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

Anonymous Author(s)
Affiliation
Address
email

Abstract

Many problems in machine learning can be formulated as optimizing a convex functional over a space of measures. This paper studies the convergence of the mirror descent algorithm in this infinite-dimensional setting. Defining Bregman divergences through directional derivatives, we derive the convergence of the scheme for relatively smooth and strongly convex pairs of functionals. Applying our result to joint distributions and the Kullback–Leibler (KL) divergence, we show that Sinkhorn's primal iterations for entropic optimal transport in the continuous setting correspond to a mirror descent, and we obtain a new proof of its (sub)linear convergence. We also show that Expectation Maximization (EM) can always formally be written as a mirror descent, and, when optimizing on the latent distribution while fixing the mixtures, we derive sublinear rates of convergence.

1 Introduction

Many important problems in machine learning and computational statistics can be cast as an optimization problem over the space of probability distributions, where the objective functional measures the dissimilarity to a target distribution $\bar{\mu}$ on $\mathbb{R}^d$. Classical dissimilarities include f-divergences, Integral Probability Metrics (IPMs), or optimal transport distances among others. In Bayesian inference for instance, it is common to optimize the Kullback–Leibler (KL) divergence to the target, which corresponds to the posterior distribution of the parameters of interest. In generative modelling, the goal is to generate data whose distribution is similar to the training set distribution defined by samples of the target, where the similarity is often measured by an integral probability metric or an optimal transport distance (Arjovsky et al., 2017; Dziugaite et al., 2015). In supervised learning, optimizing an infinite-width one hidden layer neural network, e.g. through the mean squared error, corresponds to minimizing a functional over the space of probability distributions over the parameters of the network (Chizat and Bach, 2018; Mei et al., 2018; Rotskoff and Vanden-Eijnden, 2018). In particular, the objective functional can be identified to a Maximum Mean Discrepancy (MMD) in the well-specified setting (Arbel et al., 2019). Many other problems in machine learning can be formalized in this framework (Chu et al., 2019).

Once the objective functional is chosen, one has to select an optimization algorithm that is well-suited to the geometry of the problem. In this article, we consider the widely used mirror descent scheme, a first-order optimization method based on Bregman divergences. While the traditional smoothness and strong convexity assumptions, required by standard convergence analysis, do not always hold over measure spaces, their "relative" versions have received increased interest. The relative smoothness assumption was first suggested by Birnbaum et al. (2011) in the context of algorithmic game theory, but remained unnoticed by the optimization community, until Bauschke et al. (2017) discovered the same concept independently, while Lu et al. (2018) coupled it with relative strong convexity. We extend their work to the infinite dimensional setting. When using the KL divergence as Bregman

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

divergence, mirror descent yields multiplicative updates, such as Sinkhorn's algorithm, widely used to solve entropic optimal transport (Peyré and Cuturi, 2019). A prominent alternative scheme to mirror descent for optimizing over measures is based on (grid-free) Wasserstein gradient flows, which we do not cover here. An elegant comparison of these settings can be found in Trillos and Sanz-Alonso (2020) for Bayesian inference. On the other hand the EM algorithm is a very common approach to fitting probabilistic models.

Related work. Chizat (2021) gave sharp convergence rates of mirror descent on measure spaces for smooth integral functionals, working mostly in the L^1 space and without leveraging relative smoothness. However his assumptions do not cover f-divergences, such as the ubiquitous KL, nor entropic regularized transport as in Léger (2020). His setting also implies Gâteaux differentiability, which is classical for mirror descent in Banach spaces (Bauschke et al., 2003). As the entropy is not differentiable in infinite dimensions, directional derivatives were instead used by Resmerita (2005) for mirror descent, and applied by Chu et al. (2019) to gradient descent. Connections between mirror descent and Sinkhorn iterations for the entropic optimal transport problem were first investigated in Mishchenko (2019); Mensch and Peyré (2020); Léger (2020). Our framework is closest to Léger (2020), which we simplify by considering primal rather than dual iterations, and extend by also deriving a linear convergence rate. Kunstner et al. (2021) recently showed that EM over parametric exponential mixtures corresponds to a mirror descent scheme with relative smoothness properties, in a setting complementary to ours since we consider instead fixed mixtures and a nonparametric latent distribution.

Contributions. We propose a rigourous framework for the analysis of the infinite-dimensional version of mirror descent over measure spaces. In this setting, we recover the rate of convergence of mirror descent under relative smoothness and convexity, previously shown in finite dimensions. Focusing on optimization over joint distributions, we show that both Sinkhorn's primal iterations for entropic optimal transport in the continuous setting and EM can be written as a mirror descent. We then obtain a new proof of Sinkhorn's (sub)linear convergence. For EM, when optimizing on the latent distribution while fixing the mixtures, a choice which coincides with Richardson–Lucy deconvolution, we derive new sublinear rates of convergence.

This paper is organized as follows. Section 2 introduces the necessary background on derivatives in measure spaces and relative smoothness and convexity. Section 3 discusses the well-posedness of the mirror descent scheme and provides our proof of convergence adapted from Lu et al. (2018). Section 4 recovers the convergence of algorithms such as Sinkhorn's iterations and latent EM as special cases of mirror descent with relative smoothness and convexity.

2 Background and definitions

In this section, we set the mathematical framework in which we will rigourously reformulate relative smoothness and convexity on a space of measures.

Notation. Given a topological vector space $\mathcal{Y}$ with topology τ, the domain $\mathrm{dom}(f)$ of an extended-valued function $f : \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$ is the set of points of $\mathcal{Y}$ where f takes finite values. The function f is said to be proper if $\mathrm{dom}(f)$ is non-empty and if f never takes the value $-\infty$. It is τ-lower semicontinuous (l.s.c.) if its sublevel sets are τ-closed. We consider a dual pair $(\mathcal{Y}, \mathcal{Y}^*)$ with duality product $\langle \cdot, \cdot \rangle_{\mathcal{Y}^* \times \mathcal{Y}}$ which induces a $\mathcal{Y}^*$-weak topology on $\mathcal{Y}$. We write $\mathrm{Int}\, C$ for the interior of a set $C \subset \mathcal{Y}$. We refer to Aliprantis and Border (2006) for more on these classical concepts. Let $\mathcal{X} \subset \mathbb{R}^d$, and fix a locally convex topological vector space of measures $(\mathcal{M}(\mathcal{X}), \tau)$, which could be for instance $L^1(\mathrm{d}\rho)$, $L^2(\mathrm{d}\rho)$ where ρ is a reference measure, or the space of Radon measures $\mathcal{M}_r(\mathcal{X})$ with the total variation (TV) norm. Fix $\mathcal{M}^*(\mathcal{X})$ a topological dual of $\mathcal{M}(\mathcal{X})$. For $\mu \in \mathcal{M}(\mathcal{X})$ and $f \in \mathcal{M}^*(\mathcal{X})$, we use the shorthand $\langle f, \mu \rangle = \langle f, \mu \rangle_{\mathcal{M}^*(\mathcal{X}) \times \mathcal{M}(\mathcal{X})}$, formally equal to $\int_{\mathcal{X}} f(x)\mu(dx)$. We denote by $\mathcal{P}(\mathcal{X})$ the subset of measures $\mu \in \mathcal{M}(X)$ with mass 1, and write $\mu \ll \nu$ when μ is absolutely continuous with respect to ν, $d\mu/d\nu$ their Radon–Nikodym density for any $\mu, \nu \in \mathcal{M}(\mathcal{X})$.

