Journal of Machine Learning Research 25 (2024) 1-32 Submitted 11/21; Revised 9/23; Published 7/24

Understanding Entropic Regularization in GANs

Daria Reshetova RESH@STANFORD.EDU
Department of Electrical Engineering
Stanford University
Stanford, CA 94305, USA

Yikun Bai BAI@UDEL.EDU
Department of Electrical and Computer Engineering
University of Delaware
Newark, DE 19716, USA

Xiugang Wu XWU@UDEL.EDU
Department of Electrical and Computer Engineering
University of Delaware
Newark, DE 19716, USA

Ayfer Özgür AOZGUR@STANFORD.EDU
Department of Electrical Engineering
Stanford University
Stanford, CA 94305, USA

Editor: Marco Cuturi

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.

Keywords: Generative Adversarial Networks, Wasserstein GANs, Optimal Transport, Entropic Regularization, Sinkhorn Divergence

1. 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, including image super resolution (Ledig et al., 2017), image-to-image translation (Isola et al., 2017), text to image synthesis (Reed et al., 2016) and analyzing social networks (De et al., 2016). 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 (Goodfellow et al., 2014). This measure was later replaced by the Wasserstein distance (Arjovsky et al., 2017), and the follow-up works showed that Wasserstein GANs can help resolve several issues related to the original formulation, such as the lack of continuity, mode collapse (Arjovsky et al., 2017) and vanishing gradients (Gulrajani et al., 2017).

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 the computation of the Wasserstein distance is to regularize it with an entropic term which makes the problem strongly convex and hence solvable by matrix scaling algorithms (Cuturi, 2013; Balaji et al., 2019). More recent results have shown that this also results in faster convergence and stability of the first-order methods used for optimizing Wasserstein GANs (Sanjabi et al., 2018).

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 by Feizi et al. (2017) and focus on a simple benchmark setting where solutions can be explicitly characterized and compared. Following Feizi et al. (2017), we assume that the generator is linear and the target distribution is high-dimensional Gaussian. The population solution for the Wasserstein GAN in this setup has been characterized by Feizi et al. (2017), who further showed that even in this simple setting the learning problem suffers from the curse of dimesionality—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 issue, Feizi et al. (2017) then propose to restrict the discriminator to be quadratic. This insight is arguably based on knowing that the sought target distribution is Gaussian, in which case the optimal discriminator is indeed quadratic and this restriction does not impact the optimal generator. However, this insight does not generalize beyond the linear/Gaussian setting as for non-Gaussian data the generator obtained under a quadratic discriminator is not necessarily the one minimizing the Wasserstein distance between the generated and the target distributions.

In this paper, by focusing on the linear generator and Gaussian distribution setting (Feizi et al., 2017), we explore how regularization impacts what generator is learned and how it leads to better generalization. We study two slightly different ways of regularizing: entropic regularization (Cuturi, 2013) and Sinkhorn divergence (Genevay et al., 2018). Extending our previous results (Reshetova et al., 2021), we show that the former introduces bias to the solution as if one were to constrain the nuclear norm of the covariance matrix of generator's output distribution, while Sinkhorn divergence results in the same solution as the unregularized Wasserstein GAN (Feizi et al., 2017). 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 (Feizi et al., 2017) without explicitly constraining the discriminator. This indicates that adding regularization implicitly constrains the discriminator in a way suitable for a large class of distributions.

2. Preliminaries

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

2.1 Wasserstein GANs

Let $\mathcal{P}(\mathcal{X})$ be the set of all probability measures with support $\mathcal{X}\subseteq \mathbb{R}^d$ and finite second moments. For $\mathcal{Z}, \mathcal{Y} \subseteq \mathbb{R}^d$, $P_Z \in \mathcal{P}(\mathcal{Z})$ 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, that is all joint probability measures from $\mathcal{P}(\mathcal{Z} \times \mathcal{Y})$ with marginal distributions being P_Z and P_Y. The 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

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

Here we denote $\mathbb{E}_\pi$ the expectation with respect to the measure π. Since π is a coupling of P_Z, P_Y, the marginals of $(Z, Y) \sim \pi$ are correspondingly P_Z and P_Y and (1) is well-defined. It can be verified that 2-Wasserstein distance is a metric between probability distributions in $\mathcal{P}(\mathbb{R}^d)$; in particular, it is symmetric with respect to its two arguments, satisfies the triangle inequality, and $W_2(P_Y, P_Y) = 0$.

The main objective of GANs is to find a mapping $G(\cdot)$, called generator, that comes from a set of functions $\mathcal{G} \subseteq \{G : \mathcal{X} \to \mathcal{Y}\}$ and maps a latent random variable $X \in \mathcal{X}$ with some known distribution to a variable $Y \in \mathcal{Y}$ with some target probability measure P_Y. In the population case, we assume that we have access to P_Y, the true distribution of Y, while in the empirical case one has access to only a finite sample $\{Y_i\}_{i=1}^n$, hence the empirical distribution of Y. Using the squared 2-Wasserstein distance to measure the dissimilarity between the generated and target distribution leads to the following learning problem of GAN, referred to as *W2GAN*:

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

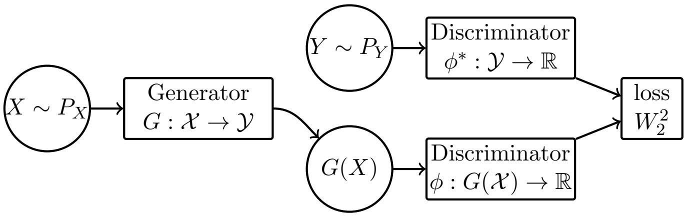
A remarkable feature of the Wasserstein distance is that strong duality holds for the minimization problem described in (1), and hence the objective, squared 2-Wasserstein distance, in (2) can be equivalently written in its dual form (Villani, 2009, Theorem 5.10 and equation 5.12):

$$\begin{aligned} W_2^2(P_{G(X)}, P_Y) &= \sup_{\substack{\psi \in L^1(P_{G(X)}), \phi \in L^1(P_Y) \\ \psi(G(x)) - \phi(y) \leq \|G(x)-y\|^2}} \mathbb{E}\left[\psi(G(X)) - \phi(Y)\right] \\ &= \sup_{\phi \in \mathrm{Conv}(G(\mathcal{X}))} \mathbb{E}\left[\|G(X)\|^2 - 2\phi(G(X)) + \|Y\|^2 - 2\phi^*(Y)\right] \end{aligned} \tag{3}$$

where $\mathrm{Conv}(\mathcal{Z})$ is the set of all (lower semicontinuous) convex functions on $\mathcal{Z}$ and $L^1(P_Z)$ is the set of all functions whose absolute value has a finite expectation: $\phi \in L^1(P_Z) \iff \mathbb{E}[|\phi(Z)|] < \infty$.

Note that the above optimization problem is 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 GANs, 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.

The function ϕ can then be parametrized by a neural network resulting in the following architecture


$Y \sim P_Y$
Discriminator
$\phi^* : \mathcal{Y} \to \mathbb{R}$
$X \sim P_X$
Generator
$G : \mathcal{X} \to \mathcal{Y}$
loss
W_2^2
$G(X)$
Discriminator
$\phi : G(\mathcal{X}) \to \mathbb{R}$

2.2 Entropic Wasserstein GANs

In practice, the Wasserstein distance in (1) is often regularized to facilitate its computation leading to the *entropy regularized 2-Wasserstein distance* (Cuturi, 2013):

$$W_{2,\lambda}^2(P_Z, P_Y) = \inf_{\pi \in \Pi(P_Z, P_Y)} \mathbb{E}_\pi\left[\|Z - Y\|^2\right] + \lambda I_\pi(Z;Y) \tag{4}$$

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

$$\min_{G \in \mathcal{G}} W_{2,\lambda}^2\left(P_{G(X)}, P_Y\right). \tag{5}$$

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, that is $W_{2,\lambda}^2(P_Y, P_Y) \neq 0$. This issue can be resolved by adding corrective terms to (4) (see Genevay et al., 2018), which leads to the Sinkhorn divergence:

$$S_\lambda(P_{G(X)}, P_Y) = W_{2,\lambda}^2(P_{G(X)}, P_Y) - \left(W_{2,\lambda}^2(P_{G(X)}, P_{G(X)}) + W_{2,\lambda}^2(P_Y, P_Y)\right)/2. \tag{6}$$

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

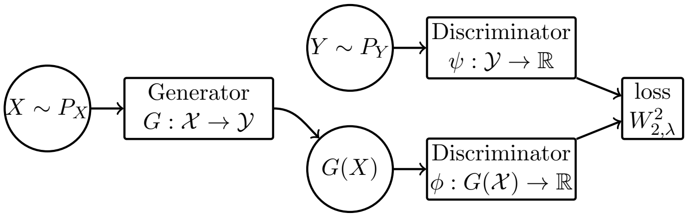
$$\min_{G\in\mathcal{G}} S_\lambda(P_{G(X)}, P_Y). \tag{7}$$

Analogous to the case of the Wasserstein distance, the entropic Wasserstein distance also has a dual formulation which makes it suitable for GAN optimization problems. This dual formulation does not involve optimizing over all couplings, but instead the search space is the set of all essentially bounded functions (Chizat et al., 2018):

$$\begin{aligned} W_{2,\lambda}^2(P_{G(X)}, P_Y) = &\sup_{\psi\in L_\infty(P_Y),\phi\in L_\infty(P_{G(\mathcal{X})})} \mathbb{E}\left[\psi(Y) + \phi(G(X))\right] + \lambda \\ &- \lambda\mathbb{E}_{(X,Y)\sim P_X\times P_Y}\left[e^{\frac{\phi(G(X))+\psi(Y)-\|G(X)-Y\|_2^2}{\lambda}}\right], \end{aligned} \tag{8}$$

where $L_\infty(P_Y)$ is the set of all essentially bounded functions, $\phi \in L_\infty(P_Y) \iff \exists C > 0 : P\{\phi(Y) > C\} = 0$.

The so-called dual potentials ϕ, ψ can be parametrized by neural networks resulting in the following architecture


$Y \sim P_Y$
Discriminator
$\psi : \mathcal{Y} \to \mathbb{R}$
$X \sim P_X$
Generator
$G : \mathcal{X} \to \mathcal{Y}$
loss
$W_{2,\lambda}^2$
$G(X)$
Discriminator
$\phi : G(\mathcal{X}) \to \mathbb{R}$

Note that in (8), there are no constraints on the dual potentials, which makes the dual form suitable to implement with Neural networks, while 2-Wasserstein distance requires convexity/quadratically bounded differences for the dual potential and 1-Wasserstein distance, another popular metric used in GANs, requires Lipschitz continuity of the dual potential. The constraints on the discriminators then give rise to various heuristics (Korotin et al. 2019; Liu et al. 2019 for 2-Wasserstein GANs; Arjovsky et al. 2017; Wei et al. 2018 for 1-Wasserstein GANs) since the constraints cannot be handled exactly.

When one of the measures is an empirical distribution supported on $\{y_i\}_{i=1}^n$, which is often the case in GANs, only the values of ψ on the empirical samples influence the solution, thus letting $\psi_i = \psi(y_i)$ and plugging in the empirical measure in place of P_Y simplifies (8) to

$$\begin{aligned} W_{2,\lambda}^2(P_{G(X)}, P_Y) = &\sup_{\psi\in\mathbb{R}^n,\phi\in L_\infty(G(\mathcal{X}))} \sum_{i=1}^n \frac{\psi_i}{n} + \mathbb{E}\left[\phi(G(X))\right] + \lambda \\ &- \lambda\mathbb{E}\left[\sum_{i=1}^n e^{\frac{\phi(G(X))+\psi_i-\|G(X)-y_i\|_2^2}{\lambda}}/n\right] \end{aligned} \tag{9}$$

The above form is especially useful for optimization since one of the parametric functions becomes a vector.

