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 $\int \int W(x - y)d\mu(x)d\mu(y)$, negative entropy $\int \log(\mu(x))d\mu(x)$
- Distance or divergence to a target probability distribution $\mu^*$ (e.g. $W_2(\mu, \mu^*)$...)

$$^*W_2^2(\nu, \mu) = \inf_{s \in \Pi(\nu, \mu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 ds(x, y), \text{ where } \Pi(\nu, \mu) \text{= couplings between } \nu, \mu.$$
Outline

1. Background on Wasserstein geometry
2. Mirror descent
3. Preconditioned gradient descent
4. Applications and Experiments
5. Conclusion
Brenier’s theorem. Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ s.t. $\mu \ll \text{Leb}$. Then, there exists a unique $T^{\mu,\nu} : \mathbb{R}^d \to \mathbb{R}^d$ such that

1. $T^{\mu,\nu}_# \mu = \nu$

2. $W_2^2(\mu, \nu) = \| \text{Id} - T^{\mu,\nu} \|_{L^2(\mu)}^2 \overset{\text{def.}}{=} \int \| x - T^{\mu,\nu}(x) \|^2 d\mu(x)$.

and $T^{\mu,\nu}$ is called the Optimal Transport map between $\mu$ and $\nu$. The path

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

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

$F$ is said to be $\alpha$-geodesically (or displacement) convex if it is convex along the curves $\rho_t$ defined as above:

$$F(\rho_t) \leq (1 - t)F(\mu) + tF(\nu) - \frac{\alpha t(1 - t)}{2} W_2^2(\mu, \nu),$$
Equipped with the Wasserstein-2 ($W_2$) distance, the metric space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ 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 $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $T : \mathbb{R}^d \to \mathbb{R}^d$ a measurable map. The pushforward measure $T_# \mu$ is characterized by: $X \sim \mu \implies T(X) \sim T_# \mu$. If $T \in L^2(\mu)$, then $T_# \mu \in \mathcal{P}_2(\mathbb{R}^d)$.
Equipped with the Wasserstein-2 ($W_2$) distance, the metric space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ has a convenient **Riemannian structure** [Otto and Villani, 2000].

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

where $L^2(\mu) = \{ f : \mathbb{R}^d \to \mathbb{R}^d, \int_{\mathbb{R}^d} \| f(x) \|^2 \, d\mu(x) < \infty \}$. 
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 $\mathcal{F}'(\mu) : \mathbb{R}^d \rightarrow \mathbb{R}$ is the **First Variation** of $\mathcal{F}$ at $\mu$.

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

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

$$\mathcal{F}((\text{Id} + \varepsilon h) \# \mu) = \mathcal{F}(\mu) + \varepsilon \langle \nabla_{W_2} \mathcal{F}(\mu), h \rangle_{L^2(\mu)} + o(\varepsilon),$$

then $\nabla_{W_2} \mathcal{F}(\mu) \in L^2(\mu)$ is a **Wasserstein gradient** of $\mathcal{F}$ at $\mu$. Typically, $\nabla_{W_2} \mathcal{F}(\mu) = \nabla \mathcal{F}'(\mu)$. 

More formally. Notice that \((\text{Id} + \varepsilon h)\) generate optimal transport maps for \(\varepsilon\) small. In the following, we use the differential structure of \((\mathcal{P}_2(\mathbb{R}^d), W_2)\) introduced in [Bonnet, 2019, Lanzetti et al., 2022].

We say that \(\nabla_{W_2} F(\mu)\) is a Wasserstein gradient of \(F\) at \(\mu \in \text{Dom}(F)\) if for any \(\nu \in \mathcal{P}_2(\mathbb{R}^d)\) and any optimal coupling \(\gamma \in \Pi_o(\mu, \nu)\),

\[
F(\nu) = F(\mu) + \int \langle \nabla_{W_2} F(\mu)(x), y - x \rangle \ d\gamma(x, y) + o(W_2(\mu, \nu)).
\] (1)

If such a gradient exists, then we say that \(F\) is \(W_2\)-differentiable at \(\mu\).

- 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 \(\mathcal{V}(\mu) = \int V d\mu\) or interaction energies \(\mathcal{W}(\mu) = \int \int W(x - y) \ d\mu(x) d\mu(y)\) for \(V\) and \(W\) differentiable and \(L\)-smooth.

