



Squared quadratic Wasserstein distance: optimal couplings and Lions differentiability

Aurélien Alfonsi, Benjamin Jourdain

► To cite this version:

Aurélien Alfonsi, Benjamin Jourdain. Squared quadratic Wasserstein distance: optimal couplings and Lions differentiability. ESAIM: Probability and Statistics, 2020, 24, pp.703-717. 10.1051/ps/2020013 . hal-01934705v2

HAL Id: hal-01934705

<https://hal.science/hal-01934705v2>

Submitted on 16 Nov 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

SQUARED QUADRATIC WASSERSTEIN DISTANCE: OPTIMAL COUPLINGS AND LIONS DIFFERENTIABILITY*

AURÉLIEN ALFONSI^{1,2,**} AND BENJAMIN JOURDAIN^{1,2}

Abstract. In this paper, we remark that any optimal coupling for the quadratic Wasserstein distance $W_2^2(\mu, \nu)$ between two probability measures μ and ν with finite second order moments on $\mathbb{R}^d$ is the composition of a martingale coupling with an optimal transport map $\mathcal{T}$. We check the existence of an optimal coupling in which this map gives the unique optimal coupling between μ and $\mathcal{T}\#\mu$. Next, we give a direct proof that $\sigma \mapsto W_2^2(\sigma, \nu)$ is differentiable at μ in the Lions (Cours au Collège de France, 2008) sense iff there is a unique optimal coupling between μ and ν and this coupling is given by a map. It was known combining results by Ambrosio, Gigli and Savaré (Lectures in Mathematics ETH Zürich, Birkhäuser Verlag, Basel, 2005) and Ambrosio and Gangbo (Comm. Pure Appl. Math., 61:18–53, 2008) that, under the latter condition, geometric differentiability holds. Moreover, the two notions of differentiability are equivalent according to the recent paper of Gangbo and Tudorascu (J. Math. Pures Appl. 125:119–174, 2019). Besides, we give a self-contained probabilistic proof that mere Fréchet differentiability of a law invariant function F on $L^2(\Omega, \mathbb{P}; \mathbb{R}^d)$ is enough for the Fréchet differential at X to be a measurable function of X .

Mathematics Subject Classification. 90C08, 60G42, 60E15, 58B10, 49J50.

Received December 23, 2019. Accepted March 4, 2020.

1. INTRODUCTION

In this paper, we are interested in the structure of optimal couplings for the squared quadratic Wasserstein distance $W_2^2(\mu, \nu)$ between μ and ν in the set $\mathcal{P}_2(\mathbb{R}^d)$ of probability measures with finite second order moments on $\mathbb{R}^d$, and in the differentiability of $W_2^2(\mu, \nu)$ with respect to μ . By definition, $W_2^2(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int |y - x|^2 \pi(dx, dy)$ where $\Pi(\mu, \nu)$ denotes the set of coupling measures on $\mathbb{R}^d \times \mathbb{R}^d$ with first and second marginals respectively equal to μ and ν and $|\cdot|$ denotes the Euclidean norm on $\mathbb{R}^d$. There always exists an optimal coupling and we denote by $\Pi^{opt}(\mu, \nu)$ the set of optimal couplings. According to [11], there exists only one W_2 -optimal coupling π between μ and each $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ and this coupling is given by a map T (i.e. $\pi = (I_d, T)\#\mu$ where I_d denotes the identity function on $\mathbb{R}^d$) iff μ gives 0 mass to the $c - c$ hypersurfaces of dimension $d - 1$. Even when μ does not satisfy this condition which is implied by absolute continuity with respect to the Lebesgue measure, according to Proposition 5.13 [8], if $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ is a C^2 strictly convex function such

*This research benefited from the support of the “Chaire Risques Financiers”, Fondation du Risque.

Keywords and phrases: Optimal transport, Wasserstein distance, differentiability, couplings of probability measures, convex order.

¹ CERMICS, Ecole des Ponts, Marne-la-Vallée, France.

² MathRisk, Inria, Paris, France.

** Corresponding author: aurelien.alfonsi@enpc.fr

that $\int_{\mathbb{R}^d} |\nabla \varphi(x)|^2 \mu(dx) < \infty$, then there is a unique W_2 -optimal coupling between μ and $\nu = \nabla \varphi \# \mu$ and this coupling is given by the map $\nabla \varphi$. But there also exist measures $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ such that either the unique optimal coupling (uniqueness holds in dimension $d = 1$ for instance) is not given by a map or there exist distinct optimal couplings. In the latter case, any strictly convex combination of these couplings is an optimal coupling which is not given by a map.

In Section 2, we study optimal couplings π which are not given by a map. By disintegration, $\pi(dx, dy) = \mu(dx)k(x, dy)$ for some Markov kernel k on $\mathbb{R}^d$ (which is $\mu(dx)$ a.e. unique). Setting $\mathcal{T}(x) = \int_{\mathbb{R}^d} yk(x, dy)$ and using the bias-variance decomposition under the kernel k , we obtain that π is the composition of a martingale coupling between $\mathcal{T}\#\mu$ and ν with the map $\mathcal{T}$ which gives a W_2 -optimal coupling between μ and $\mathcal{T}\#\mu$. Note that couplings of this form have recently been studied by Gozlan and Juillet [12] when considering the barycentric optimal cost problem. For $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ a strictly convex function such that $\int_{\mathbb{R}^d} \phi(y)\nu(dy) < \infty$, by minimizing $\int_{\mathbb{R}^d} \phi(\mathcal{T}(x))\mu(dx)$ over the W_2 -optimal couplings between μ and ν , we obtain optimal couplings such that the associated map $\mathcal{T}_\phi$ gives the only optimal coupling between μ and $\mathcal{T}_\phi\#\mu$. There is a unique such coupling when $\phi(x) = |x|^2$.

In Section 3, we are interested in the differentiability of $W_2^2(\mu, \nu)$ in the Lions sense with respect to μ . Gangbo and Tudorascu have recently proved in Corollary 3.22 [10] that the Lions differentiability [15] of a function $f : \mathcal{P}_2(\mathbb{R}^d) \rightarrow \mathbb{R}$ is equivalent to the geometric differentiability and that the Fréchet derivative of the lift at $X \sim \mu$ is then given by $\nabla_\mu f(X)$ where $\nabla_\mu f \in L^2(\mathbb{R}^d, \mu; \mathbb{R}^d)$ is the geometric (or Wasserstein) gradient of f at μ . While the lifted space that they consider is the ball centered at the origin of unit volume in $\mathbb{R}^d$ endowed with the Lebesgue measure, the result can be transferred to any atomless lifted space by considering an almost isomorphism between those spaces¹. Theorem 10.2.6 [4] states that $\sigma \mapsto W_2^2(\sigma, \nu)$ is subdifferentiable in the geometric sense at μ when $\Pi^{opt}(\mu, \nu) = \{(I_d, T)\#\mu\}$ for some measurable transport map $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$. On the other hand, Proposition 4.3 [3] states that $\sigma \mapsto W_2^2(\sigma, \nu)$ is always superdifferentiable in the geometric sense at μ with $x \mapsto 2(x - \int_{\mathbb{R}^d} yk(x, dy))$ belonging to the superdifferential for each Markov kernel k on $\mathbb{R}^d$ such that $\mu(dx)k(x, dy) \in \Pi^{opt}(\mu, \nu)$. Since geometric differentiability amounts to simultaneous geometric sub- and superdifferentiability, as soon as $\Pi^{opt}(\mu, \nu) = \{(I_d, T)\#\mu\}$, then $\sigma \mapsto W_2^2(\sigma, \nu)$ is differentiable in the geometric sense at μ . On the other hand, geometric differentiability implies that the geometric sub- and superdifferential considered as subsets of $L^2(\mathbb{R}^d, \mu; \mathbb{R}^d)$ coincide and contain one element only (see for instance [8], Prop. 5.63). The fact that the quotient of $\{x \mapsto \int_{\mathbb{R}^d} yk(x, dy) : \mu(dx)k(x, dy) \in \Pi^{opt}(\mu, \nu)\}$ for the $\mu(dx)$ a.e. equality is a singleton is therefore necessary for the geometric differentiability of $\sigma \mapsto W_2^2(\sigma, \nu)$ to hold at μ .

We prove that $\sigma \mapsto W_2^2(\sigma, \nu)$ is differentiable at μ in the Lions sense iff $\Pi^{opt}(\mu, \nu) = \{(I_d, T)\#\mu\}$. We give a direct probabilistic proof of the sufficient condition which also follows from the just mentioned results. To prove the necessary condition, we use that the Fréchet differentiability at $X \sim \mu$ of the lift on an atomless probability space is enough for the Fréchet derivative at X to be a.s. equal to a measurable function of X , a consequence of [10] that we show again using simple probabilistic arguments. Let us emphasize that the quotient of $\{x \mapsto \int_{\mathbb{R}^d} yk(x, dy) : \mu(dx)k(x, dy) \in \Pi^{opt}(\mu, \nu)\}$ for the $\mu(dx)$ a.e. equality may be a singleton while $\Pi^{opt}(\mu, \nu)$ is not equal to $\{(I_d, T)\#\mu\}$ for some measurable map $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ (see, in dimension $d = 1$, Rem. 2.4 below).

2. STRUCTURE OF QUADRATIC WASSERSTEIN OPTIMAL COUPLINGS

In this section, we are interested in characterizing the set

$$\begin{aligned} \Pi^{opt}(\mu, \nu) &= \{\pi(dx, dy) \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d) : \mu(dx) = \int_{y \in \mathbb{R}^d} \pi(dx, dy), \nu(dy) = \int_{x \in \mathbb{R}^d} \pi(dx, dy) \\ &\text{and } W_2^2(\mu, \nu) = \int_{\mathbb{R}^d \times \mathbb{R}^d} |y - x|^2 \pi(dx, dy)\}. \end{aligned}$$

¹We thank one of the referees for pointing out this argument to us.

of optimal couplings between two probability measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ for the quadratic cost. This set is not empty : see *e.g.* [4], page 133.

The refined version of the Brenier theorem in [11] ensures that $\Pi^{opt}(\mu, \nu)$ contains a single element $(I_d, T) \# \mu$ which is given by a measurable transport map $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ for each $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ iff μ does not give mass to the $c - c$ hypersurfaces parametrized by an index $i \in \{0, \dots, d-1\}$ and two convex functions f and g from $\mathbb{R}^{d-1}$ to $\mathbb{R}$:

$$\{(x_1, \dots, x_i, f(x) - g(x), x_{i+1}, \dots, x_{d-1}) : x = (x_1, \dots, x_{d-1}) \in \mathbb{R}^{d-1}\}.$$

The next lemma deals with the case where $\Pi^{opt}(\mu, \nu) \neq \{(I_d, T) \# \mu\}$ for some measurable transport map.

Lemma 2.1. *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$. One of the two conditions holds:*

- $\Pi^{opt}(\mu, \nu) = \{(I_d, T) \# \mu\}$ for some measurable transport map $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$,
- $\exists \mu(dx)k(x, dy) \in \Pi^{opt}(\mu, \nu)$ such that $\int_{\mathbb{R}^d \times \mathbb{R}^d} |y - \int_{\mathbb{R}^d} zk(x, dz)|^2 k(x, dy) \mu(dx) > 0$.

Moreover, if any coupling in $\Pi^{opt}(\mu, \nu)$ is given by a map i.e. writes $(I_d, T) \# \mu$ for some measurable function $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$, then $\Pi^{opt}(\mu, \nu)$ is a singleton.

