Optimal 1-Wasserstein distance for WGANs

Arthur Stéphanovitch, Ugo Tanielian, Benoît Cadre, Nicolas Klutchnikoff, Gérard Biau

Optimal 1-Wasserstein distance for WGANs

Arthur Stéphanovitch STEPHANOVITCH@LPSM.PARIS
Université Paris Cité, CNRS, LPSM
F-75013 Paris, France

Ugo Tanielian UGO.TANIELIAN@GMAIL.COM
Criteo AI Lab, Paris, France

Benoît Cadre BENOIT.CADRE@UNIV-RENNES2.FR
Univ Rennes, CNRS, IRMAR - UMR 6625
F-35000 Rennes, France

Nicolas Klutchnikoff NICOLAS.KLUTCHNIKOFF@UNIV-RENNES2.FR
Univ Rennes, CNRS, IRMAR - UMR 6625
F-35000 Rennes, France

Gérard Biau GERARD.BIAU@SORBONNE-UNIVERSITE.FR
Sorbonne Université, CNRS, LPSM
F-75005 Paris, France

Abstract

The mathematical forces at work behind Generative Adversarial Networks raise challenging theoretical issues. Motivated by the important question of characterizing the geometrical properties of the generated distributions, we provide a thorough analysis of Wasserstein GANs (WGANs) in both the finite sample and asymptotic regimes. We study the specific case where the latent space is univariate and derive results valid regardless of the dimension of the output space. We show in particular that for a fixed sample size, the optimal WGANs are closely linked with connected paths minimizing the sum of the squared Euclidean distances between the sample points. We also highlight the fact that WGANs are able to approach (for the 1-Wasserstein distance) the target distribution as the sample size tends to infinity, at a given convergence rate and provided the family of generative Lipschitz functions grows appropriately. We derive in passing new results on optimal transport theory in the semi-discrete setting.

Keywords: Wasserstein Generative Adversarial Networks, Wasserstein distance, optimal distribution, shortest path, rate of convergence, optimal transport theory

1. Introduction

Recent years have witnessed the advent of generative methodologies based on Generative Adversarial Networks (GANs, Goodfellow et al., 2014), with outstanding achievements in the fields of image (Radford et al., 2016; Karras et al., 2018), video (Vondrick et al., 2016), and text generation (Yu et al., 2017), just to name a few. The surveys by Lucic et al. (2018) and Borji (2019) cover the different GANs techniques together with a comparison

of their performances. We are concerned with the Wasserstein GAN (WGAN) approach of Arjovsky et al. (2017), which uses the 1-Wasserstein distance as an alternative to the Jensen-Shannon divergence implemented in traditional GANs. Over the years, WGANs and their derivatives have gained popularity in the machine learning community. They are today considered as one of the most successful generative techniques, achieving state-of-the art results in difficult problems (Karras et al., 2018, 2019) while improving the stability and getting rid of unpleasant issues such as mode collapse (Gulrajani et al., 2017).

To get started, let us properly define WGANs. Assume that we are given a sample $X_1, \ldots, X_n$ of independent $\mathbb{R}^d$-valued random variables, identically distributed according to some unknown distribution μ. Throughout the manuscript, the space $\mathbb{R}^d$ as well as all other spaces $\mathbb{R}^k$ are equipped with the Euclidean norm $\|\cdot\|$, with no reference to d or k as the context is clear. The generative problem is to use the sample $X_1, \ldots, X_n$ to learn μ and, simultaneously, generate new "fake" data that look "similar" to the X_i's. In the WGAN framework, this problem is addressed by minimizing the 1-Wasserstein distance between a family of candidate distributions and the empirical measure of the sample. Recall here that for two probability measures π_1 and π_2 on $\mathbb{R}^d$, the 1-Wasserstein distance $W_1(\pi_1, \pi_2)$ between π_1 and π_2 is defined by

$$W_1(\pi_1, \pi_2) = \inf_{\pi \in \Pi(\pi_1, \pi_2)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\| \mathrm{d}\pi(x, y),$$

where $\Pi(\pi_1, \pi_2)$ denotes the collection of all joint probability measures π on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals π_1 and π_2 (e.g., Villani, 2008). Notice that $W_1(\cdot, \cdot)$ is not a distance in the strict sense, because it may take the value $+\infty$. We also recall that the empirical measure μ_n based on $X_1, \ldots, X_n$ is defined, for any Borel set $A \subseteq \mathbb{R}^d$, by $\mu_n(A) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{X_i \in A\}$. Now, let U be a uniform random variable on $[0,1]^p$ and, for $K > 0$, let $\mathrm{Lip}_K(E, E')$ be the set of K-Lipschitz continuous functions from $E \subseteq \mathbb{R}^k$ to $E' \subseteq \mathbb{R}^{k'}$, equipped with their respective Euclidean norms, that is

$$\mathrm{Lip}_K(E, E') = \{G : E \to E' : \|G(x) - G(y)\| \leqslant K\|x - y\|, \ (x, y) \in E^2\}.$$

For $G \in \mathrm{Lip}_K([0,1]^p, \mathbb{R}^d)$, we denote by $G_\sharp U$ the pushforward distribution of U by G, that is, for any Borel set $A \subseteq \mathbb{R}^d$, $G_\sharp U(A) = \lambda_p(G^{-1}(A))$, where λ_p is the Lebesgue measure on $\mathbb{R}^p$. In their abstract formulation, WGANs use the family of pushforward distributions $\{G_\sharp U : G \in \mathrm{Lip}_K([0,1]^p, \mathbb{R}^d)\}$ as candidate distributions to estimate μ, with the objective of finding the best function G that minimizes the 1-Wasserstein distance between $G_\sharp U$ and the empirical measure μ_n. In other words, one seeks to find an optimal $\widehat{G}_K \in \mathrm{Lip}_K([0,1]^p, \mathbb{R}^d)$ such that

$$W_1(\widehat{G}_{K\sharp} U, \mu_n) = \inf_{G \in \mathrm{Lip}_K([0,1]^p, \mathbb{R}^d)} W_1(G_\sharp U, \mu_n). \tag{1}$$

Once a minimizer $\widehat{G}_K$ has been found, it is easy to generate "fake" observations, by simply taking a uniform i.i.d. sample $U_1, \ldots, U_m$ and computing $\widehat{G}_K(U_1), \ldots, \widehat{G}_K(U_m)$. In the GAN literature, the space $[0,1]^p$ is called the latent space and the distribution of the random variable U the latent distribution. It should be stressed that assuming Lipschitz continuous candidate functions G is classical when defining WGANs (e.g., Zhou et al., 2019). However, some authors have also considered smoother classes, such as for example functions with

Lipschitz partial derivatives up to some order (e.g., Luise et al., 2020; Schreuder et al., 2021). To keep things as simple as possible, we do not make further assumptions on the generative functions other than their Lipschitz property.

The key to approach the infimum in (1) is to use the dual formulation of the 1-Wasserstein distance (Kantorovich and Rubinstein, 1958). Indeed, one has

$$\begin{aligned} W_1(G_\sharp U, \mu_n) &= \sup_{f \in \mathrm{Lip}_1(\mathbb{R}^d, \mathbb{R})} \int_{\mathbb{R}^d} f \mathrm{d}G_\sharp U - \int_{\mathbb{R}^d} f \mathrm{d}\mu_n \\ &= \sup_{f \in \mathrm{Lip}_1(\mathbb{R}^d, \mathbb{R})} \int_{[0,1]^p} f(G(u))\mathrm{d}u - \frac{1}{n}\sum_{i=1}^n f(X_i), \end{aligned}$$

so that the WGAN optimization Problem (1) takes the min-max form

$$W_1(\widehat{G}_K \sharp U, \mu_n) = \inf_{G \in \mathrm{Lip}_K([0,1]^p, \mathbb{R}^d)} \sup_{f \in \mathrm{Lip}_1(\mathbb{R}^d, \mathbb{R})} \int_{[0,1]^p} f(G(u))\mathrm{d}u - \frac{1}{n}\sum_{i=1}^n f(X_i). \quad (2)$$

Since the nonparametric classes $\mathrm{Lip}_K([0,1]^p, \mathbb{R}^d)$ and $\mathrm{Lip}_1(\mathbb{R}^d, \mathbb{R})$ are too large to be implemented, they are replaced in practice by parametric models, respectively called the generator and the discriminator. In most applications, these parametric models take the form of multilayer neural networks, either feedforward or convolutional, hence the name WGANs. It is also important to note that in practice the function $G_\sharp U$ in (1) is estimated by random samples $G(U_1), \ldots, G(U_m)$ drawn from U. In other words, there exists an estimation error—on top of the approximation error by neural networks—between the optimum $\inf_G W_1(G_\sharp U, \mu_n)$ and any simulation. However, sampling from U is easy and one can take sufficiently large m. From an optimization perspective, the training of (W)GANs is challenging. The min-max optimum in (2) is usually found by using stochastic gradient descent, alternatively on the generator's and the discriminator's parameters. Studying the convergence of the different learning procedures is an interesting question, tackled for example by Kodali et al. (2017) and Mescheder et al. (2018).

In addition to the numerous empirical research studies, several theoretical articles aimed at understanding the mathematical and statistical properties of the adversarial problem (2) and its extensions to integral probability metrics (IPM, Müller, 1997). For example, leveraging the approximation properties of some family of neural networks, Biau et al. (2021) study the convergence of the model as the sample size tends to infinity, and clarify the respective effects of the generator and the discriminator by underlining some trade-off properties. Assuming smoothness properties on the generator and the discriminator, Liang (2021) and Singh et al. (2018) exhibit rates of convergence under an IPM-based loss for estimating densities that live in Sobolev spaces, while Uppal et al. (2019) explore the case of Besov spaces. More recently, Schreuder et al. (2021) have stressed the properties of IPM losses defined with smooth functions on a compact set. Remarkably, Liang (2021) discusses bounds for the Kullback-Leibler divergence, the Hellinger distance, and the 1-Wasserstein distance. Studying a different facet of the problem, Luise et al. (2020) analyze the interplay between the latent distribution and the complexity of the pushforward map, and how it affects the overall performance.

In this paper, we seek to describe the properties of the K-Lipschitz continuous functions that achieve the infimum in (1). Our approach is motivated by an active line of experimental

research, which aims at characterizing the distributions output by GANs, typically the geometry of their supports. For example, when dealing with the learning of disconnected manifolds, Tanielian et al. (2020) derived lower-bounds on the measure of the proposal distribution that lies out of the target manifold. Another much-debated question is to understand to what extent GANs memorize the dataset Vaishnavh et al. (2018). In this regard, Gulrajani et al. (2019) stress their tendency to memorize, and, in turn, propose a new evaluation protocol that enhances generalization. Yet, most of the conclusions on this subject are of an experimental nature, without clear theoretical arguments regarding the statistical properties of the distribution produced by GANs.

Motivated by the above, we provide in the present article a thorough analysis of Problem (1). Since this question is highly nontrivial, we deeply study the univariate latent setting ($p = 1$). Beyond the technical aspects, the motivation to study the univariate case is related to the so-called manifold hypothesis (Fefferman et al., 2016; Facco et al., 2017), which states that high-dimensional datasets may lay on manifolds of lower dimensions. For instance, YoonHaeng et al. (2021) show that using a latent dimension $p = 2$ is already sufficient to generate high-quality images for the MNIST dataset. We later give intuitions for the case $p > 1$.

Our contributions are the following:

1. To grasp how WGANs can approach the distribution μ, we start in Section 2 by an asymptotic analysis of $W_1(\widehat{G}_{K\sharp U}, \mu)$ as the sample size n tends to infinity, assuming that the Lipschitz constant K is kept fixed, independent of the data. We show in particular that in most situations, and independently of the dimension d, one has $\liminf_{n\to\infty} W_1(\widehat{G}_{K\sharp U}, \mu) > 0$ a.s.

2. Next, we provide in Section 3 a thorough finite sample analysis of the case $d = 1$, that is, whenever the output space is univariate. In this context, the Lipschitz constant K is allowed to depend on the sample $X_1, \dots, X_n$. We explicitly describe the (two) functions achieving the infimum in (1), give the exact value of the infimum, and show that the corresponding optimal distributions have atoms at the X_i's. Finally, taking an asymptotic point of view, we prove that $\lim_{n\to\infty} W_1(\widehat{G}_{K\sharp U}, \mu) = 0$ and offer convergence rates.

3. We then discuss in Section 4 new existence results on transport maps in semi-discrete optimal transport theory, for measures that are non necessarily absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$. This step is necessary before diving into the analysis of Problem (1) for $d > 1$.

4. In Section 5, we move to the case where the observations are multivariate ($d > 1$) and derive a finite sample bound on the infimum in (1). We show in particular, provided K is allowed to depend on the sample, that the bound is achieved by a distribution concentrated on a shortest-path-type graph constructed on the X_i's. Up to our knowledge, this is the first time that such bounds are available in the literature. Taking neural networks for the generator and the discriminator classes, we illustrate the results empirically. Similarly to Section 3, we also provide convergence rates for $\lim_{n\to\infty} W_1(\widehat{G}_{K\sharp U}, \mu)$.

All the proofs are gathered in the Annex (Stéphanovitch et al., 2023), with the exception of the proofs of Theorem 9 and Theorem 12.

2. Asymptotic analysis

The study begins with an asymptotic analysis of Problem (1), when the sample size n tends to infinity and the Lipschitz constant K is assumed to be fixed. For more clarity, the univariate case $d = 1$ is handled in Theorem 2 and the multivariate case $d > 1$ in Theorem 3. Recall that the latent variable U is assumed to be uniformly distributed on $[0, 1]$, and that the data $X_1, \ldots, X_n$ are i.i.d with unknown distribution μ. Throughout, we let

$$\widehat{\mathscr{G}_K} = \underset{G \in \mathrm{Lip}_K([0,1],\mathbb{R}^d)}{\arg\min} W_1(G_{\sharp U}, \mu_n)$$

be the set of minimizers of Problem (1), that is,

$$\widehat{\mathscr{G}_K} = \{\widehat{G}_K \in \mathrm{Lip}_K([0,1],\mathbb{R}^d) : W_1(\widehat{G}_{K\sharp U}, \mu_n) = \inf_{G \in \mathrm{Lip}_K([0,1],\mathbb{R}^d)} W_1(G_{\sharp U}, \mu_n)\}.$$

Observe that $\{\widehat{G}_{K\sharp U} : \widehat{G}_K \in \widehat{\mathscr{G}_K}\}$ is the collection of optimal distribution(s). Whenever μ is of order 1, i.e., $\mathbb{E}\|X_1\| = \int_{\mathbb{R}^d} \|x\|\mu(\mathrm{d}x) < \infty$, it is convenient to consider $\mathscr{G}_K$, the population version of $\widehat{\mathscr{G}_K}$ defined by

$$\mathscr{G}_K = \underset{G_K \in \mathrm{Lip}_K([0,1],\mathbb{R}^d)}{\arg\min} W_1(G_{K\sharp U}, \mu).$$

We start with the following simple but useful lemma.

Lemma 1 *The set $\widehat{\mathscr{G}_K}$ is not empty. In addition, assuming that μ is of order 1, the set $\mathscr{G}_K$ is not empty.*

In the sequel, we let $S(\mu)$ be the support of μ, i.e.,

$$S(\mu) = \{x \in \mathbb{R}^d : \mu(B(x, \varepsilon)) > 0 \text{ for all } \varepsilon > 0\},$$

where $B(x, \varepsilon)$ is the closed ball in $\mathbb{R}^d$ centered at x of radius ε. We are now ready to state the first theorem, which reveals the different behaviors of the quantity $W_1(\widehat{G}_{K\sharp U}, \mu)$ in dimension $d = 1$, provided $\widehat{G}_K$ is any minimizer in $\widehat{\mathscr{G}_K}$. Interestingly, we distinguish different cases depending on both the smoothness of the distribution function of μ and the boundedness of its support $S(\mu)$.

Theorem 2 (Case $d = 1$) *Let $\widehat{G}_K \in \widehat{\mathscr{G}_K}$. Assume that μ is of order 1, and let F^{-1} be the generalized inverse of the distribution function F of μ, i.e., for all $u \in (0, 1)$, $F^{-1}(u) = \inf\{x \in \mathbb{R} : F(x) \geqslant u\}$.*

1. *Assume that $S(\mu)$ is bounded.*

   (*i*) *If $F^{-1} \in \mathrm{Lip}_{K_0}([0, 1], \mathbb{R})$ for some $K_0 > 0$, then, for all $K \geqslant K_0$,*

   $$\lim_{n\to\infty} W_1(\widehat{G}_{K\sharp U}, \mu) = 0 \ \textit{a.s.}$$

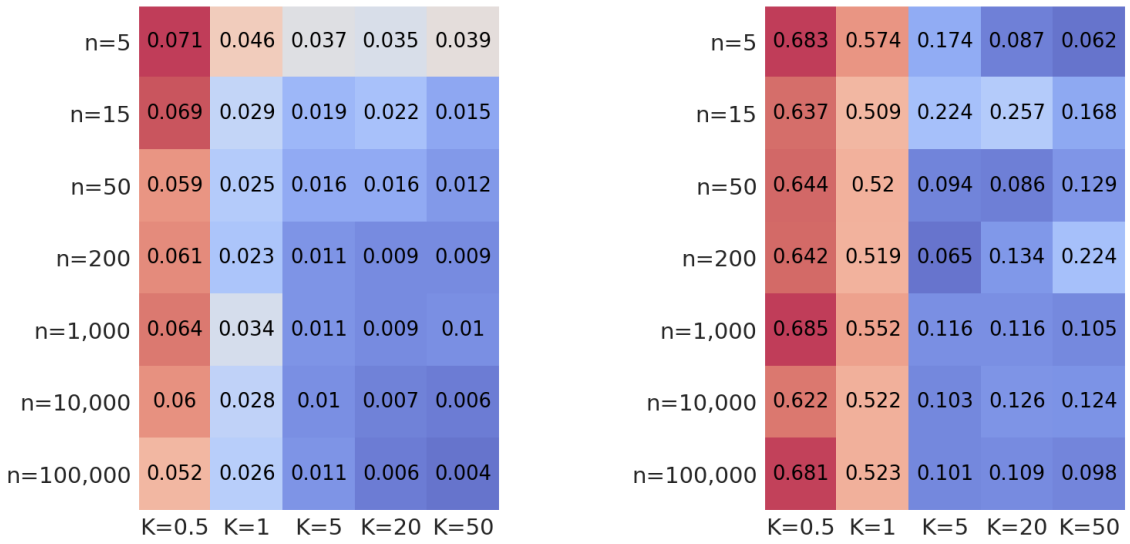
(*ii*) *If* $F \in \mathrm{Lip}_{K_1}(\mathbb{R}, [0,1])$ *for some* $K_1 > 0$*, then, for all* $K < 1/K_1$*,*

$$\liminf_{n\to\infty} \; W_1(\widehat{G}_{K\sharp U}, \mu) > 0 \;\; a.s.$$

2. *Assume that* $S(\mu)$ *is unbounded. Then, for all* $K > 0$*,*

$$\liminf_{n\to\infty} \; W_1(\widehat{G}_{K\sharp U}, \mu) > 0 \;\; a.s.$$

A first remark could be that, in $1(i)$, the support $S(\mu)$ is necessarily bounded since F^{-1} is assumed to be a K_0-Lipschitz function on $[0,1]$. Next, note that both conditions in $1(i)$ and $1(ii)$ may be satisfied simultaneously or not. For example, when μ is the uniform distribution on $[0,1]$, they are both verified with $K_0 = K_1 = 1$. Also, observing that $K_0 K_1 \geqslant 1$ (since $F \circ F^{-1}$ is the identity function), these two conditions focus in fact on different regimes. The first one pertains to the case where the set of generative functions ought to be big while the second one claims that a smaller class cannot recover the target distribution μ. In 2, we notice however that independently of the smoothness of μ and the magnitude of K, WGANs cannot recover the target distribution. This is for example the case when μ is a standard Gaussian distribution on the real line. The mechanism is illustrated in Figure 1, which shows the values of $W_1(\widehat{G}_{K\sharp U}, \mu)$ as a function of both n and K, when the target distribution μ is either uniform (left) or Gaussian (right). In the uniform case, as predicted by Theorem 2, we see that $W_1(\widehat{G}_{K\sharp U}, \mu)$ significantly decreases for K larger than 1 and stays rather constant for smaller K. In the Gaussian setting, 1-Wasserstein distances are far from zero, independently of the value of n and K. In the experiment, the generator is a 3-layer feedforward neural network while the discriminator is a 5-layer network.



