Difference of Convex Programming in the Wasserstein Space with Applications to MMD Optimization

Clément Bonet[1], Pierre-Cyril Aubin-Frankowski[2], Youssef Mroueh[3]

[1]Ecole Polytechnique, CMAP, Institut Polytechnique de Paris
[2]CERMICS, ENPC, Institut Polytechnique de Paris
[3]IBM Research

LOL 2026
01/07/2026

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)

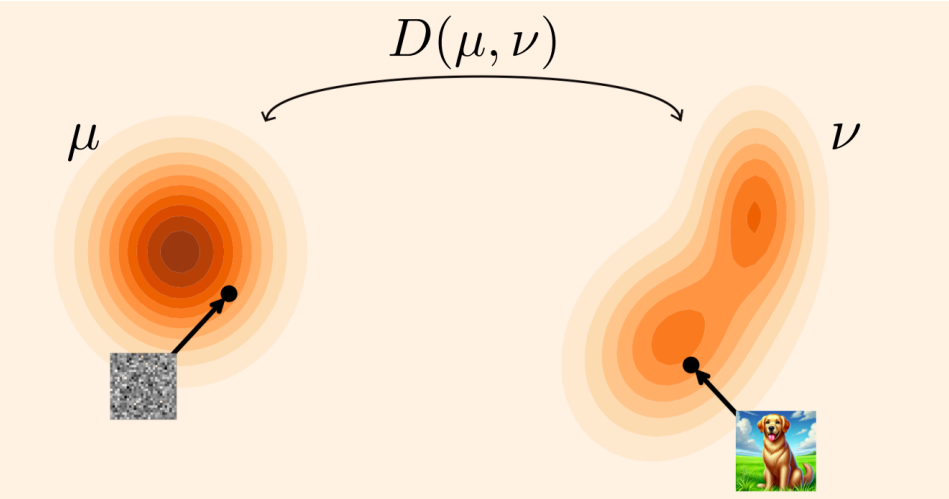
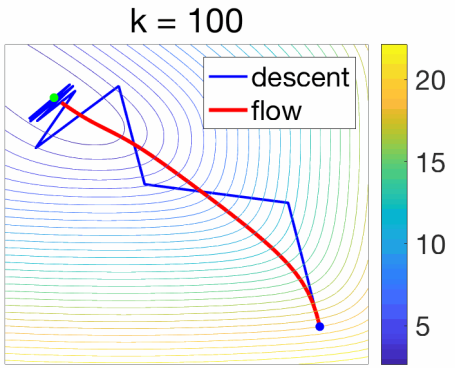


Table of Contents

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

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

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



From (Bach, 2020)

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

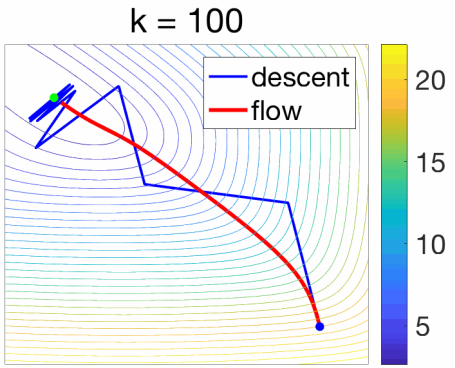
$$\forall \tau > 0, \; \forall k \geq 0, \; x_{k+1} = x_k - \tau \nabla f(x_k)$$

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

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

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

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



From (Bach, 2020)

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

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

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)

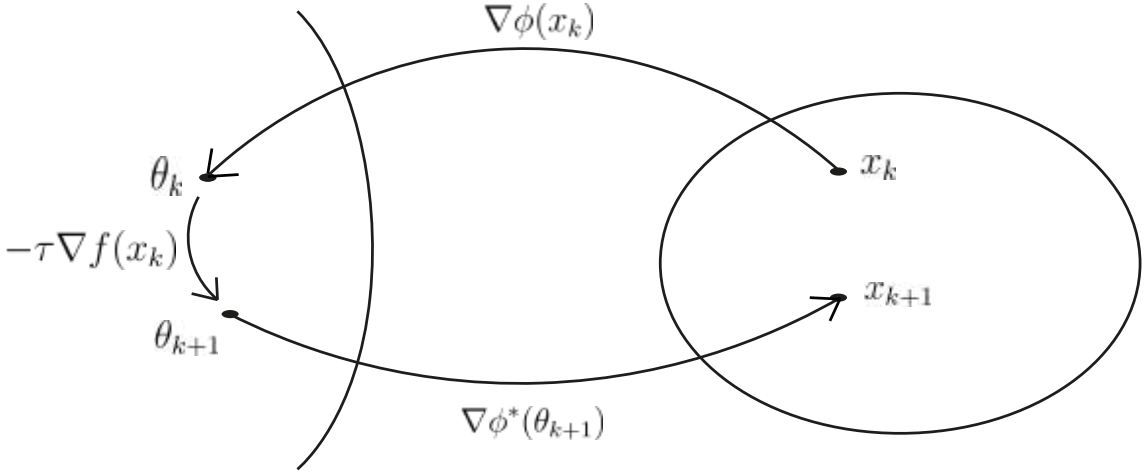
Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be convex, then the Bregman divergence is defined as

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

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) $\implies$ $f(x_k) - f(x^*) \leq \frac{\beta - \alpha}{k} \mathrm{D}_\phi(x^*, x_0)$

CCCP on $\mathbb{R}^d$ (Yuille and Rangarajan, 2001)

f DC (i.e. difference-of-convex) if there exists f^-, f^+ convex such that

$$f = f^+ - f^-$$

Remark: Every C^1 function with Lipschitz gradient is DC (Hiriart-Urruty, 1985)

CCCP on $\mathbb{R}^d$ (Yuille and Rangarajan, 2001)

f DC (i.e. difference-of-convex) if there exists f^-, f^+ convex such that

$$f = f^+ - f^-$$

Remark: Every C^1 function with Lipschitz gradient is DC (Hiriart-Urruty, 1985)

At iteration k, given $x_k \in \mathbb{R}^d$,

$$\begin{aligned} f^- \text{ convex} &\implies \forall x \in \mathbb{R}^d,\ f^-(x) \geq f^-(x_k) + \langle \nabla f^-(x_k), x - x_k \rangle \\ &\implies f(x) = f^+(x) - f^-(x) \leq f^+(x) - f^-(x_k) - \langle \nabla f^-(x_k), x - x_k \rangle \end{aligned}$$

CCCP on $\mathbb{R}^d$ (Yuille and Rangarajan, 2001)

f DC (i.e. difference-of-convex) if there exists f^-, f^+ convex such that

$$f = f^+ - f^-$$

Remark: Every C^1 function with Lipschitz gradient is DC (Hiriart-Urruty, 1985)

At iteration k, given $x_k \in \mathbb{R}^d$,

$$\begin{aligned} f^- \text{ convex} &\implies \forall x \in \mathbb{R}^d,\ f^-(x) \geq f^-(x_k) + \langle \nabla f^-(x_k), x - x_k \rangle \\ &\implies f(x) = f^+(x) - f^-(x) \leq f^+(x) - f^-(x_k) - \langle \nabla f^-(x_k), x - x_k \rangle \end{aligned}$$

Convex Concave Procedure (CCCP)

$$\forall k \geq 0,\ x_{k+1} = \operatorname*{argmin}_{x \in \mathbb{R}^d}\ f^+(x) - \langle \nabla f^-(x_k), x - x_k \rangle$$

$\rightarrow$ Majorization-Minimization algorithm

$\rightarrow$ Objective convex at each iteration

Table of Contents

Wasserstein Geometry (Ambrosio et al., 2008)

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., 2008)

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:

