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

Pierre-Cyril Aubin-Frankowski
DI ENS, Ecole normale supérieure,
Université PSL, CNRS, INRIA Paris
pierre-cyril.aubin@inria.fr

Anna Korba
CREST, ENSAE
IP Paris
anna.korba@ensae.fr

Flavien Léger
INRIA Paris
flavien.leger@inria.fr

Abstract

Many problems in machine learning can be formulated as optimizing a convex functional over a vector 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 convex pairs of functionals. Such assumptions allow to handle non-smooth functionals such as the Kullback–Leibler (KL) divergence. Applying our result to joint distributions and KL, 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. When optimizing only on the latent distribution while fixing the mixtures parameters – which corresponds to the Richardson–Lucy deconvolution scheme in signal processing – 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 assesses 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 on 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

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

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 and target specifically the KL divergence. Indeed, when using the entropy as Bregman divergence, mirror descent is known to yield multiplicative updates for the measures. In this paper we study two such schemes in machine learning, i.e. Sinkhorn's algorithm, widely used to solve entropic optimal transport (Peyré and Cuturi, 2019), and the EM algorithm, a very common approach to fitting probabilistic models.

Related work. Chizat (2021) gave convergence rates of mirror descent on measure spaces for integral functionals with Lipschitz gradients, 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. Their setting is complementary to ours since we consider instead fixed mixtures and a nonparametric latent distribution. A prominent alternative setting to ours when optimizing over measures is based on (grid-free) Wasserstein gradient flows, which we do not cover here (see Appendix E for a discussion on the difference of geometries). A remarkable discussion on the optimization over measures with different geometries can be found in Trillos and Sanz-Alonso (2020).

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. Defining Bregman divergences through directional derivatives, we are able to consider objective functionals over measures that are not smooth in the "standard" sense, but satisfy relative smoothness and/or convexity with respect to a Bregman divergence, e.g. KL. 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 $\mathcal{Y}$ (see Aliprantis and Border, 2006). We write $\mathrm{Int}\, C$ for the interior of a set $C \subset \mathcal{Y}$. 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, for any $\mu, \nu \in \mathcal{M}(\mathcal{X})$, we write $\mu \ll \nu$ when μ is absolutely continuous w.r.t ν, i.e. when it has a Radon–Nikodym derivative $d\mu/d\nu$.

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. In infinite dimensions, for functionals defined on positive measures, such as the negative entropy, $\mathrm{Int}(\mathrm{dom}(\mathcal{F}))$ is however empty[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, whenever the directional derivative is a linear function in a restricted set of directions, the notion of first variation that comes next will enable us 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 $\mathbb{R}_+^* \ni h \mapsto \frac{\mathcal{F}(\nu+h\mu)-\mathcal{F}(\nu)}{h}$ 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 $\mathrm{span}(\mathrm{dom}(\mathcal{F}) \cap C - \mu)$, such that:

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

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, we instead take it most in often in $L^\infty(\mathcal{X})$. We now introduce Bregman divergences over measures through directional derivatives.

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*.

[1] Intuitively if $\mathcal{X}$ contains an open set, then any positive measure can be infinitesimally perturbed to have negative values. For finite sets $\mathcal{X}$, this phenomenon does not occur, see also Example 2.

[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.

In the previous definition, the restriction to $\mathrm{dom}(\phi)$ is necessary to avoid the substraction of infinite values. As a direct consequence of Lemma 12 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 the last term in (4) being linear, 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. for two convex ϕ, ψ, $D_{\phi+\psi} = D_\phi + D_\psi$. Moreover, it is idempotent, as shown in the following lemma.

Lemma 1 (Idempotence of Bregman divergence)**.** 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 $\nabla_C \phi(\xi)$ exists, then, for all $\mu, \nu \in C \cap \mathrm{dom}(\phi)$, $D_{D_\phi(\cdot|\xi)}(\nu|\mu) = D_\phi(\nu|\mu)$.

Proof. Since $\psi : \tilde{\mu} \mapsto -d^+\phi(\xi)(\tilde{\mu} - \xi) = -\langle \nabla_C \phi(\xi), \tilde{\mu} - \xi \rangle$ is convex over $C \cap \mathrm{dom}(\phi)$, we can apply the linearity of the Bregman divergence:

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

since the Bregman divergence of a constant or of a linear form is null. □

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

Definition 4. (Relative smoothness and convexity) 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

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

Example 1 (L-smoothness)**.** Choosing $\phi(\mu) = \|\mu\|^2_{\mathcal{M}(\mathcal{X})}$ the square norm on $\mathcal{M}(\mathcal{X})$ shows that relative smoothness extends the notion of L-smooth functionals (see for instance Chizat (2021, Lemma 3.1)), i.e. functionals with L-Lipschitz Gâteaux derivative, thus satisfying:

$$\mathcal{F}(\nu) - \mathcal{F}(\mu) - d^+\mathcal{F}(\mu)(\nu - \mu) \leq L\|\nu - \mu\|^2. \tag{7}$$

Notice that by Lemma 1, provided $\nabla_C \phi(\xi)$ is well-defined, a Bregman divergence objective $D_\phi(\cdot|\xi)$ is always 1-relatively smooth and strongly convex w.r.t. ϕ. This we will heavily exploit for mirror descent schemes that involve the KL divergence both as an objective and Bregman divergence in Section 4. Interestingly, relative smoothness and convexity can be characterized in different, equivalent ways, see Lemma 13 in the Appendix. We now turn to the analysis of the mirror descent scheme using the above framework.

3 Mirror descent over measures and convergence

In the following, ϕ is assumed to be strictly convex. The relative smoothness assumption (5) of the convex objective functional $\mathcal{F}$ w.r.t. 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

$$\mu_{n+1} = \operatorname*{argmin}_{\nu \in C} \{d^+\mathcal{F}(\mu_n)(\nu - \mu_n) + L D_\phi(\nu|\mu_n)\}. \tag{8}$$

Let $\mathcal{R} \subset C$ be a given subset. As proven later in this section, sufficient conditions for the convergence of the scheme (8) are:

- ($\mathbf{A}_1$) (Existence) The sequence of iterates $(\mu_n)_{n\in\mathbb{N}}$ defined by (8) exist, belong to $\mathcal{R}$, and are unique.
- ($\mathbf{A}_2$) (Relative smoothness/convexity) For some $l, L \geq 0$, the functional $\mathcal{F}$ is L-smooth and l-strongly convex relative to ϕ as in Definition 4 for elements of $\mathcal{R}$.

($\mathbf{A}_3$) (Existence of first variation of ϕ) For each $n \geq 0$, the first variation $\nabla_C \phi(\mu_n)$ exists.

These assumptions have to be verified on a case-by-case basis. In the simplest case, one can take $\mathcal{R} = C$. However the set $\mathcal{R}$ does not have to be convex (see Section 4). On the other hand, Assumption (**A'**$_1$) below, ensures that the iterates in (8) are well-defined, uniqueness resulting from the strict 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. (iii) The iterates belong to $\mathcal{R}$.

By Attouch et al. (2014, Theorem 3.2.2), Assumption (**A'**$_1$) implies ($\mathbf{A}_1$). Indeed, (i) τ-lower semicontinuity and τ-compactness guarantee the existence of minimizers of the objective (8), while (ii) the existence of $\nabla_C \phi(\mu_n)$ 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. We refer to Appendix B for more details on proving that Assumption (**A'**$_1$) holds for some $\mathcal{F}$, ϕ and C.

Mirror descent can also be defined through a subdifferential constraint if $\mathcal{F}$ also has first variations.

Lemma 2 (Mirror descent dual iteration)**.** If $\nabla_C \mathcal{F}(\mu_n)$ and $\nabla_C \phi(\mu_n)$ exist for all $n \geq 0$, then (8) is equivalent to

$$\nabla_C \phi(\mu_n) - \frac{1}{L} \nabla_C \mathcal{F}(\mu_n) \in \partial_C \phi(\mu_{n+1}) := \{p \,|\, \forall \nu \in C,\ \langle p, \nu - \mu_{n+1} \rangle \leq d^+\phi(\mu_{n+1})(\nu - \mu_{n+1})\} \tag{9}$$

Thus, if $\partial_C \phi(\mu_{n+1}) = \{\nabla_C \phi(\mu_{n+1})\}$, (8) corresponds to $\nabla_C \phi(\mu_{n+1}) - \nabla_C \phi(\mu_n) = -\frac{1}{L} \nabla \mathcal{F}(\mu_n)$.

Proof. The minimization (8) is equivalent to having, for all $\nu \in C$,

$$\begin{gathered} d^+\mathcal{F}(\mu_n)(\nu - \mu_n) + L D_\phi(\nu|\mu_n) \geq d^+\mathcal{F}(\mu_n)(\mu_{n+1} - \mu_n) + L D_\phi(\mu_{n+1}|\mu_n) \\ \langle \nabla_C \mathcal{F}(\mu_n) - L \nabla_C \phi(\mu_n), \nu - \mu_{n+1} \rangle + L(\phi(\nu) - \phi(\mu_{n+1})) \geq 0. \end{gathered}$$

Take $\tilde{\nu} \in C$, set $\nu = \mu_{n+1} + t(\tilde{\nu} - \mu_{n+1})$ for $t \in [0, 1]$. Taking the limit $t \to 0^+$ yields the result. □

In the general case, as discussed in finite dimensions in Bauschke et al. (2017, Remark 3), one needs extra assumptions to justify that μ_{n+1} exists in (9), 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 (9), we stick with the minimal formulation (8) as was also done by Bauschke et al. (2017); Lu et al. (2018).

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

Lemma 3 (Three-point inequality)**.** 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 $\bar{\nu} = \operatorname{argmin}_{\nu \in C} \{\mathcal{G}(\nu) + D_\phi(\nu|\mu)\}$, then for all $\nu \in C \cap \operatorname{dom}(\phi) \cap \operatorname{dom}(\mathcal{G})$:

$$\mathcal{G}(\nu) + D_\phi(\nu|\mu) \geq \mathcal{G}(\bar{\nu}) + D_\phi(\bar{\nu}|\mu) + D_\phi(\nu|\bar{\nu}). \tag{10}$$

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

$$D_f(\nu|\bar{\nu}) = D_{\mathcal{G}}(\nu|\bar{\nu}) + D_{D_\phi(\cdot|\mu)}(\nu|\bar{\nu}) = D_{\mathcal{G}}(\nu|\bar{\nu}) + D_\phi(\nu|\bar{\nu}) \geq D_\phi(\nu|\bar{\nu}). \tag{11}$$

By optimality of $\bar{\nu}$, for all $\nu \in C$, $d^+(f)(\bar{\nu})(\nu - \bar{\nu}) = \lim_{h \to 0^+} (f((1-h)\bar{\nu} + h\nu) - f(\bar{\nu}))/h \geq 0$; which is equivalent to $f(\nu) \geq f(\bar{\nu}) + D_f(\nu|\bar{\nu})$. We conclude using (11) and the definition of f. □

[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 $\mathcal{M}(\mathcal{X})$.

