Econometric Theory, **38**, 2022, 381–417.
doi:10.1017/S0266466621000037

ON THE CONVERGENCE RATE OF POTENTIALS OF BRENIER MAPS*

FLORIAN F. GUNSILIUS
University of Michigan

The theory of optimal transportation has experienced a sharp increase in interest in many areas of economic research such as optimal matching theory and econometric identification. A particularly valuable tool, due to its convenient representation as the gradient of a convex function, has been the Brenier map: the matching obtained as the optimizer of the Monge–Kantorovich optimal transportation problem with the euclidean distance as the cost function. Despite its popularity, the statistical properties of the Brenier map have yet to be fully established, which impedes its practical use for estimation and inference. This article takes a first step in this direction by deriving a convergence rate for the simple plug-in estimator of the potential of the Brenier map via the semi-dual Monge–Kantorovich problem. Relying on classical results for the convergence of smoothed empirical processes, it is shown that this plug-in estimator converges in standard deviation to its population counterpart under the minimax rate of convergence of kernel density estimators if one of the probability measures satisfies the Poincaré inequality. Under a normalization of the potential, the result extends to convergence in the L^2 norm, while the Poincaré inequality is automatically satisfied. The main mathematical contribution of this article is an analysis of the second variation of the semi-dual Monge–Kantorovich problem, which is of independent interest.

1. INTRODUCTION

Optimal transport theory has been an active area of research in applied mathematics, machine learning, statistics, and economics. Its applications include optimal matching and Hedonic models (e.g., Chiappori, McCann, and Nesheim (2010) Chiappori et al. (2010); Chernozhukov et al. (2017) Chernozhukov et al. (2019); Lindenlaub (2017) Lindenlaub (2017)), partial identification of economic models (e.g., Ekeland, Galichon, and Henry (2010) Ekeland et al. (2010)), model-free hedging (e.g., Henry-Labordère (2017) Henry-Labordère (2017)), the definition of multivariate quantiles and quantile regression (Carlier, Chernozhukov, and Galichon, 2016; Chernozhukov et al., 2017), and nonlinear principal component analysis (Gunsilius and Schennach, 2019). For a recent overview of applications

*I would like to thank the Editor Peter Phillips, the Co-Editor Ivan Fernandez-Val, and especially two referees for their effort and important comments which helped me improve the paper substantially. I would also like to thank Susanne Schennach for helpful feedback. I wrote the initial version of this paper while a graduate student in the Economics Department at Brown University. All errors are mine. Address correspondence to Florian F. Gunsilius, University of Michigan, Ann Arbor, MI 48109, USA; e-mail: ffg@umich.edu .

in the econometrics literature, see Galichon (2017) and Galichon (2016) for an introduction to the subject from an economic perspective.

Despite its many uses and applications, the large sample properties of the classical Monge–Kantorovich optimal transport problem have yet to be fully established in the statistical literature. Recently, Sommerfeld and Munk (2018) proved asymptotic normality of its value function in the case where the respective probability measures have finite support; in a seminal contribution, del Barrio and Loubes (2019) proved central limit theorems for the value function for general probability measures under a weak moment condition using the Efron–Stein inequality. These results are important, as the value functions of the Monge–Kantorovich problem under L^p-cost functions are the Wasserstein distances on the space of signed Borel measures, which metrize weak convergence plus convergence in pth mean (Santambrogio, 2015, Thm. 5.11). However, especially in economic applications, one is often rather interested in the optimizer of these problems than the value function: the optimizer is the induced "optimal matching" between the two probability measures for the respective cost function. This optimal matching induces copulas which can potentially be useful in many multivariate settings, for instance, in the characterization of higher-order Markov processes (Ibragimov, 2009). In this respect, the *Brenier map*, the minimizer of the Monge–Kantorovich problem under the euclidean distance as a cost function, has received the most attention, because it possesses many interesting properties as an optimal matching, like monotonicity. In fact, Galichon (2016, p. 64) states that establishing the large sample properties of (the potential function of) the Brenier map is an important open problem.

This paper takes a first step to analyze this problem: by deriving and analyzing the second variation of the semi-dual Monge–Kantorovich problem, we derive a convergence rate of the simple plug-in estimator of the potential function of the Brenier map for smoothed empirical measures (i.e., integrated kernel density estimators) in the sense of Yukich (1992) and van der Vaart (1994).[1] Brenier (1991) showed that the Brenier map in this setting takes the form of the gradient $D\varphi$ of a convex function φ. We consider the Kantorovich potential function φ of the Brenier map an M-estimator and show that a natural sample counterpart $\hat{\varphi}_n$ converges to φ in standard deviation on every compact subset of the interior of its support under appropriate smoothness assumptions on the densities. Without a normalization on φ, at least one of the two probability measures between which the Brenier map acts has to satisfy *Poincaré's inequality* for this result to hold. However, under a normalization of φ that fixes its average value to zero (Lee, 2018), the convergence holds in the L^2 norm without requiring Poincaré's inequality. This complements the consistency proof for (the potential function of) the Brenier map established in Chernozhukov et al. (2017).

[1] The results we derive also hold for the standard empirical measure. In this case, the regularity of the potential function in finite samples does not follow from the regularity of the smoothed empirical measures, and one needs to make assumptions on the potential directly. See the discussion in Section 2.4.

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

Since the initial submission of this article, Hütter and Rigollet (2019) have derived the minimax rate of convergence for the Brenier map by constructing a theoretical wavelet estimator. We find that the rate of convergence we obtain for the potential function corresponds to the minimax rate of the kernel density estimators for the respective measures. This rate is slightly suboptimal compared to the minimax rate of the Brenier map derived in Hütter and Rigollet (2019). The reason for this is not our derived results of the semi-dual Monge–Kantorovich problem but is due to the fact that we rely on classical results for the rate of convergence of smoothed empirical processes in Giné and Nickl (2008) and Radulović and Wegkamp (2000). These results derive the stochastic equicontinuity of the M-estimator in question via an approximation of the smoothed empirical process by the standard empirical process. It turns out that this imposes too strong restrictions on the bandwidth h_n for our problem, and in fact ignores some of the additional regularity of the Brenier map, which makes our rate of convergence suboptimal.

Despite this, the current article shows that the simple plug-in estimator of the semi-dual Monge–Kantorovich problem, which is easier to implement and conceptually simpler than the theoretical wavelet estimator from Hütter and Rigollet (2019), performs well, and possesses enough regularity to potentially reach the minimax rate of convergence. In fact, as the main mathematical result of this article, we show that the second variation of the semi-dual Monge–Kantorovich problem takes the form of a Dirichlet energy functional weighted by the density function of the source measure. This problem has the same regularity properties in multiple dimensions as the infinitesimal generators of ergodic diffusions, which has been shown to be higher than the regularity of classical smoothed empirical processes by the seminal result Dalalyan and Reiß (2007, Prop. 1), see the analysis in Rohde and Strauch (2010). This additional regularity also exactly coincides with the additional regularity found in Hütter and Rigollet (2019). This strongly suggests that an application of these results in place of the classical results for smoothed empirical processes can lead to fewer restrictions on the admissible bandwidth which would imply the minimax rate of convergence found in Hütter and Rigollet (2019) for the simple plug-in estimator, without changing the estimator or mathematical results of this paper.

Our focus is to examine the rate of convergence of the potential function in general dimension $d \geq 2$, i.e., when the probability measures between which the Brenier map acts are supported in $\mathbb{R}^d$. The one-dimensional case is simpler and has already been solved, showing that the Brenier map converges at the parametric rate in this case; in the econometrics literature, this has been done—without mentioning the connection to optimal transport theory—in Athey and Imbens (2006) for instance. The reason for why the one-dimensional case is easier to handle lies in the fact that the Brenier map between the distributions F_X and F_Y has a closed form expression as the monotone rearrangement $D\varphi(x) := F_Y^{-1}(F_X(x))$; this is not true in higher dimensions, which is why one has to resort to more general tools in order to tackle this question. In particular, our main mathematical result is Lemma 1,

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

which provides the regularity conditions of the first and second variation of the semi-dual problem of the Monge–Kantorovich problem in our setting.

The structure of this article is as follows: Section 2 contains all theoretical results of this article. We state the lemma about the analysis of the second variation of the semi-dual Monge–Kantorovich problem in Section 2.1 and our assumptions in Section 2.2. Section 2.3 introduces the results for the rate of convergence of the kernel density estimator of the semi-dual of the optimal transport problem with and without normalization. We also state a simple suboptimal rate of convergence for the Brenier map based on the bandwidth used for the potential function. Section 2.4 contains a brief discussion about computational issues. We conclude in Section 3. Appendixes A and B contain a brief review of the Monge–Kantorovich problem and all proofs, including additional lemmas.

2. CONVERGENCE RATES

In this section, we state the rate of convergence of the natural plug-in estimators of φ and $D\varphi$ by considering the *semi-dual* problem to the Monge–Kantorovich problem, which is the dual to the optimal transport problem with the cost function $c(x,y) := -\langle x,y\rangle$:

$$\begin{aligned} &\min_{\varphi,\psi} \int_{\mathcal{X}} \varphi(x) dP_X(x) + \int_{\mathcal{Y}} \psi(y) dP_Y(y) \\ &\text{s.t.} \quad \varphi(x) + \psi(y) \geq \langle x,y\rangle, \quad \varphi \in C(\mathcal{X}), \psi \in C(\mathcal{Y}), \end{aligned} \tag{1}$$

where $C(\mathcal{X})$ is the space of all bounded continuous functions on the support $\mathcal{X}$ of P_X, endowed with the standard supremum norm $\|f\|_\infty := \sup_{x\in\mathcal{X}} |f(x)|$, and where $\langle x,y\rangle$ denotes the standard scalar product of vectors $x, y \in \mathbb{R}^d$.[2] Throughout, we assume that P_X and P_Y are absolutely continuous with respect to Lebesgue measure with densities f_X and f_Y, respectively. Theorem 1.3 in Villani (2003) shows that there is no duality gap for the Monge–Kantorovich problem in this setting, so that we can in fact use the dual problem to derive the rate of convergence of the estimator of φ. Note that the optimal solution of (1) always consists of conjugate duals, i.e., $\psi = \varphi^*$, where the convex conjugate φ^* of φ is defined as

$$\varphi^*(y) = \sup\{\langle y,x\rangle - \varphi(x) : x \in \mathcal{X}\},$$

see the first step in the proof of Proposition 3.1 in Brenier (1991). Therefore, the problem reduces to simply estimating φ, and we can write (1) as

$$M(\varphi) = \min_{\varphi} \int_{\mathcal{X}} \varphi(x) dP_X(x) + \int_{\mathcal{Y}} \varphi^*(y) dP_Y(y), \quad \varphi \in C(\mathcal{X}). \tag{2}$$

[2]The Monge–Kantorovich problem with cost function $c(x,y) := -\langle x,y\rangle$ is equivalent to the Monge–Kantorovich problem with quadratic cost function $c(x,y) = \|x-y\|_2^2/2$, where $\|\cdot\|_2$ denotes the euclidean norm, in the sense that they generate the same solution. We focus on the former cost function, because it allows us to write its dual problem in the above convenient manner. The dual to this problem is known as the semi-dual problem in the literature. For further information, we refer to Appendix A.

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

In practice, there are many ways to estimate the Brenier map, see, for instance, Benamou and Brenier (2000), Benamou, Froese, and Oberman (2014), and Chartrand et al. (2009) among others. The latter approach introduces an infinite dimensional gradient descent method to calculate the Brenier map, using ideas from the calculus of variations to derive this result. We go a very similar route in this article. We use ideas from the calculus of variations in conjunction with the fact that φ is an M-estimator to derive a rate of convergence for its natural plug-in estimator, i.e., its smoothed sample analog $\hat{\varphi}_n$, which is the solution to

$$\underset{\varphi \in C(\mathcal{X})}{\operatorname{argmin}} \int_{\mathcal{X}} \varphi(x) d\mathbb{P}_n^X * K_{h_n}(x) + \int_{\mathcal{Y}} \varphi^*(y) d\mathbb{P}_n^Y * K_{h'_n}(y), \tag{3}$$

where we use the notation from Giné and Nickl (2008) and define the standard kernel density estimator by

$$\hat{f}_{h_n}^X(x) \equiv \mathbb{P}_n^X * K_{h_n}(x) := \frac{1}{nh_n^d} \sum_{i=1}^n K\left(\frac{X_i - x}{h_n}\right).$$

Here, $K\left(\frac{X_i - x}{h_n}\right) := K'\left(\frac{X_{i1} - x_1}{h_n}\right) \cdots K'\left(\frac{X_{id} - x_1}{h_n}\right)$ denotes a product smoothing kernel, h_n denotes the bandwidth, which for ease of notation we assume to be the same for all dimensions, and $f * g$ denotes the convolution between two functions f and g. The same notation holds for $\mathbb{P}_n^Y * K_{h'_n}(y)$. n denotes the number of observations of the sample throughout.[3]

For this rate of convergence result to be applicable in practice, we must assume that computational implementation of (3) delivers a strictly convex solution $\hat{\varphi}_n$. This can be achieved by introducing a strict convexity penalty term that enforces convexity of $\hat{\varphi}_n$ for all iterations of the algorithm. The penalty term can usually be chosen such that it gradually disappears with the number of iterations, so that it is not binding at the solution. We found that this type of regularization prevents algorithms like the one in Chartrand et al. (2009) from overfitting and hence makes the program more robust, which is another reason to use this in implementations, see, for instance, Gunsilius and Schennach (2019).

2.1. Main Lemma

In order to derive the rate of convergence of $\hat{\varphi}_n$, we need to make regularity assumptions on the densities f_X and f_Y as well as the kernel $K\left(\frac{X_i - x}{h_n}\right)$ and the bandwidth h_n. In addition, we make use of Caffarelli's regularity theory for optimal transport maps, in particular the a priori interior estimates established in Caffarelli (1990) for which we have to make further regularity assumptions. Throughout, we will be working in Hölder spaces. Let $k := (k_1, \ldots, k_d)$ be a multi-index of

[3] One can extend this result to the two-sample case straightforwardly.

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

nonnegative integers $k_1, \dots, k_d$. Set $|k| = \sum_{i=1}^{d} k_i$ and write

$$D^k := \frac{\partial^{|k|}}{(\partial x_1)^{k_1} \cdots (\partial x_d)^{k_d}}.$$

We define the Hölder norm $\|f_X\|_{C^{s,\alpha}}$ for $s \in \mathbb{N}_0$ and $\alpha \in (0,1]$ of the Hölder space $C^{s,\alpha}(\mathcal{X})$ as

$$\|f\|_{C^{s,\alpha}} := \sum_{0 \le |k| \le s} \|D^k f\|_\infty + \sum_{k:|k|=s} \sup_{x_1, x_2 \in \mathcal{X}: x_1 \neq x_2} \frac{\|D^k f(x_1) - D^k f(x_2)\|}{|x_1 - x_2|^\alpha},$$

where $\|\cdot\|$ stands again for the euclidean norm in $\mathbb{R}^d$. In this sense, a function $f_X \in C^{s,\alpha}(\mathcal{X})$ if there exists a constant $c < +\infty$ such that $\|f_X\|_{C^{s,\alpha}} \le c$. Based on this, we say that a function f_X is *locally Hölder continuous* on $\mathcal{X}$ if it is Hölder continuous on every compact subset of the interior of $\mathcal{X}$, which we denote by $f \in C^{s,\alpha}_{loc}(\mathcal{X}^\circ)$.[4]

The main mathematical result of this article upon which all statistical derivations rely is the following lemma, which provides the regularity conditions of the first and second variation of (2). It might be of independent interest.[5]

LEMMA 1. *Let P_X and P_Y be absolutely continuous measures with respect to Lebesgue measure with densities f_X and f_Y on convex and compact supports $\mathcal{X}$ and $\mathcal{Y}$. Then, the functional M is convex, Lipschitz-continuous, and Hadamard differentiable on $C(\mathcal{X})$. The first variation $\delta M_\varphi(v)$ of $M(\varphi)$ on $C(\mathcal{X})$ is*

$$\delta M_\varphi(v) = \int_{\mathcal{X}} v(x) f_X(x) dx + \int_{\mathcal{Y}} v\left(D\varphi^*(y)\right) f_Y(y) dy.$$

Based on this, the Hadamard differential takes the form

$$\langle M'(\varphi), v \rangle \equiv \int_{\mathcal{X}} \left[f_X(x) - f_Y(D\varphi^{**}(x)) \det(D^2\varphi^{**}(x)) \right] v(x) dx.$$

The second variation $\delta^2 M_\varphi(v,u)$ on $C^1(\mathcal{X})$ is

$$\begin{aligned} \delta^2 M_\varphi(v,u) &= \int_{\mathcal{Y}} \langle Dv(D\varphi^*(y)), Du\left(D\varphi^*(y)\right) \rangle f_Y(y) dy \\ &= \int_{\mathcal{X}} \langle Dv(x), Du(x) \rangle f_Y(D\varphi^{**}(x)) \det(D^2\varphi^{**}(x)) dx. \end{aligned}$$

If we restrict $M(\cdot)$ to $C^2(\mathcal{X}^\circ)$, then there exists a neighborhood $\mathcal{N}_{\varphi_0} \subset C^2(K)$ around the strictly convex φ_0 such that $M'(\varphi)$ coincides with the Fréchet derivative on $\mathcal{N}_{\varphi_0}$, where $K \subset \mathcal{X}^\circ$ is an arbitrary compact set. Moreover, the second variation $\delta^2 M_\varphi(v,u)$ is continuous in φ on $\mathcal{N}_{\varphi_0}$ and a bounded bilinear functional in $u, v \in$

[4] We denote the topological interior of a set A by A°.

