UNDERSTANDING GANS: THE LQG SETTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative Adversarial Networks (GANs) have become a popular method to learn a probability model from data. Many GAN architectures with different optimization metrics have been introduced recently. Instead of proposing yet another architecture, this paper aims to provide an understanding of some of the basic issues surrounding GANs. First, we propose a natural way of specifying the loss function for GANs by drawing a connection with supervised learning. Second, we shed light on the statistical performance of GANs through the analysis of a simple LQG setting: the generator is **linear**, the loss function is **quadratic** and the data is drawn from a **Gaussian** distribution. We show that in this setting: 1) the optimal GAN solution converges to population Principal Component Analysis (PCA) as the number of training samples increases; 2) the number of samples required scales exponentially with the dimension of the data; 3) the number of samples scales almost linearly if the discriminator is constrained to be quadratic. Moreover, under this quadratic constraint on the discriminator, the optimal finite-sample GAN performs simply empirical PCA.

1 INTRODUCTION

Learning a probability model from data is a fundamental problem in statistics and machine learning. Building off the success of deep learning methods, Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) have given this age-old problem a face-lift. In contrast to traditional methods of parameter fitting like maximum likelihood estimation, the GAN approach views the problem as a *game* between a *generator* whose goal is to generate fake samples that are close to the real data traaining samples and a *discriminator* whose goal is to distinguish between the real and fake samples. The generator and the discriminators are typically implemented by deep neural networks. GANs have achieved impressive performance in several domains (e.g., (Ledig et al., 2016; Reed et al., 2016)). Since (Goodfellow et al., 2014), many variations of GANs have been developed, including f-GAN (Nowozin et al., 2016), MMD-GAN (Dziugaite et al., 2015; Li et al., 2015), WGAN (Arjovsky et al., 2017), improved WGAN (Gulrajani et al., 2017), relaxed WGAN (Guo et al., 2017), Least-Squares GAN (Mao et al., 2016), Boundary equilibrium GAN (Berthelot et al., 2017), etc. These GANs use different metrics in the optimization for training the generator and discriminator networks (Liu et al., 2017).

The game theoretic formulation in GANs can be viewed as the dual of an optimization that minimizes a distance measure between the empirical distributions of the fake and real samples. In the population limit where there are infinite number of samples, this optimization minimizes the distance between the generated distribution and the true distribution from which the data is drawn. In the original GAN framework, this distance measure is the Jenson Shannon-divergence. However, Arjovsky et al (Arjovsky et al., 2017) noted that this distance does not depend on the generated distribution whenever its dimension is smaller than that of the true distribution. In this typical case, the Jenson Shannon divergence does not serve as a useful criterion in choosing the appropriate generated distribution. To resolve this issue, (Arjovsky et al., 2017) proposed the Wasserstein GAN (WGAN) which uses the first-order Wasserstein distance instead of Jensen-Shannon divergence. This distance is meaningful even when the dimension of the generated distribution is less than the true distribution. Nevertheless there are many other distance measures that satisfy this criterion and it is not clear how to choose among them. This is responsible in part for the fact that there are so many different GAN architectures. In fact, there is currently some confusion in the literature even on the basic issue of how to specify the loss function for GANs. For example, while the "Wasserstein" in Wasserstein

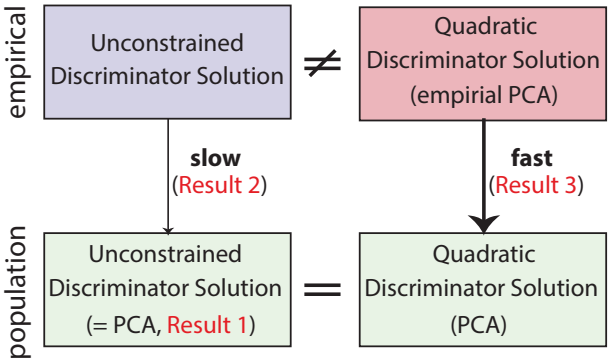

Figure 1: Summary of main results. In the LQG setting, the population optimal GAN solution is PCA when the discriminator is unconstrained and when the discriminator is constrained to be quadratic. But the convergence to the population optimal is exponentially faster under the quadratic constraint on the discriminator.

GAN refers to the use of Wasserstein distance in the *distance measure* between the generated and true distributions. the "Least Squares" in Least-Squares GAN (Mao et al., 2016) refers to the use of squared error in the *discriminator optimization objective*. These are two totally different types of objects. This is in contrast to the situation in supervised learning, where how the loss function is specified in the formulation is clear.

A central issue in any learning problem is *generalization*: how close a model learnt from a finite amount of data is to the true distribution. Or, in statistical terms, how fast is the rate of convergence of the learnt model to the true model as a function of number of samples? Arora et al (Arora et al., 2017) have recently studied this problem for GANs. They observed that for Wasserstein GAN, if there are no constraints on the generator or the discriminator, the number of samples required to converge scales exponentially in the data dimension. They then showed that if the discriminator is constrained to be in a parametric family, then one can achieve convergence almost linearly in the number of parameters in that family (Theorem 3.1 in (Arora et al., 2017)). However, the convergence is *no longer* measured in the Wasserstein distance but in a new distance measure they defined (the neural network distance). The result is interesting as it highlights the role of the discriminator in generalization, but it is somewhat unsatisfactory in that the distance measure needs to be modified to tailor to the specific constraints on the discriminator. Moreover, the result requires the invention of (yet) another family of distance measures for GANs.

In this paper, we first argue that there is a natural way to specify the loss function for GANs, in an analogous way as in the supervised learning setup. The resulting optimal GAN minimizes a generalized loss-function dependent Wasserstein distance between the generated distribution and the true distribution, and the dual formulation of this Wasserstein distance leads to a loss-function dependent discriminator architecture. To study the impact of the generator and the discriminator on the generalization performance in this distance measure, we focus on a basic *Linear-Quadratic-Gaussian* setting: the generator is constrained to be linear of a given rank k, the loss function is quadratic, and the true distribution is Gaussian. In this setting, the optimal GAN minimizes the *second-order* Wasserstein distance between the generated distribution and the empirical data distribution among all linear generators of a given rank. We show the following results:

1 in the population limit as the number of data samples grow, the optimal generated distribution is the rank k Gaussian distribution retaining exactly the top k principal components of the true distribution, i.e. GAN performs PCA in the population limit;

2 The number of samples required for convergence in (second-order) Wasserstein distance however scales exponentially with the dimension of the data distribution.

3 Under a further constraint that the discriminator is *quadratic*, GAN converges to the same population-optimal PCA limit, but with the number of samples scaling almost linearly with the dimension. The constrained GAN simply performs empirical PCA, and in the case

when the rank k of the generator is the same as the dimension of the data distribution, GAN is equivalent to maximum likelihood estimation of the underlying Gaussian model.

These results are summarized in Figure 1.

(Arora et al., 2017) observed that the number of samples required to generalize for GAN is exponential in the number of samples when there are no constraints on either the generator or the discriminator. (They proved the result for first-order Wasserstein distance but a similar result holds for second-order Wasserstein distance as well, see Lemma 2 in Section 3.) Result 2 above says that even constraining the generator drastically to be linear cannot escape this exponential scaling. Result 3 says that this exponential scaling is not due to statistical limitation, but much better inference can be obtained by constraining the discriminator appropriately. Similar to Theorem 3.1 in (Arora et al., 2017), Result 3 points to the importance of constraining the discriminator. But there are two key differences. First, the convergence in Result 3 is with respect to the original (second-order) Wasserstein distance, not another distance measure tailored to the constraint on the discriminator as in Theorem 3.1 in (Arora et al., 2017). Thus, the original loss function is respected. Second, the population limit is the same as the PCA limit achieved without constraints on the discriminator. Thus, by imposing a discriminator constraint, the rate of convergence is drastically improved without sacrificing the limiting performance. There is no such guarantee in (Arora et al., 2017). On the other hand, our results are obtained in a specific setting.

The LQG setting, dating back to at least Gauss, has been widely used across many fields, including statistics, machine learning, control, signal processing and communication. It has resulted in celebrated successes such as linear regression, the Wiener filter, the Kalman filter, PCA, etc., and is often used to establish a baseline to understand more complex models. We believe it serves a similar role here for GANs[1]. Indeed it allows us to make a clear connection between GAN and PCA, perhaps the most basic of unsupervised learning methods.

