Understanding The Role of Adversarial Regularization in Supervised Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Despite numerous attempts sought to provide empirical evidence of adversarial regularization outperforming sole supervision in various inverse problems, the theoretical understanding of such phenomena 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 into fruition, we study vanishing gradient issue, asymptotic iteration complexity and gradient flow in the context of sole supervision and adversarial regularization. The key ingredient is a theoretical justification supported by empirical evidence of adversarial acceleration in gradient descent. In addition, motivated by a recently introduced unit-wise capacity based generalization bound, we analyze the generalization error in adversarial framework. Guided by our observation, we cast doubts on the ability of this measure to explain generalization. We therefore leave as open questions to explore new measures that can explain generalization behavior in adversarial learning.

1 Introduction

At a fundamental level, we study the role of adversarial regularization in supervised learning through the lens of theoretical justification. We intend to resolve the mystery of why supervised learning with adversarial regularization accelerates gradient updates as compared to sole supervision. In light of deeper understanding, we explore several crucial properties pertaining to adversarial acceleration in gradient descent.

In recent years, the research community has witnessed pervasive use of Generative Adversarial Networks (GANs) on a wide variety of complex tasks [1, 2, 3, 4]. 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. 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 [5]. In photo-realistic image super resolution, Euclidean or supremum norm is used to minimize the distance between reconstructed and original high resolution image [6, 7]. In medical image segmentation, multi-scale L_1-loss with adversarial regularization is shown to outperform sole supervision [8].

The authors of [1] use L_1-loss as a supervision signal and adversarial regularization as a continuously evolving loss function. Because GANs learn a loss that adapts to data, they fairly solve multitude of tasks which would otherwise require hand-engineered loss. The authors of [9] 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 operate under the synonym conditional GAN where a convex composition of pixel and adversarial loss is primarily optimized [10, 11, 12]. Karacan et al. [13] use this technique to efficiently generate images of outdoor scenes. The authors of [14] combined spatial and Laplacian spectral channel attention in regularized adversarial learning to synthesize high resolution images.

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

Furthermore, Emami et al. [15] coalesce spatial attention with adversarial regularization and feature map loss to achieve state-of-the-art image to image translation.

Additionally, the spectral and spatial super resolution based on adversarial regularization [16, 17] is proven to achieve faster convergence and better empirical risk compared to purely supervised learning [18]. Further, the authors of [6] showed improvement in perceptual quality of high resolution images in adversarial setting [19]. Despite superior empirical performance of adversarial regularization in diverse domains, the theoretical understanding of such phenomena remains elusive. So far the theoretical analysis suggests that there is a constant that bounds the total empirical risk above [8]. As a consequence, this inhibits erroneous gradient estimation by the discriminator which apparently improves perceptual quality. However, these benign properties of loss surface do not fully explain this phenomenon at a fundamental level.

As per these prior works [17, 8, 16, 19, 20, 21], it is understandable that supervised learning with adversarial regularization boosts empirical performance. In addition, this improvement is consistent across a wide variety of inverse problems and network configurations. As much beneficial as this regularization has been so far, to our knowledge, the theoretical understanding still remains relatively less explored. Aiming to bridge this gap, we provide both theoretical and empirical evidence of faster convergence due to adversarial regularization.

2 Preliminaries

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 distributions of X and Y are denoted by $\mathcal{P}_X$ and $\mathcal{P}_Y$. Given an input $x \in X$, $f(\theta; x) : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$ represents a common neural network architecture with rectified linear unit (ReLU) activation for both supervised and adversarial learning. Here, θ denotes the trainable parameters of the generator $f(\theta; .)$. The discriminator, $g(\psi; .)$ has trainable parameters collected by ψ. For $g : \mathbb{R}^{d_y} \to \mathbb{R}$, ∇g denotes its gradient and $\nabla^2 g$ denotes its Hessian. Given a vector x, $\|x\|$ represents the Euclidean norm. Given a matrix M, $\|M\|$ and $\|M\|_F$ represent 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. (β-Smoothness) *A function f is β-smooth if $\forall\theta$, $\left\|\nabla^2 f(\theta)\right\| \leq \beta$*

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

$$\arg\min_{\theta} -\mathbb{E}_{x\sim\mathcal{P}_X}\left[g\left(\psi; f\left(\theta; x\right)\right)\right] \tag{1}$$

and the discriminator cost function,

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

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 $f_\theta(.)$ in augmented objective has access to a feedback signal from the discriminator. Thus, the optimization in supervised learning with adversarial regularization is carried out by

$$\arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[l\left(f(\theta; x); y\right) - g\left(\psi; f\left(\theta; x\right)\right)\right]. \tag{3}$$

The discriminator cost function remains identical to Wasserstein discriminator as given by equation (2). Here, $\mathcal{P}$ denotes the joint empirical distribution over X and Y.

3 Theoretical Analysis

This section clearly states the assumptions and justifies their fidelity in the context of adversarial regularization. The theoretical findings are intended to provide a reasonable justification to multitude of tasks that owe the benefits to adversarial training. The technical overview begins with exploiting vanishing gradient issue in the near optimal region. It then presents the main results by estimating the iteration complexity and sub-optimality gap.

3.1 Warm-Up: Mitigating Vanishing Gradient in Near Optimal Region

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

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

Assumption 1 is a mild requirement that is easily satisfied in near optimal region. Different from standard smoothness in optimization, it is trivial to justify **Assumption 2** by relating it to a quadratic loss function as followed by most in practice.

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

Proof. Refer to Appendix D.1. **Lemma 1** provides an upper bound on the expected gradient over empirical distribution $\mathcal{P}$ in near optimal region. As the intermediate iterates (θ) move closer to the optima (θ^*), i.e. $\epsilon \to 0$, 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 near optimal region. We therefore ask a fundamental question: can we attain faster convergence without loosing any empirical risk benefits? The following sections are intended to shed some 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 $\|-\nabla_\theta \mathbb{E}_{x\sim\mathcal{P}_X}\left[g\left(\psi; f\left(\theta; x\right)\right)\right]\| \leq L\delta$.*

Proof. Refer to Appendix D.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 [8].

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

Proof. Refer to Appendix D.3. To focus more on the empirical success of adversarial regularization, we study simple convex-concave minimax optimization. It will certainly be interesting to borrow some ideas from the vast minimax optimization literature in various other settings [22, 23, 24, 25]. 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 $\left\|\nabla_\theta \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[l\left(f(\theta; x); y\right) - g\left(\psi; f\left(\theta; x\right)\right)\right]\right\| \geq \left\|\nabla_\theta \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[l\left(f(\theta; x); y\right)\right]\right\|$. The upper and lower bounds of the intermediate iterates justify non-vanishing gradient in near optimal region. Having proven the contribution of discriminator in mitigating vanishing gradient, it seems natural to wonder whether adversarial regularization improves the iteration complexity, which we discuss in the following Section 3.2.

3.2 Main Results: Asymptotic Iteration Complexity

In this section, we analyze global iteration complexity of both sole supervision and adversarial regularization [26, 27]. We restrict our analysis to a deterministic setting. For a deterministic sequence of parameters $\{\theta_k\}_{k\in\mathbb{N}}$, the complexity of $\{\theta_k\}_{k\in\mathbb{N}}$ for a function $l(\theta)$ is defined as

$$\mathcal{T}_\epsilon\left(\{\theta_k\}_{k\in\mathbb{N}}, l\right) := \inf\left\{k \in \mathbb{N} \mid \|\nabla l\left(\theta_k\right)\| \leq \epsilon\right\}.$$

For a given initialization θ_0, risk function l and algorithm A_ϕ, where ϕ denotes hyperparameters of training algorithm, such as learning rate and momentum coefficient, $A_\phi\left[l, \theta_0\right]$ denotes the sequence of iterates generated during training. We compute iteration complexity of an algorithm class parameterized by p hyperparameters, $\mathcal{A} = \{A_\phi\}_{\phi\in\mathbb{R}^p}$ on a function class, $\mathscr{L}$ as

$$\mathcal{N}\left(\mathcal{A}, \mathscr{L}, \epsilon\right) := \inf_{A_\phi\in\mathcal{A}} \sup_{\theta_0\in\left\{\mathbb{R}^{h\times d_x}, \mathbb{R}^{d_y\times h}\right\}, l\in\mathscr{L}} \mathcal{T}_\epsilon\left(A_\phi\left[l, \theta_0\right], l\right).$$

We derive the asymptotic bounds under a less restrictive setting as introduced in [26]. The new condition is weaker than commonly used Lipschitz smoothness assumption. Under this condition, the authors of [26] aim to resolve the mystery of why adaptive gradient methods converge faster. We use this theoretical tool to study the asymptotic convergence of sole supervision and adversarial regularization in near optimal region. To circumvent the tractability issues in non-convex optimization, we follow the common practice of seeking an ϵ-stationary point, i.e. $\|\nabla l\left(\theta\right)\| < \epsilon$. We start by analyzing the iteration complexity of gradient descent with fixed step size. In this regard, we build on the assumptions made in [26]. 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. ((L_0, L_1)-Smoothness). The function is (L_0, L_1)-smooth, i.e. there exist positive constants L_0 and L_1 such that $\left\|\nabla^2 l\left(\theta\right)\right\| \leq L_0 + L_1 \left\|\nabla l\left(\theta\right)\right\|$.

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

Proof. Refer to Appendix D.4.

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

Proof. Refer to Appendix D.5.

Assumption 6. (Existence of useful gradients) *For arbitrarily small $\zeta > 0$, the norm of the gradients provided by discriminator is lower bounded by ζ, i.e. $\|\nabla g\left(\psi; f\left(\theta; x\right)\right)\| \geq \zeta$.*

Assumption 6 requires discriminator to provide useful gradients until convergence. This is a valid assumption in convex-concave minimax optimization problems. 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 analysis for the sole pupose of mathematical simplicity. Keeping this in mind, we analyze the global iteration complexity in adversarial setting.

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

Proof. Refer to Appendix D.6.

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

Proof. Refer to Appendix D.7. Since $2\epsilon\zeta - L^2\delta^2 \geq 0$, the supervised learning with adversarial regularization has a *tighter* global iteration complexity compared to sole supervision. In a simplified setup, one can easily verify this hypothesis by following the proof using first order Taylor's approximation as given by **Corollary 1** and **2**. In this case, $h\zeta\epsilon > 0$ ensures *tighter* iteration complexity bound. This result is significant because it improves the convergence rates from $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$ to $\mathcal{O}\left(\frac{1}{\epsilon^2+\epsilon\zeta}\right)$. 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 δ 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. $\|g - g^*\| \leq \delta \leq \frac{\sqrt{2\epsilon\zeta}}{L}$, the adversarial regularization accelerates gradient updates. Notably, both the empirical risk and iteration complexity benefit from this provided the discriminator and generator are trained alternatively as typically followed in practice[1].

