# Towards a Deeper Understanding of Adversarial Losses

**Anonymous authors**
**Paper under double-blind review**

## Abstract

Recent years have seen adversarial losses been applied to many fields. Their applications extend beyond the originally proposed generative modeling to conditional generative and discriminative settings. While prior work has proposed various activation functions and regularization approaches, some open questions still remain unanswered. In this paper, we aim to study the following two research questions: 1) What types of activation functions form a well-behaved adversarial loss? 2) How different combinations of activation functions and regularization approaches perform empirically against one another? To answer the first question, we adopt the perspective of variational divergence minimization and consider an adversarial loss well-behaved if it behaves as a divergence-like measure between the data and model distributions. Using a generalized formulation for adversarial losses, we derive the necessary and sufficient conditions of a well-behaved adversarial loss. Our analysis reveals a large class of theoretically valid adversarial losses. For the second question, we propose a simple comparative framework for adversarial losses using discriminative adversarial networks. The proposed framework allows us to efficiently evaluate adversarial losses using a standard evaluation metric such as the classification accuracy. With the proposed framework, we evaluate a comprehensive set of 168 combinations of twelve activation functions and fourteen regularization approaches on the handwritten digit classification problem to decouple their effects. Our empirical findings suggest that there is no single winning combination of activation functions and regularization approaches across all settings. Our theoretical and empirical results may together serve as a reference for choosing or designing adversarial losses in future research.

## 1 Introduction

Generative adversarial networks (GANs) (Goodfellow et al., 2014) are a class of unsupervised machine learning algorithms. The core of a GAN is an *adversarial loss* that learns to measure the discrepancy between the real data distribution and the distribution of the generated samples. While adversarial losses are originally used to train a generative model (Goodfellow et al., 2014), their applications extend beyond unsupervised generative modeling to other supervised tasks such as conditional generative modeling (Mirza & Osindero, 2014) and discriminative modeling (dos Santos et al., 2017). Moreover, prior work has proposed a range of adversarial losses in different forms, motivated and derived from various existing divergences or distances. For example, the classic minimax GANs (Goodfellow et al., 2014) is motivated from the Jensen–Shannon divergence, Wasserstein GANs (Arjovsky et al., 2017) from the Wasserstein distance, least squares GANs (Mao et al., 2017) from the Pearson $\chi^2$ divergence, and $f$-GANs (Nowozin et al., 2016) from the $f$-divergence. Notably, these GAN formulations adopt different activation functions in their training objectives. However, it remains unclear what types of activation functions form a well-behaved adversarial loss. Moreover, apart from the activation functions, the adversarial losses proposed in prior work differ also in the regularization approach adopted. Some representative examples include the gradient penalties (Gulrajani et al., 2017; Kodali et al., 2017; Mescheder et al., 2018) and the spectral normalization (Miyato et al., 2018). However, it remains unclear how the activation functions and the regularization approach contribute respectively to the empirical performance of an adversarial loss.

In this paper, we aim to study the following two research questions:

**Question 1.** *What types of activation functions form a well-behaved adversarial loss?*

**Question 2.** *How different combinations of activation functions and regularization approaches perform empirically against one another?*

Our contributions to Question 1 are based on the *variational divergence minimization* perspective proposed by Nowozin et al. (2016) and the concept of *adversarial divergence* proposed by Liu et al. (2017). Specifically, we consider an adversarial loss *well-behaved* if it behaves as a divergence-like measure between the data and model distributions.[1] Using a generalized formulation proposed by Jolicoeur-Martineau (2019), we derive the necessary and sufficient conditions for an adversarial loss to be well-behaved. Our theoretical analysis reveals a large class of theoretically valid adversarial losses. Based on our findings, we propose two new adversarial losses—the *absolute* loss and the *asymmetric* loss. The absolute loss features a mixture of characteristics from two existing adversarial losses, while the asymmetric loss possesses distinct characteristics from other adversarial losses. While both losses do not originate from any underlying divergence, we will show in our empirical study that they can achieve similar performance against other adversarial losses.

For Question 2, some attempts have been made towards decoupling the effects of activation functions and regularization approaches. For example, Fedus et al. (2018) showed that the original nonsaturating GAN (Goodfellow et al., 2014) can achieve comparable performance as the WGAN-GP model (Gulrajani et al., 2017) when using the same gradient penalty. However, due to the high computational cost for conducting a large-scale comparative study, Fedus et al. (2018); Lucic et al. (2018); Kurach et al. (2019) evaluate only five, seven and ten, respectively, combinations of activation functions and regularization approaches. This necessitates the needs for an efficient way to compare adversarial losses. In view of the various tasks that adversarial losses are applied to, we propose to adopt the discriminative adversarial networks for our comparative study as we can efficiently evaluate different adversarial losses using a standard evaluation metric such as the classification accuracy. For simplicity and computation cost concerns, we consider the MNIST handwritten digit classification problem (LeCun et al., 1998). With the proposed framework, we systematically evaluate 168 adversarial losses featuring the combinations of twelve activation functions and fourteen regularization approaches. Moreover, we empirically study the effects of the Lipschitz constant (Arjovsky et al., 2017), penalty weights (Mescheder et al., 2018), momentum terms (Kingma & Ba, 2014). We also study the robustness of different regularization approaches to data imbalance. Our findings suggest that *there is no single winning combination of activation functions and regularization approaches across all settings*. Although it is unclear whether the empirical results we obtain for DANs can generalize to unconditional or conditional GANs,[2] we argue that the idea of adversarial losses is fairly generic and they can be used to train either generative or discriminative models. Hence, our empirical results still contribute to the goal towards a deeper understanding of adversarial losses.

## 2 Related Work

Prior work has studied the working of adversarial losses from different perspectives. Some view adversarial losses from the perspective of variational divergence minimization (Nowozin et al., 2016). From this view, Nowozin et al. (2016) derived the $f$-GAN family based on the $f$-divergences. However, the activation functions used for the data and model distribution need to be the conjugate function of each other in their formulations. In our theoretical analysis, we will relax this restriction and work on a more generalized formulation of adversarial losses. Further, Liu et al. (2017) proposed a broader concept of adversarial divergence and showed that if the objective function is an adversarial divergence with some conditions, then using a restricted discriminator family has a moment-matching effect. However, they did not show what activation functions lead to a valid adversarial divergence. In this paper, we will derive the necessary and sufficient conditions for a well-behaved adversarial loss.

Some view adversarial losses from the perspective of integral probability metric view (Arjovsky & Bottou, 2017; Mroueh et al., 2017; Li et al., 2017; Mroueh & Sercu, 2017). In particular, Arjovsky & Bottou (2017) showed that by imposing a Lipschitz constraint to the discriminator, the training objective for the generator

[1] The model distribution is the distribution of the generated samples. It is implicitly defined by the generator (see Section 3.1).

[2] We note that even though the generator is no longer a generative model in a DAN, the discriminator is still trained to minimize the same loss function as that in a conditional GAN (see Section 3.6).

corresponds to minimizing the Wasserstein distance between the data and model distributions. In this formulation, the activation functions used for the data and model distribution need to be the same. Later, some follow-up work proposed to enforce the Lipschitz constraint using gradient penalties (Gulrajani et al., 2017; Kodali et al., 2017; Wei et al., 2018; Mescheder et al., 2018) and spectral normalization (Miyato et al., 2018). Moreover, Fedus et al. (2018) suggested that these regularization approaches extend beyond the integral probability metric view and can be applied to other GAN formulations. Motivated by this observation, we will examine in our empirical study combining different activation functions and regularization approaches to decouple their effects. Moreover, they suggested that the training dynamic of a GAN is closer to approaching a Nash equilibrium of the adversarial game than minimizing a specific divergence. Nonetheless, we can view an adversarial loss as a dynamic, divergence-like measure that quantifies the discrepancy between the data and model distributions. This view will serve as the foundation of our theoretical analysis.

Another relevant line of research are the comparative studies of adversarial losses. Fedus et al. (2018) showed that both the gradient penalty (Gulrajani et al., 2017) and spectral normalization (Miyato et al., 2018) can also improve the performance for a nonsaturating GAN (Goodfellow et al., 2014) in addition to the Waasserstein GAN (Arjovsky et al., 2017) they are orginally proposed for. Later, Lucic et al. (2018) conducted a large-scale empirical study and showed that the seven GAN models considered in their study can reach similar scores given enough computation budgets. Similarly, Kurach et al. (2019) showed that the spectral normalization (Miyato et al., 2018) consistently improves the sample quality for the four activation functions considered, while the gradient penalty (Gulrajani et al., 2017) require a large computation budget to find a good hyperparameter setting. However, due to the high computational cost for conducting a large-scale experiment, Fedus et al. (2018); Lucic et al. (2018); Kurach et al. (2019) examine only five, seven and ten, respectively, combinations of activation functions and regularization approaches. In this paper, we will propose a simple comparative framework for efficiently comparing adversarial losses based on a discriminative adversarial network. With the proposed framework, we evaluate a comprehensive set of 168 combinations of twelve activation functions and fourteen regularization approaches.

## 3 Background

In this section, we provide key background knowledge useful for our discussions in the rest of the paper.

### 3.1 Generative adversarial networks

A generative adversarial network (GAN) (Goodfellow et al., 2014) is a generative latent variable model that aims to learn a generative model $G : \mathcal{Z} \mapsto \mathcal{X}$ that maps some latent space $\mathcal{Z}$ to the data space $\mathcal{X}$. A discriminative model $D : \mathcal{X} \mapsto \mathbb{R}$ defined on $\mathcal{X}$ is trained alongside $G$ to provide guidance for $G$. Now, let $P_d$ denote the data distribution and $P_z$ the latent distribution. The original GAN is formulated as a two-player minimax game between the generator $G$ and discriminator $D$ as follows.

$$\min_G \max_D \mathbb{E}_{\boldsymbol{x}\sim P_d}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim P_z}[\log(1 - D(G(\boldsymbol{z})))] \,. \tag{1}$$

Alternatively, let $P_g$ be the *model distribution* implicitly defined by $G(\boldsymbol{z})$ with $\boldsymbol{z} \sim P_z$.

$$\min_G \max_D \mathbb{E}_{\boldsymbol{x}\sim P_d}[\log D(\boldsymbol{x})] + \mathbb{E}_{\tilde{\boldsymbol{x}}\sim P_g}[\log(1 - D(\tilde{\boldsymbol{x}}))] \,. \tag{2}$$

We will refer to this original formulation to the classic minimax GAN for the rest of this paper.

### 3.2 A variational divergence minimization view of adversarial losses

Some prior work investigated adversarial losses from the perspective of variational divergence minimization. Nowozin et al. (2016) proposed to minimize the following variational lower bound of a $f$-divergence derived by Nguyen et al. (2010):

$$D_f(P_d, P_g) \geq \sup_{D\in\mathcal{D}} \mathbb{E}_{\boldsymbol{x}\sim P_d}[D(\boldsymbol{x})] - \mathbb{E}_{\tilde{\boldsymbol{x}}\sim P_g}[f^*(D(\tilde{\boldsymbol{x}}))] \,, \tag{3}$$

Table 1: Distribution $p_{\hat{\boldsymbol{x}}}$ and function $R$ in Equation (7) for common gradient penalties, where $k \in \mathbb{R}$ is the Lipschitz constant and $c \in \mathbb{R}$ is a hyperparameter.

| | $p_{\hat{\boldsymbol{x}}}$ | $R(r)$ |
|---|---|---|
| Two-side coupled gradient penalty (Gulrajani et al., 2017) | $P_d + U[0,1]\,(P_g - P_d)$ | $(r-k)^2$ |
| One-side coupled gradient penalty (Gulrajani et al., 2017) | $P_d + U[0,1]\,(P_g - P_d)$ | $\max(r,k)$ |
| Two-side local gradient penalty (Kodali et al., 2017) | $P_d + c\,N(0,I)$ | $(r-k)^2$ |
| One-side local gradient penalty (Kodali et al., 2017) | $P_d + c\,N(0,I)$ | $\max(r,k)$ |
| $\mathrm{R}_1$ gradient penalty (Mescheder et al., 2018) | $P_d$ | $r$ |
| $\mathrm{R}_2$ gradient penalty (Mescheder et al., 2018) | $P_g$ | $r$ |

where $\mathcal{D}$ is an arbitrary set of functions $D : \mathcal{X} \mapsto \mathbb{R}$ and $f^*$ is the corresponding conjugate function for an $f$-divergence. Nowozin et al. (2016) showed that this can be formulated as an adversarial minimax game similar to a GAN:

$$\min_G \max_D \mathbb{E}_{\boldsymbol{x}\sim P_d}[g_f(D(\boldsymbol{x}))] + \mathbb{E}_{\tilde{\boldsymbol{x}}\sim P_g}[-f^*(g_f(D(\tilde{\boldsymbol{x}}))]\,, \tag{4}$$

where $g_f : \mathbb{R} \mapsto dom(f^*)$ an output activation function specific to the $f$-divergence used. The $f$-GAN family given by Equation (4) includes some representative GAN formulations, including the classic minimax GANs (Goodfellow et al., 2014) and least-square GANs (Mao et al., 2017).