When the discriminator is not constrained, the number of samples scale exponentially with the dimension. It turns out that a simple characterization of the performance of GAN in that regime can be obtained using rate-distortion theory (Cover & Thomas, 2012). In the population limit, quadratic GAN minimizes the Wasserstein distance between the generated distribution and the true data distribution. This distance is the solution of an unconstrained optimal transport problem. It turns out that when there are a finite number of samples, the performance is again characterized by an optimal transport between the generated and true distributions, but with a constraint on the mutual information of the transport to be less than the logarithm of the number of samples. This constraint reflects the fact that only empirical samples from the true distribution are seen and therefore the generated distribution cannot be too strongly coupled to the true distribution. We discuss this story in the appendix.

The rest of the paper is organized as follows. In Section 2, we discuss a formulation of the GAN problem for general loss functions. In Section 3, we specialize to the LQG setting and analyze the generalization performance of GAN. In Section 4, we analyze the performance of GAN under a quadratic constraint on the discriminator. In Section 5, we present some experimental results.

2 A General Formulation for GANs

Let $\{\mathbf{y}_i\}_{i=1}^n$ be n observed data points in $\mathbb{R}^d$ drawn i.i.d. from the distribution $\mathbb{P}_Y$. Let $\mathbb{Q}_Y^n$ be the empirical distribution of these observed samples. Moreover, let $\mathbb{P}_X$ be a normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I}_k)$. GANs can be viewed as an optimization that minimizes a distance between the observed empirical distribution $\mathbb{Q}_Y^n$ and the generated distribution $\mathbb{P}_{g(X)}$. The *population* GAN optimization replaces $\mathbb{Q}_Y^n$ with $\mathbb{P}_Y$. The question we ask in this section is: what is a natural way of specifying a loss function ℓ for GANs and how it determines the distance?

[1]The importance of establishing simple baselines in machine learning has also been expressed by Ben Recht (personal communication, October, 2017).

2.1 WGAN REVISITED

Let us start with the WGAN optimization (Arjovsky et al., 2017):

$$\inf_{g(.)\in\mathcal{G}} W_1(\mathbb{P}_Y, \mathbb{P}_{g(X)}), \tag{1}$$

where $\mathcal{G}$ is the set of generator functions, and the p-th order Wasserstein distance between distributions $\mathbb{P}_{Z_1}$ and $\mathbb{P}_{Z_2}$ is defined as (Villani, 2008)

$$W_p^p(\mathbb{P}_{Z_1}, \mathbb{P}_{Z_2}) := \inf_{\mathbb{P}_{Z_1,Z_2}} \mathbb{E}\left[\|Z_1 - Z_2\|^p\right], \tag{2}$$

where the minimization is over all joint distributions with marginals fixed. Replacing (2) in (1), the WGAN optimization can be re-written as

$$\inf_{g(.)\in\mathcal{G}} \inf_{\mathbb{P}_{g(X),Y}} \mathbb{E}\left[\|Y - g(X)\|\right]. \tag{3}$$

or equivalently:

$$\inf_{g(.)\in\mathcal{G}} \inf_{\mathbb{P}_{X,Y}} \mathbb{E}\left[\|Y - g(X)\|\right], \tag{4}$$

where the second minimization is over all joint distributions $\mathbb{P}_{X,Y}$ with fixed marginals $\mathbb{P}_X$ and $\mathbb{P}_Y$.

We now connect (4) to the *supervised learning* setup. In supervised learning, the joint distribution $\mathbb{P}_{X,Y}$ is fixed and the goal is to learn a relationship between the feature variable represented by $X \in \mathbb{R}^k$, and the target variable represented by $Y \in \mathbb{R}^d$, according to the following optimization:

$$\inf_{g(.)\in\mathcal{G}} \mathbb{E}\left[\ell\left(Y, g(X)\right)\right], \tag{5}$$

where ℓ is the *loss* function. Assuming the marginal distribution of X is the same in both optimizations (4) and (5), we can connect the two optimization problems by setting $\ell(y, y') = \|y - y'\|$ in optimization (5). Note that for every fixed $\mathbb{P}_{X,Y}$, the solution of the supervised learning problem (5) yields a predictor g which is a feasible solution to the WGAN optimization problem 4. Therefore, the WGAN optimization (3) can be re-interpreted as solving the *easiest* such supervised learning, over all possible joint distributions $\mathbb{P}_{X,Y}$ with fixed $\mathbb{P}_X$ and $\mathbb{P}_Y$.

2.2 FROM SUPERVISED TO UNSUPERVISED LEARNING

GAN is a solution to an unsupervised learning problem. What we are establishing above is a general connection between supervised and unsupervised learning problems: a good predictor g learnt in a supervised learning problem can be used to generate samples of the target variable Y. Hence, to solve an unsupervised learning problem for Y with distribution $\mathbb{P}_Y$, one should solve the easiest supervised learning problem $\mathbb{P}_{X,Y}$ with given marginal $\mathbb{P}_Y$ (and $\mathbb{P}_X$, the randomness generating distribution). This is in contrast to the traditional view of the unsupervised learning problem as observing the feature variable X without the label Y. (Thus in this paper we break with tradition and use Y to denote the unknown distribution and X as randomness for the generator in stating the GAN problem.)

This connection between supervised and unsupervised learning leads to a natural way of specifying the loss function in GANs: we simply replace the l_2 Euclidean norm in (3) with a general loss function ℓ:

$$\inf_{g(.)\in\mathcal{G}} \inf_{\mathbb{P}_{g(X),Y}} \mathbb{E}\left[l\left(Y, g(X)\right)\right]. \tag{6}$$

The inner optimization is the optimal transport problem between distributions of $g(X)$ and Y (Villani, 2008) with general cost ℓ. This is a linear programming problem for general cost, so there is always a dual formulation (the Kantorovich dual). The dual formulation can be interpreted as a generalized discriminator optimization problem for the cost ℓ. (For example, in the case of ℓ being the Euclidean norm, we get the WGAN architecture; see Figure 2(a).) Hence, we propose (6) as a formulation of GANs for general loss functions.

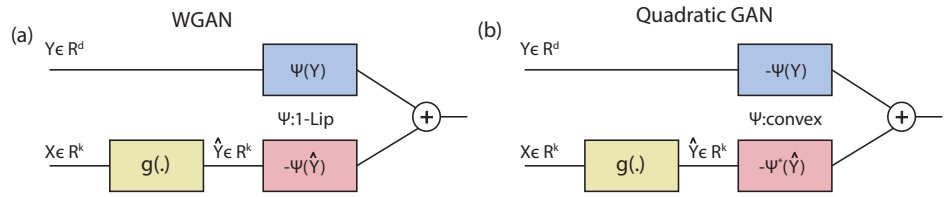

Figure 2: Dual (min-max) formulations of (a) WGAN, and (b) Quadratic GAN.

2.3 Quadratic Loss

The most widely used loss function in supervised learning is the quadratic loss: $l(y, y') = \|y - y'\|^2$ (*squared* Euclidean norm). Across many fields its use had led to many important discoveries. With the connection between supervised and unsupervised learning in mind, this loss function should be a prime choice to consider in GANs as well. This choice of the loss function leads to the *quadratic GAN* optimization:

$$\inf_{g(.)\in\mathcal{G}} W_2^2(\mathbb{P}_Y, \mathbb{P}_{g(X)}). \tag{7}$$

Since Wasserstein distances are weakly continuous measures in the probability space (Villani, 2008), similar to WGAN, the optimization of the quadratic GAN is well-defined even if $k < d$. The dual formulation (discriminator) for W_2 is shown in Figure 2. Note that in this dual, the discriminator applies Ψ to the real data and the convex conjugate Ψ^* to the generated (fake) data.

The *empirical* quadratic GAN optimization can be formulated by replacing $\mathbb{P}_Y$ with the empirical distribution $\mathbb{Q}_Y^n$ of the data as follows:

$$\inf_{g(.)\in\mathcal{G}} W_2^2(\mathbb{Q}_Y^n, \mathbb{P}_{g(X)}). \tag{8}$$

Note that while in practice one generates fake samples from X, we will keep the notations simpler in this paper by assuming we can generate the exact distribution $g(X)$, i.e. we can generate as many fake samples as we wish. Almost all our results can be extended to the case when we have finite number of samples from X.

