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

Anna Korba

CREST, ENSAE, Institut Polytechnique de Paris

Yes workshop, Eurandom, 2022 - Optimal Transport, Statistics, Machine Learning and moving in between.

Joint work with Pierre-Cyril Aubin-Frankowski and Flavien Léger (INRIA).

Outline

Optimisation over the space of measures

Let $\mathcal{X} \subset \mathbb{R}^d$ and consider $\mathcal{P}(\mathcal{X})$ the space of probability measures on $\mathcal{X}$

Let $\mathcal{F} : \mathcal{P}(\mathcal{X}) \to \mathbb{R} \cup \{+\infty\}$ convex and $C \subset \mathcal{M}(\mathcal{X})$ is a convex set:

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

Many problems in machine learning can be cast as the latter optimization problem, where $\mathcal{F}(\cdot) = \mathrm{D}(\cdot|\bar{\mu})$ where $\bar{\mu}$ is a fixed target distribution on $\mathbb{R}^d$.

Example 1 and 2

We will consider the following examples:

- Sinkhorn's algorithm
- Expectation-Maximization algorithm

Example 3 - Bayesian inference

Goal of Bayesian inference: learn the best distribution over a parameter x to fit observed data.

Example 3 - Bayesian inference

Goal of Bayesian inference: learn the best distribution over a parameter x to fit observed data.

(1) Let $\mathcal{D} = (w_i, y_i)_{i=1}^{p}$ a dataset of i.i.d. examples with features w, label y.

(2) Assume an underlying model parametrized by $x \in \mathbb{R}^d$, e.g.:

$$y = g(w, x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathrm{Id}).$$

Example 3 - Bayesian inference

Goal of Bayesian inference: learn the best distribution over a parameter x to fit observed data.

(1) Let $\mathcal{D} = (w_i, y_i)_{i=1}^{p}$ a dataset of i.i.d. examples with features w, label y.

(2) Assume an underlying model parametrized by $x \in \mathbb{R}^d$, e.g.:

$$y = g(w, x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathrm{Id}).$$

Step 1. Compute the Likelihood:

$$p(\mathcal{D}|x) \overset{(1)}{\propto} \prod_{i=1}^{p} p(y_i|x, w_i) \overset{(2)}{\propto} \exp\left(-\frac{1}{2}\sum_{i=1}^{p} \|y_i - g(w_i, x)\|^2\right).$$

Step 2. Choose a prior distribution (initial guess) on the parameter:

$$x \sim p_0, \quad \text{e.g. } p_0(x) \propto \exp\left(-\frac{\|x\|^2}{2}\right).$$

Step 2. Choose a prior distribution (initial guess) on the parameter:

$$x \sim p_0, \quad \text{e.g. } p_0(x) \propto \exp\left(-\frac{\|x\|^2}{2}\right).$$

Step 3. Bayes' rule yields the formula for the posterior distribution over the parameter x:

$$p(x|\mathcal{D}) = \frac{p(\mathcal{D}|x)p_0(x)}{Z} \quad \text{where} \quad Z = \int_{\mathbb{R}^d} p(\mathcal{D}|x)p_0(x)dx$$

is called the normalization constant and is **intractable**.

Step 2. Choose a prior distribution (initial guess) on the parameter:

$$x \sim p_0, \quad \text{e.g. } p_0(x) \propto \exp\left(-\frac{\|x\|^2}{2}\right).$$

Step 3. Bayes' rule yields the formula for the posterior distribution over the parameter x:

$$p(x|\mathcal{D}) = \frac{p(\mathcal{D}|x)p_0(x)}{Z} \quad \text{where} \quad Z = \int_{\mathbb{R}^d} p(\mathcal{D}|x)p_0(x)dx$$

is called the normalization constant and is **intractable.**

Denoting $\bar{\mu} := p(\cdot|\mathcal{D})$ the posterior on parameters $x \in \mathbb{R}^d$, we have:

$$\bar{\mu}(x) \propto \exp\left(-V(x)\right), \quad V(x) = \frac{1}{2}\sum_{i=1}^{p}\|y_i - g(w_i, x)\|^2 + \frac{\|x\|^2}{2}.$$

i.e. $\bar{\mu}$'s density is known "up to a normalization constant".

