On characterizing optimal Wasserstein GAN solutions for non-Gaussian data

Yu-Jui Huang†, Shih-Chun Lin*, Yu-Chih Huang§, Wen-Yi Zeng**, 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
** National Taiwan University of Science and Technology, Department of ECE , Taipei, Taiwan
‡ Bosch Center for AI, USA, wan-yi.lin@us.bosch.com

***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. Prior work on optimal parameters for Wasserstein GAN (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] and covert communication [3].

Traditionally, both the generator and discriminator in GAN are approximated by neural networks (NNs) [1] [4]. 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) [5] that replaces the JSD by the Wasserstein distance widely adopted in the optimal transport problem [6]. 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 [5].

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 [7] [8], which limit the development of GAN beyond heuristic methods in [1] [4]. This lack of rigorous analysis also restricts the evaluation of GANs' performance to subjective terms. One exception is [7] 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 [7, 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 [4]. Thus optimal WGAN solutions under LQG settings, with additional entropic and Sinkhorn regularizers, are also studied [8].

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 [7] [8]. Also, we attempt to exactly solve WGANs rather than their regularized versions as [8]. 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 [6]. To overcome the challenge, we mainly focus on one-dimensional data and generator in Sec. III, where we provide our main results: closed-form solutions for optimal generators in one-dimensional WGANs defined in Sec. II. The proofs are presented in Sec. IV, where we leverage result in [9] 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 and hence provide additional benefit for training WGAN with a decentralized system [10]. Empirical studies in Sec. V-A show that our closed-form WGAN parameters have good convergence behavior with synthetic data under both Gaussian and Laplace distributions. Finally, extensions to high-dimensional WGANs are discussed

in Sec. V-B.

II. Problem Formulation

For the WGAN considered in this paper, the overall transfer function of the generator NN is denoted as $G_\theta(.)$, 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 qth-order WGAN setting, one aims to solve

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

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 $G_\theta(Z)$ for a given θ and Gaussian input Z, respectively. Also, $\Pi(\mu, \nu^\theta)$ 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 $\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^q d\mu(x) d\nu^\theta(y) < \infty$. The WGAN problem described in (1) can be equivalently written as $\min_\theta \mathbb{E}_\mu[P(x, \theta)]$ where the inner-discriminator problem is defined as

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

Note that (2) belongs to the family of the optimal transport problems with q-th order Wasserstein distance, $q \in \{1, 2\}$ [6, Proposition 2.2].

Following [7], 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 problem for $d = 1$ 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-A.

III. Main Results

With $d = 1$, we present our main results for $q = 2$ in Sec. III-A and that for $q = 1$ Sec. III-B. It is worth mentioning that for applications in communication systems [2] [3], low-dimensional (even with $d = 1$) results can be very useful.

A. Results for $q = 2$

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

$$G_\theta(Z) = \theta_1 + \theta_2 h(Z), \tag{3}$$

where $h : \mathbb{R} \to \mathbb{R}$ and $Z \sim N(0, 1)$, also $(\theta_1, \theta_2) \in \mathbb{R} \times \mathbb{R}$ 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 data distribution μ has no atom (continuous), 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 $(\theta_1, \theta_2) \in \mathbb{R} \times \mathbb{R}$ as*

$$\begin{aligned} \theta_2^* &= \frac{\mathrm{Cov}\left(X, \Psi^{-1}(F_\mu(X))\right)}{\mathrm{Var}(h(Z))} \geq 0, \\ \theta_1^* &= \mathbb{E}_\mu[X] - \theta_2^* \mathbb{E}_g\left[h(Z)\right]; \end{aligned} \tag{5}$$

if (4) *is not met, (θ_1^*, θ_2^*) is given by replacing θ_2^* in* (5) *by*

$$\theta_2^* = \frac{\mathrm{Cov}\left(X, \Psi^{-1}(1 - F_\mu(X))\right)}{\mathrm{Var}(h(Z))} \leq 0, \tag{6}$$

where $\mathbb{E}_g$ is taking expectation over Gaussian $Z \sim \mathcal{N}(0, 1)$.

