A Primal-Dual link between GANs and Autoencoders

Hisham Husain[‡,†] **Richard Nock**[†,‡,♣] **Robert C. Williamson**[‡,†]

[‡]The Australian National University, [†]Data61, [♣]The University of Sydney

`firstname.lastname@{data61.csiro.au,anu.edu.au}`

Abstract

Since the introduction of Generative Adversarial Networks (GANs) and Variational Autoencoders (VAE), the literature on generative modelling has witnessed an overwhelming resurgence. The impressive, yet elusive empirical performance of GANs has lead to the rise of many GAN-VAE hybrids, with the hopes of GAN level performance and additional benefits of VAE, such as an encoder for feature reduction, which is not offered by GANs. Recently, the Wasserstein Autoencoder (WAE) was proposed, achieving performance similar to that of GANs, yet it is still unclear whether the two are fundamentally different or can be further improved into a unified model. In this work, we study the f-GAN and WAE models and make two main discoveries. First, we find that the f-GAN and WAE objectives partake in a primal-dual relationship and are equivalent under some assumptions, which then allows us to explicate the success of WAE. Second, the equivalence result allows us to, for the first time, prove generalization bounds for Autoencoder models, which is a pertinent problem when it comes to theoretical analyses of generative models. Furthermore, we show that the WAE objective is related to other statistical quantities such as the f-divergence and in particular, upper bounded by the Wasserstein distance, which then allows us to tap into existing efficient (regularized) optimal transport solvers. Our findings thus present the first primal-dual relationship between GANs and Autoencoder models, comment on generalization abilities and make a step towards unifying these models.

1 Introduction

Implicit probabilistic models [1] are defined to be the pushforward of a simple distribution P_Z over a latent space $\mathcal{Z}$ through a map $G : \mathcal{Z} \to \mathcal{X}$, where $\mathcal{X}$ is the space of the input data. Such models allow easy sampling, but the computation of the corresponding probability density function is intractable. The goal of these methods is to match $G\#P_Z$ to a target distribution P_X by minimizing $D(P_X, G\#P_Z)$, for some discrepancy $D(\cdot,\cdot)$ between distributions. An overwhelming number of methods have emerged after the introduction of Generative Adversarial Networks [2, 3] and Variational Autoencoders [4] (GANs and VAEs), which have established two distinct paradigms: Adversarial (networks) training and Autoencoders respectively. Adversarial training involves a set of functions $\mathcal{D}$, referred to as *discriminators*, with an objective of the form

$$D(P_X, G\#P_Z) = \max_{d \in \mathcal{D}} \left\{ \mathbb{E}_{x \sim P_X}[a(d(x))] - \mathbb{E}_{x \sim G\#P_Z}[b(d(x))] \right\}, \tag{1}$$

for some functions $a : \mathbb{R} \to \mathbb{R}$ and $b : \mathbb{R} \to \mathbb{R}$. Autoencoder methods are concerned with finding a function $E : \mathcal{X} \to \mathcal{Z}$, referred to as an *encoder*, whose goal is to reverse G, and learn a feature space with the objective

$$D(P_X, G\#P_Z) = \min_{E} \left\{ \mathcal{R}(G, E) + \Omega(E) \right\}, \tag{2}$$

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

where $\mathcal{R}(G, E)$ is the *reconstruction* loss and acts to ensure G and E reverse each other and $\Omega(E)$ is a regularization term. Much work on Autoencoder methods has focused upon the choice of Ω.

In practice, the two methods demonstrate contrasting abilities in their strengths and limitations, which have resulted in differing directions of progress. Indeed, there is a lack of theoretical understanding of how these frameworks are parametrized and it is not clear whether the methods are fundamentally different. For example, Adversarial training based methods have empirically demonstrated high performance when it comes to producing realistic looking samples from P_X. However, GANs often have problems in convergence and stability of training [5]. Autoencoders, on the other hand, deal with a more well behaved objective and learn an encoder in the process, making them useful for feature representation. However in practice, Autoencoder based methods have reported shortfalls, such as producing blurry samples for image based datasets [6]. This has motivated researchers to adapt Autoencoder models by borrowing elements from Adversarial networks in the hopes of GAN level performance whilst learning an encoder. Examples include replacing Ω with Adversarial objectives [7, 8] or replacing the reconstruction loss with an adversarial objective [9, 10]. Recently, the Wasserstein Autoencoder (WAE) [6] has been shown to subsume these two methods with an Adversarial based Ω, and has demonstrated performance similar to that of Adversarial methods.

Understanding the connection between the two paradigms is important for not only the practical purposes outlined above but for the inheritance of theoretical analyses from one another. For example, when it comes to directions of progress, Adversarial training methods now have theoretical guarantees on generalization performance [11], however no such theoretical results have been obtained to date for autoencoders. Indeed, generalization performance is a pressing concern, since both techniques implicitly assume the samples represent the target distribution [12] and eventually leads to memorizing training data.