Given the optimal dual potentials, the optimal coupling can be found as (Janati et al., 2020),

$$\pi(G(x), y) = P_{G(X)}(G(x))P_Y(y)e^{\frac{\phi(G(x))+\psi(y)-\|G(x)-y\|_2^2}{\lambda}}. \tag{10}$$

Even though we are less interested in the computational aspects of optimal transport in this paper, we note that the optimal dual potentials for entropy regularized 2-Wasserstein distance can be shown to satisfy the following equations

$$\phi^*(G(x)) = -\lambda \ln \mathbb{E}\left[e^{(\psi^*(Y)-\|Y-G(x)\|_2^2)/\lambda}\right] \tag{11}$$

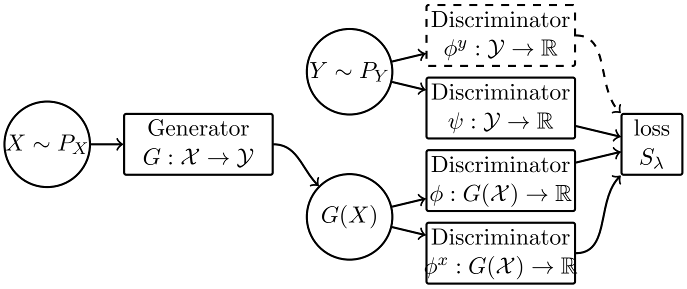
$$\psi^*(y) = -\lambda \ln \mathbb{E}\left[e^{(\phi^*(G(X))-\|y-G(X)\|_2^2)/\lambda}\right] \tag{12}$$

Equations (11),(12) give rise to the celebrated Sinkhorn-Knopp algorithm that allows for fast computation of entropic optimal transport via iterative updates: at iteration t we set,

$$\phi^t(G(x)) = -\lambda \ln \mathbb{E}\left[e^{(\psi^{t-1}(Y)-\|Y-G(x)\|_2^2)/\lambda}\right] \tag{13}$$

$$\psi^t(y) = -\lambda \ln \mathbb{E}\left[e^{(\phi^t(G(X))-\|y-G(X)\|_2^2)/\lambda}\right]. \tag{14}$$

Since Sinkhorn divergence is a linear combination of entropy-regularized Wasserstein distances, it also has a dual form and strong duality holds. The dual form of entropy-regularized 2-Wasserstein distance gives rise to an equivalent formulation of the Sinkhorn divergence as a linear combination of the dual formulations of entropy-regularized 2-Wasserstein distances. Since $W_{2,\lambda}^2(P_{G(X)}, P_{G(X)})$ and $W_{2,\lambda}^2(P_Y, P_Y)$ are symmetric in the dual potentials and concave, the optimal dual potentials will be equal, that is $\phi^x(G(x)) = \psi^x(G(x))$ and $\phi^y(y) = \psi^y(y)$ resulting in the following architecture


$X \sim P_X$
Generator
$G: \mathcal{X} \to \mathcal{Y}$
$Y \sim P_Y$
$G(X)$
Discriminator
$\phi^y : \mathcal{Y} \to \mathbb{R}$
Discriminator
$\psi : \mathcal{Y} \to \mathbb{R}$
Discriminator
$\phi : G(\mathcal{X}) \to \mathbb{R}$
Discriminator
$\phi^x : G(\mathcal{X}) \to \mathbb{R}$
loss
S_λ

The Discriminator ϕ^y is dashed since it depends only on the distribution P_Y and does not influence the generator (see Feydy et al., 2019).

3. Population Solution for the Linear/Gaussian Setting

In this section, we aim to compare the optimal solution for GANs when we use the different measures introduced in the previous section for quantifying the dissimilarity between the generated and target probability distributions. For this purpose, we focus on the benchmark setting considered by Feizi et al. (2017), where the generator is linear and the target distribution is Gaussian. The population solution in this setting is the low-dimensional Gaussian approximation of a higher-dimensional Gaussian with the regularized 2-Wasserstein distance as the approximation measure. We can rewrite the general formulation of (2) as:

$$\min_{G\in\mathbb{R}^{d\times r}} W_2^2\left(P_{GX}, P_Y\right), \tag{15}$$

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}.

While this is a toy setting that is almost never encountered in applications, even in this case the unregularized Wasserstein GAN suffers from the curse of dimensionality (Feizi et al., 2017). In the case of entropic regularization, the setting leads to a closed-form solution and an explicit characterization of the impact of the regularization on the generator. The population solution to the above W2GAN problem has been characterized by Feizi et al. (2017) as the r-PCA solution of Y, i.e. the covariance matrix K_{G^*X} for P_{G^*X}, where G^* denotes the minimizer of (15), 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*

$$\min_{G\in\mathbb{R}^{d\times r}} W_{2,\lambda}^2\left(P_{GX}, P_Y\right), \tag{16}$$

is given by a soft-thresholded r-PCA solution of Y, i.e., the covariance matrix K_{G^*X} for P_{G^*X}, where G^* denotes now the minimizer of (16), is a rank-r matrix whose top r eigenvectors are the same as those of K_Y and the top r eigenvalues are

$$\sigma_i^2 = (\lambda_i(K_Y) - \lambda/2)_+ \quad \textit{for} \quad i \in [1:r],$$

where $(x)_+ := \max\{x, 0\}$ and $\{\lambda_i(K_Y)\}_{i=1}^r$ are the top r eigenvalues of K_Y.

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. Note that the soft-thresholding promotes sparsity in the eigenvalues of the covariance matrix of the generated distribution since if many of the eigenvalues of K_Y are below the threshold $\lambda/2$ the rank of K_{G^*X} can be significantly smaller than K_Y.

We would like to emphasize that the sparsity is with respect to the generator function and thus the output distribution covariance matrix. If sparsity is considered in terms of the optimal transport plan, we note that entropic regularization can compromise the sparsity that would appear in the unregularized optimal transport plan as shown by Blondel et al. (2018, Figure 1). This is intuitive as entropic regularization forces a certain amount of randomization in the transport plan.

We note that soft thresholding of singular values arises as the optimal solution to a different problem that has been studied in the context of low rank matrix completion. Consider the problem:

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

where $\|\cdot\|_*$ is the nuclear norm, i.e. the sum of all singular values of a matrix, which can be regarded as a relaxation of a low rank constraint. The solution of this problem is shown to be the soft thresholded PCA solution for $r = d$, (see Cai et al., 2010, Theorem 2.1), namely K_Z and K_Y share the same eigenvectors with corresponding eigenvalues thresholded as in Theorem 1. We note that even though the two problems interestingly lead to the same solution, they are different problems and this result cannot be applied to a more general or the empirical setting.

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 a simple consequence of 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 sample complexity, as we investigate 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 given by*

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

is given by the r-PCA solution of Y.

3.1 Proofs of Theorems 1 and 2

Proof [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 $Y = Y_{\mathcal{S}_Z} + Y_{\mathcal{S}_Z^\perp}$. Note that for a fixed G, $Y_{\mathcal{S}_Z}$ and $Y_{\mathcal{S}_Z^\perp}$ can be computed given Y. The entropy regularized 2-Wasserstein distance is then

$$\begin{aligned}
W_{2,\lambda}^2(P_Y, P_Z) &= \min_{\pi \in \Pi(P_Y, P_Z)} \mathbb{E}_\pi\big[\|Z - Y\|^2\big] + \lambda I_\pi(Z; Y) \\
&= \min_{\pi \in \Pi(P_Y, P_Z)} \mathbb{E}_\pi\big[\|(Z - Y_{\mathcal{S}_Z}) - Y_{\mathcal{S}_Z^\perp}\|^2\big] + \lambda I_\pi(Z; Y) \\
&= \min_{\pi \in \Pi(P_Y, P_Z)} \mathbb{E}_\pi\big[\|Z - Y_{\mathcal{S}_Z}\|^2\big] + \mathbb{E}\big[Y_{\mathcal{S}_Z^\perp}\|^2\big] + \lambda I_\pi(Z; Y) \qquad (17) \\
&= \min_{\pi \in \Pi(P_Y, P_Z)} \mathbb{E}_\pi\big[\|Z - Y_{\mathcal{S}_Z}\|^2\big] + \mathbb{E}\big[Y_{\mathcal{S}_Z^\perp}\|^2\big] + \lambda I_\pi(Z; Y_{\mathcal{S}_Z}). \qquad (18)
\end{aligned}$$

The last equality above holds because $I_\pi(Z;Y) = I_\pi(Z;Y_{\mathcal{S}_Z}, Y_{\mathcal{S}_Z^\perp}) \geq I_\pi(Z;Y_{\mathcal{S}_Z})$ and moreover, for any coupling π, one can construct π' such that $\pi'(Z, Y_{\mathcal{S}_Z}, Y_{\mathcal{S}_Z^\perp}) = \pi(Z, Y_{\mathcal{S}_Z})\pi(Y_{\mathcal{S}_Z^\perp}|Y_{\mathcal{S}_Z})$, i.e. $Z - Y_{\mathcal{S}_Z} - Y_{\mathcal{S}_Z^\perp}$ forms a Markov chain. Note that π' preserves the values of the first two terms in (17) while $I_{\pi'}(Z;Y_{\mathcal{S}_Z}, Y_{\mathcal{S}_Z^\perp}) = I_\pi(Z;Y_{\mathcal{S}_Z})$.

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 (18), the above is

$$\min_{\substack{\mathcal{S}\in\mathbb{S}_d:\dim(\mathcal{S})\leq r\\ P_Z\in\mathcal{N}_{d,r}:\, Z\in\mathcal{S}\\ \pi\in\Pi(P_Y,P_Z)}} \mathbb{E}_\pi[\|Z - Y_\mathcal{S}\|^2] + \mathbb{E}[\|Y_{\mathcal{S}^\perp}\|^2] + \lambda I_\pi(Z;Y_\mathcal{S}), \tag{19}$$

where $\mathbb{S}_d$ is the set of all subspaces of $\mathbb{R}^d$. To solve (19) 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 $U^T U = I_r$, we replace Z and $Y_\mathcal{S}$ in (19) by $U^T Z$ and $U^T Y$ respectively. To find optimal π, P_Z for $\mathcal{S}$ we then solve

$$\min_{\substack{P_Z\in\mathcal{N}_{d,r}:Z\in\operatorname{Im}(U)\\ \pi\in\Pi(P_Z,P_Y)}} \mathbb{E}_\pi[\|U^T Z - U^T Y\|^2] + \lambda I_\pi(U^T Z; U^T Y) - \mathbb{E}[\|U^T Y\|^2] + \mathbb{E}[\|Y\|^2] \tag{20}$$

Let $\bar{Z} = U^T Z$ and $\bar{Y} = U^T Y$, and let $\mathcal{N}_{r,r}$ be the set of all r-dimensional Gaussian distributions. Then solving Problem (20) is equivalent to solving

$$\min_{P_{\bar{Z}}\in\mathcal{N}_{r,r}} \min_{\pi\in\Pi(P_{\bar{Z}},P_{\bar{Y}})} \mathbb{E}_\pi[\|\bar{Z} - \bar{Y}\|^2] + \lambda I_\pi(\bar{Z};\bar{Y}) \tag{21}$$

We will proceed by first lower bounding (21) and then providing the coupling and $P_{\bar{Z}}$ that achieve the lower bound.

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. Since $\bar{Y}$ is also Gaussian this implies that its components are independent. This in turn implies that $I_\pi(\bar{Z};\bar{Y}) \geq \sum_{i=1}^r I_\pi(\bar{Z}_i;\bar{Y}_i)$, and hence (21) can be lower bounded by

$$\begin{aligned}
&\min_{P_{\bar{Z}}\in\mathcal{N}_{r,r}} \min_{\pi\in\Pi(P_{\bar{Z}},P_{\bar{Y}})} \mathbb{E}_\pi[\|\bar{Z} - \bar{Y}\|^2] + \lambda I_\pi(\bar{Z};\bar{Y})\\
&\qquad\geq \min_{P_{\bar{Z}}\in\mathcal{N}_{r,r}} \min_{\pi\in\Pi(P_{\bar{Z}},P_{\bar{Y}})} \sum_{i=1}^r \mathbb{E}_{\pi_i}[(\bar{Z}_i - \bar{Y}_i)^2] + \lambda I_{\pi_i}(\bar{Z}_i;\bar{Y}_i)
\end{aligned} \tag{22}$$

Note that for fixed Gaussian $P_{\bar{Z}}, P_{\bar{Y}}$ and cross-covariance matrix $K_{\bar{Z}\bar{Y}}$ the first term in (21) is fixed and the mutual information term is minimized when π is jointly Gaussian, i.e. $\pi \in \mathcal{N}(P_{\bar{Z}}, P_{\bar{Y}})$, where $\mathcal{N}(\mu,\nu)$ denotes a set of jointly Gaussian distributions with marginals μ,ν. Therefore the minimization in (22) can be restricted to $\pi \in \mathcal{N}(P_{\bar{Z}}, P_{\bar{Y}})$. If we let $D_i = \mathbb{E}(\bar{Y}_i - \bar{Z}_i)^2$, this in turn yields

$$\begin{aligned}
I_{\pi_i}(\bar{Z}_i;\bar{Y}_i) &= h(\bar{Y}_i) - h(\bar{Y}_i|\bar{Z}_i)\\
&= h(\bar{Y}_i) - h(\bar{Y}_i - \bar{Z}_i|\bar{Z}_i)\\
&\geq h(\bar{Y}_i) - h(\bar{Y}_i - \bar{Z}_i)\\
&= (1/2)\ln(\Lambda_i/D_i),
\end{aligned} \tag{23}$$

where (23) follows from the fact that conditioning reduces entropy. Since mutual information is non-negative, we can tighten the bound to $I_{\pi_i}(\bar{Z}_i;\bar{Y}_i) \geq \max\{0, (1/2)\ln(\Lambda_i/D_i\}$.

