Understanding Entropic Regularization in GANs

Daria Reshetova*, Yikun Bai†, Xiugang Wu† and Ayfer Özgür*
*Department of Electrical Engineering, Stanford University, Email: {resh,aozgur}@stanford.edu
†Department of Electrical and Computer Engineering, University of Delaware, Email: {bai,xwu}@udel.edu

***Abstract*—Generative Adversarial Networks (GANs) are a popular method for learning distributions from data by modeling the target distribution as a function of a known distribution. The function, often referred to as the generator, is optimized to minimize a chosen distance measure between the generated and target distributions. One commonly used measure for this purpose is the Wasserstein distance. However, Wasserstein distance is hard to compute and optimize, and in practice entropic regularization techniques are used to facilitate its computation and improve numerical convergence. The influence of regularization on the learned solution, however, remains not well-understood. In this paper, we study how several popular entropic regularizations of Wasserstein distance impact the solution learned by a Wasserstein GAN in a simple benchmark setting where the generator is linear and the target distribution is high-dimensional Gaussian. We show that entropy regularization of Wasserstein distance promotes sparsification of the solution, while replacing the Wasserstein distance with the Sinkhorn divergence recovers the unregularized solution. The significant benefit of both regularization techniques is that they remove the curse of dimensionality suffered by Wasserstein distance. We show that in both cases the optimal generator can be learned to accuracy ϵ with $O(1/\epsilon^2)$ samples from the target distribution without requiring to constrain the discriminator. We thus conclude that these regularization techniques can improve the quality of the generator learned from empirical data in a way that is applicable for a large class of distributions.**

I. Introduction

Generative Adversarial Networks (GANs) have become a popular framework for learning data distributions and sampling as they have achieved impressive results in various domains [1], [2], [3], [4]. As opposed to traditional methods of fitting a parametric distribution, GANs' objective is to find a mapping from a known distribution to the unknown data distribution or its empirical approximation. The mapping is set to a minimizer of a chosen distance measure between the generated and target distribution.

In the original GAN framework, the distance measure is the Jensen-Shannon divergence [5]. This measure was later replaced by the Wasserstein distance in [6] and follow-up work, which showed that Wasserstein GANs can help resolve several issues related to the original formulation, such as the lack of continuity, mode collapse [6] and vanishing gradients [7].

Despite these advantages, minimizing the Wasserstein distance between the target (data) and the generated distribution is a computationally challenging task. Indeed, computing the Wasserstein distance between two empirical distributions involves the resolution of a linear program whose cost can quickly become prohibitive whenever the size of the support of these measures or the number of samples exceeds several hundreds. A popular approach to facilitate computation of Wasserstein distance is to regularize it with an entropic term which makes the problem strongly convex and hence solvable by matrix scaling algorithms [8], [9]. More recent results have shown that this also results in faster convergence and stability of the first-order methods used for optimizing Wasserstein GANs [10].

However, the impact of these regularization methods on the generator learned by the Wasserstein GAN remains poorly understood. This is partly due to the fact that GANs are primarily evaluated on real data, typically images, and although clearly valuable, such evaluations are often subjective due to lack of clear baselines for benchmarking. In this paper, we follow the philosophy advocated in [11] and focus on a simple benchmark setting where solutions can be explicitly characterized and compared. Following their work, we assume that the generator is linear and the target distribution is high-dimensional Gaussian. [11] characterizes the population solution for the Wasserstein GAN in this setup, and show that, even in this simple setting the problem suffers from the curse of dimensionality. The empirical solution learned on n samples of the target distribution converges to the population solution as $\Omega(n^{-2/d})$, where d is the dimension of the target distribution support. To resolve this sample complexity problem, [11] proposes to restrict the discriminator to be quadratic, however this insight comes from knowing that the sought target distribution is Gaussian and hence does not generalize beyond the considered linear/Gaussian setting.

In this paper, by focusing on the linear generator and Gaussian distribution setting of [11], we explore how regularization impacts what generator is learnt and how it leads to better generalization. We study two slightly different ways of regularizing: entropic regularization [8] and Sinkhorn divergence [12]. We show that the former introduces bias to the solution by constraining nuclear norm of the covariance matrix of generator's output distribution, while Sinkhorn divergence results in the same solution as the unregularized Wasserstein GAN in [11]. We then show, in the more general case of sub-gaussian distributions and Lipschitz generators, that these regularizations result in sample complexity of $O_d(1/\sqrt{n})$, thus overcoming the curse of dimensionality in [11] without explicit constraints on the discriminator. This indicates that adding regularization implicitly constrains the discriminator in a way suitable for a large class of distributions.

II. Preliminaries

In this section, we provide some background on optimal transport and optimal transport GANs.

Let $\mathcal{P}(\mathcal{X})$ be the set of all probability measures with support $\mathcal{X}$ and finite second moments. For $P_Z \in \mathcal{P}(\mathcal{Z})$ and $P_Y \in \mathcal{P}(\mathcal{Y})$, denote by $\Pi(P_Z, P_Y)$ the set of all couplings of P_Z and P_Y, i.e. all joint probability measures from $\mathcal{P}(\mathcal{Z}\times\mathcal{Y})$ with marginal distributions being P_Z and P_Y. Squared Wasserstein distance between $P_Z, P_Y \in \mathcal{P}(\mathbb{R}^d)$ under ℓ_2 metric, or simply the *squared 2-Wasserstein distance*, is defined as

$$W_2^2(P_Z, P_Y) = \inf_{\pi\in\Pi(P_Z,P_Y)} \mathbb{E}_\pi \left[\|Z - Y\|^2\right]. \quad (1)$$

