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

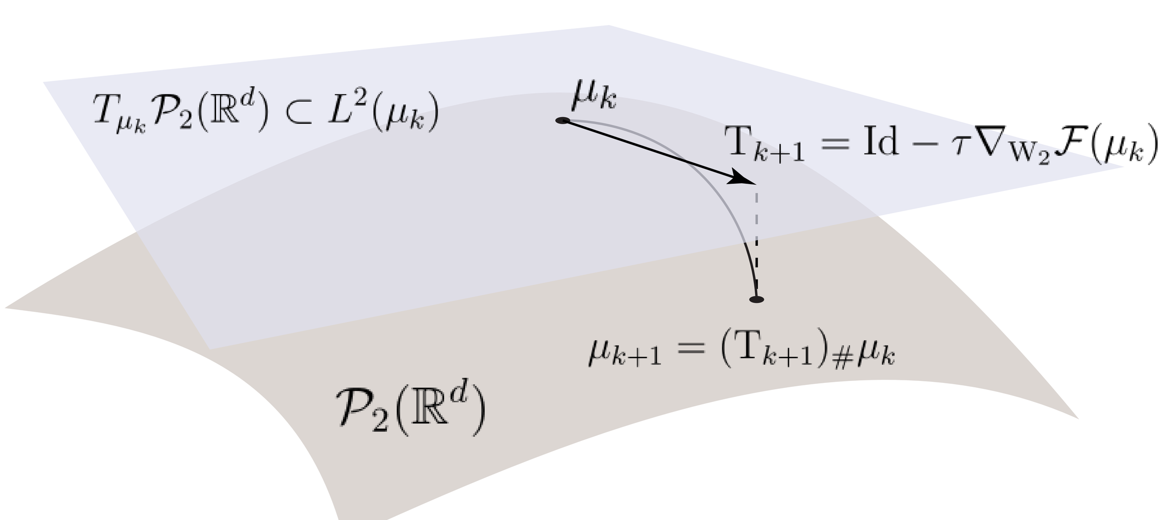
Contributions

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

- Study two optimization schemes of the form

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

- Provide descent and convergence conditions
- Verification of the benefit on experiments



Wasserstein Space

Wasserstein gradient: For $\mathcal{F}:\mathcal{P}_2(\mathbb{R}^d)\to\mathbb{R}$, $\gamma\in\Pi_o(\mu,\nu)$,

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

For $\mathcal{F}:\mathcal{P}_2(\mathbb{R}^d)\to\mathbb{R}$, define $\tilde{\mathcal{F}}_\mu(\mathrm{T}) := \mathcal{F}(\mathrm{T}_\#\mu)$.
If $\mathcal{F}$ W_2-differentiable, $\nabla\tilde{\mathcal{F}}_\mu(\mathrm{T}) = \nabla_{\mathrm{W}_2}\mathcal{F}(\mathrm{T}_\#\mu)\circ\mathrm{T}$.

Examples: potentials $\mathcal{V}_V(\mu) = \int V\,\mathrm{d}\mu$, interactions $\mathcal{W}_W(\mu) = \iint W(x-y)\,\mathrm{d}\mu(x)\mathrm{d}\mu(y)$, entropy $\mathcal{H}(\mu) = \int \log\big(\mu(x)\big)\mathrm{d}\mu(x)$.
$\nabla_{\mathrm{W}_2}\mathcal{V}_V(\mu) = \nabla V$, $\nabla_{\mathrm{W}_2}\mathcal{W}_W(\mu) = \nabla W\star\mu$, $\nabla_{\mathrm{W}_2}\mathcal{H}(\mu) = \nabla\log\mu$

Bregman Divergence and Convexity

Bregman divergence: Let $\phi_\mu: L^2(\mu)\to\mathbb{R}$, $\mathrm{T},\mathrm{S}\in L^2(\mu)$,

$$\mathrm{d}_{\phi_\mu}(\mathrm{T},\mathrm{S}) = \phi_\mu(\mathrm{T}) - \phi_\mu(\mathrm{S}) - \langle\nabla\phi_\mu(\mathrm{S}), \mathrm{T}-\mathrm{S}\rangle_{L^2(\mu)}$$

