Quantitative Stability of the Pushforward Operation by an Optimal Transport Map

Guillaume Carlier, Alex Delalande, Quentin Mérigot

QUANTITATIVE STABILITY OF THE PUSHFORWARD OPERATION BY AN OPTIMAL TRANSPORT MAP

GUILLAUME CARLIER, ALEX DELALANDE, AND QUENTIN MÉRIGOT

Abstract. We study the quantitative stability of the mapping that to a measure associates its pushforward measure by a fixed (non-smooth) optimal transport map. We exhibit a tight Hölder-behavior for this operation under minimal assumptions. Our proof essentially relies on a new bound that quantifies the size of the singular sets of a convex and Lipschitz continuous function on a bounded domain.

Keywords: Optimal transport, Pushforward measure, Singularities of convex functions.

2020 Mathematics Subject Classification: 49Q22, 49K40, 26B05.

1. Introduction

The optimal transport problem is a two-century old foundational optimization problem of optimal mass allocation in geometric domains [38]. The theoretical study of this problem has allowed to define a natural geometry on spaces of probability measures that offers precious tools for tackling both theoretical and numerical questions involving probability measures [53, 4, 47, 45]. The main feature of this geometry is the *Wasserstein distance*: on the set $\mathcal{P}_2(\mathbb{R}^d)$ of probability measures with finite second moment over $\mathbb{R}^d$, the (2-)Wasserstein distance between two measures $\rho, \mu \in \mathcal{P}_2(\mathbb{R}^d)$, denoted $\mathrm{W}_2(\rho, \mu)$, is defined as the square-root of the value of the following minimization problem:

$$\min_{\gamma \in \Gamma(\rho,\mu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, \mathrm{d}\gamma(x, y), \tag{1}$$

where $\Gamma(\rho, \mu)$ denotes the set of *transport plans* or *couplings* between ρ and μ, that is the set of probability measures over $\mathbb{R}^d \times \mathbb{R}^d$ with first marginal ρ and second marginal μ. Endowed with the Wasserstein distance, the metric space $(\mathcal{P}_2(\mathbb{R}^d), \mathrm{W}_2)$ is a geodesic space referred to as the *Wasserstein space*. In this space, the (constant-speed) geodesics connecting two measures ρ and μ in $\mathcal{P}_2(\mathbb{R}^d)$ are given by the paths $(((1-t)p_1 + tp_2)_{\#}\gamma)_{t\in[0,1]}$ for any $\gamma \in \Gamma(\rho, \mu)$ that minimizes (1), where $p_1 : (x, y) \mapsto x$ and $p_2 : (x, y) \mapsto y$ are the projections onto the first and second coordinates respectively and where $f_{\#}\nu$ denotes the image measure of a measure ν under a map f. Interestingly, the Wasserstein space has found a physically-relevant pseudo-Riemannian structure, which has been leveraged to describe some well known evolution PDEs (such as the Fokker-Planck or porous medium equations) as gradient flows of some energy functionals on the space of probability distributions [42, 27, 43, 4]. In this formal Riemannian interpretation (formal because $(\mathcal{P}_2(\mathbb{R}^d), \mathrm{W}_2)$ is not locally homeomorphic to a Euclidean space or even a Hilbert space), the geometric tangent cone $\mathcal{T}_\rho\mathcal{P}_2(\mathbb{R}^d)$ to $\mathcal{P}_2(\mathbb{R}^d)$ at a measure

$\rho \in \mathcal{P}_2(\mathbb{R}^d)$ can be described as the closure of the set

$$\{(p_1, \lambda(p_2 - p_1))_{\#}\gamma \mid \lambda > 0, \gamma \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d), (p_1)_{\#}\gamma = \rho, \mathrm{spt}(\gamma) \subset \partial\phi,\ \phi \text{ convex}\}$$

with respect to an appropriately chosen Riemannian metric (see Chapter 12 of [4]). In this expression, $\mathrm{spt}(\gamma)$ denotes the support of γ and $\partial\phi$ denotes the subdifferential of the (proper and continuous) convex function $\phi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, that is the set

$$\partial\phi = \{(x, y) \in \mathbb{R}^d \times \mathbb{R}^d \mid \phi(x) + \phi^*(y) = \langle x|y\rangle\},$$

where $\phi^*(\cdot) = \sup_{z \in \mathbb{R}^d} \langle z|\cdot\rangle - \phi(z)$ corresponds to the convex conjugate or Legendre transform of ϕ.

1.1. **Problem statement.** From above, it appears that the *directions* of the elements of the tangent cone $\mathcal{T}_\rho\mathcal{P}_2(\mathbb{R}^d)$ to $\mathcal{P}_2(\mathbb{R}^d)$ at a probability measure ρ are prescribed with convex functions. In the spirit of building a Riemannian logarithmic map, being given a new measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, one may wonder what are the possible *directions* ϕ of the elements of $\mathcal{T}_\rho\mathcal{P}_2(\mathbb{R}^d)$ that support the Wasserstein geodesics connecting ρ to μ. These can be recovered from the convex functions ϕ or ψ^* that solve the following Kantorovich dual problems, which essentially correspond to the convex dual problems of (1) (see e.g. Particular Case 5.16 in [53]):

$$\min_{\phi:\mathbb{R}^d \to \mathbb{R}\cup\{\infty\}} \int_{\mathbb{R}^d} \phi \mathrm{d}\rho + \int_{\mathbb{R}^d} \phi^* \mathrm{d}\mu = \min_{\psi:\mathbb{R}^d \to \mathbb{R}\cup\{\infty\}} \int_{\mathbb{R}^d} \psi^* \mathrm{d}\rho + \int_{\mathbb{R}^d} \psi \mathrm{d}\mu. \tag{2}$$

Conversely, in the spirit of building a Riemannian exponential map, being given a *direction* from a convex function $\phi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, one may wonder what are the possible $t = 1$ endpoints of the geodesics starting from ρ and with initial velocities *directed* by ϕ, that is of the form $(p_1, p_2 - p_1)_{\#}\gamma$ for a coupling γ with first marginal equal to ρ and with support included in $\partial\phi$. Here, the answer is simple and the corresponding endpoint is the measure $(p_2)_{\#}\gamma$. Note that whenever ϕ is differentiable ρ-almost-everywhere, there is only one coupling γ with first marginal equal to ρ and support in $\partial\phi$: it is given by the coupling $\gamma = (\mathrm{id}, \nabla\phi)_{\#}\rho$ and in this case $\nabla\phi$ corresponds to the optimal transport map from ρ to $\mu = (\nabla\phi)_{\#}\rho$ in the Brenier sense [7]. In this setting, the *exponential map* from the base point ρ applied to the direction ϕ reduces to the pushforward measure $(\nabla\phi)_{\#}\rho$.

In this article, we are concerned with the quantitative stability with respect to the base point of the above-described *exponential mapping* (or pushforward operation) in a fixed direction. Namely, we investigate the following problem:

Problem 1.1. *Let $\phi : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a* fixed, *proper and continuous convex function. Let $\rho, \tilde{\rho} \in \mathcal{P}_2(\mathbb{R}^d)$ and consider $\gamma, \tilde{\gamma} \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$ that are such that $(p_1)_{\#}\gamma = \rho$, $(p_1)_{\#}\tilde{\gamma} = \tilde{\rho}$, $\mathrm{spt}(\gamma) \subset \partial\phi$ and $\mathrm{spt}(\tilde{\gamma}) \subset \partial\phi$. Under what conditions on $\phi, \rho, \tilde{\rho}$ and how can one upper bound $\mathrm{W}_2((p_2)_{\#}\gamma, (p_2)_{\#}\tilde{\gamma})$ in terms of $\mathrm{W}_2(\rho, \tilde{\rho})$?*

As mentioned above, whenever the function ϕ is differentiable ρ- and $\tilde{\rho}$-almost-everywhere in Problem 1.1, the question it raises becomes that of upper bounding $\mathrm{W}_2((\nabla\phi)_{\#}\rho, (\nabla\phi)_{\#}\tilde{\rho})$ in terms of $\mathrm{W}_2(\rho, \tilde{\rho})$, which is the question of the quantitative stability of the pushforward operation by an optimal transport map.

1.2. **Motivations.** While of a theoretical nature, Problem 1.1 finds its relevance in several applied contexts. In order to motivate our study, we mention some of these contexts in what follows.

1.2.1. *Numerical resolution of the Kantorovich dual.* Many numerical methods that aim at solving the optimal transport problem (1) between two measures ρ and μ in $\mathcal{P}_2(\mathbb{R}^d)$ rely on the dual problems exposed in (2) (see [45, 39] for surveys on such methods). Focusing for instance on the right-hand side problem in (2), it is possible to add the constraint $\int_{\mathbb{R}^d} \psi \mathrm{d}\mu = 0$ in this problem without altering its value. The resolution of (1) can thus be reduced to the minimization of the *Kantorovich functional* $\mathcal{K}_\rho : \psi \mapsto \int_{\mathbb{R}^d} \psi^* \mathrm{d}\rho$ under the constraint $\int_{\mathbb{R}^d} \psi \mathrm{d}\mu = 0$. The functional $\mathcal{K}_\rho$ being convex, its minimization is amenable to first- and second-order optimization methods. In these methods, the user must be able to evaluate the *gradient* of $\mathcal{K}_\rho$ at a given ψ. Formally, this gradient reads

$$\nabla \mathcal{K}_\rho(\psi) = -(\nabla \psi^*)_\# \rho.$$

Whenever ρ is absolutely continuous, the numerical computation of such a gradient can be challenging in dimension $d \geq 3$. This happens for instance in the setting of semi-discrete optimal transport (where in addition the target μ is assumed to be discrete) that is used to model Euler incompressible equations [20, 24], in computational geometry [34], optics design [37] or in cosmology [41]. In this case, the user might instead consider a finitely supported approximation $\tilde{\rho} = \frac{1}{N} \sum_i \delta_{x_i}$ of ρ, and set

$$\tilde{\mu} := -\frac{1}{N} \sum_i \delta_{y_i}$$

as an approximation for $\nabla \mathcal{K}_\rho(\psi)$, where in this definition each y_i is chosen as an element of the subdifferential $\partial \psi^*(x_i)$. This measure $\tilde{\mu}$ is very easy to compute in practice (one only needs to compute elements of the subdifferential of ψ^*). This procedure then raises the question of the quality of the approximation of $\nabla \mathcal{K}_\rho(\psi)$ offered by $\tilde{\mu}$ in terms of the quality of the approximation of ρ given by $\tilde{\rho}$, which is an instance of Problem 1.1.

