Mirror (and Preconditioned Gradient) Descent on the Wasserstein space

Anna Korba

ENSAE, CREST, Institut Polytechnique de Paris

PDE Methods in Machine Learning: from Continuum Dynamics to Algorithms - BIRS-IMAG workshop, Granada

Joint work with Clément Bonet, Théo Uscidda (CREST), Adam David (TU Berlin), Pierre-Cyril Aubin-Frankowski (TU Wien)

Problem - optimization over $\mathcal{P}_2(\mathbb{R}^d)$

Consider the following optimization problem:

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

where $\mathcal{P}_2(\mathbb{R}^d) = \{\mu \in \mathcal{P}(\mathbb{R}^d), \int \|x\|^2 d\mu(x) < \infty\}$, equipped with the W_2 distance*.

Applications include:

- Sampling (from a target probability distribution whose density is known up to a normalization constant)
- Generative Modeling
- Learning neural networks

Examples of functionals:

- Free energies: potential energy $\int V(x) d\mu(x)$, interaction energy $\iint W(x-y) d\mu(x) d\mu(y)$, negative entropy $\int \log(\mu(x)) d\mu(x)$
- Distance or divergence to a target probability distribution μ^* (e.g. $\mathrm{W}_2(\mu, \mu^*)$...)

*$\mathrm{W}_2^2(\nu, \mu) = \inf_{s \in \Pi(\nu,\mu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, ds(x, y)$, where $\Pi(\nu, \mu)$= couplings between ν, μ.

Outline

- **Brenier's theorem.** Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ s.t. $\mu \ll$ Leb. Then, there exists a unique $\mathrm{T}^{\mu,\nu} : \mathbb{R}^d \to \mathbb{R}^d$ such that

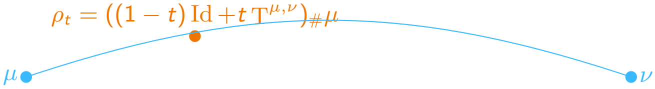
  1. $\mathrm{T}^{\mu,\nu}_{\#}\mu = \nu$
  2. $\mathrm{W}_2^2(\mu,\nu) = \|\,\mathrm{Id} - \mathrm{T}^{\mu,\nu}\,\|^2_{L^2(\mu)} \overset{\textbf{def.}}{=} \int \|x - \mathrm{T}^{\mu,\nu}(x)\|^2 d\mu(x)$.

  and $\mathrm{T}^{\mu,\nu}$ is called the Optimal Transport map between μ and ν.

  The path

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

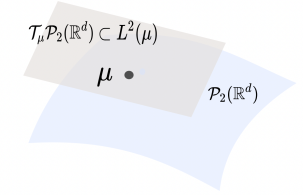
  is the Wasserstein geodesic between $\rho_0 = \mu$ and $\rho_1 = \nu$.



- $\mathcal{F}$ is said to be α-geodesically (or displacement) convex if it is convex along the curves ρ_t defined as above:

$$\mathcal{F}(\rho_t) \le (1-t)\mathcal{F}(\mu) + t\mathcal{F}(\nu) - \frac{\alpha t(1-t)}{2}\mathrm{W}_2^2(\mu,\nu),$$

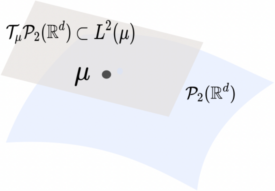
- Equipped with the Wasserstein-2 (W_2) distance, the metric space $(\mathcal{P}_2(\mathbb{R}^d), \mathrm{W}_2)$ has a convenient **Riemannian structure** [Otto and Villani, 2000].



where $L^2(\mu) = \{f : \mathbb{R}^d \to \mathbb{R}^d, \int_{\mathbb{R}^d} \|f(x)\|^2 d\mu(x) < \infty\}$.

- Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $\mathrm{T} : \mathbb{R}^d \to \mathbb{R}^d$ a measurable map. The pushforward measure $T_\# \mu$ is characterized by: $X \sim \mu \implies \mathrm{T}(X) \sim \mathrm{T}_\# \mu$. If $\mathrm{T} \in L^2(\mu)$, then $\mathrm{T}_\# \mu \in \mathcal{P}_2(\mathbb{R}^d)$.