Consider a convex functional $\mathcal{F} : \mathcal{M}(\mathcal{X}) \to \mathbb{R} \cup \{+\infty\}$ and the following minimization problem

$$\min_{\nu \in C} \mathcal{F}(\nu) \tag{1}$$

where $C \subset \mathcal{M}(\mathcal{X})$ is a convex set. To solve this convex optimization problem, a classical choice is to resort to a mirror descent scheme (see e.g. Beck and Teboulle, 2003). The latter is a first-order

optimization scheme based on the knowledge of the "derivative" of the objective functional $\mathcal{F}$ at each iteration. The difficulty is to choose the appropriate notion of derivative. Indeed, Gâteaux and Fréchet derivatives have to be defined in every direction, see Appendix A, thus requiring that the points of differentiability belong to the interior of the domain $\mathrm{Int}(\mathrm{dom}(\mathcal{F}))$ of the functional $\mathcal{F}$ considered (as it is the case for finite dimensional $\mathcal{M}(\mathcal{X})$, when $\mathcal{X}$ is finite). In infinite dimensions, $\mathrm{Int}(\mathrm{dom}(\mathcal{F}))$ is however often empty, most of all for the negative entropy[1]. Consequently, following Resmerita (2005), we favor a weaker notion, that of directional derivatives.[2] This comes at the price of manipulating $\pm\infty$ values but ensures that the considered derivatives are always well-defined for convex functionals. Besides, the notion of first variation will allow to perform all the computations we need, as if the function was Gâteaux differentiable.

Definition 1 (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

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

Remark 1. In particular, for convex and proper functions, $d^+\mathcal{F}(\nu)(\mu)$ exists and belongs to $\mathbb{R} \cup \{\pm\infty\}$ (Aliprantis and Border, 2006, Lemma 7.14). This is a direct consequence of the nondecreasingness of $h \in \mathbb{R}_+^* \mapsto \frac{\mathcal{F}(\nu+h\mu)-\mathcal{F}(\nu)}{h}$ for any ν, μ and convex $\mathcal{F}$. The monotonicity also entails that $d^+\mathcal{F}(\nu)(\mu) \leq \mathcal{F}(\nu + \mu) - \mathcal{F}(\nu) < \infty$ whenever ν and $\nu + \mu$ belong to $\mathrm{dom}(\mathcal{F})$ and that d^+ is a linear operation over the cone of convex functions. Note that $\mu \mapsto d^+\mathcal{F}(\nu)(\mu)$ is not always τ-l.s.c. although it is always positively homogeneous, and, whenever $\mathcal{F}$ is convex, it is convex.

Definition 2 (First variation)**.** Let $\mathcal{F} : \mathcal{M}(\mathcal{X}) \to \mathbb{R} \cup \{+\infty\}$ be a functional and C be a subset of $\mathcal{M}(\mathcal{X})$. If it exists, the *first variation* of $\mathcal{F}$ over C evaluated at $\mu \in \mathrm{dom}(\mathcal{F}) \cap C$ is the element $\nabla_C \mathcal{F}(\mu) \in \mathcal{M}^*(\mathcal{X})$, unique up to orthogonal components to $\mathrm{span}(\mathrm{dom}(\mathcal{F}) \cap C - \mu)$, s.t.:

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

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

By Remark 1, we have that $d^+\mathcal{F}(\mu)(\xi) \in [-\infty, \infty)$, since $\nu, \mu \in \mathrm{dom}(\mathcal{F}) \cap C$. Naturally, if $\mathcal{F}$ has a Fréchet or Gâteaux derivative at μ, which implies that $\mu \in \mathrm{Int}(\mathrm{dom}(\mathcal{F}))$, then it coincides with the first variation at μ. Calling first variations the derivatives of functionals stems from the field of calculus of variations; a pragmatic approach when dealing with probability measures can be found in Santambrogio (2015, Definition 7.12), where $\nabla_{\mathcal{P}(\mathcal{X})}\mathcal{F}(\mu)$ is defined as a measurable function.

Definition 3. (Bregman divergence) Let $\phi : \mathcal{M}(\mathcal{X}) \to \mathbb{R} \cup \{+\infty\}$ be a convex functional. For $\mu \in \mathrm{dom}(\phi)$, the *$\phi$-Bregman divergence* is defined for all $\nu \in \mathrm{dom}(\phi)$ by

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

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

In the previous definition, the restriction to $\mathrm{dom}(\phi)$ is necessary to avoid the substraction of infinite values. Moreover, as a direct consequence of Lemma 11 in Appendix, a stricly convex ϕ entails that the Bregman divergence D_ϕ separates measures, i.e. $D_\phi(\nu|\mu) = 0$ if and only if $\nu = \mu$. Note that $D_\phi(\cdot|\mu)$ is a difference of convex functions, so it is not convex in general. Nevertheless the existence of a first variation (3) of ϕ over C, resulting in a linear term in (4), is sufficient to ensure the convexity of the restriction of $D_\phi(\cdot|\mu)$ to C. Bregman divergences have useful immediate properties: since d^+ is a linear operation over convex functions, so is the Bregman divergence, i.e. $D_{\phi+\psi} = D_\phi + D_\psi$ for two Bregman potentials ϕ, ψ. Moreover, it is idempotent, as shown in the following lemma.

Lemma 1. Let $\phi : \mathcal{M}(\mathcal{X}) \to \mathbb{R} \cup \{+\infty\}$ be a convex functional. Assume that given $\xi \in \mathrm{dom}(\phi)$, the first variation $\nabla_{\mathrm{dom}(\phi)}\phi(\xi)$ exists, then, for all $\mu, \nu \in \mathrm{dom}(\phi)$, $D_{D_\phi(\cdot|\xi)}(\nu|\mu) = D_\phi(\nu|\mu)$.

Proof. It is straightforward to show that the Bregman divergence of a constant or of a linear form is zero. Since $\psi : \tilde{\mu} \mapsto -d^+\phi(\xi)(\tilde{\mu} - \xi) = -\langle \nabla_{\mathrm{dom}(\phi)}\phi(\xi), \tilde{\mu} - \xi \rangle$ is convex over $\mathrm{dom}(\phi)$, we can apply the linearity of the Bregman divergence:

$$D_{D_\phi(\cdot|\xi)}(\nu|\mu) = D_\phi(\nu|\mu) + D_{-\phi(\xi)}(\nu|\mu) + D_\psi(\nu|\mu) = D_\phi(\nu|\mu). \qquad \square$$

[1]The non-differentiability of the entropy is recalled for instance in Butnariu and Resmerita (2006, p12) and Santambrogio (Remark 7.13 2015, p239).