### 3.3 An integral probability metric view of adversarial losses

Some prior work has investigated viewing an adversarial loss as an estimator for an integral probability metric (IPM) (Müller, 1997). Arjovsky et al. (2017) considered the Wasserstein distance and proposed to minimize the following dual form given by the Kantorovich–Rubinstein duality:

$$W(P_d, P_g) = \sup_{||f||_L \leq 1} \mathbb{E}_{\boldsymbol{x}\sim P_d}[f(\boldsymbol{x})] - \mathbb{E}_{\tilde{\boldsymbol{x}}\sim P_g}[f(\tilde{\boldsymbol{x}})]\,, \tag{5}$$

Arjovsky et al. (2017) then proposed the following Wasserstein GAN formulation:

$$\min_G \max_{D\in\mathcal{D}} \mathbb{E}_{\boldsymbol{x}\sim P_d}[f(\boldsymbol{x})] - \mathbb{E}_{\tilde{\boldsymbol{x}}\sim P_g}[f(\tilde{\boldsymbol{x}})]\,, \tag{6}$$

where $\mathcal{D}$ is the set of all 1-Lipschitz functions. Other examples of IPM-based GANs include McGANs (Mroueh et al., 2017), MMD GANs (Li et al., 2017) and Fisher GANs (Mroueh & Sercu, 2017).

### 3.4 Gradient penalties

Originally proposed in (Gulrajani et al., 2017), a gradient penalty is a regularization approach to impose a soft Lipschitz constraint on the discriminator in a Wasserstein GAN (Arjovsky et al., 2017). Prior work has proposed different variants of gradient penalties (Kodali et al., 2017; Mescheder et al., 2018). In general, they take the following generalized form as a regularization term added to the discriminator loss function:

$$\lambda\, \mathbb{E}_{\hat{\boldsymbol{x}}\sim p_{\hat{\boldsymbol{x}}}}[R(||\nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}})||)]\,, \tag{7}$$

where $\lambda \in \mathbb{R}$ is a hyperparameter, which we will refer to as the *penalty weight*, and $R : \mathbb{R} \mapsto \mathbb{R}$ is a real function. The distribution $p_{\hat{\boldsymbol{x}}}$ defines where the gradient penalty is enforced. This formulation covers some representative gradient penalties proposed in prior work, including the coupled gradient penalty used in the WGAN-GP model (Gulrajani et al., 2017), the local gradient penalty used in the DRAGAN model (Kodali et al., 2017), and the $R_1$ and $R_2$ gradient penalties (Mescheder et al., 2018), as summarized in Table 1. In addition, Figure 1 illustrates the support of $p_{\hat{\boldsymbol{x}}}$, i.e., where gradient penalty is imposed.

(a) Coupled gradient penalty (b) Local gradient penalty (c) $R_1$ gradient penalty (d) $R_2$ gradient penalty

Figure 1: Illustrations of the support of $p_{\tilde{\boldsymbol{x}}}$, i.e., the region where the gradient penalty is imposed, for the gradient penalties considered in this paper, shown as the red shaded area in (a) and (b) and the red curves in (c) and (d). The blue and black curves denote the model and data manifolds, respectively. The right part of (a) shows the case when the generator perfectly fabricates the data distribution, i.e., $P_g = P_d$. For (c) and (d), the gradient penalty is enforced directly on the model and data manifolds, respectively.

### 3.5 Spectral normalization

Another regularization approach for Wasserstein GANs (Arjovsky et al., 2017) is the spectral normalization proposed by Miyato et al. (2018). By normalizing the spectral norm of the weight matrix for each layer of the discriminator, the spectral normalization imposes a hard Lipschitz constraint on the discriminator. As compared to a gradient penalty, which imposes a local regularization as illustrated in Figure 1, the spectral normalization imposes a global regularization for the discriminator.

### 3.6 Discriminative adversarial networks

A discriminative adversarial network (dos Santos et al., 2017) is essentially a conditional GAN (Mirza & Osindero, 2014) where both the generator and discriminator are discriminative models. In a DAN, the generator aims to predict the label of a real data sample, whereas the discriminator takes as input either a real pair—"(real data, real label)" or a fake pair—"(real data, fake label)", and aims to examine its authenticity. Mathematically, the generator in a DAN learns the mapping from the data space to the label space, i.e., $\mathcal{X} \mapsto \mathcal{Y}$, where $\mathcal{Y}$ denotes the label space. This is different from a conditional GAN, where the generator learns a mapping from the label space to the data space, i.e., $\mathcal{Y} \mapsto \mathcal{X}$. However, the discriminator in either a DAN or a conditional GAN learns the same mapping $\mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$.

## 4 Theoretical Results

In this section, we present our theoretical results on Question 1: What types of activation functions form a well-behaved adversarial loss? For the rest of the paper, we will adopt the following notations: $D$ denotes the discriminator and $G$ denotes the generator in a GAN. $P_d$ denotes the data distribution defined on the data space $\mathcal{X}$. $P_z$ denotes the latent distribution defined on the latent space $\mathcal{Z}$. $P_g$ denotes the *model distribution* implicitly defined by $G(\boldsymbol{z})$ with $\boldsymbol{z} \sim p_z$.

### 4.1 Generalized formulation of adversarial losses

Following Jolicoeur-Martineau (2019), we adopt a generalized form of adversarial losses as follows:

$$\max_D \mathbb{E}_{\boldsymbol{x}\sim P_d}[f(D(\boldsymbol{x}))] + \mathbb{E}_{\tilde{\boldsymbol{x}}\sim P_g}[g(D(\tilde{\boldsymbol{x}}))]\,, \tag{8}$$

$$\min_G \mathbb{E}_{\tilde{\boldsymbol{x}}\sim P_g}[h(D(\tilde{\boldsymbol{x}}))]\,, \tag{9}$$

where $f$, $g$ and $h$ are all real functions, i.e., $\mathbb{R} \mapsto \mathbb{R}$. We will refer to $f$, $g$ and $h$ as the *activation functions*. This generalized formulation covers a wide range of adversarial losses proposed in prior work, including the classic minimax and nonsaturating GANs (Goodfellow et al., 2014), Wasserstein GANs (Arjovsky et al., 2017), least squares GANs (Mao et al., 2017) and hinge GANs (Lim & Ye, 2017; Tran et al., 2017), as summarized in Table 2.[3]

[3]Notably, this formulation does not cover the relativistic GANs proposed by Jolicoeur-Martineau (2019), which we will include in our empirical comparative study in Section 6.

Table 2: Activation functions for the adversarial losses discussed in this paper (see Equation (8) and Equation (9) for the definitions of $f$, $g$ and $h$). In the rightmost column, $y^*$ denotes the common root of $f(y) = g(y)$ and $f'(y) = -g'(y) \neq 0$ (see Section 4.4 for more details).

| | $f$ | $g$ | $h$ | $y^*$ |
|---|---|---|---|---|
| Classic minimax (2014) | $-\log(1+e^{-y})$ | $-y-\log(1+e^{-y})$ | $-y-\log(1+e^{-y})$ | 0 |
| Classic nonsaturating (2014) | $-\log(1+e^{-y})$ | $-y-\log(1+e^{-y})$ | $\log(1+e^{-y})$ | 0 |
| Wasserstein (2017) | $y$ | $-y$ | $-y$ | 0 |
| Least squares (2017) | $-(y-1)^2$ | $-y^2$ | $(y-1)^2$ | $\frac{1}{2}$ |
| Hinge (2017; 2017) | $\min(0, y-1)$ | $\min(0, -y-1)$ | $-y$ | 0 |
| Absolute | $-\lvert 1-y\rvert$ | $-\lvert y\rvert$ | $\lvert 1-y\rvert$ | $\frac{1}{2}$ |
| Asymmetric | $-\lvert y\rvert$ | $-y$ | $-y$ | 0 |

In the following theoretical analysis, we will focus only on the activation functions $f$ and $g$. Specifically, we consider the minimax formulation where $h = g$:

$$\min_G \max_D \mathbb{E}_{\boldsymbol{x}\sim P_d}[f(D(\boldsymbol{x}))] + \mathbb{E}_{\tilde{\boldsymbol{x}}\sim P_g}[g(D(\tilde{\boldsymbol{x}}))]\,. \tag{10}$$

We note that from the perspective of either variational divergence minimization or integral probability metric, the generator should minimize the divergence-like measure estimated by the discriminator, and accordingly we have $h = g$. However, some prior work has investigated setting $h$ different from $g$. For example, the classic nonsaturating loss proposed by Goodfellow et al. (2014) has a different $h$ with opposite concavity as compared to the classic minimax loss (see Table 2). While our theoretical analysis concerns with only $f$ and $g$, we will empirically examine the effects of the activation function $h$ in Section 6.2.

### 4.2 Desirable properties for adversarial losses

Following the perspective of variational divergence minimization, we can see that if the discriminator is able to reach optimality, the training objective for the generator is

$$L_G = \max_D \mathbb{E}_{\boldsymbol{x}\sim P_d}[f(D(\boldsymbol{x}))] + \mathbb{E}_{\tilde{\boldsymbol{x}}\sim P_g}[g(D(\tilde{\boldsymbol{x}}))]\,. \tag{11}$$

In principle, the discriminator in this formulation is responsible for providing a measure of the discrepancy between $P_d$ and $P_g$, which will then serve as the training objective for the generator to push $P_g$ towards $P_d$. Hence, we would like such an adversarial loss to be a divergence-like measure between $P_g$ and $P_d$. Following this view of adversarial divergence (Liu et al., 2017), we can define the following two properties of a well-behaved adversarial losses.

**Property 1** (Weak desired property of a adversarial loss)**.** *For any data distribution $P_d$, the generator loss $L_G$ has a global minimum at $P_g = P_d$ with respect to the model distribution $P_g$.*

**Property 2** (Strong desired property of a adversarial loss)**.** *For any data distribution $P_d$, the generator loss $L_G$ has a unique global minimum at $P_g = P_d$ with respect to the model distribution $P_g$.*

We can see that Property 2 makes $L_G - L_G^*$ a divergence between $P_d$ and $P_g$ for any fixed $P_d$, where $L_G^* = L_G\big|_{P_g=P_d}$ is a constant term irrelevant to optimization. In contrast, Property 1 provides a weaker version when the uniqueness of the global minimum is not guaranteed.