The following theorem gives the rate of convergence of mirror descent for relatively smooth and convex pairs of functionals, and extends to infinite dimensions the convergence result of Lu et al. (2018, Theorem 3.1). Its proof can be found in Appendix F.1.

Theorem 4 (Convergence rate)**.** Assume that Assumptions ($\mathbf{A}_1$), ($\mathbf{A}_2$) and ($\mathbf{A}_3$) hold. Consider the mirror descent scheme (8), then for all $n \geq 0$ and all $\nu \in \mathrm{dom}(\mathcal{F}) \cap \mathrm{dom}(\phi) \cap \mathcal{R}$, we have

$$\mathcal{F}(\mu_n) - \mathcal{F}(\nu) \leq \frac{l D_\phi(\nu|\mu_0)}{\left(1 + \frac{l}{L-l}\right)^n - 1} \leq \frac{L}{n} D_\phi(\nu|\mu_0), \tag{12}$$

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

Remark 2 (About the proof of convergence)**.** Our proof resembles the one of Lu et al. (2018), which also relies on a three-point inequality as stated in Lemma 3. However, the proof of the latter inequality in finite dimensions relies on a sum of subdifferentials formula, which is hard to verify for general functionals and in particular for the KL divergeence defined below (see also Remark 6 in Appendix). On the contrary, by working with directional derivatives and first variations, we circumvent most of the difficulties related to (sub)differentiability.

An important example is the one discussed below where ϕ is chosen to be the negative entropy ϕ_e.

Example 2 (The KL divergence and negative entropy)**.** 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

$$\mathrm{KL}(\mu|\bar{\mu}) = \int_{\mathcal{X}} \ln\left({}^{d\mu}/_{d\bar{\mu}}(x)\right) d\mu(x), \quad \phi_e(\mu) = \int_{\mathcal{X}} \ln(\mu(x))\mu(x) d\rho(x) = \mathrm{KL}(\mu|\rho), \tag{13}$$

where ρ is some reference finite measure on $\mathcal{X}$. 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. $D_{\phi_e}(\mu|\bar{\mu}) = \mathrm{KL}(\mu|\bar{\mu})$; hence one can choose $\phi = \phi_e$ for the mirror descent scheme (8). Assumption (**A'**$_1$), guaranteeing that the iterates (8) are well-posed, is satisfied for instance when $\mathcal{M}(\mathcal{X}) = L^1(\mathcal{X})$ and if there exists $\kappa_0, \kappa_1 > 0$ such that $\kappa_0 \leq {}^{d\mu_n}/_{d\bar{\mu}}(x) \leq \kappa_1$ almost everywhere over $\mathcal{X}$ for any n, which is the case for Sinkhorn's and EM iterates. This implies that the first variations of the negative entropy or KL belong to L^∞ at μ_n, see also Remark 5 in Appendix for more details. Moreover, by Lemma 2, exponentiating the dual iteration (9) recovers the classical multiplicative scheme: $\mu_{n+1} = \mu_n e^{-\frac{1}{L}\nabla\mathcal{F}(\mu_n)}$, $n \geq 0$. This observation generalizes to the Iterative Proportional Fitting Procedure, also known as Sinkhorn's algorithm.

KL is a strong Bregman divergence, in the sense that it dominates a wide range of objective functionals $\mathcal{F}$. Indeed, we already know from Section 2 and the idempotence property that $\mathcal{F} = \mathrm{KL}(\cdot|\bar{\mu})$ is 1-relatively smooth w.r.t. ϕ_e, since for any $\mu, \nu \in \mathrm{dom}(\mathcal{F})$, $D_{\mathcal{F}}(\nu|\mu) = D_{\phi_e}(\nu|\mu) = \mathrm{KL}(\nu|\mu)$. In Section 4 we will extensively use this fact (sometimes applying KL to joint distributions rather than marginals, e.g. for Sinkhorn's algorithm). This is of crucial importance because $\mathcal{F} = \mathrm{KL}(\cdot|\bar{\mu})$ is not a smooth objective in the "standard" sense (see Example 1) - hence convergence proofs requiring the latter cannot apply -, but it is relatively smooth w.r.t. itself. Indeed the KL diverges for Dirac masses, so is unbounded over the bounded set $\mathcal{P}(\mathcal{X})$, and thus KL does not have subquadratic growth (7) w.r.t. any norm on measures. Other objective functionals can be dominated by KL, such as the Maximum Mean Discrepancy (MMD) for bounded kernels, see Proposition 14 in Appendix D.

4 Applications to Sinkhorn and EM with convergence rates

We now analyze the convergence of two algorithms, Sinkhorn and Expectation-Maximization (EM), by showing that they can be written as mirror descent schemes based on the KL divergence, in order to apply the results of Section 3. Note that the 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 $L^\infty(\mathcal{X} \times \mathcal{Y})$ 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:

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

$$\mathrm{KL}(\pi|\bar{\pi}) = \mathrm{KL}(p_{\mathcal{X}}\pi|p_{\mathcal{X}}\bar{\pi}) + \int_{\mathcal{X}} \mathrm{KL}(K_\pi|K_{\bar{\pi}})\, dp_{\mathcal{X}}\pi = \mathrm{KL}(p_{\mathcal{X}}\pi|p_{\mathcal{X}}\bar{\pi}) + \mathrm{KL}(\pi|p_{\mathcal{X}}\pi \otimes K_{\bar{\pi}}). \quad (14)$$

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

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

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, (15) admits a unique solution π_*. We use a characterization (Nutz, 2021, Theorem 4.2, Lemma 4.9, Section 6) of the set of cyclically invariant couplings, and define it as follows, as the set of couplings π that solve an entropic optimal transport problem for their own marginals,

$$\Pi_c = \{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \mid \mathrm{KL}(\pi|e^{-c/\epsilon}\mu \otimes \nu) = \min_{\tilde{\pi} \in \Pi(\mu,\nu)} \mathrm{KL}(\tilde{\pi}|e^{-c/\epsilon}\mu \otimes \nu),\ (\mu, \nu) = (p_{\mathcal{X}}\pi, p_{\mathcal{Y}}\pi)\}. \quad (16)$$

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 solves (15) by alternating (entropic) projections on $\Pi(\bar{\mu}, *)$ and $\Pi(*, \bar{\nu})$ (Ruschendorf, 1995), i.e. initializing with $\pi_0 \in \Pi_c$, iterate

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

$$\pi_{n+1} = \operatorname*{argmin}_{\pi \in \Pi(*,\bar{\nu})} \mathrm{KL}(\pi|\pi_{n+\frac{1}{2}}). \quad (18)$$

Let $\mu_n = p_{\mathcal{X}}\pi_n$. More explicitly, (17) 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 (18) 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 (14), as in (17) the first marginal is fixed, so the optimum of (17) is such that the integral term of (14) vanishes. One can also show recursively that $\pi_n \in \Pi_c$ (see Nutz, 2021, Section 6, Lemma 6.22).

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

$$F_{\mathrm{S}}(\pi) = \mathrm{KL}(p_{\mathcal{X}}\pi|\bar{\mu}). \quad (19)$$

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 (17)–(18) directly instead of introducing dual potentials. This is stated in the following Proposition, whose complete proof can be found in Appendix F.2.

Proposition 5 (Sinkhorn as mirror descent)**.** The Sinkhorn iterations (17) 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 (17). 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 $\pi \in C$. □

We first show relative smoothness of F_{S} relatively to ϕ_e, as a consequence of the standard KL data processing inequality, i.e. KL of the marginals is smaller than Kl of the plans.

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

Lemma 6. 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. Then (14) results in $D_{F_{\mathrm{S}}}(\tilde{\pi}|\pi) \leq \mathrm{KL}(\tilde{\pi}|\pi)$. □

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 $\tilde{\pi}, \pi \in C$; i.e. that relative strong convexity of F_{S} relatively to ϕ_e does not hold *over all* C. However, we show in the next Proposition that this inequality actually holds over $\mathcal{R} = \Pi_c \cap C$. Its complete proof can be found in Appendix F.3.

Proposition 7. 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

$$\mathrm{KL}(\tilde{\pi}|\pi) \leq (1 + 4e^{3D_c/\epsilon})\,\mathrm{KL}(p_{\mathcal{X}}\tilde{\pi}|p_{\mathcal{X}}\pi), \tag{20}$$

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

$$\epsilon\,\mathrm{KL}(\tilde{\pi}|\pi) \leq \|\tilde{f} - f\|_{\mathrm{var}}\|\tilde{\mu} - \mu\|_{\mathrm{TV}} + \epsilon\,\mathrm{KL}(\tilde{\mu}|\mu). \tag{21}$$

We can then bound the potentials by the marginals using Proposition 15 in Appendix:

$$\|f - \tilde{f}\|_{\mathrm{var}} + \|g - \tilde{g}\|_{\mathrm{var}} \leq 2\epsilon\, e^{3D_c/\epsilon}\big(\|\mu - \tilde{\mu}\|_{\mathrm{TV}} + \|\bar{\nu} - \bar{\nu}\|_{\mathrm{TV}}\big). \tag{22}$$

Then, chaining (21) and (22) we conclude using Pinsker's inequality. □

We are now ready to recover convergence rates for Sinkhorn leveraging relative smoothness and strong convexity.

