Understanding the Role of Adversarial Regularization in Supervised Learning

Litu Rout
Space Applications Centre
Indian Space Research Organisation
lr@sac.isro.gov.in

Abstract

Despite numerous attempts sought to provide empirical evidence of adversarial regularization outperforming sole supervision, the theoretical understanding of such a phenomenon remains elusive. In this study, we aim to resolve whether adversarial regularization indeed performs better than sole supervision at a fundamental level. To bring this insight to fruition, we study vanishing gradient issue, asymptotic iteration complexity, sub-optimality gap, and provable convergence in the context of sole supervision and adversarial regularization. While the main results revolve around the central theme, the reported derivations rely on different theoretic tools to maintain consistency with existing literature. The key ingredient is a theoretical justification supported by empirical evidence of adversarial acceleration in gradient descent. Also, motivated by a recently introduced unit-wise capacity-based generalization bound, we analyze the generalization error in an adversarial framework.

1. Introduction

At a fundamental level, we study the role of adversarial regularization in supervised learning. We intend to resolve the mystery of why conditional generative adversarial networks accelerate gradient updates when compared with sole supervision. In light of deeper understanding, we explore several crucial properties pertaining to adversarial acceleration.

Over the years several variants of gradient descent algorithms have emerged. In various tasks, adaptive methods including Adagrad [6], RMSProp [38], and ADAM [16] perform better than classical gradient descent. Of particular interest, stochastic version of gradient descent, namely SGD with momentum has enjoyed great success in neural network optimization. Its simplicity, superior performance [42], and theoretical guarantees [2] often provide an edge over other algorithms. This motivates us to choose SGD as our primary learning algorithm [26, 29]. Despite superior empirical performance by SGD, we observe vanishing gradient issue in near optimal region. This is mirrored by poor practical performance when compared with adversarial regularization [4, 40, 21, 41, 44]. We identify the root cause of this issue to be the primary objective function. Since these methods rely on some form of gradients estimated from the supervised objective, the issue of vanishing gradient inherently resides in the near optimal region.

In recent years, the research community has witnessed pervasive use of Generative Adversarial Networks (GANs) on a wide variety of complex tasks [13, 49, 30, 15]. Among many applications, some require generation of a particular sample subject to a conditional input. For this reason, there has been a surge in designing conditional adversarial networks [25]. In visual object tracking via adversarial learning, Euclidean norm is used to regulate the generation process so that the generated mask falls within a small neighborhood of actual mask [36]. In photo-realistic image super resolution, Euclidean or supremum norm is used to minimize the distance between reconstructed and original image [21, 41]. In medical image segmentation, multi-scale $L_1$-loss with adversarial regularization is shown to outperform sole supervision [44]. In medical image analysis, a 3d conditional GAN along with $L_1$-distance is used to super resolve CT scan imagery [18].

Furthermore, Isola et al. [13] use $L_1$-loss as a supervision signal and adversarial regularization as a continuously evolving loss function. Because GANs can learn a loss that adapts to data, they fairly solve multitude of tasks that would otherwise require hand-engineered loss. Xian et al. [43] use adversarial loss on top of pixel, style, and feature loss to restrict the generated images on a manifold of real data. Prior works on this fall under the category of conditional GAN where a composition of pixel and adversarial loss is primarily optimized [25, 4, 40]. Karacan et al. [14] use this technique to efficiently generate images of outdoor scenes. Rout et al. [33] combine spatial and Laplacian spectral channel attention in regularized adversarial learning to synthesize high resolution images. Emami et al. [7] coalesce spatial attention with adversarial regularization and
feature map loss to perform image-to-image translation.

As per these prior works [44, 5, 12, 34, 32], it is understandable that supervised learning with adversarial regularization boosts empirical performance. More importantly, this behavior is consistent across a wide variety of tasks. As much beneficial as this has been so far, to our knowledge, the theoretical understanding still remains relatively less explored. This paper aims to bridge the gap by providing theoretical justification and empirical evidence on the role of adversarial regularization in supervised learning.