The posterior $\bar{\mu}$ is interesting for

- measuring uncertainty on prediction through the distribution of $g(w, \cdot)$, $x \sim \bar{\mu}$.
- prediction for a new input w:

$$\hat{y} = \underbrace{\int_{\mathbb{R}^d} g(w, x) d\bar{\mu}(x)}_{\text{"Bayesian model averaging"}}$$

i.e. predictions of models parametrized by $x \in \mathbb{R}^d$ are reweighted by $\bar{\mu}(x)$.

Can be cast as:

$$\min_{\nu \in C} \mathrm{KL}(\nu|\bar{\mu})$$

where KL is the "Kullback-Leibler divergence" or relative entropy":

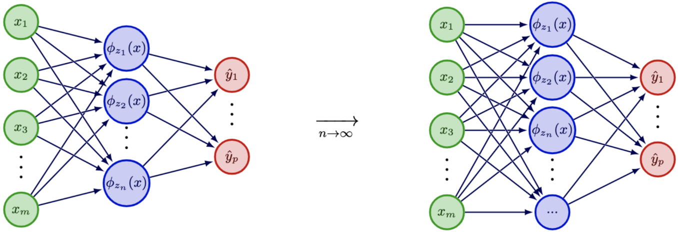
$$\mathrm{KL}(\mu|\bar{\mu}) = \begin{cases} \int_{\mathbb{R}^d} \log\left(\frac{\mu}{\bar{\mu}}(x)\right) d\mu(x) & \text{if } \mu \ll \bar{\mu} \\ +\infty & \text{else.} \end{cases}$$

The KL as an objective is convenient since it **does not depend on the normalization constant** Z (unknown in Bayesian inference)!

Recall that writing $\bar{\mu}(x) = e^{-V(x)}/Z$ we have:

$$\mathrm{KL}(\mu|\bar{\mu}) = \int_{\mathbb{R}^d} \log\left(\frac{\mu}{e^{-V}}(x)\right) d\mu(x) + \log(Z).$$

Example 4 - Optimisation of 1 hidden layer neural networks



$$\min_{z_1,\dots,z_n \in \mathbb{R}^d} \mathbb{E}_{(x,y)\sim P_{data}} \left[\left\| y - \frac{1}{n} \sum_{i=1}^{n} \phi_{z_i}(x) \right\|^2 \right] \xrightarrow[n\to\infty]{} \min_{\nu \in \mathcal{P}(\mathbb{R}^d)} \mathbb{E}_{(x,y)\sim P_{data}} \left[\left\| y - \int_{\mathbb{R}^d} \phi_z(x) d\mu(z) \right\|^2 \right]$$

Assume $\exists \bar{\mu}$, $\mathbb{E}[y|X = x] = \int \phi_z(x) d\bar{\mu}(z)$.

The problem can be cast as:

$$\min_{\nu \in C} \mathrm{MMD}^2(\nu, \bar{\mu})$$

where MMD is the Maximum Mean Discrepancy:

$$\mathrm{MMD}^2(\mu, \pi) = \mathbb{E}_{\substack{z \sim \mu \\ z' \sim \mu}}[k(z, z')] + \mathbb{E}_{\substack{z \sim \pi \\ z' \sim \bar{\mu}}}[k(z, z')] - 2\mathbb{E}_{\substack{z \sim \mu \\ z' \sim \bar{\mu}}}[k(z, z')],$$

with $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a kernel.

Mirror Descent with relative smoothness over the space of measures

To solve

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

we consider the **mirror descent algorithm** [Beck and Teboulle, 2003], a first-order optimization method based on **Bregman divergences**.

Its convergence analysis classically requires **strong convexity and smoothness**.

However, the latter is not satisfied for the KL, hence we consider **relative convexity and smoothness**.

For now assume $C = \mathcal{M}(\mathcal{X})$.

Outline

Space of measures

Let $\mathcal{X} \subset \mathbb{R}^d$, and fix a vector space of (signed) measures $\mathcal{M}(\mathcal{X})$.