Proof. If the set $\Pi^{opt}(\mu, \nu)$ has a single element $\mu(dx)k(x, dy)$, defining $\mathcal{T}(x) = \int_{\mathbb{R}^d} yk(x, dy)$ we either have $\int_{\mathbb{R}^d \times \mathbb{R}^d} |y - \mathcal{T}(x)|^2 k(x, dy) \mu(dx) > 0$ or $\mu(dx)k(x, dy) = \mu(dx)\delta_{\mathcal{T}(x)}(dy)$. Otherwise, we can pick two distinct elements $k_1, k_2 \in \Pi^{opt}(\mu, \nu)$ and $k(x, dy) = \frac{1}{2}(k_1(x, dy) + k_2(x, dy))$ is such that $\mu(dx)k(x, dy) \in \Pi^{opt}(\mu, \nu)$ and $\int_{\mathbb{R}^d \times \mathbb{R}^d} |y - \int_{\mathbb{R}^d} zk(x, dz)|^2 k(x, dy) \mu(dx) > 0$. The second statement easily follows. $\square$

Remarking that if ν is the Dirac mass at $x \in \mathbb{R}^d$ and ν_ε the uniform distribution on the ball centered at x with radius ε , then $W_2(\nu, \nu_\varepsilon) \leq \varepsilon$, we deduce from the next proposition that for any $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, we can always find $\mu_\varepsilon, \nu_\varepsilon \in \mathcal{P}_2(\mathbb{R}^d)$ such that $W_2(\mu, \mu_\varepsilon) \leq \varepsilon$, $W_2(\nu, \nu_\varepsilon) \leq \varepsilon$ and $\exists \mu_\varepsilon(dx)k_\varepsilon(x, dy) \in \Pi^{opt}(\mu_\varepsilon, \nu_\varepsilon)$ such that $\int_{\mathbb{R}^d \times \mathbb{R}^d} |y - \int_{\mathbb{R}^d} zk_\varepsilon(x, dz)|^2 k_\varepsilon(x, dy) \mu_\varepsilon(dx) > 0$.

Proposition 2.2. *Assume that $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ is not a Dirac mass. Then for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, there exists a sequence $(\mu_n)_n$ of elements of $\mathcal{P}_2(\mathbb{R}^d)$ such that $\lim_{n \rightarrow \infty} W_2(\mu_n, \mu) = 0$ and for each n , there does not exist $T_n : \mathbb{R}^d \rightarrow \mathbb{R}^d$ measurable such that $\Pi^{opt}(\mu_n, \nu) = \{(I_d, T_n) \# \mu_n\}$.*

Proof. Let $(X_i)_{i \geq 1}$ be an i.i.d. sequence of random variables with law μ , and $(Y_i)_{i \geq 1}$ an independent i.i.d. sequence of uniform random variables on the unit ball $\{x \in \mathbb{R}^d, |x| \leq 1\}$. We set $\tilde{\mu}_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ the empirical measure and $\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i + Y_i/n}$. By construction, we have $W_2^2(\mu_n, \tilde{\mu}_n) \leq \frac{1}{n} \sum_{i=1}^n |Y_i/n|^2 \leq 1/n^2$ and $\mathbb{P}(\exists i \neq j, X_i + Y_i/n = X_j + Y_j/n) = 0$, which means that a.s. for each $n \in \mathbb{N}^*$, μ_n weights a.s. exactly n points. The law of large numbers gives the almost sure weak convergence of $\tilde{\mu}_n$ towards μ and the almost sure convergence of $\frac{1}{n} \sum_{i=1}^n |X_i|^2$ to $\mathbb{E}[|X_1|^2]$. Proposition 7.1.5 in [4] ensures that $W_2(\tilde{\mu}_n, \mu) \xrightarrow{n \rightarrow +\infty} 0$ almost surely. By the triangle inequality, we get $W_2(\mu_n, \mu) \xrightarrow{n \rightarrow +\infty} 0$ almost surely.

Now, we consider $(p_n)_{n \geq 1}$ the increasing sequence of prime numbers. Suppose that $\exists n_0 \in \mathbb{N}^*$, such that $T \# \mu_{p_{n_0}} = \nu$. Then, ν weights at most p_{n_0} points and the masses are equal to k/p_{n_0} with $1 \leq k \leq p_{n_0} - 1$ since ν is not a Dirac mass. Then, if we had $T \# \mu_{p_n} = \nu$ for some $n > n_0$, we would have $k/p_{n_0} = k'/p_n$ with $1 \leq k' \leq p_n - 1$. This would imply that p_{n_0} divides kp_n and thus k , which is impossible since $1 \leq k \leq p_{n_0} - 1$. Thus, there is at most one $n_0 \in \mathbb{N}^*$ such that there is a transport map T_{n_0} satisfying $T_{n_0} \# \mu_{p_{n_0}} = \nu$. $\square$

Let us now give a necessary and sufficient condition for the existence of an optimal transport map in dimension $d = 1$. We denote $F_\eta(x) = \eta((-\infty, x])$ and $F_\eta^{-1}(u) = \inf\{x \in \mathbb{R} : \eta((-\infty, x]) \geq u\}$ the cumulative distribution function and the quantile function of a probability measure η on $\mathbb{R}$. For $\mu, \nu \in \mathcal{P}_2(\mathbb{R})$, by Theorem 2.9 in [16], the only element of $\Pi^{opt}(\mu, \nu)$ is the image of the Lebesgue measure on $[0, 1]$ by (F_μ^{-1}, F_ν^{-1}) . The next lemma characterizes the case when this coupling is given by a map.

Lemma 2.3. *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R})$. There exists $T \in L^2(\mathbb{R}, \mu; \mathbb{R})$ such that $\Pi^{opt}(\mu, \nu) = \{(I_1, T) \# \mu\}$ iff for all $x \in \mathbb{R}$ such that $\mu(\{x\}) > 0$, F_ν^{-1} is constant on $(F_\mu(x-), F_\mu(x)]$. Then, the unique optimal transport map is $T(x) = F_\nu^{-1}(F_\mu(x))$.*

Remark 2.4. When F_ν^{-1} is not constant on $(F_\mu(x-), F_\mu(x)]$ for some $x \in \mathbb{R}$ such that $\mu(\{x\}) > 0$, then $\Pi^{opt}(\mu, \nu)$ is not equal to $\{(I_1, T) \# \mu\}$ for some measurable map $T : \mathbb{R} \rightarrow \mathbb{R}$ while, since $\Pi^{opt}(\mu, \nu)$ is a singleton, the quotient of $\{x \mapsto \int_{\mathbb{R}^d} yk(x, dy) : \mu(dx)k(x, dy) \in \Pi^{opt}(\mu, \nu)\}$ for the $\mu(dx)$ a.e. equality is a singleton.

Proof. Let $X \sim \mu$ and U be an independent random variable uniform on $[0, 1]$. The random variable $V = F_\mu(X-) + U(F_\mu(X) - F_\mu(X-))$ is such that $\mathbb{P}(\{F_\mu(X-) < V \leq F_\mu(X)\} \cup \{F_\mu(X-) = V = F_\mu(X)\}) = 1$. This is an uniform random variable on $[0, 1]$: for $u \in (0, 1)$, $u \in [F_\mu(x-), F_\mu(x)]$ for some $x \in \mathbb{R}$ and $\mathbb{P}(V \leq u) = \mathbb{P}(X < x) + \mathbb{P}\left(X = x, U \leq \frac{u - F_\mu(x-)}{F_\mu(x) - F_\mu(x-)}\right) = u$ since X is independent of U . Since $F_\mu^{-1}(V) = X$ for $V \in (F_\mu(X-), F_\mu(X)]$ and $F_\mu^{-1}(V) \leq X$ for $V = F_\mu(X-) = F_\mu(X)$, we have $F_\mu^{-1}(V) \leq X$ a.s.. Since $F_\mu^{-1}(V)$ and X have the same law, we necessarily have $F_\mu^{-1}(V) = X$ a.s.. By the inverse transform sampling, $F_\nu^{-1}(V)$ is distributed according to ν . Let us assume that F_ν^{-1} is constant on $(F_\mu(x-), F_\mu(x)]$ for all $x \in \mathbb{R}$ such that $\mu(\{x\}) > 0$. Then $F_\nu^{-1}(V) = F_\nu^{-1}(F_\mu(X))$ a.s., $F_\nu^{-1} \circ F_\mu \# \mu = \nu$ and

$$\int_0^1 (F_\mu^{-1}(v) - F_\nu^{-1}(v))^2 dv = \mathbb{E}[(X - F_\nu^{-1}(F_\mu(X)))^2] = \int_{\mathbb{R}} (x - F_\nu^{-1}(F_\mu(x)))^2 \mu(dx).$$

Hence $T(x) = F_\nu^{-1}(F_\mu(x))$ is an optimal transport map. Conversely, if T is an optimal transport map such that $T \# \mu = \nu$, we have $T(F_\mu^{-1}(v)) = F_\nu^{-1}(v)$, dv -a.e. For $x \in \mathbb{R}$ such that $\mu(\{x\}) > 0$, F_μ^{-1} is constant on $(F_\mu(x-), F_\mu(x)]$, and therefore F_ν^{-1} is necessarily constant on $(F_\mu(x-), F_\mu(x)]$. $\square$

Remark 2.5. Lemma 2.3 still holds true for μ, ν probability measures on $\mathbb{R}$ with finite moments of order $\rho \geq 1$, and a transport cost $c(x, y) = h(|y - x|)$, with $h : \mathbb{R}_+ \rightarrow \mathbb{R}$ strictly convex such that $\exists C < \infty$, $\forall x \in \mathbb{R}$, $h(|x|) \leq C(1 + |x|^\rho)$. The same proof applies since, by Theorem 2.9 in [16], the only optimal coupling for such a cost is the image of the Lebesgue measure on $[0, 1]$ by (F_μ^{-1}, F_ν^{-1}) .

The next proposition, which is one of the main results of this section, shows that any W_2 -optimal coupling can be written as the composition of a transport map and a martingale kernel *i.e.* a Markov kernel k such that for all $x \in \mathbb{R}^d$, $\int_{\mathbb{R}^d} |y|k(x, dy) < \infty$ and $\int_{\mathbb{R}^d} yk(x, dy) = x$. Let us now give the definition of the convex order on probability measures before recalling its link with the existence of martingale couplings.

Definition 2.6. Let η, ν be two probability measures on $\mathbb{R}^d$. We say that η is smaller than ν in the convex order and write $\eta \leq_{cx} \nu$ if for each convex function $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ such that the integrals make sense,

$$\int_{\mathbb{R}^d} \phi(x) \eta(dx) \leq \int_{\mathbb{R}^d} \phi(y) \nu(dy).$$

Notice that since a convex function ϕ on $\mathbb{R}^d$ is bounded from below by an affine function, for a probability measure η on $\mathbb{R}^d$ with finite first order moment (and in particular for $\eta \in \mathcal{P}_2(\mathbb{R}^d)$), $\int_{\mathbb{R}^d} \phi(x) \eta(dx)$ always makes sense possibly equal to $+\infty$.

Theorem 8 in Strassen [17] ensures that, when $\int_{\mathbb{R}^d} |y| \nu(dy) < \infty$, $\eta \leq_{cx} \nu$ iff there exists a martingale Markov kernel k such that $\eta(dx)k(x, dy) \in \Pi(\eta, \nu)$.

Proposition 2.7. *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, $\mu(dx)k(x, dy) \in \Pi^{opt}(\mu, \nu)$, $\mathcal{T}(x) = \int_{\mathbb{R}^d} yk(x, dy)$ and $\eta = \mathcal{T}\#\mu$. Then $\eta \leq_{cx} \nu$,*

$$W_2^2(\mu, \nu) = W_2^2(\mu, \eta) + \int_{\mathbb{R}^d} |y|^2 \nu(dy) - \int_{\mathbb{R}^d} |z|^2 \eta(dz) \quad (2.1)$$

and $(I_d, \mathcal{T})\#\mu \in \Pi^{opt}(\mu, \eta)$.

On the other hand, if $\eta \leq_{cx} \nu$ is such that (2.1) holds, then combining $\mu(dx)q(x, dz) \in \Pi^{opt}(\mu, \eta)$ with any martingale coupling $\eta(dz)m(z, dy)$ between η and ν , we obtain a W_2 -optimal coupling $\mu(dx)qm(x, dy)$ (where, as usual, $qm(x, dy) = \int_{z \in \mathbb{R}^d} q(x, dz)m(z, dy)$) between μ and ν .