4 Discussion

In this study, we investigated the reason behind slow convergence of purely supervised learning in near optimal region, and how adversarial regularization circumvents 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. In theoretical analysis, the asymptotic iteration complexity, gradient flow, provable convergence guarantee and the analysis of generalization error provided further insights to the empirical findings of adversarial regularization as reported in copious literature.

[1] Refer to Appendix for further analysis and experimental results.

References

[1] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

[2] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

[3] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, “Semantic image synthesis with spatially-adaptive normalization,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2337–2346, 2019.

[4] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

[5] Y. Song, C. Ma, X. Wu, L. Gong, L. Bao, W. Zuo, C. Shen, R. W. Lau, and M.-H. Yang, “Vital: Visual tracking via adversarial learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8990–8999, 2018.

[6] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, *et al.*, “Photo-realistic single image super-resolution using a generative adversarial network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, 2017.

[7] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, “Esrgan: Enhanced super-resolution generative adversarial networks,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 0–0, 2018.

[8] Y. Xue, T. Xu, H. Zhang, L. R. Long, and X. Huang, “Segan: Adversarial network with multi-scale l-1 loss for medical image segmentation,” *Neuroinformatics*, vol. 16, no. 3-4, pp. 383–392, 2018.

[9] W. Xian, P. Sangkloy, V. Agrawal, A. Raj, J. Lu, C. Fang, F. Yu, and J. Hays, “Texturegan: Controlling deep image synthesis with texture patches,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8456–8465, 2018.

[10] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” *arXiv preprint arXiv:1411.1784*, 2014.

[11] E. L. Denton, S. Chintala, R. Fergus, *et al.*, “Deep generative image models using laplacian pyramid of adversarial networks,” in *Advances in neural information processing systems*, pp. 1486–1494, 2015.

[12] X. Wang and A. Gupta, “Generative image modeling using style and structure adversarial networks,” in *European Conference on Computer Vision*, pp. 318–335, Springer, 2016.

[13] L. Karacan, Z. Akata, A. Erdem, and E. Erdem, “Learning to generate images of outdoor scenes from attributes and semantic layouts,” *arXiv preprint arXiv:1612.00215*, 2016.

[14] L. Rout, I. Misra, S. M. Moorthi, and D. Dhar, “S2a: Wasserstein gan with spatio-spectral laplacian attention for multi-spectral band synthesis,” in *Proceedings of the IEEE conference on computer vision and pattern recognition workshop*, 2020.

[15] H. Emami, M. M. Aliabadi, M. Dong, and R. B. Chinnam, “Spa-gan: Spatial attention gan for image-to-image translation,” *arXiv preprint arXiv:1908.06616*, 2019.

[16] L. Rout, “Alert: Adversarial learning with expert regularization using tikhonov operator for missing band reconstruction,” *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[17] A. Rangnekar, N. Mokashi, E. Ientilucci, C. Kanan, and M. Hoffman, “Aerial spectral super-resolution using conditional adversarial networks,” *arXiv preprint arXiv:1712.08690*, 2017.

[18] C. Lanaras, J. Bioucas-Dias, S. Galliani, E. Baltsavias, and K. Schindler, “Super-resolution of sentinel-2 images: Learning a globally applicable deep neural network,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 146, pp. 305–319, 2018.

[19] C. Dong, C. C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[20] M. Henaff, A. Canziani, and Y. LeCun, "Model-predictive policy learning with uncertainty regularization for driving in dense traffic," *arXiv preprint arXiv:1901.02705*, 2019.

[21] M. Sarmad, H. J. Lee, and Y. M. Kim, "Rl-gan-net: A reinforcement learning agent controlled gan network for real-time point cloud shape completion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5898–5907, 2019.

[22] T. Lin, C. Jin, and M. I. Jordan, "On gradient descent ascent for nonconvex-concave minimax problems," *arXiv preprint arXiv:1906.00331*, 2019.

[23] T. Lin, C. Jin, M. Jordan, *et al.*, "Near-optimal algorithms for minimax optimization," *arXiv preprint arXiv:2002.02417*, 2020.

[24] C. Jin, P. Netrapalli, and M. I. Jordan, "What is local optimality in nonconvex-nonconcave minimax optimization?," *arXiv preprint arXiv:1902.00618*, 2019.

[25] P. Mertikopoulos, C. Papadimitriou, and G. Piliouras, "Cycles in adversarial regularized learning," in *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 2703–2717, SIAM, 2018.

[26] J. Zhang, T. He, S. Sra, and A. Jadbabaie, "Why gradient clipping accelerates training: A theoretical justification for adaptivity," in *International Conference on Learning Representations*, 2019.

[27] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford, "Lower bounds for finding stationary points i," *Mathematical Programming*, pp. 1–50, 2019.

[28] A. Kudo, Y. Kitamura, Y. Li, S. Iizuka, and E. Simo-Serra, "Virtual thin slice: 3d conditional gan-based super-resolution for ct slice interval," in *International Workshop on Machine Learning for Medical Image Reconstruction*, pp. 91–100, Springer, 2019.

[29] M. Schmidt, N. Le Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," *Mathematical Programming*, vol. 162, no. 1-2, pp. 83–112, 2017.

[30] D. Zhou, P. Xu, and Q. Gu, "Stochastic nested variance reduction for nonconvex optimization," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 3925–3936, Curran Associates Inc., 2018.

[31] Y. Nesterov, "Efficiency of coordinate descent methods on huge-scale optimization problems," *SIAM Journal on Optimization*, vol. 22, no. 2, pp. 341–362, 2012.

[32] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford, "Accelerated methods for nonconvex optimization," *SIAM Journal on Optimization*, vol. 28, no. 2, pp. 1751–1772, 2018.

[33] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2121–2159, 2011.

[34] M. Staib, S. J. Reddi, S. Kale, S. Kumar, and S. Sra, "Escaping saddle points with adaptive gradient methods," *arXiv preprint arXiv:1901.09149*, 2019.

[35] D. Zhou, Y. Tang, Z. Yang, Y. Cao, and Q. Gu, "On the convergence of adaptive gradient methods for nonconvex optimization," *arXiv preprint arXiv:1808.05671*, 2018.

[36] J. Zhang, S. P. Karimireddy, A. Veit, S. Kim, S. J. Reddi, S. Kumar, and S. Sra, "Why adam beats sgd for attention models," *arXiv preprint arXiv:1912.03194*, 2019.

[37] S. Lacoste-Julien, M. Schmidt, and F. Bach, "A simpler approach to obtaining an o (1/t) convergence rate for the projected stochastic subgradient method," *arXiv preprint arXiv:1212.2002*, 2012.

[38] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro, "Exploring generalization in deep learning," in *Advances in Neural Information Processing Systems*, pp. 5947–5956, 2017.

[39] V. Nagarajan and J. Z. Kolter, "Generalization in deep networks: The role of distance from initialization," in *Neural Information Processing Systems (NeurIPS) Workshop, Deep Learning: Bridging Theory and Practice*, 2017.

[40] B. Neyshabur, Z. Li, S. Bhojanapalli, Y. LeCun, and N. Srebro, "The role of over-parametrization in generalization of neural networks," in *Proceedings of Intenational Conference on Learning Represenations (ICLR)*, 2019.

A More Related Works

A.1 Adversarial Regularization

The notion of adversarial regularization has also been studied in Reinforcement Learning (RL). The authors of [20] use adversarial learning with expert regularization to learn a predictive policy that allows to drive in simulated dense traffic. In [21], the authors use RL agent controlled GAN along with L_2-distance between global feature vectors to convert noisy, partial point cloud into high-fidelity data. In medical image analysis, a 3d conditional GAN along with L_1-distance is used to super resolve CT scan imagery [28].

A.2 Accelerated Gradients

The idea of accelerated training has long been studied. An elegant line of research focuses on variance reduction that aims to address stochastic and finite sum problems by averaging the stochastic noise [29, 30]. Among momentum based acceleration, much theoretical progress has been made to accelerate any smooth convex optimization [31, 32]. Further, many efforts have been made towards changing the step size across iterations based on estimated gradient norm [33, 34, 35].

B More Theoretical Analysis

B.1 Main Results: Sub-Optimality Gap

Here, we analyze the continuous time gradient flow in both approaches. In this analysis, we define each iterate, $\theta(t)$ at a continuous time, t. The optimal set of parameters is denoted by θ^*. The sub-optimality gap of generator and discriminator are defined by $\kappa(t) = \kappa(\theta(t)) := l\left(\theta(t)\right) - l\left(\theta^*\right)$ and $\pi(t) = \pi(\theta(t)) := g\left(\theta^*\right) - g\left(\theta(t)\right)$, respectively. In adversarial setting, $l(.)$ is a convex downward and $g(.)$ is a convex upward function. For clarity, we first analyze the gradient flow in sole supervision using common theoretic tools and then extend this analysis to adversarial regularization.

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*

$$\mathcal{O}\left(\frac{\|\theta(0)-\theta^*\|^2}{2T}\right).$$

Proof. Refer to Appendix D.8.

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*

$$\mathcal{O}\left(\frac{\|\theta(0)-\theta^*\|^2}{2T} - \pi\left(\frac{1}{T}\int_0^T \theta(t)dt\right)\right).$$

Proof. Refer to Appendix D.9.

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

B.2 Main Results: Provable Convergence

In this section, we analyze the convergence guarantee of the minimax adversarial training under strongly-convex-strongly-concave and smooth nonconvex-nonconcave criteria. In this regard, we assume finite α-moment of estimated stochastic gradients as the unbounded variance has a profound impact on optimization process [36, 37]. At each iteration $k = 1, \ldots, T$, we denote unbiased stochastic gradient by $\mathfrak{g}_k = \mathfrak{g}(\theta_k) \coloneqq \nabla l(\theta_k, \xi) - \nabla g(\theta_k, \xi)$, where ξ represents stochasticity. Here, we analyze rates for global clipping. Similar analyses can also be made for coordinate-wise clipping [36].

Assumption 7. (Existence of α-moment) Suppose we have access to gradients at each iteration. There exist positive real numbers $\alpha \in (1, 2]$ and $G > 0$, such that $\mathbb{E}\left[\|\mathfrak{g}(\theta)\|^{\alpha}\right] \leq G^{\alpha}$ for all θ.

Theorem 6. (Strongly-convex-strongly-concave convergence) *Suppose* ***Assumption 7*** *holds. Let $\mathfrak{l}(\theta_k) \triangleq l(\theta_k) - g(\theta_k)$ is a μ-strongly convex function. Let $\{\theta_k\}$ be the sequence of iterates obtained using global clipping on SGD with momentum $\beta = 0$. Define the output to be k-weighted combination of iterates: $\bar{\theta} = \frac{\sum_{k=1}^{T} k\theta_{k-1}}{\sum_{k=1}^{T} k}$. If adaptive clipping $\tau_k = Gk^{\frac{1}{\alpha}}\mu^{\frac{1}{\alpha}}$ and step size $\eta_k = \frac{5}{2\mu(k+1)}$, then the output iterate $\bar{\theta}$ satisfies*

$$\mathbb{E}\left[l\left(\bar{\theta}\right)\right] - l\left(\theta^*\right) \leq \mathcal{O}\left(G^2\left(\mu\left(T+1\right)\right)^{\frac{2-2\alpha}{\alpha}} - \left(g\left(\theta^*\right) - \mathbb{E}\left[g\left(\bar{\theta}\right)\right]\right)\right).$$

Proof. Refer to Appendix D.10.