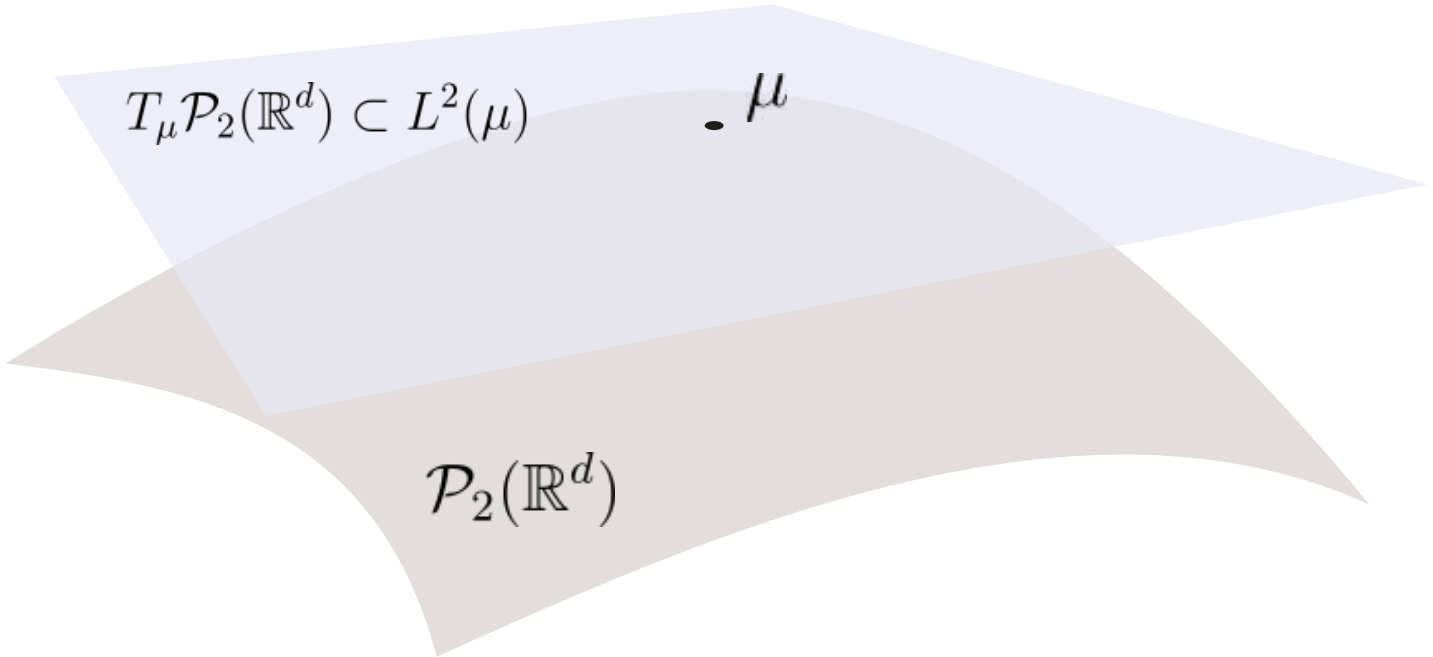
- W_2 distance, $(\mathcal{P}_2(\mathbb{R}^d), \mathrm{W}_2)$: Wasserstein space
- $\mathrm{W}_2(\delta_x, \delta_y) = \|x - y\|_2$
- **Riemannian structure**

Tangent Space on the Wasserstein Space

Tangent space at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ (Ambrosio et al., 2008):

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

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., 2008)

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., 2008)

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 (Ambrosio et al., 2008)

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 $\gamma \in \Pi_o(\mu, \nu)$,

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

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,

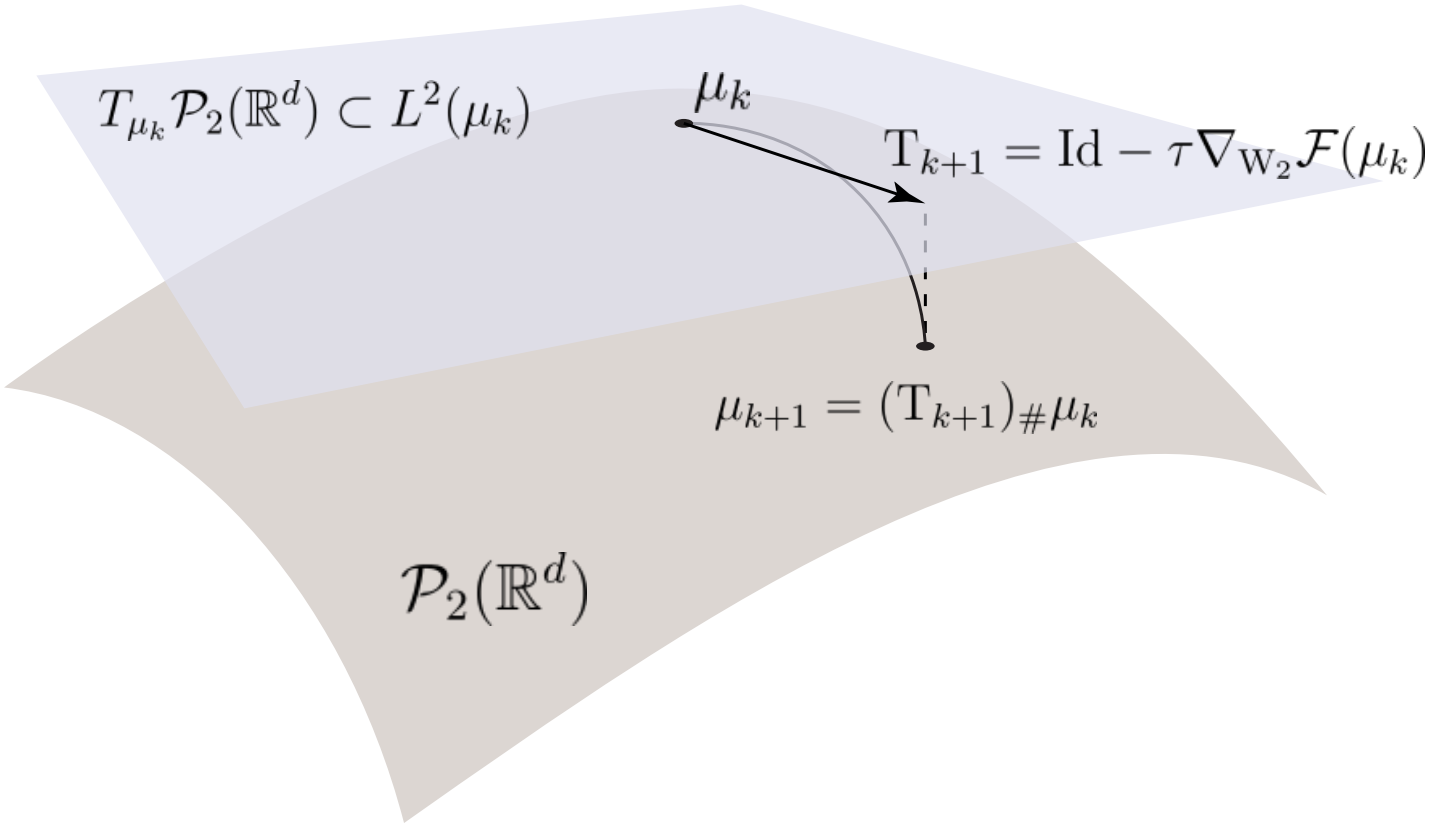
$$\nabla_{\mathrm{W}_2}\mathcal{W}(\mu) = \nabla W \star \mu$$