The first part of this proposition is also a consequence of Theorem 12.4.4 in [4]: the barycentric projection of $\mu(x)k(x, dy)$ is precisely $(I_d, \mathcal{T})\#\mu$. Here, we present this result with a probabilistic fashion. For $\mu(dx)k(x, dy)$ as in the first statement and $(X, Y) \sim \mu(dx)k(x, dy)$, by definition of $\mathcal{T}$, $\mathbb{E}[Y|X] = \mathcal{T}(X)$ a.s. so that $\mathbb{E}[Y|\mathcal{T}(X)] = \mathcal{T}(X)$ a.s. and this optimal coupling is the composition of the martingale coupling given by the law of $(\mathcal{T}(X), Y)$ and the transport map $\mathcal{T}$. Notice that since it relies on the bias-variance decomposition, this structure of optimal couplings does not seem to generalize to other Wasserstein distances $W_\rho(\mu, \nu) = (\inf_{\pi \in \Pi(\mu, \nu)} \int |y - x|^\rho \pi(dx, dy))^{1/\rho}$, $\rho \in [1, \infty) \setminus \{2\}$. Nevertheless, Gozlan and Juillet [12] have recently obtained optimal couplings that are the composition of a martingale coupling and a deterministic transport map by considering the barycentric optimal cost problem, which consists in minimizing for a given cost function $\theta : \mathbb{R}^d \rightarrow \mathbb{R}_+$ the quantity $\int_{\mathbb{R}^d} \theta(x - \int_{\mathbb{R}^d} yk(x, dy))\mu(dx)$ among all couplings $\mu(dx)k(x, dy)$ between μ and ν .

Proof. Let us first prove the second statement. Let $\eta \leq_{cx} \nu$, q be a Markov kernel such that $\mu(dx)q(x, dz) \in \Pi^{opt}(\mu, \eta)$ and m be any martingale kernel such that $\eta m = \nu$. Then $\mu(dx)qm(x, dy)$ is a coupling between μ and ν such that

$$\begin{aligned} W_2^2(\mu, \nu) &\leq \int_{\mathbb{R}^d \times \mathbb{R}^d} |y - x|^2 \mu(dx)qm(x, dy) = \int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} |y - z + z - x|^2 \mu(dx)q(x, dz)m(z, dy) \\ &= \int_{\mathbb{R}^d \times \mathbb{R}^d} |y - z|^2 \eta(dz)m(z, dy) + \int_{\mathbb{R}^d \times \mathbb{R}^d} |z - x|^2 \mu(dx)q(x, dz) \\ &= \int_{\mathbb{R}^d} |y|^2 \nu(dy) - \int_{\mathbb{R}^d} |z|^2 \eta(dz) + W_2^2(\mu, \eta) \end{aligned} \quad (2.2)$$

where we used the variance-bias decomposition under the martingale kernel m for the third equality. Hence, if (2.1) holds, then $\mu(dx)qm(x, dy) \in \Pi^{opt}(\mu, \nu)$.

Let now $\mu(dx)k(x, dy) \in \Pi^{opt}(\mu, \nu)$, $\mathcal{T}(x) = \int_{\mathbb{R}^d} yk(x, dy)$ and $\eta = \mathcal{T}\#\mu$. Jensen's inequality immediately gives $\eta \leq_{cx} \nu$ and thus $\eta \in \mathcal{P}_2(\mathbb{R}^d)$. We have

$$\begin{aligned} W_2^2(\mu, \nu) &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |y - \mathcal{T}(x) + \mathcal{T}(x) - x|^2 \mu(dx)k(x, dy) \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |y - \mathcal{T}(x)|^2 \mu(dx)k(x, dy) + \int_{\mathbb{R}^d} |\mathcal{T}(x) - x|^2 \mu(dx) \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} (|y|^2 - |\mathcal{T}(x)|^2) \mu(dx)k(x, dy) + \int_{\mathbb{R}^d} |\mathcal{T}(x) - x|^2 \mu(dx) \\ &= \int_{\mathbb{R}^d} |y|^2 \nu(dy) - \int_{\mathbb{R}^d} |z|^2 \eta(dz) + \int_{\mathbb{R}^d} |\mathcal{T}(x) - x|^2 \mu(dx), \end{aligned}$$

where we used the variance-bias decomposition with respect to $k(x, \cdot)$ for the second equality. With (2.2), we deduce that $\int_{\mathbb{R}^d} |\mathcal{T}(x) - x|^2 \mu(dx) \leq W_2^2(\mu, \eta)$ and $\mathcal{T}$ is a W_2 -optimal transport map between μ and η . $\square$

For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, let us define the sets

$$\begin{aligned}\mathcal{I}_\mu^\nu &= \{\eta \in \mathcal{P}_2(\mathbb{R}^d) : \eta \leq_{\text{cx}} \nu \text{ and } W_2^2(\mu, \nu) = W_2^2(\mu, \eta) + \int_{\mathbb{R}^d} |y|^2 \nu(dy) - \int_{\mathbb{R}^d} |z|^2 \eta(dz)\}, \\ \tilde{\mathcal{I}}_\mu^\nu &= \left\{ \mathcal{T} \# \mu : \exists \mu(dx)k(x, dy) \in \Pi^{\text{opt}}(\mu, \nu), \mathcal{T}(x) = \int_{\mathbb{R}^d} yk(x, dy) \right\}.\end{aligned}$$

By Proposition 2.7, we have $\tilde{\mathcal{I}}_\mu^\nu \subset \mathcal{I}_\mu^\nu$ and $\tilde{\mathcal{I}}_\mu^\nu \neq \emptyset$ since $\Pi^{\text{opt}}(\mu, \nu) \neq \emptyset$. Moreover, there exists an optimal transport map between μ and any element of $\tilde{\mathcal{I}}_\mu^\nu$. The measure $\mathcal{T} \# \mu$ associated with an optimal coupling in $\Pi^{\text{opt}}(\mu, \nu)$ is possibly equal to ν , which always belongs to $\mathcal{I}_\mu^\nu$.

Lemma 2.8. *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$. If $\eta \in \mathcal{I}_\mu^\nu$, then for any $\tilde{\eta}$ such that $\eta \leq_{\text{cx}} \tilde{\eta} \leq_{\text{cx}} \nu$, $\tilde{\eta} \in \mathcal{I}_\mu^\nu$ and $\eta \in \mathcal{I}_\mu^{\tilde{\eta}}$. Moreover, $\mathcal{I}_\mu^\nu = \{\eta \in \mathcal{P}_2(\mathbb{R}^d) : \exists \tilde{\eta} \in \tilde{\mathcal{I}}_\mu^\nu, \tilde{\eta} \leq_{\text{cx}} \eta \leq_{\text{cx}} \nu\}$. Last, the set $\mathcal{I}_\mu^\nu$ is convex.*

Proof. Let $\eta \in \mathcal{I}_\mu^\nu$ and $\tilde{\eta}$ be such that $\eta \leq_{\text{cx}} \tilde{\eta} \leq_{\text{cx}} \nu$. We have

$$W_2^2(\mu, \nu) = W_2^2(\mu, \eta) + \int_{\mathbb{R}^d} |y|^2 \nu(dy) - \int_{\mathbb{R}^d} |\tilde{z}|^2 \tilde{\eta}(d\tilde{z}) + \int_{\mathbb{R}^d} |\tilde{z}|^2 \tilde{\eta}(d\tilde{z}) - \int_{\mathbb{R}^d} |z|^2 \eta(dz). \quad (2.3)$$

Now, we consider $\mu(dx)k(x, dz) \in \Pi^{\text{opt}}(\mu, \eta)$ and $\eta(dz)m(z, d\tilde{z})$ a martingale coupling between η and $\tilde{\eta}$. Then, $W_2^2(\mu, \tilde{\eta}) \leq \int_{(\mathbb{R}^d)^3} |\tilde{z} - z + z - x|^2 \mu(dx)k(x, dz)m(z, d\tilde{z}) = W_2^2(\mu, \eta) + \int_{\mathbb{R}^d} |\tilde{z}|^2 \tilde{\eta}(d\tilde{z}) - \int_{\mathbb{R}^d} |z|^2 \eta(dz)$. This inequality cannot be strict: otherwise, by combining an optimal coupling between μ and $\tilde{\eta}$ and a martingale coupling between $\tilde{\eta}$ and ν , we would contradict (2.3). The equality gives $\eta \in \mathcal{I}_\mu^{\tilde{\eta}}$ and $\tilde{\eta} \in \mathcal{I}_\mu^\nu$ by using (2.3).

If $\tilde{\eta} \in \tilde{\mathcal{I}}_\mu^\nu$, since $\tilde{\mathcal{I}}_\mu^\nu \subset \mathcal{I}_\mu^\nu$, by the first statement, each probability measure η such that $\tilde{\eta} \leq_{\text{cx}} \eta \leq_{\text{cx}} \nu$ belongs to $\mathcal{I}_\mu^\nu$. Hence $\{\eta \in \mathcal{P}_2(\mathbb{R}^d) : \exists \tilde{\eta} \in \tilde{\mathcal{I}}_\mu^\nu, \tilde{\eta} \leq_{\text{cx}} \eta \leq_{\text{cx}} \nu\} \subset \mathcal{I}_\mu^\nu$. On the other hand, for $\eta \in \mathcal{I}_\mu^\nu$, $\mu(dx)q(x, dz) \in \Pi^{\text{opt}}(\mu, \eta)$ and a martingale coupling $\eta(dz)m(z, dy)$ between η and ν , we have $\mu(dx)qm(x, dy) \in \Pi^{\text{opt}}(\mu, \nu)$, by the second assertion in Proposition 2.7. Since, by the martingale property, $\int_{\mathbb{R}^d} yqm(x, dy) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} ym(z, dy)q(x, dz) = \int_{\mathbb{R}^d} zq(x, dz)$ setting $\mathcal{T}(x) = \int_{\mathbb{R}^d} zq(x, dz)$, we have $\mathcal{T} \# \mu \in \tilde{\mathcal{I}}_\mu^\nu$, by the first assertion in Proposition 2.7. Since $\mathcal{T} \# \mu \leq_{\text{cx}} \eta$, we conclude that $\mathcal{I}_\mu^\nu \subset \{\eta \in \mathcal{P}_2(\mathbb{R}^d) : \exists \tilde{\eta} \in \tilde{\mathcal{I}}_\mu^\nu, \tilde{\eta} \leq_{\text{cx}} \eta \leq_{\text{cx}} \nu\}$.

Last, let us consider $\eta_1, \eta_2 \in \mathcal{I}_\mu^\nu$ and $\lambda \in (0, 1)$. Using a convex combination of couplings in $\Pi^{\text{opt}}(\mu, \eta_1)$ and $\Pi^{\text{opt}}(\mu, \eta_2)$, we obtain that $W_2^2(\mu, \lambda\eta_1 + (1 - \lambda)\eta_2) \leq \lambda W_2^2(\mu, \eta_1) + (1 - \lambda)W_2^2(\mu, \eta_2)$. Since $\eta_1, \eta_2 \in \mathcal{I}_\mu^\nu$, we deduce that

$$W_2^2(\mu, \nu) \geq W_2^2(\mu, \lambda\eta_1 + (1 - \lambda)\eta_2) + \int_{\mathbb{R}^d} |y|^2 \nu(dy) - \int_{\mathbb{R}^d} |z|^2 (\lambda\eta_1 + (1 - \lambda)\eta_2)(dz).$$

Since $\lambda\eta_1 + (1 - \lambda)\eta_2 \leq_{\text{cx}} \nu$, there exists a martingale coupling between $\lambda\eta_1 + (1 - \lambda)\eta_2$ and ν . Composing it with an element of $\Pi^{\text{opt}}(\mu, \lambda\eta_1 + (1 - \lambda)\eta_2)$, we obtain a coupling between μ and ν which ensures that

$$W_2^2(\mu, \nu) \leq W_2^2(\mu, \lambda\eta_1 + (1 - \lambda)\eta_2) + \int_{\mathbb{R}^d} |y|^2 \nu(dy) - \int_{\mathbb{R}^d} |z|^2 (\lambda\eta_1 + (1 - \lambda)\eta_2)(dz).$$

Hence $\lambda\eta_1 + (1 - \lambda)\eta_2 \in \mathcal{I}_\mu^\nu$. □

In dimension $d = 1$, since $\Pi^{\text{opt}}(\mu, \nu)$ is a singleton, we can specify the sets $\mathcal{I}_\mu^\nu$ and $\tilde{\mathcal{I}}_\mu^\nu$.