### 4.3 The $\Psi$ and $\psi$ functions

In order to derive the necessary and sufficient conditions for Property 1 and Property 2, we first observe from Equation (11) that

$$L_G = \max_D \int_{\boldsymbol{x}} P_d(\boldsymbol{x})\, f(D(\boldsymbol{x})) + P_g(\boldsymbol{x})\, g(D(\boldsymbol{x}))\, d\boldsymbol{x}\,, \tag{12}$$

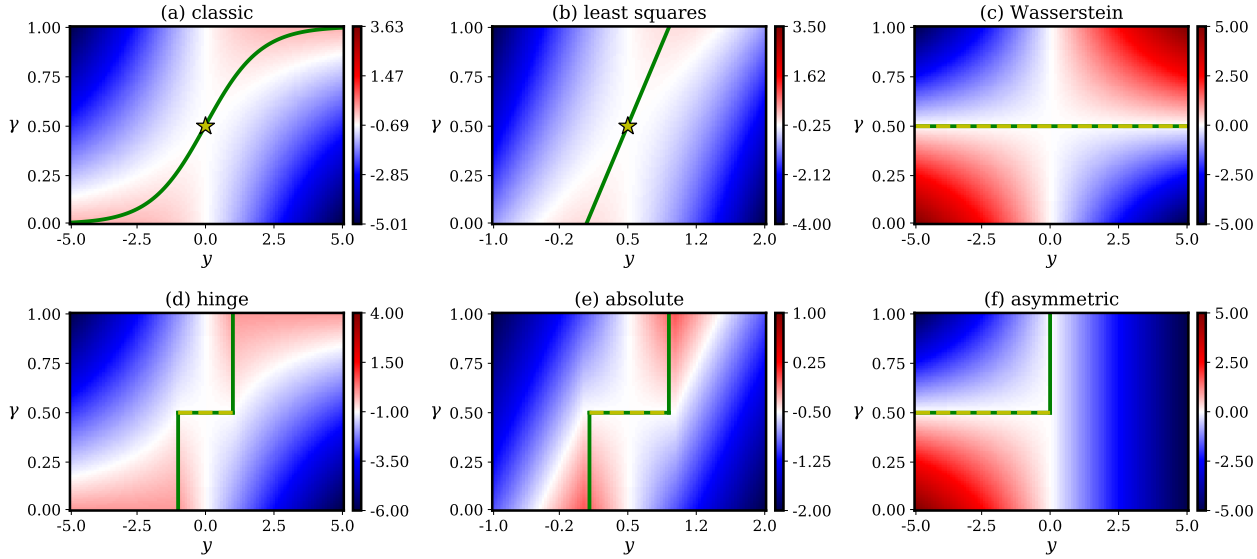



Figure 2: Graphs of the $\Psi$ functions for the adversarial losses discussed in this paper (see Equation (16) for the definition of $\Psi$). The green lines show the domains of $\psi$, i.e., the values that $y$ can take for different $\gamma$. The stars and the yellow dashed lines indicate the global minima of $\psi$. The midpoint of each color map is intentionally set to the minimum of $\psi$, i.e., the value taken at the star marks or the yellow segments. Note that as $y \in \mathbb{R}$, we plot different portions of $y$ where the characteristics of $\Psi$ can be best observed. (Best viewed in color.)

Assuming $D$ can be any function that maps from $\mathcal{X}$ to $\mathbb{R}$,[4] we can let $y = D(\boldsymbol{x})$ be a variable independent of $\boldsymbol{x}$. Then, we can exchange the order of the integral and maximum in Equation (12) and obtain

$$L_G = \int_{\boldsymbol{x}} \max_D \Big( P_d(\boldsymbol{x})\, f(D(\boldsymbol{x})) + P_g(\boldsymbol{x})\, g(D(\boldsymbol{x})) \Big)\, d\boldsymbol{x}\,, \tag{13}$$

Next, rearranging Equation (13) we obtain

$$L_G = \int_{\boldsymbol{x}} \big( P_d(\boldsymbol{x}) + P_g(\boldsymbol{x}) \big) \max_D \left( \frac{P_d(\boldsymbol{x})\, f(D(\boldsymbol{x}))}{P_d(\boldsymbol{x}) + P_g(\boldsymbol{x})} + \frac{P_g(\boldsymbol{x})\, g(D(\boldsymbol{x}))}{P_d(\boldsymbol{x}) + P_g(\boldsymbol{x})} \right) d\boldsymbol{x}\,. \tag{14}$$

Now, let $y = D(\boldsymbol{x})$ and $\gamma = \frac{P_d(\boldsymbol{x})}{P_d(\boldsymbol{x}) + P_g(\boldsymbol{x})}$ (note that $\gamma = \frac{1}{2}$ if and only if $P_d(\boldsymbol{x}) = P_g(\boldsymbol{x})$). We then obatin

$$L_G = \int_{\boldsymbol{x}} \big( P_d(\boldsymbol{x}) + P_g(\boldsymbol{x}) \big) \max_y \; \gamma\, f(y) + (1-\gamma)\, g(y)\, d\boldsymbol{x}\,. \tag{15}$$

Let us consider the latter term inside the integral and define the following two functions:

$$\Psi(\gamma, y) = \gamma\, f(y) + (1-\gamma)\, g(y)\,, \tag{16}$$

$$\psi(\gamma) = \max_y \; \Psi(\gamma, y)\,, \tag{17}$$

where $\gamma \in [0, 1]$ and $y \in \mathbb{R}$. We visualize in Figure 2 the $\Psi$ functions and Figure 3 the $\psi$ functions for the adversarial losses considered in this paper.

### 4.4 Necessary and sufficient conditions for a well-behaved adversarial loss

With the $\psi$ function defined by Equation (17), we can now derive the necessary and sufficient conditions for a well-behaved adversarial loss. All proofs can be found in Appendix A. First, Theorems 1 and 2 provide the necessary conditions for a well-behaved adversarial loss.

[4]This assumption does not hold if $D$ is parameterized as a neural network. See the discussions in Section 4.7.

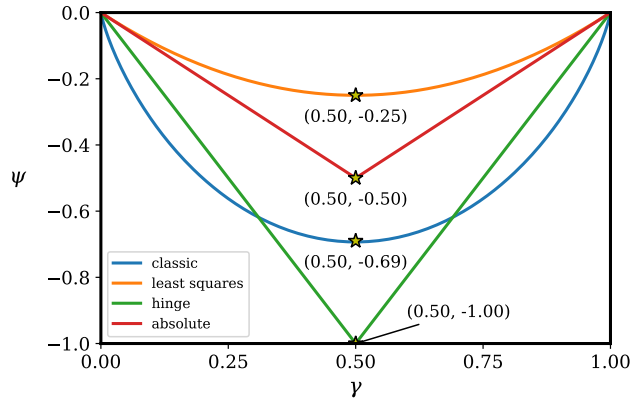



Figure 3: Graphs of the $\psi$ functions for the adversarial losses discussed in this paper (see Equation (17) for the definition of $\psi$). The stars indicate their global minima. For the Wasserstein loss, $\psi$ is only defined at $\gamma = 0.5$, where it takes a value of zero, and thus we do not include it in this plot.

**Theorem 1** (Necessary conditions of the weak desired property of an adversarial loss)**.** *If Property 1 holds, then for any* $\gamma \in [0, 1]$, $\psi(\gamma) + \psi(1-\gamma) \geq 2\,\psi(\frac{1}{2})$.

**Theorem 2** (Necessary conditions of the strong desired property of an adversarial loss)**.** *If Property 2 holds, then for any* $\gamma \in [0, \frac{1}{2}) \cup (\frac{1}{2}, 1]$, $\psi(\gamma) + \psi(1-\gamma) > 2\,\psi(\frac{1}{2})$.

For the sufficient conditions, we have Theorems 3 and 4 as follows.

**Theorem 3** (Sufficient conditions of the weak desired property of an adversarial loss)**.** *If* $\psi(\gamma)$ *has a global minimum at* $\gamma = \frac{1}{2}$, *then Property 1 holds.*

**Theorem 4** (Sufficient conditions of the strong desired property of an adversarial loss)**.** *If* $\psi(\gamma)$ *has a unique global minimum at* $\gamma = \frac{1}{2}$, *then Property 2 holds.*

With Theorems 3 and 4, we can check if a pair of activation functions $f$ and $g$ form a well-behaved adversarial loss by finding their corresponding $\psi$ function. Further, we have Theorem 5 for a more specific guideline for choosing or designing the activation functions $f$ and $g$.

**Theorem 5** (Alternative sufficient conditions of the strong desired property of an adversarial loss)**.** *If (1)* $f'' + g'' \leq 0$ *and (2) there exists some* $y^* \in \mathbb{R}$ *such that* $f(y^*) = g(y^*)$ *and* $f'(y^*) = -g'(y^*) \neq 0$, *then Property 2 holds.*

In other word, a pair of activation functions $f$ and $g$ form a well-behaved adversarial loss as long as (1) the sum of $f$ and $g$ is concave downward and (2) we can find a point where they take the same value while having exact opposite nonzero derivatives. Intuitively, the second part of the prerequisites reflects the adversarial nature of the game between the discriminator and generator as it requires their objectives are opposite at some $y^* \in \mathbb{R}$.

Theorem 5 reveals a large class of theoretically valid adversarial losses. While such a theoretical analysis has not been done on the generalized form of adversarial losses in Equations (8) and (9), it happens that all the adversarial losses listed in Table 2 have such desirable properties, and we show in Table 2 their corresponding $y^*$ in Theorem 5. Moreover, we note that Nowozin et al. (2016) also derived a large class of adversarial losses based on the $f$-divergence. However, in their formulations, $f$ and $g$ need to be the conjugate function of each other. Similarly, IPM-based GAN Arjovsky et al. (2017) require $f$ and $g$ to be the same. In contrast, the prerequisites of Theorem 5 is less restricted. For example, the hinge loss (Lim & Ye, 2017; Tran et al., 2017) does not fall into either the $f$-GAN or IPM-based GAN classes, while it satisfies the prerequisites of Theorem 5 and is thus a well-behaved adversarial loss based on our theoretical results. Theorem 5 suggests that the

### 4.5 Analyzing the adversarial game by the $\Psi$ functions

Apart from deriving the necessary and sufficient conditions for a well-behaved adversarial loss, the $\Psi$ and $\psi$ functions defined in Equations (16) and (17) also provide some new insights into the adversarial behaviors of the generator and discriminator. If we follow Equation (15) and consider $y = D(\boldsymbol{x})$ and $\gamma(\boldsymbol{x}) = \frac{P_d(\boldsymbol{x})}{P_d(\boldsymbol{x})+P_g(\boldsymbol{x})}$, then the discriminator can be viewed as maximizing $\Psi$ along the $y$-axis in Figure 2. On the other hand, since the generator is trained to push $P_g$ towards $P_d$, it can be viewed as minimizing $\Psi$ along the $\gamma$-axis. Following this view, the saddle-shaped landscape of the $\Psi$ functions presented in Figure 2 reflects the minimax nature of the adversarial game. Moreover, they all have saddle points at $\gamma = \frac{1}{2}$, which happens when $P_d(\boldsymbol{x}) = P_g(\boldsymbol{x})$. In the ideal case when the discriminator can be trained till optimality in each iteration, we will always stay on the domain of $\psi$, i.e., the green line. In this case, the generator can be viewed as minimizing $\Psi$ along the green line.[5] Finally, the neighborhoods around the saddle points of $\Psi$ for different adversarial losses possess distinct characteristics, and such difference may affect the game dynamics and optimization stability. Further investigation on the properties of $\Psi$ and $\psi$ may lead insights into why and how different activation functions perform differently in practice, which we will leave as future work.

### 4.6 Designing adversarial losses by the $\Psi$-landscape

From the analysis presented in the previous sections, we see that the $\Psi$ function reflects some important characteristics of an adversarial loss. By observing and designing the landscape of the $\Psi$ function, we propose the following two new adversarial losses that comply with Theorem 5.