Observe that when we eliminate adversarial regularization and set $\alpha = 2$, we recover exactly the SGD rate, i.e., $\mathcal{O}\left(\frac{G^2}{\mu T}\right)$ [37]. Thus, adversarial regularization does converge under strongly-convex-strongly-concave criterion. However, it is determined by the convergence of the inner maximization loop in minimax optimization.

Theorem 7. (Nonconvex-nonconcave convergence) *Suppose* ***Assumption 3.1*** *and* ***3.2*** *hold. Let $\mathfrak{l}(\theta_k) \triangleq l(\theta_k) - g(\theta_k)$ is a possible L-smooth function and $\{\theta_k\}$ be the sequence of iterates obtained using global clipping on SGD with momentum, $\beta = 0$. Given constant clipping $\tau_k = G\left(\eta_k L\right)^{\frac{-1}{\alpha}}$ and constant step size $\eta_k = \left(\frac{R_0^{\alpha} L^{2-2\alpha}}{G^2 T^{\alpha}}\right)^{\frac{1}{3\alpha-2}}$, where $R_0 = l(\theta_0) - l(\theta^*)$, the sequence $\{\theta_k\}$ satisfies*

$$\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\left\|\nabla l\left(\theta_{k-1}\right)\right\|^2\right] \leq \mathcal{O}\left(G^{\frac{2\alpha}{3\alpha-2}}\left(\frac{R_0 L}{T}\right)^{\frac{2\alpha-2}{3\alpha-2}} - \frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\left\|\nabla g(\theta_{k-1})\right\|^2\right]\right).$$

Proof. Refer to Appendix D.11.

By setting $\alpha = 2$ and discarding adversarial acceleration, we obtain the standard SGD rate, $\mathcal{O}\left(\frac{G}{\sqrt{T}}\right)$. It is important to heed the fact adversarial regularization converges under nonconvex-nonconcave criterion as well. To this end, we have established that augmented objective is *guaranteed* to converge under strongly-convex-strongly-concave and nonconvex-noncave criteria provided the assumptions are satisfied. These convergence guarantees provide additional insights to our understanding of adversarial regularization in practice. It is necessary to hightlight the fact that such analysis is far from being conclusive. While this paper studies minimax optimization under nonconvex-smooth settings, it will be interesting to derive convergence guarantees under nonconvex-nonsmooth settings.

B.3 Main Results: Generalization Error

Motivated by the role of over-parametrization in generalization [38, 39, 40], we study the generalization behavior of adversarial regularization. We use Rademacher complexity to get a bound on generalization error. Since it depends on hypothesis class, we use a set of restricted parameters of trained networks to get a tighter bound on generalization. The restricted set of parameters is defined as

$$\mathcal{W} = \left\{(V, U)\,|V \in \mathbb{R}^{d_y \times h}, U \in \mathbb{R}^{h \times d_x}, \|v_i\| \leq \alpha_i, \left\|u_i - u_i^0\right\| \leq \beta_i\right\},$$

where $i = 1, 2, \ldots, h$. Here, $v_i \in \mathbb{R}^{dy}$ and $u_i \in \mathbb{R}^{dx}$ denote vector representation of each neuron in the top layer and hidden layer, respectively. Thus, the restricted hypothesis class becomes

$$\mathcal{F}_{\mathcal{W}} = \left\{V[Ux]_+ |\,(V, U) \in \mathcal{W}\right\},$$

where $[.]_+$ represents ReLU activation. For any hypothesis class $\mathcal{F}$, let $l\ o\ \mathcal{F}$ denote the composition of loss function and hypothesis class. The following generalization bound holds for any $f \in \mathcal{F}_\mathcal{W}$ over m training samples with probability $1-\delta$.

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[l\ o\ f\right] \leq \frac{1}{m}\sum_{i=1}^{m} l\left(f(x);y\right) + 2\mathcal{R}_\mathcal{S}\left(l\ o\ \mathcal{F}_\mathcal{W}\right) + 3\sqrt{\frac{ln(2/\delta)}{2m}},$$

where $\mathcal{R}_\mathcal{S}(\mathcal{H})$ is the Rademacher complexity of a hypothesis class $\mathcal{H}$ with respect to training set $\mathcal{S}$.

$$\mathcal{R}_\mathcal{S}\left(\mathcal{H}\right) = \frac{1}{m}\mathbb{E}_{\xi_i\in\{\pm 1\}^m}\left[\sup_{f\in\mathcal{H}}\sum_{i=1}^{m}\xi_i f(x_i)\right].$$

Relative Generalization Error: We define relative generalization error as

$$e_{gen,r} = \left(\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[l\ o\ f\right] - \frac{1}{m}\sum_{i=1}^{m} l\left(f(x);y\right)\right)\times N^*.$$

To be consistent with [40] while studying generalization, we assume $l(f(\theta;x);y)$ be a locally K-Lipschitz function, i.e. given $y\in Y$, $\|\nabla l(f(\theta;x);y)\| \leq K$, $\forall\ \theta$. Using K-Lipschitz property of loss function l in **Lemma 9** of [40], one can easily prove that the Rademacher complexity of $l\ o\ \mathcal{F}_\mathcal{W}$ is bounded as

$$\begin{aligned}
&\mathcal{R}_\mathcal{S}\left(l\ o\ \mathcal{F}_\mathcal{W}\right)\\
&\leq \frac{2K\sqrt{d_y}}{m}\sum_{j=1}^{h}\alpha_j\left(\beta_j\left\|X\right\|_F + \left\|u_j^0 X\right\|_2\right)\\
&\leq \frac{2K\sqrt{d_y}}{\sqrt{m}}\left\|\alpha\right\|_2\left(\left\|\beta\right\|_2\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left\|x_i\right\|_2^2} + \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left\|U^0 x_i\right\|_2^2}\right).
\end{aligned}$$

Adapted to current setting, the generalization error becomes

$$\mathcal{O}\left(\left\|U^0\right\|_2\left\|V\right\|_F + \left\|U-U^0\right\|_F\left\|V\right\|_F + \sqrt{h}\right).$$

Next, we empirically verify the required assumptions and corresponding theoretical results.

C Experiments

This section contains empirical results to support the theoretical findings on sole supervision and adversarial regularization.

C.1 Training Details

The majority of the experiments are conducted on two layer neural networks with ReLU activation function. For completeness however, we experiment with practical neural network architectures. We do not use weight decay, dropout or normalization in these networks. We experiment on both MNIST and CIFAR10 datasets. We use SGD with momentum 0.9, batch size 64 and fixed learning rate of 0.01 for MNIST and CIFAR10. We use mean square error of 0.001 for MNIST and 0.02 for CIFAR10 as our convergence criteria. We train on both datasets for a maximum of 1000 epochs, or until convergence. In these settings, we train 13 architectures on both datasets in which the number of hidden units (h) range from 2^3 to 2^{15}. We use PyTorch to design all these experiments. All parameters are initialized with uniform distribution.

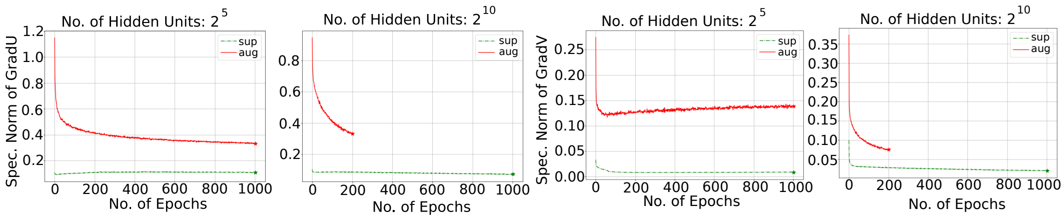

Figure 1: Comparison of gradient updates between supervised (sup) and augmented (aug) objective in the *hidden layer* (left) and *top layer* (right) on MNIST.

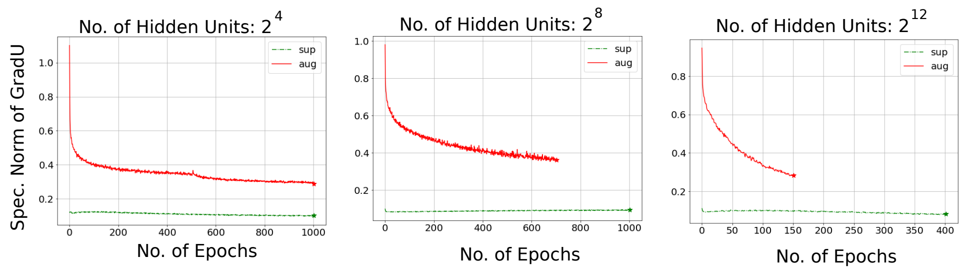

Figure 2: Comparison of gradient updates between supervised and augmented objective as observed in the *hidden layer* on MNIST.

C.2 Experimental Results

C.2.1 Results on MNIST

Figure 1 provides empirical evidence of the vanishing gradient issue and how adversarial regularization helps circumvent this. In all the experimented architectures, the spectral norm of gradients estimated by purely supervised objective is smaller than adversarial learning. This is consistent with the theoretical analysis in Section 3. The main reason for such non-vanishing gradient is the feedback signal from discriminator. Further, we observe that the rate of convergence is at least as good as sole supervision, as marked by $\star$ in Figure 1.

Figure 5 offers experimental support to better empirical risk in adversarial setting. We observe the significance of near optimal region, i.e. ϵ with 32 hidden units in Figure 5. Since the expressive power of such networks is very small in both approaches, evidently neither meets the convergence criteria. However, as the capacity increases the supervised cost, which is common in both approaches, guides them to a tiny landscape around optimal risk and thereby, it satisfies the condition of **Theorem 1**. Under this circumstance, the optimal empirical risk attained by augmented objective can be provably better than sole supervision as predicted by our theory. Figure 5 and 4 supports this theory as augmented objective consistently achieves better performance either by risk or by rate of convergence for networks with sufficient expressive power.

As shown in Figure 2 and 3, the estimated gradient in SGD+momentum vanishes within the tiny landscape of optimal empirical risk. Further, the adversarial regularization accelerates gradient updates and attains minimal empirical risk compared to sole supervision. It is evident from Figure 4 where we observe this particular phenomenon across a wide variety of architectures.