For the rest of the paper, we will focus on the problem (8) for the particular case of Y Gaussian of dimension d, and g linear of rank $k \leq d$. This is the **LQG** setting for GANs.

3 GANs under the LQG setup

3.1 The Population GAN Optimization

First, we analyze the population GAN optimization under the LQG setup. We have the following lemma:

Lemma 1 *Let $\mathcal{S}$ be a k dimensional subspace in $\mathbb{R}^d$. Let $\hat{Y}$ be a random variable whose support lies in $\mathcal{S}$. Then, $\hat{Y}^*$, the optimal solution of the optimization*

$$\inf_{\mathbb{P}_{\hat{Y}}} W_2^2(\mathbb{P}_Y, \mathbb{P}_{\hat{Y}}), \tag{9}$$

is the projection of Y to $\mathcal{S}$.

Proof 1 *See Section C.2.*

This Lemma holds even if $\mathbb{P}_Y$ is a non-Gaussian distribution. However, the optimal solution of GAN may not be generated as $g(X)$ when $\mathbb{P}_X \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_k)$ and $g(.)$ is restricted to be linear.

Using Lemma 1 and under the LQG setup, we show that the optimal solution for the population GAN optimization is the same as the PCA solution. PCA is the most standard unsupervised dimensionality

reduction approach (Jolliffe, 2002). PCA computes an optimal linear mapping from Y to $\hat{Y}$ under the rank constraint on the covariance matrix of $\hat{Y}$ ($\mathbf{K}_{\hat{Y}}$). We say $\hat{Y}$ is the k-PCA solution of Y if $\mathbf{K}_{\hat{Y}}$ is a rank k matrix whose top k eigenvalues and eigenvectors are the same as top k eigenvalues and eigenvectors of the covariance matrix of Y ($\mathbf{K}_Y$).

Theorem 1 *Let $Y \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_Y)$ where $\mathbf{K}_Y$ is full-rank. Let $X \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_k)$ where $k \leq d$. The optimal population GAN solution of optimization* (7) *under linear $\mathcal{G}$ is the k-PCA solution of Y.*

Proof 2 *See Section C.3.*

Lemma 1 holds if we replace W_2 with W_1. However, the conclusion of Theorem 1 is tied to the W_2 distance because the PCA optimization also considers the quadratic projection loss.

3.2 The Empirical GAN Optimization

In practice, one solves the GAN optimization over the empirical distribution of the data $\mathbb{Q}_Y^n$, not the population distribution $\mathbb{P}_Y$. Thus, it is important to analyze how close optimal empirical and population GAN solutions are in a given sample size n. This notion is captured in the generalization error of the GAN optimization, defined as follows:

Definition 1 (Generalization of GANs) *Let n be the number of observed samples from Y. Let $\hat{g}(.)$ and $g^*(.)$ be optimal solutions for empirical and population GAN optimizations. Then,*

$$d_{\mathcal{G}}(\mathbb{P}_Y, \mathbb{Q}_Y^n) := W_2^2(\mathbb{P}_Y, \mathbb{P}_{\hat{g}(X)}) - W_2^2(\mathbb{P}_Y, \mathbb{P}_{g^*(X)}), \tag{10}$$

is a random variable representing the excess error of the empirical solution.

$d_{\mathcal{G}}(\mathbb{P}_Y, \mathbb{Q}_Y^n)$ can be viewed as a distance between $\mathbb{P}_Y$ and $\mathbb{Q}_Y^n$ which depends on $\mathcal{G}$. To have a proper generalization property, one needs to have $d_{\mathcal{G}}(\mathbb{P}_Y, \mathbb{Q}_Y) \to 0$ quickly as $n \to \infty$. Before analyzing the convergence rate of $d_{\mathcal{G}}(\mathbb{P}_Y, \mathbb{Q}_Y^n)$ for linear $\mathcal{G}$, we characterize this rate for an unconstrained $\mathcal{G}$. For an unconstrained $\mathcal{G}$, the second term of (10) is zero (this can be seen using a space filling generator function (Cannon & Thurston, 1987)). Moreover, $\mathbb{P}_{\hat{g}(X)}$ can be arbitrarily close to $\mathbb{Q}_Y^n$. Thus, we have

Lemma 2 *If $\mathcal{G}$ is unconstrained, we have*

$$d_{\mathcal{G}}(\mathbb{P}_Y, \mathbb{Q}_Y^n) = W_2^2(\mathbb{P}_Y, \mathbb{Q}_Y^n), \tag{11}$$

which goes to zero with the rate of $\mathcal{O}(n^{-2/d})$.

The approach described for the unconstrained $\mathcal{G}$ corresponds to the *memorization* of the empirical distribution $\mathbb{Q}_Y^n$ using the trained model. Note that one can write

$$n^{-\frac{2}{d}} = 2^{-\frac{2\log(n)}{d}}.$$

Thus, to have a small $W_2^2(\mathbb{P}_Y, \mathbb{Q}_Y^n)$, the number of samples n should be exponentially large in d (Canas & Rosasco, 2012). It is possible that for some distributions $\mathbb{P}_Y$, the convergence rate of $W_2^2(\mathbb{P}_Y, \mathbb{Q}_Y^n)$ is much faster than $\mathcal{O}(n^{-2/d})$. For example, reference (Weed & Bach, 2017) shows that if $\mathbb{P}_Y$ is clusterable (i.e., Y lies in a fixed number of separate balls with fixed radii), then the convergence of $W_2^2(\mathbb{P}_Y, \mathbb{Q}_Y^n)$ is fast. However, even in this case, one optimal strategy would be to memorize observed samples which is not the mechanism used in GANs.

A natural question is whether constraining the family of generator functions $\mathcal{G}$ can improve the generalization of GANs. Below we study this question in the LQG setup as follows. To simplify calculations, we assume that $\mathbb{P}_Y \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and $d = k$. Under these assumptions, the GAN optimization (8) can be re-written as

$$\min_{\mu, \mathbf{K}} \; W_2^2(\mathcal{N}(\mu, \mathbf{\Sigma}), \mathbb{Q}_Y^n), \tag{12}$$

where $\mathbf{K}$ is the covariance matrix with the eigen decomposition $\mathbf{K} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^t$. The optimal population solution of this optimization is $\mu_{pop}^* = \mathbf{0}$ and $\mathbf{K}_{pop}^* = \mathbf{I}$, which provides a zero Wasserstein loss.

Theorem 2 *Let μ_n^* and $\mathbf{K}_n^*$ be optimal solutions for optimization* (12)*. Then, $\mu_n^* \to \mathbf{0}$ with the rate of $\mathcal{O}(d/n)$ and $Tr(\mathbf{\Sigma}_n^*) \to d$ with the rate of $\mathcal{O}(n^{-2/d})$.*

Proof 3 *See Section C.4.*

Theorem 2 indicates that there is a bias in GAN's estimation of the true distribution which translates to the slow convergence of the generalization error. Note that in the Wasserstein space, the empirical distribution $\mathbb{Q}_Y^n$ and the population distribution $\mathbb{P}_Y$ are far from each other (the distance between them concentrates around $n^{-2/d}$ (Canas & Rosasco, 2012)). Thus, if there exists another Gaussian distribution within the sphere around $\mathbb{Q}_Y^n$ with the radius of $n^{-2/d}$, the Wasserstein-based learning method will converge to the wrong Gaussian distribution. This phenomenon causes a bias term in estimating the true distribution. In Section A, we characterize this bias term explicitly in an asymptotic regime by drawing a connection between GANs and the rate-distortion optimization in information theory.

Theorem 2 considers the regime where $k = d$. In practice, the dimension of the generated distribution is often much smaller than that of the true one (i.e., $k \ll d$). In this case, GAN's convergence rate can be increased from $\mathcal{O}(n^{-2/d})$ to $\mathcal{O}(n^{-2/k})$. However, this faster convergence comes at the expense of the increased bias term of the excess error (the second term of (10)). The trade off is favorable if Y is near low rank.

Theorem 2 suggests that GAN's convergence rate is slow. In practice, however, GANs have demonstrated impressive performances. In the next section, we show that by suitably constraining the GAN optimization, the convergence rate can be enhanced significantly.

