On characterizing optimal Wasserstein GAN solutions for non-Gaussian data

Yu-Jui Huang†, Shih-Chun Lin*, Yu-Chih Huang§, Kuan-Hui Lyu*, Hsin-Hua Shen*, and Wan-Yi Lin ‡
† Univ. of Colorado, Dept. of Applied Math., Boulder, CO 80309, USA, yujui.huang@colorado.edu
* National Taiwan University, Department of EE and GICE, Taipei, Taiwan, sclin2@ntu.edu.tw
§National Yang Ming Chiao Tung University, Institute of CM, HsinChu, Taiwan, jerryhuang@nctu.edu.tw
‡ Bosch Center for AI, USA, wan-yi.lin@us.bosch.com

2023 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-7554-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISIT54713.2023.10206785

Abstract—The generative adversarial network (GAN) aims to approximate an unknown distribution via a parameterized neural network (NN). While GANs have been widely applied in reinforcement and semi-supervised learning as well as computer vision tasks, selecting their parameters often needs an exhaustive search and only a few selection methods can be proved to be theoretically optimal. One of the most promising GAN variants is the Wasserstein GAN (WGAN). Prior work on optimal parameters for WGAN is limited to the linear-quadratic-Gaussian (LQG) setting, where the NN is linear and the data is Gaussian. In this paper, we focus on the characterization of optimal WGAN parameters beyond the LQG setting. We derive closed-form optimal parameters for one-dimensional WGANs with non-linear sigmoid and ReLU activation functions. Extensions to high-dimensional WGANs are also discussed. Empirical studies show that our closed-form WGAN parameters have good convergence behavior with data under both Gaussian and Laplace distributions.

I. Introduction

Generative adversarial networks (GANs) are a new class of machine learning frameworks put forth by Goodfellow *et al.* [1]. A GAN aims to learn an unknown distribution from training data via two competing components, namely the generator and the discriminator. The former tries to mimic the distribution of training data while the latter discriminates between true data and generated data. Besides computer vision tasks, applications of GANs to communication systems have also received a lot of attentions. For example, GANs have been applied to autonomous wireless channel modeling [2] [3] and covert communication [4].

Traditionally, both the generator and discriminator in GAN are approximated by neural networks (NNs) [1], [5]. By removing NN restrictions and under an optimal unconstrained discriminator, the minimax problem associated with a GAN becomes a minimization of the Jensen-Shannon divergence (JSD) between the distributions of true and generated data. However, due to the nature of this minimax game, the GAN suffers from several problems including vanishing gradient and mode collapse. Many variants of GAN have been proposed to solve these problems, and one of the most promising variants is the Wasserstein GAN (WGAN) [6] that replaces the JSD by the Wasserstein distance widely adopted in the optimal transport problem [7]. The WGAN is differentiable with respect to the generator parameters almost everywhere, which benefits the convergence of stochastic gradient descent (SGD) usually adopted for training NNs [6].

Despite the many successes of applying WGAN to learn distributions in real applications, there are only a few GAN parameter selection algorithms proved to be theoretically optimal [8], [9], which limits the development of GAN beyond heuristic methods in [1], [5]. This lack of rigorous analysis also restricts the evaluation of GANs' performance to subjective terms. One exception is [8] where Feizi *et al.* attempted to theoretically understand WGANs on a simple linear quadratic Gaussian (LQG) setting. In this benchmark setting, the synthetic data is generated by a Gaussian distribution, the generator NN is restricted to be linear, and the loss function is quadratic. It is shown in [8, Theorem 1] that for this simple setting, the optimal GAN solution happens to be the principal component analysis (PCA) solution. Regularized versions of GANs are also well-adopted [5]. Optimal WGAN solutions under LQG settings, with additional entropic and Sinkhorn regularizers, are also studied [9].

In this paper, we aim to analytically solve WGANs beyond the LQG setting. As described in Sec. II, our setting allows non-Gaussian data distribution and non-linear generators including sigmoid and ReLU, and is more general than [8]–[10]. Also, we attempt to exactly solve WGANs rather than their regularized versions as [9]. All of these make our problem exceedingly challenging since even for the inner discriminator problem, which is an optimal transport problem, the solution in most cases is numerically approximated but not analytically characterized [7]. To overcome the challenge, we first focus on one-dimensional data and generator in Sec. III-A and III-B, where we provide our closed-form solutions for optimal generators in one-dimensional WGANs defined in Sec. II. Extensions to high-dimensional WGANs for results with linear generators are given in Sec. III-C. The proofs are presented in Sec. IV, where we leverage result in [11] to solve the inner discriminator problem in closed-form which greatly simplifies the necessary conditions of optimal WGAN parameters. Moreover, our closed-form solutions do not need any training for the discriminator as [12] and hence provide additional benefit for training WGAN with a decentralized system [13]. Empirical studies in Sec. V show that our closed-form WGAN parameters have good convergence behavior with synthetic data under both Gaussian and Laplace distributions.

This work was supported in part by NSF under Grant DMS-2109002; in part by the National Science and Technology Council, Taiwan, under Grant 111-2221-E-002-099-MY2, 111-3114-E-002-001 and MOST 111-2221-E-A49 -069 -MY3.

