Semi-Dual Unbalanced Quadratic Optimal Transport: Fast Statistical Rates and Convergent Algorithm

Adrien Vacher [1] **François-Xavier Vialard** [1]

Abstract

In this paper, we derive a semi-dual formulation for the problem of unbalanced quadratic optimal transport and we study its stability properties, namely we give upper and lower bounds for the Bregman divergence of the new objective that hold globally. We observe that the new objective gains even more convexity than in the balanced case. We use this formulation to prove the first results on statistical estimation of UOT potentials and we leverage the extra convexity to recover super-parametric rates. Interestingly, unlike in the balanced case, we do not require the potentials to be smooth. Then, use variable metric descent to solve the semi-dual problem for which we prove convergence at a $1/k$ rate for strongly convex potentials and exponential convergence in the balanced case when potentials are also smooth. We emphasize that our convergence results has an interest on its own as it generalizes previous convergence results to non-equivalent metrics. Last, we instantiate a proof-of-concept tractable version of our theoretical algorithm that we benchmark on a 2D experiment in the balanced case and on a medium dimension synthetic experiment in the unbalanced case.

1. Introduction

In its original formulation, OT is a tool to compare probability distributions: it seeks a map that optimally transports one distribution μ to an other distribution ν with respect to some fixed cost c and it returns the associated transport cost. This problem was later relaxed into a linear program by Kantorovitch and its primal formulation consists into seeking a coupling instead of a map with minimal cost and whose marginals are constrained to be μ and ν. Quite recently, OT was extended to arbitrary positive measures (Chizat, 2017), with possibly different masses, thus the name Unbalanced Optimal Transport (UOT). On the primal problem, the hard marginal constraints are relaxed by soft entropic penalties.

Currently, the methods to estimate UOT potentials mostly rely on the dual formulation of the problem (Chizat, 2017; Séjourné et al., 2019). Yet, just as in the balanced case, the raw dual formulations suffers two major drawback: first the discretisation of the infinite cost inequality constraint strongly bias the estimators, especially when the dimension is large (Vacher et al., 2021) and second, the lack of *strong* convexity of the objective leads to algorithms that require many iterations (Léger, 2021; Pham et al., 2020). One way to circumvent this issue is to pre-optimize on one potential in the dual formulation to get rid of the cost constraint and obtain the so-called *semi-dual* formulation of optimal transport. From a statistical point of view, this new formulation now benefits from the underlying regularity of the problem, even leading to superparametric rates of estimation under smoothness hypothesis (Hütter & Rigollet, 2021). Concerning numerical experiments, it was shown empirically to produce very sharp transport maps on grids with algorithms converging in just a few iterations (Jacobs & Léger, 2020). The key element behind these successes is the fact that the semi-dual formulation gains in convexity with respect to the previous linear objective of OT; around the optimum, its controls the L^2 distance between the gradient of the potential and the gradient of the optimal solution.

In this article, we propose to continue this line of study and derive a semi-dual formulation for quadratic UOT. Unlike previous works (Hütter & Rigollet, 2021; Manole et al., 2021), we derive stability bounds that hold globally and not simply around the optimum. First, we observe that in the unbalanced case, there is a gain of convexity with respect to the balanced case that allows us to derive fast statistical rates (van de Geer, 2002) even when the potentials are not assumed to be smooth. As a corollary, we obtain the first statistical rates for the problem for UOT potentials estimation. Then we derive a provably convergent algorithm to solve theoretically our semi-dual formulation. To this

[1]LIGM, Université Gustave Eiffel, Marne-La-Vallée, France. Correspondence to: Adrien Vacher <adrien.vacher@univ-eiffel.fr>, François-Xavier Vialar <francois-xavier.vialard@univ-eiffel.fr>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

end, we extend the convergence results on gradient descents with variable, yet equivalent, metrics (also known as *preconditioned* gradient descent) to the infinite dimensional case, with non-equivalent metrics. As a result, we obtain a $O(1/k)$ convergence when the potentials are assumed to be strongly convex and exponential convergence in the balanced case for smooth strongly convex potentials; crucially, we relied on the global nature of our estimates to obtain those rates. Finally, we instantiate a tractable version of our algorithm that we benchmark in the balanced case on a stochastic 2D shape matching experiment and on a medium dimension experiment in the unbalanced case that aims to recover potentials from samples. For these tasks, our model is competitive (UOT) Sinkhorn.

Assumptions and notations In what follows, μ and ν are two positive Radon measures supported on X, Y subsets of $\mathbb{R}^d$ included in B_R, the euclidean ball of radius R centered in 0. The support of a measure μ is denoted $\text{supp}(\mu)$. The duality pairing between a Radon measure α and a continuous function f is denoted $\langle f, \alpha\rangle = \int f(x)d\alpha(x)$; when f is squared integrable and α is positive, we shall denote $\|f\|^2_\beta = \langle f^2, \beta\rangle$. The quadratic function $x \mapsto \|x\|^2/2$ is denoted q and for any Gateaux differentiable function h with differential Dh, we shall denote Δ_h the Bregman divergence associated to h, defined as $\Delta_h(x, y) = h(x) - h(y) - Dh(y)(x - y)$. For any convex function f we shall denote by ∇f a subgradient of f, by f^* the conjugate (or Legendre transform) $f^*(y) = \sup_x x^\top y - f(x)$ and we call f an M-smooth function whenever the gradient of f is M-lipschitz. Given μ a positive Radon measure and a map T differentiable on the support of μ, we denote $T_\#(\mu)$ the *pushforward* of the measure μ by the map T, defined as $T_\#(\mu)(A) = \mu(T^{-1}(A))$ for all Borel sets A. Finally, for a positive Radon measure μ and a positive function h, $\mu \cdot h$ is the measure such that for all Borel A, $(\mu \cdot h)(A) = \int_A h(x)\mu(dx)$.

2. Semi-dual Unbalanced Quadratic Optimal Transport

2.1. Semi-dual formulation

Unbalanced optimal transport is a relaxation of the hard marginal constraints of optimal transport with so called Csizár divergences D_ϕ associated to some entropy function ϕ defined as follows.

Definition 2.1 (Csizár divergences)**.** An entropy function $\phi : \mathbb{R}_+ \mapsto \mathbb{R}_+ \cup \{+\infty\}$ is a convex lower semicontinuous function such that $\phi(1) = 0$. Its recession constant is $\phi'_\infty = \lim_\infty \frac{\phi(r)}{r}$. Let μ, ν be nonnegative Radon measures on a convex domain Ω in $\mathbb{R}^d$. The *Csiszàr divergence* associated with ϕ is $D_\phi(\mu, \nu) = \int_\Omega \phi\left(\frac{\mathrm{d}\mu(x)}{\mathrm{d}\nu(x)}\right) \mathrm{d}\nu(x) + \phi'_\infty \int_\Omega \mathrm{d}\mu^\perp$ where $\mu^\perp$ is the orthogonal part of the Lebesgue decomposition of μ with respect to ν.

For instance, if one takes the entropy $\phi(t) = t\log(t) - t + 1$, we recover the Kulback Leibler case $D_\phi(\mu, \nu) = \text{KL}(\mu, \nu)$. The primal formulation of Unbalanced Optimal transport reads

$$\text{UOT}(\mu, \nu) = \inf_{\pi \in \mathcal{M}_+(X\times Y)} D_\phi(\pi_0, \mu) + D_\phi(\pi_1, \nu) + \langle c, \pi\rangle \,,$$

where c is the ground cost, $M_+(X \times Y)$ is the space of positive Radon measures over $X \times Y$ and π_i is the i-th marginal of π; note that standard OT is recovered for the entropy function $\phi(x) = \iota_{\{1\}}(x)$ the convex indicator function of $\{1\}$ and that the Gaussian-Hellinger metric is recovered for $\phi(t) = t\log(t) - t + 1$ and a quadratic cost. When strong duality holds (Chizat, 2017), the dual formulation of UOT reads

$$\begin{aligned}\text{UOT}(\mu, \nu) = \sup_{z_0, z_1} &- \langle \phi^*(-z_0), \mu\rangle - \langle \phi^*(-z_1), \mu\rangle \\ &+ \iota(z_0 \oplus z_1 \leq c)\,,\end{aligned} \tag{1}$$

where for all $(x, y) \in X \times Y$, $z_0 \oplus z_1(x, y) = z_0(x) + z_1(y)$ and where the functions (z_0, z_1) will be referred as *potentials* throughout the rest of the paper. We shall assume throughout the paper that the cost c is the quadratic cost $c(x, y) = \frac{\|x-y\|^2}{2} := q(x - y)$. Similarly to standard optimal transport, UOT often leads to a transport map between (rescaled) versions of the marginals, as shown in the Proposition below.

Proposition 2.2 (Gallouët et al. (2021))**.** *Assuming that ϕ^* is differentiable and that the optimum of problem (1) is attained for the potentials (z_0, z_1), the measures $\tilde{\mu} = \mu\cdot(\phi^*)'\circ(z_0 - q)$ and $\tilde{\nu} = \nu\cdot(\phi^*)'\circ(z_1 - q)$ have equal mass. Furthermore, if (z_0, z_1) are differentiable, these measures are related by $\tilde{\mu} = \nabla(q - z_1)_\#(\tilde{\nu})$ and $\tilde{\nu} = \nabla(q - z_0)_\#(\tilde{\mu})$.*

One can check that at optimum, the potentials (z_0, z_1) are related as $z_1 = q - (q - z_0)^*$. If one plugs this optimality condition in the dual formulation, one obtains the semi-dual formulation of UOT.

Proposition 2.3. *The semi-dual formulation reads $\text{UOT}(\mu, \nu) = -\inf_z J_{\mu,\nu}(z)$ with $J_{\mu,\nu}(z) := \langle \phi^*(z - q), \mu\rangle + \langle \phi^*(z^* - q), \nu\rangle + \iota(z \in CVX)$ where CVX is the set of convex functions over $\mathbb{R}^d$. The functional $J_{\mu,\nu}$ is convex and furthermore, if g is strongly convex and ϕ^* is differentiable, its differential reads $DJ_{\mu,\nu}(g) = \mu \cdot (\phi^*)' \circ (g - q) - (\nabla g^*)_\#(\nu \cdot (\phi^*)' \circ (g^* - q))$.*

The proof is left in Appendix. When confusion is possible we shall denote $J_{\mu,\nu}$ by J. We shall make the assumption that ϕ^* is differentiable throughout the rest of the paper; note that this assumption is standard and is fulfilled for balanced OT where $\phi^* = id$ and for unbalanced OT with $D_\phi = \text{KL}$.

2.2. Stability estimates

Just as in the balanced case, the semi-dual formulation gains *stability* with respect to the dual-formulation; namely, under certain regularity conditions on the potentials (f, g), one can bound from under and below $\Delta_J(f, g)$ by some quadratic function of $f - g$. While previous works mainly focus on stability around the optimum (Manole et al., 2021; Hütter & Rigollet, 2021), that is $\Delta_J(f, g_0)$ with $g = g_0$ the ground truth balanced transport potential, we focus on global stability: for *any* (f, g), we derive upper and lower bounds of $\Delta_J(f, g)$ in the unbalanced case. As we shall demonstrate in Sec. 4, we need our estimates to hold globally to derive provably convergent algorithms.

Proposition 2.4. *Let g be a strongly convex function such that (f, g) are bounded by K_R over B_R and (f^*, g^*) are bounded by K_R^* over B_R. If f is λ-strongly convex and ϕ^* is twice differentiable then, denoting $K = \max(K_R, K_R^*)$, the Bregman divergence $\Delta_J(f, g) = J(f) - J(g) - \langle DJ(g), f - g\rangle$ is bounded as*

$$\begin{aligned}\frac{\lambda}{2}\|\nabla(f^* - g^*)\|^2_{[\nu]_{g^*}} + \frac{I_K}{2}H_{\mu,\nu}(f, g)^2 \le \Delta_J(f, g)\\ \le \frac{1}{2\lambda}\|\nabla(f - g)\|^2_{T([\nu]_{g^*})} + \frac{S_K}{2}H_{\mu,\nu}(f, g)^2\,,\end{aligned} \tag{2}$$

where $T(\nu) = (\nabla g^*)_\#(\nu)$, $[\beta]_h = \beta \cdot (\phi^*)' \circ (h - q)$ with $I_\zeta \coloneqq \inf (\phi^*)''(B_{\zeta+R^2/2}), S_\zeta \coloneqq \sup (\phi^*)''(B_{\zeta+R^2/2})$ and $H_{\mu,\nu}(f, g)^2 = \|f - g\|^2_\mu + \|f^* - g^*\|^2_\nu$. Conversely, if f is convex and M-smooth, then $\Delta_J(f, g)$ is bounded as

$$\begin{aligned}\frac{1}{2M}\|\nabla(f - g)\|^2_{T([\nu]_{g^*})} + \frac{I_K}{2}H_{\mu,\nu}(f, g)^2 \le \Delta_J(f, g)\\ \le \frac{M}{2}\|\nabla(f^* - g^*)\|^2_{[\nu]_{g^*}} + \frac{S_K}{2}H_{\mu,\nu}(f, g)^2\,.\end{aligned}$$

In particular, for balanced OT, we have $\phi^* = id$, hence $I_K = S_K = 0$ and we recover the (standard) estimates

$$\frac{\lambda}{2}\|\nabla(f^* - g^*)\|^2_{L^2(\nu)} \le \Delta_J(f, g) \le \frac{\|\nabla(f - g)\|^2_{L^2(T(\nu))}}{2\lambda}\,,$$

in the λ-strongly convex case.

The proof is left in Appendix. In the case of balanced OT, when f is only assumed to be λ-strongly convex, Δ_J only controls $\nabla f^* - \nabla g^*$ while in unbalanced OT, if we pick a locally strongly convex entropy ϕ, which occurs in the standard case $D_\phi = \mathrm{KL}$, we have $I_K > 0$ hence Δ_J also controls the difference of conjugates $f^* - g^*$ and the potentials themselves $f - g$; unlike balanced OT, we do not need to make a smoothness assumption on f to gain control over $f - g$.

In the following corollary, we derive an upper-bound on $\Delta_J(f, g)$ that only depends on $f - g$ and no longer on the difference of the conjugates $f^* - g^*$. Indeed, as we show in Sec. 4, removing this dependency is necessary to derive provably convergent algorithms.

Corollary 2.5. *Under the same assumptions as in Prop. 2.4, if f is λ-strongly convex and if there exists R^* such that $\nabla f^*(B_R), \nabla g^*(B_R) \subset B_{R^*}$ with f being L-lipschitz on B_{R^*} then, denoting $\tilde{H}^g_{\mu,\nu}(h) = \|h\|^2_\mu + \|h\|^2_{(\nabla g^*)_\#(\nu)}$, $\Delta_J(f, g)$ is upper-bounded as*

$$\begin{aligned}\Delta_J(f, g) \le &\frac{\|\nabla(f - g)\|^2_{(\nabla g^*)_\#([\nu]_{g^*})}}{2\lambda} + \frac{3S_K}{2}\tilde{H}^g_{\mu,\nu}(f - g)\\ &+ \frac{3S_K}{2}\frac{R^2 + L^2}{\lambda^2}\|\nabla(f - g)\|^2_{(\nabla g^*)_\#(\nu)}\,.\end{aligned}$$

The proof is left in Appendix. Whether a similar lower bound on Δ_J depending only on the difference $f - g$ still holds is an open question that we postpone for future works.

3. Statistical Rates

In this section, we restrict ourselves to the case where μ, ν are measures that we can only access in a stochastic setting through their (possibly weighted) n-independent samples denoted by $\hat{\mu}, \hat{\nu}$. In this setting, a natural way to estimate UOT map is to solve the empirical semi-dual over some space C.

Definition 3.1 (Stochastic Semi-Dual Unbalanced OT)**.** Let C be a set of real-valued function, we define $\widehat{\mathrm{UOT}}_C = -\inf_{z\in C}\hat{J}(z)$, where $\hat{J} = J_{\hat{\mu},\hat{\nu}}$. Conversely, we define an empirical potential $\hat{z}_C = \arg\min_{z\in C}\hat{J}(z)$. When no confusion is possible, we shall simply denote it $\hat{z}$.

