

# Mirror (and Preconditioned Gradient) Descent on the Wasserstein Space

Clément Bonet<sup>1</sup>

Joint work with Théo Uscidda<sup>1</sup>, Adam David<sup>2</sup>,  
Pierre-Cyril Aubin-Frankowski<sup>3</sup>, Anna Korba<sup>1</sup>

<sup>1</sup>ENSAE, CREST, Institut Polytechnique de Paris

<sup>2</sup>TU Berlin

<sup>3</sup>TU Wien

Level Set Seminar

05/08/2024

# Motivations

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

**Goal:**

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

**Applications:**

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

## Example of functionals

- Free energies:  $\mathcal{F}(\mu) = \int V d\mu + \iint W(x, y) d\mu(x)d\mu(y) + \mathcal{H}(\mu)$  where  $\mathcal{H}(\mu) = \int \log(\mu(x)) d\mu(x)$  for  $\mu \ll \text{Leb}$
- $\mathcal{F}(\mu) = \text{KL}(\mu||\nu) = \int V d\mu + \mathcal{H}(\mu)$  for sampling from  $\nu \propto e^{-V(x)}$
- $\mathcal{F}(\mu) = D(\mu, \nu)$  for sampling from  $\nu$

# Table of Contents

Detour by  $\mathbb{R}^d$

Wasserstein Gradient Flows

Mirror Descent

Preconditioned Gradient Descent

Applications

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

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

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

$$\frac{dx_t}{dt} = -\nabla f(x_t), \quad x_0 = x_0$$

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

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

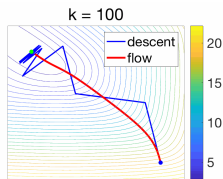
**Goal:**  $\min_{x \in \mathbb{R}^d} f(x)$  via gradient flow

$$\frac{dx_t}{dt} = -\nabla f(x_t), \quad x_0 = x_0$$

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

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

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



From (Bach, 2020)

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

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

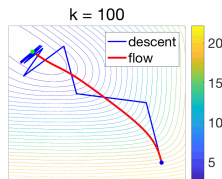
**Goal:**  $\min_{x \in \mathbb{R}^d} f(x)$  via gradient flow

$$\frac{dx_t}{dt} = -\nabla f(x_t), \quad x_0 = x_0$$

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

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

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



From (Bach, 2020)

## Convergence Analysis (Garrigos and Gower, 2023)

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

Reminder:

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

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

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

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

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

## Definition (Bregman Divergence)

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

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

**Mirror Descent** algorithm:

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

Remark: For  $\phi(x) = \frac{1}{2} \|x\|_2^2$ , MD = GD and  $d_\phi(x, y) = \frac{1}{2} \|x - y\|_2^2$



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

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

## Definition (Bregman Divergence)

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

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

**Mirror Descent** algorithm:

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

Remark: For  $\phi(x) = \frac{1}{2} \|x\|_2^2$ , MD = GD and  $d_\phi(x, y) = \frac{1}{2} \|x - y\|_2^2$

## Convergence analysis (Lu et al., 2018)

- $f$   $\beta$ -smooth relative to  $\phi$ , i.e.  $d_f(x, y) \leq \beta d_\phi(x, y)$  (equivalently  $\beta\phi - f$  convex)  $\implies f(x_{k+1}) \leq f(x_k) - \beta d_\phi(x_k, x_{k+1})$
- $f$   $\beta$ -smooth and  $\alpha$ -convex relative to  $\phi$ , i.e.  $\alpha d_\phi(x, y) \leq d_f(x, y)$  (equivalently  $f - \alpha\phi$  convex)  $\implies f(x_k) - f(x^*) \leq \frac{\beta - \alpha}{k} d_\phi(x^*, x_0)$

# Preconditioned Gradient Descent (Maddison et al., 2021)

Let  $h : \mathbb{R}^d \rightarrow \mathbb{R}$  strictly convex,  $g : \mathbb{R}^d \rightarrow \mathbb{R}$ .

**Preconditioned Gradient Descent** scheme:

$$\begin{aligned} \forall k \geq 0, y_{k+1} &= y_k - \tau \nabla h^*(\nabla g(y_k)) \\ &= \operatorname{argmin}_{y \in \mathbb{R}^d} h\left(\frac{y_k - y}{\tau}\right) \tau + \langle \nabla g(y_k), y - y_k \rangle \end{aligned}$$

# Preconditioned Gradient Descent (Maddison et al., 2021)

Let  $h : \mathbb{R}^d \rightarrow \mathbb{R}$  strictly convex,  $g : \mathbb{R}^d \rightarrow \mathbb{R}$ .

**Preconditioned Gradient Descent** scheme:

$$\begin{aligned}\forall k \geq 0, y_{k+1} &= y_k - \tau \nabla h^*(\nabla g(y_k)) \\ &= \operatorname{argmin}_{y \in \mathbb{R}^d} h\left(\frac{y_k - y}{\tau}\right) \tau + \langle \nabla g(y_k), y - y_k \rangle\end{aligned}$$

Closely related to MD (Kim et al., 2023) as for  $g = \phi^*$ ,  $h^* = f$ ,  $y = \nabla \phi(x)$ ,

$$\nabla \phi(x_{k+1}) = \nabla \phi(x_k) - \tau \nabla f(x_k) \iff x_{k+1} = \nabla \phi^*(\nabla \phi(x_k) - \tau \nabla f(x_k)).$$

# Preconditioned Gradient Descent (Maddison et al., 2021)

Let  $h : \mathbb{R}^d \rightarrow \mathbb{R}$  strictly convex,  $g : \mathbb{R}^d \rightarrow \mathbb{R}$ .