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., $\min_{\theta_1, \theta_2 \in \mathbb{R}} \mathbb{E}_\mu[P(X, \theta_1, \theta_2)]$ equals to

$$\min\left(\min_{\theta_1 \in \mathbb{R}, \theta_2 \geq 0} \mathbb{E}_\mu[P(X, \theta_1, \theta_2)], \min_{\theta_1 \in \mathbb{R}, \theta_2 \leq 0} \mathbb{E}_\mu[P(X, \theta_1, \theta_2)]\right) \tag{7}$$

We show that the solution of the first sub-problem is (5) while that for the 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 $\theta_2 \in \mathbb{R}_+$ ($\theta_2 \geq 0$), is given in Sec. IV-A, while those of other parts are omitted. ■

Let us now look at some specific cases of $h(z)$. For the sigmoid function $h(z)$, recall that the logit function

$$\mathrm{logit}(p) := \ln\left(\frac{p}{1-p}\right) \quad \text{for } p \in (0, 1),$$

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

$$\Psi(v) = \begin{cases} 0, & \text{for } v < 0, \\ \Phi(v), & \text{for } v \geq 0, \end{cases} \tag{8}$$

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 $(\theta_1, \theta_2) \in \mathbb{R} \times \mathbb{R}$ as*

$$\begin{aligned} \theta_2^* &= \frac{2\pi}{\pi - 1} \mathrm{Cov}\left(X, \Phi^{-1}(F_\mu(X)) 1_{\{F_\mu(X) > 1/2\}}\right) \\ \theta_1^* &= \mathbb{E}[X] - \theta_2^* / \sqrt{2\pi}; \end{aligned} \tag{10}$$

if (9) *is not met,* (θ_1^*, θ_2^*) *is given by replacing* θ_2^* *in* (10) *by*

$$\theta_2^* = -\frac{2\pi}{\pi - 1} \operatorname{Cov}\left(X, \Phi^{-1}(F_\mu(X)) 1_{\{F_\mu(X) \le 1/2\}}\right). \quad (11)$$

Proof: The proof is sketched in Sec. IV-B ∎

For the linear generator as [4, (27)] [7] [8], one can simplify the results in Theorem 1 and also get alternative closed-form formula for the optimal parameters 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* $(\theta_1, \theta_2) \in \mathbb{R} \times \mathbb{R}$ *as*

$$\theta_1^* = \mathbb{E}_\mu[X] \quad \textit{and} \quad \theta_2^* = \mathbb{E}_\mu[X \cdot \Phi^{-1}(F_\mu(X))], \quad (12)$$

also equivalently

$$\theta_2^* = \mathbb{E}_g[F_\mu^{-1}(\Phi(Z)) \cdot Z], \quad (13)$$

under $q = 2, d = 1$

Proof: Besides checking (4), it is easy to see that one can limit $\theta_2 \in \mathbb{R}_+$ 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 $G_\theta(Z) = \theta_1 + (-\theta_2) * (-Z)$, and absorb the case for $\theta_2 < 0$ into that for $\theta_2 \ge 0$. The rest of proof is omitted. ∎

Here we briefly compare our proposed WGAN solutions with the results for the LQG setting in [7, Theorem 1]. In the LQG setting, the d-dimensional synthetic data is generated by Gaussian distribution, i.e., $X \sim \mathcal{N}(0, \mathbf{K})$, the generator is restricted to be a linear generator of the form

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

with $Z \sim \mathcal{N}(0, \mathbf{I}_r)$ a Gaussian vector, and the loss function is quadratic (i.e., second-order WGAN). Since WGAN output is already Gaussian $\Theta Z \sim \mathcal{N}(0, \mathbf{K}')$, it is shown in [7, 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 $\mathbf{K}' = \Theta\Theta^T$ is a rank r matrix and $\mathbf{K}'$ and $\mathbf{K}$ share the same largest r eigenvalues and the corresponding eigenvectors. Unlike [7], our results can deal with non-Gaussian data distribution. Moreover, we can recover the result in [7] when $q = 2, d = 1$. Indeed, if $X \sim \mathcal{N}(0, \sigma^2)$, by looking into (12), we have

$$\Phi^{-1}(F_\mu(X)) = \Phi^{-1}(\Phi(X/\sigma)) = X/\sigma. \quad (15)$$

Plugging this into our solution in (12) shows that $\theta_1^* = 0$ and $\theta_2^* = \mathbb{E}[X \frac{X}{\sigma}] = \sigma$, coinciding with the result in [7] for $d = r = 1$. We must emphasize that neither of our and the results in [7] subsume the other as a special case as [7] considers general dimension d while the present work is not restricted to Gaussian data distribution.