[2]In finite dimensions, Maddison et al. (2021) also defined Bregman divergences through directional derivatives, but under the stringent assumption of essentially smooth convex functions, an assumption which does not extend well to infinite dimensions.

We are now ready to introduce the notions of relative smoothness (Bauschke et al., 2017) and convexity (Lu et al., 2018) of a functional with respect to a Bregman potential.

Definition 4. Let $\mathcal{F} : \mathcal{M}(\mathcal{X}) \to \mathbb{R} \cup \{+\infty\}$ be a convex proper functional. Given a scalar $L \geq 0$, we say that $\mathcal{F}$ is L-smooth relative to ϕ over C if, for any $\mu, \nu \in \mathrm{dom}(\mathcal{F}) \cap \mathrm{dom}(\phi) \cap C$, we have

$$D_{\mathcal{F}}(\nu|\mu) = \mathcal{F}(\nu) - \mathcal{F}(\mu) - d^{+}\mathcal{F}(\mu)(\nu - \mu) \leq L D_{\phi}(\nu|\mu). \tag{5}$$

Conversely, we say that $\mathcal{F}$ is l-strongly convex relative to ϕ over C, for some scalar $l \geq 0$, if, for any $\mu, \nu \in \mathrm{dom}(\mathcal{F}) \cap \mathrm{dom}(\phi) \cap C$, we have

$$D_{\mathcal{F}}(\nu|\mu) \geq l D_{\phi}(\nu|\mu). \tag{6}$$

By Lemma 1, provided $\nabla_{\mathrm{dom}(\phi)}\phi(\xi)$ is well-defined, a Bregman divergence objective $D_{\phi}(\cdot|\xi)$ is always 1-relatively smooth and strongly convex w.r.t. ϕ. Relative smoothness and convexity can be characterized in different, equivalent ways, see Lemma 12 in the Appendix. We now turn to the analysis of the mirror descent scheme using the above framework. In the following, ϕ is assumed to be strictly convex. Proving that the functionals of interest have a first variation at some μ has to be done on a case-by-case basis, but all the ones we are interested in will share this property.

3 Mirror descent over measures and convergence

The relative smoothness assumption (5) of the convex objective functional $\mathcal{F}$ with respect to a Bregman potential ϕ suggests to minimize iteratively over $\nu \in C$ the function $\nu \mapsto \mathcal{F}(\mu) + d^{+}\mathcal{F}(\mu)(\nu - \mu) + L D_{\phi}(\nu|\mu)$, acting as an upper approximation of $\mathcal{F}(\nu)$. Starting from a given $\mu_0 \in \mathcal{M}(\mathcal{X})$, the mirror descent iterates are thus defined at each time $n \geq 0$ as

$$\mu_{n+1} = \underset{\nu \in C}{\mathrm{argmin}}\{d^{+}\mathcal{F}(\mu_n)(\nu - \mu_n) + L D_{\phi}(\nu|\mu_n)\}, \tag{7}$$

for which we want to prove that for all $n \geq 0$ and ν in some set $\mathcal{R} \subset C$ containing a minimum of (1), the scheme (7) converges. As proven later in this section, sufficient conditions for the latter are:

($\mathbf{A}_1$) (Existence) The sequence of iterates $(\mu_n)_{n \in \mathbb{N}}$ defined by (7) exist, belong to $\mathcal{R}$, and are unique.

($\mathbf{A}_2$) (Relative smoothness) The functional $\mathcal{F}$ is L-smooth relative to ϕ as in (5) for elements of $\mathcal{R}$.

Assumption ($\mathbf{A}_2$) has to be verified on a case-by-case basis, in particular the set $\mathcal{R}$ does not have to be convex for relative smoothness or convexity as we will see in Section 4 when dealing with transport plans. On the other hand Assumption (**A'**$_1$), below, ensures that the iterates in (7) are well-defined, uniqueness resulting from the strict of convexity of ϕ.

(**A'**$_1$) (Lower semicontinuity and coercivity) (i) the set C is τ-closed in $\mathcal{M}(\mathcal{X})$, the functionals $\mathcal{G}_n(\cdot) := d^{+}\mathcal{F}(\mu_n)(\cdot - \mu_n)$ and $D_{\phi}(\cdot|\mu_n)$ are proper and τ-l.s.c. when restricted to C, and the functional $\mathcal{G}_n + D_{\phi}(\cdot|\mu_n) + i_C$[3] has at least one τ-compact sublevel set. (ii) For each $n \geq 0$, the first variations $\nabla_C \phi(\mu_n)$ exist.

By Attouch et al. (2014, Theorem 3.2.2), Assumption (**A'**$_1$) implies Assumption ($\mathbf{A}_1$). Indeed, (i) τ-lower semicontinuity and τ-compactness guarantee the existence of minimizers of the objective (7), while (ii) the existence of a first variation of ϕ at each iterate guarantees that $D_{\phi}(\cdot|\mu_n)$ is strictly convex (since ϕ is assumed strictly convex) hence unicity of the minimizer. Regarding Assumption (**A'**$_1$)(i), notice that $\mathcal{G}_n$ is proper and $\mathcal{M}^*(\mathcal{X})$-weak-l.s.c. as soon as $\mathcal{F}$ has a first variation at μ_n, since in this case $\mathcal{G}_n$ is linear on C. See Appendix B for more details on proving that Assumption (**A'**$_1$) holds for some $\mathcal{F}$, ϕ and C.

We now state a preliminary result, (8), known as the "three-point property" or "Bregman proximal inequality" (Chen and Teboulle, 1993, Lemma 3.2), (Lan et al., 2011, Lemma 1), which we use to prove the convergence of the mirror descent scheme, similarly to Lu et al. (2018).

Lemma 2. Given $\mu \in \mathcal{M}(\mathcal{X})$ and some proper convex functional $\mathcal{G} : \mathcal{M}(\mathcal{X}) \to \mathbb{R} \cup \{+\infty\}$, if $\nabla_C \phi(\mu)$ exists, as well as $\bar{\nu} = \mathrm{argmin}_{\nu \in C}\{\mathcal{G}(\nu) + D_{\phi}(\nu|\mu)\}$, then for all $\nu \in C \cap \mathrm{dom}(\phi) \cap \mathrm{dom}(\mathcal{G})$:

$$\mathcal{G}(\nu) + D_{\phi}(\nu|\mu) \geq \mathcal{G}(\bar{\nu}) + D_{\phi}(\bar{\nu}|\mu) + D_{\phi}(\nu|\bar{\nu}). \tag{8}$$

[3] i_C denotes the indicator function of the set C, defined by $i_C(\mu) = 0$ if $\mu \in C$, $+\infty$ otherwise for any $\mu \in \mathcal{M}(\mathcal{X})$. Notice that i_C being τ-l.s.c. is equivalent to C being τ-closed in $\mathcal{M}(\mathcal{X})$.