Thus, continuing the lower bound from (22) we get

$$\begin{aligned}
&\min_{P_{\bar{Z}}\in\mathcal{N}_{r,r}} \min_{\pi\in\Pi(P_{\bar{Z}},P_{\bar{Y}})} \mathbb{E}_\pi[\|\bar{Z}-\bar{Y}\|^2] + \lambda I_\pi(\bar{Z};\bar{Y}) \\
&\quad\geq \sum_{i=1}^r \min_{D_i\geq 0} D_i + (\lambda/2)\max\{\ln(\Lambda_i/D_i), 0\} \\
&\quad= \sum_{i=1}^r \min_{0\leq D_i\leq\Lambda_i} D_i + (\lambda/2)\ln(\Lambda_i/D_i) \qquad (24)\\
&\quad= \sum_{i=1}^r \min_{0\leq D_i\leq\Lambda_i} g(D_i) + (\lambda/2)\ln(\Lambda_i),
\end{aligned}$$

where (24) follows from the fact that increasing D_i beyond Λ_i leaves the second summand the same, while increases the first one, so the minimum is attained at $D_i \leq \Lambda_i$ and $g(x) = x - \lambda/2\ln x$. As $g'(x) = 1 - \lambda/(2x) < 0$ for $x < \lambda/2$, the optimal value is attained at $D_i = D_i^* = \min\{\lambda/2, \Lambda_i\}$ and plugging it into the bound we get

$$\begin{aligned}
&\min_{P_{\bar{Z}}\in\mathcal{N}_{r,r}} \min_{\pi\in\Pi(P_{\bar{Z}},P_{\bar{Y}})} \mathbb{E}_\pi[\|\bar{Z}-\bar{Y}\|^2] + \lambda I_\pi(\bar{Z};\bar{Y}) \qquad (25)\\
&\quad\geq \sum_{i=1}^r \min\{\lambda/2, \Lambda_i\} + (\lambda/2)\ln(\Lambda_i/\min\{\lambda/2,\Lambda_i\}) \qquad (26)
\end{aligned}$$

To prove that the lower bound holds with equality we need (23) to hold with equality, i.e. $\bar{Y}_i - \bar{Z}_i$ be independent of $\bar{Z}_i$, so we can choose $\bar{Y} = \bar{Z} + N$, where $N \sim \mathcal{N}(0, \text{diag}(\{\min\{\lambda/2, \Lambda_i\}\}_{i=1}^r)$ independent of $\bar{Z}$. Note that since the distribution of $\bar{Y}$ is fixed, this in turn fixes the distribution of $\bar{Z}$. With this choice of a coupling

$$\begin{aligned}
\mathbb{E}_\pi[\|\bar{Z}-\bar{Y}\|^2] &= \mathbb{E}_\pi[\|N\|^2] = \sum_{i=1}^r D_i^* \\
I(\bar{Z};\bar{Y}) &= h(\bar{Y}) - h(\bar{Y}-\bar{Z} \mid \bar{Z}) \\
&= h(\bar{Y}) - h(N) \\
&= (1/2)\sum_{i=1}^r \ln(\Lambda_i) - \ln(D_i^*)
\end{aligned}$$

Combining the above we get

$$\mathbb{E}_\pi[\|\bar{Z}-\bar{Y}\|^2] + \lambda I_\pi(\bar{Z};\bar{Y}) = \sum_{i=1}^r \min\{\lambda/2,\Lambda_i\} + (\lambda/2)\sum_{i=1}^r \ln(\Lambda_i/\min\{\lambda/2,\Lambda_i\}),$$

which matches the lower bound in (25), so $\bar{Y} = \bar{Z} + N$ with $N \sim \mathcal{N}(0, \text{diag}(\{D_i\}_{i=1}^r))$ independent of Z is the optimal coupling.

Before we continue with the proof we make the following remark. Note that when the condition $P_{\bar{Z}} \in \mathcal{N}_{r,r}$ is dropped we get a lower bound on (21):

$$W_{lower} = \min_{\pi_{\bar{Z}|\bar{Y}}} \mathbb{E}_\pi[\|\bar{Z} - \bar{Y}\|^2] + \lambda I_\pi(\bar{Z}; \bar{Y}) \tag{27}$$

Introduction of a new variable $D = \mathbb{E}\|\bar{Z} - \bar{Y}\|^2$ leads to the following optimization problem:

$$\begin{array}{ll} \min\limits_{\pi_{\bar{Z}|\bar{Y}}, D} & D + \lambda I_\pi(\bar{Z}; \bar{Y}) \\ \text{subject to} & D = \mathbb{E}[\|\bar{Z} - \bar{Y}\|^2] \\ & D \geq 0 \end{array}$$

Note that the equality in the constraints can be relaxed to an inequality since the objective is linear in D, which leads to

$$\begin{array}{ll} \min\limits_{\pi_{\bar{Z}|\bar{Y}}, D} & D + \lambda I_\pi(\bar{Z}; \bar{Y}) \\ \text{subject to} & D \geq \mathbb{E}[\|\bar{Z} - \bar{Y}\|^2] \\ & D \geq 0 \end{array} \tag{28}$$

Finally we can rewrite the full minimization over $\pi_{\bar{Z}|\bar{Y}}, D$ as a consecutive one, which leads to

$$W_{lower} = \min_{D \geq 0} \left\{ D + \lambda \min_{\pi_{\bar{Z}|\bar{Y}} : \mathbb{E}[\|\bar{Z} - \bar{Y}\|^2] \leq D} I_\pi(\bar{Z}; \bar{Y}) \right\} \tag{29}$$

The inner minimization problem is exactly the Gaussian rate distortion problem (Cover, 1999, Eq. 10.38) that can be solved by noting that the mutual information term is minimized for a Gaussian distribution, plugging the value of the mutual information in and writing down the Karush-Kuhn-Tucker optimality conditions. The solution for this problem is given by reverse waterfilling (Cover, 1999, Theorem 10.3.3), under which the optimal $P_{\bar{Z}}$ is an r-dimensional Gaussian which matches the solution we have obtained in (26).

The entropic W2GAN optimization problem (19) is then equivalent to minimizing (26) over the set of all $r-$dimensional subspaces of $\mathbb{R}^d$:

$$\begin{aligned} &\min_{U \in \mathbb{R}^{d \times r}} \sum_{i=1}^r \min\{\Lambda_i, \lambda/2\} + \frac{\lambda}{2} \ln \frac{\Lambda_i}{\min\{\Lambda_i, \lambda/2\}} + \mathbb{E}[\|Y_{(\operatorname{Im} U)^\perp}\|^2] \\ &\quad = \min_{U \in \mathbb{R}^{d \times r}} \sum_{i=1}^r \left(\Lambda_i + \lambda/2 - \max\{\Lambda_i, \lambda/2\} + \frac{\lambda}{2} \ln \frac{\max\{\Lambda_i, \lambda/2\}}{\lambda/2} \right) + \mathbb{E}[\|Y_{(\operatorname{Im} U)^\perp}\|^2] \\ &\quad = \min_{U \in \mathbb{R}^{d \times r}} \sum_{i=1}^r \left(\frac{\lambda}{2} \ln \frac{\max\{\Lambda_i, \lambda/2\}}{\lambda/2} - \max\{\Lambda_i, \lambda/2\} \right) + \frac{r\lambda}{2} + \mathbb{E}[\|Y\|^2] \end{aligned}$$

where the optimization is over all $U \in \mathbb{R}^{d \times r}$ such that $U^T U = I_r$ and $U^T K_Y U = \operatorname{diag}(\Lambda_1, \ldots, \Lambda_r)$. We now let

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

and complete the proof by showing

$$\min_{U\in\mathbb{R}^{d\times r}:U^T K_Y U=\text{diag}(\Lambda_i,\dots,\Lambda_r)} \sum_{i=1}^{r} f(\Lambda_i) = \sum_{i=1}^{r} f(\lambda_i(K_Y)).,\tag{30}$$

where $[\lambda_1(K_Y),\dots,\lambda_r(K_Y)]$ are the largest r eigenvalues of K_Y.

Indeed, for each U we can construct an orthogonal matrix $U' = [U\ U_\perp] \in \mathbb{R}^{d\times d}$ with the first r columns equal to U. Then the first r diagonal elements of $U'^T K_Y U'$ are $\Lambda_1 \dots \Lambda_r$, let the rest be $\Lambda_{r+1}\dots\Lambda_d$. The eigenvalues of $U'^T K_Y U'$ are $\lambda_1(K_Y)\dots\lambda_d(K_Y)$. By the fact that the diagonal entries of a symmetric matrix are majorized by its eigenvalues (Marshall et al., 1979, Theorem 9.B.1), we have

$$\{\Lambda_i\}_{i=1}^{d} \prec \{\lambda_i(K_Y)\}_{i=1}^{d},$$

where $\prec$ denotes majorization, i.e. for $x, y \in \mathbb{R}^d$:

$$x \prec y \iff \forall r \le d : \sum_{i=1}^{r} x_{(i)} \le \sum_{i=1}^{r} y_{(i)} \text{ and } \sum_{i=1}^{d} x_i = \sum_{i=1}^{d} y_i,$$

where $x_{(i)}$ is the i-th largest element of the vector x.

Therefore,

$$\sum_{i=1}^{r'} \Lambda_i \le \sum_{i=1}^{r'} \Lambda_{(i)} \le \sum_{i=1}^{r'} \lambda_i(K_Y), \quad \forall r' \le r$$

and

$$\{\Lambda_i\}_{i=1}^{r} \prec_w \{\lambda_i(K_Y)\}_{i=1}^{r},$$

where $\prec_w$ denotes weak majorization, i.e. for $x, y \in \mathbb{R}^r$:

$$x \prec_w y \iff \forall r' \le r : \sum_{i=1}^{r'} x_i \le \sum_{i=1}^{r'} y_i.$$

We can now use the majorizing inequality (see Marshall et al., 1979, Proposition 4.B.2) to complete the proof.

Proposition 3 (Majorizing inequality) *The inequality*

$$\sum g(x_i) \le \sum g(y_i)\tag{31}$$

holds for all continuous non-decreasing convex functions g if and only if $x \prec_w y$.

Note that $-f$ is a continuous non-decreasing convex function and

$$\{\Lambda_i\}_{i=1}^{r} \prec_w \{\lambda_i(K_Y)\}_{i=1}^{r},$$

thus by the proposition

$$\sum_{i=1}^{r} f(\lambda_i(K_Y)) \le \sum_{i=1}^{r} f(\Lambda_i).$$

Therefore, columns of the optimal U are the top r eigenvectors of K_Y, and the optimal P_Z has covariance matrix given by

$$K_Z = U[\operatorname{diag}(\sigma_1^2, \dots, \sigma_r^2)|0_{r\times(d-r)}]U^T$$

where $\sigma_i^2 = (\lambda_i(K_Y) - \lambda/2)_+$. ■

Proof [Proof of Theorem 2] From (18) in the proof of Theorem 1, we have for given G and $\mathcal{S} = \operatorname{Im} G$,

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

and therefore for the Sinkhorn divergence,

$$\begin{aligned} S_\lambda(P_Z, P_Y) - \mathbb{E}\|Y_{\mathcal{S}^\perp}\|_2^2 &= W_{2,\lambda}^2(P_Z, P_{Y_\mathcal{S}}) - \big(W_{2,\lambda}^2(P_Z, P_Z) + W_{2,\lambda}^2(P_Y, P_Y)\big)/2 \\ &= S_\lambda(P_Z, P_{Y_\mathcal{S}}) + \big(W_{2,\lambda}^2(P_{Y_\mathcal{S}}, P_{Y_\mathcal{S}}) - W_{2,\lambda}^2(P_Y, P_Y)\big)/2, \end{aligned} \tag{32}$$

which follows from the definition of Sinkhorn divergence.

Consider the optimization problem in the Sinkhorn divergence GAN, i.e. $\min_{P_Z} S_\lambda(P_Z, P_Y)$. In light of (32), given $Z \in \mathcal{S}$ the optimal P_Z should be $P_{Y_\mathcal{S}}$, which makes the first term in (32) zero while the remaining terms do not depend on P_Z. Therefore, it only remains to optimize over $\mathcal{S}$, and in particular, the problem reduces to

$$\min_{\mathcal{S}\in\mathbb{S}_d : \dim(\mathcal{S})\le r} W_{2,\lambda}^2(P_{Y_\mathcal{S}}, P_{Y_\mathcal{S}})/2 + \mathbb{E}\|Y_{\mathcal{S}^\perp}\|_2^2. \tag{33}$$

To calculate $W_{2,\lambda}^2(P_{Y_\mathcal{S}}, P_{Y_\mathcal{S}})$ we use the result of Janati et al. (2020, Theorem 1) stated below.

