First Order Generative Adversarial Networks

Calvin Seward [1 2] **Thomas Unterthiner** [2] **Urs Bergmann** [1] **Nikolay Jetchev** [1] **Sepp Hochreiter** [2]

Abstract

GANs excel at learning high dimensional distributions, but they can update generator parameters in directions that do not correspond to the steepest descent direction of the objective. Prominent examples of problematic update directions include those used in both Goodfellow's original GAN and the WGAN-GP. To formally describe an optimal update direction, we introduce a theoretical framework which allows the derivation of requirements on both the divergence and corresponding method for determining an update direction, with these requirements guaranteeing unbiased minibatch updates in the direction of steepest descent. We propose a novel divergence which approximates the Wasserstein distance while regularizing the critic's first order information. Together with an accompanying update direction, this divergence fulfills the requirements for unbiased steepest descent updates. We verify our method, the First Order GAN, with image generation on CelebA, LSUN and CIFAR-10 and set a new state of the art on the One Billion Word language generation task.

1. Introduction

Generative adversarial networks (GANs) (Goodfellow et al., 2014) excel at learning generative models of complex distributions, such as images (Radford et al., 2016; Ledig et al., 2017), textures (Jetchev et al., 2016; Bergmann et al., 2017; Jetchev et al., 2017), and even texts (Gulrajani et al., 2017; Heusel et al., 2017).

GANs learn a generative model G that maps samples from multivariate random noise into a high dimensional space. The goal of GAN training is to update G such that the

[1]Zalando Research, Mühlenstraße 25, 10243 Berlin, Germany [2]LIT AI Lab & Institute of Bioinformatics, Johannes Kepler University Linz, Austria. Correspondence to: Calvin Seward <calvin.seward@zalando.de>.

generative model approximates a target probability distribution. In order to determine how close the generated and target distributions are, a class of divergences, the so-called adversarial divergences was defined and explored by (Liu et al., 2017). This class is broad enough to encompass most popular GAN methods such as the original GAN (Goodfellow et al., 2014), f-GANs (Nowozin et al., 2016), moment matching networks (Li et al., 2015), Wasserstein GANs (Arjovsky et al., 2017) and the tractable version thereof, the WGAN-GP (Gulrajani et al., 2017).

GANs learn a generative model with distribution $\mathbb{Q}$ by minimizing an objective function $\tau(\mathbb{P}\|\mathbb{Q})$ measuring the similarity between target and generated distributions $\mathbb{P}$ and $\mathbb{Q}$. In most GAN settings, the objective function to be minimized is an adversarial divergence (Liu et al., 2017), where a critic function is learned that distinguishes between target and generated data. For example, in the classic GAN (Goodfellow et al., 2014) the critic f classifies data as real or generated, and the generator G is encouraged to generate samples that f will classify as real.

Unfortunately in GAN training, the generated distribution often fails to converge to the target distribution. Many popular GAN methods are unsuccessful with toy examples, for example failing to generate all modes of a mixture of Gaussians (Srivastava et al., 2017; Metz et al., 2017) or failing to learn the distribution of data on a one-dimensional line in a high dimensional space (Fedus et al., 2017). In these situations, updates to the generator don't significantly reduce the divergence between generated and target distributions; if there always was a significant reduction in the divergence then the generated distribution would converge to the target.

The key to successful neural network training lies in the ability to efficiently obtain unbiased estimates of the gradients of a network's parameters with respect to some loss. With GANs, this idea can be applied to the generative setting. There, the generator G is parameterized by some values $\theta \in \mathbb{R}^m$. If an unbiased estimate of the gradient of the divergence between target and generated distributions with respect to θ can be obtained during mini-batch learning, then SGD can be applied to learn G.

In GAN learning, intuition would dictate updating the generated distribution by moving θ in the direction of steepest descent $-\nabla_\theta \tau(\mathbb{P}\|\mathbb{Q}_\theta)$. Unfortunately, $-\nabla_\theta \tau(\mathbb{P}\|\mathbb{Q}_\theta)$ is

generally intractable, therefore θ is updated according to a tractable method; in most cases a critic f is learned and the gradient of the expected critic value $\nabla_\theta \mathbb{E}_{\mathbb{Q}_\theta}[f]$ is used as the update direction for θ. Usually, this update direction and the direction of steepest descent $-\nabla_\theta \tau(\mathbb{P}\|\mathbb{Q}_\theta)$, don't coincide and therefore learning isn't optimal. As we see later, popular methods such as WGAN-GP (Gulrajani et al., 2017) are affected by this issue.

Therefore we set out to answer a simple but fundamental question: Is there an adversarial divergence and corresponding method that produces unbiased estimates of the direction of steepest descent in a mini-batch setting?

In this paper, under reasonable assumptions, we identify a path to such an adversarial divergence and accompanying update method. Similar to the WGAN-GP, this divergence also penalizes a critic's gradients, and thereby ensures that the critic's first order information can be used directly to obtain an update direction in the direction of steepest descent.

This program places four requirements on the adversarial divergence and the accompanying update rule for calculating the update direction that haven't to the best of our knowledge been formulated together. This paper will give rigorous definitions of these requirements, but for now we suffice with intuitive and informal definitions:

A. the divergence used must decrease as the target and generated distributions approach each other. For example, if we define the trivial distance between two probability distribution to be 0 if the distributions are equal, and 1 otherwise, i.e.

$$\tau_{\text{trivial}}(\mathbb{P}\|\mathbb{Q}) := \begin{cases} 0 & \mathbb{P} = \mathbb{Q} \\ 1 & \text{otherwise} \end{cases}$$

then even as $\mathbb{Q}$ gets close to $\mathbb{P}$, $\tau_{\text{trivial}}(\mathbb{P}\|\mathbb{Q})$ doesn't change. Without this requirement, $\nabla_\theta \tau(\mathbb{P}\|\mathbb{Q}_\theta) = 0$ and every direction is a "direction of steepest descent,"

B. critic learning must be tractable,

C. the gradient $\nabla_\theta \tau(\mathbb{P}\|\mathbb{Q}_\theta)$ and the result of an update rule must be well defined,

D. the optimal critic enables an update which is an estimate of $-\nabla_\theta \tau(\mathbb{P}\|\mathbb{Q}_\theta)$.

In order to formalize these requirements, we define in Section 2 the notions of adversarial divergences and optimal critics. In Section 3 we will apply the adversarial divergence paradigm and begin to formalize the requirements above and better understand existing GAN methods. The last requirement is defined precisely in Section 4 where we explore criteria for an update rule guaranteeing a low variance unbiased estimate of the true gradient $-\nabla_\theta \tau(\mathbb{P}\|\mathbb{Q}_\theta)$.

After stating these conditions, we devote Section 5 to defining a divergence, the Penalized Wasserstein Divergence that fulfills the first two basic requirements. In this setting, a critic is learned, that similarly to the WGAN-GP critic, pushes real and generated data as far apart as possible while being penalized if the critic violates a Lipschitz condition.

As we will discover, an optimal critic for the Penalized Wasserstein Divergence between two distributions need not be unique. In fact, this divergence only specifies the values that the optimal critic assumes on the supports of generated and target distributions. Therefore, for many distributions, multiple critics with different gradients on the support of the generated distribution can all be optimal.

We apply this insight in Section 6 and add a gradient penalty to define the First Order Penalized Wasserstein Divergence. This divergence enforces not just correct values for the critic, but also ensures that the critic's gradient, its first order information, assumes values that allow for an easy formulation of an update rule. Together, this divergence and update rule fulfill all four requirements.