2. Preliminaries

2.0.1 Notations

Let \( X \subset \mathbb{R}^{d_x} \) and \( Y \subset \mathbb{R}^{d_y} \), where \( d_x \) and \( d_y \) denote input and output dimensions, respectively. The empirical distribution of \( X \) and \( Y \) are denoted by \( \mathcal{P}_X \) and \( \mathcal{P}_Y \). Given an input \( x \in X, f(\theta; x) : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_y} \) is a neural network with rectified linear unit (ReLU) activation, which is common for both supervised and adversarial learning. Here, \( \theta \) denotes the trainable parameters of the generator, \( f(\cdot; \cdot) \). On the other hand, the discriminator, \( g(\psi; \cdot) \) has trainable parameters collected by \( \psi \). The optimal values of these parameters are represented by \( \theta^* \) and \( \psi^* \). For \( g : \mathbb{R}^{d_y} \rightarrow \mathbb{R}, \nabla g \) denotes its gradient and \( \nabla^2 g \) denotes its Hessian. Given a vector \( x, ||x|| \) represents its Euclidean norm. Given a matrix \( M, ||M|| \) and \( ||M||_F \) denote its spectral and Frobenius norm, respectively.

Definition 1 (L-Lipschitz). A function \( f \) is L-Lipschitz if \( \forall \theta, ||\nabla f(\theta)|| \leq L. \)

Definition 2 (\( \beta \)-Smoothness). A function \( f \) is \( \beta \)-smooth if \( \forall \theta, ||\nabla^2 f(\theta)|| \leq \beta. \)

2.0.2 Problem Setup

In sole supervision, the goal is to optimize the following:

\[
\arg\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{P}} [l(f(\theta; x); y)].
\]  

(1)

In Wasserstein GAN (WGAN) + Gradient Penalty (GP), the generator cost function is given by

\[
\arg\min_{\theta} -\mathbb{E}_{x \sim \mathcal{P}_X} [g(\psi; f(\theta; x))]
\]

(2)

and the discriminator cost function is given by:

\[
\arg\min_{\psi} \mathbb{E}_{x \sim \mathcal{P}_X} [g(\psi; f(\theta; x))] - \mathbb{E}_{y \sim \mathcal{P}_Y} [g(\psi; y)]
+ \lambda_{\text{GP}} \mathbb{E}_{z \sim \mathcal{P}_Z} \left[ \|\nabla z g(\psi; z)\| - 1 \right]^2.
\]

(3)

Here, \( \mathcal{P}_Z \) represents the distribution over samples along the line joining samples from real and generator distribution. Unlike sole supervision, the mapping function \( f_0(\cdot) \) in an augmented objective has access to feedback signals from the discriminator. Thus, the optimization in supervised learning with adversarial regularization is given by

\[
\arg\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{P}} [l(f(\theta; x); y) - g(\psi; f(\theta; x))].
\]

(4)

Here, \( \mathcal{P} \) denotes the joint empirical distribution over \( X \) and \( Y \). The discriminator cost function remains identical to the Wasserstein discriminator as given by equation (3).

3. Theoretical Analysis

This section states the assumptions and their justifications in the context of adversarial regularization. It is intended to justify a multitude of tasks that owe the benefits to adversarial training. The technical overview begins with vanishing gradient issue in the near optimal region. It then presents the main results of this paper. The bounds may appear weak to some readers, but note that the goal of this study is not to provide a tighter bound individually for sole supervision and adversarial regularization. Rather, the goal is to understand the role of adversarial regularization in supervised learning — whether adverserial regularization helps tighten the existing bounds in supervised learning literature. Thus, the emphasis is on providing a theoretical justification to the practical success of supervised learning with adversarial regularization.

3.1. Mitigating Vanishing Gradient

The primary assumptions are stated as following.

Assumption 1. The function \( f(\theta; x) \) is L-Lipschitz in \( \theta \).

Assumption 2. The loss function \( l(p; y) \), where \( p = f(\theta; x) \), is \( \beta \)-smooth in \( p \).

Assumption 1 is a mild requirement that is easily satisfied in the near optimal region. Different from standard smoothness in optimization, it is trivial to justify Assumption 2 by relating it to a quadratic loss function\(^1\).

