A Standard results

Before detailing our proofs, we first recall some well-known results regarding the Kullback-Leibler divergence and the SoftMin operator defined in (12).

A.1 The Kullback-Leibler divergence

First properties. For any pair of Radon measures $\alpha, \beta \in \mathcal{M}^+(\mathcal{X})$ on the compact metric set $(\mathcal{X}, \mathrm{d})$, the Kullback-Leibler divergence is defined through

$$\mathrm{KL}(\alpha,\beta) \overset{\text{def.}}{=} \begin{cases} \langle \alpha,\, \log \frac{\mathrm{d}\alpha}{\mathrm{d}\beta} - 1 \rangle + \langle \beta,\, 1 \rangle & \text{if } \alpha \ll \beta \\ +\infty & \text{otherwise.} \end{cases}$$

It can be rewritten as an f-divergence associated to

$$\psi : x \in \mathbb{R}_{\geqslant 0} \mapsto x\,\log(x) - x + 1 \;\in \mathbb{R}_{\geqslant 0},$$

with $0 \cdot \log(0) = 0$, as

$$\mathrm{KL}(\alpha,\beta) \;=\; \begin{cases} \langle \beta,\, \psi(\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}) \rangle & \text{if } \alpha \ll \beta \\ +\infty & \text{otherwise.} \end{cases} \tag{28}$$

Since ψ is a strictly convex function with a unique global minimum at $\psi(1) = 0$, we thus get that $\mathrm{KL}(\alpha,\beta) \geqslant 0$ with equality iff. $\alpha = \beta$.

Dual formulation. The convex conjugate of ψ is defined for $u \in \mathbb{R}$ by

$$\begin{aligned} \psi^*(u) &\overset{\text{def.}}{=} \sup_{x>0} \left(xu - \psi(x)\right) \\ &= \quad e^u - 1, \\ \text{and we have} \quad \psi(x) + \psi^*(u) &\;\geqslant\; xu \end{aligned} \tag{29}$$

for all $(x,u) \in \mathbb{R}_{\geqslant 0} \times \mathbb{R}$, with equality if $x > 0$ and $u = \log(x)$. This allows us to rewrite the Kullback-Leibler divergence as the solution of a dual concave problem:

Proposition 5 (Dual formulation of KL)**.** *Under the assumptions above,*

$$\mathrm{KL}(\alpha,\beta) \;=\; \sup_{h \in \mathcal{F}_b(\mathcal{X},\mathbb{R})} \langle \alpha,\, h \rangle - \langle \beta,\, e^h - 1 \rangle \tag{30}$$

where $\mathcal{F}_b(\mathcal{X},\mathbb{R})$ is the space of bounded measurable functions from $\mathcal{X}$ to $\mathbb{R}$.

Proof. Lower bound on the sup. If α is not absolutely continuous with respect to β, there exists a Borel set A such that $\alpha(A) > 0$ and $\beta(A) = 0$. Consequently, for $h = \lambda\, \mathbf{1}_A$,

$$\langle \alpha,\, h \rangle - \langle \beta,\, e^h - 1 \rangle \;=\; \lambda\, \alpha(A) \;\xrightarrow{\lambda \to +\infty}\; +\infty.$$

Otherwise, if $\alpha \ll \beta$, we define $h_* = \log \frac{\mathrm{d}\alpha}{\mathrm{d}\beta}$ and see that

$$\langle \alpha,\, h_* \rangle - \langle \beta,\, e^{h_*} - 1 \rangle \;=\; \mathrm{KL}(\alpha,\beta).$$

If $h_n = \log(\frac{\mathrm{d}\alpha}{\mathrm{d}\beta})\, \mathbf{1}_{1/n \leqslant \mathrm{d}\alpha/\mathrm{d}\beta \leqslant n} \in \mathcal{F}_b(\mathcal{X},\mathbb{R})$, the monotone and dominated convergence theorems then allow us to show that

$$\langle \alpha,\, h_n \rangle - \langle \beta,\, e^{h_n} - 1 \rangle \;\xrightarrow{n \to +\infty}\; \mathrm{KL}(\alpha,\beta).$$

