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

Adrien Vacher, François-Xavier Vialard

Stability of Semi-Dual Unbalanced Optimal Transport: fast statistical rates and convergent algorithm.

Adrien Vacher
LIGM, Univ. Gustave Eiffel, CNRS
INRIA
adrien.vacher@u-pem.fr

François-Xavier Vialard
LIGM, Univ. Gustave Eiffel, CNRS
INRIA
francois-xavier.vialard@u-pem.fr

June 15, 2022

Abstract

In this paper, we derive stability results for the semi-dual formulation of unbalanced optimal transport. From a statistical point of view, the gain of stability with respect to the balanced case allows to employ localization arguments while only assuming strong convexity of potentials and recover superparametric rates. Then we derive a provably convergent theoretical algorithm to minimize the semi-dual: if the potentials are constrained to be strongly convex, both the values and minimizers converge at a $1/k$ rate. Under an additional smoothness assumption, the convergence is exponential in the balanced case. Finally we instantiate a tractable version of our theoretical algorithm in the case of strongly convex, possibly smooth potentials. We benchmark the method in the balanced case on a 2D experiment and in the unbalanced case on a medium dimension synthetic experiment.

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: the discretisation of the infinite cost constraint strongly bias the estimators, especially when the dimension is large (Vacher et al., 2021) and 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 can benefit from the underlying regularity of the problem leading to superparametric rates under smoothness hypothesis (Hütter and Rigollet, 2021). Numerically speaking, it was shown empirically to produce very sharp transport maps on grids with algorithms converging in just a few iterations (Jacobs and 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 UOT. Unlike previous works (Hütter and 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 and use in particular localization arguments (van de Geer, 2002) even when no smoothness is assumed (more details on the localization technique are given in Sec. 3). As a corollary, we obtain the first statistical rates for the problem for UOT potentials estimation. Then we derive an algorithm to solve theoretically our semi-dual formulation. To this end, we design a variable metric gradient scheme that we believe has an interest in itself as it generalizes concepts of relatively smooth and relatively strongly convex optimization (Lu et al., 2018; Bauschke et al., 2017). 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. The results are competitive with the Sinkhorn model and its generalization in UOT.

Assumptions and notations In what follows, μ and ν are two positive Radon supported on X, Y subsets of $\mathbb{R}^d$ included in some centered ball B_R. For a probability measure β, we shall denote for $p \in [1, +\infty]$, $\|g\|_{L^p(\beta)} = (\int_x |g(x)|^p \,\mathrm{d}\beta(x))^{\frac{1}{p}}$. We shall denote by q the quadratic function $q(x) = \|x\|^2/2$ and for any Gateaux differentiable function h, we shall denote Δ_h the Bregman divergence associated to h, defined as $\Delta_h(x, y) = h(x) - h(y) - dh(y)(x - y)$. Finally, 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.

2 Semi-dual Unbalanced 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 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.*

Its primal formulation reads $\mathrm{UOT}(\mu, \nu) = \inf_{\pi \in \mathcal{M}_+(X \times Y)} D_\phi(\pi_0, \mu) + D_\phi(\pi_1, \nu) + \int_{X \times Y} c(x, y) \,\mathrm{d}\gamma(x, y)$ where c is the ground cost. 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. We shall assume throughout the paper that $c(x, y) = q(x - y)$ and for this cost, we derive a *semi-dual* formulation of the problem.

Proposition 1. *The demi-dual formulation reads*

$$\mathrm{UOT}(\mu,\nu) = -\inf_{z\in C_b(X)} J_{\mu,\nu}(z) + \iota_{CVX}(z)$$

, with $J_{\mu,\nu}(z) = \langle \phi^*(z-q),\mu\rangle + \langle \phi^*(z^*-q),\mu\rangle$ *and CVX is the set of convex functions.*

The proof is left in Appendix and is a direct application of optimality conditions. When confusion is possible we shall denote $J_{\mu,\nu}$ by J. The next proposition shows that J is convex and differentiable on a subset of convex functions.

Proposition 2. *The functional* J *is convex. Furthermore, if* g *is such that* g^* *is differentiable over the support of* ν *and* ϕ *is differentiable, its differential reads* $DJ(g) = \mu(\phi^*)'(g-q) - \nabla g^*(\nu)(\phi^*)'(g^*-q)$.

The proof is left in Appendix. In particular, if g is convex, $DJ(g)$ can be computed pointwise via convex programming. This observation is the key to derive tractable algorithms as shown in Sec. 5.

2.2 Stability estimates

We study in this paragraph upper and lower bounds on the Bregman divergence associated to the semi-dual Δ_J that we shall refer to as *stability estimates*. While previous works mainly focus on stability around the optimum (Manole et al., 2021; Hütter and Rigollet, 2021), that is $\Delta_J(f,g_0) = J(f) - J(g_0) - \langle DJ(g_0), f-g_0\rangle = J(f) - J(g_0)$ 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, the global nature of our estimates is a crucial point to derive provably convergent algorithms.

Proposition 3. *Let* (f,g) *be two convex functions bounded by* K_R *over* B_R *and such that* (f^*,g^*) *are bounded by* K_R^* *over* B_R *and* g^* *is differentiable over the support of* ν*. If* f *is* λ*-strongly convex and* ϕ^* *is twice differentiable then, denoting* $K = \max(K_R, K_R^*)$*, the Bregman divergence* $\Delta_J(f,g) = J(f) - J(g) - DJ(g)(f-g)$ *is bounded as*

$$\frac{\lambda}{2}\|\nabla f^* - \nabla g^*\|^2_{L^2(\tilde{\nu}_g)} + \frac{I_K}{2}H_{\mu,\nu}(f,g)^2 \leq \Delta_J(f,g) \leq \frac{1}{2\lambda}\|\nabla f - \nabla g\|^2_{L^2(\widetilde{\pi(g)}_g)} + \frac{S_K}{2}H_{\mu,\nu}(f,g)^2, \quad (1)$$

where $\pi(g) = \nabla g^*(\nu)$, $\tilde{\beta}_g = (\phi^*)'(g-q)\beta$, $I_\zeta = \inf_{|z|\leq\zeta+\frac{R^2}{2}} (\phi^*)''(z), S_\zeta = \sup_{|z|\leq\zeta+\frac{R^2}{2}} (\phi^*)''(z)$ *and* $H_{\mu,\nu}(f,g)^2 = \|f-g\|^2_{L^2(\mu)} + \|f^*-g^*\|^2_{L^2(\nu)}$. *Conversely, if* f *is* M*-smooth, then* $\Delta_J(f,g)$ *is bounded as*

$$\frac{1}{2M}\|\nabla f - \nabla g\|^2_{L^2(\widetilde{\pi(g)}_g)} + \frac{I_K}{2} H_{\mu,\nu}(f,g)^2 \leq \Delta_J(f,g) \leq \frac{M}{2}\|\nabla f^* - \nabla g^*\|^2_{L^2(\tilde{\nu}_g)} + \frac{S_K}{2} H_{\mu,\nu}(f,g)^2 \,.$$

The proof is left in Appendix. It consists into making a second order Taylor expansion of the conjugate of the entropy ϕ^* and to remark that the resulting expansion allows us to write $\Delta_J(f,g)$ either as $\Delta_f(\nabla f^*, \nabla g^*)$ integrated over the measure $(\phi^*)'(g-q)\nu$ or either as $\Delta_{f^*}(\nabla f, \nabla g)$ integrated over the measure $(\phi^*)'(g^*-q)\nabla g^*(\nu)$. Interestingly, when ϕ^* is locally strongly convex, as it is the case in the Gaussian-Hellinger metric, we gain in stability with respect to the balanced case on two levels. First, as shown in the left hand side of (1), Δ_J does not only control $\nabla f^* - \nabla g^*$ but it also controls $f^* - g^*$, which is coherent with the fact that in the balanced case, the potentials are no longer defined up to a constant. Second, it not only controls the difference of the conjugates $f^* - g^*$ but also the difference of the potentials themselves $f - g$.

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, we need to remove the dependency in the conjugate to derive provably convergent algorithms.

Corollary 1. *Under the same assumptions as in Prop. 3, 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* $\tilde{H}_{\mu,\nu}(f,g) = \|f-g\|^2_{L^2(\mu)} + \|f-g\|^2_{L^2(\nabla g^*(\nu))}$, $\Delta_J(f,g)$ *is upper-bounded as*

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

The proof is left in Appendix and is based on the following remark: when we apply Proposition 3 in the balanced case with $\phi^* = id$, we get under strong convexity the upper bound $\|\nabla f^* - \nabla g^*\|^2_{L^2(\nu)} \leq \frac{1}{\lambda^2}\|\nabla f - \nabla g\|^2_{\nabla g^*(\nu)}$. We do

not know if a similar lower bound on Δ_J depending only on the difference $f - g$ still holds. As shown Sec. 4, such a lower-bound would be sufficient to derive an exponentially convergent algorithm.