Lemma 1. Let Assumption 1 and Assumption 2 hold. If \( \|\theta - \theta^*\| \leq \epsilon \), then \( \|\nabla \theta \mathbb{E}_{(x,y) \sim \mathcal{P}} [l(f(\theta; x); y)]\| \leq L^2 \beta \epsilon. \)

Proof. Refer to Appendix C.1. \(\square\)

Lemma 1 provides an upper bound on the expected gradient over empirical distribution \( \mathcal{P} \) in the near optimal region. As the intermediate iterates \( (\theta) \) move closer to the optima \( (\theta^*) \), i.e., \( \epsilon \rightarrow 0 \), the gradient norm vanishes in expectation. This essentially resonates with the intuitive understanding of gradient descent. From another perspective, the issue of gradient descent inherently resides in the near

\(^1\)Please refer to Appendix D for numerical experiments confirming these assumptions in practice.
optimal region. We therefore ask a fundamental question: can we attain faster convergence without having to loose any empirical risk benefits? The following sections are intended to shed light in this direction.

**Lemma 2.** Suppose Assumption 1 holds. For a differentiable discriminator \( g(\psi; y) \), if \( \|g - g^*\| \leq \delta \), where \( g^* \triangleq g(\psi^*) \) denote optimal discriminator, then\[ \| - \nabla_\theta E_{x \sim P_X} [g(\psi; f(\theta; x))]) \| \leq L\delta. \]

*Proof.* Refer to Appendix C.2.

**Lemma 2** indicates that the expected gradient of purely adversarial generator does not produce erroneous gradients in the near optimal region, suggesting well behaved composite empirical risk [44].

**Theorem 1.** Let us suppose Assumption 1 and Assumption 2 hold. If \( \|\theta - \theta^*\| \leq \epsilon \) and \( \|g - g^*\| \leq \delta \), then\[ \|\nabla_p E_{(x,y) \sim P} [l(f(\theta; x); y) - g(\psi; f(\theta; x))])\| \leq (L^2\beta\epsilon + L\delta). \]

*Proof.* Refer to Appendix C.3.

To focus more on the empirical success of adversarial regularization, we study a simple convex-concave minimax optimization problem. It will certainly be interesting to borrow some ideas from the vast minimax optimization literature in various other settings [22, 24]. According to **Theorem 1**, the expected gradient of augmented objective does not vanish in the near optimal region, i.e., \( \|\Delta\theta\| \to L\delta \) as \( \epsilon \to 0 \). In the current setting, the estimated gradients of \( l(\theta) \) and \(-g(\theta)\) at any instant during the optimization process are positively correlated. Thus, the gradients of augmented objective is lower bounded by\[ \|\nabla_p E_{(x,y) \sim P} [l(f(\theta; x); y) - g(\psi; f(\theta; x))])\| \geq \|\nabla_p E_{(x,y) \sim P} [l(f(\theta; x); y))]\|. \]

The upper and lower bounds of the intermediate iterates justify non-vanishing gradient in the near optimal region. It is important to heed the fact that supervised learning with adversarial regularization sets a more stringent criterion, which requires convergence of both primary and secondary objectives. In a smooth-convex-concave setting, which is not necessarily true in the deep learning paradigm, \( \epsilon \to 0 \) promotes the reduction of \( \delta \) that makes the generator close to optimal generator. Although this results in vanishing gradients, the stringent convergence criterion would have already accelerated gradient updates in the augmented objective. This will be verified in the following sections. Having mitigated the vanishing gradient issue, it seems natural to wonder whether adversarial regularization improves iteration complexity.

### 3.2. Asymptotic Iteration Complexity

In this section, we analyze global iteration complexity of sole supervision and the augmented objective [45, 3].