In this work, we study the two paradigms and in particular focus on the f-GANs [3] for Adversarial training and Wasserstein Autoencoders (WAE) for Autoencoders, which generalize the original GAN and VAE models respectively. We prove that the f-GAN objective with Lipschitz (with respect to a metric c) discriminators is equivalent to the WAE objective with cost c. In particular, we show that the WAE objective is an upper bound; schematically we get

$$\boxed{f\text{-GAN} \leq \text{WAE}}$$

and discuss the tightness of this bound. Our result is a generalization of the Kantorovich-Rubinstein duality and thus suggests a primal-dual relationship between Adversarial and Autoencoder methods. Consequently we show, to the best of our knowledge, the first generalization bounds for autoencoders. Furthermore, using this equivalence, we show that the WAE objective is related to key statistical quantities such as the f-divergence and Wasserstein distance, which allows us to tap into efficient (regularized) OT solvers.

The main contributions can be summarized as the following:

▷ (Theorem 8) Establishes an equivalence between Adversarial training and Wasserstein Autoencoders, showing conditions under which the f-GAN and WAE coincide. This further justifies the similar performance of WAE to GAN based methods. When the conditions are not met, we have an inequality, which allows us to comment on the behavior of the methods.

▷ (Theorem 9, 10 and 14) Show that the WAE objective is related to other statistical quantities such as f-divergence and Wasserstein distance.

▷ (Theorem 13) Provide generalization bounds for WAE. In particular, this focuses on the empirical variant of the WAE objective, which allows the use of Optimal Transport (OT) solvers as they are concerned with discrete distributions. This allows one to employ efficient (regularized) OT solvers for the estimation of WAE, f-GANs and the generalization bounds.

2 Preliminaries

2.1 Notation

We will use $\mathcal{X}$ to denote the input space (a Polish space), typically taken to be a Euclidean space. We use $\mathcal{Z}$ to denote the latent space, also taken to be Euclidean. We use $\mathbb{N}_*$ to denote the natural numbers without 0: $\mathbb{N} \setminus \{0\}$. We denote by $\mathscr{P}$ the set of probability measures over $\mathcal{X}$, and elements of this set will be referred to as *distributions*. If $P \in \mathscr{P}(\mathcal{X})$ happens to be absolutely continuous with respect to the Lebesgue measure then we will use $dP/d\lambda$ to refer to the *density* function (Radon-Nikodym derivative with respect to the Lebesgue measure). For any $T \in \mathscr{F}(\mathcal{X}, \mathcal{Z})$, for any measure $\mu \in \mathscr{P}(\mathcal{X})$, the pushforward measure of μ through T denoted $T\#\mu \in \mathscr{P}(\mathcal{Z})$ is such that $T\#\mu(A) = \mu(T^{-1}(A))$ for any measurable set $A \subset \mathcal{Z}$. The set $\mathscr{F}(\mathcal{X}, \mathbb{R})$ refers to all measurable functions from $\mathcal{X}$ into the set $\mathbb{R}$. We will use functions to represent conditional distributions over a space $\mathcal{Z}$ conditioned on elements $\mathcal{X}$, for example $P \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z}))$ so that for any $x \in \mathcal{X}$, $P(x) = P(\cdot|x) \in \mathscr{P}(\mathcal{Z})$. For any $P \in \mathscr{P}(\mathcal{X})$, the *support* of P is $\text{supp}(P) = \{x \in \mathcal{X} : \text{ if } x \in N_x \text{ open } \implies P(N_x) > 0\}$. In any metric space $(\mathcal{X}, c)$, for any set $S \subseteq \mathcal{X}$, we define the *diameter* of S to be $\text{diam}_c(S) = \sup_{x,x' \in S} c(x, x')$. Given a metric c over $\mathcal{X}$, for any $f \in \mathscr{F}(\mathcal{X}, \mathbb{R})$, $\text{Lip}_c(f)$ denotes the Lipschitz constant of f with respect to c and $\mathcal{H}_c = \{g \in \mathscr{F}(\mathcal{X}, \mathbb{R}) : \text{Lip}_c(g) \leq 1\}$. For some set $S \subseteq \mathbb{R}$, $\mathbf{1}_S$ corresponds to the convex *indicator function*, ie. $\mathbf{1}_S(x) = 0$ if $x \in S$ and $\mathbf{1}_S(x) = \infty$ otherwise. For any $x \in \mathcal{X}$, $\delta_x : \mathcal{X} \to \{0, 1\}$ corresponds to the *characteristic function*, with $\delta_x(0) = 1$ if $x = 0$ and $\delta_x(0) = 0$ if $x \neq 0$.

2.2 Background

2.2.1 Probability Discrepancies

Probability discrepancies are central to the objective of finding the best fitting model. We introduce some key discrepancies and their notation, which will appear later.

Definition 1 (f-Divergence) *For a convex function $f : \mathbb{R} \to (-\infty, \infty]$ with $f(1) = 0$, for any $P, Q \in \mathscr{P}(\mathcal{X})$ with P absolutely continuous with respect to Q, the f-Divergence between P and Q is*

$$D_f(P, Q) := \int_{\mathcal{X}} f\left(\frac{dP}{dQ}\right) dQ,$$

with $D_f(P, Q) = \infty$ if P is note absolutely continuous with respect to Q.