(a) The distribution μ is uniform.

(b) The distribution μ is standard Gaussian.

Figure 1: 1-Wasserstein distance $W_1(\widehat{G}_{K\sharp U}, \mu)$ as a function of n and K (the bluer the lower and the redder the higher). Results are averaged over 2 runs.

These results should of course be appreciated in the light of the specific case where both the latent space and the target distribution μ share the same dimension 1. In the case where μ lies on a space of dimension strictly larger than 1, then the minimizers in $\widehat{\mathscr{G}}_K$ cannot reconstruct μ, as stated by the following theorem.

Theorem 3 (Case $d > 1$) *Let $\widehat{G}_K \in \widehat{\mathscr{G}}_K$. Assume that μ is of order 1 and that $\lambda_d(S(\mu)) > 0$, where λ_d denotes the Lebesgue measure on $\mathbb{R}^d$. Then, for all $K > 0$,*

$$\liminf_{n\to\infty} \; W_1(\widehat{G}_{K\sharp}U, \mu) > 0 \;\; a.s.$$

The condition on the support of μ states that μ is a "true" measure on $\mathbb{R}^d$. We leave it as an exercise to prove that the same result holds by assuming that the Hausdorff dimension of $S(\mu)$ is strictly larger than 1.

3. Finite sample analysis in a univariate output space

The topic of the present section is to fully describe the set of minimizers $\widehat{\mathscr{G}}_K$ (Lemma 1), in the specific setting where both the output and the latent spaces are univariate. We denote by $X_{(1)}, \ldots, X_{(n)}$ the reordering of $X_1, \ldots, X_n$ according to their increasing values, that is $X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)}$, where ties are broken arbitrarily. Importantly, the Lipschitz constant K is now allowed to depend upon the sample and is chosen to satisfy the constraint $K \geqslant n \max_{i=1,\ldots,n-1}(X_{(i+1)} - X_{(i)})$.

The analysis starts by introducing the following function $\widehat{G}^\star_K : [0,1] \to \mathbb{R}$, which will play a key role in solving Problem (1): for all $u \in [0,1]$,

$$\widehat{G}^\star_K(u) = \begin{cases} X_{(1)} & \text{if } u \in \big[0, \frac{1}{n} - \frac{X_{(2)} - X_{(1)}}{2K}\big] \\ X_{(i)} + K\big(u - (\frac{i}{n} - \frac{X_{(i+1)} - X_{(i)}}{2K})\big) & \text{if } u \in \big[\frac{i}{n} - \frac{X_{(i+1)} - X_{(i)}}{2K}, \frac{i}{n} + \frac{X_{(i+1)} - X_{(i)}}{2K}\big] \\ & \text{for } 1 \leqslant i \leqslant n-1 \\ X_{(i+1)} & \text{if } u \in \big[\frac{i}{n} + \frac{X_{(i+1)} - X_{(i)}}{2K}, \frac{i+1}{n} - \frac{X_{(i+2)} - X_{(i+1)}}{2K}\big] \\ & \text{for } 1 \leqslant i \leqslant n-2 \\ X_{(n)} & \text{if } u \in \big[\frac{n-1}{n} + \frac{X_{(n)} - X_{(n-1)}}{2K}, 1\big]. \end{cases} \tag{3}$$

Observe that $\widehat{G}^\star_K$ is piecewise linear and that the condition $K \geqslant n \max_{i=1,\ldots,n-1}(X_{(i+1)} - X_{(i)})$ ensures that this function is well-defined. We also note that $\widehat{G}^\star_K \in \mathrm{Lip}_K([0,1], \mathbb{R})$ and that it visits each data point, going iteratively from $X_{(i)}$ to $X_{(i+1)}$. A typical example is shown in Figure 2. Observe that, for each $i \in \{1, \ldots, n\}$,

$$\lambda_1\big(\{u \in [0,1] : |\widehat{G}^\star_K(u) - X_i| \leqslant |\widehat{G}^\star_K(u) - X_j| \;\; : \;\; j = 1, \ldots, n\}\big) = \frac{1}{n}. \tag{4}$$

This geometric feature has an interpretation in terms of Voronoi cells and will play an important role in the multivariate extension of $\widehat{G}^\star_K$, as we will see in Section 5.

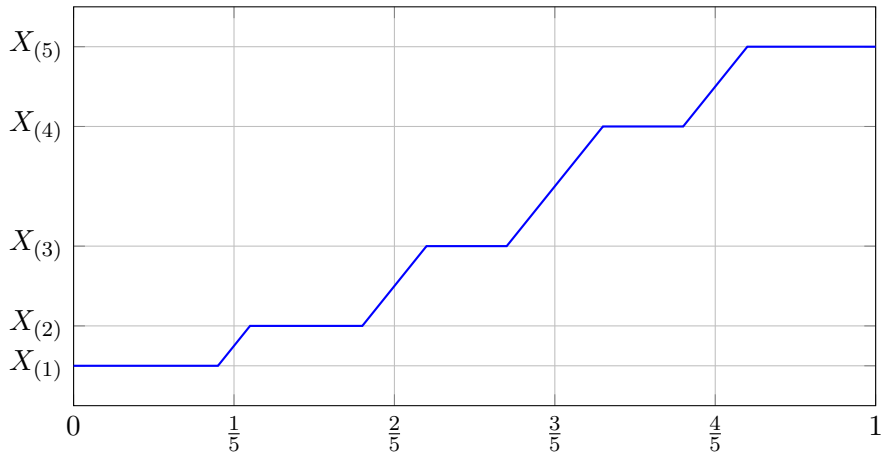


Figure 2: An example of function $\widehat{G}_K^\star$, with $n = 5$ and $K = 25$.

Proposition 4 *Assume that* $K \geqslant n \max_{i=1,\dots,n-1}(X_{(i+1)} - X_{(i)})$, *and let* $\widehat{G}_K^\star \in \mathrm{Lip}_K([0,1],\mathbb{R})$ *be defined in* (3). *Then*

$$W_1(\widehat{G}^\star_{K\sharp U}, \mu_n) = \frac{1}{4K}\sum_{i=1}^{n-1}(X_{(i+1)} - X_{(i)})^2.$$

The key message of Proposition 4 is that the 1-Wasserstein distance between $\widehat{G}^\star_{K\sharp U}$ and μ_n depends on the sum of the *squared* distances $(X_{(i+1)} - X_{(i)})^2$. We are now in a position to state the main result of the section.

Theorem 5 *Assume that* $K \geqslant n \max_{i=1,\dots,n-1}(X_{(i+1)} - X_{(i)})$, *and let the function* $\widehat{G}_K^\star \in \mathrm{Lip}_K([0,1],\mathbb{R})$ *be defined in* (3). *Then*

$$W_1(\widehat{G}^\star_{K\sharp U}, \mu_n) = \inf_{G\in \mathrm{Lip}_K([0,1],\mathbb{R})} W_1(G_\sharp U, \mu_n) = \frac{1}{4K}\sum_{i=1}^{n-1}(X_{(i+1)} - X_{(i)})^2.$$

Moreover, $\widehat{\mathscr{G}}_K = \{\widehat{G}_K^\star, \widehat{G}_K^\star \circ S\}$, *where* $S(u) = 1 - u$, $u \in [0,1]$.

Theorem 5 states that there are only two minimizers in $\widehat{\mathscr{G}}_K$. Moreover, the two distributions $\widehat{G}^\star_{K\sharp U}$ and $(\widehat{G}_K^\star \circ S)_{\sharp U}$ are identical. We thus conclude that in the univariate setting, the distribution output by the WGAN Problem (1) exists and is unique, provided K is large enough. It is important to note that the distribution $\widehat{G}^\star_{K\sharp U}$ has atoms at the X_i's, of respective sizes

$$\begin{aligned} &\frac{1}{n} - \frac{X_{(2)} - X_{(1)}}{2K} \quad \text{for } X_{(1)}, \quad \frac{1}{n} - \frac{X_{(n)} - X_{(n-1)}}{2K} \quad \text{for } X_{(n)}, \\ &\qquad \frac{1}{n} - \frac{X_{(i+1)} - X_{(i-1)}}{2K} \quad \text{for } X_{(i)},\ i = 2,\dots,n-1, \end{aligned} \tag{5}$$

and that it is absolutely continuous with respect to the Lebesgue measure elsewhere.

Being able to describe the minimizers of Problem (1) helps us to better understand the overall objective of WGANs when playing with different parameters. For example, when the dataset (and thus the sample size n) is kept fixed, the 1-Wasserstein distance $W_1(\widehat{G}^\star_{K\sharp U}, \mu_n)$ decreases towards 0 as the Lipschitz constant K gets bigger. This is easily explained by the fact that when K increases, the class of generative distributions increases as well, and the measure of the atoms in (5) of $\widehat{G}^\star_{K\sharp U}$ grows towards 1. In this regime, the optimal distribution $\widehat{G}^\star_{K\sharp U}$ tends to memorize the data samples, that is the WGAN overfits the data. On the opposite, the measures of the different atoms $X_{(i)}$, $i \in \{1, \dots, n\}$, decrease with the distance $X_{(i+1)} - X_{(i-1)}$. Consequently, any outlier data, far from its nearest neighbors, will be less sampled by the optimal distribution.

In order to illustrate the result of Theorem 5, we consider a synthetic setting where both the class of generative and discriminative functions are replaced by parametric neural networks. The generator is composed of ReLU neural networks of respective depths 3 (Figure 3a and 3c) and 5 (Figure 3b and 3d), with a width 100, while the discriminator is composed of ReLU neural networks of depth 5 and width 100. The true distribution is assumed to be uniform on $[0, 10]$. We train a WGAN architecture in the setting of both $n = 5$ and $n = 9$, with the choice $K = 50$ (we choose K big enough such that $K \geqslant n \max_{i=1,\dots,n-1}(X_{(i+1)} - X_{(i)})$). The Lipschitz constraint on the generator is implemented using a gradient penalty similar to the one used for the discriminator in Gulrajani et al. (2017). The obtained results are depicted in Figure 3.

We see that the parametric WGANs (denoted by G^θ) get close to the optimal function $\widehat{G}^\star_K$ while operating some smoothing. This smoothing is due to the fact that the networks cannot replicate all Lipschitz functions. Therefore, the optimal parametric WGANs have a higher 1-Wasserstein distance to the empirical distribution than $\widehat{G}^\star_K$. Interestingly, as the number of samples increases and the architecture remains fixed, it gets more complicated for the generator to memorize the dataset. As expected, the results of the parametric WGANs get better as the depth of the generator increases.

Changing a little bit the way of looking at the problem, one may take an asymptotic point of view in the sample size n and analyze the asymptotic behavior of the 1-Wasserstein distance $W_1(\widehat{G}^\star_{K\sharp U}, \mu)$, as done in Section 1. However, a major difference is that, in accordance with Theorem 5, the Lipschitz constant K is now viewed as a data-dependent random variable larger than $\underline{K}_1$, where

$$\underline{K}_1 := n \max_{i=1,\dots,n-1}(X_{(i+1)} - X_{(i)}).$$

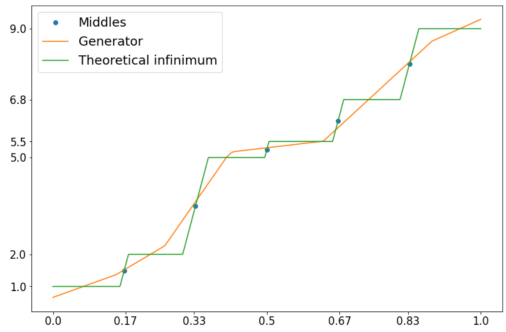
Proposition 6 *Assume that $S(\mu) = [A, B]$, where $-\infty < A < B < \infty$.*

1. *If μ admits a strictly positive probability density on $[A, B]$, continuously differentiable, with a unique minimum on $[A, B]$, then*

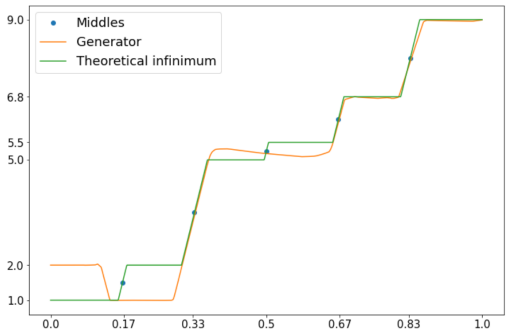
$$\frac{1}{\underline{K}_1} = \mathscr{O}((\log n)^{-1}) \text{ a.s.}$$

2. *For all $K \geqslant \underline{K}_1$,*

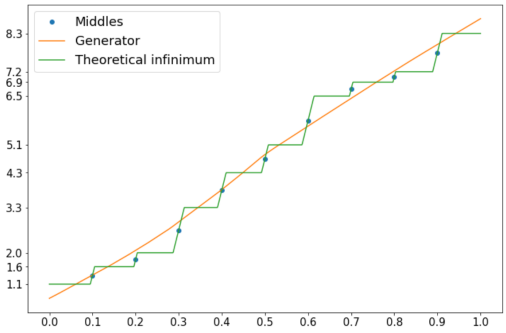
$$W_1(\widehat{G}^\star_{K\sharp U}, \mu) = \mathscr{O}(n^{-1/2}) \text{ in probability.}$$



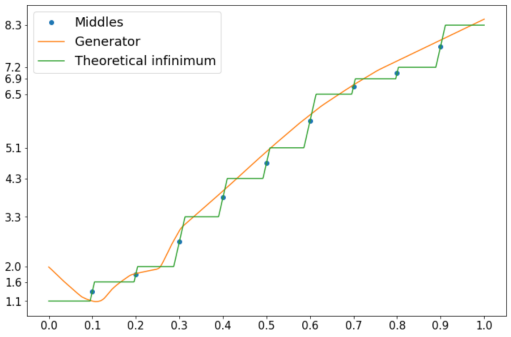
(a) Fitting $n = 5$ data points with a generator depth equal to 3. $W_1(\widehat{G}^\star_{K\sharp U}, \mu_n) = 0.080$ and $W_1(G^\theta_{\sharp U}, \mu_n) = 0.501$.



(b) Fitting $n = 5$ data points with a generator depth equal to 5. $W_1(\widehat{G}^\star_{K\sharp U}, \mu_n) = 0.080$ and $W_1(G^\theta_{\sharp U}, \mu_n) = 0.165$.



(c) Fitting $n = 9$ data points with a generator depth equal to 3. $W_1(\widehat{G}^\star_{K\sharp U}, \mu_n) = 0.033$ and $W_1(G^\theta_{\sharp U}, \mu_n) = 0.280$.



(d) Fitting $n = 9$ data points with a generator depth equal to 5. $W_1(\widehat{G}^\star_{K\sharp U}, \mu_n) = 0.033$ and $W_1(G^\theta_{\sharp U}, \mu_n) = 0.210$.

Figure 3: Output functions G^θ of the WGANs compared with the optimal $\widehat{G}^\star_K$.

The proof of Proposition 6 reveals that $W_1(\widehat{G}^\star_{K\sharp U}, \mu_n) = \mathcal{O}(n^{-1})$ in probability, which should be compared with the rate $W_1(\mu, \mu_n) = \mathcal{O}(n^{-1/2})$ (Fournier and Guillin, 2015, Theorem 1). Therefore, the speed of convergence to 0 of $W_1(\widehat{G}^\star_{K\sharp U}, \mu)$ is significantly slowed down by the term $W_1(\mu, \mu_n)$. Besides, the assumptions on μ are made here for simplicity, and many other cases may be handled similarly by connecting $\underline{K}_1$ to statistical results regarding the analysis of maximal spacings. For example, built on results from Extreme Values Theory, Deheuvels (1986, Theorem 1 or Example 1) entails that when μ is standard Gaussian, then, in probability,

$$\frac{1}{\underline{K}_1} = \mathcal{O}\Big(\frac{\sqrt{\log n}}{n}\Big) \quad \text{and} \quad W_1(\widehat{G}^\star_{K\sharp U}, \mu) = \mathcal{O}(n^{-1/2}).$$

Similar results, yet with different rates, may be obtained for the Cauchy and Gamma distributions (Deheuvels, 1986, Example 2 and Example 3). The general message is that, provided the class of candidate distributions grows with the sample size n, then the WGANs can asymptotically recover the target distribution μ.