Proposition 2.9. *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R})$ and*

$$\mathcal{T}(x) = \int_0^1 F_\nu^{-1}(F_\mu(x-) + u[F_\mu(x) - F_\mu(x-)])du. \quad (2.4)$$

We have $\tilde{\mathcal{I}}_\mu^\nu = \{\mathcal{T} \# \mu\}$ and $\mathcal{I}_\mu^\nu = \{\eta \in \mathcal{P}_2(\mathbb{R}) : \mathcal{T} \# \mu \leq_{\text{cx}} \eta \leq_{\text{cx}} \nu\}$. Moreover, $\Pi^{\text{opt}}(\mu, \mathcal{T} \# \mu) = \{(I_1, \mathcal{T}) \# \mu\}$ and there is a unique martingale coupling between $\mathcal{T} \# \mu$ and ν and it is W_2 -optimal.

Proof. By the second assertion in Lemma 2.8, the characterization of $\mathcal{I}_\mu^\nu$ easily follows from the one of $\tilde{\mathcal{I}}_\mu^\nu$, which, with the definition of $\tilde{\mathcal{I}}_\mu^\nu$, the first statement in Proposition 2.7 and the uniqueness of the optimal coupling in dimension $d = 1$, also implies that $\Pi^{\text{opt}}(\mu, \mathcal{T} \# \mu) = \{(I_1, \mathcal{T}) \# \mu\}$. Let U, U' be two independent uniform random variables on $[0, 1]$. We define

$$V = F_\mu(F_\mu^{-1}(U)-) + U'[F_\mu(F_\mu^{-1}(U)) - F_\mu(F_\mu^{-1}(U)-)], \quad (2.5)$$

and have by construction

$$F_\mu^{-1}(V) = F_\mu^{-1}(U) \text{ a.s.} \quad (2.6)$$

For $u \in (0, 1)$, $u \in [F_\mu(x-), F_\mu(x)]$ for some $x \in \mathbb{R}$ and

$$\mathbb{P}(V \leq u) = \mathbb{P}(F_\mu^{-1}(U) < x) + \mathbb{P}\left(F_\mu^{-1}(U) = x, U' \leq \frac{u - F_\mu(x-)}{F_\mu(x) - F_\mu(x-)}\right) = u$$

since U' is independent of U . Hence V is uniformly distributed on $[0, 1]$. According to Theorem 2.9 [16], the law of $(F_\mu^{-1}(V), F_\nu^{-1}(V))$ is the unique element of $\Pi^{\text{opt}}(\mu, \nu)$. From (2.5), we get $\mathbb{E}[F_\nu^{-1}(V)|U] = \mathcal{T}(F_\mu^{-1}(U))$ and by (2.6),

$$\mathbb{E}[F_\nu^{-1}(V)|F_\mu^{-1}(V)] = \mathbb{E}[\mathbb{E}[F_\nu^{-1}(V)|U]|F_\mu^{-1}(V)] = \mathbb{E}[\mathcal{T}(F_\mu^{-1}(V))|F_\mu^{-1}(V)] = \mathcal{T}(F_\mu^{-1}(V)).$$

Hence the single element of $\tilde{\mathcal{I}}_\mu^\nu$ is the law $\mathcal{T} \# \mu$ of $\mathcal{T}(F_\mu^{-1}(V))$. Since $\mathcal{T}$ is nondecreasing, $\mathcal{T}(F_\mu^{-1}(V)) = F_{\mathcal{T} \# \mu}^{-1}(V)$ a.s. and $\mathbb{E}[F_\nu^{-1}(V)|F_{\mathcal{T} \# \mu}^{-1}(V)] = F_{\mathcal{T} \# \mu}^{-1}(V)$ a.s.. Hence the law of $(F_{\mathcal{T} \# \mu}^{-1}(V), F_\nu^{-1}(V))$, which is the single element of $\Pi^{\text{opt}}(\mathcal{T} \# \mu, \nu)$, is a martingale coupling. Since all the martingale couplings share the quadratic cost $\int_{\mathbb{R}} y^2 \nu(dy) - \int_{\mathbb{R}} \mathcal{T}(x)^2 \mu(dx)$, each martingale coupling belongs to $\Pi^{\text{opt}}(\mathcal{T} \# \mu, \nu)$ and is therefore equal to the previous one. $\square$

In dimension $d = 1$, there is a single element $\eta \in \tilde{\mathcal{I}}_\mu^\nu$, a unique element in $\Pi^{\text{opt}}(\mu, \eta)$ and the unique martingale coupling between η and ν is W_2 -optimal. We now provide an example in dimension $d = 2$ where these properties fail.

Example 2.10. Let $\mu = \frac{1}{2}(\delta_{(-1,0)} + \delta_{(1,0)})$ and $\nu = \frac{1}{2}(\delta_{(0,-1)} + \delta_{(0,1)})$. Since $|(0, -1) - (-1, 0)| = |(0, 1) - (-1, 0)| = |(0, -1) - (1, 0)| = |(0, 1) - (1, 0)|$, any coupling between μ and ν is W_2 -optimal. The couplings write $\mu(dx)k_p(x, dy)$ with $k_p((-1, 0), dy) = (p\delta_{(0,-1)} + (1-p)\delta_{(0,1)})(dy)$ and $k_p((1, 0), dy) = ((1-p)\delta_{(0,-1)} + p\delta_{(0,1)})(dy)$ for $p \in (0, 1)$. One has $\mathcal{T}_p((-1, 0)) = (0, 1 - 2p)$, $\mathcal{T}_p((1, 0)) = (0, 2p - 1)$, and $\eta_p = \frac{1}{2}(\delta_{(0,1-2p)} + \delta_{(0,2p-1)})$. Any coupling between μ and η_p is W_2 -optimal and as soon as $p \neq 1/2$, there is an optimal coupling different from $(I_2, \mathcal{T}_p) \# \mu$. Moreover, unless $p \in \{0, 1/2, 1\}$, the martingale coupling between η_p and ν is not W_2 -optimal.

According to the next theorem, we can find elements η in $\tilde{\mathcal{I}}_\mu^\nu$ such that $\Pi^{\text{opt}}(\mu, \eta) = \{(I_d, T) \# \mu\}$ for some measurable transport map T by minimizing over $\mathcal{I}_\mu^\nu$ the integral of a strictly convex function.

Theorem 2.11. Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ be strictly convex such that $\int_{\mathbb{R}^d} \phi(y) \nu(dy) < \infty$ and $\mathcal{I}_{\mu, \phi}^\nu := \{\eta \in \mathcal{I}_\mu^\nu : \int_{\mathbb{R}^d} \phi(z) \eta(dz) = \inf_{\eta \in \mathcal{I}_\mu^\nu} \int_{\mathbb{R}^d} \phi(z) \eta(dz)\}$. We have $\emptyset \neq \mathcal{I}_{\mu, \phi}^\nu \subset \tilde{\mathcal{I}}_\mu^\nu$ and for each $\eta \in \mathcal{I}_{\mu, \phi}^\nu$, $\Pi^{\text{opt}}(\mu, \eta) = \{(I_d, T) \# \mu\}$ for some measurable transport map $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$. Moreover, there is a single $\eta_\phi \in \mathcal{I}_{\mu, \phi}^\nu$ such that $\int_{\mathbb{R}^d} |z|^2 \eta_\phi(dz) = \inf_{\eta \in \mathcal{I}_{\mu, \phi}^\nu} \int_{\mathbb{R}^d} |z|^2 \eta(dz)$. Last, there is a single element $\underline{\eta}$ in $\mathcal{I}_{\mu, |x|^2}^\nu$.

This theorem permits to select extreme elements of $\mathcal{I}_\mu^\nu$ and provides the following characterization of the existence of a minimal element for the convex order in this set.

Corollary 2.12. *For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, there exists $\eta_0 \in \mathcal{P}_2(\mathbb{R}^d)$ such that $\mathcal{I}_\mu^\nu = \{\eta_0 \leq_{\text{cx}} \eta \leq_{\text{cx}} \nu\}$ if and only if $\{\eta_\phi : \phi : \mathbb{R}^d \rightarrow \mathbb{R} \text{ strictly convex and such that } \int_{\mathbb{R}^d} \phi(y)\nu(dy) < \infty\} = \{\underline{\eta}\}$ and then $\eta_0 = \underline{\eta}$.*

Let us show the corollary before proving the theorem.

Proof of Corollary 2.12. The necessary condition is obvious. Let us show that it is sufficient. It is enough to check that for any $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ convex such that $\exists C < \infty, \forall x \in \mathbb{R}^d, |\phi(x)| \leq C(1 + |x|)$, we have $\forall \eta \in \mathcal{I}_\mu^\nu, \int_{\mathbb{R}^d} \phi(x)\underline{\eta}(dx) \leq \int_{\mathbb{R}^d} \phi(x)\eta(dx)$ (see e.g. [1], Lem. A.1). For such a function ϕ and for $\varepsilon > 0$, $\phi_\varepsilon(x) := \phi(x) + \varepsilon|x|^2$ is strictly convex and, since $\eta_{\phi_\varepsilon} = \underline{\eta}$, we have

$$\forall \eta \in \mathcal{I}_\mu^\nu, \int_{\mathbb{R}^d} \phi_\varepsilon(x)\underline{\eta}(dx) \leq \int_{\mathbb{R}^d} \phi_\varepsilon(x)\eta(dx).$$

We conclude by letting $\varepsilon \rightarrow 0$ using the dominated convergence theorem. $\square$

To prove Theorem 2.11, we will need the following Lemma

Lemma 2.13. *Let ν be a probability measure on $\mathbb{R}^d$ such that $\int_{\mathbb{R}^d} |y|\nu(dy) < \infty$ and $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ a convex function such that $\int_{\mathbb{R}^d} \phi(y)\nu(dy) < \infty$. Then the family of probability measures $\{\phi\#\eta : \eta \leq_{\text{cx}} \nu\}$ is uniformly integrable.*

Proof of Lemma 2.13. Let us first suppose that ϕ is nonnegative. Let $M \in (0, +\infty)$, $\eta \leq_{\text{cx}} \nu$ and m be a martingale kernel such that $\int_{x \in \mathbb{R}^d} \eta(dx)m(x, dy) = \nu(dy)$. Using Jensen's inequality for the first inequality and the Markov inequality combined with $\eta \leq_{\text{cx}} \nu$ for the third one, we obtain that

$$\begin{aligned} \int_{\mathbb{R}^d} \phi(x)1_{\{\phi(x) \geq M\}}\eta(dx) &\leq \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \phi(y)m(x, dy)1_{\{\phi(x) \geq M\}}\eta(dx) \\ &\leq \int_{\mathbb{R}^d \times \mathbb{R}^d} \left(\phi(y)1_{\{\phi(y) \geq \sqrt{M}\}} + \sqrt{M}1_{\{\phi(y) < \sqrt{M}\}} \right) m(x, dy)\eta(dx) \\ &= \int_{\mathbb{R}^d} \phi(y)1_{\{\phi(y) \geq \sqrt{M}\}}\nu(dy) + \sqrt{M} \int_{\mathbb{R}^d} 1_{\{\phi(x) \geq M\}}\eta(dx) \\ &\leq \int_{\mathbb{R}^d} \phi(y)1_{\{\phi(y) \geq \sqrt{M}\}}\nu(dy) + \frac{1}{\sqrt{M}} \int_{\mathbb{R}^d} \phi(y)\nu(dy). \end{aligned}$$