Upper bound on the sup. If $h \in \mathcal{F}_b(\mathcal{X},\mathbb{R})$ and $\alpha \ll \beta$, combining (28) and (29) allow us to show that

$$\begin{aligned} &\mathrm{KL}(\alpha,\beta) - \langle \alpha,\, h \rangle + \langle \beta,\, e^h - 1 \rangle \\ &= \langle \beta,\, \psi(\tfrac{\mathrm{d}\alpha}{\mathrm{d}\beta}) + \psi^*(h) - h\tfrac{\mathrm{d}\alpha}{\mathrm{d}\beta} \rangle \;\geqslant\; 0. \end{aligned}$$

The optimal value of $\langle \alpha,\, h \rangle - \langle \beta,\, e^h - 1 \rangle$ is bounded above and below by $\mathrm{KL}(\alpha,\beta)$: we get (30). □

Since $\langle \alpha,\, h \rangle - \langle \beta,\, e^h - 1 \rangle$ is a convex function of (α,β), taking the supremum over test functions $h \in \mathcal{F}_b(\mathcal{X},\mathbb{R})$ defines a *convex* divergence:

Proposition 6. *The* KL *divergence is a (jointly) convex function on $\mathcal{M}^+(\mathcal{X}) \times \mathcal{M}^+(\mathcal{X})$.*

Going further, the density of continuous functions in the space of bounded measurable functions allows us to restrict the optimization domain:

Proposition 7. *Under the same assumptions,*

$$\mathrm{KL}(\alpha,\beta) \;=\; \sup_{h \in \mathcal{C}(\mathcal{X},\mathbb{R})} \langle \alpha,\, h \rangle - \langle \beta,\, e^h - 1 \rangle \tag{31}$$

where $\mathcal{C}(\mathcal{X},\mathbb{R})$ is the space of (bounded) continuous functions on the compact set $\mathcal{X}$.

Proof. Let $h = \sum_{i \in I} h_i\, \mathbf{1}_{A_i}$ be a simple Borel function on $\mathcal{X}$, and let us choose some error margin $\delta > 0$. Since α and β are Radon measures, for any i in the finite set of indices I, there exists a compact set K_i and an open set V_i such that $K_i \subset A_i \subset V_i$ and

$$\sum_{i \in I} \max[\, \alpha(V_i \backslash K_i)\,,\, \beta(V_i \backslash K_i)\,] \leqslant \delta.$$

Moreover, for any $i \in I$, there exists a continuous function φ_i such that $\mathbf{1}_{K_i} \leqslant \varphi_i \leqslant \mathbf{1}_{V_i}$. The continuous function $g = \sum_{i \in I} h_i \varphi_i$ is then such that

$$|\langle \alpha,\, g - h \rangle| \leqslant \|h\|_\infty\, \delta \quad \text{and} \quad |\langle \beta,\, e^g - e^h \rangle| \leqslant \|e^h\|_\infty\, \delta$$

so that

$$\begin{aligned} &|\, (\langle \alpha,\, h \rangle - \langle \beta,\, e^h - 1 \rangle) \;-\; (\langle \alpha,\, g \rangle - \langle \beta,\, e^g - 1 \rangle)\,| \\ &\qquad\qquad \leqslant (\|h\|_\infty + \|e^h\|_\infty)\, \delta. \end{aligned}$$

As we let our simple function approach any measurable function in $\mathcal{F}_b(\mathcal{X},\mathbb{R})$, choosing δ arbitrarily small, we then get (31) through (30). □

We can then show that the Kullback-Leibler divergence is weakly lower semi-continuous:

Proposition 8. *If $\alpha_n \rightharpoonup \alpha$ and $\beta_n \rightharpoonup \beta$ are weakly converging sequences in $\mathcal{M}^+(\mathcal{X})$, we get*

$$\liminf_{n\to+\infty} \mathrm{KL}(\alpha_n, \beta_n) \geqslant \mathrm{KL}(\alpha, \beta).$$

Proof. According to (31), the KL divergence is defined as a pointwise supremum of weakly continuous applications

$$\varphi_h : (\alpha, \beta) \mapsto \langle \alpha, h\rangle - \langle \beta, e^h - 1\rangle,$$

for $h \in \mathcal{C}(\mathcal{X}, \mathbb{R})$. It is thus lower semi-continuous for the convergence in law. □

A.2 SoftMin Operator

Proposition 9 (The SoftMin interpolates between a minimum and a sum)**.** *Under the assumptions of the definition* (12)*, we get that*

$$\min_{x\sim\alpha}{}_\varepsilon\ \varphi(x) \xrightarrow{\varepsilon\to 0} \min_{x\in \mathrm{Supp}(\alpha)} \varphi(x) \xrightarrow{\varepsilon\to+\infty} \langle \alpha, \varphi\rangle.$$

