Mirror and Preconditioned Gradient Descent in Wasserstein Space

Clément Bonet[1], Théo Uscidda[2], Adam David[3],
Pierre-Cyril Aubin-Frankowski[4], Anna Korba[2]

[1]Ecole Polytechnique, CMAP, Institut Polytechnique de Paris
[2]ENSAE, CREST, Institut Polytechnique de Paris
[3]TU Berlin
[4]CERMICS, ENPC

PGMO Days
19/11/2025

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:

- 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: $\mathcal{F}(\mu) = \int V \mathrm{d}\mu + \iint W(x,y)\ \mathrm{d}\mu(x)\mathrm{d}\mu(y) + \mathcal{H}(\mu)$ where $\mathcal{H}(\mu) = \int \log\big(\mu(x)\big)\ \mathrm{d}\mu(x)$ for $\mu \ll \mathrm{Leb}$
- $\mathcal{F}(\mu) = \mathrm{KL}(\mu||\nu) = \int V \mathrm{d}\mu + \mathcal{H}(\mu) + \mathrm{cst}$ 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 and Preconditioned Gradient Descent on $\mathcal{P}_2(\mathbb{R}^d)$

Applications

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

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

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

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

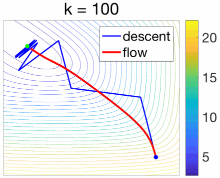
Gradient Descent on $\mathbb{R}^d$

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

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

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

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



From (Bach, 2020)

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

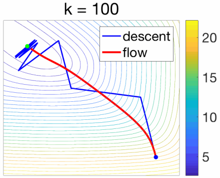
$$= \operatorname*{argmin}_{x \in \mathbb{R}^d} \ \frac{1}{2}\|x - x_k\|_2^2 + \tau \langle \nabla f(x_k), x - x_k \rangle$$

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

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

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

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



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

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

From (Bach, 2020)

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

Convergence Analysis

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

Reminder:

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

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

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

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

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

Definition (Bregman Divergence)

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

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

Properties:

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

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

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

Definition (Bregman Divergence)

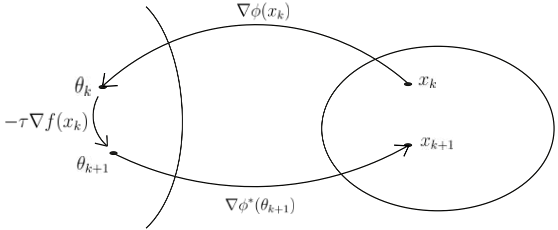
Let $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be convex, then the Bregman divergence is defined as

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

Mirror Descent (MD) algorithm:

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

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



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

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

Definition (Bregman Divergence)

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

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

Mirror Descent (MD) algorithm:

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

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

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

Preconditioned Gradient Descent (Maddison et al., 2021)

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

Preconditioned Gradient Descent scheme for $\min_y \; g(y)$:

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

Preconditioned Gradient Descent (Maddison et al., 2021)

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

Preconditioned Gradient Descent scheme for $\min_y \ g(y)$:

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

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

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

Preconditioned Gradient Descent (Maddison et al., 2021)

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

Preconditioned Gradient Descent scheme for $\min_y \ g(y)$:

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

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

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

Convergence analysis (Maddison et al., 2021)

