Mirror Descent with Relative Smoothness in Measure Spaces, with application to Sinkhorn and EM

Pierre-Cyril Aubin-Frankowski[1], Anna Korba[2], Flavien Léger[1]

[1]INRIA Paris [2]CREST, ENSAE, IP Paris

Quick Summary

- Rigorous proof of convergence of Mirror Descent (MD) under relative smoothness and convexity, in the infinite-dimensional setting of optimization over measure spaces
- New and simple way to derive rates of convergence for Sinkhorn's algorithm as an MD over transport plans
- New expression as MD for EM, convergence rates when restricted to the latent distribution, coincides with Lucy-Richardson

Mirror descent over measures

Let $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{M}(\mathcal{X})$ the space of Radon measures on $\mathcal{X}$, convex functionals $\mathcal{F}, \phi : \mathcal{M}(\mathcal{X}) \to \mathbb{R} \cup \{+\infty\}$, convex $C \subset \mathcal{M}(\mathcal{X})$, consider mirror descent:

$$\min_{\mu \in C} \mathcal{F}(\mu)$$

$$\mu_{n+1} = \operatorname*{argmin}_{\nu \in C} \{ d^+\mathcal{F}(\mu_n)(\nu - \mu_n) + L D_\phi(\nu|\mu_n) \} \tag{1}$$

Under which assumptions does it converge and for which rate?

Examples of optimization of measures

The "Kullback-Leibler divergence" or relative entropy is

$$\mathrm{KL}(\mu|\bar{\mu}) = \begin{cases} \int_{\mathbb{R}^d} \log\left(\frac{\mu}{\bar{\mu}}(x)\right) d\mu(x) & \text{if } \mu \ll \bar{\mu} \\ +\infty & \text{else.} \end{cases}$$

- Entropic optimal transport $\min_{\pi \in \Pi(\mu,\nu)} \mathrm{KL}(\pi|R)$ for $R \propto \exp(-c(x,y)/\epsilon)\mu\otimes\nu$
- Expectation-Maximization $\min_{q\in\mathcal{Q}} \mathrm{KL}(\bar{\nu}|p_{\mathcal{Y}}p_q)$ with the observations $\bar{\nu}$
- Bayesian inference $\min_{\mu\in\mathcal{P}(\mathcal{X})} \mathrm{KL}(\mu|\bar{\mu})$ with the posterior $\bar{\mu} \propto \exp(-V)$
- Optimization of 1-hidden layer neural network $\min_{\mu\in C} \mathrm{MMD}^2(\mu|\bar{\mu})$

Definitions of derivatives and relative smoothness

The KL does not have a Gâteaux derivative! Need for weaker notions:

$$\text{(directional derivative)} \quad d^+\mathcal{F}(\nu)(\mu) = \lim_{h\to 0^+} \frac{\mathcal{F}(\nu + h\mu) - \mathcal{F}(\nu)}{h}, \tag{2}$$

$$\text{(first variation)} \quad \langle \nabla_C \mathcal{F}(\mu), \xi \rangle = d^+\mathcal{F}(\mu)(\xi) \quad \xi - \mu \in \mathrm{dom}(\mathcal{F}) \cap C, \tag{3}$$

$$\text{(Bregman divergence)} \quad D_\phi(\nu|\mu) = \phi(\nu) - \phi(\mu) - d^+\phi(\mu)(\nu - \mu). \tag{4}$$

$\mathcal{F}$ is L-smooth relative to ϕ for $L \geq 0$ if, for any $\mu, \nu \in \mathrm{dom}(\mathcal{F}) \cap \mathrm{dom}(\phi)$,

$$D_{\mathcal{F}}(\nu|\mu) = \mathcal{F}(\nu) - \mathcal{F}(\mu) - d^+\mathcal{F}(\mu)(\nu - \mu) \leq L D_\phi(\nu|\mu).$$

Conversely, $\mathcal{F}$ is l-strongly convex relative to ϕ, for $l \geq 0$, if we have

$$D_{\mathcal{F}}(\nu|\mu) \geq l D_\phi(\nu|\mu).$$

Convergence result for mirror descent

Theorem: Assume that $\mathcal{F}$ is l-strongly convex and L-smooth relative to ϕ, with $l, L \geq 0$. Consider the mirror descent scheme (1), and assume that for each $n \geq 0$, $\nabla\phi(\mu_n)$ exists. Then for all $n \geq 0$ and all $\nu \in \mathrm{dom}(\mathcal{F}) \cap \mathrm{dom}(\phi)$:

$$\mathcal{F}(\mu_n) - \mathcal{F}(\nu) \leq \frac{l D_\phi(\nu|\mu_0)}{\left(1 + \frac{l}{L-l}\right)^n - 1} \leq \frac{L}{n} D_\phi(\nu|\mu_0)$$