Using 2-Wasserstein distance to measure the dissimilarity between the generated and target distributions leads to the following learning problem of GAN, referred to as *W2GAN*:

$$\min_{G\in\mathcal{G}} W_2^2\left(P_{G(X)}, P_Y\right). \quad (2)$$

Here X is a latent random variable, $G(\cdot)$ is a generator that comes from a set of functions $\mathcal{G} \subseteq \{G : \mathcal{X} \to \mathcal{Y}\}$, and P_Y is the target probability measure, which could be, e.g., either the true distribution of Y in the population case, or the empirical distribution of Y when one has access to only a finite sample $\{Y_i\}_{i=1}^n$. A remarkable feature of the Wasserstein distance is that strong duality holds [13], [14] for (1). Thus (1) is equivalent to maximizing a concave objective over a set of functions, discriminators, instead of optimizing over couplings as in the primal form (1). This naturally leads to the min-max game formulation of GAN, where the generator seeks to generate samples that are close to the real data training samples, and it competes with a discriminator that seeks to distinguish between real and generated samples.

In practice, the Wasserstein distance is often regularized to facilitate its computation leading to the *entropy regularized 2-Wasserstein distance* [8]:

$$W_{2,\lambda}^2(P_Z, P_Y) = \inf_{\pi\in\Pi(P_Z,P_Y)} \mathbb{E}_\pi \left[\|Z - Y\|^2\right] + \lambda I_\pi(Z;Y) \quad (3)$$

where the regularization term is the mutual information $I_\pi(Z;Y)$ calculated according to the the joint distribution π. The corresponding *entropic W2GAN* is defined as

$$\min_{G\in\mathcal{G}} W_{2,\lambda}^2\left(P_{G(X)}, P_Y\right). \quad (4)$$

While the entropic Wasserstein distance allows for faster computation, note that it can be strictly larger than zero even if the generated distribution is exactly the same as the target distribution, i.e. $W_{2,\lambda}^2(P_Y, P_Y) \neq 0$. This issue can be resolved by adding corrective terms to (3) [12], which leads to the Sinkhorn divergence:

$$S_\lambda(P_{G(X)}, P_Y) = W_{2,\lambda}^2(P_{G(X)}, P_Y) - \left(W_{2,\lambda}^2(P_{G(X)}, P_{G(X)}) + W_{2,\lambda}^2(P_Y, P_Y)\right)/2. \quad (5)$$

One can easily check that $S_\lambda(P_Y, P_Y) = 0$ for any P_Y. The corresponding *Sinkhorn W2GAN* is given by:

$$\min_{G\in\mathcal{G}} S_\lambda(P_{G(X)}, P_Y). \quad (6)$$

We would like to emphasize that strong duality holds for (3) and (5), and they can be reformulated as maximizing a strongly concave objective over a set of functions from $\mathcal{Y}$ to $\mathbb{R}$ (discriminators) [15].

III. Population Solution For the Linear/Gaussian Setting

In this section, we focus on the benchmark setting considered in [11], where the generator is linear and the target distribution is Gaussian, in which case we can rewrite the general formulation of (2) as:

$$\min_{G\in\mathbb{R}^{d\times r}} W_2^2\left(P_{GX}, P_Y\right),$$

where the latent random variable $X \in \mathbb{R}^r$ follows the standard Gaussian distribution $\mathcal{N}(0, I_r)$, the underlying distribution of data $Y \in \mathbb{R}^d$ is $\mathcal{N}(0, K_Y)$, and the optimization is over all matrices $G \in \mathbb{R}^{d\times r}$ with $d \geq r$ so that the generated distribution is P_{GX}. The population solution to the above W2GAN problem has been characterized in [11] as the r-PCA solution of Y, i.e. the covariance matrix K_{G^*X} for P_{G^*X} is a rank-r matrix whose top r eigenvalues and eigenvectors are the same as those of K_Y.

We next show that adding entropic regularization to the W2GAN objective changes this solution to a soft-thresholded r-PCA solution of Y as shown by the following theorem.

***Theorem** 1:* Let $Y \sim \mathcal{N}(0, K_Y)$ and $X \sim \mathcal{N}(0, I_r)$ where $r \leq d$. The population solution P_{G^*X} to the entropic W2GAN problem is given by a soft-thresholded r-PCA solution of Y, i.e., the covariance matrix K_{G^*X} for P_{G^*X} is a rank-r matrix whose top r eigenvectors are the same as those of K_Y and the top r eigenvalues are $\sigma_i^2 = (\lambda_i(K_Y) - \lambda/2)_+$ for $i \in [1:r]$, where $(x)_+ := \max\{x, 0\}$ and $\{\lambda_i(K_Y)\}_{i=1}^r$ are the top r eigenvalues of K_Y.

This theorem connects entropic W2GAN to a version of PCA with soft thresholding of singular values, which is the solution for the matrix completion problem [16, Theorem 2.1]:

$$\min_{G\in\mathbb{R}^{d\times r}} \|K_Z - K_Y\|_F^2 + \lambda\|K_Z\|_*,$$

where $K_Z = GG^T$ corresponds to the covariance matrix of the generated distribution P_{GX} in the GAN problem, and $\|\cdot\|_*$ is the nuclear norm, i.e. the sum of all singular values of a matrix. Thus, entropic regularization promotes sparsity in the singular values of the covariance matrix of the generated distribution.