4. A general result in semi-discrete optimal transport

We now turn to the multivariate case, assuming that the observations $X_1, \ldots, X_n$ are i.i.d. according to an unknown distribution μ on $\mathbb{R}^d$, $d > 1$. As we will see below, characterizing the optimal transport problem is much more complicated in the multivariate setting, and requires a more involved analysis. The key is to better describe the optimal transport function between $G_{\sharp U}$ and μ_n, keeping in mind that for $d > 1$, $G_{\sharp U}$ is never absolutely continuous with respect to the Lebesgue measure λ_d on $\mathbb{R}^d$. Therefore, we need to extend existing results to larger classes of distributions.

Recall, as we saw in the introduction, that for two probability measures π_1 and π_2 on $\mathbb{R}^d$,

$$W_1(\pi_1, \pi_2) = \inf_{\pi \in \Pi(\pi_1, \pi_2)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\| \mathrm{d}\pi(x, y),$$

where $\Pi(\pi_1, \pi_2)$ denotes the collection of all transport plans between π_1 and π_2, that is, the joint probability measures π on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals π_1 and π_2. When π_1 is *nonatomic*, then, according to Pratelli (2007, Theorem B),

$$W_1(\pi_1, \pi_2) = \inf_T \int_{\mathbb{R}^d} \|x - T(x)\| \mathrm{d}\pi_1(x), \tag{6}$$

where the infimum is taken over all measurable functions $T : \mathbb{R}^d \to \mathbb{R}^d$ satisfying $T_{\sharp \pi_1} = \pi_2$. Such a function T is called a transport map from π_1 to π_2, and (6) is referred to as the Monge formulation of the 1-Wasserstein distance. Providing existence and unicity results for transport maps is, in general, a difficult question. It turns out however that in the so-called semi-discrete setting, where π_1 is absolutely continuous with respect to the Lebesgue measure and $\pi_2 = \sum_{i=1}^n \alpha_i \delta_{x_i}$ is discrete ($\alpha_i \geqslant 0$, $\sum_{i=1}^n \alpha_i = 1$), the Monge problem has a simple and elegant solution in terms of additively weighted Voronoi diagram of $\mathbb{R}^d$ (e.g., Aurenhammer et al., 1998) around the atoms $\{x_1, \ldots, x_n\}$ of π_2. Recall that for a vector $w = (w_1, \ldots, w_n) \in \mathbb{R}^n$ that assigns to each x_i a weight w_i, the additively weighted Voronoi tessellation is the set of cells

$$\mathrm{Vor}^w(i) = \big\{x \in \mathbb{R}^d : \|x - x_i\| - w_i \leqslant \|x - x_j\| - w_j \text{ for all } j \neq i\big\}, \quad i = 1, \ldots, n.$$

Now, according to Hartmann and Schuhmacher (2020, Theorem 2 and Theorem 3) (see also Geiß et al., 2013), there exists in this semi-discrete setting a π_1-almost surely unique transport map $T^\star$ such that $W_1(\pi_1, \pi_2) = \int_{\mathbb{R}^d} \|x - T^\star(x)\| \mathrm{d}\pi_1(x)$. Noting that the intersection of two boundaries has Lebesgue measure zero (and thus π_1-measure zero by absolute continuity), this optimal function $T^\star$ is defined λ_d-almost surely and has the form

$$T^\star(x) = \sum_{i=1}^n x_i \mathbb{1}\{x \in \mathrm{Vor}^{w^\star}(i)\}, \tag{7}$$

where the weight vector $w^\star$ is *adapted* to (π_1, π_2) in the sense that $\pi_1(\mathrm{Vor}^{w^\star}(i)) = \alpha_i$ for all $i \in \{1, \ldots, n\}$. The existence of such an adapted vector is guaranteed by Theorem 3 of Hartmann and Schuhmacher, 2020, who also provide an algorithm to compute it.

Returning to the WGAN problem, it seems natural to consider the semi-discrete setting with $\pi_1 = G_{\sharp U}$ and $\pi_2 = \mu_n$, and to describe the optimal transport maps between these

two distributions in order to gain information on $W_1(G_\sharp U, \mu_n)$. Unfortunately, there is no reason for $G_\sharp U$ to be nonatomic and, even if this is the case, it is impossible for this distribution to be absolutely continuous with respect to the Lebesgue measure as soon as $d > 1$. We therefore conclude that none of the above results can be used to characterize the infimum in (1) and that some extensions are needed. In the rest of this section, we address this issue and offer a solution in two steps. First, we prove in Proposition 7 that the WGAN optimization in Problem (1) can be safely restricted to distributions $G_\sharp U$ that are nonatomic. Second, we provide in Theorem 9 a solution to the Monge problem under the sole assumption that π_1 is nonatomic with compact support, getting rid of the absolute continuity requirement. To the extent of our knowledge, this is the first time that such a theorem has been proved, and it therefore provides a new resource in the toolbox of optimal transport theory.

Proposition 7 *Let* $\mathrm{Lip}_K^-([0,1],\mathbb{R}^d) = \{G \in \mathrm{Lip}_K([0,1],\mathbb{R}^d) : G_\sharp U \text{ is nonatomic}\}$. *Then*

$$\inf_{G\in \mathrm{Lip}_K([0,1],\mathbb{R}^d)} W_1(G_\sharp U, \mu_n) = \inf_{G\in \mathrm{Lip}_K^-([0,1],\mathbb{R}^d)} W_1(G_\sharp U, \mu_n).$$

In the following, for $w \in \mathbb{R}^n$ and $i \in \{1,\dots,n\}$, we let $\mathrm{Vor}^w(i)$ be the i-th weighted Voronoi cell associated with the sample $X_1,\dots,X_n$. We denote by $\partial\mathrm{Vor}^w(i)$ the boundary of $\mathrm{Vor}^w(i)$ and let $\mathrm{Vor}^w(i)^\circ = \mathrm{Vor}^w(i) \setminus \partial\mathrm{Vor}^w(i)$ be its interior. For any $p \in \{1,\dots,n\}$ and any set $\{j_1,\dots,j_p\}$ where the j_k's are all different and in $\{1,\dots,n\}^p$, we let

$$\Gamma^w_{j_1\dots j_p} = \bigcap_{k=1}^p \mathrm{Vor}^w(j_k) \setminus \Big(\bigcup_{\ell\notin\{j_1,\dots,j_p\}} \mathrm{Vor}^w(\ell)\Big). \tag{8}$$

Observe that each set $\Gamma^w_{j_1\dots j_p}$ above is the subset of the common boundary of the Voronoi cells $\mathrm{Vor}^w(j_1),\dots,\mathrm{Vor}^w(j_p)$ that has no intersection with any other cell $\mathrm{Vor}^w(l)$, for all $l \notin \{j_1,\dots,j_p\}$. Note also that together, the $\Gamma^w_{j_1\dots j_p}$ (for all p and all different sets $\{j_1,\dots,j_p\}$) form a partition of the set of the boundaries of the Voronoi cells. For a given w, we will be interested in the class $\mathscr{H}^w$ of functions taking values in the sample $X_1,\dots,X_n$ defined by

$$\begin{aligned}\mathscr{H}^w = \big\{T : \mathbb{R}^d \to \{X_1,\dots,X_n\} \ : \ &\forall x \in \mathrm{Vor}^w(i)^\circ, T(x) = X_i \\ &\text{and } \forall x \in \Gamma^w_{j_1\dots j_p}, T(x) \in \{X_{j_1},\dots,X_{j_p}\}\big\}.\end{aligned} \tag{9}$$

The following result states under which assumptions we can find an optimal transport map from a nonatomic probability measure ν to the empirical measure μ_n.

Proposition 8 *Let* ν *be a probability measure on* $\mathbb{R}^d$ *with finite first moment. If there exists* $w^\star \in \mathbb{R}^n$ *and* $T^\star \in \mathscr{H}^{w^\star}$ *such that* $T^\star_\sharp \nu = \mu_n$, *then* $T^\star$ *is an optimal transport map from* ν *to* μ_n.

We deduce from Proposition 8 that in order to state the existence of an optimal transport map, it is enough to show that there exist $w^\star \in \mathbb{R}^n$ and $T^\star \in \mathscr{H}^{w^\star}$ such that $T^\star_\sharp \nu = \mu_n$. This result plays a key role in the proof of the next theorem, which guarantees the existence of an optimal transport map between *any* nonatomic probability measure ν (so, non necessarily absolutely continuous with respect to the Lebesgue measure) and the empirical measure μ_n. It should be stressed that Theorem 9 also holds if the empirical measure μ_n is replaced by a more general discrete measure, with a finite number of atoms. The adaptation is easy and is left to the reader.

Theorem 9 *Let ν be a nonatomic probability measure on $\mathbb{R}^d$ with compact support. Then there exists an optimal transport map from ν to μ_n, which is defined λ_d-almost everywhere by*

$$T^\star(x) = \sum_{i=1}^{n} X_i \mathbb{1}\{x \in \mathrm{Vor}^{w^\star}(i)\},$$

for some $w^\star \in \mathbb{R}^n$.

Proof Let $K \subseteq \mathbb{R}^d$ be the compact support of ν. For $\varepsilon \in (0,1]$, we let ν_ε be the probability measure on $\mathbb{R}^d$ defined for any Borel subset A by

$$\nu_\varepsilon(A) = \int_{\mathbb{R}^d} \frac{\lambda_d(A \cap B(x,\varepsilon))}{\lambda_d(B(x,\varepsilon))} \mathrm{d}\nu(x),$$

where $B(x,\varepsilon)$ stands for the closed ball centered at x of radius ε. Observe that ν_ε has compact support K_ε, where

$$K_\varepsilon = \{x \in \mathbb{R}^d : \exists z \in K \text{ such that } \|x - z\| \leqslant \varepsilon\}.$$

Since, for any Borel subset A such that $\lambda_d(A) = 0$ one has $\nu_\varepsilon(A) = 0$, we see that ν_ε is absolutely continuous with respect to the Lebesgue measure. Thus, according to Hartmann and Schuhmacher (2020), there exists $w_\varepsilon = (w_{\varepsilon_1}, \ldots, w_{\varepsilon_n}) \in \mathbb{R}^n$ solution to the Monge problem between ν_ε and μ_n. In particular, for each $i \in \{1,\ldots,n\}$, $\nu_\varepsilon(\mathrm{Vor}^{w_\varepsilon}(i)) = \frac{1}{n}$.

Clearly, adding a constant to each w_{ε_i} does not change the definition of the cells. Thus, in the sequel, it is assumed that $w_{\varepsilon_1} = 0$. Let $C = \max_{i=1,\ldots,n} \max_{z \in K_\varepsilon} \|X_i - z\|$. If there exists $i \in \{2,\ldots,n\}$ such that $w_{\varepsilon_i} > C$, then $\mathrm{Vor}^{w_\varepsilon}(1) \cap K_\varepsilon = \emptyset$. This is not possible since $\nu_\varepsilon(\mathrm{Vor}^{w_\varepsilon}(1)) = \frac{1}{n}$. Likewise, if $w_{\varepsilon_i} < -C$, then $\mathrm{Vor}^{w_\varepsilon}(i) \cap K_\varepsilon = \emptyset$. Therefore, we may only consider w_ε's such that $\|w_\varepsilon\|_\infty \leqslant C$, where $\|\cdot\|_\infty$ stands for the supremum norm on $\mathbb{R}^n$.

Consider now the sequence $(w_{1/m})_{m \in \mathbb{N}^\star}$, which, as we have seen, takes its values in a compact set. Thus, there exists a subsequence $(w_{1/\varphi(m)})_{m \in \mathbb{N}^\star}$ that converges to some $w^\star \in \mathbb{R}^n$. As for now, to lighten the notation, we let $\psi(m) = 1/\varphi(m)$ for all $m \in \mathbb{N}^\star$. Clearly, we have

$$\begin{aligned}
\nu_{\psi(m)}(\mathrm{Vor}^{w_{\psi(m)}}(i)) = &\int_{\mathbb{R}^d} \mathbb{1}\{x \in \mathrm{Vor}^{w^\star}(i)^c\} \frac{\lambda_d\big(\mathrm{Vor}^{w_{\psi(m)}}(i) \cap B(x,\psi(m))\big)}{\lambda_d\big(B(x,\psi(m))\big)} \mathrm{d}\nu(x) \\
&+ \int_{\mathbb{R}^d} \mathbb{1}\{x \in \mathrm{Vor}^{w^\star}(i)^\circ\} \frac{\lambda_d\big(\mathrm{Vor}^{w_{\psi(m)}}(i) \cap B(x,\psi(m))\big)}{\lambda_d\big(B(x,\psi(m))\big)} \mathrm{d}\nu(x) \\
&+ \int_{\mathbb{R}^d} \mathbb{1}\{x \in \partial\mathrm{Vor}^{w^\star}(i)\} \frac{\lambda_d\big(\mathrm{Vor}^{w_{\psi(m)}}(i) \cap B(x,\psi(m))\big)}{\lambda_d\big(B(x,\psi(m))\big)} \mathrm{d}\nu(x). \quad (10)
\end{aligned}$$

For $x \in \mathrm{Vor}^{w^\star}(i)^c$ and all m large enough,

$$\frac{\lambda_d\big(\mathrm{Vor}^{w_{\psi(m)}}(i) \cap B(x,\psi(m))\big)}{\lambda_d\big(B(x,\psi(m))\big)} = 0.$$

Therefore, by dominated convergence, the first integral in identity (10) tends to 0 as m tends to infinity. Similarly, for $x \in \mathrm{Vor}^{w^\star}(i)^\circ$ and all m large enough,

$$\frac{\lambda_d\big(\mathrm{Vor}^{w_{\psi(m)}}(i) \cap B(x,\psi(m))\big)}{\lambda_d\big(B(x,\psi(m))\big)} = 1.$$

Thus, by dominated convergence, the second integral tends to $\nu(\mathrm{Vor}^{w^\star}(i)^\circ)$. The analysis of the third integral in (10) is more delicate and is done by carefully studying each part of the boundary $\partial\mathrm{Vor}^{w^\star}(i)$. For any $p \in \{1,\dots,n\}$ and $j_1,\dots,j_p$ all different, let

$$\Gamma^{w^\star}_{j_1\dots j_p} = \bigcap_{k=1}^{p} \mathrm{Vor}^{w^\star}(j_k) \setminus \Big(\bigcup_{\ell \notin \{j_1,\dots,j_p\}} \mathrm{Vor}^{w^\star}(\ell) \Big).$$

Using the notation

$$\alpha_{j_1\dots j_p}(i)_{\psi(m)} := \int_{\mathbb{R}^d} \mathbb{1}\{x \in \Gamma^{w^\star}_{j_1\dots j_p}\} \frac{\lambda_d\big(\mathrm{Vor}^{w_{\psi(m)}}(i) \cap B(x,\psi(m))\big)}{\lambda_d\big(B(x,\psi(m))\big)} \mathrm{d}\nu(x),$$

we see that

$$\nu(\Gamma^{w^\star}_{j_1\dots j_p}) = \sum_{i=1}^{n} \alpha_{j_1\dots j_p}(i)_{\psi(m)}.$$

Observe that for $i \notin \{j_1,\dots,j_p\}$, $\alpha_{j_1\dots j_p}(i)_{\psi(m)} \to 0$ as m tends to infinity, since for all $x \in \Gamma^{w^\star}_{j_1\dots j_p}$,

$$\mathrm{Vor}^{w_{\psi(m)}}(i) \cap B(x,\psi(m)) = \emptyset$$

for all m large enough. Moreover, $\alpha_{j_1\dots j_p}(j_1)_{\psi(m)} \in [0,1/n]$. Thus, we can extract a subsequence $(\psi_1(m))_{m\in\mathbb{N}^\star}$ such that $\alpha_{j_1\dots j_p}(j_1)_{\psi_1(m)}$ converges to some $\alpha_{j_1\dots j_p}(j_1)$ as m tends to infinity. Likewise, we can extract a subsequence $\psi_{12}(m)$ such that $\alpha_{j_1\dots j_p}(j_2)_{\psi_{12}(m)}$ converges to some $\alpha_{j_1\dots j_p}(j_2)$. Repeating the same procedure, we obtain a subsequence $\psi_{1\dots p}(m)$ such that each $\alpha_{j_1\dots j_p}(j_k)_{\psi_{1\dots p}(m)}$, $k \in \{1,\dots,p\}$, converges to $\alpha_{j_1\dots j_p}(j_k)$ as m tends to infinity. In particular,

$$\nu(\Gamma^{w\star}_{j_1\dots j_p}) = \sum_{k=1}^{p} \alpha_{j_1\dots j_p}(j_k).$$

Starting from the subsequence $\psi_{1\dots p}(m)$, we may repeat the previous exercise for all sets $\Gamma^{w^\star}_{j_1\dots j_p}$, where $p \in \{1,\dots,n\}$, $j_1,\dots,j_p$ are all different, and the subsequence $\Psi(m)$ is such that any $\alpha_{j_1\dots j_p}(j_k)_{\Psi(m)}$ converges to some $\alpha_{j_1\dots j_p}(j_k)$, for all $j_1,\dots,j_p$. We conclude that there exists a subsequence of the third integral in (10) that converges to $\sum_{p=1}^{n} \sum_{j_1,\dots,j_p} \alpha_{j_1\dots j_p}(i)$. Since, for $i \in \{1,\dots,n\}$, $\nu_{\Psi(m)}(\mathrm{Vor}^{w_{\Psi(m)}}(i)) = \frac{1}{n}$ for all m, we have, letting $m \to \infty$,

$$\nu(\mathrm{Vor}^{w^\star}(i)^\circ) + \sum_{p=1}^{n} \sum_{j_1,\dots,j_p} \alpha_{j_1\dots j_p}(i) = \frac{1}{n}. \tag{11}$$

Now, cut each $\Gamma^{w^\star}_{j_1\dots j_p}$ into arbitrarily p disjoints parts $A_{j_1\dots j_p}(j_k)$ such that $\nu(A_{j_1\dots j_p}(j_k)) = \alpha_{j_1\dots j_p}(j_k)$ (this is always possible since ν is nonatomic). Let $T^\star : \mathbb{R}^d \to \{X_1, \dots, X_n\}$ be defined by

$$T^\star(x) = \begin{cases} X_i & \text{for } x \in \mathrm{Vor}^{w^\star}(i)^\circ \\ X_{j_k} & \text{for } x \in A_{j_1\dots j_p}(j_k). \end{cases}$$

Then $T^\star \in \mathscr{H}^{w^\star}$ and, by identity (11), $T^\star_{\sharp\nu} = \mu_n$. This, together with Proposition 8, concludes the proof of the theorem. ■

While the expression of the optimal transport map as given in Theorem 9 (for nonatomic source measure) and the one from Hartmann and Schuhmacher (2020), as recalled in equation (7) (for absolutely continuous source measure), are the same, there is a significant difference in their definition at the boundaries of the cells. Indeed, these boundaries have Lebesgue measure zero. Therefore, when the source measure is absolutely continuous, the optimal mapping can take any values at the boundaries. However, when the source is assumed to be only nonatomic, the boundaries may have strictly positive measure. Consequently, the choice of values for an optimal mapping at the boundaries should be made with care. In the proof of Theorem 9, we show that for each set $\Gamma^{w^\star}_{j_1\dots j_p}$ and each $i \in \{1, \dots n\}$, there exist weights $\alpha_{j_1\dots j_p}(i) \geq 0$ such that

$$\nu(\Gamma^{w\star}_{j_1\dots j_p}) = \sum_{k=1}^{p} \alpha_{j_1\dots j_p}(j_k)$$

and

$$\nu(\mathrm{Vor}^{w^\star}(i)^\circ) + \sum_{p=1}^{n} \sum_{j_1,\dots,j_p} \alpha_{j_1\dots j_p}(i) = \frac{1}{n}.$$

Then, cutting each subset $\Gamma^{w^\star}_{j_1\dots j_p}$ into p arbitrarily disjoint parts $A_{j_1\dots j_p}(j_k)$ such that $\nu(A_{j_1\dots j_p}(j_k)) = \alpha_{j_1\dots j_p}(j_k)$ and defining $T^\star : \mathbb{R}^d \to \{X_1, \dots, X_n\}$ by

$$T^\star(x) = \begin{cases} X_i & \text{for } x \in \mathrm{Vor}^{w^\star}(i)^\circ \\ X_{j_k} & \text{for } x \in A_{j_1\dots j_p}(j_k), \end{cases}$$

we obtain that $T^\star$ is an optimal transport map.