Defining the pseudo distance $d^\lambda_\phi(z, z_0)^2 = \frac{\lambda}{2}\|\nabla(z^* - z_0^*)\|^2_{[\nu]_{z_0^*}} + \frac{I_K}{2}H^2_{\mu,\nu}(z, z_0)$ where I_K and $H^2_{\mu,\nu}(z, z_0)$ are defined in Proposition 2.4, we shall prove under suitable assumptions detailed below that, in the absence of bias *i.e.* if $z_0 \in C$, $\hat{z}$ converges toward z_0 with respect to d^λ_ϕ. For the sake of simplicity, we chose to make an *unbiased* analysis, which are known to be sub-optimal. Similar results do hold with a bias measured in terms of d^λ_ϕ.

Assumption 3.2. (i) The measures μ, ν have support included in B_R for some $R > 0$. (ii) The measures μ, ν have continuous densities with respect to the Lebesgue measure on B_R. (iii) There exists $\tilde{z}_0 \in C$ such that $\tilde{z}_0$ coincides with z_0 on $\mathrm{supp}(\mu)$ and with $\tilde{z}_0^*$ coincides with z_0^* on $\mathrm{supp}(\nu)$. (iv) The functions in C are uniformly bounded by $b(r)$ over B_r, uniformly lower bounded by l and are λ-strongly convex. (v) The conjugate of the entropy ϕ^* is strongly convex on every compact.

Note that assumptions (i), (ii) are standard in statistical OT estimation (Hütter & Rigollet, 2021; Pooladian & Niles-Weed, 2021) and that assumption (iii) ensures the absence of bias in the model. While (ii) can hardly be removed

as it ensures the existence of optimal solutions in (1) and that these solutions can be extended outside the support of the measures, we believe that under a finer analysis, (iv) can re replaced by a sub-gaussian assumption. The goal of assumptions (iv) and (v) is to avoid the smoothness assumption previously made in classical OT to attain fast statistical rates; note that assumption (v) is verified for the standard UOT case $D_\phi = \mathrm{KL}$. However, similar results would hold if (iv) and (v) were replaced by smoothness and strong-convexity.

Theorem 3.3. *Under Assumptions (i)-(iv), it holds for all* $\delta \leq \frac{M'}{L}$

$$\mathbb{E}[d_\phi^\lambda(\hat{z}_C, z_0)^2] \lesssim \delta + \frac{1}{\sqrt{n}} \int_{\frac{\delta}{4}}^{\frac{b'}{P}} \sqrt{n(C, L^\infty(B_{R'}), Pu)} \mathrm{d}u \,,$$

where $n(C, \|\cdot\|, u)$ *is the logarithm of the covering number, also called the metric entropy, of* C *with respect to the* $\|\cdot\|$ *(semi)-norm at scale* u*,* $b' = (b, R, \lambda, l, \phi)$*,* $R' = (b, R, \lambda, l)$*,* $L^\infty(B_{R'})$ *is the supremum norm restricted to* $B_{R'}$*,* $P = (b, R, \lambda, l, \phi)$ *and* $\lesssim$ *hides a factor* 64*. If we further assume (v) and that there exists* (P_μ, P_ν) *and* $\alpha < 2$ *such that for every* $u \in \mathbb{R}_{\geq 0}$*,* $n(C, L^2(\mu), u) \leq P_\mu u^{-\alpha}$ *and* $n(C, L^2(\nu), u) \leq P_\nu u^{-\alpha}$ *then* $\forall n \geq 1$*,*

$$\mathbb{E}[d_\phi^\lambda(\hat{z}, z_0)^2] \lesssim n^{-\frac{1}{1+\alpha/2}} \,, \tag{3}$$

where $\lesssim$ *hides constants that do not depend on* n*.*

The proof is left in Appendix. First we observe that in both cases the speed of convergence does improve when the metric entropy of C decreases. As shown in Vacher et al. (2021), had we relied on the raw dual formulation of UOT, we would have obtained rates scaling in $n^{-2/d}$ independently of C because of the discretization of the cost constraint. Second, we note that if the metric entropy of C is sufficiently low, we can recover rates that are faster than $1/\sqrt{n}$ and up to $O(1/n)$: again this is thanks to the semi-dual formulation, which brings extra convexity and can localize the solution as it was observed in the balanced case by Hütter & Rigollet (2021) (see the proof for more details on the localization technique). Finally, note that unlike previous works (Hütter & Rigollet, 2021; Pooladian & Niles-Weed, 2021; Manole et al., 2021), we do not require the functions in C to be smooth. An interesting example is the case of Input Convex Neural Networks (Amos et al., 2017). To go further, an analysis including a bias term is necessary as z_0 may not be represented by an ICNN. We postpone this interesting question for future work and derive instead upper-bounds for the problem estimating smooth UOT potentials where we leverage the recent results of Gallouët et al. (2021).

Corollary 3.4. *Assume that* μ *and* ν *have compact and convex support with densities* (ρ_1, ρ_2) *bounded away from zero and infinity and assume that* ϕ *is strictly convex with infinite slope at* 0*. If* (ρ_1, ρ_2) *are* k*-times continuously differentiable with* $k \in \mathbb{N}^\star$ *then, denoting* z_0 *an optimal unbalanced OT potential and* $\alpha_{k,d} = \frac{k+2}{d}$*, there exists* C *such that the empirical potential* $\hat{z}_C$ *satisfies*

$$\begin{cases} \mathbb{E}[d_\phi^\lambda(\hat{z}_C, z_0)^2] \lesssim n^{-\alpha_{k,d}} \text{ if } \alpha_{k,d} \leq 1/2, \\ \mathbb{E}[d_\phi^\lambda(\hat{z}_C, z_0)^2] \lesssim n^{-\frac{1}{1+\alpha_{k,d}^{-1}/2}} \quad \text{if } \alpha_{k,d} > 1/2 \,. \end{cases}$$

The proof is left in Appendix. Note that in that setting, d_ϕ^λ is finer than $h \mapsto \|\nabla h\|_{L^2(\mu)}$, hence it is legit to compare our rates to the ones of Hütter & Rigollet (2021) who study the problem of minimax estimation of smooth balanced transport map w.r.t. the latter pseudo-distance. As shown on Fig. 1, our rates are very close to their statistical lower bounds. Whether these lower bounds are still sharp in the unbalanced case is an open question yet we believe that their proof method remains valid in this setting. As a consequence, we conjecture that our estimator is nearly statistically optimal under the hypothesis of Corollary 3.4 in the highly smooth regime $\alpha_{k,d} > 1/2$. We claim that the larger gap in the lowly smooth regime $\alpha_{k,d} << 1/2$ is an artefact of our proof which relies on an unbiased analysis. We postpone this question for future work.

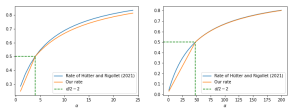

Figure 1. Comparison of our rates against the statistical lower bounds of Hütter & Rigollet (2021): on the left for $d = 12$ and on the right for $d = 100$.

4. A Provably Convergent Algorithm

In this section, we provide a theoretical algorithm to estimate unbalanced transport potentials and solve for arbitrary positive measures μ, ν the problem $\min_{f \in C} J_{\mu,\nu}(f)$ where C is a convex set of functions. Had we been in a finite dimensional setting, we could have directly applied gradient based methods to solve this problem with updates of the form $f_{k+1} = f_k - \beta DJ(f_k)$. However in our infinite dimensional setting, the gradient $DJ(f_k)$ is a measure, not a function. Frank-Wolfe algorithm provides implicit updates of the form of a convex combination of linear oracles $\arg\min_{f \in C} \langle f, DJ(f_k) \rangle$. This scheme provably converges (Dunn, 1980) yet it can be slow in practice. One way to improve convergence in practice is to recall the variational formulation of gradient descent and generate updates as $f_{k+1} = \arg\min_{f \in C} \langle f, DJ(f_k) \rangle + \beta \|f - f_k\|^2$ where $\|\cdot\|$ is a well-chosen *fixed* norm. Under the relative smoothness assumption (Bauschke et al., 2017; Lu et al., 2018) $\Delta_J(f, g) \leq \frac{1}{\beta}\|f - g\|^2$, this scheme provably converges at a $O(1/k)$ rate. Yet in our setting $\Delta_J(f, f_k)$ depends on

the *varying* pseudo-norm $L^2(\nabla f_k^*(\nu))$. Hence we study the guarantees we can obtain with a *variable metric* scheme (Frankel et al., 2015)

$$f_{k+1} = \arg\min_{f\in C} \langle f, DJ(f_k)\rangle + \frac{\beta}{2}\|f - f_k\|_{f_k}^2 . \quad (4)$$

Note that formally, (4) is a particular case of the forward-backward algorithm with variable metric (Chen & Rockafellar, 1997) that minimizes functions of the form $F = f + g$ and for which convergence rates were proven. Yet to the best of our knowledge, all the previous works either made the assumption that the metrics $\|\cdot\|_{f_k}$ are all equivalent to some *fixed* metric (Chen & Rockafellar, 1997; Uschmajew & Vandereycken, 2022), either they assume that the metrics remain at least as fine throughout the iterations $\|\cdot\|_{f_k} \lesssim \|\cdot\|_{f_{k+1}}$ (Combettes & Vũ, 2014; Cui et al., 2019). As none of these assumptions hold in our setting, we give convergence guarantees for the more general case of *non-equivalent* and *non-monotone* metrics. We prove in Appendix that under the assumption of variable relative smoothness $\Delta_J(f,g) \le \frac{\beta}{2}\|f-g\|_g^2$ where $\|\cdot\|_g$ is a pseudo-norm depending on g, we obtain a $O(1/k)$ convergence of (4); under the additional assumption that there exists $\alpha > 0$ such that $\Delta_J(f,g) \ge \frac{\alpha}{2}\|f-g\|_g^2$, we obtain exponential convergence.

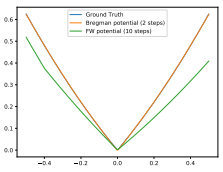

Figure 2. Potential generated by Frank-Wolfe with 10 steps (in green) vs generated by our algorithm with 2 steps (in orange) vs ground truth (in blue).

In what follows, we focus on the instantiation of (4) and the guarantees we obtain for the minimization of the semi-dual. Nevertheless, for a general result, we encourage the reader interested in convex optimization in Banach spaces to read Appendix 9.1.

Theorem 4.1 (Strongly convex case)**.** *Let C be a closed convex set of λ-strongly convex functions, $L(r)$-Lipschitz over B_r and such that for all $f \in C$, $|f(0)| \le b$. The minimum $\bar{f} = \arg\min_{f\in C} J(f)$ exists. Furthermore, for*

$$\|h\|_g^2 = 3S_K\left[\frac{R^2 + L(R^*)^2}{\lambda^2}\|\nabla h\|^2_{(\nabla g^*)_\#(\nu)} + \tilde{H}^g_{\mu,\nu}(h)\right] + \frac{1}{\lambda}\|\nabla h\|^2_{(\nabla g^*)_\#([\nu]_{g^*})},$$

where $R^* = \frac{R}{\lambda} + 2\sqrt{\frac{b}{\lambda}}$, $K = \max(b + LR, R^*(R + b + L))$ and S_K, $[\cdot]_g$ are defined in Proposition 2.4 and $\tilde{H}^g_{\mu,\nu}(h)$ is defined in Corollary 2.5, the iterates $f_{k+1} = \arg\min_{f\in C}\langle DJ(f_k), f - f_k\rangle + \frac{1}{2}\|f - f_k\|_{f_k}^2$ are well defined and they verify

$$J(f_k) - J(\bar{f}) \le \frac{I(\mu(\mathbb{R}^d) + \nu(\mathbb{R}^d))}{k\lambda^2},$$

where I is a constant that does not depend on k, λ, μ, ν.

The proof is left in Appendix. The main argument consists into remarking with this choice of varying pseudo-norm $\|h\|_g$, the global estimates of Sec.2 show that J is variably relatively smooth with respect to $\|\cdot\|_g$. In particular, the quantity minimized at each step is an upper-bound of J. Since this upper bound is quadratic, each iteration should be at least as good as Frank-Wolfe which uses a linear proxy. In practice, we observe on Fig. 2 that in the convergence of our scheme is actually faster than Frank-Wolfe.

We emphasize that our algorithm converges at a $O(1/k)$ rate. As a comparison, for the Sinkhorn algorithm (Cuturi, 2013), the best known rates in the balanced case are $O(1/(k\varepsilon))$ (Léger, 2021) or $O((1 - e^{-1/\varepsilon})^{2k})$ (Peyré et al., 2019); in particular, when ε is small (which is the case in practice), our algorithm requires fewer iterations to converge for a fixed precision.

In the next theorem, we show that in the balanced case, when the potentials of C are both strongly convex and smooth, the convergence is further improved to an exponential rate.

Theorem 4.2 (Balanced case)**.** *Let C be a set of λ-strongly convex, M-smooth function that are $L(r)$-Lipschitz over B_r and such that for all $f \in C, |f(0)| \le b$. The minimum $\min_{f\in C} J(f)$ is attained at $\bar{f} \in C$. Furthermore, using the pseudo-norm $\|h\|_g^2 = \|h\|^2_{(\nabla g^*)_\#(\nu)}$, the iterates $f_{k+1} = \arg\min_{f\in C}\langle DJ(f_k), f - f_k\rangle + \frac{1}{2\lambda}\|f - f_k\|_{f_k}^2$ are well defined and verify $J(f_k) - J(\bar{f}) \le (1 - \frac{\alpha}{\beta})^k (J(f_0) - J(\bar{f}))$.*

The proof is left in Appendix. Again, the global stability results shown in Sec. 2 are the key to prove the convergence rate: the additional lower bound that we proved in the balanced case enables us to prove that the semi-dual is Polyak-Lojaseiwicz in a generalized sense (Karimi et al., 2016) which gives the exponential convergence. Indeed, Sinkhorn algorithm also enjoys exponential convergence yet with rate $e^{-1/\epsilon} \to_{\varepsilon\to 0} 0$ as shown above whereas in our case, the rate is $\alpha/\beta = O(1)$. In comparison with the strongly convex case, Fig.3 shows that when we explicitly constrain our potentials to also be smooth (orange curve), we observe an exponential convergence up to the numerical precision. On the other hand we observe that when the potentials are only constrained to be strongly convex (blue curve), there is a first phase where the convergence is exponential and then the convergence slows down to a sublinear rate. More details on the practical instantiation of

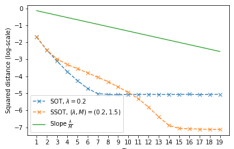

Figure 3. Convergence of $\log(\|\nabla f_k - \nabla \bar{f}\|^2_{L^2(\mu)})$ in the strongly convex case (in blue) vs the smooth and strongly convex case (in orange).

the algorithm are provided in the next section.

Whether exponential convergence can hold in the unbalanced case is an open question. On a simple example where μ a one dimensional uniform distribution and z_0 is quadratic, Fig.4 shows that convergence does occur at a linear rate in the unbalanced case (orange curve) yet significantly slower than in balanced case (blue curve) for which only several steps are necessary to reach numerical precision. This discrepancy might be due to the conditioning of the semi-dual in the unbalanced case which involves a ratio of the form $\frac{\sup(\phi^*)''}{\inf(\phi^*)''}$ that can be very large, in the KL case for instance where $\phi^*(t) = e^t - 1$. For future works, we plan to investigate other semi-dual formulations of unbalanced OT that may be better suited to the choice of the entropy function ϕ and improve the conditioning of the problem.

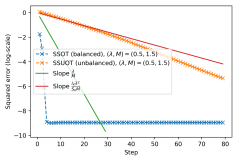

Figure 4. Convergence of $\log(\|\nabla f_k - \nabla \bar{f}\|^2_{L^2(\mu)})$ in the balanced case (blue curve) vs in the unbalanced case (orange curve).

5. Our Tractable Model

At first glimpse, even when the measures μ and ν are discrete, problem (4) seems hard to compute: though it is convex and quadratic, it nevertheless optimizes over C, a possibly infinite dimensional functional space. In this section, relying on the results of Taylor et al. (2017) on convex interpolation, we provide a tractable version of our algorithm when C is chosen to be the space of λ-strongly convex, possibly M-smooth functions.

We place ourselves in the setting where $\hat{\mu}, \hat{\nu}$ are n-samples discrete empirical measures with weights (ω^μ, ω^ν) and we chose C to be the set of λ-strongly convex functions with other minor regularity conditions detailed below. We show how to solve (4) to compute this model as well as its resulting algorithmic complexity and statistical behavior.