An example of a method to compute the f-divergence is to first compute dP/dQ and estimate the integral empirically using samples from Q.

Definition 2 (Integral Probability Metric) *For a fixed function class $\mathcal{F} \subseteq \mathscr{F}(\mathcal{X}, \mathbb{R})$, the Integral Probability Metric (IPM) based on $\mathcal{F}$ between $P, Q \in \mathscr{P}(\mathcal{X})$ is defined as*

$$\text{IPM}_{\mathcal{F}}(P, Q) := \sup_{f \in \mathcal{F}} \left\{ \int_{\mathcal{X}} f(x) dP(x) - \int_{\mathcal{X}} f(x) dQ(x) \right\}.$$

If we have that $-\mathcal{F} = \mathcal{F}$ then $\text{IPM}_{\mathcal{F}}$ forms a metric over $\mathscr{P}(\mathcal{X})$ [13]. A particular IPM we will make use of is Total Variation (TV): $\text{TV}(P, Q) = \text{IPM}_{\mathcal{V}}(P, Q)$ where $\mathcal{V} = \{h \in \mathscr{F}(\mathcal{X}, \mathbb{R}) : |h| \leq 1\}$. We also note that when $f(x) = |x - 1|$ then $\text{TV} = D_f$ and thus TV is both an IPM and an f-divergence.

Definition 3 *For any $P, Q \in \mathscr{P}(\mathcal{X})$, define the* set of couplings *between P and Q to be*

$$\Pi(P, Q) = \left\{ \pi \in \mathscr{P}(\mathcal{X} \times \mathcal{X}) : \int_{\mathcal{X}} \pi(x, y) dx = P, \int_{\mathcal{X}} \pi(x, y) dy = Q \right\}.$$

For a cost $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, the Wasserstein distance *between P and Q is*

$$W_c(P, Q) := \inf_{\pi \in \Pi(P,Q)} \left\{ \int_{\mathcal{X} \times \mathcal{X}} c(x, y) d\pi(x, y) \right\}.$$

The Wasserstein distance can be regarded as an infinite linear program and thus admits a dual form, and in the case of c being a metric, belongs to the class of IPMs. We summarize this fact the following lemma [14].

Lemma 4 (Wasserstein Duality) *Let $(\mathcal{X}, c)$ be a metric space, and suppose $\mathcal{H}_c$ is the set of all 1-Lipschitz functions with respect to c. Then for any $P, Q \in \mathscr{P}(\mathcal{X})$, we have*

$$\begin{aligned} W_c(P, Q) &= \sup_{h \in \mathcal{H}_c} \left\{ \int_{\mathcal{X}} h(x) dP(x) - \int_{\mathcal{X}} h(x) dQ(x) \right\} \\ &= \text{IPM}_{\mathcal{H}_c}(P, Q). \end{aligned}$$

2.3 Generative Models

In both GAN and VAE models, we have a latent space $\mathcal{Z}$ (typically taken to be $\mathbb{R}^d$, with d being small) and a prior distribution $P_Z \in \mathscr{P}(\mathcal{Z})$ (e.g. unit variance Gaussian). We have a function referred to as the generator $G : \mathcal{Z} \to \mathcal{X}$, which induces the *generated* distribution, denoted by $P_G \in \mathscr{P}(\mathcal{X})$, as the pushforward of P_Z through G: $P_G = G\#P_Z$. The true data distribution will be referred to as $P_X \in \mathscr{P}(\mathcal{X})$. The common goal between the two methods is to find a generator G such that the samples generated by pushing forward P_Z through G ($G\#P_Z$) are close to the true data distribution (P_X). More formally, one can cast this as an optimization problem by finding the best G such that $D(P_G, P_X)$ is minimized, where $D(\cdot, \cdot)$ is some discrepancy between distributions. Both methods (as we outline below) utilize their own discrepancies between P_X and P_G, which offer their own benefits and weaknesses.

2.3.1 Wasserstein Autoencoder

Let $E : \mathcal{X} \to \mathscr{P}(\mathcal{Z})$ denote a probabilistic *encoder* [1], which maps each point x to a conditional distribution $E(x) \in \mathscr{P}(\mathcal{Z})$, denoted as the *posterior* distribution. The pushforward of P_X through E: $E\#P_X$, will be referred to as the *aggregated posterior*.

Definition 5 (Wasserstein Autoencoder [6]) *Let $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$, $\lambda > 0$ and $\Omega : \mathscr{P}(\mathcal{Z}) \times \mathscr{P}(\mathcal{Z}) \to \mathbb{R}_{\geq 0}$ with $\Omega(P, P) = 0$ for all $P \in \mathscr{P}(\mathcal{Z})$. The Wasserstein Autoencoder objective is*

$$\mathrm{WAE}_{c,\lambda\cdot\Omega}(P_X, G) = \inf_{E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z}))} \left\{ \int_{\mathcal{X}} \mathbb{E}_{z\sim E(x)}[c(x, G(z))] dP_X(x) + \lambda \cdot \Omega(E\#P_X, P_Z) \right\}.$$

We remark that there are various choices of c and $\lambda \cdot \Omega$. [6] select these by tuning λ and selecting different measures of discrepancies between probability distortions for Ω.