Relative smoothness/convexity along $t\mapsto\mu_t$ with $\mu_t = (\mathrm{T}_t)_\#\mu$, $\mathrm{T}_t = (1-t)\mathrm{S} + t\mathrm{T}$ for $\mathrm{S},\mathrm{T}\in L^2(\mu)$. $\mathcal{F}$ is β-smooth (resp. α-convex) relative to $\mathcal{G}$ along $t\mapsto\mu_t$ if for all $s,t\in[0,1]$, $\mathrm{d}_{\tilde{\mathcal{F}}_\mu}(\mathrm{T}_s,\mathrm{T}_t)\le\beta\mathrm{d}_{\tilde{\mathcal{G}}_\mu}(\mathrm{T}_s,\mathrm{T}_t)$ (resp. $\mathrm{d}_{\tilde{\mathcal{F}}_\mu}(\mathrm{T}_s,\mathrm{T}_t)\ge\alpha\mathrm{d}_{\tilde{\mathcal{G}}_\mu}(\mathrm{T}_s,\mathrm{T}_t)$).

- For $\mathcal{F}=\mathcal{V}_V$, $\mathcal{G}=\mathcal{V}_U$: holds provided V β-smooth (resp. α-convex) relative to U
- For $\mathcal{F}=\mathcal{W}_W$, $\mathcal{G}=\mathcal{W}_K$: holds provided W β-smooth (resp. α-convex) relative to K
- $\mathcal{F}=\mathcal{V}_V+\mathcal{H}$ 1-convex relative to $\mathcal{V}_V$ and $\mathcal{H}$

Implementation of the Schemes

Mirror descent: $\mathrm{d} = \frac{1}{\tau}\mathrm{d}_{\phi_\mu}$, by FOC: $\nabla\phi_\mu(\mathrm{T}_{k+1}) = \nabla\phi_\mu(\mathrm{Id}) - \tau\nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k)$
For $\phi_\mu(\mathrm{T}) = \int V\circ\mathrm{T}\,\mathrm{d}\mu = \mathcal{V}_V(\mathrm{T}_\#\mu)$, $\mathrm{T}_{k+1} = \nabla V^*\circ\big(\nabla V - \tau\nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k)\big)$
In general: Newton method

Preconditioned gradient descent:

$$\mathrm{d}(\mathrm{T},\mathrm{S}) = \phi_\mu^h\big((\mathrm{S}-\mathrm{T})/\tau\big)\tau = \int h\big((\mathrm{S}(x)-\mathrm{T}(x))/\tau\big)\tau\,\mathrm{d}\mu(x)$$

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

For $\mu_k = \frac{1}{n}\sum_{i=1}^n\delta_{x_i^k}$, for all $k\ge0$, $i\in\{1,\dots,n\}$, $x_i^{k+1} = \mathrm{T}_{k+1}(x_i^k)$.

Theory of Mirror Descent in Wasserstein Space

Let $\beta>0$, $\tau\le\frac{1}{\beta}$. For any $\mu\in\mathcal{P}_2(\mathbb{R}^d)$, let $\phi_\mu: L^2(\mu)\to\mathbb{R}$ be strictly convex, proper and differentiable. Assume $\phi_\mu(\mathrm{T}) = \phi(\mathrm{T}_\#\mu)$ for $\phi:\mathcal{P}_2(\mathbb{R}^d)\to\mathbb{R}$.
Define $\mathrm{W}_\phi(\mu,\nu) = \inf_{\gamma\in\Pi(\mu,\nu)}\ \phi(\nu) - \phi(\mu) - \int\langle\nabla_{\mathrm{W}_2}\phi(\mu)(y), x-y\rangle\,\mathrm{d}\gamma(x,y)$.