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 2 (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 $\hat{z}$.*

If the true unbalanced potential z_0 belongs to C, we can prove that the empirical potential $\hat{z}$ converges toward z_0 with respect the pseudo distance $d_\phi^\lambda(z, z_0)^2 = \frac{\lambda}{2}\|\nabla z^* - \nabla z_0^*\|^2_{L^2(\widetilde{\pi(z_0)}_{z_0})} + \frac{I_K}{2} H^2_{\mu,\nu}(z, z_0)$ where $\widetilde{\pi(z_0)}_{z_0}, I_K$ and $H^2_{\mu,\nu}(z, z_0)$ are defined in Proposition 3. Under suitable assumptions detailed below, the Legendre transform is Lipschitz and as in standard regression problems, the convergence rate is given by the growth rate of the metric entropy of the search space C; the lower, the faster. Furthermore, if $I_K > 0$, d_ϕ^λ not only controls $z^* - z_0^*$ but also $z - z_0$. This enables us to apply a localization argument without the C^2 smoothness assumption used in previous works. Informally, the localization argument is a bootstrap reasoning working as follows: denoting $\tau = d_\phi^\lambda(\hat{z}, z_0)$, the upper-bound (1) combined with standard statistical learning results gives $d_\phi^\lambda(\hat{z}, z_0)^2 \leq \tau^{1-\alpha/2}/\sqrt{n}$ where α is the growth rate of the *metric entropy* of C (Yukich, 1986). In particular, it constrains τ as $\tau^2 \leq n^{-\frac{1}{1+\alpha/2}}$; hence, when $\alpha < 2$, we recover a *superparametric* rate strictly faster than $1/\sqrt{n}$.

For the sake of simplicity, we chose to make an *unbiased* analysis, that is we make the assumption that $z_0 \in C$, even though unbiased statistical estimators are known to be suboptimal. Yet similar results hold with a bias measured in terms of d_ϕ^λ pseudo-distance.

Assumption 1. *(i) The measures μ, ν have support included in B_R, where B_r is the euclidean ball of $\mathbb{R}^d$ centered in 0 and of radius r. (ii) The measures μ, ν have 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* $supp(\mu)$ *and with* $\tilde{z}_0^*$ *coincides with* z_0^* *on* $supp(\nu)$. *(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.*

The goal of Assumption (ii) is to ensure the existence of the unbalanced transport map between μ, ν. The goal of Assumption (iii) is to ensure the absence of bias in the model. We believe that under a finer analysis, Assumption (i) could be replaced with sub-gaussian measures. Assumption (iv) ensures that the Legendre transform is Lipschitz for the supremum.

Proposition 4. *Under Assumptions (i)-(iv), if the unbalanced optimal transport potential* z_0 *between* μ *and* ν *belongs to* C, *then we have for all* $\delta \leq \frac{M'}{L}$

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

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, $b' = (b, R, \lambda, l, \varphi)$, $R' = (b, R, \lambda, l)$, $P = (b, R, \lambda, l, \varphi)$ *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.

Note that Proposition 4 does 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 2. *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* $\hat{z}_C$

verifies

$$\mathbb{E}[d_\phi^\lambda(\hat{z}_C, z_0)^2] \lesssim n^{-\alpha_{k,d}} \text{ if } \alpha_{k,d} \leq 1/2, \mathbb{E}[d_\phi^\lambda(\hat{z}_C, z_0)^2] \lesssim n^{-\frac{1}{1+\alpha_{k,d}^{-1}/2}} \text{ if } \alpha_{k,d} > 1/2\,. \quad (4)$$

Hence, we obtain a rate of $n^{-\frac{k+2}{d}}$ when $k+2 < d/2$ and $n^{-\frac{1}{1+\frac{d}{2(k+2)}}}$ when $k+2 > d/2$; note that we continuously transition from one rate to another when $k+2 = d/2$ where we recover the parametric rate $1/\sqrt{n}$. We conjecture that the minimax rates derived in Hütter and Rigollet (2021) still hold in the unbalanced setting and we compare them to our upper-bounds. If densities are k-times differentiable, the minimax rate is $n^{-\frac{k+1}{k+d/2}}$. As shown in Fig. 1, this rate is faster for any $k, d > 0$ yet when we transition in the highly smooth regime $k+2 > d/2$, our rate closely matches it. This discrepancy is due to the fact that we have no bias in our model *i.e.* we assumed $z_0 \in C$. In Hütter and Rigollet (2021), the authors fixed C to be a finite wavelet basis that does not necessarily contain z_0. In particular, they improve the bias variance trade-off as they benefit from the localization argument for any k.

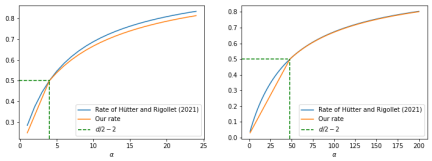

Figure 1: Comparison of our rates against the rates of Hütter and Rigollet (2021): on the left for $d = 12$ and on the right for $d = 100$.

4 Provably convergent minimization algorithm

In this section, we provide a theoretical algorithm to estimate unbalanced transport potentials and solve for arbitrary positive measures μ, ν $\min_{f \in C} J_{\mu,\nu}(f)$ 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 $f_{k+1} = f_k - \alpha DJ(f_k)$. However in our infinite dimensional setting, the gradient $DJ(f_k)$ is a measure, not a function. In a Banach setting such as ours, the 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 may converge slowly in practice. One way to improve convergence in practice is to recall the variational formulation of gradient descent and generate updates as $f_{k+1} = \arg\min_{f \in C} \langle f, DJ(f_k) \rangle + \alpha \|f - f_k\|^2$ where $\|\cdot\|$

is a well-chosen *fixed* norm. This update method was used in Jacobs and Léger (2020) to optimize the semi-dual in the balanced case on a fixed grid μ with the norm $L^2(\mu)$ leading to updates of the form $f_{k+1} = \mathcal{L}_\mu^{-1} DJ(f_k)$ where $\mathcal{L}_\mu^{-1}$ is the inverse Laplacian operator over μ. More broadly, updates generated as $f_{k+1} = \arg\min_{f \in C} \langle f, DJ(f_k) \rangle + \alpha \Delta_h(f, f_k)$, where h is some *fixed* convex function, were shown to converge at a $O(1/k)$ rate under *the relative smoothness assumption* $\Delta_J(f, g) \leq \frac{1}{\alpha} \Delta_h(f, g)$ (Bauschke et al., 2017; Lu et al., 2018). Unfortunately, we cannot benefit from these guarantees in our setting as our upper bound on $\Delta_J(f, f_k)$ depends on the *varying* pseudo-norm $L^2(\nabla f_k^*(\nu))$.

Hence, we study in this section the guarantees we can obtain on updates of the form $f_{k+1} = \arg\min_{f \in C} \langle f, DJ(f_k) \rangle + \alpha \|f - f_k\|_{f_k}^2$ where $\|\cdot\|_{f_k}$ is a pseudo-norm depending on the current iterate f_k. Our guarantees hold in a quite general setting and may have an interest on its own, in particular in the context of infinite-dimensional optimization.

4.1 The strongly convex case: sublinear rates

If all the functions in the set C are λ-strongly convex functions, recall that Sec. 2 provides a global quadratic upper bound of J: $J(f) \leq J(g_0) + \langle f - g_0, DJ(g_0) \rangle + \frac{1}{2\lambda} A^{g_0}(f - g_0)$ for any fixed g_0. Hence it is natural to minimize at each step this global upper-bound proxy and compute iterates as $f_{k+1} = \arg\min_{f \in C} \langle f - f_k, DJ(f_k) \rangle + \frac{1}{2\lambda} A^{f_k}(f - f_k)$. Yet we draw the attention on the fact that this particular scheme has not been studied so far and does not fit standard convex optimization settings: it is not a Newton scheme as the quadratic form $A^{f_k}(\cdot)$ is not associated to the hessian of J nor it is similar to a Bregman scheme as the quadratic form depends on the current iterate f_k. Had we been in a finite dimensional setting, we could have used the equivalence of norms to reduce to a Bregman setting. However in the context of functional optimization, we cannot make such a reduction. Still, since our proxy is sharper than a linear approximation, we can expect this minimization scheme to perform at least as well as the Frank-Wolfe algorithm which provably converges at a $O(1/k)$ rate. This observation is the key to prove our next result which holds in any Banach space and for any 2-homogeneous second-order upper-bound.