Furthermore, we compare the optimal empirical risk and iteration complexity with different number of hidden units in Figure 6. One can infer from Figure 6 (a) (left) that the value of ϵ in **Theorem 1** is approximately equal to 0.005[2]. The number of epochs required to attain optimum in adversarial learning is always less than or equal to supervised learning, which validates our theorems.

[2]The value of ϵ is more relevant to the present body of analysis as it performs the inverse mapping in practical scenarios. Moreover, it is not hard to estimate δ where discriminator acts as the mapping function.

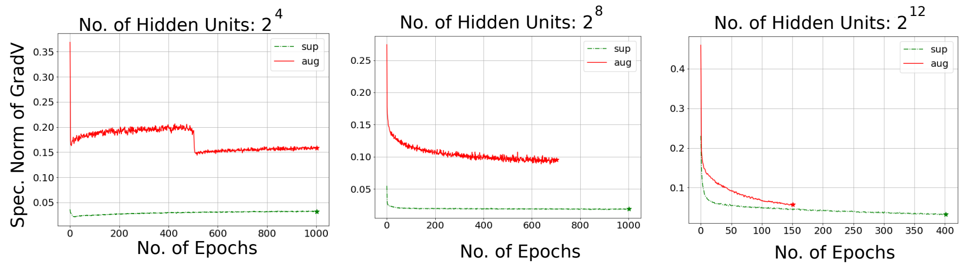

Figure 3: Comparison of gradient updates between supervised and augmented objective as observed in the *top layer* on MNIST.

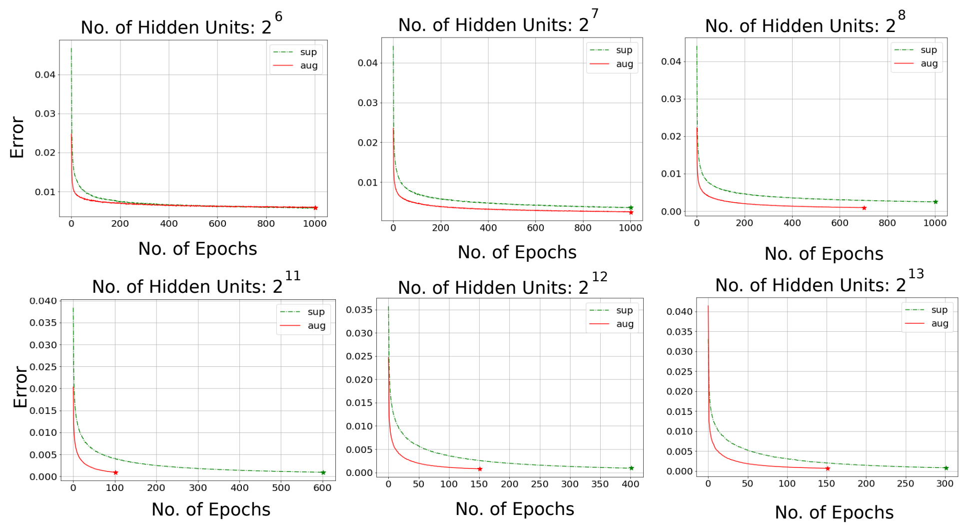

Figure 4: Comparison of optimal empirical risk on MNIST.

C.2.2 Results on CIFAR10

These theorems are also verified on CIFAR10 dataset. As shown in Figure 6 (right), supervised learning with adversarial regularization performs better than sole supervision both in terms of optimal empirical risk and rate of convergence. Here, we find ϵ to be approximately equal to 0.06.

Similar to our analysis on MNIST, we also observe vanishing gradient issue on CIFAR10 which is shown in Figure 7 and 8. Figure 9 illustrates how model capacity correlates with empirical risk and thereby, satisfies the condition of **Theorem 1**. Across a wide variety of architectures, we verify that supervised learning with adversarial regularization can be better than sole supervision both in terms of optimal empirical risk and iteration complexity as predicted by our theory. As shown in Figure 9, though both methods start with almost same initial empirical risk, augmented objective traverses through a shorter path and attains minimal risk upon convergence.

C.2.3 Results on Various Network Configurations

To study the impact of these findings on more realistic scenarios, we experiment on various network configurations. As shown in Figure 10 and 11, the issue of vanishing gradient is persistent across these experimented configurations. Furthermore, the discussion on adversarial acceleration is also supported by Figure 12. In addition, Table 1 shows that the proposed hypothesis: *adversarial regularization achieves tighter ϵ-stationary point at an optimal rate* holds under practical circumstances. More specifically, we observe accelerated gradient updates not only in two layer ReLU networks, but also in deep MLP with exponential linear activations, convolution layers, skip connections, dense connections, L_1 regularized networks, and L_2 regularized networks. Thus, augmented objective owes its performance benefits to adversarial learning at a fundamental level.

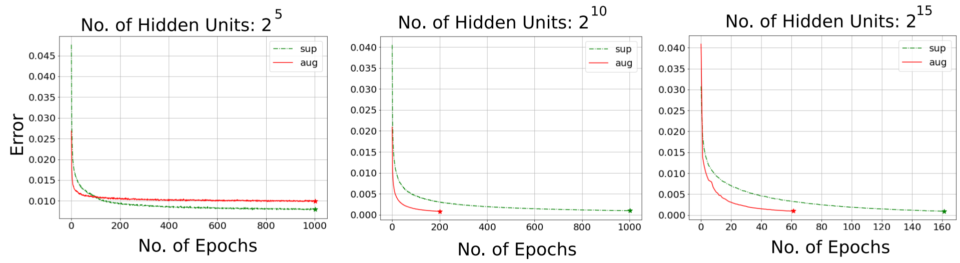

Figure 5: Comparison of optimal empirical risk on MNIST.

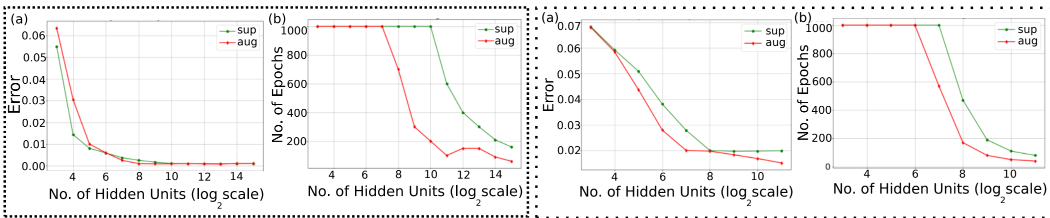

Figure 6: Comparison on MNIST (left) and CIFAR10 (right). (a) Optimal empirical risk. (b) Iteration Complexity. Adversarial regularization attains tighter ϵ-stationary point at an optimal rate.

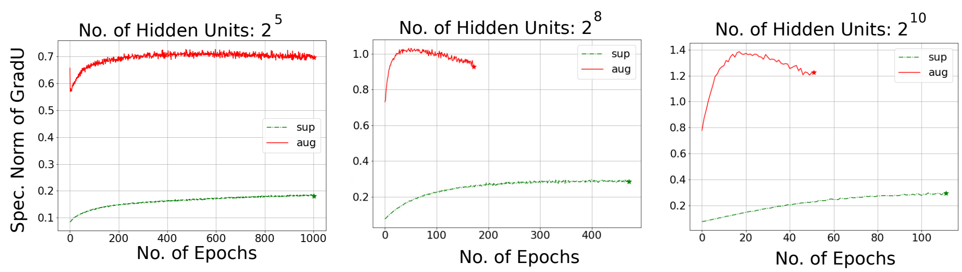

Figure 7: Comparison of gradient updates between supervised and augmented objective as observed in the *hidden layer* on CIFAR10.

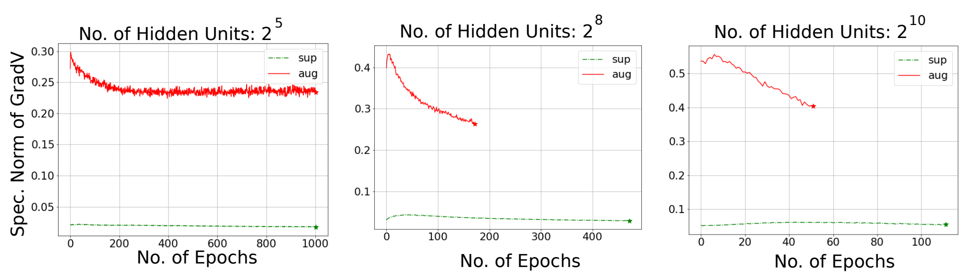

Figure 8: Comparison of gradient updates between supervised and augmented objective as observed in the *top layer* on CIFAR10.

C.3 Results on Generalization Error

The generalization trend in sole supervision is shown in Figure 13(a) and 13(c). As per equation (B.3), the combined measure of Frobenius norm of top layer, i.e. $\|V\|_F$ and distance from initialization of hidden layer, i.e. $\left\|U - U^0\right\|_F$ explains the generalization gap on MNIST and CIFAR10. We verify this measure in our experimental setting and study whether it can explain generalization in adversarial learning. Note that adversarial learning and sole supervision share exactly same mapping function (f), learning algorithm (SGD+momentum) and empirical data distribution (S). Thus the generalization bound, which is derived for a purely supervised objective, is expected to explain the generalization

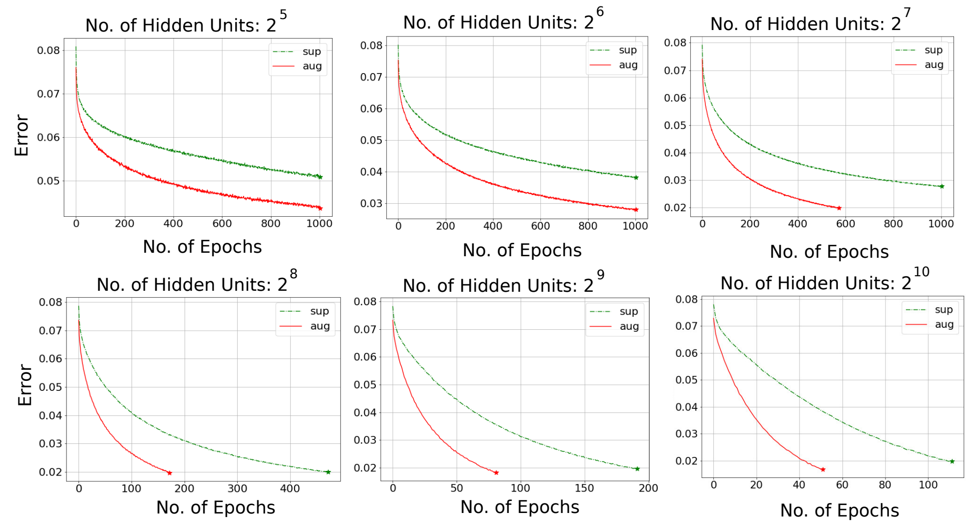

Figure 9: Comparison of optimal empirical risk on CIFAR10.