- First, the *absolute loss* is obtained by replacing the square functions in the least squares loss with absolute functions. Specifically, we have $f(y) = -h(y) = -|1-y|$, $g(y) = -|y|$. As shown in Figure 2(e), its $\Psi$-landscape features a bounded positive-valued region (i.e., the red-colored region) of similar shape to that of the least squares loss. It also has a bounded global minima region (i.e., the yellow dashed lines) similar to that of the hinge loss.
- Second, the *asymmetric loss* is obtained by negating the activation function $f$ of a Wasserstein for $y \in \mathbb{R}^+$. Specifically, we have $f(y) = -|y|$, $g(y) = h(y) = -y$. As can be seen from Figure 2(f), it features an asymmetric, non-saddle shaped $\Psi$-landscape distinct from other adversarial losses. Its $\Psi$-landscape is similar to that of the Wasserstein loss, but the positive part of $y$ is all negative-valued.

Moreover, we note that both losses feature lower computation costs as their activation functions are all piecewise linear. We will include these two new losses in our empirical study in Section 6.2. While these two losses do not originate from any underlying divergence, we will show in Section 6.2 that they can achieve similar empirical performance against other adversarial losses, and that an adversarial loss does not necessarily need to have a saddle-shaped $\Psi$-landscape.

### 4.7 Limitations of the theoretical results

We want to discuss some of the limitations of our theoretical analysis presented in the previous sections. First, while our analysis originates from a variational divergence minimization perspective, the GAN training dynamic in practice has been shown to be closer to approaching a Nash equilibrium of the adversarial game than minimizing a specific divergence, as suggested by Fedus et al. (2018). While Theorem 5 reveals a large class of theoretically valid adversarial losses, their optimization and convergence properties remain unclear and require further investigations. Second, our derivation in Section 4.3 is based on the assumption that the discriminator $D$ can be any function. However, this does not hold in general. In practice, the discriminator is usually parameterized as a neural network, thus having limited expressiveness. Moreover, the regularization approaches adopted may further constraint the set of functions that $D$ can possibly be. Finally, we note that the sufficient conditions provided in Theorem 5 are not tight. To examine the tightness of these sufficient conditions, we will empirically examine in Section 6.6 some cases when the prerequisites of Theorem 5 do not hold.

[5] Note that $L_G$ is an integral over all possible $\boldsymbol{x} \in \mathcal{X}$, and thus the adversarial game is played in a high dimensional space.

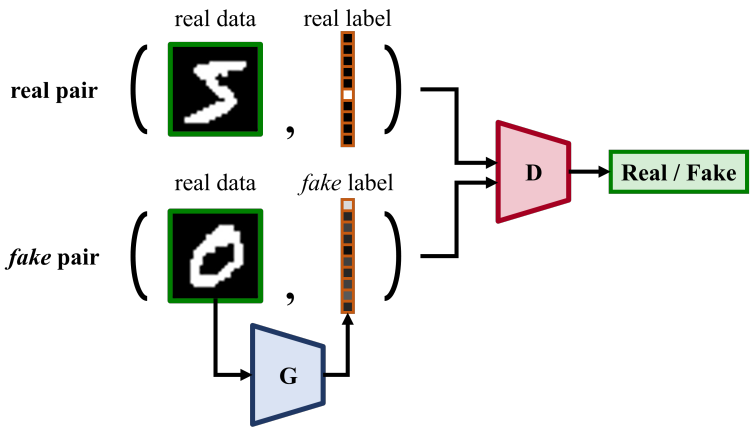



Figure 4: An example of a discriminative adversarial network (DAN) for MNIST digit classification.

## 5 A Comparative Framework for Adversarial Losses

In Section 4, we derive the necessary and sufficient conditions for a well-behaved adversarial loss. While the analysis reveals a large class of theoretical valid adversarial losses, it brings us to the next question—how does these different adversarial losses perform against one another? Although we focus only one the activation functions in our theoretical analysis, a pair of adversarial losses can differ not only in their activation functions but also the regularization approaches adopted. While Fedus et al. (2018); Lucic et al. (2018); Kurach et al. (2019) has attempted to decouple their effects, they evaluate only five, seven and ten, respectively, combinations of activation functions and regularization approaches in their experiments due to the high computational cost for conducting the comparative study. This necessitates the needs for an efficient way to empirically compare different adversarial losses. Motivated by the various domains that the adversarial losses have been applied to, we propose to compare adversarial losses through using discriminative adversarial networks (DANs).

In this paper, we will work on the MNIST handwritten digit classification problem (LeCun et al., 1998). As illustrated in Figure 4, the generator $G$ in a DAN takes a sample $\boldsymbol{x} \in \mathcal{X}$ as input and predict a fake label $c \in C$, where $C$ is the set of all possible classes. The discriminator takes a pair $(\boldsymbol{x}, c)$ ($\boldsymbol{x} \in \mathcal{X}$, $c \in C$) as input, where the label can either be the real label or a label predicted by the generator, and aims to examine if it is a real pair. Mathematically, the discriminator of a DAN learns the same mapping $\mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ as that of a conditional GAN does. The generator, a classification model in this example, is then trained by the supervision signal provided by the discriminator.

Although we work on a classification task in this paper, the proposed framework is generic to the underlying task as long as the evaluation metrics for the task are well defined. Moreover, adopting a DAN in this framework offers two additional advantages over a GAN. First, it avoids the high computation cost for training a GAN as the generator in a DAN is a discriminative model rather than a generative model in a GAN. Second, we can use standard evaluation metrics to evaluate a DAN. Hence, its performance can be easier evaluated than a GAN, which is usually assessed via more complex metrics such as the Inception Score (Salimans et al., 2016) and Fréchet Inception Distance (Heusel et al., 2017).

## 6 Empirical Results

Based on the comparative framework proposed in Section 5, we present in this section our empirical results on Question 2: How different combinations of activation functions and regularization approaches perform empirically against one another? In addition, we use the proposed framework to examine the effects of penalty weights, Lipschitz constants and momentum terms in the optimizer. Moreover, we examine the tightness of the sufficient conditions provided by Theorem 5 and the robustness of different regularization approaches to data imbalance.

Table 3: Network architectures for the (a) generator $G$ and (b) discriminator $D$ used in our experiments.

(a) Generator

| Layer type | Filters/neurons | Kernel | Stride |
|---|---|---|---|
| Convolutional | 32 | $3 \times 3$ | $3 \times 3$ |
| Convolutional | 64 | $3 \times 3$ | $3 \times 3$ |
| Max pooling | - | $2 \times 2$ | $2 \times 2$ |
| Dense | 128 | - | - |
| Dense | 10 | - | - |

(b) Discriminator

| Layer type | Filters/neurons | Kernel | Stride |
|---|---|---|---|
| Convolutional | 32 | $3 \times 3$ | $3 \times 3$ |
| Convolutional | 64 | $3 \times 3$ | $3 \times 3$ |
| Max pooling | - | $2 \times 2$ | $2 \times 2$ |
| Dense | 128 | - | - |
| Dense | 1 | - | - |

### 6.1 Implementation details

We implement the comparative framework described in Section 5 on the MNIST handwritten digit classification problem (LeCun et al., 1998). We implement both the generator $G$ and discriminator $D$ as convolutional neural networks (CNNs). In addition, we use the same network architectures across all experiments, as summarized in Table 3. Moreover, we concatenate the label vector to the input of each layer in the discriminator. To speed up and stabilize the training, we use the batch normalization (Ioffe & Szegedy, 2015) in the generator, and either the layer normalization (Ba et al., 2016) in the discriminator, which is disabled when spectral normalization (Miyato et al., 2018) is used. For all the gradient penalties, we use Euclidean norms and a penalty weight $\lambda = 10$ in Equation (7). In addition, we use a Lipschitz constant $k = 1$ for the coupled and local gradient penalties (Gulrajani et al., 2017; Kodali et al., 2017). We set the hyperparameter $c$ for the local gradient penalties (Kodali et al., 2017) to 0.01 (see Table 1). For the training, we alternatively update $G$ and $D$ once in each iteration and train the model for 100,000 generator steps. We use ReLUs (Glorot et al., 2011) as the activation functions for all layers except the last layer of the generator, which uses a softmax function, and the last layer of the discriminator, which has no activation function.[6]. We use the Adam optimizer (Kingma & Ba, 2014) with $\alpha = 0.001$, $\beta_1 = 0$ and $\beta_2 = 0.9$. We implement all the models in Python and TensorFlow (Abadi et al., 2016). All experiments are run on NVIDIA 1080 Tis with a batch size of 64. Finally, each experiment is repeated ten times, and we report the mean error rate and its 95% confidence interval. For reproducibility, we will release all the source code upon acceptance.

### 6.2 Decoupling the effects of activation functions and regularization approaches

In the first experiment, we aim to study how different activation functions and regularization approaches perform against one another. Moreover, in order to decouple their effects, we evaluate a comprehensive set of 168 combinations of twelve activation functions and fourteen regularization approaches. For the activation functions, we consider the classic minimax and nonsaturating losses (Goodfellow et al., 2014), Wasserstein loss (Arjovsky et al., 2017), least squares loss (Mao et al., 2017), hinge loss (Lim & Ye, 2017; Tran et al., 2017), relativistic average and relativistic average hinge losses (Jolicoeur-Martineau, 2019), as well as the absolute and asymmetric losses we propose in Section 4.5. Moreover, while our theoretical analysis focuses only on the activation functions $f$ and $g$, we also aim to examine the effects of the activation function $h$ for the generator. Specifically, we examine pairing the classic and hinge losses with the following three variants of activation function $h$: *minimax*—$h(y) = g(y)$, *nonsaturating*—$h(y) = \log(1 + e^{-y})$, and *linear*—$h(y) = -y$. For the regularization approaches, we consider the coupled gradient penalties (Gulrajani et al., 2017), local gradient penalties (Kodali et al., 2017), $\mathrm{R}_1$ and $\mathrm{R}_2$ gradient penalties (Mescheder et al., 2018) and the spectral normalization (Miyato et al., 2018). For the coupled and local gradient penalties, we examine both the two-side and one-side versions (see Table 1). In addition, we also examine the combinations of the spectral normalization with different gradient penalties.s

We report in Table 4 the results of the comparative study. Our empirical results suggest that *there is no single winning combination of activation functions and regularization approach across all settings*. The global

[6]We do not apply any activation function to the output layer of the discriminator because the activation function is instead implemented in the loss terms in our formulation (see Equations (8) and (9)).

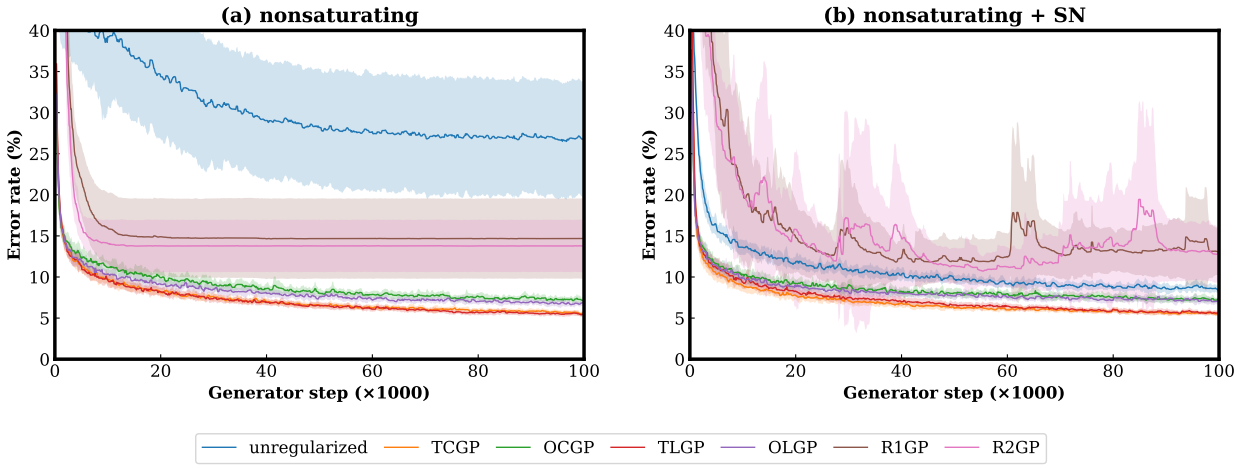



Figure 5: Error rates along the training progress for the classic nonsaturating loss with different regularization approaches. The models are evaluated every 100 steps, and the curves are smoothed by a median filter with a window size of 5. The shaded regions represent the standard deviations over ten runs. (Abbreviations: GP—gradient penalty, SN—spectral normalization, TCGP—two-side coupled gradient penalty, TLGP—two-side local gradient penalty; best viewed in color.)

lowest error rate is achieved by the combination of the relativistic hinge loss and two-side local gradient penalty. However, either relativisitic hinge loss or two-side local gradient penalty does not always achieve the lowest error rate when combined with other activation functions or regularization approaches. Morover, some combinations result in a notable high error rates, e.g., "classic minimax loss + $R_1$ gradient penalty," "least squares loss + $R_1$ gradient penalty," and "least squares loss + $R_2$ gradient penalty."