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

Relation between MD and Preconditoned GD

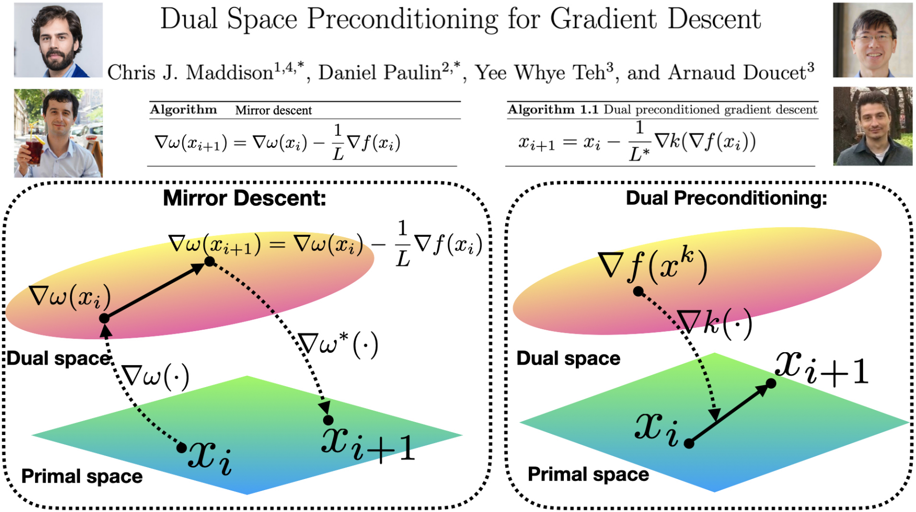


Figure: Taken from a tweet of Konstantin Mishchenko [1]

[1] https://mobile.x.com/konstmish/status/1431983100561592323/photo/1

Table of Contents

Detour by $\mathbb{R}^d$

Wasserstein Gradient Flows

Mirror Descent and Preconditioned Gradient Descent on $\mathcal{P}_2(\mathbb{R}^d)$

Applications

Wasserstein Geometry (Ambrosio et al., 2005)

Endow $\mathcal{P}_2(\mathbb{R}^d)$ with the Wasserstein distance,

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

Wasserstein Geometry (Ambrosio et al., 2005)

Endow $\mathcal{P}_2(\mathbb{R}^d)$ with the Wasserstein distance,

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

Properties:

- W_2 distance, $(\mathcal{P}_2(\mathbb{R}^d), \mathrm{W}_2)$: Wasserstein space
- Riemannian structure (with geodesics and tangent space $\mathcal{T}_\mu\mathcal{P}_2(\mathbb{R}^d) \subset L^2(\mu)$)

Wasserstein gradient $\nabla_{\mathrm{W}_2}\mathcal{F}(\mu) : \mathbb{R}^d \to \mathbb{R}^d \in \mathcal{T}_\mu\mathcal{P}_2(\mathbb{R}^d)$ of $\mathcal{F}$ at μ satisfies

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

with $\mathrm{T}_\#\mu(A) = \mu\big(\mathrm{T}^{-1}(A)\big)$ for all $A \in \mathcal{B}(\mathbb{R}^d)$.

Wasserstein Geometry (Ambrosio et al., 2005)

Endow $\mathcal{P}_2(\mathbb{R}^d)$ with the Wasserstein distance,

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

Properties:

- W_2 distance, $(\mathcal{P}_2(\mathbb{R}^d), \mathrm{W}_2)$: Wasserstein space
- Riemannian structure (with geodesics and tangent space $\mathcal{T}_\mu \mathcal{P}_2(\mathbb{R}^d) \subset L^2(\mu)$)

Wasserstein gradient $\nabla_{\mathrm{W}_2}\mathcal{F}(\mu) : \mathbb{R}^d \to \mathbb{R}^d \in \mathcal{T}_\mu \mathcal{P}_2(\mathbb{R}^d)$ of $\mathcal{F}$ at μ satisfies

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

with $\mathrm{T}_\#\mu(A) = \mu\big(\mathrm{T}^{-1}(A)\big)$ for all $A \in \mathcal{B}(\mathbb{R}^d)$.