Assumptions: Let $\mathrm{T}_{\phi_{\mu_k}}^{\mu_k,\mu^*} = \mathrm{argmin}_{\mathrm{T},\mathrm{T}_\#\mu_k=\mu^*}\ \mathrm{d}_{\phi_{\mu_k}}(\mathrm{T},\mathrm{Id})$. For all $k\ge0$,

1. $\mathcal{F}$ is β-smooth relative to ϕ along $t\mapsto\big((1-t)\mathrm{Id}+t\mathrm{T}_{k+1}\big)_\#\mu_k$
2. $\mathcal{F}$ is α-convex relative to ϕ along the curves $t\mapsto\big((1-t)\mathrm{Id}+t\mathrm{T}_{\phi_{\mu_k}}^{\mu_k,\mu^*}\big)_\#\mu_k$
3. $\mathrm{d}_{\phi_{\mu_k}}(\mathrm{T}_{\phi_{\mu_k}}^{\mu_k,\mu^*},\mathrm{Id}) = \mathrm{W}_\phi(\mu^*,\mu_k)$ and $\mathrm{d}_{\phi_{\mu_k}}(\mathrm{T}_{\phi_{\mu_k}}^{\mu_k,\mu^*},\mathrm{T}_{k+1})\ge\mathrm{W}_\phi(\mu^*,\mu_{k+1})$ (True *e.g.* if $\mu_k,\mu_{k+1}\in\mathcal{P}_{2,ac}(\mathbb{R}^d)$ and $\nabla_{\mathrm{W}_2}\phi(\mu_k)$, $\nabla_{\mathrm{W}_2}\phi(\mu_{k+1})$ invertibles)

Convergence Results

- Under 1), for all $k\ge0$ $\mathcal{F}(\mu_{k+1})\le\mathcal{F}(\mu_k) - \frac{1}{\tau}\mathrm{d}_{\phi_{\mu_k}}(\mathrm{Id},\mathrm{T}_{k+1})$
- Under 1), 2), 3), for all $k\ge1$, $\mathcal{F}(\mu_k) - \mathcal{F}(\mu^*)\le\frac{1-\alpha\tau}{k\tau}\mathrm{W}_\phi(\mu^*,\mu_0)$

Theory of Preconditioned GD in Wasserstein Space

Let $\beta>0$, $\tau\le\frac{1}{\beta}$ and $\bar{\mathrm{T}} = \mathrm{argmin}_{\mathrm{T},\mathrm{T}_\#\mu_k=\mu^*}\ \mathrm{d}_{\tilde{\mathcal{F}}_{\mu_k}}(\mathrm{Id},\bar{\mathrm{T}})$.

Assumptions: For all $k\ge0$,

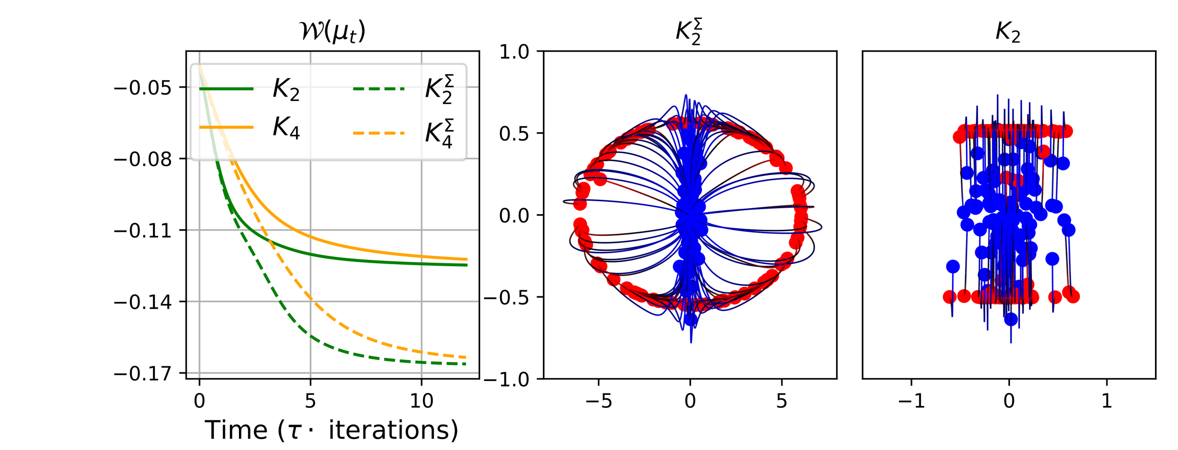
1. $\mathcal{F}$ convex along $t\mapsto\big((1-t)\mathrm{Id}+t\mathrm{T}_{k+1}\big)_\#\mu_k$
2. $\mathrm{d}_{\phi_{\mu_k}^{h^*}}\big(\nabla_{\mathrm{W}_2}\mathcal{F}(\mu_{k+1})\circ\mathrm{T}_{k+1}, \nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k)\big)\le\beta\mathrm{d}_{\tilde{\mathcal{F}}_{\mu_k}}(\mathrm{Id},\mathrm{T}_{k+1})$
3. $\alpha\mathrm{d}_{\tilde{\mathcal{F}}_{\mu_k}}(\mathrm{Id},\bar{\mathrm{T}})\le\mathrm{d}_{\phi_{\mu_k}^{h^*}}\big(\nabla_{\mathrm{W}_2}\mathcal{F}(\bar{\mathrm{T}}_\#\mu_k)\circ\bar{\mathrm{T}}, \nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k)\big)$