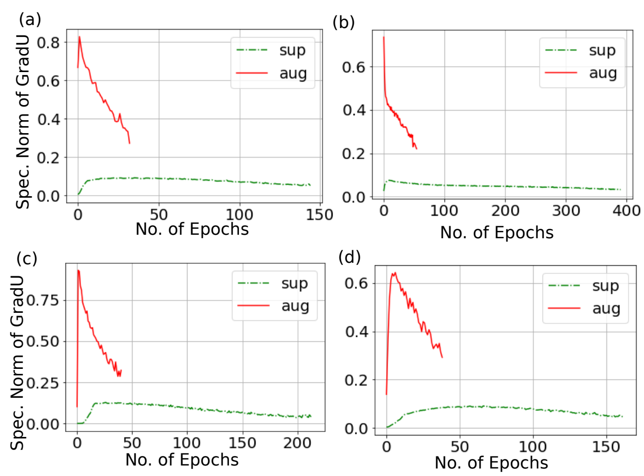

Figure 10: Comparison of gradient updates between supervised and augmented objective as observed in the *first layer* on MNIST. (a) Multi-Layer Perceptron. (b) Exponential Activation. (c) Residual Network. (d) Dense Network.

error in adversarial learning with expert regularization. However, as shown in Figure 13(b) and 13(d), this bound does not fully explain the generalization error observed in adversarial learning.

In Figure 14, we observe that the relative generalization error of adversarial regularization can be better than sole supervision. This is feasible for a network with sufficient expressive power to achieve near optimal convergence.

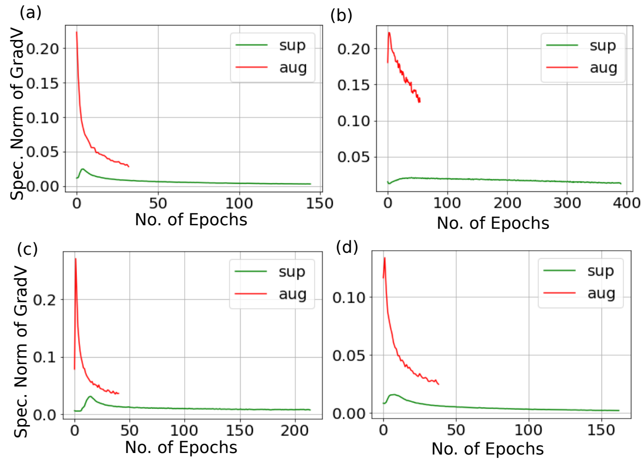

Figure 11: Comparison of gradient updates between supervised and augmented objective as observed in the *last layer* on MNIST. (a) Multi-Layer Perceptron. (b) Exponential Activation. (c) Residual Network. (d) Dense Network.

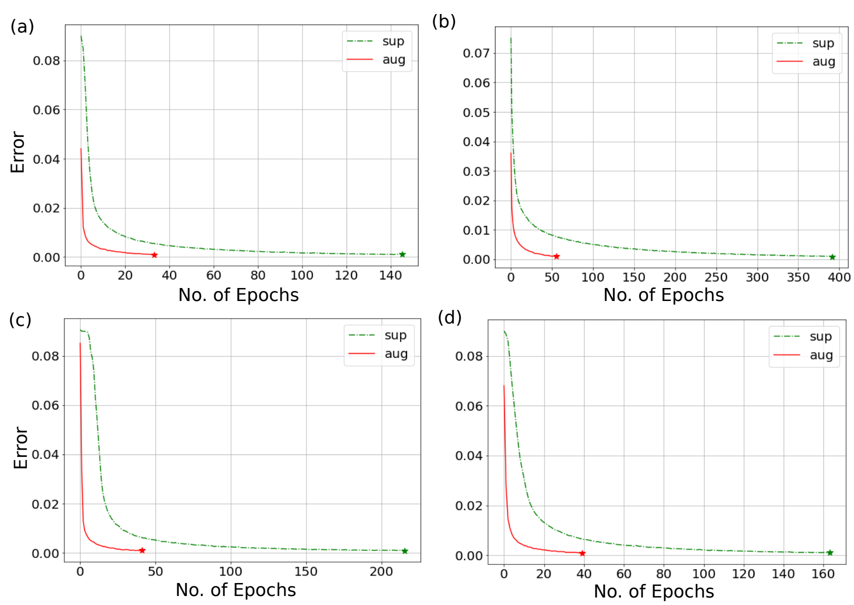

Figure 12: Comparison of optimal empirical risk on MNIST. (a) Multi-Layer Perceptron. (b) Exponential Activation. (c) Residual Network. (d) Dense Network.

Table 1: Hypothesis Testing on Various Network Configurations

Architecture	No. Layer	Activation	No. ResBlock	No. DenseBlock	No. Epoch Sup	No. Epoch Aug	Hypothesis
MLP-Deep	6	ELU	2	0	391	55	✓
CNN-ResNet	6	ReLU	2	0	215	41	✓
CNN-DenseNet	6	ReLU	2	1	163	39	✓
CNN-DenseNet-L1	6	ReLU	2	1	1000	39	✓
CNN-DenseNet-L2	6	ReLU	2	1	155	39	✓
CNN-ResNet-AvgPool	6	ReLU	2	0	109	29	✓

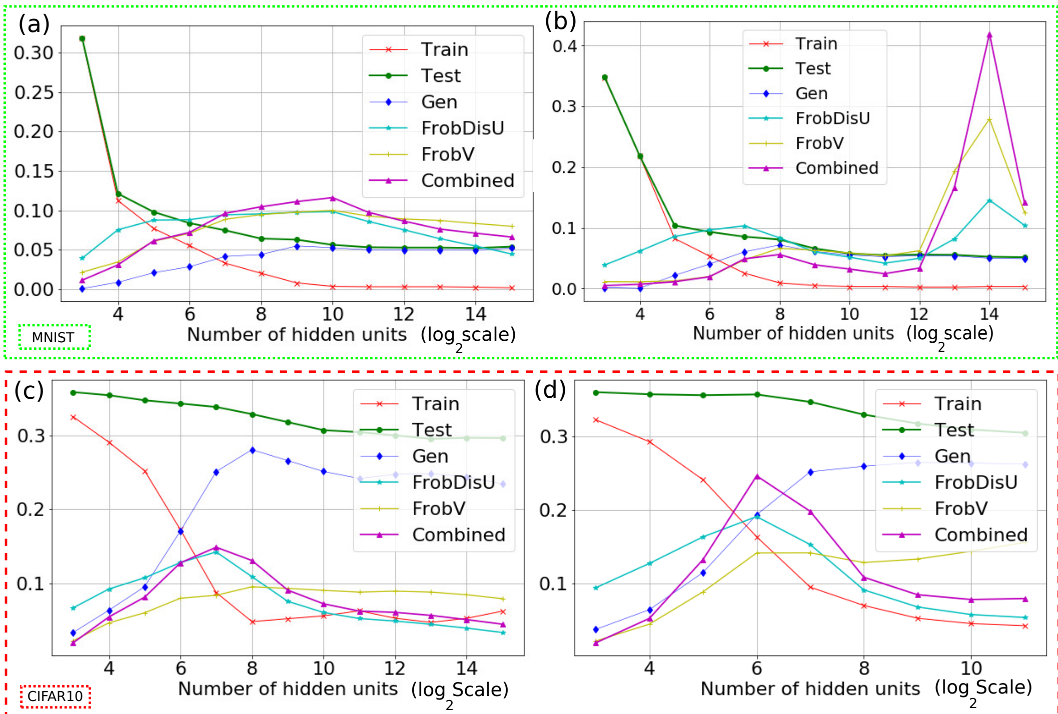

Figure 13: Generalization error on MNIST and CIFAR10.

D Technical Proofs

D.1 Proof of Lemma 1

Using Jensen's inequality,

$$\begin{aligned}
\left\|\nabla_\theta \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[l\left(f(\theta;x);y\right)\right]\right\|^2 &\leq \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[\left\|\nabla_\theta l\left(f(\theta;x);y\right)\right\|^2\right] \\
&\leq \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[\left\|\nabla_p l\left(p;y\right)\nabla_\theta f(\theta;x)\right\|^2\right], \text{where } p = f(\theta;x) \\
&\leq \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[\underbrace{\left\|\nabla_p l\left(p;y\right)\right\|^2\left\|\nabla_\theta f(\theta;x)\right\|^2}_{\text{Cauchy-Schwarz inequality}}\right] \\
&\leq L^2\mathbb{E}_{(x,y)\sim\mathcal{P}}\left[\left\|\nabla_p l\left(p;y\right)\right\|^2\right]
\end{aligned}$$

Let $p = f(\theta;x)$ and $q = f(\theta^*;y)$. Using β-smoothness and L-Lipschitz property, we get

$$\left\|\nabla_p l\left(p;y\right)\right\| - \left\|\nabla_q l\left(q;y\right)\right\| \leq \left\|\nabla_p l\left(p;y\right) - \nabla_q l\left(q;y\right)\right\| \leq \beta\left\|p-q\right\| \leq \beta L\left\|\theta-\theta^*\right\|.$$

Since $\|\theta-\theta^*\| \leq \epsilon$,

$$\left\|\nabla_\theta \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[l\left(f(\theta;x);y\right)\right]\right\|^2 \leq L^2\mathbb{E}_{(x,y)\sim\mathcal{P}}\left[\left(\left\|\nabla_q l\left(q;y\right)\right\| + L\beta\epsilon\right)^2\right].$$

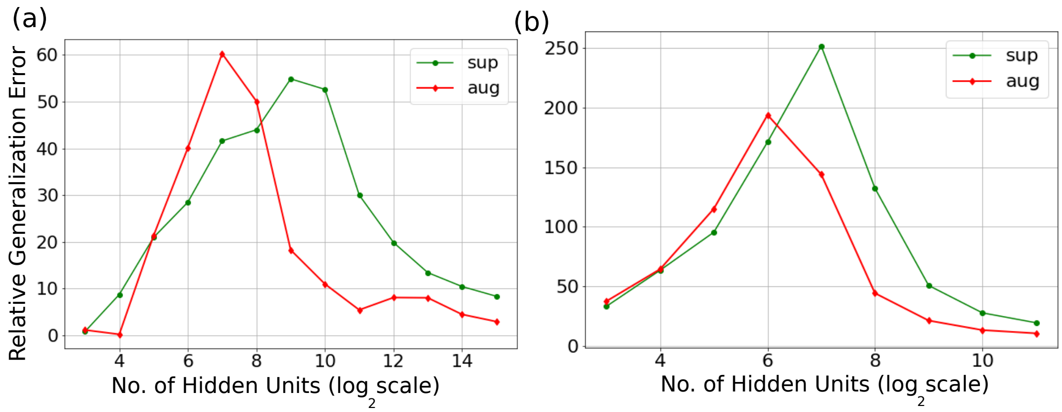

Figure 14: Relative generalization. (a) MNIST. (b) CIFAR10.