B. Results for $q = 1$

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* (θ_1^*, θ_2^*) *of the population WGAN* (1) *with* $q = 1, d = 1$ *meet the following conditions*

$$\mathbb{E}_\mu\left[\textit{sign}\Big(\theta_1^* + \theta_2^* \Psi^{-1}(F_\mu(X)) - X\Big)\right] = 0,$$
$$\mathbb{E}_\mu\left[\textit{sign}\Big(\theta_1^* + \theta_2^* \Psi^{-1}(F_\mu(X)) - X\Big)\Psi^{-1}(F_\mu(X))\right] = 0,$$

when $\theta_2^* > 0$*, where* $\textit{sign}(x) = 1, 0, -1$ *for* $x > 0$*,* $x = 0$*,* $x < 0$*, respectively.*

Proof: Omitted. ∎

IV. The Proofs

A. Proof sketch for Theorem 1

In the following, we focus on solving the first sub-problem with $\theta_2 \in \mathbb{R}_+$ in (7). With $d = 1$, we recall the following result for the inner discriminator problem (2) of (1), by taking $\nu^\theta := P_{G_\theta(z)}$, the measure generated by the generator NN (with parameter $\theta := (\theta_1, \theta_2)$ 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 ([9], Theorem 5.1)**.** *Define* $t^\theta : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ *by*

$$t^\theta(x) := \sup\{y \in \mathbb{R} : F_{\nu^\theta}(y) \le F_\mu(x)\}. \quad (16)$$

Then, if μ *has no atom (*μ *is a continuous distribution),*

$$\inf_{\pi \in \Pi(\mu, \nu^\theta)} \int_{\mathbb{R}\times\mathbb{R}} |x - y|^2 d\pi(x, y) = \int_{\mathbb{R}} |x - t^\theta(x)|^2 d\mu(x),$$
$$\inf_{\pi \in \Pi(\mu, \nu^\theta)} \int_{\mathbb{R}\times\mathbb{R}} |x - y| d\pi(x, y) = \int_{\mathbb{R}} |x - t^\theta(x)| d\mu(x). \quad (17)$$

For the inner discriminator problem (2) with $q = 2, d = 1$, $\mathbb{E}_\mu[P(X, \theta_1, \theta_2)]$ for given $(\mu, \theta_1, \theta_2)$ equals to (17), where $t^\theta(x)$ is defined as in (16). Now we need to find a closed-form $t^\theta(x)$ to continue. From (3), let Ψ denote the CDF of $h(Z)$. If Ψ is continuous and strictly increasing and $\theta_2 \in \mathbb{R}_+$, (16) can be expressed in closed-form as

$$t^\theta(x) = \theta_1 + \theta_2 \Psi^{-1}(F_\mu(x)). \quad (18)$$

To see this, since $\theta_2 > 0$, observe that

$$\Psi\left(\frac{t^\theta(x) - \theta_1}{\theta_2}\right) = \mathbb{P}\left(h(Z) \le \frac{t^\theta(x) - \theta_1}{\theta_2}\right)$$
$$= \mathbb{P}\left(G_\theta(Z) \le t^\theta(x)\right) = F_{\nu^\theta}\left(t^\theta(x)\right) = F_\mu(x),$$

where the last equality follows from (16) and that F_{ν^θ} is continuous and strictly increasing; here, F_{ν^θ} inherits the same properties from Ψ thanks to (3). It follows that $\frac{t^\theta(x) - \theta_1}{\theta_2} = \Psi^{-1}(F_\mu(x))$, which yields (18). On the other hand, if $\theta_2 = 0$, since $G_\theta(Z) \equiv \theta_1 \in \mathbb{R}$, we have $F_{\nu^\theta}(y) = 1_{\{y \ge \theta_1\}}$. Plugging this into (16) directly gives $t^\theta(x) = \theta_1$ for all $x \in \mathbb{R}$. This particularly shows that (18) is also satisfied for the case $\theta_2 = 0$.