Combining the result of Proposition 7 with Theorem 9, we can now properly characterize the 1-Wasserstein distance between the optimal distribution $G_{\sharp U}$ derived from $\mathrm{Lip}_K([0,1], \mathbb{R}^d)$ and the empirical measure μ_n.

5. Finite sample analysis in a multivariate output space

We are now ready to analyze Problem (1) in the more realistic multivariate setting. In the remainder of the section, it is therefore assumed that the observations $X_1, \dots, X_n$ take their values in $\mathbb{R}^d$ with $d > 1$, while the latent space still has dimension 1. Following the schema of Section 3, we first define a candidate function $\widehat{G}^\star_K \in \mathrm{Lip}_K([0,1], \mathbb{R}^d)$, compute $W_1(\widehat{G}^\star_{K\sharp U}, \mu_n)$ in Proposition 10, and finally show in Theorem 9 that $\widehat{G}^\star_K$ solves Problem (1) in a large subset of $\mathrm{Lip}_K([0,1], \mathbb{R}^d)$. Finally, similarly to Section 3, we conclude with an asymptotic analysis of $W_1(\widehat{G}^\star_{K\sharp U}, \mu_n)$ when K is a function of the sample size n.

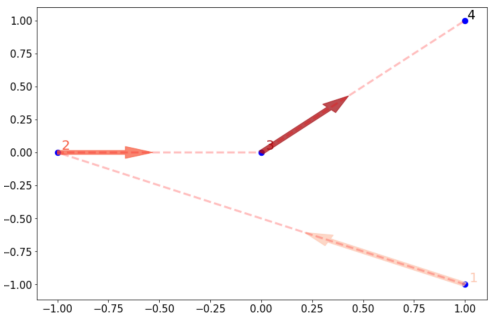
5.1 Construction of $\widehat{G}_K^\star$

In the multivariate setting, the shortest path among the n data samples X_i, $i \in \{1, \dots, n\}$, plays an essential role in the definition of the optimal $\widehat{G}_K^\star$. The set of paths connecting all data points $X_1, \dots, X_n$, while minimizing the sum of the squared Euclidean distances, is defined as follows:

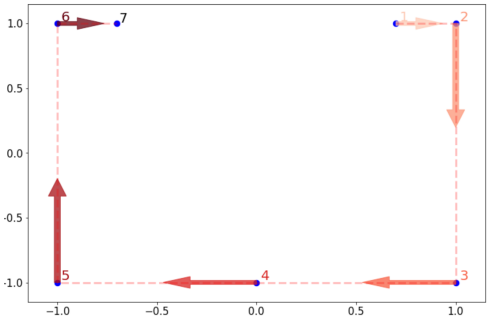
$$(k, \sigma) \in \arg\min \left\{ \sum_{i=1}^{n+k'-1} \|X_{\sigma'(i+1)} - X_{\sigma'(i)}\|^2 : k' \in \mathbb{N}, \sigma' \in \mathscr{S}_{k'} \right\}, \tag{12}$$

where $\mathscr{S}_{k'}$ denotes the set of all discrete functions $\sigma' : \{1, \dots, n + k'\} \to \{1, \dots, n\}$ such that $\sigma'(\{1, \dots, n+k'\}) = \{1, \dots, n\}$ and $\sigma'(j) \neq \sigma'(j+1)$. Note that such a pair (k, σ) may not be unique and keep in mind that σ depends on k.

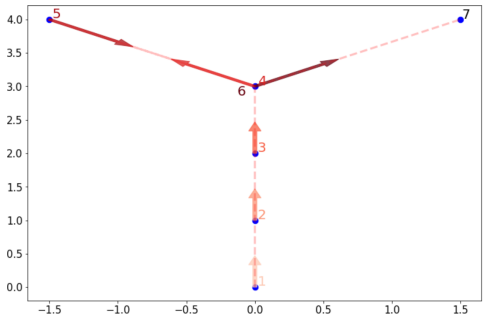
An important remark is that any shortest path (with a squared norm) is allowed to visit several times the same point (i.e., $k > 0$). This is a consequence of the fact that the squared Euclidean distance does not verify the triangle inequality. Note also that the number of visits to the point X_i is equal to $|\sigma^{-1}(i)|$. An illustration of four shortest paths in dimension 2 is provided in Figure 4. On the top, every single data point is visited once (i.e., $k = 0$ in formula (12)), contrary to the two examples in the bottom, where a point is visited twice (i.e., $k = 1$).



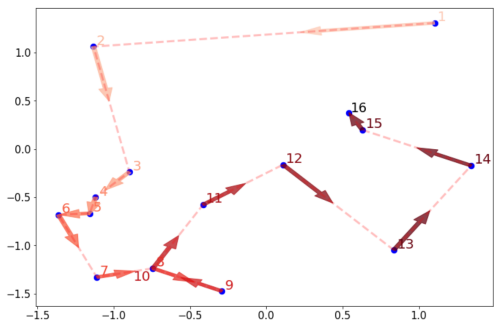
(a) Shortest path with $n = 4$, $k = 0$ in (12).



(b) Shortest path with $n = 7$, $k = 0$ in (12).



(c) Shortest path with $n = 6$, $k = 1$ in (12).



(d) Shortest path with $n = 15$, $k = 1$ in (12).

Figure 4: Examples of shortest paths in dimension 2, with the squared Euclidean distance.

Let us now provide some intuition on the way the optimal function $\widehat{G}_K^\star : [0,1] \to \mathbb{R}^d$ is obtained. In a nutshell, this function strictly follows σ, one of the optimal paths in (12). Thus, there exist some $0 \leqslant t_1 < \cdots < t_{n+k} \leqslant 1$ such that $\widehat{G}_K^\star(t_j) = X_{\sigma(j)}$, $j \in \{1, \ldots, n+k\}$. Since the optimal path (and therefore $\widehat{G}_K^\star$) can visit several times each sample point X_i, we need to take into account how long $\widehat{G}_K^\star$ stays constant at X_i, whenever it visits this data point. This period of time is denoted by $\varphi(i)$ and chosen to be equal to

$$\varphi(i) = \frac{1}{|\sigma^{-1}(i)|}\Big(\frac{1}{n} - \sum_{j \in \sigma^{-1}(i)} \frac{1}{2K}\big(\|X_{\sigma(j-1)} - X_i\| + \|X_{\sigma(j+1)} - X_i\|\big)\Big)$$

(by convention, $X_{\sigma(0)} = X_{\sigma(1)}$ and $X_{\sigma(n+k+1)} = X_{\sigma(n+k)}$). The quantity $|\sigma^{-1}(i)| \times \varphi(i)$ thus corresponds to the total measure of the atoms X_i under the distribution $\widehat{G}_{K\sharp U}^\star$. Finally, for any $j \in \{1, \ldots, n+k\}$, we let

$$V_j = \sum_{\ell=1}^{j-1} \Big(\varphi(\sigma(\ell)) + \frac{\|X_{\sigma(\ell+1)} - X_{\sigma(\ell)}\|}{K}\Big) = V_{j-1} + \varphi(\sigma(j-1)) + \frac{\|X_{\sigma(j)} - X_{\sigma(j-1)}\|}{K}.$$

This quantity V_j is more complicated to grasp, but intuitively, it corresponds to the time steps where the function $\widehat{G}_K^\star$ has arrived on a sample point $X_{\sigma(j)}$ and will pause at $X_{\sigma(j)}$ for a time equal to $\varphi(\sigma(j))$.

A visual explanation of the construction mechanism of $\widehat{G}_K^\star$ is depicted in Figure 5. The top shows the trajectory of $\widehat{G}_K^\star$ following an optimal path σ in (12). The bottom shows the succession of time steps at which $\widehat{G}_K^\star$ passes from one point to another.

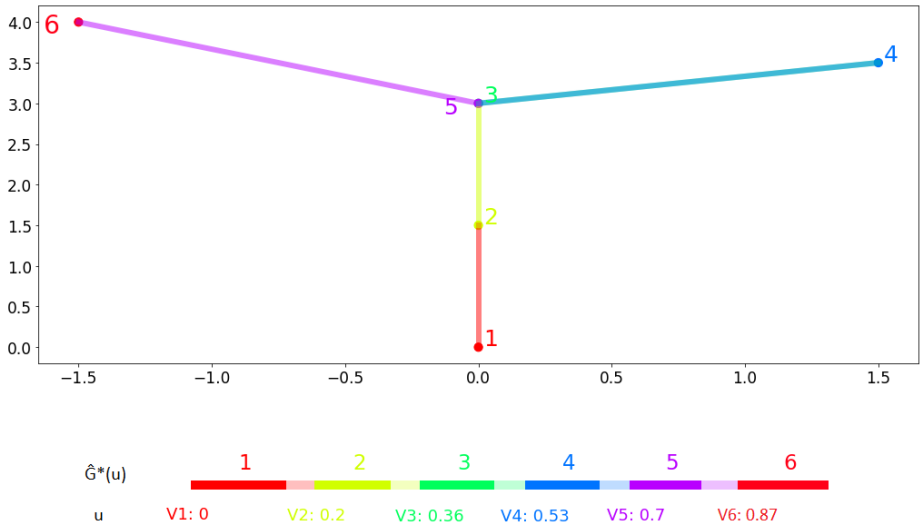


Figure 5: $\widehat{G}_K^\star$ explained. For each data point $X_{\sigma(j)}$ (embedded by a specific color), the bold part of the interval symbolizes the time $\widehat{G}_K^\star$ is equal to $X_{\sigma(j)}$, while the light part refers to the jump from $X_{\sigma(j)}$ to $X_{\sigma(j+1)}$. Note that $\widehat{G}_K^\star$ follows the optimal path under the squared Euclidean norm.

Equipped with this notation, we may now properly define the function $\widehat{G}_K^\star : [0,1] \to \mathbb{R}^d$, as follows:

$$\widehat{G}_K^\star(u) = \begin{cases} X_{\sigma(j)} & \text{if } u \in [V_j, V_j + \varphi(\sigma(j))] \\ & \text{for } 1 \leqslant j \leqslant n+k \\ X_{\sigma(j)} + \big(u - (V_j + \varphi(\sigma(j)))\big) K \frac{X_{\sigma(j+1)} - X_{\sigma(j)}}{\|X_{\sigma(j+1)} - X_{\sigma(j)}\|} & \text{if } u \in [V_j + \varphi(\sigma(j)), V_{j+1}] \\ & \text{for } 1 \leqslant j \leqslant n+k-1. \end{cases} \tag{13}$$

Observe that the function $\widehat{G}_K^\star$ is well-defined as soon as

$$K \geqslant n \max_{i=1,\dots,n} \sum_{j \in \sigma^{-1}(i)} \frac{1}{2}(\|X_{\sigma(j-1)} - X_i\| + \|X_{\sigma(j+1)} - X_i\|),$$

and that it belongs to $\mathrm{Lip}_K([0,1], \mathbb{R}^d)$. Making the connection with the univariate case of Section 3, we have that if $d = 1$, then each point is visited only once, so that $k = 0$ and, for each $i \in \{1, \dots, n\}$, $X_{\sigma(i)} = X_{(i)}$ (or $X_{\sigma(i)} = X_{(n-i+1)}$). Besides, $|\sigma^{-1}(i)| = 1$ and $\varphi(i) = 1/n - (X_{(i+1)} - X_{(i-1)})/(2K)$. We thus recover the univariate function defined in (3).

5.2 Optimality properties

In this subsection, we first compute the 1-Wasserstein distance between $\widehat{G}^\star_{K\sharp U}$ and μ_n, and then prove that this value minimizes Problem (1) in the multivariate setting, under a mild assumption.

Proposition 10 *Assume that*

$$K \geqslant n \max_{i=1,\dots,n} \sum_{j \in \sigma^{-1}(i)} \frac{1}{2}(\|X_{\sigma(j-1)} - X_i\| + \|X_{\sigma(j+1)} - X_i\|),$$

and let $\widehat{G}_K^\star \in \mathrm{Lip}_K([0,1], \mathbb{R}^d)$ *be defined in* (13). *Then*

$$W_1(\widehat{G}^\star_{K\sharp U}, \mu_n) = \frac{1}{4K} \sum_{j=1}^{n+k-1} \|X_{\sigma(j+1)} - X_{\sigma(j)}\|^2.$$

By construction, it is clear that $\widehat{G}_K^\star$ visits each data point, following the optimal path σ. The proof of Proposition 10 reveals that $\widehat{G}_K^\star$ spends a time $1/n$ (in terms of Lebesgue measure) in each "standard" Voronoi cell $\mathrm{Vor}(i)$, that is

$$\mathrm{Vor}(i) = \big\{x \in \mathbb{R}^d : \|x - X_i\| \leqslant \|x - X_j\| \text{ for all } j \neq i\big\}, \quad i = 1, \dots, n.$$

These cells correspond to additively weighted Voronoi cells with weight $w = (0, \dots, 0)$. We define in the same way $\Gamma^0_{j_1 \dots j_p}$ and $\mathscr{H}^0$ as in (8) and (9), respectively.

In the remainder of the subsection, we prove the optimality of $\widehat{G}_K^\star$ on a subset smaller than $\mathrm{Lip}_K([0,1], \mathbb{R}^d)$. This subset is denoted by $\mathrm{Lip}^\circ_K([0,1], \mathbb{R}^d)$ and is defined below. Recall

that for any $G \in \mathrm{Lip}_K([0,1],\mathbb{R}^d)$ such that $G_\sharp U$ is nonatomic, there exists according to Theorem 9 a weight $w \in \mathbb{R}^n$ and an optimal transport map T^w from $G_\sharp U$ to μ_n such that

$$W_1(G_\sharp U, \mu_n) = \int_{\mathbb{R}^d} \|x - T^w(x)\| \mathrm{d}G_\sharp U(x) = \int_0^1 \|G(u) - T^w(G(u))\| \mathrm{d}u.$$

Definition 11 *Let $G \in \mathrm{Lip}_K([0,1],\mathbb{R}^d)$. We say that G is in $\mathrm{Lip}_K^\circ([0,1],\mathbb{R}^d)$ if $G_\sharp U$ is nonatomic and, for all $u \in [0,1]$ and all $i \in \{1,\dots,n\}$ such that $T^w(G(u)) = X_i$, there exists $v \in [0,1]$ such that $G(v) = X_i$ and $\forall x \in [u,v]$ (or $[v,u]$), $T^w(G(x)) = X_i$.*

Definition 11 means that as soon as the function G enters a weighted Voronoi cell, then it must passes through its center. Even though $\widehat{G}^\star_{K\sharp U}$ has atoms, the following theorem shows that $W_1(\widehat{G}^\star_{K\sharp U}, \mu_n)$ achieves the infimum of Problem (1) over $\mathrm{Lip}_K^\circ([0,1],\mathbb{R}^d)$.

Theorem 12 *Assume that $K \geqslant n \max_{i=1,\dots,n} \sum_{j\in\sigma^{-1}(i)} \frac{1}{2}(\|X_{\sigma(j-1)} - X_i\| + \|X_{\sigma(j+1)} - X_i\|)$, and let the function $\widehat{G}^\star_K \in \mathrm{Lip}_K([0,1],\mathbb{R}^d)$ be defined in (13). Then*

$$W_1(\widehat{G}^\star_{K\sharp U}, \mu_n) = \inf_{G\in\mathrm{Lip}_K^\circ([0,1],\mathbb{R}^d)} W_1(G_\sharp U, \mu_n) = \frac{1}{4K} \sum_{j=1}^{n+k-1} \|X_{\sigma(j+1)} - X_{\sigma(j)}\|^2.$$

Proof Let $G \in \mathrm{Lip}_K^\circ([0,1],\mathbb{R}^d)$. According to Theorem 9, there exists a weight $w \in \mathbb{R}^n$ and an optimal transport map T^w from $G_\sharp U$ to μ_n. We denote by $[a_1,a_2],[a_2,a_3],\dots,[a_{p-1},a_p]$ the intervals such that $a_1 = 0$, $a_p = 1$, and, for all $j \in \{1,\dots,p-1\}$, there exists $\tau(j) \in \{1,\dots,n\}$ such that $u \in [a_j, a_{j+1}]$ implies that $T^w(G(u)) = X_{\tau(j)}$ (with $\tau(j) \neq \tau(j+1)$).

