Mirror Descent with Relative Smoothness in Measure Spaces, with application to Sinkhorn and Expectation-Maximization (EM)

Pierre-Cyril Aubin-Frankowski[1], Anna Korba[2], Flavien Léger[3]

[1] DI, Ecole normale supérieure, Université PSL, CNRS, INRIA SIERRA, Paris, France
[2] CREST, ENSAE, IP Paris
[3] INRIA MOKAPLAN Paris

NeurIPS 2022

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 of Expectation-Maximization (EM) as MD, convergence rates when restricted to the latent distribution, coincides with Lucy-Richardson's algorithm in signal processing

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

Definitions of derivatives

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

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

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

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

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

Convergence result for mirror descent under relative smoothness

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

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

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

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

Theorem 1

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 $n \geq 0$, $\nabla_C \phi(\mu_n)$ exists. Then for all $n \geq 0$ and all $\nu \in C \cap \mathrm{dom}(\mathcal{F}) \cap \mathrm{dom}(\phi)$:

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

Entropic optimal transport and Sinkhorn

$$\text{Entropic optimal transport} \quad \min_{\pi \in \Pi(\bar{\mu},\bar{\nu})} \mathrm{KL}(\pi | e^{-c/\epsilon} \bar{\mu} \otimes \bar{\nu})$$

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

$$\pi_{n+\frac{1}{2}} = \operatorname*{argmin}_{\pi \in \Pi(\bar{\mu}, *)} \mathrm{KL}(\pi | \pi_n), \tag{5}$$

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

For $c \in L^\infty$, define $C = \Pi(*, \bar{\nu})$ and the objective function $F_{\mathrm{S}}(\pi) = \mathrm{KL}(p_{\mathcal{X}}\pi | \bar{\mu})$.

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 $\nabla F_{\mathrm{S}}(\pi_n) = \ln(d\mu_n / d\bar{\mu}) \in L^\infty(\mathcal{X} \times \mathcal{Y})$, $\mu_n = p_{\mathcal{X}}\pi_n$

$$\pi_{n+1} = \operatorname*{argmin}_{\pi \in C} \langle \nabla F_{\mathrm{S}}(\pi_n), \pi - \pi_n \rangle + \mathrm{KL}(\pi | \pi_n) \tag{7}$$

Entropic optimal transport and Sinkhorn (cont.)

The functional $F_{\mathrm{S}}(\pi) = \mathrm{KL}(p_{\mathcal{X}}\pi|\bar{\mu})$ is convex and is 1-relatively smooth w.r.t. KL over $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$.

$D_c := \frac{1}{2}\sup_{x,y,x',y'}[c(x,y) + c(x',y') - c(x,y') - c(x',y)$. 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 $\Pi_c \cap C$ (cyclically invariant).

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

$$\min_{q \in \mathcal{Q}} \mathrm{KL}(\bar{\nu} | p_{\mathcal{Y}} p_q), \tag{8}$$

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

$$q_n = \operatorname*{argmin}_{q \in \mathcal{Q}} \mathrm{KL}(\pi_n | p_q), \tag{9}$$

$$\pi_{n+1} = \operatorname*{argmin}_{\pi \in \Pi(*, \bar{\nu})} \mathrm{KL}(\pi | p_{q_n}). \tag{10}$$

Define the constraint set $C = \Pi(*, \bar{\nu})$ and $F_{\mathrm{EM}}(\pi) = \inf_{q \in \mathcal{Q}} \mathrm{KL}(\pi | p_q)$.

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

$$\pi_{n+1} = \operatorname*{argmin}_{\pi \in C} \langle \nabla F_{\mathrm{EM}}(\pi_n), \pi - \pi_n \rangle + \mathrm{KL}(\pi | \pi_n) \tag{11}$$

EM and latent EM (cont.)

$F_{\mathrm{EM}} = \inf_{q\in\mathcal{Q}} \mathrm{KL}(\pi|p_q)$ 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 $F_{\mathrm{LEM}}(\pi) := \mathrm{KL}(\pi|p_{\mathcal{X}}\pi \otimes K) = \inf_{\mu\in\mathcal{P}(\mathcal{X})} \mathrm{KL}(\pi|\mu \otimes K)$

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

$$\pi_{n+1} = \operatorname*{argmin}_{\pi\in C} \langle \nabla F_{\mathrm{LEM}}(\pi_n), \pi - \pi_n\rangle + \mathrm{KL}(\pi|\pi_n) \tag{12}$$

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 $\pi_0 \in \Pi(*, \bar{\nu})$,

$$\mathrm{KL}(\bar{\nu}|T_K\mu_n) \le \mathrm{KL}(\bar{\nu}|T_K\mu_*) + \frac{\mathrm{KL}(\mu_*|\mu_0) + \mathrm{KL}(\bar{\nu}|T_K\mu_*) - \mathrm{KL}(\bar{\nu}|T_K\mu_0)}{n}. \tag{13}$$