4 GANs with Constrained Discriminators

In this section, first we review the min-max (dual) formulation of WGAN (Arjovsky et al., 2017). Then, we characterize the min-max formulation of the quadratic GAN. Finally, we show that a properly *constrained* quadratic GAN achieves the empirical PCA solution, which converges to the population optimal with an exponentially faster rate of convergence compared to the case when the discriminator is unconstrained.

Using the Kantorovich duality (Villani, 2008), the first-order Wasserstein distance $W_1(\mathbb{P}_Y, \mathbb{P}_{g(X)})$ can be written as the following optimization (Arjovsky et al., 2017):

$$W_1(\mathbb{P}_Y, \mathbb{P}_{g(X)}) = \sup_{\psi(.):\text{1-Lip}} \mathbb{E}\left[\psi(Y) - \psi(\hat{Y})\right], \tag{13}$$

where the function $\psi(.)$ is restricted to be 1-Lipschitz. This dual formulation of W_1 is then used in optimization (1) to implement WGAN in a min-max architecture similar to the one of the original GAN (Figure 2). In this architecture, $\psi(.)$ is implemented by deep neural networks.

In a similar way, one can write the second-order Wasserstein distance $W_2^2(\mathbb{P}_Y, \mathbb{P}_{g(X)})$ as the following optimization (Villani, 2008):

$$W_2^2(\mathbb{P}_Y, \mathbb{P}_{g(X)}) = \mathbb{E}[\|Y\|^2] + \mathbb{E}[\|g(X)\|^2] + 2 \sup_{\psi(.):\text{convex}} -\mathbb{E}\left[\psi(Y)\right] - \mathbb{E}\left[\psi^*(g(X))\right], \tag{14}$$

where $\psi^*(\hat{\mathbf{y}}) := \sup_{\mathbf{v}} (\mathbf{v}^t \hat{\mathbf{y}} - \psi(\mathbf{v}))$ is the convex-conjugate of the function $\psi(.)$. Similarly, this dual formulation of W_2^2 can be used to implement the quadratic GAN optimization (7) in a min-max architecture which can be interpreted as a game between optimizing two functions $g(.)$ and $\psi(.)$ (Figure 2).

The following lemma characterizes the optimal discriminator of optimization (14) (Chernozhukov et al., 2017):

Lemma 3 *let $\mathbb{P}_Y$ be absolutely continuous whose support contained in a convex set in $\mathbb{R}^d$. For a fixed $g(.) \in \mathcal{G}$, let ψ^{opt} be the optimal discriminator function of optimization* (14)*. This solution is unique. Moreover, we have*

$$\bigtriangledown \psi^{opt}(Y) \overset{dist}{=} g(X), \tag{15}$$

where $\overset{dist}{=}$ means matching distributions.

In the LQG setup, since $g(X)$ is Gaussian, $\bigtriangledown\psi^{opt}$ is a linear function. Thus, without loss of generality, $\psi(.)$ in the discriminator optimization can be constrained to $\psi(\mathbf{y}) = \mathbf{y}^t\mathbf{A}\mathbf{y}$ where $\mathbf{A}$ is positive semidefinite. Therefore, we have

$$\begin{aligned} W_2^2\left(\mathbb{P}_Y, \mathbb{P}_{g(X)}\right) &= \mathbb{E}[\|Y\|^2] + \mathbb{E}[\|g(X)\|^2] + 2 \sup_{\psi(\mathbf{y})=\mathbf{y}^t\mathbf{A}\mathbf{y}, \mathbf{A}\succeq 0} -\mathbb{E}\left[\psi(Y)\right] - \mathbb{E}\left[\psi^*(g(X))\right] \\ &= Tr(\mathbf{K}_Y) + Tr(\mathbf{K}_{g(X)}) + 2 \sup_{\mathbf{A}\succeq 0} -Tr(\mathbf{A}\mathbf{K}_Y) - Tr(\mathbf{A}^\dagger\mathbf{K}_{g(X)}), \end{aligned} \tag{16}$$

where $\mathbf{A}^\dagger$ is the pseudo inverse of the matrix $\mathbf{A}$.

If we compute the right hand side of (16) on the empirical distribution $\mathbb{Q}_Y^n$, it will *not* be equal to $W_2^2(\mathbb{Q}_Y^n, \mathbb{P}_{g(X)})$ $\mathbb{Q}_Y^n$ is not linear and therefore the optimal discriminator is not quadratic. Nevertheless, by computing the quadratically constrained discriminator optimization on the empirical distribution, we obtain

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n \|\mathbf{y}_i\|^2 + \mathbb{E}[\|g(X)\|^2] + 2 \sup_{\psi(\mathbf{y})=\mathbf{y}^t\mathbf{A}\mathbf{y}, \mathbf{A}\succeq 0} -\frac{1}{n}\sum_{i=1}^n \psi(\mathbf{y}_i) - \mathbb{E}\left[\psi^*(g(X))\right] \\ &= Tr(\hat{\mathbf{K}}_Y) + Tr(\mathbf{K}_{g(X)}) + 2 \sup_{\mathbf{A}\succeq 0} -Tr(\mathbf{A}\hat{\mathbf{K}}_Y) - Tr(\mathbf{A}^\dagger\mathbf{K}_{g(X)}) \\ &= W_2^2(\mathbb{P}_Z, \mathbb{P}_{g(X)}), \end{aligned} \tag{17}$$

where $\hat{\mathbf{K}}_Y$ is the sample covariance matrix and $Z \sim \mathcal{N}(\mathbf{0}, \hat{\mathbf{K}}_Y)$ [2]. Therefore, the empirical constrained quadratic GAN solves the following optimization:

$$\inf_{g(.)\in\mathcal{G}} W_2^2(\mathbb{P}_Z, \mathbb{P}_{g(X)}). \tag{18}$$

Using Theorem 1, the optimal $g(X)$ to this problem is the empirical PCA solution, i.e. keeping the top k principal components of the *empirical covarance matrix.*

Theorem 3 *Under the LQG setup, the empirical constrained quadratic GAN optimization is equivalent to the empirical PCA.*

Consider the case where $d = k$ (the case $k < d$ is similar). The second term in the generalization distance $d_{\mathcal{G}}(\mathbb{Q}_Y^n, \mathbb{P}_Y)$ (10) is zero. Therefore, we have

$$d_{\mathcal{G}}(\mathbb{P}_Y, \mathbb{Q}_Y^n) = W_2^2(\mathbb{P}_Z, \mathbb{P}_Y) = Tr(\mathbf{K}_Y) + Tr(\hat{\mathbf{K}}_Y) - 2Tr\left(\left(\mathbf{K}_Y^{1/2}\hat{\mathbf{K}}_Y\mathbf{K}_Y^{1/2}\right)^{1/2}\right). \tag{19}$$

Note that convergence of $d_{\mathcal{G}}(\mathbb{P}_Y, \mathbb{Q}_Y^n)$ only depends on the convergence of the sample covariance to the population covariance, which occurs at a quick rate of $\tilde{\mathcal{O}}(\sqrt{d/n})$ (Rippl et al., 2016).

5 Experimental Results

For experiments, we generate n i.i.d. samples from $\mathbb{P}_Y \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, represented as Q_Y^n. We then fit a d dimensional Gaussian distribution $\mathcal{N}(\mu, \mathbf{K})$ to $\mathbb{Q}_Y^n$ using two methods: Maximum Likelihood (ML) estimation, which computes the sample mean and the sample covariance; and WGAN (Arjovsky et al., 2017) with an affine generator function. Note that according to Theorem 3, ML is equivalent to the constrained quadratic GAN (18)*. Moreover, note that the WGAN implementation uses W_1 and not W_2 in its optimization. Although analyzing GANs with W_2 is more tractable than that of W_1, in practice we do not expect a significant difference between their performance. Considering this and owing to the lack of an implementation of GANs with W_2, we perform numerical experiments using the WGAN implementation. Details of the experiments can be found in Section B.*

[2]Note that by considering $\psi(\mathbf{y}) = \mathbf{y}^t\mathbf{A}\mathbf{y} + \mathbf{a}^t\mathbf{y}$ where $\mathbf{A} \succeq 0$, we would have obtained $Z \sim \mathcal{N}\left(\hat{\mu}_Y, \hat{\mathbf{K}}_Y\right)$ where $\hat{\mu}_Y$ is the sample mean. For simplicity, we ignore the affine terms.