Hence $\lim_{M \rightarrow \infty} \sup_{\eta \leq_{\text{cx}} \nu} \int_{\mathbb{R}^d} \phi(x)1_{\{\phi(x) \geq M\}}\eta(dx) = 0$. In particular, the family $\{|\phi|\#\eta : \eta \leq_{\text{cx}} \nu\}$ is uniformly integrable. When the sign of ϕ is not constant, we obtain a nonnegative convex function $\tilde{\phi}$ such that $\int_{\mathbb{R}^d} \tilde{\phi}(y)\nu(dy) < \infty$ by addition to ϕ of a suitable affine function ψ . The conclusion follows from the uniform integrability of both the families $\{\psi\#\eta : \eta \leq_{\text{cx}} \nu\}$ and $\{\tilde{\phi}\#\eta : \eta \leq_{\text{cx}} \nu\}$. $\square$

Proof of Theorem 2.11. Let $(\eta_n)_{n \in \mathbb{N}}$ be a sequence in $\mathcal{I}_\mu^\nu$ minimizing $\int_{\mathbb{R}^d} \phi(z)\eta_n(dz)$. For $n \in \mathbb{N}$, let $\mu(dx)q_n(x, dz) \in \Pi^{\text{opt}}(\mu, \eta_n)$ and $\eta_n(dz)m_n(z, dy)$ be a martingale coupling between η_n and ν . By the second part in Proposition 2.7, $\mu(dx)q_n m_n(x, dy) \in \Pi^{\text{opt}}(\mu, \nu)$. Up to extracting a subsequence, we may suppose that $(\mu(dx)q_n(x, dz)m_n(z, dy))_n$ converges weakly to $\mu(dx)r_\infty(x, dz, dy)$ where $\mu(dx) \int_{z \in \mathbb{R}^d} r_\infty(x, dz, dy) \in \Pi^{\text{opt}}(\mu, \nu)$. Let $\mathcal{T}_\infty(x) = \int_{\mathbb{R}^d \times \mathbb{R}^d} y r_\infty(x, dz, dy)$ and $\eta_\infty = \mathcal{T}_\infty \# \mu$. By the first part of Proposition 2.7, $\eta_\infty \in \tilde{\mathcal{I}}_\mu^\nu$. Moreover, by the above weak convergence and the uniform integrability deduced from Lemma 2.13,

$$\int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} \phi(z)\mu(dx)r_\infty(x, dz, dy) = \lim_{n \rightarrow \infty} \int_{\mathbb{R}^d} \phi(z)\eta_n(dz).$$

Taking the limit $n \rightarrow \infty$ in the equality $\int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} \varphi(x, z)(y - z)\mu(dx)q_n(x, dz)m_n(z, dy) = 0$, we obtain that $\int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} \varphi(x, z)(y - z)\mu(dx)r_\infty(x, dz, dy) = 0$ for any continuous and bounded function $\varphi : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$. Hence, for (X, Z, Y) distributed according to $\mu(dx)r_\infty(x, dz, dy)$, $Z = \mathbb{E}[Y|(X, Z)]$ and $\mathcal{T}_\infty(X) = \mathbb{E}[Y|X] = \mathbb{E}[\mathbb{E}[Y|(X, Z)]|X] = \mathbb{E}[Z|X]$. By using Jensen inequality for the conditional expectation, we get

$$\int_{\mathbb{R}^d} \phi(z)\eta_\infty(dz) \leq \int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} \phi(z)\mu(dx)r_\infty(x, dz, dy) = \lim_{n \rightarrow \infty} \int_{\mathbb{R}^d} \phi(z)\eta_n(dz).$$

Thus, η_∞ satisfies $\int_{\mathbb{R}^d} \phi(z)\eta_\infty(dz) = \inf_{\eta \in \mathcal{I}_\mu^\nu} \int_{\mathbb{R}^d} \phi(z)\eta(dz)$. Hence $\mathcal{I}_{\mu, \phi}^\nu \neq \emptyset$.

Let $\eta \in \mathcal{I}_{\mu, \phi}^\nu$. We now check that $\eta \in \tilde{\mathcal{I}}_\mu^\nu$ and $\Pi^{opt}(\mu, \eta)$ is a singleton. Let $\mu(dx)q(x, dz) \in \Pi^{opt}(\mu, \eta)$ and $\eta(dz)m(z, dy)$ be a martingale coupling between η and ν . By the second assertion in Proposition 2.7, $\mu(dx)qm(x, dy) \in \Pi^{opt}(\mu, \nu)$ and, by the first assertion, for $\mathcal{T}(x) = \int_{\mathbb{R}^d} yqm(x, dy)$, $\mathcal{T}\#\mu \in \tilde{\mathcal{I}}_\mu^\nu$. By the martingale property of m , $\mathcal{T}(x) = \int_{\mathbb{R}^d} zq(x, dz)$ so that $\mathcal{T}\#\mu \leq_{cx} \eta$. Since $\mathcal{T}\#\mu \in \mathcal{I}_\mu^\nu$ and $\eta \in \mathcal{I}_{\mu, \phi}^\nu$ implies that $\int_{\mathbb{R}^d} \phi(z)\mathcal{T}\#\mu(dz) \geq \int_{\mathbb{R}^d} \phi(z)\eta(dz)$, we deduce with the strict convexity of ϕ that $\eta = \mathcal{T}\#\mu$ and $\mu(dx)q(x, dz) = \mu(dx)\delta_{\mathcal{T}(x)}(dz)$. Hence any coupling in $\Pi^{opt}(\mu, \eta)$ is given by a map. By the second statement in Lemma 2.1, we conclude that this set is a singleton.

By repeating the first argument with $(\phi, \mathcal{I}_\mu^\nu)$ replaced by $(|x|^2, \mathcal{I}_{\mu, \phi}^\nu)$, we obtain the existence of $\eta_\phi \in \mathcal{I}_\mu^\nu$ such that $\int_{\mathbb{R}^d} |z|^2\eta_\phi(dz) \leq \inf_{\eta \in \mathcal{I}_{\mu, \phi}^\nu} \int_{\mathbb{R}^d} |z|^2\eta(dz)$. Since the construction also reduces the integral of ϕ , $\eta_\phi \in \mathcal{I}_{\mu, \phi}^\nu$.

Let us now check that if $\tilde{\eta} \in \mathcal{I}_{\mu, \phi}^\nu$ is such that $\int_{\mathbb{R}^d} |z|^2\tilde{\eta}(dz) = \inf_{\eta \in \mathcal{I}_{\mu, \phi}^\nu} \int_{\mathbb{R}^d} |z|^2\eta(dz)$, then $\tilde{\eta} = \eta_\phi$. By the first statement, $\Pi^{opt}(\mu, \eta_\phi) = \{(I_d, T_\phi)\#\mu\}$ and $\Pi^{opt}(\mu, \tilde{\eta}) = \{(I_d, \tilde{T})\#\mu\}$ for measurable transport maps T_ϕ and $\tilde{T} : \mathbb{R}^d \rightarrow \mathbb{R}^d$. One has $\int_{\mathbb{R}^d} |z|^2\eta_\infty(dz) = \int_{\mathbb{R}^d} |z|^2\tilde{\eta}(dz)$ and therefore, since $\eta_\phi, \tilde{\eta} \in \mathcal{I}_\mu^\nu$, $W_2^2(\mu, \eta_\phi) = W_2^2(\mu, \tilde{\eta})$. Let now $\bar{\eta} = \frac{\eta_\phi + \tilde{\eta}}{2}$. One has $\int_{\mathbb{R}^d} |z|^2\bar{\eta}(dz) = \int_{\mathbb{R}^d} |z|^2\eta_\phi(dz) = \int_{\mathbb{R}^d} |z|^2\tilde{\eta}(dz)$. The coupling $\mu(dx)\frac{1}{2}(\delta_{T_\phi(x)}(dz) + \delta_{\tilde{T}(x)}(dz))$ between μ and $\bar{\eta}$ implies that $W_2^2(\mu, \bar{\eta}) \leq W_2^2(\mu, \eta_\phi) = W_2^2(\mu, \tilde{\eta})$. Since $\eta_\phi \in \mathcal{I}_\mu^\nu$, we deduce that

$$W_2^2(\mu, \nu) \geq W_2^2(\mu, \bar{\eta}) + \int_{\mathbb{R}^d} |y|^2\nu(dy) - \int_{\mathbb{R}^d} |z|^2\bar{\eta}(dz).$$

Moreover, $\bar{\eta} \leq_{cx} \nu$ and combining a coupling in $\Pi^{opt}(\mu, \bar{\eta})$ with a martingale coupling between $\bar{\eta}$ and ν , we deduce that the previous inequality is an equality so that $\bar{\eta} \in \mathcal{I}_\mu^\nu$ and $\mu(dx)\frac{1}{2}(\delta_{T_\phi(x)}(dz) + \delta_{\tilde{T}(x)}(dz)) \in \Pi^{opt}(\mu, \bar{\eta})$. As $\eta_\phi, \tilde{\eta} \in \mathcal{I}_{\mu, \phi}^\nu$, $\int_{\mathbb{R}^d} \phi(z)\bar{\eta}(dz) = \inf_{\eta \in \mathcal{I}_{\mu, \phi}^\nu} \int_{\mathbb{R}^d} \phi(z)\eta(dz)$ and $\bar{\eta} \in \mathcal{I}_{\mu, \phi}^\nu$. By the first assertion, $\Pi^{opt}(\mu, \bar{\eta}) = \{(I_d, \bar{T})\#\mu\}$ for some measurable transport map $\bar{T} : \mathbb{R}^d \rightarrow \mathbb{R}$. Therefore $\mu(dx)$ a.e., $T_\phi(x) = \bar{T}(x)$ and $\eta_\phi = \tilde{\eta}$. For the choice $\phi(x) = |x|^2$, we deduce that $\mathcal{I}_{\mu, |x|^2}^\nu$ is a singleton. $\square$

From the equality $W_2^2(\mu, \nu) = W_2^2(\mu, \eta) + \int_{\mathbb{R}^d} |y|^2\nu(dy) - \int_{\mathbb{R}^d} |z|^2\eta(dz)$ valid for $\eta \in \mathcal{I}_\mu^\nu$, we see that minimizing $\int_{\mathbb{R}^d} |z|^2\eta(dz)$ over $\mathcal{I}_\mu^\nu$ is equivalent to minimizing $W_2^2(\mu, \eta)$. Therefore the probability measure $\underline{\eta}$ can be seen as the W_2 -projection of μ on the set $\mathcal{I}_\mu^\nu$. It is in general different from the W_2 -projection $\mu_{\underline{\mathcal{P}}(\nu)}$ of μ on the set $\underline{\mathcal{P}}(\nu) := \{\eta : \eta \leq_{cx} \nu\}$, which has been studied recently in dimension $d = 1$ by Gozlan *et al.* [13] and in general dimension d by Alfonsi *et al.* [1] (who also give an explicit formula for the antiderivative of the quantile function of this projection when $d = 1$), Alibert *et al.* [2], Gozlan and Juillet [12] and Backhoff-Veraguas *et al.* [5]. Notice that since $\mathcal{I}_\mu^\nu \subset \underline{\mathcal{P}}(\nu)$, one always has $W_2(\mu, \mu_{\underline{\mathcal{P}}(\nu)}) \leq W_2(\mu, \underline{\eta})$.

Example 2.14. For μ and ν the respective uniform distributions on $[0, 1]$ and $[0, 2]$, we have $\mathcal{I}_\mu^\nu = \{\nu\}$ and thus $\underline{\eta} = \nu$. By using the characterization in Theorem 2.6 [1], we obtain that the W_2 -projection $\mu_{\underline{\mathcal{P}}(\nu)}$ of μ on the set $\underline{\mathcal{P}}(\nu)$ is the uniform distribution on $[1/2, 3/2]$.