If φ and ψ are two continuous functions in $\mathcal{C}(\mathcal{X})$ such that $\varphi \leqslant \psi$,

$$\min_{x\sim\alpha}{}_\varepsilon\ \varphi(x) \leqslant \min_{x\sim\alpha}{}_\varepsilon\ \psi(x). \tag{32}$$

Finally, if $K \in \mathbb{R}$ is constant with respect to x, we have that

$$\min_{x\sim\alpha}{}_\varepsilon\ \big[K + \varphi(x)\big] = K + \min_{x\sim\alpha}{}_\varepsilon\ \big[\varphi(x)\big]. \tag{33}$$

Proposition 10 (The SoftMin operator is continuous)**.** *Let (α_n) be a sequence of probability measures converging* weakly *towards α, and (φ_n) be a sequence of continuous functions that converges* uniformly *towards φ. Then, for $\varepsilon > 0$, the SoftMin of the values of φ_n on α_n converges towards the SoftMin of the values of φ on α, i.e.*

$$\big(\alpha_n \rightharpoonup \alpha,\ \varphi_n \xrightarrow{\|\cdot\|_\infty} \varphi\big) \implies \min_{x\sim\alpha_n}{}_\varepsilon\ \varphi_n(x) \to \min_{x\sim\alpha}{}_\varepsilon\ \varphi(x).$$

B Proofs

B.1 Dual Potentials

We first state some important properties of solutions (f, g) to the dual problem (8). Please note that these results hold under the assumption that $(\mathcal{X}, \mathrm{d})$ is a compact metric space, endowed with a *ground cost* function $\mathrm{C} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that is κ-Lipschitz with respect to both of its input variables.

The existence of an optimal pair (f, g) of potentials that reaches the maximal value of the dual objective is proved using the contractance of the Sinkhorn map T, defined in (11), for the Hilbert projective metric (Franklin and Lorenz, 1989).

While optimal potentials are only defined (α, β)-a.e., as highlighted in Proposition 1, they are extended to the whole domain $\mathcal{X}$ by imposing, similarly to the classical theory of OT (Santambrogio, 2015, Remark 1.13), that they satisfy

$$f = \mathrm{T}(\beta, g) \quad \text{and} \quad g = \mathrm{T}(\alpha, f), \tag{34}$$

with T defined in (11). We thus assume in the following that this condition holds. The following propositions studies the uniqueness and the smoothness (with respect to the spacial position and with respect to the input measures) of these functions (f, g) defined on the whole space.

Proposition 11 (Uniqueness of the dual potentials up to an additive constant)**.** *Let (f_0, g_0) and (f_1, g_1) be two optimal pairs of dual potentials for a problem $\mathrm{OT}_\varepsilon(\alpha, \beta)$ that satisfy* (34)*. Then, there exists a constant $K \in \mathbb{R}$ such that*

$$f_0 = f_1 + K \text{ and } g_0 = g_1 - K. \tag{35}$$

Proof. For $t \in [0, 1]$, let us define $f_t = f_0 + t(f_1 - f_0)$, $g_t = g_0 + t(g_1 - g_0)$ and

$$\begin{aligned}\varphi(t) =& \langle \alpha, f_t\rangle + \langle \beta, g_t\rangle \\ &- \varepsilon\langle \alpha\otimes\beta, \exp\big(\tfrac{1}{\varepsilon}(f_t \oplus g_t - \mathrm{C})\big) - 1\rangle,\end{aligned}$$

the value of the dual objective between the two optimal pairs. As φ is a concave function bounded above by $\varphi(0) = \varphi(1) = \mathrm{OT}_\varepsilon(\alpha, \beta)$, it is *constant* with respect to t. Hence, for all t in $[0, 1]$,

$$\begin{aligned}0 &= \varphi''(t) \\ &= -\tfrac{1}{\varepsilon}\langle \alpha\otimes\beta, e^{(f_t\oplus g_t - \mathrm{C})/\varepsilon}((f_1 - f_0)\oplus(g_1 - g_0))^2\rangle.\end{aligned}$$