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

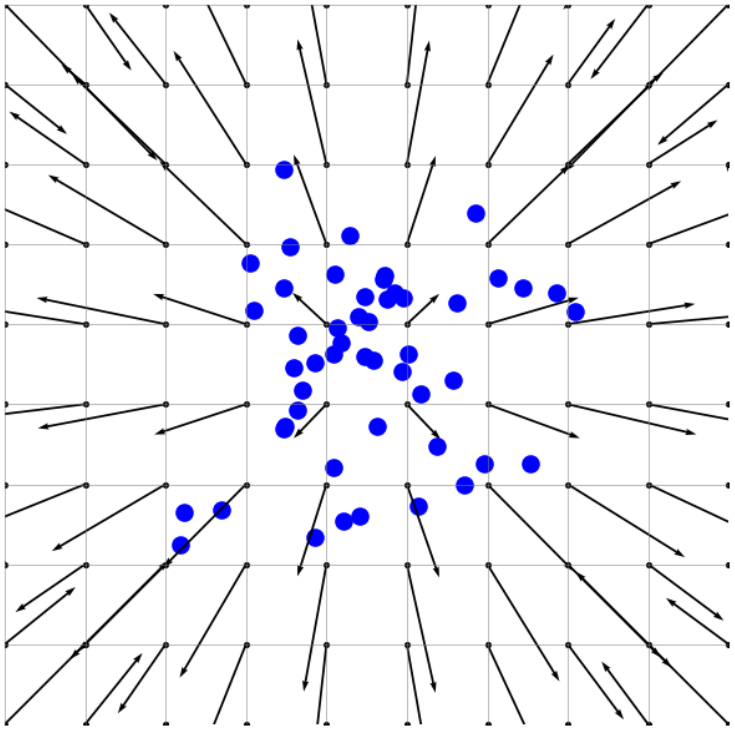


Wasserstein Gradient Descent in Practice

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

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 $\mathrm{T}_{k+1} = \mathrm{Id} - \tau\nabla_{\mathrm{W}_2}\mathcal{F}(\hat{\mu}_k^n) = \mathrm{Id} - \tau\int \nabla W(\cdot - y)\ \mathrm{d}\hat{\mu}_k^n(y)$
- Update particles: $\forall i \in \{1,\dots,n\},\ x_i^{k+1} = \mathrm{T}_{k+1}(x_i^k)$

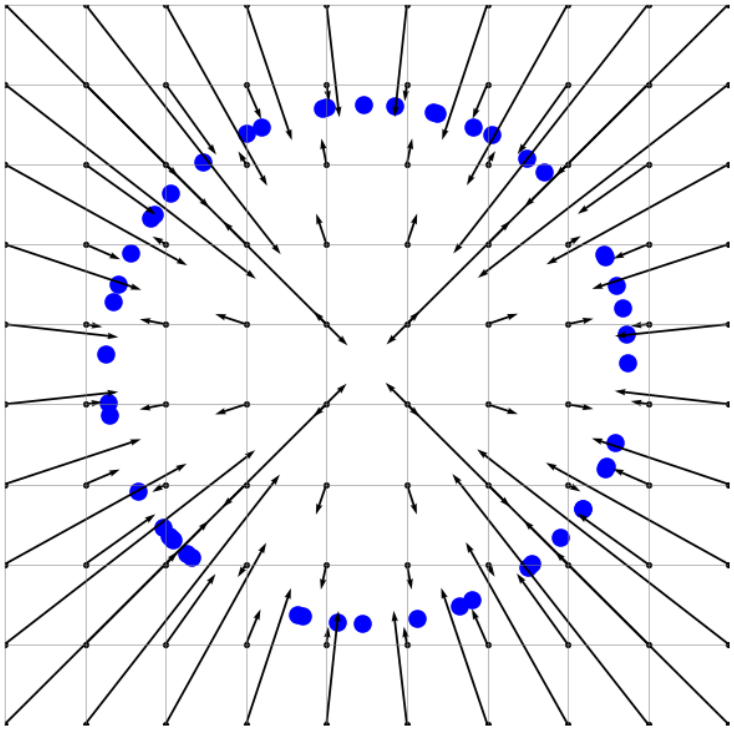


Wasserstein Gradient Descent in Practice

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

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 $\mathrm{T}_{k+1} = \mathrm{Id} - \tau \nabla_{\mathrm{W}_2} \mathcal{F}(\hat{\mu}_k^n) = \mathrm{Id} - \tau \int \nabla W(\cdot - y)\ \mathrm{d}\hat{\mu}_k^n(y)$
- Update particles: $\forall i \in \{1, \dots, n\},\ x_i^{k+1} = \mathrm{T}_{k+1}(x_i^k)$



Wasserstein Mirror Descent (Bonet et al., 2024)

Mirror Descent

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

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 $\mathcal{P}_2(\mathbb{R}^d)$.

Theoretical analysis: requires (relative) smoothness and convexity along curves

Table of Contents

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,

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

- Given $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$, define the Bregman divergence associated to $\tilde{\mathcal{F}}_\mu : \mathrm{T} \mapsto \mathcal{F}(\mathrm{T}_{\#}\mu)$ as

$$\mathrm{D}^\mu_{\mathcal{F}}(\mathrm{T}, \mathrm{S}) = \mathcal{F}(\mathrm{T}_{\#}\mu) - \mathcal{F}(\mathrm{S}_{\#}\mu) - \langle \nabla_{\mathrm{W}_2}\mathcal{F}(\mathrm{S}_{\#}\mu) \circ \mathrm{S}, \mathrm{T} - \mathrm{S} \rangle_{L^2(\mu)}.$$

Convexity along a curve on $\mathcal{P}_2(\mathbb{R}^d)$

Let $\mu_t = (\mathrm{T}_t)_{\#}\mu$ with $\mathrm{T}_t = (1-t)\mathrm{S} + t\mathrm{T}$ for all $t \in [0,1]$, $\mathrm{T}, \mathrm{S} \in L^2(\mu)$.

Relative smoothness/convexity along $t \mapsto \mu_t$:

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

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

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

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

Convexity along a curve on $\mathcal{P}_2(\mathbb{R}^d)$

Let $\mu_t = (\mathrm{T}_t)_{\#}\mu$ with $\mathrm{T}_t = (1-t)\mathrm{S} + t\mathrm{T}$ for all $t \in [0,1]$, $\mathrm{T}, \mathrm{S} \in L^2(\mu)$.

Relative smoothness/convexity along $t \mapsto \mu_t$:

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

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

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

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

Total convexity

Let $\alpha \geq 0$. A functional $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is α-**totally convex** if for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $\mathrm{T}, \mathrm{S} \in L^2(\mu)$,

$$\mathrm{D}^{\mu}_{\mathcal{F}}(\mathrm{T}, \mathrm{S}) \geq \tfrac{\alpha}{2}\|\mathrm{T} - \mathrm{S}\|^2_{L^2(\mu)} = \alpha \mathrm{D}^{\mu}_{\mathcal{G}}(\mathrm{T}, \mathrm{S}) \quad \text{for} \quad \mathcal{G}(\mu) = \int \tfrac{1}{2}\|\cdot\|_2^2 \, \mathrm{d}\mu.$$

Equivalently, for all $t \in [0,1]$,

$$\mathcal{F}\big(\big((1-t)\mathrm{S} + t\mathrm{T}\big)_{\#}\mu\big) \leq (1-t)\mathcal{F}(\mathrm{S}_{\#}\mu) + t\mathcal{F}(\mathrm{T}_{\#}\mu) - \alpha \frac{t(1-t)}{2}\|\mathrm{T} - \mathrm{S}\|^2_{L^2(\mu)}$$

Convex Functionals (Ambrosio et al., 2008, Chapter 9)