1.2.2. *Computation of geodesics and barycenters in the Linearized Optimal Transport framework.* In [54], the above-described pseudo-Riemannian structure of the Wasserstein space was leveraged to *linearize* the optimal transport geometry in order to perform tractable data analysis tasks on measure-like data, giving birth to the *Linearized Optimal Transport* (LOT) framework. In this framework, an absolutely continuous reference measure $\rho \in \mathcal{P}_2(\mathbb{R}^d)$ is chosen and fixed. Because ρ is absolutely continuous, any continuous convex function $\phi : \mathbb{R}^d \to \mathbb{R}$ is differentiable ρ-almost everywhere, so that the tangent bundle $\mathcal{T}_\rho \mathcal{P}_2(\mathbb{R}^d)$ to $\mathcal{P}_2(\mathbb{R}^d)$ at ρ can be regarded (see Chapter 8 of [4]) as the $\mathrm{L}^2(\rho; \mathbb{R}^d)$ closure of

$$\{\lambda(\nabla \phi - \mathrm{id}) \mid \lambda > 0, \phi \text{ convex}\}.$$

Then, the LOT framework maps any new measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ to $\mathrm{L}^2(\rho; \mathbb{R}^d)$ via the embedding $\mu \mapsto \nabla \phi_\mu - \mathrm{id} \in \mathrm{L}^2(\rho; \mathbb{R}^d)$ where ϕ_μ is any minimizer of the left-hand side dual problem in (2). This can be seen as sending $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ into the (linear) tangent space $\mathcal{T}_\rho \mathcal{P}_2(\mathbb{R}^d) \subset \mathrm{L}^2(\rho; \mathbb{R}^d)$ via a Riemannian logarithmic map. The advantage of employing this embedding is to enable the use of all the Hilbertian tools of statistics and machine learning on datasets of probability measures, somehow consistently with the Wasserstein geometry. Note that working with this embedding is equivalent to replacing the Wasserstein distance with the distance

$$\mathrm{W}_{2,\rho}(\mu, \nu) := \|\nabla \phi_\mu - \nabla \phi_\nu\|_{\mathrm{L}^2(\rho; \mathbb{R}^d)}.$$

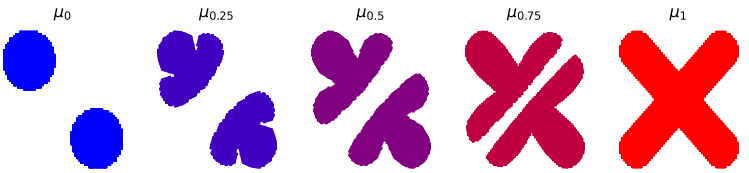


FIGURE 1. *Linearized Optimal Transport* barycenters, or *generalized geodesic*, between the discrete probability measures $\mu_0, \mu_1 \in \mathcal{P}([0,1]^2)$ (colored pixels indicate the position of support points). For $k \in \{0,1\}$, the optimal transport map $\nabla\phi_k$ between the Lebesgue measure ρ on $[0,1]^2$ and μ_k is computed using the Python package `pysdot` and [28]. Then for $N = 70^2$, define $\frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ to be a discrete approximation of ρ on a uniform grid. For $t \in (0,1)$, the interpolant is defined as $\mu_t = \frac{1}{N}\sum_{i=1}^N \delta_{y_i^t}$, where each y_i^t is chosen in $\partial((1-t)\phi_0 + t\phi_1)(x_i)$.

This distance, with respect to which geodesics are called *generalized geodesics* in [4], has been shown to be Hölder-equivalent in some settings to the original Wasserstein distance W_2 in [36, 21], justifying to some extent the successes of the LOT framework witnessed on tasks of pattern recognition [54, 30, 6, 11], generative modeling [44] or image processing [29]. A key advantage of the LOT embedding is that its image is convex, in the sense that any convex combination of embeddings provide a *valid new embedding*. More precisely, for a dataset $(\mu_i)_{1\leq i\leq N}$ of $N \geq 1$ probability measures in $\mathcal{P}_2(\mathbb{R}^d)$ and a set $(\alpha_i)_{1\leq i\leq N}$ of non-negative weights summing to one, the $\mathrm{L}^2(\rho, \mathbb{R}^d)$-barycenter $\sum_{i=1}^N \alpha_i(\nabla\phi_{\mu_i} - \mathrm{id})$ of the embeddings of each μ_i is a valid element of $\mathcal{T}_\rho\mathcal{P}_2(\mathbb{R}^d)$ since it reads as the gradient of a convex function minus identity. One can thus apply the above-described exponential map to this barycenter of embeddings in order to define the measure

$$\bar{\mu} := \left(\sum_{i=1}^N \alpha_i \nabla\phi_{\mu_i}\right)_{\#} \rho.$$

The measure $\bar{\mu}$ gives a notion of average of the dataset $(\mu_i)_{1\leq i\leq N}$ with respect to the weights $(\alpha_i)_{1\leq i\leq N}$ (see Figure 1 for an illustration in the case $N = 2$ with varying weights). It may be used in place of the notion of Wasserstein barycenter [1], which is defined as any minimizer of

$$\min_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \sum_{i=1}^N \alpha_i \mathrm{W}_2^2(\mu, \mu_i). \tag{3}$$

Wasserstein barycenters provide geometrically meaningful notions of averages of datasets of probability measures and have found many successful applications [46, 48, 22, 18, 32, 49, 19, 25]. However, the numerical resolution of (3) is often tedious and working with the proxy $\bar{\mu}$ is often preferable since it essentially requires solving N optimal transport problems between ρ and each μ_i. Nonetheless, it also requires computing the pushforward of ρ by the map $\sum_{i=1}^N \alpha_i\nabla\phi_{\mu_i}$, which can be difficult in dimension $d \geq 3$. In practice, as for the computation of the gradient of the Kantorovich functional above, the user may approximate $\bar{\mu}$ by first discretizing

ρ and then pushing forward this discretization by the the map $\sum_{i=1}^N \alpha_i \nabla\phi_{\mu_i}$. The problem of controlling the bias induced by this process then gives another instance of Problem 1.1.

1.2.3. *Generative modeling with ICNNs.* Over the last decade, tools from optimal transport have made an increasing number of successful incursions in large-scale machine learning problems. These incursions are in part due to the introduction in [5] of the *Input Convex Neural Networks* (ICNNs). These are neural networks whose architecture constraints them to be convex with respect to their input. Such networks were shown to be able to approximate arbitrarily well in supremum norm any convex Lipschitz function on a bounded domain [17]. ICNNs have been used in the context of generative modeling through optimal transport, where one typically wants to learn a *model* for a data probability distribution μ through the observation of samples from it. The optimal transport approach of this problem generally sees μ as the pushforward of a chosen simple probability distribution ρ (typically a Gaussian) by an optimal transport map. This transport map must be learned: in [50, 35, 31, 8], it is parametrized as the gradient of an ICNN ϕ_θ parametrized by θ, and the loss functions used in these works to find the right parameter are essentially proxies for the loss

$$\mathcal{L}(\theta) = \mathrm{W}_2((\nabla\phi_\theta)_\#\rho, \mu).$$

In practice however, neither the source nor target distributions ρ and μ are directly usable or known, and the user has to deal with statistical approximations $\hat\rho$ and $\hat\mu$ instead. This leads to the minimization of the empirical loss function

$$\hat{\mathcal{L}}(\theta) = \mathrm{W}_2((\nabla\phi_\theta)_\#\hat\rho, \hat\mu)$$

in place of the original loss function $\mathcal{L}$. In order to derive convergence rates for this empirical risk minimization problem, one may want to upper bound $|\hat{\mathcal{L}}(\theta) - \mathcal{L}(\theta)|$ in terms of $\mathrm{W}_2(\hat\rho, \rho)$ and $\mathrm{W}_2(\hat\mu, \mu)$. This reduces to yet another instance of Problem 1.1 after a use of the triangle inequality.

1.3. **Positive and negative results.** We expose here a positive result for Problem 1.1 in the case where ϕ is assumed to be regular. We also expose negative results in the case where no assumptions are made on ϕ, ρ and $\tilde\rho$, justifying this way the necessity for minimal assumptions.

1.3.1. *A positive result in the regular case.* In the case where the convex function ϕ of Problem 1.1 is of class $\mathcal{C}^{1,\alpha}$ for some $\alpha > 0$, the answer to the question raised in this problem is trivial:

Proposition. *Let $\alpha \in (0, 1]$ and $\phi \in \mathcal{C}^{1,\alpha}(\mathbb{R}^d)$ convex. Then for any $\rho, \tilde\rho \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$\mathrm{W}_2((\nabla\phi)_\#\rho, (\nabla\phi)_\#\tilde\rho) \leq \|\nabla\phi\|_{\mathcal{C}^{0,\alpha}} \mathrm{W}_2(\rho, \tilde\rho)^\alpha.$$

This follows from Jensen's inequality and the fact that for any $\gamma \in \Gamma(\rho, \tilde\rho)$, $(\nabla\phi, \nabla\phi)_\#\gamma$ is a valid coupling between $(\nabla\phi)_\#\rho$ and $(\nabla\phi)_\#\tilde\rho$. Even though this proposition brings an answer to Problem 1.1, its outreach is limited. Indeed, when ϕ is an optimal transport potential (i.e. a solution to a dual problem of the type of (2)), getting regularity estimates for ϕ requires in general to make strong regularity assumptions on the involved measures in order to be able to apply Caffarelli's regularity theory results [9, 10], assumptions that are rarely satisfied in applications where at least one of the considered measures is often discrete. For instance, when ϕ is the dual solution of a semi-discrete optimal transport problem (with absolutely

continuous source and discrete target), ϕ corresponds to a maximum of affine functions and as such it has many singularities. These singularities are actually often desirable, as for instance in the context of generative modeling of a data probability distribution μ with disconnected support as the pushforward of a Gaussian ρ by the gradient of a convex function [35].

1.3.2. *Negative results.* Whenever ϕ has singularities, it is very easy to build measures $\rho, \tilde{\rho}$ and couplings $\gamma, \tilde{\gamma}$ in Problem 1.1 that are such that it is not possible to control $\mathrm{W}_2((p_2)_\#\gamma, (p_2)_\#\tilde{\gamma})$ in terms of $\mathrm{W}_2(\rho, \tilde{\rho})$. Consider for instance in dimension $d = 1$ the case where $\phi = |\cdot|$ is the absolute value. Then ϕ has a singularity at 0:

$$\partial\phi(0) = [-1, 1].$$

A first negative result is available when both ρ and $\tilde{\rho}$ are allowed to be discrete:

Example 1.2. Let $\phi = |\cdot|$ on $\mathbb{R}$. Let $\rho = \tilde{\rho} = \delta_0$ be the the Dirac mass at zero and let $\gamma = \delta_{(0,1)}$ and $\tilde{\gamma} = \delta_{(0,-1)}$. Then $(p_1)_\#\gamma = \rho$, $(p_1)_\#\tilde{\gamma} = \tilde{\rho}$, $\mathrm{spt}(\gamma) \subset \partial\phi$ and $\mathrm{spt}(\tilde{\gamma}) \subset \partial\phi$. However $\mathrm{W}_2((p_2)_\#\gamma, (p_2)_\#\tilde{\gamma}) = 2$ while $\mathrm{W}_2(\rho, \tilde{\rho}) = 0$.

This example relies on placing both ρ and $\tilde{\rho}$ at singularities of the convex function ϕ. The set of singular points (i.e. points of non-differentiability) of a convex function defined on $\mathbb{R}$ being at most countable, one can wonder what happens if we constraint one of the source measures in Problem 1.1 to be absolutely continuous with respect to the Lebesgue measure. Under such a constraint, it is still possible to build source measures that are arbitrarily close from each other but with pushforwards that are at a fixed non-zero distance from each other:

Example 1.3. Let $\phi = |\cdot|$ on $\mathbb{R}$ and let $\varepsilon > 0$. Let $\rho = \delta_0$ and let $\rho^\varepsilon = \frac{1}{\varepsilon}\lambda_{|[-\frac{\varepsilon}{2}, \frac{\varepsilon}{2}]}$ be the rescaled Lebesgue measure restricted to $[-\frac{\varepsilon}{2}, \frac{\varepsilon}{2}]$. Let $\gamma = \delta_{(0,1)}$ and $\gamma^\varepsilon = (\mathrm{id}, \nabla\phi)_\#\rho^\varepsilon$. Then $(p_1)_\#\gamma = \rho$, $(p_1)_\#\gamma^\varepsilon = \rho^\varepsilon$, $\mathrm{spt}(\gamma) \subset \partial\phi$ and $\mathrm{spt}(\gamma^\varepsilon) \subset \partial\phi$. However $\mathrm{W}_2((p_2)_\#\gamma, (p_2)_\#\gamma^\varepsilon) = \sqrt{2}$ while $\mathrm{W}_2(\rho, \rho^\varepsilon) = \varepsilon/2\sqrt{3}$.