The analysis is restricted to a deterministic setting. For a sequence of parameters \( \{\theta_k\}_{k \in \mathbb{N}} \), the complexity of a function \( l(\theta) \) is defined as\[ T_\epsilon (\{\theta_k\}_{k \in \mathbb{N}}, l) := \inf \{ k \in \mathbb{N} | \|\nabla l(\theta_k)\| \leq \epsilon \}. \]

For a given initialization \( \theta_0 \), risk function \( l \) and algorithm \( A_\phi \), where \( \phi \) denotes hyperparameters of training algorithm, such as learning rate and momentum coefficient, \( A_\phi [l, \theta_0] \) denotes the sequence of iterates generated during training. We compute iteration complexity of an algorithm class parameterized by \( p \) hyperparameters, \( A = \{A_\phi\}_{\phi \in \mathbb{R}^p} \) on a function class, \( \mathcal{L} \) as\[ N_\epsilon (\mathcal{A}, \mathcal{L}, \epsilon) := \inf A_\phi \in \mathcal{A} \sup_{\theta_0 \in \{\mathbb{R}^h \times \mathbb{R}^h \times \mathbb{R}^h\},l \in \mathcal{L}} T_\epsilon (A_\phi [l, \theta_0], l). \]

We derive asymptotic bounds under a less restrictive setting as introduced by Zhang et al. [45]. The new condition is weaker than commonly used Lipschitz smoothness assumption. Under this condition, Zhang et al. [45] aim to resolve the mystery of why adaptive gradient methods converge faster. We use this theoretical tool to study the asymptotic convergence bounds. To circumvent tractability issues in non-convex optimization, we follow the common practice of seeking an \( \epsilon \)-stationary point, i.e., \( \|\nabla l(\theta)\| < \epsilon \).

We start by analyzing the iteration complexity of gradient descent with fixed step size. In this regard, we build upon the assumptions made in [45]. To put more succinctly, let us recall the assumptions.

**Assumption 3.** The loss \( l \) is lower bounded by \( l^* > -\infty \).

**Assumption 4.** The function is twice differentiable.

**Assumption 5 ((L₀, L₁)-Smoothness).** The function is \((L₀, L₁)\)-smooth, i.e., there exist positive constants \( L₀ \) and \( L₁ \) such that \( \|\nabla^2 l(\theta)\| \leq L₀ + L₁ \|\nabla l(\theta)\|. \)

**Theorem 2.** Suppose the functions in \( \mathcal{L} \) satisfy **Assumption 3, 4 and 5**. Given \( \epsilon > 0 \), the iteration complexity in sole supervision is upper bounded by \( \mathcal{O} \left( \frac{((l(\theta_0) - l^*))}{\epsilon} (L₀ + L₁ L^2\beta\epsilon) \right) \).

*Proof.* Refer to Appendix C.4.

**Corollary 1.** Using first order Taylor series, the upper bound in **Theorem 2** becomes \( \mathcal{O} \left( \frac{((l(\theta_0) - l^*))}{\epsilon^2} \right) \).

*Proof.* Refer to Appendix C.5.
**Assumption 6** (Existence of useful gradients). For arbitrarily small $\zeta > 0$, the norm of the gradients of the discriminator is lower bounded by $\zeta$, i.e., $\|\nabla g(\psi; f(\theta); x)\| \geq \zeta$.

Assumption 6 requires the discriminator to provide useful gradients until convergence. It is a valid assumption in minimax optimization problems. Also, it is trivial to prove this in the inner maximization loop under concave setting. In other words, the stated assumptions are mild and derived from prior analyses for the purpose of maintaining consistency with existing literature. Next, we analyze the global iteration complexity in the adversarial setting.

**Theorem 3.** Suppose the functions in $L$ satisfy Assumption 3, 4 and 5. Given Assumption 6 holds, $\epsilon > 0$ and $\delta \leq \frac{\sqrt{2\epsilon}}{L}$, the iteration complexity in adversarial regularization is upper bounded by $O\left(\frac{(l(\theta_0) - l^*) (L_0 + L_1 L^2 \beta^2)}{\epsilon^2 + 2\epsilon^2 - L^2 \beta^2}\right)$.

**Proof.** Refer to Appendix C.6.\qed

**Corollary 2.** Using first order Taylor series, the upper bound in Theorem 3 becomes $O\left(\frac{l(\theta_0) - l^*}{\epsilon^2 + \epsilon\zeta}\right)$.