Using the fact that G is K-Lipschitz and satisfies Definition 11, it is easy to see that

$$a_{j+1} - a_j \geqslant \frac{1}{K}\big(\|G(a_j) - X_{\tau(j)}\| + \|G(a_{j+1}) - X_{\tau(j)}\|\big).$$

Observe that

$$\begin{aligned} W_1(G_\sharp U, \mu_n) &= \int_{\mathbb{R}^d} \|x - T^w(x)\| \mathrm{d}G_\sharp U(x) = \int_0^1 \|G(u) - T^w(G(u))\| \mathrm{d}u \\ &= \sum_{j=1}^{p-1} \int_{a_j}^{a_{j+1}} \|G(u) - X_{\tau(j)}\| \mathrm{d}u. \end{aligned}$$

Therefore,

$$\begin{aligned} W_1(G_\sharp U, \mu_n) \geqslant & \sum_{j=1}^{p-1} \int_{a_j}^{a_j + \frac{\|G(a_j) - X_{\tau(j)}\|}{K}} (\|G(a_j) - X_{\tau(j)}\| - K(u - a_j)) \mathrm{d}u \\ & + \int_{a_{j+1} - \frac{\|G(a_{j+1}) - X_{\tau(j)}\|}{K}}^{a_{j+1}} K\Big(u - \Big(a_{j+1} - \frac{\|G(a_{j+1}) - X_{\tau(j)}\|}{K}\Big)\Big) \mathrm{d}u \\ = & \frac{1}{2K} \sum_{j=1}^{p-1} \big(\|G(a_j) - X_{\tau(j)}\|^2 + \|G(a_{j+1}) - X_{\tau(j)}\|^2\big). \end{aligned}$$

Observe that, by the triangle inequality,

$$\|G(a_j) - X_{\tau(j-1)}\| + \|G(a_j) - X_{\tau(j)}\| \geqslant \|X_{\tau(j-1)} - X_{\tau(j)}\|.$$

So,

$$\|G(a_j) - X_{\tau(j-1)}\|^2 + \|G(a_j) - X_{\tau(j)}\|^2 \geqslant \frac{1}{2}\|X_{\tau(j-1)} - X_{\tau(j)}\|^2.$$

Using (12) (Main Document), we conclude that

$$W_1(G_{\sharp U}, \mu_n) \geqslant \frac{1}{4K} \sum_{j=1}^{p-1} \|X_{\tau(j-1)} - X_{\tau(j)}\|^2 \geqslant W_1(\widehat{G}^\star_{K\sharp U}, \mu_n).$$

Therefore,

$$\inf_{G \in \mathrm{Lip}^\circ_K([0,1],\mathbb{R}^d)} W_1(G_{\sharp U}, \mu_n) \geqslant W_1(\widehat{G}^\star_{K\sharp U}, \mu_n).$$

Finally, a slight adaptation of Proposition 7 shows that

$$W_1(\widehat{G}^\star_{K\sharp U}, \mu_n) \geqslant \inf_{G \in \mathrm{Lip}^\circ_K([0,1],\mathbb{R}^d)} W_1(G_{\sharp U}, \mu_n),$$

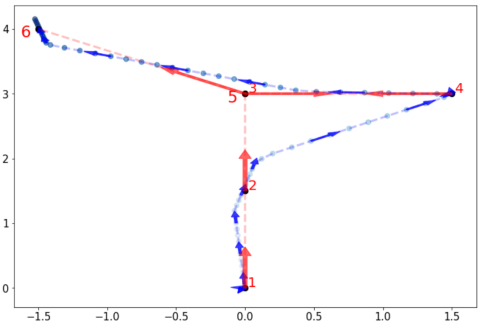
and the theorem is proved. ■

Note that we could not show the optimality of $\widehat{G}^\star_K$ on $\mathrm{Lip}_K([0,1],\mathbb{R}^d)$. However, all the numerical experiments indicate that the generative functions G^θ output by WGANs satisfy $W_1(\widehat{G}^\star_{K\sharp U}, \mu_n) < W_1(G^\theta_{\sharp U}, \mu_n)$. Consequently, restricting the set of Lipschitz continuous functions to $\mathrm{Lip}^\circ_K([0,1],\mathbb{R}^d)$ might not be necessary. We leave it as an open problem to prove that $\widehat{G}^\star_K$ is indeed the infimum over the whole set $\mathrm{Lip}_K([0,1],\mathbb{R}^d)$. Similarly to the univariate case, the distribution $\widehat{G}^\star_{K\sharp U}$ has atoms located at the sample points X_i, with respective mass $|\sigma^{-1}(i)| \times \varphi(i)$. It is also noteworthy that the minimizer $\widehat{G}^\star_K$ is not necessarily unique, because there may be different paths σ minimizing the sum of the squared Euclidean distances in (12). Furthermore, if $|\sigma^{-1}(i)| \geqslant 2$, one can arbitrarily choose how to split the time period $|\sigma^{-1}(i)| \times \varphi(i) = \widehat{G}^\star_{K\sharp U}(\{X_i\})$ according to the different moments $\widehat{G}^\star_K$ passes by X_i.

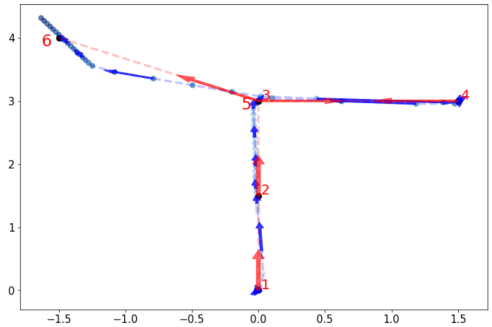
In order to illustrate Theorem 12, we propose in Figure 6 a 2-dimensional experiment that compares the 1-Wasserstein distance $W_1(\widehat{G}^\star_{K\sharp U}, \mu_n)$ with the results of parametric WGANs. The generator is composed of ReLU neural networks of depth 3 and 6, and width 100, while the discriminator is composed of ReLU neural networks of depth 5 and width 100. We train a WGAN architecture on two different configurations, $n = 5$ for the first and $n = 10$ for the second, both with the choice $K = 50$ (compatible with the assumption on K in Theorem 12). We see, as expected, that the parametric WGAN (denoted by $G^\theta_{\sharp U}$) gets close to the optimal function $\widehat{G}^\star_K$. However, since neural networks lack capacity and cannot replicate all Lipschitz functions, they operate some smoothing. Finally, observe that as n grows, mimicking the optimal function $\widehat{G}^\star_K$ is harder, while increasing the depth can help.

Theorem 12 is valid under the condition $K \geqslant \underline{K}_2$, where

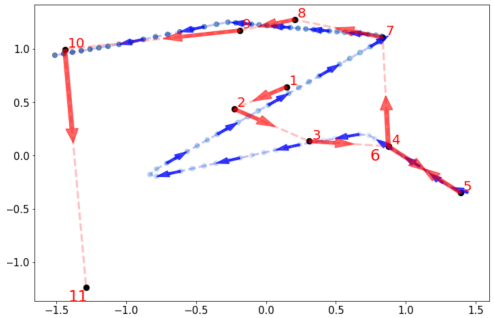
$$\underline{K}_2 := n \max_{i=1,\dots,n} \sum_{j \in \sigma^{-1}(i)} \frac{1}{2}(\|X_{\sigma(j-1)} - X_i\| + \|X_{\sigma(j+1)} - X_i\|).$$



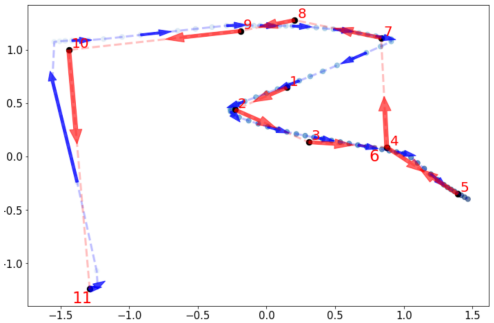
(a) The sample size is $n = 5$ and the depth of the generator is equal to 3. The WGAN misses the shortest path leading to a deteriorated 1-Wasserstein distance: $W_1(\widehat{G}^\star_{K\sharp U}, \mu_n) = 0.030$ and $W_1(G^\theta_{\sharp U}, \mu_n) = 0.286$.



(b) The sample size is $n = 5$ and the depth of the generator is equal to 6. The WGAN is closer to the shortest path: $W_1(\widehat{G}^\star_{K\sharp U}, \mu_n) = 0.018$ and $W_1(G^\theta_{\sharp U}, \mu_n) = 0.174$.



(c) The sample size is $n = 10$ and the depth of the generator is equal to 3. The WGAN misses the shortest path, with a worsened 1-Wasserstein distance: $W_1(\widehat{G}^\star_{K\sharp U}, \mu_n) = 0.025$ and $W_1(G^\theta_{\sharp U}, \mu_n) = 0.321$.



(d) The sample size is $n = 10$ and the depth of the generator is equal to 6. The WGAN is closer to the shortest path: $W_1(\widehat{G}^\star_{K\sharp U}, \mu_n) = 0.025$ and $W_1(G^\theta_{\sharp U}, \mu_n) = 0.160$.

Figure 6: Fitting 2-dimensional data points with a univariate latent space. The blue curves are the ones reached after optimization with WGANs $G^\theta_{\sharp U}$ and the red curves are the constructed ones $G^\star_{K\sharp U}$

.

As $\underline{K}_2$ (and thus K) is a function of n, it is therefore natural to understand the behavior of $W_1(\widehat{G}^\star_{K\sharp U}, \mu)$ when n tends to infinity.

Proposition 13 *Assume that μ has a probability density with respect to the Lebesgue measure on $\mathbb{R}^d$ and that $S(\mu)$ is bounded.*

1. *We have*

$$\frac{1}{\underline{K}_2} = \mathcal{O}(n^{-1+1/d}) \ a.s.$$

2. *If, in addition, the density of μ is bounded away from 0 on $S(\mu)$, then, for all $K \geqslant \underline{K}_2$, in probability,*

$$W_1(\widehat{G}^\star_{K\sharp U}, \mu) = \begin{cases} \mathscr{O}(\frac{\log n}{\sqrt{n}}) & \text{for } d = 2 \\ \mathscr{O}(n^{-1/d}) & \text{for } d \geqslant 3. \end{cases}$$

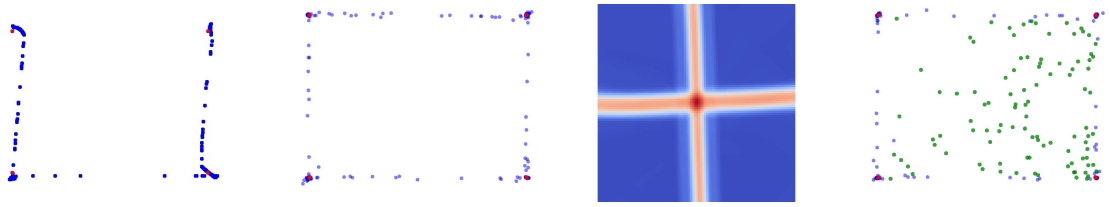
The proof of Proposition 13 reveals that, for $d \geqslant 2$, $W_1(\widehat{G}^\star_{K\sharp U}, \mu_n) = \mathscr{O}(n^{-1/d})$ in probability, which coincides with the rate of $W_1(\mu, \mu_n)$ for $d \geqslant 3$ (Fournier and Guillin, 2015, Theorem 1). However, for $d = 2$, $W_1(\mu, \mu_n) = \mathscr{O}(\frac{\log n}{\sqrt{n}})$, and the speed of convergence to 0 of $W_1(\widehat{G}^\star_{K\sharp U}, \mu)$ is therefore slightly slowed down by the term $W_1(\mu, \mu_n)$. In essence, this proposition states that while $K \geqslant \underline{K}_2$ tends to infinity as n grows, the infimum is taken over a larger collection of functions, which enables $\widehat{G}^\star_{K\sharp U}$ to get closer to the target distribution μ for the 1-Wasserstein distance. Liang (2021) derived minimax-type results for classes of absolutely continuous distributions defined with Sobolev constraints.

6. Conclusion

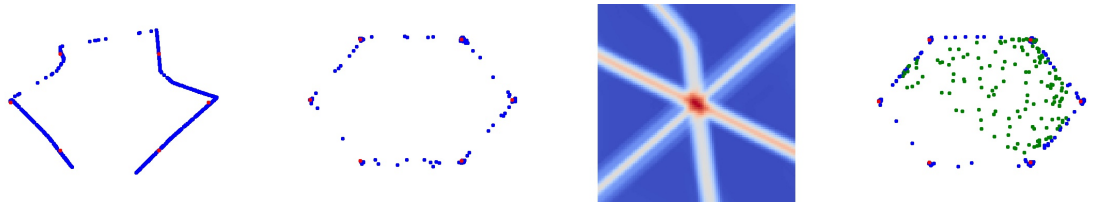
We provided in this paper a thorough analysis of the properties of WGANs, in both the finite sample and asymptotic regimes. Although the dimension of the latent space is assumed to be equal to 1, the results are valid regardless of the dimension d of the output space. In this setting, we showed that for a fixed sample size n, optimal WGANs are closely linked with connected paths minimizing the sum of the squared Euclidean distances between the sample points. We also highlighted the fact that WGANs are able to approach (for the 1-Wasserstein distance) the target distribution as n tends to infinity, at a given convergence rate and provided the family of Lipschitz functions grows with K. We derived in passing new results on optimal transport theory in the semi-discrete setting. In a nutshell, the main message is that WGANs generate data that lie on very specific regions of the ambient space—thus showing some limited "creativity"— while being able to asymptotically recover the unknown distribution of the observations under appropriate assumptions.

Nevertheless, many questions remain open and should, in our eyes, be given special attention. First, the current approach is based on a somewhat ideal definition of WGANs, in the sense that we use $\mathrm{Lip}_K([0,1], \mathbb{R}^d)$ and $\mathrm{Lip}_1(\mathbb{R}^d, \mathbb{R})$ for, respectively, the generator and the discriminator. However, one should keep in mind that in practice both the generator and the discriminator are implemented by deep neural networks. It follows that the results of the paper have to be appreciated in light of the approximation capabilities of neural networks. In particular, larger datasets will require deeper and more expressive networks to reconstruct the optimal functions $\widehat{G}^\star_K$. Also, using neural networks, the sample points in the dataset are less likely to be overfitted, thus getting closer to the true purpose of generative models, which is to mimic the observations without resampling from the learning database. We believe that studying the potential benefits of this regularization effect is an interesting problem. Next, it was assumed throughout that the latent random variable U is uniform. The extension to latent variables with unbounded support, such as Gaussian distributions, is not straightforward and requires careful investigation.

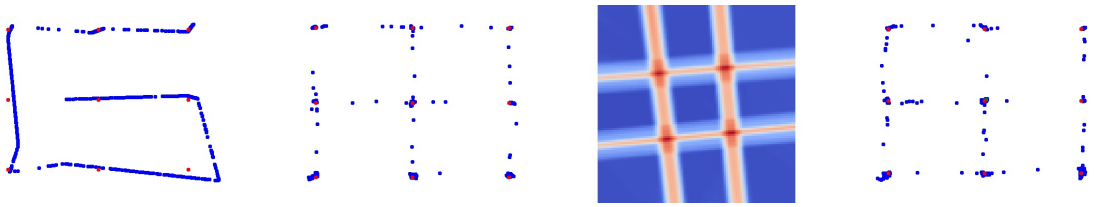
Finally, an interesting research direction is to understand and analyze the mechanisms of WGANs when the dimension p of the latent space is strictly larger than 1. In this context, the univariate shortest paths will be replaced by surfaces, and the interesting question



(a) $W_1(\widehat{G}^\star_{K\sharp U}, \mu_n) = 0.27$. (b) $W_1(\bar{G}^\star_{K\sharp U}, \mu_n) = 0.15$. (c) Latent space heatmap. (d) Supp. green points.



(e) $W_1(\widehat{G}^\star_{K\sharp U}, \mu_n) = 0.16$. (f) $W_1(\bar{G}^\star_{K\sharp U}, \mu_n) = 0.10$. (g) Latent space heatmap. (h) Supp. green points.



(i) $W_1(\widehat{G}^\star_{K\sharp U}, \mu_n) = 0.42$. (j) $W_1(\bar{G}^\star_{K\sharp U}, \mu_n) = 0.13$. (k) Latent space heatmap. (l) Supp. green points.

Figure 7: Influence of the dimension p of the latent space. In the left column, we use a uniform latent distribution on $[0, 1]$ (target points in red, sampled points in blue). In the second column, we use a uniform latent distribution on $[0, 1]^2$ (the optimal generator is denoted by $\bar{G}^\star_K$). The third column shows heatmaps of the gradients' norm of the optimal generator (the bluer the lower and the redder the higher). Finally, the last column shows supplementary points sampled close to $(\frac{1}{2}, \frac{1}{2})$ (in the latent space).

will then be to understand the driving forces of WGANs when $p < d$ and $p = d$. As a teaser, we show in Figure 7 the impact of increasing the dimension of the latent space from $p = 1$ to $p = 2$, in the case where data (in red) lie in dimension $d = 2$. We note that when $p = 1$, the WGAN is able to find the shortest paths for the squared Euclidean distance, as predicted by the theory. For $p = 2$, the situation is quite intriguing since the 1-Wasserstein distance between the empirical measure and the pushforward distribution of U by the optimal function G is decreasing. Besides, the generated distributions seem to be concentrated with positive mass on the data points and, with decreasing probabilities, on a path—theoretically undetermined—linking them. Note however that it seems also possible to generate samples anywhere in the convex hull of the data points. This is illustrated in the fourth column of the figure, where we voluntarily sample latent vectors close to the center $(\frac{1}{2}, \frac{1}{2})$. We visualize on the heatmaps in the third column the appearance of areas with high gradients of the optimal generator, dividing the latent space. Analyzing the geometrical properties of these latent configurations is a very exciting avenue for future research.

Acknowledgments

The authors thank J. Lambolley and M. Pierre for stimulating and fruitful discussions. They also thank the Editor-in-Chief and two anonymous referees for their careful reading of the paper and constructive comments, which led to a substantial improvement of the document.

References

L. Ambrosio and N. Gigli. A user's guide to optimal transport. In B. Piccoli and M. Rascle, editors, *Modelling and Optimisation of Flows on Networks: Cetraro, Italy 2009*, Berlin, 2013. Springer.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In D. Precup and Y.W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 214–223. PMLR, 2017.