Example 1.3 relies on an absolutely continuous source measure ρ^ε whose density is allowed to explode so as to recover in the limit $\varepsilon \to 0$ the pathological case of Example 1.2 with only discrete sources. In order to avoid this problem, we will make from now on the following minimal assumption in Problem 1.1: one of the probability measures, say ρ, is absolutely continuous with respect to the Lebesgue measure and its density is upper bounded by some finite constant $M_\rho > 0$. Under this assumption, it is still possible to build an example (not as bad as Examples 1.2 and 1.3) showing that one cannot expect better than a Hölder-behavior for the pushforward operation:

Example 1.4. (See Figure 2 for an illustration.) Let $\phi = |\cdot|$ on $\mathbb{R}$ and let $\varepsilon \in (0, \frac{1}{2})$. Let $\rho = \lambda_{[-\frac{1}{2}, \frac{1}{2}]}$ and let $\rho^\varepsilon = \lambda_{|[-\frac{1}{2}, -\frac{\varepsilon}{2}]\cup[\frac{\varepsilon}{2}, \frac{1}{2}]} + \varepsilon\delta_0$. Let $\gamma = (\mathrm{id}, \nabla\phi)_\#\rho$ and let $\gamma^\varepsilon = \int_{[-\frac{1}{2}, -\frac{\varepsilon}{2}]\cup[\frac{\varepsilon}{2}, \frac{1}{2}]} \delta_x \otimes \delta_{\nabla\phi(x)} \mathrm{d}x + \varepsilon\delta_{(0,1)}$. Then $(p_1)_\#\gamma = \rho$, $(p_1)_\#\gamma^\varepsilon = \rho^\varepsilon$, $\mathrm{spt}(\gamma) \subset \partial\phi$ and $\mathrm{spt}(\gamma^\varepsilon) \subset \partial\phi$. Moreover, $\mathrm{W}_2((p_2)_\#\gamma, (p_2)_\#\gamma^\varepsilon) = (2\varepsilon)^{1/2}$ while $\mathrm{W}_2(\rho, \rho^\varepsilon) = (\varepsilon^3/12)^{1/2}$, so that $\mathrm{W}_2((p_2)_\#\gamma, (p_2)_\#\gamma^\varepsilon) \sim \mathrm{W}_2(\rho, \rho^\varepsilon)^{1/3}$.

1.4. **Contributions and outline.** In this article, we limit ourselves to a compact setting and work with probability measures supported in a ball $\Omega = B(0, R) \subset \mathbb{R}^d$ centered at zero and of radius $R > 0$. In this context, we assume that the convex function ϕ of Problem 1.1 is an R-Lipschitz continuous convex function in order to

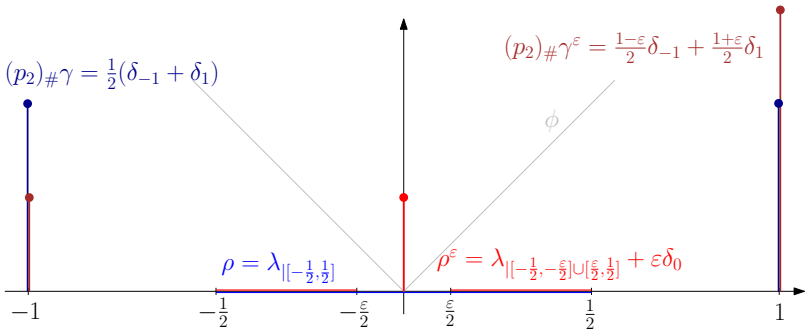


FIGURE 2. Illustration of Example 1.4.

ensure that the *pushforward* measures $(p_2)_\#\gamma$ and $(p_2)_\#\tilde{\gamma}$ of this problem also live in Ω. We emphasize on the fact that we make *no regularity assumption* on $\nabla\phi$.

Our main result shows that, perhaps surprisingly, the situation described in Example 1.4 is as bad as it could get, and the Hölder-behavior being observed in this example is a general phenomenon:

Theorem (Theorem 3.2)**.** *Let $R > 0$ and let $\Omega = B(0, R) \subset \mathbb{R}^d$. Let $\phi : \Omega \to \mathbb{R}$ be an R-Lipschitz continuous convex function. Let $\rho, \tilde{\rho} \in \mathcal{P}(\Omega)$ and assume that ρ is absolutely continuous with density upped bounded by a constant $M_\rho \in (0, \infty)$. Then for any $\tilde{\gamma} \in \mathcal{P}(\Omega \times \Omega)$ such that $(p_1)_\#\tilde{\gamma} = \tilde{\rho}$ and $\mathrm{spt}(\tilde{\gamma}) \subset \partial\phi$,*

$$\mathrm{W}_2((\nabla\phi)_\#\rho, (p_2)_\#\tilde{\gamma}) \lesssim \mathrm{W}_2(\rho, \tilde{\rho})^{1/3},$$

where $\lesssim$ hides an explicit multiplicative constant that depends on d, R and M_ρ.

We refer to Theorem 3.2 in Section 3 for a more precise statement, with in particular an explicit expression for the hidden constant. Note also that the statement of Theorem 3.2 is not limited to the context of quadratic optimal transport (i.e. optimal transport with respect to the cost $c(x, y) = \|x - y\|^2$) but deals with the more general case of pushforwards by transport maps that are optimal with respect to the *p-cost* $c(x, y) = \|x - y\|^p$ for $p \geq 2$; and that the bounds are expressed in W_q and W_r distances (with q and r parameters to be chosen) in order to ensure the highest generality.

We are now in place of sketching the proof of our main result and the outline of the rest of the article. In Examples 1.2, 1.3 and 1.4, we have seen that the *instabilities* in the pushforward operation by the gradient of a convex Lipschitz function arise from the singularities of this function. Our main technical result, presented in Theorem 2.1 of Section 2 and which might be of independent interest, shows that on a bounded domain the number of singularities of such a function can be explicitly bounded. More precisely, for ϕ a convex Lipschitz function defined on $\mathbb{R}^d$, we present in Theorem 2.1 a tight upper bound on the covering numbers of the singular sets

$$\Sigma_{\eta,\alpha} = \{x \in \Omega \mid \mathrm{diam}(\partial\phi(B(x, \eta))) \geq \alpha\},$$

where $\alpha > 0$ and $\eta > 0$. In Remark 2.1, we note that the bound of Theorem 2.1 may be seen as a refinement of a well-known result of Alberti, Ambrosio and Cannarsa [2], who derived upper bounds on the dimension of the singular sets of semi-convex functions using measure-theoretic arguments, falling into the long line of works that

studied the structure of the singularities of solutions to Hamilton-Jacobi equations [51, 52, 26, 13, 14, 40, 3] – see [12] for a survey. As an immediate corollary to Theorem 2.1 – presented in Corollary 2.2 of Section 2 – we deduce that the function ϕ from this theorem satisfies the following integral estimate for any $\eta > 0$:

$$\int_\Omega \operatorname{diam}(\partial\phi(B(x,\eta)))^2 \mathrm{d}x \lesssim \eta. \tag{4}$$

In Section 3, after recalling some facts on the optimal transport problem with a general ground cost, we state and prove the main result Theorem 3.2 that brings a tight answer to Problem 1.1. This result essentially relies on (4) and a Markov bound. Let us sketch here the main idea: denote $S : \Omega \to \Omega$ the optimal transport map from ρ to $\tilde{\rho}$ and for $\eta > 0$, introduce the set $\Omega_\eta = \{x \in \Omega \mid \|S(x) - x\| \leq \eta\}$. Then, one has that $\mathrm{W}_2(\rho, \tilde{\rho}) = \|S - \mathrm{id}\|_{\mathrm{L}^2(\rho,\mathbb{R}^d)}$, so that Markov's inequality entails

$$\rho(\Omega \setminus \Omega_\eta) \lesssim \frac{\mathrm{W}_2^2(\rho,\tilde{\rho})}{\eta^2}. \tag{5}$$

Bounds (4) and (5) allow to conclude: assuming here for simplicity that ϕ is differentiable $\tilde{\rho}$-almost-everywhere, we have for any $\eta > 0$

$$\begin{aligned}
\mathrm{W}_2^2((\nabla\phi)_\#\rho, (\nabla\phi)_\#\tilde{\rho}) &\leq \int_{\Omega_\eta} \|\nabla\phi - \nabla\phi \circ S\|^2 \,\mathrm{d}\rho + \int_{\Omega\setminus\Omega_\eta} \|\nabla\phi - \nabla\phi\circ S\|^2 \,\mathrm{d}\rho \\
&\leq M_\rho \int_{\Omega_\eta} \operatorname{diam}(\partial\phi(B(x,\eta)))^2 \mathrm{d}x + \int_{\Omega\setminus\Omega_\eta} (2R)^2 \mathrm{d}\rho \\
&\lesssim \eta + \frac{\mathrm{W}_2^2(\rho,\tilde{\rho})}{\eta^2}.
\end{aligned}$$

Setting $\eta = \mathrm{W}_2^{2/3}(\rho,\tilde{\rho})$ allows to reach the conclusion of Theorem 3.2.

2. Covering number of near-singularity sets of convex functions

The following result allows to quantify the size of the singular sets (i.e. points of non-differentiability) of a convex Lipschitz function on a bounded domain. We bound here the covering numbers of the sets of points x in the domain for which there exist two *nearby* points $x^\pm$, i.e. such that $\|x - x^\pm\| \leq \eta$, where the gradients of the convex function ϕ are *far* from each other, i.e. such that $\|\nabla\phi(x^+) - \nabla\phi(x^-)\| \geq \alpha$. In this statement, $\mathcal{N}(K,\eta)$ denotes the minimum number of balls of radius $\eta > 0$ that are needed to cover a compact set $K \subset \mathbb{R}^d$.

Theorem 2.1. *Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be a convex and Lipschitz continuous function. Denote*

$$\begin{aligned}
\Sigma_{\eta,\alpha} &= \{x \in \mathbb{R}^d \mid \operatorname{diam}(\partial\phi(B(x,\eta)) \geq \alpha\}, \\
\Sigma_\alpha &= \{x \in \mathbb{R}^d \mid \operatorname{diam}(\partial\phi(x)) \geq \alpha\}.
\end{aligned}$$

Then, for all $R > 0$, α and $\eta > 0$, we have

$$\mathcal{N}(\Sigma_{\eta,\alpha} \cap B(0,R), 8\eta) \leq c_{d,R,\eta} \frac{\mathrm{Lip}(\phi)}{\alpha\eta^{d-1}},$$

with $c_{d,R,\eta} = 48d^2(R+4\eta)^{d-1}$. In particular, there exists a dimensional constant c_d such that

$$\mathcal{H}^{d-1}(\Sigma_\alpha \cap B(0,R)) \leq c_d \frac{\mathrm{Lip}(\phi) R^{d-1}}{\alpha}.$$

As a corollary to this result, we get the following estimate that will prove useful in Section 3 for the study of the stability of the pushforward operation by an optimal transport map.

Corollary 2.2. *Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be a convex and Lipschitz continuous function. Then for any $\eta, R > 0$ and $q > 1$,*

$$\int_{B(0,R)} \operatorname{diam}(\partial\phi(B(x,\eta)))^q \mathrm{d}x \leq c_{d,q,R,\eta} \operatorname{Lip}(\phi)^q \eta,$$

with $c_{d,q,R,\eta} = 48d^2\beta_d 2^{3d+q-1}\frac{q}{q-1}(R+4\eta)^{d-1}$, where β_d denotes the volume of the unit ball of $\mathbb{R}^d$.

