PROCEEDINGS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 151, Number 9, September 2023, Pages 4089–4098
https://doi.org/10.1090/proc/16466
Article electronically published on May 25, 2023

A SMOOTH VARIATIONAL PRINCIPLE ON WASSERSTEIN SPACE

ERHAN BAYRAKTAR, IBRAHIM EKREN, AND XIN ZHANG

(Communicated by Amarjit Budhiraja)

ABSTRACT. In this note, we provide a smooth variational principle on Wasserstein space by constructing a smooth gauge-type function using the sliced Wasserstein distance. This function is a crucial tool for optimization problems and in viscosity theory of PDEs on Wasserstein space.

1. INTRODUCTION

This note is devoted to proving a smooth variational principle on Wasserstein space. Due to the lack of local compactness, a continuous function on an infinite dimensional space may not attain its local maxima/minima, which becomes an issue when dealing with optimization problems. Smooth variational principle provides a way to perturb the function smoothly so that its perturbation can attain its local extremas. Recently, smooth variational principles on Wasserstein space appeared in the study of viscosity solution of partial differential equations on Wasserstein space. A major effort in this direction was performed by [6].

For a continuous function on a separable Hilbert space, Ekeland's variational principle provides a smooth variation so that the perturbation attains local extremas; see e.g. [7]. However, in the Wasserstein space the variation part is given by the Wasserstein metric which is not smooth anymore. One of the observations in [6] was to use the Borwein-Preiss variational principle [3, Theorem 2.5.2], to have smooth variations on the Wasserstein space, which states that it is sufficient to construct a topologically equivalent complete metric which is differentiable in the sense of [4]. In this note, we achieve this using the sliced Wasserstein distance, which defines a metric between high dimensional probability distributions using their one dimensional projections; see e.g. [2] and page 214 of [8]. The advantage of our choice is that the optimal transport map in one dimension can be explicitly written down, and is regular after a Gaussian convolution. As such, our choice of the sliced Wasserstein distance allows a simple construction of smooth gauge type function compared to the alternative in [6]; see in particular Lemma 4.4 therein.

In the next subsection, we recall the definition of Wasserstein distance, and the L-derivative. Then in Section 2, we analyze the differential properties of Gaussian

Received by the editors September 29, 2022, and, in revised form, November 15, 2022, and February 9, 2023.

2020 *Mathematics Subject Classification.* Primary 58E30, 90C05.

The first author was partially supported by the National Science Foundation under grant DMS-2106556 and by the Susan M. Smith chair.

The second author was supported in part by NSF Grant DMS 2007826.

regularized sliced Wasserstein distance, and finally prove the smooth variational principle in Proposition 2.5.

1.1. **Wasserstein distance and L derivative.** We denote by $\mathcal{P}_2(\mathbb{R}^k)$ the set of Borel probability measures μ such that $\int |x|^2 \mu(dx) < \infty$. We endow the space $\mathcal{P}_2(\mathbb{R}^k)$ with the 2-Wasserstein distance W_2, i.e., for any $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^k)$

$$W_2(\mu,\nu)^2 := \inf_{\pi\in\Pi(\mu,\nu)} \int \frac{1}{2}|x-y|^2 \,\pi(dx,dy), \tag{1.1}$$

where $\Pi(\mu,\nu)$ denotes the collection of probability measures on $\mathbb{R}^k \times \mathbb{R}^k$ with first and second marginals μ and ν respectively.

Let us now present the L-derivative introduced in [4]; see [5, Chapter 5] for a survey. Let $u : \mathcal{P}_2(\mathbb{R}^k) \to \mathbb{R}$, and $(\Omega, \mathcal{P}, \mathbb{P})$ be an atomless probability space. The lifting U of u on the Hilbert space $L^2(\Omega, \mathcal{P}, \mathbb{P}; \mathbb{R}^k)$ is defined via

$$U(X) := u(\mathbb{P}_X), \quad \forall\, X \in L^2(\Omega, \mathcal{P}, \mathbb{P}; \mathbb{R}^k),$$

where $\mathbb{P}_X$ stands for the distribution of X. Recall that U is said to be Fréchet differentiable at some random variable $X \in L^2(\Omega, \mathcal{P}, \mathbb{P}; \mathbb{R}^k)$ if there exists a random variable $Z \in L^2(\Omega, \mathcal{P}, \mathbb{P}; \mathbb{R}^k)$ such that

$$\lim_{t\to 0} \frac{U(X+tY)-U(X)}{t} = \mathbb{E}[ZY], \quad \forall\, Y \in L^2(\Omega, \mathcal{P}, \mathbb{P}; \mathbb{R}^k),$$

and we denote this derivative Z by $DU(X)$.

Definition 1.1. A function $u : \mathcal{P}_2(\mathbb{R}^k) \to \mathbb{R}$ is said to be L-differentiable at μ if there exists some $X \in L^2(\Omega, \mathcal{P}, \mathbb{P}; \mathbb{R}^k)$ such that $\mathbb{P}_X = \mu$ and U is Fréchet differentiable at X. And u is said to be L-differentiable if there exists a jointly measurable function $D_\mu u : \mathcal{P}_2(\mathbb{R}^k) \times \mathbb{R}^k \to \mathbb{R}$ such that the lifting U is Fréchet differentiable at any $X \in L^2(\Omega, \mathcal{P}, \mathbb{P}; \mathbb{R}^k)$ and $DU(X) = D_\mu u(\mu, X)$.