Proof. The existence of $\nabla_C \phi(\mu)$ entails that $C \cap \mathrm{dom}(D_\phi(\cdot|\mu)) = C \cap \mathrm{dom}(\phi)$. Let $f(\cdot) = \mathcal{G}(\cdot) + D_\phi(\cdot|\mu)$ be defined over C, then by linearity of the Bregman divergence and Lemma 1, we obtain that, for $\nu \in C$,

$$D_f(\nu|\bar{\nu}) = D_{\mathcal{G}}(\nu|\bar{\nu}) + D_{D_\phi(\cdot|\mu)}(\nu|\bar{\nu}) = D_{\mathcal{G}}(\nu|\bar{\nu}) + D_\phi(\nu|\bar{\nu}) \geq D_\phi(\nu|\bar{\nu}). \tag{9}$$

Since $\bar{\nu}$ is optimal, we have that, for all $\nu \in C$, $d^+(f)(\bar{\nu})(\nu - \bar{\nu}) = \lim_{h \to 0^+}(f((1-h)\bar{\nu} + h\nu) - f(\bar{\nu}))/h \geq 0$, hence $f(\nu) \geq f(\bar{\nu}) + D_f(\nu|\bar{\nu})$. We conclude using (9) and the definition of f. □

The following theorem, whose proof can be found in Appendix D.1, extends to infinite dimensions the result of Lu et al. (2018, Theorem 3.1) which was stated for differentiable functions.

Theorem 3. Assume that $\mathcal{F}$ is l-strongly convex and L-smooth relative to ϕ over $\mathcal{R}$, with $l, L \geq 0$, and that Assumptions ($\mathbf{A}_1$) and ($\mathbf{A'}_1$)-(ii) hold. Consider the mirror descent scheme (7), then for all $n \geq 0$ and all $\nu \in \mathrm{dom}(\mathcal{F}) \cap \mathrm{dom}(\phi) \cap C \cap \mathcal{R}$, we have

$$\mathcal{F}(\mu_n) - \mathcal{F}(\nu) \leq \frac{l D_\phi(\nu|\mu_0)}{\left(1 + \frac{l}{L-l}\right)^n - 1} \leq \frac{L}{n} D_\phi(\nu|\mu_0), \tag{10}$$

where, in the case $l = 0$, the middle expression is defined in the limit as $l \to 0^+$.

Remark 2. While our proof resembles the one of Lu et al. (2018), the latter relies on the three point inequality (Lemma 2), which proof in finite dimensions uses a formula of the form $\partial(\mathcal{G} + D_\phi) = \partial\mathcal{G} + \partial D_\phi$ as in Chen and Teboulle (1993, Lemma 3.2). Such an equality between subdifferentials can be obtained typically under at least three (non-equivalent) conditions for convex and l.s.c. $\mathcal{G}$ and ϕ over a Banach space $\mathcal{M}(\mathcal{X})$: (i) having $\cup_{\lambda \geq 0} \lambda(\mathrm{dom}(\mathcal{G}) - \mathrm{dom}(D_\phi))$ to be a closed vector space of $\mathcal{M}(\mathcal{X})$ (Attouch and Brezis, 1986) (ii) having a non-empty relative interior of $(\mathrm{dom}(\mathcal{G}) - \mathrm{dom}(D_\phi)$ (iii) continuity of D_ϕ or $\mathcal{G}$ at least at some $\mu \in \mathrm{dom}(\mathcal{G}) \cap \mathrm{dom}(D_\phi)$ Peypouquet (2015, Theorem 3.30). Condition (iii) does not hold for the entropy ϕ_e (defined below in (11)) in none of the spaces we consider. It is known that the other conditions are difficult to verify for given functionals. For instance the domain $\mathrm{dom}(\phi)$ is not explicit for the entropy and $\mathcal{M}(\mathcal{X}) \supset L^1(\mathcal{X})$ as per Example 1 below. On the contrary, by favoring directional derivatives and first variations, we circumvent most of the difficulties related to (sub)differentiability.

Example 1 (KL as objective or Bregman divergence)**.** The Kullback–Leibler (KL) divergence and the negative entropy are defined for $\mu \ll \bar{\mu}$ and $\mu \ll \rho$, writing $\mu(x) = {}^{d\mu}/_{d\rho}(x)$, respectively as

$$\mathrm{KL}(\mu|\bar{\mu}) = \int_{\mathcal{X}} \ln\left({}^{d\mu}/_{d\bar{\mu}}(x)\right) d\mu(x), \quad \phi_e(\mu) = \int_{\mathcal{X}} \ln(\mu(x))\mu(x) d\rho(x) = \mathrm{KL}(\mu|\rho), \tag{11}$$

where ρ is some finite reference measure on $\mathcal{X}$. Indeed, it is straightforward to show that KL can be written as a Bregman divergence of ϕ_e if $\mu \ll \bar{\mu} \ll \rho$, i.e. $D_{\phi_e}(\mu|\bar{\mu}) = \mathrm{KL}(\mu|\bar{\mu})$. For compact $\mathcal{X}$, the domain of the negative entropy ϕ_e is strictly included in $L^1_+(\mathcal{X})$, contains $L^q_+(\mathcal{X})$ for $q > 1$, and is of empty interior for the norm/strong topology of $L^p(\mathcal{X})$ (Resmerita, 2005, Lemma 4.1). In the next Section we will apply KL to joint distributions rather than marginals, but this does not change the formalism. Regarding the use of KL as a Bregman divergence (i.e. ϕ_e as potential) in a mirror descent scheme, one can for instance take $\mathcal{M}(\mathcal{X}) = L^1(\mathcal{X})$ equipped with the weak topology induced by $L^\infty(\mathcal{X})$ and the Lebesgue measure as reference. We have that $\mathcal{P}(X) \cap L^1(\mathcal{X})$ is weakly closed and that KL and ϕ_e are strictly convex, weakly l.s.c. and have weakly compact sublevel sets in $L^1(\mathcal{X})$ by Eggermont (1993, Lemma 2.1, 2.3) (see also Resmerita and Anderssen, 2007, Section 3). By Resmerita (2005, Lemma 4.1), a sufficient condition for KL (resp. ϕ_e) to have a first variation in L^∞ at μ is that $\exists\ \kappa_0, \kappa_1 > 0$ such that $\kappa_0 \leq {}^{d\mu}/_{d\bar{\mu}}(x) \leq \kappa_1$ almost everywhere over $\mathcal{X}$ (resp. ϕ_e for $\bar{\mu} = \rho$). The KL is a strong Bregman divergence, and in particular the MMD for bounded kernels is relatively smooth with respect to ϕ_e, see Proposition 13 in Appendix.