Proof. From Theorem 2.1, we directly get

$$\begin{aligned}\int_{B(0,R)} \operatorname{diam}(\partial\phi(B(x,\eta)))^q \mathrm{d}x &= \int_0^\infty |\{x \in B(0,R) \mid \operatorname{diam}(\partial\phi(B(x,\eta)))^q \geq t\}| \,\mathrm{d}t \\ &\leq \int_0^{(2\operatorname{Lip}(\phi))^q} 48d^2(R+4\eta)^{d-1}\frac{\operatorname{Lip}(\phi)}{t^{1/q}\eta^{d-1}}\beta_d(8\eta)^d \mathrm{d}t \\ &= c_{d,q,R,\eta}\operatorname{Lip}(\phi)^q\eta. \end{aligned}$$

□

Remark 2.1 (Singular sets of a convex Lipschitz function)**.** For $k \geq 1$, the k-singular set Σ^k of $\phi : \mathbb{R}^d \to \mathbb{R}$ corresponds to the set of points x in $\mathbb{R}^d$ such that the Hausdorff dimension of $\partial\phi(x)$ is greater than or equal to k. The fact that the k-singular set of a convex Lipschitz function $\phi : \mathbb{R}^d \to \mathbb{R}$ is countably $\mathcal{H}^{d-k}$-rectifiable was established by Alberti, Ambrosio and Cannarsa in [2]. They also established the following estimate on the size of Σ^k:

$$\int_{\Sigma^k \cap B(0,R)} \mathcal{H}^k(\partial\phi(x))\mathrm{d}\mathcal{H}^{d-k}(x) \leq c_d \left(\operatorname{Lip}(\phi) + 2R\right)^d,$$

where c_d is a dimensional constant. With the notation of Theorem 2.1, taking $k = 1$ in this estimate and using Markov's inequality allows to get the bound

$$\mathcal{H}^{d-1}(\Sigma_\alpha \cap B(0,R)) \leq c_d \frac{(\operatorname{Lip}(\phi) + 2R)^d}{\alpha},$$

that is similar in spirit to the bound we present in Theorem 2.1. However, the approach in [2] does not give an estimate on the covering numbers of $\Sigma_{\eta,\alpha}$, which may prove necessary in specific contexts (see for instance the proof of Theorem 3.2 in the next section that relies on Corollary 2.2). In this sense, the quantitative estimate of Theorem 2.1 can be seen as a refinement of the estimate from [2] on the size of the set of non-differentiability points of a convex Lipschitz function on a bounded domain.

Remark 2.2 (Tightness)**.** The bounds presented in Theorem 2.1 are tight. Indeed, in dimension $d = 1$, let $N \in \mathbb{N}^*$ and $L, R > 0$ and define on $\mathbb{R}$ the function

$$\xi : x \mapsto \max_{i=0,\dots,N} \left(\frac{2i}{N} - 1\right) Lx + \frac{2LR}{N(N+1)} i(N-i).$$

Then ξ is convex and L-Lipschitz continuous (see Figure 3 for an illustration of the graph of ξ when $N = 4$). Moreover, denoting $x_i = \left(\frac{2i}{N+1} - 1\right) R$ for all i in

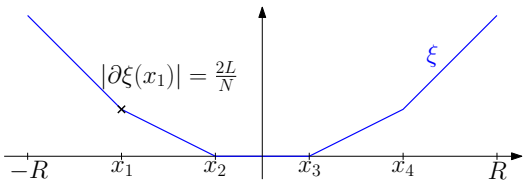


FIGURE 3. Graph of ξ for $N = 4$.

$\{1, \dots, N\}$, this function satisfies for all such i

$$\partial\xi(x_i) = \left[\left(\frac{2i}{N} - 1\right) L, \left(\frac{2(i+1)}{N} - 1\right) L\right],$$

and ξ is differentiable everywhere else in $\mathbb{R} \setminus \{x_i\}_{\{1 \le i \le N\}}$. In particular, setting $\alpha = \frac{2L}{N}$, one can observe that for $\eta > 0$,

$$\Sigma_{\eta,\alpha} = \{x \in \mathbb{R} \mid \operatorname{diam}(\partial\xi(B(x,\eta)) \ge \alpha\} = \bigcup_{1 \le i \le N} [x_i - \eta; x_i + \eta],$$
$$\text{and } \Sigma_\alpha = \{x \in \mathbb{R} \mid \operatorname{diam}(\partial\xi(x)) \ge \alpha\} = \{x_i \mid 1 \le i \le N\},$$

so that for $\eta \in (0, \frac{R}{8(N+1)})$,

$$\mathcal{H}^0(\Sigma_\alpha \cap [-R, R]) = \mathcal{N}(\Sigma_{\eta,\alpha} \cap [-R, R], 8\eta) = N = \frac{2L}{\alpha},$$

which shows the tightness of the bounds of Theorem 2.1 in dimension $d = 1$. In dimension $d \ge 1$, one may generalize this example by defining

$$\phi : x \mapsto \xi(x^1),$$

where x^1 is the projection of $x \in \mathbb{R}^d$ on its first coordinate. Then, with the notations of Theorem 2.1, there are dimensional constants $c_d, \tilde{c_d}$ such that the convex and L-Lipschitz continuous function ϕ verifies for $\alpha = \frac{2L}{N}$ and $\eta \in (0, \frac{R}{8(N+1)})$:

$$\mathcal{N}(\Sigma_{\eta,\alpha} \cap B(0,R), 8\eta) \ge c_d \frac{R^{d-1}}{\eta^{d-1}} \frac{2L}{\alpha}$$
$$\mathcal{H}^{d-1}(\Sigma_\alpha \cap B(0,R)) \ge \tilde{c_d} R^{d-1} \frac{2L}{\alpha}.$$

These last inequalities also show the tightness of the bounds of Theorem 2.1 with respect to R and $\operatorname{Lip}(\phi)$. In comparison to the estimate obtained on the $(d-1)$-Hausdorff measure of $\Sigma_\alpha \cap B(0,R)$ deduced from [2] in Remark 2.1, this tightness corresponds to yet another refinement of the estimates from [2].

The proof of Theorem 2.1 uses the following lemma, similar to [15, Lemma 3.2], and whose proof is postponed after the proof of Theorem 2.1.

Lemma 2.3. *Let ϕ be a convex function over $\mathbb{R}^d$. Then for any $x \in \mathbb{R}^d$ and $\eta > 0$,*

$$\operatorname{diam}(\partial\phi(B(x,\eta))) \le \frac{12}{\beta_d \eta^d} \|\nabla\phi\|_{\mathrm{L}^1(B(x,4\eta))},$$

where β_d denotes the volume of the unit ball of $\mathbb{R}^d$.

With Lemma 2.3 in hand, we are now ready to prove Theorem 2.1.

Proof of Theorem 2.1. Let $\Sigma = \Sigma_{\eta,\alpha}$, and let $Z \subseteq \Sigma$ be a maximal ε-packing of Σ with $\varepsilon = 4\eta$, i.e. a finite subset of Σ satisfying $\forall y \neq z \in Z,\ B(y,\varepsilon) \cap B(z,\varepsilon) = \emptyset$ and which is maximal with respect to the inclusion in the class of subsets of Σ satisfying this assumption. We denote by N the cardinal number of Z. For any $x \in Z$, Lemma 2.3 gives us for any $c \in \mathbb{R}^d$

$$\alpha \leq \operatorname{diam}\left(\partial\phi(B(x,\eta))\right) \leq \frac{12}{\beta_d \eta^d} \left\|\nabla\phi - c\right\|_{\mathrm{L}^1(B(x,4\eta))}. \tag{6}$$

Choosing $c = \frac{1}{|B(x,4\eta)|}\int_{B(x,4\eta)} \nabla\phi(u)\mathrm{d}u$, the Poincaré-Wirtinger inequality then ensures

$$\left\|\nabla\phi - c\right\|_{\mathrm{L}^1(B(x,4\eta))} \leq 4\eta \int_{B(x,4\eta)} \left\|\mathrm{D}^2\phi(u)\right\|_{1,1} \mathrm{d}u.$$

Using that for any positive semi-definite $d \times d$ matrix M, $\|M\|_{1,1} \leq d\mathrm{tr}(M)$, we then have

$$\left\|\nabla\phi - c\right\|_{\mathrm{L}^1(B(x,4\eta))} \leq 4\eta d \int_{B(x,4\eta)} \Delta\phi(u)\mathrm{d}u,$$

where Δ stands for the Laplace operator. Injecting this last bound into (6) yields

$$\alpha \leq \frac{48d}{\beta_d}\frac{1}{\eta^{d-1}} \int_{B(x,4\eta)} \Delta\phi(u)\mathrm{d}u,$$

Summing the last bound over $x \in Z$ and using that the balls of radius $4\eta \leq \varepsilon$ centered at points of Z do not intersect, we get

$$\begin{aligned}
\alpha N &\leq \frac{48d}{\beta_d}\frac{1}{\eta^{d-1}} \sum_{x\in Z} \int_{B(x,4\eta)} \Delta\phi(u)\mathrm{d}u \\
&\leq \frac{48d}{\beta_d}\frac{1}{\eta^{d-1}} \int_{B(0,R+4\eta)} \Delta\phi(u)\mathrm{d}u \\
&\leq \frac{48d}{\beta_d}\omega_{d-1}(R+4\eta)^{d-1}\frac{\mathrm{Lip}(\phi)}{\eta^{d-1}}
\end{aligned}$$

where we used an integration by part to get the last inequality and where $\omega_{d-1} = d\beta_d$ denotes the surface area of the $(d-1)$-unit sphere. Finally, we can easily check that Z is a 2ε-covering of $\Sigma_{\alpha,\eta}$, implying the first bound of the statement.

To prove the second inequality, first note that $\Sigma_\alpha \subseteq \Sigma_{\alpha,\eta}$, so that for any $\eta \leq R$ one has

$$\mathcal{N}(\Sigma_\alpha \cap B(0,R),\eta) \leq c_d \frac{\mathrm{Lip}(\phi)R^{d-1}}{\alpha\eta^{d-1}}.$$

We conclude using $\mathcal{H}^{d-1}(X) \leq c_d \liminf_{\eta\to 0} \eta^{d-1}\mathcal{N}(X,\eta)$, where c_d is a dimensional constant. □

We finally prove Lemma 2.3.

Proof of Lemma 2.3. Let $x \in \mathbb{R}^d$ and $\eta > 0$. One has by definition:

$$\begin{aligned}
\operatorname{diam}(\partial\phi(B(x,\eta))) &= \sup_{y,y'\in B(x,\eta)} \sup_{g\in\partial\phi(y),g'\in\partial\phi(y')} \|g - g'\| \\
&\leq \sup_{y,y'\in B(x,\eta)} \sup_{g\in\partial\phi(y),g'\in\partial\phi(y')} \|g\| + \|g'\| \\
&= 2 \sup_{y\in B(x,\eta)} \sup_{g\in\partial\phi(y)} \|g\| \\
&= 2\left\|\partial\phi\right\|_{\mathrm{L}^\infty(B(x,\eta))}.
\end{aligned}$$

But for any $y, y' \in \mathbb{R}^d$ and $g \in \partial\phi(y)$, the convexity of ϕ entails

$$\langle g | y' - y \rangle \leq |\phi(y') - \phi(y)| .$$

Therefore, choosing $y \in B(x,\eta)$ and $g \in \partial\phi(y)$ such that $\|\partial\phi\|_{\mathrm{L}^\infty(B(x,\eta))} = \|g\|$, one has for $y' = y + \eta \frac{g}{\|g\|} \in B(y,\eta) \subset B(x,2\eta)$ the following bound:

$$\eta \|g\| \leq |\phi(y') - \phi(y)| \leq \mathrm{osc}_{B(x,2\eta)}(\phi),$$

where $\mathrm{osc}_K(f) = \sup_{u,v\in K} |f(u) - f(v)|$. We thus have shown

$$\mathrm{diam}(\partial\phi(B(x,\eta))) \leq \frac{2}{\eta} \mathrm{osc}_{B(x,2\eta)}(\phi). \tag{7}$$

We conclude exactly as in the proof of Lemma 3.2 of [15], that we report here only for completeness: let $y_0 \in \arg\min_{B(x,2\eta)} \phi, y_1 \in \arg\max_{B(x,2\eta)} \phi, g_1 \in \partial\phi(y_1)$. Then by convexity of ϕ, for any $y \in \mathbb{R}^d$ and $g \in \partial\phi(y)$ one has

$$\phi(y_1) + \langle g_1 | y - y_1 \rangle \leq \phi(y) \leq \phi(y_0) + \langle g | y - y_0 \rangle.$$