2.3.2 f-Generative Adversarial Network

Let $d : \mathcal{X} \to \mathbb{R}$ denote a *discriminator* function.

Definition 6 (f-GAN [3]) *Let $f : \mathbb{R} \to (-\infty, \infty]$ denote a convex function with property $f(1) = 0$ and $\mathcal{D} \subset \mathscr{F}(\mathcal{X}, \mathbb{R})$ a set of discriminators. The f-GAN model minimizes the following objective for a generator $G : \mathcal{Z} \to \mathcal{X}$*

$$\mathrm{GAN}_f(P_X, G; \mathcal{D}) := \sup_{d\in\mathcal{D}} \left\{ \mathbb{E}_{x\sim P_X}[d(x)] - \mathbb{E}_{z\sim P_Z}[f^*(d(G(z)))] \right\}, \tag{3}$$

where $f^\star(x) = \sup_y \{x \cdot y - f(y)\}$ is the convex conjugate of f.

There are two knobs in this method, namely $\mathcal{D}$, the set of discriminators, and the convex function f. The objective in (3) is a variational approximation to D_f [3]; if $\mathcal{D} = \mathscr{F}(\mathcal{X}, \mathbb{R})$, then $\mathrm{GAN}_f(P_X, G; \mathcal{D}) = D_f(P_X, P_G)$ [15]. In the case of $f(x) = x\log(x) - (x+1)\log(x+1) + 2\log 2$, we recover the original GAN [2].

3 Related Work

Current attempts at building a taxonomy for generative models have largely been within each paradigm or the proposal of hybrid methods that borrow elements from the two. We first review major and relevant advances in each paradigm, and then move on to discuss results that are close to the technical contributions of our work.

The line of Autoencoders begin with $\Omega = 0$, which is the original autoencoder concerned only with reconstruction loss. VAE then introduced a non-zero Ω, along with implementing Gaussian encoders [4]. This was then replaced by an adversarial objective [7], which is sample based and consequently allows arbitrary encoders. In the spirit of unification, Adversarial Autoencoders (AAE) [8] proposed Ω to be a discrepancy between the pushforward of the target distribution through the encoder ($E\#P_X$)

[1] We remark that this is not standard notation in the VAE and Variational Inference literature.

and the prior distribution (P_Z) in the latent space, which was then showed to be equivalent to the VAE Ω minus a mutual information term [16]. Independently, InfoVAE [17] proposed a similar objective, which was subsequently shown to be equivalent to adding mutual information. [6] reparametrized the Wasserstein distance into an Autoencoder objective (WAE) where the Ω term generalizes AAE, and has reported performance comparable to that of Adversarial methods. Other attempts also include adjusting the reconstruction loss to be adversarial as well [9, 10]. Another work that focuses on WAE is the Sinkhorn Autoencoders (SAE) [18], which select Ω to be the Wasserstein distance and show that the overall objective is an upper bound to the Wasserstein distance between P_X and P_G.

[19] discussed the two paradigms and their unification by interpretting GANs from the perspective of variational inference, which allowed a connection to VAE, resulting in a GAN implemented with importance weighting techniques. While this approach is the closest to our work in forming a link, their results apply to standard VAE (and not other AE methods such as WAE) and cannot be extended to all f-GANs. [20] introduced the notion of an Adversarial divergence, which subsumed mainstream adversarial based methods. This also led to the formal understanding of how the selected discriminator set $\mathcal{D}$ affects the final generator G learned. However, this approach is silent with regard to Autoencoder based methods. [11] established the tradeoff between the Rademacher complexity of the discriminator class $\mathcal{D}$ and generalization performance of G, with no results present for Autoencoders. These theoretical advances in Adversarial training methods are inherited by Autoencoders as a consequence of the equivalence presented in our work.

One key point in the proof of our equivalence is the use of a result that decomposes the GAN objective into an f-divergence and an IPM for a restricted class of discriminators (which we used for Lipschitz functions). This decomposition is used in [21] and applied to linear f-GANs, showing that the adversarial training objective decomposes into a mixture of maximum likelihood and moment matching. [22] used this decomposition with Lipschitz discriminators like our work, however does not make any extension or further progress to establish the link to WAE. Indeed, GANs with Lipschitz discriminators have been independently studied in [23], which suggest that one should enforce Lipschitz constraints to provide useful gradients.

4 f-Wasserstein Autoencoders

We define a new objective, that will help us in the proof of the main theorems of this paper.

Definition 7 (f-Wasserstein Autoencoder) *Let $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $\lambda > 0$, $f : \mathbb{R} \to (-\infty, \infty]$ be a convex function (with $f(1) = 0$), the f-Wasserstein Autoencoder (f-WAE) objective is*

$$\overline{W}_{c,\lambda\cdot f}(P_X, G) = \inf_{E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z}))} \{W_c(P_X, (G \circ E)\#P_X) + \lambda D_f(E\#P_X, P_Z)\} \tag{4}$$

In the proof of the main result, we will show that the f-WAE objective is indeed the same as the WAE objective when using the same cost c and selecting the regularizer to be $\lambda \cdot \Omega = D_{\lambda f} = \lambda D_f$. The only difference between this and the standard WAE is the use of $W_c(P_X, (G \circ E)\#P_X)$ as the reconstruction loss instead of the standard cost which is an upper bound (Lemma 18), and the regularizer is chosen to be $\lambda \cdot \Omega = D_{\lambda f} = \lambda D_f$. We now present the main theorem that captures the relationship between f-GAN and WAE.