- Equipped with the Wasserstein-2 (W_2) distance, the metric space $(\mathcal{P}_2(\mathbb{R}^d), \mathrm{W}_2)$ has a convenient **Riemannian structure** [Otto and Villani, 2000].



  where $L^2(\mu) = \{f : \mathbb{R}^d \to \mathbb{R}^d, \int_{\mathbb{R}^d} \|f(x)\|^2 d\mu(x) < \infty\}$.

- Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $\mathrm{T} : \mathbb{R}^d \to \mathbb{R}^d$ a measurable map. The pushforward measure $T_\#\mu$ is characterized by: $X \sim \mu \Longrightarrow \mathrm{T}(X) \sim \mathrm{T}_\# \mu$. If $\mathrm{T} \in L^2(\mu)$, then $\mathrm{T}_\# \mu \in \mathcal{P}_2(\mathbb{R}^d)$.

Wasserstein gradient

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

Definition: (First variation) Consider a linear perturbation $\mu + \varepsilon\xi \in \mathcal{P}_2(\mathbb{R}^d)$ for a perturbation $\xi = \nu - \mu$, $\nu \in \mathcal{P}_2(\mathbb{R}^d)$.

If a Taylor expansion of $\mathcal{F}$ yields:

$$\mathcal{F}(\mu + \varepsilon\xi) = \mathcal{F}(\mu) + \varepsilon \int \mathcal{F}'(\mu)(x) d\xi(x) + o(\varepsilon),$$

then $\mathcal{F}'(\mu) : \mathbb{R}^d \to \mathbb{R}$ is the First Variation of $\mathcal{F}$ at μ.

Definition: (informal) Consider a perturbation on the Wasserstein space $(\mathrm{Id} + \varepsilon h)_{\#}\mu$ for $h \in L^2(\mu)$.

If a Taylor expansion of $\mathcal{F}$ yields:

$$\boxed{\mathcal{F}((\mathrm{Id} + \varepsilon h)_{\#}\mu) = \mathcal{F}(\mu) + \varepsilon \langle \nabla_{\mathrm{W}_2}\mathcal{F}(\mu), h \rangle_{L^2(\mu)} + o(\varepsilon),}$$

then $\nabla_{\mathrm{W}_2}\mathcal{F}(\mu) \in L^2(\mu)$ is a Wasserstein gradient of $\mathcal{F}$ at μ. Typically, $\nabla_{\mathrm{W}_2}\mathcal{F}(\mu) = \nabla \mathcal{F}'(\mu)$.

More formally. Notice that $(\mathrm{Id} + \varepsilon h)$ generate optimal transport maps for ϵ small. In the following, we use the differential structure of $(\mathcal{P}_2(\mathbb{R}^d), \mathrm{W}_2)$ introduced in [Bonnet, 2019, Lanzetti et al., 2022].

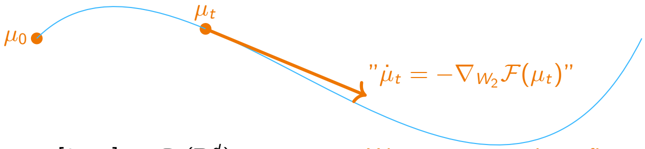
We say that $\nabla_{\mathrm{W}_2}\mathcal{F}(\mu)$ is a Wasserstein gradient of $\mathcal{F}$ at $\mu \in Dom(\mathcal{F})$ if for any $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ and any optimal coupling $\gamma \in \Pi_o(\mu, \nu)$,

$$\mathcal{F}(\nu) = \mathcal{F}(\mu) + \int \langle \nabla_{\mathrm{W}_2}\mathcal{F}(\mu)(x), y - x\rangle \, \mathrm{d}\gamma(x, y) + o\big(\mathrm{W}_2(\mu, \nu)\big). \tag{1}$$

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

- There is a unique gradient belonging to the tangent space of $\mathcal{P}_2(\mathbb{R}^d)$ verifying (1) .
- W_2-differentiable functionals include c-Wasserstein costs, potential energies $\mathcal{V}(\mu) = \int V \mathrm{d}\mu$ or interaction energies $\mathcal{W}(\mu) = \iint W(x - y) \, \mathrm{d}\mu(x)\mathrm{d}\mu(y)$ for V and W differentiable and L-smooth.
- the negative entropy defined as $\mathcal{H}(\mu) = \int \log\big(\mu(x)\big)\mathrm{d}\mu(x)$ is not W_2-differentiable. In this case, we can consider subgradients $\nabla_{\mathrm{W}_2}\mathcal{F}(\mu)$ at μ for which (1) becomes an inequality.