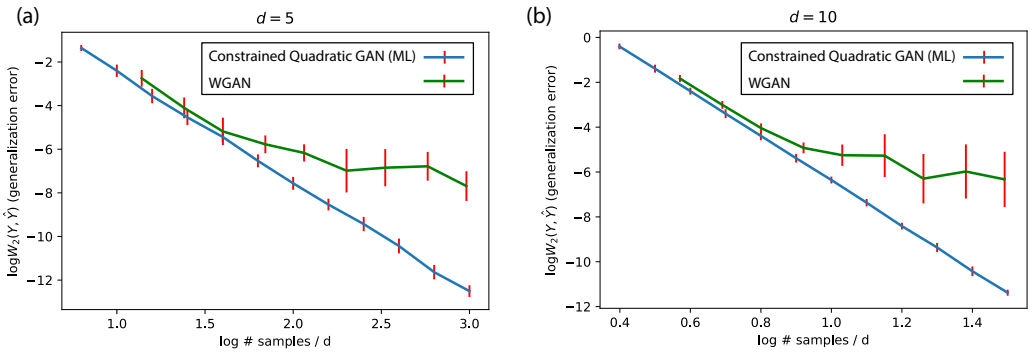

Figure 3: Generalization errors of constrained quadratic GAN (ML) and WGAN under the LQG setup.

Let $\hat{\mu}$ and $\hat{\mathbf{K}}$ be the estimated mean and the covariance. For evaluation, we compute

$$\|\hat{\mu}\|^2 + \|\mathbf{I} - \hat{\mathbf{K}}^{1/2}\|_F^2, \tag{20}$$

which is the W_2^2 distance between $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and $\mathcal{N}\left(\hat{\mu}, \hat{\mathbf{K}}\right)$ (see Lemma 4 in Section C).

Figure 3 demonstrates the estimation error of the ML (constrained quadratic GAN)and WGAN methods for $d = 5$ and $d = 10$ and in different sample sizes. These figures are consistent with Theorem 2 and results of Section A which suggest that GAN's convergence can be slow owing to a bias in its optimal empirical solution with respect to the population one. Moreover, this figure shows that the convergence of the constrained quadratic GAN (ML) is fast. Finally, in our experimental results of Figure 3, one should take into the consideration practical issues of the WGAN implementation such as the use of the stochastic gradient descent, convergence to bad locals, etc.

REFERENCES

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. arXiv preprint arXiv:1701.07875, *2017.*

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (gans). arXiv preprint arXiv:1703.00573, *2017.*

David Berthelot, Tom Schumm, and Luke Metz. Began: Boundary equilibrium generative adversarial networks. arXiv preprint arXiv:1703.10717, *2017.*

Guillermo Canas and Lorenzo Rosasco. Learning probability measures with respect to optimal transport metrics. In Advances in Neural Information Processing Systems, *pp. 2492–2500, 2012.*

James W Cannon and William P Thurston. Group invariant peano curves. 1987.

Victor Chernozhukov, Alfred Galichon, Marc Hallin, Marc Henry, et al. Monge–kantorovich depth, quantiles, ranks and signs. The Annals of Statistics, *45(1):223–256, 2017.*

Thomas M Cover and Joy A Thomas. Elements of information theory. *John Wiley & Sons, 2012.*

Gintare Karolina Dziugaite, Daniel M Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. arXiv preprint arXiv:1505.03906, *2015.*

Clark R Givens, Rae Michael Shortt, et al. A class of wasserstein metrics for probability distributions. The Michigan Mathematical Journal, *31(2):231–240, 1984.*

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in neural information processing systems, *pp. 2672–2680, 2014.*

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. arXiv preprint arXiv:1704.00028*, 2017.*

Xin Guo, Johnny Hong, Tianyi Lin, and Nan Yang. Relaxed wasserstein with applications to gans. arXiv preprint arXiv:1705.07164*, 2017.*

Ian Jolliffe. Principal Component Analysis*. Wiley Online Library, 2002.*

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. arXiv preprint arXiv:1609.04802*, 2016.*

Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pp. 1718–1727, 2015.*

Shuang Liu, Olivier Bousquet, and Kamalika Chaudhuri. Approximation and convergence properties of generative adversarial learning. arXiv preprint arXiv:1705.08991*, 2017.*

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, and Zhen Wang. Multi-class generative adversarial networks with the l2 loss function. arXiv preprint arXiv:1611.04076*, 2016.*

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In Advances in Neural Information Processing Systems*, pp. 271–279, 2016.*

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. arXiv preprint arXiv:1605.05396*, 2016.*

Thomas Rippl, Axel Munk, and Anja Sturm. Limit laws of the empirical wasserstein distance: Gaussian distributions. Journal of Multivariate Analysis*, 151:90–109, 2016.*

Cédric Villani. Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.*

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. arXiv preprint arXiv:1707.00087*, 2017.*

Appendix

A A RATE-DISTORTION VIEW OF GANS

As (Arjovsky et al., 2017) suggested, there is a strong connection between GANs and the optimal transport problem (e.g., this connection is reflected in the use of the Wasserstein distance in GAN's optimization). Moreover, the optimal transport theory and the optimal quantization theory are strongly tied to one other (e.g., (Villani, 2008; Canas & Rosasco, 2012)). On the other hand, in the regime where the sample size n grows exponentially in d, Shannon has noted that the optimal quantization is strongly related to the rate-distortion optimization (Cover & Thomas, 2012). In this section, we exploit these connections to better understand GANs through the rate-distortion analysis. Note that in the light of Theorem 2 and without properly constraining the discriminator function, the regime where n is exponential in d provides an appropriate resolution to study GAN's problem.

Consider the following optimization:

$$\min_{\mathbb{P}_{Y,Z}} \ \mathbb{E}\left[\|Y-Z\|^2\right], \quad \mathbb{P}_Y : \textit{fixed}, \quad I(Y;Z) \leq Rd, \tag{21}$$

where $I(Y;Z)$ is the mutual information between Y and Z. Let $\mathbb{P}_Z^*$ be an optimal solution of this optimization. Shannon has shown that in the regime where $n = 2^{Rd}$, generating n i.i.d. samples from $\mathbb{P}_Z^*$ is an optimal quantization strategy of $\mathbb{P}_Y$ (Cover & Thomas, 2012). We refer to this optimization as the rate-distortion optimization.

Using the rate-distortion theory, our goal is to understand W_2 and therefore GANs when n grows exponentially in d. Consider the following optimization:

$$\min_{\mathbb{P}_{Y,Z}} \ \mathbb{E}\left[\|Y-Z\|^2\right], \quad \mathbb{P}_Y : \textit{fixed}, \quad \mathbb{P}_Z = \mathbb{P}_Y, \quad I(Y;Z) \leq Rd. \tag{22}$$

Note that unlike optimization (21) *where $\mathbb{P}_Z$ can be an arbitrary distribution, in optimization* (22) *$\mathbb{P}_Z$ is constrained to be equal to $\mathbb{P}_Y$. This is to have n i.i.d. samples from $\mathbb{P}_Z = \mathbb{P}_Y$ (represented as $\mathbb{Q}_Y^n$) be an optimal quantizer of $\mathbb{P}_Y$.*

Theorem 4 *Let $D(R)$ be the optimal value of optimization* (22) *where $\mathbb{P}_Y$ is absolutely continuous and L Lipschitz. Then, for a given $\epsilon > 0$, we have*

$$D(R) \dot{\leq} W_2^2(\mathbb{P}_Y, \mathbb{Q}_Y^n) \dot{\leq} D(R-\epsilon), \tag{23}$$

where $\dot{\leq}$ means inequality in the asymptotic regime of $n = 2^{dR}$.