Example of Convex functionals

- Potential energies $\mathcal{V}(\mu) = \int V \mathrm{d}\mu$ with V α-convex

$$\begin{aligned}\mathcal{V}(\mu_t) &= \int V\big((1-t)\mathrm{S}(x) + t\mathrm{T}(x)\big)\ \mathrm{d}\mu(x) \\ &\leq (1-t)\int V(\mathrm{S}(x))\ \mathrm{d}\mu(x) + t\int V(\mathrm{T}(x))\ \mathrm{d}\mu(x) \\ &\quad - \alpha\frac{t(1-t)}{2}\int \|\mathrm{S}(x) - \mathrm{T}(x)\|_2^2\ \mathrm{d}\mu \\ &\leq (1-t)\mathcal{V}(\mu_0) + t\mathcal{V}(\mu_1) - \alpha\frac{t(1-t)}{2}\mathrm{W}_2^2(\mu_0,\mu_1)\end{aligned}$$

- Interaction energies $\mathcal{W}(\mu) = \frac{1}{2}\iint W(x-y)\ \mathrm{d}\mu(x)\mathrm{d}\mu(y)$ with W convex
- Negative entropy $\mathcal{H}(\mu) = \int \rho\log\rho$ with $\mathrm{d}\mu = \rho\mathrm{dLeb}$
- $\mathcal{F}(\mu) = \mathrm{KL}(\mu||\nu) = \mathcal{H}(\mu) + \int V\mathrm{d}\mu + \mathrm{cst}$ with $\nu \propto e^{-V}$ log-concave

Functionals not Convex

Examples of functionals not convex (in general)

- $\mathcal{F}(\mu) = \frac{1}{2}\mathrm{W}_2^2(\mu, \nu)$: convex along only one curve
- KL with ν non log-concave
- Maximum Mean Discrepancy (MMD) (Arbel et al., 2019)

Functionals not Convex

Examples of functionals not convex (in general)

- $\mathcal{F}(\mu) = \frac{1}{2}\mathrm{W}_2^2(\mu, \nu)$: convex along only one curve
- KL with ν non log-concave
- Maximum Mean Discrepancy (MMD) (Arbel et al., 2019)

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

$$\mathcal{F}(\mu) = \frac{1}{2}\mathrm{MMD}^2(\mu, \nu) = \frac{1}{2}\iint k(x, y)\ \mathrm{d}(\mu - \nu)(x)\mathrm{d}(\mu - \nu)(y)$$
$$= \mathcal{V}_k(\mu) + \mathcal{W}_k(\mu) + \mathrm{cst},$$

with

$$\mathcal{V}_k(\mu) = \int V_k(x)\ \mathrm{d}\mu(x), \quad V_k(x) = -\int k(x, y)\ \mathrm{d}\nu(y),$$

$$\mathcal{W}_k(\mu) = \frac{1}{2}\iint k(x, y)\ \mathrm{d}\mu(x)\mathrm{d}\mu(y).$$

DC Decomposition of the KL

Goal: Given $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$, find $\mathcal{F}^+, \mathcal{F}^-$ totally convex such that

$$\mathcal{F}(\mu) = \mathcal{F}^+(\mu) - \mathcal{F}^-(\mu)$$

DC Decomposition of the KL

Goal: Given $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$, find $\mathcal{F}^+, \mathcal{F}^-$ totally convex such that

$$\mathcal{F}(\mu) = \mathcal{F}^+(\mu) - \mathcal{F}^-(\mu)$$

For $\nu \propto e^{-V}$,

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

If V DC, there exists V^+, V^- convex such that $V = V^+ - V^-$, and we can write

$$\mathcal{F}(\mu) = \mathcal{F}^+(\mu) - \mathcal{F}^-(\mu),$$

with

$$\mathcal{F}^+(\mu) = \mathcal{H}(\mu) + \int V^+ \mathrm{d}\mu + \mathrm{cst}, \quad \mathcal{F}^-(\mu) = \int V^- \mathrm{d}\mu$$

$\to \mathcal{F}^+, \mathcal{F}^-$ totally-convex

A first DC Decomposition of the MMD (Luu et al., 2024)

$$\mathcal{F}(\mu) = \tfrac{1}{2}\mathrm{MMD}^2(\mu, \nu) = \mathcal{V}_k(\mu) + \mathcal{W}_k(\mu) + c(\nu)$$

Assumptions:

- k L-smooth, *i.e.*

$$\forall x, x', y, y' \in \mathbb{R}^d,\ \|\nabla k(x, y) - \nabla k(x', y')\|_2 \leq L\big(\|x - x'\|_2 + \|y - y'\|_2\big)$$

- Choose $\alpha \geq L$

A DC decomposition is given by:

$$\mathcal{F}^+(\mu) = \alpha \int \|x\|_2^2 \ \mathrm{d}\mu(x) + \frac{1}{2} \iint k(x, y) \ \mathrm{d}\mu(x)\mathrm{d}\mu(y) + c(\nu),$$

$$\mathcal{F}^-(\mu) = \int \big(\alpha\|x\|_2^2 - V_k(x)\big) \ \mathrm{d}\mu(x)$$

$\rightarrow \mathcal{F}^+, \mathcal{F}^-$ totally convex

Other DC Decomposition of the MMD

$$\mathcal{F}(\mu) = \tfrac{1}{2}\mathrm{MMD}^2(\mu,\nu) = \mathcal{V}_k(\mu) + \mathcal{W}_k(\mu) + c(\nu)$$

Assumptions:

- $k(x,y) = \psi(x-y)$
- $\psi = \psi_+ - \psi_-$ with ψ_+, ψ_- respectively $\alpha^+, \alpha^- \geq 0$ convex

A DC decomposition is given by $\mathcal{F}(\mu) = \mathcal{F}^+(\mu) - \mathcal{F}^-(\mu)$ where

$$\begin{cases} \mathcal{F}^+(\mu) = \frac{1}{2}\iint \psi_+(x-y)\ \mathrm{d}\mu(x)\mathrm{d}\mu(y) + \int V^-\mathrm{d}\mu + c(\nu), \\ V^-(\cdot) = \int \psi_-(\cdot - y)\ \mathrm{d}\nu(y) \end{cases}$$

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

$\rightarrow \mathcal{F}^+, \mathcal{F}^-$ are respectively α^-, α^+ totally convex

Decomposition of ψ for Radial Kernels

2 possible decompositions for $\psi(z) = q(\|z\|_2^2)$, $q : \mathbb{R}_+ \to \mathbb{R}$:

1. Jordan decomposition: For $A \geq \max\big(0, -q'(0)\big)$ (to ensure $q_+, q_- \nearrow$),

$$\begin{cases} q_+(x) = q(0) + (q'(0) + A)x + \int_0^x (x-t) \max\big(0, q''(t)\big)\, \mathrm{d}t, \\ q_-(x) = Ax - \int_0^x (x-t) \min\big(0, q''(t)\big)\, \mathrm{d}t, \end{cases}$$

Gaussian kernel, *i.e.* $q(x) = e^{-x/(2h)}$

$$q_+(x) = e^{-x/(2h)} + \frac{x}{2h}, \quad q_-(x) = \frac{x}{2h}$$

Decomposition of ψ for Radial Kernels

2 possible decompositions for $\psi(z) = q(\|z\|_2^2)$, $q : \mathbb{R}_+ \to \mathbb{R}$:

1. Jordan decomposition: For $A \geq \max\big(0, -q'(0)\big)$ (to ensure $q_+, q_- \nearrow$),

$$\begin{cases} q_+(x) = q(0) + (q'(0) + A)x + \int_0^x (x-t) \max\big(0, q''(t)\big)\, \mathrm{d}t, \\ q_-(x) = Ax - \int_0^x (x-t) \min\big(0, q''(t)\big)\, \mathrm{d}t, \end{cases}$$

2. If $q(x) = \sum_{i \in \mathbb{N}} a_i x^i$ for all $x \in \mathbb{R}_+$,

$$q_+(x) = \sum_{i \in \mathbb{N}, a_i > 0} a_i x^i, \quad q_-(x) = -\sum_{i \in \mathbb{N}, a_i < 0} a_i x^i.$$

