Mirror and Preconditioned Gradient Descent in Wasserstein Space

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

[1]ENSAE, CREST, Institut Polytechnique de Paris
[2]TU Berlin
[3]TU Wien

NeurIPS in Paris 2024
05/12/2024

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)
- Modeling dynamic of population of cells (Schiebinger et al., 2019)
- 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

Detour by $\mathbb{R}^d$

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

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

- Gradient descent:

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

  - Non-increasing if f β-smooth
  - Converge if f β-smooth and α-strongly convex (*i.e.* $f - \alpha\frac{\|\cdot\|_2^2}{2}$ convex)
- Mirror descent (Lu et al., 2018):

$$\forall k \geq 0, \; x_{k+1} = \nabla\phi^*\big(\nabla\phi(x_k) - \tau\nabla f(x_k)\big)$$

  - Non-increasing if f β-smooth relative to ϕ (*i.e.* $\beta\phi - f$ convex)
  - Converge if f β-smooth and α-convex relative to ϕ (*i.e.* $f - \alpha\phi$ convex)
- Preconditioned gradient descent (Maddison et al., 2021):

$$\forall k \geq 0, \; x_{k+1} = x_k - \tau\nabla h^*\big(\nabla f(x_k)\big)$$

  - Non-increasing if h^* β-smooth relative to f^* (with f^* the Legendre transform)
  - Converge if h^* β-smooth and α-convex relative to f^*

Wasserstein Geometry (Ambrosio et al., 2005)

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

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

Properties:

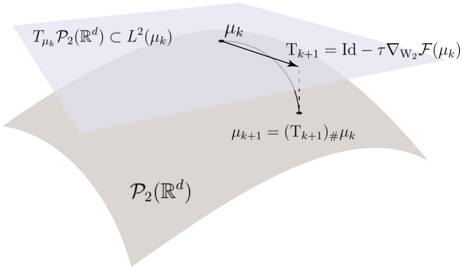
- W_2 distance, $(\mathcal{P}_2(\mathbb{R}^d), \mathrm{W}_2)$: Wasserstein space
- Riemannian structure (with geodesics and tangent space $\mathcal{T}_\mu\mathcal{P}_2(\mathbb{R}^d) \subset L^2(\mu)$)
- Wasserstein gradient $\nabla_{\mathrm{W}_2}\mathcal{F}(\mu) \in \mathcal{T}_\mu\mathcal{P}_2(\mathbb{R}^d)$ of $\mathcal{F}$ at μ satisfies for all $\mathrm{T} \in L^2(\mu)$,

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

Example

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

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

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

Relative Convexity and Smoothness

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

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

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

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

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 $\mu \ll \mathrm{Leb}$ and $\nabla_{\mathrm{W}_2}\phi(\mu)$ is invertible, then $\gamma^* = (\mathrm{T}_{\phi_\mu}^{\mu,\nu}, \mathrm{Id})_\#\mu$, and $\mathrm{W}_\phi(\nu,\mu) = \mathrm{d}_{\phi_\mu}(\mathrm{T}_{\phi_\mu}^{\mu,\nu}, \mathrm{Id})$.

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 $\mathcal{F}^*$:

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

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

Showing Relative Smoothness and Convexity

Smoothness and convexity of $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ relative to $\phi : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$?

- 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 $\mathcal{H}$
- In general: look at the Hessian

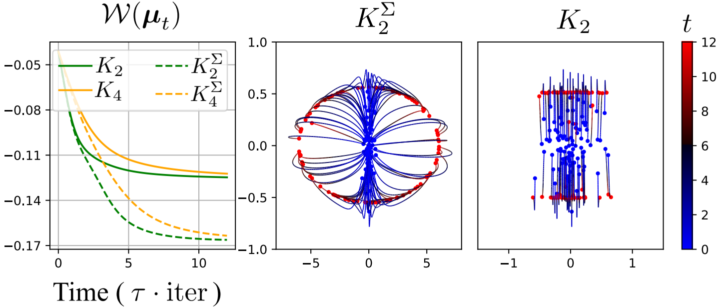
Mirror Descent on Interaction Energy

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

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

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

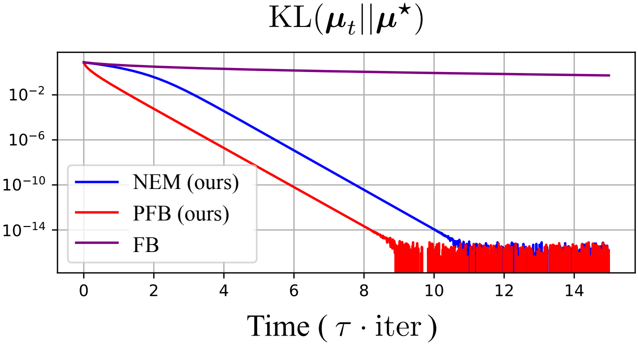
Goal:

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

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

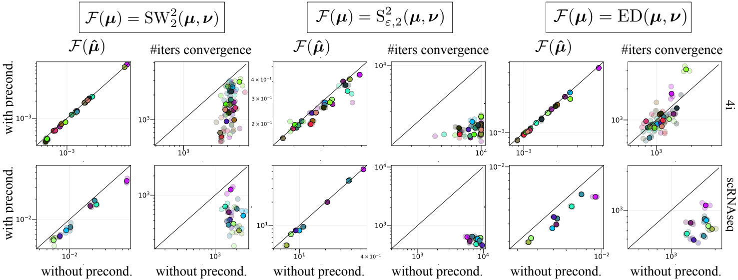
Comparison between:

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

$\rightarrow$ **Points below the diagonal: PGD provides a better minimum or converges faster**

Conclusion

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.

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.

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.

References II

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.

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.

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.