Wasserstein gradient flows (WGF) [Ambrosio et al., 2008]



The curve $\mu : [0, \infty] \to \mathcal{P}_2(\mathbb{R}^d)$, $t \mapsto \mu_t$ is a Wasserstein gradient flow of $\mathcal{F}$ if:

$$\boxed{\frac{\partial \mu_t}{\partial t} = \boldsymbol{\nabla} \cdot \left(\mu_t \nabla_{W_2}\mathcal{F}(\mu_t)\right),}$$

where $\nabla_{W_2}\mathcal{F}(\mu) \in L^2(\mu)$ denotes a Wasserstein (sub)gradient of $\mathcal{F}$.

Wasserstein Gradient Descent (WGD)

Let $\tau > 0$ a step-size. 2 possibles time-discretizations:

- Implicit (JKO [Jordan et al., 1998])

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

- Explicit (WGD)

$$\mathrm{T}_{k+1} = \operatorname*{arg\,min}_{\mathrm{T} \in L^2(\mu_k)} \langle \nabla_{W_2} \mathcal{F}(\mu_k), \mathrm{T} - \mathrm{Id} \rangle_{L^2(\mu_k)} + \frac{1}{2\tau} \| \mathrm{T} - \mathrm{Id} \|^2_{L^2(\mu_k)}$$

and $\mu_{k+1} = \mathrm{T}_{k+1\#}\, \mu_k = (\mathrm{Id} - \tau \nabla_{W_2} \mathcal{F}(\mu_k))_\# \mu_k$.

Space discretization: Let $x_0^1, \dots, x_0^n \sim \mu_0$, at each time $k \geq 0$ we have:

$$x_{k+1}^i = x_k^i - \tau \nabla_{W_2} \mathcal{F}(\hat{\mu}_k)(x_k^i) \quad \text{for } i = 1, \dots, n, \text{ where } \hat{\mu}_k = \frac{1}{n} \sum_{i=1}^n \delta_{x_k^i}. \tag{2}$$

In particular, if $\mathcal{F}(\mu)$ is well-defined for discrete measures μ, Algorithm (2) simply corresponds to gradient descent of $F : \mathbb{R}^{n \times d} \to \mathbb{R}$, $F(x^1, \dots, x^n) := \mathcal{F}(\mu^n)$ where $\mu^n = \frac{1}{n} \sum_{i=1}^n \delta_{x^i}$.

Outline

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

Let $f : \mathbb{R}^d \to \mathbb{R}$. Mirror descent [Beck and Teboulle, 2003] writes for each $k \geq 0$:

$$x_{k+1} = \underset{x \in \mathbb{R}^d}{\arg\min} \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{\tau} \mathrm{D}_\phi(x, x_k) \tag{3}$$

where D_ϕ is a Bregman divergence, i.e.

$$\mathrm{D}_\phi(x, x_k) = \phi(x) - \phi(x_k) - \langle \nabla \phi(x_k), x - x_k \rangle$$

for ϕ a strictly convex function (taking $\phi(x) = \frac{1}{2}\|x\|^2$ recovers gradient descent).

Implementation. FOC of (3):

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

where ϕ^* is the Legendre transform of ϕ.

Guarantees. [Lu et al., 2018] obtained rates for relatively smooth and convex functions, i.e. $\alpha \, \mathrm{D}_\phi(x, y) \leq \mathrm{D}_f(x, y) \leq \beta \, \mathrm{D}_\phi(x, y)$ (equivalently, $f - \alpha\phi$ and $\beta\phi - f$ are convex).

This work - MD and PGD on $\mathcal{P}_2(\mathbb{R}^d)$

We are interested in minimizing a functional $\mathcal{F}:\mathcal{P}_2(\mathbb{R}^d)\to\mathbb{R}\cup\{+\infty\}$ over probability distributions, through schemes of the form, for $k\geq 0$,

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

with different
costs $\mathrm{D}: L^2(\mu_k)\times L^2(\mu_k)\to\mathbb{R}_+$, and in providing convergence conditions.