Note that the population solution for the entropic W2GAN is not the same as that for the unregularized W2GAN, which is not surprising as they optimize two different objective functions. Nevertheless, Theorem 1 reveals that in the linear/Gaussian case, there is a natural relationship between the two solutions as the former turns out to be a soft-thresholded version of the latter. We next investigate the population solution for the Sinkhorn W2GAN and show that, while it is not the case in general, when restricted to the linear/Gaussian benchmark, surprisingly Sinkhorn W2GAN does recover the regular PCA solution as shown in the following theorem. We remark that this is not ensured by the property $S_\lambda(P_Y, P_Y) = 0$ for any P_Y of the Sinkhorn divergence, as in the current setting the Sinkhorn divergence between the optimal generated and target distributions is non-zero. However, it does suggest that the Sinkhorn divergence can lead to solutions closer to the target distribution, while also

possessing other favorable qualities like unbiasedness and the one described in the following section.

***Theorem** 2:* Let $Y \sim \mathcal{N}(0, K_Y)$ and $X \sim \mathcal{N}(0, I_r)$ where $r \leq d$. The population solution P_{G^*X} to the Sinkhorn W2GAN problem is given by the r-PCA solution of Y.

A. Proofs of Theorems 1 and 2

Proof of Theorem 1: Let $Z = GX$, where $G \in \mathbb{R}^{d\times r}$. Since $X \sim \mathcal{N}(0, I_r)$, P_Z is a d-dimensional Gaussian distribution whose covariance matrix K_Z has rank less than or equal to r. For any such P_Z, denote by $\mathcal{S}_Z$ the r-dimensional subspace that contains the support of Z. For any $Y \in \mathbb{R}^d$, let $Y_{\mathcal{S}_Z}$ and $Y_{\mathcal{S}_Z^\perp}$ be respectively the projections of Y onto $\mathcal{S}_Z$ and its orthogonal complement $\mathcal{S}_Z^\perp$ so that $Y = Y_{\mathcal{S}_Z} + Y_{\mathcal{S}_Z^\perp}$. The entropy regularized 2-Wasserstein distance is then

$$\begin{aligned}
&W_{2,\lambda}^2(P_Y, P_Z) = \min_{\pi\in\Pi(P_Y,P_Z)} \mathbb{E}_\pi\big[\|Z - Y\|^2\big] + \lambda I_\pi(Z;Y)\\
&= \min_{\pi\in\Pi(P_Y,P_Z)} \mathbb{E}_\pi\big[\|(Z - Y_{\mathcal{S}_Z}) - Y_{\mathcal{S}_Z^\perp}\|^2\big] + \lambda I_\pi(Z;Y)\\
&= \min_{\pi\in\Pi(P_Y,P_Z)} \mathbb{E}_\pi\big[\|Z - Y_{\mathcal{S}_Z}\|^2\big] + \mathbb{E}\big[Y_{\mathcal{S}_Z^\perp}\|^2\big] + \lambda I_\pi(Z;Y) \quad (7)\\
&= \min_{\pi\in\Pi(P_Y,P_Z)} \mathbb{E}_\pi\big[\|Z - Y_{\mathcal{S}_Z}\|^2\big] + \mathbb{E}\big[Y_{\mathcal{S}_Z^\perp}\|^2\big] + \lambda I_\pi(Z;Y_{\mathcal{S}_Z}) \quad (8)
\end{aligned}$$

The last equality above holds because the optimal coupling should make $Z - Y_{\mathcal{S}_Z} - Y_{\mathcal{S}_Z^\perp}$ a Markov chain, in which case $I_\pi(Z;Y) = I_\pi(Z;Y_{\mathcal{S}_Z}, Y_{\mathcal{S}_Z^\perp}) = I_\pi(Z;Y_{\mathcal{S}_Z})$; indeed, for any coupling π, one can construct π' such that $\pi'(Z, Y_{\mathcal{S}_Z}, Y_{\mathcal{S}_Z^\perp}) = \pi(Z, Y_{\mathcal{S}_Z})\pi(Y_{\mathcal{S}_Z^\perp}|Y_{\mathcal{S}_Z})$, and π' preserves the values of the first two terms in (7) while decreasing the value of the third term.

Consider the optimization problem in the entropic W2GAN, i.e. $\min_{P_Z\in\mathcal{N}_{d,r}} W_{2,\lambda}^2(P_Y, P_Z)$, where the optimization is over the set $\mathcal{N}_{d,r}$ of all d-dimensional Gaussian distributions with rank not exceeding r. In light of (8), the above is

$$\min_{\substack{\mathcal{S}\in\mathbb{S}_d:\dim(\mathcal{S})\leq r\\ P_Z\in\mathcal{N}_{d,r}:\, Z\in\mathcal{S}\\ \pi\in\Pi(P_Y,P_Z)}} \mathbb{E}_\pi[\|Z - Y_{\mathcal{S}}\|^2] + \mathbb{E}[\|Y_{\mathcal{S}^\perp}\|^2] + \lambda I_\pi(Z;Y_{\mathcal{S}}), \quad (9)$$

where $\mathbb{S}_d$ is the set of all subspaces of $\mathbb{R}^d$. To solve (9) we first fix $\mathcal{S}$. If columns of $U \in \mathbb{R}^{d\times r}$ form an orthonormal basis of $\mathcal{S}$, i.e. $\mathcal{S} = \operatorname{Im} U$ and $U^TU = I_r$, we replace Z and $Y_{\mathcal{S}}$ in (9) by U^TZ and U^TY respectively. To find optimal π, P_Z for $\mathcal{S}$ we then solve

$$\begin{aligned}
\min_{\substack{P_Z\in\mathcal{N}_{d,r}:\, Z\in\operatorname{Im}(U)\\ \pi\in\Pi(P_Z,P_Y)}} &\mathbb{E}_\pi[\|U^TZ - U^TY\|^2] + \lambda I_\pi(U^TZ;U^TY)\\
&- \mathbb{E}[\|U^TY\|^2] + \mathbb{E}[\|Y\|^2] \quad (10)
\end{aligned}$$