We hope that this gradient penalty trick will be applied to other popular GAN methods and ensure that they too return better generator updates. Indeed, (Fedus et al., 2017) improves existing GAN methods by adding a gradient penalty.

Finally in Section 7, the effectiveness of our method is demonstrated by generating images and texts.

2. Notation, Definitions and Assumptions

In (Liu et al., 2017) an adversarial divergence is defined:

Definition 1 (Adversarial Divergence). *Let X be a topological space, $C(X^2)$ the set of all continuous real valued functions over the Cartesian product of X with itself and set $\mathcal{G} \subseteq C(X^2)$, $\mathcal{G} \neq \emptyset$. An adversarial divergence $\tau(\cdot\|\cdot)$ over X is a function*

$$\mathcal{P}(X) \times \mathcal{P}(X) \to \mathbb{R} \cup \{+\infty\}$$
$$(\mathbb{P}, \mathbb{Q}) \mapsto \tau(\mathbb{P}\|\mathbb{Q}) = \sup_{g \in \mathcal{G}} \mathbb{E}_{\mathbb{P}\otimes\mathbb{Q}}[g].$$

The function class $\mathcal{G} \subseteq C(X^2)$ must be carefully selected if $\tau(\cdot\|\cdot)$ is to be reasonable. For example, if $\mathcal{G} = C(X^2)$ then the divergence between two Dirac distributions $\tau(\delta_0\|\delta_1) = \infty$, and if $\mathcal{G} = \{0\}$, i.e. $\mathcal{G}$ contains only the constant function which assumes zero everywhere, then $\tau(\cdot\|\cdot) = 0$.

Many existing GAN procedures can be formulated as an adversarial divergence. For example, setting

$$\mathcal{G} = \{x, y \mapsto \log(u(x)) + \log(1 - u(y)) \mid u \in \mathcal{V}\}$$
$$\mathcal{V} = (0,1)^X \cap C(X)^1$$

---

[1] $(0,1)^X$ denotes all functions mapping X to $(0,1)$.

results in $\tau_G(\mathbb{P}\|\mathbb{Q}) = \sup_{g \in \mathcal{G}} \mathbb{E}_{\mathbb{P} \otimes \mathbb{Q}}[g]$, the divergence in Goodfellow's original GAN (Goodfellow et al., 2014). See (Liu et al., 2017) for further examples.

For convenience, we'll restrict ourselves to analyzing a special case of the adversarial divergence (similar to Theorem 4 of (Liu et al., 2017)), and use the notation:

Definition 2 (Critic Based Adversarial Divergence). *Let X be a topological space, $\mathcal{F} \subseteq C(X)$, $\mathcal{F} \neq \emptyset$. Further let $f \in \mathcal{F}$, $m_f : X \times X \to \mathbb{R}$, $m_f : (x,y) \mapsto m_1(f(x)) - m_2(f(y))$ and $r_f \in C(X^2)$. Then define*

$$\tau : \mathcal{P}(X) \times \mathcal{P}(X) \times \mathcal{F} \to \mathbb{R} \cup \{+\infty\}$$
$$(\mathbb{P}, \mathbb{Q}, f) \mapsto \tau(\mathbb{P}\|\mathbb{Q}; f) = \mathbb{E}_{\mathbb{P} \otimes \mathbb{Q}}[m_f - r_f] \quad (1)$$

and set $\tau(\mathbb{P}\|\mathbb{Q}) = \sup_{f \in \mathcal{F}} \tau(\mathbb{P}\|\mathbb{Q}; f)$.

For example, the τ_G from above can be equivalently defined by setting $\mathcal{F} = (0,1)^X \cap C(X)$, $m_1(x) = \log(x)$, $m_2(x) = \log(1-x)$ and $r_f = 0$. Then

$$\tau_G(\mathbb{P}\|\mathbb{Q}) = \sup_{f \in \mathcal{F}} \mathbb{E}_{\mathbb{P}}[m_f - r_f] \quad (2)$$

is a critic based adversarial divergence.

An example with a non-zero r_f is the WGAN-GP (Gulrajani et al., 2017), which is a critic based adversarial divergence when $\mathcal{F} = C^1(X)$, the set of all differentiable real functions on X, $m_1(x) = m_2(x) = x$, $\lambda > 0$ and

$$r_f(x,y) = \lambda \mathbb{E}_{\alpha \sim \mathcal{U}([0,1])}[(\|\nabla_z f(z)|_{\alpha x + (1-\alpha)y}\| - 1)^2].$$

Then the WGAN-GP divergence $\tau_I(\mathbb{P}\|\mathbb{Q})$ is:

$$\tau_I(\mathbb{P}\|\mathbb{Q}) = \sup_{f \in \mathcal{F}} \tau_I(\mathbb{P}\|\mathbb{Q}; f) = \sup_{f \in \mathcal{F}} \mathbb{E}_{\mathbb{P} \otimes \mathbb{Q}}[m_f - r_f]. \quad (3)$$

While Definition 1 is more general, Definition 2 is more in line with most GAN models. In most GAN settings, a critic in the simpler $C(X)$ space is learned that separates real and generated data while reducing some penalty term r_f which depends on both real and generated data. For this reason, we use exclusively the notation from Definition 2.

One desirable property of an adversarial divergence is that $\tau(\mathbb{P}^*\|\mathbb{P})$ obtains its infimum if and only if $\mathbb{P}^* = \mathbb{P}$, leading to the following definition adapted from (Liu et al., 2017):

Definition 3 (Strict adversarial divergence). *Let τ be an adversarial divergence over a topological space X. τ is called a strict adversarial divergence if for any $\mathbb{P}, \mathbb{P}^* \in \mathcal{P}(X)$,*

$$\tau(\mathbb{P}^*\|\mathbb{P}) = \inf_{\mathbb{P}' \in \mathcal{P}(X)} \tau(\mathbb{P}^*\|\mathbb{P}') \Rightarrow \mathbb{P}^* = \mathbb{P}$$

In order to analyze GANs that minimize a critic based adversarial divergence, we introduce the set of optimal critics.

Definition 4 (Optimal Critic, $\mathrm{OC}_\tau(\mathbb{P}, \mathbb{Q})$). *Let τ be a critic based adversarial divergence over a topological space X and $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(X)$, $\mathcal{F} \subseteq C(X)$, $\mathcal{F} \neq \emptyset$. Define $\mathrm{OC}_\tau(\mathbb{P}, \mathbb{Q})$ to be the set of critics in $\mathcal{F}$ that maximize $\tau(\mathbb{P}\|\mathbb{Q}; \cdot)$. That is*

$$\mathrm{OC}_\tau(\mathbb{P}, \mathbb{Q}) := \{f \in \mathcal{F} \mid \tau(\mathbb{P}\|\mathbb{Q}; f) = \tau(\mathbb{P}\|\mathbb{Q})\}.$$

Note that $\mathrm{OC}_\tau(\mathbb{P}, \mathbb{Q}) = \emptyset$ is possible, (Arjovsky & Bottou, 2017). In this paper, we will always assume that if $\mathrm{OC}_\tau(\mathbb{P}, \mathbb{Q}) \neq \emptyset$, then an optimal critic $f^* \in \mathrm{OC}_\tau(\mathbb{P}, \mathbb{Q})$ is known. Although is an unrealistic assumption, see (Bińkowski et al., 2018), it is a good starting point for a rigorous GAN analysis. We hope further works can extend our insights to more realistic cases of approximate critics.