For D, we consider:

- Bregman divergences on $L^2(\mu)$ (extending MD to $\mathcal{P}_2(\mathbb{R}^d)$)
- c-Wasserstein costs with c translation-invariant (extending PGD to $\mathcal{P}_2(\mathbb{R}^d)$)

PGD = Preconditioned Gradient Descent [Maddison et al., 2021]

$$y_{k+1}-y_k = -\tau\nabla h^*\big(\nabla g(y_k)\big)$$

for some objective g and (strictly convex) regularizer h. Setting $g=\phi^*$ and $h^*=f$, we see that, for $y=\nabla\phi(x)$, the two schemes are equivalent when permuting the roles of the objective and of the regularizer.

Bregman on L^2, Rel. smoothness and convexity on $\mathcal{P}_2(\mathbb{R}^d)$

Definition (Bregman potential and divergence)

Let $\phi_\mu : L^2(\mu) \to \mathbb{R}$ be strictly convex and continuously Gâteaux differentiable. The Bregman divergence is defined for all $\mathrm{T}, \mathrm{S} \in L^2(\mu)$ as $\mathrm{D}_{\phi_\mu}(\mathrm{T}, \mathrm{S}) = \phi_\mu(\mathrm{T}) - \phi_\mu(\mathrm{S}) - \langle \nabla \phi_\mu(\mathrm{S}), \mathrm{T} - \mathrm{S} \rangle_{L^2(\mu)}$.

In particular, for $\phi_\mu(\mathrm{T}) = \frac{1}{2} \| \mathrm{T} \|^2_{L^2(\mu)}$, we recover the L^2 norm as a divergence $\mathrm{D}_{\phi_\mu}(\mathrm{T}, \mathrm{S}) = \frac{1}{2} \| \mathrm{T} - \mathrm{S} \|^2_{L^2(\mu)}$.

Definition (Relative smoothness and convexity)

Let $\psi_\mu, \phi_\mu : L^2(\mu) \to \mathbb{R}$ strictly convex and continuously Gâteaux differentiable. We say that ψ is β-smooth (respectively α-convex) relative to ϕ if and only if for all $\mathrm{T}, \mathrm{S} \in L^2(\mu)$, $\mathrm{D}_{\psi_\mu}(\mathrm{T}, \mathrm{S}) \leq \beta \, \mathrm{D}_{\phi_\mu}(\mathrm{T}, \mathrm{S})$ (respectively $\mathrm{D}_{\psi_\mu}(\mathrm{T}, \mathrm{S}) \geq \alpha \, \mathrm{D}_{\phi_\mu}(\mathrm{T}, \mathrm{S})$).

- if ψ_μ, ϕ_μ are potential energies, relative notions on $\mathbb{R}^d$ translate directly.
- geodesic convexity corresponds to choosing ϕ_μ the L^2 norm, ψ_μ the objective functional and considering OT maps and identity.

Mirror descent on $\mathcal{P}_2(\mathbb{R}^d)$

$$\mathrm{T}_{k+1} = \operatorname*{arg\,min}_{\mathrm{T}\in L^2(\mu_k)} \mathrm{D}_{\phi_{\mu_k}}(\mathrm{T}, \mathrm{Id}) + \tau\langle \nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k), \mathrm{T}-\mathrm{Id}\rangle_{L^2(\mu_k)}, \quad \mu_{k+1} = (\mathrm{T}_{k+1})_{\#}\mu_k.$$

FOC lead to

$$\nabla\phi_{\mu_k}(\mathrm{T}_{k+1}) = \nabla\phi_{\mu_k}(\mathrm{Id}) - \tau\nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k) \iff \mathrm{T}_{k+1} = \nabla\phi^*_{\mu_k}\big(\nabla\phi_{\mu_k}(\mathrm{Id}) - \tau\nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k)\big).$$

which recovers Wasserstein gradient descent if $\phi_\mu = \frac{1}{2}\|\mathrm{T}\|^2_{L^2(\mu)}$.

Implementation. Let ϕ_μ be a **pushforward compatible** functional, *i.e.* there exists $\phi : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ such that for all $\mathrm{T} \in L^2(\mu)$, $\phi_\mu(\mathrm{T}) = \phi(\mathrm{T}_{\#}\mu)$. In that case $\nabla\phi_{\mu_k}(\mathrm{T}_{k+1}) = \nabla_{\mathrm{W}_2}\phi\big((\mathrm{T}_{k+1})_{\#}\mu_k\big) \circ \mathrm{T}_{k+1}$.