The next example shows that the set

$$\left\{ \eta_\phi : \phi : \mathbb{R}^d \rightarrow \mathbb{R}^d \text{ strictly convex and such that } \int_{\mathbb{R}^d} \phi(y) \nu(dy) < \infty \right\}$$

may contain distinct elements.

Example 2.15. Let $\mu = \frac{1}{2}(\delta_{(-1,0)} + \delta_{(1,0)})$ and $\nu = \frac{1}{4}(\delta_{(-1,-1)} + \delta_{(0,-1)} + \delta_{(0,1)} + \delta_{(1,1)})$. Any optimal coupling between μ and ν can be written as $\mu(dx)k_p(x, dy)$ with $k_p((-1, 0), dy) = \frac{1}{2}(\delta_{(-1,-1)} + p\delta_{(0,-1)} + (1-p)\delta_{(0,1)})(dy)$ and $k_p((1, 0), dy) = \frac{1}{2}(\delta_{(1,1)} + (1-p)\delta_{(0,-1)} + p\delta_{(0,1)})(dy)$ for $p \in [0, 1]$. One has $\mathcal{T}_p((-1, 0)) = (-1/2, -p)$ and $\mathcal{T}_p((1, 0)) = (1/2, p)$. The measures $\eta_p = \frac{1}{2}(\delta_{(-1/2,-p)} + \delta_{(1/2,p)})$ are not comparable for the convex order since for $p \neq p'$ there is no martingale coupling between η_p and $\eta_{p'}$. Moreover, for each $p \in [0, 1]$ the unique optimal transport plan $\delta_{((-1,0),(-1/2,-p))} + \delta_{((1,0),(1/2,p))}$ between μ and η_p is given by a map. For this example, $\underline{\eta} = \eta_0 = \frac{1}{2}(\delta_{(-1/2,0)} + \delta_{(1/2,0)})$ and $\eta_p = \eta_{\phi_p}$, with $\phi_p(x) = x_1^2 + (x_2 - 2px_1)^2$. The W_2 -optimal couplings between $\underline{\eta}$ and ν can be written as $\eta_0(dz)k_p(2z, dy)$ for $p \in [0, 1]$ and in particular the unique martingale coupling $\eta_0(dz)k_0(2z, dy)$ is optimal.

The last example shows that, unlike in the previous one, the martingale couplings between $\underline{\eta}$ and ν are not necessarily W_2 -optimal (even when $\Pi^{opt}(\mu, \nu)$ is a singleton).

Example 2.16. Let $\mu = \frac{1}{2}(\delta_{(-1,0)} + \delta_{(1,0)})$, $\nu_a = \frac{1}{4}(\delta_{(-1,-1)} + \delta_{(-1,2a+1)} + \delta_{(1,-2a-1)} + \delta_{(1,1)})$ with $a \in \mathbb{R}$. The unique W_2 -optimal coupling between μ and ν_a is $\mu(dx)k_a(x, dy)$ with $k_a((-1, 0), dy) = \frac{1}{2}(\delta_{(-1,-1)} + \delta_{(-1,2a+1)})(dy)$ and $k_a((1, 0), dy) = \frac{1}{2}(\delta_{(1,-2a-1)} + \delta_{(1,1)})(dy)$ so that $\eta_a = \frac{1}{2}(\delta_{(-1,a)} + \delta_{(1,-a)})$. Since $|(-1, -1) - (-1, a)|^2 - |(1, 1) - (-1, a)|^2 = (a+1)^2 - 4 - (a-1)^2 = 4(a-1)$, for $a > 1$,

$$\begin{aligned} W_2^2(\eta_a, \nu_a) &= \frac{1}{2}((a+1)^2 + 4 + (a-1)^2) < (a+1)^2 = \frac{1}{2}(3 + (2a+1)^2) - (1+a^2) \\ &= \int |y|^2 \nu_a(dy) - \int |z|^2 \eta_a(dz), \end{aligned}$$

so that the martingale coupling between η_a and ν_a is not W_2 -optimal.

3. DIFFERENTIABILITY OF THE SQUARED QUADRATIC WASSERSTEIN DISTANCE

We now present the notion of differentiability introduced by Lions [15]. Let $f : \mathcal{P}_2(\mathbb{R}^d) \rightarrow \mathbb{R}$. We consider an atomless probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and denote by $L^2(\Omega, \mathbb{P}; \mathbb{R}^d)$ the set of $\mathbb{R}^d$ -valued square integrable random variables on this space. The lift of the function f on $L^2(\Omega, \mathbb{P}; \mathbb{R}^d)$ is the function $F : L^2(\Omega, \mathbb{P}; \mathbb{R}^d) \rightarrow \mathbb{R}$ such that

$$\forall X \in L^2(\Omega, \mathbb{P}; \mathbb{R}^d), \quad F(X) = f(\mathcal{L}(X)),$$

where $\mathcal{L}(X) \in \mathcal{P}_2(\mathbb{R}^d)$ is the probability distribution of X . The atomless property is equivalent to the existence of a random variable $U : \Omega \rightarrow \mathbb{R}$ uniformly distributed on $[0, 1]$ (see *e.g.* [9], Prop. A.27). By the fundamental Theorem of simulation (see *e.g.* Bouleau and Lépingle [6], Thm. A.3.1, p. 38), it ensures the existence on $(\Omega, \mathcal{A}, \mathbb{P})$ of a random variable distributed according to each probability measure on each Polish space, and in particular of $X : \Omega \rightarrow \mathbb{R}^d$ distributed according to μ , for each $\mu \in \mathcal{P}_2(\mathbb{R}^d)$.

Definition 3.1. A function $f : \mathcal{P}_2(\mathbb{R}^d) \rightarrow \mathbb{R}$ is L -differentiable at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ if there exists $X \in L^2(\Omega, \mathbb{P}; \mathbb{R}^d)$ such that $X \sim \mu$ and F is Fréchet differentiable at X .

Let $f : \mathcal{P}_2(\mathbb{R}^d) \rightarrow \mathbb{R}$ and $F(X) = f(\mathcal{L}(X))$ for $X \in L^2(\Omega, \mathbb{P}; \mathbb{R}^d)$. The Fréchet differentiability of F at X amounts to the existence of a bounded linear operator $D_X^F : L^2(\Omega, \mathbb{P}; \mathbb{R}^d) \rightarrow \mathbb{R}$ such that $F(X + Y) = F(X) +$

$D_X^F(Y) + \|Y\|_{2\varepsilon_X(Y)}$, where $\varepsilon_X(Y) \rightarrow 0$ as $\|Y\|_2 \rightarrow 0$. By the Riesz representation theorem, there is a unique $DF(X) \in L^2(\Omega, \mathbb{P}; \mathbb{R}^d)$ such that $\forall Y \in L^2(\Omega, \mathbb{P}; \mathbb{R}^d)$, $D_X^F(Y) = \mathbb{E}[DF(X).Y]$, and we will call later on $DF(X)$ the Fréchet derivative of F at X . From Theorem 6.2 in [7], if f is L -differentiable at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, then F is Fréchet differentiable at X for all $X \in L^2(\Omega, \mathbb{P}; \mathbb{R}^d)$ such that $\mu = \mathcal{L}(X)$. Besides, the law of $(X, DF(X))$ does not depend on X by Proposition 5.24 in [8]. According to Theorem 6.5 [7], under the additional continuous differentiability assumption, the Fréchet derivative $DF(X)$ is equal to $g(X)$ for some measurable function g . According to Corollary 3.22 [10], the continuous differentiability assumption is not needed. We will provide a simple and direct proof of this result, see Lemma 3.4.

We now state the main result of this section that characterizes the differentiability of the square quadratic Wasserstein distance. To do so, we first exhibit the lift of the Wasserstein distance. Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$. From the atomless property, there exist random variables $X \sim \mu$ and $Y \sim \nu$ on $(\Omega, \mathcal{A}, \mathbb{P})$. The dual formulation (see for instance [18], Thm. 5.10)

$$W_2^2(\mu, \nu) = \sup_{\psi \in L^1(\mu), \tilde{\psi} \in L^1(\nu) : \psi(x) + \tilde{\psi}(y) \leq |x-y|^2} \mathbb{E} [\psi(X) + \tilde{\psi}(Y)] =: \mathbb{W}_2^2(X, Y) \quad (3.1)$$

permits to lift W_2^2 to $L^2(\Omega, \mathbb{P}; \mathbb{R}^d)$.

Theorem 3.2. *For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, the mapping $\mathcal{P}_2(\mathbb{R}^d) \ni \sigma \mapsto W_2^2(\sigma, \nu)$ is L -differentiable at μ iff there exists a measurable function $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ such that $\Pi^{opt}(\mu, \nu) = \{(I_d, T) \# \mu\}$ and then the Fréchet derivative of the function $Z \mapsto \mathbb{W}_2^2(Z, Y)$ at $X \sim \mu$ is given by $2(X - T(X))$.*

Remark 3.3.

- In particular, since the only coupling $\pi \in \Pi^{opt}(\nu, \nu)$ is $(I_d, I_d) \# \nu$, $\mathcal{P}_2(\mathbb{R}^d) \ni \sigma \mapsto W_2^2(\sigma, \nu)$ is differentiable at ν with a vanishing Fréchet derivative.
- According to Proposition 2.2, if ν is not a Dirac mass, then there is no $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ such that $\mathcal{P}_2(\mathbb{R}^d) \ni \sigma \mapsto W_2^2(\sigma, \nu)$ is differentiable in a neighbourhood of μ .

The L -differentiability is equivalent to the geometric differentiability ([10], Cor. 3.22). As explained in the introduction, the sufficient condition in Theorem 3.2 can be deduced from this equivalence, Theorem 10.2.6 [4] and Proposition 4.3 [3].

We are going to give a probabilistic proof of Theorem 3.2 by working with the L -differentiability. The two following lemmas are needed: the first one is used to get the necessary condition while the second is used for the sufficient condition. Their proofs are postponed after the proof of the theorem.

Lemma 3.4. *Let $F : L^2(\Omega, \mathbb{P}; \mathbb{R}^d) \rightarrow \mathbb{R}$ be law invariant. If F is Fréchet differentiable at $X \sim \mu$, then its Fréchet derivative is equal to $g(X)$ for some measurable function $g \in L^2(\mathbb{R}^d, \mu; \mathbb{R}^d)$ and it is differentiable with Fréchet derivative $g(\tilde{X})$ at each $\tilde{X} \sim \mu$ in $L^2(\Omega, \mathbb{P}; \mathbb{R}^d)$.*

Let us note that this result is also a consequence of the work by Gangbo and Tudorascu [10]. Here, we provide an alternative simple probabilistic proof of this fact. Wu and Zhang ([19], Prop. 1) already gave a different probabilistic proof when X is discrete.

Lemma 3.5. *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ be such that there exists $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ measurable such that $\Pi^{opt}(\mu, \nu) = \{(I_d, T) \# \mu\}$. Let also $(\mu_n)_n$ be a sequence of elements of $\mathcal{P}_2(\mathbb{R}^d)$ converging weakly to μ and such $\lim_{n \rightarrow \infty} W_2(\mu_n, \nu) = W_2(\mu, \nu)$. If (on a single probability space), $X \sim \mu$ and for $n \in \mathbb{N}$, (X_n, Y_n) is such that $X_n \sim \mu_n$, $Y_n \sim \nu$, $W_2^2(\mu_n, \nu) = \mathbb{E}[|X_n - Y_n|^2]$ and $X_n \xrightarrow{\text{Pr}} X$ as $n \rightarrow \infty$, then*

$$\lim_{n \rightarrow \infty} \mathbb{E}[|X_n - X|^2 + |Y_n - T(X)|^2] = 0.$$

Remark 3.6. The fact that $\lim_{n \rightarrow \infty} \mathbb{E}[|X_n - X|^2] = 0$ implies that $\lim_{n \rightarrow \infty} W_2(\mu_n, \mu) = 0$.