Algorithmic resolution We remind the reader that the iterates of the algorithm are generated as $f_{k+1} = \arg\min_{f \in C} \langle f, DJ(f_k)\rangle + \alpha \|f - f_k\|^2_{f_k}$. Now recall that for f^* differentiable on $\hat{\nu}$, the gradient of the semi-dual reads $DJ(f) = \hat{\mu} \cdot (\phi^*)' \circ (f - q) - (\nabla f^*)_\# (\hat{\nu} \cdot (\phi^*)' \circ (f^* - q))$. In particular, when f is convex, $(\nabla f^*)_\# (\hat{\nu} \cdot (\phi^*)' \circ (f^* - q))$ can be computed pointwise. Conversely, the term $\|f - f_k\|^2_{f_k}$ can also be computed exactly for a given f as it is simply an integral over the discrete measures $(\nabla g^*)_\# (\hat{\nu})$ and $\hat{\mu}$ (see Sec. 4). Hence, the objective can be computed exactly for any given (convex) f yet the constraint $f \in C$ may remain infinite dimensional. We show in the next proposition that when $C_{\lambda,L,b} = \{g + \lambda q \mid g \text{ convex}, L - \text{lipschitz}, |g(0)| \leq b\}$, the problem admits a finite quadratic reformulation.

Proposition 5.1 (Taylor et al. (2017))**.** *For $C = C_{\lambda,L,b}$, problem (4) can be reformulated as*

$$\inf_{\substack{y \in \mathbb{R}^{2n+1} \\ \mathbf{z} \in \mathbb{R}^{(2n+1)\times d}}} c^\top y + \frac{1}{2}[y - f, \mathbf{z} - \mathbf{g}]^\top Q [y - f, \mathbf{z} - \mathbf{g}]\,,$$
$$y_i - y_j \geq \mathbf{z}_j^\top (\mathbf{x}_i - \mathbf{x}_j) \qquad (5)$$
$$\|\mathbf{z}_i\| \leq L, |y_0| \leq b$$

where $f = f_k(\hat{\mu}) - \lambda q(\hat{\mu})$, $c = [0, \omega^\mu (\phi^*)'((f_k - q)(\hat{\mu})), -\omega^\nu (\phi^*)'((f_k^* - q)(\hat{\nu}))]$, $\mathbf{g} = [\vec{0}, \hat{\nu} - \lambda \nabla f_k^*(\hat{\nu})]$ and Q is a diagonal matrix with diagonal $3S_K[0, \omega^\mu, \omega^\nu, \vec{0}, (\omega^\nu \zeta)^{*d}]$ with $\zeta := \frac{1}{3S_K \lambda} + (\phi^*)'((f_k^* - q)(\hat{\nu}))(R^2 + L(R^*)^2)/\lambda^2$ where for some vector $v \in \mathbb{R}^p$ we define $(v)^{*d} = (v_1, \cdots, v_1, \cdots, v_p, \cdots v_p) \in \mathbb{R}^{d \times p}$; in the balanced case, the diagonal of Q is simply $[\vec{0}, \frac{\omega^\nu}{\lambda}]$. Furthermore, the cost to solve (5) with an Interior Point Method (Nemirovski, 2004) requires $O(n^3)$ operations.

The proof is left in Appendix where we also show how to compute pointwise the potential f_{k+1} and its conjugate, given $(y^{k+1}, \mathbf{z}^{k+1})$ the solution of (5). Furthermore, we also prove in Appendix the same type of reformulation when the potentials are also constrained to be smooth; the resulting reformulation is a Quadratically Constrained Quadratic Program with the same complexity.

Remark 5.2. In the case where we are in a low dimensional setting ($d = 1, d = 2$), it was shown that the convexity constraint can be reduced to $\tilde{O}(n)$ sparse inequalities (Mirebeau, 2016) instead of $O(n^2)$, reducing the IPM complexity to $\tilde{O}(n\sqrt{n})$, where the notation $\tilde{O}$ hides polylog factors.

Statistical behavior We give the statistical behavior of the estimator $\hat{z} = \arg\min_{z \in C_{\lambda,L,b}} \hat{J}(z)$ under the assumptions made in Sec. 3.

Proposition 5.3. *Under Assumption 3.2 (i)-(iv), it holds*

$$\begin{cases} \mathbb{E}[d_\phi^\lambda(\hat{z}, z_0)^2] \lesssim 1/\sqrt{n} & \text{if } d < 4 \\ \mathbb{E}[d_\phi^\lambda(\hat{z}, z_0)^2] \lesssim \frac{\log(n)}{\sqrt{n}} & \text{if } d = 4 \\ \mathbb{E}[d_\phi^\lambda(\hat{z}, z_0)^2] \lesssim n^{-2/d} & \text{if } d > 4\,. \end{cases}$$

If we further assume (v), we recover for $d < 4$: $\mathbb{E}[d_\phi^\lambda(\hat{z}, z_0)^2] \lesssim n^{-1/(1+d/4)}$.

Note that the no-bias assumption (iii) is not empty since when ν has a density with compact and convex support, z_0 is strongly convex (Gallouët et al., 2021) and in particular, there exists $\tilde{z}_0 \in C_{\lambda,L,b}$ such that $\tilde{z}_0$ coincides with z_0 on the support of μ and $\tilde{z}_0^*$ coincides with z_0^* on the support of ν for λ sufficiently small and L, b sufficiently large.

Overall performance comparison with Sinkhorn Using the two previous paragraphs, let us compare the performance of our model with Sinkhorn in the smooth, strongly convex balanced case [1]. In the case of balanced optimal transport, the performance of a model can be measured as the number of operations necessary to compute an estimator $\hat{z}$ of the original transport potential z_0 than achieves an error τ in average: $\mathbb{E}[\|\nabla(\hat{z}-z_0)\|^2_{L^2(\mu)}] \leq \tau$. Denoting $\hat{z}^\varepsilon_n$ the estimator obtained when using Sinkhorn with regularization ε on n-samples empirical measures $\hat{\mu}, \hat{\nu}$, Pooladian & Niles-Weed (2021) showed that $\mathbb{E}[\|\nabla(\hat{z}^\varepsilon_n - z_0)\|^2_{L^2(\mu)}] = O(\varepsilon + \frac{\varepsilon^{d/2}}{\sqrt{n}})$. Hence, after optimizing on ε w.r.t. n, we recover an average error $\mathbb{E}[\|\nabla(\hat{z}^\varepsilon_n - z_0)\|^2_{L^2(\mu)}] = O(n^{-1/d+2})$ while the error of our model is no worse than $O(n^{-2/d})$. In particular, if we denote n^m_τ (resp. n^S_τ) the number of samples required to reach an average τ error with our model (resp. with Sinkhorn), we have $n^m_\tau << n^S_\tau$ thanks to the better statistical behavior of our model. However, in order to compute our estimator from these samples, we need $O((n^m_\tau)^3)$ operations per iteration and few iterations (see Sec. 4) while Sinkhorn requires $O((n^S_\tau)^2)$ operations per iteration yet many iterations (see Sec. 4). As often in computational statistics, there is a trade-off between computational and statistical efficiency. The following result quantifies this trade-off for our model and for Sinkhorn.

Proposition 5.4 (Overall performance, informal)**.** *Under Assumption 3.2 (i)-(iv) and the additional assumption that z_0 is M-smooth, then our model requires $\tilde{O}(\tau^{-\max(7, \frac{3d}{2}+1)})$ operations to compute an estimator $\hat{z}$ such that $\nabla\hat{z}$ is a τ approximation in squared $L^2(\mu)$ norm of the original OT map ∇z_0 in average, i.e., $\mathbb{E}[\|\nabla(\hat{z} - z_0)\|^2_{L^2(\mu)}] \leq \tau$. If we further assume that integrated Fisher information is bonded along the Wasserstein-2 geodesic between μ and ν (see Chizat et al. (2020)[Equation 5] for a formal definition), the Sinkhorn model requires $O(\tau^{-2(d+1)-7})$ to compute a τ approximation of the original OT map in average.*

The proof is left in Appendix. Proposition 5.4 shows that for any dimension, our estimator has a better performance than Sinkhorn yet, in order to fulfill the no-bias assumption (iii), it requires *a priori* knowledge on z_0: indeed the strong convexity parameter λ must satisfy $\lambda \geq \inf \|\nabla^2 z_0\|$ and the parameter L must satisfy $L \geq \|\nabla(z_0 - \lambda q)\|_{L^\infty(\text{supp}(\mu))}$; we postpone these questions for future works. In the next section, our numerical experiments shall only focus on the analysis of the statistical error of the models, that is computing $\mathbb{E}[d^\lambda_\phi(\hat{z}, z_0)^2]$ for different models when given the same amount of samples, without consideration for the computations time.

[1]The statistical properties of Sinkhorn are only well studied in the balanced, smooth and strongly convex regime, hence our restriction.

6. Numerical Experiments

In this section we first present two other popular model of (unbalanced) OT potentials. Then, we compare our model and the two others on a 2D shape matching experiment and on a medium dimension experiment.

6.1. Other models

Sinkhorn The well-known Sinkhorn model (Cuturi, 2013) can be extended to the unbalanced case and its primal objective reads $\mathcal{S}^\phi_\varepsilon(\mu,\nu) = \inf_{\pi\geq 0}\langle\pi, C\rangle + D_\phi(\pi_1|\mu) + D_\phi(\pi_2|\nu) + \varepsilon\text{KL}(\pi|\mu\otimes\nu)$ where C is the ground cost (see Chizat (2017)). We use Séjourné et al. (2019, Proposition 7) to extend the discrete Sinkhorn potentials to the whole domain $\mathbb{R}^d$.

SSNB The Smooth Strongly convex Nearest Brenier model (Paty et al., 2020) was only defined for balanced optimal transport and is formulated as $\arg\min_{f\in C_{\lambda,M}} W_2^2(\nabla f(\mu), \nu)$ where $C_{\lambda,M}$ is the space of λ-strongly convex, M-smooth functions and W_2^2 is the squared Wasserstein distance. Indeed, there is a strong connection between this model and ours as the search space of the potentials is the same. However the objective differs and crucially, while the semi-dual is convex, the function $f \mapsto W_2^2(\nabla f(\mu), \nu)$ is not. The authors propose a sequence of two stages optimization to solve the problem yet no convergence guarantees are provided and as we shall see in the first experiment, their method performs less well than ours, probably because the algorithm is stuck in a local minimum.

6.2. 2D shape matching

In this experiment the models are trained on $\hat{\mu}_t$ (that we shall refer to as the ellipse) and $\hat{\nu}_t$ (that we shall refer to as the saxophone) with 700 points each. The distributions are represented on Fig. 5 [2]. As we can observe, since the ellipse is convex, we expect the pushforward from the saxophone to be smooth and conversely, we expect the pushforward from the ellipse to be strongly convex. On the other hand, we expect the pushforward from the ellipse to be discontinuous on its center to match the upper and lower parts of the saxophone respectively. We train the models on $(\hat{\mu}_t, \hat{\nu}_t)$ and we recover potentials $\hat{f}$. Then we sample 2000 points from the ellipse $\hat{\mu}_{test}$ and we visualize on Fig. 6 the pushforwards $\nabla\hat{f}(\hat{\mu}_{test})$. We observe that for a large value of ε, the potential given by

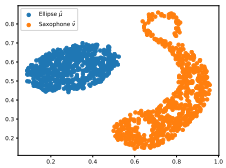

Figure 5. 2D experiment: Ellipse (in blue) and Saxophone (in orange).

[2]This data is borrowed from Feydy et al. (2017).

Sinkhorn is too smooth and cannot sufficiently deform the ellipsoid to obtain the curved shape of the saxophone. For the SSNB model, the shape of the pushforward roughly corresponds to the saxophone however the top of quite fuzzy. The shape is sharper for the semi-dual model and holes start to appear. We emphasize that for the semi-dual and SSNB models, the same search space was used yet we suspect that because of the non-convexity of its objective, the SSNB was stuck in a suboptimal local minimum; indeed, when we computed $W_2^2(\nabla \hat{f}_{sd}(\hat{\mu}_t), \hat{\nu}_t)$ we obtained a smaller value than $W_2^2(\nabla \hat{f}_{SSNB}(\hat{\mu}_t), \hat{\nu}_t)$. Finally, when ε is small enough, the Sinkhorn model recovers a very sharp pushforward. We believe that the discrepancy between the performance of our model and Sinkhorn can be explained by the $O(1/k)$ convergence rate of the semi-dual for the non-smooth case. In particular, when we computed $J_{\hat{\mu}_t,\hat{\nu}_t}(\hat{f}_\varepsilon + \lambda q)$ with $\lambda > 0$, we managed to slightly decrease the value of the semi-dual, thus proving that the optimal potential was not recovered yet. In future works, we hope to derive more efficient algorithms when smoothness is not assumed.

6.3. Medium dimension synthetic experiment

In this paragraph, we study the ability of models to recover the ground truth unbalanced transport potential z_0 between $\tilde{\mu}$ and $\tilde{\nu}$ from sampled measures $(\widehat{\tilde{\mu}}, \widehat{\tilde{\nu}})$. Using Proposition 2.2, if we take μ a Lebesgue continuous probability distribution and z_0 a strongly convex function, we have that z_0 is the solution of the unbalanced problem $\inf_z J_{\tilde{\mu},\tilde{\nu}}(z)$ with $\tilde{\mu} = \mu/(\phi^*)' \circ (z_0 - q)$ and $\tilde{\nu} = (\nabla z_0)_\#(\mu)/(\phi^*)' \circ (z_0^* - q)$. In this example, we take $\mu \sim \mathcal{U}([-0.5, 0.5]^6)$, $z_0(x) = |x| + q(x)$ and $D_\phi = \rho\text{KL}$. We chose $\rho = 5$ to avoid extreme values of $(\phi^*)'(t) = e^{t/\rho}$. Because of the low scalabilty of our model, we sampled only $n = 400$ from $\tilde{\mu}$ and $n = 400$ from $\tilde{\nu}$. We trained an unbalanced Sinkhorn model $\hat{z}_\varepsilon$ for several values of ε and an unbalanced semi-dual model $\hat{z}_\lambda$ for several values of λ; the parameters L and R were set as $1.1\|\widehat{\tilde{\mu}}\|_\infty$ and $1.1\|\widehat{\tilde{\mu}}\|_2$ respectively and S was set to 0.5. Fig. 7 plots the error $\|\hat{z} - z_0\|^2_{\tilde{\mu}}$ computed on 5000 samples of $\tilde{\mu}$; the training and computation of the error were repeated 20 independent times and the vertical bars represent the confidence interval. It shows that for $\lambda = 0.2, 0.5, 1.0$, the semi-dual model consistently outperforms Sinkhorn for any value of ε. Indeed as we could expect, the value $\lambda = 2.0$ performs the least well as it generates 2-strongly convex solutions while z_0 is only 1-strongly convex. Conversely, the value of ε needs to be sufficiently small to recover the discontinuity of $|x|$ and to reduce the bias of the model yet it should not be too small in order to mitigate the variance that behaves poorly as the dimension grows.

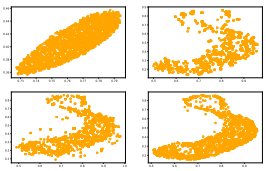

Figure 6. Pushforwards $\nabla \hat{f}(\hat{\mu}_{test})$. From top left to bottom right: Sinkhorn ($\varepsilon = 0.1$), SSNB ($\lambda = 0.2$, $M = +\infty$), Semi-dual OT ($\lambda = 0.2, L = 1$), Sinkhorn ($\varepsilon = 0.0001$).

Despite the superiority of our model in terms of statistical error, we acknowledge that it was much slower to compute in practice than Sinkhorn. While the training of Sinkhorn took at worse one minute for the lowest values of ε, our model took about 30 minutes. In particular we reckon that, whenever a reasonable precision τ is sought after, Sinkhorn achieves in practice a better statistical/computational trade-off than our model (see Overall performance paragraph in Sec. 5 for more details on the notion of statistical/computational trade-off). However we highlight the fact that our choice of search space C should be thought as a proof-of-concept model rather than as the new golden standard for UOT. It aims at illustrating the concrete usage of the semi-dual functional and what benefits we can expect when relying on this formulation instead of the vanilla dual formulation. We hope that this work will encourage UOT practitioners to design more computationally efficient, yet expressive, models C while retaining the nice properties offered by the semi-dual tool: namely, a potentially superparametric rate of statistical estimation in $O(1/n)$ and a fast algorithmic scheme (see Appendix 9.2 for an example).