But if $\nabla\phi^*_\mu$ is unknown, the scheme is implicit in T_{k+1}, and we can solve it with Newton's method.

- in the special case $\phi^V_\mu(\mathrm{T}) = \int V \circ \mathrm{T}\; \mathrm{d}\mu$ the scheme reads as $\mathrm{T}_{k+1} = \nabla V^* \circ \big(\nabla V - \tau\nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k)\big)$, which recovers (standard) mirror descent.
- the scheme is also implementable for ϕ_μ's that are not pushforward compatible (e.g. SVGD [Liu et al., 2016], EKS [Garbuno-Inigo et al., 2020] algorithms pick $\phi_\mu(\mathrm{T}) = \frac{1}{2}\|P_\mu \mathrm{T}\|^2_{L^2(\mu)}$)

Continuous time

Informally, in continuous time we have:

$$\frac{\mathrm{d}}{\mathrm{d}t}\nabla_{\mathrm{W}_2}\phi(\mu_t) = -\nabla_{\mathrm{W}_2}\mathcal{F}(\mu_t).$$

However, $\frac{\mathrm{d}}{\mathrm{d}t}\nabla_{\mathrm{W}_2}\phi(\mu_t) = \mathrm{H}\phi_{\mu_t}(v_t)$ where $\mathrm{H}\phi_{\mu_t} : L^2(\mu_t) \to L^2(\mu_t)$ is the Hessian operator defined such that $\frac{\mathrm{d}^2}{\mathrm{d}t^2}\phi(\mu_t) = \langle \mathrm{H}\phi_{\mu_t}(v_t), v_t\rangle_{L^2(\mu_t)}$ and $v_t \in L^2(\mu_t)$ is a velocity field satisfying $\partial_t\mu_t + \mathrm{div}(\mu_t v_t) = 0$. Thus, the continuity equation followed by the Mirror Flow is given by

$$\partial_t\mu_t + \mathrm{div}\left(\mu_t(\mathrm{H}\phi_{\mu_t})^{-1}(-\nabla_{\mathrm{W}_2}\mathcal{F}(\mu_t))\right) = 0. \tag{4}$$

For **specific choices** of ϕ and $\mathcal{F}$, this continuous formulation coincides with

- mirror Langevin [Ahn and Chewi, 2021, Wibisono, 2019] ($\mathcal{F}(\mu) = \mathrm{KL}(\mu|\mu^*)$, $\phi(\mu) = \int V d\mu$)
- Information Newton's flows [Wang and Li, 2020] ($\phi = \mathcal{F}$)
- Sinkhorn's flow [Deb et al., 2023] ($\mathcal{F}(\mu) = \mathrm{KL}(\mu|\mu^*), \phi(\mu) = \mathrm{W}_2^2(\mu, \nu)$)

Main assumptions

Recall we optimize $\mathcal{F}$ on $\mathcal{P}_2(\mathbb{R}^d)$ and we defined $\tilde{\mathcal{F}}_\mu(T) = \mathcal{F}(T_\# \mu)$ on $L^2(\mu)$, similarly for ϕ on $\mathcal{P}_2(\mathbb{R}^d)$ we denote $\phi_\mu(T) = \phi(T_\# \mu)$.

If $\mathcal{F}$ is Wasserstein differentiable, then $\tilde{\mathcal{F}}_\mu$ is Fréchet differentiable, and for all $\mathrm{S} \in Dom(\tilde{\mathcal{F}}_\mu)$, $\nabla \tilde{\mathcal{F}}_\mu(\mathrm{S}) = \nabla_{\mathrm{W}_2} \mathcal{F}(\mathrm{S}_\# \mu) \circ \mathrm{S}$.

Definition (Rel. smoothness and convexity, restricted)

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

We say that $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ **is α-convex (resp. β-smooth) relative to** $\phi : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ **along** $t \mapsto \mu_t$ if for all $s, t \in [0,1]$, $\mathrm{D}_{\tilde{\mathcal{F}}_\mu}(\mathrm{T}_s, \mathrm{T}_t) \geq \alpha\, \mathrm{D}_{\phi_\mu}(\mathrm{T}_s, \mathrm{T}_t)$ (resp. $\mathrm{D}_{\tilde{\mathcal{F}}_\mu}(\mathrm{T}_s, \mathrm{T}_t) \leq \beta\, \mathrm{D}_{\phi_\mu}(\mathrm{T}_s, \mathrm{T}_t)$).