Gaussian kernel, *i.e.* $q(x) = e^{-x/(2h)}$

$$q_+(x) = \cosh\big(x/(2h)\big), \quad q_-(x) = \sinh\big(x/(2h)\big)$$

Decomposition of ψ for Radial Kernels

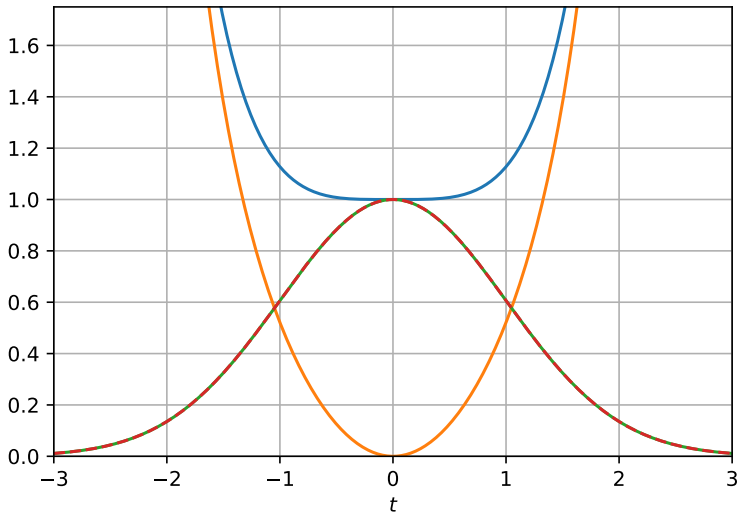
2 possible decompositions for $\psi(z) = q(\|z\|_2^2)$, $q : \mathbb{R}_+ \to \mathbb{R}$:

1. Jordan decomposition: For $A \geq \max\big(0, -q'(0)\big)$ (to ensure $q_+, q_- \nearrow$),

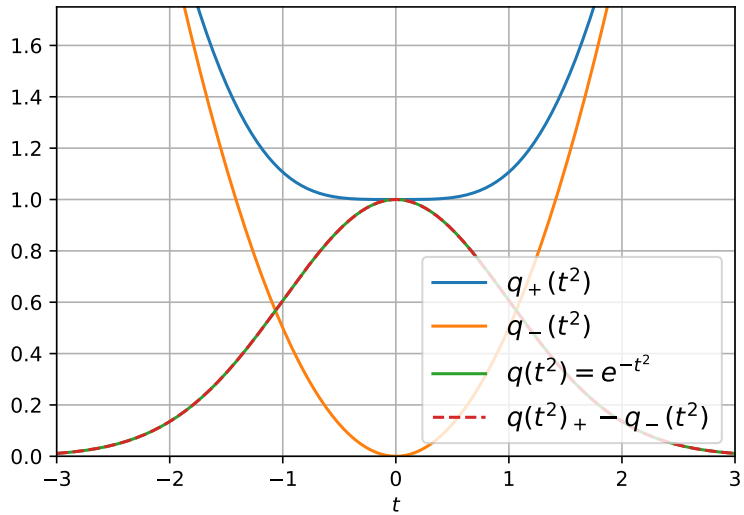
$$\begin{cases} q_+(x) = q(0) + (q'(0) + A)x + \int_0^x (x-t) \max\big(0, q''(t)\big)\, \mathrm{d}t, \\ q_-(x) = Ax - \int_0^x (x-t) \min\big(0, q''(t)\big)\, \mathrm{d}t, \end{cases}$$

2. If $q(x) = \sum_{i \in \mathbb{N}} a_i x^i$ for all $x \in \mathbb{R}_+$,

$$q_+(x) = \sum_{i \in \mathbb{N}, a_i > 0} a_i x^i, \quad q_-(x) = -\sum_{i \in \mathbb{N}, a_i < 0} a_i x^i.$$



$q_+(t) = \cosh(t),\ q_-(t) = \sinh(t)$



$q_+(t) = e^{-\alpha t} + \alpha t,\ q_-(t) = \alpha t$

Table of Contents

Wasserstein CCCP

Let $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ be a functional that can be decomposed as $\mathcal{F} = \mathcal{F}^+ - \mathcal{F}^-$ with $\mathcal{F}^+, \mathcal{F}^-$ totally convex.

For any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $\mathrm{T} \in L^2(\mu)$,

$$\begin{aligned} \mathcal{F}^- \text{ totally-convex} &\implies \mathrm{D}^{\mu}_{\mathcal{F}^-}(\mathrm{T}, \mathrm{Id}) \geq 0 \\ &\implies \mathcal{F}^-(\mathrm{T}_{\#}\mu) \geq \mathcal{F}^-(\mu) + \langle \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu), \mathrm{T} - \mathrm{Id} \rangle_{L^2(\mu)} \\ &\implies \mathcal{F}(\mathrm{T}_{\#}\mu) \leq \mathcal{F}^+(\mathrm{T}_{\#}\mu) - \mathcal{F}^-(\mu) - \langle \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu), \mathrm{T} - \mathrm{Id} \rangle_{L^2(\mu)} \end{aligned}$$

Wasserstein CCCP

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

$\to$ At each iteration, can perform gradient descent on $L^2(\mu_k)$,

$$\tilde{\mathrm{T}}_k^{\ell+1} = \tilde{\mathrm{T}}_k^{\ell} - \tau\big(\nabla_{\mathrm{W}_2}\mathcal{F}^+(\mu_k^{\ell}) \circ \tilde{\mathrm{T}}_k^{\ell} - \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k)\big)$$

Relation to Bregman and Mirror Descent

- **Equivalence with the Bregman Proximal Descent**:

$$\begin{aligned}
&\mathcal{F}^+(\mathrm{T}_{\#}\mu_k) - \mathcal{F}^-(\mu_k) - \langle \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k), \mathrm{T} - \mathrm{Id}\rangle_{L^2(\mu_k)} \\
&= \mathcal{F}^+(\mathrm{T}_{\#}\mu_k) - \mathcal{F}^-(\mathrm{T}_{\#}\mu_k) + \mathcal{F}^-(\mathrm{T}_{\#}\mu_k) - \mathcal{F}^-(\mu_k) - \langle \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k), \mathrm{T} - \mathrm{Id}\rangle_{L^2(\mu_k)} \\
&= \mathcal{F}(\mathrm{T}_{\#}\mu_k) + \mathrm{D}^{\mu_k}_{\mathcal{F}^-}(\mathrm{T}, \mathrm{Id})
\end{aligned}$$

$$\mathrm{T}_{k+1} = \underset{\mathrm{T}\in L^2(\mu_k)}{\operatorname{argmin}} \ \mathrm{D}^{\mu_k}_{\mathcal{F}^-}(\mathrm{T}, \mathrm{Id}) + \mathcal{F}(\mathrm{T}_{\#}\mu_k)$$

$\rightarrow$ Linear rate for $\mathcal{F}$ α-convex relative to $\mathcal{F}^-$ along a suitable curve

- **Equivalence with Mirror Descent**:

$$\mathrm{T}_{k+1} = \underset{\mathrm{T}\in L^2(\mu_k)}{\operatorname{argmin}} \ \mathrm{D}^{\mu_k}_{\mathcal{F}^+}(\mathrm{T}, \mathrm{Id}) + \langle \nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k), \mathrm{T} - \mathrm{Id}\rangle_{L^2(\mu_k)}$$

$\rightarrow$ Sub-linear ($\alpha = 0$) and linear rates ($\alpha > 0$) for $\mathcal{F}$ β-smooth ($\beta \leq 1$) and α-convex relative to $\mathcal{F}^+$ along suitable curves (Bonet et al., 2024)