Example

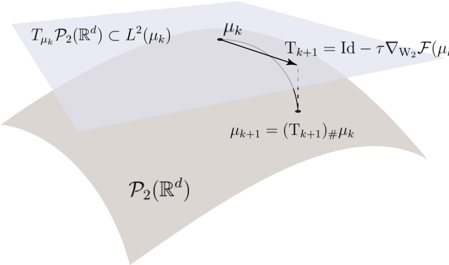
- $\mathcal{V}(\mu) = \int V \mathrm{d}\mu$, $\nabla_{\mathrm{W}_2}\mathcal{V}(\mu) = \nabla V$
- $\mathcal{W}(\mu) = \frac{1}{2}\iint W(x-y)\ \mathrm{d}\mu(x)\mathrm{d}\mu(y)$, $\nabla_{\mathrm{W}_2}\mathcal{W}(\mu) = \nabla W \star \mu$

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

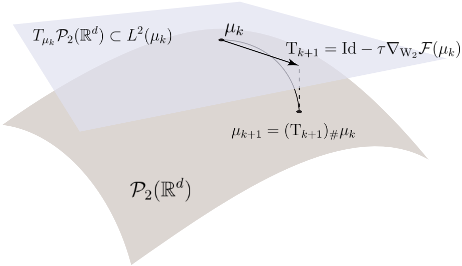


Wasserstein Gradient Descent

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



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

Contributions

Study schemes of the form

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

and provide **convergence conditions**.

Considered divergences:

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

Table of Contents

Detour by $\mathbb{R}^d$

Wasserstein Gradient Flows

Mirror Descent and Preconditioned Gradient Descent on $\mathcal{P}_2(\mathbb{R}^d)$

Applications

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,

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

Relative Convexity and Smoothness

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

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

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

Relative Convexity and Smoothness

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

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

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

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

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

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

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

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

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

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

Mirror Descent on the Wasserstein Space

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

Mirror Descent scheme:

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

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

Mirror Descent on the Wasserstein Space

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

Mirror Descent scheme:

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

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

Computing the scheme:

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

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

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

Descent Lemma

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

Mirror Descent scheme:

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

Proposition (Descent Lemma)

Assumptions:

- *For all $k \geq 0$, $\mathcal{F}$ is β-smooth relative to ϕ along $t \mapsto \big((1-t)\mathrm{Id} + t\mathrm{T}_{k+1}\big)_\# \mu_k$*

Then, for all $k \geq 0$,

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

Convergence

Proposition

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

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

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

Convergence

Proposition

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

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

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

Let ϕ_μ be pushforward compatible. Define the OT problem:

$$\begin{aligned}\mathrm{W}_\phi(\nu,\mu) &= \inf_{\gamma\in\Pi(\nu,\mu)}\ \phi(\nu) - \phi(\mu) - \int \langle \nabla_{\mathrm{W}_2}\phi(\mu)(y), x-y\rangle\ \mathrm{d}\gamma(x,y) \\ &\leq \mathrm{d}_{\phi_\eta}(\mathrm{T},\mathrm{S}) \quad \text{for} \quad (\mathrm{T},\mathrm{S})_\#\eta \in \Pi(\nu,\mu)\end{aligned}$$

Property: If $\mu_{k+1} \ll \mathrm{Leb}$ and $\nabla_{\mathrm{W}_2}\phi(\mu_{k+1})$ is invertible, then

$$\mathrm{d}_{\phi_{\mu_k}}(\mathrm{T}_{\phi_{\mu_k}}^{\mu_k,\mu^*}, \mathrm{T}_{k+1}) \geq \mathrm{d}_{\phi_{\mu_{k+1}}}(\mathrm{T}_{\phi_{\mu_{k+1}}}^{\mu_{k+1},\mu^*}, \mathrm{Id}) = \mathrm{W}_\phi(\mu^*, \mu_{k+1})$$

Preconditioned GD

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

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

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

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

Under relative smoothness and convexity of $\phi_\mu^{h^*}$ relative to $\mathcal{F}^*$:

$$\forall k \geq 0,\ \phi_{\mu_{k+1}}^{h^*}\big(\nabla_{\mathrm{W}_2}\mathcal{F}(\mu_{k+1})\big) \leq \phi_{\mu_k}^{h^*}\big(\nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k)\big) - \beta \mathrm{d}_{\tilde{\mathcal{F}}_{\mu_k}}(\mathrm{T}_{k+1}, \mathrm{Id}),$$

$$\forall k \geq 1,\ \phi_{\mu_k}^{h^*}\big(\nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k)\big) - h^*(0) \leq \frac{\beta - \alpha}{k}\big(\mathcal{F}(\mu_0) - \mathcal{F}(\mu^*)\big).$$