We define the "appropriate OT problem": for all $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\mathrm{W}_\phi(\nu, \mu) = \inf_{\gamma \in \Pi(\nu,\mu)} \phi(\nu) - \phi(\mu) - \int \langle \nabla_{\mathrm{W}_2} \phi(\mu)(y), x - y \rangle \, \mathrm{d}\gamma(x,y). \quad (5)$$

It coincides with the Bregman-Wasserstein divergence [Rankin and Wong, 2023] in the case where ϕ is a potential (linear) energy, but is strictly more general. We need to assume $\nabla_{W_2} \phi(\mu)$ invertible.

Results

In this case we can leverage Brenier's theorem [Brenier, 1991], and show that the optimal coupling of (5) is of the form $(\mathrm{T}_{\phi_\mu}^{\mu,\nu}, \mathrm{Id})_{\#}\mu$ with $\mathrm{T}_{\phi_\mu}^{\mu,\nu} = \arg\min_{\mathrm{T}_{\#}\mu=\nu} \mathrm{D}_{\phi_\mu}(\mathrm{T}, \mathrm{Id})$.

This is needed in the proof to telescope consecutive distances between iterates and the global minimizer. It is not as direct as in $\mathbb{R}^d$, because in our case the minimization problem of each iteration happens in a different space $L^2(\mu_k)$.

Theorem (Rates of convergence)

Let $\beta \geq \alpha > 0$, $\tau \leq \frac{1}{\beta}$. Assume for all $k \geq 0$, $\mathcal{F}$ is β-smooth relative to ϕ along $t \mapsto \big((1-t)\mathrm{Id} + t\,\mathrm{T}_{k+1}\big)_{\#}\mu_k$; and that $\mathcal{F}$ is α-convex relative to ϕ along the curves $t \mapsto ((1-t)\mathrm{Id} + t\,\mathrm{T}_{\phi_{\mu_k}}^{\mu_k,\nu})_{\#}\mu_k$. Then, for all $k \geq 1$,

$$\mathcal{F}(\mu_k) - \mathcal{F}(\nu) \leq \alpha\big((1-\tau\alpha)^{-k} - 1\big)^{-1}\mathrm{W}_\phi(\nu, \mu_0) \leq \frac{1-\alpha\tau}{k\tau}\mathrm{W}_\phi(\nu, \mu_0). \quad (6)$$

Moreover, if $\alpha > 0$, taking $\nu = \mu^*$ the minimizer of $\mathcal{F}$, we obtain a linear rate: for all $k \geq 0$, $\mathrm{W}_\phi(\mu^*, \mu_k) \leq (1-\tau\alpha)^k\, \mathrm{W}_\phi(\mu^*, \mu_0)$.

Outline

Recall we are interested in:

$$\mathrm{T}_{k+1} = \operatorname*{arg\,min}_{\mathrm{T}\in L^2(\mu_k)} \langle \nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k), \mathrm{T}-\mathrm{Id}\rangle_{L^2(\mu_k)} + \frac{1}{\tau}\mathrm{D}(\mathrm{T},\mathrm{Id}),$$
$$\mu_{k+1} = (\mathrm{T}_{k+1})_\#\mu_k.$$

Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $h : \mathbb{R}^d \to \mathbb{R}$ proper and strictly convex on $\mathbb{R}^d$. We consider in this section $\phi_\mu^h(\mathrm{T}) = \int h \circ \mathrm{T}\ \mathrm{d}\mu$ and

$$\mathrm{D}(\mathrm{T},\mathsf{Id}) = \phi_{\mu_k}^h\big((\mathsf{Id}-\mathrm{T})/\tau\big)\tau = \int h\big((x-\mathrm{T}(x))/\tau\big)\tau\ \mathrm{d}\mu_k(x).$$

This type of discrepancy is **analogous to OT costs** with translation-invariant ground cost $c(x,y) = h(x-y)$.