Theorem 8 (f-GAN and WAE equivalence) *Suppose $(\mathcal{X}, c)$ is a metric space and let $\mathcal{H}_c$ denote the set of all functions from $\mathcal{X} \to \mathbb{R}$ that are 1-Lipschitz (with respect to c). Let $f : \mathbb{R} \to (-\infty, \infty]$ be a convex function with $f(1) = 0$. Then for all $\lambda > 0$,*

$$\mathrm{GAN}_{\lambda f}(P_X, G; \mathcal{H}_c) \leq \mathrm{WAE}_{c,\lambda\cdot D_f}(P_X, G), \tag{5}$$

with equality if G is invertible.

Proof (This is a sketch, see Section A.1 for full proof). The proof begins by proving certain properties of $\mathcal{H}_c$ (Lemma 16), allowing us to use the dual form of restricted GANs (Theorem 15),

$$\begin{aligned} \mathrm{GAN}_f(P_X, G; \mathcal{H}_c) &= \inf_{P' \in \mathscr{P}(\mathcal{X})} \left\{ D_f(P', P_G) + \sup_{h \in \mathcal{H}_c} \{\mathbb{E}_{P_X}[h] - \mathbb{E}_{P'}[h]\} \right\} \\ &= \inf_{P' \in \mathscr{P}(\mathcal{X})} \{D_f(P', P_G) + W_c(P', P_X)\}. \end{aligned} \tag{6}$$

The key is to reparametrize (6) by optimizing over couplings. By rewriting $P' = (G \circ E)\#P_X$ for some $E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z}))$ and rewriting (6) as an optimization over E (Lemma 20), we obtain

$$\begin{aligned} &\inf_{P' \in \mathscr{P}(\mathcal{X})} \{D_f(P', P_G) + W_c(P', P_X)\} \\ &= \inf_{E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z}))} \{D_f((G \circ E)\#P_X, P_G) + W_c((G \circ E)\#P_X, P_X)\} \end{aligned} \tag{7}$$

We then have

$$D_f((G \circ E)\#P_X, P_G) = D_f(G\#(E\#P_X), G\#P_Z) \overset{(*)}{\leq} D_f(E\#P_X, P_Z),$$

with equality in $(*)$ if G is invertible (Lemma 17). A weaker condition is required if f is differentiable, namely if G is invertible with respect to $f' \circ d(E\#P_X)/dP_Z$ in the sense that

$$G(z) = G(z') \implies f' \circ (d(E\#P_X)/dP_Z)(z) = f' \circ (d(E\#P_X)/dP_Z)(z'), \tag{8}$$

noting that an invertible G trivially satisfies this requirement. Letting $f \leftarrow \lambda f$, we have $D_f(\cdot, \cdot) \leftarrow \lambda D_f(\cdot, \cdot)$, and so from Equation 7, we have

$$\begin{aligned} \mathrm{GAN}_{\lambda f}(P_X, G; \mathcal{H}_c) &\overset{(*)}{\leq} \inf_{E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z}))} \{\lambda D_f(E\#P_X, P_Z) + W_c((G \circ E)\#P_X, P_X)\} \\ &= \overline{W}_{c, \lambda \cdot f}(P_X, G) \\ &\leq \inf_{E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z}))} \left\{\lambda D_f(E\#P_X, P_Z) + \int_{\mathcal{X}} \mathbb{E}_{z \sim E(x)}[c(x, G(z))] dP_X(x)\right\} \\ &= \mathrm{WAE}_{c, \lambda \cdot D_f}(P_X, G), \end{aligned}$$

where the final inequality follows from the fact that $W_c(P, Q) \leq \int_{\mathcal{X}} \mathbb{E}_{z \sim E(x)}[c(x, G(z))] dP_X(x)$ (Lemma 18). Using the fact that $\overline{W} \geq \mathrm{WAE}$ (Lemma 19) completes the proof. ■