Finally, we assume that generated data is distributed according to a probability distribution $\mathbb{Q}_\theta$ parameterized by $\theta \in \Theta \subseteq \mathbb{R}^m$ satisfying the mild regularity Assumption 1. Furthermore, we assume that $\mathbb{P}$ and $\mathbb{Q}$ both have compact and disjoint support in Assumption 2. Although we conjecture that weaker assumptions can be made, we decide for the stronger assumptions to simplify the proofs.

Assumption 1 (Adapted from (Arjovsky et al., 2017)). *Let $\Theta \subseteq \mathbb{R}^m$. We say $\mathbb{Q}_\theta \in \mathcal{P}(X)$, $\theta \in \Theta$ satisfies assumption 1 if there is a locally Lipschitz function $g : \Theta \times \mathbb{R}^d \to X$ which is differentiable in the first argument and a distribution $\mathbb{Z}$ with bounded support in $\mathbb{R}^d$ such that for all $\theta \in \Theta$ it holds $\mathbb{Q}_\theta \sim g(\theta, z)$ where $z \sim \mathbb{Z}$.*

Assumption 2 (Compact and Disjoint Distributions). *Using $\Theta \subseteq \mathbb{R}^m$ from Assumption 1, we say that $\mathbb{P}$ and $(\mathbb{Q}_\theta)_{\theta \in \Theta}$ satisfies Assumption 2 if for all $\theta \in \Theta$ it holds that the supports of $\mathbb{P}$ and $\mathbb{Q}_\theta$ are compact and disjoint.*

3. Requirements Derived From Related Work

With the concept of an Adversarial Divergence now formally defined, we can investigate existing GAN methods from an Adversarial Divergence minimization standpoint. During the last few years, weaknesses in existing GAN frameworks have been highlighted and new frameworks have been proposed to mitigate or eliminate these weaknesses. In this section we'll trace this history and formalize requirements for adversarial divergences and optimal updates.

Although using two competing neural networks for unsupervised learning isn't a new concept (Schmidhuber, 1992), recent interest in the field started when (Goodfellow et al., 2014) generated images with the divergence τ_G defined in Eq. 2. However, (Arjovsky & Bottou, 2017) shows if $\mathbb{P}, \mathbb{Q}_\theta$ have compact disjoint support then $\nabla_\theta \tau_G(\mathbb{P}\|\mathbb{Q}_\theta) = 0$, preventing the use of gradient based learning methods.

In response to this impediment, the Wasserstein GAN was proposed in (Arjovsky et al., 2017) with the divergence:

$$\tau_W(\mathbb{P}\|\mathbb{Q}) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{x \sim \mathbb{P}}[f(x)] - \mathbb{E}_{x' \sim \mathbb{Q}}[f(x')]$$

where $\|f\|_L$ is the Lipschitz constant of f. The following example shows the advantage of τ_W. Consider a series of Dirac measures $(\delta_{\frac{1}{n}})_{n>0}$. Then $\tau_W(\delta_0\|\delta_{\frac{1}{n}}) = \frac{1}{n}$ while $\tau_G(\delta_0\|\delta_{\frac{1}{n}}) = 1$. As $\delta_{\frac{1}{n}}$ approaches δ_0, the Wasserstein divergence decreases while $\tau_G(\delta_0\|\delta_{\frac{1}{n}})$ remains constant.

This issue is explored in (Liu et al., 2017) by creating a weak ordering, the so-called strength, of divergences. A divergence τ_1 is said to be stronger than τ_2 if for any sequence of probability measures $(\mathbb{P}_n)_{n\in\mathbb{N}}$ and any target probability measure $\mathbb{P}^*$ the convergence $\tau_1(\mathbb{P}^*\|\mathbb{P}_n) \xrightarrow{n\to\infty} \inf_{\mathbb{P}\in\mathcal{P}(X)} \tau_1(\mathbb{P}^*\|\mathbb{P})$ implies $\tau_2(\mathbb{P}^*\|\mathbb{P}_n) \xrightarrow{n\to\infty} \inf_{\mathbb{P}\in\mathcal{P}(X)} \tau_2(\mathbb{P}^*\|\mathbb{P})$. The divergences τ_1 and τ_2 are equivalent if τ_1 is stronger than τ_2 and τ_2 is stronger than τ_1. The Wasserstein distance τ_W is the weakest divergence in the class of strict adversarial divergences (Liu et al., 2017), leading to the following requirement:

Requirement 1 (Equivalence to τ_W). *An adversarial divergence τ is said to fulfill Requirement 1 if τ is a strict adversarial divergence which is weaker than τ_W.*

The issue of the zero gradients was side stepped in (Goodfellow et al., 2014) (and the option more rigorously explored in (Fedus et al., 2017)) by not updating with $-\nabla_\theta \mathbb{E}_{x'\sim\mathbb{Q}_\theta}[\log(1 - f(x'))]$ but instead using the gradient $\nabla_\theta \mathbb{E}_{x'\sim\mathbb{Q}_\theta}[f(x')]$. As will be shown in Section 4, this update direction doesn't generally move θ in the direction of steepest descent.

Although using the Wasserstein distance as a divergence between probability measures solves many theoretical problems, it requires that critics are Lipschitz continuous with Lipschitz constant 1. Unfortunately, no tractable algorithm has yet been found that is able to learn the optimal Lipschitz continuous critic (or a close approximation thereof).

This is due in part to the fact that if the critic is parameterized by a neural network f_ϑ, $\vartheta \in \Theta_C \subseteq \mathbb{R}^c$, then the set of admissible parameters $\{\vartheta \in \Theta_C \mid \|f_\vartheta\|_L \leq 1\}$ is highly non-convex. Thus critic learning is a non-convex optimization problem (as is generally the case in neural network learning) with non-convex constraints on the parameters. Since neural network learning is generally an unconstrained optimization problem, adding complex non-convex constraints makes learning intractable with current methods. Thus, finding an optimal Lipschitz continuous critic is a problem that can not yet be solved, leading to the second requirement:

Requirement 2 (Convex Admissible Critic Parameter Set). *Assume τ is a critic based adversarial divergence where critics are chosen from a set $\mathcal{F}$. Assume further that in training, a parameterization $\vartheta \in \mathbb{R}^c$ of the critic function f_ϑ is learned. The critic based adversarial divergence τ is said to fulfill requirement 2 if the set of admissible parameters $\{\vartheta \in \mathbb{R}^c \mid f_\vartheta \in \mathcal{F}\}$ is convex.*

Table 1. Comparing existing GAN methods with regard to the four Requirements formulated in this paper. The methods compared are the classic GAN (Goodfellow et al., 2014), WGAN (Arjovsky et al., 2017), WGAN-GP (Gulrajani et al., 2017), WGAN-LP (Petzka et al., 2018), DRAGAN (Kodali et al., 2017), PWGAN (our method) and FOGAN (our method).

	Req. 1	Req. 2	Req. 3	Req. 4
GAN	no	✓	✓	no
WGAN	✓	no	✓	✓
WGAN-GP	✓	✓	no	no
WGAN-LP	✓	✓	no	no
DRAGAN	✓	✓	✓	no
PWGAN	✓	✓	no	no
FOGAN	✓	✓	✓	✓

It was reasoned in (Gulrajani et al., 2017) that since a Wasserstein critic must have gradients of norm at most 1 everywhere, a reasonable strategy would be to transform the constrained optimization into an unconstrained optimization problem by penalizing the divergence when a critic has non-unit gradients. With this strategy, the so-called Improved Wasserstein GAN or WGAN-GP divergence defined in Eq. 3 is obtained.