7. Conclusion

In this article, we derived a semi-dual formulation of unbalanced optimal transport and provided stability bounds for its associated Bregman divergence, generalizing the results known in the balanced case. This new objective provides a natural and well-behaved estimator of unbalanced transport potentials, leading to superparametric rates of estimation even when the search space is not assumed to contain smooth functions. From an optimization point of view, our global stability results allowed to derive $O(1/k)$ and exponential rates for the balanced case using a variable metric gradient scheme with non equivalent metrics, a result that has an interest of its own. Finally, we instantiated a tractable,

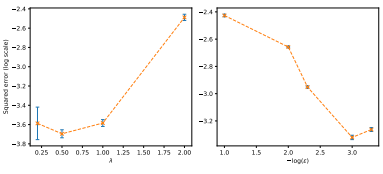

Figure 7. 6D experiment: On the left, $\|\hat{z}_\lambda - z_0\|^2_{\tilde{\mu}}$ (our model) and on the right, $\|\hat{z}_\varepsilon - z_0\|^2_{\tilde{\mu}}$ (Sinkhorn).

proof-of-concept version of our algorithm that has a more favorable statistical behavior than the well-known Sinkhorn algorithm. For future works, we shall focus on the design of a more computationally efficient search space C and the generalization of the unbalanced semi-dual to other costs.

References

Amos, B., Xu, L., and Kolter, J. Z. Input convex neural networks. In *ICML*, 2017.

Azagra, D. and Mudarra, C. Smooth convex extensions of convex functions. *Calculus of Variations and Partial Differential Equations*, 2019.

Bauschke, H. H., Bolte, J., and Teboulle, M. A descent lemma beyond lipschitz gradient continuity: First-order methods revisited and applications. *Mathematics of Operations Research*, 2017.

Bronshtein, E. M. ϵ-entropy of convex sets and functions. *Siberian Mathematical Journal*, 1976.

Bubeck, S. Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 2015.

Chen, G. H. and Rockafellar, R. T. Convergence rates in forward–backward splitting. *SIAM Journal on Optimization*, 1997.

Chizat, L. *Unbalanced optimal transport: Models, numerical methods, applications*. PhD thesis, Université Paris sciences et lettres, 2017.

Chizat, L., Roussillon, P., Léger, F., Vialard, F.-X., and Peyré, G. Faster wasserstein distance estimation with the sinkhorn divergence. In *NeurIPS*, 2020.

Combettes, P. L. and Vũ, B. C. Variable metric forward–backward splitting with applications to monotone inclusions in duality. *Optimization*, 2014.

Cui, F., Tang, Y., and Zhu, C. Convergence analysis of a variable metric forward–backward splitting algorithm with applications. *Journal of Inequalities and Applications*, 2019.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, 2013.

Dunn, J. C. Convergence rates for conditional gradient sequences generated by implicit step length rules. *SIAM Journal on Control and Optimization*, 1980.

Feydy, J., Charlier, B., Vialard, F.-X., and Peyré, G. Optimal transport for diffeomorphic registration. In *MICCAI*, 2017.

Frankel, P., Garrigos, G., and Peypouquet, J. Splitting methods with variable metric for kurdyka–Łojasiewicz functions and general convergence rates. *Journal of Optimization Theory and Applications*, 2015.

Gallouët, T., Ghezzi, R., and Vialard, F.-X. Regularity theory and geometry of unbalanced optimal transport, 2021.

Hütter, J.-C. and Rigollet, P. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 2021.

Jacobs, M. and Léger, F. A fast approach to optimal transport: the back-and-forth method. *Numerische Mathematik*, 2020.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *ECML*, 2016.

Léger, F. A gradient descent perspective on sinkhorn. *Appl. Math. Optim.*, 2021.

Lu, H., Freund, R. M., and Nesterov, Y. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 2018.

Luxburg, U. v. and Bousquet, O. Distance–based classification with lipschitz functions. *JMLR*, 2004.

Manole, T., Balakrishnan, S., Niles-Weed, J., and Wasserman, L. Plugin estimation of smooth optimal transport maps. *arXiv*, 2021.

Mirebeau, J.-M. Adaptive, anisotropic and hierarchical cones of discrete convex functions. *Numerische Mathematik*, 2016.

Nemirovski, A. Interior point polynomial time methods in convex programming. *Lecture notes*, 2004.

Paty, F.-P., d'Aspremont, A., and Cuturi, M. Regularity as regularization: Smooth and strongly convex brenier potentials in optimal transport. In *AISTATS*, 2020.

Peyré, G., Cuturi, M., et al. Computational optimal transport: With applications to data science. *Found. Trends Mach. Learn.*, 2019.

Pham, K., Le, K., Ho, N., Pham, T., and Bui, H. On unbalanced optimal transport: An analysis of sinkhorn algorithm. *ICML*, 2020.

Pooladian, A.-A. and Niles-Weed, J. Entropic estimation of optimal transport maps, 2021.

Séjourné, T., Feydy, J., Vialard, F.-X., Trouvé, A., and Peyré, G. Sinkhorn divergences for unbalanced optimal transport. *arXiv*, 2019.

Taylor, A. B., Hendrickx, J. M., and Glineur, F. Exact worst-case performance of first-order methods for composite convex optimization. *SIAM Journal on Optimization*, 2017.

Uschmajew, A. and Vandereycken, B. Variable metric forward–backward splitting with applications to monotone inclusions in duality. *Journal of Optimization Theory and Applications*, 2022.

Vacher, A., Muzellec, B., Rudi, A., Bach, F., and Vialard, F.-X. A dimension-free computational upper-bound for smooth optimal transport estimation. In *COLT*, 2021.

van de Geer, S. M-estimation using penalties or sieves. *Journal of Statistical Planning and Inference*, 2002.

van der Vaart, A. W. and Wellner, J. A. *Weak convergence and empirical processes*. Springer, 1996.

8. Proofs

8.1. Proposition 2.3

Proof. The dual formulation of UOT reads

$$\begin{aligned} \mathrm{UOT}(\mu,\nu) = \sup_{z_0,z_1} \langle -\phi^*(-z_0),\mu\rangle + \langle -\phi^*(-z_1),\nu\rangle \\ \text{s.t. } z_0(x)+z_1(y)\le q(x-y)\,, \end{aligned} \tag{6}$$

Defining $\tilde{z}_i = q - z_i$, we rewrite the problem as

$$\begin{aligned} \mathrm{UOT}(\mu,\nu) = \sup_{\tilde{z}_0,\tilde{z}_1} \langle -\phi^*(\tilde{z}_0-q),\mu\rangle + \langle -\phi^*(\tilde{z}_1-q),\nu\rangle \\ \text{s.t. } \tilde{z}_0(x)+\tilde{z}_1(y)\ge x^\top y\,. \end{aligned} \tag{7}$$

Recalling that ϕ^* is non-decreasing (Séjourné et al., 2019, Proposition 2), we can replace at the optimum $\tilde{z}_1$ by $\tilde{z}_0^*$ the Legendre of $\tilde{z}_0$. Hence we obtain the semi-dual reformulation

$$\mathrm{UOT}(\mu,\nu) = \sup_{\tilde{z}_0} \langle -\phi^*(\tilde{z}_0-q),\mu\rangle + \langle -\phi^*(\tilde{z}_0^*-q),\nu\rangle \tag{8}$$

Conversely, we can replace $\tilde{z}_0$ by its double Legendre transform $\tilde{z}_0^{**}$ which is convex. Hence, we can enforce the convexity constraint at the optimum and obtain

$$\mathrm{UOT}(\mu,\nu) = \sup_{\tilde{z}_0} \langle -\phi^*(\tilde{z}_0-q),\mu\rangle + \langle -\phi^*(\tilde{z}_0^*-q),\nu\rangle + \iota_{\mathrm{CVX}}(z_0) \tag{9}$$

The Legendre transform $z \mapsto z^*$ is itself pointwise convex. Indeed $(tz_0+(1-t)z_1)^*(y) = \sup_x x^\top y - tz_0(y) - (1-t)z_1(y) = \sup_x t(x^\top y - z_0(x)) + (1-t)(x^\top y - z_1(x)) \le tz_0^*(y) + (1-t)z_1^*(y)$. Using again the fact that ϕ^* is non-decreasing, we have

$$\begin{aligned} J(tz_0+(1-t)z_1) &= \langle \phi^*(tz_0+(1-t)z_1-q),\mu\rangle + \langle \phi^*((tz_0+(1-t)z_1)^*-q),\nu\rangle \\ &\le \langle \phi^*(t(z_0-q)+(1-t)(z_1-q)),\mu\rangle + \langle \phi^*(t(z_0^*-q)+(1-t)(z_1^*-q)),\nu\rangle\,. \end{aligned}$$

Using the convexity of ϕ^*, we recover

$$J(tz_0+(1-t)z_1) \le tJ(z_0)+(1-t)J(z_1)\,. \tag{10}$$

The formula for the first derivative comes from the differentiation of the Legendre transform w.r.t. to z, the envelope theorem gives the result. Indeed, one has $z^*(p) = \sup_x p^\top x - z(x)$. Assuming that z is strongly convex, it defines a unique supremum $\nabla z^*(p)$ and the envelope theorem gives

$$\frac{\delta z^*}{\delta z} = -(\delta z)(\nabla z^*(p))\,. \tag{11}$$

Now, one has, ϕ being differentiable, $DJ(z)(\delta z) = \langle \phi^*(z-q)\delta z,\mu\rangle + \langle -\phi^*(z^*-q)(\delta z)(\nabla z^*(p)),\nu\rangle = \langle \delta z, \phi^*(z-q)\mu - (\nabla z^*)_\#(\phi^*(z^*-q)\nu)\rangle$. Note that the measures $\phi^*(z^*-q)\nu$ and $\phi^*(z-q)\mu$ are well defined since $\phi^*(z-q)$ and $\phi^*(z^*-q)$ are continuous functions and ν,μ Radon measures.

□

8.2. Proposition 2.4

Proof. Let us start by computing the Bregman divergence $\Delta_J(f,g)$. Since we assumed g^* differentiable over the support of ν, we have $DJ(g) = [\mu]_g - \nabla g^*([\nu]_{g^*})$ where for a measure β we denote $[\beta]_h = (\phi^*)'(h-q)$. Hence, we can write

$$\begin{aligned}\Delta_J(f,g) &= J(f) - J(g) - \langle DJ(g), f-g\rangle \\ &= \langle \phi^*(f-q),\mu\rangle + \langle \phi^*(f^*-q),\nu\rangle - \langle \phi^*(g-q),\mu\rangle + \langle \phi^*(g^*-q),\nu\rangle \\ &\quad - \langle f-g, [\mu]_g - (\nabla g^*)_\#([\nu]_{g^*})\rangle \\ &= \langle \phi^*(f-q) - \phi^*(g-q),\mu\rangle + \langle \phi^*(f^*-q) - \phi^*(g^*-q),\nu\rangle \\ &\quad - \langle f-g, (\phi^*)'(g-q)\cdot\mu - (\nabla g^*)_\#([\nu]_{g^*})\rangle\,.\end{aligned}$$

Recall that μ,ν have their support included in some (centered) ball B_R and that (f,g) (resp. (f^*,g^*)) are bounded by K_R (resp. K_R^*) on B_R. Denoting $K = \max(K_R, K_R^*)$, the Taylor-Lagrange theorem applied to ϕ^* at order 2 gives the upper-bounds

$$\begin{cases}\phi^*(f-q) - \phi^*(g-q) \le (\phi^*)'(g-q)(f-g) + \frac{S_K}{2}(f-g)^2 \\ \phi^*(f^*-q) - \phi^*(g^*-q) \le (\phi^*)'(g^*-q)(f^*-g^*) + \frac{S_K}{2}(f^*-g^*)^2\,,\end{cases}$$

where $S_K = \sup_{|t|\le K+q(R)}(\phi^*)''(t)$, and the lower bounds

$$\begin{cases}\phi^*(f-q) - \phi^*(g-q) \ge (\phi^*)'(g-q)(f-g) + \frac{I_K}{2}(f-g)^2 \\ \phi^*(f^*-q) - \phi^*(g^*-q) \ge (\phi^*)'(g^*-q)(f^*-g^*) + \frac{I_K}{2}(f^*-g^*)^2\,,\end{cases}$$

where $I_K = \inf_{|t|\le K+q(R)}(\phi^*)''(t)$. We inject these bounds in Δ_J and with the cancellation of the linear term $(\phi^*)'(g-q)(f-g)$, we obtain as a lower bound on Δ_J

$$\frac{I_K}{2}H_{\mu,\nu}(f,g)^2 + \langle (\phi^*)'(g^*-q)(f^*-g^*),\nu\rangle + \langle f-g, (\nabla g^*)_\#([\nu]_{g^*})\rangle\,, \tag{LB}$$

and the upper-bound

$$\frac{S_K}{2}H_{\mu,\nu}(f,g)^2 + \langle (\phi^*)'(g^*-q)(f^*-g^*),\nu\rangle + \langle f-g, (\nabla g^*)_\#([\nu]_{g^*})\rangle\,, \tag{UB}$$

where we denoted $H_{\mu,\nu}(f,g)^2 = \|f-g\|_\mu^2 + \|f^*-g^*\|_\nu^2$.

We now focus on the term $\langle (\phi^*)'(g^*-q)(f^*-g^*),\nu\rangle + \langle f-g, (\nabla g^*)_\#([\nu]_{g^*})\rangle$. We can re-write it as $\langle f^*\circ\nabla g - g^*\circ\nabla g + (f-g), (\nabla g^*)_\#([\nu]_{g^*})\rangle$ and we denote the pointwise integrand $\Gamma_{f,g}(x) = f^*(\nabla g(x)) - g^*(\nabla g(x)) + (f(x)-g(x))$. Now recall that the Legendre identity gives $g^*(\nabla g(x)) = \nabla g(x)^\top x - g(x)$ and $f(x) = x^\top\nabla f(x) - f^*(\nabla f(x))$, hence we have

$$\begin{aligned}\Gamma_{f,g}(x) &= f^*(\nabla g(x)) - \nabla g(x)^\top x + g(x) + x^\top\nabla f(x) - f^*(\nabla f(x)) - g(x) \\ &= f^*(\nabla g(x)) - \nabla g(x)^\top x + x^\top\nabla f(x) - f^*(\nabla f(x)) \\ &= f^*(\nabla g(x)) - f^*(\nabla f(x)) - x^\top(\nabla g(x) - \nabla f(x))\,.\end{aligned}$$

Finally, recalling $x = \nabla f^*(\nabla f(x))$, we can re-write $\Gamma_{f,g}(x)$ as a Bregman divergence

$$\Gamma_{f,g}(x) = \Delta_{f^*}(\nabla g(x), \nabla f(x))\,. \tag{12}$$

Conversely, the term $\langle (\phi^*)'(g^*-q)(f^*-g^*),\nu\rangle + \langle f-g, (\nabla g^*)_\#([\nu]_{g^*})\rangle$ can be re-written as $\langle f^*-g^* + f\circ\nabla g^* - g\circ\nabla g^*, [\nu]_{g^*}\rangle$. We observe that the integrand can be written $\Gamma_{f^*,g^*}(y) = \Delta_f(\nabla g^*(y), \nabla f^*(y))$. Hence when f is λ-strongly convex $\Gamma_{f,g}(x) \le \frac{1}{2\lambda}\|\nabla g(x) - \nabla f(x)\|^2$ and $\Gamma_{f^*,g^*}(y) \ge \frac{\lambda}{2}\|\nabla g^*(y) - \nabla f^*(y)\|^2$ which yields the following bound on Δ_J

$$\frac{\lambda}{2}\|\nabla f^* - \nabla g^*\|^2_{(\nabla g^*)_\#([\nu]_{g^*})} + \frac{I_K}{2}H_{\mu,\nu}(f,g)^2 \le \Delta_J(f,g) \le \frac{1}{2\lambda}\|\nabla f - \nabla g\|^2_{(\nabla g^*)_\#([\nu]_{g^*})} + \frac{S_K}{2}H_{\mu,\nu}(f,g)^2\,.$$

Conversely, when f is M-smooth

$$\frac{1}{2M}\|\nabla f - \nabla g\|^2_{(\nabla g^*)_\#([\nu]_{g^*})} + \frac{I_K}{2}H_{\mu,\nu}(f,g)^2 \le \Delta_J(f,g) \le \frac{M}{2}\|\nabla f^* - \nabla g^*\|^2_{(\nabla g^*)_\#([\nu]_{g^*})} + \frac{S_K}{2}H_{\mu,\nu}(f,g)^2\,.$$