Proof 4 *See Section C.5.*

Connecting W_2 to the rate-distortion optimization allows us to exploit standard rate-distortion results to better understand Wasserstein distances and in turn GANs. For example,

Theorem 5 *Let $\mathbb{P}_Y \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Then, the minimum value of optimization* (22) *is*

$$2d\left(1 - \sqrt{1 - 2^{-2R}}\right). \tag{24}$$

Proof 5 *See Section C.6.*

For a large value of R, the first-order tailor expansion of (24) *leads to an expression of $d2^{-2R}$ which provides a $2^{-2R} \doteq n^{-2/d}$ distortion per dimension.*

Next, we consider the empirical GAN optimization (8) *when* $d = k$*. This optimization can be written*

$$\min_{\mathbb{P}_{\hat{Y}}} \; W_2^2(\mathbb{Q}_Y^n, \mathbb{P}_{\hat{Y}}), \tag{25}$$

where $\hat{Y} = g(X)$*. Although Theorem 4 indicates a point-wise convergence of the objective of this optimization to a rate-distortion optimization, for an unconstrained* $\mathbb{P}_{\hat{Y}}$ *and an arbitrary* $\mathbb{P}_Y$*, a uniform convergence may not occur. However, we conjecture that such a uniform convergence will occur under the LQG setup. In such a case, the GAN optimization of* (25) *can be written as the following optimization:*

$$\begin{aligned} \min_{\mathbb{P}_{\hat{Y}}} \; \min_{\mathbb{P}_{Z,\hat{Y}}} \; & \mathbb{E}\left[\|\hat{Y} - Z\|^2\right], \\ & \mathbb{P}_{\hat{Y}} : \textit{fixed}, \\ & \mathbb{P}_Z = \mathbb{P}_Y, \\ & I(\hat{Y}; Z) \leq Rd. \end{aligned} \tag{26}$$

Conjecture 1 *Under the LQG setup, the empirical GAN optimization* (25) *and the rate-distortion optimization* (26) *are equivalent.*

Note that unlike optimization (22) *where both marginal distributions are fixed (and thus, it is a constrained rate-distortion optimization), in optimization* (26)*, one of the marginal distributions (*$\mathbb{P}_{\hat{Y}}$*) is not fixed due to the generator optimization. Therefore, optimization* (26) *is the same as the rate-distortion optimization* (21)*.*

Next, we analyze optimization (26) *when* $\mathbb{P}_Y \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ *which is the rate distortion problem for an i.i.d. Gaussian source (Cover & Thomas, 2012). In this case, the optimal* $\mathbb{P}^*_{\hat{Y}}$ *of optimization* (25) *is*

$$\mathbb{P}^*_{\hat{Y}} = \mathcal{N}\left(\mathbf{0}, (1 - D^*)\mathbf{I}_d\right), \tag{27}$$

where $D^* = 2^{-2R}$*.*

Given conjecture 1, this result has two main consequences:

- *In the context of GANs, with* $X \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ *and generator functions* $g \in \mathcal{G}$*, this result suggests that linear generators are optimal when* $\mathbb{P}_Y$ *is Gaussian (in the asymptotic regime). This can be viewed as an unsupervised analogue of the fact that the Minimum Mean Squared Error (MMSE) estimator is linear for the Gaussian supervised learning problem.*
- *The optimal generated distribution* $\mathbb{P}^*_{\hat{Y}}$ *which is the minimum* W_2 *distance fitting distribution to* n *i.i.d. samples of* Y *is always a biased estimate of the true distribution* $\mathbb{P}_Y \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$*. The bias is in fact simple: it is a shrinkage to the origin by the same amount on every dimension. The amount of shrinkage is precisely* $D^* = 2^{-2R}$*.*

So far in this section, we have considered the regime where $k = d$*. The analysis of* $k < d$ *can be done similarly. For example, consider the following GAN optimization where* $\hat{Y}$ *is restricted to be in a subspace* $\mathcal{S}$ *whose rank is* $k \leq d$*:*

$$\min_{\mathbb{P}_{\hat{Y}}} \; W_2^2(\mathbb{P}_{\hat{Y}}, \mathbb{Q}_Y^n). \tag{28}$$

One can decompose $Y = Y_{S'} + Y_S$ *where* Y_S *is the projection of* Y *onto the subspace* $\mathcal{S}$*. Thus, we can re-write optimization* (28) *as*

$$\mathbb{E}_{\mathbb{Q}_{Y_{S'}}}\left[\|Y_{S'}\|^2\right] + \min_{\mathbb{P}_{\hat{Y}}} \; W_2^2(\mathbb{P}_{\hat{Y}}, \mathbb{Q}_{Y_S}^n). \tag{29}$$

Similar to optimization (26) *and given conjecture 1, we can re-write this optimization as*

$$\begin{aligned} \mathbb{E}_{\mathbb{Q}_{Y_{S'}}}\left[\|Y_{S'}\|^2\right] + \min_{\mathbb{P}_{Y_S,\hat{Y}}} \; & \mathbb{E}\left[\|\hat{Y} - Y_S\|^2\right], \\ & \mathbb{P}_{Y_S} : \textit{fixed}, \\ & I(\hat{Y}; Y_S) \leq Rk. \end{aligned} \tag{30}$$

Let $\mathbb{P}_Y \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and $\mathbb{P}_{Y_S} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_S)$ where the eigen decomposition of $\mathbf{K}_S$ is $\mathbf{K}_S = \mathbf{U}_S \mathbf{\Sigma}_S \mathbf{U}_S^t$. Moreover, suppose diagonal elements of $\mathbf{\Sigma}_S$ are $\sigma_i^2(S)$. Then, optimal $\mathbb{P}^*_{\hat{Y}}$ for optimization (30) *is*

$$\mathbb{P}^*_{\hat{Y}} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{U}_S \begin{bmatrix} \sigma_1^2 - D_1^* & \cdots & 0 \\ 0 & \sigma_2^2 - D_2^* & 0 \\ \vdots & \ddots & \vdots \\ 0 & 0 & \sigma_k^2 - D_k^* \end{bmatrix} \mathbf{U}_S^t\right), \tag{31}$$

where $D_i^* = \min_i(\lambda, \sigma_i^2)$ and λ is chosen so that $R = \frac{1}{k}\sum_{i=1}^k \frac{1}{2}\log\left(\sigma_i^2/D_i^*\right)$ (the reverse water-filling solution (Cover & Thomas, 2012)). Similar to the $d = k$ case, even through the generator functions $\mathcal{G}$ were unconstrained, for Gaussian $\mathbb{P}_Y$, the optimal generator is linear. Therefore, the objective function of optimization (30) *can be re-written as*

$$\left(d - \sum_{i=1}^k \sigma_i^2(S)\right) + \sum_{i=1}^k D_i^*. \tag{32}$$

The GAN optimization aims to select a subspace to minimize this objective, or alternatively to solve

$$\max_{S} \quad \sum_{i=1}^k \sigma_i^2(S) - \sum_{i=1}^k D_i^*. \tag{33}$$

The first term is the energy of the projected signal onto the subspace while the second term is the quantization cost (or, the optimal transport cost) in the subspace. The larger the energy of the signal is, the larger the quantization cost is, and vice versa. This tarde-off can lead to having an optimal solution of optimization (33) *which is not equal to the PCA subspace. However, for a low-rank distribution $\mathbb{P}_Y$, we show that optimization* (33) *always selects the PCA subspace:*

Theorem 6 *Let $\mathbb{P}_Y \sim \mathcal{N}(\mathbf{0}, diag(1, ..., 1, 0, ..., 0))$ whose covariance has k non-zero eigenvalues. Then, the PCA-subspace is the optimal solution of optimization* (33)*.*

Proof 6 *See Section C.7.*

B Details of Experiments

The WGAN is implemented in pytorch. Denote fully connected layer with the input dimension d_{in} and the output dimension d_{out} as $FC(d_{in}, d_{out})$. The generator can be represented as $FC(d, d)$; and the discriminator can be represented as $FC(d, n_f) - ReLU - FC(n_f, n_f) - ReLU - FC(n_f, n_f) - ReLU - FC(n_f, 1)$. The model is trained 100k iterations with batch size 128 with Adam optimizer. The learning rate is set to 2×10^{-4}. As for hyper parameters, n_f is set to 128, the ratio of iterations between discriminator and generator is set to 5, and the weight clipping threshold is set to -0.02 and 0.02. Both ML and WGAN are repeated 10 times for each setting, and the mean and standard deviation is calculated and plotted (68.3% confidence interval).