Proof of Theorem 3.2. Let us first assume $\Pi^{opt}(\mu, \nu) \neq \{(I_d, T) \# \mu\}$. The existence on the lifted probability space of a random variable uniformly distributed on $[0, 1]$ combined with [6] Theorem A.3.1 and Lemma 2.1 ensures the existence on this space of (X, Y) with $X \sim \mu$, $Y \sim \nu$, $W_2^2(\mu, \nu) = \mathbb{E}[|Y - X|^2]$ and $\mathbb{E}[|Y - \mathbb{E}[Y|X]|^2] > 0$. Let $\xi = Y - \mathbb{E}[Y|X]$. One has $\mathbb{E}[\xi|X] = 0$ a.s. so that for $h : \mathbb{R}^d \rightarrow \mathbb{R}^d$ measurable and such that $h(X)$ is square integrable,

$$\mathbb{E}[h(X) \cdot \xi] = \mathbb{E}[h(X) \cdot \mathbb{E}[\xi|X]] = 0. \quad (3.2)$$

On the other hand, denoting by μ_n the distribution of $X + \xi_n$ where $\xi_n = \frac{\xi}{n}$, we have

$$\begin{aligned} \mathbb{W}_2^2(X + \xi_n, Y) &= W_2^2(\mu_n, \nu) \leq \mathbb{E}[|X + \xi_n - Y|^2] = \mathbb{E}[|X - Y|^2] + \frac{2}{n} \mathbb{E}[(X - Y) \cdot \xi] + \frac{\mathbb{E}[|\xi|^2]}{n^2} \\ &= W_2^2(\mu, \nu) - \frac{2}{n} \mathbb{E}[Y \cdot \xi] + \frac{\mathbb{E}[|\xi|^2]}{n^2} \\ &= \mathbb{W}_2^2(X, Y) - \left(\frac{2}{n} - \frac{1}{n^2} \right) \mathbb{E}[|Y - \mathbb{E}[Y|X]|^2], \end{aligned}$$

where we used (3.2) for the second equality and the definition of ξ for the third. If $\sigma \mapsto W_2^2(\sigma, \nu)$ were L -differentiable at μ , then (3.2) combined with Lemma 3.4 would imply that, as $n \rightarrow \infty$,

$$\mathbb{W}_2^2(X + \xi_n, Y) - \mathbb{W}_2^2(X, Y) = o(\|\xi_n\|_2),$$

which does not hold since $\|\xi_n\|_2 = \frac{\mathbb{E}^{1/2}[|Y - \mathbb{E}[Y|X]|^2]}{n}$.

Now, we assume that $\Pi^{opt}(\mu, \nu) = \{(I_d, T) \# \mu\}$ for some measurable transport map $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$. Let, on the lifted probability space, $X \sim \mu$, $Y \sim \nu$ and $(\xi_n)_n$ be a sequence of square integrable $\mathbb{R}^d$ -valued random vectors such that $\|\xi_n\|_2 := \mathbb{E}^{1/2}[|\xi_n|^2]$ tends to 0 as $n \rightarrow \infty$. We denote by μ_n the law of $X + \xi_n$. Let $Y_n \sim \nu$ such that $W_2^2(\mu_n, \nu) = \mathbb{E}[|X + \xi_n - Y_n|^2]$ be defined on a possible enlargement of the lifted probability space. We have

$$\begin{aligned} W_2^2(\mu_n, \nu) &\leq \mathbb{E}[|X + \xi_n - T(X)|^2] = \mathbb{E}[|X - T(X)|^2] + 2\mathbb{E}[(X - T(X)) \cdot \xi_n] + \mathbb{E}[|\xi_n|^2] \\ &= W_2^2(\mu, \nu) + 2\mathbb{E}[(X - T(X)) \cdot \xi_n] + \mathbb{E}[|\xi_n|^2]. \end{aligned}$$

On the other hand,

$$\begin{aligned} W_2^2(\mu, \nu) &\leq \mathbb{E}[|X - Y_n|^2] = \mathbb{E}[|X + \xi_n - Y_n|^2] - 2\mathbb{E}[(X - Y_n) \cdot \xi_n] - \mathbb{E}[|\xi_n|^2] \\ &= W_2^2(\mu_n, \nu) - 2\mathbb{E}[(X - T(X)) \cdot \xi_n] - \mathbb{E}[|\xi_n|^2] + 2\mathbb{E}[(Y_n - T(X)) \cdot \xi_n]. \end{aligned}$$

With Cauchy-Schwarz inequality, we deduce that

$$|W_2^2(\mu_n, \nu) - W_2^2(\mu, \nu) - 2\mathbb{E}[(X - T(X)) \cdot \xi_n]| \leq \|\xi_n\|_2 (\|\xi_n\|_2 + \|Y_n - T(X)\|_2).$$

Note that from (3.1) the left-hand side is equal to $|\mathbb{W}_2^2(X + \xi_n, Y) - \mathbb{W}_2^2(X, Y) - 2\mathbb{E}[(X - T(X)) \cdot \xi_n]|$ and is thus well defined on the original lifted probability space, as required by the definition of the Lions derivative. Now, Lemma 3.5 applied with $X_n = X + \xi_n$ ensures that $\lim_{n \rightarrow \infty} \|Y_n - T(X)\|_2 = 0$ so that $\sigma \mapsto W_2^2(\sigma, \nu)$ is L -differentiable at μ with $\partial_\mu W_2^2(\mu, \nu)(x) = 2(x - T(x))$. $\square$

Proof of Lemma 3.4. By the fundamental theorem of simulation (see e.g. Bouleau and Lépingle [6], Thm. A.3.1, p. 38), since the lifted probability space supports a random variable with uniform distribution on $[0, 1]$, it also supports a couple $(\tilde{X}, U)$ with $\tilde{X} \sim \mu$ and U an independent random variable uniformly distributed on $[0, 1]$. Since F is Fréchet differentiable at $X \sim \mu$, by Theorem 6.2 [7] it is also Fréchet differentiable at

$\tilde{X}$. Let for $i \in \{1, \dots, d\}$, $DF(\tilde{X})_i$ denote the i th coordinate of $DF(\tilde{X})$ and $P_i(x, dz, du)$ with respective marginals $Q_i(x, dz)$ and $R_i(x, du)$ denote a regular version of the conditional law of $(DF(\tilde{X})_i, U)$ given $\tilde{X} = x$. Let $g_i(x) = \inf\{z \in \mathbb{R} : Q_i(x, (-\infty, z]) \geq 1/2\}$ be the median of $Q_i(x, dz)$. Notice that, by property of the median, $\mathbb{E}[|DF(\tilde{X})_i - g_i(\tilde{X})| | \tilde{X}] \leq \mathbb{E}[|DF(\tilde{X})_i - \mathbb{E}[DF(\tilde{X})_i | \tilde{X}]| | \tilde{X}]$ so that

$$\begin{aligned} \mathbb{E}[|g_i(\tilde{X})|] &\leq \mathbb{E}[|DF(\tilde{X})_i|] + \mathbb{E}[|DF(\tilde{X})_i - g_i(\tilde{X})|] \\ &\leq \mathbb{E}[|DF(\tilde{X})_i|] + \mathbb{E}[|DF(\tilde{X})_i - \mathbb{E}[DF(\tilde{X})_i | \tilde{X}]|] < \infty. \end{aligned}$$

Let

$$v_i^\pm(x) = \inf\{u \in [0, 1] : P_i(x, \{g_i(x)\} \times [0, u]) \geq (Q_i(x, (-\infty, g_i(x))) - Q_i(x, (g_i(x), +\infty)))^\pm\}.$$

By independence of $\tilde{X}$ and U , there is a Borel subset A of $\mathbb{R}^d$ with $\mu(A) = 0$ such that for $x \notin A$, $R_i(x, du)$ is the Lebesgue measure on $[0, 1]$. Since

$$Q_i(x, (-\infty, g_i(x))) \vee Q_i(x, (g_i(x), +\infty)) \leq \frac{1}{2} \leq Q_i(x, (-\infty, g_i(x))) \wedge Q_i(x, (g_i(x), +\infty)) + Q_i(x, \{g_i(x)\}),$$

for $x \notin A$, $P_i(x, \{g_i(x)\} \times [0, v_i^\pm(x)]) = (Q_i(x, (-\infty, g_i(x))) - Q_i(x, (g_i(x), +\infty)))^\pm$.

The random variables $\xi_+^i = 1_{\{DF(\tilde{X})_i > g_i(\tilde{X})\}} + 1_{\{DF(\tilde{X})_i = g_i(\tilde{X}), U \leq v_i^+(\tilde{X})\}}$ and $\xi_-^i = 1_{\{DF(\tilde{X})_i < g_i(\tilde{X})\}} + 1_{\{DF(\tilde{X})_i = g_i(\tilde{X}), U \leq v_i^-(\tilde{X})\}}$ are such that $(\tilde{X}, \xi_+^i)$ and $(\tilde{X}, \xi_-^i)$ have the same distribution: indeed, conditionally on $\tilde{X} = x$, these are Bernoulli random variables of parameter $Q_i(x, (-\infty, g_i(x))) \vee Q_i(x, (g_i(x), +\infty))$. Therefore $\mathbb{E}[g_i(\tilde{X})\xi_+^i] = \mathbb{E}[g_i(\tilde{X})\xi_-^i]$ and, denoting by e_i the i th vector of the canonical basis of $\mathbb{R}^d$, for each $\varepsilon \in [0, 1]$, $\tilde{X} + \varepsilon\xi_+^i e_i$ and $\tilde{X} + \varepsilon\xi_-^i e_i$ have the same distribution so that $F(\tilde{X} + \varepsilon\xi_+^i e_i) = F(\tilde{X} + \varepsilon\xi_-^i e_i)$. Hence $\mathbb{E}(\xi_+^i DF(\tilde{X})_i) = \mathbb{E}(\xi_-^i DF(\tilde{X})_i)$. We deduce that

$$0 = \mathbb{E}[(DF(\tilde{X})_i - g_i(\tilde{X}))(\xi_+^i - \xi_-^i)] = \mathbb{E}[|DF(\tilde{X})_i - g_i(\tilde{X})|]$$

and conclude that $\mathbb{P}(DF(\tilde{X}) = g(\tilde{X})) = 1$. Proposition 5.24 [8] ensures that the couples $(X, DF(X))$ and $(\tilde{X}, DF(\tilde{X}))$ share the same distribution and therefore $\mathbb{P}(DF(X) = g(X)) = 1$. $\square$

Proof of Lemma 3.5. Let η_n and η_{34}^n respectively denote the distributions of $(X, T(X), X_n, Y_n)$ and (X_n, Y_n) . Since $(\mu_n)_n$ converges weakly to μ , this sequence is tight and we deduce that $(\eta_n)_n$ is tight. Let us consider a subsequence weakly converging to η^∞ and that we still index by n for notational simplicity. From the convergence $X_n \xrightarrow{\text{Pr}} X$ as $n \rightarrow \infty$, we deduce that $(X, T(X), X_n) \xrightarrow{\text{Pr}} (X, T(X), X)$. Hence the marginal η_{123}^∞ of the triplet of the three first coordinates under η^∞ is $\eta_{123}^\infty = (I_d, T, I_d) \# \mu$. Next, the marginal η_{34}^∞ of the couple of the two last coordinates is a coupling between μ and ν such that $\int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^2 \eta_{34}^\infty(dx, dy) \leq \liminf_{n \rightarrow \infty} \int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^2 \eta_{34}^n(dx, dy)$. Since $\int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^2 \eta_{34}^n(dx, dy) = \mathbb{E}[|X_n - Y_n|^2] = W_2^2(\mu_n, \nu)$ converges to $W_2^2(\mu, \nu)$ as $n \rightarrow \infty$, $\eta_{34}^\infty \in \Pi^{opt}(\mu, \nu)$. Therefore $\eta_{34}^\infty = (I_d, T) \# \mu$ and $\mu(dw)$ a.e. the conditional law of the fourth coordinate given that the third is equal to w is $\delta_{T(w)}(dz)$. Since under η_{123}^∞ the two first coordinates are a function of the third one, this is also the conditional law of the fourth coordinate given that the three first coordinates are equal to $(w, T(w), w)$. Hence $\eta^\infty(dx, dy, dw, dz) = \mu(dw) \delta_{(w, T(w))}(dx, dy) \delta_{T(w)}(dz)$ so that $\eta^\infty = (I_d, T, I_d, T) \# \mu$. Since the weak limit does not depend on the subsequence, the whole sequence $(\eta_n)_n$ converges weakly to $(I_d, T, I_d, T) \# \mu$.

