

Mirror and Preconditioned Gradient Descent in Wasserstein Space

Clément Bonet¹

Joint work with Théo Uscidda¹, Adam David²,
Pierre-Cyril Aubin-Frankowski³, Anna Korba¹

¹ENSAE, CREST, Institut Polytechnique de Paris
²TU Berlin, ³CERMICS, ENPC



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](#))
- Modeling dynamic of populations of cells ([Schiebinger et al., 2019](#))

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 $\mu \ll \text{Leb}$
- $\mathcal{F}(\mu) = \text{KL}(\mu||\nu) = \int V d\mu + \mathcal{H}(\mu) (+C)$ for sampling from $\nu \propto e^{-V(x)}$
- $\mathcal{F}(\mu) = D(\mu, \nu)$ for sampling from ν

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

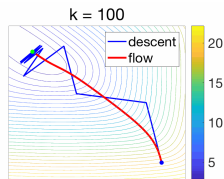
Goal: $\min_{x \in \mathbb{R}^d} f(x)$ via gradient flow

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

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

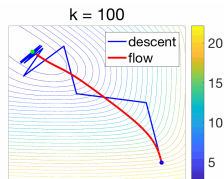
Goal: $\min_{x \in \mathbb{R}^d} f(x)$ via gradient flow

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

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

- 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 $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ 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 on \mathbb{R}^d (Beck and Teboulle, 2003)

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

Definition (Bregman Divergence)

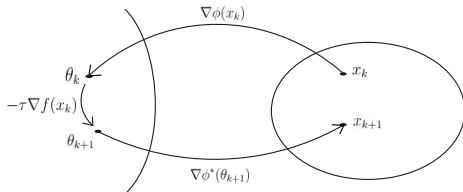
Let $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be strictly convex, then the Bregman divergence is defined as

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

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

Proof of convergence

Lemma (Three-Point Inequality)

Let $g : \mathbb{R}^d \rightarrow \mathbb{R}$ be a convex function, $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ strictly convex. Let $z_0 \in \mathbb{R}^d$ and $z^* = \operatorname{argmin}_z d_\phi(z, z_0) + g(z)$. Then, for all $z \in \mathbb{R}^d$,

$$g(z) + d_\phi(z, z_0) \geq g(z^*) + d_\phi(z^*, z_0) + d_\phi(z, z^*).$$

Proof of convergence

Lemma (Three-Point Inequality)

Let $g : \mathbb{R}^d \rightarrow \mathbb{R}$ be a convex function, $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ strictly convex. Let $z_0 \in \mathbb{R}^d$ and $z^* = \operatorname{argmin}_z d_\phi(z, z_0) + g(z)$. Then, for all $z \in \mathbb{R}^d$,

$$g(z) + d_\phi(z, z_0) \geq g(z^*) + d_\phi(z^*, z_0) + d_\phi(z, z^*).$$

Using smoothness, Three-point inequality on $g(x) = \frac{1}{\beta} \langle \nabla f(x_k), x - x_k \rangle$ and strong convexity we get for all $x \in \mathbb{R}^d$,

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \beta d_\phi(x_{k+1}, x_k) \\ &\leq f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \beta d_\phi(x, x_k) - \beta d_\phi(x, x_{k+1}) \\ &\leq f(x) + (\beta - \alpha) d_\phi(x, x_k) - \beta d_\phi(x, x_{k+1}). \end{aligned}$$

By induction and using the monotonicity of f , and taking $x = x^*$, we get the desired rates: $f(x_k) - f(x) \leq \frac{\alpha d_\phi(x, x_0)}{(1 + \frac{\alpha}{\beta - \alpha})^k - 1} \leq \frac{\beta - \alpha}{k} d_\phi(x, x_0)$.