It could be $L^1(\mathrm{d}\rho)$, $L^2(\mathrm{d}\rho)$ where ρ is a reference measure, or the space of Radon measures $\mathcal{M}_r(\mathcal{X})$ with the total variation (TV) norm.

Let $\mathcal{M}^*(\mathcal{X})$ the dual of $\mathcal{M}(\mathcal{X})$.

For $\mu \in \mathcal{M}(\mathcal{X})$ and $f \in \mathcal{M}^*(\mathcal{X})$, we denote

$$\langle f, \mu \rangle = \langle f, \mu \rangle_{\mathcal{M}^*(\mathcal{X}) \times \mathcal{M}(\mathcal{X})} = \int_{\mathcal{X}} f(x) \mu(dx).$$

Derivative of $\mathcal{F}$

Mirror Descent is a first-order optimization scheme based on the knowledge of the "derivative" of the objective functional $\mathcal{F}$.

Derivative of $\mathcal{F}$

Mirror Descent is a first-order optimization scheme based on the knowledge of the "derivative" of the objective functional $\mathcal{F}$.

The difficulty is to choose the right notion of derivative.

Recall that Gâteaux and Fréchet derivatives have to be defined in every direction:

Definition 1

The function $\mathcal{F}$ is said to be Gâteaux differentiable at ν if there exists a linear operator $\nabla F(\nu) : \mathcal{M}(\mathcal{X}) \to \mathbb{R}$ such that for any direction $\mu \in \mathcal{M}(\mathcal{X})$:

$$\nabla \mathcal{F}(\nu)(\mu) = \lim_{h \to 0} \frac{\mathcal{F}(\nu + h\mu) - \mathcal{F}(\nu)}{h}. \quad (1)$$

The operator $\nabla \mathcal{F}(\nu)$ is called the Gâteaux derivative of $\mathcal{F}$ at ν, and if it exists, it is unique.

However in infinite dimensions, $\mathrm{Int}(\mathrm{dom}(\mathcal{F}))$ is however often empty (most of all for the negative entropy $\mathcal{F}(\mu) = \int \log(\mu) d\mu$)

However in infinite dimensions, $\mathrm{Int}(\mathrm{dom}(\mathcal{F}))$ is however often empty (most of all for the negative entropy $\mathcal{F}(\mu) = \int \log(\mu) d\mu$)

We thus consider first a weaker notion of directional derivatives.

Then, the notion of first variation will allow to perform all the computations we need, as if the function was Gâteaux differentiable.

Definition 2 (Directional derivative)

If it exists, the *directional derivative* of $\mathcal{F} : \mathcal{M}(\mathcal{X}) \to \mathbb{R} \cup \{\pm\infty\}$ at a point $\nu \in \mathrm{dom}(\mathcal{F})$ in the direction $\mu \in \mathcal{M}(\mathcal{X})$ is defined as

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

Definition 3 (First variation)

If it exists, the *first variation* of $\mathcal{F}$ evaluated at $\mu \in \mathrm{dom}(\mathcal{F})$ is the element $\nabla\mathcal{F}(\mu) \in \mathcal{M}^*(\mathcal{X})$, unique up to orthogonal components to $\mathrm{span}(\mathrm{dom}(\mathcal{F}) - \mu)$, s.t.:

$$\langle \nabla\mathcal{F}(\mu), \xi \rangle = d^+\mathcal{F}(\mu)(\xi) \tag{3}$$

for all $\xi = \nu - \mu \in \mathcal{M}(\mathcal{X})$, where $\nu \in \mathrm{dom}(\mathcal{F})$.

Bregman divergences

Let $\phi : \mathcal{M}(\mathcal{X}) \to \mathbb{R} \cup \{+\infty\}$ be a convex functional. For $\mu \in \text{dom}(\phi)$, the ϕ-*Bregman divergence* is defined for all $\nu \in \text{dom}(\phi)$ by

$$D_\phi(\nu|\mu) = \phi(\nu) - \phi(\mu) - d^+\phi(\mu)(\nu - \mu) \in [0, +\infty], \qquad (4)$$

and $+\infty$ elsewhere. The function ϕ is referred to as *the Bregman potential*.