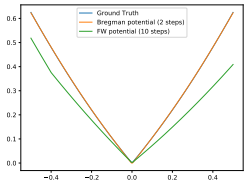

Figure 2: Potential generated by Frank-Wolfe with 10 steps (in green) vs generated by our algorithm with 2 steps (in orange) vs ground truth (in blue).

Proposition 5. *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 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*

$$x_{k+1} \in \arg\min_{x \in C} dF(x_k)(x - x_k) + \frac{\beta}{2} A^{x_k}(x - x_k)\,, \tag{5}$$

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

The proof is left in Appendix. As a Corollary, we show in Appendix that when we apply the scheme (5) to the semi dual with A^g and β the upper-bounding quantities of Corollary 1, we generate updates f_k such that $J(f_k) - J(\bar{f}) = O(1/k)$ where $\bar{f}$ is the optimum

With our stability results, we obtain not only the convergence of the values but also of the minimizers themselves with respect to d_ϕ^λ. Indeed, we could have achieved the same rate with a Frank-Wolfe algorithm, which only requires to solve a linear problem instead of a non-linear one. However, numerically, it has two drawbacks: at every step k, the current potential f_k is a barycenter of the k previous linear oracles, making the fenchel-transform f_k^* harder and harder to compute. Furthermore, Fig. 2 shows that in practice Frank-Wofle requires much more iterations for convergence than our algorithm.

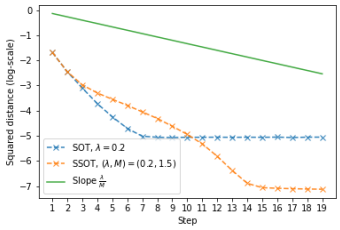

Figure 3: Convergence of $\log(\|\nabla f_k - \nabla \bar{f}\|^2_{L^2(\mu)})$ in the strongly convex case (in blue) vs the smooth and strongly convex case (in orange).

4.2 Smooth and strongly convex case: exponential rates for balanced optimal transport

In the M-smooth, λ-strongly convex case, we can bound from above and below the Bregman divergence of the semi-dual in the balanced case as $\frac{1}{2M} A^{g_0}(f - g_0) \leq \Delta_J(f, g_0) \leq \frac{1}{2\lambda} A^{g_0}(f - g_0)$. As in relatively smooth and strongly-convex optimization (Lu et al., 2018; Gutman and Pena, 2019), we can expect to obtain an exponential convergence when minimizing at each step the upper-bound $\langle DJ(f_k), f - f_k\rangle + \frac{1}{2\lambda} A^{f_k}(f - f_k)$. However, the dependency in the current point of the quadratic form $A^{f_k}(\cdot)$ prevents us from applying Lu et al. (2018); Gutman and Pena (2019). To recover exponential rates,

we rely on a proximal-PL (Karimi et al., 2016) analysis for the convergence which alleviates the dependency problem.

Proposition 6. *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 minimzer $\bar{x} \in C$ and the iterates $x_0 \in C$, (x_k) generated by 5 exist, we have* $F(x_k) - F(\bar{x}) \leq \left(1 - \frac{\alpha}{\beta}\right)^k [F(x_0) - F(\bar{x})]$.

The proof is left in Appendix. As a corollary, we show in Appendix that under M-smoothness and λ-strong convexity, taking $\beta = \frac{1}{\lambda}$ and $A^g(h) \equiv \|\nabla h\|^2_{L^2(\nabla g^*(\nu))}$, we obtain iterates verifying $J(f_k) - J(\bar{f}) \leq \left(1 - \frac{\lambda}{M}\right)^k [F(x_0) - F(\bar{x})]$. Fig.3 shows that when we explicitly constrain our potentials to be smooth, we observe an exponential convergence up to the numerical precision. On the other hand we observe that when the potentials are not explicitly constrained to be smooth, 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 the algorithm are provided in the next section.

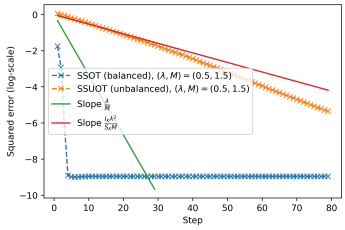

Figure 4: Convergence of $\log(\|\nabla f_k - \nabla \bar{f}\|^2_{L^2(\mu)})$ in the balanced case (in blue) vs the unbalanced case (in orange).

There remains a gap in the unbalanced case where, because we did not manage to remove the dependency in the conjugates in the lower bound of $\Delta_J(f, g)$, we cannot claim an exponential convergence of the algorithm even under smoothness condition. In fact, Fig. 4 empirically suggests that the convergence does occur at a linear rate yet significantly slower than in balanced case for which only several steps are necessary to reach numerical precision. This might be due to the poor 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 $\phi^*(t) = e^t - 1$. This opens the question of whether we can find a semi-dual formulation of unbalanced that is more suited to a specific choice of the entropy ϕ. For instance, in the KL case, the dual problem can be reformulated as $\sup_{f,g \leq 1} \langle f, \mu \rangle + \langle g, \nu \rangle + \iota((1 - f(x))(1 - g(y)) \geq e^{-\|x-y\|^2})$, see Chizat

(2017) for more details. Defining the transformation $\tilde{f}(y) = 1 + \sup_x \frac{e^{-\|x-y\|^2}}{1-f(x)}$, we can get a new semi-dual formulation $\sup_{f \leq 1} \langle f, \mu \rangle + \langle \tilde{f}, \mu \rangle$ that gets rid of the entropy in the objective. We believe such a formulation possesses better conditioning and can be tractable for a well-chosen class of potentials. We postpone this direction for future works.

5 Numerical Experiments

In this section, we show how to instantiate and implement the algorithm (5) when C is the set of function $g + \lambda q$ with g is convex and L-lipschitz or when C is the set of λ-strongly convex, M-smooth functions. Then we benchmark this model in the balanced case in a 2D stochastic shape-matching experiment against the SSNB model Paty et al. (2020) and the Sinkhorn model Cuturi (2013) and in the unbalanced case for a problem of potential estimation in dimension 6 against Sinkhorn.

5.1 The model

We place ourselves in the setting where $\hat{\mu}, \hat{\nu}$ are n-samples discrete empirical measures $(x_i), (y_i)$ with weights (ω^μ, ω^ν). Recall that for f^* differentiable on $\hat{\nu}$, the gradient of the semi-dual reads $DJ(f) = \hat{\mu} - \nabla f^*(\hat{\nu})$. In particular, when f is convex, $\nabla f^*(\hat{\nu}) = (z_i)$ can be computed pointwise and the infinite dimensional problem (5) can be cast as a finite interpolation problem with quadratic objective $\inf_{\xi^\mu, \xi^\nu, \zeta^\nu} (\xi^\mu)^\top \omega^\mu - (\xi^\nu)^\top \omega^\nu + \frac{\beta}{2} \Big[(\zeta^\nu - y) \Omega^{\tilde{\nu}} (\zeta^\nu - y) + (\xi^\mu - f_k(\hat{\mu})) \Omega^\mu (\xi^\mu - f_k(\hat{\mu})) + (\xi^\nu - f_k(\hat{\nu})) \Omega^\nu (\xi^\nu - f_k(\hat{\nu})) \Big]$ where $\Omega^{\tilde{\nu}}$ is a diagonal matrix of size nd with diagonal given by $\phi^*(f_k^*(y_i) - q(y_i))\omega_i^\nu$ (each entry is repeated d-times), Ω^μ, Ω^ν are diagonal matrices of size n with diagonal given by ω^μ, ω^ν respectively, under the constraint that there exists $f \in C$ such that for all i,

$$f(x_i) = \xi_i^\mu, f(z_i) = \xi_i^\nu, \nabla f(z_i) = \zeta_i^\nu \,. \tag{6}$$

When $C_{\lambda,L} = \{g + \lambda q \mid g \text{ convex}, L-\text{lipschitz}\}$, the constraint (6) admits a finite reformulation of $O(n^2)$ linear (sparse) constraints and that the minimizers can be extrapolated in closed form.