Remark 3 (Mirror descent dual iteration)**.** If $C = \mathcal{M}(\mathcal{X})$ and $\mathcal{F}$ and ϕ were Gâteaux differentiable at μ_n, then provided μ_{n+1} exists, the first-order optimality condition for (7) would give

$$\nabla\phi(\mu_{n+1}) - \nabla\phi(\mu_n) = -\frac{1}{L}\nabla\mathcal{F}(\mu_n). \tag{12}$$

Note that any μ_{n+1} satisfying (12) is optimal for (7) owing to Lemma 11-ii). In the general case, as discussed in finite dimensions in Bauschke et al. (2017, Remark 3), one needs extra assumptions to justify (12), akin to the invertibility of $\nabla\phi$ or that ϕ is essentially smooth or of Legendre type. To avoid any restrictive assumption required to use (12), we stick with the minimal formulation (7) as was also done by Bauschke et al. (2017); Lu et al. (2018). In particular, setting $\phi = \phi_e$ in (12), then exponentiating, leads to the classical multiplicative scheme $\mu_{n+1} = \mu_n e^{-\frac{1}{L}\nabla\mathcal{F}(\mu_n)}$. This observation generalizes to the Iterative Proportional Fitting Procedure, also known as Sinkhorn's algorithm.

4 Applications to Sinkhorn and EM with convergence rates

Applying the results of Section 3, we can now analyze the convergence of two algorithms, Sinkhorn and Expectation-Maximization (EM), by showing that they can be written as mirror descent schemes. Note that the notion of relative smoothness was first introduced by Birnbaum et al. (2011) for this very purpose, to study the convergence of Proportional Response Dynamics.

In both the Sinkhorn and EM settings we will be given two probability spaces $(\mathcal{X}, \bar{\mu})$ and $(\mathcal{Y}, \bar{\nu})$. We recall that $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ denotes the subset of Radon measures $\mathcal{M}_r(\mathcal{X} \times \mathcal{Y})$ with mass 1. We equip $\mathcal{M}_r$ with L^∞ as dual space. A joint measure $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ is also called a *coupling* between its first $p_{\mathcal{X}}\pi$ and second $p_{\mathcal{Y}}\pi$ marginals. We denote by $\Pi(\bar{\mu}, *)$ the set of couplings having first marginal $\bar{\mu}$ and $\Pi(*, \bar{\nu})$ the set of couplings having second marginal $\bar{\nu}$, and $\Pi(\bar{\mu}, \bar{\nu}) = \Pi(\bar{\mu}, *) \cap \Pi(*, \bar{\nu})$ the couplings with marginals $(\bar{\mu}, \bar{\nu})$. We now recall an instrumental disintegration formula:

Fact: Let $\pi, \bar{\pi} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ with $\pi \ll \bar{\pi}$ and $K_{\bar{\pi}}(x, dy) = {}^{\bar{\pi}(dx,dy)}/_{p_{\mathcal{X}}\bar{\pi}(dx)}$. We can write $\bar{\pi} = p_{\mathcal{X}}\bar{\pi} \otimes K_{\bar{\pi}}$ and[4]

$$\mathrm{KL}(\pi|\bar{\pi}) = \mathrm{KL}(p_{\mathcal{X}}\pi|p_{\mathcal{X}}\bar{\pi}) + \int_{\mathcal{X}} \mathrm{KL}(K_\pi|K_{\bar{\pi}})\, dp_{\mathcal{X}}\pi = \mathrm{KL}(p_{\mathcal{X}}\pi|p_{\mathcal{X}}\bar{\pi}) + \mathrm{KL}(\pi|p_{\mathcal{X}}\pi \otimes K_{\bar{\pi}}). \quad (13)$$

This decomposition is at the heart of the two objective functions F_{S} and F_{EM} considered below.

4.1 Sinkhorn

To describe the entropic optimal transport problem we follow Nutz (2021). 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

$$\mathrm{OT}_\epsilon(\bar{\mu}, \bar{\nu}) = \min_{\pi \in \Pi(\bar{\mu}, \bar{\nu})} \mathrm{KL}(\pi|e^{-c/\epsilon}\bar{\mu} \otimes \bar{\nu}). \quad (14)$$

By adding a constant to c we can assume without loss of generality that $e^{-c/\epsilon}\bar{\mu} \otimes \bar{\nu}$ has mass 1. Since c is bounded, (14) admits a unique solution π_*. 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 (Nutz, 2021, Theorem 4.2, Lemma 4.9, Section 6)

$$\mathrm{KL}(\pi|e^{-c/\epsilon}\mu \otimes \nu) = \min_{\tilde{\pi} \in \Pi(\mu, \nu)} \mathrm{KL}(\tilde{\pi}|e^{-c/\epsilon}\mu \otimes \nu). \quad (15)$$

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 (14) by alternative (entropic) projections on $\Pi(\bar{\mu}, *)$ and $\Pi(*, \bar{\nu})$ (Ruschendorf, 1995), i.e. initializing with $\pi_0 \in \Pi_c$, iterate

$$\pi_{n+\frac{1}{2}} = \operatorname*{argmin}_{\pi \in \Pi(\bar{\mu}, *)} \mathrm{KL}(\pi|\pi_n), \quad (16)$$

$$\pi_{n+1} = \operatorname*{argmin}_{\pi \in \Pi(*, \bar{\nu})} \mathrm{KL}(\pi|\pi_{n+\frac{1}{2}}). \quad (17)$$

Let $\mu_n = p_{\mathcal{X}}\pi_n$. More explicitly, (16) is a "rescaling of the rows", $\pi_{n+\frac{1}{2}}(dx, dy) = \pi_n(dx, dy)\bar{\mu}(dx)/p_{\mathcal{X}}\pi_n(dx)$, while (17) is a "rescaling of the columns", $\pi_{n+1}(dx, dy) = \pi_{n+\frac{1}{2}}(dx, dy)\bar{\nu}(dy)/p_{\mathcal{Y}}\pi_{n+\frac{1}{2}}(dy)$. This can be seen as a consequence of (13), as in (16) the first marginal is fixed, so the optimum of (16) is such that the integral term of (13) vanishes. One can also show recursively that $\pi_n \in \Pi_c$ (see Nutz, 2021, Section 6, Lemma 6.22).

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

$$F_{\mathrm{S}}(\pi) = \mathrm{KL}(p_{\mathcal{X}}\pi|\bar{\mu}). \quad (18)$$

Connections between mirror descent and Sinkhorn iterations for the entropic regularized optimal transport problem were first discovered in Mishchenko (2019); Mensch and Peyré (2020); Léger (2020). We propose yet another mirror descent interpretation of Sinkhorn in the spirit of Léger (2020), and use the primal formulation (16)–(17) directly instead of introducing dual potentials. This is stated in the following Proposition, whose complete proof can be found in Appendix D.2.

[4] The last equality stems from: $\mathrm{KL}(\pi|p_{\mathcal{X}}\pi \otimes K_{\bar{\pi}}) = \int_{\mathcal{X}\times\mathcal{Y}} \ln\left(\frac{p_{\mathcal{X}}\pi \otimes K_\pi}{p_{\mathcal{X}}\pi \otimes K_{\bar{\pi}}}\right) dp_{\mathcal{X}}\pi \otimes K_\pi = \int_{\mathcal{X}} \mathrm{KL}(K_\pi|K_{\bar{\pi}})\, dp_{\mathcal{X}}\pi$.