Upon substituting optimality condition, i.e. $\|\nabla_q l\,(q; y)\| = 0$, the above expression simplifies to

$$\left\|\nabla_\theta \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[l\left(f(\theta; x); y\right)\right]\right\| \leq L^2\beta\epsilon.$$

This completes the proof of the theorem. □

D.2 Proof of Lemma 2

Using similar arguments from Appendix D.1,

$$\begin{aligned}\left\|-\nabla_\theta \mathbb{E}_{x\sim\mathcal{P}_X}\left[g\left(\psi; f\left(\theta; x\right)\right)\right]\right\|^2 &\leq \mathbb{E}_{x\sim\mathcal{P}_X}\left[\left\|\nabla_\theta g\left(\psi; f\left(\theta; x\right)\right)\right\|^2\right] \\ &\leq \mathbb{E}_{x\sim\mathcal{P}_X}\left[\left\|\nabla_p g\left(\psi; p\right)\right\|^2 \left\|\nabla_\theta f\left(\theta; x\right)\right\|^2\right], \text{ where } p = f\left(\theta; x\right) \\ &\leq L^2 \mathbb{E}_{x\sim\mathcal{P}_X}\left[\left\|\nabla_p g\left(\psi; p\right)\right\|^2\right] \\ &\leq L^2 \mathbb{E}_{x\sim\mathcal{P}_X}\left[\left(\left\|\nabla_p g\left(\psi; p\right)\right\| + \delta\right)^2\right] \\ &\leq L^2\delta^2\end{aligned}$$

Taking square root, $\left\|-\nabla_\theta \mathbb{E}_{x\sim\mathcal{P}_X}\left[g\left(\psi; f\left(\theta; x\right)\right)\right]\right\| \leq L\delta$, which finishes the proof. □

D.3 Proof of Theorem 1

By applying triangle inequality after simplification,

$$\begin{aligned}\left\|\nabla_\theta \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[l\left(f(\theta; x); y\right) - g\left(\psi; f\left(\theta; x\right)\right)\right]\right\| &\leq \left\|\nabla_\theta \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[l\left(f(\theta; x); y\right)\right]\right\| + \left\|-\nabla_\theta \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[\left(\psi; f\left(\theta; x\right)\right)\right]\right\| \\ &\leq L^2\beta\epsilon + L\delta \text{ } (\textbf{Lemma 1} \text{ and } \textbf{Lemma 2}),\end{aligned}$$

which completes the statement of the theorem. □

D.4 Proof of Theorem 2

We parameterize the path between θ_k and θ_{k+1} as following:

$$\gamma(t) = t\theta_{k+1} + (1-t)\theta_k \forall t \in [0, 1]. \quad (4)$$

By fixed step gradient descent, the iterate $\theta_{k+1} = \theta_k - h_k\nabla l(\theta_k)$. Using Taylor's expansion,

$$\begin{aligned}l\left(\theta_{k+1}\right) &= l\left(\theta_k\right) + \nabla l\left(\theta_k\right)\left(\theta_{k+1} - \theta_k\right) + \frac{1}{2}\left(\theta_{k+1} - \theta_k\right)^T \nabla^2 l\left(\theta_k\right)\left(\theta_{k+1} - \theta_k\right) \\ &= l\left(\theta_k\right) - h_k\left\|\nabla l\left(\theta_k\right)\right\|^2 + \frac{1}{2}\left(\theta_{k+1} - \theta_k\right)^T \nabla^2 l\left(\theta_k\right)\left(\theta_{k+1} - \theta_k\right), \; \left(\because \theta_{k+1} - \theta_k = -h_k\nabla l(\theta_k)\right).\end{aligned}$$

Using Cauchy-Schwarz inequality and integrating over parameterized curve $\gamma(t)$,

$$\begin{aligned} l(\theta_{k+1}) &\leq l(\theta_k) - h_k \|\nabla l(\theta_k)\|^2 + \frac{1}{2}\|(\theta_{k+1}-\theta_k)\| \left\|\nabla^2 l(\theta_k)(\theta_{k+1}-\theta_k)\right\| \\ &\leq l(\theta_k) - h_k \|\nabla l(\theta_k)\|^2 + \frac{1}{2}\|(\theta_{k+1}-\theta_k)\|^2 \int_0^1 \left\|\nabla^2 l(\gamma(t))\right\| dt. \end{aligned}$$

We know by **Assumption 5**

$$\left\|\nabla^2 l(\theta)\right\| \leq L_0 + L_1 \|\nabla l(\theta)\|.$$

Then using descent rule and arguments of **Theorem 1**, we obtain the following inequality:

$$\begin{aligned} l(\theta_{k+1}) &\leq l(\theta_k) - h_k \|\nabla l(\theta_k)\|^2 + \frac{h_k^2 \|\nabla l(\theta_k)\|^2}{2} \int_0^1 (L_0 + L_1 \|\nabla l(\gamma(t))\|)\, dt \\ &\leq l(\theta_k) - h_k \|\nabla l(\theta_k)\|^2 + \frac{h_k^2 \|\nabla l(\theta_k)\|^2}{2} \int_0^1 \left(L_0 + L_1 L^2 \beta\epsilon\right) dt \\ &\leq l(\theta_k) - h_k \|\nabla l(\theta_k)\|^2 + \frac{h_k^2 \|\nabla l(\theta_k)\|^2 \left(L_0 + L_1 L^2 \beta\epsilon\right)}{2}. \end{aligned}$$

Let us choose $h_k = \frac{1}{L_0 + L_1 L^2 \beta\epsilon}$. Now,

$$\begin{aligned} l(\theta_{k+1}) &\leq l(\theta_k) - \frac{h_k \|\nabla l(\theta_k)\|^2}{2} \\ &\leq l(\theta_k) - \frac{\|\nabla l(\theta_k)\|^2}{2(L_0 + L_1 \lambda M)}. \end{aligned}$$

Assume that it takes T iterations to reach ϵ-stationary point, i.e., $\epsilon \leq \|\nabla l(\theta_k)\|$ for $k \leq T$. By a telescopic sum over k,

$$\begin{aligned} \sum_{k=0}^{T-1} l(\theta_{k+1}) - l(\theta_k) &\leq \frac{-T\epsilon^2}{2(L_0 + L_1 \lambda M)} \\ \implies T &\leq \frac{2(l(\theta_0) - l^*)\left(L_0 + L_1 L^2 \beta\epsilon\right)}{\epsilon^2}. \end{aligned}$$

Therefore, we get

$$\sup_{\theta_0 \in \left\{\mathbb{R}^{h \times d_x}, \mathbb{R}^{d_y \times h}\right\}, l \in \mathscr{L}} \mathcal{T}_\epsilon(A_h[l, \theta_0], l) = \mathcal{O}\left(\frac{(l(\theta_0) - l^*)\left(L_0 + L_1 L^2 \beta\epsilon\right)}{\epsilon^2}\right)$$

which finishes the proof. □

D.5 Proof of Corollary 1

Using the arguments made in the proof of **Theorem 2** and first-order Taylor's expansion, we get

$$\begin{aligned} l(\theta_{k+1}) &= l(\theta_k) - h_k \|\nabla l(\theta_k)\|^2 \\ &\leq l(\theta_k) - h_k \epsilon^2. \end{aligned}$$

By telescopic sum,

$$\begin{aligned} \sum_{k=0}^{T-1} l(\theta_{k+1}) - l(\theta_k) &\leq -T h_k \epsilon^2 \\ \implies T &\leq \frac{(l(\theta_0) - l^*)}{h_k \epsilon^2}. \end{aligned}$$

So,

$$\sup_{\theta_0 \in \left\{\mathbb{R}^{h \times d_x}, \mathbb{R}^{d_y \times h}\right\}, l \in \mathscr{L}} \mathcal{T}_\epsilon(A_h[l, \theta_0], l) = \mathcal{O}\left(\frac{(l(\theta_0) - l^*)}{h\epsilon^2}\right)$$

which finishes the proof. □

D.6 Proof of Theorem 3

Recall that the target function $l(\theta)$ remains identical in both settings except for additional cost of discriminator over generator in augmented objective. In this setting, the parameters are updated as

$$\theta_{k+1} = \theta_k - h_k \nabla \left(l\left(\theta_k\right) - g\left(\psi; f\left(\theta_k; x\right)\right)\right). \tag{5}$$

Using Taylor's expansion, the triangle and Cauchy-Schwarz inequality as in Appendix D.4, we obtain

$$l\left(\theta_{k+1}\right) \leq l\left(\theta_k\right) - h_k \left\|\nabla l\left(\theta_k\right)\right\|^2 - h_k \left\|\nabla l\left(\theta_k\right)\right\| \left\|\nabla g\left(\psi; f\left(\theta_k; x\right)\right)\right\| + \frac{h_k^2 \left\|\nabla\left(l\left(\theta_k\right) - g\left(\psi; f\left(\theta_k; x\right)\right)\right)\right\|^2}{2} \int_0^1 \left\|\nabla^2 l(\gamma(t))\right\| dt.$$

By **Assumption 5** and **6**,

$$l\left(\theta_{k+1}\right) \leq l\left(\theta_k\right) - h_k \left\|\nabla l\left(\theta_k\right)\right\|^2 - h_k \left\|\nabla l\left(\theta_k\right)\right\| \zeta + \frac{h_k^2 \left\|\nabla l\left(\theta_k\right) - \nabla g\left(\psi; f\left(\theta_k; x\right)\right)\right\|^2}{2} \int_0^1 \left(L_0 + L_1 \left\|\nabla l(\gamma(t))\right\|\right) dt.$$

Upon simplification using arguments of Appendix D.4 and applying Minkowski's inequality,

$$l\left(\theta_{k+1}\right) \leq l\left(\theta_k\right) - h_k \left\|\nabla l\left(\theta_k\right)\right\|^2 - h_k \left\|\nabla l\left(\theta_k\right)\right\| \zeta + \frac{h_k^2 \left(\left\|\nabla l\left(\theta_k\right)\right\|^2 + \left\|\nabla g\left(\psi; f\left(\theta_k; x\right)\right)\right\|^2\right)}{2} \left(L_0 + L_1 \lambda M\right).$$

Using $h_k = \frac{1}{L_0 + L_1 L^2 \beta \epsilon}$, we get

$$\begin{aligned} l\left(\theta_{k+1}\right) &\leq l\left(\theta_k\right) - \frac{h_k \left\|\nabla l\left(\theta_k\right)\right\|^2}{2} - h_k \left\|\nabla l\left(\theta_k\right)\right\| \zeta + \frac{h_k \left\|\nabla g\left(\psi; f\left(\theta_k; x\right)\right)\right\|^2}{2} \\ &\leq l\left(\theta_k\right) - \frac{h_k \left\|\nabla l\left(\theta_k\right)\right\|^2}{2} - h_k \left\|\nabla l\left(\theta_k\right)\right\| \zeta + \frac{h_k L^2 \delta^2}{2}, \text{ (from } \textbf{Lemma 2}\text{)}. \end{aligned}$$