Proposition 7 ((Taylor et al., 2017)). *For $C = C_{\lambda,L}$, the constraint 6 admits the following finite reformulation: for all $1 \leq i, j \leq 2n$, $i \neq j$ $\tilde{\xi}_i \geq \tilde{\xi}_j + (\tilde{z}_i - \tilde{z}_j)^\top \tilde{\zeta}_j + \lambda(q(\tilde{z}_i) - q(\tilde{z}_j) + (\tilde{z}_i - \tilde{z}_j)^\top \tilde{z}_i)$, $\|\tilde{\zeta}_j\|_\infty \leq L$ where $\tilde{\xi} = [\xi^\mu, \xi^\nu]$,*

$\tilde{\zeta} = [\zeta^\mu, \zeta^\nu]$ *and* $\tilde{z} = [x, z]$ *and the potential* f_{k+1} *can be extended on every point* $x \in \mathbb{R}^d$ *as* $f_{k+1}(x) = \max_i \ \tilde{\xi}_i^k - \lambda q(\tilde{z}_i^k) + (x - \tilde{z}_i^k)^\top(\tilde{\zeta}_i^k - \lambda \tilde{z}_i^k) + \lambda q(x)$. *Furthemore, its conjugate gradient can be computed pointwise* $\nabla f_{k+1}^*(y)$ *as the program* $\inf_{t,x} t - x^\top y$ *under the constraint* $t \geq \tilde{\xi}_i^k - \lambda q(\tilde{z}_i^k) + (x - \tilde{z}_i^k)^\top(\tilde{\zeta}_i^k - \lambda \tilde{z}_i^k) + \lambda q(x), \ \forall \ 1 \leq i \leq 2n$.

As shown in Nemirovski (2004, Section 10.1), the cost to solve the resulting finite reformulation of (5) is $O(n^3)$. We show in Appendix that if the ground truth UOT potential z_0 is λ-strongly convex and L-lipschitz on B_R, then denoting $\hat{z}_{\lambda,L} = \inf_{z \in C_{\lambda,L}} \hat{J}(z)$, we have $\mathbb{E}[d_\phi^\lambda(\hat{z}_{\lambda,L}, z_0)] \lesssim n^{-\frac{2}{d}}$. We state in Appendix similar results for $C = C_{\lambda,M}$ the set of λ-strongly convex, M-smooth functions.

5.2 Other models

Sinkhorn The well-known Sinkhorn model (Cuturi, 2013) can be extended to the unbalanced case and its primal objective reads $\mathcal{S}_\varepsilon^\phi(\mu, \nu) = \inf_{\pi \geq 0} \langle \pi, C \rangle + D_\phi(\pi_1|\mu) + D_\phi(\pi_2|\nu) + \varepsilon \mathrm{KL}(\pi|\mu \otimes \nu)$ 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 spaces 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.

5.3 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 [1], 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.

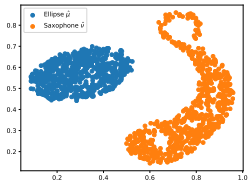

Figure 5: 2D experiment: Ellipse (in blue) and Saxophone (in orange).

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 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.

5.4 Medium dimension synthetic experiment

In this paragraph, we study the ability of models to recover the ground truth unbalanced transport potential z_0 between μ and ν (i.e. the solution of $\inf_z J_{\mu,\nu}(z)$) from sampled measures $(\hat{\mu}, \hat{\nu})$. However, even if μ and ν are available in closed forms, z_0 generally isn't. In the following proposition, we derive a Brenier-like result that will allow us to easily generate ground truths.

[1] This data was borrowed from Feydy et al. (2017).

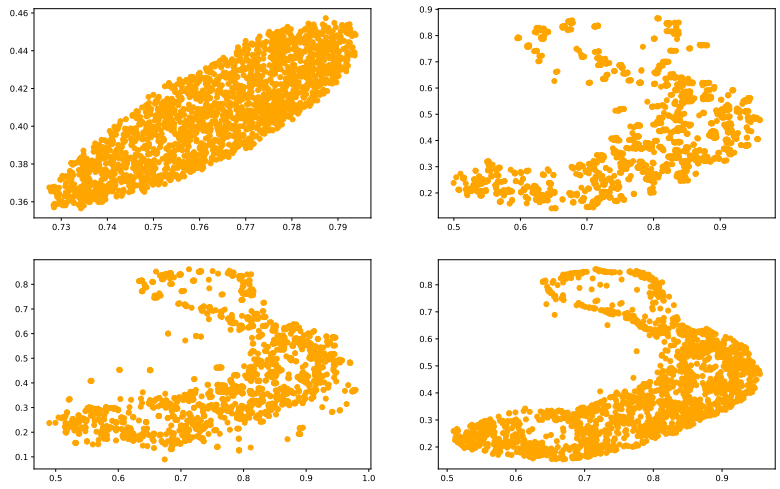

Figure 6: Pushforwards $\nabla \hat{f}(\hat{\mu}_{test})$. From top left to bottom right: Sinkhorn ($\varepsilon = 0.1$), SSNB ($\lambda = 0.2$, $M = +\infty$), Semi-dual OT ($\lambda = 0.2, L = 1$), Sinkhorn ($\varepsilon = 0.0001$).

Proposition 8. *Let μ be a probability measure, z_0 be a convex function and ϕ an entropy function such that ϕ^* is strictly convex. If we take $\tilde{\mu} = \mu/(\phi^*)'(z_0 - q)$ and $\tilde{\nu} = \nabla z_0(\mu)/(\phi^*)'(z_0^* - q)$, then z_0 is solution of $\inf_z J_{\tilde{\mu},\tilde{\nu}}(z)$.*

The proof is left in Appendix. In our setting 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 scalibilty 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 $\|\hat{z} - z_0\|^2_{L^2(\tilde{\mu})}$ 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 reduce the bias of the model yet not too small in order to mitigate the variance that behaves poorly as the dimension grows.

6 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 our new variable metric gradient scheme, that we believe has an interest of its own. There remains a theoretical gap in the unbalanced case where we did not manage to prove the exponential convergence observed in practice. Finally, we instantiated a tractable, proof-of-concept version of our algorithm that is competitive with the well-known Sinkhorn algorithm yet it poorly scales in $O(n^3)$. For future works, we shall focus on two directions: first the design of a search space C with better scaling and alternative semi-dual formulations with improved stability bounds and conditioning to hopefully attain faster theoretical and practical convergence.

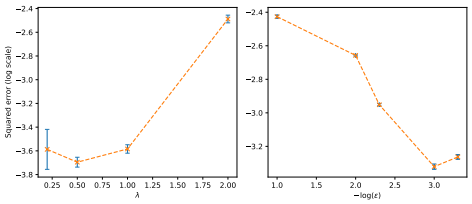

Figure 7: 6D experiment: On the left, $\|\hat{z}_\lambda - z_0\|^2_{L^2(\tilde{\mu})}$ (our model) and on the right, $\|\hat{z}_\varepsilon - z_0\|^2_{L^2(\tilde{\mu})}$ (Sinkhorn).

References

Brandon Amos, Lei Xu, and J Zico Kolter. Input convex neural networks. In *ICML*, 2017.

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

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

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

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

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

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

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

Jean Feydy, Benjamin Charlier, François-Xavier Vialard, and Gabriel Peyré. Optimal transport for diffeomorphic registration. In *MICCAI*, 2017.

Thomas Gallouët, Roberta Ghezzi, and Francois-Xavier Vialard. Regularity theory and geometry of unbalanced optimal transport, 2021.

David H. Gutman and Javier F. Pena. The condition number of a function relative to a set. In *Mathematical Programming*, 2019.

Jan-Christian Hütter and Philippe Rigollet. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 2021.

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

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *ECML*, 2016.

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

Haihao Lu, Robert M. Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 2018.

Ulrike von Luxburg and Olivier Bousquet. Distance–based classification with lipschitz functions. *JMLR*, 2004.

Tudor Manole, Sivaraman Balakrishnan, Jonathan Niles-Weed, and Larry Wasserman. Plugin estimation of smooth optimal transport maps. *arXiv*, 2021.

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

François-Pierre Paty, Alexandre d'Aspremont, and Marco Cuturi. Regularity as regularization: Smooth and strongly convex brenier potentials in optimal transport. In *AISTATS*, 2020.

Khiem Pham, Khang Le, Nhat Ho, Tung Pham, and Hung Bui. On unbalanced optimal transport: An analysis of sinkhorn algorithm. *ICML*, 2020.

Thibault Séjourné, Jean Feydy, François-Xavier Vialard, Alain Trouvé, and Gabriel Peyré. Sinkhorn divergences for unbalanced optimal transport. *arXiv*, 2019.