It was proven in [1, Theorem 2.2] that if the optimizers of (1.1) are reduced to the set $\{(Id, T)\sharp\mu\}$ for some measurable function T, then $\mu \mapsto W_2(\mu,\nu)^2$ is L-differentiable at μ. Additionally, the Fréchet derivative of its lift at X with $X \sim \mu$ is $X - T(X)$. By the Brenier theorem the condition on the uniqueness of the optimizer is satisfied when μ is absolutely continuous.

2. Gaussian regularized sliced Wasserstein distance

Denote by $\mathcal{S}^{k-1}$ the unit sphere in $\mathbb{R}^k$. For any $\mu \in \mathcal{P}_2(\mathbb{R}^k)$, and $\theta \in \mathcal{S}^{k-1}$, define the mapping $P_\theta : \mathbb{R}^k \to \mathbb{R}$ by the expression $P_\theta(x) = x^\top\theta$ and the pushforward measure $\mu_\theta := P_\theta\sharp\mu \in \mathcal{P}_2(\mathbb{R})$. For any $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^k)$, the sliced Wasserstein distance is defined via

$$SW_2(\mu,\nu)^2 = \int W_2(\mu_\theta, \nu_\theta)^2 \, d\theta, \tag{2.1}$$

where the integration is with respect to the standard spherical measure on $\mathcal{S}^{k-1}$; see e.g. [2] and page 214 of [8]. Moreover, we consider the Gaussian regularized version

$$SW_2^\sigma(\mu,\nu) := SW_2(\mu^\sigma, \nu^\sigma), \tag{2.2}$$

where $\mu^\sigma := \mu * \mathcal{N}_\sigma$ and $\mathcal{N}_\sigma \in \mathcal{P}_2(\mathbb{R}^k)$ is the Normal distribution with variance $\sigma^2 I_k$ for some $\sigma \in (0,\infty)$. By abuse of notation, $\mathcal{N}_\sigma$ also denotes the one dimensional normal distribution (and its density) with mean 0 and variance σ^2, and then we have that $(\mu^\sigma)_\theta = (\mu * \mathcal{N}_\sigma)_\theta = \mu_\theta * (\mathcal{N}_\sigma)_\theta = \mu_\theta * \mathcal{N}_\sigma$ where the last Gaussian is one-dimensional and the previous one is k-dimensional.

Lemma 2.1. *For any $\sigma \geq 0$, $(\mathcal{P}_2(\mathbb{R}^k), SW_2^\sigma)$ is a complete metric space, and it is equal to $(\mathcal{P}_2(\mathbb{R}^k), W_2)$ as a topological space.*

Proof. Let us first prove the first claim. Take any Cauchy sequence $(\mu^n)_{n\geq 1}$ in $(\mathcal{P}_2(\mathbb{R}^k), SW_2^\sigma)$, and we can assume without loss of generality by taking a subsequence that

$$\sum_{n\geq 1} SW_2^\sigma(\mu^n, \mu^{n+1})^2 = \int \sum_{n\geq 1} W_2(\mu_\theta^n * \mathcal{N}_\sigma, \mu_\theta^{n+1} * \mathcal{N}_\sigma)^2 \, d\theta < +\infty.$$

Define $\mathcal{S} \subset \mathcal{S}^{k-1}$ to be the set of θ such that $\sum_{n\geq 1} W_2(\mu_\theta^n * \mathcal{N}_\sigma, \mu_\theta^{n+1} * \mathcal{N}_\sigma)^2$ is finite. Then it is clear that $\mathcal{S} \subset \mathcal{S}^{k-1}$ is of full spherical measure. Choose a finite subset $\{\theta(1), \ldots, \theta(F)\} \subset \mathcal{S}$ with the property that

$$|x|^2 \leq 2 \max_{i=1,\ldots,F} |x^\top \theta(i)|^2, \quad \forall\, x \in \mathbb{R}^k.$$

Then it can be easily seen that

$$\lim_{R\to\infty} \sup_{n\in\mathbb{N}} \int_{\{x\in\mathbb{R}^k : |x|\geq R\}} |x|^2 \, \mu^n * \mathcal{N}_\sigma(dx) = 0,$$

and hence $\{\mu^n * \mathcal{N}_\sigma\}_{n\geq 1}$ is tight with respect to W_2 topology. As in [6, Lemma 4.2], it can be shown that $(\mu^n)_{n\geq 1}$ is also tight, and has a limit $\nu \in \mathcal{P}_2(\mathbb{R}^k)$ with respect to the W_2 metric. Due to the inequality

$$W_2(\mu_\theta^n * \mathcal{N}_\sigma, \nu_\theta * \mathcal{N}_\sigma) \leq W_2(\mu_\theta^n, \nu_\theta) \leq W_2(\mu^n, \nu),$$

we conclude that μ^n converges to ν in SW_2^σ distance.