□

8.3. Corollary 2.5

Proof. We derive an upper-bound of $\|f^* - g^*\|_\nu^2$ that solely depends on the difference $f-g$. We start to re-write this quantity as $\|f^* \circ \nabla g - g^* \circ \nabla g\|^2_{(\nabla g^*)_\#(\nu)}$ and use the Legendre identities $g^*(\nabla g(x)) = \nabla g(x)^\top x - g(x)$ and $f^*(y) = y^\top \nabla f^*(y) - f(\nabla f^*(y))$. Let us denote again the integrand $\Gamma(x) = [\nabla g(x)^\top \nabla f^*(\nabla g(x)) - f(\nabla f^*(\nabla g(x))) - \nabla g(x)^\top x + g(x)]^2$ and re-write Γ as

$$\begin{aligned}\Gamma(x) &= [\nabla g(x)^\top \nabla f^*(\nabla g(x)) - f(\nabla f^*(\nabla g(x))) - \nabla g(x)^\top x + g(x)]^2 \\ &= [\nabla g(x)^\top (\nabla f^*(\nabla g(x)) - x) + g(x) - f(x) + f(x) - f(\nabla f^*(\nabla g(x)))]^2 \\ &\leq 3[(\nabla g(x)^\top (\nabla f^*(\nabla g(x)) - x))^2 + (g(x) - f(x))^2 + (f(x) - f(\nabla f^*(\nabla g(x))))^2]\end{aligned}$$

The integration middle term readily gives $\|f - g\|^2_{(\nabla g^*)_\#(\nu)}$. Using Cauchy-Schwartz and the fact that measures are supported over B_R, the integration of the first term can be upper-bounded as

$$\begin{aligned}\int (\nabla g(x)^\top (\nabla f^*(\nabla g(x)) - x))^2 (\,\mathrm{d}\nabla g^*(\nu))(x) &= \int (y^\top (\nabla f^*(y) - \nabla g^*(y)))^2 \,\mathrm{d}\nu(y) \\ &\leq R^2 \|\nabla f^* - \nabla g^*\|_\nu^2 \,.\end{aligned}$$

The previous results give in the balanced case $\|\nabla f^* - \nabla g^*\|_\nu^2 \leq \frac{1}{\lambda^2}\|\nabla f - \nabla g\|^2_{(\nabla g^*)_\#(\nu)}$ which yields the upper bound on the first term

$$\int (\nabla g(x)^\top (\nabla f^*(\nabla g(x)) - x))^2 (\,\mathrm{d}\nabla g^*(\nu))(x) \leq \frac{R^2}{\lambda^2}\|\nabla f - \nabla g\|^2_{(\nabla g^*)_\#(\nu)} \,. \tag{13}$$

Using the fact that $\nabla f^*(B_R), \nabla g^*(B_R) \subset B_{R^*}$ and that f is L lipschitz over B_{R^*}, we can bound the integration of the third term of $\Gamma(x)$ as

$$\begin{aligned}\int (f(x) - f(\nabla f^*(\nabla g(x))))^2 (\,\mathrm{d}\nabla g^*(\nu))(x) &= \int (f(\nabla g^*(y)) - f(\nabla f^*(y)))^2 \,\mathrm{d}\nu(y) && (14)\\ &\leq L^2 \|\nabla g^* - \nabla f^*\|_\nu^2 && (15)\\ &\leq \frac{L^2}{\lambda^2}\|\nabla f - \nabla g\|^2_{(\nabla g^*)_\#(\nu)} \,. && (16)\end{aligned}$$

□

8.4. Theorem 3.3

In this paragraph, to avoid confusions, we shall denote $\|\cdot\|_{L^2(\mu)}$ the L^2 norm and $\|\cdot\|_{L^\infty(\mu)}$ the supremum norm over a measure μ ; by abuse of notation, the supremum norm over some set S shall be denoted $\|\cdot\|_{L^\infty(S)}$.

Proof. To prove Theorem 3.3 we need to ensure that the Legendre transform is Lipschitz with respect to the supremum on a certain ball. The following lemma explicitly gives the ball to consider.

Lemma 8.1. *For all z that are λ-strongly convex and such that $z \geq l$, $\|z\|_{L^\infty(B_R)} \leq b(r)$, we have $\|\nabla z^*\|_{L^\infty(B_R)} \leq G(r) \coloneqq \frac{r}{\lambda} + \sqrt{\frac{2(b(0)-l)}{\lambda}}$ and $\|z^*\|_{L^\infty(B_R)} \leq b'(r) \coloneqq rG(r) + b(G(r))$.*

Proof. For $z \in C$, we have that z^* is $\frac{1}{\lambda}$-smooth. In particular, for $x \in B_r$

$$\begin{aligned}\|\nabla z^*(x)\| &= \|\nabla z^*(x) - \nabla z^*(0) + \nabla z^*(0)\| && (17)\\ &\leq \|\nabla z^*(x) - \nabla z^*(0)\| + \|\nabla z^*(0)\| && (18)\\ &\leq \frac{r}{\lambda} + \|\nabla z^*(0)\| \,. && (19)\end{aligned}$$

Now recall that $\nabla z^*(0) = \arg\min_{x \in \mathbb{R}^d} z(x)$. Since z is λ-strongly convex, we have the following inequality

$$z(0) \geq z(x_*) + \frac{\lambda}{2}\|x_*\|^2 \,, \tag{20}$$

where $x_* = \arg\min_{x\in\mathbb{R}^d} z(x)$. Using that $z(0) \leq b(0)$ and $-z \leq -l$, we recover

$$\|x_*\| \leq \sqrt{\frac{2(b(0)-l)}{\lambda}}\,. \tag{21}$$

The bound on $\|z^*\|_{L^\infty(B_R)}$ follows the definition of the Fenchel-Legendre transform

$$z^*(x) = x^\top \nabla z^*(x) - z(\nabla z^*(x))\,. \tag{22}$$

□

Using the previous estimates, we can now prove that the Legendre transform is Lipschitz.

Lemma 8.2. *Let z_1, z_2 be λ-strongly convex functions such that z_1, z_2 are lower-bounded by l and bounded by $b(r)$ on B_r. We have $\|z_1^* - z_2^*\|_{L^\infty(B_R)} \leq \|z_1 - z_2\|_{L^\infty(B_{G(R)})}$, where $G(r) \coloneqq \frac{r}{\lambda} + \sqrt{\frac{2(b(0)-l)}{\lambda}}$ as in Lemma 8.1.*

Proof. Let $x \in B_R$. By definition of the Fenchel transform, we have for all $y \in \mathbb{R}^d$

$$z_1^*(x) \geq x^\top y - z_1(y)\,, \tag{23}$$

with equality when $y = \nabla z_1^*(x)$. Hence, we have for all y

$$z_1^*(x) - z_2^*(x) \geq x^\top y - z_1(y) + z_2(\nabla z_2^*(x)) - x^\top \nabla z_2^*(x)\,. \tag{24}$$

In particular, for $y = \nabla z_2^*(x)$, we obtain

$$z_1^*(x) - z_2^*(x) \geq z_2(\nabla z_2^*(x)) - z_1(\nabla z_2^*(x))\,, \tag{25}$$

and applying Lemma 8.1 yields $z_1^*(x) - z_2^*(x) \geq -\|z_1 - z_2\|_{L^\infty(B_{G(R)})}$. Conversely, flipping the role of z_1, z_2, we obtain

$$z_2^*(x) - z_1^*(x) \geq z_1(\nabla z_1^*(x)) - z_2(\nabla z_1^*(x))\,, \tag{26}$$

which yields $|z_1^*(x) - z_2^*(x)| \leq \|z_1 - z_2\|_{L^\infty(B_G(R))}$. □

We have now all the ingredients to prove the first part of Theorem 3.3.

Proof. We start by applying the strong convexity inequality of the semi-dual and the optimality conditions

$$\begin{aligned} d_\phi^\lambda(\hat{z}, z_0)^2 &\leq J(\hat{z}) - J(z_0) && (27)\\ &= J(\hat{z}) - \hat{J}(\hat{z}) + \hat{J}(\hat{z}) - \hat{J}(\tilde{z}_0) + \hat{J}(z_0) - J(z_0)\,. && (28)\end{aligned}$$

Using Assumption (iii), the term $\hat{J}(\hat{z}) - \hat{J}(\tilde{z}_0)$ is negative hence we have

$$\begin{aligned} d_\phi^\lambda(\hat{z}, z_0)^2 &\leq J(\hat{z}) - \hat{J}(\hat{z}) + \hat{J}(z_0) - J(z_0) && (29)\\ &\leq \sup_{z\in C} \langle \phi^*(z-q), \mu - \hat{\mu}\rangle && (30)\\ &\quad + \sup_{z\in C^*} \langle \phi^*(z-q), \nu - \hat{\nu}\rangle && (31)\\ &\quad + \hat{J}(z_0) - J(z_0)\,, && (32)\end{aligned}$$

where we denoted $C^* = \{z^*, z \in C\}$.

Bound on term (30) Denoting $C_0 = \{\phi^*(g-q), g \in C\}$, we apply Luxburg & Bousquet (2004, Theorem 16) to bound our empirical process

$$W \coloneqq \sup_{z \in C} \langle \phi^*(z-q), \mu - \hat{\mu} \rangle \,,$$

and we obtain for all $\delta > 0$

$$W \leq 2\delta + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\frac{\delta}{4}}^{\infty} \sqrt{n(C_0, L^2(\hat{\mu}), u)}\, \mathrm{d}u \,. \tag{33}$$

Noting that $\|g\|_{L^2(\hat{\mu})} \leq \|g\|_{L^\infty(\mu)}$ almost surely, we recover the upper bound

$$\mathbb{E}[W] \leq 2\delta + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\frac{\delta}{4}}^{\infty} \sqrt{n(C_0, L^\infty(\mu), u)}\, \mathrm{d}u \,. \tag{34}$$

Since the functions in C are uniformly bounded by $b(R)$ on B_R and that μ is supported on B_R, we have $\forall (g_1, g_2) \in C^2$,

$$\|\phi^*(g_1 - q) - \phi^*(g_2 - q)\|_{L^\infty(\mu)} \leq L^1_{\phi^*} \|g_1 - g_2\|_{L^\infty(\mu)} \,, \tag{35}$$

where $L^1_{\phi^*}$ is defined as

$$L^1_{\phi^*} \coloneqq \sup_{x \in [-M_1, M_1]} |\partial \phi^*(x)| \,, \tag{36}$$

and $M_1 = 2b(R) + R^2$. In particular, we get the new upper-bound for all $\frac{\delta}{4} \leq \frac{2b(R)}{L^1_{\phi^*}}$

$$\begin{aligned}
\mathbb{E}[W] &\leq 2\delta + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\frac{\delta}{4}}^{\frac{2b(R)}{L^1_{\phi^*}}} \sqrt{n(C, L^\infty(\mu), L^1_{\phi^*} u)}\, \mathrm{d}u \\
&\leq 2\delta + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\frac{\delta}{4}}^{\frac{2b(R)}{L^1_{\phi^*}}} \sqrt{n(C, L^\infty(B_R), L^1_{\phi^*} u)}\, \mathrm{d}u \,.
\end{aligned}$$

Bound on term (31) Lemma 8.1 ensures that the functions in C^* are uniformly bounded on every ball B_r by some constant $b'(r)$. In particular, we can proceed as in the last paragraph and obtain

$$\mathbb{E}[W^*] \leq 2\delta + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\frac{\delta}{4}}^{\frac{2b'(R)}{L^2_{\phi^*}}} \sqrt{n(C^*, L^\infty(B_R), L^2_{\phi^*} u)}\, \mathrm{d}u \,,$$

where $W^* \coloneqq \sup_{z \in C^*} \langle z, \nu - \hat{\nu} \rangle$ and $L^2_{\phi^*}$ is defined as

$$L^2_{\phi^*} \coloneqq \sup_{x \in [-M_2, M_2]} |\partial \phi^*(x)| \,, \tag{37}$$

with $M_2 = 2b'(R) + R^2$. Using Lemma 8.2 that states

$$\|z_1^* - z_2^*\|_{L^\infty(B_R)} \leq \|z_1 - z_2\|_{L^\infty(B_{G(R)})} \,, \tag{38}$$

for some constant $G(R)$, we can control the covering number of C^* with respect to the $L^\infty(B_R)$ and we have the upper-bound for $\frac{\delta}{4} \leq \frac{2b'(R)}{L^2_{\phi^*}}$

$$\mathbb{E}[W^*] \leq 2\delta + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\frac{\delta}{4}}^{\frac{2b'(R)}{L^2_{\phi^*}}} \sqrt{n(C, L^\infty(B_{G(R)}), L^2_{\phi^*} u)}\, \mathrm{d}u \,.$$

Final upper bound Since the term (32) is zero in average, we obtain our final bound

$$d^\lambda_\phi(\hat{z}, z_0)^2 \leq 4\delta + \frac{8\sqrt{2}}{\sqrt{n}} \int_{\frac{\delta}{4}}^{\frac{M'}{L}} \sqrt{n(C, L^\infty(B_{R'}), Lu)} \mathrm{d}u \,,$$

where $M' = 2\max(b(R), b'(R))$ and $L = \max(L^1_{\phi^*}, L^2_{\phi^*})$ □

We now prove the second part of Proposition 3.3. For this we need to control the *localized* empirical process

$$W(\tau) \coloneqq \sup_{z \in C \cap B^\circ(z_0,\tau)} \langle \phi^*(z-q) - \phi^*(z_0-q), \mu - \hat{\mu} \rangle \,, \tag{39}$$

and

$$W^*(\tau) \coloneqq \sup_{z \in C \cap B^\circ(z_0,\tau)} \langle \phi^*(z^*-q) - \phi^*(z_0^*-q), \nu - \hat{\nu} \rangle \,, \tag{40}$$

where $B^\circ(z_0,\tau)$ is the ball centered on z_0 of radius τ with respect to the d_ϕ^λ pseudo-norm.

Lemma 8.3. *Under Assumptions (iv)-(v), if we assume that there exists (P_μ, P_ν) and $\alpha < 2$ such that for every $u \in \mathbb{R}_{\geq 0}$, $n(C, L^2(\mu), u) \leq P_\mu u^{-\alpha}$ and $n(C, L^2(\nu), u) \leq P_\nu u^{-\alpha}$, it holds with probability at least $1 - e^{-t}$*

$$\begin{cases} W(\tau) \leq \frac{\sqrt{P_\mu}(K\tau)^{1-\alpha/2}}{(1-\alpha/2)\sqrt{n(L^1_{\phi^*})^\alpha}} \left(1 + \frac{M\sqrt{P_\mu}}{(1-\alpha/2)\sqrt{n(L^1_{\phi^*})^\alpha}(K\tau)^{1+\alpha/2}}\right) + K\tau\sqrt{\frac{2t}{n}} + \frac{2b(R)L^1_{\phi^*}}{n} \\ W^*(\tau) \leq \frac{\sqrt{P_\nu}(K'\tau)^{1-\alpha/2}}{(1-\alpha/2)\sqrt{n(L^2_{\phi^*})^\alpha}} \left(1 + \frac{M\sqrt{P_\mu}}{(1-\alpha/2)\sqrt{n(L^2_{\phi^*})^\alpha}(K\tau)^{1+\alpha/2}}\right) + K'\tau\sqrt{\frac{2t}{n}} + \frac{2b'(R)L^2_{\phi^*}}{n} \,, \end{cases} \tag{41}$$

where $L^1_{\phi^*}, L^2_{\phi^*}$ are defined in Equations (36) and (37) respectively and measure local lipschitz behaviors of φ^*, $b(R)$ is defined in Assumption (iv) and is a uniform bound over B_R of the potentials in C, $b'(R)$ is defined in Lemma 8.1 and is a uniform bound over B_R of the conjugate of the potentials in C, and $K = K(R, M, \phi^*)$, $K' = K'(R, b, \phi^*, \lambda, l)$ are such that for $(f,g) \in C$, $\|f - g\|_{L^2(\mu)} \leq K d_\phi^\lambda(f,g)$ and $\|f^* - g^*\|_{L^2(\nu)} \leq K' d_\phi^\lambda(f,g)$.

Proof. The proof relies on the Lipschitz behavior of the Legendre transform that preserves the metric entropy of C and on the Bousquet concentration inequality. We start by analyzing the term $W(\tau)$.