II. Problem Formulation

For the WGAN considered in this paper, the overall transfer function of the generator NN is denoted as $G_\vartheta(.)$, where ϑ is the generator NN parameter (weights). In this paper, we consider the following popular activation functions as examples: 1) the linear function; 2) the rectified linear unit (ReLU) function $\max(0; z)$; and 3) the sigmoid function $1/(1 + \exp(-z))$. Let $q \in \{1, 2\}$ represent the order of the Wasserstein distance. In a q^{th}-order WGAN setting, one aims to solve

$$\min_{\vartheta} \inf_{\pi \in \Pi(\mu, \nu^\vartheta)} \mathsf{E}_\pi[\|X - G_\vartheta(Z)\|^q]^{1/q} \tag{1}$$

for an optimal parameter ϑ, where $\|\cdot\|$ denotes the ℓ_2-norm in $\mathbb{R}^d$, μ and ν^ϑ are probability measures on $\mathbb{R}^d$, generated by the data X and the generator output $G_\vartheta(Z)$ for a given ϑ and Gaussian input Z, respectively. Also, $\Pi(\mu, \nu^\vartheta)$ is the set of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ whose marginals on the first and second coordinates are μ and ν^ϑ, respectively, satisfying $\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^q d\mu(x) d\nu^\vartheta(y) < \infty$. The WGAN problem described in (1) can be equivalently written as $\min_\vartheta (\mathsf{E}_\mu[P(x, \vartheta)])^{1/q}$ where the inner-discriminator problem is defined as

$$\mathsf{E}_\mu[P(X, \vartheta)] := \inf_{\pi \in \Pi(\mu, \nu^\vartheta)} \mathsf{E}_\pi[\|X - Y\|^q]. \tag{2}$$

Note that (2) belongs to the family of the optimal transport problems with q^{th}-order Wasserstein distance, $q \in \{1, 2\}$ [7, Proposition 2.2].

Following [8], we call (1) the population GAN problem, where ϑ can be optimized with the true data distribution μ. In the next Sec. III, we first theoretically solve the population GAN problems whose solutions will depend on μ. Similarly to almost every work in the GAN literature, in practice, we use empirical data to get an estimate of the statistics we need in our solution, as detailed in Sec. V.

III. Main Results

With $d = 1$, we present our main results for $q = 2$ in Sec. III-A and that for $q = 1$ in Sec. III-B. Note that for applications in communication systems [2] [4], low-dimensional (even with $d = 1$) results can be very useful. For other applications, extensions to $d > 1$ are given in Sec. III-C.

A. Results for second-order WGAN under one dimension

First, we consider the quadratic case $q = 2$ with a non-linear generator

$$G_\vartheta(Z) = \vartheta_1 + \vartheta_2 h(Z), \tag{3}$$

where $h : \mathbb{R} \to \mathbb{R}$ and $Z \sim N(0, 1)$, also $(\vartheta_1, \vartheta_2) \in \mathbb{R} \times \mathbb{R}$ are parameters of the generator NN to be selected. Let Ψ denote the cumulative distribution function (CDF) of $h(Z)$, for any continuous data distribution μ, our closed-form WGAN parameters are as follows:

Theorem 1. *Assume CDF F_μ of μ and CDF Ψ of $h(Z)$ in* (3) *are continuous and strictly increasing, and variance* $\mathrm{Var}(h(Z)) > 0$. *If*

$$\mathrm{Cov}\left(X, \Psi^{-1}(F_\mu(X)) + \Psi^{-1}(1 - F_\mu(X))\right) \geq 0, \tag{4}$$

the population WGAN (1) *with $q = 2, d = 1$ has a unique minimizer for $(\vartheta_1, \vartheta_2) \in \mathbb{R} \times \mathbb{R}$ as*

$$\begin{aligned} \vartheta_2^* &= \frac{\mathrm{Cov}\left(X, \Psi^{-1}(F_\mu(X))\right)}{\mathrm{Var}(h(Z))} \geq 0, \\ \vartheta_1^* &= \mathsf{E}_\mu[X] - \vartheta_2^* \mathsf{E}_g[h(Z)]; \end{aligned} \tag{5}$$

if (4) *is not met, $(\vartheta_1^*, \vartheta_2^*)$ is given by replacing ϑ_2^* in* (5) *by*

$$\vartheta_2^* = \frac{\mathrm{Cov}\left(X, \Psi^{-1}(1 - F_\mu(X))\right)}{\mathrm{Var}(h(Z))} \leq 0, \tag{6}$$

where E_g is taking expectation over Gaussian $Z \sim N(0, 1)$.

Proof: To solve the inner discriminator problem (2), we break (1) down into two sub-problems depending on the sign of ϑ_2, i.e., $\min_{\vartheta_1, \vartheta_2 \in \mathbb{R}} \mathsf{E}_\mu[P(X, \vartheta_1, \vartheta_2)]$ equals to

$$\min\left\{\min_{\vartheta_1 \in \mathbb{R}, \vartheta_2 \geq 0} \mathsf{E}_\mu[P(X, \vartheta_1, \vartheta_2)], \min_{\vartheta_1 \in \mathbb{R}, \vartheta_2 \leq 0} \mathsf{E}_\mu[P(X, \vartheta_1, \vartheta_2)]\right\} \tag{7}$$

The solution of the first sub-problem is indeed (5) while that for second sub-problem is (6). The condition (4) is obtained by comparing the values of the two subproblems. The proof of the first sub-problem, where $\vartheta_2 \geq 0$, is given in Sec. IV-A, while the other proofs are omitted due to space limit. ∎