This is only possible if, $\alpha\otimes\beta$-a.e. in (x, y),

$$\big(f_1(x) - f_0(x) + g_1(y) - g_0(y)\big)^2 = 0,$$

i.e. there exists a constant $K \in \mathbb{R}$ such that

$$\begin{aligned}f_1(x) - f_0(x) &= +K \quad \alpha\text{-a.e.} \\ g_1(y) - g_0(y) &= -K \quad \beta\text{-a.e.}\end{aligned}$$

As we extend the potentials through (34), the SoftMin operator commutes with the addition of K (33) and lets our result hold on the whole feature space. □

Proposition 12 (Lipschitz property)**.** *The optimal potentials (f, g) of the dual problem* (8) *are both κ-Lipschitz functions on the feature space $(\mathcal{X}, \mathrm{d})$.*

Proof. According to (34), f is a SoftMin combination of κ-Lipschitz functions of the variable x; using the algebraic properties of the SoftMin operator detailed in (32-33), one can thus show that f is a κ-Lipschitz function on the feature space. The same argument holds for g. □

Proposition 13 (The dual potentials vary continuously with the input measures)**.** *Let $\alpha_n \rightharpoonup \alpha$ and $\beta_n \rightharpoonup \beta$ be weakly converging sequences of measures in $\mathcal{M}_1^+(\mathcal{X})$. Given some arbitrary anchor point $x_o \in \mathcal{X}$, let us denote by (f_n, g_n) the (unique) sequence of optimal potentials for $\mathrm{OT}_\varepsilon(\alpha_n, \beta_n)$ such that $f_n(x_o) = 0$.*

Then, f_n and g_n converge uniformly towards the unique pair of optimal potentials (f, g) for $\mathrm{OT}_\varepsilon(\alpha, \beta)$ such that $f(x_o) = 0$. Up to the value at the anchor point x_o, we thus have that

$$\big(\alpha_n \rightharpoonup \alpha,\, \beta_n \rightharpoonup \beta\big) \Longrightarrow \big(f_n \xrightarrow{\|\cdot\|_\infty} f,\, g_n \xrightarrow{\|\cdot\|_\infty} g\big).$$

Proof. For all n in $\mathbb{N}$, the potentials f_n and g_n are κ-Lipschitz functions on the compact, bounded set $\mathcal{X}$. As $f_n(x_o)$ is set to zero, we can bound $|f_n|$ on $\mathcal{X}$ by κ times the diameter of $\mathcal{X}$; combining this with (34), we can then produce a uniform bound on both f_n and g_n: there exists a constant $M \in \mathbb{R}$ such that

$$\forall\, n \in \mathbb{N}, \forall\, x \in \mathcal{X}, -M \leqslant f_n(x), g_n(x) \leqslant +M.$$

Being equicontinuous and uniformly bounded on the compact set $\mathcal{X}$, the sequence $(f_n, g_n)_n$ satisfies the hypotheses of the Ascoli-Arzela theorem: there exists a subsequence $(f_{n_k}, g_{n_k})_k$ that converges uniformly towards a pair (f, g) of continuous functions. k tend to infinity, we see that $f(x_o) = 0$ and, using the continuity of the SoftMin operator (Proposition 10) on the optimality equations (10), we show that (f, g) is an optimal pair for $\mathrm{OT}_\varepsilon(\alpha, \beta)$.

Now, according to Proposition 11, such a limit pair of optimal potentials (f, g) is *unique*. $(f_n, g_n)_n$ is thus a *compact* sequence with a *single* possible adherence value: it has to converge, uniformly, towards (f, g). □

B.2 Proof of Proposition 2

The proof is mainly inspired from (Santambrogio, 2015, Proposition 7.17). Let us consider α, $\delta\alpha$, β, $\delta\beta$ and times t in a neighborhood of 0, as in the statement above. We define $\alpha_t = \alpha + t\delta\alpha$, $\beta_t = \beta + t\delta\beta$ and the variation ratio Δ_t given by

$$\Delta_t \overset{\text{def.}}{=} \frac{\mathrm{OT}_\varepsilon(\alpha_t, \beta_t) - \mathrm{OT}_\varepsilon(\alpha, \beta)}{t}.$$

Using the very definition of OT_ε and the continuity property of Proposition 13, we now provide lower and upper bounds on Δ_t as t goes to 0.

Weak* continuity. As written in (13), $\mathrm{OT}_\varepsilon(\alpha, \beta)$ can be computed through a straightforward, *continuous* expression that does not depend on the value of the optimal dual potentials (f, g) at the anchor point x_o:

$$\mathrm{OT}_\varepsilon(\alpha, \beta) = \langle \alpha, f \rangle + \langle \beta, g \rangle.$$

Combining this equation with Proposition 13 (that guarantees the *uniform* convergence of potentials for weakly converging sequences of probability measures) allows us to conclude.