The generator parameters are updated by training an optimal critic f^* and updating with $\nabla_\theta \mathbb{E}_{\mathbb{Q}_\theta}[f^*]$. Although this method has impressive experimental results, it is not yet ideal. (Petzka et al., 2018) showed that an optimal critic for τ_I has undefined gradients on the support of the generated distribution $\mathbb{Q}_\theta$. Thus, the update direction $\nabla_\theta \mathbb{E}_{\mathbb{Q}_\theta}[f^*]$ is undefined; even if a direction was chosen from the sub-gradient field (meaning the update direction is defined but random) the update direction won't generally point in the direction of steepest gradient descent. This naturally leads to the next requirement:

Requirement 3 (Well Defined Update Rule). *An update rule is said to fulfill Requirement 3 on a target distribution $\mathbb{P}$ and a family of generated distributions $(\mathbb{Q}_\theta)_{\theta\in\Theta}$ if for every $\theta \in \Theta$ the update rule at $\mathbb{P}$ and $\mathbb{Q}_\theta$ is well defined.*

Note that kernel methods such as (Dziugaite et al., 2015) and (Li et al., 2015) provide exciting theoretical guarantees and may well fulfill all four requirements. Since these guarantees come at a cost in scalability, we won't analyze them further.

4. Correct Update Rule Requirement

In the previous section, we stated a bare minimum requirement for an update rule (namely that it is well defined). In this section, we'll go further and explore criteria for a "good" update rule. For example in Lemma 8 in Section A of Appendix, it is shown that there exists a target $\mathbb{P}$ and a family

of generated distributions $(\mathbb{Q}_\theta)_{\theta \in \Theta}$ fulfilling Assumptions 1 and 2 such that for the optimal critic $f_{\theta_0}^* \in \mathrm{OC}_{\tau_I}(\mathbb{P}, \mathbb{Q}_{\theta_0})$ there is no $\gamma \in \mathbb{R}$ so that

$$\nabla_\theta \tau_I(\mathbb{P}\|\mathbb{Q}_\theta)|_{\theta_0} = \gamma \nabla_\theta \mathbb{E}_{\mathbb{Q}_\theta}[f_{\theta_0}^*]|_{\theta_0}$$

for all $\theta_0 \in \Theta$ if all terms are well defined. Thus, the update rule used in the WGAN-GP setting, although well defined for this specific $\mathbb{P}$ and $\mathbb{Q}_{\theta_0}$, isn't guaranteed to move θ in the direction of steepest descent. In fact, (Mescheder et al., 2018) shows that the WGAN-GP does not converge for specific classes of distributions. Therefore, the question arises what well defined update rule also moves θ in the direction of steepest descent?

The most obvious candidate for an update rule is simply use the direction $-\nabla_\theta \tau(\mathbb{P}\|\mathbb{Q}_\theta)$, but since in the adversarial divergence setting $\tau(\mathbb{P}\|\mathbb{Q}_\theta)$ is the supremum over a set of infinitely many possible critics, calculating $-\nabla_\theta \tau(\mathbb{P}\|\mathbb{Q}_\theta)$ directly is generally intractable.

One strategy to address this issue is to use an envelope theorem (Milgrom & Segal, 2002). Assuming all terms are well defined, then for every optimal critic $f^* \in \mathrm{OC}_\tau(\mathbb{P}, \mathbb{Q}_{\theta_0})$ it holds $\nabla_\theta \tau(\mathbb{P}\|\mathbb{Q}_\theta)|_{\theta_0} = \nabla_\theta \tau(\mathbb{P}\|\mathbb{Q}_\theta; f^*)|_{\theta_0}$. This strategy is outlined in detail in (Arjovsky et al., 2017) when proving the Wasserstein GAN update rule, and explored in the context of the classic GAN divergence τ_G in (Arjovsky & Bottou, 2017).

Yet in many GAN settings, (Goodfellow et al., 2014; Arjovsky et al., 2017; Salimans et al., 2016; Petzka et al., 2018), the update rule is to train an optimal critic f^* and then take a step in the direction of $\nabla_\theta \mathbb{E}_{\mathbb{Q}_\theta}[f^*]$. In the critic based adversarial divergence setting (Definition 2), a direct result of Eq. 1 together with Theorem 1 from (Milgrom & Segal, 2002) is that for every $f^* \in \mathrm{OC}_\tau(\mathbb{P}, \mathbb{Q}_{\theta_0})$

$$\begin{aligned}
\nabla_\theta \tau(\mathbb{P}\|\mathbb{Q}_\theta)|_{\theta_0} &= \nabla_\theta \tau(\mathbb{P}\|\mathbb{Q}; f^*) \\
&= -\nabla_\theta (\mathbb{E}_{\mathbb{Q}_\theta}[m_2(f^*)] + \mathbb{E}_{\mathbb{P}\otimes\mathbb{Q}_\theta}[r_{f^*}])|_{\theta_0}
\end{aligned}$$
(4)

when all terms are well defined. Thus, the update direction $\nabla_\theta \mathbb{E}_{\mathbb{Q}_\theta}[f^*]$ only points in the direction of steepest descent for special choices of m_2 and r_f. One such example is the Wasserstein GAN where $m_2(x) = x$ and $r_f = 0$.

Most popular GAN methods don't employ functions m_2 and r_f such that the update direction $\nabla_\theta \mathbb{E}_{\mathbb{Q}_\theta}[f^*]$ points in the direction of steepest descent. For example, with the classic GAN, $m_2(x) = \log(1-x)$ and $r_f = 0$, so the update direction $\nabla_\theta \mathbb{E}_{\mathbb{Q}_\theta}[f^*]$ clearly is not oriented in the direction of steepest descent $\nabla_\theta \mathbb{E}_{\mathbb{Q}_\theta}[\log(1-f^*)]$. The WGAN-GP is similar, since as we see in Lemma 8 in Appendix, Section A, $\nabla_\theta \mathbb{E}_{\mathbb{P}\otimes\mathbb{Q}_\theta}[r_{f^*}]$ is not generally a multiple of $\nabla_\theta \mathbb{E}_{\mathbb{Q}_\theta}[f^*]$.

The question arises why this direction is used instead of directly calculating the direction of steepest descent? Using

the correct update rule in Eq. 4 above involves estimating $\nabla_\theta \mathbb{E}_{\mathbb{P}\otimes\mathbb{Q}_\theta}[r_{f^*}]$, which requires sampling from both $\mathbb{P}$ and $\mathbb{Q}_\theta$. GAN learning happens in mini-batches, therefore $\nabla_\theta \mathbb{E}_{\mathbb{P}\otimes\mathbb{Q}_\theta}[r_{f^*}]$ isn't calculated directly, but estimated based on samples which can lead to variance in the estimate.