Let us now look at some specific cases of $h(z)$. For the sigmoid function $h(z)$, recall that the logit function $\mathrm{logit}(p) := \ln(p/(1 - p)), p \in (0, 1)$ which is the inverse function $h^{-1}(z)$. The random variable $h(Z)$ has a logit-normal distribution, i.e. $\mathrm{logit}(h(Z))$ is normally distributed, with CDF

$$\Psi(v) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{\mathrm{logit}(v)}{\sqrt{2}}\right)\right) \quad \text{for } v \in (0, 1).$$

For the ReLU function, the CDF of $h(Z) = \max\{Z, 0\}$ is given by

$$\Psi(v) = \Phi(v) \cdot 1_{\{v \geq 0\}}, \tag{8}$$

where Φ is the CDF of Gaussian $N(0, 1)$. However, now Ψ has a jump from 0 to $1/2$ at $v = 0$ and does not meet the setting of Theorem 1 since it is neither continuous nor strictly increasing. We need the following modification.

Theorem 2. *Assume μ has the setting as in Theorem 1 and Ψ is given by* (8). *If*

$$\begin{aligned} &\mathrm{Cov}\left(X, \Phi^{-1}(F_\mu(X)) 1_{\{F_\mu(X) > 1/2\}}\right) \\ &\geq \mathrm{Cov}\left(X, \Phi^{-1}(F_\mu(X)) 1_{\{F_\mu(X) \leq 1/2\}}\right), \end{aligned} \tag{9}$$

the population WGAN (1) *with $q = 2, d = 1$ has a unique minimizer for $(\vartheta^1, \vartheta^2) \in \mathbb{R} \times \mathbb{R}$ as*

$$\begin{aligned} \vartheta_2^* &= \frac{2\pi}{\pi - 1} \mathrm{Cov}\left(X, \Phi^{-1}(F_\mu(X)) 1_{\{F_\mu(X) > 1/2\}}\right) \\ \vartheta_1^* &= \mathsf{E}[X] - \vartheta_2^*/\sqrt{2\pi}; \end{aligned} \tag{10}$$

if (9) *is not met, $(\vartheta_1^*, \vartheta_2^*)$ is given by replacing ϑ_2^* in* (10) *by*

$$\vartheta_2^* = \frac{-2\pi}{\pi - 1} \mathrm{Cov}\left(X, \Phi^{-1}(F_\mu(X)) 1_{\{F_\mu(X) \leq 1/2\}}\right). \tag{11}$$

For the linear generator $h(Z) = Z$ as [5, eqn. (27)] [8] [9], one can simplify the results in Theorem 1 and also get alternative closed-form formula as follows

Corollary 1. *Assume μ has the setting as in Theorem 1 and consider the linear case $h(Z) = Z$ in* (3)*. Then population WGAN* (1) *has a unique minimizer for $(\vartheta^1, \vartheta^2) \in \mathbb{R} \times \mathbb{R}$ as*

$$\vartheta_1^* = \mathbb{E}_\mu[X] \quad \text{and} \quad \vartheta_2^* = \mathbb{E}_\mu[X \cdot \Phi^{-1}(F_\mu(X))], \tag{12}$$

also equivalently

$$\vartheta_2^* = \mathbb{E}_g[F_\mu^{-1}(\Phi(Z)) \cdot Z], \tag{13}$$

Proof: Besides checking (4), it is easy to see that one can limit $\vartheta_2 \in \mathbb{R}_+$ such that the optimal parameter is (5) when $h(Z) = Z$. Now the distribution of $-h(Z)$ is still the same as $h(Z)$. Then one can rewrite (3) as $G_\vartheta(Z) = \vartheta_1 + (-\vartheta_2)\cdot(-Z)$, and absorb the case for $\vartheta_2 < 0$ into that for $\vartheta_2 \geq 0$. The rest of proof is omitted. ∎

Here we briefly compare our proposed WGAN solutions with the results for the LQG setting in [8, Theorem 1]. In the LQG setting, the d-dimensional synthetic data is generated by Gaussian distribution, i.e., $X \sim \mathcal{N}(0, \mathsf{K})$, the generator is restricted to be a linear generator of the form

$$\Theta Z, \quad \Theta \in \mathbb{R}^{d \times r}, \tag{14}$$

with $Z \sim \mathcal{N}(0, \mathsf{I}_r)$ a Gaussian vector, and the loss function is quadratic (i.e., second-order WGAN). Since WGAN output is already Gaussian $\Theta Z \sim \mathcal{N}(0, \mathsf{K})$, it is shown in [8, Theorem 1] that for this benchmark, the optimal WGAN solution happens to be the r-PCA solution of Gaussian X. That is, optimal generator matrix Θ fulfills that $\mathsf{K} = \Theta\Theta^T$ is a rank r matrix and K and K share the same largest r eigenvalues and the corresponding eigenvectors. Unlike [8], our results can deal with non-Gaussian data distribution. Moreover, we can recover the result in [8] when $q = 2, d = 1$. Indeed, if $X \sim \mathcal{N}(0, \sigma^2)$, by looking into (12), we have $\Phi^{-1}(F_\mu(X)) = \Phi^{-1}(\Phi(X/\sigma)) = X/\sigma$. Plugging this into our solution in (12) shows that $\vartheta_1^* = 0$ and $\vartheta_2^* = \mathbb{E}[X\frac{X}{\sigma}] = \sigma$, coinciding with the result in [8] for $d = r = 1$. We emphasize that neither of our and the results in [8] subsume the other as a special case as [8] considers general dimension d while our work is not restricted to Gaussian data distribution.