Properties:

- $D_\phi(\cdot|\mu)$ is convex if ϕ has a first variation (last term is linear)
- D_ϕ separates measures for ϕ strictly convex
- linearity $D_{\phi+\psi} = D_\phi + D_\psi$ (since d^+ is linear)
- idempotence: $D_{D_\phi(\cdot|\xi)}(\nu|\mu) = D_\phi(\nu|\mu)$ for any $\xi \in \text{dom}(\phi)$ assuming $\nabla\phi(\xi)$ exists.

Relative smoothness and convexity

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

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

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

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

- Since $D_{\mathcal{F}}(\nu|\mu) = \mathcal{F}(\nu) - \mathcal{F}(\mu) - d^+\mathcal{F}(\mu)(\nu - \mu)$, convexity of $\mathcal{F}$ writes $D_{\mathcal{F}}(\nu|\mu) \geq 0$.
- Smoothness can be written as

$$\|\nabla \mathcal{F}(\mu) - \nabla \mathcal{F}(\nu)\| \leq L\|\mu - \nu\|$$

which implies

$$\mathcal{F}(\nu) - \mathcal{F}(\mu) - d^+\mathcal{F}(\mu)(\nu - \mu) \leq L\|\nu - \mu\|^2$$

- A Bregman divergence objective $\mathcal{F}(\cdot) = D_\phi(\cdot|\xi)$ is always 1-relatively smooth and strongly convex w.r.t. ϕ (due to the idempotence: $D_{D_\phi(\cdot|\xi)}(\nu|\mu) = D_\phi(\nu|\mu)$)

Case of the KL

The KL is not smooth:

- the "gradient of the KL": $\mu \mapsto \log(\mu|\bar{\mu})(.)$ typically is not Lipschitz
- traditional smoothness cannot hold because KL diverges for Dirac masses, thus does not have subquadratic growth with respect to any norm on measures.

Case of the KL

The KL is not smooth:

- the "gradient of the KL": $\mu \mapsto \log(\mu|\bar{\mu})(.)$ typically is not Lipschitz
- traditional smoothness cannot hold because KL diverges for Dirac masses, thus does not have subquadratic growth with respect to any norm on measures.

Fact: Let $\phi_e(\mu) = \int_{\mathcal{X}} \ln(\mu(x))\mu(x)d\rho(x)$ the **negative entropy**. The KL can be written as a Bregman divergence of ϕ_e, if $\mu \ll \bar{\mu} \ll \rho$, i.e.

$$D_{\phi_e}(\mu|\bar{\mu}) = \mathrm{KL}(\mu|\bar{\mu}).$$

Hence the KL is always 1-relatively smooth with respect to the negative entropy.

Case of the KL

The KL is not smooth:

- the "gradient of the KL": $\mu \mapsto \log(\mu|\bar{\mu})(.)$ typically is not Lipschitz
- traditional smoothness cannot hold because KL diverges for Dirac masses, thus does not have subquadratic growth with respect to any norm on measures.

Fact: Let $\phi_e(\mu) = \int_{\mathcal{X}} \ln(\mu(x))\mu(x)d\rho(x)$ the **negative entropy**. The KL can be written as a Bregman divergence of ϕ_e, if $\mu \ll \bar{\mu} \ll \rho$, i.e.

$$D_{\phi_e}(\mu|\bar{\mu}) = \mathrm{KL}(\mu|\bar{\mu}).$$

Hence the KL is always 1-relatively smooth with respect to the negative entropy.

Remark: It is a strong Bregman divergence. For instance, for a bounded kernel k, $\mathrm{MMD}(\mu, \nu) \leq c_k \,\mathrm{KL}(\mu|\nu)$.

Outline

Relative smoothness :
$\mathcal{F}(\nu) \leq \mathcal{F}(\mu) + d^+\mathcal{F}(\mu)(\nu - \mu) + LD_\phi(\nu|\mu)$.