To analyze this issue, we use the notation from (Bellemare et al., 2017) where $\mathbf{X}_m := X_1, X_2, \ldots, X_m$ are samples from $\mathbb{P}$ and the empirical distribution $\hat{\mathbb{P}}_m$ is defined by $\hat{\mathbb{P}}_m := \hat{\mathbb{P}}_m(\mathbf{X}_m) := \frac{1}{m} \sum_{i=1}^m \delta_{X_i}$. Further let $\mathbb{V}_{\mathbf{X}_m \sim \mathbb{P}}$ be the element-wise variance. Now with mini-batch learning we get[2]

$$\begin{aligned}
&\mathbb{V}_{\mathbf{X}_m \sim \mathbb{P}}[\nabla_\theta \mathbb{E}_{\hat{\mathbb{P}}_m \otimes \mathbb{Q}_\theta}[m_{f^*} - r_{f^*}]|_{\theta_0}] \\
&= \mathbb{V}_{\mathbf{X}_m \sim \mathbb{P}}[\nabla_\theta (\mathbb{E}_{\hat{\mathbb{P}}_m}[m_1(f^*)] - \mathbb{E}_{\mathbb{Q}_\theta}[m_2(f^*)] \\
&\qquad\qquad - \mathbb{E}_{\hat{\mathbb{P}}_m \otimes \mathbb{Q}_\theta}[r_{f^*}])|_{\theta_0}] \\
&= \mathbb{V}_{\mathbf{X}_m \sim \mathbb{P}}[\nabla_\theta \mathbb{E}_{\hat{\mathbb{P}}_m \otimes \mathbb{Q}_\theta}[r_{f^*}]|_{\theta_0}].
\end{aligned}$$

Therefore, estimation of $\nabla_\theta \mathbb{E}_{\mathbb{P}\otimes\mathbb{Q}_\theta}[r_{f^*}]$ is an extra source of variance.

Our solution to both these problems chooses the critic based adversarial divergence τ in such a way that there exists a $\gamma \in \mathbb{R}$ so that for all optimal critics $f^* \in \mathrm{OC}_\tau(\mathbb{P}, \mathbb{Q}_{\theta_0})$ it holds

$$\nabla_\theta \mathbb{E}_{\mathbb{P}\otimes\mathbb{Q}_\theta}[r_{f^*}]|_{\theta_0} \approx \gamma \nabla_\theta \mathbb{E}_{\mathbb{Q}_\theta}[m_2(f^*)]|_{\theta_0}. \quad (5)$$

In Theorem 2 we see conditions on $\mathbb{P}, \mathbb{Q}_\theta$ such that equality holds. Now using Eq. 5 we see that

$$\begin{aligned}
\nabla_\theta \tau(\mathbb{P}\|\mathbb{Q}_\theta)|_{\theta_0} &= -\nabla_\theta (\mathbb{E}_{\mathbb{Q}_\theta}[m_2(f^*)] + \mathbb{E}_{\mathbb{P}\otimes\mathbb{Q}_\theta}[r_{f^*}])|_{\theta_0} \\
&\approx -\nabla_\theta (\mathbb{E}_{\mathbb{Q}_\theta}[m_2(f^*)] + \gamma \mathbb{E}_{\mathbb{Q}_\theta}[m_2(f^*)])|_{\theta_0} \\
&= -(1+\gamma)\nabla_\theta \mathbb{E}_{\mathbb{Q}_\theta}[m_2(f^*)]|_{\theta_0}
\end{aligned}$$

making $(1+\gamma)\nabla_\theta \mathbb{E}_{\mathbb{Q}_\theta}[m_2(f^*)]$ a low variance update approximation of the direction of steepest descent.

We're then able to have the best of both worlds. On the one hand, when r_f serves as a penalty term, training of a critic neural network can happen in an unconstrained optimization fashion like with the WGAN-GP. At the same time, the direction of steepest descent can be approximated by calculating $\nabla_\theta \mathbb{E}_{\mathbb{Q}_\theta}[m_2(f^*)]$, and as in the Wasserstein GAN we get reliable gradient update steps.

With this motivation, Eq. 5 forms the basis of our final requirement:

Requirement 4 (Low Variance Update Rule). *An adversarial divergence τ is said to fulfill requirement 4 if τ is a critic based adversarial divergence and for every optimal critic $f^* \in \mathrm{OC}_\tau(\mathbb{P}, \mathbb{Q}_{\theta_0})$ fulfills Eq. 5.*

---

[2]Because the first expectation doesn't depend on θ, $\nabla_\theta \mathbb{E}_{\hat{\mathbb{P}}_m}[m_1(f^*)] = 0$. In the same way, because the second expectation doesn't depend on the mini-batch $\mathbf{X}_m$ sampled, $\mathbb{V}_{\mathbf{X}_m \sim \mathbb{P}}[\mathbb{E}_{\mathbb{Q}_\theta}[m_2(f^*)]] = 0$.

It should be noted that the WGAN-GP achieves impressive experimental results; we conjecture that in many cases $\nabla_\theta \mathbb{E}_{\mathbb{Q}_\theta}[f^*]$ close enough to the true direction of steepest descent. Nevertheless, as the experiments in Section 7 show, our gradient estimates lead to better convergence in a challenging language modeling task.

5. Penalized Wasserstein Divergence

We now attempt to find an adversarial divergence that fulfills all four requirements. We start by formulating an adversarial divergence τ_P and a corresponding update rule than can be shown to comply with Requirements 1 and 2. Subsequently in Section 6, τ_P will be refined to make its update rule practical and conform to all four requirements.

The divergence τ_P is inspired by the Wasserstein distance, there for an optimal critic between two Dirac distributions $f^* \in \mathrm{OC}_\tau(\delta_a, \delta_b)$ it holds $f(a) - f(b) = |a - b|$. Now if we look at

$$\tau_{\mathrm{simple}}(\delta_a \| \delta_b) := \sup_{f \in \mathcal{F}} f(a) - f(b) - \frac{(f(a) - f(b))^2}{|a - b|} \quad (6)$$

it's easy to calculate that $\tau_{\mathrm{simple}}(\delta_a \| \delta_b) = \frac{1}{4}|a - b|$, which is the same up to a constant (in this simple setting) as the Wasserstein distance, without being a constrained optimization problem. See Figure 1 for an example.

This has another intuitive explanation. Because Eq. 6 can be reformulated as

$$\tau_{\mathrm{simple}}(\delta_a \| \delta_b) = \sup_{f \in \mathcal{F}} f(a) - f(b) - |a - b| \left(\frac{f(a) - f(b)}{|a - b|} \right)^2$$

which is a tug of war between the objective $f(a) - f(b)$ and the squared Lipschitz penalty $\frac{|f(a) - f(b)|}{|a - b|}$ weighted by $|a - b|$. This $|a - b|$ term is important (and missing from (Gulrajani et al., 2017), (Petzka et al., 2018)) because otherwise the slope of the optimal critic between a and b will depend on $|a - b|$.

The penalized Wasserstein divergence τ_P is a straightforward adaptation of τ_{simple} to the multi dimensional case.

Definition 5 (Penalized Wasserstein Divergence). *Assume $X \subseteq \mathbb{R}^n$ and $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(X)$ are probability measures over X, $\lambda > 0$ and $\mathcal{F} = C^1(X)$. Set*

$$\tau_P(\mathbb{P} \| \mathbb{Q}; f) := \mathbb{E}_{x \sim \mathbb{P}}[f(x)] - \mathbb{E}_{x' \sim \mathbb{Q}}[f(x')]$$
$$- \lambda \, \mathbb{E}_{x \sim \mathbb{P}, x' \sim \mathbb{Q}} \left[\frac{(f(x) - f(x'))^2}{\|x - x'\|} \right].$$

Define the penalized Wasserstein divergence as

$$\tau_P(\mathbb{P} \| \mathbb{Q}) = \sup_{f \in \mathcal{F}} \tau_P(\mathbb{P} \| \mathbb{Q}; f).$$

This divergence is updated by picking an optimal critic $f^* \in \mathrm{OC}_{\tau_P}(\mathbb{P}, \mathbb{Q}_{\theta_0})$ and taking a step in the direction of $\nabla_\theta \mathbb{E}_{\mathbb{Q}_\theta}[f^*]|_{\theta_0}$.