Let $\bar{Z} = U^TZ$ and $\bar{Y} = U^TY$, and let $\mathcal{N}_{r,r}$ be the set of all r-dimensional Gaussian distributions. Then Problem (10) is

$$\min_{P_{\bar{Z}}\in\mathcal{N}_{r,r}} \min_{\pi\in\Pi(P_{\bar{Z}},P_{\bar{Y}})} \mathbb{E}_\pi[\|\bar{Z} - \bar{Y}\|^2] + \lambda I_\pi(\bar{Z};\bar{Y}) \quad (11)$$

For fixed $P_{\bar{Z}}, P_{\bar{Y}}$ and cross-covariance matrix $K_{\bar{Z}\bar{Y}}$ the first term in (12) is fixed and the mutual information term is minimized when the π is jointly Gaussian. Therefore, (11) is simply a rate distortion problem with source distribution $P_{\bar{Y}}$ being Gaussian, i.e.

$$\min_{\pi\in\mathcal{N}(P_{\bar{Z}},P_{\bar{Y}})} \mathbb{E}_\pi[\|\bar{Z} - \bar{Y}\|^2] + \lambda I_\pi(\bar{Z};\bar{Y}) \quad (12)$$

WLOG, we can assume $\bar{Y}$ has diagonal covariance matrix $K_{\bar{Y}} = \operatorname{diag}(\Lambda_1, \ldots, \Lambda_r)$, where the diagonal elements are in decreasing order. The solution for this problem is given by reverse waterfilling [17], under which the optimal $P_{\bar{Z}}$ has covariance matrix $K_{\bar{Z}} = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_r^2)$ where $\sigma_i^2 = (\Lambda_i - \frac{\lambda}{2})_+$, and the optimal value is given by

$$\sum_{i=1}^r \left(\frac{\lambda}{2}\ln\frac{\max\{\Lambda_i, \lambda/2\}}{\lambda/2} - \max\left\{\Lambda_i, \frac{\lambda}{2}\right\}\right) + \frac{r\lambda}{2} + \mathbb{E}[\|\bar{Y}\|^2]$$

The entropic W2GAN optimization problem (9) is then equivalent to:

$$\min_{U\in\mathbb{R}^{d\times r}} \sum_{i=1}^r \left(\frac{\lambda}{2}\ln\frac{\max\{\Lambda_i, \lambda/2\}}{\lambda/2} - \max\{\Lambda_i, \lambda/2\}\right)$$

where the optimization is over all $U \in \mathbb{R}^{d\times r}$ such that $U^TU = I_r$ and $U^TK_YU = \operatorname{diag}(\Lambda_i, \ldots, \Lambda_r)$. We now let

$$f(\Lambda_i) = (\lambda/2)\ln\left(\max\{\Lambda_i, \lambda/2\}/\left(\lambda/2\right)\right) - \max\{\Lambda_i, \lambda/2\},$$

and complete the proof by showing

$$[\lambda_1(K_Y), \ldots, \lambda_r(K_Y)] = \arg\min_{[\Lambda_1,\ldots,\Lambda_r]} \sum_{i=1}^r f(\Lambda_i) \quad (13)$$

Indeed, for each U we have $\sum_{i=1}^{r'} \Lambda_i \leq \sum_{i=1}^{r'} \lambda_i, \forall r' \leq r$ (all the sequences are in decreasing order). Using the majorizing inequality (see, e.g., Lemma 2.2 of [18]) and the fact that f is concave non-increasing function, we have $\sum_{i=1}^r f(\lambda_i(K_Y)) \leq \sum_{i=1}^r f(\Lambda_i)$. Therefore, columns of the optimal U are the top r eigenvectors of K_Y, and the optimal P_Z has covariance matrix given by $K_Z = U[\operatorname{diag}(\sigma_1^2, \ldots, \sigma_r^2)|0_{r\times(d-r)}]U^T$ where $\sigma_i^2 = (\lambda_i(K_Y) - \lambda/2)_+$. ∎

Proof of Theorem 2: From (8) in the proof of Theorem 1, we have for given $Z = GX$ and $\mathcal{S} = \operatorname{Im} G$,

$$\begin{aligned}
&W_{2,\lambda}^2(P_Z, P_Y) - \mathbb{E}[\|Y_{\mathcal{S}^\perp}\|^2]\\
&= \min_{\pi\in\Pi(P_Z,P_{Y_{\mathcal{S}}})} \mathbb{E}[\|Z - Y_{\mathcal{S}}\|^2] + \lambda I(Z;Y_{\mathcal{S}}) = W_{2,\lambda}^2(P_Z, P_{Y_{\mathcal{S}}}),
\end{aligned}$$

and therefore for the Sinkhorn divergence,

$$\begin{aligned}
&S_\lambda(P_Z, P_Y) - \mathbb{E}\|Y_{\mathcal{S}^\perp}\|_2^2\\
&= W_{2,\lambda}^2(P_Z, P_{Y_{\mathcal{S}}}) - \big(W_{2,\lambda}^2(P_Z, P_Z) + W_{2,\lambda}^2(P_Y, P_Y)\big)/2\\
&= S_\lambda(P_Z, P_{Y_{\mathcal{S}}}) + \big(W_{2,\lambda}^2(P_{Y_{\mathcal{S}}}, P_{Y_{\mathcal{S}}}) - W_{2,\lambda}^2(P_Y, P_Y)\big)/2 \quad (14)
\end{aligned}$$