Proposition 8 (Sinkhorn convergence)**.** For all $n \geq 0$, the Sinkhorn iterates verify, for π_* the optimum of (15) and μ_* its first marginal,

$$\mathrm{KL}(\mu_n|\mu_*) \leq \frac{\mathrm{KL}(\pi_*|\pi_0)}{(1 + 4e^{\frac{3D_c}{\epsilon}})\left(\left(1 + 4e^{-\frac{3D_c}{\epsilon}}\right)^n - 1\right)} \leq \frac{\mathrm{KL}(\pi_*|\pi_0)}{n}. \tag{23}$$

Proof. Fix $n \geq 0$, we know that $\pi_n, \pi_* \in \mathcal{R} := C \cap \Pi_c$. We conclude by applying Theorem 4 to $\mathcal{F} = F_{\mathrm{S}}$ and $\phi = \mathrm{KL}(\cdot|\pi_*)$, leveraging the results of Lemma 6 and Proposition 7. □

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). To derive this linear rate (the first inequality in (23)), we require relative strong convexity of the objective - hence we also fundamentally rely on the boundedness of the cost, as one can see from the assumptions of Proposition 7. Indeed the proof of Proposition 7 relies on the classical result that the soft c-transforms are contractions in the Hilbert metric; a result also at the heart of other proofs for the linear rate (see Franklin and Lorenz, 1989; Chen et al., 2016, for a proof). Regarding the sublinear convergence (the second inequality in (23)), Léger (2020) first obtained sublinear rates for unbounded costs leveraging relative smoothness, using (9) formally and through dual iterations on the potentials. In this Section we assumed $c \in L^\infty(\mathcal{X} \times \mathcal{Y})$, in order to manipulate finite quantities in our computations (e.g. first variations in the proof of Proposition 5) - hence the latter can be seen as a convenient working hypothesis. In this paper we derive the same rate as Léger (2020) 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

$$\min_{q \in \mathcal{Q}} \mathrm{KL}(\bar{\nu} | p_{\mathcal{Y}} p_q), \tag{24}$$

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, as a consequence of the disintegration formula (14), there is equality if and only if

$$\pi(dx, dy) = p_q(dx, dy) \bar{\nu}(dy) / p_{\mathcal{Y}} p_q(dy). \tag{25}$$

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

$$q_n = \operatorname*{argmin}_{q \in \mathcal{Q}} \mathrm{KL}(\pi_n | p_q), \tag{26}$$

$$\pi_{n+1} = \operatorname*{argmin}_{\pi \in \Pi(*, \bar{\nu})} \mathrm{KL}(\pi | p_{q_n}). \tag{27}$$

The above formulation consists in (26), optimizing the parameters q_n at step n (M-step), and then (27), optimizing the joint distribution π_{n+1} at step $n+1$ (E-step). We choose this order to highlight the analogy of (27) with Sinkhorn's (18). The minimization (27) corresponds to taking an explicit expectation, according to (25) 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

$$F_{\mathrm{EM}}(\pi) = \inf_{q \in \mathcal{Q}} \mathrm{KL}(\pi | p_q). \tag{28}$$

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 F.4.

Proposition 9 (EM as mirror descent, formal)**.** The EM iterations (26)–(27) can be written as a mirror descent iteration with objective function F_{EM}, Bregman potential ϕ_e and constraints C,

$$\pi_{n+1} = \operatorname*{argmin}_{\pi \in C} \langle \nabla_C F_{\mathrm{EM}}(\pi_n), \pi - \pi_n \rangle + \mathrm{KL}(\pi | \pi_n) \text{ with } \nabla_C F_{\mathrm{EM}}(\pi_n) = \ln(d\pi_n / dp_{q_n}). \tag{29}$$

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 F.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 (29) matches (27). □

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 (15).

Remark 3 (Various EM)**.** The general goal of EM is to fit, through the objective function F_{EM}, a parametric distribution, e.g. a mixture of Gaussians, to some observed data Y. One needs to estimate both the latent variable distribution on X (i.e. weights of each Gaussian) and the parameters of conditionals $P(Y | X = x)$ (e.g. means and covariances of each Gaussian). Latent EM focuses on learning the mixture weights, since it consists in optimizing over the nonparametric latent distribution μ, which can be continuous or discrete. In contrast, 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 $q = \theta$.

Parametrizing by the first marginal, EM iterations (26)–(27) takes the following form for Latent EM:

$$\mu_n = \operatorname*{argmin}_{\mu \in \mathcal{P}(\mathcal{X})} \mathrm{KL}(\pi_n | \mu \otimes K), \tag{30}$$

$$\pi_{n+1} = \operatorname*{argmin}_{\pi \in \Pi(*, \bar{\nu})} \mathrm{KL}(\pi | \mu_n \otimes K). \tag{31}$$

First, we necessarily have from (30) that $\mu_n = p_{\mathcal{X}} \pi_n$. Indeed, from the disintegration formula (14), (30) corresponds to minimizing over first marginals. Then, since the E-step (31) corresponds to computing (25), we can rewrite (30)–(31) as:

$$\mu_{n+1}(\cdot) = \int_{\mathcal{Y}} \pi_{n+1}(\cdot, dy) = \mu_n(\cdot) \int_{\mathcal{Y}} \frac{K(\cdot, dy)\bar{\nu}(dy)}{\int_{\mathcal{X}} K(x, dy)\mu_n(dx)}. \tag{32}$$

Define $F_{\text{LEM}}(\pi) := \inf_{\mu \in \mathcal{P}(\mathcal{X})} \mathrm{KL}(\pi | \mu \otimes K)$. Notice that by disintegration (14), 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 10 (Latent EM as mirror descent)**.** The latent EM iterations (30)–(31) 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 9, 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 F.5 for rigorous justifications. Since $\mu_n = p_{\mathcal{X}} \pi_n$ due to (30), 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 F.6 for a complete proof.

Proposition 11 (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 (14), we can decompose F_{LEM} as

$$F_{\text{LEM}}(\pi) = \mathrm{KL}(\bar{\nu} | p_{\mathcal{Y}}(p_{\mathcal{X}} \pi \otimes K)) + \int \mathrm{KL}(\pi / \bar{\nu} | (p_{\mathcal{X}} \pi \otimes K) / p_{\mathcal{Y}}(p_{\mathcal{X}} \pi \otimes K)) d\bar{\nu}, \tag{33}$$

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 (14) 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 4 yields:

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

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

Remark 4 (Richardson–Lucy)**.** Interestingly, the iterations (32) of latent EM correspond precisely to that of Richardson–Lucy deconvolution (Richardson, 1972; Lucy, 1974) where K is a known convolution and one aims at recovering the original signal μ_* based on the observations $\bar{\nu}$. Thus our proof yields rates of convergence for this other algorithm from signal processing, a novel result to the best of our knowledge.

Conclusion: 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. The latter condition can handle objective functionals that are not smooth in the standard sense, such as the ubiquitous KL. 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 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.

Ambrosio, L., Gigli, N., and Savaré, G. (2008). *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media.

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.

Borwein, J. and Goebel, R. (2003). Notions of relative interior in banach spaces. *Journal of Mathematical Sciences*, 115(4).

Braides, A. (2002). Γ*-convergence for Beginners*, volume 22. Oxford University 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.

Dal Maso, G. (1987). γ-convergence and μ-capacities. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 14(3):423–464.

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.

Korba, A., Aubin-Frankowski, P.-C., Majewski, S., and Ablin, P. (2021). Kernel stein discrepancy descent. In *International Conference on Machine Learning*, pages 5719–5730. PMLR.

Korba, A., Salim, A., Arbel, M., Luise, G., and Gretton, A. (2020). A non-asymptotic analysis for stein variational gradient descent. *Advances in Neural Information Processing Systems*, 33:4672–4682.

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, NeurIPS 2019 OTML Workshop*.

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`.

Otto, F. (2001). The geometry of dissipative evolution equations: the porous medium equation.

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).

Salim, A., Korba, A., and Luise, G. (2020). The wasserstein proximal gradient algorithm. *Advances in Neural Information Processing Systems*, 33:12356–12366.

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.

Villani, C. (2003). *Topics in optimal transportation*, volume 58. American Mathematical Soc.

Wibisono, A. (2018). Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem. In *Conference on Learning Theory*, pages 2093–3027. PMLR.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Purely theoretical work to prove the convergence of algorithms
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

A Definition of Gâteaux and Fréchet derivatives

We first recall the notion of Gâteaux and Fréchet derivatives for $\mathcal{F} : \mathcal{M}(\mathcal{X}) \to \mathbb{R} \cup \{\pm\infty\}$ where $\mathcal{M}(\mathcal{X})$ is a topological vector space (Aliprantis and Border, 2006, Chapter 7, pp.267,273), see also Phelps (1989, Section 1) for Banach spaces.

Definition 5. 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 $\mu \in \mathcal{M}(\mathcal{X})$:

$$\nabla \mathcal{F}(\nu)(\mu) = \lim_{h \to 0} \frac{\mathcal{F}(\nu + h\mu) - \mathcal{F}(\nu)}{h}. \tag{34}$$

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

Definition 6. If $\mathcal{M}(\mathcal{X})$ is a normed space, the function $\mathcal{F}$ is said to be Fréchet differentiable at ν if there exists a bounded linear form $\delta\mathcal{F}(\nu, \cdot) : \mathcal{M}(\mathcal{X}) \to \mathbb{R}$ such that

$$\mathcal{F}(\nu + h\mu) = \mathcal{F}(\mu) + h\delta\mathcal{F}(\nu, \mu) + ho(\|\mu\|_{\mathcal{M}(\mathcal{X})}) \tag{35}$$

Equivalently, the operator $\delta\mathcal{F}(\nu, \cdot)$ is called the Fréchet derivative of $\mathcal{F}$ at ν if it is a Gâteaux derivative of $\mathcal{F}$ at ν and the limit (34) holds uniformly in μ in the unit ball (or unit sphere) in $\mathcal{M}(\mathcal{X})$.

If $\mathcal{F}$ is Fréchet differentiable, then it is also Gâteaux differentiable, and its Fréchet and Gâteaux derivatives agree: $\nabla \mathcal{F}(\nu)(\mu) = \delta\mathcal{F}(\nu, \mu)$.

B Additional details on the well-posedness of the mirror descent scheme

Recall Assumption (**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 $\mathcal{G}_n + D_\phi(\cdot|\mu_n) + i_C$[5] has at least one τ-compact sublevel set. (ii) For each $n \geq 0$, the first variations $\nabla_C \phi(\mu_n)$ exist. (iii) The iterates belong to $\mathcal{R}$.

Case where $\mathcal{F}$ has first variations. Equip $\mathcal{M}(\mathcal{X})$ with a topology τ that is stronger than the $\mathcal{M}(\mathcal{X})^*$-weak topology. If $\mathcal{F}$ has first variations, then we can even remove $\mathcal{G}_n(\cdot)$ from Assumption (**A'**$_1$), since $\mathcal{G}_n(\cdot)$ is linear on the set of interest and τ-l.s.c. Whence we get the simpler assumption

(**A''**$_1$) (Lower semicontinuity and coercivity) For each $n \geq 0$, the iterates belong to $\mathcal{R}$ and the first variations $\nabla_C \phi(\mu_n)$ and $\nabla_C \mathcal{F}(\mu_n)$ exist. Moreover the set C is τ-closed in $\mathcal{M}(\mathcal{X})$, the functional ϕ is proper and τ-l.s.c. when restricted to C, and $\phi(\cdot)$ has at least one τ-compact sublevel set when restricted to $C \cap \mathrm{dom}(\mathcal{F})$.