**Preconditioned Gradient Descent** scheme:

$$\begin{aligned} \forall k \geq 0, y_{k+1} &= y_k - \tau \nabla h^*(\nabla g(y_k)) \\ &= \operatorname{argmin}_{y \in \mathbb{R}^d} h\left(\frac{y_k - y}{\tau}\right) \tau + \langle \nabla g(y_k), y - y_k \rangle \end{aligned}$$

Closely related to MD (Kim et al., 2023) as for  $g = \phi^*$ ,  $h^* = f$ ,  $y = \nabla \phi(x)$ ,

$$\nabla \phi(x_{k+1}) = \nabla \phi(x_k) - \tau \nabla f(x_k) \iff x_{k+1} = \nabla \phi^*(\nabla \phi(x_k) - \tau \nabla f(x_k)).$$

## Convergence analysis (Maddison et al., 2021)

- $h^*$   $\beta$ -smooth relative to  $g^*$   $\implies h^*(\nabla g(y_{k+1})) \leq h^*(\nabla g(y_k)) - \beta d_g(y_{k+1}, y_k)$
- $h^*$   $\beta$ -smooth and  $\alpha$ -convex relative to  $g^*$ 
  - $\implies \forall k \geq 1, h^*(\nabla g(y_k)) - h^*(0) \leq \frac{\alpha - \beta}{k} (g(y_0) - g(y^*))$
  - $\implies \forall k \geq 0, g(y_k) - g(y^*) \leq (1 - \alpha/\beta)^k (g(y_0) - g(y^*))$

# Table of Contents

Detour by  $\mathbb{R}^d$

**Wasserstein Gradient Flows**

Mirror Descent

Preconditioned Gradient Descent

Applications

# Wasserstein Geometry

## Definition (Wasserstein distance)

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

$$W_2^2(\mu, \nu) = \inf_{\gamma \in \Pi(\mu, \nu)} \int \|x - y\|_2^2 d\gamma(x, y).$$

## Properties:

- $W_2$  distance,  $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ : Wasserstein space
- **Brenier's theorem:** If  $\mu \ll \text{Leb}$ , then there exists a unique  $T_\mu^\nu$  such that
  1.  $(T_\mu^\nu)_\# \mu = \nu$  ( $T_\# \mu(A) = \mu(T^{-1}(A))$ ) for all  $A \subset \mathbb{R}^d$
  2.  $W_2^2(\mu, \nu) = \int \|x - T_\mu^\nu(x)\|_2^2 d\mu(x) = \|\text{Id} - T_\mu^\nu\|_{L^2(\mu)}^2$
- Riemannian structure:
  - Geodesics between  $\mu \ll \text{Leb}$  and  $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ :  $\forall t \in [0, 1]$ ,  $\mu_t = ((1-t)\text{Id} + tT_\mu^\nu)_\# \mu$
  - 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  $L^2(\mu) = \{f \in \mathbb{R}^d \rightarrow \mathbb{R}^d, \int \|f(x)\|_2^2 d\mu(x) < \infty\}$ .

# Wasserstein Gradient

## Definition (Wasserstein gradient (Bonnet, 2019))

Let  $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ .  $\nabla_{W_2} \mathcal{F}(\mu) \in L^2(\mu)$  is a Wasserstein gradient of  $\mathcal{F}$  at  $\mu$  if for any  $\nu \in \mathcal{P}_2(\mathbb{R}^d)$  and any optimal coupling  $\gamma \in \Pi_o(\mu, \nu)$ ,

$$\mathcal{F}(\nu) = \mathcal{F}(\mu) + \int \langle \nabla_{W_2} \mathcal{F}(\mu)(x), y - x \rangle d\gamma(x, y) + o(W_2(\mu, \nu)).$$

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

### 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  $\gamma \in \Pi(\mu, \nu)$ ,

$$\mathcal{F}(\nu) = \mathcal{F}(\mu) + \int \langle \nabla_{W_2} \mathcal{F}(\mu)(x), y - x \rangle d\gamma(x, y) + o\left(\sqrt{\int \|x - y\|_2^2 d\gamma(x, y)}\right).$$

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

$$\mathcal{F}(T_\# \mu) = \mathcal{F}(\mu) + \langle \nabla_{W_2} \mathcal{F}(\mu), T - \text{Id} \rangle_{L^2(\mu)} + o(\|T - \text{Id}\|_{L^2(\mu)})$$

# Wasserstein Gradient

## Example of functionals

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

$$\nabla_{W_2} \mathcal{V}(\mu) = \nabla V$$

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

$$\nabla_{W_2} \mathcal{W}(\mu) = \nabla W \star \mu$$

## Negative entropy

$\mathcal{H}(\mu) = \int \log(\mu(x)) d\mu(x)$  not  $W_2$ -differentiable but can consider subgradients under regularity assumptions:

$$\forall x \in \mathbb{R}^d, \nabla_{W_2} \mathcal{H}(\mu)(x) = \nabla \log \mu(x)$$



# Wasserstein Gradient Flows (Ambrosio et al., 2005)

**Wasserstein gradient flow of  $\mathcal{F}$ :** curve  $t \mapsto \mu_t$  satisfying (weakly)

$$\partial_t \mu_t = \operatorname{div}(\mu_t \nabla_{W_2} \mathcal{F}(\mu_t)).$$

Particles:  $x_t \sim \mu_t \iff \frac{dx_t}{dt} = -\nabla_{W_2} \mathcal{F}(\mu_t)(x_t)$ .

# Wasserstein Gradient Flows (Ambrosio et al., 2005)

**Wasserstein gradient flow of  $\mathcal{F}$ :** curve  $t \mapsto \mu_t$  satisfying (weakly)

$$\partial_t \mu_t = \operatorname{div}(\mu_t \nabla_{W_2} \mathcal{F}(\mu_t)).$$

Particles:  $x_t \sim \mu_t \iff \frac{dx_t}{dt} = -\nabla_{W_2} \mathcal{F}(\mu_t)(x_t)$ .