Convergence Results

- Under 1), 2), $\phi_{\mu_{k+1}}^{h^*}\big(\nabla_{\mathrm{W}_2}\mathcal{F}(\mu_{k+1})\big)\le\phi_{\mu_k}^{h^*}\big(\nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k)\big) - \frac{1}{\tau}\mathrm{d}_{\tilde{\mathcal{F}}_{\mu_k}}(\mathrm{T}_{k+1},\mathrm{Id})$
- Under 1), 2), 3), $\phi_{\mu_k}^{h^*}\big(\nabla_{\mathrm{W}_2}\mathcal{F}(\mu_k)\big) - h^*(0)\le\frac{1-\tau\alpha}{\tau k}\big(\mathcal{F}(\mu_0)-\mathcal{F}(\mu^*)\big)$

Mirror Descent Experiments

Minimization of an interaction energy $\mathcal{F}(\mu) = \mathcal{W}_W(\mu)$ with $W(z) = \frac{1}{4}\|z\|_{\Sigma^{-1}}^4 - \frac{1}{2}\|z\|_{\Sigma^{-1}}^2$ and $\phi(\mu) = \mathcal{W}_K(\mu)$ with $K_2^\Sigma(z) = \frac{1}{2}\|z\|_{\Sigma^{-1}}^2$, $K_2 = K_2^{I_2}$, $K_4^\Sigma(z) = \frac{1}{4}\|z\|_{\Sigma^{-1}}^4 + \frac{1}{2}\|z\|_{\Sigma^{-1}}^2$, $K_4 = K_4^{I_2}$.