With closed-form representation (17)(18) for the inner problem (2), WGAN (1) with $q = 2, d = 1$ can be simplified to be the following stochastic minimization problem

$$\min_{\theta_1\in\mathbb{R},\theta_2\in\mathbb{R}_+} \mathbb{E}_\mu\left[\left|X-\theta_1+\theta_2\Psi^{-1}(F_\mu(X))\right|^2\right]. \tag{19}$$

Together with (3), (19) becomes the constrained optimization problem

$$\begin{aligned}\min_{\theta_1,\theta_2\in\mathbb{R}} \quad & J(\theta_1,\theta_2) := \int_{\mathbb{R}}\Big(\theta_1+\theta_2\Psi^{-1}(F_\mu(x))-x\Big)^2 F'_\mu(x)dx\\ \text{subject to} \quad & g(\theta_1,\theta_2) := -\theta_2 \le 0.\end{aligned} \tag{20}$$

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

$$\nabla J(\theta_1,\theta_2) + \lambda\nabla g(\theta_1,\theta_2) = 0, \tag{21}$$

$$\lambda g(\theta_1,\theta_2) = 0, \tag{22}$$

where $\lambda \ge 0$ is the Lagrange multiplier. By direct calculation, (21) becomes

$$\begin{aligned}&\int_{\mathbb{R}}\Big(\theta_1+\theta_2\Psi^{-1}(F_\mu(x))-x\Big)F'_\mu(x)dx = 0,\\ &\int_{\mathbb{R}}\Big(\theta_1+\theta_2\Psi^{-1}(F_\mu(x))-x\Big)\Psi^{-1}(F_\mu(x))F'_\mu(x)dx = \frac{\lambda}{2}.\end{aligned}$$

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

$$\begin{aligned}&\theta_1+\theta_2\mathbb{E}\left[\Psi^{-1}(F_\mu(X))\right] = \mathbb{E}[X],\\ &\theta_1\mathbb{E}\left[\Psi^{-1}(F_\mu(X))\right] + \theta_2\mathbb{E}\left[(\Psi^{-1}(F_\mu(X)))^2\right]\\ &-\mathbb{E}\left[X\Psi^{-1}(F_\mu(X))\right] = \lambda/2.\end{aligned}$$

Note that $F_\mu(X) \sim \text{Uniform}[0,1]$ from [11, 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

$$\begin{aligned}&\theta_1+\theta_2\mathbb{E}\left[h(Z)\right]-\mathbb{E}[X] = 0,\\ &\theta_1\mathbb{E}\left[h(Z)\right]+\theta_2\mathbb{E}\left[(h(Z))^2\right]-\mathbb{E}\left[X\Psi^{-1}(F_\mu(X))\right] = \frac{\lambda}{2}.\end{aligned}$$

Plugging $\theta_1 = \mathbb{E}[X] - \theta_2\mathbb{E}\left[h(Z)\right]$ from the first equality into the second one, we obtain

$$\mathbb{E}[X]\mathbb{E}\left[h(Z)\right]+\theta_2\text{Var}(h(Z))-\mathbb{E}\left[X\Psi^{-1}(F_\mu(X))\right]-\lambda/2 = 0.$$

Hence, a solution $(\theta_1,\theta_2,\lambda)$ to (21) must equivalently satisfy

$$\begin{aligned}&\theta_1 = \mathbb{E}[X] - \theta_2\mathbb{E}\left[h(Z)\right],\\ &\lambda/2 = \theta_2\text{Var}(h(Z)) - \text{Cov}(X,\Psi^{-1}(F_\mu(X))).\end{aligned} \tag{23}$$

To solve the KKT condition (21)-(22) for candidate minimizers, we already know that (21) boils down to (23), while (22) simply implies either $\lambda = 0$ or $\theta_2 = 0$. Also, because Ψ^{-1} is strictly increasing and F_μ is nondecreasing, the map $x \mapsto \Psi^{-1}(F_\mu(x))$ is nondecreasing. This readily implies

$$\text{Cov}\left(X,\Psi^{-1}(F_\mu(X))\right) \ge 0 \tag{24}$$

in (23) since $(x-x')(\Psi^{-1}(F_\mu(x)) - \Psi^{-1}(F_\mu(x'))) \ge 0$ for all $x, x' \in \mathbb{R}$. Specifically, by taking an independent but same distribution copy X' of X,

$$\begin{aligned}0 &\le \mathbb{E}\left[(X-X')\left(\Psi^{-1}(F_\mu(X)) - \Psi^{-1}(F_\mu(X'))\right)\right]\\ &= 2\,\text{Cov}\left(X,\Psi^{-1}(F_\mu(X))\right).\end{aligned}$$

We therefore separate the proof into two cases from (24).