Assuming T iterations to reach ϵ-stationary point, i.e., $\epsilon \leq \left\|\nabla l\left(\theta_k\right)\right\|$ for $k \leq T$. By a telescopic sum over k,

$$\begin{aligned} \sum_{k=0}^{T-1} l\left(\theta_{k+1}\right) - l\left(\theta_k\right) &\leq \frac{-T\left(\epsilon^2 + 2\epsilon\zeta - L^2\delta^2\right)}{2\left(L_0 + L_1 L^2 \beta \epsilon\right)} \\ \implies T &\leq \frac{2\left(l(\theta_0) - l^*\right)\left(L_0 + L_1 L^2 \beta \epsilon\right)}{\epsilon^2 + 2\epsilon\zeta - L^2\delta^2}. \end{aligned}$$

Therefore, we obtain

$$\sup_{\theta_0 \in \left\{\mathbb{R}^{h \times d_x}, \mathbb{R}^{d_y \times h}\right\}, l \in \mathscr{L}} \mathcal{T}_\epsilon\left(A_h\left[l, \theta_0\right], l\right) = \mathcal{O}\left(\frac{\left(l(\theta_0) - l^*\right)\left(L_0 + L_1 \lambda M\right)}{\epsilon^2 + 2\epsilon\zeta - \delta^2 M^2}\right)$$

which finishes the proof. □

D.7 Proof of Corollary 2

Using the arguments made in the proof of **Theorem 3** and first-order Taylor's approximation, we get

$$\begin{aligned} l\left(\theta_{k+1}\right) &= l\left(\theta_k\right) - h_k \left\|\nabla l\left(\theta_k\right)\right\|^2 - h_k \left\|\nabla l\left(\theta_k\right)\right\| \left\|\nabla g\left(\psi; f\left(\theta_k; x\right)\right)\right\| \\ &\leq l\left(\theta_k\right) - h_k \epsilon^2 - h_k \epsilon \zeta. \end{aligned}$$

By telescopic sum,

$$\begin{aligned} \sum_{k=0}^{T-1} l\left(\theta_{k+1}\right) - l\left(\theta_k\right) &\leq -T h_k \epsilon^2 - T h_k \epsilon \zeta \\ \implies T &\leq \frac{\left(l\left(\theta_0\right) - l^*\right)}{h_k \epsilon^2 + h_k \epsilon \zeta}. \end{aligned}$$

Therefore,

$$\sup_{\theta_0 \in \left\{\mathbb{R}^{h \times d_x}, \mathbb{R}^{d_y \times h}\right\}, l \in \mathscr{L}} \mathcal{T}_\epsilon\left(A_h\left[l, \theta_0\right], l\right) = \mathcal{O}\left(\frac{\left(l\left(\theta_0\right) - l^*\right)}{h\epsilon^2 + h\epsilon\zeta}\right)$$

which finishes the proof. □

D.8 Proof of Theorem 4

In sole supervision, the parameters are updated by $\frac{d\theta(t)}{dt} = -\nabla l(\theta(t))$. We define distance to optimal solution as $r^2(t) = \frac{1}{2}\|\theta(t) - \theta^*\|^2$. Now differentiating both sides, we get

$$\begin{aligned}\frac{dr^2(t)}{dt} &= \left\langle \frac{d\theta(t)}{dt}, \theta(t) - \theta^* \right\rangle \\ &= \langle -\nabla l(\theta(t)), \theta(t) - \theta^* \rangle .\end{aligned}$$

Using convexity and integrating over all iterates in a trajectory of T time steps,

$$\begin{aligned}\frac{1}{T}\int_0^T \frac{dr^2(t)}{dt}dt &\le \frac{1}{T}\int_0^T -\kappa(t)dt \\ \implies \frac{1}{T}\left(r^2(T) - r^2(0)\right) &\le -\frac{1}{T}\int_0^T \kappa(t)dt \\ \implies \frac{1}{T}\int_0^T \kappa(\theta(t))dt &\le \frac{r^2(0)}{T}.\end{aligned}$$

By Jensen's inequality,

$$\kappa\left(\frac{1}{T}\int_0^T \theta(t)dt\right) \le \frac{1}{T}\int_0^T \kappa(\theta(t))dt.$$

Therefore, $\kappa\left(\frac{1}{T}\int_0^T \theta(t)dt\right) = \mathcal{O}\left(\frac{\|\theta(0)-\theta^*\|^2}{2T}\right)$ which finishes the proof. □

D.9 Proof of Theorem 5

In supervised learning with adversarial regularization, the parameters are updated by $\frac{d\theta(t)}{dt} = -\nabla l(\theta(t)) + \nabla g(\theta(t))$. Using arguments of Appendix D.8, we obtain

$$\frac{dr^2(t)}{dt} = \langle -\nabla l(\theta(t)), \theta(t) - \theta^* \rangle + \langle \nabla g(\theta(t)), \theta(t) - \theta^* \rangle .$$

Since $l(.)$ is a convex downward and $g(.)$ is a convex upward function, we get

$$\begin{aligned}\frac{1}{T}\int_0^T \frac{dr^2(t)}{dt}dt &\le -\frac{1}{T}\int_0^T \kappa(t)dt - \frac{1}{T}\int_0^T \pi(t)dt \\ \implies \frac{1}{T}\left(r^2(T) - r^2(0)\right) &\le -\frac{1}{T}\int_0^T \kappa(t)dt - \frac{1}{T}\int_0^T \pi(t)dt \\ \implies \frac{1}{T}\int_0^T \kappa(\theta(t))dt &\le \frac{r^2(0)}{T} - \frac{1}{T}\int_0^T \pi(\theta(t))dt.\end{aligned}$$

Now, using Jensen's inequality on both $\kappa(.)$ and $\pi(.)$

$$\kappa\left(\frac{1}{T}\int_0^T \theta(t)dt\right) = \mathcal{O}\left(\frac{\|\theta(0) - \theta^*\|^2}{2T} - \pi\left(\frac{1}{T}\int_0^T \theta(t)dt\right)\right)$$

which finishes the proof. □

D.10 Proof of Theorem 6

For simplicity, let us denote the bias $b_k = \mathbb{E}\left[\hat{\mathfrak{g}}_k\right] - \nabla\mathfrak{l}(\theta_k)$.

$$\begin{aligned}\|\theta_k - \theta^*\|^2 &= \|\theta_{k-1} - \eta_k\hat{\mathfrak{g}}_{k-1} - \theta^*\|^2 \\ &= \|\theta_{k-1} - \theta^*\|^2 - 2\eta_k\langle\theta_{k-1} - \theta^*, \hat{\mathfrak{g}}_{k-1}\rangle + \eta_k^2\|\hat{\mathfrak{g}}_{k-1}\|^2 \\ &= \|\theta_{k-1} - \theta^*\|^2 - 2\eta_k\langle\theta_{k-1} - \theta^*, \nabla\mathfrak{l}(\theta_{k-1})\rangle - 2\eta_k\langle\theta_{k-1} - \theta^*, b_{k-1}\rangle + \eta_k^2\|\hat{\mathfrak{g}}_{k-1}\|^2 \\ &\le \|\theta_{k-1} - \theta^*\|^2 - 2\eta_k\langle\theta_{k-1} - \theta^*, \nabla\mathfrak{l}(\theta_{k-1})\rangle + \underbrace{2\eta_k\|\theta_{k-1} - \theta^*\|\|b_{k-1}\|}_{\text{By Cauchy-Schwarz inequality}} + \eta_k^2\|\hat{\mathfrak{g}}_{k-1}\|^2 \\ &\le \|\theta_{k-1} - \theta^*\|^2 - 2\eta_k\langle\theta_{k-1} - \theta^*, \nabla\mathfrak{l}(\theta_{k-1})\rangle + \underbrace{\eta_k\left(\|\theta_{k-1} - \theta^*\|^2 + \|b_{k-1}\|^2\right)}_{\text{By AM-GM inequality}} + \eta_k^2\|\hat{\mathfrak{g}}_{k-1}\|^2\end{aligned}$$

By μ-strong convexity, it is required that there exist positive constants μ such that for all (x, y), $\mathfrak{l}(y) \geq \mathfrak{l}(x) + \langle y - x, \nabla \mathfrak{l}(x) \rangle + \frac{\mu}{2} \|y - x\|^2$. Using strong-convexity at θ_{k-1} and θ^*, we get

$$\begin{aligned}\|\theta_k - \theta^*\|^2 &\leq \|\theta_{k-1} - \theta^*\|^2 - 2\eta_k \left(\mathfrak{l}(\theta_{k-1}) - \mathfrak{l}(\theta^*)\right) - \eta_k \mu \|\theta_{k-1} - \theta^*\|^2 + \eta_k \left(\|\theta_{k-1} - \theta^*\|^2 + \|b_{k-1}\|^2\right) + \eta_k^2 \|\hat{\mathfrak{g}}_{k-1}\|^2 \\ &\leq \|\theta_{k-1} - \theta^*\|^2 (1 - \eta_k \mu + \eta_k) - 2\eta_k \left(\mathfrak{l}(\theta_{k-1}) - \mathfrak{l}(\theta^*)\right) + \eta_k \|b_{k-1}\|^2 + \eta_k^2 \|\hat{\mathfrak{g}}_{k-1}\|^2 .\end{aligned}$$

Lemma 3. *Suppose **Assumption 7** holds for any $\mathfrak{g}(\theta)$ and $\alpha \in (1, 2]$. With global clipping parameter $\tau \geq 0$, the variance and bias of the estimator $\hat{\mathfrak{g}}$ are upper bounded as:*

$$\mathbb{E}\left[\|\hat{\mathfrak{g}}(\theta)\|^2\right] \leq G^\alpha \tau^{2-\alpha} \textit{and} \, \|\mathbb{E}\left[\hat{\mathfrak{g}}(\theta)\right] - \nabla l(\theta) + \nabla g(\theta)\|^2 \leq G^{2\alpha} \tau^{2-2\alpha}.$$

One can easily prove this using **Lemma 2** of [36]. Upon rearranging, taking expectation of both sides, and using **Lemma 3**,

$$\mathbb{E}\left[\mathfrak{l}(\theta_{k-1})\right] - \mathfrak{l}(\theta^*) \leq \mathbb{E}\left[\left(\frac{\eta_k^{-1} - \mu + 1}{2}\right) \|\theta_{k-1} - \theta^*\|^2 - \frac{\eta_k^{-1}}{2} \|\theta_k - \theta^*\|^2\right] + \frac{1}{2} G^{2\alpha} \tau^{2-2\alpha} + \frac{\eta_k}{2} G^\alpha \tau^{2-\alpha}.$$