*With respect to the activation functions,* we find that the classic minimax and nonsaturating losses never achieve the lowest error rates with any regularization approach. In general, The hinge minimax and relativistic hinge losses achieve lower error rates than other losses, achieving the lowest error rates for three and four regularization approaches, respectively. The relativistic average loss outperforms both the classic minimax and nonsaturating losses across all regularization approaches, while the relativistic hinge loss does not always outperform the standard hinge losses. The proposed absolute and asymmetric losses are more robust to different regularization approaches, representing the only two losses that achieve a mean error rate under 12% with any regularization approach. Regarding the activation function $h$ for the generator, the three variants perform similarly and the differences do not reach statistical significance for most regularization approaches except the unregularized case. For the unregaularized case, the minimax version significantly outperforms the other two variants.

*With respect to the regularization approaches,* we find that the coupled and local gradient penalties outperform the $R_1$ and $R_2$ gradient penalties across nearly all activation functions, no matter whether the spectral normalization is used or not. The two-side local penalty achieves the lowest error rate for six activation functions. Combining either the coupled or local gradient penalty with the spectral normalization usually leads to higher error rates than using the gradient penalty only. The one-side gradient penalties lead to higher error rates than their two-side counterparts for all activation functions.

To further investigate how the training dynamic is affected by the regularization approach, we present in Figure 5 the training progress for the nonsaturating loss with different regularization approaches. We can see that the coupled and local gradient penalties result in a stable training dynamic, no matter whether spectral normalization is used or not. The $R_1$ and $R_2$ gradient penalties are too strong and halt the training early. This is possibly because they encourage the discriminator to have small gradients, and thus the gradients for both the generator and discriminator may vanish when the data and model distributions are close enough. In addition, combining either the $R_1$ or $R_2$ gradient penalty with the spectral normalization leads to an unstable training dynamic.

Table 4: Error rates (%) for different combinations of activation functions and regularization approaches. Mean values and 95% confidence intervals are reported. Underlined and bold fonts indicate the lowest error rates per column and per row, respectively. (Abbreviations: GP—gradient penalty, SN—spectral normalization, TCGP—two-side coupled gradient penalty, TLGP—two-side local gradient penalty, OCGP—one-side coupled gradient penalty, OLGP—one-side local gradient penalty.)

| | Unregularized | TCGP | TLGP | $R_1$ GP | $R_2$ GP |
|---|---|---|---|---|---|
| Classic minimax (2014) | 9.11 ± 0.39 | 5.65 ± 0.17 | **5.42 ± 0.11** | 19.01 ± 2.31 | 12.91 ± 0.70 |
| Classic nonsaturating (2014) | 26.83 ± 4.44 | 5.64 ± 0.14 | 5.56 ± 0.19 | 14.67 ± 3.01 | 13.80 ± 1.98 |
| Classic linear | 17.38 ± 3.20 | 5.66 ± 0.22 | 5.55 ± 0.10 | 18.49 ± 3.42 | 14.92 ± 3.22 |
| Hinge minimax | 5.57 ± 0.16 | **4.83 ± 0.21** | 4.88 ± 0.15 | 7.31 ± 0.92 | 9.49 ± 3.28 |
| Hinge nonsaturating | 37.55 ± 12.53 | 5.00 ± 0.15 | **4.97 ± 0.15** | 7.34 ± 1.13 | 7.54 ± 0.81 |
| Hinge linear (2017; 2017) | 11.50 ± 3.30 | 5.01 ± 0.16 | **4.89 ± 0.11** | 8.96 ± 2.20 | 7.71 ± 1.13 |
| Wasserstein (2017) | 7.69 ± 0.20 | 5.04 ± 0.12 | **4.92 ± 0.14** | 13.89 ± 12.79 | 7.25 ± 0.74 |
| Least squares (2017) | 7.15 ± 0.29 | 7.27 ± 0.27 | 6.70 ± 0.27 | 30.12 ± 17.62 | 32.44 ± 13.05 |
| Relativistic (2019) | 90.20 ± 0.00 | 5.25 ± 0.15 | **5.01 ± 0.19** | 8.00 ± 1.01 | 8.75 ± 3.61 |
| Relativistic hinge (2019) | 52.01 ± 5.81 | 8.28 ± 6.36 | **4.71 ± 0.07** | 8.39 ± 1.19 | 7.67 ± 1.13 |
| Absolute | 6.69 ± 0.15 | 5.23 ± 0.18 | 5.20 ± 0.16 | 8.01 ± 1.21 | 6.64 ± 0.32 |
| Asymmetric | 7.81 ± 0.17 | **4.77 ± 0.21** | 4.94 ± 0.09 | 8.79 ± 1.97 | 7.33 ± 0.63 |

| | SN | SN + TCGP | SN + TLGP | SN + $R_1$ GP | SN + $R_2$ GP |
|---|---|---|---|---|---|
| Classic minimax (2014) | 7.37 ± 0.32 | 5.55 ± 0.23 | 5.57 ± 0.17 | 11.16 ± 1.65 | 14.00 ± 1.54 |
| Classic nonsaturating (2014) | 8.25 ± 0.22 | **5.52 ± 0.10** | 5.61 ± 0.31 | 12.98 ± 1.68 | 13.50 ± 2.34 |
| Classic linear | 7.98 ± 0.22 | 5.70 ± 0.22 | **5.48 ± 0.18** | 15.45 ± 4.05 | 17.61 ± 4.71 |
| Hinge minimax | 6.22 ± 0.14 | 4.93 ± 0.12 | 5.06 ± 0.20 | 10.62 ± 1.30 | 12.91 ± 2.66 |
| Hinge nonsaturating | 6.90 ± 0.20 | 5.05 ± 0.14 | 5.06 ± 0.24 | 11.91 ± 2.49 | 12.10 ± 2.93 |
| Hinge linear (2017; 2017) | 6.59 ± 0.19 | 4.97 ± 0.12 | 5.18 ± 0.17 | 13.63 ± 2.56 | 11.35 ± 2.11 |
| Wasserstein (2017) | 5.89 ± 0.16 | 5.50 ± 0.11 | 5.76 ± 0.43 | 13.74 ± 3.39 | 13.82 ± 3.06 |
| Least squares (2017) | 7.88 ± 0.28 | **6.69 ± 0.15** | 7.11 ± 0.23 | 9.91 ± 0.96 | 11.56 ± 2.54 |
| Relativistic (2019) | 7.14 ± 0.24 | 5.35 ± 0.18 | 5.25 ± 0.16 | 9.31 ± 1.25 | 8.62 ± 0.37 |
| Relativistic hinge (2019) | 6.44 ± 0.10 | 5.02 ± 0.19 | 5.03 ± 0.13 | 12.56 ± 2.74 | 12.40 ± 2.82 |
| Absolute | 6.79 ± 0.28 | 5.23 ± 0.08 | **5.18 ± 0.22** | 10.42 ± 1.90 | 9.93 ± 1.41 |
| Asymmetric | 5.98 ± 0.25 | 5.60 ± 0.18 | 5.82 ± 0.27 | 8.46 ± 0.27 | 8.80 ± 0.73 |

| | OCGP | OLGP | SN + OCGP | SN + OLGP |
|---|---|---|---|---|
| Classic minimax (2014) | 7.15 ± 0.48 | 6.95 ± 0.32 | 7.16 ± 0.16 | 6.86 ± 0.18 |
| Classic nonsaturating (2014) | 7.20 ± 0.24 | 6.98 ± 0.14 | 7.47 ± 0.38 | 7.15 ± 0.22 |
| Classic linear | 7.12 ± 0.38 | 7.00 ± 0.62 | 7.29 ± 0.22 | 7.18 ± 0.33 |
| Hinge minimax | 5.82 ± 0.19 | 7.33 ± 0.84 | 5.80 ± 0.15 | 5.83 ± 0.12 |
| Hinge nonsaturating | 5.69 ± 0.19 | 7.88 ± 0.82 | 5.92 ± 0.22 | 5.74 ± 0.17 |
| Hinge linear (2017; 2017) | 5.77 ± 0.18 | 6.22 ± 0.64 | 5.77 ± 0.19 | 5.82 ± 0.12 |
| Wasserstein (2017) | 7.60 ± 1.87 | 13.34 ± 0.92 | 6.35 ± 0.27 | 6.06 ± 0.28 |
| Least squares (2017) | 7.99 ± 0.22 | 8.06 ± 0.30 | 8.43 ± 0.31 | 8.31 ± 0.32 |
| Relativistic (2019) | 8.03 ± 2.06 | 9.41 ± 1.80 | 6.18 ± 0.18 | 6.03 ± 0.15 |
| Relativistic hinge (2019) | 10.70 ± 1.56 | 14.17 ± 1.11 | 5.42 ± 0.20 | 5.42 ± 0.20 |
| Absolute | 5.95 ± 0.12 | 5.88 ± 0.25 | 6.22 ± 0.15 | 6.08 ± 0.20 |
| Asymmetric | 5.85 ± 0.22 | 7.57 ± 0.61 | 6.21 ± 0.21 | 5.92 ± 0.23 |

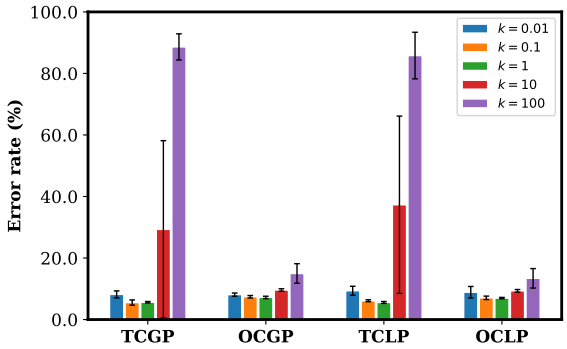
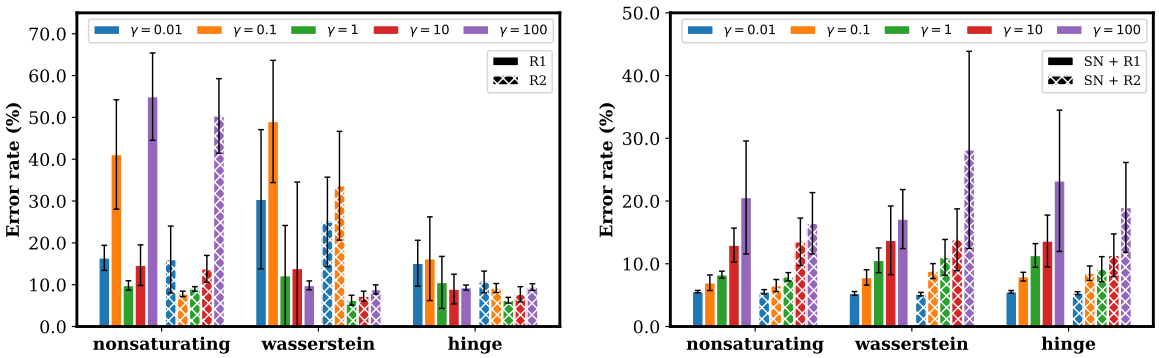



Figure 6: Error rates for different Lipschitz constants $k$ in the coupled and local gradient penalties, using the classic nonsaturating loss. (Abbreviations: TCGP—two-side coupled gradient penalty, OCGP—one-side coupled gradient penalty, TLGP—two-side local gradient penalty, OLGP—one-side local gradient penalty; Best viewed in color.)