When G is invertible, we remark that P_G can still be expressive and capable of modelling complex distributions in WAE and GAN models. For example, if G is implemented with feedforward neural networks, and G is invertible then P_G can model *deformed* exponential families [24], which encompasses a large class appearing in statistical physics and information geometry [25, 26]. There exists many invertible activation functions under which G will be invertible. Furthermore, in the proof of the Theorem it is clear that $\overline{W}$ and WAE are the same objective (from Lemma 18 and Lemma 19). When using $f = \mathbf{1}_{\{1\}}$ ($f(x) = 0$ if $x = 1$ and $f(x) = \infty$ otherwise), and noting that $f^\star(x) = x$, meaning that Theorem 8 (with $\lambda = 1$) reduces to

$$\begin{aligned} \sup_{h \in \mathcal{H}_c} \{\mathbb{E}_{x \sim P_X}[h(x)] - \mathbb{E}_{x \sim P_G}[h(x)]\} &= \mathrm{GAN}_f(P_X, G; \mathcal{H}_c) \\ &\leq \overline{W}_{c,f}(P_X, P_G) \\ &= \inf_{E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z})): E\#P_X = P_Z} \{W_c(P_X, (G \circ E)\#P_X)\} \\ &= \inf_{E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z})): E\#P_X = P_Z} \{W_c(P_X, G\#P_Z\} \\ &= W_c(P_X, P_G), \end{aligned}$$

which is the standard primal-dual relation between Wasserstein distances as in Lemma 4. Hence, Theorem 8 can be viewed as a generalization of this primal-dual relationship, where Autoencoder and Adversarial objectives represent primal and dual forms respectively.

We note that the left hand of Equation (5) does not explicitly engage the prior space Z as much as the right hand side in the sense that one can set $Z = \mathcal{X}$, $G = \mathrm{Id}$ (which is invertible) and $P_Z = P_G$ and indeed results in the exact same f-GAN objective since $G\#P_Z = \mathrm{Id}\#P_G = P_G$, yet the equivalent f-WAE objective (from Theorem 8) will be different. This makes the Theorem versatile in reparametrizations, which we exploit in the proof of Theorem 10. We now consider weighting the reconstruction along with the regularization term in $\overline{W}$ (which is equivalent to weighting WAE), which simply amounts to re-weighting the cost since for any $\gamma > 0$,

$$\overline{W}_{\gamma \cdot c, \lambda \cdot f}(P_X, G) = \inf_{E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z}))} \{\gamma W_c((G \circ E)\#P_X, P_X) + \lambda D_f(E\#P_X, P_Z)\}.$$

The idea of weighting the regularization term by λ was introduced by [27] and furthermore studied empirically, showing that the choice of λ influences learning disentanglement in the latent space. [28]. We show that if $\lambda = 1$ and γ is larger than some γ^* then $\overline{W}$ will become an f-divergence (Theorem 9). On the other hand if we fix $\gamma = 1$ and take λ is larger than some λ^*, then $\overline{W}$ becomes the Wasserstein distance and in particular, equality holds in (5) (Theorem 10). We show explicitly how high γ and λ need to be for such equalities to occur. This is surprising since f-divergence and Wasserstein distance are quite different distortions.

We begin with the f-divergence case. Consider $f : \mathbb{R} \to (-\infty, \infty]$ convex, differentiable and $f(1) = 0$ and assume that P_X is absolutely continuous with respect to P_G, so that $D_f(P_X, P_G) < \infty$.

Theorem 9 *Set $c(x, y) = \delta_{x-y}$ and let $f : \mathbb{R} \to (-\infty, \infty]$ be a convex function (with $f(1) = 0$) and differentiable. Let $\gamma^* = \sup_{x,x' \in \mathcal{X}} \left| f'\left(\frac{dP_X}{dP_G}\right) - f'(\frac{dP_X}{dP_G})(x') \right|$ and suppose P_G is absolutely continuous with respect to P_X and that G is invertible, then we have for all $\gamma \geq \gamma^*$*

$$\overline{W}_{\gamma \cdot c, f}(P_X, G) = D_f(P_X, P_G).$$

(Proof in Appendix, Section A.3). It is important to note that $W_c(P_X, P_G) = \mathrm{TV}(P_X, P_G)$ when $c(x, y) = \delta_{x-y}$ and so Theorem 9 tells us that the objective with a weighted total variation reconstruction loss with a f-divergence prior regularization amounts to the f-divergence. It was shown that in [24] that when G is an invertible feedforward neural network then $D_f(P_X, P_G)$ is a *Bregman* divergence (a well regarded quantity in information geometry) between the parametrizations of the network for a fixed choice of activation function of G, which depends on f. Hence, a practioner should design G with such activation function when using f-WAE under the above setting ($c(x, y) = \delta_{x-y}$ and $\gamma = \gamma^*$) with G being invertible, so that the information theoretic divergence (D_f) between the distributions becomes an information geometric divergence involving the network parameters.

We now show that if λ is selected higher than $\lambda^* := \sup_{P' \in \mathscr{P}(\mathcal{X})} (W_c(P', P_G)/D_f(P', P_G))$, then $\overline{W}$ becomes W_c and furthermore we have equality between f-GAN, f-WAE and WAE.