- the negative entropy defined as \(\mathcal{H}(\mu) = \int \log (\mu(x)) d\mu(x)\) is not \(W_2\)-differentiable. In this case, we can consider subgradients \(\nabla_{W_2} F(\mu)\) at \(\mu\) for which (1) becomes an inequality.
Wasserstein gradient flows (WGF) [Ambrosio et al., 2008]

The curve $\mu : [0, \infty] \to P_2(\mathbb{R}^d)$, $t \mapsto \mu_t$ is a Wasserstein gradient flow of $\mathcal{F}$ if:

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

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

- **Explicit (WGD)**
  \[
  T_{k+1} = \arg\min_{T \in L^2(\mu_k)} \langle \nabla W_2(\mu_k), T - \text{Id} \rangle_{L^2(\mu_k)} + \frac{1}{2\tau} \| T - \text{Id} \|_{L^2(\mu_k)}^2
  \]
  and $\mu_{k+1} = T_{k+1}\# \mu_k = (\text{Id} - \tau \nabla W_2 F(\mu_k)) \# \mu_k$.

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

\[
  x_{k+1}^i = x_k^i - \tau \nabla W_2 F(\hat{\mu}_k)(x_k^i) \quad \text{for } i = 1, \ldots, n, \text{ where } \hat{\mu}_k = \frac{1}{n} \sum_{i=1}^n \delta_{x_k^i}. \quad (2)
\]

In particular, if $F(\mu)$ is well-defined for discrete measures $\mu$, Algorithm (2) simply corresponds to gradient descent of $F : \mathbb{R}^{n \times d} \to \mathbb{R}$,

$F(x^1, \ldots, x^n) := F(\mu^n)$ where $\mu^n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$. 
Outline

1. Background on Wasserstein geometry
2. Mirror descent
3. Preconditioned gradient descent
4. Applications and Experiments
5. Conclusion
Mirror Descent on $\mathbb{R}^d$

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

$$x_{k+1} = \arg \min_{x \in \mathbb{R}^d} \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{\tau} D_\phi(x, x_k)$$

(3)

where $D_\phi$ is a Bregman divergence, i.e.

$$D_\phi(x, x_k) = \phi(x) - \phi(x_k) - \langle \nabla \phi(x_k), x - x_k \rangle$$

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

Implementation. FOC of (3):

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

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

where $\phi^*$ is the Legendre transform of $\phi$.

Guarantees. [Lu et al., 2018] obtained rates for relatively smooth and convex functions, i.e. $\alpha D_\phi(x, y) \leq D_f(x, y) \leq \beta D_\phi(x, y)$ (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 $k \geq 0$,

$$T_{k+1} = \arg \min_{T \in L^2(\mu_k)} \langle \nabla W_2 \mathcal{F}(\mu_k), T - \text{Id} \rangle_{L^2(\mu_k)} + \frac{1}{\tau} D(T, \text{Id}),$$

$$\mu_{k+1} = (T_{k+1}) \# \mu_k,$$

with different costs $D : L^2(\mu_k) \times L^2(\mu_k) \to \mathbb{R}_+$, 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]

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

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 $T, S \in L^2(\mu)$ as

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

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

$$D_{\phi_\mu}(T, S) = \frac{1}{2} \| T - S \|_{L^2(\mu)}^2.$$

**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 $\psi$ is $\beta$-smooth (respectively $\alpha$-convex) relative to $\phi$ if and only if for all $T, S \in L^2(\mu)$,

$$D_{\psi_\mu}(T, S) \leq \beta D_{\phi_\mu}(T, S) \quad \text{(respectively)} \quad D_{\psi_\mu}(T, S) \geq \alpha D_{\phi_\mu}(T, S).$$

- if $\psi_\mu, \phi_\mu$ are potential energies, relative notions on $\mathbb{R}^d$ translate directly.
- geodesic convexity corresponds to choosing $\phi_\mu$ the $L^2$ norm, $\psi_\mu$ the objective functional and considering OT maps and identity.
Mirror descent on $\mathcal{P}_2(\mathbb{R}^d)$

\[ T_{k+1} = \arg\min_{T \in L^2(\mu_k)} D_{\phi_{\mu_k}}(T, \text{Id}) + \tau \langle \nabla W_2 \mathcal{F}(\mu_k), T - \text{Id} \rangle_{L^2(\mu_k)}, \quad \mu_{k+1} = (T_{k+1})_\# \mu_k. \]

FOC lead to

\[ \nabla \phi_{\mu_k}(T_{k+1}) = \nabla \phi_{\mu_k}(\text{Id}) - \tau \nabla W_2 \mathcal{F}(\mu_k) \iff T_{k+1} = \nabla \phi^*_{\mu_k} \left( \nabla \phi_{\mu_k}(\text{Id}) - \tau \nabla W_2 \mathcal{F}(\mu_k) \right). \]

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

**Implementation.** Let $\phi_{\mu}$ be a **pushforward compatible** functional, i.e. there exists $\phi : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ such that for all $T \in L^2(\mu)$, $\phi_{\mu}(T) = \phi(T_\# \mu)$. In that case $\nabla \phi_{\mu_k}(T_{k+1}) = \nabla W_2 \phi((T_{k+1})_\# \mu_k) \circ T_{k+1}$.