(a) Without the spectral normalization

(b) With the spectral normalization

Figure 7: Error rates for different penalty weights $\lambda$ in the $R_1$ and $R_2$ gradient penalties. Mean values and standard deviations are reported. (Abbreviation: SN—spectral normalization; best viewed in color.)

### 6.3 Effects of the Lipschitz constants

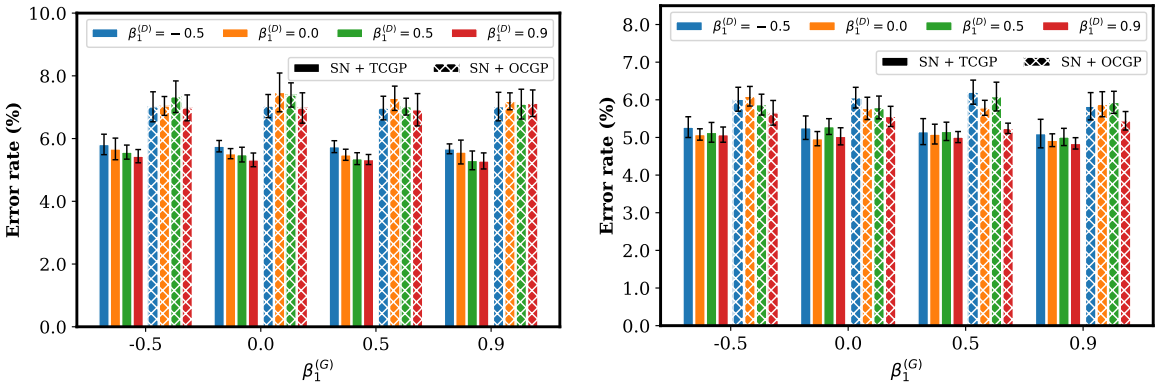
In this experiment, we examine the effects of the Lipschitz constant $k$ in the coupled and local gradient penalties (Gulrajani et al., 2017; Kodali et al., 2017). We report in Figure 6 the results for $k = 0.01$, 0.1, 1, 10, 100 using the classic nonsaturating loss. We can see that the error rate increases as $k$ goes away from 1.0, suggesting that $k = 1$ is a good default value. In addition, the two-side gradient penalties achieve lower error rates when $k$ is set properly, while they are more sensitive to $k$ than their one-side counterparts.

### 6.4 Effects of the penalty weights

We examine in this experiment the effects of the penalty weights $\lambda$ for the $R_1$ and $R_2$ gradient penalties (Mescheder et al., 2018). We show in Figure 7 the results for $\lambda = 0.01, 0.1, 1, 10, 100$ using the classic nonsaturating, Wasserstein and hinge losses. We can see that the $R_2$ gradient penalty generally outperform the $R_1$ gradient penalty. However, they are both sensitive to the value of $\lambda$. Moreover, when spectral normalization is used, the error rate increases as $\lambda$ increases, which implies that the $R_1$ and $R_2$ gradient penalties do not work well with the spectral normalization.

### 6.5 Effects of the momentum terms of the optimizers

We observe a trend in prior work towards using a smaller momentum (Radford et al., 2016), no momentum (Arjovsky et al., 2017; Gulrajani et al., 2017; Miyato et al., 2018; Brock et al., 2019) or a negative momentum




(a) Classic loss

(b) Hinge loss

Figure 8: Error rates for different momentum terms $\beta_1$ in the Adam optimizer. $\beta_1^{(G)}$ and $\beta_1^{(D)}$ denote the momentum terms for the generator and discriminator, respectively. Mean values and standard deviations are reported. (Abbreviations: SN—spectral normalization, TCGP—two-side coupled gradient penalty, OCGP—one-side coupled gradient penalty; best viewed in color.)

(Gidel et al., 2019) in GAN training. Hence, we want to examine in this experiment the effects of the momentum term $\beta_1$ in an Adam optimizer (Kingma & Ba, 2014) with our proposed framework. Further, since the generator and discriminator are trained by distinct objectives, we also investigate setting different momentum values for them. We present in Figure 8 the results for all combinations of $\beta_1 = -0.5, 0.0, 0.5, 0.9$ for the generator and discriminator, using the classic nonsaturating loss along with the spectral normalization and coupled gradient penalties for regularization. We can see that for the two-side coupled gradient penalty, using a larger momentum in both the generator and discriminator leads to lower error rates. However, there is no specific trend for the one-side coupled gradient penalty.

### 6.6 Examining the necessary conditions for desirable adversarial loss functions

As discussed in Section 4.4, the sufficient conditions provided by Theorem 5 are not tight. In order to examine the tightness of these sufficient conditions, we examine in this experiment the cases when the prerequisites in Theorem 5 are violated. Specifically, we change the training objective for the discriminator into the following *$\omega$-weighted loss*:

$$\max_{D} \; \mathbb{E}_{\boldsymbol{x}\sim P_d}[\omega f(D(\boldsymbol{x}))] + \mathbb{E}_{\tilde{\boldsymbol{x}}\sim P_g}[g(D(\tilde{\boldsymbol{x}}))] \,, \tag{18}$$

where $\omega \in \mathbb{R}^+ \setminus \{1\}$ is a constant. Note that the training objective for the generator remains the same. We can see that the prerequisites in Theorem 5 do not hold when $\omega \neq 1$.[7] We consider the classic nonsaturating, the Wasserstein and hinge losses, using the spectral normalization for regularization. We show in Table 5 the results for $\omega = 0.5, 0.9, 1.0, 1.1, 2.0$. We can see that the original losses result in the lowest error rates as compared to the $\omega$-weighted loss. While the Wasserstein loss fails with error rates over 50% when $\omega \neq 1$, the $\omega$-weighted versions of the nonsaturing and hinge losses achieve a higher yet acceptable error rate. Moreover, the error rates generally increase as $\omega$ goes away from 1.0. These results suggest that although the Theorem 5 reveals a large class of theoretical valid adversarial losses, the sufficient conditions it provide are not tight.

### 6.7 Limitations of the empirical results

While the proposed comparative framework allows us to efficiently compare adversarial losses without high computation costs, one of the main limitations of our empirical results is that it is unclear whether the empirical results we obtain for DANs can be generalized to conditional and unconditional GANs. Although the discriminators in a DAN and a conditional GAN are trained to optimize the same objective, learning the same mapping $\mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$, the generators are trained to learn the opposite mappings—the generator in

[7]See Appendix B for the graphs of the corresponding $\Psi$ and $\psi$ functions for $\omega = 0.5, 0.9, 1.0, 1.1, 2.0$.

Table 5: Error rates (%) for the nonsaturating, the Wasserstein and hinge losses along with their $\omega$-weighted versions (see Equation (18) for the definition). Mean values and 95% confidence intervals are reported.

| | $\omega = 0.5$ | $\omega = 0.9$ | Original | $\omega = 1.1$ | $\omega = 2.0$ |
|---|---|---|---|---|---|
| Nonsaturating | $8.47 \pm 0.22$ | $8.96 \pm 0.39$ | $\mathbf{8.25 \pm 0.22}$ | $8.62 \pm 0.28$ | $9.18 \pm 0.58$ |
| Wasserstein | $73.16 \pm 3.94$ | $57.66 \pm 3.18$ | $\mathbf{5.89 \pm 0.16}$ | $60.30 \pm 4.72$ | $69.54 \pm 3.33$ |
| Hinge | $15.20 \pm 1.52$ | $8.94 \pm 0.54$ | $\mathbf{6.59 \pm 0.19}$ | $8.02 \pm 0.22$ | $11.87 \pm 0.53$ |

a DAN learns the mapping $\mathcal{X} \mapsto \mathcal{Y}$, while the generator in a conditional GAN learns the mapping $\mathcal{Y} \mapsto \mathcal{X}$. Nonetheless, we argue that the idea of adversarial losses is rather generic and can be applied to various tasks, either generative or discriminative. Our empirical results still contribute to the goal towards a deeper understanding of adversarial losses.

## 7 Conclusion

In the first part of this paper, following the perspective of variational divergence minimization, we derived the necessary and sufficient conditions of a well-behaved adversarial loss in a generalized formulation. Our theoretical analysis reveals a large class of adversarial losses that includes a range of adversarial losses proposed in prior work. In the second part of this paper, we proposed a simple comparative framework for adversarial losses based on discriminative adversarial networks. With the proposed framework, we evaluated a comprehensive set of 168 adversarial losses featuring the combinations of twelve activation functions and fourteen regularization approaches. Our empirical findings suggest that there is no single winning combination of activation functions and regularization approaches across all settings. The theoretical and empirical results presented in this paper may together serve as a reference for choosing or designing adversarial losses in future research.

## References

Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 2016. (Cited on page 11.)

Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017. (Cited on page 2.)

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. (Cited on pages 1, 2, 3, 4, 5, 6, 8, 11, 13, and 14.)

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. (Cited on page 11.)

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019. (Cited on page 14.)

Cicero Nogueira dos Santos, Kahini Wadhawan, and Bowen Zhou. Learning loss functions for semi-supervised learning via discriminative adversarial networks. *NeurIPS Workshop on Learning with Limited Labeled Data*, 2017. (Cited on pages 1 and 5.)

William Fedus, Mihaela Rosca, Balaji Lakshminarayanan, Andrew M. Dai, Shakir Mohamed, and Ian Goodfellow. Many paths to equilibrium: GANs do not need to decrease a divergence at every step. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018. (Cited on pages 2, 3, 9, and 10.)

Gauthier Gidel, Reyhane Askari Hemmat, Mohammad Pezeshki, Remi Lepriol, Gabriel Huang, Simon Lacoste-Julien, and Ioannis Mitliagkas. Negative momentum for improved game dynamics. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019. (Cited on page 15.)

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 315–323, 2011. (Cited on page 11.)

Ian J. 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 27 (NeurIPS)*, 2014. (Cited on pages 1, 2, 3, 4, 5, 6, 11, and 13.)

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017. (Cited on pages 1, 2, 3, 4, 11, and 14.)

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017. (Cited on page 10.)

Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2015. (Cited on page 11.)

Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard GAN. *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019. (Cited on pages 2, 5, 11, and 13.)

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2014. (Cited on pages 2, 11, and 15.)

Naveen Kodali, Jacob Abernethy, James Hays, and Zsolt Kira. On convergence and stability of GANs. *arXiv preprint arXiv:1705.07215*, 2017. (Cited on pages 1, 3, 4, 11, and 14.)

Karol Kurach, Mario Lucic, Xiaohua Zhai, Marcin Michalski, and Sylvain Gelly. A large-scale study on regularization and normalization in gans. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2019. (Cited on pages 2, 3, and 10.)

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. (Cited on pages 2, 10, and 11.)

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. MMD GAN: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 2203–2213, 2017. (Cited on pages 2 and 4.)

Jae Hyun Lim and Jong Chul Ye. Geometric GAN. *arXiv preprint arXiv:1705.02894*, 2017. (Cited on pages 5, 6, 8, 11, and 13.)

Shuang Liu, Olivier Bousquet, and Kamalika Chaudhuri. Approximation and convergence properties of generative adversarial learning. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017. (Cited on pages 2 and 6.)

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are GANs created equal? a large-scale study. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2018. (Cited on pages 2, 3, and 10.)

Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. (Cited on pages 1, 4, 5, 6, 11, and 13.)

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for GANs do actually converge? In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018. (Cited on pages 1, 2, 3, 4, 11, and 14.)

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. (Cited on pages 1 and 5.)

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018. (Cited on pages 1, 3, 5, 11, and 14.)