**Time discretization** of the flow:

- Implicit/Backward (JKO) scheme (Jordan et al., 1998):

$$\mu_{k+1} = \operatorname{argmin}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \frac{1}{2} W_2^2(\mu, \mu_k) + \tau \mathcal{F}(\mu)$$

If  $\mu_k \ll \operatorname{Leb}$ ,  $\mu_{k+1} = T_{\#} \mu_k$  with

$$T = \operatorname{argmin}_{T \in L^2(\mu_k)} \frac{1}{2} \|T - \operatorname{Id}\|_{L^2(\mu_k)}^2 + \tau \mathcal{F}(T_{\#} \mu_k)$$

# Wasserstein Gradient Flows (Ambrosio et al., 2005)

**Wasserstein gradient flow of  $\mathcal{F}$ :** curve  $t \mapsto \mu_t$  satisfying (weakly)

$$\partial_t \mu_t = \operatorname{div}(\mu_t \nabla_{W_2} \mathcal{F}(\mu_t)).$$

Particles:  $x_t \sim \mu_t \iff \frac{dx_t}{dt} = -\nabla_{W_2} \mathcal{F}(\mu_t)(x_t)$ .

**Time discretization** of the flow:

- Implicit/Backward (JKO) scheme (Jordan et al., 1998):

$$\mu_{k+1} = \operatorname{argmin}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \frac{1}{2} W_2^2(\mu, \mu_k) + \tau \mathcal{F}(\mu)$$

If  $\mu_k \ll \operatorname{Leb}$ ,  $\mu_{k+1} = \mathbb{T}_{\#} \mu_k$  with

$$\mathbb{T} = \operatorname{argmin}_{\mathbb{T} \in L^2(\mu_k)} \frac{1}{2} \|\mathbb{T} - \operatorname{Id}\|_{L^2(\mu_k)}^2 + \tau \mathcal{F}(\mathbb{T}_{\#} \mu_k)$$

- Explicit/Forward scheme

$$\begin{cases} \mathbb{T}_{k+1} = \operatorname{argmin}_{\mathbb{T} \in L^2(\mu_k)} \frac{1}{2} \|\mathbb{T} - \operatorname{Id}\|_{L^2(\mu_k)}^2 + \tau \langle \nabla_{W_2} \mathcal{F}(\mu_k), \mathbb{T} - \operatorname{Id} \rangle_{L^2(\mu_k)} \\ \mu_{k+1} = (\mathbb{T}_{k+1})_{\#} \mu_k \end{cases}$$

Taking the FOC:  $\mathbb{T}_{k+1} = \operatorname{Id} - \tau \nabla_{W_2} \mathcal{F}(\mu_k)$

# Wasserstein Gradient Flows (Ambrosio et al., 2005)

**Wasserstein gradient flow of  $\mathcal{F}$ :** curve  $t \mapsto \mu_t$  satisfying (weakly)

$$\partial_t \mu_t = \operatorname{div}(\mu_t \nabla_{W_2} \mathcal{F}(\mu_t)).$$

Particles:  $x_t \sim \mu_t \iff \frac{dx_t}{dt} = -\nabla_{W_2} \mathcal{F}(\mu_t)(x_t)$ .

**Time discretization** of the flow:

- Implicit/Backward (JKO) scheme (Jordan et al., 1998):

$$\mu_{k+1} = \operatorname{argmin}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \frac{1}{2} W_2^2(\mu, \mu_k) + \tau \mathcal{F}(\mu)$$

If  $\mu_k \ll \operatorname{Leb}$ ,  $\mu_{k+1} = T_{\#} \mu_k$  with

$$T = \operatorname{argmin}_{T \in L^2(\mu_k)} \frac{1}{2} \|T - \operatorname{Id}\|_{L^2(\mu_k)}^2 + \tau \mathcal{F}(T_{\#} \mu_k)$$

- Explicit/Forward scheme

$$\begin{cases} T_{k+1} = \operatorname{Id} - \tau \nabla_{W_2} \mathcal{F}(\mu_k) \\ \mu_{k+1} = (T_{k+1})_{\#} \mu_k \end{cases}$$

**Particle approximation:**  $\hat{\mu}_k^n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i^k}$ .

# Contributions

Study schemes of the form

$$\begin{cases} \mathbb{T}_{k+1} = \operatorname{argmin}_{\mathbb{T} \in L^2(\mu_k)} d(\mathbb{T}, \operatorname{Id}) + \tau \langle \nabla_{W_2} \mathcal{F}(\mu_k), \mathbb{T} - \operatorname{Id} \rangle_{L^2(\mu_k)} \\ \mu_{k+1} = (\mathbb{T}_{k+1}) \# \mu_k, \end{cases}$$

and provide **convergence conditions**.