Proposition 4 (Entropy-regularized Wasserstein distance for Gaussian measures) *Let $K_X, K_Y \in \mathbb{R}^{d\times d}$ be positive definite and $X \sim \mathcal{N}(\mu_X, K_X)$ and $Y \sim \mathcal{N}(\mu_Y, K_Y)$. Define $\mathbf{D}_\lambda = (4A^{\frac{1}{2}} K_Y A^{\frac{1}{2}} + \lambda^2 I/4)^{\frac{1}{2}}$. Then,*

$$\begin{aligned} W_{2,\lambda}^2(\alpha, \beta) =& \|\mu_X - \mu_Y\|^2 + \operatorname{Tr}(K_X) + \operatorname{Tr}(K_Y) - \operatorname{Tr}(\mathbf{D}_\lambda) \\ &+ \frac{\lambda}{2}\left(d(1 - \log\lambda) + \log\det\left(\mathbf{D}_\lambda + \frac{\lambda}{2}\right)\right) \end{aligned}$$

With this proposition the objective function in (33) becomes

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

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 depending only on λ and d and $C' = \operatorname{Tr} K + C$. The above is minimized when $\Lambda_i = \lambda_i(K_Y)$ using a similar argument as the one for showing (30), i.e., by using the majorizing inequality and noting that for the function

$$f(x) = \sqrt{4x^2 + \lambda^2/4}/2 - \lambda \ln(\sqrt{4x^2 + \lambda^2/4} + \lambda/2)/4,$$

$-f(x)$ is convex and non-decreasing for $\lambda > 0$ and $x \geq 0$. ■

4. Generalization of the 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), (5) and (7). 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), (5) and (7) 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 by Feizi et al. (2017) 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 excess risk

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

That is, to decrease the excess risk 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, Feizi et al. (2017) 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 in the Gaussian setting because the optimal discriminator for W2GAN is indeed quadratic in this case. On the other hand, it was shown that this constraint will lead to fast convergence to the optimal solution of (1) when the generator is linear and the data distribution is high-dimensional Gaussian. More precisely, the convergence is of order $O_d(n^{-1/2})$ and hence the issue of curse of dimensionality is resolved in this case.

While constraining the discriminator to be quadratic as done by Feizi et al. (2017) is conceptually appealing and works for the setup of linear generators and Gaussian data, this insight does not generalize beyond this special case, 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 and in general can be far from optimality.

In this section, We show in theorems 6 and 7 that under mild conditions on the underlying distribution of data, the latent random variable and the set of generators, convergence of order $O_d(n^{-1/2})$ can be achieved for entropic W2GAN and Sinkhorn W2GAN without the need to constrain the discriminator.

The parametric rate of convergence of entropy-regularized 2-Wasserstein distance (4) was first discovered in the seminal work of Mena and Niles-Weed (2019). Our work extends their result to the setting of GANs, the entropic W2GAN, and the Sinkhorn W2GAN in particular. In the remainder of this section, we first state the result of Mena and Niles-Weed (2019), then we formally state our generalization results and overview the proof technique discussing why these results cannot be directly applied to the framework of GANs. Finally, we provide the full formal proof in subsection 4.4.

4.1 Convergence Rate of Entropic 2-Wasserstein Distance

To formally state the results for the convergence rate, let us first recall the definition of sub-Gaussianity. A distribution P_X is σ^2 sub-gaussian for $\sigma \geq 0$ if

$$\mathbb{E} \exp\left(\|X\|^2/(2r\sigma^2)\right) \leq 2.$$

Let

$$\sigma^2(X) {=} \min\{\sigma {\geq}\ 0 \big| \mathbb{E} \exp(\|X\|^2/(2r\sigma^2)) \leq 2\,\},$$

denote the sub-gaussian parameter of the distribution of X.

We note that the definition of the Entropic Wasserstein distance in this work differs from the definition of Mena and Niles-Weed (2019) by a factor of 1/2 in the cost function: they define an entropy-regularized 2-Wasserstein distance between P_X and P_Y as

$$\inf_{\pi \in \Pi(P_X, P_Y)} \frac{1}{2}\mathbb{E}[\|X-Y\|^2] + \lambda I_\pi(X,Y),$$

so all the results of Mena and Niles-Weed (2019) apply to $(1/2)W^2_{2,2\lambda}(P_X, P_Y)$ in our setting. We state their main result below by modifying it for this factor of 1/2 in the regularization strength and cost function.

Proposition 5 (Mena and Niles-Weed 2019,Corollary 1 of Theorem 2) *If P_Z and P_Y are σ^2 sub-gaussian, then*

$$\mathbb{E}\left[\left|W^2_{2,\lambda}(P_Z, Q^n_Y) - W^2_{2,\lambda}(P_Z, P_Y)\right|\right] \leq K_d \lambda n^{-1/2}\left(1 + (2\sigma^2/\lambda)^{\lceil 5d/4 \rceil + 3}\right), \tag{34}$$

where K_d is a constant depending on the dimension.

This result establishes the convergence rate of the entropy-regularized Wasserstein distance to be of order $O_d(1/\sqrt{n})$ (ignoring the dimension-dependent constants) compared to $\Omega(n^{-2/d})$ known for the unregularized version of the Wasserstein distance (Dudley, 1969). It is also worth noting, that though the dimension was removed from the exponent of n, the dependence of the constant K_d in (34) still remains exponential in the dimension.

4.2 Generalization Results

We show the generalization of entropic W2GANs and Sinkhorn GANs by providing an upper bound on their excess risk, which shows how far the true loss of the empirical solution G_n (the solution found from the empirical distribution) is from the loss function of the population solution G^*. We consider sub-gaussian latent and target distribution and star-shaped generator sets, the definition for which we provide below.

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.

$$G \in \mathcal{G} \;\Rightarrow \alpha G \in \mathcal{G}, \forall \alpha \in [0,1]. \tag{35}$$

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. This also includes the set of all L-Lipschitz functions as another example.

Theorem 6 *Let $P_{K_X^{-1/2}X}$ and P_Y be sub-gaussian and the generator set $\mathcal{G}$ be a set of linear functions satisfying condition* (35)*. Then the excess risk for entropic W2GAN can be bounded by*

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

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.

Similar results also hold for the set of Lipschitz generators and extend to the Sinkhorn W2GAN.

Theorem 7 *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* (35)*. Then the excess risk 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

$$K_d \lambda n^{-1/2}\big(1 + (2\tau^2/\lambda)^{\lceil 5d/4 \rceil + 3}\big) \tag{36}$$

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

The above theorems essentially say that under certain mild conditions on the latent and target distributions and the set of generators $\mathcal{G}$, the excess risks for entropic W2GAN and Sinkhorn W2GAN converge to zero at speed $O_d(1/\sqrt{n})$. From the perspective of Feizi et al. (2017) mentioned earlier, this may suggest that the set of possible discriminators is implicitly constrained due to the entropic regularization term used in the primal form of entropy regularized 2-Wasserstein distance.

It is worth mentioning that a related result was proved by Luise et al. (2020, Theorem 2). However, their framework is different from ours as they focus on latent distribution learning and assume that the target distribution is supported on a low-dimensional manifold of dimension r, in which case they are able to provide convergence rates that depend on the latent dimension r rather than the ambient dimension d. We note that both results can be applied to the special case $r = d$ in our setting, i.e., when the target and latent distributions are of the same dimensionality. In this special case, both the results of Luise et al. (2020, Theorem 2) and our results in Theorems 6 and 7 above yield similar convergence rates but Luise et al. (2020, Theorem 2) require significantly stronger conditions for the set of generator functions $\mathcal{G}$. In particular, they require any $G \in \mathcal{G}$ to be $\lceil d/2 \rceil + 1$ times continuously differentiable with all partial derivatives uniformly bounded with some constant $\tau : \|G\|_{\lceil d/2 \rceil + 1, \infty} \leq \tau$. Namely, it does not apply to $\mathcal{G}$ being the set of all linear functions or $L-$Lipschitz functions unless $r = 1$.

4.3 Overview of Proofs of Theorems 6 and 7

We first note that Proposition 5 (Mena and Niles-Weed, 2019, Theorem 2) cannot be directly used to bound the excess risk for GANs. Consider the standard way to bound it, which is to decompose the quantity under the expectation into two following terms:

$$\begin{aligned} W_{2,\lambda}^2(P_{G_n(X)}, P_Y) - W_{2,\lambda}^2(P_{G^*(X)}, P_Y) &= \left(W_{2,\lambda}^2(P_{G^*(X)}, Q_Y^n) - W_{2,\lambda}^2(P_{G^*(X)}, P_Y)\right) \\ &\quad + \left(W_{2,\lambda}^2(P_{G_n(X)}, P_Y) - W_{2,\lambda}^2(P_{G^*(X)}, Q_Y^n)\right) \\ &\leq \left(W_{2,\lambda}^2(P_{G^*(X)}, Q_Y^n) - W_{2,\lambda}^2(P_{G^*(X)}, P_Y)\right) \qquad (37) \\ &\quad + \left(W_{2,\lambda}^2(P_{G_n(X)}, P_Y) - W_{2,\lambda}^2(P_{G_n(X)}, Q_Y^n)\right), \end{aligned}$$

where the second inequality holds due to the optimality of G_n. Now, the expected value of the first summand in (37) can be upper bounded with (34) provided that $P_{G^*(X)}$ and P_Y are σ^2 sub-gaussian. However, the expectation of the second summand in (37) cannot be bounded in the same way, primarily because the distribution $P_{G_n(X)}$ now depends on Q_Y^n. Additionally, in the case of linear generators discussed in Section 3, there is no known finite bound on the sub-gaussian norm of the generated distribution as $\sup_{G_n \in \mathcal{G}} \sigma^2(G_n(X)) = \infty$, so even if the dependence of G_n and the sample Q_Y^n was negligible, further bounds on the sub-gaussian parameter of empirical solution would be necessary.

Instead of following this approach, we rely on the fact that Theorem 2 of Mena and Niles-Weed (2019) itself relies on a covering number bound of the set of admissible dual potentials (8), which can be used to separate the dependent G_n and the sample through a union bound over the covering. In particular, Mena and Niles-Weed (2019) show that if P_Z, P_Y, and Q_Y^n are all $\tilde{\sigma}^2$ sub-gaussian distributions, equivalently $\tilde{\sigma}^2$ is the maximum sub-gaussian parameter for these three distributions, then for $s = \lceil d/2 \rceil + 1$ and for a certain set of smooth functions $\mathcal{F}^s$ defined formally in proposition 9 it holds almost surely that

$$\left|W_{2,2}^2(P_Z, Q_Y^n) - W_{2,2}^2(P_Z, P_Y)\right| \leq 4 \sup_{f \in \mathcal{F}^s} \left|\mathbb{E}_{Y \sim P_Y} f(Y) - \mathbb{E}_{\hat{Y} \sim Q_Y^n} f(\hat{Y})\right| (1 + \tilde{\sigma}^{3s}). \qquad (38)$$

The sub-Gaussian parameter in (38) is random and depends on the sample. We can set $P_Z = P_{G_n(X)}$ as long as $P_{G_n(X)}$ is guaranteed to be sub-gaussian with some random parameter $\tilde{\sigma}$, which holds for any Lipschitz G_n. This bound comes from the dual form and the fact that one can choose the dual potential f that is smooth and satisfies $f/(1 + \tilde{\sigma}^{3s}) \in \mathcal{F}^s$. It is also important to note that $\mathcal{F}^s$ depends on s (and hence the dimension) but not on the sub-Gaussian parameter $\tilde{\sigma}$. We note that we used $\lambda = 2$ in (38), because by rescaling the distributions and the loss function one can obtain a bound for any λ based on the one for $\lambda = 2$ as we will show in the proof of Proposition 5.

Now, one can take the expectation of both sides of (38) and apply Cauchy-Schwartz inequality to completely separate the dependent $\tilde{\sigma}$ and Q_Y^n, which leads to

$$\begin{aligned} &\mathbb{E}\left|W_{2,2}^2(P_{G_n(X)}, Q_Y^n) - W_{2,2}^2(P_{G_n(Z)}, P_Y)\right| \\ &\quad \leq 4\sqrt{\mathbb{E}\left[\sup_{f \in \mathcal{F}^s} \left|\mathbb{E}_{Y \sim P_Y} f(Y) - \mathbb{E}_{\hat{Y} \sim Q_Y^n} f(\hat{Y})\right|^2\right] \mathbb{E}\left[(1 + \tilde{\sigma}^{3s})^2\right]}. \qquad (39) \end{aligned}$$

Here the first term in the product also appears in the proof of Theorem 2 by Mena and Niles-Weed (2019), for which the authors provide an upper bound which we state as Proposition 10:

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}^s}\left|\mathbb{E}_{Y\sim P_Y}f(Y)-\mathbb{E}_{\hat{Y}\sim Q_Y^n}f(\hat{Y})\right|^2\right]\leq C_d\frac{1}{n}(1+\sigma^{2d+4}),$$

where σ is the sub-gaussian parameter of the distribution P_Y and C_d is the dimension-dependent constant. Finally, redefining the constant, simplifying, and plugging the above into (38) gives

$$\begin{aligned}&\mathbb{E}\left|W_{2,2}^2(P_{G_n(X)},Q_Y^n)-W_{2,2}^2(P_{G_n(Z)},P_Y)\right|\\&\qquad\leq C_d\frac{1}{\sqrt{n}}(1+\sigma^{2d+4}\sqrt{\mathbb{E}\left[\tilde{\sigma}^{6s}\right]}).\end{aligned}$$

Now, the only thing left is to bound the $6s$-order moment of the sub-Gaussian parameter $\tilde{\sigma}^2=\max\{\sigma^2(G_n(X)),\sigma^2(Y),\sigma^2(\hat{Y})\}$ with $\hat{Y}\sim Q_Y^n$. By redefining C_d again one can simplify the above to

$$\begin{aligned}&\mathbb{E}\left|W_{2,2}^2(P_{G_n(X)},Q_Y^n)-W_{2,2}^2(P_{G_n(Z)},P_Y)\right|\\&\quad\leq C_d\frac{1}{\sqrt{n}}\left(1+\sigma^{2d+4}\max\left\{\mathbb{E}\left[\sigma^2(G_n(X))^{3s}\right],\mathbb{E}\left[\sigma^2(\hat{Y})^{3s}\right],\sigma^2(Y)^{3s}\right\}^{1/2}\right).\end{aligned}\tag{40}$$

The moments of the empirical distribution's sub-Gaussian parameter $\sigma^2(\hat{Y})$ were bounded by Mena and Niles-Weed (2019, Lemma 4) stated here as Proposition 11 and give the following: $\mathbb{E}\left[\sigma^2(\hat{Y})^{3s}\right]\leq 2(3s)^{3s}\sigma^2(Y)^{3s}$.