Proof of convergence

Lemma (Three-Point Inequality)

Let $g : \mathbb{R}^d \rightarrow \mathbb{R}$ be a convex function, $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ strictly convex. Let $z_0 \in \mathbb{R}^d$ and $z^* = \operatorname{argmin}_z d_\phi(z, z_0) + g(z)$. Then, for all $z \in \mathbb{R}^d$,

$$g(z) + d_\phi(z, z_0) \geq g(z^*) + d_\phi(z^*, z_0) + d_\phi(z, z^*).$$

Using smoothness, Three-point inequality on $g(x) = \frac{1}{\beta} \langle \nabla f(x_k), x - x_k \rangle$ and strong convexity we get for all $x \in \mathbb{R}^d$,

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \beta d_\phi(x_{k+1}, x_k) \\ &\leq f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \beta d_\phi(x, x_k) - \beta d_\phi(x, x_{k+1}) \\ &\leq f(x) + (\beta - \alpha) d_\phi(x, x_k) - \beta d_\phi(x, x_{k+1}). \end{aligned}$$

By induction and using the monotonicity of f , and taking $x = x^*$, we get the desired rates: $f(x_k) - f(x) \leq \frac{\alpha d_\phi(x, x_0)}{(1 + \frac{\alpha}{\beta - \alpha})^k - 1} \leq \frac{\beta - \alpha}{k} d_\phi(x, x_0)$.

Remarks: (1) the proof works replacing β with $1/\tau$ if $\tau \leq 1/\beta$, (2) we need smoothness and convexity in specific directions.

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

Relation between MD and Preconditioned GD



Dual Space Preconditioning for Gradient Descent



Chris J. Maddison^{1,4,*}, Daniel Paulin^{2,*}, Yee Whye Teh³, and Arnaud Doucet³



Algorithm Mirror descent

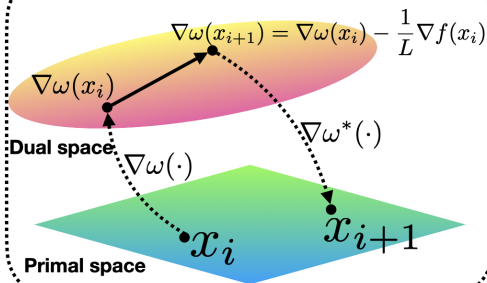
$$\nabla\omega(x_{i+1}) = \nabla\omega(x_i) - \frac{1}{L}\nabla f(x_i)$$

Algorithm 1.1 Dual preconditioned gradient descent

$$x_{i+1} = x_i - \frac{1}{L^*}\nabla k(\nabla f(x_i))$$



Mirror Descent:



Dual Preconditioning:

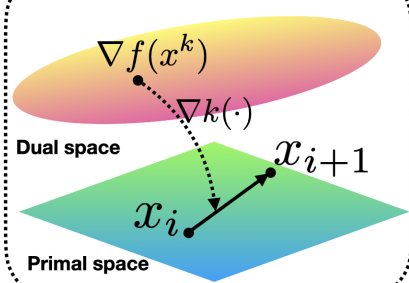


Figure: Taken from a tweet of Konstantin Mishchenko ¹

¹<https://mobile.x.com/konstmish/status/1431983100561592323/photo/1>

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 μ, ν . Then, the Wasserstein distance is

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

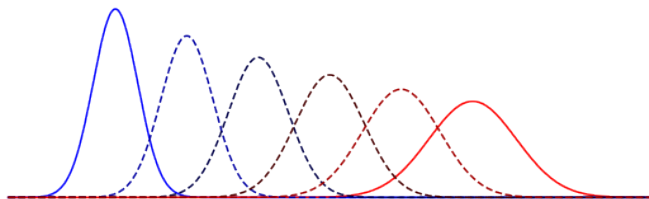
Properties:

- 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_μ^ν 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**

Riemannian Structure of the Wasserstein Space

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

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



Riemannian Structure of the Wasserstein Space

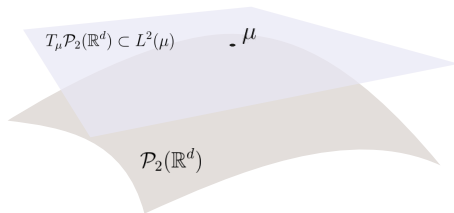
- Geodesics between $\mu \ll \text{Leb}$ and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$:

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

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



Riemannian Structure of the Wasserstein Space

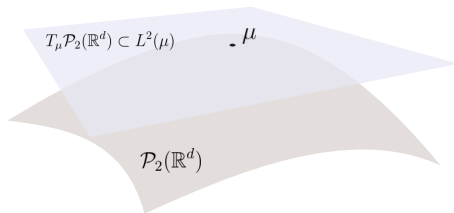
- Geodesics between $\mu \ll \text{Leb}$ and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$:

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

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



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

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

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 μ if for any $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ and any optimal coupling $\gamma \in \Pi_o(\mu, \nu)$,

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

If such a gradient exists, then we say that \mathcal{F} is W_2 -differentiable at 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 $\gamma \in \Pi(\mu, \nu)$,

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

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 smooth,

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

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

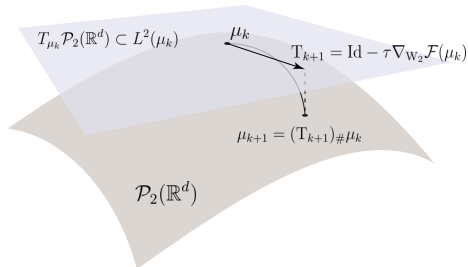
$$\mathbb{T} = \operatorname{argmin}_{\mathbb{T} \in L^2(\mu_k)} \frac{1}{2} \|\mathbb{T} - \operatorname{Id}\|_{L^2(\mu_k)}^2 + \tau \mathcal{F}(\mathbb{T}_{\#} \mu_k)$$

- 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 Descent



Wasserstein Gradient Descent:

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

Taking the FOC: $T_{k+1} = \text{Id} - \tau \nabla_{W_2} \mathcal{F}(\mu_k)$

Particle approximation: $\hat{\mu}_k^n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i^k}$, $x_i^{k+1} = T_{k+1}(x_i^k)$ for all $i \in \{1, \dots, n\}$.

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:

- For $d(\mathbb{T}, \operatorname{Id}) = \frac{1}{2} \|\mathbb{T} - \operatorname{Id}\|_{L^2(\mu)}^2$: **Wasserstein gradient descent**
- For $d_{\phi_{\mu}}(\mathbb{T}, \operatorname{Id}) = \phi_{\mu}(\mathbb{T}) - \phi_{\mu}(\operatorname{Id}) - \langle \nabla \phi_{\mu}(\operatorname{Id}), \mathbb{T} - \operatorname{Id} \rangle_{L^2(\mu)}$ (**Bregman divergence** on $L^2(\mu)$): extends **Mirror Descent** ([Beck and Teboulle, 2003](#)) to $\mathcal{P}_2(\mathbb{R}^d)$.
- For $d(\mathbb{T}, \operatorname{Id}) = \int h(\mathbb{T}(x) - x) d\mu(x)$: extends **Preconditioned Gradient Descent** ([Maddison et al., 2021](#)) 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 ϕ_μ 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 ϕ is W_2 -differentiable, then ϕ_μ 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.
 - ϕ_μ is β -smooth relative to ψ_μ if for all $T, S \in L^2(\mu)$, $d_{\phi_\mu}(T, S) \leq \beta d_{\psi_\mu}(T, S)$.
 - ϕ_μ is α -convex relative to ψ_μ 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 $\mu_t = ((1-t)\text{Id} + t\text{T}_{\mu_0}^{\mu_1})\# \mu_0$.

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

$$\mathcal{F}(\mu_t) \leq (1-t)\mathcal{F}(\mu_0) + t\mathcal{F}(\mu_1) - \frac{\alpha t(1-t)}{2} \text{W}_2^2(\mu_0, \mu_1),$$

or equivalently

$$\begin{aligned} \frac{\alpha}{2} \text{W}_2^2(\mu_0, \mu_1) &= \frac{\alpha}{2} \|\text{T}_{\mu_0}^{\mu_1} - \text{Id}\|_{L^2(\mu_0)}^2 \leq \mathcal{F}(\mu_1) - \mathcal{F}(\mu_0) - \langle \nabla_{\text{W}_2} \mathcal{F}(\mu_0), \text{T}_{\mu_0}^{\mu_1} - \text{Id} \rangle_{L^2(\mu_0)} \\ &= d_{\tilde{\mathcal{F}}_{\mu_0}}(\text{T}_{\mu_0}^{\mu_1}, \text{Id}) \end{aligned}$$

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

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 α -geodesically convex if $t \mapsto \mathcal{F}(\mu_t)$ is α -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} β -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} α -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 ϕ_μ 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).