It follows that

$$\|g\| \geq \frac{\mathrm{osc}_{B(x,2\eta)}(\phi) + \langle g_1 | y - y_1 \rangle}{\|y - y_0\|}.$$

Introducing $W_\eta(y_1, g_1) = \{y \in B(y_1, 2\eta) | \langle g_1 | y - y_1 \rangle \geq 0\} \subset B(x, 4\eta)$, one then has

$$\begin{aligned}
\|\nabla\phi\|_{\mathrm{L}^1(B(x,4\eta))} &\geq \int_{W_\eta(y_1,g_1)} \|\nabla\phi\| \, \mathrm{d}y \\
&\geq \int_{W_\eta(y_1,g_1)} \frac{\mathrm{osc}_{B(x,2\eta)}(\phi)}{\|y - y_0\|} \mathrm{d}y \\
&\geq \mathrm{osc}_{B(x,2\eta)}(\phi) \int_{W_\eta(y_1,g_1)} \frac{1}{\|y - y_1\| + \|y_1 - y_0\|} \mathrm{d}y \\
&\geq \frac{\mathrm{osc}_{B(x,2\eta)}(\phi)}{6\eta} \int_{B(y_1 + \eta\frac{g_1}{\|g_1\|}, r)} \mathrm{d}y \\
&= \beta_d \frac{\eta^{d-1}}{6} \mathrm{osc}_{B(x,2\eta)}(\phi),
\end{aligned}$$

where β_d denotes the volume of the unit ball of $\mathbb{R}^d$ and where we used the fact that $B(y_1 + \eta\frac{g_1}{\|g_1\|}, \eta) \subset W_\eta(y_1, g_1)$. Plugging this last bound into (7) finally yields

$$\mathrm{diam}(\partial\phi(B(x,\eta))) \leq \frac{12}{\omega_d \eta^d} \|\nabla\phi\|_{\mathrm{L}^1(B(x,4\eta))} . \qquad \square$$

3. Stability of the pushforward by an optimal transport map

3.1. **Optimal transportation problem.** We start this section by recalling some facts about the optimal transport problem with a general ground cost and discuss the existence and properties of optimal transport maps.

3.1.1. *Primal and dual formulations.* Let $\Omega = B(0,R)$ be the open ball of $\mathbb{R}^d$ centered at zero and of radius $R > 0$. For $\rho, \mu \in \mathcal{P}(\Omega)$ two probability measures supported over Ω, Kantorovich's formulation of the optimal transport problem between ρ and μ with respect to a continuous cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ corresponds to the following minimization problem:

$$\inf_{\Gamma(\rho,\mu)} \int_{\Omega\times\Omega} c(x,y)\mathrm{d}\gamma(x,y). \tag{8}$$

In this problem, the optimization is over the set $\Gamma(\rho,\mu)$ of couplings (or transport plans) between ρ and μ, i.e. the set of probability measures over $\Omega \times \Omega$ with first marginal ρ and second marginal μ. It is well-known (see e.g. Chapter 1 of [47]) that problem (8) always admits a minimizer (possibly non-unique) and that it enjoys the following dual formulation, holding with strong-duality:

$$\sup_{\varphi:\Omega\to\mathbb{R}} \int_\Omega \varphi \mathrm{d}\rho + \int_\Omega \varphi^c \mathrm{d}\mu, \tag{9}$$

where $\varphi^c(\cdot) = \inf_{x\in\Omega} c(x,\cdot) - \varphi(x)$ corresponds to the c-transform of φ. In turn, problem (9) always admits a maximizer φ (non-unique), which is referred to as a *Kantorovich potential* and which must verify $(\varphi^c)^{\bar{c}} = \varphi$, where $\psi^{\bar{c}}(\cdot) = \inf_{y\in\mathbb{R}^d} c(\cdot,y) - \psi(y)$ is a $\bar{c}$-transform.

3.1.2. *Wasserstein distances.* Whenever the cost function corresponds to the p-cost $c(x,y) = \xi_p(x-y)$ where $\xi_p(z) = \|z\|^p$ for some $p \geq 1$, the p-th root of the value of problem (8) defines the p-Wasserstein distance between the probability measures ρ and μ, denoted $\mathrm{W}_p(\rho,\mu)$. Wasserstein distances come with strong geometrical and physical interpretations that have made their success in many theoretical and applied contexts, see e.g. [53, 47, 45] for references.

When $p \geq 2$, the p-cost satisfies some immediate but strong regularity properties that we will exploit. In the following statement (whose proof can be found in the appendix), f is said λ-concave with $\lambda \in \mathbb{R}$ if $f + \frac{\lambda}{2}\|\cdot\|^2$ is a concave function.

Lemma 3.1 (Properties of p-cost)**.** *Let $p \geq 2$. On $\Omega = B(0,R)$, the mapping $z \mapsto \xi_p(z) = \|z\|^p$ is strictly convex, of class $\mathcal{C}^2$, (pR^{p-1})-Lipschitz continuous and $(-p(p-1)R^{p-2})$-concave. The mapping $z \mapsto (\nabla\xi_p)^{-1}(z)$ is well-defined: for any $z \in \mathbb{R}^d \setminus \{0\}$,*

$$(\nabla\xi_p)^{-1}(z) = \frac{1}{p^{\frac{1}{p-1}}\|z\|^{\frac{p-2}{p-1}}} z,$$

and $(\nabla\xi_p)^{-1}(0) = 0$. In particular, $(\nabla\xi_p)^{-1}$ is $\frac{1}{p-1}$-Hölder continuous:

$$\forall x,y \in \Omega, \quad \left\|(\nabla\xi_p)^{-1}(y) - (\nabla\xi_p)^{-1}(x)\right\| \leq \frac{3}{p^{\frac{1}{p-1}}}\|y-x\|^{\frac{1}{p-1}}.$$

3.1.3. *Optimal transport maps.* By duality, one can observe that any $\gamma \in \Gamma(\rho,\mu)$ and $\varphi : \Omega \to \mathbb{R}$ are respective solutions of problems (8) and (9) if and only if

$$\mathrm{spt}(\gamma) \subset \partial^c\varphi := \{(x,y) \mid \varphi(x) + \varphi^c(y) = c(x,y)\}, \tag{10}$$

where $\mathrm{spt}(\gamma)$ denotes the support of γ. Incidentally, this observation allows to characterize cases of uniqueness of the solutions to problem (8) depending on the choice of cost function c and the assumptions made on the involved measures ρ and μ. Choose for instance the p-cost $c = \xi_p$ with $p \geq 2$ (see Section 1.3 of [47] for more

general costs). Lemma 3.1 ensures that ξ_p is Lipschitz continuous and λ-concave with some explicit constants. These regularity properties are transmitted, with the same constants, to any Kantorovich potential $\varphi : \Omega \to \mathbb{R}$ solution to (9). This follows from the fact that any such φ corresponds to the $\bar{c}$-transform of the function φ^c. In turn, the Lipschitz behavior of ξ_p and φ allows to ensure their differentiability almost-everywhere using Rademacher's theorem. Consider now an optimal transport plan γ minimizer of (8) and a Kantorovich potential φ maximizer of (9). The primal-dual relationship (10) ensures that for any $(x_0, y_0) \in \mathrm{spt}(\gamma)$, the function $x \mapsto \xi_p(x - y_0) - \varphi(x)$ is minimized in x_0. Thus, almost-every $(x_0, y_0) \in \mathrm{spt}(\gamma)$ satisfies the optimality condition $\nabla\xi_p(x_0 - y_0) - \nabla\varphi(x_0) = 0$, which leads to

$$y_0 = T(x_0) := x_0 - (\nabla\xi_p)^{-1}(\nabla\varphi(x_0)). \tag{11}$$

The mapping $T : \Omega \to \Omega$ is well defined almost-everywhere. These considerations show that if ρ is absolutely continuous with respect to the Lebesgue measure, γ is induced by the map T defined in (11), i.e. $\gamma = (\mathrm{id}, T)_{\#}\rho$. Because the choices of γ and φ were not made depending on each other, these ideas also show the uniqueness of γ and of $\nabla\varphi$. The map T is referred to as the *optimal transport map with respect to the ground cost* $c = \xi_p$ in the transport between ρ and μ.

3.2. **Stability estimate for the pushforward operation.** We now state our main result, that brings a tight answer to Problem 1.1:

Theorem 3.2. *Let $p \geq 2$ and consider the p-cost $c(x, y) = \xi_p(x - y) = \|x - y\|^p$. Let $\rho, \tilde{\rho} \in \mathcal{P}(\Omega)$ where $\Omega = B(0, R)$ with $R > 0$. Assume that ρ is absolutely continuous with density bounded from above by $M_\rho \in (0, +\infty)$. Let $\varphi \in \mathcal{C}(\Omega)$ satisfying $\varphi = (\varphi^c)^{\bar{c}}$. Let $\tilde{\gamma} \in \mathcal{P}(\Omega \times \Omega)$ be such that $(p_1)_{\#}\tilde{\gamma} = \tilde{\rho}$ and assume that $\mathrm{spt}(\tilde{\gamma}) \subset \partial^c\varphi$. Introduce the optimal transport map $T_\varphi : \Omega \to \Omega$ which satisfies for almost-every $x \in \Omega$,*

$$T_\varphi(x) = x - (\nabla\xi_p)^{-1}(\nabla\varphi(x)).$$

Then, for any $q \in (p - 1, \infty)$ and $r \in (1, \infty)$,

$$\mathrm{W}_q((T_\varphi)_{\#}\rho, (p_2)_{\#}\tilde{\gamma}) \leq c_{d,q,p,R,M_\rho}\mathrm{W}_r(\rho, \tilde{\rho})^{\frac{r}{q(r+1)}},$$

where $c_{d,q,p,R,M_\rho} = 2^{8(d+1)}p^3\left(\frac{q}{q-p+1}\right)^{1/q} d^2(1 + \beta_d)(1 + M_\rho)(1 + R)^{2+p+d}$, with β_d denoting the volume of the unit ball of $\mathbb{R}^d$. It also holds, with the same constant,

$$\mathrm{W}_q((T_\varphi)_{\#}\rho, (p_2)_{\#}\tilde{\gamma}) \leq c_{d,q,p,R,M_\rho}\mathrm{W}_\infty(\rho, \tilde{\rho})^{\frac{1}{q}}.$$

Remark 3.1 (Case of $\varphi \in \mathcal{C}^1$)**.** Whenever φ is differentiable $\tilde{\rho}$-almost-everywhere, Theorem 3.2 ensures for all $q > p - 1$ and $r > 1$ the following stability result for the pushforward operation by T_φ:

$$\mathrm{W}_q((T_\varphi)_{\#}\rho, (T_\varphi)_{\#}\tilde{\rho}) \leq c_{d,q,p,R,M_\rho}\mathrm{W}_r(\rho, \tilde{\rho})^{\frac{r}{q(r+1)}}.$$

Remark 3.2 (Case of $\varphi \in \mathcal{C}^{1,\alpha}$)**.** If the potential φ was regular in the previous proposition, e.g. $\varphi \in \mathcal{C}^{1,\alpha}(\Omega)$, one would trivially get an estimate of the form

$$\mathrm{W}_q((T_\varphi)_{\#}\rho, (T_\varphi)_{\#}\tilde{\rho}) \leq C\mathrm{W}_r(\rho, \tilde{\rho})^{\frac{\alpha}{p-1}},$$

relying on Lemma 3.1 and for any $q, r \geq 1$ that are such that $r \geq \frac{\alpha q}{p-1}$. However, as noticed in the introduction, even for $p = 2$, getting regularity estimates for optimal transport potentials requires to make strong regularity assumptions on the involved measures which are rarely satisfied in applications. When $p \neq 2$, the situation is

even worse since the cost fails to satisfy the so-called Ma-Trudinger-Wang condition which, as shown in Theorem 3.1 in [33], is in fact necessary for the C^1 regularity of optimal potentials .