Proposition 4 (Sinkhorn as mirror descent)**.** The Sinkhorn iterations (16) can be written as a mirror descent with objective F_{S} and Bregman divergence KL over the constraint $C = \Pi(*, \bar{\nu})$,

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

Sketch of proof. Let $\mu_n = p_{\mathcal{X}} \pi_n$ where π_n is defined in (16). We have the identity:

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

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

The relative smoothness below is a consequence of the standard KL data processing inequality.

Lemma 5. The functional F_{S} is convex and is 1-relatively smooth w.r.t. 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 (13), we obtain that $D_{F_{\mathrm{S}}}(\tilde{\pi}|\pi) \leq \mathrm{KL}(\tilde{\pi}|\pi)$. □

By considering singular first marginals, it is obvious that there exists no $l > 0$ such that $D_{F_{\mathrm{S}}}(\tilde{\pi}|\pi) \geq l\,\mathrm{KL}(\tilde{\pi}|\pi)$ for all $\tilde{\pi}, \pi \in C$. However, we show in the next Proposition that this inequality actually holds over $\Pi_c \cap C$. This result follows from quantitative stability estimates in the Hilbert metric.

Proposition 6. 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

$$\mathrm{KL}(\tilde{\pi}|\pi) \leq (1 + 4e^{3D_c/\epsilon})\,\mathrm{KL}(p_{\mathcal{X}} \tilde{\pi}|p_{\mathcal{X}} \pi), \tag{19}$$

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

Sketch of proof. For $\pi, \tilde{\pi} \in \Pi_c \cap C$ with their potentials and marginals $(f, g, \mu, \bar{\nu})$ and $(\tilde{f}, \tilde{g}, \tilde{\mu}, \bar{\nu})$ respectively, setting $\|f\|_{\mathrm{var}} = (\sup_{\mathcal{X}} f) - (\inf_{\mathcal{X}} f)$, we can derive the bound

$$\epsilon\,\mathrm{KL}(\tilde{\pi}|\pi) \leq \|\tilde{f} - f\|_{\mathrm{var}} \|\tilde{\mu} - \mu\|_{\mathrm{TV}} + \epsilon\,\mathrm{KL}(\tilde{\mu}|\mu). \tag{20}$$

We can then bound the potentials by the marginals using Proposition 14 in Appendix:

$$\|f - \tilde{f}\|_{\mathrm{var}} + \|g - \tilde{g}\|_{\mathrm{var}} \leq 2\epsilon\, e^{3D_c/\epsilon} \big(\|\mu - \tilde{\mu}\|_{\mathrm{TV}} + \|\bar{\nu} - \bar{\nu}\|_{\mathrm{TV}} \big). \tag{21}$$

Then, chaining (20) and (21) we conclude using Pinsker's inequality. □

The complete proof of Proposition 6 can be found in Appendix D.3. We are now ready to recover convergence rates for Sinkhorn leveraging relative smoothness and strong convexity.

Proposition 7 (Sinkhorn convergence)**.** For all $n \geq 0$, the Sinkhorn iterates verify, for π_* the optimum of (14) and μ_* its first marginal,

$$\mathrm{KL}(\mu_n|\mu_*) \leq \frac{\mathrm{KL}(\pi_*|\pi_0)}{(1 + 4e^{\frac{3D_c}{\epsilon}}) \left(\left(1 + 4e^{-\frac{3D_c}{\epsilon}}\right)^n - 1 \right)} \leq \frac{\mathrm{KL}(\pi_*|\pi_0)}{n}. \tag{22}$$

Proof. Fix $n \geq 0$, we know that $\pi_n, \pi_* \in \mathcal{R} := C \cap \Pi_c$. We conclude by applying Theorem 3 to $\mathcal{F} = F_{\mathrm{S}}$ and $\phi = \mathrm{KL}(\cdot|\pi_*)$, leveraging the results of Lemma 5 and Proposition 6. □

The linear convergence of Sinkhorn for bounded costs c has been known since at least Franklin and Lorenz (1989) and has then been derived also in the non-discrete case and in multimarginal settings (see Carlier, 2022, and references therein). Léger (2020) first obtained sublinear rates for unbounded costs leveraging relative smoothness, using (12) formally and through dual iterations on the potentials. Here we derive the same rate rigorously with a more direct proof using primal iterations, and complete the picture by recovering linear rates of convergence.

4.2 Expectation-Maximization

In this subsection, we show how the EM algorithm can always formally be written as a mirror descent scheme, and, when optimizing the latent variable distribution, results in a convex problem with sublinear convergence rates. 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 $\bar{\nu}$, and we posit a joint distribution $p_q(dx, dy)$ parametrized by an element q of some given set $\mathcal{Q}$. As presented in Neal and Hinton (1998), the goal is to infer q by solving

$$\min_{q \in \mathcal{Q}} \mathrm{KL}(\bar{\nu} | p_{\mathcal{Y}} p_q), \tag{23}$$

where $p_{\mathcal{Y}} p_q(dy) = \int_{\mathcal{X}} p_q(dx, dy)$. The EM approach starts by minimizing a surrogate function of q upperbounding $\mathrm{KL}(\bar{\nu} | p_{\mathcal{Y}} p_q)$. For any $\pi \in \Pi(*, \bar{\nu})$, by the data processing inequality,

$$\mathrm{KL}(\bar{\nu} | p_{\mathcal{Y}} p_q) \leq \mathrm{KL}(\pi | p_q) =: L(\pi, q).$$

Again a consequence of the disintegration formula (13), there is equality if and only if

$$\pi(dx, dy) = p_q(dx, dy) \bar{\nu}(dy) / p_{\mathcal{Y}} p_q(dy). \tag{24}$$

EM then proceeds by alternate minimizations of $L(\pi, q)$ (see Neal and Hinton, 1998, Theorem 1):

$$q_n = \operatorname*{argmin}_{q \in \mathcal{Q}} \mathrm{KL}(\pi_n | p_q), \tag{25}$$

$$\pi_{n+1} = \operatorname*{argmin}_{\pi \in \Pi(*, \bar{\nu})} \mathrm{KL}(\pi | p_{q_n}). \tag{26}$$

The above formulation consists in (25), optimizing the parameters q_n at step n (M-step), and then (26), optimizing the joint distribution π_{n+1} at step $n+1$ (E-step). We choose this order to highlight the analogy of (26) with Sinkhorn's (17). The minimization (26) corresponds to taking an explicit expectation, according to (24) justifying the denomination. On the contrary making explicit the M-step is often difficult.

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

$$F_{\mathrm{EM}}(\pi) = \inf_{q \in \mathcal{Q}} \mathrm{KL}(\pi | p_q). \tag{27}$$

We first show that EM can be formally written as a mirror descent scheme in the following Proposition, whose precise statement with additional assumptions can be found in Appendix D.5-Proposition 17.