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

Adrien Vacher, Boris Muzellec, Alessandro Rudi, Francis Bach, and Francois-Xavier Vialard. A dimension-free computational upper-bound for smooth optimal transport estimation. In *COLT*, 2021.

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

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

J. E. Yukich. Metric entropy and the central limit theorem in banach spaces. In *Geometrical and Statistical Aspects of Probability in Banach Spaces*, 1986.

A Additional results

A.1 Section 4

In this paragraph, we prove that the generic algorithms of Sec. 4 do apply in the unbalanced case for well chosen forms $A^{\cdot}(\cdot)$.

Corollary 3. *Let C be a closed convex set of λ-strongly convex functions, $L(r)$-lipschitz over B_r and such that for all $f \in C$, $|f(0)| \leq b$. The minimum $\bar{f} = \arg\min_{f\in C} J(f)$ exists. Furthermore, for the choice of form*

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

where $R^* = \frac{R}{\lambda}+2\sqrt{\frac{b}{\lambda}}$, $K = \max(b+LR, R^*(R+b+L))$, $S_\zeta = \sup_{|t|\leq\zeta+q(R)}(\phi^*)''(t)$, $\pi(g) = \nabla g^*(\nu)(\phi^*)'(g^* - q)(\nu)$ and $\tilde{H}^g_{\mu,\nu}(h) = \|h\|^2_{L^2(\mu)} + \|h\|^2_{L^2(\nabla g^*(\nu))}$, the iterates $x_{k+1} = \arg\min_{f\in C}\langle DJ(f_k), f - f_x\rangle + \frac{1}{2}A^{f_k}(f - f_k)$ are well defined and they verify

$$J(f_k) - J(\bar{f}) \leq \frac{4\lambda IL(R^*) + 3S_K\left[m_\nu L(R^*)(R^2 + L(R^*)^2 + 1) + m_\mu(b + L(R)) + m_\nu b\right]}{\lambda^2}, \tag{7}$$

where $I = \int(\phi^*)'(K + q(y))$ and m_μ, m_ν are the total masses of μ, ν respectively.

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 1 that yields for $f \in C$

$$\|\nabla f^*\|_{L^\infty(B_R)} \leq R^* \coloneqq \frac{R}{\lambda} + 4\sqrt{\frac{b}{\lambda}} \quad \|f^*\|_{L^\infty(B_R)} \leq RR^* + b + R^* L(R^*)\,. \tag{8}$$

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 1, we have indeed for all $(f, g) \in C$

$$\begin{aligned}\Delta_J(f, g) \leq \frac{1}{\lambda} \|\nabla f - \nabla g\|^2_{L^2(\pi(g))} \\ + 3S_K \left[\frac{R^2 + L(R^*)^2}{\lambda^2} \|\nabla f - \nabla g\|^2_{L^2(\nabla g^*(\nu))} + \tilde{H}^g_{\mu,\nu}(f - g) \right],\end{aligned} \tag{9}$$

where $\pi(g) = \nabla g^*(\nu)(\phi^*)'(g^* - q)$ and $\tilde{H}^g_{\mu,\nu}(h) = \|h\|^2_{L^2(\mu)} + \|h\|^2_{L^2(\nabla g^*(\nu))}$. All that remains to prove is the boundedness of $A^g(h)$ now. Since $\nabla g^*(\nu) \subset B_{R^*}$ we have $\|\nabla f - \nabla g\|^2_{L^2(\pi(g))} \leq 2L(R^*) \int (\phi^*)'(K + q(y))\, \mathrm{d}\nu(y)$ (recall that $(\phi^*)'$ is a non-decreasing function). And conversely, $\|\nabla f - \nabla g\|^2_{L^2(\nabla g^*(\nu))} \leq 2m_\nu L(R^*)$. Finally, $\tilde{H}^g_{\mu,\nu}(f - g) \leq 2m_\mu(b + L(R)) + 2m_\nu(b + L(R^*))$. Using Proposition 5, we do recover

$$\begin{aligned}&J(f_k) - \bar{J} \leq \\ &4 \frac{\lambda I L(R^*) + 3S_K \left[(R^2 + L(R^*)^2 + 1) m_\nu L(R^*) + m_\mu(b + L(R)) + b m_\nu \right]}{\lambda^2},\end{aligned} \tag{10}$$

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

Corollary 4. *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 $f \in C, |f(0)| \leq b$. Using*

the form $A^g(h) = \|h\|_{\nabla g^*(\nu)}$ *and* $\beta = \frac{1}{\lambda}$*, we have in the balanced case that* $\bar{f} = \arg\min_{f \in C} J(f)$ *as well as the iterates* f_k *are well-defined and that* $J(f_k)_J(\bar{f}) \leq (1 - \frac{\lambda}{M})^k (J(f_0) - J(\bar{f}))$.

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

$$\frac{1}{2M}\|\nabla f - \nabla g\|_{L^2(\nabla g^*(\nu))} \leq \Delta_J(f, g) \leq \frac{1}{2\lambda}\|\nabla f - \nabla g\|_{L^2(\nabla g^*(\nu))}\,. \tag{11}$$

□

A.2 Section 5

We derive a sample complexity result for $C_{\lambda,L,b} = \{\lambda q + g | g \text{ convex}, L - \text{ lipschitz }, |g(0)| \leq b\}$.

Corollary 5. *If* z_0 *the ground truth UOT potential belongs to* C*, then under Assumptions 1 (i)-(v),*

$$\begin{cases} \mathbb{E}[d_\phi^\lambda(\hat{z}_{C_{\lambda,L,b}}, z_0)^2] \lesssim n^{-1/(1+d/4)} \text{ if } d < 4 \\ \mathbb{E}[d_\phi^\lambda(\hat{z}_{C_{\lambda,L,b}}, z_0)^2] \lesssim \frac{\log(n)}{\sqrt{n}} \text{ if } d = 4 \\ \mathbb{E}[d_\phi^\lambda(\hat{z}_{C_{\lambda,L,b}}, z_0)^2] \lesssim n^{-2/d} \text{ if } d > 4\,. \end{cases} \tag{12}$$

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

$$n(C_{\lambda,L,b}, L^\infty(B_{R'}), u) \lesssim u^{-d/2}\,, \tag{13}$$

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 4 and recover

$$\mathbb{E}[d_\phi^\lambda(\hat{z}_{C_{\lambda,L,b}}, z_0)^2] \lesssim n^{-1/(1+d/4)}\,. \tag{14}$$

If $d = 4$, applying the first part of Proposition 4 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

$$\mathbb{E}[d_\phi^\lambda(\hat{z}_{C_{\lambda,L,b}}, z_0)^2] \lesssim n^{-2/d}\,. \tag{15}$$

□

Next, we state how the interpolability constraints (6) can be reformulated as a convex finite problem using again Taylor et al. (2017).

Proposition 9 ((Taylor et al., 2017)). *For $C = C_{\lambda,M}$ the set of λ-strongly convex, M-smooth functions, the constraint 6 admits the following finite reformulation: for all $1 \leq i, j \leq 2n$, $i \neq j$ $\tilde{\xi}_i \geq \tilde{\xi}_j + (\tilde{z}_i - \tilde{z}_j)^\top \tilde{\zeta}_j + \frac{1}{2(1-\lambda/M)} \Big(\frac{1}{M} \|\tilde{\zeta}_j - \tilde{\zeta}_i\|^2 + \lambda \|\tilde{z}_i - \tilde{z}_j\|^2 - 2\frac{\lambda}{M} (\tilde{\zeta}_j - \tilde{\zeta}_i)^\top (\tilde{z}_j - \tilde{z}_i) \Big)$, where $\tilde{\xi} = [\xi^\mu, \xi^\nu]$, $\tilde{\zeta} = [\zeta^\mu, \zeta^\nu]$ and $\tilde{z} = [x, z]$.*

The potential f_{k+1} can be extended on every point $x \in \mathbb{R}^d$ as $f_{k+1}(x) = (\max_i \ h_i)^*(x) + \lambda q(x)$. where $h_i(y) = \tilde{\xi}_i' + (\tilde{\zeta}_i')^\top (y - \tilde{z}_i') + \frac{1}{2(M-\lambda)} \|y - \tilde{z}_i'\|^2$ with $\tilde{\xi}_i' = \tilde{z}_i^\top \tilde{\zeta}_i - \tilde{\xi}_i - \frac{\lambda}{2}\|\tilde{z}_i\|^2$, $\tilde{\zeta}_i' = \tilde{z}_i$ and $\tilde{z}_i' = \tilde{\zeta}_i - \lambda \tilde{z}_i$. Furthemore, its conjugate gradient can be computed pointwise $\nabla f_{k+1}^*(y)$ as the program $\nabla f_{k+1}^*(y) = \inf_x \max_i \ h_i(x) + \frac{1}{\lambda} q(y - x)$.