This formulation is similar to the WGAN-GP (Gulrajani et al., 2017), restated here in Eq. 3.

Theorem 1. *Assume $X \subseteq \mathbb{R}^n$, and $\mathbb{P}, \mathbb{Q}_\theta \in \mathcal{P}(X)$ are probability measures over X fulfilling Assumptions 1 and 2. Then for every $\theta_0 \in \Theta$ the Penalized Wasserstein Divergence with it's corresponding update direction fulfills Requirements 1 and 2.*

Further, there exists an optimal critic $f^* \in \mathrm{OC}_{\tau_P}(\mathbb{P}, \mathbb{Q}_{\theta_0})$ that fulfills Eq. 5 and thus Requirement 4.

Proof. See Appendix, Section A. □

Note that this theorem isn't unique to τ_P. For example, for the penalty in Eq. 8 of (Petzka et al., 2018) we conjecture that a similar result can be shown. The divergence τ_P is still very useful because, as will be shown in the next section, τ_P can be modified slightly to obtain a new critic τ_F, where every optimal critic fulfills Requirements 1 to 4.

Since τ_P only constrains the value of a critic on the supports of $\mathbb{P}$ and $\mathbb{Q}_\theta$, many different critics are optimal, and in general $\nabla_\theta \mathbb{E}_{\mathbb{Q}_\theta}[f^*]$ depends on the optimal critic choice and is thus is not well defined. With this, Requirements 3 and 4 are not fulfilled. See Figure 1 for a simple example.

In theory, τ_P's critic could be trained with a modified sampling procedure so that $\nabla_\theta \mathbb{E}_{\mathbb{Q}_\theta}[f^*]$ is well defined and Eq. 5 holds, as is done in both (Kodali et al., 2017) and (Unterthiner et al., 2018). By using a method similar to (Bishop et al., 1998), one can minimize the divergence $\tau_P(\mathbb{P}, \hat{\mathbb{Q}}_\theta)$ where $\hat{\mathbb{Q}}_\theta$ is data equal to $x' + \epsilon$ where x' is sampled from $\mathbb{Q}_\theta$ and ϵ is some zero-mean uniform distributed noise. In this way the support of $\hat{\mathbb{Q}}_\theta$ lives in the full space X and not the submanifold $\mathrm{supp}(\mathbb{Q}_\theta)$. Unfortunately, while this method works in theory, the number of samples required for accurate gradient estimates scales with the dimensionality of the underlying space X, not with the dimensionality of data or generated submanifolds $\mathrm{supp}(\mathbb{P})$ or $\mathrm{supp}(\mathbb{Q}_\theta)$. In response, we propose the First Order Penalized Wasserstein Divergence.

6. First Order Penalized Wasserstein Divergence

As was seen in the last section, since τ_P only constrains the value of optimal critics on the supports of $\mathbb{P}$ and $\mathbb{Q}_\theta$, the gradient $\nabla_\theta \mathbb{E}_{\mathbb{Q}_\theta}[f^*]$ is not well defined. A natural method to refine τ_P to achieve a well defined gradient is to enforce two things:
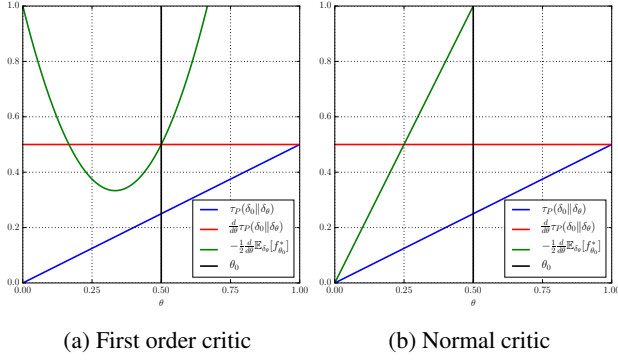
(a) First order critic   (b) Normal critic

Figure 1. Comparison of τ_P update rule given different optimal critics. Consider the simple example of divergence τ_P from Definition 5 between Dirac measures with update rule $\frac{1}{2}\frac{d}{d\theta}\mathbb{E}_{\delta_\theta}[f]$ (the update rule is from Lemma 7 in Appendix, Section A). Recall that $\tau_P(\delta_0\|\delta_\theta; f) = -f(\theta) - \frac{(f(\theta))^2}{2\theta}$, and that so $\frac{d}{d\theta}\tau_P(\delta_0\|\delta_\theta) = \frac{1}{2}$. Let $\theta_0 = 0.5$; our goal is to calculate $\frac{d}{d\theta}\tau_P(\delta_0\|\delta_\theta)|_{\theta=\theta_0}$ via our update rule. Since multiple critics are optimal for $\tau_P(\delta_0\|\delta_\theta)$, we will explore how the choice of optimal critic affects the update. In Subfigure 1a, we chose the first order optimal critic $f_{\theta_0}^*(x) = x(-4x^2 + 4x - 2)$, and $\frac{d}{d\theta}\tau_P(\delta_0\|\delta_\theta)|_{\theta=\theta_0} = -\frac{1}{2}\frac{d}{d\theta}\mathbb{E}_{\delta_\theta}[f_{\theta_0}^*]|_{\theta=\theta_0}$ and the update rule is correct (see how the red, black and green lines all intersect in one point). In Subfigure 1b, the optimal critic is set to $f_{\vartheta_0}^*(x) = -2x^2$ which is not a first order critic resulting in the update rule calculating an incorrect update.

- f^* should be optimal on a larger manifold, namely the manifold $\mathbb{Q}'_\theta$ that is created by "stretching" $\mathbb{Q}_\theta$ bit in the direction of $\mathbb{P}$ (the formal definition is below).

- The norm of the gradient of the optimal critic, $\|\nabla_x f^*(x)\|$ on $\text{supp}(\mathbb{Q}'_\theta)$ should be equal to the norm of the maximal directional derivative in the support of $\mathbb{Q}'_\theta$ (see Eq. 15 in Appendix).

By enforcing these two points, we assure that $\nabla_x f^*(x)$ is well defined and points towards the real data $\mathbb{P}$. Thus, the following definition emerges (see proof of Lemma 7 in Appendix, Section A for details).