Proposition 8 (EM as mirror descent, formal)**.** The EM iterations (25)–(26) can be written as a mirror descent iteration with objective function F_{EM}, Bregman potential ϕ_e and constraints C,

$$\pi_{n+1} = \operatorname*{argmin}_{\pi \in C} \langle \nabla_C F_{\mathrm{EM}}(\pi_n), \pi - \pi_n \rangle + \mathrm{KL}(\pi | \pi_n) \text{ with } \nabla_C F_{\mathrm{EM}}(\pi_n) = \ln(d\pi_n / dp_{q_n}). \tag{28}$$

Sketch of proof. Let π_n be the current EM iterate. Formally, we use the envelope theorem to differentiate F_{EM} and find that $\nabla_C F_{\mathrm{EM}}(\pi_n) = \ln(d\pi_n / dp_{q_n})$ (see Appendix D.4 for a justification based on directional derivatives and Milgrom and Segal (2002, Theorem 3)). Then for any coupling π, we have the identity

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

Thus (28) matches (26). □

Since F_{EM} is in general non-convex, we cannot apply outright the framework developed in Section 3. There is however one direct case of p_q making F_{EM} convex, by optimizing only over its first marginal.

Latent EM. We consider the case where $p_q(dx, dy) = \mu(dx) K(x, dy)$, i.e. p_q is of the form $\mu \otimes K$, with $\mathcal{Q} = \mathcal{P}(\mathcal{X})$ and K kept fixed along iterations. In other words, we choose to only optimize over the density of the latent variable, and keep the mixture parameters fixed. Here K can be interpreted as the conditional distribution of Y given X, $K(x, dy) = \mathbb{P}(Y = y \mid X = x)$. We consider in the following Gibbs distributions with $K(x, dy) = e^{-c(x,y)} \bar{\nu}(dy)$ with c uniformly bounded. The term c can be interpreted as a cost similarly to the entropic optimal transport (14). The parametric setting

considered by Kunstner et al. (2021), who obtained $\mathcal{O}(1/n)$ rates of convergence in KL for EM, can be seen as complementary to ours since they consider a fixed μ and a variable exponential mixture K_θ, with $q = \theta$. In our case, parametrizing by the first marginal, EM (25)–(26) takes the form:

$$\mu_n = \operatorname*{argmin}_{\mu \in \mathcal{P}(\mathcal{X})} \mathrm{KL}(\pi_n | \mu \otimes K), \tag{29}$$

$$\pi_{n+1} = \operatorname*{argmin}_{\pi \in \Pi(*,\bar{\nu})} \mathrm{KL}(\pi | \mu_n \otimes K). \tag{30}$$

First, we necessarily have from (29) that $\mu_n = p_{\mathcal{X}} \pi_n$. Indeed, from the disintegration formula (13), (29) corresponds to minimizing over first marginals. Then, since the E-step (30) corresponds to computing (24), we can rewrite (29)–(30) as:

$$\mu_{n+1}(\cdot) = \int_{\mathcal{Y}} \pi_{n+1}(\cdot, dy) = \mu_n(\cdot) \int_{\mathcal{Y}} \frac{K(\cdot, dy)\bar{\nu}(dy)}{\int_{\mathcal{X}} K(x, dy)\mu_n(dx)}. \tag{31}$$

Define $F_{\text{LEM}}(\pi) := \inf_{\mu \in \mathcal{P}(\mathcal{X})} \mathrm{KL}(\pi | \mu \otimes K)$. Notice that by disintegration (13), F_{LEM} takes the form $F_{\text{LEM}}(\pi) = \mathrm{KL}(\pi | p_{\mathcal{X}} \pi \otimes K)$. To take care of the initialization, we define the operator $T_K : \mu \in \mathcal{P}(\mathcal{X}) \mapsto \int_{\mathcal{X}} \mu(dx) K(x, \cdot) \in \mathcal{M}_r(\mathcal{Y})$, with $K(x, dy) = k(x, y)\bar{\nu}(dy)$ and take $\mu_0 = e^{f_0}\bar{\mu}$ with $f_0 \in L^\infty$ and assume that $T_K \bar{\mu} \gg \bar{\nu}$ (in other words we assume that the mixture applied to the latent space is compatible with all the observations).

Proposition 9 (Latent EM as mirror descent)**.** The latent EM iterations (29)–(30) can be written as mirror descent with objective F_{LEM}, Bregman potential ϕ_e and the constraints $C = \Pi(*, \bar{\nu})$,

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

Proof. Similarly to Proposition 8, we have the identity

$$F_{\text{LEM}}(\pi_n) + \langle \nabla_C F_{\text{LEM}}(\pi_n), \pi - \pi_n \rangle + \mathrm{KL}(\pi | \pi_n) = \mathrm{KL}(\pi | p_{\mathcal{X}} \pi_n \otimes K),$$

where $\nabla_C F_{\text{LEM}}(\pi_n) = \ln(\pi_n / p_{\mathcal{X}} \pi_n \otimes K) \in L^\infty(\mathcal{X}, \mathbb{R})$, see Appendix D.5 for rigorous justifications. Since $\mu_n = p_{\mathcal{X}} \pi_n$ due to (29), we conclude by minimizing over $\pi \in C$. □

We are now ready to state convergence rates of latent EM in the following proposition. The reader may refer to Appendix D.6 for a complete proof.

Proposition 10 (Convergence rate for Latent EM)**.** Set $\mu_* \in \operatorname{argmin}_{\mu \in \mathcal{P}(\mathcal{X})} \mathrm{KL}(\bar{\nu} | T_K(\mu))$. The functional F_{LEM} is convex and 1-smooth relative to ϕ_e. Moreover for $\pi_0 \in \Pi(*, \bar{\nu})$,

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

Sketch of proof. By the disintegration formula (13), we can decompose F_{LEM} as

$$F_{\text{LEM}}(\pi) = \mathrm{KL}(\bar{\nu} | p_{\mathcal{Y}}(p_{\mathcal{X}} \pi \otimes K)) + \int \mathrm{KL}(\pi / \bar{\nu} | (p_{\mathcal{X}} \pi \otimes K) / p_{\mathcal{Y}}(p_{\mathcal{X}} \pi \otimes K)) d\bar{\nu}, \tag{32}$$

and show then straightforwardly that π_* defined by $\pi_*(dx, dy) = \mu_*(dx) k(x, dy) \bar{\nu}(dy) / (T_K \mu_*)(dy)$ is a minimizer of F_{LEM} with $F_{\text{LEM}}(\pi^*) = \mathrm{KL}(\bar{\nu} | T_K \mu_*)$. Then, by the disintegration formula (13) and linearity of the Bregman divergence, $\mathrm{KL}(\pi | \tilde{\pi}) = D_{F_{\text{S}}}(\pi | \tilde{\pi}) + D_{F_{\text{LEM}}}(\pi | \tilde{\pi})$, hence F_{LEM} is 1-relatively smooth w.r.t. ϕ_e. Consequently, Theorem 3 yields:

$$F_{\text{LEM}}(\pi_n) \leq F_{\text{LEM}}(\pi_*) + \frac{\mathrm{KL}(\pi_* | \pi_0)}{n},$$

Finally, we have: $\mathrm{KL}(\bar{\nu} | T_K \mu_n) = \mathrm{KL}(p_{\mathcal{Y}} \pi_n | p_{\mathcal{Y}}(p_{\mathcal{X}} \pi_n \otimes K)) \leq \mathrm{KL}(\pi_n | p_{\mathcal{X}} \pi_n \otimes K) = F_{\text{EM}}(\pi_n)$. □

Richardson–Lucy. Interestingly, the iterations (31) of latent EM correspond precisely to that of Richardson–Lucy deconvolution (Richardson, 1972; Lucy, 1974) where K is a known convolution and one has to recover the original signal μ_* based on the observations $\bar{\nu}$. Thus our proof also gives rates of convergence for this other algorithm from signal processing.