B Proofs of Sec. 2

B.1 Proof of Proposition 1

The dual formulation of UOT reads

$$\begin{aligned} \mathrm{UOT}(\mu, \nu) = \sup_{z_0, z_1} & \langle -\phi^*(-z_0), \mu \rangle + \langle -\phi^*(-z_1), \nu \rangle \\ & \text{s.t. } z_0(x) + z_1(y) \leq q(x - y)\,, \end{aligned} \tag{16}$$

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

$$\begin{aligned} \mathrm{UOT}(\mu, \nu) = \sup_{\tilde{z}_0, \tilde{z}_1} & \langle -\phi^*(\tilde{z}_0 - q), \mu \rangle + \langle -\phi^*(\tilde{z}_1 - q), \nu \rangle \\ & \text{s.t. } \tilde{z}_0(x) + \tilde{z}_1(y) \geq x^\top y\,. \end{aligned} \tag{17}$$

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

$$\mathrm{UOT}(\mu, \nu) = \sup_{\tilde{z}_0} \langle -\phi^*(\tilde{z}_0 - q), \mu \rangle + \langle -\phi^*(\tilde{z}_0^* - q), \nu \rangle \tag{18}$$

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

$$\mathrm{UOT}(\mu,\nu) = \sup_{\tilde{z}_0} \langle -\phi^*(\tilde{z}_0 - q), \mu\rangle + \langle -\phi^*(\tilde{z}_0^* - q), \nu\rangle + \iota_{\mathrm{CVX}}(z_0)\,. \tag{19}$$

B.2 Proof of Proposition 2

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

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

Using the convexity of ϕ^*, we recover

$$J(tz_0 + (1-t)z_1) \leq tJ(z_0) + (1-t)J(z_1)\,. \tag{20}$$

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 f is strongly convex, it defines a unique supremum $\nabla z^*(p)$ and the envelope theorem gives

$$\frac{\delta z^*}{\delta z} = -(\delta z)(\nabla z^*(p))\,. \tag{21}$$

Now, one has, ϕ being differentiable,

$$\begin{aligned} DJ(z)(\delta z) &= \langle \phi^*(z)\delta z, \mu\rangle + \langle -\phi^*(z^*)(\delta z)(\nabla z^*(p)), \nu\rangle \\ &= \langle \delta z, \phi^*(z)\mu - [\nabla z^*]_\sharp(\phi^*(z^*)\nu)\rangle\,. \end{aligned} \tag{22}$$

Note that the measures $\phi^*(z^*)\nu$ and $\phi^*(z)\mu$ are well defined since $\phi^*(z)$ and $\phi^*(z^*)$ are continuous functions and ν, μ Radon measures.

B.3 Proof of Proposition 3

Let us start by computing the Bregman divergence $\Delta_J(f,g)$. Since we assumed g^* differentiable over the support of ν, we have $DJ(g) = (\phi^*)'(g-q)\mu - (\phi^*)'(g^*-q)\nabla g^*(\nu)$. Hence, we can write

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

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

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

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

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

where $I_K = \inf_{|t| \leq K+q(R)} (\phi^*)''(t)$. 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

$$\frac{I_K}{2} H_{\mu,\nu}(f,g)^2 + \langle (\phi^*)'(g^*-q)(f^*-g^*), \nu \rangle + \langle f-g, (\phi^*)'(g^*-q) \nabla g^*(\nu) \rangle \,, \quad \text{(LB)}$$

and the upper-bound

$$\frac{S_K}{2} H_{\mu,\nu}(f,g)^2 + \langle (\phi^*)'(g^*-q)(f^*-g^*), \nu \rangle + \langle f-g, (\phi^*)'(g^*-q) \nabla g^*(\nu) \rangle \,, \quad \text{(UB)}$$

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

We now focus on the term $\langle (\phi^*)'(g^*-q)(f^*-g^*), \nu \rangle + \langle f-g, (\phi^*)'(g^*-q) \nabla g^*(\nu) \rangle$. We can re-write it as $\langle f^* \circ \nabla g - g^* \circ \nabla g + (f-g), (\phi^*)'(g^*-q) \nabla g^*(\nu) \rangle$ 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

$$\Gamma_{f,g}(x) = \Delta_{f^*}(\nabla g(x), \nabla f(x)) \,. \quad (23)$$

Conversely, the term $\langle (\phi^*)'(g^*-q)(f^*-g^*), \nu \rangle + \langle f-g, (\phi^*)'(g^*-q) \nabla g^*(\nu) \rangle$ can be re-written as $\langle f^* - g^* + f \circ \nabla g^* - g \circ \nabla g^*, (\phi^*)'(g^*-q)\nu \rangle$. 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 $\Gamma_{f,g}(x) \leq \frac{1}{2\lambda}\|\nabla g(x) - \nabla f(x)\|^2$ and $\Gamma_{f^*,g^*}(y) \geq \frac{\lambda}{2}\|\nabla g^*(y) - \nabla f^*(y)\|^2$ which yields the following bound on Δ_J

$$\frac{\lambda}{2}\|\nabla f^* - \nabla g^*\|^2_{L^2(\tilde{\nu}_g)} + \frac{I_K}{2}H_{\mu,\nu}(f,g)^2 \leq \Delta_J(f,g) \\ \leq \frac{1}{2\lambda}\|\nabla f - \nabla g\|^2_{L^2(\tilde{\pi}(g)_g)} + \frac{S_K}{2}H_{\mu,\nu}(f,g)^2\,,$$

where $\pi(g) = \nabla g^*(\nu)$, $\tilde{\beta}_g = (\phi^*)'(g-q)\beta$. Conversely, when f is M-smooth

$$\frac{1}{2M}\|\nabla f - \nabla g\|^2_{L^2(\tilde{\pi}(g)_g)} + \frac{I_K}{2}H_{\mu,\nu}(f,g)^2 \leq \Delta_J(f,g) \\ \leq \frac{M}{2}\|\nabla f^* - \nabla g^*\|^2_{L^2(\tilde{\nu}_g)} + \frac{S_K}{2}H_{\mu,\nu}(f,g)^2\,.$$

B.4 Proof of Corollary 1

We derive an upper-bound of $\|f^* - g^*\|^2_{L^2(\nu)}$ that solely depends on the difference $f - g$. We start to re-write this quantity as $\|f^* \circ \nabla g - g^* \circ \nabla g\|^2_{L^2(\nabla g^*(\nu))}$ 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 $\|f - g\|^2_{L^2(\nabla g^*(\nu))}$. Using Cauchy-Schwartz and the fact that measures are supported over B_R, the integration of the first term can be upper-bounded as

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

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

$$\int (\nabla g(x)^\top (\nabla f^*(\nabla g(x)) - x))^2 (\,\mathrm{d}\nabla g^*(\nu))(x) \leq \frac{R^2}{\lambda^2}\|\nabla f - \nabla g\|^2_{L^2(\nabla g^*(\nu))}\,. \tag{24}$$

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

$$\int (f(x) - f(\nabla f^*(\nabla g(x))))^2 (\,\mathrm{d}\nabla g^*(\nu))(x) = \int (f(\nabla g^*(y)) - f(\nabla f^*(y)))^2 \,\mathrm{d}\nu(y) \tag{25}$$

$$\leq L^2 \|\nabla g^* - \nabla f^*\|^2_{L^2(\nu)} \tag{26}$$

$$\leq \frac{L^2}{\lambda^2} \|\nabla f - \nabla g\|^2_{L^2(\nabla g^*(\nu))} \,. \tag{27}$$

C Proofs of Section 3

C.1 Proof of Proposition 4

To prove Proposition 4 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 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 $x \in B_r$

$$\|\nabla z^*(x)\| = \|\nabla z^*(x) - \nabla z^*(0) + \nabla z^*(0)\| \tag{28}$$

$$\leq \|\nabla z^*(x) - \nabla z^*(0)\| + \|\nabla z^*(0)\| \tag{29}$$

$$\leq \frac{r}{\lambda} + \|\nabla z^*(0)\| \,. \tag{30}$$

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

$$z(0) \geq z(x_*) + \frac{\lambda}{2} \|x_*\|^2 \,, \tag{31}$$

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

$$\|x_*\| \leq \sqrt{\frac{2(b(0) - l)}{\lambda}} \,. \tag{32}$$

The bound on $\|z^*\|_{L^\infty_{B_r}}$ follows the definition of the Fenchel-Legendre transform

$$z^*(x) = x^\top \nabla z^*(x) - z(\nabla z^*(x)) \,. \tag{33}$$

□

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

Lemma 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 1.*

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

$$z_1^*(x) \geq x^\top y - z_1(y)\,, \tag{34}$$

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

$$z_1^*(x) - z_2^*(x) \geq x^\top y - z_1(y) + z_2(\nabla z_2^*(x)) - x^\top \nabla z_2^*(x)\,. \tag{35}$$

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

$$z_1^*(x) - z_2^*(x) \geq z_2(\nabla z_2^*(x)) - z_1(\nabla z_2^*(x))\,, \tag{36}$$

and applying Lemma 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

$$z_2^*(x) - z_1^*(x) \geq z_1(\nabla z_1^*(x)) - z_2(\nabla z_1^*(x))\,, \tag{37}$$

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 the first part of Proposition 4.