Here, the scheme writes:

$$\mathrm{T}_{k+1} = \operatorname*{arg\,min}_{\mathrm{T}\in L^2(\mu_k)} \langle \nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k), \mathrm{T}-\mathsf{Id}\rangle_{L^2(\mu_k)} + \int h\left(\frac{x-\mathrm{T}(x)}{\tau}\right)\tau\ \mathrm{d}\mu_k(x).$$

Deriving the first order conditions, we obtain the following update

$$\forall k\geq 0,\ \mathrm{T}_{k+1} = \mathsf{Id} - \tau(\nabla\phi_{\mu_k}^h)^{-1}\big(\nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k)\big) = \mathsf{Id} - \tau\nabla h^* \circ \nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k).$$

More generally, for ϕ_μ strictly convex, proper, differentiable and superlinear, we have $(\nabla\phi_\mu)^{-1} = \nabla\phi_\mu^*$.

Outline

Mirror Descent

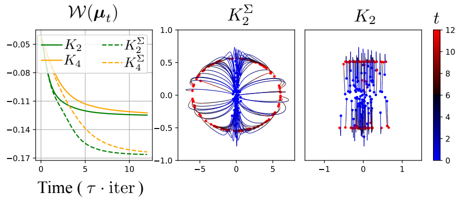


Figure: **(Left)** Value of $\mathcal{W}$ along the flow for two difference interaction Bregman potentials, **(Middle and Right)** Trajectories of particles to minimize $\mathcal{W}$.

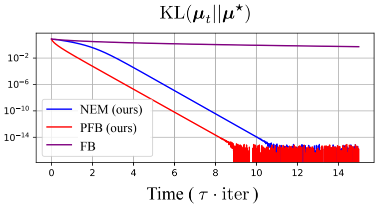


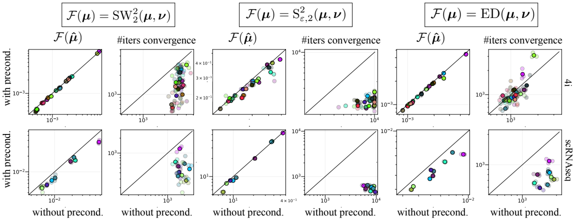
Figure: Convergence towards Gaussians $\mathcal{N}(0, UDU^T)$ averaged over 20 covariances, with $U \sim \mathrm{Unif}(O_{10}(\mathbb{R}))$ and D fixed.

Left figure. both $\mathcal{F} = \mathcal{W}$ and ϕ are interaction energies with kernel W and K respectively. $W(x) = \frac{1}{4}\|x\|^4_{\Sigma^{-1}} - \frac{1}{2}\|x\|^2_{\Sigma^{-1}}$ with $\Sigma \in S_d^{++}(\mathbb{R})$, $K_4(x) = \frac{1}{4}\|x\|_2^4 + \frac{1}{2}\|x\|_2^2$, $K_2(x) = \frac{1}{2}\|x\|_2^2$, $K_4^\Sigma(x) = \frac{1}{4}\|x\|^4_{\Sigma^{-1}} + \frac{1}{2}\|x\|^2_{\Sigma^{-1}}$, $K_2^\Sigma(x) = \frac{1}{2}\|x\|^2_{\Sigma^{-1}}$.

Right figure. $\mathcal{F}(\mu) = \int V \mathrm{d}\mu + \mathcal{H}(\mu)$ for $V(x) = \frac{1}{2}x^T\Sigma^{-1}x$ with $\Sigma = UDU^T$ ill-conditioned. NEM = MD with $\phi(\mu) = \int \log(\mu) d\mu$, PFB = Forward-Backward scheme (PFB) with Bregman potential $\phi(\mu) = \int V d\mu$, FB = standard FB schemes on Gaussians [Diao et al., 2023].

Predicting responses of cells to treatment with PGD

Idea: match a population of control cells μ to treated cells ν minimizing $\mathcal{F} = D(\mu, \nu)$. Prediction $\hat{\mu} = \min_\mu \mathcal{F}(\mu)$. We use $h^*(x) = (\|x\|_2^a + 1)^{1/a} - 1$ with $a \in \{1.25, 1.5, 1.75\}$, which is well suited to minimize functions which grow in $\|x - x^*\|^{a/(a-1)}$ near x^*.