Lower bound. First, let us remark that (f, g) is a *suboptimal* pair of dual potentials for $\mathrm{OT}_\varepsilon(\alpha_t, \beta_t)$. Hence,

$$\begin{aligned}\mathrm{OT}_\varepsilon(\alpha_t, \beta_t) \geqslant\ & \langle \alpha_t, f \rangle + \langle \beta_t, g \rangle \\ & - \varepsilon \langle \alpha_t \otimes \beta_t, \exp\big(\tfrac{1}{\varepsilon}(f \oplus g - \mathrm{C})\big) - 1 \rangle\end{aligned}$$

and thus, since

$$\begin{aligned}\mathrm{OT}_\varepsilon(\alpha, \beta) =\ & \langle \alpha, f \rangle + \langle \beta, g \rangle \\ & - \varepsilon \langle \alpha \otimes \beta, \exp(\tfrac{1}{\varepsilon}(f \oplus g - \mathrm{C})) - 1 \rangle,\end{aligned}$$

one has

$$\begin{aligned}\Delta_t &\geqslant \langle \delta\alpha, f \rangle + \langle \delta\beta, g \rangle \\ &\quad - \varepsilon \langle \delta\alpha \otimes \beta + \alpha \otimes \delta\beta, \exp(\tfrac{1}{\varepsilon}(f \oplus g - \mathrm{C})) \rangle + o(1) \\ &\geqslant \langle \delta\alpha, f - \varepsilon \rangle + \langle \delta\beta, g - \varepsilon \rangle + o(1),\end{aligned}$$

since g and f satisfy the optimality equations (10).

Upper bound. Conversely, let us denote by (g_t, f_t) the optimal pair of potentials for $\mathrm{OT}_\varepsilon(\alpha_t, \beta_t)$ satisfying $g_t(x_o) = 0$ for some arbitrary anchor point $x_o \in \mathcal{X}$. As (f_t, g_t) are suboptimal potentials for $\mathrm{OT}_\varepsilon(\alpha, \beta)$, we get that

$$\begin{aligned}\mathrm{OT}_\varepsilon(\alpha, \beta) \geqslant\ & \langle \alpha, f_t \rangle + \langle \beta, g_t \rangle \\ & - \varepsilon \langle \alpha \otimes \beta, \exp\big(\tfrac{1}{\varepsilon}(f_t \oplus g_t - \mathrm{C})\big) - 1 \rangle\end{aligned}$$

and thus, since

$$\begin{aligned}\mathrm{OT}_\varepsilon(\alpha_t, \beta_t) =\ & \langle \alpha_t, f_t \rangle + \langle \beta_t, g_t \rangle \\ & - \varepsilon \langle \alpha_t \otimes \beta_t, \exp(\tfrac{1}{\varepsilon}(f_t \oplus g_t - \mathrm{C})) - 1 \rangle,\end{aligned}$$

$$\begin{aligned}\Delta_t &\leqslant \langle \delta\alpha, f_t \rangle + \langle \delta\beta, g_t \rangle \\ &\quad - \varepsilon \langle \delta\alpha \otimes \beta_t + \alpha_t \otimes \delta\beta, \exp(\tfrac{1}{\varepsilon}(f_t \oplus g_t - \mathrm{C})) \rangle + o(1) \\ &\leqslant \langle \delta\alpha, f_t - \varepsilon \rangle + \langle \delta\beta, g_t - \varepsilon \rangle + o(1).\end{aligned}$$

Conclusion. Now, let us remark that as t goes to 0

$$\alpha + t\delta\alpha \rightharpoonup \alpha \qquad \text{and} \qquad \beta + t\delta\beta \rightharpoonup \beta.$$

Thanks to Proposition 13, we thus know that f_t and g_t converge uniformly towards f and g. Combining the lower and upper bound, we get

$$\Delta_t \xrightarrow{t \to 0} \langle \delta\alpha, f - \varepsilon \rangle + \langle \delta\beta, g - \varepsilon \rangle = \langle \delta\alpha, f \rangle + \langle \delta\beta, g \rangle,$$

since $\delta\alpha$ and $\delta\beta$ both have an overall mass that sums up to zero.