[5] The integral of some function g with respect to Lebesgue measure is denoted by $\int g(x)dx$.

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

$C^2(\mathcal{X})$. *In particular, for* $u \equiv v$, *it holds that*

$$\delta^2 M_\varphi(v,v) = \int_{\mathcal{X}} \|Dv(x)\|^2 f_Y(D\varphi^{**}(x)) \det(D^2\varphi^{**}(x))dx$$

and, for every $\varepsilon > 0$ *and convex compact* $K \subset \mathcal{X}^\circ$, *there exists an* $\eta(\varepsilon) > 0$ *such that*

$$\left|\delta^2 M_\varphi(v,v) - \delta^2 M_{\varphi_0}(v,v)\right| \leq \varepsilon \|v\|^2_{C^2(K),}$$

for all $v, \varphi \in C^2(\mathcal{X})$ *with* $\|\varphi - \varphi_0\|_{C^2(K)} \leq \eta(\varepsilon)$.

Lemma 1 is needed for a Taylor-expansion argument. The apparently novel result is the derivation of the formula for the second variation in all directions $C^1(\mathcal{X})$, along with its properties.[6]

2.2. Assumptions

We can now state the required assumptions for the statistical result.

Assumption 1 (Regularity of the densities) The densities f_X and f_Y have the following properties:

(i) The supports $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$ of f_X and f_Y are compact and convex.
(ii) f_X and f_Y are bounded above and below, i.e., there exists $0 < \gamma < +\infty$ such that $\gamma^{-1} \leq f_Y(y) \leq \gamma$ and $\gamma^{-1} \leq f_X(x) \leq \gamma$, for all $y \in \mathcal{Y}$ and $x \in \mathcal{X}$.
(iii) $f_X \in C^{s,\alpha}_{loc}(\mathcal{X}^\circ)$ and $f_Y \in C^{s,\alpha}_{loc}(\mathcal{Y}^\circ)$, with degree $s \geq 0$, for some $\alpha \in (0,1)$, which is fixed throughout.

The assumptions on the support of the densities are required for Caffarelli's interior regularity theory (Caffarelli, 1990), see Lemma 2 in Appendix B.2. The boundedness assumption of the density will make the Monge–Ampère operator associated with the Monge–Kantorovich problem (2) an elliptic operator. Ellipticity is important, as it allows us to prove convergence of $\hat{\varphi}_n$ to φ_0 in Hölder norm using standard Schauder estimates of elliptic second-order partial differential equations (PDEs; Gilbarg and Trudinger, 1998, Chap. 6). See Lemma 3 in Appendix B.2.

Assumption 2 (Order of kernels and bandwidths) The kernel K is nonnegative and of second order. The bandwidths $h_n, h'_n > 0$ satisfy $h_n^2 n^{1/2} \to 0$ and $(h'_n)^2 n^{1/2} \to 0$ as $n \to +\infty$.

Assumption 2 is a standard assumption on the bandwidths and the kernel, taken directly from Theorem 6 in Giné and Nickl (2008), which provides the asymptotic

[6] By a simple formal differentiability argument (e.g., Rubinstein (2008)), one can derive the formula of the second variation in convex directions. The contribution here is the rigorous derivation in all directions $C^1(\mathcal{X})$ and the analysis of the regularity properties.

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

distribution result for smoothed empirical processes indexed by univariate Hölder functions. In particular, the upper bound on the bandwidth makes the bias of the smoothed empirical process asymptotically negligible as proved in Theorem 6 of Giné and Nickl (2008). We adapt their result to the multivariate setting in a straightforward way Giné and Nickl (2008, p. 369). We are more restrictive than this result by requiring nonnegative kernels, as they allow us to relate bracketing numbers of the empirical processes to their smoothed empirical counterparts. It is possible to allow for higher-order kernels like the ones in Giné and Nickl (2008) if one replaces our bracketing entropy approach by uniform entropy under some formal complications. Note also that we do not require a lower bound on the amount of smoothing needed. This is because the potential φ will always be in a Donsker class, so that the empirical process will converge for this class even without smoothing (Giné and Nickl, 2008, p. 344).

Without a normalization on the potential φ, we need to uphold another assumption on P_X. This assumption will make the objective function of the dual of the Kantorovich problem well-separated at the optimal φ_0 in an appropriately chosen semi-metric. In fact, we require that P_X satisfy the *Poincaré inequality,* for any $v : \mathbb{R}^d \to \mathbb{R}$, which reads

$$\int_{\mathcal{X}} \|Dv(x)\|^2 dP_X(x) \geq c\mathrm{Var}_{P_X}(v), \tag{4}$$

where

$$\begin{aligned}\mathrm{Var}_{P_X}(v) :&= \int_{\mathcal{X}} |v(x)|^2 dP_X(x) - \left(\int_{\mathcal{X}} v(x) dP_X(x)\right)^2 \\ &= \|v\|^2_{L^2(P_X)} - \left(\int_{\mathcal{X}} v(x) dP_X(x)\right)^2 \end{aligned} \tag{5}$$

is the variance of v with respect to the measure P_X, $\|Dv(x)\|^2$ is the squared euclidean norm of the gradient of $v(x)$, and $c > 0$ is some constant.

Assumption 3 The probability measure P_X satisfies the Poincaré inequality (4).

2.3. Convergence Rates for the Potential

The following is our main result.

THEOREM 1 (Rate of convergence of φ). *Under Assumptions 1–3, it holds that*

$$r_n \sqrt{Var_{P_X}(\hat{\varphi}_n - \varphi_0)} = O_{P^*}(1), \tag{6}$$

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

for

$$r_n = \begin{cases} n^{\frac{s+\alpha}{2(s+\alpha)+d}} & \text{for } s+\alpha > \dfrac{d}{2} \\ \dfrac{n^{1/4}}{(\log(n))^{1/2}} & \text{for } s+\alpha = \dfrac{d}{2} \\ n^{\frac{1}{(s+\alpha)(2(s+\alpha)+d)}} & \text{for } s+\alpha < \dfrac{d}{2} \end{cases}.$$

Here, φ_0 is the minimizer of (2) and $\hat{\varphi}_n$ is the minimizer of (3).[7]

Theorem 1 implies a curse of dimensionality of the same form as the one for the square root of the asymptotic integrated mean-square error for kernel density estimators. This is not surprising due to the fact that kernel density estimators are used to estimate the probability measures. Note that, for $s+\alpha > d/2$, this is the standard minimax rate of estimation of kernel density estimators (Tsybakov, 2008, Thm. 1.1).

As mentioned in the Introduction, the rate for the case $s+\alpha > \frac{d}{2}$ is slightly worse than the minimax rate for estimators of the Brenier map in Hütter and Rigollet (2019); indeed, a rate of the potential of $n^{\frac{s+\alpha}{2(s+\alpha)+d}}$ would in the best possible case correspond to a rate of convergence of $n^{\frac{s+\alpha}{2(s+\alpha)+2+d}}$, because one would sacrifice "one degree of smoothness" when estimating the Brenier map directly, i.e., the gradient of the potential function. Hütter and Rigollet (2019) derive a minimax rate of $n^{\frac{s+\alpha}{2(s+\alpha)-2+d}}$ for the Brenier map, which is slightly faster than the potential rate achievable by the standard plug-in estimator for our bandwidth h_n. The reason is that we rely on classical asymptotic results for smoothed empirical processes which do not take into account the additional regularity of the Brenier map and hence impose the strong upper bound on the bandwidth $h_n^4 n \to 0$.

As for Assumption 3, the requirement that a probability measure satisfies Poincaré's inequality is a high-level requirement, and it is an active area of research linked to the concentration of measure phenomenon to establish when this requirement is satisfied. In particular, if P_X satisfies the Poincaré inequality, then it satisfies an exponential concentration inequality (Ledoux, 2001, Cor. 3.2). One result establishes the reverse connection between the Poincaré inequality and a dimension-free concentration inequality (Gozlan, Roberto, and Samson, 2015, Thm. 1.2). Many important probability measures satisfy Poincaré's inequality, such as the exponential and uniform distribution, most notably the set of log-concave probability measures, which are measures possessing density functions $e^{-V(x)}$ for some convex and twice continuously differentiable $V(x)$, but there are many more classes (e.g., Bakry et al. (2008)).

Intuitively, Poincaré's inequality allows us to work with the standard deviation as a semi-norm on the Hölder space defined above. This semi-norm creates an

[7] P^* denotes outer probability, which we use to circumvent measurability issues in the nonseparable space $\ell^\infty(C^{s,\alpha}_{loc}(\mathcal{X}^\circ))$ of all bounded functionals on the Hölder space $C^{s,\alpha}_{loc}(\mathcal{X}^\circ)$.

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

equivalence class of functions of the form $u \sim v$ if and only if $u = v + c$, where c is a constant function on $\mathcal{X}$. Under this equivalence class, the solution to the dual problem of the Monge–Kantorovich problem is unique and well-separated, which is one of two main requirements for our derivation of the rate of convergence of the plug-in estimator.[8] In general, the potential function of the Brenier map is only identified up to an additive constant, which would make the optimum not well-separated under a different (semi-) metric.

Since the main quantity of interest in general is the Brenier map, a usual approach in the literature on optimal transport theory is to fix the potential function by fixing the coordinate system. The standard normalization for this is $\varphi(0) = 0$.[9] However, Lee (2018), who relies on some of the results in the present article to derive a sieve estimation procedure for optimal transport maps, introduces the following helpful and natural normalization for the potential function:

$$\int_{\mathcal{X}} \varphi(x)dx = 0, \tag{7}$$

i.e., fixing the mean to be zero. This normalization is rather helpful, because it makes the standard deviation coincide with the L^2 norm, so that all convergence arguments in the proof go through in the L^2 norm. Furthermore, under this normalization, we can replace the generalized Poincaré inequality with respect to the general probability measure P_X by the classical Poincaré inequality with respect to Lebesgue measure. In other words, if the potential function φ is normalized to have zero mean and is defined on a compact support, then it automatically satisfies the Poincaré inequality with respect to P_X. The classical Poincaré inequality satisfies

$$\int_{\mathcal{X}} \|Dv(x)\|^2 dx \geq c\mathrm{Var}(v) = c\|v\|^2_{L^2(\mathcal{X})} - c\left(\int_{\mathcal{X}} v(x)dx\right)^2 = c\|v\|^2_{L^2(\mathcal{X})} \tag{8}$$

for some constant $c < +\infty$ and where the last inequality follows from the normalization (7).[10] Now, by Hölder's inequality and the fact that f_X is bounded above by $\gamma < +\infty$, it follows that

$$\|v\|^2_{L^2(P_X)} = \int_{\mathcal{X}} v(x)^2 f_X(x)dx \leq \gamma \int_{\mathcal{X}} v(x)^2 dx = \gamma \|v\|^2_{L^2(\mathcal{X})}. \tag{9}$$

Since f_X is bounded above and below by γ and γ^{-1} and $\mathcal{X}$ is compact by Assumption 1, it follows that

$$\begin{aligned}\int_{\mathcal{X}} \|Dv(x)\|^2 dx &\leq \frac{1}{\inf_{x\in\mathcal{X}} f_X(x)} \int_{\mathcal{X}} \|Dv(x)\|^2 f_X(x)dx \\ &\leq \gamma \int_{\mathcal{X}} \|Dv(x)\|^2 f_X(x)dx < +\infty.\end{aligned} \tag{10}$$

[8] Also note that working with the standard-deviation semi-metric on $C^{s,\alpha}(\mathcal{X})$ does not pose problems, as $C^{s,\alpha}(\mathcal{X})$ embeds compactly into $L^2(P_X)$ by Hölder's inequality and the Arzelà–Ascoli theorem.

[9] An example for this normalization is Chernozhukov et al. (2017).

[10] We denote by $\|\cdot\|_{L^2(\mathcal{X})}$ the L^2 norm with respect to Lebesgue measure on $\mathcal{X}$.

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

Putting equations 8–10 together implies

$$\int_{\mathcal{X}} \|Dv(x)\|^2 f_X(x)dx \geq \frac{c}{\gamma}\|v\|^2_{L^2(P_X)}.$$

This implies that the optimal φ_0 is well-separated with respect to $\|\cdot\|_{L^2(P_X)}$, and we immediately obtain the following.

COROLLARY 1 (Rate of convergence under normalization). *Under Assumptions 1 and 2, it holds that*

$$r_n \left\|\hat{\varphi}_n - \varphi_0\right\|_{L^2(P_X)} = O_{P^*}(1), \tag{11}$$

for

$$r_n = \begin{cases} n^{\frac{s+\alpha}{2(s+\alpha)+d}} & \text{for } s+\alpha > \dfrac{d}{2} \\ \dfrac{n^{1/4}}{(\log(n))^{1/2}} & \text{for } s+\alpha = \dfrac{d}{2} \\ n^{\frac{1}{(s+\alpha)(2(s+\alpha)+d)}} & \text{for } s+\alpha < \dfrac{d}{2} \end{cases}.$$

Here, φ_0 is the minimizer of (2), $\hat{\varphi}_n$ is the minimizer of (3), and where we uphold the normalization (7).

Based on the results established in the previous section, we can derive a rate of convergence for the Brenier map $D\varphi$ by subtracting one degree of smoothness in our rate of convergence for the potential function in the case where $s+\alpha > \frac{d}{2}$. This is a classical result in the setting where $D\varphi$ has enough smoothness, but requires that we change the bandwidth h_n from the estimation of the potential function. Moreover, in the other two regimes, it is less clear how the rate of convergence behaves. The following proposition therefore provides a crude upper bound for the rate of convergence of the Brenier map based on the convergence rate of its potential by bounding the squared euclidean distance of gradients of convex functions by the distance of their potential functions as in the proof of Theorem 2.33 in Attouch and Wets (1986). It provides the rate of convergence of the Brenier map *for the bandwidth used to estimate the potential function* and is therefore by definition suboptimal compared to the rate of convergence where we are allowed to change h_n. However, this result is applicable in all three different cases for the smoothness by relying on the bandwidth used for the potential function and even holds for functions φ that are not twice continuously differentiable, but only possess a Lipschitz-continuous first derivative.

PROPOSITION 1 (Rate of convergence of $D\varphi$ for the bandwidth of the potential). *Under Assumptions 1 and 2 and by using the optimal bandwidth h_n for obtaining the rate of convergence of the potential function φ, it holds that*

$$r_n \left\|D\hat{\varphi}_n - D\varphi_0\right\|_{L^2(P_X)} = O_{P^*}(1), \tag{12}$$

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

for

$$r_n = \begin{cases} n^{\frac{s+\alpha}{4(s+\alpha)+2d}} & \text{for } s+\alpha > \frac{d}{2} \\ \frac{n^{1/8}}{(\log(n))^{1/4}} & \text{for } s+\alpha = \frac{d}{2} \\ n^{\frac{1}{(s+\alpha)(4(s+\alpha)+2d)}} & \text{for } s+\alpha < \frac{d}{2} \end{cases},$$

where φ_0 is the minimizer of (2), $\hat{\varphi}_n$ is the minimizer of (3), and where we uphold the normalization (7).