Theorem 10 *Let $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a metric. For any $f : \mathbb{R} \to (-\infty, \infty]$ convex function (with $f(1) = 0$), we have for all $\lambda \geq \lambda^*$*

$$\mathrm{GAN}_{\lambda f}(P_X, G; \mathcal{H}_c) = \overline{W}_{c, \lambda \cdot f}(P_X, G) = \mathrm{WAE}_{c, \lambda \cdot D_f}(P_X, G) = W_c(P_X, P_G).$$

(Proof in Appendix, Section A.4). Note that Theorem 10 holds for any f (satisfying properties of the Theorem) and so one can estimate the Wasserstein distance using any f as long as λ is scaled to λ^*. In order to understand how high λ^* can be,there are two extremes in which the supremum may be unbounded. The first case is when P' is taken far from P_G so that $W_c(P', P_G)$ increases, however one should note that in the case when $\Delta = \max_{x,x' \in \mathcal{X}} c(x, x') < \infty$ then $W_c \in [0, \Delta]$ and so W_c will be finite whereas $D_f(P', P_G)$ can possibly diverge to ∞, making $\lambda^* \to 0$. The other case is when P' is made close to P_G, in which case $\frac{1}{D_f(P', P_G)} \to \infty$ however $W_c(P', P_G) \to 0$ so the quantity λ^* can still be small in this case, depending on the rate of decrease between W_c and D_f. Now suppose that $f(x) = |x - 1|$ and $c(x, y) = \delta_{x-y}$, in which case $D_f = W_c$ and thus $\lambda^* = 1$. In this case, Theorem 10 reduces to the standard result [29] regarding the equivalence between Wasserstein distance and f-divergence intersecting at the variational divergence under these conditions.

5 Generalization bounds

We present generalization bounds using machinery developed in [30] with the following definitions.

Definition 11 (Covering Numbers) *For a set $S \subseteq \mathcal{X}$, we denote $N_\eta(S)$ to be the η-*covering number *of S, which is the smallest $m \in \mathbb{N}_*$ such that there exists closed balls $B_1, \ldots, B_m$ of radius η with $S \subseteq \bigcup_{i=1}^m B_i$. For any $P \in \mathscr{P}(\mathcal{X})$, the (η, τ)-*dimension *is $d_\eta(P, \tau) := \frac{\log N_\eta(P, \tau)}{-\log \eta}$, where $N_\eta(P, \tau) := \inf \{N_\eta(S) : P(S) \geq 1 - \tau\}$.*

Definition 12 (1**-Upper Wasserstein Dimension)** *The* 1-Upper Wasserstein dimension *of any* $P \in \mathscr{P}(\mathcal{X})$ *is* $d^*(P) := \inf\left\{s \in (2, \infty) : \limsup_{\eta \to 0} d_\eta(P, \eta^{\frac{s}{s-2}}) \leq s\right\}$.

We make an assumption of P_X and P_G having bounded support to achieve the following bounds. For any $P \in \mathscr{P}(\mathcal{X})$ in a metric space $(\mathcal{X}, c)$, we use define $\Delta_{P,c} = \mathrm{diam}_c(\mathrm{supp}(P))$.

Theorem 13 *Let* $(\mathcal{X}, c)$ *be a metric space and suppose* $\Delta := \max\left\{\Delta_{c,P_X}, \Delta_{c,P_G}\right\} < \infty$*. For any* $n \in \mathbb{N}_*$*, let* $\hat{P}_X$ *and* $\hat{P}_G$ *denote the empirical distribution with* n *samples drawn i.i.d from* P_X *and* P_G *respectively. Let* $s_X > d^*(P_X)$ *and* $s_G > d^*(P_G)$*. For all* $f : \mathbb{R} \to (-\infty, \infty]$ *convex functions,* $f(1) = 0$*,* $\lambda > 0$ *and* $\delta \in (0, 1)$*, then with probability at least* $1 - \delta$*, we have*

$$\mathrm{GAN}_{\lambda f}(P_X, G; \mathcal{H}_c) \leq \overline{W}_{c,\lambda\cdot f}(\hat{P}_X, P_G) + O\left(n^{-1/s_X} + \Delta\sqrt{\frac{1}{n}\ln\left(\frac{1}{\delta}\right)}\right), \tag{9}$$

and if $f(x) = |x - 1|$ *is chosen then*

$$\mathrm{GAN}_{\lambda f}(P_X, G; \mathcal{H}_c) \leq \overline{W}_{c,\lambda\cdot f}(\hat{P}_X, \hat{P}_G) + O\left(n^{-1/s_X} + n^{-1/s_G} + \Delta\sqrt{\frac{1}{n}\ln\left(\frac{1}{\delta}\right)}\right). \tag{10}$$

(Proof in Appendix, Section A.2). First note that there is no requirement on G to be invertible and no restriction on λ. Second, there are the quantities s_X,s_G and Δ that are influenced by the distributions P_X and P_G. It is interesting to note that d^* is related to fractal dimensions [31] and thus relates the convergence of GANs to statistical geometry. If G is invertible in the above then the left hand side of both bounds becomes $\overline{W}_{c,\lambda\cdot f}(P_X, G)$ by Theorem 8. In general, $\hat{P}_X$ and $\hat{P}_G$ will not share the same support, in which case $D_f(\hat{P}_X, \hat{P}_G) = \infty$ – This would lead one to suspect the same from $\overline{W}_{c,\lambda\cdot f}(\hat{P}_X, \hat{P}_G)$, however this is not the case since

$$\overline{W}_{c,\lambda\cdot f}(\hat{P}_X, \hat{P}_G) \leq \inf_{E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{X}))}\left\{W_c((G \circ E)\#P_X, P_X) + \lambda D_f(E\#\hat{P}_X, \hat{P}_Z)\right\},$$

and so $E \in \mathscr{F}(\mathcal{X}, \mathscr{P}(\mathcal{Z}))$ would be selected such that $E\#\hat{P}_X$ shares the support of $\hat{P}_Z$, resulting in a bounded value. We now show the relationship between $\overline{W}$ and W_c.

