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

Pierre-Cyril Aubin-Frankowski\textsuperscript{1}, Anna Korba\textsuperscript{2}, Flavien Léger\textsuperscript{3}

\textsuperscript{1} DI, Ecole normale supérieure, Université PSL, CNRS, INRIA SIERRA, Paris, France
\textsuperscript{2} CREST, ENSAE, IP Paris
\textsuperscript{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:

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

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

(1)

Under which assumptions does it converge and at which rate?
Examples of optimization of measures

The “Kullback-Leibler divergence” or relative entropy is

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

- Entropic optimal transport
  $$\min_{\pi \in \Pi(\mu, \nu)} KL(\pi | R) \text{ for } R \propto \exp(-c(x, y)/\epsilon) \mu \otimes \nu$$

- Expectation-Maximization
  $$\min_{q \in \mathcal{Q}} KL(\bar{\nu} | p_Y p_q) \text{ with the observations } \bar{\nu}$$

- Bayesian inference
  $$\min_{\mu \in \mathcal{P}(X)} KL(\mu | \bar{\mu}) \text{ with the posterior } \bar{\mu} \propto \exp(-V)$$

- Optimization of 1-hidden layer neural network
  $$\min_{\mu \in \mathcal{C}} \text{MMD}^2(\mu | \bar{\mu})$$
Definitions of derivatives

\[ \mu_{n+1} = \arg\min_{\nu \in C} \left\{ d^+ F(\mu_n)(\nu - \mu_n) + LD_\phi(\nu|\mu_n) \right\} \]

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

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

(\textit{first variation}) \quad \langle \nabla C F(\mu), \xi \rangle = d^+ F(\mu)(\xi) \quad \xi + \mu \in \text{dom}(F) \cap C, \quad (3)

(\textit{Bregman divergence}) \quad D_\phi(\nu|\mu) = \phi(\nu) - \phi(\mu) - d^+ \phi(\mu)(\nu - \mu). \quad (4)
Convergence result for mirror descent under relative smoothness

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

\[
D_{\mathcal{F}}(\nu|\mu) = \mathcal{F}(\nu) - \mathcal{F}(\mu) - d^+\mathcal{F}(\mu)(\nu - \mu) \leq LD_\phi(\nu|\mu).
\]

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

\[
D_{\mathcal{F}}(\nu|\mu) \geq lD_\phi(\nu|\mu).
\]

Theorem 1

Assume that \( \mathcal{F} \) is \( l \)-strongly convex and \( L \)-smooth relative to \( \phi \), 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 \( \nu \in C \cap \text{dom}(\mathcal{F}) \cap \text{dom}(\phi) \):

\[
\mathcal{F}(\mu_n) - \mathcal{F}(\nu) \leq \frac{ID_\phi(\nu|\mu_0)}{(1 + \frac{l}{L-l})^n - 1} \leq \frac{L}{n}D_\phi(\nu|\mu_0).
\]
Entropic optimal transport and Sinkhorn

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

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

\[ \pi_{n+\frac{1}{2}} = \arg\min_{\pi \in \Pi(\bar{\mu}, \ast)} \text{KL}(\pi | \pi_n), \]  
(5)

\[ \pi_{n+1} = \arg\min_{\pi \in \Pi(\ast, \bar{\nu})} \text{KL}(\pi | \pi_{n+\frac{1}{2}}). \]  
(6)

For \( c \in L^\infty \), define \( C = \Pi(\ast, \bar{\nu}) \) and the objective function \( F_S(\pi) = \text{KL}(p_X \pi | \bar{\mu}) \).

The Sinkhorn iterations can be written as a mirror descent with objective \( F_S \) and Bregman divergence \( \text{KL} \) over the constraint \( C = \Pi(\ast, \bar{\nu}) \), with \( \nabla F_S(\pi_n) = \ln(d_{\mu_n}/d\bar{\mu}) \in L^\infty(\mathcal{X} \times \mathcal{Y}) \),

\[ \mu_n = p_X \pi_n \]

\[ \pi_{n+1} = \arg\min_{\pi \in C} \langle \nabla F_S(\pi_n), \pi - \pi_n \rangle + \text{KL}(\pi | \pi_n) \]  
(7)
Entropic optimal transport and Sinkhorn (cont.)

The functional $F_S(\pi) = KL(p_{\mathcal{X}} \pi | \bar{\mu})$ 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

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

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 $\pi_*$ the optimum of EOT and $\mu_*$ its first marginal,

$$KL(\mu_n | \mu_*) \leq \frac{KL(\pi_* | \pi_0)}{(1 + 4e^{3D_c / \epsilon}) \left( \left( 1 + 4e^{-3D_c / \epsilon} \right)^n - 1 \right)} \leq \frac{KL(\pi_* | \pi_0)}{n}.$$
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 $p_Y p_q(dy) = \int_X p_q(dx, dy)$, the goal is to infer $q$ by solving

$$\min_{q \in \mathcal{Q}} KL(\bar{\nu} | p_Y p_q),$$

(8)

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

$$q_n = \arg\min_{q \in \mathcal{Q}} KL(\pi_n | p_q),$$

(9)

$$\pi_{n+1} = \arg\min_{\pi \in \Pi(\ast, \bar{\nu})} KL(\pi | p_{q_n}).$$

(10)

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

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

$$\pi_{n+1} = \arg\min_{\pi \in C} \langle \nabla F_{EM}(\pi_n), \pi - \pi_n \rangle + KL(\pi | \pi_n)$$

(11)
EM and latent EM (cont.)

\[ F_{EM} = \inf_{q \in Q} \text{KL}(\pi | p_q) \] 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 \( F_{LEM}(\pi) := \text{KL}(\pi | p_X \pi \otimes K) = \inf_{\mu \in \mathcal{P}(X)} \text{KL}(\pi | \mu \otimes K) \)

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

\[ \pi_{n+1} = \arg\min_{\pi \in C} \langle \nabla F_{LEM}(\pi_n), \pi - \pi_n \rangle + \text{KL}(\pi | \pi_n) \quad (12) \]

Set \( \mu_\ast \in \arg\min_{\mu \in \mathcal{P}(X)} \text{KL}(\bar{\nu} | T_K(\mu)) \) where \( T_K : \mu \in \mathcal{P}(X) \mapsto \int_X \mu(dx)K(x, \cdot) \in \mathcal{M}(Y) \). The functional \( F_{LEM} \) is convex and 1-smooth relative to \( \phi_e \). For \( \pi_0 \in \Pi(\ast, \bar{\nu}) \),

\[ \text{KL}(\bar{\nu} | T_K\mu_n) \leq \text{KL}(\bar{\nu} | T_K\mu_\ast) + \frac{\text{KL}(\mu_\ast | \mu_0) + \text{KL}(\bar{\nu} | T_K\mu_\ast) - \text{KL}(\bar{\nu} | T_K\mu_0)}{n} \quad (13) \]