Note that this rate is rather slow and far from optimal. In particular, it is the square root of the rate of the potential. This makes it much slower than the minimax rate obtained in Hütter and Rigollet (2019). This suboptimality should not be surprising, as this rate of convergence is based on the optimal bandwidth used for the potential function and not optimized for estimating the Brenier map directly, as mentioned. The novelty of Proposition 1 is that it provides a direct connection between the Brenier map and its potential function for a fixed bandwidth.

2.4. Practical Considerations

The rate obtained holds for the optimization problem (3). The first Hadamard derivative in Lemma 1 suggests the following simple functional gradient descent approach which was suggested in Chartrand et al. (2009):

$$\varphi_{k+1} = \varphi_k - \alpha_k M'(\varphi_k),$$

where $k = 1, \dots$ denotes the iterations, α_n is a step size, and M' is the Hadamard derivative derived in Lemma 1. In practice, this approach can be implemented by a grid approach (Gunsilius and Schennach, 2019): place a grid on your data and estimate both densities via a kernel density estimator on the grid. Then, compute the optimal transport map by computing the gradient descent step at each point of the grid.

This approach clearly suffers from a computational curse of dimensionality, not least since it relies on kernel density estimation. The number of grid points grows exponentially with the dimension, so that this approach is only applicable in lower dimensions. Also, in all of our analysis, we do not account for the computational approximation error induced via the specific optimization procedure induced by the gradient descent algorithm, and we exclusively focus on the statistical rate of convergence. Let us also address the implementation of the normalization (7). In a practical grid-approach, it is not clear a priori how to include this normalization, as it requires knowledge of the potential function, which is the object we want to estimate. In contrast, in sieve estimation procedures like Lee (2018), one can include the normalization straightforwardly, by only focusing on sets of potentials that are normalized in this way.

Lastly, note that all results presented in this article work for the standard empirical measure, without smoothing; this also manifests itself in the fact that

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

we do not require a lower bound on the bandwidth in Assumption 2. The simple model for this would be

$$\operatorname*{argmin}_{\varphi \in C^{s,\alpha}_{loc}(\mathcal{X}^\circ)} \sum_{i=1}^{n} \varphi(X_i) + \sum_{i=1}^{n} \varphi^*(Y_i).$$

In this case, one needs to make the smoothness assumptions on $\hat{\varphi}_n$ directly, as they do not follow from the kernel density estimates in finite samples. In practice, this optimization is still hard to solve, even though it is not based on a kernel density estimation. This is not just a problem in our setting, but for the computational theory of optimal transport more generally (see, for instance, the discussions in other implementations such as Benamou et al. (2014) and Seguy et al. (2017)). One way to circumvent this curse is to use the Sinkhorn divergence introduced in Cuturi (2013), which penalizes the optimal transport problem with an entropy term and can be solved efficiently via Sinkhorn iterations (Sinkhorn, 1967). The optimizer does not coincide with the optimal transport map in this case, however. Computing the optimal transport map efficiently in practice in high dimensions is therefore still an open problem.

3. CONCLUSION

We have derived a convergence rate for the simple plug-in estimator of the semi-dual Monge–Kantorovich problem for the potential function of the Brenier map. It coincides with the minimax rate of convergence of the kernel density estimator of Hölder classes if the respective probability densities are smooth enough. The idea is to use the semi-dual problem of the Monge–Kantorovich problem and analyze the second variation of this problem for a Taylor-expansion argument. Without a normalization on the potential, this rate will be achieved for the standard deviation and requires that at least one of the measures satisfies the Poincaré inequality. Under the normalization that the first moment of the potential is zero, this convergence takes place in L^2-norm and does not require the measure to satisfy the Poincaré inequality anymore.

The obtained rate of convergence is suboptimal compared to the minimax rate of convergence recently derived via a wavelet estimator in Hütter and Rigollet (2019). This suboptimality follows from the fact that we rely on classical asymptotic results for empirical processes as derived in Giné and Nickl (2008), which requires the strict bound $h^4 n \to 0$ on the bandwidth. The main mathematical result in this article shows, however, that the second variation of the semi-dual Monge–Kantorovich problem takes the form of a classical Dirichlet energy functional weighted by the source measure. This problem has the same regularity properties in multiple dimensions as the infinitesimal generators of ergodic diffusions, which has been shown to be higher than the regularity of classical smoothed empirical processes by the seminal result Dalalyan and Reiß (2007, Prop. 1), see the analysis in Rohde and Strauch (2010). This additional regularity also exactly coincides with the additional regularity found in Hütter and Rigollet (2019) and implies that an

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

application of these results in place of the classical results for smoothed empirical processes can lead to fewer restrictions on the admissible bandwidth which would imply the minimax rate of convergence found in Hütter and Rigollet (2019).

The results derived in this article can potentially be used to derive the asymptotic distribution of the potential function of the Brenier map by making use of the second variation of the semi-dual Monge–Kantorovich problem derived in Lemma 1.

APPENDIX

A. THE MONGE–KANTOROVICH PROBLEM AND THE BRENIER MAP

This section is designed to give the reader a very brief overview of the Monge–Kantorovich problem and the Brenier map. For more information, we refer to the introductory text Villani (2003).

The goal of the theory of optimal transport is to analyze maps T which transport one probability measure P_X onto another probability measure P_Y in a "cost-efficient way." The setup for this is the *Monge–Kantorovich problem*. To be precise, the Monge and Kantorovich problems are actually two different problems, the latter being the convex relaxation of the former. Monge's problem asks for an optimal transport map between two (probability) measures, P_X and P_Y defined on their supports $\mathcal{X}$ and $\mathcal{Y}$, where optimality is measured with respect to some cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. This problem can be stated as

$$\text{minimize} \int_{\mathcal{X}} c\,(x, T(x))\, dP_X(x) \quad T : \mathcal{X} \to \mathcal{Y} \text{ measurable.} \tag{A.1}$$

In words, the Monge problem asks for an explicit measurable map $T : \mathcal{X} \to \mathcal{Y}$ which turns out to take any Borel set $E_x \subset \mathcal{X}$ and maps it to some Borel set $TE_x \subset \mathcal{Y}$ "of the same size.". The formal property is that T *preserves measure* in the sense that $P_X(T^{-1}E_y) = P_Y(E_y)$, i.e., it needs to "match" every Borel set E_y of size $P_Y(E_y)$ to a corresponding Borel set $E_x = T^{-1}E_y$ of the same size $P_X(T^{-1}E_x) = P_Y(E_y)$.

The Kantorovich problem between (probability) measures P_X and P_Y under some cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is the convex relaxation of the Monge problem; it only asks for an optimal transport *plan* in the sense that the transport does not have to be accomplished through a function as in the Monge problem, but is concentrated on the support Γ of a joint probability distribution γ which has P_X and P_Y as its marginals

$$\min_{\pi \in \Pi(P_X, P_Y)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\pi\,(x, y), \tag{A.2}$$

where $y \in \mathcal{Y}$, $x \in \mathcal{X}$, and $\Pi(P_X, P_X)$ is the set of all probability measures on $\mathcal{X} \times \mathcal{Y}$ such that the marginal distributions of some $\pi \in \Pi(P_X, P_Y)$ are precisely P_X and P_Y.

For many cost functions c, the solution to the Monge and Kantorovich problems actually coincides under the assumption that P_X is absolutely continuous with respect to Lebesgue measure, so that it is legitimate to speak of the Monge–Kantorovich problem in these cases. Moreover, this solution is *unique* for many important cost functions. All of these statements can be found in Chapters 1 and 2 of Villani (2003). Different cost functions c lead to different transport maps or transport plans, many of which occur naturally in economics and econometrics. The Brenier map results from solving the Monge–Kantorovich problem

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

under the standard squared euclidean distance as a cost function, i.e.,

$$c(x-y) = \|x-y\|^2 = \sum_{i=1}^{d} |x_i - y_i|^2.$$

In fact, an equivalent formulation of the Monge–Kantorovich problem under quadratic cost in the sense that they generate is

$$\sup_{\pi \in \Pi(P_X, P_Y)} \int \langle x, y \rangle d\pi(x,y), \tag{A.3}$$

whose dual problem we consider in this article. The importance lies in the fact that the dual to the Monge–Kantorovich under the quadratic cost does not admit a solution of the form (φ, φ^*), whereas the latter does.[11]

Brenier (1991) first proved that if P_X and P_Y possess finite second-order moments and if P_X is absolutely continuous with respect to Lebesgue measure, then the Monge and Kantorovich problems coincide, and the *unique* solution to

$$\operatorname*{argmin}_{T} \int_{\mathcal{X}} \|x - T(x)\|^2 dP_X(x), \quad T \text{ measurable} \tag{A.4}$$

takes the form of the *gradient of a convex function*, i.e., $T_0(x) = D\varphi(x)$, for some convex φ. Moreover, one can choose φ as the optimal *Kantorovich potential*, which is the solution of (2).

B. PROOFS

B.1. Proof of Lemma 1.

Proof The result for convexity and Lipschitz continuity follow straightforwardly from the properties of the Legendre–Fenchel transform, see Theorem 3.1 in Chartrand et al. (2009). So let us turn to deriving the variations and showing Hadamard differentiability in all directions $v \in C(\mathcal{X})$.

We do this by a simple argument using subgradients which has already been derived in Gangbo (1994) and used in Chartrand et al. (2009) and Villani (2003, p. 74).[12] The idea is to use the limit definition of the directional derivative, which is

$$\begin{aligned}\delta M_\varphi(v) &= \lim_{t \to 0} \frac{M(\varphi + tv) - M(\varphi)}{t} \\ &= \int_{\mathcal{X}} v(x) f_X(x) dx + \lim_{t \to 0} \int_{\mathcal{Y}} \frac{(\varphi + tv)^*(y) - \varphi^*(y)}{t} f_Y(y) dy.\end{aligned}$$

Then, since φ is convex and hence differentiable almost everywhere, we can fix a $y_0 \in \mathcal{Y}$ such that $D\varphi^*(y_0) = x_0$, for some x_0. Note on the other hand that $\varphi + tv$ need not be convex; we therefore fix some $x_t \in \partial(\varphi + tv)^*(y_0) = \operatorname{argmin}_{x \in \mathcal{X}} \{\langle x, y_0 \rangle - \varphi(x) - tv(x)\}$, where $\partial(\varphi + tv)^*$ denotes the subdifferential of $(\varphi + tv)^*$. This x_t is finite since $\mathcal{X}$ is compact.

[11] I thank a referee for pointing this out.

[12] A simple differentiability argument only calculates the variation in strictly convex directions v, which is not enough for our purposes. I thank a referee for pointing this out.

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

Then, we can write

$$(\varphi + tv)^*(y) - \varphi^*(y) = \langle x_t, y_0\rangle - \varphi(x_t) - tv(x_t) - \langle x_0, y_0\rangle + \varphi(x_0).$$

Now, the reasoning is exactly analogous to the reasoning in both Chartrand et al. (2009) and Villani (2003, p. 74). In fact, replacing x_t by x_0 in the right-hand side of the last equation gives a smaller quantity, whereas replacing x_0 by x_t gives a larger quantity, so that we can bound

$$\begin{aligned} &-tv(x_0) \le (\varphi + tv)^*(y) - \varphi^*(y) \le -tv(x_t) \Leftrightarrow \\ &\quad 0 \le \frac{(\varphi + tv)^*(y) - \varphi^*(y)}{t} + v(x_0) \le v(x_0) - v(x_t). \end{aligned}$$

Since v is bounded and continuous, we can pass the limit into the integral by the Dominated Convergence Theorem, so that

$$\delta M_\varphi(v) = \int_{\mathcal{X}} v(x) f_X(x)dx + \int_{\mathcal{Y}} \lim_{t\to 0} \frac{(\varphi + tv)^*(y) - \varphi^*(y)}{t} f_Y(y)dy.$$

The fact that v is uniformly continuous on $\mathcal{X}$ implies that the sequence tv converges uniformly to 0. Therefore any convergent subsequence will converge to a maximizer of $\langle x, y_0\rangle - \varphi(x)$; since x_0 is the unique minimizer (the gradient exists), it follows that $x_t \to x_0$ and hence $v(x_0) - v(x_t) \to 0$. This in turn implies that

$$\delta M_\varphi(v) = \int_{\mathcal{X}} v(x) f_X(x)dx + \int_{\mathcal{Y}} v\big(D\varphi^*(y)\big) f_Y(y)dy$$

and proves the form of the first variation in all directions $v \in C(\mathcal{X})$ at a convex φ.

Finally, note that Hadamard differentiability of M at φ is equivalent to the existence of a measure $\mu \in C^*(\mathcal{X})$ such that $\delta M_\varphi(v) = \int_{\mathcal{X}} v(x)\mu(dx)$, since M is Lipschitz continuous and hence automatically Hadamard differentiable if the directional derivative is linear (also see Chartrand et al. (2009)). We can derive this form of the first variation by the change of variables $y = D\varphi^{**}(x)$, in which case we get

$$\begin{aligned} \delta M_\varphi(v) &= \int_{\mathcal{X}} v(x) f_X(x)dx + \int_{\mathcal{Y}} \lim_{t\to 0} \frac{(\varphi + tv)^*(y) - \varphi^*(y)}{t} f_Y(y)dy \\ &= \int_{\mathcal{X}} v(x) f_X(x)dx - \int_{\mathcal{X}} v(D\varphi^*(D\varphi^{**}(x))) f_Y(D\varphi^{**}(x)) \det(D^2\varphi^{**}(x))dx \\ &= \int_{\mathcal{X}} \Big[f_X(x) - f_Y(D\varphi^{**}(x)) \det(D^2\varphi^{**}(x))\Big] v(x)dx, \end{aligned}$$

where the last line follows from the fact that $x = D\varphi^*(D\varphi^{**}(x))$, so that we can define $\mu(dx) := [f_X(x) - f_Y(D\varphi^{**}(x)) \det(D^2\varphi^{**}(x))]dx$ which establishes the Hadamard derivative.

Let us now turn to the second variation and compute it in all possible directions $v, u \in C^1(\mathcal{X})$. In general, we can find the second variation by taking the variation of the first variation. We denote the first variation of the Legendre–Fenchel transform in direction v at point $y \in \mathcal{Y}$ by

$$\delta(\varphi; v)^*(y) = \lim_{t\to 0} \frac{(\varphi + tv)^*(y) - \varphi^*(y)}{t}.$$

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

Therefore, we can calculate the second variation as

$$\lim_{s\downarrow 0}\frac{\delta M_{\varphi+su}(v)-\delta M_{\varphi}(v)}{s}=\lim_{s\downarrow 0}\int_{\mathcal{Y}}\frac{\delta(\varphi+su;v)^*(y)-\delta(\varphi;v)^*(y)}{s}f_Y(y)dy, \tag{B.1}$$

where $s\downarrow 0$ means $s\to 0$ and $s>0$.

To do so, we need to consider the map

$$\lim_{s\downarrow 0}\frac{\delta(\varphi+su;v)^*(y_0)-\delta(\varphi;v)^*(y_0)}{s}.$$

Recall that $\delta(\varphi;v)^*(y_0)=v(D\varphi^*(y_0))$, so that $\delta(\varphi+su;v)^*(y_0)=v(x_{s,u})$, for some $x_{s,u}\in\partial(\varphi+su)^*(y_0)$. Now, let us work from the inside out. Let us first show that

$$\lim_{s\downarrow 0}\frac{\partial(\varphi+su)^*(y_0)-D\varphi^*(y_0)}{s}=Du(D\varphi^*(y_0)). \tag{B.2}$$

Note that since u need not be convex, we can only work with the subgradient $\partial(\varphi+su)^*$ even though $u\in C^1(\mathcal{X})$, since the gradient $D(\varphi+su)(x)$ is not invertible in general. In this respect, it is important to note, however, that $(\varphi+su)^*$ is subdifferentiable, for all s, since su is differentiable everywhere and φ, being strictly convex, is differentiable almost everywhere and subdifferentiable everywhere. Therefore, the inverse of the subgradient, which coincides with the subgradient of the convex conjugate, is always nonempty.