Since $|Y_n - T(X)| \leq 2|Y_n|1_{\{|Y_n| \geq |T(X)|\}} + 2|T(X)|1_{\{|Y_n| < |T(X)|\}}$, for $m > 0$, we have

$$|Y_n - T(X)|^2 1_{\{|Y_n - T(X)| \geq m\}} \leq 4|Y_n|^2 1_{\{|Y_n| \geq m/2\}} + 4|T(X)|^2 1_{\{|T(X)| \geq m/2\}}.$$

Hence

$$\mathbb{E} [|Y_n - T(X)|^2 1_{\{|Y_n - T(X)| \geq m\}}] \leq 8 \int_{\mathbb{R}^d} |x|^2 1_{\{|x| \geq m/2\}} \nu(dx)$$

which provides the uniform integrability needed to conclude that $\lim_{n \rightarrow \infty} \mathbb{E} [|Y_n - T(X)|^2] = 0$. The convergence of $\mathbb{E} [|X_n - Y_n|^2] = W_2^2(\mu_n, \nu)$ to $W_2^2(\mu, \nu) = \mathbb{E} [|X - T(X)|^2]$ as $n \rightarrow \infty$ together with the convergence in probability of (X_n, Y_n) to $(X, T(X))$ implies that

$$\lim_{n \rightarrow \infty} \mathbb{E} [|X - T(X)|^2 - |X_n - Y_n|^2] = 0$$

and therefore that the random variables $(|X_n - Y_n|^2)_n$ are uniformly integrable. From the inequality $|X_n - X|^2 \leq 3(|X_n - Y_n|^2 + |Y_n|^2 + |X|^2)$ and the convergence of $(Y_n)_n$ to $T(X)$ in quadratic mean, we deduce that the random variables $(|X_n - X|^2)_n$ are uniformly integrable. With the convergence in probability of $(X_n)_n$ to X , we conclude that $\lim_{n \rightarrow \infty} \mathbb{E} [|X_n - X|^2] = 0$. $\square$

Lemma 3.4 is also useful to check that the notion of L -differentiability does not depend on the choice of the atomless lifted probability space $(\Omega, \mathcal{A}, \mathbb{P})$. We will say that a function $f : \mathcal{P}_2(\mathbb{R}^d) \rightarrow \mathbb{R}$ is L_Ω -differentiable at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ if there exists $X \in L^2(\Omega, \mathbb{P}; \mathbb{R}^d)$ such that $\mu = \mathcal{L}(X)$ and $F_\Omega(X) = f(\mathcal{L}(X))$ is Fréchet differentiable at X . The function F_Ω is called the lift of f on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$.

Proposition 3.7. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ and $(\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{\mathbb{P}})$ be two atomless probability spaces. The function $f : \mathcal{P}_2(\mathbb{R}^d) \rightarrow \mathbb{R}$ is L_Ω -differentiable at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ iff it is $L_{\tilde{\Omega}}$ -differentiable at μ .*

Proof. By symmetry, it is enough to prove that the L_Ω -differentiability implies the $L_{\tilde{\Omega}}$ -differentiability and the Fréchet derivatives are given by the same function in $L^2(\mathbb{R}^d, \mu; \mathbb{R}^d)$. Let us assume that f is L_Ω -differentiable at μ . The atomless property and the fundamental theorem of simulation ensure the existence on the original lifted space $(\Omega, \mathcal{A}, \mathbb{P})$ of random variables (U, X) such that U is uniformly distributed on $[0, 1]$ and independent from $X \sim \mu$. Then, F_Ω is Fréchet differentiable at X and there exists a measurable function $g \in L^2(\mathbb{R}^d, \mu; \mathbb{R}^d)$ such that $DF_\Omega(X) = g(X)$ by Lemma 3.4. We consider $F_{\tilde{\Omega}} : L^2(\tilde{\Omega}, \tilde{\mathbb{P}}; \mathbb{R}^d) \rightarrow \mathbb{R}$ the lift of f on an atomless probability space $(\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{\mathbb{P}})$ and $\tilde{X} \sim \mu$ under $\tilde{\mathbb{P}}$. Let $\tilde{Y} \in L^2(\tilde{\Omega}, \tilde{\mathbb{P}}; \mathbb{R}^d)$ and $R(x, dy)$ denote a regular version of the conditional law of $\tilde{Y}$ given $\tilde{X} = x$. By Lemma 2.22 [14], there exists a measurable function $\rho : \mathbb{R}^d \times [0, 1] \rightarrow \mathbb{R}^d$ such that for all $x \in \mathbb{R}^d$, $\rho(x, U)$ is distributed according to $R(x, dy)$. Then, $Y = \rho(X, U)$ is such that (X, Y) has the same law under $\mathbb{P}$ as $(\tilde{X}, \tilde{Y})$ under $\tilde{\mathbb{P}}$, and therefore $Y \in L^2(\Omega, \mathbb{P}; \mathbb{R}^d)$. We then have

$$\begin{aligned} F_{\tilde{\Omega}}(\tilde{X} + \tilde{Y}) - F_{\tilde{\Omega}}(\tilde{X}) - \tilde{\mathbb{E}}[g(\tilde{X}) \cdot \tilde{Y}] &= F_\Omega(X + Y) - F_\Omega(X) - \mathbb{E}[g(X) \cdot Y] \\ &= F_\Omega(X + Y) - F_\Omega(X) - \mathbb{E}[DF_\Omega(X) \cdot Y]. \end{aligned}$$

With $\mathbb{E}[|Y|^2] = \tilde{\mathbb{E}}[|\tilde{Y}|^2]$, we deduce that the Fréchet differentiability of F_Ω at X implies the Fréchet differentiability of $F_{\tilde{\Omega}}$ at $\tilde{X}$ and $DF_{\tilde{\Omega}}[\tilde{X}] = g(\tilde{X})$. $\square$

By considering the atomless probability space $(B_d, \mathcal{B}(B_d), \text{Leb})$ where B_d the ball centered at the origin of unit volume in $\mathbb{R}^d$ endowed with the Borel sets and the Lebesgue measure, Gangbo and Tudorascu ([10], Cor. 3.22) have proved the equivalence between the L_{B_d} -differentiability (and therefore the L -differentiability) and the geometric differentiability, as well as the relation between the two derivatives.

As pointed by one of the referees, one can also deduce Proposition 3.7 from Corollary 3.22 [10] by Gangbo and Tudorascu: f is geometrically differentiable at μ if, and only if, F_{B_d} is Fréchet differentiable at any $X \sim \mu$, and in this case we have $DF_{B_d}(X) = \nabla_\mu f(X)$, where $\nabla_\mu f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is the Wasserstein gradient of f at μ . Let $(\Omega, \mathcal{A}, \mathbb{P})$ be an atomless probability space and F_Ω denote the corresponding lift of f . Then, there exists an almost isomorphism $i : B_d \rightarrow \Omega$ that pushes forward the Lebesgue measure to $\mathbb{P}$. We have $F_{B_d}(X \circ i) = F_\Omega(X)$ for $X \in L^2(\Omega, \mathbb{P}; \mathbb{R}^d)$. If $X \sim \mu$, then $X \circ i \sim \mu$ and one easily deduces that F_Ω is Fréchet differentiable at X with

$DF_{\Omega}(X) = DF_{B_d}(X \circ i) \circ i^{-1} = \nabla_{\mu} f(X \circ i) \circ i^{-1} = \nabla_{\mu} f(X)$. In a symmetric way, if F_{Ω} is Fréchet differentiable at $X \sim \mu$, then F_{B_d} is Fréchet differentiable at $X \circ i$ and $DF_{B_d}(X \circ i) = DF_{\Omega}(X) \circ i = \nabla_{\mu} f(X \circ i)$, so that $DF_{\Omega}(X) = \nabla_{\mu} f(X)$.

Acknowledgements. We thank Pierre Cardaliaguet for pointing out the reference [10] to us.

REFERENCES

- [1] A. Alfonsi, J. Corbetta and B. Jourdain, Sampling of probability measures in the convex order by Wasserstein projection. *Ann. Inst. Henri Poincaré Probab. Stat.* **56** (2020) 1706–1729.
- [2] J.-J. Alibert, G. Bouchitté and T. Champion, A new class of costs for optimal transport planning. *Eur. J. Appl. Math.* **30** (2019) 1229–1263.
- [3] L. Ambrosio and W. Gangbo, Hamiltonian ODEs in the Wasserstein space of probability measures. *Commun. Pure Appl. Math.* **61** (2008) 18–53.
- [4] L. Ambrosio, N. Gigli and G. Savaré, Gradient flows in metric spaces and in the space of probability measures. *Lectures in Mathematics* ETH Zürich. Birkhäuser Verlag, Basel (2005).
- [5] J. Backhoff-Veraguas, M. Beiglböck, G. Pammer, Existence, duality, and cyclical monotonicity for weak transport costs. *Calc. Var. Partial Differ. Equ.* **58** (2019) Art. 203.
- [6] N. Bouleau and D. Lépine, Numerical methods for stochastic processes. *Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. John Wiley & Sons, Inc., New York (1994).
- [7] P. Cardaliaguet, Notes on Mean-Field Games (from P.-L. Lions lectures at Collège de France). Available at: <https://www.ceremade.dauphine.fr/~cardaliaguet/MFG20130420.pdf> (2013).
- [8] R. Carmona and F. Delarue, Probabilistic theory of mean field games with applications. I. Mean field FBSDEs, control, and games. Vol. 83 of *Probability Theory and Stochastic Modelling*. Springer, Cham (2018).
- [9] H. Föllmer and A. Schied, Stochastic finance, An introduction in discrete time. Walter de Gruyter & Co., Berlin, third edition (2011).
- [10] W. Gangbo and A. Tudorascu, On differentiability in the Wasserstein space and well-posedness for Hamilton-Jacobi equations. *J. Math. Pures Appl.* **125** (2019) 119–174.
- [11] N. Gigli, On the inverse implication of Brenier-McCann theorems and the structure of $(\mathcal{P}_2(M), W_2)$. *Methods Appl. Anal.* **18** (2011) 127–158.
- [12] N. Gozlan and N. Juillet, On a mixture of Brenier and Strassen theorems. *Proc. Lond. Math. Soc.* **120** (2020) 434–463.
- [13] N. Gozlan, C. Roberto, P.-M. Samson, Y. Shu and P. Tetali, Characterization of a class of weak transport-entropy inequalities on the line. *Ann. Inst. Henri Poincaré Probab. Stat.* **54** (2018) 1667–1693.
- [14] O. Kallenberg, Foundations of modern probability. Probability and its Applications (New York). Springer-Verlag, New York (1997).
- [15] P.-L. Lions, Cours au Collège de France (2008).
- [16] F. Santambrogio, Optimal transport for applied mathematicians. In Vol. 87 of *Progress in Nonlinear Differential Equations and their Applications*. Birkhäuser/Springer (2015).
- [17] V. Strassen, The existence of probability measures with given marginals. *Ann. Math. Statist.* **36** (1965) 423–439.
- [18] C. Villani, Vol. 338 of Optimal transport, Old and New. Springer-Verlag (2009).
- [19] C. Wu and J. Zhang, An Elementary Proof for the Structure of Derivatives in Probability Measures. Preprint ArXiv [1705.08046](https://arxiv.org/abs/1705.08046) (2017).