B.3 Proof of Proposition 3

The definition of $\mathrm{OT}_\varepsilon(\alpha,\alpha)$ is that

$$\mathrm{OT}_\varepsilon(\alpha,\alpha) = \max_{(f,g)\in\mathcal{C}(\mathcal{X})^2} \langle \alpha, f+g\rangle - \varepsilon\langle \alpha\otimes\alpha, e^{(f\oplus g-\mathrm{C})/\varepsilon}-1\rangle.$$

Reduction of the problem. Thanks to the symmetry of this concave problem with respect to the variables f and g, we know that there exists a pair $(f, g=f)$ of optimal potentials on the diagonal, and

$$\mathrm{OT}_\varepsilon(\alpha,\alpha) = \max_{f\in\mathcal{C}(\mathcal{X})} 2\langle \alpha, f\rangle - \varepsilon\langle \alpha\otimes\alpha, e^{(f\oplus f-\mathrm{C})/\varepsilon}-1\rangle.$$

Thanks to the density of continuous functions in the set of simple measurable functions, just as in the proof of Proposition 7, we show that this maximization can be done in the full set of measurable functions $\mathcal{F}_b(\mathcal{X},\mathbb{R})$:

$$\begin{aligned}\mathrm{OT}_\varepsilon(\alpha,\alpha) &= \max_{f\in\mathcal{F}_b(\mathcal{X},\mathbb{R})} 2\langle \alpha, f\rangle - \varepsilon\langle \alpha\otimes\alpha, e^{(f\oplus f-\mathrm{C})/\varepsilon}-1\rangle \\ &= \max_{f\in\mathcal{F}_b(\mathcal{X},\mathbb{R})} 2\langle \alpha, f\rangle - \varepsilon\langle \exp(f/\varepsilon)\alpha, k_\varepsilon\star\exp(f/\varepsilon)\alpha\rangle + \varepsilon,\end{aligned}$$

where $\star$ denotes the smoothing (convolution) operator defined through

$$[k\star\mu](x) \;=\; \int_{\mathcal{X}} k(x,y)\,\mathrm{d}\mu(y)$$

for $k\in\mathcal{C}(\mathcal{X}\times\mathcal{X})$ and $\mu\in\mathcal{M}^+(\mathcal{X})$.

Optimizing on measures. Through a change of variables

$$\mu = \exp(f/\varepsilon)\,\alpha \qquad \text{i.e.} \qquad f = \varepsilon\log\tfrac{\mathrm{d}\mu}{\mathrm{d}\alpha},$$

keeping in mind that α is a probability measure, we then get that

$$\begin{aligned}\mathrm{OT}_\varepsilon(\alpha,\alpha) &= \varepsilon \max_{\mu\in\mathcal{M}^+(\mathcal{X}),\alpha\ll\mu\ll\alpha} 2\langle\alpha,\log\tfrac{\mathrm{d}\mu}{\mathrm{d}\alpha}\rangle - \langle\mu, k_\varepsilon\star\mu\rangle + 1 \\ -\tfrac{1}{2}\mathrm{OT}_\varepsilon(\alpha,\alpha) &= \varepsilon \min_{\mu\in\mathcal{M}^+(\mathcal{X}),\alpha\ll\mu\ll\alpha} \langle\alpha,\log\tfrac{\mathrm{d}\alpha}{\mathrm{d}\mu}\rangle + \tfrac{1}{2}\langle\mu, k_\varepsilon\star\mu\rangle - \tfrac{1}{2},\end{aligned}$$

where we optimize on positive measures $\mu\in\mathcal{M}^+(\mathcal{X})$ such that $\alpha\ll\mu$ and $\mu\ll\alpha$.

Expansion of the problem. As $k_\varepsilon(x,y) = \exp(-\mathrm{C}(x,y)/\varepsilon)$ is positive for all x and y in $\mathcal{X}$, we can remove the $\mu\ll\alpha$ constraint from the optimization problem:

$$-\tfrac{1}{2}\mathrm{OT}_\varepsilon(\alpha,\alpha) = \varepsilon \min_{\mu\in\mathcal{M}^+(\mathcal{X}),\alpha\ll\mu} \langle\alpha,\log\tfrac{\mathrm{d}\alpha}{\mathrm{d}\mu}\rangle + \tfrac{1}{2}\langle\mu, k_\varepsilon\star\mu\rangle - \tfrac{1}{2}.$$