Conclusion. In this paper, we have provided a rigorous proof of convergence of mirror descent under relative smoothness and convexity, which holds in the infinite-dimensional setting of optimization over measure spaces. It enabled us to provide a new and simple way to derive rates of convergence for Sinkhorn's algorithm. We also derived new convergence rates for EM when restricted to the latent distribution, obtaining similar but complementary rates to Kunstner et al. (2021).

References

Aliprantis, C. and Border, K. C. (2006). *Infinite dimensional analysis : a hitchhiker's guide*. Springer, Berlin New York.

Arbel, M., Korba, A., Salim, A., and Gretton, A. (2019). Maximum mean discrepancy gradient flow. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6481–6491.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR.

Attouch, H. and Brezis, H. (1986). Duality for the sum of convex functions in general Banach spaces. In *Aspects of Mathematics and its Applications*, volume 34, pages 125–133. Elsevier.

Attouch, H., Buttazzo, G., and Michaille, G. (2014). *Variational Analysis in Sobolev and BV Spaces*. Society for Industrial and Applied Mathematics.

Bauschke, H. H., Bolte, J., and Teboulle, M. (2017). A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348.

Bauschke, H. H., Borwein, J. M., and Combettes, P. L. (2003). Bregman monotone optimization algorithms. *SIAM Journal on Control and Optimization*, 42(2):596–636.

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

Birnbaum, B., Devanur, N. R., and Xiao, L. (2011). Distributed algorithms via gradient descent for fisher markets. In *Proceedings of the 12th ACM conference on Electronic commerce - EC '11*. ACM Press.

Butnariu, D. and Resmerita, E. (2006). Bregman distances, totally convex functions, and a method for solving operator equations in Banach spaces. *Abstract and Applied Analysis*, 2006:1–39.

Carlier, G. (2022). On the linear convergence of the multimarginal sinkhorn algorithm. *SIAM Journal on Optimization*.

Chen, G. and Teboulle, M. (1993). Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM Journal on Optimization*, 3(3):538–543.

Chen, Y., Georgiou, T., and Pavon, M. (2016). Entropic and displacement interpolation: a computational approach using the Hilbert metric. *SIAM Journal on Applied Mathematics*, 76(6):2375–2396.

Chizat, L. (2021). Convergence rates of gradient methods for convex optimization in the space of measures. *arXiv:2105.08368 [math]*. arXiv: 2105.08368.

Chizat, L. and Bach, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31.

Chu, C., Blanchet, J., and Glynn, P. (2019). Probability functional descent: A unifying perspective on GANs, variational inference, and reinforcement learning. In *International Conference on Machine Learning (ICML)*, volume 97, pages 1213–1222.

Csiszar, I. (1975). I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158.

Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. (2015). Training generative neural networks via maximum mean discrepancy optimization. *Uncertainty in Artificial Intelligence*.

Eggermont, P. P. B. (1993). Maximum entropy regularization for Fredholm integral equations of the first kind. *SIAM Journal on Mathematical Analysis*, 24(6):1557–1576.

Franklin, J. and Lorenz, J. (1989). On the scaling of multidimensional matrices. *Linear Algebra and its Applications*, 114-115:717–735.

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*.

Lan, G., Lu, Z., and Monteiro, R. D. C. (2011). Primal-dual first-order methods with $o(1/e)$ iteration-complexity for cone programming. *Mathematical Programming*, 126:1–29.

Léger, F. (2020). A gradient descent perspective on Sinkhorn. *Applied Mathematics & Optimization*, 84(2):1843–1855.

Lu, H., Freund, R. M., and Nesterov, Y. (2018). Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354.

Lucy, L. B. (1974). An iterative technique for the rectification of observed distributions. *The Astronomical Journal*, 79:745.

Luise, G., Salzo, S., Pontil, M., and Ciliberto, C. (2019). Sinkhorn barycenters with free support via Frank-Wolfe algorithm. *Advances in neural information processing systems*, 32.

Maddison, C. J., Paulin, D., Teh, Y. W., and Doucet, A. (2021). Dual space preconditioning for gradient descent. *SIAM Journal on Optimization*, 31(1):991–1016.

Mei, S., Montanari, A., and Nguyen, P.-M. (2018). A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671.

Mensch, A. and Peyré, G. (2020). Online Sinkhorn: Optimal transport distances from sample streams. *Advances in Neural Information Processing Systems*, 33:1657–1667.

Milgrom, P. and Segal, I. (2002). Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–610.

Mishchenko, K. (2019). Sinkhorn algorithm as a special case of stochastic mirror descent. *arXiv preprint arXiv:1909.06918*.

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.

Nutz, M. (2021). Introduction to entropic optimal transport. `http://www.math.columbia.edu/~mnutz/docs/EOT_lecture_notes.pdf`.

Peypouquet, J. (2015). *Convex optimization in normed spaces: theory, methods and examples*. Springer.

Peyré, G. and Cuturi, M. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Phelps, R. R. (1989). *Convex Functions, Monotone Operators and Differentiability*. Springer Berlin Heidelberg.

Resmerita, E. (2005). Regularization of ill-posed problems in Banach spaces: convergence rates. *Inverse Problems*, 21(4):1303–1314.

Resmerita, E. and Anderssen, R. S. (2007). Joint additive Kullback–Leibler residual minimization and regularization for linear inverse problems. *Mathematical Methods in the Applied Sciences*, 30(13):1527–1544.

Richardson, W. H. (1972). Bayesian-based iterative method of image restoration. *Journal of the Optical Society of America*, 62(1):55.

Rotskoff, G. M. and Vanden-Eijnden, E. (2018). Trainability and accuracy of neural networks: An interacting particle system approach. Technical report. (`https://arxiv.org/abs/1805.00915`).

Ruschendorf, L. (1995). Convergence of the iterative proportional fitting procedure. *The Annals of Statistics*, 23(4).

Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians*. Springer International Publishing.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007). A Hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31. Springer.

Steinwart, I. and Christmann, A. (2008). *Support vector machines*. Springer Science & Business Media.

Trillos, N. G. and Sanz-Alonso, D. (2020). The Bayesian update: variational formulations and gradient flows. *Bayesian Analysis*, 15(1):29–56.