Table of Contents

Detour by $\mathbb{R}^d$

Wasserstein Gradient Flows

Mirror Descent and Preconditioned Gradient Descent on $\mathcal{P}_2(\mathbb{R}^d)$

Applications

Showing Relative Smoothness and Convexity

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

Showing Relative Smoothness and Convexity

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

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

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

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

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

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

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

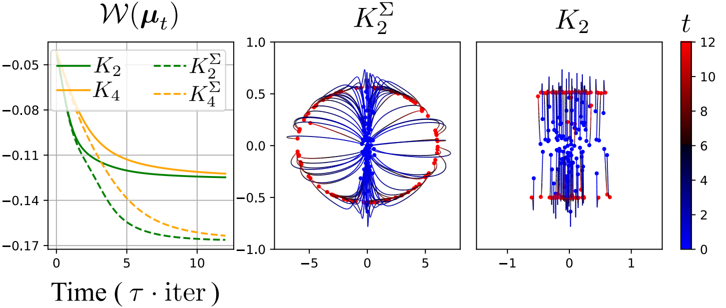
Mirror Descent on Interaction Energy

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

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

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

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



Mirror Descent on Gaussian

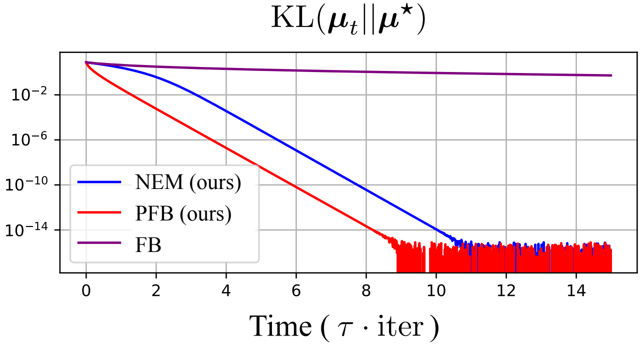
Goal:

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

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

Comparison between:

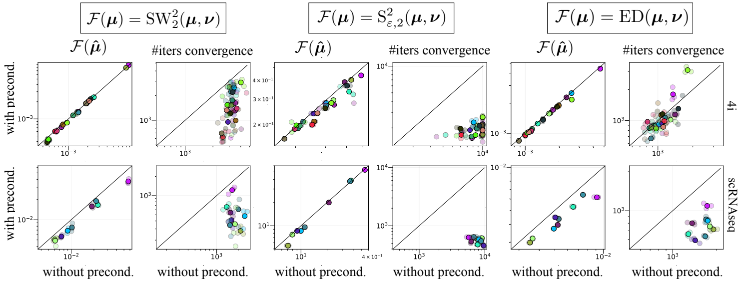
- Forward-Backward (FB) on the Bures-Wasserstein space (Diao et al., 2023)
- Preconditioned Forward-Backward (PFB) scheme with $\phi(\mu) = \int V \mathrm{d}\mu$
- NEM: MD with $\phi(\mu) = \mathcal{H}(\mu)$ and restriction to Gaussian



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 $a \in \{1.25, 1.5, 1.75\}$



- Rows: 2 profiling technologies
- Columns: Different objectives $\mathcal{F}$
- Subcolumns: Measure of convergence and number of iterations to converge

$\rightarrow$ **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:

- Find more examples satisfying the conditions
- Analyze the Gaussian MD scheme

Thank you!

Paper: https://arxiv.org/abs/2406.08938

References I

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

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

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

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

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

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

References II

Jaeyeon Kim, Chanwoo Park, Asuman Ozdaglar, Jelena Diakonikolas, and Ernest K Ryu. Mirror Duality in Convex Optimization. *arXiv preprint arXiv:2311.17296*, 2023.

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.

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.