Definition 6 (First Order Penalized Wasserstein Divergence (FOGAN)). *Assume $X \subseteq \mathbb{R}^n$ and $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(X)$ are probability measures over X. Set $\mathcal{F} = C^1(X)$, $\lambda, \mu > 0$ and*

$$\tau_F(\mathbb{P}\|\mathbb{Q}; f) := \mathbb{E}_{x\sim\mathbb{P}}[f(x)] - \mathbb{E}_{x'\sim\mathbb{Q}}[f(x')]$$

$$- \lambda \mathop{\mathbb{E}}_{x\sim\mathbb{P}, x'\sim\mathbb{Q}}\left[\frac{(f(x) - f(x'))^2}{\|x - x'\|}\right]$$

$$- \mu \mathop{\mathbb{E}}_{x\sim\mathbb{P}, x'\sim\mathbb{Q}}\left(\left\|\nabla_x f(x)\big|_{x'}\right\| - \frac{\left\|\mathbb{E}_{\tilde{x}\sim\mathbb{P}}[(\tilde{x} - x')\frac{f(\tilde{x})-f(x')}{\|x'-\tilde{x}\|^3}]\right\|}{\mathbb{E}_{\tilde{x}\sim\mathbb{P}}[\frac{1}{\|x'-\tilde{x}\|}]}\right)^2$$

Define the First Order Penalized Wasserstein Divergence as

$$\tau_F(\mathbb{P}\|\mathbb{Q}) = \sup_{f\in\mathcal{F}} \tau_F(\mathbb{P}\|\mathbb{Q}; f).$$

This divergence is updated by picking an optimal critic $f^* \in \text{OC}_{\tau_P}(\mathbb{P}, \mathbb{Q}_{\theta_0})$ and taking a step in the direction of $\nabla_\theta \mathbb{E}_{\mathbb{Q}_\theta}[f^*]|_{\theta_0}$.

In order to define a GAN from the First Order Penalized Wasserstein Divergence, we must define a slight modification of the generated distribution $\mathbb{Q}_\theta$ to obtain $\mathbb{Q}'_\theta$. Similar to the WGAN-GP setting, samples from $\mathbb{Q}'_\theta$ are obtained by $x' - \alpha(x' - x)$ where $x \sim \mathbb{P}$ and $x' \sim \mathbb{Q}_\theta$. The difference is that $\alpha \sim \mathcal{U}([0, \varepsilon])$, with ε chosen small, making $\mathbb{Q}_\theta$ and $\mathbb{Q}'_\theta$ quite similar. Therefore updates to θ that reduce $\tau_F(\mathbb{P}\|\mathbb{Q}'_\theta)$ also reduce $\tau_F(\mathbb{P}\|\mathbb{Q}_\theta)$.

Conveniently, as is shown in Lemma 5 in Appendix, Section A, any optimal critic for the First Order Penalized Wasserstein divergence is also an optimal critic for the Penalized Wasserstein Divergence. The key advantage to the First Order Penalized Wasserstein Divergence is that for any $\mathbb{P}$, $\mathbb{Q}_\theta$ fulfilling Assumptions 1 and 2, $\tau_F(\mathbb{P}\|\mathbb{Q}'_\theta)$ with its corresponding update rule $\nabla_\theta \mathbb{E}_{\mathbb{Q}'_\theta}[f^*]$ on the slightly modified probability distribution $\mathbb{Q}'_\theta$ fulfills requirements 3 and 4.

Theorem 2. *Assume $X \subseteq \mathbb{R}^n$, and $\mathbb{P}, \mathbb{Q}_\theta \in \mathcal{P}(X)$ are probability measures over X fulfilling Assumptions 1 and 2 and $\mathbb{Q}'_\theta$ is $\mathbb{Q}_\theta$ modified using the method above. Then for every $\theta_0 \in \Theta$ there exists at least one optimal critic $f^* \in \text{OC}_{\tau_F}(\mathbb{P}, \mathbb{Q}'_{\theta_0})$ and τ_F combined with update direction $\nabla_\theta \mathbb{E}_{\mathbb{Q}_\theta}[f^*]|_{\theta_0}$ fulfills Requirements 1 to 4. If $\mathbb{P}, \mathbb{Q}'_\theta$ are such that $\forall x, x' \in \text{supp}(\mathbb{P}), \text{supp}(\mathbb{Q}'_\theta)$ it holds $f^*(x) - f^*(x') = c\|x - x'\|$ for some constant c, then equality holds for Eq. 5.*

Proof. See Appendix, Section A   □

Note that adding a gradient penalty, other than being a necessary step for the WGAN-GP (Gulrajani et al., 2017), DRAGAN (Kodali et al., 2017) and Consensus Optimization GAN (Mescheder et al., 2017), has also been shown empirically to improve the performance the original GAN method (Eq. 2), see (Fedus et al., 2017). In addition, using stricter assumptions on the critic, (Nagarajan & Kolter, 2017) provides a theoretical justification for use of a gradient penalty in GAN learning. The analysis of Theorem 2 in Appendix, Section A provides a theoretical understanding why in the Penalized Wasserstein GAN setting adding a gradient penalty causes $\nabla_\theta \mathbb{E}_{\mathbb{Q}'_\theta}[f^*]$ to be an update rule that points in the direction of steepest descent, and may provide a path for other GAN methods to make similar assurances.

7. Experimental Results

7.1. Image Generation

We begin by testing the FOGAN on the CelebA image generation task (Liu et al., 2015), training a generative model with the DCGAN architecture (Radford et al., 2016) and

Table 2. Comparison of different GAN methods for image and text generation. We measure performance with respect to the FID on the image datasets and JSD between n-grams for text generation.

Task	BEGAN	DCGAN	Coulomb	WGAN-GP	FOGAN
CelebA	28.5	12.5	9.3	4.2	6.0
LSUN	112	57.5	31.2	9.5	11.4
CIFAR-10	-	-	27.3	24.8	27.4
4-gram	-	-	-	$.220 \pm .006$	$.226 \pm .006$
6-gram	-	-	-	$.573 \pm .009$	$.556 \pm .004$

obtaining Fréchet Inception Distance (FID) scores (Heusel et al., 2017) competitive with state of the art methods without doing a tuning parameter search. Similarly, we show competitive results on LSUN (Yu et al., 2015) and CIFAR-10 (Krizhevsky & Hinton, 2009). See Table 2, Appendix B.1 and released code [3].

7.2. One Billion Word

Finally, we use the First Order Penalized Wasserstein Divergence to train a character level generative language model on the One Billion Word Benchmark (Chelba et al., 2013). In this setting, a 1D CNN deterministically transforms a latent vector into a $32 \times C$ matrix, where C is the number of possible characters. A softmax nonlinearity is applied to this output, and given to the critic. Real data is one-hot encoding of 32 character texts sampled from the true data.

We conjecture this is an especially difficult task for GANs, since data in the target distribution lies in just a few corners of the $32 \times C$ dimensional unit hypercube. As the generator is updated, it must push mass from one corner to another, passing through the interior of the hypercube far from any real data. Methods other than Coulomb GAN (Unterthiner et al., 2018) WGAN-GP (Gulrajani et al., 2017; Heusel et al., 2017) and the Sobolev GAN (Mroueh et al., 2018) have not been shown to be successful at this task.

We use the same setup as in both (Gulrajani et al., 2017; Heusel et al., 2017) with two differences. First, we train to minimize our divergence from Definition 6 with parameters $\lambda = 0.1$ and $\mu = 1.0$ instead of the WGAN-GP divergence. Second, we use batch normalization in the generator, both for training our FOGAN method and the benchmark WGAN-GP; we do this because batch normalization improved performance and stability of both models.