We next overview how we bound $\sigma^2(G_n)$ under the conditions of Theorem 6. First, we upper bound $\mathbb{E}_{X\sim P_X}[\|G_n(X)\|^2]$. Since G_n is the empirical solution, for any $G\in\mathcal{G}$ it holds that $W_{2,2}^2(P_{G_n(X)},Q_Y^n)\leq W_{2,2}^2(P_{G(X)},Q_Y^n)$. Now since $\alpha G_n\in\mathcal{G}$ is a feasible point by assumption (35), it holds that

$$1=\alpha^*=\operatorname{argmin}_{\alpha\in[0,1]}W_{2,2}^2(P_{\alpha G_n(X)},Q_Y^n)$$

as if $\alpha^*\neq 1$ then $G=\alpha G_n$ gives a smaller loss and G_n is not optimal. The optimality condition leads to $\mathbb{E}\left[\|G_n(X)\|^2\right]\leq\operatorname{Tr}K_{\hat{Y}}$, where $K_{\hat{Y}}$ is the covariance matrix of $\hat{Y}$ as we show in Lemma 12. Since G_n is also linear plugging it into the expectation gives a bound of $\sigma^2(G_n(X))\leq\operatorname{Tr}K_{\hat{Y}}\sigma^2(K_X^{-1/2}X)$ as also shown in Lemma 12. Finally the fact that $\operatorname{Tr}(K_{\hat{Y}})\leq 2d\sigma^2(\hat{Y})$ obtained by Jensen's inequality as shown in lemma 13 allows us to bound $\sigma^2(G_n(X))\leq\sigma^2(K_X^{-1/2}X)\sigma^2(\hat{Y})$. Plugging these bounds into (40) proves theorem 6. The bound of $\sigma^2(G_n(X))$ under the conditions of Theorem 7 is similar, but requires separating out the $G_n(0)$ in $W_{2,2}^2(P_{G_n(X)},Q_Y^n)$ and bounding the difference for the translated generator function $G_n(X)-G_n(0)$ and then establishing the rate for the term added by $G_n(0)$. We next proceed with the formal version of the proofs.

4.4 Proofs

In this section we use Q_Y^n to denote the random empirical distribution extracted from a sample $\hat{\mathcal{Y}} = \{y_i\}_{i=1}^n \sim P_Y^{\otimes n}$ unless stated otherwise. We also use $\mathbb{E}_{\hat{\mathcal{Y}}}[\cdot]$ denotes the expectation conditioned on the sample $\hat{\mathcal{Y}}$ and let

$$\sigma^2_{\hat{\mathcal{Y}}}(\hat{Y}) = \min\{\sigma \geq 0 \,\Big|\, \mathbb{E}_{\hat{\mathcal{Y}}} \exp(\|\hat{Y}\|^2/(2r\sigma^2)) \leq 2\}$$

be the sub-gaussian parameter of the distribution of $\hat{Y}$ conditioned on the sample.

Note that to be able to apply the result of Mena and Niles-Weed (2019) we first need to bound the sub-gaussian norm of the output distributions $\sigma^2(G^*(X))$ and $\sigma^2(G_n(X))$ since $\sup_{G\in\mathcal{G}} \sigma^2(G(X)) = \infty$. We do this in lemmas 12, 13. The rest of the proof follows the proof of Theorem 2 by Mena and Niles-Weed (2019) with the additional constraint that $\sigma^2_{\hat{\mathcal{Y}}}(G_n(X))$ and Q_Y^n are dependent.

To prove the theorem we will need several intermediate results from the proof of Theorem 2 by Mena and Niles-Weed (2019) and two lemmas specific to our setting, and we state them in the following subsection.

4.4.1 Facts Introduced by Mena and Niles-Weed (2019) and Additional Lemmas

Proposition 8 *(Mena and Niles-Weed, 2019, Proposition 2) Let P_X, P_Y and Q_Y^n all be $\tilde{\sigma}^2$ sub-gaussian distributions for a possibly random $\tilde{\sigma}^2 \in [0; +\infty)$. Then for a set of functions F denoting $\|P_Y - Q_Y^n\|_F = \sup_{f\in F} |\mathbb{E}_{Y\sim P_Y}[f(Y)] - \mathbb{E}_{\hat{Y}\sim Q_Y^n}[f(\hat{Y})]|$ it holds that*

$$(1/2)\left|W_{2,2}^2(P_X, Q_Y^n) - W_{2,2}^2(P_X, P_Y)\right| \leq 2\|P_Y - Q_Y^n\|_{\mathcal{F}_{\tilde{\sigma}}},$$

where $\mathcal{F}_{\tilde{\sigma}}$ is a set of functions satisfying for some constants $C_{k,d}$, depending on k and d only and for any multi-index α with $|\alpha| = k$

$$|D^\alpha(f - (1/2)\|\cdot\|^2)(x)| \leq C'_{k,d} \begin{cases} 1+\sigma^4 & \textit{if } k=0 \\ \sigma^k(\sigma+\sigma^2)^k & \textit{otherwise} \end{cases}$$

if $\|x\| \leq \sqrt{d}\sigma$ and

$$\begin{aligned} &|D^\alpha(f - (1/2)\|\cdot\|^2)(x)| \\ &\qquad \leq C'_{k,d} \begin{cases} 1+(1+\sigma^2)\|x\|^2 & \textit{if } k=0 \\ \sigma^k(\sqrt{\sigma\|x\|}+\sigma\|x\|)^k & \textit{otherwise} \end{cases} \end{aligned}$$

if $\|x\| > \sqrt{d}\sigma$.

Note that we also included the definition of $\mathcal{F}_{\tilde{\sigma}}$ from Proposition 1 of Mena and Niles-Weed (2019) into the above. The proposition cannot be used directly for proving the result since the norm depends on the random sub-gaussian parameter $\tilde{\sigma}$. To overcome that we use the following proposition that will help decouple the norm and the random sub-gaussian parameter.

Proposition 9 *Let P_X, P_Y and Q_Y^n all be $\tilde{\sigma}^2$ sub-gaussian distributions for a possibly random $\tilde{\sigma}^2 \in [0; +\infty)$. Let for $s \geq 2$ $\mathcal{F}^s$ be a set of functions satisfying*

$$|f(x)| \leq C_{s,d}(1 + \|x\|^2)$$
$$|D^\alpha f(x)| \leq C_{s,d}(1 + \|x\|^s)\, \forall \alpha : |\alpha| \leq s$$

for some constant $C_{s,d}$ that depends only on s, d. Then

$$\left|W_{2,2}^2(P_X, Q_Y^n) - W_{2,2}^2(P_X, P_Y)\right| \leq 4\|P_Y - Q_Y^n\|_{\mathcal{F}^s}(1 + \tilde{\sigma}^{3s})$$

The proof of the proposition follows some of the steps of the proof of Theorem 2 by Mena and Niles-Weed (2019) and is provided here for completeness.

Proof Note that for large enough constants $C_{s,d}$ ($C_{s,d} \propto d + \max_{k \leq s} 2^k C'_{k,d}$, where $C'_{k,d}$ come from the definition of $\mathcal{F}_{\tilde{\sigma}^2}$) for any $f \in \mathcal{F}_{\tilde{\sigma}^2}$ it holds that $\frac{1}{1+\tilde{\sigma}^{3s}} f \in \mathcal{F}^s$. Combining this with Proposition 8 we get

$$\begin{aligned}
&\left|W_{2,2}^2(P_X, Q_Y^n) - W_{2,2}^2(P_X, P_Y)\right| \\
&\quad \leq 4\|P_Y - Q_Y^n\|_{\mathcal{F}_{\tilde{\sigma}}} \\
&\quad = (1 + \tilde{\sigma}^{3s}) \sup_{f \in \mathcal{F}_{\tilde{\sigma}}} \left|\mathbb{E}_{Y \sim P_Y}\left[\frac{f(Y)}{1 + \tilde{\sigma}^{3s}}\right] - \mathbb{E}_{\hat{Y} \sim Q_Y^n}\left[\frac{f(\hat{Y})}{1 + \tilde{\sigma}^{3s}}\right]\right| \\
&\quad \leq (1 + \tilde{\sigma}^{3s}) \sup_{f \in \mathcal{F}^s} |\mathbb{E}_{Y \sim P_Y}[f(Y)] - \mathbb{E}_{\hat{Y} \sim Q_Y^n}[f(\hat{Y})]| \\
&\quad = 4\|P_Y - Q_Y^n\|_{\mathcal{F}^s}(1 + \tilde{\sigma}^{3s})
\end{aligned}$$

■

The proof of Mena and Niles-Weed (2019, Theorem 2) also uses a covering number for $\mathcal{F}_s$ to bound $\mathbb{E}[\|P_Y - Q_Y^n\|_{\mathcal{F}^s}^2]$. Since the result will be used in the proofs of Theorems 6 and 7, we will state it here.

Proposition 10 *(Mena and Niles-Weed 2019, Proof of Theorem 2, page 8; Giné and Nickl 2021) For $s = \lceil d/2 \rceil + 1$, for P_Y being σ^2 sub-gaussian and $\mathcal{F}_s$ defined in proposition 9 it holds that*

$$\mathbb{E}[\|P_Y - Q_Y^n\|_{\mathcal{F}^s}^2] \leq C_d \frac{1}{n}(1 + \sigma^{2d+4})$$

Finally, we state here (Mena and Niles-Weed, 2019, Lemma 4) that helps bound the even moments of the sub-gaussian parameter $\tilde{\sigma}^2$ of the (random) empirical distribution Q_Y^n.

Proposition 11 *(Mena and Niles-Weed, 2019, Lemma 4) If Y is σ^2 sub-gaussian then Q_Y^n is $\tilde{\sigma}^2$ sub-gaussian with*

$$\mathbb{E}[\tilde{\sigma}^{2k}] \leq 2k^k \sigma^{2k}$$

for any positive integer k,

To prove the theorem we will also need the following lemmas connected to the properties of $G(X)$.

Lemma 12 *Under the assumption* (35) *the optimal generator*

$$G^* = \arg\min_{G\in\mathcal{G}} W_{2,\lambda}^2(P_{G(X)}, P_Z)$$

for $Z \in \mathbb{R}^d$ *satisfies* $\mathbb{E}\left[\|G^*(X)\|_2^2\right] \leq \operatorname{Tr} K_Z$, *and if* G^* *is linear then* $G^*(X)$ *is sub-Gaussian with* $\sigma^2(G^*(X)) \leq rd^{-1} \operatorname{Tr} K_Z \sigma^2(K_X^{-1/2}X)$.

Proof Assume that $g^2 = \mathbb{E}\left[\|G^*(X)\|_2^2\right] > 0$. If assumption (35) holds then for any $\alpha \in [0,1] : \alpha G^* \in \mathcal{G}$. Consider $\tilde{G}^*(X) = G^*(X)/g$. By optimality of G^* for the optimal coupling π^* :

$$\begin{aligned} g &= \arg\min_{\alpha\in[0,g]} W_{2,\lambda}^2(P_{\alpha\tilde{G}^*(X)}, P_Z) \\ &= \arg\min_{\alpha\in[0,g]} \mathbb{E}_{X,Z\sim\pi^*}\left[\|\alpha\tilde{G}^*(X) - Z\|_2^2\right] + \lambda I(\tilde{G}(X); Z) \\ &= \arg\min_{\alpha\in[0,g]} \alpha^2 + \mathbb{E}\left[\|Z\|^2\right] - 2\alpha\mathbb{E}_{X,Z\sim\pi^*}\left[\tilde{G}^*(X)^T Z\right] + \lambda I(\tilde{G}(X); Z) \end{aligned}$$

The above problem is minimization of a quadratic function thus

$$g = \alpha^* = \min\left\{g, \mathbb{E}_{X,Z\sim\pi^*}\left[\tilde{G}^*(X)^T Z\right]\right\} \leq \sqrt{\operatorname{Tr} K_Z},$$

so $\mathbb{E}\left[\|G^*(X)\|_2^2\right] \leq \operatorname{Tr} K_Z$. For a linear $G^* : \mathbb{E}\left[\|G^* X\|_2^2\right] = \operatorname{Tr} G^{*T} G^* K_X = \|G^* K_X^{1/2}\|_F^2 \leq \operatorname{Tr} K_Z$ for $\tau^2 = \frac{\operatorname{Tr} K_Z r}{d}\sigma^2\left(K_X^{-1/2}X\right)$

$$\mathbb{E}\, e^{\frac{\|G^* X\|_2^2}{2d\tau^2}} = \mathbb{E}\, e^{\frac{\|G^* K_X^{1/2} K_X^{-1/2} X\|_2^2}{2d\tau^2}} \leq \mathbb{E}\, e^{\frac{\operatorname{Tr} K_Z \|K_X^{-1/2} X\|_2^2}{2d\tau^2}} = \mathbb{E}\, e^{\frac{\|K_X^{-1/2} X\|_2^2}{2r\sigma^2\left(K_X^{-1/2} X\right)}} \leq 2$$

■

Lemma 13 *For a sub-gaussian* $Z \in \mathbb{R}^d$ *the covariance matrix trace is bounded as* $\operatorname{Tr} K_Z \leq 2d\sigma^2(Z)$.