Considered divergences:

- **Bregman divergence** on  $L^2(\mu)$ , extending **Mirror Descent** to  $\mathcal{P}_2(\mathbb{R}^d)$
- Divergence of the form  $d(\mathbb{T}, \mathbb{S}) = \int h(\mathbb{T}(x) - \mathbb{S}(x)) \, d\mu(x)$ , extending **Preconditioned Gradient Descent** to  $\mathcal{P}_2(\mathbb{R}^d)$

# Table of Contents

Detour by  $\mathbb{R}^d$

Wasserstein Gradient Flows

**Mirror Descent**

Preconditioned Gradient Descent

Applications

## Background on $L^2(\mu)$

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

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

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

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

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

In this case, if  $\phi$  is  $W_2$ -differentiable, then  $\phi_\mu$  is Fréchet differentiable and  $\nabla \phi_\mu(T) = \nabla_{W_2} \phi(T \# \mu) \circ T$

- Let  $\phi_\mu, \psi_\mu : L^2(\mu) \rightarrow \mathbb{R}$  convex.
  - $\phi_\mu$  is  $\beta$ -smooth relative to  $\psi_\mu$  if for all  $T, S \in L^2(\mu)$ ,  $d_{\phi_\mu}(T, S) \leq \beta d_{\psi_\mu}(T, S)$ .
  - $\phi_\mu$  is  $\alpha$ -convex relative to  $\psi_\mu$  if for all  $T, S \in L^2(\mu)$ ,  $d_{\phi_\mu}(T, S) \geq \alpha d_{\psi_\mu}(T, S)$ .

## Convexity on $\mathcal{P}_2(\mathbb{R}^d)$

Let  $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$ ,  $\mu_0 \ll \text{Leb}$ , and  $\mu_t = ((1-t)\text{Id} + t\mathbb{T}_{\mu_0}^{\mu_1})\# \mu_0$ .

- $\mathcal{F}$  is  $\alpha$ -geodesically convex if  $t \mapsto \mathcal{F}(\mu_t)$  is  $\alpha$ -convex, i.e. for all  $t \in [0, 1]$ ,

$$\mathcal{F}(\mu_t) \leq (1-t)\mathcal{F}(\mu_0) + t\mathcal{F}(\mu_1) - \frac{\alpha t(1-t)}{2} W_2^2(\mu_0, \mu_1),$$

or equivalently

$$\begin{aligned} \frac{\alpha}{2} W_2^2(\mu_0, \mu_1) &= \frac{\alpha}{2} \|\mathbb{T}_{\mu_0}^{\mu_1} - \text{Id}\|_{L^2(\mu_0)}^2 \leq \mathcal{F}(\mu_1) - \mathcal{F}(\mu_0) - \langle \nabla_{W_2} \mathcal{F}(\mu_0), \mathbb{T}_{\mu_0}^{\mu_1} - \text{Id} \rangle_{L^2(\mu_0)} \\ &= d_{\tilde{\mathcal{F}}_{\mu_0}}(\mathbb{T}_{\mu_0}^{\mu_1}, \text{Id}) \end{aligned}$$

with  $\tilde{\mathcal{F}}_{\mu}(\mathbb{T}) = \mathcal{F}(\mathbb{T}\# \mu)$ .



## Convexity on $\mathcal{P}_2(\mathbb{R}^d)$

Let  $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$ ,  $\mu_0 \ll \text{Leb}$ , and  $\mu_t = ((1-t)\text{Id} + t\mathbb{T}_{\mu_0}^{\mu_1})\# \mu_0$ .

- $\mathcal{F}$  is  $\alpha$ -geodesically convex if  $t \mapsto \mathcal{F}(\mu_t)$  is  $\alpha$ -convex, i.e. for all  $t \in [0, 1]$ ,

$$\mathcal{F}(\mu_t) \leq (1-t)\mathcal{F}(\mu_0) + t\mathcal{F}(\mu_1) - \frac{\alpha t(1-t)}{2} W_2^2(\mu_0, \mu_1),$$

or equivalently

$$\begin{aligned} \frac{\alpha}{2} W_2^2(\mu_0, \mu_1) &= \frac{\alpha}{2} \|\mathbb{T}_{\mu_0}^{\mu_1} - \text{Id}\|_{L^2(\mu_0)}^2 \leq \mathcal{F}(\mu_1) - \mathcal{F}(\mu_0) - \langle \nabla_{W_2} \mathcal{F}(\mu_0), \mathbb{T}_{\mu_0}^{\mu_1} - \text{Id} \rangle_{L^2(\mu_0)} \\ &= d_{\tilde{\mathcal{F}}_{\mu_0}}(\mathbb{T}_{\mu_0}^{\mu_1}, \text{Id}) \end{aligned}$$

with  $\tilde{\mathcal{F}}_{\mu}(\mathbb{T}) = \mathcal{F}(\mathbb{T}\#\mu)$ .

### Definition

Let  $\mathcal{F}, \mathcal{G} : \mathcal{P}_2(\mathbb{R}^d) \rightarrow \mathbb{R}$ ,  $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ ,  $\mathbb{T}, \mathbb{S} \in L^2(\mu)$ ,  $\mu_t = (\mathbb{T}_t)\#\mu$  with  $\mathbb{T}_t = (1-t)\mathbb{S} + t\mathbb{T}$  for all  $t \in [0, 1]$ .

- $\mathcal{F}$   $\beta$ -smooth relative to  $\mathcal{G}$  along  $t \mapsto \mu_t$  if  $\forall s, t \in [0, 1]$ ,  $d_{\tilde{\mathcal{F}}_{\mu}}(\mathbb{T}_s, \mathbb{T}_t) \leq \beta d_{\tilde{\mathcal{G}}_{\mu}}(\mathbb{T}_s, \mathbb{T}_t)$ .
- $\mathcal{F}$   $\alpha$ -convex relative to  $\mathcal{G}$  along  $t \mapsto \mu_t$  if  $\forall s, t \in [0, 1]$ ,  $d_{\tilde{\mathcal{F}}_{\mu}}(\mathbb{T}_s, \mathbb{T}_t) \geq \alpha d_{\tilde{\mathcal{G}}_{\mu}}(\mathbb{T}_s, \mathbb{T}_t)$ .