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

$$\begin{aligned} d_\phi^\lambda(\hat{z}, z_0)^2 &\leq J(\hat{z}) - J(z_0) && (38)\\ &= J(\hat{z}) - \hat{J}(\hat{z}) + \hat{J}(\hat{z}) - \hat{J}(z_0) + \hat{J}(z_0) - J(z_0)\,. && (39) \end{aligned}$$

Using Assumption (iii), the term $\hat{J}(\hat{z}) - \hat{J}(z_0)$ is negative hence we have

$$\begin{aligned} d_\phi^\lambda(\hat{z}, z_0)^2 &\leq J(\hat{z}) - \hat{J}(\hat{z}) + \hat{J}(z_0) - J(z_0) && (40)\\ &\leq \sup_{z \in C} \langle \phi^*(z - q), \mu - \hat{\mu} \rangle && (41)\\ &\quad + \sup_{z \in C^*} \langle \phi^*(z - q), \nu - \hat{\nu} \rangle && (42)\\ &\quad + \hat{J}(z_0) - J(z_0)\,, && (43) \end{aligned}$$

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

Bound on term (41) Denoting $C_0 = \{\phi^*(g-q), g \in C\}$, we apply Luxburg and 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 $\delta > 0$

$$\mathbb{E}[W] \leq 2\delta + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\frac{\delta}{4}}^{\infty} \sqrt{n(C_0, L^2(\hat{\mu}), u)}\, \mathrm{d}u \,. \tag{44}$$

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

$$\mathbb{E}[W] \leq 2\delta + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\frac{\delta}{4}}^{\infty} \sqrt{n(C_0, L^\infty(\mu), u)}\, \mathrm{d}u \,. \tag{45}$$

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

$$\|\phi^*(g_1 - q) - \phi^*(g_2 - q)\|_{L^\infty(\mu)} \leq L^1_{\phi^*} \|g_1 - g_2\|_{L^\infty(\mu)} \,, \tag{46}$$

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

$$L^1_{\phi^*} \coloneqq \sup_{x \in [-M_1, M_1]} |\partial \phi^*(x)| \,, \tag{47}$$

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 (42) Lemma 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

$$L^2_{\phi^*} \coloneqq \sup_{x \in [-M_2, M_2]} |\partial \phi^*(x)| \,, \tag{48}$$

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

$$\|z_1^* - z_2^*\|_{L^\infty_{B_R}} \leq \|z_1 - z_2\|_{L^\infty_{B_{G(R)}}} \,, \tag{49}$$

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 (43) 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 4. For this we need to control the *localized* empirical process

$$W(\tau) \coloneqq \sup_{z \in C \cap B^\circ(z_0, \tau)} \langle \phi^*(z - q) - \phi^*(z_0 - q), \mu - \hat{\mu} \rangle \,, \tag{50}$$

and

$$W^*(\tau) \coloneqq \sup_{z \in C \cap B^\circ(z_0, \tau)} \langle \phi^*(z^* - q) - \phi^*(z_0^* - q), \nu - \hat{\nu} \rangle \,, \tag{51}$$

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

Lemma 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 $1 - e^{-t}$*

$$\begin{cases} W(\tau) \leq \frac{8\sqrt{2P_\mu}}{(1-\frac{\alpha}{2})\sqrt{n(L^1_{\phi^*})^\alpha}} (K\tau)^{1-\alpha/2} + K\tau\sqrt{\frac{2t}{n}} + \frac{2b(R)L^1_{\phi^*}}{n} \\ W^*(\tau) \leq \frac{8\sqrt{2P_\nu}}{(1-\frac{\alpha}{2})\sqrt{n(L^2_{\phi^*})^\alpha}} (K'\tau)^{1-\alpha/2} + K'\tau\sqrt{\frac{2t}{n}} + \frac{2b'(R)L^2_{\phi^*}}{n} \,, \end{cases} \tag{52}$$

where $L^1_{\phi^*}, L^2_{\phi^*}$ are defined in Equations (47) and (48) 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 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 $(f, g) \in C$, $\|f - g\|_{L^2(\mu)} \leq K d^{\lambda}_{\phi}(f, g)$ and $\|f^* - g^*\|_{L^2(\nu)} \leq K' d^{\lambda}_{\phi}(f, g)$.

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 $x \in B_R$,

$$|g(x)| \leq L^1_{\phi^*} |z(x) - z_0(x)|\,, \tag{53}$$

where $L^1_{\phi^*} := \sup_{x \in [-M_1, M_1]} |\partial \phi^*(x)|$ 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 $\|z - z_0\|_{L^2(\mu)} \leq K d^{\lambda}_{H^\circ}(z, z_0)$ and in particular, all $g \in C_0$ verifies $\|g\|_{L^2(\mu)} \leq K\tau$. Hence, applying Luxburg and Bousquet (2004, Theorem 16), we obtain for all $\frac{\delta}{4} \leq K\tau$

$$\mathbb{E}[W(\tau)] \leq 2\delta + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\frac{\delta}{4}}^{K\tau} \sqrt{n(C_0, L^2(\mu), u)}\, \mathrm{d}u\,. \tag{54}$$

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

$$\|g_1 - g_2\|_{L^2(\mu)} \leq L^1_{\phi^*} \|z_1 - z_2\|_{L^2(\mu)}\,, \tag{55}$$

and in particular, we recover the upper-bound

$$\mathbb{E}[W(\tau)] \leq 2\delta + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\frac{\delta}{4}}^{K\tau} \sqrt{n(C, L^2(\mu), L^1_{\phi^*} u)}\, \mathrm{d}u. \tag{56}$$

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 $\alpha < 2$, we obtain taking $\delta = 0$ our final upper bound

$$\mathbb{E}[W(\tau)] \leq \frac{4\sqrt{2P_\mu}}{(1 - \frac{\alpha}{2})\sqrt{n(L^1_{\phi^*})^\alpha}} (K\tau)^{1-\alpha/2}\,. \tag{57}$$

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

Lemma 4 (Bousquet, see Theorem 26 in Hütter and 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 $1 - e^{-t}$*

$$\sup_{f\in F} \sqrt{n}|\langle f, \mu - \hat{\mu}\rangle| \leq 2\mathbb{E}[\sup_{f\in F} \sqrt{n}|\langle f, \mu - \hat{\mu}\rangle|] + \sigma\sqrt{2t} + \frac{M}{\sqrt{n}}t\,. \tag{58}$$

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

$$W(\tau) \leq \frac{8\sqrt{2P_\mu}}{(1-\frac{\alpha}{2})\sqrt{n(L^1_{\phi^*})^\alpha}}(K\tau)^{1-\alpha/2} + K\tau\sqrt{\frac{2t}{n}} + \frac{2tb(R)L^1_{\phi^*}}{n}\,, \tag{59}$$

where we used the pointwise upper-bound (53) 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 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 $z \in C, \|(z - z_0)^*\|_{L^2(\nu)} \leq K' d^\lambda_{H^\circ}(z, z_0)$. Hence we recover that with probability at least $1 - e^{-t}$,

$$W^*(\tau) \leq \frac{8\sqrt{2P_\nu}}{(1-\frac{\alpha}{2})\sqrt{n(L^2_{\phi^*})^\alpha}}(K'\tau)^{1-\alpha/2} + K'\tau\sqrt{\frac{2t}{n}} + \frac{2tb'(R)L^2_{\phi^*}}{n}\,. \tag{60}$$

□

We can now prove the second part of Proposition 4.