Now, to show (B.2), note that

$$\begin{aligned}\lim_{s\downarrow 0}\frac{\partial(\varphi+su)^*(y_0)-D\varphi^*(y_0)}{s}&=\lim_{s\downarrow 0}\partial\frac{1}{s}(\varphi+su)^*(y_0)-D\frac{1}{s}\varphi^*(y_0)\\&=\lim_{s\downarrow 0}\partial_C\left(\frac{1}{s}\left[(\varphi+su)^*-\varphi^*\right]\right)(y_0),\end{aligned} \tag{B.3}$$

where $\partial_C f(x)$ is the generalized gradient for locally Lipschitz continuous functions in the sense of Clarke (Clarke, 1975), which is the convex hull of the set of limits of the form $\lim_{i\to\infty}Df(x+h_i)$, where $h_i\to 0$ as $i\to\infty$. Note that we can use this simple form of the definition of the generalized gradient since the functions

$$f_s:=\frac{1}{s}\left[(\varphi+su)^*-\varphi^*\right]$$

are all locally Lipschitz continuous, because $(\varphi+su)^*$ and φ^* are convex and hence locally Lipschitz continuous. In particular, the functions f_s possess a gradient almost everywhere, i.e., $Df_s(y)$ is defined for almost every $y\in\mathcal{Y}$ and every s (Clarke, 1975; Hiriart-Urruty, 1985). The first equality in (B.3) follows by the positive homogeneity of the subdifferential and the fact that $s>0$. The second equality—which is in terms of sets—follows from the property of the Clarke subdifferential of difference convex functions which states that $\partial f_1(x)-\partial f_2(x)=\partial_C(f_1-f_2)(x)$ for convex functions f_1,f_2 (Bačák and Borwein, 2011; Hiriart-Urruty, 1985). In order to show (B.2), we therefore need to show that we can interchange the limit and the Clarke subdifferential in (B.3). Recall from our proof of the first variation that

$$\lim_{s\downarrow 0}f_s(y_0)\equiv\lim_{s\downarrow 0}\frac{(\varphi+su)^*(y_0)-\varphi^*(y_0)}{s}=u(D\varphi^*(y_0)).$$

Note that the sequence $\{f_s\}_{s\downarrow 0}$ is an equi-Lipschitzian sequence on the compact $\mathcal{Y}$. This follows from the fact that f_s is locally Lipschitz for all s, so that f_s is Lipschitz continuous on

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

all of the compact $\mathcal{Y}$ for every s; furthermore, f_s converges to $u \circ D\varphi^*$, which is a composition of two (almost everywhere) differentiable and hence Lipschitz continuous functions, so that there must exist a Lipschitz constant which holds uniformly for all f_s. In particular, the functions f_s converge uniformly to $u \circ D\varphi$ for almost every $y \in \mathcal{Y}$ by Lipschitz continuity and compactness of $\mathcal{Y}$.

The generalized directional derivative in the sense of Clarke (Clarke, 1975) $u^\circ(D\varphi^*(y_0), D\varphi^*(y'))$ of u at $D\varphi^*(y_0)$ in direction $D\varphi^*(y')$ takes the form

$$u^\circ(D\varphi^*(y_0), D\varphi^*(y')) := \limsup_{\substack{\lambda \downarrow 0 \\ h \to 0}} \frac{u(D\varphi^*(y_0) + h + \lambda D\varphi^*(y')) - u(D\varphi^*(y_0) + h)}{\lambda},$$

for some h, in the same space as $D\varphi^*(y_0)$. In particular, note that this generalized directional derivative coincides with the standard directional derivative

$$\begin{aligned} u'(D\varphi^*(y_0), D\varphi^*(y')) &= \lim_{\lambda \downarrow 0} \frac{u(D\varphi^*(y_0) + \lambda D\varphi^*(y')) - u(D\varphi^*(y_0))}{\lambda} \\ &= \langle Du(D\varphi^*(y_0)), D\varphi^*(y') \rangle \end{aligned}$$

at almost every $y_0 \in \mathcal{Y}$ since $u \in C^1(\mathcal{X})$ and $D\varphi^*(y)$ is differentiable at almost every $y \in \mathcal{Y}$ by a result from Aleksandrov (1939). Furthermore, by Proposition 1.11 in Clarke (1975), the generalized directional derivative defined on $\mathcal{Y} \setminus N$ for some set $N \subset \mathcal{Y}$ of measure zero coincides with the generalized directional derivative on all of $\mathcal{Y}$, so that we can assume that y_0 and y' are points where $D\varphi^*$ exists and is a single element.

Now, given any $m > u^\circ(D\varphi^*(y_0), D\varphi^*(y'))$, there exists a $\lambda > 0$ and a small enough h such that $D\varphi^*(y_0) + h + \lambda D\varphi^*(y') \in \mathcal{Y}$ by convexity of $\mathcal{Y}$ and

$$[u(D\varphi^*(y_0) + h + \lambda D\varphi^*(y')) - u(D\varphi^*(y_0) + h)] \frac{1}{\lambda} < m.$$

By the uniform convergence of f_s to u, which is a result of the Arzelà–Ascoli theorem and the compact support, it holds that

$$[f_s(y_s + h + \lambda y'_s) - f_s(y_s + h)] \frac{1}{\lambda} < m,$$

for $\mathcal{Y} \ni (y_s, y'_s) \to (y_0, y')$ and sufficiently small s. By the definition of $f_s^\circ(y_s; y'_s)$, it therefore holds that there exists a small enough s such that

$$f_s^\circ(y_s; y'_s) < m.$$

Taking the limit supremum with respect to this expression implies that

$$\limsup_{s \downarrow 0} f_s^\circ(y_s; y'_s) < m.$$

We may, by putting $y'_s \equiv y'$ for all s, conclude from this (Rockafellar, 1997, p. 234) that

$$\limsup_{s \downarrow 0} f_s^\circ(y_s; y') \leq u^\circ(D\varphi^*(y_0), D\varphi^*(y')) \qquad \forall y' \in \mathcal{Y},$$

which follows from the definition of u°. Note that the generalized directional derivative $f_s^\circ(y_s; y'_s)$ is the support function of the generalized gradient $\partial_C f_s(y_s)$ in the sense that

$$f_s^\circ(y_s; y'_s) = \max\{\langle \xi, y'_s \rangle : \xi \in \partial_C f_s(y_s)\}, \tag{B.4}$$

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

see Clarke (1975, Prop. 1.4), analogously for $u^\circ(D\varphi^*(y_0); D\varphi^*(y'))$ and $\partial_C u(D\varphi^*(y_0))$. Therefore, and by uniform convergence of f_s to u, it follows directly by (B.4) that, for every $\varepsilon > 0$, there exists a small enough s such that

$$f_s(y_0) \le u(D\varphi^*(y_0)) + \varepsilon|y_0|$$

by positive homogeneity of f_s and u and since $\mathcal{Y}$ is compact. But since the generalized derivative is the support function of the generalized gradient, it follows that

$$\partial_C f_s(y_0) \subset \partial_C u(D\varphi^*(y_0)) + \varepsilon B,$$

where B denotes the unit ball of the same dimension as $\mathcal{Y}$, see Rockafellar (1997, Thm. 24.5); this holds for every direction $y'_s \to y' \in \mathcal{Y}$ by letting $\varepsilon \to 0$.[13] But this implies that we can interchange the limit and the generalized gradient in (B.3) to obtain (B.2). Furthermore, since u is differentiable at $x_0 = D\varphi^*(y_0)$, the generalized gradient becomes a gradient by Proposition 1.13 in Clarke (1975).

We can therefore write

$$\partial(\varphi + su)^*(y_0) = D\varphi^*(y_0) + Du(D\varphi^*(y_0))s + o(s). \tag{B.5}$$

Now, define

$$x_0 + k := \partial(\varphi + su)^*(y_0) = D\varphi^*(y_0) + Du(D\varphi^*(y_0))s + o(s),$$

so that $k := Du(D\varphi^*(y_0))s + o(s)$ since $x_0 = D\varphi^*(y_0)$. Recall that $v \in C^1(\mathcal{X})$ by assumption so that

$$v(x_0 + k) = v(x_0) + \langle Dv(x_0), k\rangle + o(\|k\|),$$

and replacing $x_0 + k = \partial(\varphi + su)^*(y)$ and k by the above gives

$$v(x_{s,u}) = v(x_0) + \langle Dv(x_0), Du(D\varphi^*(y_0))s + o(s)\rangle + o(s),$$

since the inner product is continuous. In other words,

$$\lim_{s\downarrow 0} \frac{\delta(\varphi + su; v)^*(y) - \delta(\varphi; v)^*(y)}{s} = \langle Dv(D\varphi^*(y_0)), Du(D\varphi^*(y_0))\rangle,$$

which is what we wanted to show. Now, since $v \in C^1(\mathcal{X})$ and is hence uniformly bounded on $\mathcal{X}$, we can apply the Dominated Convergence Theorem to conclude that

$$\begin{aligned}
\lim_{s\downarrow 0} \frac{\delta M_{\varphi+su}(v) - \delta M_\varphi(v)}{s} &= \lim_{s\downarrow 0} \int_{\mathcal{Y}} \frac{\delta(\varphi + su; v)^*(y) - \delta(\varphi; v)^*(y)}{s} f_Y(y)dy \\
&= \int_{\mathcal{Y}} \lim_{s\downarrow 0} \frac{\delta(\varphi + su; v)^*(y) - \delta(\varphi; v)^*(y)}{s} f_Y(y)dy \\
&= \int_{\mathcal{Y}} \langle Dv(D\varphi^*(y)), Du(D\varphi^*(y))\rangle f_Y(y)dy \\
&= \int_{\mathcal{X}} \langle Dv(x), Du(x)\rangle f_Y(D\varphi^{**}(x)) \det(D^2\varphi^{**}(x))dx,
\end{aligned}$$

[13] This part of the argument is analogous to the argument in the proof of Theorem 24.5 in Rockafellar (1997), only for generalized gradients and not the convex case.

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

where the last line follows from the change of variables $y = D\varphi^{**}(x)$. This shows the formula for the second variation in directions $u, v \in C(\mathcal{X})$.

Let us now turn to the statements about the Fréchet derivative and the continuity of the second variation in φ on some neighborhood $\mathcal{N}_{\varphi_0}$. For this, we need to assume that $\varphi, v, u \in C^2(\mathcal{X})$. We can deal with both the first and second variation simultaneously. Now, in order for the Fréchet derivative to exist, $\delta M_\varphi(v)$ must exist, for all $\varphi \in C^2(K)$, in a neighborhood around the minimal $\varphi_0 \in C^2(K)$, for any compact $K \subset \mathcal{X}^\circ$, and must be continuous in φ (Zeidler, 1985, p. 192). The same holds for continuity of the second variation. Note that neither the first nor second variation is continuous in any φ everywhere on $C^2(\mathcal{X})$ in general, since the double convex conjugate φ^{**} of some function $\varphi \in C^2(\mathcal{X})$ lies in $C^{1,1}(\mathcal{X})$, but does not possess higher regularity properties in general (Griewank and Rabier, 1990; Kirchheim and Kristensen, 2001). In particular, we do need to exploit the (strict) convexity of φ_0.

The key to proving continuity of $\delta M_{\varphi_0}(v)$ and $\delta^2 M_{\varphi_0}(u, v)$ in a neighborhood around φ_0 is therefore to work with the Hessian or *Monge–Ampère* measure (Trudinger and Wang, 2008, Sect. 2.2). In fact, recall that for any differentiable convex function f at a point x, its gradient Df coincides with the Normal map at this point. Therefore, we can, for general and not necessarily differentiable convex functions f, define a measure via the Normal mapping $N_f(x)$

$$N_f(x) := \{y \in \mathcal{Y} : y \text{ is the gradient of a local supporting function of } f \text{ at } x\}.$$

For any subset $S \subset \mathcal{X}$, we then define $N_f(S) = \bigcup_{x \in S} N_f(x)$. The Monge–Ampère measure μ_f is then defined on the Borel σ-algebra on $\mathcal{X}$ as

$$\mu_f(S) := \text{Leb}(N_f(S)) \qquad \text{for all Borel sets } S \subset \mathcal{X},$$

where Leb(S) denotes Lebesgue measure of the Borel set S. It is a standard result that the Monge–Ampère measure is an actual measure (Trudinger and Wang, 2008, Sect. 2.2). If the function $f \in C^2(\mathcal{X})$, then

$$\mu_f(S) = \text{Leb}(Df(S)) = \int_S \det(D^2 f(x))dx.$$

Applying this to our setting, we set $f \equiv \varphi^{**}$. Now, fix some compact $K \subset \mathcal{X}^\circ$ and pick a sequence $\{\varphi_n\}_{n \in \mathbb{N}} \in C^2(\mathcal{X})$ which converges to φ_0 in the topology induced by the Hölder norm $\|\cdot\|_{C^2(K)}$. From this, it follows directly that $\varphi_n^{**} \in C^{1,1}(K)$, for every n (Kirchheim and Kristensen, 2001). Now, we want to prove that this implies that $\varphi_n^{**} \to \varphi_0$ uniformly on K. For this, notice that since φ_n^{**} are convex functions, they converge uniformly on K to some convex function f on K if they converge pointwise (Rockafellar, 1997, Thm. 10.8). We therefore need to show that this f must coincide with φ_0. For this, note that since $\|\varphi_n - \varphi_0\|_{C^2(K)} \to 0$, it holds that, for every point $x \in K$, there exists a large enough n_x such that the functions φ_n for $n \geq n_x$ satisfy

$$\varphi_n(x') > \varphi_n(x) + \langle D\varphi_n(x), x' - x\rangle$$

by strict convexity of φ_0. Now, by definition of the closed convex envelope of a function, i.e., the double convex conjugate, it holds that

$$\varphi_n^{**}(x) = \sup\{\langle s, x\rangle - b : \langle s, x'\rangle - b \leq \varphi_n(x') \quad \text{for all } x' \in \mathcal{X}, \text{with} b \in \mathbb{R}, s \in \mathcal{Y}\},$$

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

which directly implies that $\varphi_n^{**}(x)$ must coincide with $\varphi_n(x)$ for those points x, where φ_n is strictly convex. As $n \to \infty$, φ^{**} must therefore converge pointwise and by convexity also uniformly to the strictly convex φ_0.

Now, note that $\varphi_n^{**} \to \varphi_0$ uniformly on $K \subset \mathcal{X}^\circ$ implies that $\mu_{\varphi_n^{**}} \to \mu_{\varphi_0}$ weakly (Trudinger and Wang, 2008, Lem. 2.2). Therefore, since $v, u \in C^2(\mathcal{X})$ and f_Y and f_X are at least in $C^2(\mathcal{Y})$ and $C^2(\mathcal{X})$ by Assumption 1, it follows by the definition of weak convergence that both

$$\int_K v(x) f_Y(D\varphi_n^{**}(x)) \det(D^2\varphi_n^{**}(x))dx \to \int_K v(x) f_Y(D\varphi_0(x)) \det(D^2\varphi_0(x))dx \qquad \text{and}$$

$$\int_K \langle Dv(x), Du(x)\rangle f_Y(D\varphi_n^{**}(x)) \det(D^2\varphi_n^{**}(x))dx$$

$$\to \int_K \langle Dv(x), Du(x)\rangle f_Y(D\varphi_0(x)) \det(D^2\varphi_0(x))dx,$$

which implies that the second variation is continuous in φ in a neighborhood around φ_0, so that the Hadamard derivative $M'(\varphi_0)$ coincides with the Fréchet derivative (Zeidler, 1985, p. 192). In particular, setting $v \equiv u$, it follows from this that, for every $\varepsilon > 0$, there exists some $\eta(\varepsilon) > 0$ such that

$$\begin{aligned}
&\left|\delta^2 M_\varphi(v,v) - \delta^2 M_{\varphi_0}(v,v)\right| \\
&\quad \le \int_K \left| \|Dv(x)\|^2 \left[f_Y(D\varphi^{**}(x)) \det(D^2\varphi^{**}(x)) - f_Y(D\varphi_0(x)) \det(D^2\varphi_0(x)) \right] \right| dx \\
&\quad \le \int_K \left| f_Y(D\varphi^{**}(x)) \det(D^2\varphi^{**}(x)) - f_Y(D\varphi_0(x)) \det(D^2\varphi_0(x)) \right| dx \|Dv(x)\|_{\infty,K}^2 \\
&\quad \le \varepsilon \|v\|_{C^2(K)}^2 \qquad \text{for } \|\varphi - \varphi_0\|_{C^2(K)} \le \eta(\varepsilon),
\end{aligned}$$

where $\|\cdot\|_{\infty,K}$ denotes the supremum norm on K. The third line follows from Hölder's inequality and the fact that $\sup_x f^2(x) = \left(\sup_x f(x)\right)^2$ for a nonnegative function f. The fourth line follows from the definition of the Hölder norm and from the sequence definition of the continuity of the second variation in φ we have shown above.