The above inequality implies that the topology generated by W_2 is stronger than that generated by SW_2^σ. By the argument in the first paragraph, for any sequence $(\mu^n)_{n\geq 1}$ such that $SW_2^\sigma(\mu^n, \nu) \to 0$ with some limit $\nu \in \mathcal{P}_2(\mathbb{R}^k)$, there is a tight subsequence that converges to ν in the W_2 distance. Therefore, SW_2^σ induces the same topology as SW_2 and W_2. □

The advantage of SW_2^σ is that we can easily compute its derivatives. Denote the cumulative distribution function of μ by F_μ. Then it is well known that in the one dimensional case, the optimal transport map from μ_θ^σ to ν_θ^σ is given by

$$T_\theta^\sigma(x) := F_{\nu_\theta^\sigma}^{-1}(F_{\mu_\theta^\sigma}(x)),$$

where $F_{\nu_\theta^\sigma}^{-1}(x) := \inf\{t \in \mathbb{R} : F_{\nu_\theta^\sigma}(t) \geq x\}$, $\forall x \in [0,1]$ denotes the left-continuous inverse of $F_{\nu_\theta^\sigma}$. Moreover, we have that $W_2(\mu_\theta^\sigma, \nu_\theta^\sigma)^2 = \int \frac{1}{2}|x - T_\theta^\sigma(x)|^2 \, \mu_\theta^\sigma(dx)$.

It can be easily seen that there exists some positive constant κ such that

$$\int_{\mathcal{S}^{k-1}} \theta\theta^\top \, d\theta = \kappa I_k, \tag{2.3}$$

and hence

$$\int_{\mathcal{S}^{k-1}} |\theta^\top x|^2 \, d\theta = \kappa |x|^2, \quad \forall x \in \mathbb{R}^k.$$

Lemma 2.2. *Let $\nu \in \mathcal{P}_2(\mathbb{R}^k)$ be fixed. Suppose either $\sigma > 0$ or $F_{\mu_\theta^\sigma}$ and $F_{\nu_\theta^\sigma}$ are continuous and strictly increasing for each $\theta \in \mathcal{S}^{k-1}$. Then, the mapping*

$$\mu \in \mathcal{P}_2(\mathbb{R}^k) \mapsto SW_2^\sigma(\mu,\nu)^2$$

is L-differentiable, and

$$D_\mu SW_2^\sigma(\mu,\nu)^2(x) = \int_{\mathcal{S}^{k-1}} \int_{\mathbb{R}} \theta \left(\theta^\top x - T_\theta^\sigma(\theta^\top x + y)\mathcal{N}_\sigma(y)\right) dy d\theta. \tag{2.4}$$

Moreover, we have the estimate

$$\int_{\mathbb{R}^k} |D_\mu SW_2^\sigma(\mu,\nu)^2(x)|^2 \mu(dx) \le C \left(\int_{\mathbb{R}^k} |x|^2 \mu(dx) + \int_{\mathbb{R}^k} |y|^2 \nu^\sigma(dy) \right) \tag{2.5}$$

and the mapping $x \mapsto D_\mu SW_2^\sigma(\mu,\nu)^2(x)$ is continuous on $\mathbb{R}^k$.

Proof. The proof relies on the proof of Theorem 3.2 of [1]. We first prove that if $F_{\mu_\theta^\sigma}$ and $F_{\nu_\theta^\sigma}$ are continuous and strictly increasing functions for some $\theta \in \mathcal{S}^{k-1}$, then the function

$$\mu \in \mathcal{P}_2(\mathbb{R}^k) \mapsto W_2(\mu_\theta^\sigma, \nu_\theta^\sigma)^2$$

is L-differentiable at μ, and its L-derivatives are given by

$$D_\mu \left(W_2(\mu_\theta^\sigma,\nu_\theta^\sigma)^2\right)(x) = \theta \left(\theta^\top x - \int T_\theta^\sigma(\theta^\top x + y)\mathcal{N}_\sigma(y)\, dy \right). \tag{2.6}$$

Fix $X \in L^2(\Omega, \mathcal{P}, \mathbb{P}; \mathbb{R}^k)$ with distribution μ and $\xi \in L^2(\Omega, \mathcal{P}, \mathbb{P}; \mathbb{R}^k)$ with norm 1. Denote $N_\sigma \in L^2(\Omega, \mathcal{P}, \mathbb{P}; \mathbb{R}^k)$ which is independent of X and ξ with distribution $\mathcal{N}_\sigma$. Denote $\mu^{n,\sigma}$ the distribution of $X + \frac{\xi}{n} + N_\sigma$ and note that μ^σ is the distribution of $X + N_\sigma$. By the minimality of the 2-Wasserstein distance, we have that

$$\begin{aligned} W_2(\mu_\theta^{n,\sigma},\nu_\theta^\sigma)^2 &\le \frac{1}{2}\mathbb{E}\left[\left| X^\top\theta + \frac{\xi^\top\theta}{n} + N_\sigma^\top\theta - T_\theta^\sigma(\theta^\top(X+N_\sigma)) \right|^2 \right] \\ &\le W_2(\mu_\theta^\sigma,\nu_\theta^\sigma)^2 + \mathbb{E}\left[\frac{\xi^\top\theta}{n}\left((X+N_\sigma)^\top\theta - T_\theta^\sigma((X+N_\sigma)^\top\theta)\right) \right] + \frac{1}{2}\mathbb{E}\left[\frac{(\xi^\top\theta)^2}{n^2} \right]. \end{aligned}$$