Weakly compact sets of $\mathcal{M}(\mathcal{X})$. In finite dimensions, a set is compact iff bounded and closed; however, in infinite dimensions, characterizing compact sets is more delicate. Below we recall some classical set of conditions that guarantee (weak) compactness or lower semicontinuity.

If $\mathcal{M}(\mathcal{X})$ is a reflexive Banach space, then the weakly compact sets are just the bounded weakly closed sets, as a consequence of the Banach–Alaoglu theorem (see e.g. Attouch et al., 2014, Theorem 2.4.2). In other cases, one needs more specific theorems such as Dunford–Pettis' theorem for $L^1(\rho)$ (see e.g. Attouch et al., 2014, Theorem 2.4.5). Since we are dealing with convex functions, for normed $\mathcal{M}(\mathcal{X})$, strongly closed sublevel sets are also weakly closed, a result known as Mazur's lemma. So the notions of weakly l.s.c. and strongly l.s.c. convex functions coincide, as recalled in Attouch et al. (2014, Theorem 3.3.3).

We now regroup some known properties of KL, in particular to show that Assumption (**A'**$_1$) holds for $\phi = \phi_e$.

Remark 5 (Properties of KL)**.** 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 $L^q(\mathcal{X})$ (Resmerita, 2005, Lemma 4.1). Regarding the use of $\phi = \phi_e$ in (8), one can for instance

[5] 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 $\mathcal{M}(\mathcal{X})$.

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 there 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 $\bar{\mu} = \rho$). KL is not Gâteaux-differentiable for non-finite $\mathcal{X}$ as recalled for instance in Butnariu and Resmerita (2006, p12) and Santambrogio (Remark 7.13 2015, p239).

As a follow-up of Remark 2, we now give some known conditions for a sum of subdifferentials to be the subdifferential of the sum.

Remark 6 (About the proof of convergence in Theorem 4)**.** Our proof of Theorem 4 resembles the one of Lu et al. (2018), which also relies on a three-point inequality as stated in Lemma 3. However, the proof of the latter inequality 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), but which is harder to derive in infinite dimensions. 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 (quasi) relative interior of $(\mathrm{dom}(\mathcal{G}) - \mathrm{dom}(D_\phi))$ (Borwein and Goebel, 2003); (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 when $\mathcal{F}$ and D_ϕ are chosen as the KL divergence (defined below in Example 2) in none of the spaces we consider since KL is not continuous, its domain being of empty interior. The other conditions are difficult to verify for given functionals. For instance, $\mathrm{dom}(\phi)$ is not explicit for the negative entropy (see Example 2). On the contrary, by favoring directional derivatives and first variations, we circumvent most of the difficulties related to (sub)differentiability.

C Additional technical results

Lemma 12. Let f be a proper function over a vector space $\mathcal{Y}$ with values in $\mathbb{R} \cup \{+\infty\}$. The following conditions are equivalent:

i) f is convex;

ii) $\mathrm{dom}(f)$ is convex, and, for all $x, y \in \mathrm{dom}(f)$, $d^+f(x)(y-x)$ exists, with value in $\mathbb{R} \cup \{-\infty\}$, and we have $f(x) + d^+f(x)(y-x) \leq f(y)$, i.e. $D_f(y|x) \geq 0$;

iii) $\mathrm{dom}(f)$ is convex, and, for all $x, y \in \mathrm{dom}(f)$, $d^+f(x)(y-x)$ exists, with value in $\mathbb{R} \cup \{-\infty\}$, and we have $d^+f(x)(y-x) + d^+f(y)(x-y) \leq 0$.

The lemma immediately extends to strictly convex functions by taking strict inequalities.

Proof. Given $x, y \in \mathrm{dom}(f)$, define for any $\lambda \in [0,1]$ $u_\lambda = x + \lambda(y-x)$. Assuming (i), then

$$f(u_\lambda) \leq \lambda f(y) + (1-\lambda) f(x)$$
$$f(x) + \frac{f(u_\lambda) - f(x)}{\lambda} \leq f(y),$$

which yields (ii) by the decreasingness of differential quotients discussed in Remark 1.

Assuming (ii), we just sum the two inequalities $(f(x) + d^+f(x)(y-x) \leq f(y))$ and $(f(x) + d^+f(x)(y-x) \leq f(y))$, to derive (iii).

The last implication to show is (iii)$\Rightarrow$ (i) which requires to perform an integration. Assume that (iii) holds, we want to show that $f(u_\lambda) \leq \lambda f(y) + (1-\lambda) f(x)$. Set $g(\lambda) := f(u_\lambda)$ and denote by $g'_+(\lambda)$ (resp. $g'_-(\lambda)$) its right (resp. left) derivative, both derivatives exist with value in $\mathbb{R} \cup \{-\infty\}$ for any $\lambda \in (0,1)$ since

$$d^+f(u_\lambda)(y-x) = \lim_{h \to 0^+} \frac{f(u_\lambda + h(y-x)) - f(x)}{h} = g'_+(\lambda)$$

similarly $d^+f(u_\lambda)(x-y) = -g'_-(\lambda)$. Consequently, for all $0 < \lambda_1 < \lambda_2 < 1$, applying (iii) to u_{λ_1} and u_{λ_2}, we have that $g'_+(\lambda_1) \leq g'_-(\lambda_2)$. We now show that $\lambda \mapsto g'_-(\lambda)$ is increasing over $(0,1)$.

We just have to show that $g'_-(\lambda_1) \leq \sup_{\lambda \in (0,\lambda_2)} g'_+(\lambda)$. By contradiction, we could fix $\lambda_1 \in (0, \lambda_2)$ and $\epsilon > 0$ such that, for all $\lambda \in (0, \lambda_2)$, $g'_-(\lambda_1) \geq g'_+(\lambda) + \epsilon$. By definition of the directional derivatives, we can then fix $\delta_1 \in (0, \lambda_1)$ and $\lambda \in (\lambda_1 - \delta_1, \lambda_1)$ such that for all $h_0 \in (0, \delta_1)$

$$|g'_-(\lambda_1) + \frac{g(\lambda_1 - h_0) - g(\lambda_1)}{h_0}| \leq \epsilon/4 \quad ; \quad |g'_+(\lambda) - \frac{g(\lambda_1) - g(\lambda)}{\lambda_1 - \lambda}| \leq \epsilon/4$$

whence

$$\frac{g(\lambda_1) - g(\lambda_1 - h_0)}{h_0} \geq \frac{g(\lambda_1) - g(\lambda)}{\lambda_1 - \lambda} + \epsilon/2$$

which leads to a contradiction for $h_0 = \lambda_1 - \lambda$. Therefore $\lambda \to g'_-(\lambda)$ is increasing over $(0, 1)$, upper bounded by $g'_-(y) = d^+f(u_1)(x - y)$. Since g has both left and right derivatives, it is continuous over $[0, 1]$. We can now apply (iii) to x and u_λ, use the positive homogeneity of the directional derivative (which always holds by definition), and integrate over $(0, 1)$ since the function g'_- is Riemann-integrable,

$$\begin{aligned}
0 &\geq d^+f(x)(u_\lambda - x) + d^+f(u_\lambda)(x - u_\lambda) \\
&= \lambda d^+f(x)(y - x) + \lambda d^+f(u_\lambda)(x - y) \\
0 &\geq d^+f(x)(y - x) - \int_0^1 g'_-(\lambda) d\lambda \\
&= d^+f(x)(y - x) + g(0) - g(1) \\
&= d^+f(x)(y - x) + f(x) - f(y),
\end{aligned}$$

which concludes the proof. □

Below, we derive some useful characterizations of relative smoothness, by analogy with Bauschke et al. (2017, Proposition 1) for differentiable functions in finite dimensions. Similar results hold for relative convexity by the same arguments.

Lemma 13. The following conditions are equivalent:

(i) $\mathcal{F}$ is L-smooth relative to ϕ over C;
(ii) $L\phi - \mathcal{F}$ is convex on $C \cap \mathrm{dom}(\phi) \cap \mathrm{dom}(\mathcal{F})$;

and, if the first variations of $\mathcal{F}$ and ϕ over C evaluated at $\mu, \nu \in C \cap \mathrm{dom}(\phi) \cap \mathrm{dom}(\mathcal{F})$ exist,

(iii) $\langle \nabla_C \mathcal{F}(\mu) - \nabla_C \mathcal{F}(\nu), \mu - \nu \rangle \leq L \langle \nabla_C \phi(\mu) - \nabla_C \phi(\nu), \mu - \nu \rangle$

Proof. This is a consequence of Lemma 12 applied to $\psi(\mu) = L\phi(\mu) - \mathcal{F}(\mu)$. More precisely, condition (i) can be written as $d^+\psi(\mu)(\nu - \mu) \leq \psi(\nu) - \psi(\mu)$ which is equivalent to the convexity of ψ by Lemma 12, hence (i)⇔(ii). Provided the first variations of $\mathcal{F}$ and ϕ over C exist, assuming (i) and (iii) boils down to Lemma 12-iii). Conversely, assuming (iii), we use Lemma 12-iii) and the linearity of the first variation (3). □

D Smoothness of the Maximum Mean Discrepancy relatively to the KL divergence

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive semi-definite kernel, $\mathcal{H}_k$ its corresponding Reproducing Kernel Hilbert Space (Steinwart and Christmann, 2008). The space $\mathcal{H}_k$ is a Hilbert space with inner product and norm $\| \cdot \|_{\mathcal{H}_k}$ satisfiying the reproducing property: for all $f \in \mathcal{H}_k$ and $x \in \mathcal{X}$, $f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}_k}$. For any $\mu \in \mathcal{P}(\mathcal{X})$ such that $\int \sqrt{k(x, x)} d\mu(x) < \infty$, the kernel mean embedding of μ, $m_\mu = \int k(x, \cdot) d\mu(x)$, is well-defined, belongs to $\mathcal{H}_k$, and $\mathbb{E}_\mu[f(X)] = \langle f, m_\mu \rangle_{\mathcal{H}_k}$ (Smola et al., 2007). The kernel k is said to be characteristic when such mean embedding is injective, that is, when any probability distribution is associated to a unique mean embedding. In this case, the kernel defines a distance between probability distributions referred to as the Maximum Mean Discrepancy (MMD), defined through the square norm of the difference between mean embeddings:

$$\mathrm{MMD}^2(\mu, \bar{\mu}) = \|m_\mu - m_{\bar{\mu}}\|^2_{\mathcal{H}_k}$$

Interestingly, as soon as the kernel is bounded, the MMD is relatively smooth with respect to ϕ_e, see Proposition 14 below. Notice that, thanks to the reproducing property, $\mu \mapsto \mathrm{MMD}(\mu, \bar{\mu})$ is strictly convex whenever the kernel k is characteristic, as it is the case for the Gaussian kernel. Similarly to KL, the MMD can be written as a Bregman divergence of $\phi_k(\mu) = \|m_\mu\|^2_{\mathcal{H}_k} = \int k(x,x')d\mu(x)d\mu(x')$.