- lines: cells measured with 2 different profiling technologies
- columns/subcolumns : different objectives $\mathcal{F}$/ measures of convergence (final objective and # iters to get to fixed)
- points/colors : (i corresponds to a treatment) $z_i = (x_i, y_i)$ where (first column) y_i is the attained minima $\mathcal{F}(\hat{\mu}) = D(\hat{\mu}, \nu_i)$ with preconditioning and x_i that without preconditioning, and (second column) y_i is the number of iterations to reach convergence with preconditioning and x_i that without preconditioning.

Point below the diagonal = experiment where PGD provides a better minima or faster convergence than GD.

Outline

What is also in the paper:

- theoretical guarantees for splitting schemes

What is missing:

- more examples of relatively smooth and convex pairs of objective functionals $\mathcal{F}$ and Bregman potentials ϕ (eg when $\mathcal{F}$ is the KL, or not a free energy?)

Thank you !

What is also in the paper:

- theoretical guarantees for splitting schemes

What is missing:

- more examples of relatively smooth and convex pairs of objective functionals $\mathcal{F}$ and Bregman potentials ϕ (eg when $\mathcal{F}$ is the KL, or not a free energy?)

Thank you !

References I

Ahn, K. and Chewi, S. (2021).
Efficient constrained sampling via the mirror-langevin algorithm.
Advances in Neural Information Processing Systems, 34:28405–28418.

Ambrosio, L., Gigli, N., and Savaré, G. (2008).
Gradient flows: in metric spaces and in the space of probability measures.
Springer Science & Business Media.

Beck, A. and Teboulle, M. (2003).
Mirror descent and nonlinear projected subgradient methods for convex optimization.
Operations Research Letters, 31(3):167–175.

Bonnet, B. (2019).
A pontryagin maximum principle in wasserstein spaces for constrained optimal control problems.
ESAIM: Control, Optimisation and Calculus of Variations, 25:52.

Brenier, Y. (1991).
Polar factorization and monotone rearrangement of vector-valued functions.
Communications on pure and applied mathematics, 44(4):375–417.

Deb, N., Kim, Y.-H., Pal, S., and Schiebinger, G. (2023).
Wasserstein mirror gradient flow as the limit of the sinkhorn algorithm.
arXiv preprint arXiv:2307.16421.

Diao, M. Z., Balasubramanian, K., Chewi, S., and Salim, A. (2023).
Forward-backward gaussian variational inference via jko in the bures-wasserstein space.
In *International Conference on Machine Learning*, pages 7960–7991. PMLR.

References II

Garbuno-Inigo, A., Hoffmann, F., Li, W., and Stuart, A. M. (2020).
Interacting langevin diffusions: Gradient structure and ensemble kalman sampler.
SIAM Journal on Applied Dynamical Systems, 19(1):412–441.

Jordan, R., Kinderlehrer, D., and Otto, F. (1998).
The variational formulation of the fokker–planck equation.
SIAM journal on mathematical analysis, 29(1):1–17.

Lanzetti, N., Bolognani, S., and Dörfler, F. (2022).
First-order conditions for optimization in the wasserstein space.
arXiv preprint arXiv:2209.12197.

Liu, Q., Lee, J., and Jordan, M. (2016).
A kernelized stein discrepancy for goodness-of-fit tests.
In *International conference on machine learning*, pages 276–284.

Lu, H., Freund, R. M., and Nesterov, Y. (2018).
Relatively smooth convex optimization by first-order methods, and applications.
SIAM Journal on Optimization, 28(1):333–354.

Maddison, C. J., Paulin, D., Teh, Y. W., and Doucet, A. (2021).
Dual space preconditioning for gradient descent.
SIAM Journal on Optimization, 31(1):991–1016.

Otto, F. and Villani, C. (2000).
Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality.
Journal of Functional Analysis, 173(2):361–400.

References III

Rankin, C. and Wong, T.-K. L. (2023).
Bregman-wasserstein divergence: geometry and applications.
arXiv preprint arXiv:2302.05833.

Villani, C. (2009).
Optimal transport: old and new, volume 338.
Springer.

Wang, Y. and Li, W. (2020).
Information newton's flow: second-order optimization method in probability space.
arXiv preprint arXiv:2001.04341.

Wibisono, A. (2019).
Proximal Langevin algorithm: Rapid convergence under isoperimetry.
arXiv preprint arXiv:1911.01469.