Mirror descent can be written **in its minimal formulation** as the proximal scheme

$$\mu_{n+1} = \underset{\nu \in C}{\operatorname{argmin}}\{d^+\mathcal{F}(\mu_n)(\nu - \mu_n) + LD_\phi(\nu|\mu_n)\} \tag{5}$$

Remark: If $\mathcal{F}$ and ϕ were Gâteaux differentiable at μ_n, then provided μ_{n+1} exists, the first-order optimality condition for (5) would give

$$\nabla\phi(\mu_{n+1}) - \nabla\phi(\mu_n) = -\frac{1}{L}\nabla\mathcal{F}(\mu_n). \tag{6}$$

Remark: If $\phi = \phi_e$, $\nabla\phi_e(\mu) = \log(\mu) + 1$ which leads to the famous multiplicative update $\mu_{n+1} = \mu_n e^{-\frac{1}{L}\nabla\mathcal{F}(\mu_n)}$.

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 (5), and assume that for each $n \geq 0$, $\nabla\phi(\mu_n)$ exists. Then for all $n \geq 0$ and all $\nu \in \mathrm{dom}(\mathcal{F}) \cap \mathrm{dom}(\phi)$:

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

Remark: mirror descent rates with strong (standard) convexity and smoothness lead to $\mathcal{O}(1/\sqrt{n})$ rate with a decreasing step-size $\propto 1/\sqrt{n}$.

Outline

Preliminaries

Notations:

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

$$
\begin{aligned}
\mathrm{KL}(\pi|\bar{\pi}) &= \int \log\left(\frac{\pi}{\bar{\pi}}\right) d(p_{\mathcal{X}}\pi \otimes K_\pi) \\
&= \mathrm{KL}(p_{\mathcal{X}}\pi | p_{\mathcal{X}}\bar{\pi}) + \int_{\mathcal{X}} \mathrm{KL}(K_\pi | K_{\bar{\pi}})\, dp_{\mathcal{X}}\pi \\
&= \mathrm{KL}(p_{\mathcal{X}}\pi | p_{\mathcal{X}}\bar{\pi}) + \mathrm{KL}(\pi | p_{\mathcal{X}}\pi \otimes K_{\bar{\pi}}). \quad (7)
\end{aligned}
$$

It will be crucial for assessing the (relative) smoothness and convexity two objective functions F_{S} and F_{EM} we will consider.

Consider a cost function $c \in L^\infty(\mathcal{X} \times \mathcal{Y}, \bar{\mu} \otimes \bar{\nu})$ and a regularization parameter $\epsilon > 0$.

The **entropic optimal transport problem** is the minimization problem

$$\mathrm{OT}_\epsilon(\bar{\mu}, \bar{\nu}) = \min_{\pi \in \Pi(\bar{\mu}, \bar{\nu})} \mathrm{KL}(\pi | e^{-c/\epsilon} \bar{\mu} \otimes \bar{\nu}). \tag{8}$$

Consider a cost function $c \in L^\infty(\mathcal{X} \times \mathcal{Y}, \bar{\mu} \otimes \bar{\nu})$ and a regularization parameter $\epsilon > 0$.

The **entropic optimal transport problem** is the minimization problem

$$\mathrm{OT}_\epsilon(\bar{\mu}, \bar{\nu}) = \min_{\pi \in \Pi(\bar{\mu}, \bar{\nu})} \mathrm{KL}(\pi | e^{-c/\epsilon} \bar{\mu} \otimes \bar{\nu}). \tag{8}$$

We say that a coupling π is cyclically invariant, and write $\pi \in \Pi_c$, if denoting by $(\mu, \nu) = (p_\mathcal{X} \pi, p_\mathcal{Y} \pi)$ its marginals we have

$$\mathrm{KL}(\pi | e^{-c/\epsilon} \mu \otimes \nu) = \min_{\tilde{\pi} \in \Pi(\mu, \nu)} \mathrm{KL}(\tilde{\pi} | e^{-c/\epsilon} \mu \otimes \nu). \tag{9}$$