Finally, we can show that the second variation is a bounded bilinear functional. Since $\mathcal{X}$ is compact and P_X a probability measure, it follows that $v \in L^2(P_X)$ by Hölder's inequality, as $v \in C^2(\mathcal{X})$ and hence $\|v\|_\infty < +\infty$. Linearity of the second variation in both arguments v follows from the fact that the derivative D is a linear operator and that the inner product is linear in both arguments. To see continuity of the second variation, note that we can write $\delta^2 M_{\varphi_0}(v,v) = \|Dv\|^2_{L^2(f_Y(D\varphi_0(x)) \det(D^2\varphi_0(x))dx)}$.

Now, since the measure $f_Y(D\varphi_0(x)) \det(D^2\varphi_0(x))dx$ is a probability measure, it holds that $\|Dv\|_{L^2(f_Y(D\varphi_0(x)) \det(D^2\varphi_0(x))dx)} \le \|Dv\|_\infty$. Moreover, note that $D : C^2(K) \to C^1(K)$ is a bounded linear operator by the definition of the space $C^2(\mathcal{X})$, so that $\|Dv\|_\infty \le c\|v\|_\infty$, for all $v \in C^2(K)$ and some $0 < c < +\infty$. This implies that $\|Dv\|_{L^2(f_Y(D\varphi_0(x)) \det(D^2\varphi_0(x))dx)} \le c\|v\|_\infty$, which shows that $\delta^2 M_{\varphi_0}(v,v)$ is a continuous linear operator on $C^2(\mathcal{X}^\circ)$. The result for the boundedness of the bilinear operator $\delta^2 M_{\varphi_0}(v,u)$ follows from an analogous argument, only using the operator norm $\|\delta^2 M_{\varphi_0}\|_B = \inf\{c > 0 : |\delta^2 M_{\varphi_0}(u,v)| \le c\|u\|_\infty \|v\|_\infty\}$ for this bilinear functional. ■

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

B.2. Proof of Theorem 1.

We split its proof into several lemmas to make it easier to follow. The structure of the proof is as follows. Lemma 1 is the core of the proof, as it provides the regularity properties of the Monge–Kantorovich problem. Lemma 2 below uses Caffarelli's regularity theory to connect the smoothness of the potential function φ to the smoothness of the respective density functions. Lemma 3 below provides a connection between the convergence properties of the density functions and the potential function of the Brenier map, proving convergence in Hölder norm of $\hat{\varphi}_n$ to φ_0. Lemma 4 below lets us bound the L^2 norm of the difference of the convex conjugates of the potential function of the Brenier map and another strictly convex function by the difference of these functions themselves. We need this to bound the L^2 bracketing numbers of $\varphi^* - \varphi_1^*$ by the L^2 bracketing numbers of $\varphi_0 - \varphi_1$.

Finally, the main proof of the theorem consists of checking the stochastic equicontinuity condition and the local identification condition that the optimum is well-separated required in Theorem 3.2.5 of van der Vaart and Wellner (2013). The local identification condition follows from the Poincaré inequality and the properties of the Monge–Kantorovich problem proved in Lemma 1. The equicontinuity follows from the lemmas and the same proof as Theorem 3.2.5 in van der Vaart and Wellner (2013).

Let us also introduce some notation. The symbol $\lesssim$ means "less than or equal to a constant multiple of." We mostly work in subspaces of $L^p(P)$ or C, which are AL or AM spaces, respectively, with partial orders $g \leq f$ if $g(x) \leq f(x)$, for P-almost every x, and $g \leq f$ if $g(x) \leq f(x)$, for every x, respectively. Based on this, we can define brackets $[f_L; f_U]$ in the sense that $f \in [f_L; f_U]$ if $f_L \leq f \leq f_U$ in the respective partial order. An ε-bracket is a bracket $[f_L; f_U]$ with $\|f_L - f_U\| < \varepsilon$. The bracketing number $N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)$ is the minimal number of brackets of size $\varepsilon > 0$ which cover $\mathcal{F}$. The letters c, c' denote general constants, which can represent different values in different inequalities, $\lfloor a \rfloor$ denotes the largest integer smaller or equal to a, and $\lceil a \rceil$ denotes the smallest integer larger or equal to a.

As discussed in the main text, the optimal solution (φ_0, ψ_0) to the semi-dual problem

$$\min_{\varphi,\psi} \int_{\mathcal{X}} \varphi(x) f_X(x) dx + \int_{\mathcal{Y}} \psi(y) f_Y(y) dy$$
$$\text{s.t.} \quad \varphi(x) + \psi(y) \leq \|x - y\|^2, \quad \varphi \in C(\mathcal{X}), \psi \in C(\mathcal{Y})$$

can be shown to always be convex conjugates, i.e., the optimal solution can be written as (φ_0, φ_0^*), see the first step in the proof of Proposition 3.1 in Brenier (1991); we can therefore write the objective function as

$$M(\varphi) = \int_{\mathcal{X}} \varphi(x) f_X(x) dx + \int_{\mathcal{Y}} \varphi^*(y) f_Y(y) dy. \tag{B.6}$$

One key step in the proof is to connect the regularity of the densities f_X and f_Y to the regularity of the potential function φ. We do this by using the interior Schauder estimates proved in the seminal article Caffarelli (1990), which we state in the following.

LEMMA 2 (Regularity of φ). *Under Assumption 1, the potential function of the Brenier map, φ, between P_X and P_Y lies in $C^{s+2,\alpha}_{loc}(\mathcal{X}^\circ)$. Moreover,*

$$D\varphi^* = (D\varphi)^{-1}$$

and φ^* as a mapping between P_Y and P_X lies in $C^{s+2}_{loc}(\mathcal{Y}^\circ)$.

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

Proof. Assumption 1 guarantees that $f_X \in C^{s,\alpha}_{loc}(\mathcal{X}^\circ)$ and $f_Y \in C^{s,\alpha}_{loc}(\mathcal{Y}^\circ)$, for every $\alpha \in (0,1)$. Based on this, it follows directly from Theorem 2 in Caffarelli (1990) that $\varphi \in C^{s+2,\alpha}_{loc}(\mathcal{X}^\circ)$. Moreover, since both P_X and P_Y are absolutely continuous, it holds that

$$D\varphi^* \circ D\varphi(x) = x \quad \text{and} \quad D\varphi \circ D\varphi^*(y) = y,$$

i.e., both φ and φ^* are strictly convex functions, see Villani (2003, Thm. 2.12(iv)). From this and the fact that φ lies in $C^{s+2}_{loc}(\mathcal{X}^\circ)$, it follows from the inverse function theorem that φ^* is at least in $C^{s+2}_{loc}(\mathcal{Y}^\circ)$ (Rockafellar, 1977). ■

Another crucial step in the proof of Theorem 1 will be to show that the potentials of the empirical Brenier map, $\hat{\varphi}_n$, converge to their population counterpart φ_0 in Hölder norm. For this, we exploit recent ideas from the regularity theory of the Monge–Ampère equation, especially from de Philippis and Figalli (2013), as well as standard Schauder estimates. We do this in the next lemma.

LEMMA 3. *If* $\|\hat{f}^X_{h_n} - f_X\|_{C^{s,\alpha}_{loc}(\mathcal{X}^\circ)} \to 0$ *and* $\|\hat{f}^Y_{h'_n} - f_Y\|_{C^{s,\alpha}_{loc}(\mathcal{Y}^\circ)} \to 0$ *and the respective supports of* $\hat{f}^X_n$ *and* $\hat{f}^Y_n$ *converge in the Hausdorff metric to* $\mathcal{X}$ *and* $\mathcal{Y}$ *as* $n \to \infty$, *then*

$$\|\hat{\varphi}_n - \varphi_0\|_{C^{s+2,\alpha}_{loc}(\mathcal{X}^\circ)} \to 0 \qquad \text{as } n \to \infty.$$

Proof. We prove the result for $s = 0$, as higher orders follow straightforwardly from the "bootstrapping" procedure for uniformly elliptic second-order PDEs (Gilbarg and Trudinger, 1998, p. 4 and Chap. 8). We first want to show that

$$\left\| \frac{\hat{f}^X_{h_n}}{\hat{f}^Y_{h'_n} \circ D\hat{\varphi}_n} - \frac{f_X}{f_Y \circ D\varphi_0} \right\|_{C^{\alpha}_{loc}(\mathcal{X}^\circ)} \to 0 \qquad \text{as } n \to \infty,$$

which will turn out to be the right-hand side of the Monge–Ampère equation for the optimal transport problem. To see this, define

$$g(x) := \frac{f_X(x)}{f_Y(D\varphi_0(x))} \qquad \text{and} \qquad \hat{g}_n(x) := \frac{\hat{f}^X_{h_n}(x)}{\hat{f}^Y_{h'_n}(D\hat{\varphi}_n(x))},$$

and both $\hat{g}_n$ and g are in $C^{\alpha}_{loc}(\mathcal{X}^\circ)$, since all densities are bounded below and above on their support. From standard stability theorems on the Brenier map $D\varphi$ (Proposition 50 in Lindsey and Rubinstein (2017) Lindsey and Rubinstein (2017) and Corollary 5.23 in Villani (2003) Villani (2008)), it follows that $D\hat{\varphi}_n$ converges to $D\varphi_0$ in measure inside $\mathcal{X}$. In particular, by Cafarelli's regularity theory (Lemma 2), $D\hat{\varphi}_n$ are locally uniformly Hölder continuous maps, so they converge locally uniformly to $D\varphi_0$ (de Philippis and Figalli, 2013, p. 999). This implies that $\hat{f}^Y_{h'_n} \circ D\hat{\varphi}_n \to f_Y \circ D\varphi_0$ in $C^{\alpha}_{loc}(\mathcal{X}^\circ)$. Also, $\|\hat{f}^X_{h_n} - f_X\|_{C^{\alpha}_{loc}(\mathcal{X}^\circ)} \to 0$. This implies $\|\hat{g}_n - g\|_{C^{\alpha}_{loc}(\mathcal{X}^\circ)}$, as required.

Recall that $\hat{\varphi}_n$ and φ_0 are all *strictly* convex since the densities are all absolutely continuous with respect to Lebesgue measure. Therefore, we can use the exact same reasoning as in the proof of Theorem 1.2 in de Philippis and Figalli (2013) to write the Monge–Ampère equation for this optimal transport problem. To do so, we first fix some $x_0 \in \mathcal{X}^\circ$ and some radius $r > 0$ such that the ball $B_r(x_0)$ of radius r around x_0 lies in $\mathcal{X}^\circ$. By the strict convexity of φ_0, we can find a linear function $l(z) := a \cdot z + b$ such that the open

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

set $Z := \{z : \varphi_0(z) < \varphi_0(x_0) + l(z)\}$ is nonempty and compactly supported inside $B_{r/2}(x_0)$. Therefore, by the uniform convergence of $\hat{\varphi}_n$ to φ_0 (which follows from a normalization $\hat{\varphi}_n(x_0) = \varphi_0(x_0)$ and the fact that $D\hat{\varphi}_n$ converges to $D\varphi_0$ locally uniformly in conjunction with the convexity of φ_0 and $\hat{\varphi}_n$) and the fact that the gradient $D\varphi_0$ is normal to l on ∂Z, it holds that $Z_n := \{z : \varphi_0(z) < \varphi_0(x_0) + l(z)\}$ are nonempty convex sets which converge in the Hausdorff distance to Z. Moreover, the maps $w_n := \hat{\varphi}_n - l$ solve the following Dirichlet problem in the Aleksandrov sense:

$$\begin{cases} \det(D^2 w_n) = \hat{g}_n & \text{in } Z_n \\ w_n = 0 & \text{on } \partial Z_n. \end{cases}$$

Analogously, the map $w := \varphi_0 - l$ solves the Dirichlet problem

$$\begin{cases} \det(D^2 w) = g & \text{in } Z \\ w = 0 & \text{on } \partial Z. \end{cases}$$

Importantly, since l is linear, the Monge–Ampère measures[14] of φ and w coincide.

We can therefore subtract a scaled version, by some fixed $\varepsilon \in (0, 1)$, of the first equation from the second equation and define $v_n := w - \varepsilon w_n$. This is not a convex function in general. We therefore work with the convex envelope of $w - (1-\varepsilon)w_n$, which we denote by $\Gamma_{w-(1-\varepsilon)w_n}$.[15] Since w_n converges to the strictly convex w locally uniformly by the fact that $D\hat{\varphi}_n$ converges to $D\varphi_0$ locally uniformly, it holds that $\Gamma_{w-(1-\varepsilon)w_n}$ converges uniformly to $\Gamma_{\varepsilon w}$ inside a compact $K \subset \mathcal{X}^\circ$. Furthermore, for large enough n, the convex envelope $\Gamma_{w-(1-\varepsilon)w_n}$ almost coincides with εw, since the latter is strictly convex. Therefore, the Monge–Ampère measure of $v_n = w - (1-\varepsilon)w_n$ is uniformly elliptic away from the boundary of its support and $v_n \in C^{s,\alpha}_{loc}(\mathcal{X})$ by an application of Caffarelli's regularity theory using Lemma 2. We can hence apply standard Schauder estimates for uniformly elliptic PDEs.

In particular, by Theorem 6.2 in Gilbarg and Trudinger (1998), the following estimate holds:

$$\|v_n\|_{C^{2,\alpha}(B_{r/2}(x_0))} \leq c(\|g - \varepsilon \hat{g}_n\|_{C^\alpha(B_r(x_0))} + \|v_n\|_{L^\infty(B_r(x_0))}),$$

for some constant $0 < c < +\infty$. Now, the first term on the right-hand side converges to zero on $\mathcal{X}^\circ$ if we let $\varepsilon \to 0$ from what we have shown above. The second term also converges to zero if we let $\varepsilon \to 0$ by the fact that $D\hat{\varphi}_n$ converges uniformly to $D\varphi_0$. Since x_0 was arbitrary, this shows that

$$\|\hat{\varphi}_n - \varphi_0\|_{C^{2,\alpha}_{loc}(\mathcal{X}^\circ)} \to 0.$$

■

In addition, when deriving the degree of continuity of the objective function, we frequently need to relate the distances $\|\psi^* - \varphi^*\|_{L^p(P_Y)}$ to $\|\psi - \varphi\|_{L^p(P_X)}$, where ψ is some strictly convex function and φ is the potential function of the Brenier map transporting P_X onto P_Y. For this, we need the following lemma which lets us bound differences of convex conjugates above by their double convex conjugates in the L^p norm, $1 \leq p \leq +\infty$.

[14] For a definition of Monge–Ampère measures, see the proof of Lemma 1.

[15] We defined the convex envelope in the proof of Lemma 1 as the double convex conjugate of a function.

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

LEMMA 4. *Let $\psi : \mathbb{R}^d \to \mathbb{R}$ be a proper strictly convex function. If φ is the potential function of the Brenier map transporting P_X onto P_Y, where both P_X and P_Y are absolutely continuous with respect to Lebesgue measure, then it holds that*

$$\|\psi^* - \varphi^*\|_{L^p(P_Y)} \leq C\|\psi - \varphi\|_{L^p(P_X)}, \tag{B.7}$$

for $p \in [1, \infty]$ and a constant $C < +\infty$. In the cases $p = 1$ and $p = +\infty$, $C = 1$.