B. Results for first-order WGAN under one dimension

Here, we present our result for non-linear generators and first-order Wasserstein distance. We have

Corollary 2. *Following the settings in Theorem 1, the minimizer $(\vartheta_1^*, \vartheta_2^*)$ of the population WGAN* (1) *with $q = 1, d = 1$ meet the following necessary conditions*

$$\mathbb{E}_\mu\left[\text{sign}\left(\vartheta_1^* + \vartheta_2^*\Psi^{-1}(F_\mu(X)) - X\right)\right] = 0,$$

$$\mathbb{E}_\mu\left[\text{sign}\left(\vartheta_1^* + \vartheta_2^*\Psi^{-1}(F_\mu(X)) - X\right)\Psi^{-1}(F_\mu(X))\right] = 0, \tag{15}$$

when $\vartheta_2^* > 0$, where $\text{sign}(x) = 1, 0, -1$ for $x > 0$, $x = 0$, $x < 0$, respectively.

The proof is omitted. Note that when $X \sim \mathcal{N}(\mu, \sigma^2)$ then $F_\mu(X) = \Phi((X-\mu)/\sigma)$, It can be checked that for the linear case as Corollary 1, the optimal $(\vartheta_1^*, \vartheta_2^*) = (\mu, \sigma)$ for $q = 2$ also meets (15) and is at least a local optimum for $q = 1$.

C. From one-dimension to multi-dimension

Previously we focused on providing a closed-form solution for the population WGAN with $d = 1$. However, for some applications high-dimensional data are preferred. Here, we show how to generalize our results to cope with higher dimensional data by adopting the sliced Wasserstein distance technique [14] [15]. Let $\Omega = \{\omega \in \mathbb{R}^d : \omega = 1\}$ contain all the directions in $\mathbb{R}^d$. In sliced Wasserstein distance, we project both the data and the generator's output onto a random direction $\omega \in \Omega$ with uniform distribution and compute the Wasserstein distance with respect to the projected 1-dimensional distributions μ_ω and ν_ω^Θ, i.e., $\omega^T X$ and $\omega^T \Theta Z$ respectively follow distributions μ_ω and ν_ω^Θ. By replacing the distance of inner-discriminator problem (2) with sliced Wasserstein distance, the q^{th}-order sliced (population) WGAN with a linear generator (14) is

$$\min_\Theta \int_{\omega\in\Omega} \inf_{\pi\in\Pi(\mu_\omega, \nu_\omega^\Theta)} \mathbb{E}_\pi[|\omega^T X - \omega^T \Theta Z|^q] d\omega^{\frac{1}{q}}. \tag{16}$$

Motivated by [16] we replace the uniformly distributed $\omega \in \Omega$ with Gaussian projections $\omega \in \Omega_G$, that is, ω is generated according to a zero-mean Gaussian distribution with covariance matrix $(1/d)\mathsf{I}_d$ and aim at solving

$$\min_\Theta \int_{\omega\in\Omega_G} \inf_{\pi\in\Pi(\mu_\omega, \nu_\omega^\Theta)} \mathbb{E}_\pi[|\omega^T X - \omega^T \Theta Z|^q] d\omega^{\frac{1}{q}}, \tag{17}$$

where we use notation $\int_{\omega\in\Omega_G} a(\omega) d\omega := \int_{\omega\in\mathbb{R}^d} a(\omega) dF(\omega)$ to represent the expectation of function $a(\omega)$ over PDF $dF(\omega)$ of Gaussian random vectors. Using [17, Proposition 1], it can be easily shown that the optimal Θs for (16) and (17) are identical. Moreover, when $q = 2$, not only the optimizers but also the values in (16) and (17) become the same.

With proof sketch and comparisons with [16] [17] given in Sec IV-B, our main result for (17) with $d > 1$ is:

Theorem 3. *Let $X_G \sim \mathcal{N}(0, \tilde{\sigma}_x^2)$ independent of Gaussian vector ω with finite $\tilde{\sigma}_x^2 = \mathbb{E}_\mu[X^2]/d$. The gap of* (17) *to*

$$\min_\Theta \int_{\omega\in\Omega_G} \inf_\pi \mathbb{E}_\pi[|X_G - \omega^T \Theta Z|^q] d\omega^{1/q} \tag{18}$$

is bounded by a function $f^\mu(d) = O(d^{-1/8})$, where the marginals of π are $\mathcal{N}(0, \tilde{\sigma}_x^2)$ and ν_ω^Θ; and the optimal Θ for (18) *equals to that of*

$$\min_\Theta \mathbb{E}_\omega[|\tilde{\sigma}_x - \sqrt{\omega^T \Theta\Theta^T \omega}|^q], q = 1, 2. \tag{19}$$

Moreover, the optimal Θ when $q = 2$ is the minimizer to

$$\frac{\text{Tr}(\Theta\Theta^T)}{d} + \frac{2\tilde{\sigma}_x}{\Gamma(1/2)} \int_0^\infty z^{-1/2} \frac{\partial}{\partial z}\left|\mathsf{I} + \frac{2z}{d}\Theta\Theta^T\right|^{-1/2} dz; \tag{20}$$

where $\text{Tr}(.)$ is the matrix trace and $\Gamma(t)$ is the gamma function.