Indeed, restricting a positive measure μ to the support of α lowers the right-hand term $\langle\mu, k_\varepsilon\star\mu\rangle$ without having any influence on the density of α with respect to μ. Finally, let us remark that the $\alpha\ll\mu$ constraint is already encoded in the $\log\frac{\mathrm{d}\alpha}{\mathrm{d}\mu}$ operator, which blows up to infinity if α has no density with respect to μ; all in all, we thus have:

$$\begin{aligned}\mathrm{F}_\varepsilon(\alpha) &= -\tfrac{1}{2}\mathrm{OT}_\varepsilon(\alpha,\alpha) \\ &= \varepsilon \min_{\mu\in\mathcal{M}^+(\mathcal{X})} \langle\alpha,\log\tfrac{\mathrm{d}\alpha}{\mathrm{d}\mu}\rangle + \tfrac{1}{2}\langle\mu, k_\varepsilon\star\mu\rangle - \tfrac{1}{2},\end{aligned}$$

which is the desired result.

Existence of the optimal measure μ. In the expression above, the existence of an optimal μ is given as a consequence of the well-known fact from OT theory that optimal dual potentials f and g *exist*, so that the dual OT problem (8) is a max and not a mere supremum. Nevertheless, since this property of F_ε is key to the metrization of the convergence in law by Sinkhorn divergences, let us endow it with a direct, alternate proof:

Proposition 14. *For any $\alpha\in\mathcal{M}_1^+(\mathcal{X})$, assuming that $\mathcal{X}$ is compact, there exists a unique $\mu_\alpha\in\mathcal{M}^+(\mathcal{X})$ such that*

$$\mathrm{F}_\varepsilon(\alpha) \;=\; \varepsilon\,\big[\,\langle\alpha,\log\tfrac{\mathrm{d}\alpha}{\mathrm{d}\mu_\alpha}\rangle + \tfrac{1}{2}\langle\mu_\alpha, k_\varepsilon\star\mu_\alpha\rangle - \tfrac{1}{2}\,\big].$$

Moreover, $\alpha\ll\mu_\alpha\ll\alpha$.

Proof. Notice that for $(\alpha,\mu)\in\mathcal{M}_1^+(\mathcal{X})\times\mathcal{M}^+(\mathcal{X})$,

$$\begin{aligned}\mathrm{E}_\varepsilon(\alpha,\mu) &\overset{\text{def.}}{=} \langle\alpha,\log\tfrac{\mathrm{d}\alpha}{\mathrm{d}\mu}\rangle + \tfrac{1}{2}\langle\mu, k_\varepsilon\star\mu\rangle \\ &= \mathrm{KL}(\alpha,\mu) \;+\; \langle\alpha-\mu,\,1\rangle \;+\; \tfrac{1}{2}\|\mu\|_{k_\varepsilon}^2 - \tfrac{1}{2}.\end{aligned}$$

Since C is bounded on the compact set $\mathcal{X}\times\mathcal{X}$ and α is a probability measure, we can already say that

$$\tfrac{1}{\varepsilon}\mathrm{F}_\varepsilon(\alpha) \leqslant \mathrm{E}_\varepsilon(\alpha,\alpha) - \tfrac{1}{2} = \tfrac{1}{2}\langle\alpha\otimes\alpha,\, e^{-\mathrm{C}/\varepsilon}\rangle - \tfrac{1}{2} < +\infty.$$

Upper bound on the mass of μ. Since $\mathcal{X}\times\mathcal{X}$ is compact and $k_\varepsilon(x,y)>0$, there exists $\eta>0$ such that $k(x,y)>\eta$ for all x and y in $\mathcal{X}$. We thus get

$$\|\mu\|_{k_\varepsilon}^2 \;\geqslant\; \langle\mu,\,1\rangle^2\,\eta$$

and show that

$$\begin{aligned}\mathrm{E}_\varepsilon(\alpha,\mu) &\geqslant \langle \alpha-\mu,\,1\rangle \;+\; \tfrac{1}{2}\|\mu\|^2_{k_\varepsilon} - \tfrac{1}{2}\\ &\geqslant \langle \mu,\,1\rangle(\langle \mu,\,1\rangle\,\eta \;-\; 1) - \tfrac{1}{2}.\end{aligned}$$

As we build a minimizing sequence (μ_n) for $\mathrm{F}_\varepsilon(\alpha)$, we can thus assume that $\langle \mu_n,\,1\rangle$ is uniformly bounded by some constant $M > 0$.