Term $W(\tau)$ Let us denote $C_0 = \{\phi^*(z-q) - \phi^*(z_0-q), z \in C \cap B^\circ(z_0,\tau)\}$. For $g \in C_0$ of the form $g = \phi^*(z-q) - \phi^*(z_0-q)$ with $z \in C \cap B^\circ(z_0,\tau)$, we have the pointwise bound for all $x \in B_R$,

$$|g(x)| \leq L^1_{\phi^*} |z(x) - z_0(x)| \,, \tag{42}$$

where $L^1_{\phi^*} \coloneqq \sup_{x \in [-M_1, M_1]} |\partial \phi^*(x)|$ with $M_1 = 2b(R) + R^2$ as in the previous proof. This implies $\|g\|_{L^2(\mu)} \leq L^1_{\phi^*} \|z - z_0\|_{L^2(\mu)}$. Since we assumed ϕ^* strongly convex on every compact, there exists $K = K(R, M, \phi^*) > 0$ such that $\|z - z_0\|_{L^2(\mu)} \leq K d_\phi^\lambda(z, z_0)$ and in particular, all $g \in C_0$ verifies $\|g\|_{L^2(\mu)} \leq K\tau$. Conversely, since the functions in C are uniformly bounded over B_R, we have for all $g \in C_0$, $\|g\| \leq M$ where M is a constant. Defining $J(\sigma, C_0, L^2(\mu)) \coloneqq \int_0^\sigma \sqrt{1 + n(C_0, L^2(\mu), u)}\, \mathrm{d}u$, we apply Hütter & Rigollet (2021)[Theorem 25] and we recover

$$\mathbb{E}[W(\tau)] \lesssim \frac{J(K\tau, C_0, L^2(\mu))}{\sqrt{n}} \left(1 + \frac{M J(K\tau, C_0, L^2(\mu))}{\sqrt{n} K^2 \tau^2}\right). \tag{43}$$

Again, taking $(g_1, g_2) \in C_0^2$ of the form $g_1 = \phi^*(z_1 - q) - \phi^*(z_0 - q)$ and $g_2 = \phi^*(z_2 - q) - \phi^*(z_0 - q)$ with $(z_1, z_2) \in (C \cap B^\circ(z_0,\tau))^2$, we have

$$\|g_1 - g_2\|_{L^2(\mu)} \leq L^1_{\phi^*} \|z_1 - z_2\|_{L^2(\mu)} \,, \tag{44}$$

and in particular, we recover the upper-bound

$$\mathbb{E}[W(\tau)] \lesssim \frac{J(L^1_{\phi^*} K\tau, C, L^2(\mu))}{\sqrt{n} L^1_{\phi^*}} \left(1 + \frac{M J(L^1_{\phi^*} K\tau, C, L^2(\mu))}{L^1_{\phi^*} \sqrt{n} K^2 \tau^2}\right). \tag{45}$$

Now, we assumed that for all $u \in \mathbb{R}^+$ we had the upper-bound, $n(C, L^2(\mu), u) \leq P_\mu u^{-\alpha}$ with $\alpha < 2$, the term $J(L^1_{\phi^*} K\tau, C, L^2(\mu))$ can be upper bounded by $\frac{\sqrt{P_\mu}(L^1_{\phi^*} K\tau)^{1-\alpha/2}}{1-\alpha/2}$ hence we get

$$\mathbb{E}[W(\tau)] \lesssim \frac{\sqrt{P_\mu}(K\tau)^{1-\alpha/2}}{(1-\alpha/2)\sqrt{n(L^1_{\phi^*})^\alpha}} \left(1 + \frac{M\sqrt{P_\mu}}{(1-\alpha/2)\sqrt{n(L^1_{\phi^*})^\alpha}(K\tau)^{1+\alpha/2}}\right). \tag{46}$$

There remains to bound the process $W(\tau)$ with high probability. We use for this the Bousquet concentration inequality.

Lemma 8.4 (Bousquet, see Theorem 26 in Hütter & Rigollet (2021))**.** *Let $\mathcal{F}$ be a class of functions such that for every $f \in \mathcal{F}$, $\|f\|^2_{L^2(\mu)} \leq \sigma^2$ and $\|f\|_{L^\infty(\mu)} \leq M$, then for all $t > 0$, we have with probability at least $1 - e^{-t}$*

$$\sup_{f\in F} \sqrt{n}|\langle f, \mu - \hat{\mu}\rangle| \leq 2\mathbb{E}[\sup_{f\in F} \sqrt{n}|\langle f, \mu - \hat{\mu}\rangle|] + \sigma\sqrt{2t} + \frac{M}{\sqrt{n}}t\,. \tag{47}$$

Applying this result to $W(\tau)$ yields that with probability at least $1 - e^{-t}$,

$$W(\tau) \leq \frac{\sqrt{P_\mu}(K\tau)^{1-\alpha/2}}{(1-\alpha/2)\sqrt{n(L^1_{\phi^*})^\alpha}}\left(1 + \frac{M\sqrt{P_\mu}}{(1-\alpha/2)\sqrt{n(L^1_{\phi^*})^\alpha}(K\tau)^{1+\alpha/2}}\right) + K\tau\sqrt{\frac{2t}{n}} + \frac{2tb(R)L^1_{\phi^*}}{n}\,, \tag{48}$$

where we used the pointwise upper-bound (42) and where $b(R)$ is the constant such that $\forall z \in C, \|z\|_{L^\infty(B_R)} \leq b(R)$.

Term $W^*(\tau)$ We can apply the same reasoning as previously. Indeed, as shown in Lemma 8.1, there exists a constant $b'(R)$ such that for all $z \in C$, $\|z^*\|_{L^\infty(B_R)} \leq b'(R)$. In particular, since the potentials z^* are bounded, we can also leverage the local strong convexity of ϕ^* that yields a constant $K' = K'(R, M, \phi^*, \lambda, l) > 0$ such that for every $z \in C, \|(z - z_0)^*\|_\nu \leq K' d^\lambda_\phi(z, z_0)$. Hence we recover that with probability at least $1 - e^{-t}$,

$$W^*(\tau) \leq \frac{\sqrt{P_\nu}(K'\tau)^{1-\alpha/2}}{(1-\alpha/2)\sqrt{n(L^2_{\phi^*})^\alpha}}\left(1 + \frac{M\sqrt{P_\mu}}{(1-\alpha/2)\sqrt{n(L^2_{\phi^*})^\alpha}(K\tau)^{1+\alpha/2}}\right) + K'\tau\sqrt{\frac{2t}{n}} + \frac{2b'(R)L^2_{\phi^*}}{n}\,. \tag{49}$$

□

We can now prove the second part of Proposition 3.3.

Proof. For $\tau > 0$, define $s = \frac{\tau}{\tau + d^\lambda_\phi(\hat{z}, z_0)}$ and $\hat{z}_s = (1-s)z_0 + s\hat{z}$. By local strong convexity of J, we have

$$d^\lambda_\phi(\hat{z}_s, z_0)^2 \leq J(\hat{z}_s) - J(z_0)\,. \tag{50}$$

Let us decompose the right hand side as $J(\hat{z}_s) - \hat{J}(\hat{z}_s) - (J(z_0) - \hat{J}(z_0)) + \hat{J}(\hat{z}_s) - \hat{J}(z_0)$. By convexity of $\hat{J}$, the last term can be upper-bounded by $s\hat{J}(\hat{z}) + (1-s)\hat{J}(z_0) - \hat{J}(z_0) = s(\hat{J}(\hat{z}) - \hat{J}(z_0))$. Since $\hat{z}$ is the minimizer of the empirical semi-dual, we have in particular that $s(\hat{J}(\hat{z}) - \hat{J}(z_0)) \leq 0$ which gives

$$\begin{aligned} d^\lambda_\phi(\hat{z}_s, z_0)^2 &\leq J(\hat{z}_s) - \hat{J}(\hat{z}_s) - (J(z_0) - \hat{J}(z_0)) \\ &= \langle \phi^*(\hat{z}_s - q) - \phi^*(z_0 - q), \mu - \hat{\mu}\rangle + \langle \phi^*(\hat{z}^*_s - q) - \phi^*(z^*_0 - q), \nu - \hat{\nu}\rangle\,. \end{aligned}$$

Now, since $d^\lambda_\phi(\hat{z}_s, z_0) = \frac{\tau d^\lambda_\phi(\hat{z}, z_0)}{\tau + d^\lambda_\phi(\hat{z}, z_0)} \leq \tau$, we recover in the end $d^\lambda_\phi(\hat{z}_s, z_0)^2 \leq W(\tau) + W^*(\tau)$.

Let us now consider $A = \{\tau, d^\lambda_\phi(\hat{z}, z_0) \geq \tau\}$. We wish to recover an upper-bound on A. Remark that $A = \{\tau, d^\lambda_\phi(\hat{z}_s, z_0) \geq \frac{\tau}{2}\}$. In particular, every $\tau \in A$ verifies with probability at least $1 - e^{-t}$

$$\frac{\tau^2}{4} \leq \kappa\frac{\tau^{1-\alpha/2}}{\sqrt{n}} + \kappa'\frac{\tau^{-\alpha}}{n} + (K + K')\tau\sqrt{\frac{2t}{n}} + \frac{t\kappa''}{n}\,, \tag{51}$$

where κ and κ' are given in Lemma 8.3 defined as

$$\begin{cases} \kappa = \frac{8\sqrt{2}}{(1-\frac{\alpha}{2})}\left[\frac{\sqrt{P_\mu}K^{1-\alpha/2}}{(L^1_{\varphi^*})^{\frac{\alpha}{2}}} + \frac{\sqrt{P_\nu}(K')^{1-\alpha/2}}{(L^2_{\varphi^*})^{\frac{\alpha}{2}}}\right] \\ \kappa' = \frac{8\sqrt{2}M}{(1-\frac{\alpha}{2})^2}\left[\frac{P_\mu K^{-\alpha}}{(L^1_{\varphi^*})^\alpha} + \frac{P_\nu (K')^{-\alpha}}{(L^2_{\varphi^*})^\alpha}\right] \\ \kappa'' = 2(M(R)L^1_{\varphi^*} + M'(R)L^2_{\varphi^*})\,. \end{cases} \tag{52}$$

Let $A_n = \{\tau \in A, \tau \geq \frac{1}{\sqrt{n}}\}$. For $\tau \in A_n$, we have

$$\frac{\tau^2}{4} \leq \kappa \frac{\tau^{1-\alpha/2}}{\sqrt{n}} + \frac{\kappa' \tau^{-\alpha}}{n} + (K + K')\tau\sqrt{\frac{2t}{n}} + \frac{t\kappa'' \tau}{\sqrt{n}}\,. \tag{53}$$

Now since we assumed $\tau \geq \frac{1}{\sqrt{n}}$, we have in particular $\frac{\tau^{-\alpha}}{n} \leq \frac{\tau^{1-\alpha}}{\sqrt{n}}$ hence we get

$$\frac{\tau^2}{4} \leq (\kappa + \kappa') \frac{\tau^{1-\alpha/2}}{\sqrt{n}} + (K + K')\tau\sqrt{\frac{2t}{n}} + \frac{t\kappa'' \tau}{\sqrt{n}}\,. \tag{54}$$

Assuming that $t \geq 1$, we have two cases

Case 1 If $\tau \leq 1$, we have

$$\frac{\tau^2}{4} \leq \frac{t\eta\tau^{1-\alpha/2}}{\sqrt{n}}\,, \tag{55}$$

where $\eta = (\kappa + \kappa' + \kappa'' + \sqrt{2}(K + K'))$ and we recover $\tau \leq \frac{(4\eta t)^{\frac{1}{1+\alpha/2}}}{n^{\frac{1}{2+\alpha}}}$.

Case 2 If $\tau \geq 1$, we have $\frac{\tau^2}{4} \leq \frac{t\eta\tau}{\sqrt{n}}$ *i.e.* $\tau \leq \frac{4t\eta}{\sqrt{n}}$.

In any case, for $t \geq 1$, we have with probability at least $1 - e^{-t}$

$$\sup(A) \leq \frac{(4\eta' t)^{\frac{1}{1+\alpha/2}} + (4\eta' t)}{n^{\frac{1}{2+\alpha}}}\,, \tag{56}$$

where we defined $\eta' = \max(\eta, 1)$. Now, by definition of A, we have for all $\epsilon > 0$, $d_\phi^\lambda(\hat{z}, z_0) \leq \sup(A) + \epsilon$. Taking $\epsilon \to 0$ gives that with probability at least $1 - e^{-t}$, for $t \geq 1$

$$d_\phi^\lambda(\hat{z}, z_0) \leq \frac{(4\eta' t)^{\frac{1}{1+\alpha/2}} + (4\eta' t)}{n^{\frac{1}{2+\alpha}}} \tag{57}$$

$$\leq \frac{8\eta' t}{n^{\frac{1}{2+\alpha}}}\,. \tag{58}$$

And in particular, $d_\phi^\lambda(\hat{z}, z_0)^2 \leq \frac{64(\eta')^2 t^2}{n^{\frac{1}{1+\alpha/2}}}$ with probability at least $1 - e^{-t}$ for $t \geq 1$. We denote X the random variable $d_\phi^\lambda(\hat{z}, z_0)^2$. Since X is non-negative almost surely, we can apply Fubini's formula

$$\mathbb{E}[X] = \int_0^\infty P(X > u)\,\mathrm{d}u\,. \tag{59}$$

Let us make the change of variable $u = \frac{64(\eta')^2 t^2}{n^{\frac{1}{1+\alpha/2}}}$,

$$\mathbb{E}[X] = \frac{128(\eta')^2}{n^{\frac{1}{1+\alpha/2}}} \left(\int_0^1 tP(X > \frac{64(\eta')^2 t^2}{n^{\frac{1}{1+\alpha/2}}})\,\mathrm{d}t + \int_1^\infty tP(X > \frac{64(\eta')^2 t^2}{n^{\frac{1}{1+\alpha/2}}})\,\mathrm{d}t \right).$$

The integrand in the first term is upper-bounded by 1 and the integrand on the second term is upper bounded by te^{-t}. Hence we obtain

$$\mathbb{E}[d_\phi^\lambda(\hat{z}, z_0)^2] \leq \frac{128(\eta')^2}{n^{\frac{1}{1+\alpha/2}}}(1 + \int_1^\infty te^{-t}dt)$$

$$= \frac{128(1 + 2e^{-1})(\eta')^2}{n^{\frac{1}{1+\alpha/2}}}\,.$$

□

□

8.5. Corollary 3.4

Proof. Using the Corollary 9 of Gallouët et al. (2021), we can ensure that z_0, z_0^* are $(k+2)$-times continuously differentiable over the support of μ and ν respectively. Recalling that for all $x \in \text{supp}(\nu)$

$$\nabla^2 z_0(x) = [\nabla^2 z_0^*(\nabla z_0(x))]^{-1}\,, \tag{60}$$

and using the fact that ∇z_0 is a diffeomorphism between the support μ and ν, we recover that z_0 is λ-strongly convex over $\text{supp}(\mu)$ where we defined

$$\frac{1}{\lambda} \coloneqq \sup_{y \in \text{supp}(\nu)} \|\nabla^2 z_0^*(y)\|\,. \tag{61}$$

Now, recall that in order to apply our previous result, we need to globally bound the strong-convexity constant as well as controlling the sup norm over every ball. To achieve this, we can extend these potentials to the whole domain. Proposition 1.5 in Azagra & Mudarra (2019) provides a $(k+2)$-times continuously differentiable convex extension $\tilde{g}_0$ of $z_0 - \lambda q$ on the whole domain $\mathbb{R}^d$. Defining $\tilde{z}_0 = \tilde{g}_0 + \lambda q$, we have that $\tilde{z}_0$ coincides with z_0 on $\text{supp}(\mu)$. Using again the diffeomorphism property of ∇z_0 between $\text{supp}(\mu)$ and $\text{supp}(\nu)$, we have that $\tilde{z}_0^*$ coincides with z_0^* on $\text{supp}(\nu)$. Now let us define

$$C = \{z \mid \|z\|_{L^\infty(B_r)} \leq \|\tilde{z}_0\|_{L^\infty(B_r)}, \|\nabla^{k+2} z\|_{L^\infty(B_r)} \leq \|\nabla^{k+2}\tilde{z}_0\|_{L^\infty(B_r)}, z \geq l, z \text{ is } \lambda\text{-strongly convex}\}\,,$$

where l is the minimum of $\tilde{z}_0$. The set C indeed meets Assumption (iv) and Assumption (iii) hence we can apply Prop. 3.3 which yields

$$\mathbb{E}[d_\phi^\lambda(\hat{z}_C, z_0)^2] \lesssim \delta + \frac{1}{\sqrt{n}} \int_{\frac{\delta}{4}}^{\frac{M'}{L}} \sqrt{n(C, L^\infty(B_{R'}), Lu)}\mathrm{d}u\,. \tag{62}$$