Consider the optimization problem in the Sinkhorn divergence GAN, i.e. $\min_{P_Z} S_\lambda(P_Z, P_Y)$. In light of (14), given $Z \in \mathcal{S}$ the optimal P_Z should be $P_{Y_{\mathcal{S}}}$, which makes the first term in (14) zero. Therefore, it only remains to optimize over $\mathcal{S}$, and in particular, the problem reduces to

$$\min_{\mathcal{S}\in\mathbb{S}_d:\dim(\mathcal{S})\leq r} W_{2,\lambda}^2(P_{Y_{\mathcal{S}}}, P_{Y_{\mathcal{S}}})/2 + \mathbb{E}\|Y_{\mathcal{S}^\perp}\|_2^2.$$

Using the formula [19, Theorem 1] for entropy regularized 2-

Wasserstein distance between two Gaussian distributions, the objective function in the above problem becomes

$$\begin{aligned}
&W_{2,\lambda}^2(P_{Y_{\mathcal{S}}}, P_{Y_{\mathcal{S}}})/2 + \mathbb{E}\|Y_{\mathcal{S}^\perp}\|_2^2 \\
&\quad = \operatorname{Tr} K_{Y_{\mathcal{S}^\perp}} + \operatorname{Tr} K_{Y_{\mathcal{S}}} - \operatorname{Tr}\big((4K_{Y_{\mathcal{S}}}^2 + \lambda^2 I/4)^{1/2}\big)/2 \\
&\qquad + \lambda \log\det\big(\big(4K_{Y_{\mathcal{S}}}^2 + \lambda^2 I/4\big)^{1/2} + \lambda I/2\big)/4 + C \\
&\quad = \sum_{i=1}^{r}\left(\frac{\lambda}{4}\log\left(\sqrt{4\Lambda_i^2 + \frac{\lambda^2}{4}} + \frac{\lambda}{2}\right) - \frac{1}{2}\sqrt{4\Lambda_i^2 + \frac{\lambda^2}{4}}\right) + C'
\end{aligned}$$

where Λ_i is the ith eigenvalue of $U^T K_{Y_{\mathcal{S}}} U$ for some $U \in \mathbb{R}^{d\times r}$ such that $\operatorname{Im} U = \mathcal{S}$ and $U^T U = I_r$, C is a constant and $C' = \operatorname{Tr} K + C$. The above is minimized when $\Lambda_i = \lambda_i(K_Y)$ using the similar argument for showing (13), i.e., by using the majorizing inequality and noting that the function $f(x) = \sqrt{4x^2 + \lambda^2/4}/2 - \lambda\log(\sqrt{4x^2+\lambda^2/4} + \lambda/2)/4$ is concave and non-increasing for $\lambda > 0$ and $x \geq 0$. ∎

IV. Generalization Error of Empirical Solution

In this section we discuss the generalization capability of the empirical solutions for W2GAN, entropic W2GAN and Sinkhorn W2GAN, respectively. Note that in the population case, the underlying distribution of data P_Y was known in the GAN formulations (2), (4) and (6). In contrast, here we consider the finite sample case, where empirical distribution Q_Y^n extracted from sample $\hat{\mathcal{Y}} = \{y_i\}_{i=1}^n$ is used in the GAN objective (2), (4) and (6) to approximate P_Y. We are interested in how fast the empirical solution $P_{G_n(X)}$ converges to the population solution $P_{G^*(X)}$.

It was shown in [11] that even in our simple benchmark when generators are linear and data distribution is high-dimensional Gaussian, the convergence for W2GAN is slow in the sense that the generalization error

$$\mathbb{E}\big[W_2^2(P_{G_n(X)}, P_Y) - W_2^2(P_{G^*(X)}, P_Y)\big] = \Omega(n^{-2/d}).$$

That is to decrease the generalization error by a constant factor the number of samples has to be increased by a factor of $e^{\Omega(d)}$, and hence the generalization capability of W2GAN suffers from the curse of dimensionality. To overcome this, [11] proposed to constrain the set of discriminators for W2GAN to quadratic. This was motivated by the observation that constraining the discriminator to be quadratic will not affect the population solution because the optimal discriminator for W2GAN is indeed quadratic in the Gaussian setting. On the other hand, it was shown that this constraint will lead to fast convergence of order $O_d(n^{-1/2})$ and hence resolve the issue of curse of dimensionality.

While constraining the discriminator to be quadratic as done in [10] is conceptually appealing and works for the setup of linear generators and Gaussian data, this insight does not generalize to other distributions, i.e. for non-Gaussian data the generator obtained under a quadratic discriminator is not necessarily the one minimizing the 2-Wasserstein distance between the generated and the target distributions. Theorems 3 and 4 below show that under mild conditions on the underlying distribution of data, the latent random variable and the set of generators, similar convergence can be achieved for entropic W2GAN and Sinkhorn W2GAN without the need to constrain the discriminator.