Let us choose $\frac{\eta_k^{-1} - \mu + 1}{2} = k - 1$ and $\frac{\eta_k^{-1}}{2} = k + 1$. After simplification, $\eta_k = \frac{5}{2\mu(k+1)}$. Now, substitute $\tau_k = G k^{\frac{1}{\alpha}} \mu^{\frac{1}{\alpha}}$, $\eta_k = \frac{5}{2\mu(k+1)}$ and multiply k both sides. Thus,

$$k\mathbb{E}\left[\mathfrak{l}(\theta_{k-1})\right] - k\mathfrak{l}(\theta^*) \leq \mathbb{E}\left[k(k-1)\|\theta_{k-1} - \theta^*\|^2 - k(k+1)\|\theta_k - \theta^*\|^2\right] + \frac{G^2 k^{\frac{2-\alpha}{\alpha}} \mu^{\frac{2-2\alpha}{\alpha}}}{2} \left[\frac{5}{2}\left(\frac{k}{k+1}\right) + 1\right].$$

Since $\frac{k}{k+1} < 1$ for $k = 1, \ldots, T$, we get

$$k\mathbb{E}\left[\mathfrak{l}(\theta_{k-1})\right] - k\mathfrak{l}(\theta^*) \leq \mathbb{E}\left[k(k-1)\|\theta_{k-1} - \theta^*\|^2 - k(k+1)\|\theta_k - \theta^*\|^2\right] + \frac{7G^2 k^{\frac{2-\alpha}{\alpha}} \mu^{\frac{2-2\alpha}{\alpha}}}{4}.$$

Taking telescopic sum over $k = 1, \ldots, T$, we obtain

$$\sum_{k=1}^{T} k\mathbb{E}\left[\mathfrak{l}(\theta_{k-1})\right] - \mathfrak{l}(\theta^*) \sum_{k=1}^{T} k \leq \mathbb{E}\left[-T(T+1)\|\theta_T - \theta^*\|^2\right] + \frac{7G^2 \mu^{\frac{2-2\alpha}{\alpha}}}{4} \sum_{k=1}^{T} k^{\frac{2-\alpha}{\alpha}}.$$

Using $\sum_{k=1}^{T} k^{\frac{2-\alpha}{\alpha}} \leq \int_0^{T+1} k^{\frac{2-\alpha}{\alpha}} dk \leq (T+1)^{\frac{2}{\alpha}}$,

$$\sum_{k=1}^{T} k\mathbb{E}\left[\mathfrak{l}(\theta_{k-1})\right] - \mathfrak{l}(\theta^*) \frac{T(T+1)}{2} \leq \frac{7G^2 \mu^{\frac{2-2\alpha}{\alpha}}}{4} (T+1)^{\frac{2}{\alpha}}.$$

Now, dividing both sides by $\frac{T(T+1)}{2}$ and using $T^{-1} \leq 2(T+1)^{-1}$ for $T \geq 1$,

$$\frac{\sum_{k=1}^{T} k\mathbb{E}\left[\mathfrak{l}(\theta_{k-1})\right]}{\sum_{k=1}^{T} k} - \mathfrak{l}(\theta^*) \leq 7G^2 \mu^{\frac{2-2\alpha}{\alpha}} (T+1)^{\frac{2-2\alpha}{\alpha}}.$$

By Jensen's inequality,

$$\mathbb{E}\left[\mathfrak{l}\left(\frac{\sum_{k=1}^{T} k\theta_{k-1}}{\sum_{k=1}^{T} k}\right)\right] - \mathfrak{l}(\theta^*) \leq \mathcal{O}\left(G^2 \left(\mu(T+1)\right)^{\frac{2-2\alpha}{\alpha}}\right)$$

Substituting $\mathfrak{l}(\theta) = l(\theta) - g(\theta)$, we get

$$\mathbb{E}\left[l\left(\bar{\theta}\right)\right] - l(\theta^*) \leq \mathcal{O}\left(G^2 \left(\mu(T+1)\right)^{\frac{2-2\alpha}{\alpha}} - \left(g\left(\theta^*\right) - \mathbb{E}\left[g\left(\bar{\theta}\right)\right]\right)\right),$$

which finishes the proof. $\square$

D.11 Proof of Theorem 7

The notations of $\mathfrak{l}$ and b_k follow from Appendix D.10. Using L-smooth property of $\mathfrak{l}$, we get

$$\begin{aligned}
\mathfrak{l}\left(\theta_k\right) &\leq \mathfrak{l}\left(\theta_{k-1}\right)+\left\langle\nabla \mathfrak{l}\left(\theta_{k-1}\right), \theta_k-\theta_{k-1}\right\rangle+\frac{L}{2}\left\|\theta_k-\theta_{k-1}\right\|^2 \\
&\leq \mathfrak{l}\left(\theta_{k-1}\right)+\left\langle\nabla \mathfrak{l}\left(\theta_{k-1}\right), -\eta_k \hat{\mathfrak{g}}_{k-1}\right\rangle+\frac{\eta_k^2 L}{2}\left\|\hat{\mathfrak{g}}_{k-1}\right\|^2 \\
&\leq \mathfrak{l}\left(\theta_{k-1}\right)-\eta_k |\left\|\nabla \mathfrak{l}(\theta_{k-1})\right\|^2-\eta_k\left\langle\nabla \mathfrak{l}(\theta_{k-1}), b_{k-1}\right\rangle+\frac{\eta_k^2 L}{2}\left\|\hat{\mathfrak{g}}_{k-1}\right\|^2 \\
&\leq \mathfrak{l}\left(\theta_{k-1}\right)-\eta_k |\left\|\nabla \mathfrak{l}(\theta_{k-1})\right\|^2+\underbrace{\eta_k\left\|\nabla \mathfrak{l}(\theta_{k-1})\right\|\left\|b_{k-1}\right\|}_{\text{By Cauchy-Schwarz inequality}}+\frac{\eta_k^2 L}{2}\left\|\hat{\mathfrak{g}}_{k-1}\right\|^2 \\
&\leq \mathfrak{l}\left(\theta_{k-1}\right)-\eta_k |\left\|\nabla \mathfrak{l}(\theta_{k-1})\right\|^2+\underbrace{\frac{\eta_k}{2}\left(\left\|\nabla \mathfrak{l}(\theta_{k-1})\right\|^2+\left\|\, b_{k-1}\right\|^2\right)}_{\text{By AM-GM inequality}}+\frac{\eta_k^2 L}{2}\left\|\hat{\mathfrak{g}}_{k-1}\right\|^2
\end{aligned}$$

Taking expectation of both sides,

$$\mathbb{E}\left[\mathfrak{l}(\theta_k)-\mathfrak{l}(\theta_{k-1})\right] \leq \mathbb{E}\left[\frac{-\eta_k}{2}\left\|\nabla \mathfrak{l}(\theta_{k-1})\right\|^2\right]+\frac{\eta_k}{2} G^{2\alpha} \tau^{2-2\alpha}+\frac{\eta_k^2 L}{2} G^{\alpha} \tau^{2-\alpha}.$$

Upon rearranging and taking telescopic sum over $k=1, \ldots, T$, we obtain

$$\frac{1}{T} \sum_{k=1}^{T} \mathbb{E}\left[\left\|\nabla \mathfrak{l}(\theta_{k-1})\right\|^2\right] \leq \frac{2\eta_k^{-1}}{2}\left(\mathfrak{l}(\theta_0)-\mathfrak{l}(\theta^*)\right)+G^{2\alpha} \tau^{2-2\alpha}+\eta_k L G^{\alpha} \tau^{2-\alpha}.$$

By choosing $\tau=G\left(\eta_k L\right)^{\frac{-1}{\alpha}}$,

$$\frac{1}{T} \sum_{k=1}^{T} \mathbb{E}\left[\left\|\nabla \mathfrak{l}(\theta_{k-1})\right\|^2\right] \leq \frac{2\eta_k^{-1} R_0}{T}+2G^2\left(\eta_k L\right)^{\frac{2\alpha-2}{\alpha}}.$$

Let us choose $\eta_k=\left(\frac{R_0^{\alpha} L^{2-2\alpha}}{G^2 T^{\alpha}}\right)^{\frac{1}{3\alpha-2}}$. Thus,

$$\frac{1}{T} \sum_{k=1}^{T} \mathbb{E}\left[\left\|\nabla \mathfrak{l}(\theta_{k-1})\right\|^2\right] \leq 4G^{\frac{2\alpha}{3\alpha-2}}\left(\frac{R_0 L}{T}\right)^{\frac{2\alpha-2}{3\alpha-2}}$$

Now, substituting $\mathfrak{l}(\theta)=l(\theta)-g(\theta)$, we get

$$\frac{1}{T} \sum_{k=1}^{T} \mathbb{E}\left[\left\|\nabla l(\theta_{k-1})\right\|^2+\left\|\nabla g(\theta_{k-1})\right\|^2-2\langle\nabla l(\theta_{k-1}), \nabla g(\theta_{k-1})\rangle\right] \leq 4G^{\frac{2\alpha}{3\alpha-2}}\left(\frac{R_0 L}{T}\right)^{\frac{2\alpha-2}{3\alpha-2}}.$$

Since the gradients received from $l(\theta)$ and $g(\theta)$ are negatively correlated at any instant during the optimization process, the above expression simplifies to

$$\frac{1}{T} \sum_{k=1}^{T} \mathbb{E}\left[\left\|\nabla l(\theta_{k-1})\right\|^2+\left\|\nabla g(\theta_{k-1})\right\|^2+2\left\|\nabla l(\theta_{k-1})\right\|\left\|\nabla g(\theta_{k-1})\right\|\right] \leq 4G^{\frac{2\alpha}{3\alpha-2}}\left(\frac{R_0 L}{T}\right)^{\frac{2\alpha-2}{3\alpha-2}}.$$

Therefore,

$$\frac{1}{T} \sum_{k=1}^{T} \mathbb{E}\left[\left\|\nabla l(\theta_{k-1})\right\|^2\right]+\frac{1}{T} \sum_{k=1}^{T} \mathbb{E}\left[\left\|\nabla g(\theta_{k-1})\right\|^2\right] \leq 4G^{\frac{2\alpha}{3\alpha-2}}\left(\frac{R_0 L}{T}\right)^{\frac{2\alpha-2}{3\alpha-2}}.$$

Upon simplification,

$$\frac{1}{T} \sum_{k=1}^{T} \mathbb{E}\left[\left\|\nabla l(\theta_{k-1})\right\|^2\right] \leq \mathcal{O}\left(G^{\frac{2\alpha}{3\alpha-2}}\left(\frac{R_0 L}{T}\right)^{\frac{2\alpha-2}{3\alpha-2}}-\frac{1}{T} \sum_{k=1}^{T} \mathbb{E}\left[\left\|\nabla g(\theta_{k-1})\right\|^2\right]\right)$$

which finishes the proof. □