Descent Lemma

Define

$$\mathcal{D}^k_{\mathcal{F}^+} := \mathcal{F}(\mu_k) - \min_{\mathrm{T}\in L^2(\mu_k)} \mathcal{F}^+(\mathrm{T}_\#\mu_k) - \mathcal{F}^-(\mu_k) - \langle \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k), \mathrm{T} - \mathrm{Id}\rangle_{L^2(\mu_k)}$$
$$= \mathcal{F}^+(\mu_k) - \mathcal{F}^+(\mu_{k+1}) - \langle \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k), \mathrm{Id} - \mathrm{T}_{k+1}\rangle_{L^2(\mu_k)}$$

Properties:

- $\mathcal{D}^k_{\mathcal{F}^+} \geq 0$
- If $\mathcal{F}^+$ W_2-differentiable, $\nabla_{\mathrm{W}_2}\mathcal{F}^+(\mu_{k+1}) \circ \mathrm{T}_{k+1} = \nabla_{\mathrm{W}_2}\mathcal{F}^-(\mu_k)$ and

$$\mathcal{D}^k_{\mathcal{F}^+} = \mathrm{D}^{\mu_k}_{\mathcal{F}^+}(\mathrm{Id}, \mathrm{T}_{k+1})$$

Descent Lemma

$$\forall k \geq 0,\ \mathcal{F}(\mu_{k+1}) = \mathcal{F}(\mu_k) - \mathrm{D}^{\mu_k}_{\mathcal{F}^-}(\mathrm{T}_{k+1}, \mathrm{Id}) - \mathcal{D}^k_{\mathcal{F}^+}$$

- If $\mathcal{F}^+$ W_2-differentiable, $\mathcal{D}^k_{\mathcal{F}^+} = 0 \implies \nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k) = 0$

Convergence to Stationary Points

As $\mathcal{D}^k_{\mathcal{F}^+} = \mathcal{F}(\mu_k) - \mathcal{F}(\mu_{k+1}) - \mathrm{D}^{\mu_k}_{\mathcal{F}^-}(\mathrm{T}_{k+1}, \mathrm{Id})$,

$$
\begin{aligned}
0 \leq \min_{0 \leq k \leq K-1} \mathcal{D}^k_{\mathcal{F}^+} &\leq \frac{1}{K} \sum_{k=0}^{K-1} \mathcal{D}^k_{\mathcal{F}^+} \\
&= \frac{1}{K} \sum_{k=0}^{K-1} \left(\mathcal{F}(\mu_k) - \mathcal{F}(\mu_{k+1}) - \mathrm{D}^{\mu_k}_{\mathcal{F}^-}(\mathrm{T}_{k+1}, \mathrm{Id}) \right) \right) \\
&\leq \frac{\mathcal{F}(\mu_0) - \inf \mathcal{F}}{K}
\end{aligned}
$$

Convergence to Stationary Points

Assumptions:

- Let $\alpha^+, \alpha^- \geq 0$, $\alpha^+ + \alpha^- > 0$
- $\mathcal{F}^+$ α^+-totally convex
- $\mathcal{F}^-$ α^--totally convex

Then,

$$\min_{0 \leq k \leq K-1} \mathrm{W}_2^2(\mu_k, \mu_{k+1}) \leq \frac{2}{\alpha^- + \alpha^+} \frac{\mathcal{F}(\mu_0) - \mathcal{F}(\mu_K)}{K}$$

Moreover, if

$$\|\nabla_{\mathrm{W}_2} \mathcal{F}^+(\mu_{k+1}) \circ \mathrm{T}_{k+1} - \nabla_{\mathrm{W}_2} \mathcal{F}^+(\mu_k)\|_{L^2(\mu_k)} \leq L \|\mathrm{T}_{k+1} - \mathrm{Id}\|_{L^2(\mu_k)},$$

$$\min_{0 \leq k \leq K-1} \|\nabla_{\mathrm{W}_2} \mathcal{F}(\mu_k)\|_{L^2(\mu_k)}^2 \leq \frac{L}{2(\alpha^+ + \alpha^-)} \frac{\mathcal{F}(\mu_0) - \mathcal{F}(\mu_K)}{K}$$

Application to MMD

Let $k(x,y) = \psi(x-y) = q(\|x-y\|_2^2)$. Define

$$\begin{cases} \underline{\lambda}[q] := \inf_s \min\{2q'(s),\ 2q'(s) + 4sq''(s)\}, \\ \overline{\Lambda}[q] := \sup_s \max\{2q'(s), 2q'(s) + 4sq''(s)\}. \end{cases}$$

Assume $q = q_+ - q_-$ such that $\underline{\lambda}[q_+], \underline{\lambda}[q_-] \geq 0$,

- $\psi_+(x) = q_+(\|x\|_2^2)$ and $\psi_-(x) = q_-(\|x\|_2^2)$ are $\alpha^+ = \underline{\lambda}[q_+]$ and $\alpha^- = \underline{\lambda}[q_-]$ convex
- $\mathcal{F}^+$ and $\mathcal{F}^-$ are α^- and α^+ convex
- If $\overline{\Lambda}[q_+], \overline{\Lambda}[q_-] < \infty$, the Lipschitz condition

$$\|\nabla_{\mathrm{W}_2}\mathcal{F}^+(\mu_{k+1}) \circ \mathrm{T}_{k+1} - \nabla_{\mathrm{W}_2}\mathcal{F}^+(\mu_k)\|_{L^2(\mu_k)} \leq L\|\mathrm{T}_{k+1} - \mathrm{Id}\|_{L^2(\mu_k)}$$

holds for $L = \sqrt{2} \cdot \overline{\Lambda}[q_+] + \overline{\Lambda}[q_-]$

Application to MMD

Let $k(x,y) = \psi(x-y) = q(\|x-y\|_2^2)$. Define

$$\begin{cases} \underline{\lambda}[q] \coloneqq \inf_s \min\{2q'(s),\, 2q'(s) + 4sq''(s)\}, \\ \overline{\Lambda}[q] \coloneqq \sup_s \max\{2q'(s), 2q'(s) + 4sq''(s)\}. \end{cases}$$

Kernel	q_+	q_-	$\underline{\lambda}[q_+]$	$\underline{\lambda}[q_-]$	$\overline{\Lambda}[q_+]$	$\overline{\Lambda}[q_-]$	Ω
Gauss-Jordan	$e^{-s/(2h)} + s/2h$	$s/2h$	0	$1/h$	$\dfrac{1+2e^{-1/3}}{h}$	$1/h$	$\mathbb{R}^d$
Gauss-cosh/sinh	$\cosh(s)$	$\sinh(s)$	0	$1/h$	(49)	(50)	compact