The derivative with respect to z in (20) for $q = 2$ can be easily solved by matrix calculus [18] and the integration is over one dimension z. However, when $q = 1$, the d–dimensional integration in (19) is hard to be simplified as in (20). One can also modify the proof of Theorem 3 by adding additional approximation gap to (17) on top of $f^\mu(d)$ (but still $O(d^{-1/8})$),

and obtain a closed-form $\tilde{\sigma}_x^2 I = \Theta\Theta^T$ criteria for selecting Θ for (19) under $q = 1$.

The reason that the gap between (17) and (18) in Theorem 3 is vanishing is because for $\omega \in \Omega$, most $\omega^T X$ can be well approximated as X_G [19]. However, there exist outliers whose distributions are far from X_G. An alternative approximation could instead be restricting ω in $\Omega_e = \{e_1, \ldots, e_d\}$ where e_i is the i-th standard unit vector with the i-th entry being 1 and 0 otherwise. We have

Corollary 3. *Restrict $\omega \in \Omega_e$,* (16) *with $q = 2$ simplifies* [1] *to*

$$\min_{\Theta} \sum_{i=1}^{d} \inf_{\pi \in \Pi(\mu_{e_i}, \nu_{e_i}^{\vartheta})} \mathbb{E}[|e_i^T X - e_i^T \Theta Z|^2], \tag{21}$$

and the i,jth elements ϑ_{ij} of optimal Θ meet

$$\sum_{j=1}^{r} (\vartheta_{ij})^2 = \mathbb{E}_\mu[e_i^T X \cdot \Phi^{-1}(F_\mu(e_i^T X))]^2. \tag{22}$$

The generator in (22) depends on more statistical information of data than (20). In practice, one can regulate the two objectives (18) and (21) to balance contributions from outliers, which is left for future work.

IV. The Proofs

A. Proof sketch for Theorem 1

In the following, we focus on solving the first sub-problem with $\vartheta_2 \in \mathbb{R}_+$ in (7). With $d = 1$, we recall the following result for the inner discriminator problem (2) of (1), by taking $\nu^{\vartheta} := P_{G_\vartheta(z)}$, the measure generated by the generator NN (with parameter $\vartheta := (\vartheta^1, \vartheta_2)$ in (3)). Let F_μ and F_{ν^ϑ} denote the cumulative distribution functions (CDFs) of the measures μ and ν^{ϑ} on $\mathbb{R}$.

Lemma 1 ([11], Theorem 5.1). *Define $t^{\vartheta} : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ by*

$$t^{\vartheta}(x) := \sup\{y \in \mathbb{R} : F_{\nu^\vartheta}(y) \le F_\mu(x)\}. \tag{23}$$

For $q = 1, 2$, if μ has no atom (μ is a continuous distribution)

$$\inf_{\pi \in \Pi(\mu, \nu^{\vartheta})} \int_{\mathbb{R}\times\mathbb{R}} |x - y|^q d\pi(x, y) = \int_{\mathbb{R}} |x - t^{\vartheta}(x)|^q d\mu(x), \tag{24}$$

For the inner discriminator problem (2) with $q = 2, d = 1$, $\mathbb{E}_\mu[P(X, \vartheta^1, \vartheta^2)]$ for given $(\mu, \vartheta_1, \vartheta^2)$ equals to (24), where $t^{\vartheta}(x)$ is defined as in (23). Now we need to find a closed-form $t^{\vartheta}(x)$ to continue. From (3), let Ψ denote the CDF of $h(Z)$. If Ψ is continuous and strictly increasing and $\vartheta_2 \in \mathbb{R}_+$, (23) can be expressed in closed-form as

$$t^{\vartheta}(x) = \vartheta_1 + \vartheta_2 \Psi^{-1}(F_\mu(x)). \tag{25}$$

To see this, since $\vartheta_2 > 0$, observe that

$$\Psi\left(\frac{t^{\vartheta}(x) - \vartheta_1}{\vartheta_2}\right) = \mathbb{P}\left(h(Z) \le \frac{t^{\vartheta}(x) - \vartheta_1}{\vartheta_2}\right)$$
$$= \mathbb{P}\left(G_\vartheta(Z) \le t^{\vartheta}(x)\right) = F_{\nu^\vartheta}(t^{\vartheta}(x)) \overset{(a)}{=} F_\mu(x),$$

where the last equality follows from (23) and that F_{ν^ϑ} is continuous and strictly increasing; here, F_{ν^ϑ} inherits the same properties from Ψ thanks to (3). Though the continuity of Ψ is not needed in Lemma 1 and its Kantorovich equivalence [21, Theorem 2.18], without it equality (a) will become an inequality which harms finding closed-form ϑ. Then it follows that $\frac{t^{\vartheta}(x) - \vartheta_1}{\vartheta_2} = \Psi^{-1}(F_\mu(x))$, which yields (25). On the other hand, if $\vartheta_2 = 0$, since $G_\vartheta(Z) \equiv \vartheta_1 \in \mathbb{R}$, we have $F_{\nu^\vartheta}(y) = 1_{\{y \ge \vartheta_1\}}$. Plugging this into (23) directly gives $t^{\vartheta}(x) = \vartheta_1$ for all $x \in \mathbb{R}$. This particularly shows that (25) is also satisfied for the case $\vartheta_2 = 0$.