Finally, using van der Vaart & Wellner (1996, Theorem 2.7), we have $n(C, L^\infty(B_{R'}), Lu) \lesssim u^{-\frac{d}{k+2}}$. If $\frac{k+2}{d} < 1/2$, take $\delta = n^{-\frac{k+2}{d}}$. For this choice of δ,

$$\begin{aligned}
\frac{1}{\sqrt{n}} \int_{\frac{\delta}{4}}^{\frac{M'}{L}} \sqrt{n(C, L^\infty(B_{R'}), Lu)}\mathrm{d}u &\lesssim \frac{1}{\sqrt{n}}(n^{-\frac{k+2}{d}})^{1-\frac{d}{2(k+2)}} && (63)\\
&\lesssim \frac{1}{\sqrt{n}} n^{-\frac{2(k+2)-d}{2d}} && (64)\\
&= n^{-\frac{k+2}{d}}\,. && (65)
\end{aligned}$$

If $\frac{k+2}{d} = 1/2$, take $\delta = \frac{1}{\sqrt{n}}$. For this choice of δ, the integral is of order $\log(n)$ which yields the upper-bound

$$\mathbb{E}[d_\phi^\lambda(\hat{z}_C, z_0)^2] \lesssim \frac{\log(n)}{\sqrt{n}}\,. \tag{66}$$

Finally, if $\frac{k+2}{d} > 1/2$, we apply the second part of Propostion 3.3 and we recover the rate

$$\mathbb{E}[d_\phi^\lambda(\hat{z}_C, z_0)^2] \lesssim n^{-1/(1+d/2(k+2))}\,. \tag{67}$$

□

8.6. Theorem 4.1

Proof. First, we show that J is lower-bounded on C so that $\inf_{f\in C} J(f)$ is indeed well-defined. Recalling $J(f) = \langle \phi^*(f-q), \mu\rangle + \langle \phi^*(f^*-q), \nu\rangle$, we need to prove in particular that for $f \in C$, f^* is bounded on B_R. For f in C, denoting $x^* = \arg\min_x f(x)$, we have using the strong convexity that $f(x^*) \geq \frac{\lambda}{2}\|x^*\|^2 - f(0) \geq -b$ since we assumed $|f(0)| \leq b$. Furthermore, using the lipschitz property, we have $\|f\|_{L^\infty(B_r)} \leq b + L(r)r$. Hence we can apply Lemma 8.1 that yields for $f \in C$

$$\|\nabla f^*\|_{L^\infty(B_R)} \leq R^* \coloneqq \frac{R}{\lambda} + 4\sqrt{\frac{b}{\lambda}} \qquad\qquad \|f^*\|_{L^\infty(B_R)} \leq RR^* + b + R^* L(R^*)\,. \tag{68}$$

In particular, denoting $K = \max(b + L(R), b + R^*(R + L(R^*)))$ we have $J(f) \geq (m_\mu + m_\nu) \inf_{|t| \leq K + q(R)} \phi^*(t)$ with (m_μ, m_ν) the total masses of μ, ν respectively. Now we show the existence of a minimum. Since all functions of C are $L(R)$-lipschitz continuous over B_R and that for all $f, x \in C \times B_R$, $|f(x)| \leq b + L(R)$, we can apply the Arzela-Ascoli theorem ensuring that C is relatively compact in the set of continuous functions on B_R for the supremum topology. In particular, we can extract a minimizing suite from $\inf_{f \in C} J(f)$ that converges toward $\bar{f} \in C$ as C is assumed to be closed. Conversely, since the function $x \in C \mapsto dF(x_k)(x)$ is lower bounded by $m_\mu \inf_{|t| \leq K + q(R)} (\phi^*)'(t) - m_\nu \sup_{|t| \leq K + q(R)} (\phi^*)'(t)$, the iterates (x_k) are indeed well-defined using Arzela-Ascoli.

Now, applying Corollary 2.5, we have indeed for all $(f, g) \in C$

$$\Delta_J(f, g) \leq \frac{1}{\lambda} \|\nabla f - \nabla g\|^2_{(\nabla g^*)_\#([\nu]_{g^*})} + 3 S_K \left[\frac{R^2 + L(R^*)^2}{\lambda^2} \|\nabla f - \nabla g\|^2_{(\nabla g^*)_\#(\nu)} + \tilde{H}^g_{\mu,\nu}(f - g) \right], \tag{69}$$

where $[\beta]_h = \beta \times (\phi^*)'(h - q)$ and $\tilde{H}^g_{\mu,\nu}(h) = \|h\|^2_\mu + \|h\|^2_{(\nabla g^*)_\#(\nu)}$. All that remains to prove is the boundedness of $A^g(h)$ now. Since $(\nabla g^*)_\#(\nu) \subset B_{R^*}$ we have $\|\nabla f - \nabla g\|^2_{(\nabla g^*)_\#([\nu]_{g^*})} \leq 2L(R^*) \int (\phi^*)'(K + q(y)) \, \mathrm{d}\nu(y)$ (recall that $(\phi^*)'$ is a non-decreasing function). And conversely, $\|\nabla f - \nabla g\|^2_{(\nabla g^*)_\#(\nu)} \leq 2 m_\nu L(R^*)$. Finally, $\tilde{H}^g_{\mu,\nu}(f - g) \leq 2 m_\mu (b + L(R)) + 2 m_\nu (b + L(R^*))$. Using Theorem 9.1, we do recover

$$J(f_k) - \bar{J} \leq 4 \frac{\lambda I L(R^*) + 3 S_K \left[(R^2 + L(R^*)^2 + 1) m_\nu L(R^*) + m_\mu (b + L(R)) + b m_\nu \right]}{\lambda^2}, \tag{70}$$

where $I = \int (\phi^*)'(K + q(y))$. □

8.7. Theorem 4.2

Proof. As in the previous proof, the minimum and the the iterates are well-defined thanks to the Arzela-Ascoli theorem. The convergence rate follows the stability results in the smooth, strongly convex unbalanced case

$$\frac{1}{2M} \|\nabla f - \nabla g\|^2_{(\nabla g^*)_\#(\nu)} \leq \Delta_J(f, g) \leq \frac{1}{2\lambda} \|\nabla f - \nabla g\|^2_{(\nabla g^*)_\#(\nu)}. \tag{71}$$

□

8.8. Proposition 5.1

Proof. Let us first re-write the convexity constraint as a discrete constraint. For $f = g + \lambda q \in C$, let us denote $y \in \mathbb{R}^{2n+1}$ the value of g on the points $\mathbf{x} \coloneqq [\vec{0}, \hat{\mu}, (\nabla f_k^*)_\#(\hat{\nu})]$ and $\mathbf{z} \in \mathbb{R}^{(2n+1) \times d}$ the value of ∇g over $\mathbf{x}$. Using Taylor et al. (2017, Theorem 3.3), the convexity of g can be enforced through the discrete constraints

$$y_i - y_j \geq \mathbf{z}_j^\top (\mathbf{x}_i - \mathbf{x}_j) \;\; \forall \, 0 \leq i, j \leq 2n + 1, \tag{72}$$

with the following interpolation: $g(x) = \max_i y_i + \mathbf{z}_i^\top (x - \mathbf{x}_i)$. In particular, f^* can be computed pointwise through the quadratic program

$$\begin{aligned} f^*(y) = \; & \max_{x,t} \; y^\top x - \lambda q(x) - t \, . \\ & t \geq y_i + \mathbf{z}_i^\top (x - \mathbf{x}_i) \end{aligned}$$

The additional bound $|g(0)| \leq b$ is explicitely given by the constraint $|y_0| \leq b$. The Lipschitz constraint can first be enforced pointwise as $\|\mathbf{z}_i\| \leq L$ for all $0 \leq i \leq 2n + 1$. With the previous interpolation, the max of L-Lipschitz function being itself L-Lipschitz , g remains globally L-Lipschitz.

Now, let us write the objective of (4) with y and $\mathbf{z}$. The objective reads $U(f) = \langle DJ(f_k), f - f_k \rangle + \frac{K}{2} \|f - f_k\|^2_{f_k}$ with $\|h\|^2_{f_k} = 3 S_K (\|h\|^2_{\hat{\mu}} + \|h\|^2_{(\nabla f_k^*)_\#(\hat{\nu})}) + \frac{1}{\lambda} \|\nabla h\|^2_{(\nabla f_k^*)_\#(\hat{\nu})} + 3 S_K \frac{R^2 + L(R^*)^2}{\lambda^2} \|\nabla h\|^2_{[(\nabla f_k^*)_\#(\hat{\nu})]_{f_k}}$ where S_K and R^* are defined in Proposition 2.4. Since $DJ(f_k) = \hat{\mu} \times (\phi^*)' \circ (f_k - q) - (\nabla f_k^*)_\#(\hat{\nu}) \times (\phi^*)' \circ (f_k^* - q)$, the linear part of U can be written up to the constant terms $c^\top y$ with c defined as

$$c \coloneqq [0, \omega^\mu \times (\phi^*)'((f_k - q)(\hat{\mu})), -\omega^\nu \times (\phi^*)'((f_k^* - q)(\hat{\nu}))] \, ,$$

where ω^μ, ω^ν are the weights of the empirical measures $\hat{\mu}$ and $\hat{\nu}$ respectively. Finally, the quadratic term can be recovered as

$$\begin{aligned}\|f - f_k\|_{f_k}^2 = 3S_K \sum_{i=1}^{n+1} \omega_i^\mu (y_i + \lambda q(\mathbf{x}_i) - f_k(\mathbf{x}_i))^2 + 3S_K \sum_{i=n+1}^{2n+1} \omega_i^\nu (y_i + \lambda q(\mathbf{x}_i) - f_k(\mathbf{x}_i))^2 \\ + \frac{3S_K(R^2 + L(R^*)^2)}{\lambda^2} \sum_{i=n+1}^{2n+1} \omega_i^\nu (\phi^*)'((f_k^* - q)(\hat{\nu}_i)) \|\mathbf{z}_i + \lambda \mathbf{x}_i - \nabla f_k(\mathbf{x}_i)\|^2 \\ + \frac{1}{\lambda} \sum_{i=n+1}^{2n+1} \omega_i^\nu \|\mathbf{z}_i + \lambda \mathbf{x}_i - \nabla f_k(\mathbf{x}_i)\|^2 .\end{aligned}$$

Hence, $\|f - f_k\|_{f_k}^2$ can be re-written as

$$\|f - f_k\|_{f_k}^2 = [y - f, \mathbf{z} - \mathbf{g}]^\top Q [y - f, \mathbf{z} - \mathbf{g}] ,$$

where $f = f_k(\hat{\mu}) - \lambda q(\hat{\mu})$, $\mathbf{g} = [\vec{0}, \hat{\nu} - \lambda \nabla f_k^*(\hat{\nu})]$ and Q is a diagonal matrix with diagonal $3S_K[0, \omega^\mu, \omega^\nu, \vec{0}, (\omega^\nu(\frac{1}{3S_K\lambda} + \frac{(\phi^*)'((f_k^* - q)(\hat{\nu}))(R^2 + L(R^*)^2)}{\lambda^2}))^{*d}]$. In the end, we do recover the quadratic program

$$\begin{aligned}\inf_{y, \mathbf{z}} \; & c^\top y + \frac{1}{2}[y - f, \mathbf{z} - \mathbf{g}]^\top Q [y - f, \mathbf{z} - \mathbf{g}] . \\ & y_i - y_j \geq \mathbf{z}_j^\top (\mathbf{x}_i - \mathbf{x}_j) \;\; \forall \, 0 \leq i, j \leq 2n + 1 \\ & \|\mathbf{z}_i\| \leq L, |y_0| \leq b\end{aligned}$$

Let us derive the complexity of solving the convex problem above using an Interior Point Method (IPM).

First we need to compute the derivative $DJ(f_k)$ which involves computing Legendre of f_k transform over the support of $\hat{\nu}$. We showed in the proof of Propostion 5.1 that the Legendre transform could be computed with a linearly constrained quadratic program with $d + 1$ variable and n constraint hence the cost to assemble the KKT system is $O(nd)$ and the cost to solve it is $O(d^3)$ which make a complexity per iteration given by $O(d^3 + nd)$; since there are n constraints, we make at most $\sqrt{n}$ iterations to converge. The overall process is repeated on all the support of $\hat{\nu}$ hence the overall complexity to compute $DJ(f_k)$ is at most $O(n\sqrt{n}(nd + d^3))$.

Now, let us focus on the cost to assemble and solve the KKT system associated with (5).Let us denote $w = [y, \mathbf{z}] \in \mathbb{R}^{(2n+1)(d+1)}$. Assembling the system involves computing $DJ(f_k)$, Qw and the gradient of the constraint which is of the form $\sum_{i=1}^{n(n-1)} \frac{a_i}{a_i^\top w - b_i}$. Since Q is diagonal Qw costs $O(nd)$ and finally, since the convexity constraint is sparse, computing $a_i^\top w$ is $O(d)$ hence computing $\sum_{i=1}^{n(n-1)} \frac{a_i}{a_i^\top w - b_i}$ is $O(n^2 d)$. Hence the cost to assemble the KKT system scales in $O(n^2 d)$ per iteration with an initial $O(n\sqrt{n}(nd + d^3))$ cost. Now let us focus on the resolution of the system. It involves solving a system of the form $Pw = t$ where $P = Q + \theta_t \sum_{i=1}^{n(n-1)} \frac{a_i a_i^\top}{(a_i^\top w - b_i)^2}$ with θ_t the magnitude of the barrier at step t. If we use a conjugate gradient method to solve the system, since the evaluation $a_i a_i^\top w$ is $O(n^2 d)$, solving the system also costs $O(n^2 d)$ (recall that Qw is $O(nd)$). Now since there are $O(n^2)$ constraints, the IPM makes at most $O(n)$ iterations, leaving us with a total complexity to solve (5) of $O(n^3 d + n\sqrt{n} d^3)$. □

8.9. Proposition 5.3

Proof. We simply apply the bound on the metric entropy of uniformly Lipschitz convex functions in Bronshtein (1976) with respect to the supremum norm

$$n(C_{\lambda,L,b}, L^\infty(B_{R'}), u) \lesssim u^{-d/2} , \tag{73}$$

which implies the following growth rates with respect to to the L^2 norms $n(C_{\lambda,L,b}, L^2(\mu), u) \lesssim u^{-d/2}$ as well as $n(C_{\lambda,L,b}, L^2(\nu), u) \lesssim u^{-d/2}$. If $d < 4$, we can apply the second part of Proposition 3.3 and recover

$$\mathbb{E}[d_\phi^\lambda(\hat{z}_{C_{\lambda,L,b}}, z_0)^2] \lesssim n^{-1/(1+d/4)} . \tag{74}$$

If $d = 4$, applying the first part of Proposition 3.3 with $\delta = 1/\sqrt{n}$ yields

$$\mathbb{E}[d_\phi^\lambda(\hat{z}_{C_{\lambda,L,b}}, z_0)^2] \lesssim \frac{\log(n)}{\sqrt{n}} .$$

Finally, if $d > 4$, we pick $\delta = n^{-2/d}$ and we recover

$$\mathbb{E}[d_\phi^\lambda(\hat{z}_{C_{\lambda,L,b}}, z_0)^2] \lesssim n^{-2/d}\,. \tag{75}$$

□

8.10. Proposition 5.4

Proof. Let us denote $\hat{z}_T$ the estimator that is obtained after making T iterations of 5 and z_0 the ground truth OT potential. In virtue of the estimates of Sec. 2, we have

$$\|\nabla \hat{z}_T - \nabla z_0\|^2_{L^2(\mu)} \lesssim J(\hat{z}_T) - J(z_0)\,. \tag{76}$$

Now, denoting $\hat{J}$ the empirical semi-dual and $\hat{z}$ its minimizer, we have

$$J(\hat{z}_T) - J(z_0) = (\hat{J}(\hat{z}_T) - \hat{J}(\hat{z})) + (J(\hat{z}_T) - \hat{J}(\hat{z}_T)) + (\hat{J}(\hat{z}) - J(\hat{z})) + (J(\hat{z}) - J(z_0))\,.$$