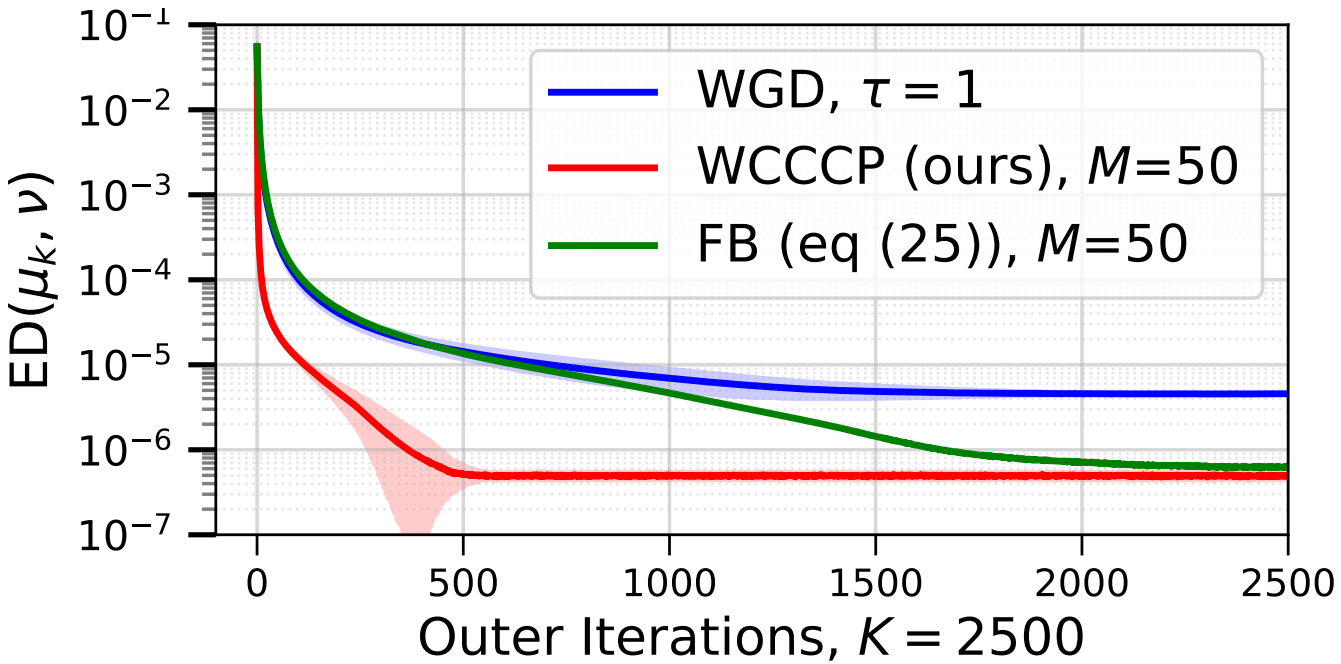
Table of Contents

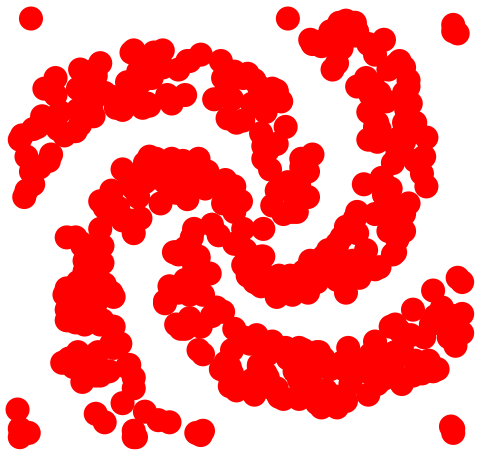
Energy Distance

$\psi(z) = -\|z\|_2$, *i.e.*

$$\frac{1}{2}\mathrm{ED}(\mu,\nu) = -\frac{1}{2}\iint \|x-y\|_2 \,\mathrm{d}\mu(x)\mathrm{d}\mu(y) + \int V \,\mathrm{d}\mu + c(\nu),$$

with $V(\cdot) = \int \|\cdot - y\|_2 \,\mathrm{d}\nu(y)$.
$\rightarrow$ DC decomposition with $\psi_+ = 0$, $\psi_-(z) = \|z\|_2$.



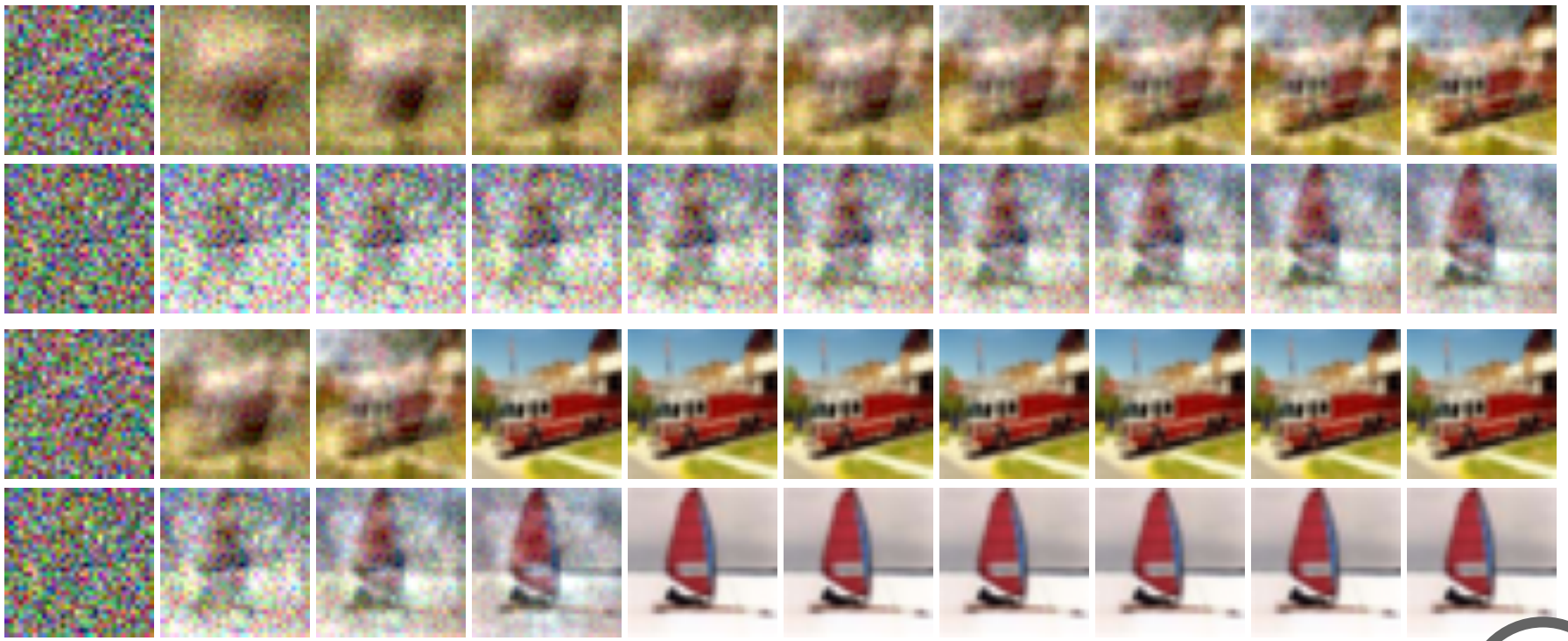


Energy Distance

$\psi(z) = -\|z\|_2$, *i.e.*

$$\frac{1}{2}\mathrm{ED}(\mu,\nu) = -\frac{1}{2}\iint \|x-y\|_2 \,\mathrm{d}\mu(x)\mathrm{d}\mu(y) + \int V \,\mathrm{d}\mu + c(\nu),$$

with $V(\cdot) = \int \|\cdot - y\|_2 \,\mathrm{d}\nu(y)$.
$\rightarrow$ DC decomposition with $\psi_+ = 0$, $\psi_-(z) = \|z\|_2$.