For the supremum norm, $C = 1$ readily follows from the definition of convex conjugates for general proper convex functions. We prove a similar result for the L^p norm, $1 \leq p < \infty$, in order to work with φ and not its convex conjugate φ^* which is more difficult to handle in our setting. Stern (2010) provides an expression for C in the case $p \in (1, \infty)$ in terms of the Jacobian.

Proof. We define the set $H := \left\{x \in \mathcal{X} : \psi^*(D\varphi(x)) > \varphi^*(D\varphi(x))\right\}$ and write

$$\begin{aligned}
\|\psi^* - \varphi^*\|^p_{L^p(P_Y)} &= \int |\psi^*(y) - \varphi^*(y)|^p dP_Y(y) \\
&= \int |\psi^*(D\varphi(x)) - \varphi^*(D\varphi(x))|^p dP_X(x) \\
&= -\int_H \left[\varphi^*(D\varphi(x)) - \psi^*(D\varphi(x))\right]^p dP_X(x) \\
&\quad - \int_{\mathcal{X}\setminus H} \left[\psi^*(D\varphi(x)) - \varphi^*(D\varphi(x))\right]^p dP_X(x),
\end{aligned}$$

where the first line follows from the fact that the Brenier map $D\varphi$ transports P_X onto P_Y. For the following, it is important to note that both integrals are nonpositive. Now, work with each term separately. We have

$$\begin{aligned}
0 \leq & -\int_H \left[\varphi^*(D\varphi(x)) - \psi^*(D\varphi(x))\right]^p dP_X(x) \\
= & -\int_H \left[\langle x, D\varphi(x)\rangle - \varphi(x) - \langle (D\psi)^{-1}(D\varphi(x)), D\varphi(x)\rangle + \psi\left(D\psi^{-1}(D\varphi(x))\right)\right]^p dP_X(x) \\
= & -\int_H \left[\langle x - (D\psi)^{-1}(D\varphi(x)), D\varphi(x)\rangle - \varphi(x) + \psi\left(D\psi^{-1}(D\varphi(x))\right)\right]^p dP_X(x) \\
= & \int_H \left[\langle (D\psi)^{-1}(D\varphi(x)) - x, D\varphi(x)\rangle + \varphi(x) - \psi\left(D\psi^{-1}(D\varphi(x))\right)\right]^p dP_X(x),
\end{aligned}$$

where the second line follows from the representation of the convex conjugate f^* of a differentiable function f, which reads

$$f^*(y) = \langle (Df)^{-1}(y), y\rangle - f\left((Df)^{-1}(y)\right),$$

see Rockafellar (1997, p. 256). Replacing f by φ and writing $y = D\varphi(x)$, we obtain

$$\varphi^*(D\varphi(x)) = \langle x, D\varphi(x)\rangle - \varphi(x);$$

analogously, replacing f by ψ, we obtain

$$\psi^*(D\varphi(x)) = \langle (D\psi)^{-1}(D\varphi(x)), D\varphi(x)\rangle - \psi\left((D\psi)^{-1}(D\varphi(x))\right).$$

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

Now, we make use of the cyclic monotonicity of the (sub)gradient of a convex function; in fact, if f is a proper convex function, then its gradient is cyclically monotone in the sense that for every set of points $\{x_i\}_{i=1,\dots,n}$ it holds that

$$\langle x_i - x_{i-1}, Df(x_{i-1})\rangle \leq f(x_i) - f(x_{i-1}),$$

see Rockafellar (1997, p. 238). Therefore,

$$\begin{aligned}&\int_H \left[\langle (D\psi)^{-1}(D\varphi(x)) - x, D\varphi(x)\rangle + \varphi(x) - \psi\left(D\psi^{-1}(D\varphi(x))\right)\right]^p dP_X(x)\\ &\quad\leq \int_H \left[\varphi\left(D\psi^{-1}(D\varphi(x))\right) - \psi\left(D\psi^{-1}(D\varphi(x))\right)\right]^p dP_X(x),\end{aligned}$$

since the integral is nonnegative. Analogously, we can write for the second term

$$\begin{aligned}0 &\leq -\int_{\mathcal{X}\setminus H}\left[\psi^*(D\varphi(x)) - \varphi^*(D\varphi(x))\right]^p dP_X(x)\\ &\quad -\int_{\mathcal{X}\setminus H}\left[\langle (D\psi)^{-1}(D\varphi(x)), D\varphi(x)\rangle - \psi\left(D\psi^{-1}(D\varphi(x))\right) - \langle x, D\varphi(x)\rangle + \varphi(x)\right]^p\\ &\qquad dP_X(x)\\ &= -\int_{\mathcal{X}\setminus H}\left[\langle (D\psi)^{-1}(D\varphi(x)) - x, D\varphi(x)\rangle + \varphi(x) - \psi\left(D\psi^{-1}(D\varphi(x))\right)\right]^p dP_X(x)\\ &= \int_{\mathcal{X}\setminus H}\left[\psi\left(D\psi^{-1}(D\varphi(x))\right) - \langle (D\psi)^{-1}(D\varphi(x)) - x, D\varphi(x)\rangle - \varphi(x)\right]^p dP_X(x),\end{aligned}$$

where the second line again follows from the representation of the convex conjugate of a differentiable function if we replace f by ψ and write $y = D\varphi(x)$ in the above formula. Now, by cyclic monotonicity just as above, we can write

$$\begin{aligned}&\int_{\mathcal{X}\setminus H}\left[\psi\left(D\psi^{-1}(D\varphi(x))\right) - \langle (D\psi)^{-1}(D\varphi(x)) - x, D\varphi(x)\rangle - \varphi(x)\right]^p dP_X(x)\\ &\quad\leq \int_{\mathcal{X}\setminus H}\left[\psi\left(D\psi^{-1}(D\varphi(x))\right) - \varphi\left(D\psi^{-1}(D\varphi(x))\right)\right]^p dP_X(x),\end{aligned}$$

since the integral is nonnegative.

Putting both terms together again, we therefore have

$$\begin{aligned}&\int |\psi^*(D\varphi(x)) - \varphi^*(D\varphi(x))|^p dP_X(x)\\ &\quad\leq \int \left|\varphi\left(D\psi^{-1}(D\varphi(x))\right) - \psi\left(D\psi^{-1}(D\varphi(x))\right)\right|^p dP_X(x).\end{aligned}$$

Now, note that we can write the last line as

$$\begin{aligned}&\int \left|\varphi\left((D\psi)^{-1}(D\varphi(x))\right) - \psi\left((D\psi)^{-1}(D\varphi(x))\right)\right|^p dP_X(x)\\ &\quad= \int |\varphi(x) - \psi(x)|^p dP_X \sharp (D\psi)^{-1}(D\varphi)(x) = \|\varphi - \psi\|^p_{L^p(P_X\sharp(D\psi)^{-1}(D\varphi))},\end{aligned}$$

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

where $P_X \sharp (D\psi)^{-1}(D\varphi)$ is the pushforward measure of P_X via $(D\psi)^{-1}(D\varphi)$. But note that we then have

$$\|\varphi - \psi\|^p_{L^p(P_X \sharp (D\psi)^{-1}(D\varphi))} \leq C\|\varphi - \psi\|^p_{L^p(P_X)},$$

which follows from Corollary 2 in Stern (2010), which also provides the expression for C in terms of the Jacobian. The fact that a change of variables is a lattice isometry between L^1 spaces (Aliprantis and Border, 2006, Cor. 13.47) implies that, for $p = 1$, we have $C = 1$. Putting everything together, we have

$$\|\psi^* - \varphi^*\|^p_{L^p(P_Y)} \leq C\|\varphi - \psi\|^p_{L^p(P_X)}$$

and taking the pth root on both sides gives the claim. ■

We are now ready to prove Theorem 1.

Proof of Theorem 1. In order to obtain the rate of convergence of the estimator, we want to apply Theorem 3.2.5 in van der Vaart and Wellner (2013). For this, we have to show that the minimum of $M(\varphi)$ is well-separated and to obtain the degree of smoothness of the objective function $M(\cdot)$. In the following, we denote by the argument that minimizes $M(\varphi)$ by φ_0. We split the proof into three parts: in the first part, we show that the minimum of M is well-separated. In the second part, we derive the degree of continuity of M. In the third part, we put everything together and derive the rate of convergence. We divide the second part, deriving the degree of continuity of M, into two further steps.

First part: The minimum is well-separated.

A sufficient condition for the minimum of the Kantorovich problem to be well-separated is that the second variation is linear in both arguments and continuous in v, as well as *strongly positive* in the sense that there exists a constant $c > 0$ such that

$$\delta^2 M_{\varphi_0}(v, v) \geq cG(v)$$

for some nonnegative function G which satisfies $G(0) = 0$, see Theorem 3.2.5 in van der Vaart and Wellner (2013). Recall from Lemma 1 that the second variation $\delta^2 M_{\varphi_0}(v, v)$ in directions $v \in C^2(\mathcal{X})$ takes the form

$$\delta^2 M_{\varphi}(v, v) = \int_{\mathcal{X}} \|Dv(x)\|^2 f_Y(D\varphi^{**}(x)) \det(D^2\varphi^{**}(x))dx.$$

Now, note that strong positivity of the second variation of $M(\cdot)$ holds by Assumption 3 if we define

$$G(v) := \mathrm{Var}_{P_X}(v) = \int |v(x)|^2 dP_X(x) - \left(\int v(x)dP_X\right)^2,$$

because then the inequality just reads

$$\delta^2 M_{\varphi_0}(v, v) = \int_{\mathcal{X}} \|Dv(x)\|^2 f_X(x)dx \geq c\mathrm{Var}_{P_X}(v), \tag{B.8}$$

as φ_0 satisfies the Monge–Ampère equation by Lemma 1. The optimum is hence well-separated. We can therefore turn to the longer part of the proof, which deals with deriving the correct modulus of continuity of the objective function $M(\varphi)$.

Second part: Finding the correct degree of continuity of $M(\varphi)$. In this part, we need to find the proper degree of continuity of the empirical process of the objective function, i.e.,

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

we need to find the correct function ϕ_n such that

$$E^* \sup_{\sqrt{\mathrm{Var}_{P_X}(\varphi-\varphi_0)}<\delta} \sqrt{n}\left|\left(\hat{M}_n - M\right)(\varphi) - \left(\hat{M}_n - M\right)(\varphi_0)\right| \lesssim \phi_n(\delta), \tag{B.9}$$

where

$$\hat{M}_n(\varphi) := \int_{\mathcal{X}} \varphi(x) d\left(\mathbb{P}_n^X * K_{h_n}\right)(x) + \int_{\mathcal{Y}} \varphi^*(y) d\left(\mathbb{P}_n^Y * K_{h_n}\right)(y)$$

is the smoothed empirical process and E^* is the outer expectation with respect to X and Y. From now on, we restrict the setting to compact subsets of $\mathcal{X}^\circ$ and $\mathcal{Y}^\circ$. To obtain the correct degree of smoothness ϕ, we use the multivariate generalization of Theorem 6 in Giné and Nickl (2008), which provides the optimal rate of convergence in terms of bandwidths for kernel density estimators. In order to do this, we need to show that φ and its convex conjugate φ^* lie in Donsker classes and that the kernel density estimators converge in the correct Hölder spaces. We start with the convergence of the kernel density estimators.

Step 1: Here, we construct appropriate function classes for $\hat{\varphi}_n - \varphi_0$ and their convolved analogs. To do so, recall that by assumption $f_X \in C_{loc}^{s,\alpha}(\mathcal{X}^\circ)$ and $f_Y \in C_{loc}^{s,\alpha}(\mathcal{Y}^\circ)$. Therefore, by Lemma 2, it follows that $\varphi \in C_{loc}^{s+2,\alpha}(\mathcal{X}^\circ)$ and $\varphi^* \in C_{loc}^{s+2}(\mathcal{Y}^\circ)$. Furthermore, Lemma 3 implies that φ_n converges in $C_{loc}^{s+2,\alpha}(\mathcal{X}^\circ)$ if f_n converges in $C_{loc}^{s,\alpha}(\mathcal{X}^\circ)$, so that we can focus on the classes of functions

$$\mathcal{F} := \{\varphi - \varphi_0 : \varphi, \varphi_0 \in C_{loc}^{s+2,\alpha}(\mathcal{X}^\circ) \text{ and strictly convex}\} \qquad \text{and}$$

$$\mathcal{F}^* := \{\varphi^* - \varphi_0^* : \varphi, \varphi_0 \in C_{loc}^{s+2,\alpha}(\mathcal{X}^\circ) \text{ and strictly convex.}\}$$

Based on this, the fact that P_X and P_Y are probability measures immediately implies by Hölder's inequality that all functions in $\mathcal{F}$ and $\mathcal{F}^*$ are square integrable.

It is hence natural to consider the classes of functions

$$\mathcal{F}_\delta := \{\|\varphi - \psi\|_{L^2(P_X)} < \delta : \varphi, \psi \in \mathcal{F}\} \quad \text{and} \quad \mathcal{F}_\delta^* := \{\|\varphi^* - \psi^*\|_{L^2(P_Y)} < \delta : \varphi^*, \psi^* \in \mathcal{F}\},$$

which form a δ-covering of $\mathcal{F}$ and $\mathcal{F}^*$. The corresponding envelope functions are

$$F_{\mathcal{F}_\delta}(x) := \sup_{\varphi,\psi\in\mathcal{F}_\delta} |\varphi(x) - \psi(x)| \quad \text{and}$$

$$F_{\mathcal{F}_\delta^*}(y) := \sup_{\varphi,\psi\in\mathcal{F}_\delta^*} |\varphi^*(y) - \psi^*(y)|$$

and are also square integrable. Note that since the variance is bounded above by the L^2 norm, it holds that the brackets measured in the variance are smaller than the brackets measured in L^2 norm, so that a δ-covering of $\mathcal{F}$ in L^2 norm is a δ-covering of $\mathcal{F}$ in variance.

Step 2: Here, we put everything together, showing that $\phi_n(\delta) = \delta^{1-\frac{d-4}{2s+2\alpha}}$ based on the previous step. To do so, we first write

$$\begin{aligned}
&\sqrt{n}\left|\left(\hat{M}_n - M\right)(\varphi) - \left(\hat{M}_n - M\right)(\varphi_0)\right| \\
&\quad = \sqrt{n}\left|\int (\varphi - \varphi_0) d\left(\mathbb{P}_n^X * K_{h_n} - P_X\right) + \int (\varphi^* - \varphi_0^*) d\left(\mathbb{P}_n^Y * K_{h_n} - P_Y\right)\right| \\
&\quad \le \sqrt{n}\left|\int (\varphi - \varphi_0) d\left(\mathbb{P}_n^X * K_{h_n} - P_X\right)\right| + \sqrt{n}\left|\int (\varphi^* - \varphi_0^*) d\left(\mathbb{P}_n^Y * K_{h_n} - P_Y\right)\right|
\end{aligned}$$

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

and consider each term separately. In fact, we focus on the first term, as the second term is completely analogous.