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

The Sinkhorn algorithm in its primal formulation searches for the solution of (8) by alternative (entropic) projections on $\Pi(\bar{\mu}, *)$ and $\Pi(*, \bar{\nu})$, i.e. initializing with $\pi_0 \in \Pi_c$, iterate

$$\pi_{n+\frac{1}{2}} = \operatorname*{argmin}_{\pi \in \Pi(\bar{\mu}, *)} \mathrm{KL}(\pi | \pi_n), \tag{10}$$

$$\pi_{n+1} = \operatorname*{argmin}_{\pi \in \Pi(*, \bar{\nu})} \mathrm{KL}(\pi | \pi_{n+\frac{1}{2}}). \tag{11}$$

The Sinkhorn algorithm in its primal formulation searches for the solution of (8) by alternative (entropic) projections on $\Pi(\bar{\mu}, *)$ and $\Pi(*, \bar{\nu})$, i.e. initializing with $\pi_0 \in \Pi_c$, iterate

$$\pi_{n+\frac{1}{2}} = \underset{\pi \in \Pi(\bar{\mu}, *)}{\operatorname{argmin}} \operatorname{KL}(\pi | \pi_n), \tag{10}$$

$$\pi_{n+1} = \underset{\pi \in \Pi(*, \bar{\nu})}{\operatorname{argmin}} \operatorname{KL}(\pi | \pi_{n+\frac{1}{2}}). \tag{11}$$

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

$$F_S(\pi) = \operatorname{KL}(p_{\mathcal{X}} \pi | \bar{\mu}). \tag{12}$$

Sinkhorn algorithm as mirror descent

Proposition: The Sinkhorn iterations (10) can be written as a mirror descent with objective F_{S} and Bregman divergence KL over the constraint $C = \Pi(*, \bar{\nu})$,

$$\pi_{n+1} = \operatorname*{argmin}_{\pi \in C} \langle \nabla F_{\mathrm{S}}(\pi_n), \pi - \pi_n \rangle + \mathrm{KL}(\pi|\pi_n)$$
$$\text{with } \nabla F_{\mathrm{S}}(\pi_n) = \ln(d\mu_n / d\bar{\mu}) \in L^\infty(\mathcal{X} \times \mathcal{Y}). \quad (13)$$

where $\mu_n = p_{\mathcal{X}} \pi_n$.

Proof: We have the identity:

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

We conclude by taking the argmin over $\pi \in C$.

(Relative) smoothness and convexity of F_{S}

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

Proof: Let $\pi, \tilde{\pi} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ with $p_{\mathcal{X}}\tilde{\pi} \ll p_{\mathcal{X}}\pi \ll \bar{\mu}$. Then:

- with straightforward computations, $D_{F_{\mathrm{S}}}(\tilde{\pi}|\pi) = \mathrm{KL}(p_{\mathcal{X}}\tilde{\pi}|p_{\mathcal{X}}\pi) \geq 0$, so F_{S} is convex
- applying the disintegration formula, we obtain that $D_{F_{\mathrm{S}}}(\tilde{\pi}|\pi) \leq \mathrm{KL}(\tilde{\pi}|\pi)$. **(KL of joint distributions is smaller than KL of marginals)**

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

(Relative) strong convexity of F_{S}

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

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

in other words 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.

We recover (known) rates for Sinkhorn

Proposition: For all $n \geq 0$, the Sinkhorn iterates verify, for π_* the optimum of:

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

and μ_* its first marginal,

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

Outline

EM

Goal: fit a parametric distribution to some observed data Y (e.g. a mixture of Gaussians approximating the data), where one needs to estimate both

- the latent variable distribution on X (e.g. weights of each Gaussian)
- parameters of conditionals $P(Y|X=x)$ (e.g. means and covariances of each Gaussian)

Consider the following probabilistic model: we have a latent, hidden random variable $X \in (\mathcal{X}, \bar{\mu})$, an observed variable $Y \in \mathcal{Y}$ distributed as $\bar{\nu}$.

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

$$\min_{q \in \mathcal{Q}} \mathrm{KL}(\bar{\nu} | p_{\mathcal{Y}} p_q), \tag{14}$$

where $p_{\mathcal{Y}} p_q(dy) = \int_{\mathcal{X}} p_q(dx, dy)$.