Weak continuity. Crucially, the Banach-Alaoglu theorem asserts that

$$\{\,\mu \in \mathcal{M}^+(\mathcal{X}) \mid \langle \mu,\,1\rangle \leqslant M\,\}$$

is weakly compact; we can thus extract a weakly converging subsequence $\mu_{n_k} \rightharpoonup \mu_\infty$ from the minimizing sequence (μ_n). Using Proposition 8 and the fact that k_ε is continuous on $\mathcal{X}\times\mathcal{X}$, we show that $\mu \mapsto \mathrm{E}_\varepsilon(\alpha,\mu)$ is a weakly lower semi-continuous function: $\mu_\infty = \mu_\alpha$ realizes the minimum of E_ε and we get our existence result.

Uniqueness. We assumed that our kernel k_ε is *positive universal*. The squared norm $\mu \mapsto \|\mu\|^2_{k_\varepsilon}$ is thus a strictly convex functional and using Proposition 6, we can show that $\mu \mapsto \mathrm{E}_\varepsilon(\alpha,\mu)$ is *strictly* convex. This ensures that μ_α is uniquely defined. □

B.4 Proof of Proposition 4

Let us take a pair of measures $\alpha_0 \neq \alpha_1$ in $\mathcal{M}^+_1(\mathcal{X})$, and $t \in (0,1)$; according to Proposition 14, there exists a pair of measures μ_0, μ_1 in $\mathcal{M}^+(\mathcal{X})$ such that

$$\begin{aligned}&(1-t)\,\mathrm{F}_\varepsilon(\alpha_0) + t\,\mathrm{F}_\varepsilon(\alpha_1)\\ &\quad= \varepsilon\,(1-t)\,\mathrm{E}_\varepsilon(\alpha_0,\mu_0) + \varepsilon\,t\,\mathrm{E}_\varepsilon(\alpha_1,\mu_1)\\ &\quad> \varepsilon\,\mathrm{E}_\varepsilon((1-t)\,\alpha_0 + t\,\alpha_1,\,(1-t)\,\mu_0 + t\,\mu_1)\\ &\quad\geqslant \mathrm{F}_\varepsilon((1-t)\,\alpha_0 + t\,\alpha_1),\end{aligned}$$

which is enough to conclude. To show the strict inequality, let us remark that

$$\begin{aligned}&(1-t)\,\mathrm{E}_\varepsilon(\alpha_0,\mu_0) + t\,\mathrm{E}_\varepsilon(\alpha_1,\mu_1)\\ =\;&\mathrm{E}_\varepsilon((1-t)\,\alpha_0 + t\,\alpha_1,\,(1-t)\,\mu_0 + t\,\mu_1)\end{aligned}$$

would imply that $\mu_0 = \mu_1$, since $\mu \mapsto \|\mu\|^2_{k_\varepsilon}$ is strictly convex. As $\alpha \mapsto \mathrm{KL}(\alpha,\beta)$ is strictly convex on the set of measures α that are absolutely continuous with respect to β, we would then have $\alpha_0 = \alpha_1$ and a contradiction with our first hypothesis.

B.5 Proof of the Metrization of the Convergence in Law

The regularized OT cost is weakly continuous, and the uniform convergence for dual potentials ensures that H_ε and S_ε are both continuous too. Paired with (6), this property guarantees the convergence towards 0 of the Hausdorff and Sinkhorn divergences, as soon as $\alpha_n \rightharpoonup \alpha$.

Conversely, let us assume that $\mathrm{S}_\varepsilon(\alpha_n,\alpha) \to 0$ (resp. $\mathrm{H}_\varepsilon(\alpha_n,\alpha)$). Any weak limit α_{n_∞} of a subsequence $(\alpha_{n_k})_k$ is equal to α: since our divergence is weakly continuous, we have $\mathrm{S}_\varepsilon(\alpha_{n_\infty},\alpha) = 0$ (resp. $\mathrm{H}_\varepsilon(\alpha_{n_\infty},\alpha)$), and positive definiteness holds through (6).

In the meantime, since $\mathcal{X}$ is compact, the set of probability Radon measures $\mathcal{M}^+_1(\mathcal{X})$ is sequentially compact for the weak-$\star$ topology. α_n is thus a compact sequence with a unique adherence value: it converges, towards α.