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 β -smooth relative to ϕ 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: β -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 β -smoothness
2. Apply 3-point inequality

Convergence

Proposition

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

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

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

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

Let ϕ_μ be pushforward compatible. Define the OT problem:

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

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

Continuous Formulation

Informally, for ϕ_μ 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 ϕ_μ 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

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

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](#))

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 $t \mapsto ((1-t)\mathbb{T}_{k+1} + t\operatorname{Id})_{\#} \mu_k$
- ϕ^{h^*} β -smooth relative to \mathcal{F}^* along a suitable curve

Then, for all $k \geq 0$,

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

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 $t \mapsto ((1-t)\mathbb{T}_{k+1} + t\operatorname{Id})_{\#} \mu_k$
- ϕ^{h^*} β -smooth relative to \mathcal{F}^* along a suitable curve
- ϕ^{h^*} α -convex relative to \mathcal{F}^* along a suitable curve

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 β -smooth relative to $U \implies \mathcal{F}$ β -smooth relative to ϕ

V α -convex relative to $U \implies \mathcal{F}$ α -convex relative to ϕ

- 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 β -smooth relative to $K \implies \mathcal{F}$ β -smooth relative to ϕ

W α -convex relative to $K \implies \mathcal{F}$ α -convex relative to ϕ

- 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

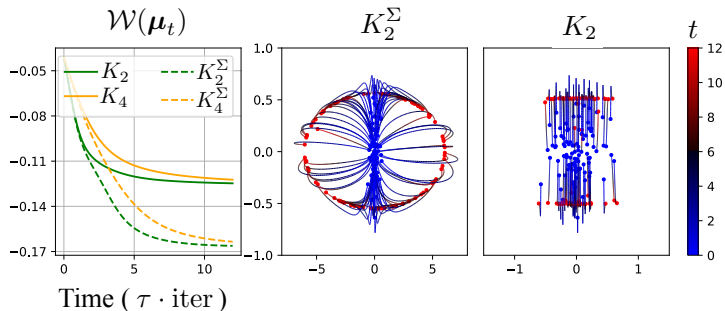
Goal: Let $\Sigma \in S_d^{++}(\mathbb{R})$ possibly ill-conditioned,

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

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:

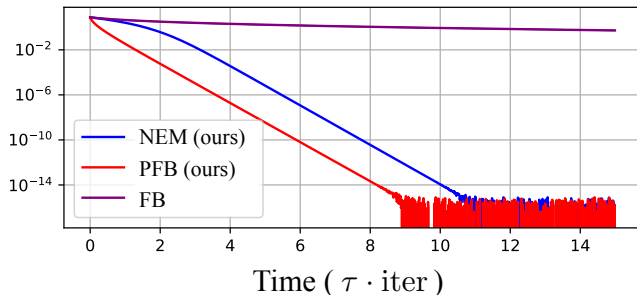
$$\min_{\mu} \mathcal{F}(\mu) = \int V d\mu + \mathcal{H}(\mu) \quad \text{with} \quad V(x) = \frac{1}{2} x^T \Sigma^{-1} x$$

→ 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

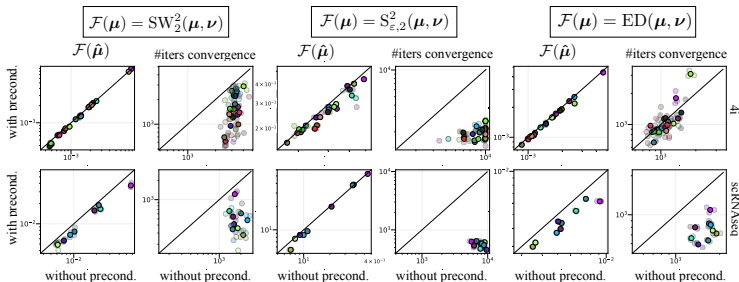
$$\text{KL}(\mu_t || \mu^*)$$



Preconditioned GD on Single-Cells

Goal: $\min_{\mu} \mathcal{F}(\mu) = D(\mu, \nu)$ with μ_0 untreated cell and ν 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.
- 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

- 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.
- 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.