Remark 3.3 (Tightness of exponents)**.** The estimate of Theorem 3.2 is tight in terms of exponents. This follows from the following generalization of Example 1.4 (Figure 2). In dimension $d = 1$, consider on $\Omega = [-1, 1]$ the probability measures $\rho = \lambda_{[-\frac{1}{2},\frac{1}{2}]}$ and $\rho^\varepsilon = \lambda_{[-\frac{1}{2},-\frac{\varepsilon}{2}]\cup[\frac{\varepsilon}{2},\frac{1}{2}]} + \varepsilon\delta_0$ where $\varepsilon \in (0, \frac{1}{2})$ and λ_I denotes the Lebesgue measure restricted to a set I. For a given $p \geq 2$, define on Ω the potential $\varphi : x \mapsto (1 - |x|)^p$. This potential satisfies $\varphi = (\varphi^c)^{\bar{c}}$ where c is the p-cost. Introduce T_φ the associated optimal transport map, which satisfies $T_\varphi(x) = \mathrm{sign}(x)$, and $\gamma^\varepsilon \in \mathcal{P}(\Omega \times \Omega)$ defined with $\gamma^\varepsilon = \int_{[-\frac{1}{2},-\frac{\varepsilon}{2}]\cup[\frac{\varepsilon}{2},\frac{1}{2}]} \delta_x \otimes \delta_{T_\varphi(x)} \mathrm{d}x + \varepsilon\delta_{(0,1)}$. Then $(p_1)_\# \gamma^\varepsilon = \rho^\varepsilon$, $\mathrm{spt}(\gamma^\varepsilon) \subset \partial^c\varphi$. One then has $(T_\varphi)_\#\rho = \frac{1}{2}(\delta_{-1} + \delta_{+1})$ and $(p_2)_\#\gamma^\varepsilon = \frac{1-\varepsilon}{2}\delta_{-1} + \frac{1+\varepsilon}{2}\delta_{+1}$. Thus for any $q \geq 1$, $\mathrm{W}_q((T_\varphi)_\#\rho, (p_2)_\#\gamma^\varepsilon) \sim \varepsilon^{1/q}$, while for any $r \geq 1$ one easily has $\mathrm{W}_r(\rho, \rho^\varepsilon) \sim \varepsilon^{\frac{r+1}{r}}$, that is

$$\mathrm{W}_q((T_\varphi)_\#\rho, (p_2)_\#\gamma^\varepsilon) \sim \mathrm{W}_r(\rho, \rho^\varepsilon)^{\frac{r}{q(r+1)}}.$$

Remark 3.4 (Comparison with stochastic approximations)**.** The tight estimates of Theorem 3.2 tend to indicate that, in dimension $d \geq 2$, the stochastic approximations of the measure $(T_\varphi)_\#\rho$ from this theorem converge more rapidly than the deterministic approximations built from ρ. Indeed, given a budget of $N \geq 1$ points, one can build an approximation $\tilde{\rho}_N \in \mathcal{P}(\Omega)$ of $\rho \in \mathcal{P}(\Omega)$ supported on a grid of N points and that satisfies

$$\mathrm{W}_\infty(\rho, \tilde{\rho}_N) \lesssim N^{-1/d}.$$

The bound in Theorem 3.2 then ensures formally for any $q \geq p - 1$:

$$\mathrm{W}_q^q((T_\varphi)_\#\rho, (T_\varphi)_\#\tilde{\rho}_N) \lesssim N^{-1/d}.$$

Meanwhile, if one samples N points $(x_i)_{1\leq i\leq N}$ from ρ and denotes $\hat{\rho}_N = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ the corresponding empirical measure, Theorem 1 of [23] ensures:

$$\mathbb{E}\mathrm{W}_q^q((T_\varphi)_\#\rho, (T_\varphi)_\#\hat{\rho}_N) \lesssim \begin{cases} N^{-1/2} & \text{if } d < 2q, \\ N^{-1/2}\log(1+N) & \text{if } d = 2q, \\ N^{-q/d} & \text{else.} \end{cases}$$

In particular, except in dimension one, the stochastic approximation $(T_\varphi)_\#\hat{\rho}_N$ converges faster (in expectation) towards $(T_\varphi)_\#\rho$ than its deterministic counterpart $(T_\varphi)_\#\tilde{\rho}_N$.

The proof of Theorem 3.2 relies on the following lemma, that is a direct consequence of Lemma 3.1 and whose proof is deferred after the proof of Theorem 3.2.

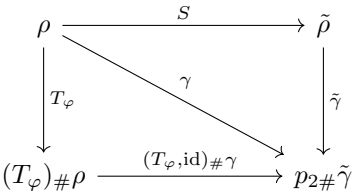
Lemma 3.3. *With the notations of Theorem 3.2, the function* $\phi : x \mapsto p(p-1)R^{p-2}\frac{\|x\|^2}{2} - \varphi(x)$ *is convex and* p^2R^{p-1}*-Lipschitz continuous on* Ω*. This function can be extended to a convex and* p^2R^{p-1}*-Lipschitz continuous function defined on* $\mathbb{R}^d$*. Moreover, for any* $x \in \Omega$ *and* $\eta > 0$*,*

$$\mathrm{diam}(\partial^c\varphi(B(x,\eta) \cap \Omega)) \leq 8p(1 + R^{\frac{p-2}{p-1}})\left(\eta^{\frac{1}{p-1}} + \mathrm{diam}(\partial\phi(B(x,\eta)))^{\frac{1}{p-1}}\right).$$

We are now ready to prove Theorem 3.2.

Proof of Theorem 3.2. In this proof, we omit for clarity the multiplicative constants that depend on d, q, p, R or M_ρ and use $\lesssim$ instead of $\leq$ for inequalities involving such constants. A close look at this proof allows to recover the multiplicative constant of the statement. Let us assume for now that $r \in (1, \infty) \cup \{\infty\}$. We will deal with each of the distinct cases $r = \infty$ and $r < \infty$ afterwards.

We first disintegrate $\tilde{\gamma}$ with respect to $\tilde{\rho}$, i.e. we let $\tilde{\gamma} = \int \delta_x \otimes \tilde{\gamma}_x \mathrm{d}\tilde{\rho}$, where $x \mapsto \tilde{\gamma}_x$ is a measurable map from Ω to $\mathcal{P}(\Omega)$. By assumption, the support of $\tilde{\gamma}$ is included in $\partial^c\varphi$. This implies that for any x in Ω, the support of $\tilde{\gamma}_x$ is included in $\partial^c\varphi(x) = \{y \in \Omega \mid \varphi(x) + \varphi^c(y) = c(x,y)\}$. We introduce $S : \mathbb{R}^d \to \mathbb{R}^d$ an optimal transport map from ρ to $\tilde{\rho}$ for the r-cost[1] and we consider the measure $\gamma = \int \delta_x \otimes \tilde{\gamma}_{S(x)} \mathrm{d}\rho(x)$. This measure γ is a coupling between ρ and $(p_2)_\# \tilde{\gamma}$, which implies that $(T_\varphi, \mathrm{id})_\# \gamma$ is a coupling between $(T_\varphi)_\# \rho$ and $(p_2)_\# \tilde{\gamma}$. These constructions may be summarized by the following diagram.



We therefore have the bound:

$$\begin{aligned} \mathrm{W}_q^q((T_\varphi)_\#\rho, (p_2)_\#\tilde{\gamma}) &\leq \int_{\Omega\times\Omega} \|T_\varphi(x) - y\|^q \,\mathrm{d}\gamma(x,y) \\ &= \int_{\Omega\times\Omega} \|T_\varphi(x) - y\|^q \,\mathrm{d}\tilde{\gamma}_{S(x)}(y)\mathrm{d}\rho(x) \\ &= \int_{x\in\Omega} \int_{y\in\partial^c\varphi(S(x))} \|T_\varphi(x) - y\|^q \,\mathrm{d}\tilde{\gamma}_{S(x)}(y)\mathrm{d}\rho(x), \end{aligned} \tag{12}$$

where we used that $\mathrm{spt}(\tilde{\gamma}_{S(x)}) \subseteq \partial^c\varphi(S(x))$ to get the last line. For a given $\eta \in (0, 2R+1]$, we will upper bound the right-hand side by splitting the integral on Ω_η and Ω_η^c, where

$$\Omega_\eta = \{x \in \mathrm{spt}(\rho) |\, \|S(x) - x\| \leq \eta\}, \qquad \Omega_\eta^c = \mathrm{spt}(\rho) \setminus \Omega_\eta.$$

Upper bound on Ω_η**.** By definition, any point x in Ω_η satisfies $\|S(x) - x\| \leq \eta$, so that $S(x)$ belongs to the ball $B(x,\eta)$ intersected with Ω. Then for any such x, $\partial^c\varphi(S(x)) \subset \partial^c\varphi(B(x,\eta) \cap \Omega)$. Therefore for any $g \in \partial^c\varphi(x)$ and $y \in \partial^c\varphi(S(x))$, one has

$$\|g - y\| \leq \mathrm{diam}\left(\partial^c\varphi(B(x,\eta) \cap \Omega)\right),$$

so that, recalling that $T_\varphi(x) \in \partial^c\varphi(x)$, the quantity

$$\int_{x\in\Omega_\eta} \int_{y\in\partial^c\varphi(S(x))} \|T_\varphi(x) - y\|^q \,\mathrm{d}\tilde{\gamma}_{S(x)}(y)\mathrm{d}\rho(x)$$

is dominated by

$$\int_{x\in\Omega_\eta} \mathrm{diam}\left(\partial^c\varphi(B(x,\eta) \cap \Omega)\right)^q \mathrm{d}\rho(x).$$

[1]For $r = \infty$, the existence of an optimal transport map was first established in [16].

Let ϕ be the convex and p^2R^{p-1}-Lipschitz function on Ω defined from φ in Lemma 3.3. This lemma ensures that:

$$\operatorname{diam}(\partial^c\varphi(B(x,\eta)\cap\Omega)) \lesssim \eta^{\frac{1}{p-1}} + \operatorname{diam}(\partial\phi(B(x,\eta)))^{\frac{1}{p-1}}.$$

We thus have the estimate:

$$\int_{x\in\Omega_\eta} \operatorname{diam}\left(\partial^c\varphi(B(x,\eta)\cap\Omega)\right)^q \mathrm{d}\rho(x) \lesssim \eta^{\frac{q}{p-1}} + \int_\Omega \operatorname{diam}(\partial\phi(B(x,\eta)))^{\frac{q}{p-1}}\mathrm{d}\rho(x).$$

Using that $\frac{q}{p-1} > 1$ and that $\eta \le 2R+1$, one has

$$\eta^{\frac{q}{p-1}} = (2R+1)^{\frac{q}{p-1}}\left(\frac{\eta}{2R+1}\right)^{\frac{q}{p-1}} \lesssim \eta.$$

Using again that $\frac{q}{p-1} > 1$, Corollary 2.2 ensures the bound:

$$\int_\Omega \operatorname{diam}(\partial\phi(B(x,\eta)))^{\frac{q}{p-1}}\mathrm{d}\rho(x) \lesssim \eta,$$

The last two bounds thus entail

$$\int_{x\in\Omega_\eta} \operatorname{diam}\left(\partial^c\varphi(B(x,\eta)\cap\Omega)\right)^q \mathrm{d}\rho(x) \lesssim \eta.$$

We therefore have the bound:

$$\int_{x\in\Omega_\eta}\int_{y\in\partial^c\varphi(S(x))} \|T_\varphi(x)-y\|^q \,\mathrm{d}\tilde\gamma_{S(x)}(y)\mathrm{d}\rho(x) \lesssim \eta. \tag{13}$$

This last bound allows to deal with the case $r=\infty$. Indeed, assuming that $r=\infty$, we get by setting $\eta = \mathrm{W}_\infty(\rho,\tilde\rho)$ that $\Omega_\eta = \Omega$, $\Omega_\eta^c = \emptyset$, and the previous inequality combined with (12) allows to reach the conclusion that

$$\mathrm{W}_q^q((T_\varphi)_\#\rho,(p_2)_\#\tilde\gamma) \lesssim \mathrm{W}_\infty(\rho,\tilde\rho).$$

We now assume that $r\in(1,+\infty)$. There remains to bound the value of the integrand in (12) on the domain Ω_η^c.