Theorem 14 *For any* $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$*,* $\lambda > 0$ *and* $f : \mathbb{R} \to (-\infty, \infty]$ *convex function (with* $f(1) = 0$*) we have* $\overline{W}_{c,\lambda\cdot f}(P_X, G) \leq W_c(P_X, P_G)$

(Proof in Appendix, Section A.5). This suggests that in order to minimize $\overline{W}$, one can minimize W_c. Indeed, majority of the solvers are concerned with discrete distributions, which is exactly what is present on the right hand side of the generalization bounds: $\overline{W}_{c,\lambda\cdot f}(\hat{P}_X, \hat{P}_G)$

6 Discussion and Conclusion

This work is the first to prove a generalized primal-dual betweenship between GANs and Autoencoders. Our result elucidated the close performance between WAE and f-GANs. Furthermore, we explored the effect of weighting the reconstruction and regularization on the WAE objective, showing relationships to both f-divergences and Wasserstein metrics along with the impact on the duality relationship. This equivalence allowed us to prove generalization results, which to the best of our knowledge, are the first bounds given for Autoencoder models. The results imply that we can employ efficient (regularized) OT solvers to approximate upper bounds on the generalization bounds, which involve discrete distributions and thus are natural for such solvers.

The consequences of unifying two paradigms are plentiful, generalization bounds being an example. One line of extending and continuing the presented work is to explore the use of a general cost c (as opposed to a metric), invoking the generalized Wasserstein dual in the goal of forming a generalized GAN. Our paper provides a basis to unify Adversarial Networks and Autoencoders through a primal-dual relationship, and opens doors for the further unification of related models.

Acknowledgments

We would like to acknowledge anonymous reviewers and the Australian Research Council of Data61.

References

[1] Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, 2016.

[2] 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*, pages 2672–2680, 2014.

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

[4] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[5] Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

[6] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

[7] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. *arXiv preprint arXiv:1701.04722*, 2017.

[8] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[9] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

[10] Aibek Alanov, Max Kochurov, Daniil Yashkov, and Dmitry Vetrov. Pairwise augmented gans with adversarial reconstruction loss. *arXiv preprint arXiv:1810.04920*, 2018.

[11] Pengchuan Zhang, Qiang Liu, Dengyong Zhou, Tao Xu, and Xiaodong He. On the discrimination-generalization tradeoff in gans. *arXiv preprint arXiv:1711.02771*, 2017.

[12] Ke Li and Jitendra Malik. On the implicit assumptions of gans. *arXiv preprint arXiv:1811.12402*, 2018.

[13] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

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

[15] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[16] Matthew D Hoffman and Matthew J Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, 2016.

[17] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*, 2017.

[18] Giorgio Patrini, Marcello Carioni, Patrick Forre, Samarth Bhargav, Max Welling, Rianne van den Berg, Tim Genewein, and Frank Nielsen. Sinkhorn autoencoders. *arXiv preprint arXiv:1810.01118*, 2018.

[19] Zhiting Hu, Zichao Yang, Ruslan Salakhutdinov, and Eric P Xing. On unifying deep generative models. *arXiv preprint arXiv:1706.00550*, 2017.

[20] Shuang Liu, Olivier Bousquet, and Kamalika Chaudhuri. Approximation and convergence properties of generative adversarial learning. In *Advances in Neural Information Processing Systems*, pages 5545–5553, 2017.

[21] Shuang Liu and Kamalika Chaudhuri. The inductive bias of restricted f-gans. *arXiv preprint arXiv:1809.04542*, 2018.

[22] Farzan Farnia and David Tse. A convex duality framework for gans. In *Advances in Neural Information Processing Systems*, pages 5254–5263, 2018.

[23] Zhiming Zhou, Yuxuan Song, Lantao Yu, Hongwei Wang, Weinan Zhang, Zhihua Zhang, and Yong Yu. Understanding the effectiveness of lipschitz-continuity in generative adversarial nets. 2018.

[24] Richard Nock, Zac Cranko, Aditya K Menon, Lizhen Qu, and Robert C Williamson. f-gans in an information geometric nutshell. In *Advances in Neural Information Processing Systems*, pages 456–464, 2017.

[25] Shun-ichi Amari. *Information geometry and its applications*. Springer, 2016.

[26] Lisa Borland. Ito-langevin equations within generalized thermostatistics. *Physics Letters A*, 245(1-2):67–72, 1998.

[27] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

[28] Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken elbo. In *International Conference on Machine Learning*, pages 159–168, 2018.

[29] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On integral probability metrics,\phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

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

[31] Kenneth Falconer. *Fractal geometry: mathematical foundations and applications*. John Wiley & Sons, 2004.

[32] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.