Case I: $\text{Cov}(X,\Psi^{-1}(F_\mu(X))) > 0$. If $\theta_2 = 0$, (23) entails $\lambda = -2\,\text{Cov}(X,\Psi^{-1}(F_\mu(X))) < 0$, which violates the requirement $\lambda \ge 0$ in (21)-(22). If $\lambda = 0$, solving (23) yields

$$\begin{aligned}\theta_1^* &= \mathbb{E}[X] - \frac{\text{Cov}(X,\Psi^{-1}(F_\mu(X)))}{\text{Var}(h(Z))}\mathbb{E}\left[h(Z)\right],\\ \theta_2^* &= \frac{\text{Cov}(X,\Psi^{-1}(F_\mu(X)))}{\text{Var}(h(Z))} \ge 0.\end{aligned} \tag{25}$$

That is, the KKT condition gives a unique candidate optimizer $(\theta_1^*,\theta_2^*,\lambda^*)$, with (θ_1^*,θ_2^*) as in (25) and $\lambda^* = 0$. To check the corresponding second-order condition, let H_J and H_g denote the Hessian matrices of $J(\theta_1,\theta_2)$ and $g(\theta_1,\theta_2)$, respectively. Clearly, $H_g = O_{2\times 2}$ and is all-zero, which implies

$$H_J + \lambda H_g = 2\begin{bmatrix}1 & \mathbb{E}[h(Z)]\\ \mathbb{E}[h(Z)] & \mathbb{E}[(h(Z))^2]\end{bmatrix}. \tag{26}$$

As $\det(H_J+\lambda H_g)/4 = \mathbb{E}[(h(Z))^2] - \mathbb{E}[(h(Z))]^2 = \text{Var}(h(Z)) > 0$, $H_J + \lambda H_g$ is positive definite. Hence, we conclude that (θ_1^*,θ_2^*) in (25) is the unique minimizer of (20).

Case II: $\text{Cov}(X,\Psi^{-1}(F_\mu(X))) = 0$. As (23) entails $\lambda/2 = \theta_2\text{Var}(h(Z))$, we have $\lambda = \theta_2 = 0$ (by recalling that either $\lambda = 0$ or $\theta_2 = 0$). That is, the KKT condition gives a unique candidate optimizer $(\theta_1^*,\theta_2^*,\lambda^*) = (\mathbb{E}[X],0,0)$. Since $H_J + \lambda H_g$ is again given by (26), which is positive definite, we conclude that $(\theta_1^*,\theta_2^*) = (\mathbb{E}[X],0)$ is the unique minimizer of (20).

Combing the results of the above two cases yields (5).

B. Proof sketch for Theorem 2

As Ψ is neither continuous nor strictly increasing, (18) does not hold in general and we cannot directly apply proofs in Sec. IV-A. Again assume $\theta_2 \in \mathbb{R}_+$, we deduce from (8) that the closed form of (16) is

$$t^\theta(x) = \begin{cases}\theta_1, & \text{if } F_\mu(x) \le 1/2,\\ \theta_1+\theta_2\Phi^{-1}(F_\mu(x)), & \text{if } F_\mu(x) > 1/2.\end{cases} \tag{27}$$

Similarly to (20), now WGAN (1) simplifies to

$$\begin{aligned}\min_{\theta_1,\theta_2\in\mathbb{R}}\Bigg\{&\int_{\{F_\mu(x)\le 1/2\}}\Big(\theta_1-x\Big)^2F'_\mu(x)dx\\ &+\int_{\{F_\mu(x)>1/2\}}\Big(\theta_1+\theta_2\Phi^{-1}(F_\mu(x))-x\Big)^2F'_\mu(x)dx\Bigg\}\\ \text{subject to} \quad & g(\theta_1,\theta_2) := -\theta_2 \le 0.\end{aligned}$$

Then optimal θ_1^*,θ_2^* in (10) can be obtained by solving this stochastic optimization problem.

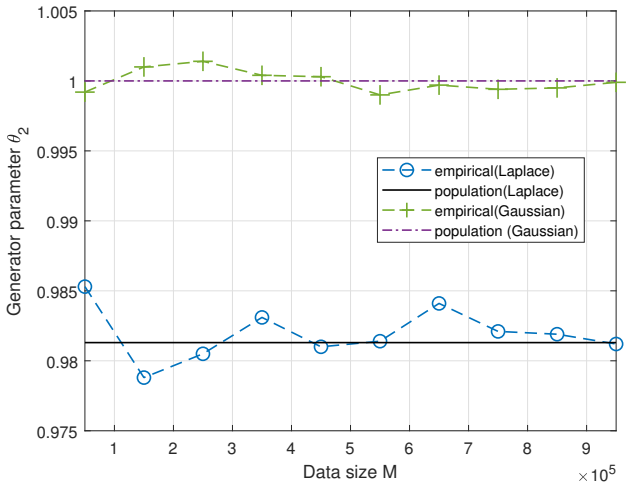

Fig. 1. Comparison of optimal parameters in (12)(13) with their estimates (28) using synthetic data.