Upper bound on Ω_η^c**.** The optimal transport map S from ρ to $\tilde\rho$ satisfies

$$\|S-\mathrm{id}\|_{\mathrm{L}^r(\rho)} = \mathrm{W}_r(\rho,\tilde\rho).$$

Then using Markov's inequality, $\eta^r\rho(\Omega_\eta^c) \le \mathrm{W}_r^r(\rho,\tilde\rho)$. The fact that T_φ is valued in Ω then implies

$$\begin{aligned}\int_{x\in\Omega_\eta^c}\int_{y\in\partial^c\varphi(S(x))} \|T_\varphi(x)-y\|^q \,\mathrm{d}\tilde\gamma_{S(x)}(y)\mathrm{d}\rho(x) &\le \int_{x\in\Omega_\eta^c}(2R)^q\mathrm{d}\rho(x)\\ &\lesssim \frac{\mathrm{W}_r^r(\rho,\tilde\rho)}{\eta^r}.\end{aligned} \tag{14}$$

Conclusion. Using bounds (13) and (14) in (12) we have for any $\eta\in(0,2R]$:

$$\mathrm{W}_q^q((T_\varphi)_\#\rho,(p_2)_\#\tilde\gamma) \lesssim \frac{\mathrm{W}_r^r(\rho,\tilde\rho)}{\eta^r} + \eta.$$

Setting $\eta = \mathrm{W}_r(\rho,\tilde\rho)^{\frac{r}{r+1}}$ then allows us to conclude:

$$\mathrm{W}_q^q((T_\varphi)_\#\rho,(p_2)_\#\tilde\gamma) \lesssim \mathrm{W}_r(\rho,\tilde\rho)^{\frac{r}{r+1}}. \qquad \square$$

We conclude this section with the proof of Lemma 3.3.

Proof of Lemma 3.3. From Lemma 3.1, we know that the p-cost ξ_p is a $-C_{p,R}$-concave function with $C_{p,R} = p(p-1)R^{p-2}$. Since φ verifies $\varphi = (\varphi^c)^{\bar{c}}$, it is also a $-C_{p,R}$-concave function as an infimum of $-C_{p,R}$-concave functions. In particular, the function ϕ is a convex function. Similarly, Lemma 3.1 ensures that ξ_p is pR^{p-1}-Lipschitz continuous on Ω and so is $\varphi = (\varphi^c)^{\bar{c}}$. An immediate computation then ensures that ϕ is p^2R^{p-1}-Lipschitz continuous on Ω. The p^2R^{p-1}-Lipschitz continuous convex function ϕ defined on Ω can be extended, by mean of double convex conjugate, as a p^2R^{p-1}-Lipschitz continuous convex function defined on the whole $\mathbb{R}^d$ and coinciding with ϕ on Ω:

$$\forall x \in \mathbb{R}^d, \quad \phi(x) := \sup_{x_\Omega \in \Omega, g \in \partial\phi(x_\Omega)} \phi(x_\Omega) + \langle g | x - x_\Omega \rangle.$$

Now consider $x \in \Omega$ and $\eta > 0$. Let $x^-, x^+ \in B(x,\eta) \cap \Omega$. Let $y^- \in \partial^c\varphi(x^-)$ and $y^+ \in \partial^c\varphi(x^+)$. We want to bound $\|y^+ - y^-\|$ in terms of η and $\operatorname{diam}(\partial\phi(B(x,\eta)))$. Let's refer for now to x^-, y^- and x^+, y^- indistinctly with $x^\pm, y^\pm$. Recall that

$$\partial^c\varphi(x^\pm) = \{y \in \Omega \mid \varphi(x^\pm) + \varphi^c(y) = \xi_p(x^\pm - y)\}.$$

Therefore, $y^\pm \in \partial^c\varphi(x^\pm)$ if and only if $z \mapsto \xi_p(z - y^\pm) - \varphi(z)$ is minimized in $x^\pm$, that is if and only if

$$z \mapsto \xi_p(z - y^\pm) - C_{p,R}\frac{\|z\|^2}{2} + \phi(z)$$

is minimized in $x^\pm$. This is possible only if there exists $g^\pm \in \partial\phi(x^\pm)$ such that

$$0 = \nabla\xi_p(x^\pm - y^\pm) - C_{p,R}x^\pm + g^\pm.$$

Hence there exists $g^\pm \in \partial\phi(x^\pm)$ such that

$$y^\pm = x^\pm - (\nabla\xi_p)^{-1}(C_{p,R}x^\pm - g^\pm).$$

Considering such subgradients $g^\pm \in \partial\phi(x^\pm)$, we thus have:

$$\left\|y^+ - y^-\right\| \leq \left\|x^+ - x^-\right\| + \left\|(\nabla\xi_p)^{-1}(C_{p,R}x^+ - g^+) - (\nabla\xi_p)^{-1}(C_{p,R}x^- - g^-)\right\|.$$

The Hölder behavior of $(\nabla\xi_p)^{-1}$ described in Lemma 3.1 then allows us to write:

$$\left\|y^+ - y^-\right\| \leq \left\|x^+ - x^-\right\| + \frac{3}{p^{\frac{1}{p-1}}}\left\|C_{p,R}(x^+ - x^-) - g^+ + g^-\right\|^{\frac{1}{p-1}}.$$

Therefore, using that $x^\pm \in B(x,\eta)$, we have that $\|g^+ - g^-\| \leq \operatorname{diam}(\partial\phi(B(x,\eta)))$ so that we have the bound

$$\left\|y^+ - y^-\right\| \leq 2\eta + \frac{3}{p^{\frac{1}{p-1}}}(C_{p,R}\eta)^{\frac{1}{p-1}} + \frac{3}{p^{\frac{1}{p-1}}}\operatorname{diam}(\partial\phi(B(x,\eta)))^{\frac{1}{p-1}}.$$

Maximizing over y^- and y^+ and using that $\eta \leq R$ leads to the bound:

$$\operatorname{diam}(\partial^c\varphi(B(x,\eta)\cap\Omega)) \leq 8p(1 + R^{\frac{p-2}{p-1}})\left(\eta^{\frac{1}{p-1}} + \operatorname{diam}(\partial\phi(B(x,\eta)))^{\frac{1}{p-1}}\right). \quad \square$$

Acknowledgement. The authors acknowledge the support of the Lagrange Mathematics and Computing Research Center and of the ANR (MAGA, ANR-16-CE40-0014).

References

[1] Martial Agueh and Guillaume Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[2] Giovanni Alberti, Luigi Ambrosio, and Piermarco Cannarsa. On the singularities of convex functions. *manuscripta mathematica*, 76(1):421–435, Dec 1992.

[3] Luigi Ambrosio, Piermarco Cannarsa, and Halil Mete Soner. On the propagation of singularities of semi-convex functions. *Annali della Scuola Normale Superiore di Pisa - Classe di Scienze*, 20(4):597–616, 1993.

[4] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[5] Brandon Amos, Lei Xu, and J. Zico Kolter. Input convex neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 146–155. PMLR, 06–11 Aug 2017.

[6] Saurav Basu, Soheil Kolouri, and Gustavo K. Rohde. Detecting and visualizing cell phenotype differences from microscopy images using transport-based morphometry. *Proceedings of the National Academy of Sciences*, 111(9):3448–3453, 2014.

[7] Yann Brenier. The least action principle and the related concept of generalized flows for incompressible perfect fluids. *Journal of the American Mathematical Society*, 2(2):225–255, 1989.

[8] Charlotte Bunne, Andreas Krause, and Marco Cuturi. Supervised training of conditional monge maps. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 6859–6872. Curran Associates, Inc., 2022.

[9] Luis A. Caffarelli. The regularity of mappings with a convex potential. *Journal of the American Mathematical Society*, 5(1):99–104, 1992.

[10] Luis A. Caffarelli. Boundary regularity of maps with convex potentials–ii. *Annals of Mathematics*, 144(3):453–496, 1996.

[11] Tianji Cai, Junyi Cheng, Nathaniel Craig, and Katy Craig. Linearized optimal transport for collider events. *Phys. Rev. D*, 102:116019, Dec 2020.

[12] Piermarco Cannarsa and Wei Cheng. Singularities of solutions of hamilton–jacobi equations. *Milan Journal of Mathematics*, 89(1):187–215, Jun 2021.

[13] Piermarco Cannarsa and Halil Mete Soner. On the singularities of the viscosity solutions to hamilton–jacobi–bellman equations. *Indiana University Mathematics Journal*, 36(3):501–524, 1987.

[14] Piermarco Cannarsa and Halil Mete Soner. Generalized one-sided estimates for solutions of hamilton-jacobi equations and applications. *Nonlinear Analysis: Theory, Methods & Applications*, 13(3):305–323, 1989.

[15] Guillaume Carlier, Katharina Eichinger, and Alexey Kroshnin. Entropic-wasserstein barycenters: Pde characterization, regularity, and clt. *SIAM Journal on Mathematical Analysis*, 53(5):5880–5914, 2021.

[16] Thierry Champion, Luigi De Pascale, and Petri Juutinen. The L^∞-Wasserstein Distance: Local Solutions and Existence of Optimal Transport Maps. *SIAM Journal on Mathematical Analysis*, 40(1):1–20, 2008.

[17] Yize Chen, Yuanyuan Shi, and Baosen Zhang. Optimal control via neural networks: A convex approach. *ArXiv, 1805.11835*, 2018.

[18] Pierre Colombo, Guillaume Staerman, Pablo Piantanida, and Chloé Clavel. Automatic Text Evaluation through the Lens of Wasserstein Barycenters. In *EMNLP 2021*, Punta Cana, Dominica, November 2021.

[19] Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32(2) of *Proceedings of Machine Learning Research*, pages 685–693, Bejing, China, 22–24 Jun 2014. PMLR.

[20] Fernando de Goes, Corentin Wallez, Jin Huang, Dmitry Pavlov, and Mathieu Desbrun. Power particles: An incompressible fluid solver based on power diagrams. *ACM Trans. Graph.*, 34(4), jul 2015.