The term $\hat{J}(\hat{z}_T) - \hat{J}(\hat{z})$ is upper-bounded by $0(1/T)$ and using the first case of Proof 8.4, the term $J(\hat{z}) - J(z_0)$ is upper bounded in average by $O(n^{-2/d})$ if $d > 4$ or $\tilde{O}(1/\sqrt{n})$ is $d \leq 4$. Now the term $J(\hat{z}_T) - \hat{J}(\hat{z}_T)$ can be decomposed as

$$J(\hat{z}_T) - \hat{J}(\hat{z}_T) = \langle \hat{z}_T, \mu - \hat{\mu}\rangle + \langle \hat{z}_T^*, \nu - \hat{\nu}\rangle\,. \tag{77}$$

Using again the same technique as in Proof 8.4, we recover using chaining bounds that in average $J(\hat{z}_T) - \hat{J}(\hat{z}_T)$ is in $O(n^{-2/d})$ if $d > 4$ and in $\tilde{O}(1\sqrt{n})$ if $d \leq 4$ and the same goes for $\hat{J}(\hat{z}) - J(\hat{z})$. Hence the total error reads

$$\begin{cases} \|\nabla \hat{z}_T - \nabla z_0\|^2_{L^2(\mu)} \lesssim \frac{1}{T} + 1/\sqrt{n} \text{ if } d \leq 4 \\ \|\nabla \hat{z}_T - \nabla z_0\|^2_{L^2(\mu)} \lesssim \frac{1}{T} + n^{-2/d} \text{ if } d > 4\,. \end{cases} \tag{78}$$

Setting $T_\tau = \frac{1}{\tau}$ and $n_\tau = 1/\tau^2$ if $d \leq 4$ and $n_\tau = 1/\tau^{d/2}$ if $d > 4$, we do recover a τ approximation. Hence, if $d \leq 4$, the number of operations to compute $\hat{z}_\tau$ scales as $\tilde{O}(\tau^{-7})$ and if $d > 4$, it scales as $\tilde{O}(\tau^{-(\frac{3d}{2}+1)})$. □

9. Additional results

9.1. Section 4

In this paragraph, we state and prove the convergence guarantees we can obtain for the variable metric gradient scheme

$$x_{k+1} = \arg\min_{x \in C} dF(x_k)(x - x_k) + \frac{\beta}{2} A^{x_k}(x - x_k)\,, \tag{79}$$

where F is a convex function over a Banach space E, A^{x_k} is a 2-homogeneous form depending on x_k and C is a convex subset of E. We first study the variable relatively smooth case $\Delta_F(x, y) \leq \frac{\beta}{2} A^{x_k}(x - x_k)$.

Theorem 9.1 (Variable relative smoothness: sub-linear convergence)**.** *Let E be a Banach space, let F be a real-valued convex function with Gateaux derivative dF satisfying for all $(x, y) \in E$, $\Delta_F(x, y) \leq \frac{\beta}{2} A^y(x - y)$ where for all $y \in E$, $A^y(\cdot)$ is a 2-homogeneous form over E depending on y and where β is a strictly positive constant and let $C \subset E$ be a closed convex subset of E. Assuming that $\sup_{(x,y)\in C^2} A^y(x - y) \leq K$, that a minimizer $\bar{x} \in C$ exists and that the iterates $x_0 \in C$, (x_k) generated as*

$$x_{k+1} \in \arg\min_{x \in C} dF(x_k)(x - x_k) + \frac{\beta}{2} A^{x_k}(x - x_k)\,, \tag{80}$$

exist, we have $F(x_k) - F(\bar{x}) \leq \frac{2\beta K}{k+1}$.

Proof. We simply adapt the proof of Bubeck (2015, Theorem 3.8). Recall that F verifies

$$F(x_{k+1}) - F(x_k) \leq dF(x_k)(x_{k+1} - x_k) + \frac{\beta}{2} A^{x_k}(x_{k+1} - x_k)\,. \tag{81}$$

Denoting $y_k = \arg\min_C dF(x_k)(y - x_k)$, we have by definition of x_{k+1} and by convexity of C,

$$\begin{aligned} dF(x_k)(x_{k+1} - x_k) + \frac{\beta}{2} A^{x_k}(x_{k+1} - x_k) &\leq dF(x_k)(s_k y_k + (1 - s_k)x_k - x_k) \\ &\quad + \frac{\beta}{2} A^{x_k}(s_k y_k + (1 - s_k)x_k - x_k) \\ &= s_k dF(x_k)(y_k - x_k) + s_k^2 \frac{\beta}{2} A^{x_k}(y_k - x_k)\,, \end{aligned}$$

where $s_k \in [0, 1]$ is a parameter that shall be defined later. Then, by definition of y_k, $dF(x_k)(y_k - x_k) \leq dF(x_k)(\bar{x} - x_k)$ hence we recover using the convexity of F

$$F(x_{k+1}) - F(x_k) \leq s_k(F(\bar{x}) - F(x_k)) + s_k^2 \frac{\beta}{2} K\,. \tag{82}$$

Denoting $\delta_k = F(\bar{x}) - F(x_k)$ we get eventually

$$\delta_{k+1} \leq (1 - s_k)\delta_k + s_k^2 K \frac{\beta}{2}\,. \tag{83}$$

Taking $s_k = 2/(k + 1)$ yields $\delta_k = \frac{2\beta K}{k+1}$ (see the proof of Bubeck (2015, Theorem 3.8) for more details).

□

Now we show that under the additional variable relative strong convexity assumption $\Delta_F(x, y) \geq \frac{\alpha}{2} A^{x_k}(x - x_k)$, we recover exponential convergence.

Theorem 9.2 (Variable relative smoothness and strong convexity: exponential convergence)**.** *Let E be a Banach space, let F be a real-valued convex function with Gateaux derivative dF and let $C \subset E$ be a closed convex subset of E. If there exists $\alpha, \beta > 0$ and $A^y(\cdot)$ a 2-homogeneous form such that for all $(x, y) \in E$, $\frac{\alpha}{2} A^y(x - y) \leq \Delta_F(x, y) \leq \frac{\beta}{2} A^y(x - y)$. If a minimizer $\bar{x} \in C$ and the iterates $x_0 \in C$, (x_k) generated by (80) exist, we have*

$$F(x_k) - F(\bar{x}) \leq \left(1 - \frac{\alpha}{\beta}\right)^k [F(x_0) - F(\bar{x})].$$

Proof. We simply adapt the proof of Karimi et al. (2016, Theorem 5). To this end, we propose to generalize the notion of being proximal PL with respect to a (convex) set C and an operator $A^{\cdot}(\cdot)$ such that for any $y \in C$, $A^y(\cdot)$ is a 2-homogeneous form. A Gateaux-differentiable function F is said to be proximal PL with respect to C, A if there exists some constants $\alpha, \beta > 0$ such that for all $x \in C$

$$\frac{1}{2} \mathcal{D}_{C,A}(x, \beta) \geq \alpha(F(x) - \bar{F})\,, \tag{84}$$

where $\bar{F} = \min_{x \in C} F(x)$ and where $\mathcal{D}_{C,A}(x, \beta)$ is defined as

$$\mathcal{D}_{C,A}(x, \beta) = -2\beta \inf_{y \in C} dF(x)(y - x) + \frac{\beta}{2} A^x(y - x)\,. \tag{85}$$

Using these notions, we show the following exponential convergence result.

Lemma 9.3. *If F verifies $\Delta_F(x, y) \leq \frac{\beta}{2} A^y(y - x)$ for all $(x, y) \in C$ and is such that*

$$\frac{1}{2} \mathcal{D}_{C,A}(x, \beta) \geq \alpha(F(x) - \bar{F})\,, \tag{86}$$

then the scheme (provided that the iterates are well-defined)

$$x_{k+1} = \arg\min_{y \in C} dF(x_k)(y - x_k) + \frac{\beta}{2} A^{x_k}(y - x_k)\,, \tag{87}$$

yields iterates that verify $F(x_k) - \bar{F} \leq (1 - \alpha/\beta)^k (F(x_0) - \bar{F})$.

Proof. By relative smoothness and definition of the iterates

$$F(x_{k+1}) \leq F(x_k) + dF(x_k)(x_{k+1} - x_k) + \frac{\beta}{2} A^{x_k}(x_{k+1} - x_k) \tag{88}$$

$$\leq F(x_k) - \frac{1}{2\beta} \mathcal{D}_{C,A}(x, \beta) \tag{89}$$

$$\leq F(x_k) - \frac{\alpha}{\beta}(F(x_k) - \bar{F})\,. \tag{90}$$

Rearanging the terms yields the desired result. □

Now we want to apply the previous result to our function F that verifies $\frac{\alpha}{2} A^y(x-y) \leq \Delta_F(x,y) \leq \frac{\beta}{2} A^y(x-y)$. The lower bound ensures $\frac{1}{2}\mathcal{D}_{C,A}(x,\alpha) \geq \alpha(F(x) - \bar{F})$. Indeed

$$\Delta_F(y,x) \geq \frac{\alpha}{2} A^x(y-x) \tag{91}$$

$$\iff F(y) - F(x) \geq dF(x)(y-x) + \frac{\alpha}{2} A^x(y-x) \tag{92}$$

$$\implies F(y) - F(x) \geq \inf_{y \in C} dF(x)(y-x) + \frac{\alpha}{2} A^x(y-x) \tag{93}$$

$$\iff 2\alpha(F(x) - F(y)) \leq \mathcal{D}_{C,A}(x,\alpha) \tag{94}$$

$$\iff \frac{1}{2}\mathcal{D}_{C,A}(x,\alpha) \geq \alpha(F(x) - F(y)) \tag{95}$$

$$\implies \frac{1}{2}\mathcal{D}_{C,A}(x,\alpha) \geq \alpha(F(x) - \bar{F})\,. \tag{96}$$

We conclude with a monotonicity lemma to recover eventually $\frac{1}{2}\mathcal{D}_{C,A}(x,\beta) \geq \alpha(F(x) - \bar{F})$.

Lemma 9.4. *For a convex set C and a 2-homogeneous form $A^y(\cdot)$, if $0 \leq \alpha \leq \beta$ then for all $x \in C$, $\mathcal{D}_{C,A}(x,\alpha) \leq \mathcal{D}_{C,A}(x,\beta)$.*

Proof. We have by definition that for all $x, y \in C$, $-2\beta(dF(x)(y-x) + \frac{\beta}{2}A^x(y-x)) \leq \mathcal{D}_{C,A}(x,\beta)$. By convexity of C, we have in particular for all $x, y \in C$,

$$-2\beta(dF(x)((1-\frac{\alpha}{\beta})x + \frac{\alpha}{\beta}y - x) + \frac{\beta}{2}A^x((1-\frac{\alpha}{\beta})x + \frac{\alpha}{\beta}y - x)) \leq \mathcal{D}_{C,A}(x,\beta) \tag{97}$$

$$\iff -2\beta(\frac{\alpha}{\beta}dF(x)(y-x) + \frac{\alpha^2}{2\beta}A^x(y-x)) \leq \mathcal{D}_{C,A}(x,\beta) \tag{98}$$

$$\iff -2\alpha(dF(x)(y-x) + \frac{\alpha}{2}A^x(y-x)) \leq \mathcal{D}_{C,A}(x,\beta)\,. \tag{99}$$

In particular, taking the supremum of the l.h.s., we do recover $D_{C,A}(x,\beta) \geq D_{C,A}(x,\alpha)$. □

We draw the attention on the fact that while Lemma 9.3 holds for any C, A, the convexity of C and the 2-homogeneity of A are crucial to derive the monotonic behavior of $\mathcal{D}_{C,A}(x,\cdot)$. □

9.2. Section 5

Smooth and strongly convex functions case We study the case where $C = C_{\lambda,M,L,b} = \{f | \lambda - \text{strongly convex}, M - \text{smooth}, \|\nabla f(0)\| \leq L, |f(0)| \leq b\}$. As in the strongly convex case, using the results of Taylor et al. (2017), the iterates (4) can be reformulated as the following Quadratically Constrained Quadratic Program

$$\begin{aligned}\inf_{\substack{y\in\mathbb{R}^{2n+1}\\ \mathbf{z}\in\mathbb{R}^{(2n+1)\times d}}} \quad & c^\top y + \frac{1}{2}[y-f, \mathbf{z}-\mathbf{g}]^\top Q[y-f, \mathbf{z}-\mathbf{g}]\,, \\ & y_i - y_j - \mathbf{z}_j^\top(\mathbf{x}_i-\mathbf{x}_j) \geq \frac{1}{2(1-\frac{\lambda}{M})}\left(\frac{1}{M}\|\mathbf{z}_i-\mathbf{z}_j\|^2 + \lambda\|\mathbf{x}_i-\mathbf{x}_j\|^2 - 2\frac{\lambda}{M}(\mathbf{x}_i-\mathbf{x}_j)^\top(\mathbf{z}_i-\mathbf{z}_j)\right) \\ & \|\mathbf{z}_0\| \leq L, |y_0| \leq b\end{aligned} \tag{100}$$

where $f = f_k(\hat{\mu})$, $c = [0, \omega^\mu(\phi^*)'((f_k-q)(\hat{\mu})), -\omega^\nu(\phi^*)'((f_k^*-q)(\hat{\nu}))]$, $\mathbf{g} = [\vec{0}, \hat{\nu}]$ and Q is a diagonal matrix with diagonal $3S_K[0, \omega^\mu, \omega^\nu, \vec{0}, \tilde{\omega}^\nu]$ with $\tilde{\omega}^\nu := \left(\omega^\nu(\frac{1}{3S_K\lambda} + \frac{(\phi^*)'((f_k^*-q)(\hat{\nu}))(R^2+L(R^*)^2)}{\lambda^2})\right)^{*d}$ where $c, f, \mathbf{g}, Q$ are defined in Proposition 4. Now, for the statistical complexity, since the functions in $C_{\lambda,M,L,b}$ are $L+MR$-Lipschitz over B_R and bounded by $b+LR$ over B_R, similar results as in Proposition 5 hold: if the ground truth potential z_0 belongs to $C_{\lambda,M,L,b}$, then denoting $\hat{z} = \arg\min_{z\in C_{\lambda,M,L,b}} \hat{J}(z)$ we have

$$\begin{cases} \mathbb{E}[d_\phi^\lambda(\hat{z}, z_0)^2] \lesssim n^{-1/(1+d/4)} & \text{if } d<4 \\ \mathbb{E}[d_\phi^\lambda(\hat{z}, z_0)^2] \lesssim \frac{\log(n)}{\sqrt{n}} & \text{if } d=4 \\ \mathbb{E}[d_\phi^\lambda(\hat{z}, z_0)^2] \lesssim n^{-2/d} & \text{if } d>4\,. \end{cases}$$

Hence, even though the model is less expressive, the statistical behavior is the same as for λ-strongly convex functions ; it is an open question whether the additional smoothness can be leveraged statistically speaking.

Parametric model case We study the computational complexity and statistical guarantees we obtain when C is chosen to be a parametric set of (strongly) convex functions.

We propose to study the following model: $C = \{\lambda q + g | g(x) = \sum_{i=1}^p w_i\sqrt{\|x-c_i\|^2+\epsilon}, 0 \leq w_i \leq L\}$ where the centroids c_i are fixed vectors and the weights w_i are to be learned. The rational behind this model is that the hessian of the components is given by $\frac{1}{\sqrt{\|x-c_i\|^2+\epsilon}}(Id - \frac{(x-c_i)(x-c_i)^\top}{\|x-c_i\|^2+\epsilon})$, hence, in a first approximation, this model enables to locally add curvature around c_i.

Since C is a parametric model, assuming that the original UOT potential z_0 belongs to C, we recover the fast statistical rate $\mathbb{E}[d_\phi^\lambda(\hat{z}_C, z_0)] = O(\frac{1}{n})$; hence we have an improvement with respect to the case where C is the whole set of λ-strongly convex function and whose statistical complexity scaled as $n^{-2/d}$. Finally, the iterates of the semi-dual involves solving a quadratic program of the form

$$\inf_{0\leq w\leq L} c^\top w + \frac{1}{2}w^\top Z^\top Z w\,,$$

where $Z \in \mathbb{R}^{2nd\times p}$. The cost to compute $Z^\top Z$ scales as $O(p^2dn)$ and the cost to solve the program is $O(p^3)$. Hence assuming that $p << n$, the overall cost is $O(p^2dn)$, which is now linear in n instead of cubic in the case where C is the set of all λ-strongly convex functions. Indeed this model is extreme since it may lack expressiveness yet it illustrates how the computational and statistical behavior can be improved.