Youssef Mroueh and Tom Sercu. Fisher GAN. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017. (Cited on pages 2 and 4.)

Youssef Mroueh, Tom Sercu, and Vaibhava Goel. McGan: Mean and covariance feature matching GAN. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. (Cited on pages 2 and 4.)

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2), 1997. (Cited on page 4.)

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), 2010. (Cited on page 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 29 (NeurIPS)*, 2016. (Cited on pages 1, 2, 3, 4, and 8.)

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *Proceedings of the 4th International Conference on Learning Representations (ICLR)*, 2016. (Cited on page 14.)

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems 29 (NeurIPS)*, 2016. (Cited on page 10.)

Dustin Tran, Rajesh Ranganath, and David M. Blei. Hierarchical implicit models and likelihood-free variational inference. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017. (Cited on pages 5, 6, 8, 11, and 13.)

Xiang Wei, Boqing Gong, Zixia Liu, Wei Lu, and Liqiang Wang. Improving the improved training of Wasserstein GANs: A consistency term and its dual effect. In *Proceedings of the 6th International Conference on Learning Representation (ICLR)*, 2018. (Cited on page 3.)

## A Proofs of Theorems

**Theorem 1** (Necessary conditions of the weak desired property of an adversarial loss)**.** *If Property 1 holds, then for any* $\gamma \in [0,1]$, $\psi(\gamma) + \psi(1-\gamma) \geq 2\,\psi(\frac{1}{2})$.

*Proof.* Since Property 1 holds, we have for any fixed $P_d$,

$$L_G \geq L_G \big|_{P_g = P_d} . \tag{19}$$

Let us consider

$$P_d(\boldsymbol{x}) = \gamma\,\delta(\boldsymbol{x}-\boldsymbol{s}) + (1-\gamma)\,\delta(\boldsymbol{x}-\boldsymbol{t})\,, \tag{20}$$
$$P_g(\boldsymbol{x}) = (1-\gamma)\,\delta(\boldsymbol{x}-\boldsymbol{s}) + \gamma\,\delta(\boldsymbol{x}-\boldsymbol{t})\,. \tag{21}$$

for some $\gamma \in [0,1]$ and $\boldsymbol{s}, \boldsymbol{t} \in \mathcal{X}, \boldsymbol{s} \neq \boldsymbol{t}$. Then, we have

$$\begin{aligned}
L_G \big|_{P_g = P_d} &= \max_D \int_{\boldsymbol{x}} P_d(\boldsymbol{x})\, f(D(\boldsymbol{x})) + P_d(\boldsymbol{x})\, g(D(\boldsymbol{x}))\, d\boldsymbol{x} && (22)\\
&= \max_D \int_{\boldsymbol{x}} P_d(\boldsymbol{x})\, (f(D(\boldsymbol{x})) + g(D(\boldsymbol{x})))\, d\boldsymbol{x} && (23)\\
&= \max_D \int_{\boldsymbol{x}} \big( (\gamma\,\delta(\boldsymbol{x}-\boldsymbol{s}) + (1-\gamma)\,\delta(\boldsymbol{x}-\boldsymbol{t}))(f(D(\boldsymbol{x})) + g(D(\boldsymbol{x}))) \big)\, d\boldsymbol{x} && (24)\\
&= \max_D \big( \gamma\,(f(D(\boldsymbol{s})) + g(D(\boldsymbol{s}))) + (1-\gamma)(f(D(\boldsymbol{t})) + g(D(\boldsymbol{t}))) \big) && (25)\\
&= \max_{y_1, y_2} \gamma\,(f(y_1) + g(y_1)) + (1-\gamma)(f(y_2) + g(y_2)) && (26)\\
&= \max_{y_1} \gamma\,(f(y_1) + g(y_1)) + \max_{y_2} (1-\gamma)(f(y_2) + g(y_2)) && (27)\\
&= \max_y f(y) + g(y) && (28)\\
&= 2\,\psi(\tfrac{1}{2})\,. && (29)
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
L_G &= \max_D \int_{\boldsymbol{x}} P_d(\boldsymbol{x})\, f(D(\boldsymbol{x})) + P_g(\boldsymbol{x})\, g(D(\boldsymbol{x}))\, d\boldsymbol{x} && (30)\\
&= \max_D \int_{\boldsymbol{x}} \big( (\gamma\,\delta(\boldsymbol{x}-\boldsymbol{s}) + (1-\gamma)\,\delta(\boldsymbol{x}-\boldsymbol{t}))\, f(D(\boldsymbol{x})) + ((1-\gamma)\,\delta(\boldsymbol{x}-\boldsymbol{s}) + \gamma\,\delta(\boldsymbol{x}-\boldsymbol{t}))\, g(D(\boldsymbol{x})) \big)\, d\boldsymbol{x} && (31)\\
&= \max_D \big( \gamma\, f(D(\boldsymbol{s})) + (1-\gamma)\, f(D(\boldsymbol{t})) + (1-\gamma)\, g(D(\boldsymbol{s})) + \gamma\, g(D(\boldsymbol{t})) \big) && (32)\\
&= \max_{y_1, y_2} \big( \gamma\, f(y_1) + (1-\gamma)\, g(y_1)) + (1-\gamma)\, f(y_2) + \gamma\, g(y_2) \big) && (33)\\
&= \max_{y_1} \gamma\, f(y_1) + (1-\gamma)\, g(y_1)) + \max_{y_2} (1-\gamma)\, f(y_2) + \gamma\, g(y_2) && (34)\\
&= \psi(\gamma) + \psi(1-\gamma)\,. && (35)
\end{aligned}$$

(Note that we can obtain Equation (26) from Equation (25) and Equation (33) from Equation (32) because $D$ can be any function and thus $D(\boldsymbol{s})$ is independent of $D(\boldsymbol{t})$.)

As Equation (19) holds for any fixed $P_d$, by substituting Equation (29) and Equation (34) into Equation (19), we get

$$\psi(\gamma) + \psi(1-\gamma) \geq 2\,\psi(\tfrac{1}{2}) \tag{36}$$

for any $\gamma \in [0,1]$, which concludes the proof. □

**Theorem 2** (Necessary conditions of the strong desired property of an adversarial loss)**.** *If Property 2 holds, then for any* $\gamma \in [0, \frac{1}{2}) \cup (\frac{1}{2}, 1]$, $\psi(\gamma) + \psi(1-\gamma) > 2\,\psi(\frac{1}{2})$.

*Proof.* Since Property 2 holds, we have for any fixed $P_d$,

$$L_G\big|_{P_g \neq P_d} > L_G\big|_{P_g = P_d}\,. \tag{37}$$

Following the proof of Theorem 1, consider

$$P_d(\boldsymbol{x}) = \gamma\,\delta(\boldsymbol{x}-\boldsymbol{s}) + (1-\gamma)\,\delta(\boldsymbol{x}-\boldsymbol{t})\,, \tag{38}$$

$$P_g(\boldsymbol{x}) = (1-\gamma)\,\delta(\boldsymbol{x}-\boldsymbol{s}) + \gamma\,\delta(\boldsymbol{x}-\boldsymbol{t})\,, \tag{39}$$

for some $\gamma \in [0,1]$ and some $\boldsymbol{s}, \boldsymbol{t} \in \mathcal{X}, \boldsymbol{s} \neq \boldsymbol{t}$. It can be easily shown that $P_g = P_d$ if and only if $\gamma = \frac{1}{2}$.

As Equation (37) holds for any fixed $P_d$, by substituting Equation (38) and Equation (39) into Equation (37), we get

$$\psi(\gamma) + \psi(1-\gamma) > 2\,\psi(\tfrac{1}{2})\,, \tag{40}$$

for any $\gamma \in [0, \frac{1}{2}) \cup (\frac{1}{2}, 1]$, concluding the proof. □

**Theorem 3** (Sufficient conditions of the weak desired property of an adversarial loss)**.** *If $\psi(\gamma)$ has a global minimum at $\gamma = \frac{1}{2}$, then Property 1 holds.*

*Proof.* First, we see that

$$\begin{aligned}
L_G\big|_{P_g=P_d} &= \max_D \int_{\boldsymbol{x}} P_d(\boldsymbol{x})\,f(D(\boldsymbol{x})) + P_d(\boldsymbol{x})\,g(D(\boldsymbol{x}))\,d\boldsymbol{x} && (41)\\
&= \max_y \int_{\boldsymbol{x}} P_d(\boldsymbol{x})\,f(y) + P_d(\boldsymbol{x})\,g(y)\,d\boldsymbol{x} && (42)\\
&= \max_y \int_{\boldsymbol{x}} P_d(\boldsymbol{x})\,\big(f(y)+g(y)\big)\,d\boldsymbol{x} && (43)\\
&= \max_y \big(f(y)+g(y)\big) \int_{\boldsymbol{x}} P_d(\boldsymbol{x})\,d\boldsymbol{x} && (44)\\
&= \max_y\; f(y)+g(y) && (45)\\
&= 2\,\psi(\tfrac{1}{2})\,. && (46)
\end{aligned}$$

On the other had, by definition we have

$$L_G = \max_D \int_{\boldsymbol{x}} P_d(\boldsymbol{x})\,f(D(\boldsymbol{x})) + P_g(\boldsymbol{x})\,g(D(\boldsymbol{x}))\,d\boldsymbol{x}\,. \tag{47}$$

Since $D$ can be any function that maps from $\mathcal{X}$ to $\mathbb{R}$, we can let $y = D(\boldsymbol{x})$ be a variable independent of $\boldsymbol{x}$. Thus, we have

$$\begin{aligned}
L_G &= \max_y \int_{\boldsymbol{x}} P_d(\boldsymbol{x})\,f(y) + P_g(\boldsymbol{x})\,g(y)\,d\boldsymbol{x} && (48)\\
&= \max_y \int_{\boldsymbol{x}} \big(P_d(\boldsymbol{x})+P_g(\boldsymbol{x})\big)\left(\frac{P_d(\boldsymbol{x})\,f(y)}{P_d(\boldsymbol{x})+P_g(\boldsymbol{x})} + \frac{P_g(\boldsymbol{x})\,g(y)}{P_d(\boldsymbol{x})+P_g(\boldsymbol{x})}\right) d\boldsymbol{x} && (49)\\
&= \int_{\boldsymbol{x}} \big(P_d(\boldsymbol{x})+P_g(\boldsymbol{x})\big) \max_y \left(\frac{P_d(\boldsymbol{x})\,f(y)}{P_d(\boldsymbol{x})+P_g(\boldsymbol{x})} + \frac{P_g(\boldsymbol{x})\,g(y)}{P_d(\boldsymbol{x})+P_g(\boldsymbol{x})}\right) d\boldsymbol{x}\,. && (50)
\end{aligned}$$

Since $\frac{P_d(\boldsymbol{x})}{P_d(\boldsymbol{x})+P_g(\boldsymbol{x})} \in [0,1]$, we have

$$L_G = \int_{\boldsymbol{x}} \big(P_d(\boldsymbol{x})+P_g(\boldsymbol{x})\big)\,\psi\left(\frac{P_d(\boldsymbol{x})}{P_d(\boldsymbol{x})+P_g(\boldsymbol{x})}\right) d\boldsymbol{x}\,. \tag{51}$$

As $\psi(\gamma)$ has a global minimum at $\gamma = \frac{1}{2}$, now we have

$$L_G \geq \int_{\boldsymbol{x}} \big(P_d(\boldsymbol{x})+P_g(\boldsymbol{x})\big)\,\psi(\tfrac{1}{2})\,d\boldsymbol{x} \tag{52}$$

$$= \psi(\tfrac{1}{2}) \int_{\boldsymbol{x}} \big(P_d(\boldsymbol{x}) + P_g(\boldsymbol{x})\big)\, d\boldsymbol{x} \tag{53}$$

$$= 2\,\psi(\tfrac{1}{2})\,. \tag{54}$$

Finally, combining Equation (46) and Equation (54) yields

$$L_G \geq L_G \big|_{P_g = P_d}\,, \tag{55}$$

which holds for any $P_d$, thus concluding the proof. □

**Theorem 4** (Sufficient conditions of the strong desired property of an adversarial loss)**.** *If $\psi(\gamma)$ has a unique global minimum at $\gamma = \frac{1}{2}$, then Property 2 holds.*