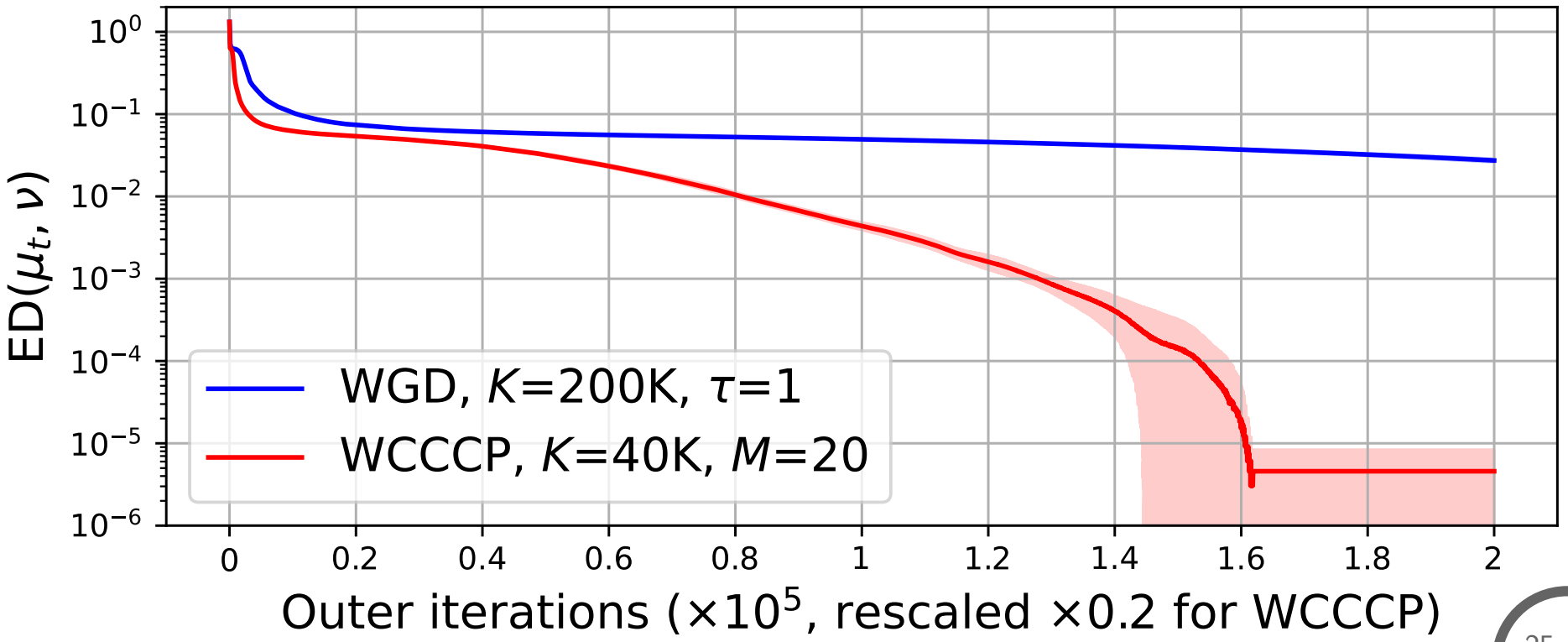


Energy Distance

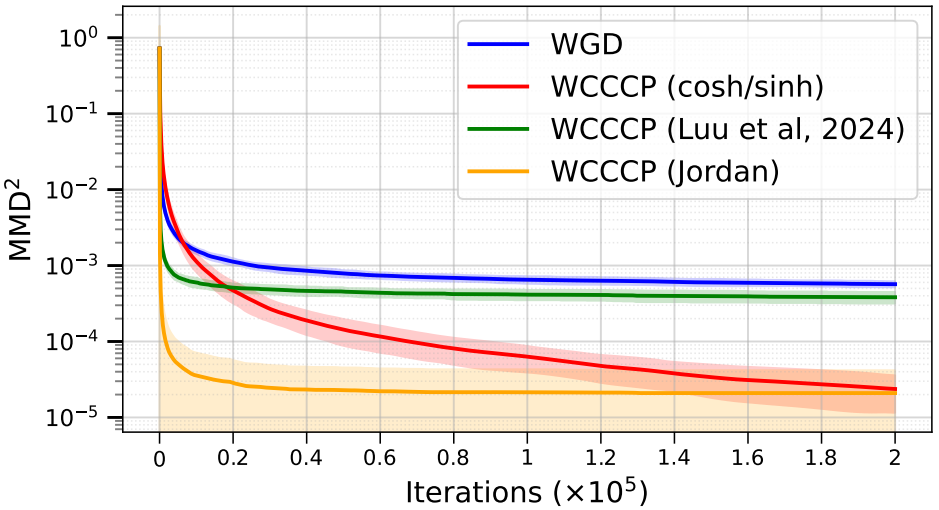
$\psi(z) = -\|z\|_2$, *i.e.*

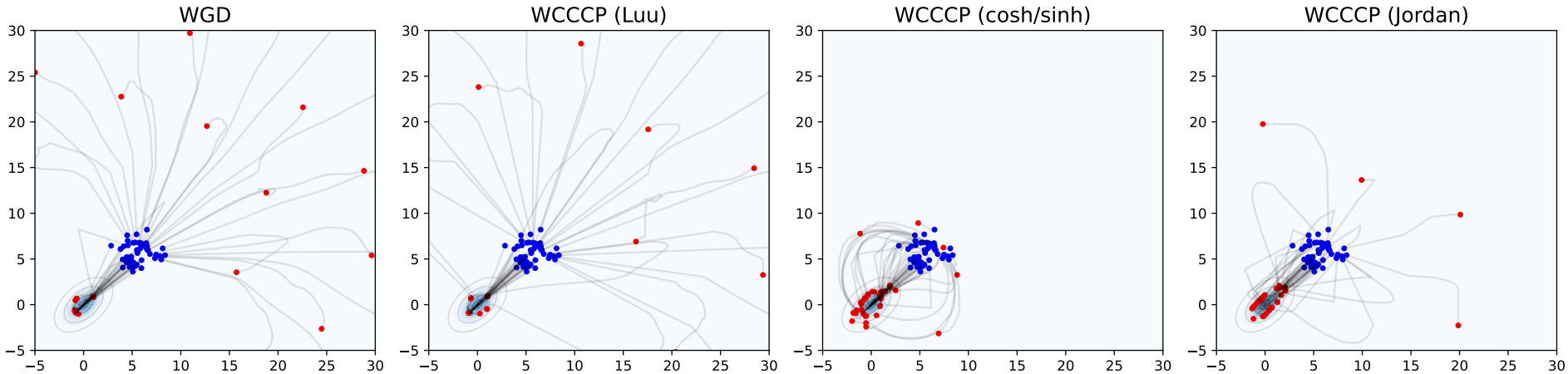
$$\frac{1}{2}\mathrm{ED}(\mu,\nu) = -\frac{1}{2}\iint \|x-y\|_2 \,\mathrm{d}\mu(x)\mathrm{d}\mu(y) + \int V \,\mathrm{d}\mu + c(\nu),$$

with $V(\cdot) = \int \|\cdot - y\|_2 \,\mathrm{d}\nu(y)$.
$\rightarrow$ DC decomposition with $\psi_+ = 0$, $\psi_-(z) = \|z\|_2$.



MMD with Gaussian Kernel





Conclusion

Conclusion:

- CCCP on $\mathcal{P}_2(\mathbb{R}^d)$
- Convergence analysis
- Application to the MMD

Perspectives:

- Better understanding the impact of the choice of DC decomposition
- Automatic, adaptative DC decomposition
- Application to other functionals

Conclusion

Conclusion:

- CCCP on $\mathcal{P}_2(\mathbb{R}^d)$
- Convergence analysis
- Application to the MMD

Perspectives:

- Better understanding the impact of the choice of DC decomposition
- Automatic, adaptative DC decomposition
- Application to other functionals

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, 2008.

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.

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.

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

References II

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.

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.

Jean-Baptiste Hiriart-Urruty. Generalized Differentiability, Duality and Optimization for Problems Dealing with Differences of Convex Functions. In J. Ponstein, editor, *Convexity and Duality in Optimization*, volume 256 of *Lecture Notes in Economics and Mathematical Systems*, pages 37–70. Springer, Berlin, 1985.

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 III

Hoang Phuc Hau Luu, Hanlin Yu, Bernardo Williams, Petrus Mikkola, Marcelo Hartmann, Kai Puolamäki, and Arto Klami. Non-geodesically-convex optimization in the Wasserstein space. *Advances in Neural Information Processing Systems*, 37:16772–16809, 2024.

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.

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.

Alan L Yuille and Anand Rangarajan. The concave-convex procedure (cccp). *Advances in neural information processing systems*, 14, 2001.