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 2024

Goal

Let $\mathcal{P}_2(\mathbb{R}^d) = \{\mu \in \mathcal{P}_2(\mathbb{R}^d),\ \int \|x\|_2^2 \, \mathrm{d}\mu(x) < \infty\}$, $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$

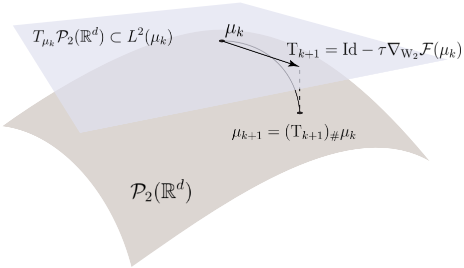
Goal:

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

Applications:

- $\mathcal{F}(\mu) = \mathrm{KL}(\mu || \mu^*)$ for sampling from $\mu^* \propto e^{-V}$
- $\mathcal{F}(\mu) = D(\mu, \nu)$ for modeling the dynamic of population of cells

Setting: **Wasserstein Gradient Descent**



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

Theoretical Results

Results: descent and convergence under relative smoothness and convexity

Mirror Descent: For any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, let $\phi_\mu \in L^2(\mu)$ be a Bregman potential. Then, under assumptions of smoothness and convexity of $\mathcal{F}$ relative to ϕ_μ, and some technical assumptions,

$$\mathcal{F}(\mu_{k+1}) \leq \mathcal{F}(\mu_k) - \beta \mathrm{d}_{\phi_{\mu_k}}(\mathrm{Id}, \mathrm{T}_{k+1}),$$
$$\mathcal{F}(\mu_k) - \mathcal{F}(\mu^*) = \mathcal{O}\left(\frac{1}{k}\right).$$

Preconditioned Gradient Descent: Let $\phi_\mu^h(\mathrm{T}) = \int h \circ \mathrm{T} \, \mathrm{d}\mu$. Under relative smoothness and convexity of $\phi_\mu^{h^*}$ relative to $\mathcal{F}^*$,

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

Implementation of the scheme

Mirror Descent:

- For $\phi_\mu(\mathrm{T}) = \int V \circ \mathrm{T} \, \mathrm{d}\mu$ (Potential energy),

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

  → Wasserstein Mirror Descent (Sharrock et al., 2023)

- For ϕ_μ pushforward compatible (*i.e.* $\phi_\mu(\mathrm{T}) = \phi(\mathrm{T}_\# \mu)$ with $\phi : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$):

$$\forall k \geq 0, \ \nabla_{\mathrm{W}_2} \phi(\mu_{k+1}) \circ \mathrm{T}_{k+1} = \nabla_{\mathrm{W}_2} \phi(\mu_k) - \tau \nabla_{\mathrm{W}_2} \mathcal{F}(\mu_k)$$

  Implicit in T_{k+1} → Newton method

  Example: $\phi_\mu(\mathrm{T}) = \iint W\big(\mathrm{T}(x) - \mathrm{T}(y)\big) \, \mathrm{d}\mu(x)\mathrm{d}\mu(y)$ (Interaction energy)

Preconditioned Gradient Descent:

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

In practice: for $\mu_k = \frac{1}{n} \sum_{i=1}^n \delta_{x_i^k}$ and for all $i \in \{1, \dots, n\}$, $x_i^{k+1} = \mathrm{T}_{k+1}(x_i^k)$.

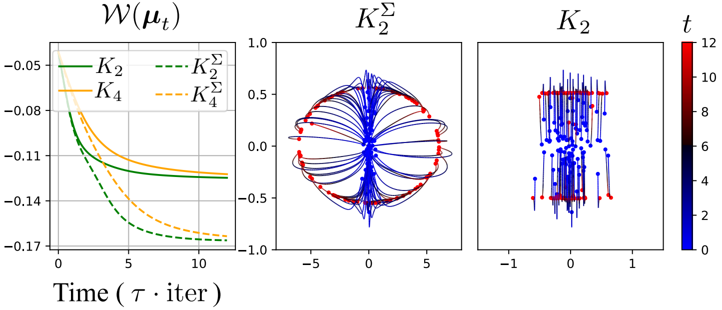
Mirror Descent on Interaction Energy

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

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

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

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



Mirror Descent on Gaussian

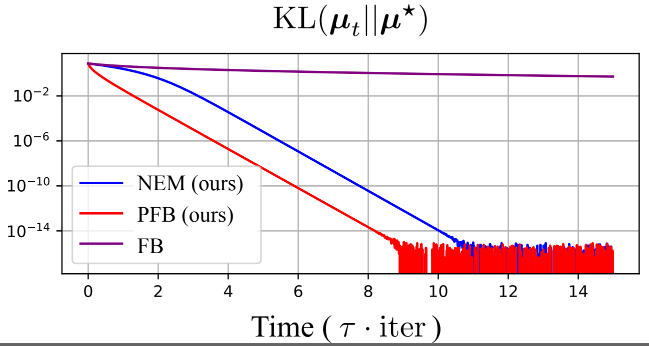
Goal:

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

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

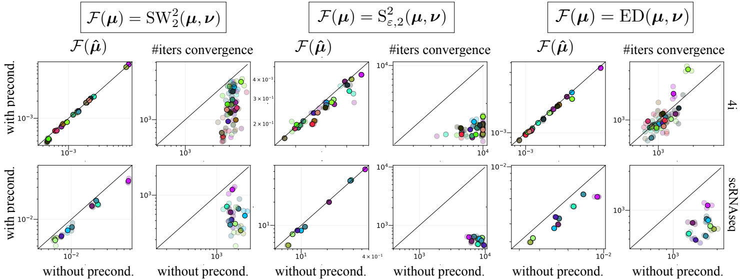
Comparison between:

- 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 $\phi(\mu) = \mathcal{H}(\mu) = \int \log\big(\mu(x)\big)\mathrm{d}\mu(x)$ 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**

Thank you!

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

References I

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

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.

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.

Louis Sharrock, Lester Mackey, and Christopher Nemeth. Learning rate free bayesian inference in constrained domains. In *NeurIPS*, 2023.