Proposition 14. Let $\phi_e : \mu \mapsto \int \log(\mu)d\mu$ and fix $\nu \in \mathcal{P}(\mathcal{X})$. Take $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ to be a bounded semipositive definite kernel, i.e. $c_k = \sup_{x\in\mathcal{X}} k(x,x) < \infty$. The squared Maximum Mean Discrepancy $\mathrm{MMD}^2(\cdot, \nu)$ is $4c_k$-smooth relative to ϕ_e.

Proof. Let $\mu, \nu \in \mathcal{P}(\mathcal{X})$ and $f_{\mu,\bar{\mu}} = \int k(x,\cdot)d\mu(x) - \int k(x,\cdot)d\bar{\mu}(x) = \frac{1}{2}\nabla \mathrm{MMD}^2(\mu, \bar{\mu})$. We have:

$$\begin{aligned}\langle \nabla \mathrm{MMD}^2(\mu,\bar{\mu}) - \nabla \mathrm{MMD}^2(\nu,\bar{\mu}), \mu - \nu\rangle &\leq \|\nabla \mathrm{MMD}^2(\mu,\bar{\mu}) - \nabla \mathrm{MMD}^2(\nu,\bar{\mu})\|_\infty \|\mu - \nu\|_{TV} \\ &\leq 2\|f_{\mu,\bar{\mu}} - f_{\nu,\bar{\mu}}\|_\infty \|\mu - \nu\|_{TV} \\ &\leq 2 \sup_{y\in\mathcal{X}} |\int k(x,y)d\mu(x) - \int k(x,y)d\nu(x)| \|\mu - \nu\|_{TV}\end{aligned}$$

since by the reproducing property and Cauchy-Schwarz inequality, $k(x,y) = \langle k(x,\cdot), k(y,\cdot)\rangle \leq \|k(x,\cdot)\|_k \|k(y,\cdot)\|_k = \sqrt{k(x,x)k(y,y)} \leq c_k$, and $y \mapsto k(x,y)$ is measurable,

$$\begin{aligned}&\leq 2c_k \sup_{\substack{f:\mathcal{X}\to[-1,1] \\ f \text{mesurable}}} |\int f(x)d\mu(x) - \int f(x)d\nu(x)| \|\mu - \nu\|_{TV} \\ &\leq 2c_k \|\mu - \nu\|^2_{TV} \leq 4c_k(\mathrm{KL}(\mu|\nu) + \mathrm{KL}(\nu|\mu)) = 4c_k \langle \nabla\phi_e(\mu) - \nabla\phi_e(\nu), \mu - \nu\rangle,\end{aligned}$$

where the last inequality results from Pinsker's inequality. We conclude by using Lemma 13. □

Remark 7. (Case of neural network optimization). It is interesting to quantify the constant c_k for some kernels of interest, for instance when optimizing an infinite-width one hidden layer neural network as in Arbel et al. (2019). Consider a regression task where the labelled data $(z, y) \sim P$ where P denotes some fixed data distribution. For any input z, the output of a single hidden layer neural network parametrized by $w \in \mathcal{X}$ can be written:

$$f_w(z) = \frac{1}{N}\sum_{j=1}^N a_j \sigma(\langle b_j, z\rangle) = \int_{\mathcal{X}} \phi(z,w)d\mu(w),$$

where a_j and b_j denote output and input weights of neuron $j = 1, \ldots, N$ respectively, $w_j = (a_j, b_j)$ and $\mu = {}^1\!/_N \sum_{j=1}^N \delta_{w_j}$. In the infinite-width setting, the limiting risk in this regression setting is written for any distribution $\mu \in \mathcal{P}(\mathcal{X})$ on the weights as $\mathbb{E}_{(z,y)\sim P}[\|y - \int \phi(z,w)d\mu(w)\|^2]$. When the model is well-posed, i.e. there exists a distribution μ^* over weights such that $\mathbb{E}[y|z = \cdot] = \int \phi(\cdot, w)d\mu^*(w)$, then the limiting risk writes as an MMD with $k(w,w') = \mathbb{E}_{z\sim P}[\phi(z,w)^T\phi(z,w)]$ (Arbel et al. (2019, Proposition 20)). Hence, bounding $c_k = \sup_{w\in\mathcal{X}} k(w,w)$ depends on the choice of the activation function σ and on bounding the output weights. If σ is bounded (e.g. σ is the sigmoid activation) then bounding c_k corresponds to bounding the output weights. If σ is the RelU activation, then bounding c_k depends on bounding both input and output weights as well, and on the data distribution P.

E Related work - Optimization over measures using the Wasserstein geometry

In this section, we attempt to clarify the differences between the (Radon) vector space geometry considered in this paper and the Wasserstein geometry, developed in particular in Otto (2001); Villani (2003); Ambrosio et al. (2008).

Given an optimisation problem over $\mathcal{P}(\mathcal{X})$ the set of probability distributions over $\mathcal{X}$, one can consider different geometries over $\mathcal{P}(\mathcal{X})$. The one adopted in our paper casts $\mathcal{P}(\mathcal{X})$ as a subset of a

normed space of measures, such as $L^2(\rho)$ where ρ is a reference measure, or Radon measures. In this space, the shortest distance paths between measures are given by their square-norm distance. Moreover in this setting, one can consider the duality of measures with continuous functions and the mirror descent algorithm, as we do in this work.

In contrast, another possibility is to restrict $\mathcal{P}(\mathcal{X})$ to the probability distributions with bounded second moments, denoted $\mathcal{P}_2(\mathcal{X})$, equipped with Wasserstein-2 (W_2) distance. The space $(\mathcal{P}_2(\mathcal{X}), W_2)$, called the Wasserstein space, is a metric space equipped with a rich Riemannian structure (often referred to as "Otto calculus") where the shortest distance paths are given by the W_2 distance and associated geodesics. In this setting, one can leverage the Riemannian structure to discretize (W_2) gradient flows and consider algorithms such as (W_2) gradient descent, in analogy with Riemannian gradient descent.

While both frameworks yield optimisation algorithms on measure spaces, the geometries and algorithms are very different. Both the notion of convexity (along L^2 versus W_2 geodesics) and of gradients (first variation vs gradient of first variation) differ; and by extension so do many definitions. Consequently, the conditions needed for the convergence of mirror descent and W_2 gradient descent over an objective functional $\mathcal{F}$ greatly differ since they rely on the chosen geometry through the definitions of convexity, smoothness, or differentiability.

Wasserstein gradient descent should be thought of the analog of Riemannian gradient descent in infinite dimensions. Consequently, mirror descent yields updates on measures allowing for change of mass (see Lemma 2), while W_2 gradient descent preserves the mass, since the updates on measures write as pushforwards (i.e., displacement of particles supporting the measures). To summarize, the mirror descent scheme we consider is very different in nature to the gradient descent schemes based on the Wasserstein geometry (e.g. Chizat and Bach (2018); Mei et al. (2018); Rotskoff and Vanden-Eijnden (2018); Wibisono (2018); Korba et al. (2020); Salim et al. (2020); Korba et al. (2021)), due to the different geometry.

F Proofs

F.1 Proof of Theorem 4

Proof. Since $\mathcal{F}$ is L-smooth relative to ϕ over $\mathcal{R}$ and we assumed that $(\mu_n)_{n\in\mathbb{N}} \in \mathcal{R}^{\mathbb{N}}$, we have

$$\mathcal{F}(\mu_{n+1}) \leq \mathcal{F}(\mu_n) + d^+\mathcal{F}(\mu_n)(\mu_{n+1} - \mu_n) + LD_\phi(\mu_{n+1}|\mu_n). \tag{36}$$

Applying Lemma 3 to the convex function $\mathcal{G}_n(\nu) = \frac{1}{L} d^+\mathcal{F}(\mu_n)(\nu - \mu_n)$, with $\mu = \mu_n$ and $\bar{\nu} = \mu_{n+1}$ yields

$$d^+\mathcal{F}(\mu_n)(\mu_{n+1} - \mu_n) + LD_\phi(\mu_{n+1}|\mu_n) \leq d^+\mathcal{F}(\mu_n)(\nu - \mu_n) + LD_\phi(\nu|\mu_n) - LD_\phi(\nu|\mu_{n+1}).$$

Fix $\nu \in \mathcal{R}$, then (36) becomes:

$$\mathcal{F}(\mu_{n+1}) \leq \mathcal{F}(\mu_n) + d^+\mathcal{F}(\mu_n)(\nu - \mu_n) + LD_\phi(\nu|\mu_n) - LD_\phi(\nu|\mu_{n+1}). \tag{37}$$

This shows in particular, by substituting $\nu = \mu_n$ and since $D_\phi(\nu|\mu_{n+1}) \geq 0$, that $\mathcal{F}(\mu_{n+1}) \leq \mathcal{F}(\mu_n)$, i.e. $\mathcal{F}$ is decreasing at each iteration. Since $\mathcal{F}$ is l-strongly convex relative to ϕ, we also have:

$$d^+\mathcal{F}(\mu_n)(\nu - \mu_n) \leq \mathcal{F}(\nu) - \mathcal{F}(\mu_n) - lD_\phi(\nu|\mu_n) \tag{38}$$

and (37) becomes:

$$\mathcal{F}(\mu_{n+1}) \leq \mathcal{F}(\nu) + (L - l)D_\phi(\nu|\mu_n) - LD_\phi(\nu|\mu_{n+1}). \tag{39}$$

By induction, similarly to Lu et al. (2018), we sum (39) over n, obtaining

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

Using the monotonicity of $(\mathcal{F}(\mu_n))_{n\geq 0}$ and the positivity of $D_\phi(\nu|\mu_n)$, we have

$$\sum_{i=1}^{n} \left(\frac{L}{L-l}\right)^i (\mathcal{F}(\mu_n) - \mathcal{F}(\nu)) \leq LD_\phi(\nu|\mu_0) - L\left(\frac{L}{L-l}\right)^n D_\phi(\nu|\mu_n) \leq LD_\phi(\nu|\mu_0). \quad \square$$

