = \frac{1}{3}\sum_{c=1}^{3} \delta_{\nu^{c,n}}$, $\nu^{c,n}$ ring.



Conclusion

Conclusion:

- Mirror and Preconditioned Gradient Descent on $\mathcal{P}_2(\mathbb{R}^d)$
- Convergence analysis of the discrete schemes
- Differential structure on $(\mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d)), \mathrm{W}_{\mathrm{W}_2})$
- Application to flowing labeled datasets (including images in the paper)

Perspectives:

- New schemes analyzed with Bregman divergences on $\mathcal{P}_2(\mathbb{R}^d)$
- Better understand the geometry of $(\mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d)), \mathrm{W}_{\mathrm{W}_2})$

Conclusion

Conclusion:

- Mirror and Preconditioned Gradient Descent on $\mathcal{P}_2(\mathbb{R}^d)$
- Convergence analysis of the discrete schemes
- Differential structure on $(\mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d)), \mathrm{W}_{\mathrm{W}_2})$
- Application to flowing labeled datasets (including images in the paper)

Perspectives:

- New schemes analyzed with Bregman divergences on $\mathcal{P}_2(\mathbb{R}^d)$
- Better understand the geometry of $(\mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d)), \mathrm{W}_{\mathrm{W}_2})$

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.

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.

Clément Bonet, Christophe Vauthier, and Anna Korba. Flowing Datasets with Wasserstein over Wasserstein Gradient Flows. In *International Conference on Machine Learning*. PMLR, 2025.

References II

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51:22–45, 2015.

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

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.

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.

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.

References III

Soheil Kolouri, Yang Zou, and Gustavo K Rohde. Sliced Wasserstein Kernels for Probability Distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5258–5267, 2016.

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.

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.

References IV

Alessandro Pinzi and Giuseppe Savaré. Totally convex functions, l^2-optimal transport for laws of random measures, and solution to the monge problem. *arXiv preprint arXiv:2509.01768*, 2025.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International conference on scale space and variational methods in computer vision*, pages 435–446. Springer, 2011.

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.

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.