# Mirror Descent on the Wasserstein Space

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

**Mirror Descent scheme:**

$$\begin{cases} \mathbb{T}_{k+1} = \operatorname{argmin}_{\mathbb{T} \in L^2(\mu_k)} d_{\phi_{\mu_k}}(\mathbb{T}, \operatorname{Id}) + \tau \langle \nabla_{W_2} \mathcal{F}(\mu_k), \mathbb{T} - \operatorname{Id} \rangle_{L^2(\mu_k)} \\ \mu_{k+1} = (\mathbb{T}_{k+1})\# \mu_k. \end{cases}$$

By FOC:  $\nabla \phi_{\mu_k}(\mathbb{T}_{k+1}) = \nabla \phi_{\mu_k}(\operatorname{Id}) - \tau \nabla_{W_2} \mathcal{F}(\mu_k)$

**Computing the scheme:**

- For  $\phi_\mu(\mathbb{T}) = \int V \circ \mathbb{T} \, d\mu$ ,  $\mathbb{T}_{k+1} = \nabla V^* \circ (\nabla V - \tau \nabla_{W_2} \mathcal{F}(\mu_k))$
- For  $\phi_\mu$  pushforward compatible:

$$\nabla_{W_2} \phi(\mu_{k+1}) \circ \mathbb{T}_{k+1} = \nabla_{W_2} \phi(\mu_k) - \tau \nabla_{W_2} \mathcal{F}(\mu_k)$$

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

- Other particular cases with closed-forms, e.g.  $\phi_\mu(\mathbb{T}) = \frac{1}{2} \|P_\mu \mathbb{T}\|_{L^2(\mu)}^2$  recovers SVGD (Liu and Wang, 2016) or EKS (Garbuno-Inigo et al., 2020).

## Continuous Formulation

Informally, for  $\phi_\mu$  pushforward compatible:

$$\begin{cases} \varphi(\mu_k) = \nabla_{\mathbb{W}_2} \phi(\mu_k) \\ \varphi(\mu_{k+1}) \circ \mathbf{T}_{k+1} = \varphi(\mu_k) - \tau \nabla_{\mathbb{W}_2} \mathcal{F}(\mu_k) \end{cases} \xrightarrow{\tau \rightarrow 0} \begin{cases} \varphi(\mu_t) = \nabla_{\mathbb{W}_2} \phi(\mu_t) \\ \frac{d}{dt} \varphi(\mu_t) = -\nabla_{\mathbb{W}_2} \mathcal{F}(\mu_t). \end{cases}$$

# Continuous Formulation

Informally, for  $\phi_\mu$  pushforward compatible:

$$\begin{cases} \varphi(\mu_k) = \nabla_{W_2} \phi(\mu_k) \\ \varphi(\mu_{k+1}) \circ T_{k+1} = \varphi(\mu_k) - \tau \nabla_{W_2} \mathcal{F}(\mu_k) \end{cases} \xrightarrow{\tau \rightarrow 0} \begin{cases} \varphi(\mu_t) = \nabla_{W_2} \phi(\mu_t) \\ \frac{d}{dt} \varphi(\mu_t) = -\nabla_{W_2} \mathcal{F}(\mu_t). \end{cases}$$

$$\frac{d}{dt} \varphi(\mu_t) = \frac{d}{dt} \nabla_{W_2} \phi(\mu_t) = H\phi_{\mu_t}(v_t),$$

with  $H\phi_{\mu_t} : L^2(\mu_t) \rightarrow L^2(\mu_t)$  Hessian operator defined such that

$$\frac{d^2}{dt^2} \varphi(\mu_t) = \langle H\phi_{\mu_t}(v_t), v_t \rangle_{L^2(\mu_t)} \quad \text{with} \quad \partial_t \mu_t + \text{div}(\mu_t v_t) = 0.$$

**Mirror flow:**

$$\partial_t \mu_t - \text{div}(\mu_t (H\phi_{\mu_t})^{-1} \nabla_{W_2} \mathcal{F}(\mu_t)) = 0.$$

**Related works:**

- For  $\phi(\mu) = \int V d\mu$ ,  $\mathcal{F}(\mu) = \text{KL}(\mu || \mu^*)$ , coincides with continuous formulation of Mirror Langevin ([Ahn and Chewi, 2021](#))
- For  $\phi = \mathcal{F}$ , coincides with Information Newton's flows ([Wang and Li, 2020](#))
- For  $\phi(\mu) = \frac{1}{2} W_2^2(\mu, \nu)$ ,  $\mathcal{F}(\mu) = \text{KL}(\mu || \mu^*)$ , coincides with Sinkhorn flows ([Deb et al., 2023](#))

## Descent Lemma

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

**Mirror Descent scheme:**

$$\begin{cases} \mathbb{T}_{k+1} = \operatorname{argmin}_{\mathbb{T} \in L^2(\mu_k)} d_{\phi_{\mu_k}}(\mathbb{T}, \operatorname{Id}) + \tau \langle \nabla_{W_2} \mathcal{F}(\mu_k), \mathbb{T} - \operatorname{Id} \rangle_{L^2(\mu_k)} \\ \mu_{k+1} = (\mathbb{T}_{k+1})_{\#} \mu_k. \end{cases}$$

### Proposition (Descent Lemma)

*Assumptions:*

- For all  $k \geq 0$ ,  $\mathcal{F}$  is  $\beta$ -smooth relative to  $\phi$  along  $t \mapsto ((1-t)\operatorname{Id} + t\mathbb{T}_{k+1})_{\#} \mu_k$

Then, for all  $k \geq 0$ ,

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

Remark:  $\beta$ -smoothness implies  $\beta d_{\phi_{\mu_k}}(\mathbb{T}_{k+1}, \operatorname{Id}) \geq d_{\tilde{\mathcal{F}}_{\mu_k}}(\mathbb{T}_{k+1}, \operatorname{Id})$