C Proofs

C.1 Notation and Preliminary Lemmas

For matrices we use bold-faced upper case letters, for vectors we use bold-faced lower case letters, and for scalars we use regular lower case letters. For example, $\mathbf{X}$ represents a matrix, $\mathbf{x}$ represents a vector, and x represents a scalar number. $\mathbf{I}_n$ is the identity matrix of size $n \times n$. $\mathbf{1}_{n_1,n_2}$ is the all one matrix of size $n_1 \times n_2$. When no confusion arises, we drop the subscripts. $\mathbf{1}\{x = y\}$ is the indicator function which is equal to one if $x = y$, otherwise it is zero. Tr$(\mathbf{X})$ and $\mathbf{X}^t$ represent the trace and the transpose of the matrix $\mathbf{X}$, respectively. $\|\mathbf{x}\|_2 = \mathbf{x}^t\mathbf{x}$ is the second norm of the vector $\mathbf{x}$. When no confusion arises, we drop the subscript. $\|\mathbf{X}\|$ is the operator (spectral) norm of the matrix $\mathbf{X}$. $< \mathbf{x}, \mathbf{y} >$ is the inner product between vectors $\mathbf{x}$ and $\mathbf{y}$. $\mathbf{A}^\dagger$ is the pseudo inverse of the matrix $\mathbf{A}$. The eigen decomposition of the matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is denoted by $\mathbf{A} = \sum_{i=1}^n \lambda_i(\mathbf{A})\mathbf{u}_i(\mathbf{A})\mathbf{u}_i(\mathbf{A})^t$,

where $\lambda_i(\mathbf{A})$ is the i-th largest eigenvalue of the matrix $\mathbf{A}$ corresponding to the eigenvector $\mathbf{u}_i(\mathbf{A})$. We have $\lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \cdots$. $\mathcal{N}(\mu, \mathbf{K})$ is the Gaussian distribution with mean μ and the covariance $\mathbf{K}$. $KL(\mathbb{P}_X, \mathbb{P}_Y)$ is the Kullback Leibler divergence between two distributions $\mathbb{P}_X$ and $\mathbb{P}_Y$. $\tilde{\mathcal{O}}(d)$ is the same as $\mathcal{O}(d\log(d))$.

Lemma 4 *Let $Y \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_Y)$ and $\hat{Y} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{\hat{Y}})$. Then,*

$$\begin{aligned} W_2^2(\mathbb{P}_Y, \mathbb{P}_{\hat{Y}}) &= Tr(\mathbf{K}_Y) + Tr(\mathbf{K}_{\hat{Y}}) - 2Tr\left(\left(\mathbf{K}_{\hat{Y}}^{1/2}\mathbf{K}_Y\mathbf{K}_{\hat{Y}}^{1/2}\right)^{1/2}\right) \\ &= Tr(\mathbf{K}_Y) + Tr(\mathbf{K}_{\hat{Y}}) - 2Tr\left(\left(\mathbf{K}_Y^{1/2}\mathbf{K}_{\hat{Y}}\mathbf{K}_Y^{1/2}\right)^{1/2}\right). \end{aligned} \tag{34}$$

Proof 7 *See reference (Givens et al., 1984).*

C.2 Proof of Lemma 1

Let $Y = Y_{\mathcal{S}'} + Y_{\mathcal{S}}$ where $Y_{\mathcal{S}}$ represents the projection of Y onto the subspace $\mathcal{S}$. Since

$$\mathbb{E}\left[\|Y - \hat{Y}\|^2\right] = \mathbb{E}\left[\|Y_{\mathcal{S}'}\|^2\right] + \mathbb{E}\left[\|Y_{\mathcal{S}} - \hat{Y}\|^2\right] \tag{35}$$

choosing $\hat{Y} = Y_{\mathcal{S}}$ achieves the minimum of optimization (9).

C.3 Proof of Theorem 1

Let $\mathcal{S}$ be a fixed subspace of rank k where $\hat{Y}$ lies on. According to Lemma 1, if $\hat{Y}$ is unconstrained, the optimal $\hat{Y}^*$ is the projection of Y onto $\mathcal{S}$ (i.e., $\hat{Y}^* = Y_{\mathcal{S}}$). Moreover, since Y is Gaussian, $\hat{Y}$ is also Gaussian. Therefore, there exists a linear $g(.)$ such that $\mathbb{P}_{\hat{Y}^*} = \mathbb{P}_{g(X)}$ where $X \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Thus, the problem simplifies to choosing a subspace where $\mathbb{E}\left[\|Y_{\mathcal{S}}\|^2\right]$ is maximized, which is the same as the PCA optimization.

C.4 Proof of Theorem 2

Let $\mathbf{y}_1$,...,$\mathbf{y}_n$ be n i.i.d. samples of $\mathbb{P}_Y$. Let $\hat{\mu}$ be the sample mean. Since $\mathbb{P}_Y$ is absolutely continuous, the optimal W_2 coupling between $\mathbb{Q}_Y^n$ and $\mathbb{P}_Y$ is deterministic (Villani, 2008). Thus, every point $\mathbf{y}_i$ is coupled with an optimal transport vornoi region with the centroid $\mathbf{c}_{y_i}^{(\mu,\mathbf{K})}$. Therefore, we have

$$\begin{aligned} & W_2^2(\mathcal{N}(\mu, \mathbf{K}), \mathbb{Q}_Y^n) \\ &= \|\mu\|^2 + Tr(\mathbf{\Sigma}) + \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{y}_i\|^2 - \frac{2}{N}\sum_{i=1}^{N}\mathbf{y}_i^t\mathbf{c}_{y_i}^{(\mu,\mathbf{K})} \\ &= \|\mu\|^2 + Tr(\mathbf{\Sigma}) + \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{y}_i\|^2 - \frac{2}{N}\sum_{i=1}^{N}\mathbf{y}_i^t\left(\mathbf{U}\mathbf{\Sigma}^{1/2}\mathbf{U}^t\mathbf{c}_{y_i}^{(\mathbf{0},\mathbf{I})} + \mu\right) \\ &= \|\mu\|^2 - 2\mu\hat{\mu} + Tr(\mathbf{\Sigma}) + \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{y}_i\|^2 - 2Tr(\mathbf{U}\mathbf{\Sigma}^{1/2}\mathbf{U}^t\mathbf{A}) \end{aligned} \tag{36}$$

where

$$\mathbf{A} := \frac{1}{N}\sum_{i=1}^{N}\mathbf{c}_{y_i}^{(\mathbf{0},\mathbf{I})}\mathbf{y}_i^t. \tag{37}$$

The first step in (36) *follows from the definition of W_2, the second step follows from the optimal coupling between $\mathcal{N}(\mu, \mathbf{K})$ and $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and the third step follows from the matrix trace equalities.*

Therefore,

$$\bigtriangledown_\mu W_2^2(\mathcal{N}(\mu, \mathbf{\Sigma}), \mathbb{Q}_Y^n) = 2\mu - 2\hat{\mu}, \tag{38}$$

which leads to $\mu_n^* = \hat{\mu}$. Moreover, each component of the sample mean is distributed according to $\mathcal{N}(\mathbf{0}, 1/n)$. Thus, $\|\mu_N^*\|^2 \sim \chi_d^2/n$ which goes to zero with the rate of $\tilde{\mathcal{O}}(d/n)$.

Let σ_i^2 be the i-th diagonal element of $\mathbf{\Sigma}$. Moreover, define $\mathbf{B} = \mathbf{U}^t\mathbf{A}\mathbf{U}$. Therefore, we have

$$\bigtriangledown_{\sigma_i} W_2^2(\mathcal{N}(\mu, \mathbf{\Sigma}), \mathbb{Q}_Y^n) = 2\sigma_i - 2b_{i,i}, \tag{39}$$

where $b_{i,i}$ is the i-th diagonal element of the matrix $\mathbf{B}$. Thus, $\sigma_i^* = b_{i,i}$ and $Tr(\mathbf{\Sigma}^*) = Tr(\mathbf{B}) = Tr(\mathbf{A})$.

Furthermore, we have

$$W_2^2(\mathbb{P}_Y, \mathbb{Q}_Y^n) = d + \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{y}_i\|^2 - 2Tr(\mathbf{A}), \tag{40}$$

which goes to zero with the rate of $\mathcal{O}(n^{-2/d})$ (Canas & Rosasco, 2012). Since $\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{y}_i\|^2$ goes to d with the rate of $\mathcal{O}(\sqrt{d/n})$ (because it has a χ-squared distribution), $Tr(\mathbf{A})$ goes to d with a rate of $\mathcal{O}(n^{-2/d})$. Combining this result with $Tr(\mathbf{\Sigma}^*) = Tr(\mathbf{A})$ completes the proof.