F. Aurenhammer, F. Hoffmann, and B. Aronov. Minkowski-type theorems and least-squares clustering. *Algorithmica*, 20:61–76, 1998.

G. Biau, M. Sangnier, and U. Tanielian. Some theoretical insights into Wasserstein GANs. *J. Mach. Learn. Res.*, 22(119):1–45, 2021.

A. Borji. Pros and cons of GAN evaluation measures. *Comput. Vis. Image Underst.*, 179: 41–65, 2019.

P. Deheuvels. Strong limit theorems for maximal spacings from a general univariate distribution. *Ann. Probab.*, 12:1181–1193, 1984.

P. Deheuvels. On the influence of the extremes of an i.i.d. sequence on the maximal spacings. *Ann. Probab.*, 14:194–208, 1986.

L.C. Evans and R.F. Gariepy. *Measure Theory and Fine Properties of Functions*. CRC Press, Boca Raton, 2015.

E. Facco, M. d'Errico, A. Rodriguez, and A. Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7:12140, 2017.

C. Fefferman, S. Mitter, and H. Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29:983–1049, 2016.

N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Related Fields*, 162:707–738, 2015.

D. Geiß, R. Klein, R. Penninger, and G. Rote. Optimally solving a transportation problem using Voronoi diagrams. *Comput. Geom.*, 46:1009–1016, 2013.

I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling,

C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680. Curran Associates, Inc., 2014.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A.C. Courville. Improved training of Wasserstein GANs. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5767–5777. Curran Associates, Inc., 2017.

I. Gulrajani, C. Raffel, and L. Metz. Towards GAN benchmarks which require generalization. In *International Conference on Learning Representations*, 2019.

V. Hartmann and D. Schuhmacher. Semi-discrete optimal transport: A solution procedure for the unsquared Euclidean distance case. *Math. Methods Oper. Res.*, 92:133–163, 2020.

L.V. Kantorovich and G.S. Rubinstein. On a space of completely additive functions. *Vestnik Leningrad Univ. Math.*, 13:52–59, 1958.

T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4396–4405, 2019.

N. Kodali, J. Abernethy, J. Hays, and Z. Kira. On convergence and stability of GANs. *arXiv.1705.07215*, 2017.

T. Liang. How well generative adversarial networks learn distributions. *J. Mach. Learn. Res.*, 22(228):1–41, 2021.

M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet. Are GANs created equal? A large-scale study. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 697–706. Curran Associates, Inc., 2018.

G. Luise, M. Pontil, and C. Ciliberto. Generalization properties of optimal transport GANs with latent distribution learning. *arXiv:2007.14641*, 2020.

L. Mescheder, A. Geiger, and S. Nowozin. Which training methods for GANs do actually converge? In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 3481–3490. PMLR, 2018.

A. Müller. Integral probability metrics and their generating classes of functions. *Adv. in Appl. Probab.*, 29:429–443, 1997.

A. Pratelli. On the equality between Monge's infimum and Kantorovich's minimum in optimal mass transportation. *Ann. Inst. Henri Poincaré Probab. Stat.*, 43:1–13, 2007.

A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In Y. Bengio and Y. LeCun, editors, *4th International Conference on Learning Representations*, 2016.

F. Santambrogio. *Optimal Transport for Applied Mathematicians*. Birkäuser, Cham, 2015.

N. Schreuder, V.-E. Brunel, and A. Dalalyan. Statistical guarantees for generative models without domination. In V. Feldman, K. Ligett, and S. Sabato, editors, *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132, pages 1051–1071. PMLR, 2021.

S. Singh, A. Uppal, B. Li, C.-L. Li, M. Zaheer, and B. Poczos. Nonparametric density estimation under adversarial losses. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 10225–10236. Curran Associates, Inc., 2018.

J.M. Steele. Growth rates of Euclidean minimal spanning trees with power weighted edges. *Ann. Probab.*, 16:1767–1787, 1988.

A. Stéphanovitch, U. Tanielian, B. Cadre, N. Klutchnikoff, and G. Biau. Supplement to "Optimal 1-Wasserstein distance for WGANs". 2023.

U. Tanielian, T. Issenhuth, E. Dohmatob, and J. Mary. Learning disconnected manifolds: A no GAN's land. In H. Daumé III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 9418–9427. PMLR, 2020.

A. Uppal, S. Singh, and B. Poczos. Nonparametric density estimation and convergence rates for GANs under Besov IPM losses. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 9089–9100. Curran Associates, Inc., 2019.

N. Vaishnavh, C. Raffel, and I.J. Goodfellow. Theoretical insights into memorization in GANs. In *Neural Information Processing Systems 2018 - Integration of Deep Learning Theories Workshop*, 2018.

C. Villani. *Optimal Transport: Old and New*. Springer, Berlin, 2008.

C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 613–621. Curran Associates, Inc., 2016.

H. YoonHaeng, W. Guo, and T. Liang. Reversible Gromov-Monge sampler for simulation-based inference. *arXiv:2109.14090*, 2021.

L. Yu, W. Zhang, J. Wang, and Y. Yu. SeqGAN: Sequence generative adversarial nets with policy gradient. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 2852–2858. AAAI Press, 2017.

J.E. Yukich. Asymptotics for weighted minimal spanning trees on random points. *Stochastic Process. Appl.*, 85:123–138, 2000.

Z. Zhou, J. Liang, Y. Song, L. Yu, H. Wang, W. Zhang, Y. Yu, and Z. Zhang. Lipschitz generative adversarial nets. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 7584–7593. PMLR, 2019.

Appendix A. Proof of Lemma 1

We only focus on the first statement since the proof of the second one is similar. Let $G, G' \in \mathrm{Lip}_K([0,1], \mathbb{R}^d)$. Observe that by the triangle inequality and the primal definition of the 1-Wasserstein distance, we have

$$\begin{aligned} |W_1(G_{\sharp U}, \mu_n) - W_1(G'_{\sharp U}, \mu_n)| &\leqslant W_1(G_{\sharp U}, G'_{\sharp U}) \\ &\leqslant \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\| \mathrm{d}\gamma(x, y), \end{aligned}$$

where γ is the pushforward distribution of U by the pair (G, G'), with marginals $G_{\sharp U}$ and $G'_{\sharp U}$. Thus,

$$\begin{aligned} |W_1(G_{\sharp U}, \mu_n) - W_1(G'_{\sharp U}, \mu_n)| &\leqslant \int_{[0,1]} \|G(u) - G'(u)\| \mathrm{d}u \\ &\leqslant \|G - G'\|_\infty, \end{aligned}$$

where $\|\cdot\|_\infty$ denotes the supremum norm of functions, i.e., for $f : [0,1] \to \mathbb{R}^d$, $\|f\|_\infty = \sup\{\|f(x)\| : x \in [0,1]\}$. Hence the map $\mathrm{Lip}_K([0,1], \mathbb{R}^d) \ni G \mapsto W_1(G_{\sharp U}, \mu_n)$ is continuous with respect to the uniform norm.

Now let $G^0 \equiv X_1$ be a constant function on $[0,1]$. Then, clearly, $W_1(G^0_{\sharp U}, \mu_n) < \infty$. Next, let G be any function in $\mathrm{Lip}_K([0,1], \mathbb{R}^d)$ such that

$$\|G\|_\infty \geqslant W_1(G^0_{\sharp U}, \mu_n) + K + \max_{i=1,\ldots,n} \|X_i\|.$$

Then, upon observing that there exists $u_0 \in [0,1]$ such that $\|G(u_0)\| = \|G\|_\infty$ and using the fact that G is K-Lipschitz continuous on $[0,1]$, we deduce that for all $u \in [0,1]$ and any $i \in \{1, \ldots, n\}$, one has

$$\|G(u) - X_i\| \geqslant \|G\|_\infty - K - \|X_i\| \geqslant \|G\|_\infty - K - \max_{i=1,\ldots,n} \|X_i\|.$$

Hence, $\|G(u) - X_i\| \geqslant W_1(G^0_{\sharp U}, \mu_n)$, which implies that $W_1(G_{\sharp U}, \mu_n) \geqslant W_1(G^0_{\sharp U}, \mu_n)$. Therefore, letting

$$\mathscr{H}_K = \{G \in \mathrm{Lip}_K([0,1], \mathbb{R}^d) \ : \ \|G\|_\infty \leqslant W_1(G^0_{\sharp U}, \mu_n) + K + \max_{i=1,\ldots,n} \|X_i\|\},$$

we see that

$$\inf_{G \in \mathrm{Lip}_K([0,1], \mathbb{R}^d)} W_1(G_{\sharp U}, \mu_n) = \inf_{G \in \mathscr{H}_K} W_1(G_{\sharp U}, \mu_n).$$

Endowed with the uniform norm, $\mathscr{H}_K$ is closed and relatively compact by the Arzelà-Ascoli theorem. It is thus a compact subset of $\mathrm{Lip}_K([0,1],\mathbb{R}^d)$. Consequently, by continuity and the above equality, $\mathrm{Lip}_K([0,1],\mathbb{R}^d) \ni G \mapsto W_1(G_{\sharp U},\mu_n)$ attains its minimum on $\mathscr{H}_K$. Therefore, $\widehat{\mathscr{G}}_K$ is not empty.

Appendix B. Proof of Theorem 2

Proof of $1(i)$

Since μ is of order 1, one has $\lim_{n\to\infty} W_1(\mu,\mu_n) = 0$ a.s. according to Villani (2008, Theorem 6.8). Hence, by the triangle inequality and because $\widehat{G}_K \in \widehat{\mathscr{G}}_K$, we only need to prove that

$$\lim_{n\to\infty} \inf_{G\in \mathrm{Lip}_K([0,1],\mathbb{R})} W_1(G_{\sharp U},\mu_n) = 0 \text{ a.s.}$$

If $K \geqslant K_0$, then $\mathrm{Lip}_{K_0}([0,1],\mathbb{R}) \subseteq \mathrm{Lip}_K([0,1],\mathbb{R})$. Therefore,

$$0 \leqslant \inf_{G\in \mathrm{Lip}_K([0,1],\mathbb{R})} W_1(G_{\sharp U},\mu_n) \leqslant \inf_{G\in \mathrm{Lip}_{K_0}([0,1],\mathbb{R})} W_1(G_{\sharp U},\mu_n) \leqslant W_1(F^{-1}_{\sharp U},\mu_n),$$

since, by assumption, $F^{-1} \in \mathrm{Lip}_{K_0}([0,1],\mathbb{R})$. But $F^{-1}(U)$ has distribution μ, and thus one has $\lim_{n\to\infty} W_1(F^{-1}_{\sharp U},\mu_n) = 0$. This proves the result.

Proof of (2)

The result is proved by contradiction. Fix $K > 0$ and assume that on an event of strictly positive probability

$$\liminf_{n\to\infty} W_1(\widehat{G}_{K\sharp U},\mu) = 0.$$

Since $\lim_{n\to\infty} W_1(\mu,\mu_n) = 0$ a.s. and $\widehat{G}_K \in \widehat{\mathscr{G}}_K$, we see that

$$\inf_{G\in \mathrm{Lip}_K([0,1],\mathbb{R})} W_1(G_{\sharp U},\mu) = 0.$$

Now, by Lemma 1, there exists $G_K \in \mathrm{Lip}_K([0,1],\mathbb{R})$ such that

$$W_1(G_{K\sharp U},\mu) = \inf_{G\in \mathrm{Lip}_K([0,1],\mathbb{R})} W_1(G_{\sharp U},\mu).$$

So, $W_1(G_{K\sharp U},\mu) = 0$ and therefore, since $F^{-1}(U)$ has distribution μ, we have

$$G_K(U) \overset{\mathscr{L}}{\sim} F^{-1}(U). \tag{14}$$

Next, by continuity of G_K, there exists a compact set $C \subseteq \mathbb{R}$ such that $\mathbb{P}(G_K(U) \in C) = 1$. But, since $S(\mu)$ is unbounded, $\mathbb{P}(F^{-1}(U) \in C) = \mu(C) < 1$, which contradicts (14).

Proof of $1(ii)$

We show the result by contradiction, assuming as in the proof of statement (2) that for $K < 1/K_1$, on an event of strictly positive probability,

$$\liminf_{n\to\infty} W_1(\widehat{G}_{K\sharp U},\mu) = 0.$$

Arguing as in the previous proof, we have that $G_K(U) \overset{\mathscr{L}}{\sim} F^{-1}(U)$. Then, it is a classical exercise to deduce from (14), since $F^{-1}(u) > -\infty$ for all $u \in (0,1)$ and F is continuous, that $F \circ G_K(U) \overset{\mathscr{L}}{\sim} U$. Iterating this relation leads to

$$(F \circ G_K)^\ell(U) \overset{\mathscr{L}}{\sim} U, \quad \forall \ell \geqslant 0. \tag{15}$$

Moreover, both assumptions $F \in \mathrm{Lip}_{K_1}(\mathbb{R}, [0,1])$ and $G_K \in \mathrm{Lip}_K([0,1], \mathbb{R})$ imply

$$|F \circ G_K(u) - F \circ G_K(v)| \leqslant KK_1|u-v| \leqslant KK_1, \quad \forall (u,v) \in [0,1]^2.$$

Repeating this inequality entails, for all $\ell \geqslant 0$,

$$|(F \circ G_K)^\ell(u) - (F \circ G_K)^\ell(v)| \leqslant (KK_1)^\ell, \quad \forall (u,v) \in [0,1]^2.$$

But, for all $u \in [0,1]$, the sequence $((F \circ G_K)^\ell(u))_{\ell \geqslant 1}$ is bounded by 1. In addition, $KK_1 < 1$ by assumption. Thus, there exist $a \in [0,1]$ and a subsequence $(\ell_q)_{q \geqslant 1}$ such that, for all $u \in [0,1]$,

$$\lim_{q \to \infty} (F \circ G_K)^{\ell_q}(u) = a.$$

Hence, as $q \to \infty$, $(F \circ G_K)^{\ell_q}(U)$ almost surely converges to a, which contradicts (15).

Appendix C. Proof of Theorem 3

Looking for a contradiction, we start as in the proof of Theorem 2, cases $(1ii)$ and (2), by assuming that on an event of strictly positive probability,

$$\liminf_{n \to \infty} \; W_1(\widehat{G}_{K\sharp U}, \mu) = 0.$$

As we have seen, this implies $W_1(G_{K\sharp U}, \mu) = 0$ and, in turn, since the support of $G_{K\sharp U}$ is included in $G_K([0,1])$, $S(\mu) \subseteq G_K([0,1])$. By our assumption on $S(\mu)$, we therefore conclude that $\lambda_d(G_K([0,1])) > 0$. Moreover, since $G_K \in \mathrm{Lip}_K([0,1], \mathbb{R}^d)$, we have that $0 < \lambda_d(G_K([0,1])) = \mathcal{H}_d(G_K([0,1])) \leqslant K^d \mathcal{H}_d([0,1])$, where $\mathcal{H}_d$ is the d-dimensional Hausdorff measure (see, e.g., Evans and Gariepy, 2015, Theorem 2.8). But this is impossible since $\mathcal{H}_d([0,1]) = 0$ as soon as $d > 1$.

Appendix D. Proof of Proposition 4

To lighten the notation, it is assumed throughout the proof that the X_i's are ordered by increasing values, i.e., $X_1 \leqslant X_2 \leqslant \cdots \leqslant X_n$. According to Santambrogio (2015, Proposition 2.17), the 1-Wasserstein distance between two probability measures π_1 and π_2 on the real line, with respective generalized inverses F_1^{-1} and F_2^{-1}, is such that

$$W_1(\pi_1, \pi_2) = \int_0^1 |F_1^{-1}(u) - F_2^{-1}(u)| \mathrm{d}u.$$

Since $\widehat{G}_K^\star$ is monotone and continuous, the generalized inverse of $\widehat{G}_{K\sharp U}^\star$ is $\widehat{G}_K^\star$. Also, denoting by $F_{\mu_n}^{-1}$ the generalized inverse of μ_n, we have $F_{\mu_n}^{-1}(u) = \sum_{i=1}^n X_i \mathbb{1}\{u \in ((i-1)/n, i/n]\}$.

Therefore,

$$\begin{aligned}
W_1(\widehat{G}^\star_{K\sharp U}, \mu_n) &= \int_0^1 |\widehat{G}^\star_K(u) - F^{-1}_{\mu_n}(u)| \mathrm{d}u \\
&= \sum_{i=1}^{n-1} \int_{i/n - \frac{X_{i+1}-X_i}{2K}}^{i/n} \Big| X_i + K\big(u - (\frac{i}{n} - \frac{X_{i+1}-X_i}{2K})\big) - X_i \Big| \mathrm{d}u \\
&\quad + \sum_{i=1}^{n-1} \int_{i/n}^{i/n + \frac{X_{i+1}-X_i}{2K}} \Big| \frac{X_{i+1}-X_i}{2K} + K(u - \frac{i}{n}) - X_{i+1} \Big| \mathrm{d}u \\
&= \sum_{i=1}^{n-1} \frac{1}{2} K \big(\frac{(X_{i+1}-X_i)^2}{4K^2} + \frac{(X_{i+1}-X_i)^2}{4K^2} \big) \\
&= \frac{1}{4K} \sum_{i=1}^{n-1} (X_{i+1} - X_i)^2,
\end{aligned}$$

as desired.