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 $\phi^V_{\mu}(T) = \int V \circ T \ d\mu$ the scheme reads as $T_{k+1} = \nabla V^* \circ (\nabla V - \tau \nabla W_2 \mathcal{F}(\mu_k))$, which recovers (standard) mirror descent.

- the scheme is also implementable for $\phi_{\mu}$’s that are not pushforward compatible (e.g. SVGD [Liu et al., 2016], EKS [Garbuno-Inigo et al., 2020] algorithms pick $\phi_{\mu}(T) = \frac{1}{2} \| P_{\mu} \ T \|^2_{L^2(\mu)}$).
Continuous time

Informally, in continuous time we have:

\[
\frac{d}{dt} \nabla_{W_2} \phi(\mu_t) = -\nabla_{W_2} F(\mu_t).
\]

However, \(\frac{d}{dt} \nabla_{W_2} \phi(\mu_t) = H\phi_{\mu_t}(\nu_t)\) where \(H\phi_{\mu_t} : L^2(\mu_t) \to L^2(\mu_t)\) is the Hessian operator defined such that \(\frac{d^2}{dt^2} \phi(\mu_t) = \langle H\phi_{\mu_t}(\nu_t), \nu_t \rangle_{L^2(\mu_t)}\) and \(\nu_t \in L^2(\mu_t)\) is a velocity field satisfying \(\partial_t \mu_t + \text{div}(\mu_t \nu_t) = 0\). Thus, the continuity equation followed by the Mirror Flow is given by

\[
\partial_t \mu_t + \text{div} \left( \mu_t (H\phi_{\mu_t})^{-1}(-\nabla_{W_2} F(\mu_t)) \right) = 0. \tag{4}
\]

For specific choices of \(\phi\) and \(F\), this continuous formulation coincides with

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

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

If $F$ is Wasserstein differentiable, then $\tilde{F}_\mu$ is Fréchet differentiable, and for all $S \in \text{Dom}(\tilde{F}_\mu)$, $\nabla \tilde{F}_\mu(S) = \nabla_{W_2} F(S#\mu) \circ S$.

**Definition (Rel. smoothness and convexity, restricted)**

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

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

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

$$W_\phi(\nu, \mu) = \inf_{\gamma \in \Pi(\nu, \mu)} \phi(\nu) - \phi(\mu) - \int \langle \nabla_{W_2} \phi(\mu)(y), x - y \rangle \, d\gamma(x, y). \quad (5)$$

It coincides with the Bregman-Wasserstein divergence [Rankin and Wong, 2023] in the case where $\phi$ is a potential (linear) energy, but is strictly more general. We need to assume $\nabla_{W_2} \phi(\mu)$ invertible.
In this case we can leverage Brenier’s theorem [Brenier, 1991], and show that the optimal coupling of (5) is of the form \((T_{\phi_{\mu}}, \text{Id})\#\mu\) with \(T_{\phi_{\mu}} = \arg\min_{T\# \mu = \nu} D_{\phi_{\mu}}(T, \text{Id})\).

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 \(k \geq 0\), \(F\) is \(\beta\)-smooth relative to \(\phi\) along \(t \mapsto ((1-t)\text{Id} + t T_{k+1})\#\mu_k\); and that \(F\) is \(\alpha\)-convex relative to \(\phi\) along the curves \(t \mapsto ((1-t)\text{Id} + t T_{\phi_{\mu_k}}^{\mu_k,\nu})\#\mu_k\). Then, for all \(k \geq 1\),

\[
F(\mu_k) - F(\nu) \leq \alpha((1 - \tau\alpha)^{-k} - 1)^{-1} W_{\phi}(\nu, \mu_0) \leq \frac{1 - \alpha \tau}{k \tau} W_{\phi}(\nu, \mu_0). \tag{6}
\]