We use a similar idea to Radulović and Wegkamp (2000) in the following. For this, we write

$$\int (\varphi-\varphi_0)d(\mathbb{P}_n^X * K_{h_n} - P_X)$$
$$= \int (\varphi-\varphi_0)d(\mathbb{P}_n^X * K_{h_n} - E\mathbb{P}_n^X * K_{h_n}) + \int (\varphi-\varphi_0)d(E\mathbb{P}_n^X * K_{h_n} - P_X).$$

The second term, the bias, goes to zero under Assumptions 1 and 2, which follows from a generalization of the proof of Theorem 6 in Giné and Nickl (2008) to the multivariate case, whose assumptions are implied by ours; the proof for this is the same as the proof of the univariate case in their paper up to some minor notational changes and is therefore omitted. We can therefore focus on the first term and establish its degree of continuity. To do so, note that we can write

$$\int_{\mathcal{X}^\circ} (\varphi(x)-\varphi_0(x))d(\mathbb{P}_n^X * K_{h_n} - E\mathbb{P}_n^X * K_{h_n})(x)$$
$$= \int_{\mathcal{X}^\circ} (\varphi(x)-\varphi_0(x))d\mathbb{P}_n^X * K_{h_n}(x) - \int_{\mathcal{X}^\circ} (\varphi(x)-\varphi_0(x))d(E\mathbb{P}_n^X * K_{h_n})(x)$$
$$= \int_{\mathcal{X}^\circ}\int (\varphi(x+x')-\varphi_0(x+x'))K_{h_n}(x')dx'd\mathbb{P}_n^X(x)$$
$$- \int_{\mathcal{X}^\circ}\int (\varphi(x+x')-\varphi_0(x+x'))K_{h_n}(x')dx'd(E\mathbb{P}_n^X)(x)$$
$$= \int_{\mathcal{X}^\circ} (\bar\varphi(x)-\bar\varphi_0(x))d\mathbb{P}_n^X(x) - \int_{\mathcal{X}^\circ} (\bar\varphi(x)-\bar\varphi_0(x))dP_X(x)$$
$$= \int_{\mathcal{X}^\circ} (\bar\varphi(x)-\bar\varphi_0(x))d(\mathbb{P}_n^X - P_X)(x),$$

where the third line follows from the definition of convolution of measures (Folland, 2013, p. 270) and the fourth line follows from the fact that the empirical measure $\mathbb{P}_n^X$ is an unbiased estimator of P_X. Here, we have defined

$$\bar\varphi(x)-\bar\varphi_0(x) := \int (\varphi(x+x')-\varphi_0(x+x'))K_{h_n}(x')dx'.$$

We therefore define the classes of functions

$$\overline{\mathcal{F}}_n := \{\bar\varphi-\bar\varphi_0 : \varphi-\varphi_0 \in \mathcal{F}\},$$

which change with n; based on these, we consider the classes

$$\overline{\mathcal{F}}_\delta^n := \{\bar\varphi-\bar\psi \in \overline{\mathcal{F}}_n : \|\varphi-\psi\|_{L^2(P_X)} < \delta\}$$

with corresponding envelope functions

$$F_{\overline{\mathcal{F}}_\delta^n}(x) := \sup_{\bar\varphi,\bar\psi\in\overline{\mathcal{F}}_\delta^n} |\bar\varphi(x)-\bar\psi(x)|.$$

By Proposition 8.10 in Folland (2013) and the fact that K is integrable and $\varphi, \varphi_0 \in C^{s+2,\alpha}$, it follows that $\bar\varphi, \bar\varphi_0 \in C^{s+2,\alpha}$.

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

Now we want to relate the bracketing entropies $\log N_{[]}\left(\varepsilon_X, \overline{\mathcal{F}}_\delta^n, \|\cdot\|_{L^2(P_X)}\right)$ to the bracketing entropy $\log N_{[]}\left(\varepsilon_X, \mathcal{F}_\delta, \|\cdot\|_{L^\infty(P_X)}\right)$. This works, because we assume the kernel K to be nonnegative on $\mathcal{X}$. In fact, a bracket $[\varphi_L;\varphi_U]$ stays a bracket (of possibly different size) $[\bar\varphi_L;\bar\varphi_U]$ if we replace φ_L by $\bar\varphi_L$ and φ_U by $\bar\varphi_U$. Indeed, by assumption, it holds that $\varphi_L(x) \le \varphi(x) \le \varphi_U(x)$, for all $x \in \mathcal{X}^\circ$ and every $\varphi \in [\varphi_L;\varphi_U]$, which is the standard partial order on $C(\mathcal{X}^\circ)$. Now, for any $\varphi \in [\varphi_L;\varphi_U]$, we have

$$\bar\varphi_L(x) - \bar\varphi(x) = \int_{\mathcal{X}^\circ} (\varphi_L(x+x') - \varphi(x)) K_{h_n}(x')dx' \le 0,$$

since $K_{h_n}(x')$ is nonnegative on $\mathcal{X}^\circ$; the analog reasoning holds for the upper bound, which shows that $[\bar\varphi_L;\bar\varphi_U]$ is a bracket when $[\varphi_L;\varphi_U]$ is. This is the reasoning used in Radulović and Wegkamp (2000) for instance.

We also need to find an appropriate bound for the brackets. To do so, first calculate[16]

$$\begin{aligned}
&\left[\int |\bar\varphi(x) - \bar\varphi_0(x)|^2 dP_X(x)\right]^{1/2} \\
&\quad \le \left[\int \left(\int \left|\varphi(x+x') - \varphi_0(x+x')\right| K_{h_n}(x')dx'\right)^2 dP_X(x)\right]^{1/2} \\
&\quad \le \int \left(\int \left|\varphi(x+x') - \varphi_0(x+x')\right|^2 dP_X(x)\right)^{1/2} K_{h_n}(x')dx',
\end{aligned}$$

where the third line follows from Minkowski's inequality for integrals (Folland, 2013, Thm. 6.19). Therefore,

$$\begin{aligned}
&\int (\bar\varphi(x) - \bar\varphi_0(x))^2 dP_X(x) \\
&\quad \le \left[\int \left(\int \left(\varphi(x+x') - \varphi_0(x+x')\right)^2 dP_X(x)\right)^{1/2} K_{h_n}(x')dx'\right]^2 \\
&\quad \le \int\int \left(\varphi(x+x') - \varphi_0(x+x')\right)^2 dP_X(x) K_{h_n}(x')dx',
\end{aligned}$$

where the third line follows from Jensen's inequality and the fact that the kernel integrates to one. Now, by a change of variables and a Taylor expansion, we have

$$\begin{aligned}
&= \int\int \left(\varphi(x+x') - \varphi_0(x+x')\right)^2 dP_X(x) K_{h_n}(x')dx' \\
&= \int\int (\varphi(x) - \varphi_0(x))^2 f_X(x-x')dx K_{h_n}(x')dx' \\
&= \int\int (\varphi(x) - \varphi_0(x))^2 \left[f_X(x) - x' \sum_{|k|=1} D^k f_X(x) \right. \\
&\qquad \left. + \frac{1}{2}(x')^2 \sum_{|k|=2} D^k f_X((1-t)x + tx')\right] dx K_{h_n}(x')dx'
\end{aligned}$$

[16] We drop the region of integration, $\mathcal{X}^\circ$ in this derivation in order to save on notation.

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

$$= \int (\varphi(x) - \varphi_0(x))^2 \, dP_X(x)$$
$$+ \int \int (\varphi(x) - \varphi_0(x))^2 \frac{1}{2} \sum_{|k|=2} D^k f_X((1-t)x + tx')(x')^2 K_{h_n}(x') dx dx',$$

where the third line follows from the fact that $P_X(dx) = f_X(x)dx$ and a change of variables and the last line follows from the fact that $\int K(x')dx' = 1$ and $\int x'K(x')dx' = 0$. Now, since $f_X \in C^{s+1}_{loc}$ and is bounded below on its support by γ, it therefore holds that

$$\int (\bar{\varphi}(x) - \bar{\varphi}_0(x))^2 \, dP_X(x)$$
$$\leq \int \int \left(\varphi(x+x') - \varphi_0(x+x')\right)^2 K_{h_n}(x') dx' dP_X(x)$$
$$\leq \int (\varphi(x) - \varphi_0(x))^2 \, dP_X(x) + \frac{1}{2\gamma} \|f\|_{C^{s+1}} h^{2d} \int (\varphi(x) - \varphi_0(x))^2 \, dP_X(x) \int (z)^2 K(z) dz$$
$$= (1 + o(1)) \int (\varphi(x) - \varphi_0(x))^2 \, dP_X(x), \tag{B.10}$$

where the last part of the third line follows from a change of variables $z_i = h^{-1}x'_i$, $i = 1, \ldots, d$, and the fact that $\int (g(x))^2 dx \leq \gamma \int (g(x))^2 f(x) dx$ for any square integrable g. The exact same reasoning as above works to show that

$$\int \left(\bar{\varphi}^*(y) - \bar{\varphi}^*_0(y)\right)^2 dP_Y(y) \leq (1 + o(1)) \int \left(\varphi^*(y) - \varphi^*_0(y)\right)^2 dP_Y(y)$$
$$\leq C(1 + o(1)) \int (\varphi(x) - \varphi_0(x))^2 \, dP_X(x), \tag{B.11}$$

for some constant $C < +\infty$, where the second inequality follows from Lemma 4. Since P_X is a probability measure, it follows from Hölder's inequality that $\|\varphi - \varphi_0\|_2 \leq \|\varphi - \varphi_0\|_\infty$. This, in combination with (B.10) and (B.11), implies that

$$N_{[]}\left(\delta, \overline{\mathcal{F}}_n, \|\cdot\|_{L^2(P_X)}\right) \lesssim N_{[]}\left(\delta, \mathcal{F}, \|\cdot\|_{L^2(P_X)}\right) \quad \text{and}$$
$$N_{[]}\left(\delta, \overline{\mathcal{F}}^*_n, \|\cdot\|_{L^2(P_Y)}\right) \lesssim N_{[]}\left(\delta, \mathcal{F}, \|\cdot\|_{L^2(P_X)}\right) \tag{B.12}$$

by the definition of bracketing numbers.

We can therefore apply Lemma 3.4.2 in van der Vaart and Wellner (2013) which uses bracketing entropy for the respective empirical processes. The bracketing integrals we consider are

$$\tilde{J}_{[]}(\delta, \mathcal{F}, L^2(P_X)) := \int_0^\delta \sqrt{1 + \log N_{[]}\left(\varepsilon_X, \mathcal{F}, \|\cdot\|_{L^2(P_X)}\right)} d\varepsilon_X.$$

Since φ is Hölder continuous, we can bound, using Corollary 2.7.2 in van der Vaart and Wellner (2013), for any $\varepsilon > 0$,

$$\log N_{[]}\left(\varepsilon, \mathcal{F}, \|\cdot\|_{L^2(P_X)}\right) \leq \overline{C}\left(\frac{1}{\varepsilon}\right)^{d/(s+\alpha)} \tag{B.13}$$

for a universal constant $0 < \overline{C} < +\infty$.

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

We can now use Lemma 3.4.2 in van der Vaart and Wellner (2013) to bound

$$
\begin{aligned}
&E^* \sup_{\bar\varphi,\bar\varphi_0\in\overline{\mathcal{F}}^n} \sqrt{n}\left|\int(\bar\varphi-\bar\varphi_0)d\left(\mathbb{P}_n^X - dP_X\right)\right| \\
&\quad\lesssim \tilde J_{[]}(\delta,\overline{\mathcal{F}}^n,L^2(P_X))\left(1+\frac{\tilde J_{[]}(\delta,\overline{\mathcal{F}}^n,L^2(P_X))}{\delta^2\sqrt{n}}\bar c\right) \\
&\quad\lesssim \tilde J_{[]}(\delta,\mathcal{F},L^2(P_X))\left(1+\frac{\tilde J_{[]}(\delta,\mathcal{F},L^2(P_X))}{\delta^2\sqrt{n}}\bar c\right) \quad \text{and} \\
&E^* \sup_{\bar\varphi^*,\bar\varphi_0^*\in\overline{\mathcal{F}}^{*,n}} \sqrt{n}\left|\int(\bar\varphi^*-\bar\varphi_0^*)d\left(\mathbb{P}_n^Y - dP_Y\right)\right| \\
&\quad\lesssim \tilde J_{[]}(\delta,\overline{\mathcal{F}}^{*,n},L^2(P_Y))\left(1+\frac{\tilde J_{[]}(\delta,\overline{\mathcal{F}}^{*,n},L^2(P_Y))}{\delta^2\sqrt{n}}\bar c'\right) \\
&\quad\lesssim \tilde J_{[]}(\delta,\mathcal{F},L^2(P_X))\left(1+\frac{\tilde J_{[]}(\delta,\mathcal{F},L^2(P_X))}{\delta^2\sqrt{n}}\bar c'\right),
\end{aligned}
$$

where the first inequalities follow directly from Lemma 3.4.3 in van der Vaart and Wellner (2013), the constants c and c' are the constants which uniformly bound the class of Hölder continuous functions we derived in step 2, and the second inequalities follow from (B.10), (B.11), and (B.12). In particular, by the fact that the L^2 norm dominates the variance, this implies that we also have a δ-covering of $\mathcal{F}$ by the sets $\mathcal{F}_\delta$ when measured in the variance.

It is here where we need to distinguish between three cases, depending on the smoothness s and the dimension d. Recalling the bound (B.13), the bracketing integrals converge for $s+\alpha > \frac{d}{2}$, i.e.,

$$
\begin{aligned}
\tilde J_{[]}(\delta,\mathcal{F},L^2(P_X)) &:= \int_0^\delta \sqrt{1+\log N_{[]}\left(\varepsilon_X,\mathcal{F},\|\cdot\|_{L^2(P_X)}\right)}d\varepsilon_X \\
&\lesssim \int_0^\delta \left(1+\bar{\bar C}\left(\frac{1}{\varepsilon_X}\right)^{d/(s+\alpha)}\right)^{1/2} d\varepsilon_X.
\end{aligned}
$$

We can bound the last line by

$$
\int_0^\delta \left(1+\bar{\bar C}\left(\frac{1}{\varepsilon_X}\right)^{d/(s+\alpha)}\right)^{1/2} d\varepsilon_X \lesssim \delta^{1-\frac{d}{2(s+\alpha)}},
$$

since we can integrate

$$
\int_0^\delta \sqrt{\left(\frac{1}{\varepsilon}\right)^{(d/(s+\alpha))}}\,d\varepsilon = \frac{1}{1-\frac{d}{2s+2\alpha}}\delta^{1-\frac{d}{2s+2\alpha}}
$$

and the fact that

$$
\sqrt{1+a} \le 1+\sqrt{a}, \qquad \text{for } a \ge 0.
$$

This shows the first case. For the other two cases, i.e., $s+\alpha = \frac{d}{2}$ and $s+\alpha < \frac{d}{2}$, the entropy integral as stated does not converge. We therefore have to adjust the bracketing

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

integral slightly, which is possible since we can allow for the limit stochastic process to have a quadratic drift as long as that drift is strictly smaller than the centering function. This implies that the lower value of the entropy integral needs not to be zero but can be replaced by $\min\{c\delta^2, \frac{\delta}{3}\}$ for some small constant c (van der Vaart and Wellner, 2013, p. 326). With this choice, both integrals converge (van der Vaart and Wellner, 2013, p. 330):

$$\int_{\min\{c\delta^2,\frac{\delta}{3}\}}^{\delta} \sqrt{\left(\frac{1}{\varepsilon}\right)^{(d/(s+\alpha))}}\, d\varepsilon \lesssim \begin{cases} \log\left(\frac{1}{\delta}\right) & \text{for } s+\alpha = \frac{d}{2} \\ \left(\frac{1}{\delta}\right)^{\frac{d-2(s+\alpha)}{2(s+\alpha)}} & \text{for } s+\alpha < \frac{d}{2} \end{cases}.$$

Part 3: Putting everything together. With these bounds and the consistency proved in Lemma 3, we can now apply Theorem 3.2.5 in van der Vaart and Wellner (2013) in all three cases to conclude. In fact, the rate of convergence r_n is obtained by the requirement $r_n^2 \phi_n\left(\frac{1}{r_n}\right) \lesssim \sqrt{n}$, where $\phi_n(\delta) = \tilde{J}_{[]}(\delta, \mathcal{F}, L^2(P_X))\left(1 + \frac{\tilde{J}_{[]}(\delta,\mathcal{F},L^2(P_X))}{\delta^2\sqrt{n}}\bar{c}\right)$. This gives the following lower bounds on the rates:

$$r_n = \begin{cases} n^{\frac{s+\alpha}{2(s+\alpha)+d}} & \text{for } s+\alpha > \frac{d}{2} \\ \frac{n^{1/4}}{(\log(n))^{1/2}} & \text{for } s+\alpha = \frac{d}{2} \\ n^{\frac{1}{(s+\alpha)(2(s+\alpha)+d)}} & \text{for } s+\alpha < \frac{d}{2} \end{cases}$$

and concludes the proof. ∎

B.3. Proof of Proposition 1.

The key for the proof is the following lemma, which is adapted from the proof of Theorem 2.33 in Attouch and Wets (1986).