*Proof.* Since $\psi(\gamma)$ has a unique global minimum at $\gamma = \frac{1}{2}$, we have for any $\gamma \in [0, \frac{1}{2}) \cup (\frac{1}{2}, 1]$,

$$\psi(\gamma) > \psi(\tfrac{1}{2})\,. \tag{56}$$

When $P_g \neq P_d$, there must be some $\boldsymbol{x}_0 \in \mathcal{X}$ such that $P_g(\boldsymbol{x}_0) \neq P_d(\boldsymbol{x}_0)$. Thus, $\frac{P_d(\boldsymbol{x}_0)}{P_d(\boldsymbol{x}_0)+P_g(\boldsymbol{x}_0)} \neq \frac{1}{2}$, and thereby $\psi\left(\frac{P_d(\boldsymbol{x}_0)}{P_d(\boldsymbol{x}_0)+P_g(\boldsymbol{x}_0)}\right) > \psi(\frac{1}{2})$. Now, by Equation (51) we have

$$L_G \big|_{P_g \neq P_d} = \int_{\boldsymbol{x}} \big(P_d(\boldsymbol{x}) + P_g(\boldsymbol{x})\big)\, \psi\left(\frac{P_d(\boldsymbol{x})}{P_d(\boldsymbol{x}) + P_g(\boldsymbol{x})}\right) d\boldsymbol{x} \tag{57}$$

$$> \int_{\boldsymbol{x}} \big(P_d(\boldsymbol{x}) + P_g(\boldsymbol{x})\big)\, \psi(\tfrac{1}{2})\, d\boldsymbol{x} \tag{58}$$

$$= \psi(\tfrac{1}{2}) \int_{\boldsymbol{x}} \big(P_d(\boldsymbol{x}) + P_g(\boldsymbol{x})\big)\, d\boldsymbol{x} \tag{59}$$

$$= 2\,\psi(\tfrac{1}{2})\,. \tag{60}$$

Finally, combining Equation (46) and Equation (60) yields

$$L_G \big|_{P_g \neq P_d} > L_G \big|_{P_g = P_d}\,, \tag{61}$$

which holds for any $P_d$, thus concluding the proof. □ □

**Theorem 5** (Alternative sufficient conditions of the strong desired property of an adversarial loss)**.** *If $f'' + g'' \leq 0$ and there exists some $y^*$ such that $f(y^*) = g(y^*)$ and $f'(y^*) = -g'(y^*) \neq 0$, then Property 2 holds.*

*Proof.* First, we have by definition

$$\Psi(\gamma, y) = \gamma\, f(y) + (1-\gamma)\, g(y)\,. \tag{62}$$

By taking the partial derivatives, we get

$$\frac{\partial \Psi}{\partial \gamma} = f(y) - g(y)\,, \tag{63}$$

$$\frac{\partial \Psi}{\partial y} = \gamma\, f'(y) + (1-\gamma)\, g'(y)\,, \tag{64}$$

$$\frac{\partial^2 \Psi}{\partial y^2} = \gamma\, f''(y) + (1-\gamma)\, g''(y)\,. \tag{65}$$

We know that there exists some $y^*$ such that

$$f(y^*) = g(y^*)\,, \tag{66}$$

$$f'(y^*) = -g'(y^*) \neq 0\,. \tag{67}$$

(i) By Equation (63) and Equation (64), we see that

$$\frac{\partial \Psi}{\partial \gamma}\Big|_{y=y^*} = 0\,, \tag{68}$$

$$\frac{\partial \Psi}{\partial y}\Big|_{(\gamma,y)=(\frac{1}{2},y^*)} = 0\,. \tag{69}$$

Now, by Equation (68) we know that $\Psi$ is constant when $y = y^*$. That is, for any $\gamma \in [0,1]$,

$$\Psi(\gamma, y^*) = \Psi(\tfrac{1}{2}, y^*)\,. \tag{70}$$

(ii) Because $f'' + g'' \leq 0$, by Equation (65) we have

$$\frac{\partial^2 \Psi}{\partial y^2}\Big|_{\gamma=\frac{1}{2}} = \tfrac{1}{2}\, f''(y) + \tfrac{1}{2}\, g''(y) \tag{71}$$

$$\leq 0\,. \tag{72}$$

By Equation (69) and Equation (71), we see that $y^*$ is a global minimum point of $\Psi\big|_{\gamma=\frac{1}{2}}$. Thus, we now have

$$\Psi(\tfrac{1}{2}, y^*) = \max_y\ \Psi(\tfrac{1}{2}, y) \tag{73}$$

$$= \psi(\tfrac{1}{2})\,. \tag{74}$$

(iii) By Equation (64), we see that

$$\frac{\partial \Psi}{\partial y}\Big|_{y=y^*} = \gamma\, f'(y^*) + (1-\gamma)\, g'(y^*) \tag{75}$$

$$= \gamma\, f'(y^*) + (1-\gamma)\,(-f'(y^*)) \tag{76}$$

$$= (2\gamma - 1)\, f'(y^*)\,. \tag{77}$$

Since $f'(y^*) \neq 0$, we have

$$\frac{\partial \Psi}{\partial y}\Big|_{y=y^*} \neq 0 \quad \forall\, \gamma \in [0, \tfrac{1}{2}) \cup (\tfrac{1}{2}, 1]\,. \tag{78}$$

This shows that for any $\gamma \in [0, \frac{1}{2}) \cup (\frac{1}{2}, 1]$, there must exists some $y^\circ$ such that

$$\Psi(\gamma, y^\circ) > \Psi(\gamma, y^*)\,. \tag{79}$$

And by definition we have

$$\Psi(\gamma, y^\circ) < \max_y\ \Psi(\gamma, y) \tag{80}$$

$$= \psi(\gamma)\,. \tag{81}$$

Hence, by Equation (79) and Equation (80) we get

$$\psi(\gamma) > \Psi(\gamma, y^*)\,. \tag{82}$$

Now, combining Equation (69), Equation (74) and Equation (82) yields

$$\psi(\gamma) > \psi(\tfrac{1}{2}) \quad \forall\, \gamma \in [0, \tfrac{1}{2}) \cup (\tfrac{1}{2}, 1]\,. \tag{83}$$

Finally, by Theorem 4, we know that Property 2 holds, which concludes the proof. □

## B Graphs of the $\Psi$ and $\psi$ Functions for the $\omega$-weighted Losses

Figures 9 and 10 show the graphs of $\Psi$ and $\psi$, respectively, for the $\omega$-weighted versions of the classic, Wasserstein and hinge losses.

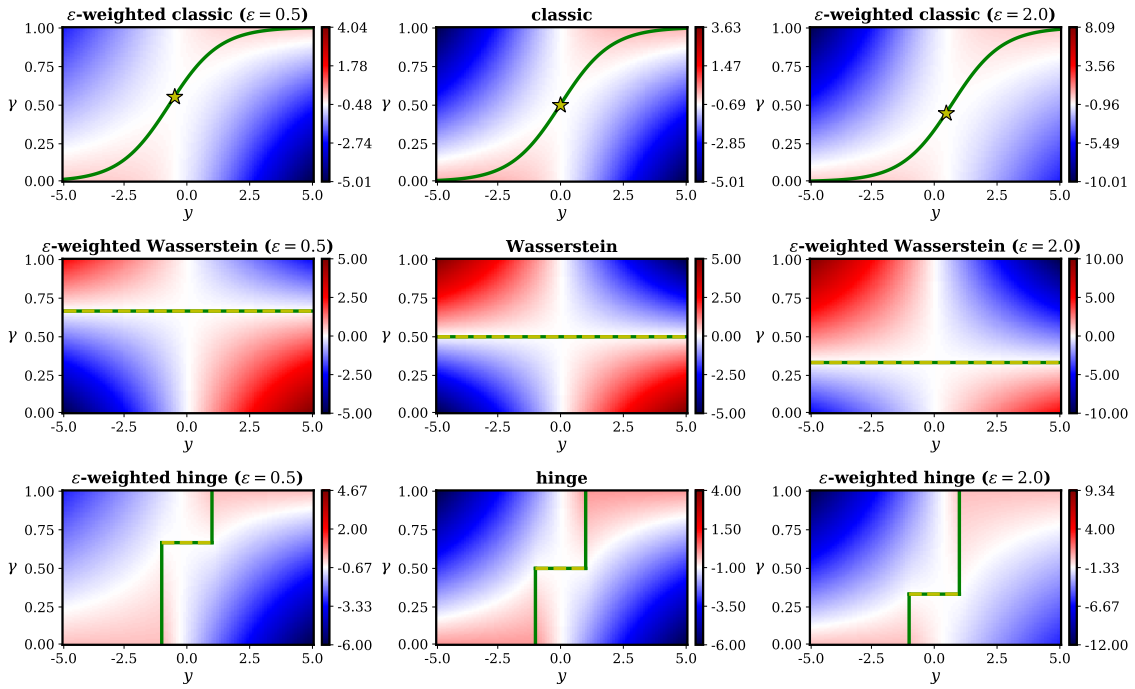



Figure 9: Graphs of the $\Psi$ functions for the $\omega$-weighted versions of the (top) classic and (middle) Wasserstein and (bottom) hinge losses (see Equation (16) for the definition of $\Psi$). The green lines show the domains of $\psi$, i.e., the values that $y$ can take for different $\gamma$. The stars and the yellow dashed lines indicate the global minima of $\psi$. The midpoints of the each color map is set to the minimum of $\psi$, i.e., the values taken at the star marks and yellow segments. Note that as $y \in \mathbb{R}$, we plot different portions of $y$ where the characteristics of $\Psi$ can be best observed. (Best viewed in color.)

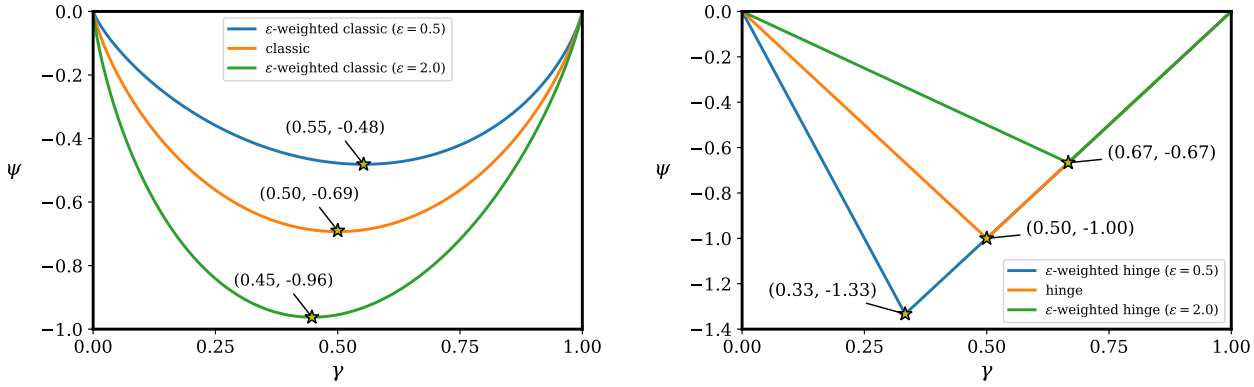



Figure 10: Graphs of the $\psi$ functions for the $\omega$-weighted versions of the (left) classic and (right) hinge losses (see Equation (17) for the definition of $\psi$). The stars indicate their minima. Note that for the Wasserstein loss, $\psi$ is only defined at $\gamma = \frac{2}{3}, \frac{1}{2}, \frac{1}{3}$ when $\omega = 0.5, 1.0, 2.0$, respectively, where it takes the value of zero, and thus we do not include the Wasserstein loss in this figure.