With closed-form representation (24)(25) for the inner problem (2), WGAN (1) with $q = 2, d = 1$ can be simplified to be the following stochastic minimization problem

$$\min_{\vartheta_1 \in \mathbb{R}, \vartheta_2 \in \mathbb{R}_+} \mathbb{E}_\mu\left[X - \vartheta_1 - \vartheta_2 \Psi^{-1}(F_\mu(X))\right]^2. \tag{26}$$

Together with (3), (26) becomes the constrained optimization problem

$$\min_{\vartheta_1, \vartheta_2 \in \mathbb{R}} J(\vartheta_1, \vartheta_2) := \int_{\mathbb{R}} \left[\vartheta_1 + \vartheta_2 \Psi^{-1}(F_\mu(x)) - x\right]^2 F_\mu(x) dx$$
$$\text{subject to} \quad g(\vartheta_1, \vartheta_2) := -\vartheta_2 \le 0. \tag{27}$$

The corresponding first-order Karush-Kuhn-Tucker (KKT) condition is

$$\nabla J(\vartheta^1, \vartheta_2) + \lambda \nabla g(\vartheta^1, \vartheta^2) = 0, \tag{28}$$
$$\lambda g(\vartheta^1, \vartheta^2) = 0, \tag{29}$$

where $\lambda \ge 0$ is the Lagrange multiplier. By direct calculation, (28) becomes

$$\int_{\mathbb{R}} \left[\vartheta_1 + \vartheta_2 \Psi^{-1}(F_\mu(x)) - x\right] F_\mu(x) dx = 0,$$
$$\int_{\mathbb{R}} \left[\vartheta_1 + \vartheta_2 \Psi^{-1}(F_\mu(x)) - x\right] \Psi^{-1}(F_\mu(x)) F_\mu(x) dx = \frac{\lambda}{2}$$

Recall X is a random variable with CDF F_μ, and the above equalities can be written as

$$\vartheta_1 + \vartheta_2 \mathbb{E}\,\Psi^{-1}(F_\mu(X)) = \mathbb{E}[X],$$
$$\vartheta_1 \mathbb{E}\,\Psi^{-1}(F_\mu(X)) + \vartheta_2 \mathbb{E}\,(\Psi^{-1}(F_\mu(X)))^2 - \mathbb{E}\,X\Psi^{-1}(F_\mu(X)) = \lambda/2.$$

Note that $F_\mu(X) \sim \text{Uniform}[0, 1]$ from [22, Lemma 1], so that the CDF of $\Psi^{-1}(F_\mu(X))$ is simply Ψ. In other words, $\Psi^{-1}(F_\mu(X))$ and $h(Z)$ have identical distribution. The formulas above thus simplify to

$$\vartheta_1 + \vartheta_2 \mathbb{E}[h(Z)] - \mathbb{E}[X] = 0,$$
$$\vartheta_1 \mathbb{E}[h(Z)] + \vartheta_2 \mathbb{E}(h(Z))^2 - \mathbb{E}\,X\Psi^{-1}(F_\mu(X)) = \frac{\lambda}{2}.$$

Plugging $\vartheta_1 = \mathbb{E}[X] - \vartheta_2 \mathbb{E}[h(Z)]$ from the first equality into the second one, we obtain

$$\mathbb{E}[X]\mathbb{E}[h(Z)] + \vartheta_2 \text{Var}(h(Z)) - \mathbb{E}\,X\Psi^{-1}(F_\mu(X)) - \lambda/2 = 0.$$

Hence, a solution $(\vartheta_1, \vartheta_2, \lambda)$ to (28) must equivalently satisfy

$$\vartheta_1 = \mathbb{E}[X] - \vartheta_2 \mathbb{E}[h(Z)],$$
$$\lambda/2 = \vartheta^2 \text{Var}(h(Z)) - \text{Cov}(X, \Psi^{-1}(F_\mu(X))). \tag{30}$$

[1] Another justification for (21) comes from [20] by setting the unknown copula of data μ same as that of the generator output ΘZ in (1). Note that Sklar's theorem ensures that there exists an unique copula function that injectively maps marginals to the joint distribution [20].

To solve the KKT condition (28)-(29) for candidate minimizers, we already know that (28) boils down to (30), while (29) simply implies either $\lambda = 0$ or $\vartheta_2 = 0$. Also, because Ψ^{-1} is strictly increasing and F_μ is nondecreasing, the map $x \to \Psi^{-1}(F_\mu(x))$ is nondecreasing. This readily implies

$$\mathrm{Cov}\ [X, \Psi^{-1}(F_\mu(X))] \geq 0 \tag{31}$$

in (30) since $(x - x)(\Psi^{-1}(F_\mu(x)) - \Psi^{-1}(F_\mu(x))) \geq 0$ for all $x, x \in \mathbb{R}$. Specifically, by taking an independent but same distribution copy X of X,

$$0 \leq \mathrm{E}\,(X - X)\,\Psi^{-1}(F_\mu(X)) - \Psi^{-1}(F_\mu(X))$$
$$= 2\,\mathrm{Cov}\ [X, \Psi^{-1}(F_\mu(X))]\ .$$

We separate the proof for solving (27) into two cases from (30)(31) and combing them yields (5). Details are omitted.