F.2 Proof of Proposition 5

Let $\pi, \bar{\pi} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, $h > 0$ and $\xi = \bar{\pi} - \pi$ hence any integral with respect to ξ of constant functions is null. We have:

$$\begin{aligned}
F_{\mathrm{S}}(\pi + h\xi) - F_{\mathrm{S}}(\pi) &= \mathrm{KL}(p_{\mathcal{X}}(\pi + h\xi)|\bar{\mu}) - \mathrm{KL}(p_{\mathcal{X}}\pi|\bar{\mu}) \\
&= \int \log\left(p_{\mathcal{X}}\pi + hp_{\mathcal{X}}\xi\right) dp_{\mathcal{X}}(\pi + h\xi) - \int \log(\bar{\mu}) dp_{\mathcal{X}}(\pi + h\xi) - \int \log\left(\frac{p_{\mathcal{X}}\pi}{\bar{\mu}}\right) dp_{\mathcal{X}}\pi \\
&= h\int \log\left(\frac{p_{\mathcal{X}}\pi}{\bar{\mu}}\right) dp_{\mathcal{X}}\xi + \underbrace{\int \log\left(1 + h\frac{p_{\mathcal{X}}\xi}{p_{\mathcal{X}}\pi}\right) dp_{\mathcal{X}}\pi}_{\approx h\int \frac{p_{\mathcal{X}}\xi}{p_{\mathcal{X}}\pi} dp_{\mathcal{X}}\pi + o(h) = 0 + o(h)} + \underbrace{h\int \log\left(1 + h\frac{p_{\mathcal{X}}\xi}{p_{\mathcal{X}}\pi}\right) dp_{\mathcal{X}}\xi}_{\approx h^2\int \frac{p_{\mathcal{X}}\xi}{p_{\mathcal{X}}\pi} dp_{\mathcal{X}}\xi + o(h^2)} \\
&= h\int \log\left(\frac{p_{\mathcal{X}}\pi}{\bar{\mu}}\right) dp_{\mathcal{X}}\xi + o(h).
\end{aligned}$$

Consequently,

$$\lim_{h \to 0^+} \frac{F_{\mathrm{S}}(\pi + h\xi) - F_{\mathrm{S}}(\pi)}{h} = \int_{\mathcal{X}} \log\left(\frac{p_{\mathcal{X}}\pi}{\bar{\mu}}\right) d\xi = \int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{p_{\mathcal{X}}\pi}{\bar{\mu}}\right) d\xi. \tag{40}$$

Hence, when it exists, $\nabla_C F_{\mathrm{S}}(\pi) = \ln(dp_{\mathcal{X}}\pi/d\bar{\mu})$. Moreover, the sets $\Pi(*, \bar{\nu})$ and $\Pi(\bar{\mu}, *)$ are L^∞-weak-* closed.[6] Besides, KL has weak-* compact sublevel sets and is weak-* l.s.c. Hence Attouch et al. (2014, Theorem 3.2.2) applies, and the iterates $(\pi_n)_{n\geq 0}$ exist, as originally shown by Csiszar (1975). As $\pi_n = e^{(f+g-c)/\epsilon}\bar{\mu} \otimes \bar{\nu}$ (Nutz, 2021, Section 6) with $f \in L^\infty(\mathcal{X})$ and $g \in L^\infty(\mathcal{Y})$, we have that $x \mapsto \ln(d\mu_n(x)/d\bar{\mu}(x)) \in L^\infty(\mathcal{X}, \mathbb{R})$; indeed as $c \in L^\infty$, the first marginal μ_n of π_n is an integral of functions bounded by strictly positive quantities.

Consider a coupling $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ with $\pi \ll \pi_n$ and denote by μ its first marginal. We have that $F_{\mathrm{S}}(\pi_n) = \int_{\mathcal{X}} \ln(\mu_n/\bar{\mu})\, d\mu_n$ and $\langle \nabla_C F_{\mathrm{S}}(\pi_n), \pi - \pi_n\rangle = \iint \ln(d\mu_n(x)/d\bar{\mu}(x))(\pi(dx, dy) - \pi_n(dx, dy))$. Simplifying and using (14) twice we obtain the identity:

$$\begin{aligned}
&F_{\mathrm{S}}(\pi_n) + \langle \nabla_C F_{\mathrm{S}}(\pi_n), \pi - \pi_n\rangle + \mathrm{KL}(\pi|\pi_n) \\
&= \int \ln(d\mu_n/d\bar{\mu}(x))\mu_n(dx) + \iint \ln(d\mu_n/d\bar{\mu}(x))\pi(dx, dy) \\
&\qquad - \iint \ln(d\mu_n/d\bar{\mu}(x))\pi_n(dx, dy) + \mathrm{KL}(\pi|\pi_n) \\
&= \int \ln(d\mu_n/d\bar{\mu}(x))\mu(dx) + \mathrm{KL}(p_{\mathcal{X}}\pi|\mu_n) + \mathrm{KL}(\pi|p_{\mathcal{X}}\pi \otimes {}^{\pi_n}/_{\mu_n}) \\
&= \int \ln(d\mu/d\bar{\mu}(x))\mu(dx) + \mathrm{KL}(\pi|p_{\mathcal{X}}\pi \otimes {}^{\pi_n}/_{\mu_n}) = \mathrm{KL}(\pi|\bar{\mu} \otimes {}^{\pi_n}/_{\mu_n}) = \mathrm{KL}(\pi|\pi_{n+\frac{1}{2}}).
\end{aligned}$$

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

F.3 Proof of Proposition 7

The proof of Proposition 7 essentially relies on bounding the entropic potentials by the marginals, as in Luise et al. (2019)[Theorem C.4]. For their purpose Luise et al. (2019) assume that $\mathcal{X} = \mathcal{Y}$ and that $c(x, y) = c(y, x)$. These assumptions are not needed here so we revisit their proof and show that their bound holds for general bounded costs. We define

$$D_c = \frac{1}{2}\sup[c(x, y) + c(x', y') - c(x, y') - c(x', y)],$$

where the supremum runs over $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$. When μ and ν are probability measures on $\mathcal{X}$ and $\mathcal{Y}$ respectively we define the soft-c transform mappings $T_\mu \colon L^\infty(\mathcal{X}) \to L^\infty(\mathcal{Y})$ and $T_\nu \colon L^\infty(\mathcal{Y}) \to L^\infty(\mathcal{X})$ by

$$T_\mu(f)(y) = -\epsilon \ln\Big(\int_{\mathcal{X}} e^{(f(x) - c(x,y))/\epsilon}\mu(dx)\Big)$$

[6] Indeed, take $(\pi_n)_{n\in N} \in \Pi(*, \bar{\nu})$ converging weakly to some $\bar{\pi}$. As $\langle g, \pi_n\rangle_{\mathcal{X}\times\mathcal{Y}} = \langle g, \bar{\nu}\rangle_{\mathcal{Y}}$ for all $g \in L^\infty(\mathcal{Y}, \mathbb{R})$, we obtain that $\langle g, \bar{\pi}\rangle_{\mathcal{X}\times\mathcal{Y}} = \langle g, \bar{\nu}\rangle_{\mathcal{Y}}$ which precisely means that $p_{\mathcal{Y}}\bar{\pi} = \bar{\nu}$.

and

$$T_\nu(g)(x) = -\epsilon \ln\Big(\int_{\mathcal{Y}} e^{(g(y)-c(x,y))/\epsilon}\nu(dy)\Big).$$

These mappings arise naturally in the context of Sinkhorn's algorithm since if $\pi \in \Pi_c$ has marginals (μ,ν), we can write $\pi(dx,dy) = e^{(f(x)+g(y)-c(x,y))/\epsilon}\mu(dx)\nu(dy)$ and taking marginals implies

$$g = T_\mu(f) \quad \text{and} \quad f = T_\nu(g). \tag{41}$$

Luise et al. (2019) use the *Hilbert metric* to prove their result, a classical tool to analyze matrix scaling problems (Franklin and Lorenz, 1989), which for our purpose here reduces to the following semi-norm.

Definition 7. When $f \in L^\infty(\mathcal{X})$ we set $\|f\|_{\text{var}} = (\sup_{\mathcal{X}} f) - (\inf_{\mathcal{X}} f)$. We similarly define $\|g\|_{\text{var}}$ for $g \in L^\infty(\mathcal{Y})$.

We are now ready to state our version of Luise et al. (2019)[Theorem C.4].

Proposition 15. Let $\pi, \tilde{\pi} \in \Pi_c$ with marginals (μ,ν) and $(\tilde{\mu},\tilde{\nu})$ respectively. Write $\pi = e^{(f+g-c)/\epsilon}\mu\otimes\nu$ and $\tilde{\pi} = e^{(\tilde{f}+\tilde{g}-c)/\epsilon}\tilde{\mu}\otimes\tilde{\nu}$. Then

$$\|f-\tilde{f}\|_{\text{var}} + \|g-\tilde{g}\|_{\text{var}} \le 2\epsilon\, e^{3D_c/\epsilon}\big(\|\mu-\tilde{\mu}\|_{\text{TV}} + \|\nu-\tilde{\nu}\|_{\text{TV}}\big).$$

The proof of this quantitative stability estimate mainly relies on the classical result that the soft c-transform mappings are contractions in the Hilbert metric; this result is at the heart of the proof of the classical linear convergence rate of Sinkhorn (see Franklin and Lorenz, 1989; Chen et al., 2016, for a proof).

Proposition 16. $\|T_\mu(\tilde{f}) - T_\mu(f)\|_{\text{var}} \le \lambda\|\tilde{f}-f\|_{\text{var}}$ with $\lambda = \frac{e^{D_c/\epsilon}-1}{e^{D_c/\epsilon}+1} < 1$.

We will also need the following lemma which is essentially contained in Luise et al. (2019).