Entropic optimal transport and Sinkhorn

$\Pi(\bar{\mu}, *), \Pi(*, \bar{\nu})$ the set of couplings having first/second marginal $\bar{\mu}$, $\bar{\nu}$

$\Pi(\bar{\mu}, \bar{\nu}) = \Pi(\bar{\mu}, *) \cap \Pi(*, \bar{\nu})$ the couplings with marginals $(\bar{\mu}, \bar{\nu})$

For any $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, we can write $\pi = p_{\mathcal{X}}\pi \otimes K_\pi$ where $K_{\bar{\pi}}(x, dy) = \bar{\pi}(dx,dy)/p_{\mathcal{X}}\bar{\pi}(dx)$. Hence we have the decomposition:

$$\mathrm{KL}(\pi|\bar{\pi}) = \int \log\left(\frac{\pi}{\bar{\pi}}\right) d(p_{\mathcal{X}}\pi \otimes K_\pi) = \mathrm{KL}(p_{\mathcal{X}}\pi | p_{\mathcal{X}}\bar{\pi}) + \mathrm{KL}(\pi | p_{\mathcal{X}}\pi \otimes K_{\bar{\pi}}).$$

Define cyclically invariant $\pi \in \Pi_c$, if for $(\mu, \nu) = (p_{\mathcal{X}}\pi, p_{\mathcal{Y}}\pi)$ its marginals,

$$\mathrm{KL}(\pi|e^{-c/\epsilon}\mu\otimes\nu) = \min_{\tilde{\pi}\in\Pi(\mu,\nu)} \mathrm{KL}(\tilde{\pi}|e^{-c/\epsilon}\mu\otimes\nu). \tag{5}$$

When $\pi \in \Pi_c$, there exist $f, g \in L^\infty(\mathcal{X}) \times L^\infty(\mathcal{Y})$ such that $\pi = e^{f+g-c/\epsilon}\mu\otimes\nu$.

The Sinkhorn algorithm in its primal formulation does alternative (entropic) projections on $\Pi(\bar{\mu}, *)$ and $\Pi(*, \bar{\nu})$, i.e. initializing with $\pi_0 \in \Pi_c$, iterate

$$\pi_{n+\frac{1}{2}} = \operatorname*{argmin}_{\pi\in\Pi(\bar{\mu},*)} \mathrm{KL}(\pi|\pi_n), \tag{6}$$

$$\pi_{n+1} = \operatorname*{argmin}_{\pi\in\Pi(*,\bar{\nu})} \mathrm{KL}(\pi|\pi_{n+\frac{1}{2}}). \tag{7}$$

For $c \in L^\infty$, define the constraint set $C = \Pi(*, \bar{\nu})$ and the objective function

$$F_{\mathrm{S}}(\pi) = \mathrm{KL}(p_{\mathcal{X}}\pi|\bar{\mu}). \tag{8}$$

Proposition: The Sinkhorn iterations can be written as a mirror descent with objective F_{S} and Bregman divergence KL over the constraint $C = \Pi(*, \bar{\nu})$, with $\nabla F_{\mathrm{S}}(\pi_n) = \ln(d\mu_n/d\bar{\mu}) \in L^\infty(\mathcal{X}\times\mathcal{Y})$, $\mu_n = p_{\mathcal{X}}\pi_n$

$$\pi_{n+1} = \operatorname*{argmin}_{\pi\in C} \langle \nabla F_{\mathrm{S}}(\pi_n), \pi - \pi_n \rangle + \mathrm{KL}(\pi|\pi_n) \tag{9}$$

Proof: We have the identity:

$$F_{\mathrm{S}}(\pi_n) + \langle \nabla F_{\mathrm{S}}(\pi_n), \pi - \pi_n \rangle + \mathrm{KL}(\pi|\pi_n) = \mathrm{KL}(\pi|\bar{\mu}\otimes{}^{\pi_n}/_{\mu_n}) = \mathrm{KL}(\pi|\pi_{n+\frac{1}{2}}).$$

Lemma: The functional F_{S} is convex and is 1-relatively smooth w.r.t. the negative entropy ϕ_e over $\mathcal{P}(\mathcal{X}\times\mathcal{Y})$.

Consequence: this already yields a $\mathcal{O}(1/n)$ rate for Sinkhorn's algorithm.

Proposition Let $D_c := \frac{1}{2}\sup_{x,y,x',y'}[c(x,y) + c(x',y') - c(x,y') - c(x',y)]$. For $\tilde{\pi}, \pi \in \Pi_c \cap C$, we have that

$$\mathrm{KL}(\tilde{\pi}|\pi) \leq (1 + 4e^{3D_c/\epsilon})\,\mathrm{KL}(p_{\mathcal{X}}\tilde{\pi}|p_{\mathcal{X}}\pi),$$

i.e. F_{S} is $(1 + 4e^{3D_c/\epsilon})^{-1}$-relatively strongly convex w.r.t. KL over $\Pi_c \cap C$.