For any $\pi \in \Pi(*, \bar{\nu})$, by the disintegration formula:

- $\mathrm{KL}(\bar{\nu}|p_{\mathcal{Y}}p_q) \leq \mathrm{KL}(\pi|p_q)$
- with equality if $\pi(dx, dy) = p_q(dx, dy)\bar{\nu}(dy)/p_{\mathcal{Y}}p_q(dy)$

For any $\pi \in \Pi(*, \bar{\nu})$, by the disintegration formula:

- $\mathrm{KL}(\bar{\nu}|p_{\mathcal{Y}}p_q) \leq \mathrm{KL}(\pi|p_q)$
- with equality if $\pi(dx, dy) = p_q(dx, dy)\bar{\nu}(dy)/p_{\mathcal{Y}}p_q(dy)$

EM then proceeds by alternate minimizations of $\mathrm{KL}(\pi, p_q)$ [Neal and Hinton, 1998]:

$$q_n = \underset{q \in \mathcal{Q}}{\mathrm{argmin}}\, \mathrm{KL}(\pi_n|p_q), \tag{15}$$

$$\pi_{n+1} = \underset{\pi \in \Pi(*, \bar{\nu})}{\mathrm{argmin}}\, \mathrm{KL}(\pi|p_{q_n}). \tag{16}$$

The above formulation consists in (15), optimizing the parameters q_n at step n (M-step), and then (16), optimizing the joint distribution π_{n+1} at step $n+1$ (E-step, which is explicit).

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

$$F_{\mathrm{EM}}(\pi) = \inf_{q \in \mathcal{Q}} \mathrm{KL}(\pi | p_q). \tag{17}$$

Proposition: EM can be written as a mirror descent iteration:

$$\pi_{n+1} = \operatorname*{argmin}_{\pi \in C} \langle \nabla F_{\mathrm{EM}}(\pi_n), \pi - \pi_n \rangle + \mathrm{KL}(\pi | \pi_n)$$
$$\text{with } \nabla F_{\mathrm{EM}}(\pi_n) = \ln(d\pi_n / dp_{q_n}). \tag{18}$$

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

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

Thus the MD iteration matches (16).

Latent EM

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 makes F_{EM} convex.

Define $F_{\text{LEM}}(\pi) := \inf_{\mu \in \mathcal{P}(\mathcal{X})} \text{KL}(\pi|\mu \otimes K)$
($F_{\text{LEM}}(\pi) = \text{KL}(\pi|p_{\mathcal{X}}\pi \otimes K)$ by the disintegration formula).

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

$$\pi_{n+1} = \operatorname*{argmin}_{\pi \in C} \langle \nabla F_{\text{LEM}}(\pi_n), \pi - \pi_n \rangle + \text{KL}(\pi|\pi_n)$$
$$\text{with } \nabla F_{\text{LEM}}(\pi_n) = \ln\left(\frac{d\pi_n}{d(\mu_n \otimes K)}\right) \in L^\infty. \quad (19)$$

Rate for Latent EM

Proposition Set $\mu_* \in \mathrm{argmin}_{\mu\in\mathcal{P}(\mathcal{X})} \mathrm{KL}(\bar{\nu}|T_K(\mu))$ 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.

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

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

Conclusion

- rigorous proof of convergence of mirror descent under relative smoothness and convexity, which holds in the infinite-dimensional setting of optimization over measure spaces
- provides a new and simple way to derive rates of convergence for Sinkhorn's algorithm
- new convergence rates for EM when restricted to the latent distribution, obtaining similar but complementary rates to [Kunstner et al., 2021].

Questions?

References I

Beck, A. and Teboulle, M. (2003).
Mirror descent and nonlinear projected subgradient methods for convex optimization.
Operations Research Letters, 31(3):167–175.

Kunstner, F., Kumar, R., and Schmidt, M. W. (2021).
Homeomorphic-invariance of EM: Non-asymptotic convergence in KL divergence for exponential families via mirror descent.
In *AISTATS*.

Neal, R. M. and Hinton, G. E. (1998).
A view of the EM algorithm that justifies incremental, sparse, and other variants.
In *Learning in Graphical Models*, pages 355–368. Springer Netherlands.