Lemma 17. Let $f = T_\nu(g)$ for some $g \in L^\infty(\mathcal{Y})$. Then

$$\|T_{\tilde{\mu}}(f) - T_\mu(f)\|_{\text{var}} \le 2\epsilon\, e^{2D_c/\epsilon}\|\mu-\tilde{\mu}\|_{\text{TV}}.$$

Likewise if $g = T_\mu(f)$ for some $f \in L^\infty(\mathcal{X})$,

$$\|T_{\tilde{\nu}}(g) - T_\nu(g)\|_{\text{var}} \le 2\epsilon\, e^{2D_c/\epsilon}\|\nu-\tilde{\nu}\|_{\text{TV}}.$$

Proof of Lemma 17. For any $f \in L^\infty(\mathcal{X})$ we have by definition

$$T_{\tilde{\mu}}(f)(y) - T_\mu(f)(y) = \epsilon\log\Big(\int_{\mathcal{X}} e^{(f(x)-c(x,y))/\epsilon}\mu(dx)\Big) - \epsilon\log\Big(\int_{\mathcal{X}} e^{(f(x)-c(x,y))/\epsilon}\tilde{\mu}(dx)\Big).$$

To control this difference of logs, Luise et al. (2019)[Lemma C.2] use the bound $|\log(a)-\log(b)| \le \max\{a^{-1},b^{-1}\}|a-b|$ (for any $a,b>0$). We have $\int_{\mathcal{X}} e^{(f(x)-c(x,y))/\epsilon}\mu(dx) \ge e^{\inf_x[f(x)-c(x,y)]/\epsilon}$ and the same lower bound holds for $\int_{\mathcal{X}} e^{(f(x)-c(x,y))/\epsilon}\tilde{\mu}(dx)$. Therefore

$$\begin{aligned}|T_{\tilde{\mu}}(f)(y) - T_\mu(f)(y)| &\le \epsilon\, e^{-\inf_x[f(x)-c(x,y)]/\epsilon}\int_{\mathcal{X}} e^{(f(x')-c(x',y))/\epsilon}|\mu-\tilde{\mu}|(dx')\\ &\le \epsilon\, e^{\sup_x[c(x,y)-f(x)]/\epsilon}e^{\sup_{x'}[f(x')-c(x',y)]/\epsilon}\|\mu-\tilde{\mu}\|_{\text{TV}}.\end{aligned}$$

This implies when taking the supremum over $y \in \mathcal{Y}$

$$\|T_{\tilde{\mu}}(f) - T_\mu(f)\|_{\text{var}} \le 2\|T_{\tilde{\mu}}(f) - T_\mu(f)\|_\infty \le 2\epsilon\, e^{\sup_{x,x',y}[f(x')-f(x)+c(x,y)-c(x',y)]/\epsilon}\|\mu-\tilde{\mu}\|_{\text{TV}}.$$

This last inequality is valid for any $f \in L^\infty(\mathcal{X})$. If in addition we take f to be an image $f = T_\nu(g)$, then we have the standard estimate for any given $x, x' \in \mathcal{X}$ and $y \in \mathcal{Y}$

$$\begin{aligned}-f(x) &= \epsilon\ln\Big(\int_{\mathcal{Y}} e^{(g(y')-c(x,y'))/\epsilon}\nu(dy')\Big)\\ &= \epsilon\ln\Big(\int_{\mathcal{Y}} e^{(c(x,y)+c(x',y')-c(x',y)-c(x,y'))/\epsilon}e^{(g(y')-c(x',y'))/\epsilon}\nu(dy')\Big) + c(x',y) - c(x,y)\\ &\le 2D_c + \epsilon\ln\Big(\int_{\mathcal{Y}} e^{(g(y')-c(x',y'))/\epsilon}\nu(dy')\Big) + c(x',y) - c(x,y)\\ &= 2D_c - f(x') + c(x',y) - c(x,y).\end{aligned}$$

This shows that $\sup_{x,x',y}[f(x') - f(x) + c(x,y) - c(x',y)] \le 2D_c$. As a consequence,

$$\|T_{\tilde\mu}(f) - T_\mu(f)\|_{\text{var}} \le 2\epsilon\, e^{2D_c/\epsilon}\|\mu - \tilde\mu\|_{\text{TV}}.$$

By symmetry the corresponding bound can be derived for quantities on $\mathcal{Y}$. □

Proof of Proposition 15. Having in mind the fixed point equations (41) for (f,g) and the corresponding ones for $(\tilde f, \tilde g)$ we write

$$\begin{aligned}\|\tilde f - f\|_{\text{var}} &= \|T_{\tilde\nu}(\tilde g) - T_\nu(g)\|_{\text{var}}\\ &\le \|T_{\tilde\nu}(\tilde g) - T_{\tilde\nu}(g)\|_{\text{var}} + \|T_{\tilde\nu}(g) - T_\nu(g)\|_{\text{var}},\end{aligned}$$

and similarly, $\|\tilde g - g\|_{\text{var}} \le \|T_{\tilde\mu}(\tilde f) - T_{\tilde\mu}(f)\|_{\text{var}} + \|T_{\tilde\mu}(f) - T_\mu(f)\|_{\text{var}}$. By Proposition 16, $\|T_{\tilde\mu}(\tilde f) - T_{\tilde\mu}(f)\|_{\text{var}} \le \lambda\|\tilde f - f\|_{\text{var}}$ and $\|T_{\tilde\nu}(\tilde g) - T_{\tilde\nu}(g)\|_{\text{var}} \le \lambda\|\tilde g - g\|_{\text{var}}$. Combining, we obtain

$$(1-\lambda)\big(\|\tilde f - f\|_{\text{var}} + \|\tilde g - g\|_{\text{var}}\big) \le \|T_{\tilde\mu}(f) - T_\mu(f)\|_{\text{var}} + \|T_{\tilde\nu}(g) - T_\nu(g)\|_{\text{var}}.$$

Lemma 17 takes care of the right-hand side, and this results in

$$(1-\lambda)\big(\|\tilde f - f\|_{\text{var}} + \|\tilde g - g\|_{\text{var}}\big) \le 2\epsilon\, e^{2D_c/\epsilon}\big(\|\mu - \tilde\mu\|_{\text{TV}} + \|\nu - \tilde\nu\|_{\text{TV}}\big).$$

Finally we divide by $1-\lambda$ and bound $(1-\lambda)^{-1} = (e^{D_c/\epsilon}+1)/2 \le e^{D_c/\epsilon}$. □

Proof of Proposition 7. Let $\pi, \tilde\pi \in \Pi_c$ with marginals $(\mu,\bar\nu)$ and $(\tilde\mu,\bar\nu)$ respectively. Write $\pi = e^{(f+g-c)/\epsilon}\mu\otimes\bar\nu$ and $\tilde\pi = e^{(\tilde f+\tilde g-c)/\epsilon}\tilde\mu\otimes\bar\nu$. We emphasize that π and $\tilde\pi$ have the same second marginal $\bar\nu$. Then

$$\begin{aligned}&\epsilon\,\mathrm{KL}(\tilde\pi|\pi) = \epsilon\,\mathrm{KL}(\tilde\pi|\pi) + \epsilon\,\mathrm{KL}(\pi|\tilde\pi) - \epsilon\,\mathrm{KL}(\pi|\tilde\pi)\\ &= \iint(\tilde f - f + \tilde g - g + \epsilon\ln\Big(\frac{d\tilde\mu}{d\mu}\Big))\,d\tilde\pi + \iint(f - \tilde f + g - \tilde g + \epsilon\ln\Big(\frac{d\mu}{d\tilde\mu}\Big))d\pi - \epsilon\,\mathrm{KL}(\pi|\tilde\pi)\\ &= \iint(\tilde f - f + \tilde g - g)\,(d\tilde\pi - d\pi) + \epsilon\,\mathrm{KL}(\tilde\mu|\mu) + \epsilon\,\mathrm{KL}(\mu|\tilde\mu) - \epsilon\,\mathrm{KL}(\pi|\tilde\pi).\end{aligned}$$

Part of the first term vanishes since $\iint(\tilde g - g)\,(d\tilde\pi - d\pi) = \int_{\mathcal{Y}}(\tilde g - g)\,(d\bar\nu - d\bar\nu) = 0$, and we can get rid of the last two terms by using the data processing inequality $\mathrm{KL}(\mu|\tilde\mu) \le \mathrm{KL}(\pi|\tilde\pi)$. Thus

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

Applying Proposition 15 we obtain

$$\epsilon\,\mathrm{KL}(\tilde\pi|\pi) \le 2\epsilon e^{3D_c/\epsilon}\|\tilde\mu - \mu\|_{\text{TV}}^2 + \epsilon\,\mathrm{KL}(\tilde\mu|\mu),$$

and after dividing by ϵ, Pinsker's inequality yields

$$\mathrm{KL}(\tilde\pi|\pi) \le (1 + 4e^{3D_c/\epsilon})\,\mathrm{KL}(\tilde\mu|\mu).$$ □

F.4 Proof of Proposition 9

Proposition 18 (EM as mirror descent)**.** Let $C = \Pi(*,\bar\nu)$. Assume that for all $\pi \in C$ there exists a $q_*(\pi) \in \mathcal{Q}$ solving (28), that, for $p_h = p_{q_*((1-h)\pi_n + h\pi)}$, ${}^{d\pi_n}/_{dp_h}$ converges pointwise to ${}^{d\pi_n}/_{dp_{q_n}}$ for $h \to 0^+$ with $|\ln({}^{d\pi_n}/_{dp_h})| \le \mathcal{G}_n$ for some $\mathcal{G}_n \in L^1(\pi + \pi_n)$, that ${}^{d\pi_n}/_{dp_{q_n}}(\cdot,\cdot) \in [a_n, b_n]$ for some $a_n > 0$ and $b_n > 0$, and that $\sup_{q\in\mathcal{Q}}|\ln({}^{d\pi_n}/_{dp_q})| < \infty$. Then the EM iterations (26)–(27) can be written as a mirror descent iteration with objective function F_{EM}, Bregman potential ϕ_e and constraints C,

$$\pi_{n+1} = \operatorname*{argmin}_{\pi\in C}\langle\nabla_C F_{\text{EM}}(\pi_n), \pi - \pi_n\rangle + \mathrm{KL}(\pi|\pi_n), \tag{42}$$

with $\nabla_C F_{\text{EM}}(\pi_n) = \ln(d\pi_n/dp_{q_n}) \in L^\infty(\mathcal{X}\times\mathcal{Y})$.

Remark 8. Note that our assumptions on the sequence $(p_h)_{h\in[0,1]}$ in Proposition 18 are very similar to what the fundamental theorem of Γ-convergence would provide (see Dal Maso (1987), Braides (2002, Theorem 2.10)). It is indeed straightforward to prove Γ-convergence (see Braides, 2002, Theorem 2.1) of the sequence $(f_{n,\pi}(\cdot,h))_{h\in[0,1]}$ in $h=0^+$ with $f_{n,\pi}(p,h) := \mathrm{KL}(\pi_n + h(\pi-\pi_n)|p)$, owing to the convexity and joint weak-* lower semicontinuity of KL. However, to prove the convergence of the sequence of minimizers $(p_h)_{h\in[0,1]}$, one would need the equicoercivity of $(\mathrm{KL}(\pi_h|p))_{h\in[0,1]}$ over $p\in\mathcal{P}_Q$ Braides (2002, Definition 2.9)), which heavily depends on the properties of $\mathcal{P}_Q$, e.g. considering a weak-* compact $\mathcal{P}_Q$ would entail equicoercivity.