We now take $Y^n \in L^2(\Omega, \mathcal{P}, \mathbb{P}; \mathbb{R}^k)$ with distribution ν_θ^σ so that the coupling $(X^\top\theta + \frac{\xi^\top\theta}{n} + N_\sigma^\top\theta, Y^n)$ yields to an optimal coupling between $\mu_\theta^{n,\sigma}$ and ν_θ^σ. We have the following estimate

$$\begin{aligned} W_2(\mu_\theta^\sigma,\nu_\theta^\sigma)^2 &\le \frac{1}{2}\mathbb{E}\left[\left| (X+N_\sigma)^\top\theta - Y^n \right|^2 \right] \\ &\le W_2(\mu_\theta^{n,\sigma},\nu_\theta^\sigma)^2 - \mathbb{E}\left[\frac{\xi^\top\theta}{n}\left((X+N_\sigma)^\top\theta - Y^n\right) \right] + \frac{1}{2}\mathbb{E}\left[\frac{(\xi^\top\theta)^2}{n^2} \right] \\ &\le W_2(\mu_\theta^{n,\sigma},\nu_\theta^\sigma)^2 - \mathbb{E}\left[\frac{\xi^\top\theta}{n}\left((X+N_\sigma)^\top\theta - T_\theta^\sigma\left((X+N_\sigma)^\top\theta\right)\right) \right] \\ &\quad - \mathbb{E}\left[\frac{\xi^\top\theta}{n}\left(T_\theta^\sigma\left((X+N_\sigma)^\top\theta\right) - Y^n\right) \right] + \frac{1}{2}\mathbb{E}\left[\frac{(\xi^\top\theta)^2}{n^2} \right]. \end{aligned}$$

Thus, we obtain the inequality

$$\begin{aligned} &n \left| W_2(\mu_\theta^{n,\sigma},\nu_\theta^\sigma)^2 - W_2(\mu_\theta^\sigma,\nu_\theta^\sigma)^2 - \mathbb{E}\left[\frac{\xi^\top\theta}{n}\left((X+N_\sigma)^\top\theta - T_\theta^\sigma\left((X+N_\sigma)^\top\theta\right)\right) \right] \right| \\ &\le \mathbb{E}\left[\left| T_\theta^\sigma\left((X+N_\sigma)^\top\theta\right) - Y^n \right|^2 \right]^{1/2} + \frac{1}{2n}, \end{aligned}$$

where the last line goes to 0 thanks to Lemma 3.5 of [1], and we obtain (2.6) thanks to the fact that N_σ has 0 mean.

Integrating (2.6) over θ, we obtain (2.4). Indeed, for each $\theta \in \mathcal{S}^{k-1}$, we have that

$$
\begin{aligned}
&n\left|W_2(\mu_\theta^{n,\sigma},\nu_\theta^\sigma)^2 - W_2(\mu_\theta^\sigma,\nu_\theta^\sigma)^2\right| \\
&\qquad \le n\left|W_2(\mu_\theta^{n,\sigma},\nu_\theta^\sigma) - W_2(\mu_\theta^\sigma,\nu_\theta^\sigma)\right|\left|W_2(\mu_\theta^{n,\sigma},\nu_\theta^\sigma) + W_2(\mu_\theta^\sigma,\nu_\theta^\sigma)\right|.
\end{aligned}
$$

By the triangle inequality of Wasserstein distance, it can be seen that

$$
n\left|W_2(\mu_\theta^{n,\sigma},\nu_\theta^\sigma) - W_2(\mu_\theta^\sigma,\nu_\theta^\sigma)\right| \le nW_2(\mu_\theta^{n,\sigma},\mu_\theta^\sigma) \le n\sqrt{\mathbb{E}[(\xi/n)^2]} = 1,
$$

and hence

$$
n\left|W_2(\mu_\theta^{n,\sigma},\nu_\theta^\sigma)^2 - W_2(\mu_\theta^\sigma,\nu_\theta^\sigma)^2\right| \le \left|W_2(\mu_\theta^{n,\sigma},\nu_\theta^\sigma) + W_2(\mu_\theta^\sigma,\nu_\theta^\sigma)\right| \le C,
$$

where the upper bound C only depends on σ,μ,ν. Therefore it follows by the dominated convergence theorem that

$$
\begin{aligned}
&\lim_{n\to\infty} n(SW_2^\sigma(\mathbb{P}_{X+(\xi/n)},\nu)^2 - SW_2^\sigma(\mathbb{P}_X,\nu)^2) \\
&\qquad = \lim_{n\to\infty}\int n(W_2(\mu_\theta^{n,\sigma},\nu_\theta^\sigma)^2 - W_2(\mu_\theta^\sigma,\nu_\theta^\sigma)^2)\,d\theta \\
&\qquad = \int \mathbb{E}\left[\xi^\top\theta\left(X^\top\theta - T_\theta^\sigma\left((X+N_\sigma)^\top\theta\right)\right)\right]d\theta \\
&\qquad = \mathbb{E}\left[\xi^\top\int\theta\left(\theta^\top X - \mathbb{E}\left[T_\theta^\sigma(\theta^\top(X+N_\sigma))\right]\right)\,d\theta\right],
\end{aligned}
$$

where the last equality is due to Fubini's theorem.