Consequence: this yields a linear rate for Sinkhorn's algorithm.

Main result for EOT/Sinkhorn

Proposition: For all $n \geq 0$, the Sinkhorn algorithm is a mirror descent and verifies, for π_* the optimum of EOT and μ_* its first marginal,

$$\mathrm{KL}(\mu_n|\mu_*) \leq \frac{\mathrm{KL}(\pi_*|\pi_0)}{(1 + 4e^{\frac{3D_c}{\epsilon}})\left(\left(1 + 4e^{-\frac{3D_c}{\epsilon}}\right)^n - 1\right)} \leq \frac{\mathrm{KL}(\pi_*|\pi_0)}{n}.$$

EM and latent EM

We posit a joint distribution $p_q(dx, dy)$ parametrized by an element q of some given set $\mathcal{Q}$. For $p_{\mathcal{Y}}p_q(dy) = \int_{\mathcal{X}} p_q(dx, dy)$, the goal is to infer q by solving

$$\min_{q\in\mathcal{Q}} \mathrm{KL}(\bar{\nu}|p_{\mathcal{Y}}p_q), \tag{10}$$

EM then proceeds by alternate minimizations of $\mathrm{KL}(\pi, p_q)$:

$$q_n = \operatorname*{argmin}_{q\in\mathcal{Q}} \mathrm{KL}(\pi_n|p_q), \tag{11}$$

$$\pi_{n+1} = \operatorname*{argmin}_{\pi\in\Pi(*,\bar{\nu})} \mathrm{KL}(\pi|p_{q_n}). \tag{12}$$

Define the constraint set $C = \Pi(*, \bar{\nu})$ and

$$F_{\mathrm{EM}}(\pi) = \inf_{q\in\mathcal{Q}} \mathrm{KL}(\pi|p_q). \tag{13}$$

Main result for general EM

Proposition: EM is a mirror descent, with $\nabla F_{\mathrm{EM}}(\pi_n) = \ln(d\pi_n/dp_{q_n})$,

$$\pi_{n+1} = \operatorname*{argmin}_{\pi\in C} \langle \nabla F_{\mathrm{EM}}(\pi_n), \pi - \pi_n \rangle + \mathrm{KL}(\pi|\pi_n) \tag{14}$$

Proof: Use the envelope theorem to differentiate F_{EM} and find that $\nabla F_{\mathrm{EM}}(\pi_n) = \ln(d\pi_n/dp_{q_n})$. Then for any coupling π, we have the identity

$$F_{\mathrm{EM}}(\pi_n) + \langle \nabla F_{\mathrm{EM}}(\pi_n), \pi - \pi_n \rangle + \mathrm{KL}(\pi|\pi_n) = \mathrm{KL}(\pi|p_{q_n}).$$

F_{EM} is in general non-convex. However, writing $p_q(dx, dy) = \mu(dx)K(x, dy)$ and optimizing only over its first marginal, i.e. $q = \mu$, makes F_{EM} convex. Define $F_{\mathrm{LEM}}(\pi) := \mathrm{KL}(\pi|p_{\mathcal{X}}\pi\otimes K) = \inf_{\mu\in\mathcal{P}(\mathcal{X})} \mathrm{KL}(\pi|\mu\otimes K)$

Main result for latent EM

Proposition: Latent EM can be written as mirror descent with objective F_{LEM}, Bregman potential ϕ_e and the constraints $C = \Pi(*, \bar{\nu})$,

$$\pi_{n+1} = \operatorname*{argmin}_{\pi\in C} \langle \nabla F_{\mathrm{LEM}}(\pi_n), \pi - \pi_n \rangle + \mathrm{KL}(\pi|\pi_n) \tag{15}$$

Proposition Set $\mu_* \in \operatorname{argmin}_{\mu\in\mathcal{P}(\mathcal{X})} \mathrm{KL}(\bar{\nu}|T_K(\mu))$ where $T_K : \mu \in \mathcal{P}(\mathcal{X}) \mapsto \int_{\mathcal{X}} \mu(dx)K(x, \cdot) \in \mathcal{M}(\mathcal{Y})$. The functional F_{LEM} is convex and 1-smooth relative to ϕ_e. For $\pi_0 \in \Pi(*, \bar{\nu})$,

$$\mathrm{KL}(\bar{\nu}|T_K\mu_n) \leq \mathrm{KL}(\bar{\nu}|T_K\mu_*) + \frac{\mathrm{KL}(\mu_*|\mu_0) + \mathrm{KL}(\bar{\nu}|T_K\mu_*) - \mathrm{KL}(\bar{\nu}|T_K\mu_0)}{n}.$$

NeurIPS | 2022 Thirty-sixth Conference on Neural Information Processing Systems, New Orleans, November 2022.