To formally state the results, let us first recall some definitions. A distribution P_X is σ^2 sub-gaussian for $\sigma \geq 0$ if $\mathbb{E}\exp\big(\|X\|^2/(2r\sigma^2)\big) \leq 2$. Let $\sigma^2(X) = \min\{\sigma \geq 0 \big| \mathbb{E}\exp(\|X\|^2/(2r\sigma^2)) \leq 2\}$, and let $\sigma_{\hat{\mathcal{Y}}}^2(Z) = \min\{\sigma \geq 0 \,\big|\, \mathbb{E}_{\hat{\mathcal{Y}}}\exp(\|Z\|^2/(2r\sigma^2)) \leq 2\}$ be the sub-gaussian parameter of the distribution of Z conditioned on the sample, where $\mathbb{E}_{\hat{\mathcal{Y}}}[\cdot]$ denotes the expectation conditioned on the sample. A set of generators $\mathcal{G}$ is said to be star-shaped with a center at 0 if a line segment between 0 and $G \in \mathcal{G}$ also lies in $\mathcal{G}$, i.e.

$$G \in \mathcal{G} \;\Rightarrow \alpha G \in \mathcal{G}, \forall \alpha \in [0,1]. \tag{15}$$

Note that this includes the set of all linear generators considered in the last section as a trivial case, as well as the set of linear functions with a bounded norm or a fixed dimension.

***Theorem** 3:* Let $P_{K_X^{-1/2}X}$ and P_Y be sub-gaussian and the generator set $\mathcal{G}$ be a set of linear function satisfying condition (15). Then the generalization error for entropic W2GAN can be bounded by

$$\begin{aligned}
&\mathbb{E}\big[W_{2,\lambda}^2(P_{G_n(X)}, P_Y) - W_{2,\lambda}^2(P_{G^*(X)}, P_Y)\big] \\
&\qquad\qquad \leq K_d \lambda n^{-1/2}\big(1 + (2\tau^2/\lambda)^{\lceil 5d/4\rceil + 3}\big),
\end{aligned}$$

where $\tau^2 = \max\{\sigma^2(K_X^{-1/2}X)\sigma^2(Y), \sigma^2(Y)\}$ and K_d is a dimension dependent constant.

Theorem 3 essentially says that under certain mild conditions, the generalization error for entropic W2GAN converges to zero at speed $O_d(1/\sqrt{n})$. This improved sample complexity suggests that the set of possible discriminators may be implicitly constrained due to the entropic regularization term used in the primal form of entropy regularized 2-Wasserstein distance. Similar results also hold for the set of Lipschitz functions and extend to the Sinkhorn distance W2GAN.

***Theorem** 4:* Let P_X and P_Y be sub-Gaussian and the set of generators $\mathcal{G}$ consist of L-Lipschitz functions, i.e. $\|G(X_1) - G(X_2)\| \leq L\|X_1 - X_2\|$ for any X_1, X_2 in the support of P_X and let $\mathcal{G}$ satisfy (15). Then the generalization error for entropic W2GAN

$$\mathbb{E}\big[W_{2,\lambda}^2(P_{G_n(X)}, P_Y) - W_{2,\lambda}^2(P_{G^*(X)}, P_Y)\big]$$

and that for Sinkhorn W2GAN

$$\mathbb{E}\big[S_\lambda(P_{G_n(X)}, P_Y) - S_\lambda(P_{G^*(X)}, P_Y)\big]$$

can be both upper bounded by

$$K_d \lambda n^{-1/2}\big(1 + (2\tau^2/\lambda)^{\lceil 5d/4\rceil + 3}\big) \tag{16}$$

with $\tau^2 = \max\{L^2\sigma^2(X), \sigma^2(Y)\}$.

It is worth mentioning that a similar result was proved in [20], however it requires significantly stronger conditions for the set of generator functions $\mathcal{G}$. In particular, it does not apply to $\mathcal{G}$ being the set of all linear functions.

A. Proofs

Due to page limit, we only provide the proof of Theorem 3. The proof of Theorem 4 follows along the similar line and is delegated to the long version of the paper. In particular, the

proof of Theorem 3 builds on a result that appears as Corollary 1 of [21] and several lemmas that we summarize below.

***Proposition** 1 (Corollary 1 of [21]):* If P_X and P_Y are σ^2 sub-gaussian, then

$$\mathbb{E}\left[\left|W^2_{2,\lambda}(P_X, Q^n_Y) - W^2_{2,\lambda}(P_X, P_Y)\right|\right] \leq K_d \lambda n^{-1/2}\left(1 + (2\sigma^2/\lambda)^{\lceil 5d/4\rceil + 3}\right), \quad (17)$$

where K_d is a constant depending on the dimension.

To prove the theorem we need the following lemmas.

***Lemma** 1:* Under the conditions of Theorem 3 for any P_Z $G^* = \arg\min_{G\in\mathcal{G}} W^2_{2,\lambda}(P_{G(X)}, P_Z)$ has $\mathbb{E}\left[\|G^*(X)\|_2^2\right] \leq \operatorname{Tr} K_Z$ and $\sigma^2(G^*(X)) \leq rd^{-1}\operatorname{Tr} K_Z \sigma^2(K_X^{-1/2}X)$.

***Lemma** 2:* For a sub-gaussian Z the covariance matrix trace is bounded as $\operatorname{Tr} K_Z \leq 4d\sigma^2(Z)$.

Lemma 1 follows from the optimality of G^*, condition (15) and the fact that mutual information is invariant to scaling of G. Lemma 2 follows directly from Jensen's inequality.