V. Discussions

In this section, we show empirical results of the derived optimal parameters and discuss how to extend our solutions to multi-dimensional WGANs.

A. 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, and its estimate with synthetic data when μ is chosen to be (a) $\mathcal{N}(0,1)$ and (b) Laplace distribution [12] with mean 0 and scale $1/\sqrt{2}$ (which also has a unit variance). Specifically, by generating training data $\{x_i\}_{i=1}^M$ according to the true distribution μ, we empirically estimate θ_2^* from (12) by

$$\frac{1}{M}\sum_{i=1}^{M} x_i \Phi^{-1}(\hat{F}_\mu(x_i)) \tag{28}$$

and the kernel density estimation [13] 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) is known to be nicely estimated by the sample mean, so the convergence is not shown in Fig. 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 [7, Theorem 2]. It is our future work to analyze the convergence rate for the setting considered in this paper, i.e., $d = 1$ but with a non-Gaussian distribution and a non-linear generator.

B. From one-dimension to multi-dimension

Throughout the paper, we focus on WGAN with $d = 1$ and provide a closed-form solution. In practice, real data is rarely 1-dimensional and extensions to a general case is called for. We can replace the cost function $\|x - G_\theta(z)\|^q$ in (1) with $c(x, G_\theta(z))$ that accepts inputs with unequal dimensions $G_\theta(z) \in \mathbb{R}, x \in \mathbb{R}^d$, as [7, eq(5)]. One application of this approach is the convolutional neural network where a linear filter $w \in \mathbb{R}^d$ is applied on x (part of the image) to get a feature. In this case

$$c(x, G_\theta(z)) = |w^T x - G_\theta(z)|^2, \tag{29}$$

by assuming a large stride and $q = 2$. Our population WGAN results can be easily extended to this case. For example, with linear generator in (3), from Corollary 1, we have

$$\theta_1^* = \mathbb{E}_\mu[w^T X], \theta_2^* = \left(\mathrm{Cov}\left(w^T X, \Phi^{-1}(F_\mu(w^T X))\right)\right)^+. \tag{30}$$

Another potential direction for future research is to approximate the d-dimensional population WGAN problem (1) by replacing the Wasserstein distance in (2) with the sliced Wasserstein distance [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 GAN output onto a direction $\omega \in \Omega$ and compute the Wasserstein distance with respect to the projected 1-dimensional distributions μ_ω and ν_ω^θ. This is done for every $\omega \in \Omega$ and the average distance is

$$\int_{\omega\in\Omega} \inf_{\pi\in\Pi(\mu_\omega,\nu_\omega^\theta)} \mathbb{E}[|X-Y|^2]\mathrm{d}\omega. \tag{31}$$

Though it looks promising to adapt our method to the WGAN problem with the sliced Wasserstein distance, one critical issue remains to be solved: our method finds the best parameter θ that minimizes the Wasserstein distance with 1-dimensional inputs, but the sliced WGAN problem looks for θ that minimizes the average Wasserstein distances of all 1-dimensional projections of inputs. A first step to resolve this issue could be combining the proposed approach with the max sliced Wasserstein distance in [14] given by

$$\max_{\omega\in\Omega} \inf_{\pi\in\Pi(\mu_\omega,\nu_\omega^\theta)} \mathbb{E}[|X-Y|^2], \tag{32}$$

which reduces the average in (31) to a single maximizing direction. Note that even with high-dimensional generator as (14), the generator output after projection ν_ω^θ is still a Gaussian random variable with variance $\omega\Theta\Theta^T\omega^T$. Then one can leverage Corollary 1 to find the necessary condition optimal Θ should satisfy for (32). Thus at least for high-dimensional linear generator (14), we expect to also provide solutions for max sliced WGAN *without* the Gaussian data assumption made in [7] [8].

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] 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.

[4] 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.

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

[6] 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.

[7] 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.

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

[9] 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.

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

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

[12] 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.

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

[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.