Proof. We will use here the envelope theorem to differentiate F_{EM} and compute its first variation. We are going to apply Milgrom and Segal (2002, Theorem 3) leveraging properties of KL. Milgrom and Segal (2002, Theorem 3) is written for the set $[0,1]\times X$, where X is some set optimized over. Here $X=\mathcal{P}_Q := \{p_q \mid q\in\mathcal{Q}\}$ and the interval $[0,1]$ will be merely the scalar of the directional derivative we consider.

Let $n\geq 0$ and $\pi\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})$. For $h\in[0,1]$, set $f_{n,\pi}(p,h) := \mathrm{KL}(\pi_n + h(\pi-\pi_n)|p)$ and $V_{n,\pi}(h) = \inf_{p\in\mathcal{P}_Q}\mathrm{KL}(\pi_n + h(\pi-\pi_n)|p)$ to match the notations of Milgrom and Segal (2002, Theorem 3). We have to show some equidifferentiability over $q\in\mathcal{P}_Q$. Notice that the following expression does not depend on p,

$$\frac{1}{h}\left[\iint \ln\left(\frac{\pi_n + h(\pi-\pi_n)}{p}\right) d(\pi_n + h(\pi-\pi_n)) - \iint \ln\left(\frac{\pi_n}{p}\right) d\pi_n\right] - \iint \ln\left(\frac{\pi_n}{p}\right) d(\pi-\pi_n)$$
$$= \frac{1}{h}\iint \ln\left(1 + h\frac{(\pi-\pi_n)}{\pi_n}\right) d\pi_n = \frac{1}{h}[h\iint d(\pi-\pi_n) + O(h^2)] = 0 + O(h),$$

so that we do have equidifferentiability when $h\to 0^+$. Our assumptions then allow to apply Milgrom and Segal (2002, Theorem 3). We thus obtain that

$$d^+F_{\text{EM}}(\pi_n)(\pi-\pi_n) = d^+V_{n,\pi}(0) = \lim_{h\to 0^+}\iint \ln\left(\frac{d\pi_n}{dp_h}\right) d\pi_n.$$

Since $|\ln(d\pi_n/dp_h)| \leq \mathcal{G}_n \in L^1(\pi+\pi_n)$ and $d\pi_n/dp_h$ converges pointwise to $d\pi_n/dp_{q_n}$ for $h\to 0^+$ (recall that $q_n = q_*(\pi_n)$ by definition), we can apply the dominated convergence theorem to interchange the limit and the integral. Consequently $d^+F_{\text{EM}}(\pi_n)(\pi-\pi_n) = \iint \ln(d\pi_n/dp_{q_n})d(\pi-\pi_n)$ proving that $\nabla F_{\text{EM}}(\pi_n) = \ln(d\pi_n/dp_{q_n}) \in L^\infty$ since $\pi_n/p_{q_n}(\cdot,\cdot)\in[a_n,b_n]$ for some $a_n>0$ and $b_n>0$.

Then, for π_n the EM iterate at time n, and for any coupling π, we have the identity:

$$F_{\text{EM}}(\pi_n) + \langle\nabla_C F_{\text{EM}}(\pi_n), \pi-\pi_n\rangle + \mathrm{KL}(\pi|\pi_n)$$
$$= \int \ln(d\pi_n/dp_{q_n}(x))\pi_n(dx) + \int \ln(d\pi_n/dp_{q_n}(x))\,(\pi-\pi_n)(dx) + \int \ln(d\pi/d\pi_n(x))\pi(dx)$$
$$= \int \ln(d\pi/dp_{q_n}(x))\pi(dx) = \mathrm{KL}(\pi|p_{q_n}).$$

Note that q_n is optimal in (28), whence (29) matches (27). □

F.5 Proof of Proposition 10

Proof. Let $\pi, \bar{\pi} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, $h > 0$ and $\xi = \bar{\pi} - \pi$, so $\iint_{\mathcal{X}\times\mathcal{Y}} \xi(dx, dy) = 0$. We have:

$$F_{\text{LEM}}(\pi + h\xi) - F_{\text{LEM}}(\pi) = \text{KL}(\pi + h\xi | p_{\mathcal{X}}(\pi + h\xi) \otimes K) - \text{KL}(\pi | p_{\mathcal{X}}\pi \otimes K)$$

$$= \int \log\left(\frac{\pi + h\xi}{p_{\mathcal{X}}(\pi + h\xi) \otimes K}\right) d(\pi + h\xi) - \int \log\left(\frac{\pi}{p_{\mathcal{X}}\pi \otimes K}\right) d\pi$$

$$= \int \log(\pi + h\xi)\, d(\pi + h\xi) - \int \log(p_{\mathcal{X}}(\pi + h\xi) \otimes K)\, d(\pi + h\xi) - \int \log(\pi)\, d\pi + \int \log(p_{\mathcal{X}}\pi \otimes K)\, d\pi$$

$$= h \int \log \pi d\xi + \int \underbrace{\log\left(1 + h\frac{\xi}{\pi}\right) d\pi}_{\approx h \int \frac{\xi}{\pi} d\pi + o(h) = 0 + o(h)} + h \underbrace{\int \log\left(1 + h\frac{\xi}{\pi}\right) d\xi}_{\approx h^2 \int \frac{\xi}{\pi} d\xi + o(h^2)}$$

$$- h \int \log(p_{\mathcal{X}}\pi \otimes K) d\xi - \underbrace{\int \log\left(1 + h\frac{p_{\mathcal{X}}\xi \otimes K}{p_{\mathcal{X}}\pi \otimes K}\right) d\pi}_{\approx h \int \frac{p_{\mathcal{X}}\xi \otimes K}{p_{\mathcal{X}}\pi \otimes K} d\pi + o(h) = h \int \frac{p_{\mathcal{X}}\xi}{p_{\mathcal{X}}\pi} d\pi + o(h) = 0 + o(h)} - h \underbrace{\int \log\left(1 + h\frac{p_{\mathcal{X}}\xi \otimes K}{p_{\mathcal{X}}\pi \otimes K}\right) d\xi}_{h^2 \int \frac{p_{\mathcal{X}}\xi \otimes K}{p_{\mathcal{X}}\pi \otimes K} d\xi + o(h^2)}$$

$$= h \int \log\left(\frac{\pi}{p_{\mathcal{X}}\pi \otimes K}\right) d\xi + o(h).$$

Hence

$$\lim_{h \to 0^+} \frac{F_{\text{LEM}}(\pi + h\xi) - F_{\text{LEM}}(\pi)}{h} = \int \log\left(\frac{\pi}{p_{\mathcal{X}}\pi \otimes K}\right) d\xi.$$

To show that $\nabla F_{\text{LEM}}(\pi_n)$ belongs to L^∞, we proceed by induction. Let $n \geq 0$ and assume that $T_K\mu_n \gg \bar{\nu}$ and $\mu_n = e^{f_n(x)}\bar{\mu}$ with f_n bounded (which we explicitly assumed for μ_0) then the multiplicative update (32) shows that f_{n+1} has the same property. Furthermore (32) gives

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

Since $K(x, dy) = e^{-c(x,y)}\bar{\nu}$ with c uniformly bounded, $\frac{\pi_{n+1}}{\mu_{n+1} \otimes K}(\cdot)$ is also bounded above and below by positive constants (depending on n). □

F.6 Proof of Proposition 11

Proof. By the disintegration formula (14),

$$F_{\text{LEM}}(\pi) = \text{KL}(\bar{\nu} | p_{\mathcal{Y}}(p_{\mathcal{X}}\pi \otimes K)) + \int \text{KL}(\pi/\bar{\nu} | (p_{\mathcal{X}}\pi \otimes K)/p_{\mathcal{Y}}(p_{\mathcal{X}}\pi \otimes K)) d\bar{\nu} \tag{43}$$

Let $\pi_* = \mu_*(dx)k(x, dy)\bar{\nu}(dy)/(T_K\mu_*)(dy)$. First, for any $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, we have $p_{\mathcal{Y}}(p_{\mathcal{X}}\pi \otimes K) = \int p_{\mathcal{X}}\pi(dx)k(x, \cdot) = T_K(p_{\mathcal{X}}\pi)$, hence by definition of $\mu^* \in \text{argmin}_\mu \text{KL}(\bar{\nu} | T_k\mu)$, π_* minimizes the first term in (43). Second, this choice leads to $\pi_*/\bar{\nu} = \mu_* \otimes K/T_K\mu^*$, cancelling the nonnegative second term in (43). Hence π_* is a minimizer of F_{LEM} and $F_{\text{LEM}}(\pi_*) = \text{KL}(\bar{\nu} | T_K\mu_*)$. Moreover, $\text{KL}(\cdot|\cdot)$ is convex in both arguments, and $\pi \mapsto p_{\mathcal{X}}\pi \otimes K$ is linear. Consequently the composition F_{LEM} is convex in π and so is F_{S} by the same arguments (see also Lemma 6). By (14) and linearity of the Bregman divergence, $\text{KL}(\pi|\tilde{\pi}) = D_{F_{\text{S}}}(\pi|\tilde{\pi}) + D_{F_{\text{LEM}}}(\pi|\tilde{\pi})$, hence F_{LEM} is 1-relatively smooth w.r.t. ϕ_e. Hence, Theorem 4 yields:

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

Since $\pi_0 \in \Pi(*, \bar{\nu})$, $\pi_0 = \mu_0(dx)k(x, dy)\bar{\nu}(dy)/(T_K\mu_*)(dy)$,

$$\text{KL}(\pi_*|\pi_0) = \text{KL}(\mu_*|\mu_0) + \iint \ln\left(\frac{k(x, dy)\bar{\nu}(dy)/(T_K\mu_*)(dy)}{k(x, dy)\bar{\nu}(dy)/(T_K\mu_0)(dy)}\right) \pi_*(dx, dy)$$

$$= \text{KL}(\mu_*|\mu_0) + \int_{\mathcal{Y}} \ln\left(\frac{\bar{\nu}(dy)/(T_K\mu_*)(dy)}{\bar{\nu}(dy)/(T_K\mu_0)(dy)}\right) \bar{\nu}(dy) = \text{KL}(\mu_*|\mu_0) + \text{KL}(\bar{\nu}|T_K\mu_*) - \text{KL}(\bar{\nu}|T_K\mu_0).$$

Finally, we use the inequality

$$\mathrm{KL}(\bar{\nu}|T_K\mu_n) = \mathrm{KL}(\pi_n|p_{\mathcal{Y}}(p_{\mathcal{X}}\pi_n \otimes K)) \leq \mathrm{KL}(p_{\mathcal{Y}}\pi_n|p_{\mathcal{X}}\pi_n \otimes K) = F_{\mathrm{LEM}}(\pi_n).$$

□