As with (Gulrajani et al., 2017; Heusel et al., 2017) we use the Jensen-Shannon-divergence (JSD) between n-grams from the model and the real world distribution as an evaluation metric. The JSD is estimated by sampling a finite number of 32 character vectors, and comparing the distributions of the n-grams from said samples and true data. This estimation is biased; smaller samples result in larger
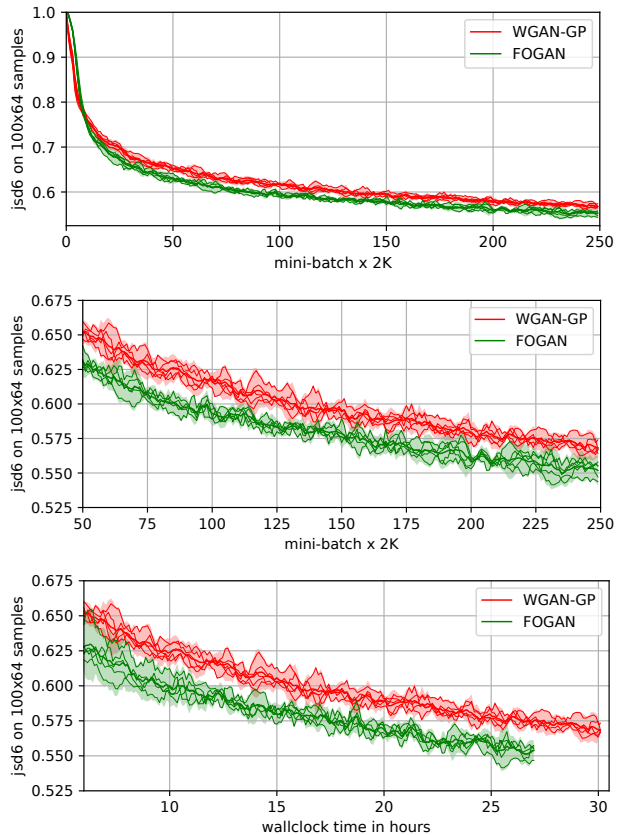
[3] https://github.com/zalandoresearch/first_order_gan

Figure 2. Five training runs of both WGAN-GP and FOGAN, with the average of all runs plotted in bold and the 2σ error margins denoted by shaded regions. For easy visualization, we plot the moving average of the last three n-gram JSD estimations. The first two plots both show training w.r.t. number of training iterations; the second plot starts at iteration 50. The last plot show training with respect to wall-clock time, starting after 6 hours of training.

JSD estimations. A Bayes limit results from this bias; even when samples are drawn from real world data and compared with real world data, small sample sizes results in large JSD estimations. In order to detect performance difference when training with the FOGAN and WGAN-GP, a low Bayes limit is necessary. Thus, to compare the methods, we sampled 6400 32 character vectors in contrast with the 640 vectors sampled in past works. Therefore, the JSD values in those papers are higher than the results here.

For our experiments we trained both models for 500, 000 iterations in 5 independent runs, estimating the JSD between 6-grams of generated and real world data every 2000 training steps, see Figure 2. The results are even more impressive when aligned with wall-clock time. Since in WGAN-GP training an extra point between real and generated distributions must be sampled, it is slower than the FOGAN training; see Figure 2 and observe the significant (2σ) drop in estimated JSD.

Acknowledgements

References

Arjovsky, M. and Bottou, L. Towards principled methods for training generative adversarial networks. In *International Conference of Learning Representations (ICLR)*, 2017.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Bellemare, M. G., Danihelka, I., Dabney, W., Mohamed, S., Lakshminarayanan, B., Hoyer, S., and Munos, R. The cramer distance as a solution to biased wasserstein gradients. *arXiv preprint arXiv:1705.10743*, 2017.

Bergmann, U., Jetchev, N., and Vollgraf, R. Learning texture manifolds with the periodic spatial GAN. In *Proceedings of the 34th International Conference on Machine Learning, (ICML)*, 2017.

Bishop, C. M., Svensén, M., and Williams, C. K. GTM: The generative topographic mapping. *Neural computation*, 10 (1):215–234, 1998.

Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. Demystifying MMD GANs. In *International Conference on Learning Representations (ICLR)*, 2018.

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., and Robinson, T. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*, 2013.

Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. Training generative neural networks via maximum mean discrepancy optimization. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence (UAI)*, 2015.

Fedus, W., Rosca, M., Lakshminarayanan, B., Dai, A. M., Mohamed, S., and Goodfellow, I. Many paths to equilibrium: GANs do not need to decrease a divergence at every step. *arXiv preprint arXiv:1710.08446*, 2017.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27 (NIPS)*, 2014.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved Training of Wasserstein GANs. In *Advances in Neural Information Processing Systems 30 (NIPS)*, 2017.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30 (NIPS)*, 2017.

Jetchev, N., Bergmann, U., and Vollgraf, R. Texture synthesis with spatial generative adversarial networks. *arXiv preprint arXiv:1611.08207*, 2016.

Jetchev, N., Bergmann, U., and Seward, C. GANosaic: Mosaic creation with generative texture manifolds. *arXiv preprint arXiv:1712.00269*, 2017.

Kodali, N., Abernethy, J., Hays, J., and Kira, Z. How to train your DRAGAN. *arXiv preprint arXiv:1705.07215*, 2017.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. 2009.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, Alejandro, Aitken, A., Tejani, A., Totz, J., Wang, Z., and Shi, W. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 105–114, July 2017. doi: 10.1109/CVPR.2017.19.

Li, Y., Swersky, K., and Zemel, R. Generative moment matching networks. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

Liu, S., Bousquet, O., and Chaudhuri, K. Approximation and convergence properties of generative adversarial learning. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pp. 5545–5553, 2017.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

Mescheder, L., Nowozin, S., and Geiger, A. The numerics of GANs. In *Advances in Neural Information Processing Systems 30 (NIPS)*, 2017.

Mescheder, L., Geiger, A., and Nowozin, S. Which training methods for gans do actually converge. *arXiv preprint arXiv:1801.04406v2*, 2018.

Metz, L., Poole, B., Pfau, D., and Sohl-Dickstein, J. Unrolled generative adversarial networks. In *International Conference of Learning Representations (ICLR)*, 2017.

Milgrom, P. and Segal, I. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601, 2002.

Mroueh, Y., Li, C.-L., Sercu, T., Raj, A., and Cheng, Y. Sobolev GAN. In *International Conference on Learning Representations (ICLR)*, 2018.

Nagarajan, V. and Kolter, J. Z. Gradient descent GAN optimization is locally stable. In *Advances in Neural Information Processing Systems 30 (NIPS)*, 2017.

Nowozin, S., Cseke, B., and Tomioka, R. F-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems 29 (NIPS)*, 2016.

Petzka, H., Fischer, A., and Lukovnicov, D. On the regularization of wasserstein GANs. In *International Conference on Learning Representations (ICLR)*, 2018.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference of Learning Representations (ICLR)*, 2016.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems 29 (NIPS)*, pp. 2234–2242, 2016.

Schmidhuber, J. Learning factorial codes by predictability minimization. *Neural Computation*, 4(6):863–879, 1992.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561, 2010.

Srivastava, A., Valkov, L., Russell, C., Gutmann, M., and Sutton, C. VEEGAN: Reducing mode collapse in GANs using implicit variational learning. In *Advances in Neural Information Processing Systems 30 (NIPS)*, 2017.

Unterthiner, T., Nessler, B., Seward, C., Klambauer, G., Heusel, M., Ramsauer, H., and Hochreiter, S. Coulomb GANs: Provably optimal nash equilibria via potential fields. *International Conference of Learning Representations (ICLR)*, 2018.

Yu, F., Zhang, Y., Song, S., Seff, A., and Xiao, J. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.