Sketch of the proof:

1. Apply  $\beta$ -smoothness
2. Apply 3-point inequality

# Convergence

Let  $\phi_\mu$  be pushforward compatible. Define the OT problem:

$$\begin{aligned} W_\phi(\nu, \mu) &= \inf_{\gamma \in \Pi(\nu, \mu)} \phi(\nu) - \phi(\mu) - \int \langle \nabla_{W_2} \phi(\mu)(y), x - y \rangle d\gamma(x, y) \\ &\leq d_{\phi_\eta}(\mathbb{T}, \mathbb{S}) \quad \text{for } (\mathbb{T}, \mathbb{S})_{\#} \eta \in \Pi(\nu, \mu) \end{aligned}$$

**Property:** If  $\mu \ll \text{Leb}$  and  $\nabla_{W_2} \phi(\mu)$  is invertible, then  $\gamma^* = (\mathbb{T}_{\phi_\mu}^{\mu, \nu}, \text{Id})_{\#} \mu$ .

## Proposition

*Assumptions:* Let  $\beta > 0, \alpha \geq 0$ . For all  $k \geq 0$ ,

- $\mathcal{F}$   $\beta$ -smooth relative to  $\phi$  along  $t \mapsto ((1-t)\text{Id} + t\mathbb{T}_{k+1})_{\#} \mu_k$
- $\mathcal{F}$   $\alpha$ -convex relative to  $\phi$  along  $t \mapsto ((1-t)\text{Id} + t\mathbb{T}_{\phi_{\mu_k}}^{\mu_k, \mu^*})_{\#} \mu_k$
- $d_{\phi_{\mu_k}}(\mathbb{T}_{\phi_{\mu_k}}^{\mu_k, \nu}, \text{Id}) = W_\phi(\nu, \mu_k)$  and  $d_{\phi_{\mu_k}}(\mathbb{T}_{\phi_{\mu_k}}^{\mu_k, \nu}, \mathbb{T}_{k+1}) \geq W_\phi(\nu, \mu_{k+1})$

Then, for all  $k \geq 1$ ,  $\mathcal{F}(\mu_k) - \mathcal{F}(\mu^*) \leq \frac{\beta - \alpha}{k} W_\phi(\mu^*, \mu_0)$ .

If  $\alpha > 0$ , for all  $k \geq 0$ ,  $W_\phi(\mu^*, \mu_k) \leq \left(1 - \frac{\alpha}{\beta}\right)^k W_\phi(\mu^*, \mu_0)$ .

3rd Assumption required to telescope  $d_{\phi_{\mu_k}}(\mathbb{T}_{\phi_{\mu_k}}^{\mu_k, \nu}, \text{Id}) = W_\phi(\nu, \mu_k)$  and

$d_{\phi_{\mu_k}}(\mathbb{T}_{\phi_{\mu_k}}^{\mu_k, \nu}, \mathbb{T}_{k+1}) \geq W_\phi(\nu, \mu_{k+1})$  as optimization done in different space  $L^2(\mu_k)$ .

# Table of Contents

Detour by  $\mathbb{R}^d$

Wasserstein Gradient Flows

Mirror Descent

**Preconditioned Gradient Descent**

Applications

## Preconditioned GD

Let  $h : \mathbb{R}^d \rightarrow \mathbb{R}$  strictly convex, proper and differentiable.

**Preconditioned Gradient Descent scheme:** Let  $\phi_{\mu}^h(\mathbb{T}) = \int h \circ \mathbb{T} \, d\mu$ ,

$$\begin{cases} \mathbb{T}_{k+1} = \operatorname{argmin}_{\mathbb{T} \in L^2(\mu_k)} \phi_{\mu_k}^h \left( \frac{\operatorname{Id} - \mathbb{T}}{\tau} \right) \tau + \langle \nabla_{W_2} \mathcal{F}(\mu_k), \mathbb{T} - \operatorname{Id} \rangle_{L^2(\mu_k)} \\ \mu_{k+1} = (\mathbb{T}_{k+1})_{\#} \mu_k \end{cases}$$

By FOC:  $\mathbb{T}_{k+1} = \operatorname{Id} - \tau \nabla h^* \circ \nabla_{W_2} \mathcal{F}(\mu_k)$



# Preconditioned GD

Let  $h : \mathbb{R}^d \rightarrow \mathbb{R}$  strictly convex, proper and differentiable.

**Preconditioned Gradient Descent scheme:** Let  $\phi_{\mu}^h(\mathbb{T}) = \int h \circ \mathbb{T} \, d\mu$ ,

$$\begin{cases} \mathbb{T}_{k+1} = \operatorname{argmin}_{\mathbb{T} \in L^2(\mu_k)} \phi_{\mu_k}^h \left( \frac{\operatorname{Id} - \mathbb{T}}{\tau} \right) \tau + \langle \nabla_{W_2} \mathcal{F}(\mu_k), \mathbb{T} - \operatorname{Id} \rangle_{L^2(\mu_k)} \\ \mu_{k+1} = (\mathbb{T}_{k+1})_{\#} \mu_k \end{cases}$$

By FOC:  $\mathbb{T}_{k+1} = \operatorname{Id} - \tau \nabla h^* \circ \nabla_{W_2} \mathcal{F}(\mu_k)$

## Proposition (Descent Lemma)

*Assumptions:* For all  $k \geq 0$ ,

- $\mathcal{F}$  convex along  $t \mapsto ((1-t)\mathbb{T}_{k+1} + t\operatorname{Id})_{\#} \mu_k$
- $d_{\phi_{\mu_k}^{h^*}}(\nabla_{W_2} \mathcal{F}(\mu_{k+1}) \circ \mathbb{T}_{k+1}, \nabla_{W_2} \mathcal{F}(\mu_k)) \leq \beta d_{\tilde{\mathcal{F}}_{\mu_k}}(\operatorname{Id}, \mathbb{T}_{k+1})$

*Then, for all  $k \geq 0$ ,*

$$\phi_{\mu_{k+1}}^{h^*}(\nabla_{W_2} \mathcal{F}(\mu_{k+1})) \leq \phi_{\mu_k}^{h^*}(\nabla_{W_2} \mathcal{F}(\mu_k)) - \beta d_{\tilde{\mathcal{F}}_{\mu_k}}(\mathbb{T}_{k+1}, \operatorname{Id}).$$

Assumptions: inequalities between  $d_{\phi} \rightarrow$  sufficient conditions using convexity?