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

1. Background on Wasserstein geometry
2. Mirror descent
3. Preconditioned gradient descent
4. Applications and Experiments
5. Conclusion
Recall we are interested in:

$$T_{k+1} = \arg\min_{T \in L^2(\mu_k)} \langle \nabla W_2 F(\mu_k), T - \text{Id} \rangle_{L^2(\mu_k)} + \frac{1}{\tau} D(T, \text{Id}),$$

$$\mu_{k+1} = (T_{k+1}) \# \mu_k.$$

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 $\phi^h_{\mu}(T) = \int h \circ T \ d\mu$ and

$$D(T, \text{Id}) = \phi^h_{\mu_k} \left( (\text{Id} - T) / \tau \right) \tau = \int h((x - T(x)) / \tau) \tau \ d\mu_k(x).$$

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

Here, the scheme writes:

$$T_{k+1} = \arg\min_{T \in L^2(\mu_k)} \langle \nabla W_2 F(\mu_k), T - \text{Id} \rangle_{L^2(\mu_k)} + \int h \left( \frac{x - T(x)}{\tau} \right) \tau \ d\mu_k(x).$$

Deriving the first order conditions, we obtain the following update

$$\forall k \geq 0, \ T_{k+1} = \text{Id} - \tau (\nabla \phi^h_{\mu_k})^{-1} (\nabla W_2 F(\mu_k)) = \text{Id} - \tau \nabla h^* \circ \nabla W_2 F(\mu_k).$$

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

15 / 18
Outline

1. Background on Wasserstein geometry
2. Mirror descent
3. Preconditioned gradient descent
4. Applications and Experiments
5. Conclusion
## Mirror Descent

![Diagram](image)

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

**Left figure.** Both $\mathcal{F} = \mathcal{W}$ and $\phi$ are interaction energies with kernel $W$ and $K$ respectively. $\mathcal{W}(x) = \frac{1}{4} ||x||^4_\Sigma - \frac{1}{2} ||x||^2_\Sigma - 1$ with $\Sigma \in S^{++}_d(\mathbb{R})$, $K_4(x) = \frac{1}{4} ||x||^4_2 + \frac{1}{2} ||x||^2_2$, $K_2(x) = \frac{1}{2} ||x||^2_2$, $K^\Sigma_4(x) = \frac{1}{4} ||x||^4_{\Sigma^{-1}} + \frac{1}{2} ||x||^2_{\Sigma^{-1}}$, $K^\Sigma_2(x) = \frac{1}{2} ||x||^2_{\Sigma^{-1}}$.

**Right figure.** $\mathcal{F}(\mu) = \int V d\mu + \mathcal{H}(\mu)$ 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].

**Figure:** Convergence towards Gaussians $\mathcal{N}(0, UDU^T)$ averaged over 20 covariances, with $U \sim \text{Unif}(O_{10}(\mathbb{R}))$ and $D$ fixed.
Predicting responses of cells to treatment with PGD

Idea: match a population of control cells $\mu$ to treated cells $\nu$ minimizing $F = D(\mu, \nu)$. Prediction $\hat{\mu} = \min_\mu F(\mu)$. We use $h^*(x) = (||x||^a + 1)^{1/a} - 1$ 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 $x^*$.

$F(\mu) = SW_2^2(\mu, \nu)$

$F(\hat{\mu})$

#iters convergence

$F(\hat{\mu})$

#iters convergence

$F(\hat{\mu})$

#iters convergence

---

- lines: cells measured with 2 different profiling technologies
- columns/subcolumns: different objectives $F$/ 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 $F(\hat{\mu}) = D(\hat{\mu}, \nu_i)$ 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

1. Background on Wasserstein geometry
2. Mirror descent
3. Preconditioned gradient descent
4. Applications and Experiments
5. Conclusion
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 $\phi$ (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 $\phi$ (e.g., when $\mathcal{F}$ is the KL, or not a free energy?)

Thank you!
Efficient constrained sampling via the mirror-langevin algorithm.

Gradient flows: in metric spaces and in the space of probability measures.

Mirror descent and nonlinear projected subgradient methods for convex optimization.

Bonnet, B. (2019).
A pontryagin maximum principle in wasserstein spaces for constrained optimal control problems.

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

Wasserstein mirror gradient flow as the limit of the sinkhorn algorithm.

Forward-backward gaussian variational inference via jko in the bures-wasserstein space.
In International Conference on Machine Learning, pages 7960–7991. PMLR.
Interacting langevin diffusions: Gradient structure and ensemble kalman sampler.

The variational formulation of the fokker–planck equation.

First-order conditions for optimization in the wasserstein space.

A kernelized stein discrepancy for goodness-of-fit tests.

Relatively smooth convex optimization by first-order methods, and applications.

Dual space preconditioning for gradient descent.

Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality.
References III