Proof of Theorem 3: The proof is based on [21, Theorem 2]. Denote $C_{d,i}$ constants depending on the dimension d as we are not aiming to find the exact dependence of the bound from the dimension. Let $\lambda = 2$ and note that the definition of entropy regularized Wasserstein distance defined in [21] and in this paper differ by a factor of $1/2$, but as all the results are stated up to a multiplicative constant, this does not influence the solution. First, we rewrite $d_\lambda(G^*, G_n)$ to fit Theorem 1:

$$\begin{aligned} d_\lambda(G^*, G_n) &= \left(W^2_{2,\lambda}(P_{G^*(X)}, Q^n_Y) - W^2_{2,\lambda}(P_{G^*(X)}, P_Y)\right) \\ &\quad + \left(W^2_{2,\lambda}(P_{G_n(X)}, P_Y) - W^2_{2,\lambda}(P_{G^*(X)}, Q^n_Y)\right) \\ &\leq \left(W^2_{2,\lambda}(P_{G^*(X)}, Q^n_Y) - W^2_{2,\lambda}(P_{G^*(X)}, P_Y)\right) \\ &\quad + \left(W^2_{2,\lambda}(P_{G_n(X)}, P_Y) - W^2_{2,\lambda}(P_{G_n(X)}, Q^n_Y)\right) \quad (18) \end{aligned}$$

Let $\nu^2 = \max\{4r\sigma^2(K_X^{-1/2}X)\sigma^2(Y), \sigma^2(Y)\} \leq 4r\tau^2$. Then

$$\begin{aligned} \sigma^2\left(G^*(X)\right) &\leq rd^{-1}\operatorname{Tr} K_Y \sigma^2\left(K_X^{-1/2}X\right) \\ &\leq 4r\sigma^2\left(K_X^{-1/2}X\right)\sigma^2(Y) \leq \nu^2, \end{aligned}$$

with the inequalities following from Lemmas 1, 2 and the definition of ν^2. By [21, Theorem 2] applied to the expectation of the first difference in (18):

$$\mathbb{E}\left[\left|W^2_{2,\lambda}(P_{G^*(X)}, Q^n_Y) - W^2_{2,\lambda}(P_{G^*(X)}, P_Y)\right|\right] \leq C_{d,2} n^{-1/2}\left(1 + \left(\nu^2\right)^{\lceil 5d/4\rceil + 3}\right), \quad (19)$$

As G_n depends on the sample, the theorem cannot be applied directly to the second difference. Following [21] for a set of functions $\mathcal{F}$ we write $\|P_Y - Q^n_Y\|_{\mathcal{F}} = \sup_{u(\cdot)\in\mathcal{F}^s}\left|\int u(y)\left(dP_Y(y) - dQ^n_Y(y)\right)\right|$. By Lemma [21, Proposition 1] for $\tilde{\sigma}^2 = \max\left\{\sigma^2_{\hat{\mathcal{Y}}}\left(G_n(X)\right), \sigma^2_{\hat{\mathcal{Y}}}(\hat{Y}), \sigma^2(Y)\right\}$ and $s = \lceil d/2\rceil + 1$:

$$\begin{aligned} &W^2_{2,\lambda}(P_{G_n(X)}, P_Y) - W^2_{2,\lambda}(P_{G_n(X)}, Q^n_Y) \\ &\leq \left(1 + \tilde{\sigma}^{3s}\right)\|P_Y - Q^n_Y\|_{\mathcal{F}^s}, . \quad (20) \end{aligned}$$

where $\mathcal{F}^s$ is a set of functions, such that $\psi/\left(1 + \tilde{\sigma}^{3s}\right) \in \mathcal{F}^s$ for all optimal dual potentials ψ. $\mathcal{F}^s$ is a larger set and its exact definition can be found in [21], but for the purpose of this proof it is important to note that $\mathcal{F}^s$ only depends on s and not on the sub-gaussian parameters of Y and GX. Taking expectation over the sample in (20) we get:

$$\begin{aligned} &\left(\mathbb{E}_{\hat{\mathcal{Y}}}\left[W^2_{2,\lambda}(P_{G_n(X)}, P_Y) - W^2_{2,\lambda}(P_{G_n(X)}, Q^n_Y)\right]\right)^2 \\ &\leq 2\mathbb{E}_{\hat{\mathcal{Y}}}\left[1 + \tilde{\sigma}^{6s}\right]\mathbb{E}_{\hat{\mathcal{Y}}}\|P_Y - Q^n_Y\|^2_{\mathcal{F}^s} \quad (21) \end{aligned}$$

In [21, Proof of Theorem 2] a covering number bound for $\mathcal{F}^s$ is used to establish that

$$\begin{aligned} &\mathbb{E}\left[\|P_Y - Q^n_Y\|^2_{\mathcal{F}^s}\right] \\ &\leq \left(1 + \sigma^2(Y)^{d+2}\right) n^{-1} C_{d,3} \leq \left(1 + \nu^{2d+4}\right) n^{-1} C_{d,3} \end{aligned}$$