# Preconditioned GD

Let  $h : \mathbb{R}^d \rightarrow \mathbb{R}$  strictly convex, proper and differentiable.

**Preconditioned Gradient Descent scheme:** Let  $\phi_\mu^h(\mathbb{T}) = \int h \circ \mathbb{T} \, d\mu$ ,

$$\begin{cases} \mathbb{T}_{k+1} = \operatorname{argmin}_{\mathbb{T} \in L^2(\mu_k)} \phi_{\mu_k}^h \left( \frac{\operatorname{Id} - \mathbb{T}}{\tau} \right) \tau + \langle \nabla_{W_2} \mathcal{F}(\mu_k), \mathbb{T} - \operatorname{Id} \rangle_{L^2(\mu_k)} \\ \mu_{k+1} = (\mathbb{T}_{k+1})_{\#} \mu_k \end{cases}$$

By FOC:  $\mathbb{T}_{k+1} = \operatorname{Id} - \tau \nabla h^* \circ \nabla_{W_2} \mathcal{F}(\mu_k)$

## Proposition

*Assumptions:* For all  $k \geq 0$ , denoting  $\bar{\mathbb{T}} = \operatorname{argmin}_{\mathbb{T}, \mathbb{T}_{\#} \mu_k = \mu^*} d_{\tilde{\mathcal{F}}_{\mu_k}}(\operatorname{Id}, \mathbb{T})$ ,

- $\mathcal{F}$  convex along  $t \mapsto ((1-t)\mathbb{T}_{k+1} + t\operatorname{Id})_{\#} \mu_k$
- $d_{\phi_{\mu_k}^{h^*}}(\nabla_{W_2} \mathcal{F}(\mu_{k+1}) \circ \mathbb{T}_{k+1}, \nabla_{W_2} \mathcal{F}(\mu_k)) \leq \beta d_{\tilde{\mathcal{F}}_{\mu_k}}(\operatorname{Id}, \mathbb{T}_{k+1})$
- $\alpha d_{\tilde{\mathcal{F}}_{\mu_k}}(\operatorname{Id}, \bar{\mathbb{T}}) \leq d_{\phi_{\mu_k}^{h^*}}(\nabla_{W_2} \mathcal{F}(\bar{\mathbb{T}}_{\#} \mu_k) \circ \bar{\mathbb{T}}, \nabla_{W_2} \mathcal{F}(\mu_k))$

Then, for all  $k \geq 1$ ,  $\phi_{\mu_k}^{h^*}(\nabla_{W_2} \mathcal{F}(\mu_k)) - h^*(0) \leq \frac{\beta - \alpha}{k} (\mathcal{F}(\mu_0) - \mathcal{F}(\mu^*))$ .

Moreover, assuming that  $h^*$  attains its minimum at 0 and  $\alpha > 0$ , for all  $k \geq 0$ ,  $\mathcal{F}(\mu_k) - \mathcal{F}(\mu^*) \leq (1 - \tau\alpha)^k (\mathcal{F}(\mu_0) - \mathcal{F}(\mu^*))$ .

# Table of Contents

Detour by  $\mathbb{R}^d$

Wasserstein Gradient Flows

Mirror Descent

Preconditioned Gradient Descent

**Applications**

# Showing Relative Smoothness and Convexity

Relative smoothness of  $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \rightarrow \mathbb{R}$  relative to  $\phi : \mathcal{P}_2(\mathbb{R}^d) \rightarrow \mathbb{R}$ ?

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

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

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

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

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

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

- For  $\mathcal{F} = \mathcal{G} + \mathcal{H}$ ,  $d_{\tilde{\mathcal{F}}_\mu} = d_{\tilde{\mathcal{G}}_\mu} + d_{\tilde{\mathcal{H}}_\mu}$  and  $\mathcal{F}$  1-convex relative to  $\mathcal{G}$  and  $\mathcal{H}$
- In general: look at the Hessian

# Mirror Descent on Interaction Energy

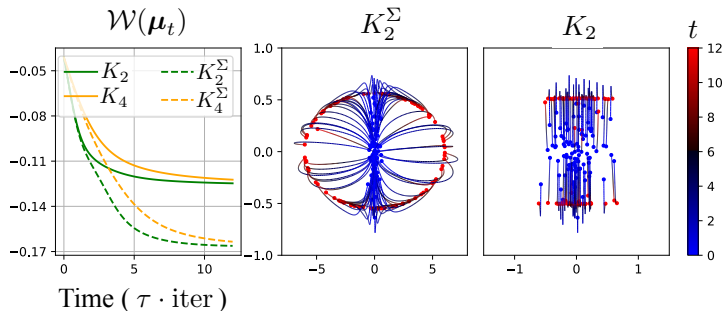
**Goal:** Let  $\Sigma \in S_d^{++}(\mathbb{R})$  possibly ill-conditioned,