Now let us show (2.5). According to $|a+b|^2 \le 2|a|^2 + 2|b|^2$, it can be easily verified that

$$
\begin{aligned}
|D_\mu SW_2^\sigma(\mu,\nu)^2(x)|^2 &\le 2\left|\int\theta\theta^\top x\,d\theta\right|^2 + 2\left|\int\theta\,\mathbb{E}\left[T_\theta^\sigma(\theta^\top(x+N_\sigma))\right]d\theta\right|^2 \\
&\le 2\kappa^2|x|^2 + 2\int|\theta|^2\,d\theta\int\left(\mathbb{E}\left[T_\theta^\sigma(\theta^\top(x+N_\sigma))\right]\right)^2\,d\theta.
\end{aligned}
$$

Integrating the above inequality with respect to μ, we conclude

$$
\begin{aligned}
\int|D_\mu SW_2^\sigma(\mu,\nu)^2(x)|^2\,\mu(dx) &\le C\left(\int|x|^2\,\mu(dx) + \int d\theta\int|T_\theta^\sigma(x)|^2\,\mu_\theta^\sigma(dx)\right) \\
&\le C\left(\int|x|^2\,\mu(dx) + \int d\theta\int|y|^2\,\nu_\theta^\sigma(dy)\right) \\
&\le C\left(\int|x|^2\,\mu(dx) + \int|y|^2\,\nu^\sigma(dy)\right).
\end{aligned}
$$

In the end, we show the continuity of $x \mapsto D_\mu SW_2^\sigma(\mu,\nu)^2(x)$. To prove this point, we only need to prove the continuity of

$$
x \mapsto \int_{\mathcal{S}^{k-1}}\int_{\mathbb{R}}\theta T_\theta^\sigma(\theta^\top x + y)e^{-\frac{y^2}{2\sigma^2}}\,dyd\theta = \int_{\mathcal{S}^{k-1}}\int_{\mathbb{R}}\theta T_\theta^\sigma(y)e^{-\frac{(y-\theta^\top x)^2}{2\sigma^2}}\,dyd\theta.
$$

Thanks to the proof of (2.5), we have that

$$
\int_{\mathbb{R}^k}\mu(dx)\int_{\mathcal{S}^{k-1}}\int_{\mathbb{R}}|T_\theta^\sigma(y)|^2e^{-\frac{(y-\theta^\top x)^2}{2\sigma^2}}dyd\theta = \int d\theta\int_{\mathbb{R}}|T_\theta^\sigma(y)|^2\,\mu_\theta^\sigma(dy) < \infty,
$$

and hence there exists $x_0 \in \mathbb{R}^k$ so that

$$\int_{\mathcal{S}^{k-1}} \int_{\mathbb{R}} |T^\sigma_\theta(y)|^2 e^{-\frac{(y-\theta^\top x_0)^2}{2\sigma^2}}\, dyd\theta < \infty. \tag{2.7}$$

Fix $M > 0$ and let $x \in \mathbb{R}^k$ so that $|x - x_0| \leq M$. Then, we have the estimate

$$\begin{aligned} \left|\theta T^\sigma_\theta(y) e^{-\frac{(y-\theta^\top x)^2}{2\sigma^2}}\right| &\leq |T^\sigma_\theta(y)| e^{\frac{M|y-\theta^\top x_0|}{\sigma^2}} e^{-\frac{(y-\theta^\top x_0)^2}{2\sigma^2}} \\ &\leq |T^\sigma_\theta(y)|^2 e^{-\frac{(y-\theta^\top x_0)^2}{2\sigma^2}} + e^{\frac{2M|y-\theta^\top x_0|}{\sigma^2}} e^{-\frac{(y-\theta^\top x_0)^2}{2\sigma^2}} \end{aligned} \tag{2.8}$$

which is independent of x and integrable. Thus, we can use the dominated convergence theorem to obtain the continuity of $x \mapsto D_\mu SW^\sigma_2(\mu,\nu)^2(x)$ on $\mathbb{R}^k$. □

Remark 2.3. Note that according to Definition 1.1, the L-derivatives of $\mu \mapsto SW^\sigma_2(\mu,\nu)^2$ may not be unique since the definition only requires the value of $D_\mu SW^\sigma_2(\mu,\nu)^2(\cdot)$ on the support of μ. However, Lemma 2.2 shows that as a consequence of the convolution with the Gaussian distribution $\mathcal{N}_\sigma$, the derivative in (2.4) admits a continuous version on $\mathbb{R}^k$. In the following lemma, we show that this function is in fact continuously differentiable and compute its derivative.