**Proof.** Refer to Appendix C.7.\qed

Since $2\epsilon\zeta - L^2 \delta^2 \geq 0$, the augmented objective has a tighter global iteration complexity compared to sole supervision. In a simplified setup, one can easily verify this hypothesis by using first order Taylor’s approximation as given by Corollary 1 and 2. In this case, $h\epsilon\zeta > 0$ ensures tighter iteration complexity bound. This result is significant because it improves the convergence rates from $O\left(\frac{1}{\epsilon^2}\right)$ to $O\left(\frac{1}{\epsilon^2 + \epsilon\zeta}\right)$. Notice that for a too strong discriminator, Assumption 6 does not hold. For a too weak discriminator, $\|g - g^*\| \leq \delta$ does not hold when $\delta$ is arbitrarily small. In these cases, the generator does not receive useful gradients from the discriminator to undergo accelerated training. However, for a sufficiently trained discriminator, i.e., $\|g - g^*\| \leq \delta \leq \frac{\sqrt{2\epsilon}}{L}$, adversarial acceleration is guaranteed. Notably, the empirical risk and iteration complexity benefit from this provided the discriminator and the generator are trained alternatively as typically followed in practice.

### 3.3. Sub-Optimality Gap

Here, we analyze continuous time gradient flow. The sub-optimality gap of the generator and the discriminator are defined by $\kappa(t) = \kappa(\theta(t)) := l(\theta(t)) - l(\theta^*)$ and $\pi(t) = \pi(\theta(t)) := g(\theta^*) - g(\theta(t))$, respectively. In the adversarial setting, $l(.)$ is a convex function, and $g(.)$ is a concave function. For clarity, we first analyze the gradient flow in sole supervision using common theoretic tools and then extend this analysis to the augmented objective.

**Theorem 4.** In purely supervised learning, the sub-optimality gap at the average over all iterates in a trajectory of $T$ time steps is upper bounded by $O\left(\frac{\|\theta(0) - \theta^*\|^2}{2T}\right)$.

**Proof.** Refer to Appendix C.8.\qed

**Theorem 5.** In supervised learning with adversarial regularization, the sub-optimality gap at the average over all iterates in a trajectory of $T$ time steps is upper bounded by

$$O\left(\frac{\|\theta(0) - \theta^*\|^2}{2T} - \frac{\pi}{T} \int_0^T \theta(t) dt\right).$$

**Proof.** Refer to Appendix C.9.\qed

According to Theorem 4 and 5, the distance to optimal solution decreases rapidly in the augmented objective when compared with the supervised objective. Since the sub-optimality gap is a non-negative quantity and $\pi \left(\frac{1}{T} \int_0^T \theta(t) dt\right) \geq 0$, the augmented objective has a tighter sub-optimality gap. The tightness is controlled by the sub-optimality gap of the adversary, $\pi(.)$ at the average over all iterates in the same trajectory. It is worth mentioning that the sub-optimality gap in the adversarial setting is at least as good as sole supervision. Also, these theorems do not require all the iterates to be within the tiny landscape of optimal empirical risk. The genericness of these theorems provides further evidence of empirical risk benefits in the augmented objective.

### 4. Concluding Remarks

In this study, we investigated the slow convergence property of sole supervision in the near optimal region, and how adversarial regularization helped circumvent this issue. Further, we explored several crucial properties at this juncture of understanding the role of adversarial regularization in supervised learning. Particularly intriguing was the genericness of these theorems around the central theme. To make a fair assessment, standard theoretic tools were employed in all the theorems. From theoretical perspective, the iteration complexity, sub-optimality gap, convergence guarantee, and the analysis of generalization error provided further insights to the empirical findings. While the sub-optimality gap proved tighter empirical risk, the iteration complexity justified adversarial acceleration. Moreover, it was shown that the learning algorithm would converge even with adversarial regularization. Although we found the improvement in empirical risk to be marginal on some datasets, the theoretical analysis justified accelerated training in conditional generative modeling, which was one of the primary subjects of investigation.
References


In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4401–4410, 2019. 1