Proof

$$\ln 2 \geq \ln \mathbb{E}\, e^{\frac{\|Z\|_2^2}{2d\sigma^2(Z)}} \geq \ln e^{\mathbb{E}\frac{\|Z\|_2^2}{2d\sigma^2(Z)}} = \operatorname{Tr} K_Z / \left(2d\sigma^2(Z)\right).$$

The first inequality follows from Z being sub-Gaussian and the second one is Jensen's inequality. ■

4.4.2 Proof of Theorems 6 and 7

Proof [Proof of Theorem 6] The proof is based on the proof of Theorem 2 by Mena and Niles-Weed (2019). 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 $\lambda = 2$, we will generalize to $\lambda \neq 2$ exacly as we did in the proof of Proposition 5. First, we rewrite $d_\lambda(G^*, G_n) = W_{2,\lambda}^2(P_{G_n(X)}, P_Y) - W_{2,\lambda}^2(P_{G^*(X)}, P_Y)$ to fit Proposition 5:

$$\begin{aligned} d_\lambda(G^*, G_n) &= \big(W_{2,\lambda}^2(P_{G^*(X)}, Q_Y^n) - W_{2,\lambda}^2(P_{G^*(X)}, P_Y)\big) \\ &\quad + \big(W_{2,\lambda}^2(P_{G_n(X)}, P_Y) - W_{2,\lambda}^2(P_{G^*(X)}, Q_Y^n)\big) \\ &\leq \big(W_{2,\lambda}^2(P_{G^*(X)}, Q_Y^n) - W_{2,\lambda}^2(P_{G^*(X)}, P_Y)\big) \\ &\quad + \big(W_{2,\lambda}^2(P_{G_n(X)}, P_Y) - W_{2,\lambda}^2(P_{G_n(X)}, Q_Y^n)\big) \end{aligned} \tag{41}$$

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

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

with the inequalities following from Lemmas 12, 13 and the definition of ν^2. By Proposition 5 applied to the expectation of the first difference in (41):

$$\begin{aligned} &\mathbb{E}\left[\left|W_{2,\lambda}^2(P_{G^*(X)}, Q_Y^n) - W_{2,\lambda}^2(P_{G^*(X)}, P_Y)\right|\right] \\ &\quad \leq C_{d,2} n^{-1/2}\big(1 + \big(\nu^2\big)^{\lceil 5d/4\rceil + 3}\big), \end{aligned} \tag{42}$$

As G_n depends on the sample, the proposition cannot be applied directly to the second difference, but by Proposition 9 for $\tilde{\sigma}^2 = \max\big\{\sigma_{\hat{\mathcal{Y}}}^2\left(G_n(X)\right), \sigma_{\hat{\mathcal{Y}}}^2(\hat{Y}), \sigma^2(Y)\big\}$ and $s = \lceil d/2 \rceil + 1$:

$$\begin{aligned} &W_{2,\lambda}^2(P_{G_n(X)}, P_Y) - W_{2,\lambda}^2(P_{G_n(X)}, Q_Y^n) \\ &\quad \leq 4\left(1 + \tilde{\sigma}^{3s}\right) \|P_Y - Q_Y^n\|_{\mathcal{F}^s}, . \end{aligned} \tag{43}$$

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 (43) we get:

$$\begin{aligned} &\big(\mathbb{E}\big[W_{2,\lambda}^2(P_{G_n(X)}, P_Y) - W_{2,\lambda}^2(P_{G_n(X)}, Q_Y^n)\big]\big)^2 \\ &\quad \leq 8\mathbb{E}\left[1 + \tilde{\sigma}^{6s}\right] \mathbb{E}\|P_Y - Q_Y^n\|_{\mathcal{F}^s}^2 \\ &\quad \leq \Big(1 + \sigma^2(Y)^{d+2}\Big) n^{-1} C_{d,3} \mathbb{E}\left[1 + \tilde{\sigma}^{6s}\right] \qquad (44) \\ &\quad \leq \Big(1 + \nu^{2d+4}\Big) n^{-1} C_{d,3} \mathbb{E}\left[1 + \tilde{\sigma}^{6s}\right], \qquad (45) \end{aligned}$$

where (44) follows from Proposition 10 and (45) from the definition of ν. By Lemma 13 we have $\operatorname{Tr} K_{\hat{Y}} \leq 2d\sigma^2(\hat{Y})$, so

$$\begin{aligned} \sigma_{\hat{\mathcal{Y}}}^2(G_n(X)) &\leq d^{-1} \operatorname{Tr} K_{\hat{Y}} r\sigma^2\big(K_X^{-1/2}X\big) \\ &\leq 2r\sigma^2\big(K_X^{-1/2}X\big)\sigma_{\hat{\mathcal{Y}}}^2\big(\hat{Y}\big) \leq \sigma_{\hat{\mathcal{Y}}}^2(\hat{Y})\nu^2/\sigma^2(Y), \end{aligned} \tag{46}$$

where the first inequality follows from Lemma 12 and the second one from Lemma 13. Taking expectation of $\tilde{\sigma}^{6s}$:

$$\begin{aligned}\mathbb{E}\big[\tilde{\sigma}^{6s}\big] &= \mathbb{E}\big[\max\{\sigma^2_{\hat{\mathcal{Y}}}(\hat{Y}), \sigma^2(Y), \sigma^2_{\hat{\mathcal{Y}}}(G_n(X))\}^{3s}\big]\\ &\leq \nu^{6s}\mathbb{E}\Big[\max\{1, \sigma^2_{\hat{\mathcal{Y}}}(\hat{Y})/\sigma^2(Y)\}^{3s}\Big] \leq 2(3s)^{3s}\nu^{6s}, \end{aligned} \tag{47}$$

where (47) is due to Proposition 11; plugging (47) in (45) gives

$$\begin{aligned}&\mathbb{E}\big[W^2_{2,\lambda}(P_{G_n(X)}, P_Y) - W^2_{2,\lambda}(P_{G_n(X)}, Q^n_Y)\big]\\ &\quad\leq \sqrt{2(1 + 2(3s)^{3s}\nu^{6s})C_{d,3}n^{-1}\left(1 + \nu^{d+2}\right)}\\ &\quad\leq C_{d,4}n^{-1/2}\big(1 + \big(\nu^2\big)^{\lceil 5d/4\rceil+3}\big) \end{aligned} \tag{48}$$

Combining (48) and (42) we get 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 $\lambda > 0$:

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

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 (48) we get

$$\mathbb{E}\left[d_\lambda(G^*, G_n)\right] \leq K_d\lambda n^{-1/2}\big(1 + (2\tau^2/\lambda)^{\lceil 5d/4\rceil+3}\big)/2. \tag{49}$$

■

Proof [Proof of Theorem 7] The proof follows the same path as the proof of Theorem 6 with the only difference being in bounding the sub-Gaussian parameters.

For $G \in \mathcal{G}$ let $\mathring{G}(X) = G(X) - G(0)$ – a shifted function. Note that $\mathring{G}$ need not be in $\mathcal{G}$. To avoid confusion we let $H^* = \operatorname{argmin}_{H\in\mathcal{G}} S_\lambda(P_{H(X)}, P_Y)$ and $H_n = \operatorname{argmin}_{H\in\mathcal{G}} S_\lambda(P_{H(X)}, Q^n_Y)$ – the population and empirical solutions to Sinkhorn W2GANs. Since

$$\begin{aligned}\mathbb{E}\left[\|G(X) - Y\|^2\right] &= \mathbb{E}\left[\|\mathring{G}(X) - Y + G(0)\|^2\right]\\ &= \mathbb{E}\left[\|\mathring{G}(X) - Y\|^2\right] + 2G(0)^T\mathbb{E}[\mathring{G}(X) - Y] + \|G(0)\|^2\end{aligned}$$

and $I(G(X),Y) = I(\mathring{G}(X),Y)$, entropy-regularized Wasserstein distance decomposes as $W_{2,\lambda}^2(P_{G(X)},P_Y) = W_{2,\lambda}^2(P_{\mathring{G}(X)},P_Y) + 2G(0)^T\mathbb{E}[\mathring{G}(X)-Y] + \|G(0)\|^2$ As in the proof of Theorem 6 we decompose the excess risk:

$$\begin{aligned}
d_\lambda(G^*,G_n) &= W_{2,\lambda}^2(P_{G_n(X)},P_Y) - W_{2,\lambda}^2(P_{G^*(X)},P_Y) \\
&\le W_{2,\lambda}^2(P_{G^*(X)},Q_Y^n) - W_{2,\lambda}^2(P_{G^*(X)},P_Y) \\
&\quad + W_{2,\lambda}^2(P_{G_n(X)},P_Y) - W_{2,\lambda}^2(P_{G_n(X)},Q_Y^n) \\
&= \Big(W_{2,\lambda}^2(P_{\mathring{G}^*(X)},Q_Y^n) - W_{2,\lambda}^2(P_{\mathring{G}^*(X)},P_Y)\Big) \\
&\quad + \Big(W_{2,\lambda}^2(P_{\mathring{G}_n(X)},P_Y) - W_{2,\lambda}^2(P_{\mathring{G}_n(X)},Q_Y^n)\Big) \\
&\quad + 2\left(G^*(0)-G_n(0)\right)^T\Big(\mathbb{E}Y - \mathbb{E}_{\hat{\mathcal{Y}}}\hat{Y}\Big), \qquad (50)
\end{aligned}$$

where the last inequality follows as (41). Similarly, for Sinkhorn W2GAN the excess risk is:

$$\begin{aligned}
d_\lambda^S(H^*,H_n) &= S_\lambda(P_{H_n(X)},P_Y) - S_\lambda(P_{H^*(X)},P_Y) \\
&\le \big(S_\lambda(P_{H^*(X)},Q_Y^n) - S_\lambda(P_{H^*(X)},P_Y)\big) \\
&\quad + \big(S_\lambda(P_{H_n(X)},P_Y) - S_\lambda(P_{H_n(X)},Q_Y^n)\big) \\
&= \big(W_{2,\lambda}^2(P_{H^*(X)},Q_Y^n) - W_{2,\lambda}^2(P_{H^*(X)},P_Y)\big) \\
&\quad + \big(W_{2,\lambda}^2(P_{H_n(X)},P_Y) - W_{2,\lambda}^2(P_{H_n(X)},Q_Y^n)\big) \\
&= \Big(W_{2,\lambda}^2(P_{\mathring{H}^*(X)},Q_Y^n) - W_{2,\lambda}^2(P_{\mathring{H}^*(X)},P_Y)\Big) \\
&\quad + \Big(W_{2,\lambda}^2(P_{\mathring{H}_n(X)},P_Y) - W_{2,\lambda}^2(P_{\mathring{H}_n(X)},Q_Y^n)\Big) \\
&\quad + 2\left(H^*(0)-H_n(0)\right)^T\Big(\mathbb{E}Y - \mathbb{E}_{\hat{\mathcal{Y}}}\hat{Y}\Big) \qquad (51)
\end{aligned}$$

The RHS of (51) and (50) are the same as the RHS of (41). Note that for any $G \in \mathcal{G}$ and for $\sigma^2 = \sigma^2(X)rL^2/d$ by the L-Lipschitzness of $G:$

$$\mathbb{E}\, e^{\frac{\|\mathring{G}(X)\|_2^2}{2d\sigma^2}} = \mathbb{E}\, e^{\frac{\|G(X)-G(0)\|_2^2}{2d\sigma^2}} \le \mathbb{E}\, e^{\frac{L^2\|X\|_2^2}{2d\sigma^2}} = \mathbb{E}\, e^{\frac{\|X\|_2^2}{2r\sigma^2(X)}} \le 2,$$

thus $\mathring{G}(X)$ and $\mathring{H}(X)$ are both sub-Gaussian, $\max\{\sigma^2(\mathring{G}(X)), \sigma^2(\mathring{H}(X))\} \le \sigma^2(X)rL^2/d \le \tau^2$.