Lemma 2.4. *For $\sigma > 0$, we have the following results for derivatives.*

$$D^2_{x\mu} SW^\sigma_2(\mu,\nu)^2(x) = \int_{\mathcal{S}^{K-1}} \int_{\mathbb{R}} \theta\theta^\top \left(1 - \frac{(\theta^\top x - y)}{\sigma^2} T^\sigma_\theta(y) \mathcal{N}_\sigma(\theta^\top x - y)\right) dyd\theta, \tag{2.9}$$

$$\int_{\mathbb{R}^k} |D^2_{x\mu} SW^\sigma_2(\mu,\nu)^2(x)|\, \mu(dx) \leq C\left(1 + \frac{1}{\sigma}\sqrt{\int |y|^2\, \nu^\sigma(dy)}\right). \tag{2.10}$$

Proof. To show (2.9), it suffices to prove that we can interchange the derivative with respect to x and the integral. To that end, we prove that for any fixed $M > 0$, there exists an integrable bound of $(\theta^\top x - y) T^\sigma_\theta(y) \mathcal{N}_\sigma(\theta^\top x - y)$ uniformly for all $|x| \leq M$. Take $x_0 \in \mathbb{R}^k$ as in (2.7) such that

$$\int_{\mathcal{S}^{k-1}} \int_{\mathbb{R}} |T^\sigma_\theta(y)|^2 \mathcal{N}_\sigma(\theta^\top x_0 - y)\, dyd\theta < \infty.$$

Without loss of generality, we assume that $|x - x_0| \leq M$. Then we have the estimate

$$\begin{aligned} &\left|(\theta^\top x - y) T^\sigma_\theta(y) \mathcal{N}_\sigma(\theta^\top x - y)\right| \leq \left(M + |\theta^\top x_0 - y|\right) |T^\sigma_\theta(y)| \mathcal{N}_\sigma(\theta^\top x_0 - y) e^{\frac{M|\theta^\top x_0 - y|}{\sigma^2}} \\ &\leq |T^\sigma_\theta(y)|^2 \mathcal{N}_\sigma(\theta^\top x_0 - y) + (M + |\theta^\top x_0 - y|)^2 e^{\frac{2M|\theta^\top x_0 - y|}{\sigma^2}} \mathcal{N}_\sigma(\theta^\top x_0 - y), \end{aligned} \tag{2.11}$$

which is independent of x and integrable with respect to variables y, θ. Therefore by the dominated convergence theorem, we can interchange the integral and derivative, and thus obtain (2.9).

Let us integrate $D^2_{x\mu} SW^\sigma_2(\mu,\nu)^2(x)$ over μ. Using (2.9), we have

$$\begin{aligned}\int |D^2_{x\mu} SW^\sigma_2(\mu,\nu)^2(x)|\, \mu(dx) &\le C + \frac{C}{\sigma^2}\int d\theta \int \mu(dx)\int |\theta^\top x - y||T^\sigma_\theta(y)|\mathcal{N}_\sigma(\theta^\top x - y)\, dy \\ &\le C + \frac{C}{\sigma^2}\int d\theta \int \mu_\theta(dx)\int |y||T^\sigma_\theta(x+y)|\mathcal{N}_\sigma(y)\, dy.\end{aligned}$$

Then the Cauchy–Schwarz inequality yields

$$\begin{aligned}&\left|\int \mu_\theta(dx)\int |yT^\sigma_\theta(x+y)|\mathcal{N}^\sigma(y)\, dy\right| \\ &\le \sqrt{\int\int |y|^2\mathcal{N}_\sigma(y)\, dy\, \mu_\theta(dx)}\sqrt{\int\int |T^\sigma_\theta(x+y)|^2\mathcal{N}_\sigma(y)\, dy\, \mu_\theta(dx)} \\ &= \sigma\sqrt{\int |x|^2\, \nu^\sigma_\theta(dx)}.\end{aligned}$$

Integrating the inequality above over θ, we conclude that

$$\int |D^2_{x\mu} SW^\sigma_2(\mu,\nu)^2(x)|\, \mu(dx) \le C\left(1 + \frac{1}{\sigma}\sqrt{\int |y|^2\, \nu^\sigma(dy)}\right).$$

□

For each $\sigma > 0$, let us define the function $\rho_\sigma : ([0,T]\times\mathcal{P}_2(\mathbb{R}^k))^2 \to \mathbb{R}$ via

$$\rho_\sigma((s,\mu),(t,\nu)) = |t-s|^2 + SW^\sigma_2(\mu,\nu)^2. \tag{2.12}$$

Then ρ_σ is a gauge type function on $(\mathcal{P}_2(\mathbb{R}^k), SW^\sigma_2)$; see [3, Definition 2.5.1]. The following smooth variational principle is the main result of this paper.

Proposition 2.5. *Fix $\delta > 0$ and let $G : [0,T]\times\mathcal{P}_2(\mathbb{R}^k) \to \mathbb{R}$ be upper semicontinuous and bounded from above. Given $\lambda > 0$, let $(t_0,\mu_0) \in [0,T]\times\mathcal{P}_2(\mathbb{R}^k)$ be such that*

$$\sup_{(t,\mu)\in[0,T]\times\mathcal{P}_2(\mathbb{R}^k)} G(t,\mu) - \lambda \le G(t_0,\mu_0).$$

Then there exist $(\tilde t,\tilde\mu) \in [0,T]\times\mathcal{P}_2(\mathbb{R}^k)$ and a sequence $\{(t_n,\mu_n)\}_{n\ge1} \subset [0,T]\times\mathcal{P}_2(\mathbb{R}^k)$ such that:

(i) *$\rho_{1/\delta}((\tilde t,\tilde\mu),(t_n,\mu_n)) \le \frac{\lambda}{2^n\delta^2}$, for every $n \ge 0$;*

(ii) *$G(t_0,\mu_0) \le G(\tilde t,\tilde\mu) - \delta^2\phi_\delta(\tilde t,\tilde\mu)$, with $\phi_\delta : [0,T]\times\mathcal{P}_2(\mathbb{R}^k) \to [0,+\infty)$ given by*

$$\phi_\delta(t,\mu) = \sum_{n=0}^{+\infty}\frac{1}{2^n}\rho_{1/\delta}((t,\mu),(t_n,\mu_n)), \quad \forall\,(t,\mu)\in[0,T]\times\mathcal{P}_2(\mathbb{R}^k); \tag{2.13}$$

(iii) *$G(t,\mu) - \delta^2\phi_\delta(t,\mu) < G(\tilde t,\tilde\mu) - \delta^2\phi_\delta(\tilde t,\tilde\mu)$, for every $(t,\mu) \in ([0,T]\times\mathcal{P}_2(\mathbb{R}^k))\setminus\{(\tilde t,\tilde\mu)\}$.*

Furthermore, the function ϕ_δ satisfies the following properties:

(1) *ϕ_δ is differentiable in time and measure;*

(2) *its time derivative is bounded by $4T$;*

(3) *its measure derivative is bounded by*

$$\int |D_\mu \phi_\delta(t,\mu)(x)|^2 \, \mu(dx) \le C\left(\int |x|^2 \, \mu(dx) + \int |x|^2 \, \tilde{\mu}(dx) + \frac{1}{\delta^2}\right);$$

(4) *the derivative* $D^2_{x\mu}\phi_\delta(t,\mu)(x)$ *satisfies*

$$\int |D^2_{x\mu}\phi_\delta(t,\mu)(x)| \, \mu(dx) \le C\left(1 + \delta\sqrt{\int |x|^2 \, \tilde{\mu}(dx)}\right).$$

Proof. Parts (i), (ii), (iii) directly follow from [3, Theorem 2.5.2]. Denoting $\sigma := 1/\delta$, let us first estimate second moments of μ_n^σ.

Recall the κ defined in (2.3). For any $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^k)$, it can be easily seen that

$$\begin{aligned}\kappa \int |x|^2 \, \nu(dx) = \int d\theta \int |y|^2 \, \nu_\theta(dy) &\le 2\int d\theta \int |y|^2 \, \mu_\theta(dy) + 2SW_2(\mu,\nu)^2 \\ &\le C\left(\int |x|^2 \, \mu(dx) + SW_2(\mu,\nu)^2\right),\end{aligned}$$

where C is twice the volume of $\mathcal{S}^{k-1}$. Replacing ν and μ with μ_n^σ and $\tilde{\mu}^\sigma$ respectively in the above inequality, by part (i) we obtain that

$$\kappa \int |x|^2 \, \mu_n^\sigma(dx) \le C\left(\int |x|^2 \tilde{\mu}^\sigma(dx) + \frac{\lambda}{2^n \delta^2}\right). \tag{2.14}$$

It is clear that ϕ_δ is differentiable in time, and we show that one can interchange the infinite sum and the derivative in measure. Supposing $X \in L^2(\Omega, \mathcal{P}, \mathbb{P}; \mathbb{R}^k)$ with distribution μ and $\xi \in L^2(\Omega, \mathcal{P}, \mathbb{P}; \mathbb{R}^k)$ with norm 1, then it can be easily seen that for any $n \ge 0$

$$\begin{aligned}&\left|\frac{\rho_\sigma((t,\mathbb{P}_{X+\epsilon\xi}),(t_n,\mu_n)) - \rho_\sigma((t,\mathbb{P}_X),(t_n,\mu_n))}{\epsilon}\right| \\ &\qquad \le \left(SW_2^\sigma(\mathbb{P}_{X+\epsilon\xi},\mu_n) + SW_2^\sigma(\mathbb{P}_X,\mu_n)\right)\frac{|SW_2^\sigma(\mathbb{P}_{X+\epsilon\xi},\mathbb{P}_X)|}{\epsilon}.\end{aligned} \tag{2.15}$$

The first term on the right is bounded by second moments of μ^σ, μ_n^σ, and hence by second moments of $\mu^\sigma, \tilde{\mu}^\sigma$ thanks to (2.14). The second term is bounded by $\int \sqrt{\mathbb{E}[|\theta^\top \xi|^2]}\, d\theta$, and thus by the volume of $\mathcal{S}^{k-1}$. Therefore, the left hand side of (2.15) is uniformly bounded in n, and we conclude that

$$\begin{aligned}\lim_{\epsilon\to 0}\frac{\phi_\delta(t,\mathbb{P}_{X+\epsilon\xi}) - \phi_\delta(t,\mathbb{P}_X)}{\epsilon} &= \sum_{n=0}^\infty \lim_{\epsilon\to 0}\frac{\rho_\sigma((t,\mathbb{P}_{X+\epsilon\xi}),(t_n,\mu_n)) - \rho_\sigma((t,\mathbb{P}_X),(t_n,\mu_n))}{2^n\epsilon} \\ &= \sum_{n=0}^\infty \frac{1}{2^n}\mathbb{E}\left[\xi^\top D_\mu SW_2^\sigma(\mu,\mu_n)^2(X)\right] \\ &= \mathbb{E}\left[\xi^\top\left(\sum_{n=0}^\infty \frac{1}{2^n} D_\mu SW_2^\sigma(\mu,\mu_n)^2(X)\right)\right].\end{aligned}$$