LEMMA 5. *Let φ_1, φ_2 be proper strictly convex and bounded potential functions on every compact subset of $\mathcal{X}^\circ$ with Lipschitz-continuous gradients $D\varphi_1$ and $D\varphi_2$ satisfying $D\varphi_1(\mathcal{X}^\circ) = \mathcal{Y}^\circ = D\varphi_2(\mathcal{X}^\circ)$. Then, it holds, for all $x \in \mathcal{X}^\circ$,*

$$\|D\varphi_1(x) - D\varphi_2(x)\|^2 \le \bar{c}(1 + \max\{L_1, L_2\})^2 \, |\varphi_1(x) - \varphi_2(x)|, \tag{B.14}$$

where $0 \le L_1, L_2 < +\infty$ are the Lipschitz constants of $D\varphi_1$ and $D\varphi_2$, respectively, $\bar{c} < +\infty$ is a constant, and $\|\cdot\|$ is the euclidean norm.

Proof. Since φ_1 and φ_2 are convex, finite everywhere, and differentiable, it follows that, for any x_1 and x_2 in $\mathcal{X}^\circ$,

$$\varphi_1(x_2) - \varphi_1(x_1) \ge \langle D\varphi_1(x_1), x_2 - x_1\rangle,$$
$$\varphi_2(x_1) - \varphi_2(x_2) \ge \langle D\varphi_2(x_2), x_1 - x_2\rangle.$$

Adding these up while noting that $\langle D\varphi_2(x_2), x_1 - x_2\rangle = -\langle D\varphi_2(x_2), x_2 - x_1\rangle$, we have, for all $x_1, x_2 \in \mathcal{X}^\circ$,

$$(\varphi_1(x_2) - \varphi_2(x_2)) - (\varphi_1(x_1) - \varphi_2(x_1)) \ge \langle D\varphi_1(x_1) - D\varphi_2(x_2), x_2 - x_1\rangle.$$

Now, fix x_1 and choose a corresponding $x_2 \in (I + D\varphi_2)^{-1}[(I + D\varphi_1)(x_1)]$, where $I(x) = x$ is the identity. To see that this is always possible in our setting, recall that both $D\varphi_1$ and

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

$D\varphi_2$ map $\mathcal{X}^\circ$ to $\mathcal{Y}^\circ$. Therefore, the Minkowski sum[17] $\mathcal{X}^\circ + \mathcal{Y}^\circ \subset \mathbb{R}^d$ is such that

$$(I + D\varphi_1)(\mathcal{X}^\circ) = \mathcal{X}^\circ + \mathcal{Y}^\circ = (I + D\varphi_2)(\mathcal{X}^\circ).$$

But since φ_1 and φ_2 are strictly convex, it follows that $D\varphi_1$ and $D\varphi_2$ are invertible (Villani, 2003, Thm. 2.12(iv)), which means that, for every $p \in \mathcal{X}^\circ + \mathcal{Y}^\circ$, there must exist $x_p \in \mathcal{X}^\circ$ such that $(I + D\varphi_2)(x_p) = p$. Now, set $x_2 = x_p$.

Thus, for all $x_1 \in \mathcal{X}^\circ$ and corresponding x_2, we have

$$\|D\varphi_1(x_1) - D\varphi_2(x_2)\| \leq |(\varphi_1(x_2) - \varphi_2(x_2)) - (\varphi_1(x_1) - \varphi_2(x_1))|^{1/2}$$

by definition of the inner product. We now obtain a lower bound for the left-hand side of the last equation by

$$\begin{aligned}\|D\varphi_1(x_1) - D\varphi_2(x_1)\| &\leq \|D\varphi_1(x_1) - D\varphi_2(x_2)\| + \|D\varphi_2(x_2) - D\varphi_2(x_1)\| \\ &\leq \|D\varphi_1(x_1) - D\varphi_2(x_2)\| + L_2\|x_1 - x_2\| \\ &= (1 + L_2)\|D\varphi_1(x_1) - D\varphi_2(x_2)\|,\end{aligned}$$

where the first line follows from the triangle inequality, the second line follows from Lipschitz continuity of $D\varphi_2$, and the third line follows from our choice of x_2: note that $x_2 \in (I + D\varphi_2)^{-1}[(I + D\varphi_1)(x_1)]$ implies that $x_1 - x_2 = D\varphi_2(x_2) - D\varphi_1(x_1)$.

Therefore,

$$\|D\varphi_1(x_1) - D\varphi_2(x_1)\| \leq (1 + L_2)\,|(\varphi_1(x_2) - \varphi_2(x_2)) - (\varphi_1(x_1) - \varphi_2(x_1))|^{1/2},$$

so that

$$\begin{aligned}\|D\varphi_1(x_1) - D\varphi_2(x_1)\|^2 &\leq (1 + L_2)^2\,|(\varphi_1(x_2) - \varphi_2(x_2)) - (\varphi_1(x_1) - \varphi_2(x_1))| \\ &\leq (1 + L_2)^2\,|\varphi_1(x_2) - \varphi_2(x_2)| + |\varphi_1(x_1) - \varphi_2(x_1)| \\ &\leq (1 + L_2)^2 \bar{c}\,|\varphi_1(x_1) - \varphi_2(x_1)|,\end{aligned}$$

where the last line follows from the fact that $\|\varphi_2\|_{C^{s+2,\alpha}} \leq \bar{c} < +\infty$. Since we can switch the roles of φ_1 and φ_2, it holds

$$\|D\varphi_1(x_1) - D\varphi_2(x_1)\|^2 \leq \bar{c}(1 + \max\{L_1, L_2\})^2\,|\varphi_1(x_1) - \varphi_2(x_1)|,$$

which concludes the proof. ■

With this lemma, the proof of the proposition is now straightforward.

Proof of Proposition 1. First, the optimum φ_0 is well-separated with respect to the L^2 norm under the normalization (7) as argued in the main text. Second, under Assumptions 1 and 2, we can apply Lemma 5, which gives

$$\|D\hat{\varphi}_n(x) - D\varphi_0(x)\|^2 \leq \bar{c}(1 + \max\{L_1, L_2\})^2\,\left|\hat{\varphi}_n(x) - \varphi_0(x)\right|.$$

Integrating both sides with respect to P_X, we obtain

$$\begin{aligned}\|D\hat{\varphi}_n - D\varphi_0\|^2_{L^2(P_X)} &= \int_{\mathcal{X}} \|D\hat{\varphi}_n(x) - D\varphi_0(x)\|^2 dP_X(x) \\ &\leq \bar{c}(1 + \max\{L_1, L_2\})^2 \int_{\mathcal{X}} \left|\hat{\varphi}_n(x) - \varphi_0(x)\right| dP_X(x).\end{aligned}$$

[17] The Minkowski sum is defined as $\mathcal{X} + \mathcal{Y} := \{x + y : x \in \mathcal{X}, y \in \mathcal{Y}\}$.

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

We can bound the right-hand side by the fact that P_X is a probability measure using Hölder's inequality, so that

$$\begin{aligned}\|D\hat{\varphi}_n - D\varphi_0\|^2_{L^2(P_X)} &\leq \overline{c}(1+\max\{L_1,L_2\})^2 \int_{\mathcal{X}} \left|\hat{\varphi}_n(x) - \varphi_0(x)\right| dP_X(x) \\ &\leq \overline{c}(1+\max\{L_1,L_2\})^2 \left(\int_{\mathcal{X}} \left|\hat{\varphi}_n(x) - \varphi_0(x)\right|^2 dP_X(x)\right)^{1/2} \\ &= \overline{c}(1+\max\{L_1,L_2\})^2 \|\hat{\varphi}_n - \varphi_0\|_{L^2(P_X)} = O_{P^*}(\frac{1}{r_n}).\end{aligned}$$

Taking square roots, this implies that

$$\|D\hat{\varphi}_n - D\varphi_0\|_{L^2(P_X)} = O_{P^*}(r_n^{-1/2})$$

and concludes the proof. ■

REFERENCES

Aleksandrov, A. (1939) Almost everywhere existence of the second differential of a convex function and some properties of convex surfaces connected with it. *Leningrad State University Annals* [*Uchenye Zapiski*] 37, 3–35.

Aliprantis, C.D. & K. Border (2006) *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer Science & Business Media.

Athey, S. & G.W. Imbens (2006) Identification and inference in nonlinear difference-in-differences models. *Econometrica* 74(2), 431–497.

Attouch, H. & R.J.B. Wets (1986) Isometries for the Legendre–Fenchel transform. *Transactions of the American Mathematical Society* 296(1), 33–60.

Bačák, M. & J.M. Borwein (2011) On difference convexity of locally Lipschitz functions. *Optimization* 60(8–9), 961–978.

Bakry, D., F. Barthe, P. Cattiaux, & A. Guillin (2008) A simple proof of the Poincaré inequality for a large class of probability measures. *Electronic Communications in Probability* 13, 60–66.

Benamou, J.D. & Y. Brenier (2000) A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. *Numerische Mathematik* 84(3), 375–393.

Benamou, J.D., B.D. Froese, & A.M. Oberman (2014) Numerical solution of the optimal transportation problem using the Monge–Ampère equation. *Journal of Computational Physics* 260, 107–126.

Brenier, Y. (1991) Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics* 44(4), 375–417.

Caffarelli, L.A. (1990). Interior $W^{2,p}$estimates for solutions of the Monge–Ampère equation. *Annals of Mathematics* 131(1), 135–150.

Carlier, G., V. Chernozhukov, & A. Galichon (2016) Vector quantile regression: An optimal transport approach. *Annals of Statistics* 44(3), 1165–1192.

Chartrand, R., B. Wohlberg, K. Vixie, & E. Bollt (2009) A gradient descent solution to the Monge–Kantorovich problem. *Applied Mathematical Sciences* 3(22), 1071–1080.

Chernozhukov, V., A. Galichon, M. Hallin, & M. Henry (2017) Monge–Kantorovich depth, quantiles, ranks and signs. *Annals of Statistics* 45(1), 223–256.

Chernozhukov, V., A. Galichon, M. Henry, & B. Pass (2019) Single Market Nonparametric Identification of Multi-Attribute Hedonic Equilibrium Models. Cemmap Working paper, CWP 27/19.

Chiappori, P.A., R.J. McCann, & L.P. Nesheim (2010) Hedonic price equilibria, stable matching, and optimal transport: Equivalence, topology, and uniqueness *Economic Theory* 42(2), 317–354.

Clarke, F.H. (1975) Generalized gradients and applications. *Transactions of the American Mathematical Society* 205, 247–262.

https://doi.org/10.1017/S0266466621000037 Published online by Cambridge University Press

Cuturi, M. (2013) Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*. Edited by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger, 26, 2292–2300. Curran Associates, Inc.

Dalalyan, A. & M. Reiß (2007) Asymptotic statistical equivalence for ergodic diffusions: The multidimensional case. *Probability Theory and Related Fields* 137(1–2), 25–47.

de Philippis, G. & A. Figalli (2013) Second order stability for the Monge–Ampère equation and strong Sobolev convergence of optimal transport maps. *Analysis & PDE* 6(4), 993–1000.

del Barrio, E. & J.-M. Loubes (2019) Central limit theorems for empirical transportation cost in general dimension. *The Annals of Probability* 47(2), 926–951.

Ekeland, I., A. Galichon, & M. Henry (2010) Optimal transportation and the falsifiability of incompletely specified economic models. *Economic Theory* 42(2), 355–374.

Folland, G.B. (2013) *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons.

Galichon, A. (2016) *Optimal Transport Methods in Economics*. Princeton University Press.

Galichon, A. (2017) A survey of some recent applications of optimal transport methods to econometrics. *The Econometrics Journal* 20(2), C1–C11.

Gangbo, W. (1994) An elementary proof of the polar factorization of vector-valued functions. *Archive for Rational Mechanics and Analysis* 128(4), 381–399.

Gilbarg, D. & N.S. Trudinger (1998) *Elliptic Partial Differential Equations of Second Order*. Springer.

Giné, E. & R. Nickl (2008) Uniform central limit theorems for kernel density estimators. *Probability Theory and Related Fields* 141(3–4), 333–387.

Gozlan, N., C. Roberto, & P.-M. Samson (2015) From dimension free concentration to the Poincaré inequality. *Calculus of Variations and Partial Differential Equations* 52(3–4), 899–925.

Griewank, A. & P. Rabier (1990) On the smoothness of convex envelopes. *Transactions of the American Mathematical Society* 322(2), 691–709.

Gunsilius, F.F. & S.M. Schennach (2019) Independent Nonlinear Component Analysis, Cemmap Working paper, CWP 46/19.

Henry-Labordère, P. (2017) *Model-Free Hedging: A Martingale Optimal Transport Viewpoint*. CRC Press.

Hiriart-Urruty, J.B. (1985) Generalized differentiability/duality and optimization for problems dealing with differences of convex functions. In *Convexity and Duality in Optimization*, pp. 37–70. Springer.

Hütter, J.C. & P. Rigollet (2019) Minimax Rates of Estimation for Smooth Optimal Transport Maps. arXiv preprint 1905.05828.

Ibragimov, R. (2009) Copula-based characterizations for higher order Markov processes. *Econometric Theory* 25(3), 819–846.

Kirchheim, B. & Kristensen, J. (2001) Differentiability of convex envelopes. *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics* 333(8), 725–728.

Ledoux, M. (2001) *The Concentration of Measure Phenomenon*. American Mathematical Society.

Lee, Y.J. (2018) Sieve Estimation of Optimal Transport with Applications to Multivariate Quantiles and Matching. UCL Job market paper.

Lindenlaub, I. (2017) Sorting multidimensional types: Theory and application. *The Review of Economic Studies* 84(2), 718–789.

Lindsey, M. & Y.A. Rubinstein (2017) Optimal transport via a Monge–Ampère optimization problem. *SIAM Journal on Mathematical Analysis* 49(4), 3073–3124.

Radulović, D. & M. Wegkamp (2000) Weak convergence of smoothed empirical processes: Beyond Donsker classes. In *High Dimensional Probability II*, pp. 89–105. Springer.

Rockafellar, R.T. (1977) Higher derivatives of conjugate convex functions. *Journal of Applied Analysis* 1(1), 41–43.

Rockafellar, R.T. (1997) *Convex Analysis*. Princeton University Press.

Rohde, A. & C. Strauch (2010) Uniform Central Limit Theorems for Multidimensional Diffusions. arXiv preprint 1010.3604.

Rubinstein, Y.A. (2008) Geometric quantization and dynamical constructions on the space of Kähler metrics. PhD thesis, Massachusetts Institute of Technology.

Santambrogio, F. (2015) *Optimal Transport for Applied Mathematicians*. Progress in Nonlinear Differential Equations and Their Applications, 87. Birkhäuser Basel.

Seguy, V., B.B. Damodaran, R. Flamary, N. Courty, A. Rolet, & M. Blondel (2017) Large-Scale Optimal Transport and Mapping Estimation. arXiv preprint 1711.02283.

Sinkhorn, R. (1967) Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly* 74(4), 402–405.

Sommerfeld, M. & Munk, A. (2018) Inference for empirical Wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80(1), 219–238.

Stern, A. (2010) L^p Change of Variables Inequalities on Manifolds. arXiv preprint 1004.0401.

Trudinger, N.S. & X.-J. Wang (2008) The Monge–Ampère equation and its geometric applications. In *Handbook of Geometric Analysis*, vol. 1, pp. 467–524. International Press.

Tsybakov, A.B. (2008) *Introduction to Nonparametric Estimation*. Springer Science & Business Media.

van der Vaart, A. (1994) Weak convergence of smoothed empirical processes. *Scandinavian Journal of Statistics* 21(4), 501–504.

van der Vaart, A. & J. Wellner (2013) *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Science & Business Media.

Villani, C. (2003) *Topics in Optimal Transportation*, Graduate Studies in Mathematics, 58. American Mathematical Society.

Villani, C. (2008) *Optimal Transport: Old and New*, Grundlehren der Mathematischen Wissenschaften, 338. Springer Science & Business Media.

Yukich, J.E. (1992) Weak convergence of smoothed empirical processes. *Scandinavian Journal of Statistics* 19, 271–279.

Zeidler, E. (1985) *Nonlinear Functional Analysis and Its Applications III: Variational Methods and Optimization*. Springer.