The next part of the proof follows the proof of Theorem 6 with $\nu^2 = \tau^2$, and $\mathring{G}_n$ and $\mathring{H}_n$ in place of G_n, $\mathring{G}^*$ and $\mathring{H}^*$ in place of G^* for the entropic and Sinkhorn W2GAN cases respectively. Indeed, for entropic W2GAN eqs. (42), (43) and (45) only require that $\max\{\sigma^2(\mathring{G}_n(X)), \sigma^2(\mathring{G}^*(X)), \sigma^2(Y)\} \le \nu^2$. As $\sigma^2_{\hat{\mathcal{Y}}}(\mathring{G}_n(X)) \le L^2\sigma^2(X)$, in place of (47) we have for $\tilde{\sigma}^2 = \max\big\{\sigma^2_{\hat{\mathcal{Y}}}\Big(\mathring{G}_n(X)\Big), \sigma^2_{\hat{\mathcal{Y}}}(\hat{Y}), \sigma^2(Y)\big\}$ and $s = \lceil d/2 \rceil + 1:$

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

where the last inequality is due to Proposition 11. So, eq. (48) and (49) hold, i.e.

$$\begin{aligned}\mathbb{E}d_\lambda(G^*, G_n) &\leq K_d \lambda n^{-1/2}\big(1+(2\tau^2/\lambda)^{\lceil 5d/4\rceil+3}\big)/2 + 2\mathbb{E}\left[(G^*(0)-G_n(0))^T\left(\mathbb{E}Y - \mathbb{E}_{\hat{\mathcal{Y}}}\hat{Y}\right)\right]\\ &\leq K_d \lambda n^{-1/2}\big(1+(2\tau^2/\lambda)^{\lceil 5d/4\rceil+3}\big)/2 + 2\sqrt{\mathbb{E}\|G_n(0)\|_2^2}\sqrt{\operatorname{Tr} K_Y/n}, \qquad (52)\end{aligned}$$

where the last inequality follows from the independence of $G^*(0)$ and the sample and the Cauchy-Schwarz inequality.

For Sinkhorn W2GAN the above results in

$$\mathbb{E}d_\lambda^S(H^*, H_n) \leq K_d \lambda n^{-1/2}\big(1+(2\tau^2/\lambda)^{\lceil 5d/4\rceil+3}\big)/2 + 2\sqrt{\mathbb{E}\|H_n(0)\|_2^2}\sqrt{\operatorname{Tr} K_Y/n} \qquad (53)$$

We will now bound the last term of (52) via Lemma 12 and Lipschitzness of G :

$$\mathbb{E}\|G_n(0)\|_2^2 \leq 2\mathbb{E}\|G_n(X)-G_n(0)\|_2^2 + 2\mathbb{E}\|G_n(X)\|_2^2 \leq 2L^2 \operatorname{Tr} K_X + 2\operatorname{Tr} K_Y \leq 8d\tau^2,$$

where the last inequality follows from Lemma 13. (52) thus becomes:

$$\begin{aligned}\mathbb{E}d_\lambda(G^*, G_n) &\leq K_d \lambda n^{-1/2}\big(1+(2\tau^2/\lambda)^{\lceil 5d/4\rceil+3}\big)/2 + 2\sqrt{8d\tau^2 \operatorname{Tr} K_Y/n}\\ &\leq K_d \lambda n^{-1/2}\big(1+(2\tau^2/\lambda)^{\lceil 5d/4\rceil+3}\big)/2 + 8d\tau^2/\sqrt{n}, \qquad (54)\end{aligned}$$

where the last inequality flows from Lemma 13. Redefining K_d completes the proof for Entropic W2GAN excess risk.

To complete the proof we need to bound $\mathbb{E}\|H_n(0)\|_2^2$. We first note that $0 \in \mathcal{G}$, so by optimality of H_n :

$$S_\lambda(P_{H_n(X)}, Q_Y^n) \leq S_\lambda(\delta_0, Q_Y^n) = \mathbb{E}_{\hat{\mathcal{Y}}}\|\hat{Y}\|_2^2 - W_{2,\lambda}^2(Q_Y^n, Q_Y^n)/2 \leq \mathbb{E}_{\hat{\mathcal{Y}}}\|\hat{Y}\|_2^2, \qquad (55)$$

where δ_0 is a point mass at 0. As $S_\lambda(P_{\mathring{H}_n(X)}, Q_Y^n) \geq 0$:

$$\begin{aligned}S_\lambda(P_{H_n(X)}, Q_Y^n) &= S_\lambda(P_{\mathring{H}_n(X)}, Q_Y^n) + \|H(0)\|_2^2 + 2H(0)^T\mathbb{E}_{\hat{\mathcal{Y}}}\left[\hat{Y} - \mathring{H}_n(X)\right]\\ &\geq \|H_n(0)\|_2^2 + 2H_n(0)^T\mathbb{E}_{\hat{\mathcal{Y}}}\left[\hat{Y} - \mathring{H}_n(X)\right] \qquad (56)\end{aligned}$$

Combining (55) and (56) and taking the expectation over the sample $\hat{\mathcal{Y}}$ we get:

$$\begin{aligned}2d\tau^2 \geq \operatorname{Tr} K_Y &= \mathbb{E}\mathbb{E}_{\hat{\mathcal{Y}}}\|\hat{Y}\|_2^2\\ &\geq \mathbb{E}\|H_n(0)\|_2^2 - 2\mathbb{E}\left[\|H_n(0)\|\mathbb{E}_{\hat{\mathcal{Y}}}\left[\|\hat{Y}\| + L\|X\|\right]\right]\\ &\geq \mathbb{E}\|H_n(0)\|_2^2 - 2\sqrt{\mathbb{E}\|H_n(0)\|_2^2}(\sqrt{\operatorname{Tr} K_Y} + L\sqrt{\operatorname{Tr} K_X})\\ &\geq \mathbb{E}\|H_n(0)\|_2^2 - 4\sqrt{\mathbb{E}\|H_n(0)\|_2^2}\sqrt{2d\tau^2}\end{aligned}$$

The above inequality implies that $\mathbb{E}\|H_n(0)\|_2^2 \leq 40d\tau^2$. From (53):

$$\begin{aligned}\mathbb{E}d_\lambda^S(H^*, H_n) &\leq K_d \lambda n^{-1/2}\big(1+(2\tau^2/\lambda)^{\lceil 5d/4\rceil+3}\big)/2 + 2\sqrt{\mathbb{E}\|H_n(0)\|_2^2}\sqrt{\operatorname{Tr} K_Y/n}\\ &\leq K_d \lambda n^{-1/2}\big(1+(2\tau^2/\lambda)^{\lceil 5d/4\rceil+3}\big)/2 + 2\sqrt{40d\tau^2}\sqrt{2d\tau^2/n}\\ &\leq K_d \lambda n^{-1/2}\big(1+(2\tau^2/\lambda)^{\lceil 5d/4\rceil+3}\big)/2 + 20d\tau^2/\sqrt{n}\end{aligned}$$

Redefining K_d completes the proof. ■

5. Computational Convergence

For the sake of completeness, in this section, we discuss some results on the computational convergence of entropic optimal transport and Sinkhorn divergence and emphasize the advantages of these regularization methods from an optimization perspective. A more detailed discussion is given by Sanjabi et al. (2018); Feydy et al. (2019). As was previously mentioned, entropic regularization makes the problem strongly convex, which in turn facilitates convergence. Note that since the optimal solution to (4) is known to satisfy (12),(11), Sinkhorn-Knopp iterates (13),(14) or any other method can be used to solve the inner problem close to optimality. In contrast, computing the unregularized optimal transport requires the use of linear programming techniques which are computationally infeasible in many machine learning applications.

Moreover, Sanjabi et al. (2018, Theorem 3.1) show that under mild conditions on the generator set $\mathcal{G}$ and the distributions of P_Y, P_X, entropy-regularized Wasserstein distance is Lipschitz smooth, i.e. has a Lipschitz continuous gradient with respect to the parameters of the generator. If we let the generator set $\mathcal{G}$ be parametrized by $\theta \in \Theta$, i.e. $\mathcal{G} = \{G_\theta \mid \theta \in \Theta\})$ then

$$| \bigtriangledown_\theta \, W_{2,\lambda}^2(P_{G_{\theta_1}(X)}, P_Y) - \bigtriangledown_\theta W_{2,\lambda}^2(P_{G_{\theta_2}(X)}, P_Y)| \leq L\|\theta_1 - \theta_2\|,$$

where L is a constant depending on $P_X, P_Y, \mathcal{G}$ and λ. Moreover, the optimal coupling $\pi^*(\theta)$ is a Lipschitz continuous function of θ :

$$\|\pi^*(\theta_1) - \pi^*(\theta_2)\|_1 \leq \frac{L_0}{\lambda}\|\theta_1 - \theta_2\|$$

The above indicates that small changes in the generator parameter θ result in small changes in the optimal coupling. Therefore, after the gradient step on the generator parameters θ, finding the regularized Wasserstein distance is easier since the discriminator parameters from the previous step are close to the optimal ones for the current step, while the Lipschitz smoothness of regularized Wasserstein distance in θ results in faster convergence of optimization.

Note that first-order optimization methods commonly used for neural network optimization require calculating the gradients $\bigtriangledown_\theta W_{2,\lambda}^2(P_{G_\theta(X)}, P_Y)$, which requires knowing the optimal dual potentials, but since they are found numerically, they can only be computed up to some positive accuracy, so the smoothness of the gradient of the entropic Wasserstein distance with respect to the accuracy up to which the dual potentials are calculated plays a crucial role in the convergence of the optimization. More precisely, if the inner problem of calculating Entropic Wasserstein distance is solved up to a certain accuracy ϵ, it can be shown that the gradient step on the outer problem of finding G is $O(\sqrt{\epsilon/\lambda})$-close to optimal, which makes training stable (see Sanjabi et al., 2018, Theorem 4.1). In contrast, the training of W2GAN, i.e. based on the unregularized squared Wasserstein distance, is known to be unstable even for the linear generator G and quadratic discriminator (Feizi et al., 2017) when $r < d$, and the training methods for Wasserstein GAN (Arjovsky et al., 2017; Gulrajani et al., 2017) do not converge locally with simultaneous or alternating gradient descent (Mescheder et al., 2018).

Finally, we note the following optimization convergence result by Sanjabi et al. (2018, Theorem 4.2.). Under mild conditions on $\mathcal{G}, P_X, P_Y$ it can be shown that when stochastic gradient is used to solve

$$\min_{\theta} f(\theta) = \min_{\theta} W_{2,\lambda}^2(P_{G_\theta(X)}, P_Y),$$

where $G(\cdot)$ is parametrized by θ, the random iterates $\theta_1, \dots, \theta_T$ satisfy

$$\min_{t=1\dots T} \mathbb{E}[\|\nabla f(\theta_t)\|^2] \leq O(1/\sqrt{T}) + O(\epsilon/\lambda).$$

Here constants in $O(\cdot)$ depend on the class of generators $G \in \mathcal{G}$ and the distributions P_X, P_Y, and T is the number of iterations of stochastic gradient descent and ϵ is the precision, to which the inner problem is solved. The expectation is over the randomness in the algorithm. The theorem implies that if there are enough iterations to get the discriminator close to optimality the training reaches a stable point of small $\mathbb{E}[\|\nabla f(\theta_t)\|^2]$. Since Sinkhorn divergence is a linear combination of entropy-regularized Wasserstein distances, a similar result holds for it and the optimization is stable.

6. Experiments

In our experiments we aim to contrast and compare the performance of Sinkhorn GAN (label: SGAN) and 1-Wasserstein GAN WGAN (label: WGAN) for linear generators. Entropic W2GAN is omitted from the comparison due to the fact that it leads to a biased solution as shown in Theorem 1. Following the experimental evaluations of Feizi et al. (2017), we generate $n = 10^5$ samples from a $d = 32$ dimensional Gaussian distribution $\mathcal{N}(0, K)$ where K is a random positive semi-definite matrix normalized to have Frobenius norm 1. We train WGAN with weight clipping (Arjovsky et al., 2017) labeled WGAN-WC, and WGAN with gradient penalty (Gulrajani et al., 2017) labeled WGAN-GP—two common methods to ensure Lipschitzness of the discriminators. We use the linear generator and a neural network discriminator with hyper-parameter settings as recommended by Gulrajani et al. (2017). The discriminator neural network has three hidden layers, each with 64 neurons and ReLU activation functions.