B. Proof sketch for Theorem 3

First we assume X is zero mean, since equivalently one can modify ΘZ in (17) with a bias $\Theta Z + \mathrm{E}[X]$ for non-zero mean X. Wasserstein distance (as RHS of (2)) of order 2 between $\omega^T X$ with $\hat{X}_G$, averaged over ω, is bounded from [16]. We modify this result to make it valid for $q = 1$, and then get desired bounded gap $f^\mu(d)$ between (17)(18) for $q = 1, 2$ as

$$f^\mu(d) := (C\,\mathrm{E}_\mu[X^2](d^{-5/4} + d^{-7/5}))^{1/2} \tag{32}$$

where C is a constant. With linear generator (14), not only $\omega^T \Theta Z$ is Gaussian as $\hat{X}_G$ given ω, on the contrary to [17, Theorem 1], but also we prevent gap $f^\mu(d)$ depending on Θ.

Next (19) is from modifying Corollary 1 and 2 to (18). Finally, solving (19) with $q = 2$ equals to solving

$$\tilde{\sigma}_x^2 - \max_{\Theta} \mathrm{E}_w[2\tilde{\sigma}_x \sqrt{\omega^T \Theta\Theta^T \omega} - \omega^T \Theta\Theta^T \omega]. \tag{33}$$

For the first term, for Gaussian $w \sim \Omega_G$, let its quadratic form $U_w := \omega^T \Theta\Theta^T \omega$ and $U_w > 0$ almost surely. Then, $\mathrm{E}_w[\sqrt{\omega^T \Theta\Theta^T \omega}] = \mathrm{E}_{U_w}[U_w U_w^{-1/2}]$ and we can prove that

$$\mathrm{E}_w[\sqrt{\omega^T \Theta\Theta^T \omega}] = \frac{-1}{\Gamma(1/2)} \int_0^\infty z^{-1/2} \frac{\partial}{\partial z} M_{U_w}(-z)\,dz \tag{34}$$

where $M_{U_w}(z) := \mathrm{E}[e^{zU_w}]$ is the moment generating function of U_w. For the second term it is easy to see that

$$\mathrm{E}_w[\omega^T \Theta\Theta^T \omega] = \mathrm{Tr}(\mathrm{E}_w[\omega\omega^T]\Theta\Theta^T) = \mathrm{Tr}(\Theta\Theta^T)/d. \tag{35}$$

Combining these results reach (20).

V. Empirical Study on Convergence with Synthetic Data

We have only considered population WGAN thus far. In practice, the distribution is estimated from training data and does not have a closed-form CDF. Therefore, it is of interest to empirically study how fast the solution converges to the population WGAN result using synthetic training data. In Fig. 1, we consider the linear activation and plot the optimal ϑ_2^* in (12)(13) obtained by solving population WGAN with $d = 1, q = 2$, and its estimate with synthetic data when μ is chosen to be (a) $N(0, 1)$ and (b) Laplace distribution [23] with mean 0 and scale $1/\sqrt{2}$ (which also has a unit variance).

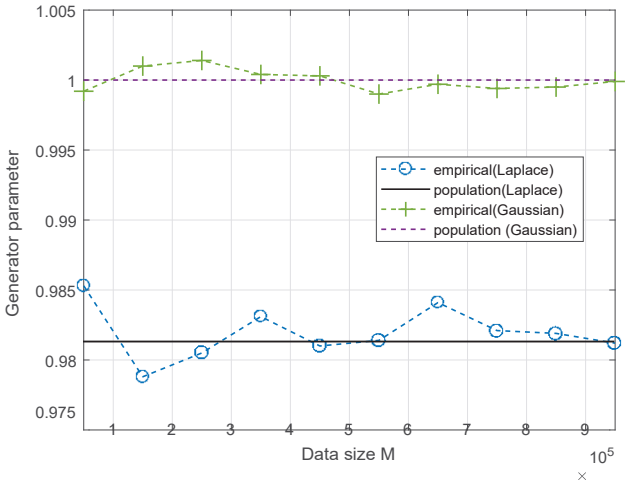

Fig. 1. Comparison of optimal parameters in (12)(13) with their estimates (36) using synthetic data.

Specifically, by generating training data $\{x_i\}_{i=1}^M$ according to the true distribution μ, we empirically estimate ϑ_2^* from (12) by

$$\frac{1}{M} \sum_{i=1}^M x_i \Phi^{-1}(F_\mu(x_i)) \tag{36}$$

and the kernel density estimation (KDE) [24] is used to replace the true CDF $F_\mu(x)$ with the estimated one $\hat{F}_\mu(x)$ from the training data. For each distribution, our result shows a nice convergence behavior where the difference becomes smaller than 0.005 with only M=50000 data size. The optimal ϑ_2^* in (12)(13) is 1 and 0.98013 for Gaussian and Laplace distributions respectively. Using linear generator, GAN output is also Gaussian and thus the convergence behavior for Gaussian data is better than that for Laplace data. Note that ϑ_1^* in (12) can be estimated by the sample mean, so the convergence is not shown in Fig. 1. We also use SGD with momentum to estimate ϑ_1 and ϑ_2 for loss function (1) under $d = 1, q = 1$, with gradient empirically obtained from (15) using KDE $\hat{F}_\mu(x)$. With $N(\mu = 1.5, \sigma^2 = 4)$ data, $(\vartheta^1, \vartheta^2)$ converges to $(1.4957, 2.0905)$ and close to the population local optimum (μ, σ) for $q = 1$.