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

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

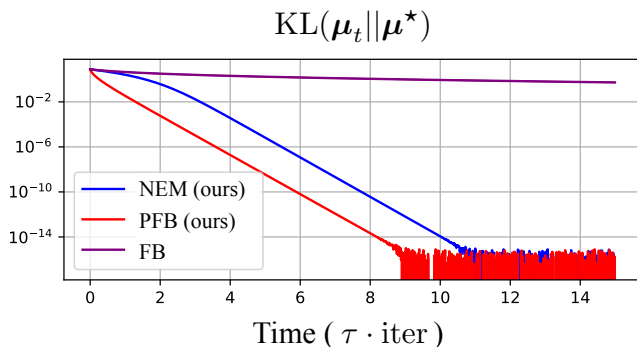
Goal:

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

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

Comparison between:

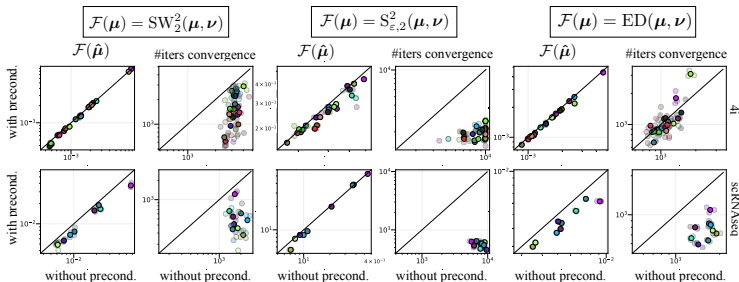
- Forward-Backward (FB) on the Bures-Wasserstein space (Diao et al., 2023)
- Preconditioned Forward-Backward (PFB) scheme with  $\phi(\mu) = \int V d\mu$
- NEM: MD with  $\phi(\mu) = \mathcal{H}(\mu)$  and restriction to Gaussian



# Preconditioned GD on Single-Cells

**Goal:**  $\min_{\mu} \mathcal{F}(\mu) = D(\mu, \nu)$  with  $\mu_0$  untreated cell and  $\nu$  perturbed cell

Use PGD with  $h^*(x) = (\|x\|_2^a + 1)^{1/a} - 1$  with  $a \in \{1.25, 1.5, 1.75\}$ , which is well suited to minimize functions growing in  $\|x - x^*\|^{a/(a-1)}$  near  $x^*$ .



- Rows: 2 profiling technologies
  - Columns/subcolumns: Different objectives  $\mathcal{F}$ /measure of convergence and number of iterations to converge
  - Points: For treatment  $i$ ,  $z_i = (x_i, y_i)$  with  $x_i$  value of  $\mathcal{F}(\hat{\mu}) = D(\hat{\mu}, \nu)$  (1st subcolumn) or number of iterations (2nd subcolumn) without preconditioning and  $y_i$  with preconditioning
  - Colors: treatments
- **Points below the diagonal: PGD provides a better minimum or converges faster**

# Conclusion

## Conclusion:

- Mirror Descent on  $\mathcal{P}_2(\mathbb{R}^d)$
- Preconditioned Gradient Descent on  $\mathcal{P}_2(\mathbb{R}^d)$
- Convergence analysis of the discrete schemes
- Also in the paper: analysis of the Bregman Forward-Backward scheme

## Perspectives:

- Better understand sufficient conditions of convergence for PGD
- Find more examples satisfying the conditions
- Analyze the Gaussian MD scheme



# Conclusion

## Conclusion:

- Mirror Descent on  $\mathcal{P}_2(\mathbb{R}^d)$
- Preconditioned Gradient Descent on  $\mathcal{P}_2(\mathbb{R}^d)$
- Convergence analysis of the discrete schemes
- Also in the paper: analysis of the Bregman Forward-Backward scheme

## Perspectives:

- Better understand sufficient conditions of convergence for PGD
- Find more examples satisfying the conditions
- Analyze the Gaussian MD scheme

Thank you for your attention!

# References I

- Kwangjun Ahn and Sinho Chewi. Efficient Constrained Sampling via the Mirror-Langevin Algorithm. *Advances in Neural Information Processing Systems*, 34:28405–28418, 2021.
- 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.
- Benoît Bonnet. A Pontryagin Maximum Principle in Wasserstein Spaces for Constrained Optimal Control Problems. *ESAIM: Control, Optimisation and Calculus of Variations*, 25:52, 2019.
- Lenaïc 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.

## References II

- Nabarun Deb, Young-Heon Kim, Soumik Pal, and Geoffrey Schiebinger. Wasserstein Mirror Gradient Flow as the Limit of the Sinkhorn Algorithm. *arXiv preprint arXiv:2307.16421*, 2023.
- 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.
- Alfredo Garbuno-Inigo, Franca Hoffmann, Wuchen Li, and Andrew M Stuart. Interacting Langevin Diffusions: Gradient Structure and Ensemble Kalman Sampler. *SIAM Journal on Applied Dynamical Systems*, 19(1):412–441, 2020.
- Guillaume Garrigos and Robert M Gower. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*, 2023.
- 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.

## References III

- Jaeyeon Kim, Chanwoo Park, Asuman Ozdaglar, Jelena Diakonikolas, and Ernest K Ryu. Mirror Duality in Convex Optimization. *arXiv preprint arXiv:2311.17296*, 2023.
- Nicolas Lanzetti, Saverio Bolognani, and Florian Dörfler. First-Order Conditions for Optimization in the Wasserstein Space. *arXiv preprint arXiv:2209.12197*, 2022.
- Qiang Liu and Dilin Wang. Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm. *Advances in neural information processing systems*, 29, 2016.
- 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

- Yifei Wang and Wuchen Li. Information Newton's Flow: Second-Order Optimization Method in Probability Space. *arXiv preprint arXiv:2001.04341*, 2020.
- 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.