The pseudocode of our optimization for Sinkhorn GAN can be found in Algorithm 1. The algorithm is similar to the algorithm of Sanjabi et al. (2018), where we assume that the generators are parametrized by θ, i.e. $G(X) = G_\theta(x)$ and we apply stochastic gradient descent on θ. Note that at every step of the gradient descent algorithm, we need to calculate the gradient of the Sinkhorn divergence, $\nabla_\theta S_\lambda(P_{G_\theta(X)}, P_Y)$. From the definition of the Sinkhorn divergence in (6), to compute $\nabla_\theta S_\lambda(P_{G_\theta(X)}, P_Y)$ we need to compute $\nabla_\theta W_{2,\lambda}^2(P_{G_\theta(X)}, P_{G_\theta(X)})$ and $\nabla_\theta W_{2,\lambda}^2(P_{G_\theta(X)}, P_Y)$. (The third term $W_{2,\lambda}^2(P_Y, P_Y)$ is irrelevant since it doesn't depend on the generator.) From (8) follows the dual representation:

$$\begin{aligned} W_{2,\lambda}^2(P_{G_\theta(X)}, P_Y) = &\sup_{\substack{\psi \in L_\infty(\mathcal{Y}) \\ \phi \in L_\infty(G_\theta(\mathcal{X}))}} \mathbb{E}\left[\psi(Y) + \phi(G_\theta(X))\right] + \lambda \\ &- \lambda \mathbb{E}_{(X,Y)\sim P_X \times P_Y}\left[e^{\frac{\phi(G_\theta(X)) + \psi(Y) - \|G_\theta(X) - Y\|_2^2}{\lambda}}\right] \end{aligned} \tag{57}$$

$$\begin{aligned} W_{2,\lambda}^2(P_{G_\theta(X)}, P_{G_\theta(X)}) = &\sup_{\phi^x \in L_\infty(G_\theta(\mathcal{X}))} 2\mathbb{E}\left[\phi^x(G_\theta(X))\right] + \lambda \\ &- \lambda \mathbb{E}_{(X_1,X_2)\sim P_X \times P_X} e^{\frac{\phi^x(G_\theta(X_1)) + \phi^x(G_\theta(X_2)) - \|G_\theta(X_1) - G_\theta(X_2)\|_2^2}{\lambda}}, \end{aligned} \tag{58}$$

Now assume that we have access to approximations of the optimal dual potentials for (57), which for simplicity we also denote by ϕ and ψ. Then by using (10), we can obtain an approximation of the optimal coupling given by

$$\begin{aligned}\pi(G_\theta(x), y) &= P_{G_\theta(X)}(G_\theta(x))P_Y(y)e^{\frac{\phi(G_\theta(x))+\psi(y)-\|G_\theta(x)-y\|_2^2}{\lambda}}\\ &= P_{G_\theta(X)}(G_\theta(x))P_Y(y)\mu(x,y). \end{aligned}\tag{59}$$

Using the above in place of the coupling in the primal formulation of the entropic 2-Wasserstein distance we can then compute an approximation of the desired gradient

$$\nabla_\theta W_{2,\lambda}^2(P_{G_\theta(X)}, P_Y) \approx \mathbb{E}_{(X,Y)\sim P_X\times P_Y}[\mu(X,Y)\nabla_\theta(\|G_\theta(X)-Y\|^2)].$$

Analogously, if $\phi^x(\cdot)$ is an approximate optimal dual potential for (58) then

$$\nabla_\theta W_{2,\lambda}^2(P_{G_\theta(X)}, P_{G_\theta(X)}) \approx \mathbb{E}_{(X_1,X_2)\sim P_X\times P_X}[\mu^x(X_1,X_2)\nabla_\theta(\|G_\theta(X_1)-G_\theta(X_2)\|^2)],$$

where

$$\mu^x(x_1,x_2) = e^{\frac{\phi^x(G_\theta(x_1))+\phi^x(G_\theta(x_2))-\|G_\theta(x_1)-G_\theta(x_2)\|_2^2}{\lambda}}. \tag{60}$$

Finally, the gradient of the Sinkhorn divergence is approximated via

$$\begin{aligned}\nabla_\theta S_\lambda(P_{G_\theta(X)}, P_Y) \approx{}& \mathbb{E}_{(X,Y)\sim P_X\times P_Y}[\mu(X,Y)\nabla_\theta(\|G_\theta(X)-Y\|^2)]\\ &- \mathbb{E}_{(X_1,X_2)\sim P_X\times P_X}[\mu^x(X_1,X_2)\nabla_\theta(\|G_\theta(X_1)-G_\theta(X_2)\|^2)]\end{aligned}$$

Since the expectations cannot be calculated exactly, we further approximate the gradient with an empirical expectation over a batch of size S, which results in a mini-batch stochastic gradient descent on $S_\lambda(P_{G(X)}, P_Y)$: for a sample $x^1, \cdots, x^S \overset{i.i.d.}{\sim} P_X, y^1, \cdots, y^S \overset{i.i.d.}{\sim} P_{\hat{Y}}$ the gradient approximation is given by

$$\begin{aligned}\nabla_\theta S_\lambda(P_{G_\theta(X)}, P_Y) \approx{}& \frac{1}{S^2}\sum_{i,j=1}^{S}[\mu(x_i,y_j)\nabla_\theta(\|G_\theta(x_i)-y_j\|^2)]\\ &- \frac{1}{S^2}\sum_{i,j=1}^{S}\mu^x(x_i,x_j)\nabla_\theta(\|G_\theta(x_i)-G_\theta(x_j)\|^2)]\end{aligned}$$

We note that the optimal dual potentials for $W_{2,\lambda}^2(P_{G(X)}, P_{G(X)})$ for Gaussian X and a linear generator can indeed be found analytically as a function of G, but since it is not possible to analytically compute the potentials in the case of a more complex $G(\cdot)$ and since we do not use the linearity of $G(\cdot)$ when computing the unregularized Wasserstein distance, to give the models a fair comparison, we find $W_{2,\lambda}^2(P_{G(X)}, P_{G(X)})$ numerically.

Note that in the above discussion, we assumed that we have access to approximations of the optimal dual potentials. These optimal dual potentials can be computed in two different ways. The first way is to compute them via the Sinkhorn-Knopp algorithm(Feydy et al., 2019), labelled SGAN-NP, which allows us to omit the discriminator network from the GAN and compute the dual potentials in a non-parametric fashion. Another way of calculating

Algorithm 1 SGD for GANs

```
INPUT: $P_X$, $P_{\hat{Y}}$, $\lambda$, $S$, $\theta_0$, step sizes $\{\alpha_t > 0\}_{t=0}^{T-1}$
for $t = 0, \cdots, T-1$ do
    Sample I.I.D. points $x_t^1, \cdots, x_t^S \sim P_X, y_t^1, \cdots, y_t^S \sim P_{\hat{Y}}$
    Call the oracle to find $\epsilon$-approximate maximizers $(\phi_t, \psi_t), \phi_t^x$ for the dual formulations (57), (58)
    Compute
        $$g_t = \frac{1}{S^2} \sum_{i,j} \Big( \mu_t(G_\theta(x_t^i), y_t^j) \nabla_\theta(\|G_\theta(x_t^i) - y_t^j\|^2) - \frac{1}{2}\mu_t^x(x_t^i, x_t^j) \nabla_\theta(\|G_\theta(x_t^i) - G_\theta(x_t^j)\|^2) \Big) \quad (61)$$
    where $\mu_t, \mu_t^x$ are computed using $(\phi_t, \psi_t)$ and $\phi_t^x$ based on (59), (60).
    Update $\theta_{t+1} \leftarrow \theta_t - \alpha_t g_t$
end for
```

approximations of the dual potentials is to represent the dual potentials as neural networks and update them using stochastic gradient descent on (8), labelled SGAN-P.

On the one hand, using neural networks helps preserve the history of the seen examples and might help the dual potentials generalize better. On the other hand, using Sinkhorn-Knopp algorithm is more precise for computing the Sinkhorn divergence between the empirical distributions. We compared the two approaches and didn't find any significant differences. To compare WGAN and SGAN as they minimize different objectives, we evaluate their performance by calculating Frobenius distance between the covariance matrix of the generated distribution $P_{G(X)}$ and the covariance matrix of the target distribution P_Y (true covariance, bottom row). We also calculate the Frobenius distance between the covariance matrix of the generated distribution $P_{G(X)}$ and the optimal covariance matrix for W2GAN (1) $P_{G^*(X)}$ (optimal covariance, bottom row) in Figure 1 for two values of the dimensions of the latent random variable, $r = 4$ and $r = 8$. We run the experiments for 500 epochs with a batch size of 200. In these experiments, we observe that different versions of SGAN enjoy similar behavior and the covariance matrix of the output distribution is closer to the one of the target distribution compared to standard Wasserstein GANs. We note that the distance to the true covariance has a higher floor here, since the error cannot be zero, i.e. the d-dimensional Gaussian distribution cannot be approximated as a function of the r-dimensional one with error converging to 0.

7. 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.

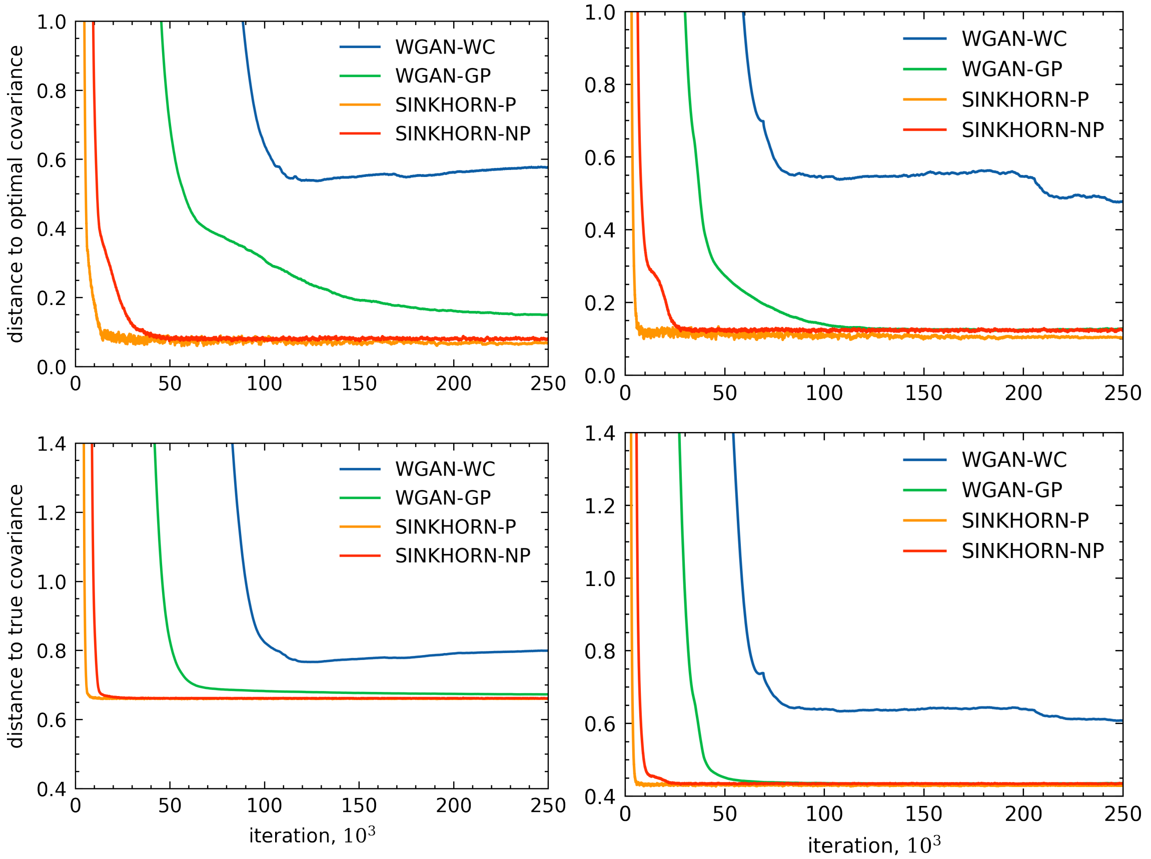


Figure 1: Training of SGAN and WGANs for latent variable dimension $r = 4$(left) and $r = 8$(right) for a linear generator. The distance is calculated to the optimal covariance (r-PCA, top) and true covariance (bottom)

Acknowledgments

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

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

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.

Mathieu Blondel, Vivien Seguy, and Antoine Rolet. Smooth and sparse optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pages 880–889. PMLR, 2018.

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.

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.

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

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

Abir De, Isabel Valera, Niloy Ganguly, Sourangshu Bhattacharya, and Manuel Gomez Rodriguez. Learning and forecasting opinion dynamics in social networks. In *Advances in Neural Information Processing Systems*, volume 29, pages 397–405. Curran Associates, Inc., 2016.

Richard Mansfield Dudley. The speed of mean glivenko-cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

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

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.

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.

Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models.* Cambridge University Press, 2021.

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*, volume 27, pages 2672–2680. Curran Associates, Inc., 2014.

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.

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.

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.

Alexander Korotin, Vage Egiazarian, Arip Asadulaev, Alexander Safin, and Evgeny Burnaev. Wasserstein-2 generative networks. *arXiv preprint arXiv:1909.13082*, 2019.

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.

Huidong Liu, Xianfeng Gu, and Dimitris Samaras. Wasserstein gan with quadratic transport cost. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4832–4841, 2019.

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

Albert W Marshall, Ingram Olkin, and Barry C Arnold. *Inequalities: theory of majorization and its applications*, volume 143. Springer, 1979.

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.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pages 3481–3490. PMLR, 2018.

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.

Daria Reshetova, Yikun Bai, Xiugang Wu, and Ayfer Özgür. Understanding entropic regularization in gans. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 825–830. IEEE, 2021.

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.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Xiang Wei, Boqing Gong, Zixia Liu, Wei Lu, and Liqiang Wang. Improving the improved training of wasserstein gans: A consistency term and its dual effect. In *International Conference on Learning Representation (ICLR)*, 2018.