Appendix E. Proof of Theorem 5

As in the proof of Proposition 4, it is assumed without loss of generality that the X_i's are ordered by increasing values, i.e., $X_1 \leqslant X_2 \leqslant \cdots \leqslant X_n$. Let $G : [0,1] \to \mathbb{R}$ be an arbitrary K-Lipschitz continuous function in $\widehat{\mathscr{G}}_K$, with $K \geqslant n \max_{i=1,\ldots,n-1} (X_{i+1} - X_i)$. According to Proposition 4, the first statement will be proven if we show that for such a function G,

$$W_1(G_{\sharp U}, \mu_n) \geqslant \sum_{i=1}^{n-1} \frac{(X_{i+1}-X_i)^2}{4K}.$$

Let $\Pi(\pi_1, \pi_2)$ be the set of couplings between two probability measures π_1 and π_2. According to Ambrosio and Gigli (2013, Lemma 2.12), for any $\pi \in \Pi(G_{\sharp U}, \mu_n)$, there exists a coupling $\gamma \in \Pi(\lambda_1, \mu_n)$ such that $\pi = (G, \mathrm{Id})_{\#\gamma}$, where λ_1 stands for the Lebesgue measure on the interval $[0,1]$ and Id is the identity function. Therefore,

$$\begin{aligned}
W_1(G_{\sharp U}, \mu_n) &= \inf_{\pi \in \Pi(G_{\sharp U}, \mu_n)} \int_{\mathbb{R}\times\mathbb{R}} |x - y| \mathrm{d}\pi(x, y) \\
&\geq \inf_{\gamma \in \Pi(\lambda_1, \mu_n)} \int_{[0,1]\times\mathbb{R}} |G(u) - y| \mathrm{d}\gamma(u, y).
\end{aligned}$$

Since the function $(u, y) \mapsto |G(u) - y|$ is continuous, then, according to Pratelli (2007, Theorem B), we have

$$\inf_{\gamma \in \Pi(\lambda_1, \mu_n)} \int_{[0,1]\times\mathbb{R}} |G(u) - y| \mathrm{d}\gamma(u, y) = \inf_T \int_0^1 |G(u) - T(u)| \mathrm{d}u,$$

where the infimum is taken over all measurable functions $T : [0,1] \to \{X_1, \ldots, X_n\}$ such that $T_{\sharp U} = \mu_n$. Any such transport map T takes the form $T(u) = \sum_{i=1}^n X_i \mathbb{1}\{u \in C_i\}$,

where $C_1, \dots, C_n$ are Borel subsets of $[0,1]$ such that $\lambda_1(C_i) = \frac{1}{n}$. We conclude that

$$W_1(G_\sharp U, \mu_n) \geqslant \inf_{C_1,\dots,C_n} \sum_{i=1}^n \int_{C_i} |G(u) - X_i| \mathrm{d}u, \tag{16}$$

where the infimum is taken over all disjoint Borel sets $C_1, \dots, C_n \subseteq [0,1]$ such that $\lambda_1(C_i) = \frac{1}{n}$. To prove the first statement of the theorem, it is therefore sufficient to lower bound the infimum above.

The case $n = 1$ is clear since the function $G(u) \equiv X_1$ satisfies $W_1(G_\sharp U, \mu_1) = 0$. Thus, in the sequel, it is assumed that $n \geqslant 2$. We let $a = \inf_{[0,1]} G$, $b = \sup_{[0,1]} G$, and $\ell_1 \leqslant \ell_2$ so that $X_{\ell_1} = \min_{X_i \geqslant a} X_i$ and $X_{\ell_2} = \max_{X_i \leqslant b} X_i$. Note that we can safely assume that ℓ_1 and ℓ_2 are well-defined, since for $\hat{G}(u) := G(u)\mathbb{1}\{G(u) \in [X_1, X_n]\} + X_1 \mathbb{1}\{G(u) < X_1\} + X_n \mathbb{1}\{G(u) > X_n\}$, we have

$$\inf_{C_1,\dots,C_n} \sum_{i=1}^n \int_{C_i} |G(u) - X_i| \mathrm{d}u \geqslant \inf_{C_1,\dots,C_n} \sum_{i=1}^n \int_{C_i} |\hat{G}(u) - X_i| \mathrm{d}u.$$

We also suppose that $n > \ell_2 \geqslant \ell_1 + 1 > 1$ and leave the other cases as straightforward adaptations. Since G is continuous, for each $i \in \{\ell_1, \dots, \ell_2 - 1\}$, there exists $u_i \in [0,1]$ such that $G(u_i) = \frac{X_i + X_{i+1}}{2}$. We let $A_i^- = [u_i - \frac{X_{i+1} - X_i}{2K}, u_i]$, $A_i^+ = [u_i, u_i + \frac{X_{i+1} - X_i}{2K}]$, and write $T(u) = \sum_{j=1}^n X_j \mathbb{1}\{u \in C_j\}$. With this notation,

$$\begin{aligned}
\int_{A_i^-} |G(u) - T(u)| \mathrm{d}u &= \sum_{j=1}^i \int_{A_i^-} (G(u) - X_i + X_i - X_j) \mathbb{1}\{u \in C_j\} \mathrm{d}u \\
&\quad + \sum_{j=i+1}^n \int_{A_i^-} (X_{i+1} - G(u) + X_j - X_{i+1}) \mathbb{1}\{u \in C_j\} \mathrm{d}u \\
&= \sum_{j=1}^i \Big[\int_{A_i^-} (G(u) - X_i) \mathbb{1}\{u \in C_j\} \mathrm{d}u + \lambda_1(C_j \cap A_i^-)(X_i - X_j) \Big] \\
&\quad + \sum_{j=i+1}^n \Big[\int_{A_i^-} (X_{i+1} - G(u)) \mathbb{1}\{u \in C_j\} \mathrm{d}u + \lambda_1(C_j \cap A_i^-)(X_j - X_{i+1}) \Big].
\end{aligned} \tag{17}$$

Exploiting the fact that the function G is K-Lipschitz continuous and $G(u_i) = \frac{X_i+X_{i+1}}{2}$, we have that for $u \in A_i^- \cup A_i^+$, $\frac{X_i+X_{i+1}}{2} - K|u_i - u| \leqslant G(u) \leqslant \frac{X_i+X_{i+1}}{2} + K|u_i - u|$. Thus,

$$\begin{aligned}
&\sum_{j=1}^{i} \int_{A_i^-} (G(u) - X_i)\mathbb{1}\{u \in C_j\}\mathrm{d}u + \sum_{j=i+1}^{n} \int_{A_i^-} (X_{i+1} - G(u))\mathbb{1}\{u \in C_j\}\mathrm{d}u \\
&\quad \geqslant \sum_{j=1}^{i} \int_{A_i^-} \Big(\frac{X_i + X_{i+1}}{2} - K(u_i - u) - X_i\Big)\mathbb{1}\{u \in C_j\}\mathrm{d}u \\
&\qquad + \sum_{j=i+1}^{n} \int_{A_i^-} \Big(X_{i+1} - \Big(\frac{X_i + X_{i+1}}{2} + K(u_i - u)\Big)\Big)\mathbb{1}\{u \in C_j\}\mathrm{d}u \\
&\quad = \sum_{j=1}^{n} \int_{A_i^-} \Big(\frac{X_{i+1} - X_i}{2} - K(u_i - u)\Big)\mathbb{1}\{u \in C_j\}\mathrm{d}u \\
&\quad = \int_{A_i^-} \Big(\frac{X_{i+1} - X_i}{2} - K(u_i - u)\Big)\mathrm{d}u \\
&\quad = \frac{(X_{i+1} - X_i)^2}{4K} - \frac{1}{2}\frac{(X_{i+1} - X_i)^2}{4K} \\
&\quad = \frac{(X_{i+1} - X_i)^2}{8K}.
\end{aligned} \tag{18}$$

Combining this inequality with (17) yields

$$\begin{aligned}
\int_{A_i^-} |G(u) - T(u)|\mathrm{d}u \geqslant \frac{(X_{i+1} - X_i)^2}{8K} \\
+ \sum_{j=1}^{i-1} \lambda_1(C_j \cap A_i^-)(X_i - X_j) + \sum_{j=i+1}^{n} \lambda_1(C_j \cap A_i^-)(X_j - X_{i+1}).
\end{aligned}$$

Employing the same technique for A_i^+, we obtain

$$\begin{aligned}
\int_{A_i^+} |G(u) - T(u)|\mathrm{d}u \geqslant \frac{(X_{i+1} - X_i)^2}{8K} \\
+ \sum_{j=1}^{i-1} \lambda_1(C_j \cap A_i^+)(X_i - X_j) + \sum_{j=i+1}^{n} \lambda_1(C_j \cap A_i^+)(X_j - X_{i+1}).
\end{aligned}$$

So, letting $A_i = A_i^- \cup A_i^+$ and using the fact that $X_{\ell+1} \geqslant X_\ell$ for all $\ell \leqslant n-1$, we are led to

$$\begin{aligned}
\int_{A_i} |G(u) - T(u)|\mathrm{d}u \geqslant \frac{(X_{i+1} - X_i)^2}{4K} \\
+ \sum_{j=1}^{i-1} \lambda_1(C_j \cap A_i)(X_{j+1} - X_j) + \sum_{j=i+2}^{n} \lambda_1(C_j \cap A_i)(X_j - X_{j-1}).
\end{aligned} \tag{19}$$

Now, let $u_{\ell_1-1} \in [0,1]$ be such that $G(u_{\ell_1-1}) = \frac{a+X_{\ell_1}}{2}$. With a slight abuse of notation, define $A^-_{\ell_1-1} = [u_{\ell_1-1} - \frac{X_{\ell_1}-a}{2K}, u_{\ell_1-1}]$ and $A^+_{\ell_1-1} = [u_{\ell_1-1}, u_{\ell_1-1} + \frac{X_{\ell_1}-a}{2K}]$. Then, using the same method as above, one easily shows that, for $A_{\ell_1-1} = A^-_{\ell_1-1} \cup A^+_{\ell_1-1}$,

$$\int_{A_{\ell_1-1}} |G(u) - T(u)| \mathrm{d}u \geqslant \frac{(X_{\ell_1} - a)^2}{4K} + \sum_{j=1}^{\ell_1-1} \lambda_1(C_j \cap A_{\ell_1-1})(a - X_j) + \sum_{j=\ell_1+1}^{n} \lambda_1(C_j \cap A_{\ell_1-1})(X_j - X_{\ell_1}).$$

In a similar fashion, for $u_{\ell_2} \in [0,1]$ such that $G(u_{\ell_2}) = \frac{X_{\ell_2}+b}{2}$ and, with a slight abuse of notation, letting $A_{\ell_2} = [u_{\ell_2} - \frac{b-X_{\ell_2+1}}{2K}, u_{\ell_2} + \frac{b-X_{\ell_2+1}}{2K}]$, we obtain

$$\int_{A_{\ell_2}} |G(u) - T(u)| \mathrm{d}u \geqslant \frac{(b - X_{\ell_2})^2}{4K} + \sum_{j=1}^{\ell_2-1} \lambda_1(C_j \cap A_{\ell_2})(X_{\ell_2} - X_j) + \sum_{j=\ell_2+1}^{n} \lambda_1(C_j \cap A_{\ell_2})(X_j - b).$$

Accordingly,

$$\begin{aligned}\int_{A_{\ell_1-1} \cup A_{\ell_2}} |G(u) - T(u)| \mathrm{d}u \geqslant{}& \frac{(X_{\ell_1} - a)^2}{4K} + \frac{(b - X_{\ell_2})^2}{4K} \\ &+ \sum_{j=1}^{\ell_1-2} \lambda_1(C_j \cap A_{\ell_1-1})(X_{j+1} - X_j) \\ &+ \lambda_1(C_{\ell_1-1} \cap A_{\ell_1-1})(a - X_{\ell_1-1}) \\ &+ \sum_{j=\ell_1+1}^{n} \lambda_1(C_j \cap A_{\ell_1-1})(X_j - X_{j-1}) \\ &+ \sum_{j=1}^{\ell_2-1} \lambda_1(C_j \cap A_{\ell_2})(X_{j+1} - X_j) \\ &+ \lambda_1(C_{\ell_2+1} \cap A_{\ell_2})(X_{\ell_2+1} - b) \\ &+ \sum_{j=\ell_2+2}^{n} \lambda_1(C_j \cap A_{\ell_2})(X_j - X_{j-1}). \end{aligned} \tag{20}$$

Let $B = \bigcup_{i=\ell_1-1}^{\ell_2} A_i$, and observe that the target integral can be decomposed in the following way:

$$\int_0^1 |G(u) - T(u)| \mathrm{d}u = \int_B |G(u) - T(u)| \mathrm{d}u + \int_{B^c} |G(u) - T(u)| \mathrm{d}u. \tag{21}$$

Inequalities (19) and (20) provide a lower bound on the first term on the right-hand side of (21). Let us now work out the second term. To this aim, observe that

$$\begin{aligned}\int_{B^c} |G(u) - T(u)| \mathrm{d}u &\geqslant \sum_{j=1}^{\ell_1 - 1} \int_{B^c} |G(u) - X_j| \mathbb{1}\{u \in C_j\} \mathrm{d}u \\ &\quad + \sum_{j=\ell_2+1}^{n} \int_{B^c} |G(u) - X_j| \mathbb{1}\{u \in C_j\} \mathrm{d}u \\ &\geqslant \sum_{j=1}^{\ell_1 - 2} \int_{B^c} (X_{\ell_1 - 1} - X_j) \mathbb{1}\{u \in C_j\} \mathrm{d}u \\ &\quad + \int_{B^c} (a - X_{\ell_1 - 1}) \mathbb{1}\{u \in C_{\ell_1 - 1}\} \mathrm{d}u \\ &\quad + \int_{B^c} (X_{\ell_2+1} - b) \mathbb{1}\{u \in C_{\ell_2+1}\} \mathrm{d}u \\ &\quad + \sum_{j=\ell_2+2}^{n} \int_{B^c} (X_j - X_{\ell_2+1}) \mathbb{1}\{u \in C_j\} \mathrm{d}u.\end{aligned}$$

Exploiting $\lambda_1(C_j) = \frac{1}{n}$ for $j \in \{1, \ldots, n\}$, we see that

$$\begin{aligned}\int_{B^c} |G(u) - T(u)| \mathrm{d}u &\geqslant \sum_{j=1}^{\ell_1 - 2} \Big(\frac{1}{n} - \sum_{i=\ell_1 - 1}^{\ell_2} \lambda_1(C_j \cap A_i) \Big) (X_{j+1} - X_j) \\ &\quad + \Big(\frac{1}{n} - \sum_{i=\ell_1 - 1}^{\ell_2} \lambda_1(C_{\ell_1 - 1} \cap A_i) \Big) (a - X_{\ell_1 - 1}) \\ &\quad + \Big(\frac{1}{n} - \sum_{i=\ell_1 - 1}^{\ell_2} \lambda_1(C_{\ell_2+1} \cap A_i) \Big) (X_{\ell_2+1} - b) \\ &\quad + \sum_{j=\ell_2+2}^{n} \Big(\frac{1}{n} - \sum_{i=\ell_1 - 1}^{\ell_2} \lambda_1(C_j \cap A_i) \Big) (X_j - X_{j-1}). \qquad (22)\end{aligned}$$

Thus, using identity (21) together with inequalities (19), (20), and (22), we are led to

$$\begin{aligned}\int_0^1 |G(u)-T(u)|\mathrm{d}u \geqslant{}& \frac{(X_{\ell_1}-a)^2}{4K}+\frac{(b-X_{\ell_2})^2}{4K}\\ &+\sum_{j=1}^{\ell_1-2}\Big(\frac{1}{n}-\sum_{i=\ell_1-1}^{\ell_2}\lambda_1(C_j\cap A_i)+\sum_{i=\ell_1-1}^{\ell_2}\lambda_1(C_j\cap A_i)\Big)(X_{j+1}-X_j)\\ &+\Big(\frac{1}{n}-\sum_{i=\ell_1-1}^{\ell_2}\lambda_1(C_{\ell_1-1}\cap A_i)+\sum_{i=\ell_1-1}^{\ell_2}\lambda_1(C_{\ell_1-1}\cap A_i)\Big)(a-X_{\ell_1-1})\\ &+\sum_{i=\ell_1}^{\ell_2-1}\frac{(X_{i+1}-X_i)^2}{4K}\\ &+\Big(\frac{1}{n}-\sum_{i=\ell_1-1}^{\ell_2}\lambda_1(C_{\ell_2+1}\cap A_i)+\sum_{i=\ell_1-1}^{\ell_2}\lambda_1(C_{\ell_2+1}\cap A_i)\Big)(X_{\ell_2+1}-b)\\ &+\sum_{j=\ell_2+2}^{n}\Big(\frac{1}{n}-\sum_{i=\ell_1-1}^{\ell_2}\lambda_1(C_j\cap A_i)+\sum_{i=\ell_1-1}^{\ell_2}\lambda_1(C_j\cap A_i)\Big)(X_j-X_{j-1}).\end{aligned}$$

So,

$$\begin{aligned}\int_0^1 |G(u)-T(u)|\mathrm{d}u \geqslant{}& \frac{(X_{\ell_1}-a)^2}{4K}+\sum_{i=\ell_1}^{\ell_2-1}\frac{(X_{i+1}-X_i)^2}{4K}+\frac{(b-X_{\ell_2})^2}{4K}\\ &+\sum_{j\in\{1,\ldots,\ell_1-2\}\cup\{\ell_2+1,\ldots,n-1\}}\frac{X_{j+1}-X_j}{n}+\frac{1}{n}(a-X_{\ell_1-1})\\ &+\frac{1}{n}(X_{\ell_2+1}-b).\end{aligned}$$

Since $K\geqslant n\max_{i=1,\ldots,n-1}(X_{i+1}-X_i)$, we have $\frac{X_{j+1}-X_j}{n}\geqslant\frac{(X_{j+1}-X_j)^2}{K}$, and thus

$$\begin{aligned}\frac{(X_{\ell_1}-a)^2}{4K}+\frac{1}{n}(a-X_{\ell_1-1}) &\geqslant \frac{1}{4K}\big((X_{\ell_1}-a)^2+4(a-X_{\ell_1-1})(X_{\ell_1}-X_{\ell_1-1})\big)\\ &= \frac{1}{4K}\big((X_{\ell_1}-a)^2+4(a-X_{\ell_1-1})(X_{\ell_1}-a)\\ &\quad +4(a-X_{\ell_1-1})^2\big)\\ &\geqslant \frac{(X_{\ell_1}-X_{\ell_1-1})^2}{4K}.\end{aligned}\tag{23}$$

Similarly,

$$\frac{(X_{\ell_2}-b)^2}{4K}+\frac{1}{n}(X_{\ell_2+1}-b)\geqslant\frac{(X_{\ell_2+1}-X_{\ell_2})^2}{4K}.$$

Using once again the assumption on K, we conclude that

$$\int_0^1 |G(u)-T(u)|\mathrm{d}u\geqslant\sum_{i=1}^{n-1}\frac{(X_{i+1}-X_i)^2}{4K}.$$

To complete the proof, it remains to show that $\widehat{G}_K^\star$ and $\widehat{G}_K^\star \circ S$ are the only minimizers of (1) (Main Document). Returning to inequality (23), we see that if the function G does not visit each data points, then

$$\int_0^1 |G(u) - T(u)| \mathrm{d}u > \sum_{i=1}^{n-1} \frac{(X_{i+1} - X_i)^2}{4K}.$$

Also, according to (18), for the function G to be optimal it needs to go at speed K between each observation. Finally, with equation (16), we have that an optimal G must be such that

$$\lambda_1\big(\{u \in [0,1] : |G(u) - X_i| \leqslant |G(u) - X_j|,\ j = 1, \dots, n\}\big) = \frac{1}{n},$$

a property satisfied by $\widehat{G}_K^\star$ and $\widehat{G}_K^\star \circ S$ according to (4) (Main Document). We conclude that $\widehat{G}_K^\star$ and $\widehat{G}_K^\star \circ S$ are the unique minimizers of Problem (1) (Main Document) as they are the only functions satisfying these three conditions.