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

$$d^\lambda_{H^\circ}(\hat{z}_s, z_0)^2 \leq J(\hat{z}_s) - J(z_0)\,. \tag{61}$$

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

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

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

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

$$\frac{\tau^2}{4} \leq \kappa \frac{\tau^{1-\alpha/2}}{\sqrt{n}} + (K + K')\tau \sqrt{\frac{2t}{n}} + \frac{t\kappa'}{n} \,, \tag{62}$$

where κ and κ' are given in Lemma 3 defined as

$$\begin{cases} \kappa = \frac{8\sqrt{2}}{(1-\frac{\alpha}{2})} \left[\frac{\sqrt{P_\mu} K^{1-\alpha/2}}{(L^1_{\varphi^*})^{\frac{\alpha}{2}}} + \frac{\sqrt{P_\nu}(K')^{1-\alpha/2}}{(L^2_{\varphi^*})^{\frac{\alpha}{2}}} \right] \\ \kappa' = 2(M(R) L^1_{\varphi^*} + M'(R) L^2_{\varphi^*}) \,. \end{cases} \tag{63}$$

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

$$\frac{\tau^2}{4} \leq \kappa \frac{\tau^{1-\alpha/2}}{\sqrt{n}} + (K + K')\tau \sqrt{\frac{2t}{n}} + \frac{t\kappa'\tau}{\sqrt{n}} \,. \tag{64}$$

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

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

$$\frac{\tau^2}{4} \leq \frac{t\eta\tau^{1-\alpha/2}}{\sqrt{n}} \,, \tag{65}$$

where $\eta = (\kappa + \kappa' + \sqrt{2}(K + K'))$ 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 $1 - e^{-t}$

$$\sup(A) \leq \frac{(4\eta' t)^{\frac{1}{1+\alpha/2}} + (4\eta' t)}{n^{\frac{1}{2+\alpha}}} \,, \tag{66}$$

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

$$d^{\lambda}_{H^\circ}(\hat{z}, z_0) \leq \frac{(4\eta' t)^{\frac{1}{1+\alpha/2}} + (4\eta' t)}{n^{\frac{1}{2+\alpha}}} \tag{67}$$

$$\leq \frac{8\eta' t}{n^{\frac{1}{2+\alpha}}}\,. \tag{68}$$

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

$$\mathbb{E}[X] = \int_0^\infty P(X > u)\,\mathrm{d}u\,. \tag{69}$$

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

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

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

□

C.2 Proof of Corollary 2

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 $x \in \mathrm{supp}(\nu)$

$$\nabla^2 z_0(x) = [\nabla^2 z_0^*(\nabla z_0(x))]^{-1}\,, \tag{70}$$

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 $\mathrm{supp}(\mu)$ where we defined

$$\frac{1}{\lambda} \coloneqq \sup_{y \in \mathrm{supp}(\nu)} \|\nabla^2 z_0^*(y)\|\,. \tag{71}$$

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 and 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 $\mathrm{supp}(\mu)$. Using again the diffeomorphism property of ∇z_0 between $\mathrm{supp}(\mu)$ and $\mathrm{supp}(\nu)$, we have that $\tilde{z}_0^*$ coincides with z_0^* on $\mathrm{supp}(\nu)$. 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. 4 which yields

$$\mathbb{E}[d_\phi^\lambda(\hat{z}_C, z_0)^2] \lesssim \delta + \frac{1}{\sqrt{n}} \int_{\frac{\delta}{4}}^{\frac{M'}{L}} \sqrt{n(C, L^\infty_{B_{R'}}, Lu)}\mathrm{d}u\,. \tag{72}$$

Finally, using van der Vaart and 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,

$$\begin{aligned}
\frac{1}{\sqrt{n}} \int_{\frac{\delta}{4}}^{\frac{M'}{L}} \sqrt{n(C, L^\infty_{B_{R'}}, Lu)}\mathrm{d}u &\lesssim \frac{1}{\sqrt{n}}(n^{-\frac{k+2}{d}})^{1-\frac{d}{2(k+2)}} && (73)\\
&\lesssim \frac{1}{\sqrt{n}} n^{-\frac{2(k+2)-d}{2d}} && (74)\\
&= n^{-\frac{k+2}{d}}\,. && (75)
\end{aligned}$$

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

$$\mathbb{E}[d_\phi^\lambda(\hat{z}_C, z_0)^2] \lesssim \frac{\log(n)}{\sqrt{n}}\,. \tag{76}$$

Finally, if $\frac{k+2}{d} > 1/2$, we apply the second part of Propostion 4 and we recover the rate

$$\mathbb{E}[d_\phi^\lambda(\hat{z}_C, z_0)^2] \lesssim n^{-1/(1+d/2(k+2))}\,. \tag{77}$$

□

D Proofs of Section 4

D.1 Proof of Proposition 5

We simply adapt the proof of Bubeck (2015, Theorem 3.8).

Proof. Recall that F verifies

$$F(x_{k+1}) - F(x_k) \leq dF(x_k)(x_{k+1} - x_k) + \frac{\beta}{2} A^{x_k}(x_{k+1} - x_k)\,. \tag{78}$$

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

$$F(x_{k+1}) - F(x_k) \leq s_k(F(\bar{x}) - F(x_k)) + s_k^2 \frac{\beta}{2} K\,. \tag{79}$$

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

$$\delta_{k+1} \leq (1 - s_k)\delta_k + s_k^2 K \frac{\beta}{2}\,. \tag{80}$$

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).

□

D.2 Proof of Proposition 6

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 $x \in C$

$$\frac{1}{2}\mathcal{D}_{C,A}(x, \beta) \geq \alpha(F(x) - \bar{F})\,, \tag{81}$$

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

$$\mathcal{D}_{C,A}(x, \beta) = -2\beta \inf_{y \in C} dF(x)(y - x) + \frac{\beta}{2} A^x(y - x)\,. \tag{82}$$

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

Lemma 5. *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*

$$\frac{1}{2}\mathcal{D}_{C,A}(x, \beta) \geq \alpha(F(x) - \bar{F})\,, \tag{83}$$

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

$$x_{k+1} = \arg\min_{y \in C} dF(x_k)(y - x_k) + \frac{\beta}{2} A^{x_k}(y - x_k)\,, \tag{84}$$

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

$$\begin{aligned} F(x_{k+1}) &\leq F(x_k) + dF(x_k)(x_{k+1} - x_k) + \frac{\beta}{2} A^{x_k}(x_{k+1} - x_k) && (85)\\ &\leq F(x_k) - \frac{1}{2\beta}\mathcal{D}_{C,A}(x, \beta) && (86)\\ &\leq F(x_k) - \frac{\alpha}{\beta}(F(x_k) - \bar{F})\,. && (87) \end{aligned}$$

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

$$\begin{aligned} & \Delta_F(y, x) \geq \frac{\alpha}{2} A^x(y - x) && (88)\\ \iff & F(y) - F(x) \geq dF(x)(y - x) + \frac{\alpha}{2} A^x(y - x) && (89)\\ \implies & F(y) - F(x) \geq \inf_{y \in C} dF(x)(y - x) + \frac{\alpha}{2} A^x(y - x) && (90)\\ \iff & 2\alpha(F(x) - F(y)) \leq \mathcal{D}_{C,A}(x, \alpha) && (91)\\ \iff & \frac{1}{2}\mathcal{D}_{C,A}(x, \alpha) \geq \alpha(F(x) - F(y)) && (92)\\ \implies & \frac{1}{2}\mathcal{D}_{C,A}(x, \alpha) \geq \alpha(F(x) - \bar{F})\,. && (93) \end{aligned}$$

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 6. *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 $x, y \in C$,

$$-2\beta(dF(x)((1-\frac{\alpha}{\beta})x + \frac{\alpha}{\beta}y - x) + \frac{\beta}{2}A^x((1-\frac{\alpha}{\beta})x + \frac{\alpha}{\beta}y - x)) \leq \mathcal{D}_{C,A}(x,\beta) \tag{94}$$

$$\iff -2\beta(\frac{\alpha}{\beta}dF(x)(y-x) + \frac{\alpha^2}{2\beta}A^x(y-x)) \leq \mathcal{D}_{C,A}(x,\beta) \tag{95}$$

$$\iff -2\alpha(dF(x)(y-x) + \frac{\alpha}{2}A^x(y-x)) \leq \mathcal{D}_{C,A}(x,\beta)\,. \tag{96}$$

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 5 holds for any C, A, the convexity of C and the 2-homogeneity of A are crucial to derive the monotonic behavior of $\mathcal{D}_{C,A}(x,\cdot)$.