Thus ϕ_δ is differentiable in measure, and

$$D_\mu \phi_\delta(t,\mu)(x) = \sum_{n=0}^\infty \frac{1}{2^n} D_\mu SW_2^\sigma(\mu,\mu_n)^2(x).$$

Part (2) is trivial. It follows from (2.14), Lemma 2.2, the Minkowski inequality, and the monotone convergence theorem that

$$\sqrt{\mathbb{E}\left[|D_\mu \phi_\delta(t,\mu)(X)|^2\right]} \leq \sum_{n=0}^{\infty} \frac{1}{2^n} \sqrt{\mathbb{E}\left[|D_\mu SW_2^\sigma(\mu,\mu_n)^2(X)|^2\right]}$$
$$\leq C\sqrt{\left(\int |x|^2\,\mu(dx) + \int |x|^2\,\tilde{\mu}(dx) + \frac{1}{\delta^2}\right)},$$

which justifies part (3).

In the end, using estimates (2.8) and (2.11), it can be easily checked that the mapping $x \mapsto D_\mu \phi_\delta(t,\mu)(x) = \sum_{n=0}^{\infty} \frac{1}{2^n} D_\mu SW_2^\sigma(\mu,\mu_n)^2(x)$ is continuous and differentiable as in Lemma 2.2 and Lemma 2.4. Then, by (2.10) and (2.14), we conclude that

$$\int \left|D^2_{\mu x}\phi_\delta(t,\mu)(x)\right| \mu(dx) \leq \sum_{n=0}^{\infty} \frac{1}{2^n} \int \left|D^2_{\mu x} SW_2^\sigma(\mu,\mu_n)^2(x)\right| \mu(dx)$$
$$\leq C\left(1 + \delta\sqrt{\int |x|^2\,\tilde{\mu}(dx)}\right).$$

□

Remark 2.6. [6, Lemma 4.4] constructed a gauge type function using dyadic partitions of the underlying space $\mathbb{R}^k$. Our construction ρ_σ is much simpler, and can serve as a substitute of [6, Lemma 4.4]. Furthermore, ρ_σ is twice differentiable with respect to μ, and thus could be useful in the study of second-order partial differential equations on Wasserstein space.

References

[1] Aurélien Alfonsi and Benjamin Jourdain, *Squared quadratic Wasserstein distance: optimal couplings and Lions differentiability*, ESAIM Probab. Stat. **24** (2020), 703–717, DOI 10.1051/ps/2020013. MR4174419

[2] Erhan Bayraktar and Gaoyue Guoï, *Strong equivalence between metrics of Wasserstein type*, Electron. Commun. Probab. **26** (2021), Paper No. 13, 13, DOI 10.3390/mca26010013. MR4236683

[3] Jonathan M. Borwein and Qiji J. Zhu, *Techniques of variational analysis*, CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, vol. 20, Springer-Verlag, New York, 2005. MR2144010

[4] Pierre Cardaliaguet, François Delarue, Jean-Michel Lasry, and Pierre-Louis Lions, *The master equation and the convergence problem in mean field games*, Annals of Mathematics Studies, vol. 201, Princeton University Press, Princeton, NJ, 2019, DOI 10.2307/j.ctvckq7qf. MR3967062

[5] René Carmona and François Delarue, *Probabilistic theory of mean field games with applications. I*, Probability Theory and Stochastic Modelling, vol. 83, Springer, Cham, 2018. Mean field FBSDEs, control, and games. MR3752669

[6] Andrea Cosso, Fausto Gozzi, Idris Kharroubi, Huyên Pham, and Mauro Rosestolato, *Master Bellman equation in the Wasserstein space: uniqueness of viscosity solutions*, arXiv:2107.10535, 2021.

[7] P.-L. Lions, *Viscosity solutions of fully nonlinear second-order equations and optimal stochastic control in infinite dimensions. III. Uniqueness of viscosity solutions for general second-order equations*, J. Funct. Anal. **86** (1989), no. 1, 1–18, DOI 10.1016/0022-1236(89)90062-1. MR1013931

[8] Filippo Santambrogio, *Optimal transport for applied mathematicians*, Progress in Nonlinear Differential Equations and their Applications, vol. 87, Birkhäuser/Springer, Cham, 2015. Calculus of variations, PDEs, and modeling, DOI 10.1007/978-3-319-20828-2. MR3409718

Department of Mathematics, University of Michigan, Ann Arbor, Michigan 48103
Email address: erhan@umich.edu

Department of Mathematics, Florida State University, Tallahassee, Florida 32304
Email address: iekren@fsu.edu

Department of Mathematics, University of Vienna, 1090 Wien, Austria
Email address: xin.zhang@univie.ac.at