C.5 Proof of Theorem 4

The lower-bound follows directly from the rate-distortion theorem (Cover & Thomas, 2012). Thus, we focus on proving the upper bound. To show this, we use the triangle inequality of the W_2 distance: $W_2(\mathbb{P}_Y, \mathbb{Q}_Y^n) \leq W_2(\mathbb{P}_Y, \mathbb{P}_Z) + W_2(\mathbb{P}_Z, \mathbb{Q}_Y^n)$, for some $\mathbb{P}_Z$ that we will construct.

We first introduce some notation: Let $A = \{\mathbf{y}_1, ..., \mathbf{y}_n\}$ be n i.i.d. samples from $\mathbb{P}_Y$, with the empirical distribution $\mathbb{Q}_Y^n$. Moreover, let $B = \{\mathbf{b}_1, ..., \mathbf{b}_{n'}\}$ be another $n' = 2^{d(R-\epsilon)}$ i.i.d. samples from $\mathbb{P}_Y$. Let $\mathcal{V}_i$ be the voronoi region assigned to point $\mathbf{b}_i$. Let

$$p_{\mathcal{V}_i} = \mathbb{P}[Y \in \mathcal{V}_i]. \tag{41}$$

Now, define A_i as the subset of samples of A that fall in the voronoi region $\mathcal{V}_i$ and let $N_i = |A_i|$. N_i can be written as the sum of indicator functions:

$$N_i = \sum_{j=1}^{n}\mathbf{1}\{\mathbf{y}_j \in \mathcal{V}_i\}. \tag{42}$$

Thus, N_i is Binomial with the parameter $p_{\mathcal{V}_i}$.

Define the probability measure $\mathbb{P}_Z$ on the set A such that for every $\mathbf{y}_i \in \mathcal{V}_j$, we have

$$\mathbb{P}_Z : \mathbf{y}_i \to p_i = \frac{p_{\mathcal{V}_j}}{N_j}. \tag{43}$$

First, we show that $W_2(\mathbb{P}_Z, \mathbb{Q}_Y^n)$ goes to zero as n increases. Let $Z_i := \sqrt{n/p_{\mathcal{V}_i}}$. Note that with high probability $|Z_i| < n2^{-d\epsilon'}$ for $0 < \epsilon' < \epsilon$. Thus, with high probability, $|p_i - 1/n| < 2^{-d\epsilon'}/n$. Therefore, with high probability, the total variation distance between $\mathbb{P}_Z$ and $\mathbb{Q}_Y^n$ can be bounded as $\|\mathbb{P}_Z - \mathbb{Q}_Y^n\|_{TV} \leq \mathcal{O}(2^{-d\epsilon'})$. Moreover, from Theorem 6.15 of (Villani, 2008), we have

$$W_2(\mathbb{P}_Z, \mathbb{Q}_Y^n) \leq r\sqrt{\|\mathbb{P}_Z - \mathbb{Q}_Y^n\|_{TV}}, \tag{44}$$

where r is the diameter of the space. In our case, $r = \mathcal{O}(d)$. Thus,

$$W_2(\mathbb{P}_Z, \mathbb{Q}_Y^n) \leq \mathcal{O}(d2^{-d\epsilon'/2}), \tag{45}$$

which goes to zero as n increases.

Next, we show that $W_2^2(\mathbb{P}_Y, \mathbb{P}_Z)$ is less than $R(D-\epsilon)$. To show this, we will construct an admissible transport plan whose cost is less than $R(D-\epsilon)$: in each region $\mathcal{V}_i$, we solve an optimal transport plan between $\mathbb{P}_{Y|Y\in\mathcal{V}_i}$ and A_i points. Combining these local transport plan provides an admissible transport plan between $\mathbb{P}_Y$ and $\mathbb{P}_Z$. Thus, $W_2^2(\mathbb{P}_Y, \mathbb{P}_Z)$ is smaller than or equal to the aggregate cost of local transport plans.

Recall that the region $\mathcal{V}_i$ contains N_i points of A. As $n \to \infty$, N_i increases as well. Since $\mathbb{P}_Y$ is absolutely continuous and Lipschitz, as $n \to \infty$, $\mathbb{P}_Y$ in each region resembles a uniform distribution. Since the number of points in the region increases, the optimal transport cost between $\mathbb{P}_{Y|Y\in\mathcal{V}_i}$ and points A_i is smaller than the transport cost between $\mathbb{P}_{Y|Y\in\mathcal{V}_i}$ and its centroid. Therefore, the overall transport cost between $\mathbb{P}_Y$ and $\mathbb{P}_Z$ (by combining local transports) is smaller than $D(R-\epsilon)$. This completes the proof.

C.6 Proof of Theorem 5

Let us first show that the claimed minimum value is achieved with a certain coupling dictated by the following virtual channel: $Y = cZ + W$ where W follows $\mathcal{N}(\mathbf{0}, (1-c^2)\mathbf{I}_d)$ and is independent of Z. Note that this coupling together with a choice $\mathbb{P}_Z \sim N(\mathbf{0}, \mathbf{I}_d)$ still respects $\mathbb{P}_Y = \mathbb{P}_Z \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Another constraint $I(Y;Z) \leq Rd$ yields a condition on c: $c \leq \sqrt{1-2^{-2R}}$. Notice that

$$\begin{aligned} Rd \geq I(Y;Z) &= h(Y) - h(Y|Z) = h(Y) - h(W) \\ &= \frac{1}{2}\log_2(2\pi e)^d - \frac{1}{2}\log_2(2\pi e(1-c^2))^d \\ &= \frac{d}{2}\log_2\frac{1}{1-c^2}. \end{aligned} \tag{46}$$

This together with the minimum value of $c = \sqrt{1-2^{-2R}}$ then gives:

$$\mathbb{E}\left[\|Y-Z\|^2\right] = (c-1)^2 + 1 - c^2 = 2(1-c) = 2(1-\sqrt{1-2^{-2R}}). \tag{47}$$

Now let us prove the converse: the optimal solution cannot go below the claimed bound. Consider a constraint: $dR \geq I(Y;Z) = h(Y) - h(Y|Z)$. Note that $h(Y|Z)$ is maximized when (Y,Z) are jointly Gaussian. So in an effort to find a lower bound on $I(Y;Z)$, we consider such a jointly Gaussian setting, i.e., $Y = cZ + W$ where $W \sim \mathcal{N}(\mathbf{0}, (1-c^2)\mathbf{I}_d)$, independent of Z. This together with a simple calculation then yields: $c \leq \sqrt{1-2^{-2R}}$. Putting this to the objective function, we get:

$$\mathbb{E}\left[\|Y-Z\|^2\right] \geq 2(1-\sqrt{1-2^{-2R}}).$$

This completes the proof.

C.7 Proof of Theorem 6

The objective function of optimization (33) *for the PCA subspace is $k(1-2^{-2R})$. Now consider a subspace $\mathcal{S}$ such that $\sum_{i=1}^k \sigma_i^2(S) = \alpha k$ where $0 < \alpha < 1$. Below, we show that this subspace cannot be the optimal solution of optimization* (33).

We have

$$\begin{aligned} 2Rk &= \sum_{i=1}^k \log\left(\sigma_i^2(\mathcal{S})/D_i^*\right) \\ &\geq \sum_{i=1}^k \sigma_i^2(\mathcal{S}) \log\left(\sigma_i^2(\mathcal{S})/D_i^*\right) \\ &\geq \alpha k \log\left(\frac{\alpha k}{\sum_i D_i^*}\right) \end{aligned} \tag{48}$$

where the first step follows from the fact that $\sigma_i^2(\mathcal{S}) \leq 1$ and the second step follows from the log-sum inequality. This inequality leads to

$$\sum_{i=1}^k D_i^* \geq \alpha k 2^{-2R/\alpha}. \tag{49}$$

Thus, for this subspace, the objective function of optimization (33) *is upper bounded by $\alpha k - \alpha k 2^{-2R/\alpha}$, which for $0 < \alpha < 1$ is always less than $k(1-2^{-2R})$, the PCA-subspace objective function. This completes the proof.*