[21] Alex Delalande and Quentin Mérigot. Quantitative Stability of Optimal Transport Maps under Variations of the Target Measure. *Duke Mathematical Journal (to appear)*, 2021.
[22] Pierre Dognin, Igor Melnyk, Youssef Mroueh, Jarret Ross, Cicero Dos Santos, and Tom Sercu. Wasserstein barycenter model ensembling. In *International Conference on Learning Representations*, 2019.
[23] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, Aug 2015.
[24] Thomas O Gallouët and Quentin Mérigot. A Lagrangian Scheme à la Brenier for the Incompressible Euler Equations. *Foundations of Computational Mathematics*, 18:835–865, 2018.
[25] Nhat Ho, XuanLong Nguyen, Mikhail Yurochkin, Hung Hai Bui, Viet Huynh, and Dinh Phung. Multilevel clustering via Wasserstein means. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1501–1509. PMLR, 06–11 Aug 2017.
[26] R. Jensen and P. E. Souganidis. A regularity result for viscosity solutions of hamilton-jacobi equations in one space dimensions. *Transactions of the American Mathematical Society*, 301(1):137–147, 1987.
[27] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.
[28] Jun Kitagawa, Quentin Mérigot, and Boris Thibert. Convergence of a Newton algorithm for semi-discrete optimal transport. *J. Eur. Math. Soc.*, 21(9):2603–2651, 2019.
[29] Soheil Kolouri and Gustavo K. Rohde. Transport-based single frame super resolution of very low resolution face images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4876–4884, 2015.
[30] Soheil Kolouri, Akif B. Tosun, John A. Ozolek, and Gustavo K. Rohde. A continuous linear optimal transport approach for pattern analysis in image datasets. *Pattern Recognition*, 51:453–462, 2016.
[31] Alexander Korotin, Vage Egiazarian, Arip Asadulaev, Alexander Safin, and Evgeny Burnaev. Wasserstein-2 generative networks. In *International Conference on Learning Representations*, 2021.
[32] Xin Lian, Kshitij Jain, Jakub Truszkowski, Pascal Poupart, and Yaoliang Yu. Unsupervised multilingual alignment using wasserstein barycenter. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3702–3708. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track.
[33] Grégoire Loeper. On the regularity of solutions of optimal transportation problems. *Acta Math.*, 202(2):241–283, 2009.
[34] Lévy, Bruno. A numerical algorithm for l2 semi-discrete optimal transport in 3d. *ESAIM: M2AN*, 49(6):1693–1715, 2015.
[35] Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6672–6681. PMLR, 13–18 Jul 2020.
[36] Quentin Mérigot, Alex Delalande, and Frederic Chazal. Quantitative stability of optimal transport maps and linearization of the 2-Wasserstein space. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108, pages 3186–3196, 26–28 Aug 2020.
[37] Jocelyn Meyron, Quentin Mérigot, and Boris Thibert. Light in power: A general and parameter-free algorithm for caustic design. *ACM Trans. Graph.*, 37(6), dec 2018.
[38] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris, avec les Mémoires de Mathématique et de Physique pour la même année*, pages 666–704, 1781.
[39] Quentin Mérigot and Boris Thibert. Chapter 2 - Optimal transport: discretization and algorithms. In Andrea Bonito and Ricardo H. Nochetto, editors, *Geometric Partial Differential Equations - Part II*, volume 22 of *Handbook of Numerical Analysis*, pages 133–212. Elsevier, 2021.
[40] Shizuo Nakane. Formation of singularities for Hamilton-Jacobi equation with several space variables. *Journal of the Mathematical Society of Japan*, 43(1):89 – 100, 1991.

[41] Farnik Nikakhtar, Ravi K. Sheth, Bruno Lévy, and Roya Mohayaee. Optimal transport reconstruction of baryon acoustic oscillations. *Phys. Rev. Lett.*, 129:251101, Dec 2022.
[42] Felix Otto. Dynamics of Labyrinthine Pattern Formation in Magnetic Fluids: A Mean-Field Theory. *Archive for Rational Mechanics and Analysis*, 141(1):63–103, Mar 1998.
[43] Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. *Communications in Partial Differential Equations*, 26:101–174, 2001.
[44] S. Park and M. Thorpe. Representing and learning high dimensional data with the optimal transport map from a probabilistic viewpoint. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7864–7872, 2018.
[45] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.
[46] Julien Rabin, Gabriel Peyré, Julie Delon, and Bernot Marc. Wasserstein Barycenter and its Application to Texture Mixing. In *SSVM'11*, pages 435–446, Israel, 2011. Springer.
[47] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55:58–63, 2015.
[48] Justin Solomon, Fernando de Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Trans. Graph.*, 34(4), jul 2015.
[49] Sanvesh Srivastava, Cheng Li, and David B. Dunson. Scalable bayes via barycenter in wasserstein space. *Journal of Machine Learning Research*, 19(8):1–35, 2018.
[50] Amirhossein Taghvaei and Amin Jalali. 2-wasserstein approximation via restricted convex potentials with application to improved training for gans, 2019.
[51] Mikio Tsuji. Formation of singularities for Hamilton-Jacobi equation, I. *Proceedings of the Japan Academy, Series A, Mathematical Sciences*, 59(2):55 – 58, 1983.
[52] Mikio Tsuji. Formation of singularities for Hamilton-Jacobi equation II. *Journal of Mathematics of Kyoto University*, 26(2):299 – 308, 1986.
[53] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.
[54] Wei Wang, Dejan Slepčev, Saurav Basu, John A. Ozolek, and Gustavo K. Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *Int. J. Comput. Vision*, 101(2):254–269, January 2013.

Appendix A. Omitted proofs

A.1. **Proof of Lemma 3.1.**

Proof of Lemma 3.1. The strict convexity of ξ_p results from the triangle inequality and the strict convexity of $u \mapsto u^p$ on $\mathbb{R}_+^*$ for $p \geq 2$.

Denoting z_i the i-th coordinate of $z \in \Omega$ in the canonical basis of $\mathbb{R}^d$, one has $\xi_p(z) = (\sum_{i=1}^d z_i^2)^{p/2}$. From this expression we deduce immediately that ξ_p is of class $\mathcal{C}^2$ and that its gradient and hessian read respectively

$$\nabla \xi_p(z) = pz \left\|z\right\|^{p-2} \quad \text{and} \quad \nabla^2 \xi_p(z) = p \left\|z\right\|^{p-2} \mathrm{id} + p(p-2) \left\|z\right\|^{p-4} zz^\top.$$

Thus for all $z \in \Omega$, $\|\nabla \xi_p(z)\| \leq pR^{p-1}$ and ξ_p is pR^{p-1}-Lipschitz continuous.

For any $v \in \mathbb{R}^d$, one has

$$v^\top \nabla^2 \xi_p(z) v = p \left\|z\right\|^{p-2} \left\|v\right\|^2 + p(p-2) \left\|z\right\|^{p-4} \langle v|z\rangle^2.$$

Cauchy-Schwartz inequality entails $\langle v|z\rangle^2 \leq \left\|v\right\|^2 \left\|z\right\|^2$, so that

$$v^\top \nabla^2 \xi_p(z) v \leq p(p-1) \left\|z\right\|^{p-2} \left\|v\right\|^2.$$

For any $z \in \Omega$ we thus have the bound

$$0 \preceq \nabla^2 \xi_p(z) \preceq p(p-1) R^{p-2},$$

from which we deduce that $z \mapsto \xi_p(z) - p(p-1)\frac{R^{p-2}}{2} \left\|z\right\|^2$ is a concave function.

Finally, the mapping $z \mapsto \nabla\xi_p(z) = pz\,\|z\|^{p-2}$ is obviously bijective on $\mathbb{R}^d$. For any $y, z \in \mathbb{R}^d \setminus \{0\}$ such that $\nabla\xi_p(y) = z$, one has

$$z = py\,\|y\|^{p-2},$$

so that $\|z\| = p\,\|y\|^{p-1}$. From this fact one deduces

$$y = (\nabla\xi_p)^{-1}(z) = \frac{1}{p^{1/(p-1)}}\frac{z}{\|z\|^{\frac{p-2}{p-1}}} = \frac{1}{p^{\beta(p)}\,\|z\|^{1-\beta(p)}}z,$$

where $\beta(p) = \frac{1}{p-1} \in (0,1]$ for $p \geq 2$.

Let's finally show the Hölder behavior of $(\nabla\xi_p)^{-1}$. Let $x, y \in \mathbb{R}^d$. If $x = 0$ and $y \neq 0$, then

$$\left\|(\nabla\xi_p)^{-1}(y) - (\nabla\xi_p)^{-1}(x)\right\| = \left\|(\nabla\xi_p)^{-1}(y)\right\| = \frac{1}{p^{\beta(p)}}\,\|y\|^{\beta(p)},$$

which corresponds to a $\beta(p)$-Hölder behavior of ξ_p near 0. Assume now that $x \neq 0$ and $y \neq 0$. Assume for now that x and y are positively linearly dependent, i.e. there exists $\lambda \geq 1$ such that $y = \lambda x$. Then:

$$\left\|(\nabla\xi_p)^{-1}(y) - (\nabla\xi_p)^{-1}(x)\right\| = \frac{1}{p^{\beta(p)}}(\lambda^{\beta(p)} - 1)\,\|x\|^{\beta(p)}.$$

Using that for any $u > 0$ and $\beta \in (0,1]$, $(1+u)^\beta - 1 \leq u^\beta$, we have $\lambda^{\beta(p)} - 1 \leq (\lambda - 1)^{\beta(p)}$. Hence we deduce:

$$\left\|(\nabla\xi_p)^{-1}(y) - (\nabla\xi_p)^{-1}(x)\right\| \leq \frac{1}{p^{\beta(p)}}(\lambda-1)^{\beta(p)}\,\|x\|^{\beta(p)} = \frac{1}{p^{\beta(p)}}\,\|y-x\|^{\beta(p)}. \quad (15)$$

Assume now that $\|x\| = \|y\|$. Then:

$$\begin{aligned}
\left\|(\nabla\xi_p)^{-1}(y) - (\nabla\xi_p)^{-1}(x)\right\| &= \frac{1}{p^{\beta(p)}\,\|x\|^{1-\beta(p)}}\,\|y-x\| \\
&= \frac{\|y-x\|^{1-\beta(p)}}{p^{\beta(p)}\,\|x\|^{1-\beta(p)}}\,\|y-x\|^{\beta(p)} \\
&\leq \frac{(\|x\|+\|y\|)^{1-\beta(p)}}{p^{\beta(p)}\,\|x\|^{1-\beta(p)}}\,\|y-x\|^{\beta(p)} \\
&= \frac{2^{1-\beta(p)}}{p^{\beta(p)}}\,\|y-x\|^{\beta(p)}. \qquad (16)
\end{aligned}$$

Finally, without making any assumption on x and y, we have:

$$\begin{aligned}
\left\|(\nabla\xi_p)^{-1}(y) - (\nabla\xi_p)^{-1}(x)\right\| \leq{}& \left\|(\nabla\xi_p)^{-1}(y) - (\nabla\xi_p)^{-1}(\frac{\|x\|}{\|y\|}y)\right\| \\
&+ \left\|(\nabla\xi_p)^{-1}(\frac{\|x\|}{\|y\|}y) - (\nabla\xi_p)^{-1}(x)\right\|.
\end{aligned}$$

Bound (15) ensures:

$$\begin{aligned}\left\|(\nabla\xi_p)^{-1}(y)-(\nabla\xi_p)^{-1}(\frac{\|x\|}{\|y\|}y)\right\| &\leq \frac{1}{p^{\beta(p)}}\left\|y-\frac{\|x\|}{\|y\|}y\right\|^{\beta(p)}\\ &= \frac{1}{p^{\beta(p)}}\left|\|y\|-\|x\|\right|^{\beta(p)}\\ &\leq \frac{1}{p^{\beta(p)}}\|y-x\|^{\beta(p)}.\end{aligned}$$

On the other hand, bound (16) ensures:

$$\begin{aligned}\left\|(\nabla\xi_p)^{-1}(\frac{\|x\|}{\|y\|}y)-(\nabla\xi_p)^{-1}(x)\right\| &\leq \frac{2^{1-\beta(p)}}{p^{\beta(p)}}\left\|\frac{\|x\|}{\|y\|}y-x\right\|^{\beta(p)}\\ &\leq \frac{2^{1-\beta(p)}}{p^{\beta(p)}}\left(\left\|\frac{\|x\|}{\|y\|}y-y\right\|+\|y-x\|\right)^{\beta(p)}\\ &= \frac{2^{1-\beta(p)}}{p^{\beta(p)}}\left(\left|\|x\|-\|y\|\right|+\|y-x\|\right)^{\beta(p)}\\ &\leq \frac{2}{p^{\beta(p)}}\|y-x\|^{\beta(p)}.\end{aligned}$$

We thus get eventually:

$$\left\|(\nabla\xi_p)^{-1}(y)-(\nabla\xi_p)^{-1}(x)\right\| \leq \frac{3}{p^{\beta(p)}}\|y-x\|^{\beta(p)}. \qquad \square$$

Ceremade, Univ. Paris-Dauphine PSL, 75775 Paris and Mokaplan, Inria Paris
Email address: carlier@ceremade.dauphine.fr

Lagrange Mathematics and Computing Research Center, 75007, Paris, France
Email address: delalande.alex@gmail.com

Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France and Institut universitaire de France
Email address: quentin.merigot@universite-paris-saclay.fr