The convergence rate of the empirical WGAN solution to that of the population WGAN problem has been theoretically analyzed to be $M^{-2/d}$ for the LQG setting in [8, Theorem 2]. Note that the convergence of iteratively solving empirical WGAN based on a good solver for (2) was given in [12]. However, this convergence heavily relies on a good semi-discrete optimal transport solver, which is still hard to design [25]. It is our future work to analyze the convergence rate for our non-Gaussian results, also compare it with the empirical simulations for $d > 1$.

Acknowledgement

We would like to acknowledge the help of Wen-Yi Tseng on the simulations.

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[2] Y. Yang, Y. Li, W. Zhang, F. Qin, P. Zhu, and C.-X. Wang, "Generative-adversarial-network-based wireless channel modeling: Challenges and opportunities," *IEEE Communications Magazine*, vol. 57, no. 3, pp. 22–27, 2019.

[3] S.-C. Lin, T.-H. Chang, E. A. Jorswieck, and P.-H. Lin, *Information theory, mathematical optimization, and their crossroads in 6G system design*, 1st ed. Springer Series in Wireless Technology, 2023.

[4] X. Liao, J. Si, J. Shi, Z. Li, and H. Ding, "Generative adversarial network assisted power allocation for cooperative cognitive covert communication system," *IEEE Communications Letters*, vol. 24, no. 7, pp. 1463–1467, 2020.

[5] M. Sanjabi, J. Ba, M. Razaviyayn, and J. D. Lee, "On the convergence and robustness of training GANs with regularized optimal transport," in *Advances in Neural Information Processing Systems*, 2018, pp. 7091–7101.

[6] S. C. M. Arjovsky and L. Bottou, "Wasserstein GAN," *arXiv preprint arXiv:1701.07875*, 2017.

[7] G. Peyré, M. Cuturi, *et al.*, "Computational optimal transport: With applications to data science," *Foundations and Trends® in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.

[8] S. Feizi, F. Farnia, T. Ginart, and D. Tse, "Understanding GANs in the LQG setting: Formulation, generalization and stability," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 304–311, 2020.

[9] D. Reshetova, Y. Bai, X. Wu, and A. Ozgur, "Understanding entropic regularization in GANs," in *IEEE International Symposium on Information Theory*, Juy 2021.

[10] B. Bailey and M. J. Telgarsky, "Size-noise tradeoffs in generative networks," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[11] L. Ambrosio and A. Pratelli, "Existence and stability results in the l1 theory of optimal transportation," *Optimal Transportation and Applications. Lecture Notes in Mathematics*, vol. 1813, 2003.

[12] Y. Chen, M. Telgarsky, C. Zhang, B. Bailey, D. Hsu, and J. Peng, "A gradual, semi-discrete approach to generative network training via explicit Wasserstein minimization," in *International Conference on Machine Learning*. PMLR, 2019, pp. 1071–1080.

[13] B. McMahan and D. Ramage, "Federated learning: Collaborative machine learning without centralized training data," *Google Research Blog*, vol. 3, 2017.

[14] I. Deshpande, Z. Zhang, and A. Schwing, "A. generative modeling using the sliced wasserstein distance," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3483–3491.

[15] S. Kolouri, K. Nadjahi, U. Simsekli, R. Badeau, and G. K. Rohde, "Generalized sliced Wasserstein distances," in *33rd Conference on Nerual Information Processing Systems*, 2019.

[16] G. Reeves, "Conditional central limit theorems for Gaussian projections," in *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017, pp. 3045–3049.

[17] K. Nadjahi, A. Durmus, P. E. Jacob, R. Badeau, and U. Simsekli, "Fast approximation of the sliced-wasserstein distance using concentration of random projections," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 12 411–12 424.

[18] A. Hjørungnes, *Complex-valued matrix derivatives: with applications in signal processing and communications*. Cambridge University Press, 2011.

[19] E. Meckes, "Approximation of projections of random vectors," *Journal of Theoretical Probability*, vol. 25, no. 2, pp. 333–352, 2012.

[20] A. Alfonsi and B. Jourdain, "A remark on the optimal transport between two probability measures sharing the same copula," *Statistics & Probability Letters*, vol. 84, pp. 131–134, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167715213003337

[21] C. Villani, *Topics in optimal transportation*. American Mathematical Soc., 2003, vol. 58.

[22] O. Shayevitz and M. Feder, "Optimal feedback communication via posterior matching," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1186–1222, 2011.

[23] H. Bauschke, C. Hamilton, M. Macklem, J. McMichael, and N. Swart, "Recompression of JPEG images by requantization," *IEEE Transactions on Image Processing* , vol. 12, no 7, pp. 843–849, 2003.

[24] H. Jiang, "Uniform convergence rates for kernel density estimation," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1694–1703.

[25] B. Taşkesen, S. Shafieezadeh-Abadeh, and D. Kuhn, "Semi-discrete optimal transport: Hardness, regularization and numerical solution," *Mathematical Programming*, pp. 1–74, 2022.

[26] C. Villani, *Optimal transport*, ser. Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, 2009, vol. 338, old and new.

[27] A. M. Mathai and S. B. Provost, *Quadratic forms in random variables*. Marcel Dekker, New York, 1992.