By Lemma 2 we have $\operatorname{Tr} K_{\hat{Y}} \leq 4d\sigma^2(\hat{Y})$, so

$$\begin{aligned} \sigma^2_{\hat{\mathcal{Y}}}(G_n(X)) &\leq d^{-1}\operatorname{Tr} K_{\hat{Y}} r\sigma^2\left(K_X^{-1/2}X\right) \\ &\leq 4r\sigma^2\left(K_X^{-1/2}X\right)\sigma^2_{\hat{\mathcal{Y}}}\left(\hat{Y}\right) \leq \sigma^2_{\hat{\mathcal{Y}}}(\hat{Y})\nu^2/\sigma^2(Y), \quad (22) \end{aligned}$$

where the last is due to Lemma 1. Taking expectation of $\tilde{\sigma}^{6s}$:

$$\begin{aligned} &\mathbb{E}\left[\tilde{\sigma}^{6s}\right] \leq \mathbb{E}\left[\max\left\{\sigma^2_{\hat{\mathcal{Y}}}\left(\hat{Y}\right), \sigma^2(Y), \sigma^2_{\hat{\mathcal{Y}}}\left(G_n(X)\right)\right\}^{3s}\right] \\ &\leq \nu^{6s}\mathbb{E}\left[\max\left\{1, \sigma^2_{\hat{\mathcal{Y}}}\left(\hat{Y}\right)/\sigma^2(Y)\right\}^{3s}\right] \leq 2(3s)^{3s}\nu^{6s}, \quad (23) \end{aligned}$$

where (23) is due [21, Lemma 4]; plugging (23) in (21) gives

$$\begin{aligned} &\mathbb{E}\left[W^2_{2,\lambda}(P_{G_n(X)}, P_Y) - W^2_{2,\lambda}(P_{G_n(X)}, Q^n_Y)\right] \\ &\leq \sqrt{2(1 + 2(3s)^{3s}\nu^{6s})C_{d,3} n^{-1}\nu^{6s}\left(1 + \nu^{d+2}\right)} \\ &\leq C_{d,4} n^{-1/2}\left(1 + \left(\nu^2\right)^{\lceil 5d/4\rceil + 3}\right) \quad (24) \end{aligned}$$

Combining (24) and (19) we get that for $\lambda = 2$:

$$\begin{aligned} \mathbb{E}\left[d_\lambda(G^*, G_n)\right] &\leq C_{d,5} n^{-1/2}(1 + (\nu^2)^{\lceil 5d/4\rceil + 3}) \\ &\leq K_d n^{-1/2}(1 + (\tau^2)^{\lceil 5d/4\rceil + 3}), \end{aligned}$$

Consider $\lambda \neq 2$. Then for any $\lambda > 0$:

$$\begin{aligned} &W^2_{2,2}(P_{Z\sqrt{2/\lambda}}, P_{Y\sqrt{2/\lambda}}) \\ &= \inf_{\pi\in\Pi((P_Z, P_Y))} 2\mathbb{E}\left[\|Z - Y\|^2\right]/\lambda + 2I(Z;Y) \\ &= 2W^2_{2,\lambda}(P_Z, P_Y)/\lambda \end{aligned}$$

Thus, noting that for a sub-gaussian Z :

$$\mathbb{E}\exp\left(\frac{\|Z\sqrt{2/\lambda}\|_2^2}{2r\sigma_Z^2 2/\lambda}\right) = \mathbb{E}\exp\left(\frac{\|Z\|_2^2}{2r\sigma_Z^2}\right) \leq 2$$

we conclude that $\sigma^2(Z\sqrt{\lambda/2}) = 2\sigma^2(Z)/\lambda$. Plugging the result into the bound (24) we get

$$\mathbb{E}\left[d_\lambda(G^*, G_n)\right] \leq K_d \lambda n^{-1/2}\left(1 + (2\tau^2/\lambda)^{\lceil 5d/4\rceil + 3}\right)/2.$$

■

V. Conclusion

In this work we provide a comprehensive complexity analysis of entropy regularized GANs and explain their robustness. Moreover, in a specific simplified setting, the linear generator and Gaussian distributions, we derive an analytic expression for the optimal generator. This results motivates further studies on model-based designing of GANs and GANs stability.

Acknowledgment

This work was partly supported by a Stanford Graduate Fellowship, NSF award CCF-1704624, and the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370.

REFERENCES

[1] 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. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[2] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[3] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069. PMLR, 2016.

[4] Abir De, Isabel Valera, Niloy Ganguly, Sourangshu Bhattacharya, and Manuel Gomez Rodriguez. Learning and forecasting opinion dynamics in social networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 397–405. Curran Associates, Inc., 2016.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680. Curran Associates, Inc., 2014.

[6] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[7] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5769–5779, 2017.

[8] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013.

[9] Yogesh Balaji, Hamed Hassani, Rama Chellappa, and Soheil Feizi. Entropic gans meet vaes: A statistical approach to compute sample likelihoods in gans. In *ICML*, 2019.

[10] Maziar Sanjabi, Jimmy Ba, Meisam Razaviyayn, and Jason D Lee. On the convergence and robustness of training gans with regularized optimal transport. In *Advances in Neural Information Processing Systems*, pages 7091–7101, 2018.

[11] Soheil Feizi, Farzan Farnia, Tony Ginart, and David Tse. Understanding gans: the lqg setting. *arXiv preprint arXiv:1710.10793*, 2017.

[12] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR, 2018.

[13] Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Advances in neural information processing systems*, pages 3440–3448, 2016.

[14] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

[15] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690. PMLR, 2019.

[16] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 20(4):1956–1982, 2010.

[17] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[18] Vesna Andova and Mirko Petrusevski. Variable zagreb indices and karamatas inequality. *MATCH Commun. Math. Comput. Chem*, 65:685–690, 2011.

[19] Hicham Janati, Boris Muzellec, Gabriel Peyré, and Marco Cuturi. Entropic optimal transport between unbalanced gaussian measures has a closed form. *Advances in Neural Information Processing Systems*, 33, 2020.

[20] Giulia Luise, Massimiliano Pontil, and Carlo Ciliberto. Generalization properties of optimal transport gans with latent distribution learning. *arXiv preprint arXiv:2007.14641*, 2020.

[21] Gonzalo Mena and Jonathan Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems*, pages 4541–4551, 2019.