Appendix F. Proof of Proposition 6

The first statement is a straightforward consequence of Deheuvels (1984, Theorem 2). Regarding the second statement, we know from Theorem 5 that, for all $K \geqslant \underline{K}_1$,

$$W_1(\widehat{G}_{K\sharp U}^\star, \mu_n) = \inf_{G \in \mathrm{Lip}_K([0,1],\mathbb{R})} W_1(G_{\sharp U}, \mu_n) = \frac{1}{4K} \sum_{i=1}^{n-1} (X_{(i+1)} - X_{(i)})^2.$$

Therefore,

$$\begin{aligned} W_1(\widehat{G}_{K\sharp U}^\star, \mu_n) &\leqslant \frac{\sum_{i=1}^{n-1} (X_{(i+1)} - X_{(i)})^2}{n \max_{i=1,\dots,n-1} (X_{(i+1)} - X_{(i)})} \\ &\leqslant \frac{1}{n} \sum_{i=1}^{n-1} (X_{(i+1)} - X_{(i)}) \\ &= \frac{1}{n} (X_{(n)} - X_{(1)}) \\ &\leqslant \frac{B - A}{n}. \end{aligned}$$

Recalling that $W_1(\mu, \mu_n) = \mathscr{O}(n^{-1/2})$ in probability (Fournier and Guillin, 2015, Theorem 1), the conclusion follows from the triangle inequality.

Appendix G. Proof of Proposition 7

The result is a consequence of the following lemma:

Lemma 14 *For each $G \in \mathrm{Lip}_K([0,1], \mathbb{R}^d)$, there exists a sequence of functions $(G_m)_{m \in \mathbb{N}}$ in $\mathrm{Lip}_K([0,1], \mathbb{R}^d)$ such that each $G_{m\sharp U}$ is nonatomic and $W_1(G_{m\sharp U}, \mu_n) \to W_1(G_{\sharp U}, \mu_n)$ as $m \to \infty$.*

Proof Let $G \in \mathrm{Lip}_K([0,1],\mathbb{R}^d)$ and $m \in \mathbb{N}$. We define G_m by slightly modifying G on each interval where it is constant. More precisely, let $\mathcal{I}$ be the set of all non degenerated connected components of $G^{-1}(\{y \in \mathbb{R}^d : \lambda_1(G^{-1}(y)) > 0\})$. This set is at most countable and, since G is continuous, it contains only disjoint closed intervals, i.e.,

$$\mathcal{I} = \{[a_\ell, b_\ell] : \ell \in \mathcal{L}\},$$

where $\mathcal{L} \subset \mathbb{N}$ and $0 \leqslant a_\ell < b_\ell \leqslant 1$. Let $K_m = \min(K, 1/m)$, $e_1 = (1, 0, \ldots, 0) \in \mathbb{R}^d$, and

$$G_m(u) = \begin{cases} G(a_\ell) + K_m\big(\frac{b_\ell - a_\ell}{2} - \big|\frac{a_\ell + b_\ell}{2} - u\big|\big)e_1 & \text{if } u \in [a_\ell, b_\ell] \text{ for some } \ell \in \mathcal{L} \\ G(u) & \text{otherwise.} \end{cases}$$

It is easy to see that $G_m \in \mathrm{Lip}_K([0,1],\mathbb{R}^d)$. Moreover, G_m is not constant over any non degenerated interval. Thus, the distribution $G_{m\sharp U}$ is nonatomic. In addition, $\|G_m - G\|_\infty \to 0$ as $m \to \infty$. In particular, for any continuous bounded function $f : \mathbb{R}^d \to \mathbb{R}$, $\|f(G_m) - f(G)\|_\infty \to 0$, so that $G_{m\sharp U} \to G_{\sharp U}$ weakly, as m tends to infinity. As the $G_{m\sharp U}$'s have supports included in the same compact set, we conclude by Villani (2008, Theorem 6.9) that $\lim_{m\to\infty} W_1(G_{m\sharp U}, G_{\sharp U}) = 0$. But, by the triangle inequality,

$$\big|W_1(G_{m\sharp U}, \mu_n) - W_1(G_{\sharp U}, \mu_n)\big| \leqslant W_1(G_{m\sharp U}, G_{\sharp U}),$$

from which $\lim_{m\to\infty} W_1(G_{m\sharp U}, \mu_n) = W_1(G_{\sharp U}, \mu_n)$ follows, as desired. ■

Appendix H. Proof of Proposition 8

Assuming that such a transport map $T^\star \in \mathscr{H}^{w^\star}$ exists, we write $w^\star_{T^\star(x)}$ instead of $w^\star_i$ whenever $T^\star(x) = X_i$, $i \in \{1, \ldots, n\}$. Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be the 1-Lipschitz map defined by

$$\varphi(x) = \|x - T^\star(x)\| - w^\star_{T^\star(x)}.$$

Since $T^\star(X_i) = X_i$ for all $i \in \{1, \ldots, n\}$, we have in particular that $\varphi(x) - \varphi(T^\star(x)) = \|x - T^\star(x)\|$. Then, denoting by

$$\partial\varphi := \{(x, y) \in \mathbb{R}^d \times \mathbb{R}^d : \varphi(x) - \varphi(y) = \|x - y\|\}$$

the superdifferential of φ (Villani, 2008, Definition 5.7), the graph of $T^\star$ is included in $\partial\varphi$. Therefore,

$$\begin{aligned} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x - T^\star(x)\| \mathrm{d}\nu(x) &= \int_{\mathbb{R}^d\times\mathbb{R}^d} (\varphi(x) - \varphi(T^\star(x))) \mathrm{d}\nu(x) \\ &= \int_{\mathbb{R}^d} \varphi(x) \mathrm{d}\nu(x) - \int_{\mathbb{R}^d} \varphi(y) \mathrm{d}\mu_n(y) \\ &\leq W_1(\nu, \mu_n). \end{aligned}$$

We conclude that $T^\star$ is an optimal transport map.

Appendix I. Proof of Proposition 10

Let us first show that, for all $i \in \{1, \dots, n+k-1\}$ and $j \notin \{\sigma(i), \sigma(i+1)\}$,

$$[V_i + \varphi(\sigma(i)), V_{i+1}] \cap \widehat{G}_K^{\star -1}(\mathrm{Vor}(j)^\circ) = \emptyset.$$

Suppose on the contrary that there exists $t \in (0,1)$ such that $Y_i := X_{\sigma(i)} + t(X_{\sigma(i+1)} - X_{\sigma(i)}) \in \mathrm{Vor}(j)^\circ$. Then

$$X_j \in B^\circ(Y_i, \|X_{\sigma(i)} - Y_i\|) \cap B^\circ(Y_i, \|X_{\sigma(i+1)} - Y_i\|),$$

where $B^\circ(x, \varepsilon)$ stands for the open ball centered at x of radius ε. Observe that for $t \leqslant 1/2$,

$$B^\circ(Y_i, \|X_{\sigma(i)} - Y_i\|) \subseteq B^\circ\Big(\frac{X_{\sigma(i)} + X_{\sigma(i+1)}}{2}, \frac{\|X_{\sigma(i+1)} - X_{\sigma(i)}\|}{2}\Big),$$

whereas for $t \geqslant 1/2$,

$$B^\circ(Y_i, \|X_{\sigma(i+1)} - Y_i\|) \subseteq B^\circ\Big(\frac{X_{\sigma(i)} + X_{\sigma(i+1)}}{2}, \frac{\|X_{\sigma(i+1)} - X_{\sigma(i)}\|}{2}\Big).$$

Consequently,

$$X_j \in B^\circ\Big(\frac{X_{\sigma(i)} + X_{\sigma(i+1)}}{2}, \frac{\|X_{\sigma(i+1)} - X_{\sigma(i)}\|}{2}\Big).$$

We deduce that $\langle X_{\sigma(i)} - X_j, X_{\sigma(i+1)} - X_j \rangle < 0$ (notation $\langle \cdot, \cdot \rangle$ means the scalar product), and so

$$\|X_{\sigma(i+1)} - X_{\sigma(i)}\|^2 > \|X_{\sigma(i+1)} - X_j\|^2 + \|X_{\sigma(i)} - X_j\|^2.$$

However, such an inequality is impossible by definition of σ. We conclude that, for all $t \in [0, 1/2]$,

$$X_{\sigma(i)} + t(X_{\sigma(i+1)} - X_{\sigma(i)}) \in \mathrm{Vor}(\sigma(i))$$

and, for all $t \in [1/2, 1]$,

$$X_{\sigma(i)} + t(X_{\sigma(i+1)} - X_{\sigma(i)}) \in \mathrm{Vor}(\sigma(i+1)).$$

Let us now turn to the computation of $W_1(\widehat{G}^\star_{K\sharp U}, \mu_n)$. First, by definition of $\varphi(i)$, for $i \in \{1, \dots, n\}$, we have

$$\begin{aligned}
&\sum_{j \in \sigma^{-1}(i)} \lambda_1\Big(\big[V_j, V_j + \varphi(i) + \frac{\|X_{\sigma(j+1)} - X_i\|}{2K}\big]\Big) \\
&+ \lambda_1\Big(\big[V_{j-1} + \varphi(\sigma(j-1)) + \frac{\|X_{\sigma(j-1)} - X_i\|}{2K}, V_{j-1} + \varphi(\sigma(j-1)) + \|X_{\sigma(j-1)} - X_i\|\big]\Big) \\
&= \sum_{j \in \sigma^{-1}(i)} \Big(\varphi(i) + \frac{\|X_{\sigma(j+1)} - X_i\|}{2K} + \frac{\|X_{\sigma(j-1)} - X_i\|}{2K}\Big) \\
&= \frac{1}{n}.
\end{aligned}$$

This shows that $\lambda_1(\widehat{G}_K^{\star -1}(\mathrm{Vor}(i))) = \frac{1}{n}$, $i \in \{1, \dots, n\}$—or, said differently, that the function $\widehat{G}_K^\star$ spends a total time $1/n$ in each Voronoi cell. Now, introduce $T^\star : \mathbb{R}^d \to \{X_1, \dots, X_n\}$ defined $\widehat{G}_{K\sharp U}$-almost everywhere by $T^\star(x) = X_i$ if $x \in \mathrm{Vor}(i)$. Then, clearly, $T^\star \in \mathscr{H}^0$, where we recall that

$$\mathscr{H}^0 = \big\{T : \mathbb{R}^d \to \{X_1, \dots, X_n\} \ : \ \forall x \in \mathrm{Vor}(i), T(x) = X_i \\ \text{and } \forall x \in \Gamma^0_{j_1 \dots j_p}, T(x) \in \{X_{j_1}, \dots, X_{j_p}\}\big\}.$$

Arguing as in the proof of Lemma 14, one shows that there exists a sequence of functions $(G_m^\star)_{m\in\mathbb{N}} \subset \mathrm{Lip}_K([0,1], \mathbb{R}^d)$ such that each $G^\star_{m\sharp U}$ is nonatomic, $W_1(G^\star_{m\sharp U}, \mu_n) \to W_1(\widehat{G}^\star_{K\sharp U}, \mu_n)$ as $m \to \infty$, and, for all m large enough, $\lambda_1(G_m^{\star -1}(\mathrm{Vor}(i))) = \frac{1}{n}$, $i \in \{1, \dots, n\}$. According to Proposition 8, we have

$$W_1(G^\star_{m\sharp U}, \mu_n) = \int_0^1 \|G_m^\star(u) - T^\star(G_m^\star(u))\| \mathrm{d}u.$$

By dominated convergence, we obtain $W_1(\widehat{G}^\star_{K\sharp U}, \mu_n) = \int_0^1 \|\widehat{G}_K^\star(u) - T^\star(\widehat{G}_K^\star(u))\| \mathrm{d}u$, so that $T^\star$ is an optimal transport map from $\widehat{G}_K^\star$ to μ_n. Finally,

$$\begin{aligned}
W_1(\widehat{G}^\star_{K\sharp U}, \mu_n) &= \int_0^1 \|\widehat{G}_K^\star(u) - T^\star(\widehat{G}_K^\star(u))\| \mathrm{d}u \\
&= \sum_{j=1}^{n+k-1} \int_{V_j}^{V_j + \varphi(\sigma(j)) + \frac{\|X_{\sigma(j+1)} - X_{\sigma(j)}\|}{2K}} \|X_{\sigma(j)} - \widehat{G}_K^\star(u)\| \mathrm{d}u \\
&\quad + \int_{V_j + \varphi(\sigma(j)) + \frac{\|X_{\sigma(j+1)} - X_{\sigma(j)}\|}{2K}}^{V_j + \varphi(\sigma(j)) + \|X_{\sigma(j+1)} - X_{\sigma(j)}\|} \|X_{\sigma(j+1)} - \widehat{G}_K^\star(u)\| \mathrm{d}u \\
&= \sum_{j=1}^{n+k-1} \int_{V_j + \varphi(\sigma(j))}^{V_j + \varphi(\sigma(j)) + \frac{\|X_{\sigma(j+1)} - X_{\sigma(j)}\|}{2K}} K\big(u - (V_j + \varphi(\sigma(j)))\big) \mathrm{d}u \\
&\quad + \int_{V_j + \varphi(\sigma(j)) + \frac{\|X_{\sigma(j+1)} - X_{\sigma(j)}\|}{2K}}^{V_j + \varphi(\sigma(j)) + \|X_{\sigma(j+1)} - X_{\sigma(j)}\|} K(V_j + \varphi(\sigma(j)) + \|X_{\sigma(j+1)} - X_{\sigma(j)}\| - u) \mathrm{d}u \\
&= \sum_{j=1}^{n+k-1} \frac{1}{8K} \|X_{\sigma(j+1)} - X_{\sigma(j)}\|^2 + \frac{1}{8K} \|X_{\sigma(j+1)} - X_{\sigma(j)}\|^2 \\
&= \frac{1}{4K} \sum_{j=1}^{n+k-1} \|X_{\sigma(j+1)} - X_{\sigma(j)}\|^2.
\end{aligned}$$

Appendix J. Proof of Proposition 13

First note, since σ is a path with points that may be visited several times, that

$$\begin{aligned}\underline{K}_2 &\geqslant \sum_{i=1}^{n} \sum_{j\in\sigma^{-1}(i)} \frac{1}{2}(\|X_{\sigma(j-1)} - X_i\| + \|X_{\sigma(j+1)} - X_i\|) \\ &\geqslant \inf_{\tau\in\mathscr{P}_n} \sum_{j=1}^{n-1} \|X_{\tau(j)} - X_{\tau(j+1)}\|, \end{aligned} \tag{24}$$

where $\mathscr{P}_n$ stands for the set of permutations of $\{1,\dots,n\}$. But, according to Steele (1988), under the conditions of the theorem, there exists a constant $C > 0$ satisfying

$$\lim_{n\to\infty} n^{-1+1/d} \inf_{\tau\in\mathscr{P}_n} \sum_{j=1}^{n-1} \|X_{\tau(j)} - X_{\tau(j+1)}\| = C \text{ a.s.}$$

This shows the first statement of the proposition.

We start the proof of the second statement by recalling that, according to Fournier and Guillin (2015, Theorem 1), one has, in probability,

$$W_1(\mu,\mu_n) = \begin{cases} \mathscr{O}(\frac{\log n}{\sqrt{n}}) & \text{for } d = 2 \\ \mathscr{O}(n^{-1/d}) & \text{for } d \geqslant 3. \end{cases}$$

Therefore, by the triangle inequality, it is enough to show that, for $d \geqslant 2$, in probability,

$$W_1(\widehat{G}^{\star}_{K\sharp U},\mu_n) = \mathscr{O}(n^{-1/d}).$$

According to Theorem 12, we only need to show that, in probability,

$$\frac{1}{4K} \sum_{j=1}^{n+k-1} \|X_{\sigma(j+1)} - X_{\sigma(j)}\|^2 = \mathscr{O}(n^{-1/d}),$$

whenever $K \geqslant \underline{K}_2$. But, by the very definition (12) (Main Document) of the pair (k,σ), we have

$$\sum_{j=1}^{n+k-1} \|X_{\sigma(j+1)} - X_{\sigma(j)}\|^2 \leqslant \sum_{j=1}^{n-1} \|X_{\tau(j+1)} - X_{\tau(j)}\|^2,$$

where $\tau \in \mathscr{P}_n$ is a permutation that minimizes the length among the whole set of paths that visit only once each data, i.e.,

$$\sum_{j=1}^{n-1} \|X_{\tau(j+1)} - X_{\tau(j)}\| \leqslant \sum_{j=1}^{n-1} \|X_{\tau'(j+1)} - X_{\tau'(j)}\|, \text{ for all } \tau' \in \mathscr{P}_n.$$

Therefore, since $K \geqslant \underline{K}_2$, we have by inequality (24) ,

$$\frac{1}{K} \sum_{j=1}^{n+k-1} \|X_{\sigma(j+1)} - X_{\sigma(j)}\|^2 \leqslant \frac{\sum_{j=1}^{n-1} \|X_{\tau(j+1)} - X_{\tau(j)}\|^2}{\sum_{j=1}^{n-1} \|X_{\tau(j+1)} - X_{\tau(j)}\|}.$$

Now, under the additional condition on the density of μ, we know by Yukich (2000, Theorem 1.3) that, for each $0 \leqslant \ell \leqslant d$, there exists $C(\ell) > 0$ such that

$$\lim_{n\to\infty} n^{-1+\ell/d} \sum_{j=1}^{n-1} \|X_{\tau(j+1)} - X_{\tau(j)}\|^{\ell} = C(\ell) \text{ a.s.}$$

By the above, we conclude that

$$\frac{1}{4K} \sum_{j=1}^{n+k-1} \|X_{\sigma(j+1)} - X_{\sigma(j)}\|^2 = \mathscr{O}(n^{-1/d}) \text{ a.s.}$$