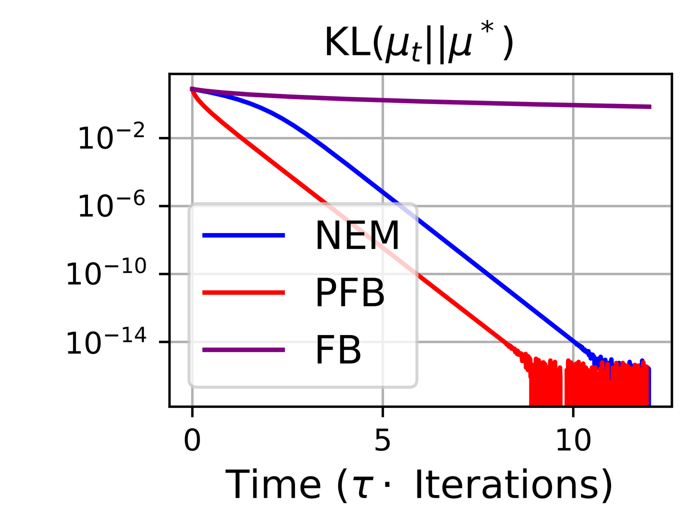


Minimization of

$$\mathcal{F}(\mu) = \mathcal{V}_V(\mu) + \mathcal{H}(\mu),$$

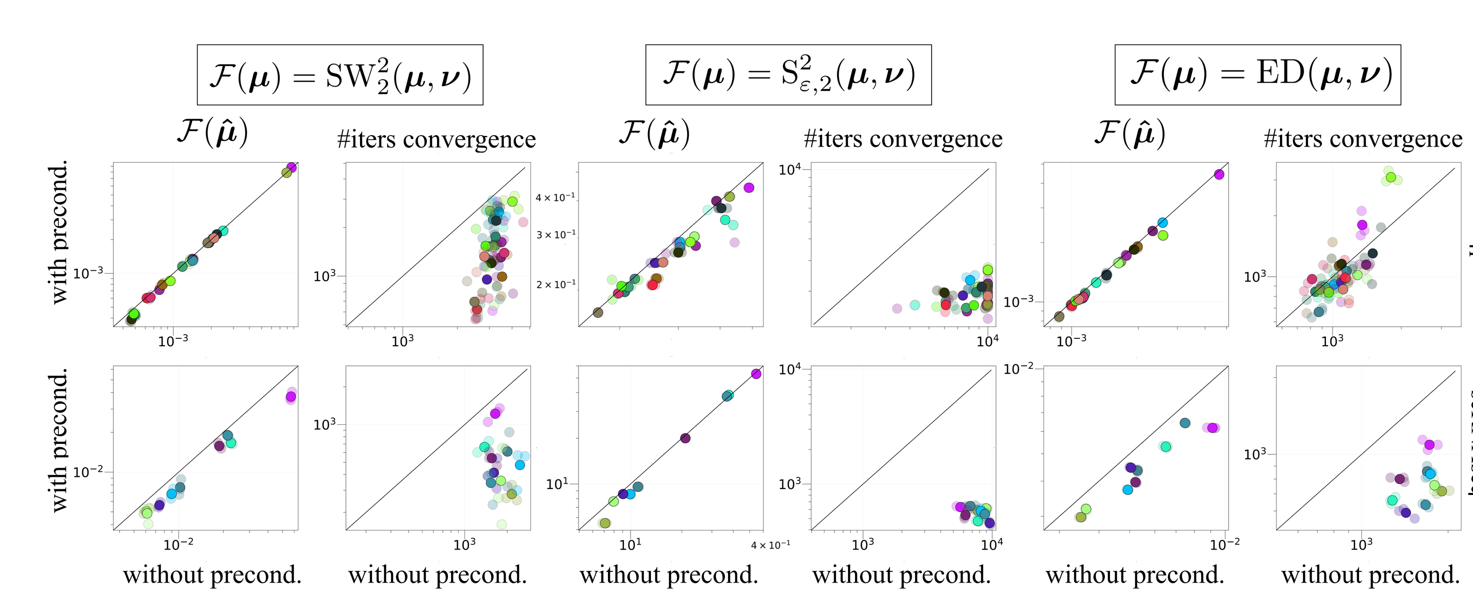
for $V(x) = \frac{1}{2}x^T\Sigma^{-1}x$ with

$\phi(\mu) = \int\frac{1}{2}\|x\|_2^2\,\mathrm{d}\mu(x)$ (FB),
$\phi(\mu) = \mathcal{V}_V(\mu)$ (PFB),
$\phi(\mu) = \mathcal{H}(\mu)$ (NEM).



Preconditioned GD for Single Cells

Minimize $\mathcal{F}(\mu) = D(\mu,\nu)$ with μ_0 untreated cells and ν perturbed cells. Use $h^*(x) = \big(\|x\|_2^a + 1\big)^{1/a} - 1$ with $a\in\{1.25, 1.5, 1.75\}$ which is well suited to minimize functions growing in $\|x-x^*\|^{a/(a-1)}$.



- Rows: 2 profiling technologies
- 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 to converge (2nd subcolumn) without preconditioning and y_i with preconditioning

→ Points below the diagonal: **Preconditioned GD provides a better minimum or converges faster**