Understanding the Role of Adversarial Regularization in Supervised Learning

Litu Rout*
Space Applications Centre
Indian Space Research Organization
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 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, gradient flow and provable convergence 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. Furthermore, we observe an intriguing phenomenon in the neural embedded vector space while contrasting adversarial learning with sole supervision.

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.

Over the years several variants of gradient descent algorithms have emerged. In various tasks, adaptive methods including Adagrad (Duchi, Hazan, and Singer 2011), Adadelta (Zeiler 2012), RMSProp (Tieleman and Hinton 2012), ADAM (Kingma and Ba 2014), and NADAM (Dozat 2016) perform relatively 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 (Wilson et al. 2017), and theoretical guarantees (Carmon et al. 2018) often provide an edge over other contemporary learning algorithms in several tasks. For this reason, we choose SGD as our primary learning algorithm to foster smooth transition from recent analyses (Nagarajan and Kolter 2017; Neyshabur et al. 2019). We argue that despite superior performance, it suffers from vanishing gradient issue in near optimal region. In fact, this is mirrored by poor practical performance when compared with adversarial regularization as independently reported in copious literature (Denton et al. 2015; Wang and Gupta 2016; Ledig et al. 2017; Rangnekar et al. 2017; Wang et al. 2018; Xue et al. 2018; Xian et al. 2018). 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 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 (Isola et al. 2017; Zhu et al. 2017; Park et al. 2019; Karras, Laine, and Aila 2019). 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 (Mirza and Osindero 2014). 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 (Song et al. 2018). In photo-realistic image super resolution, Euclidean or supremum norm is used to minimize the distance between reconstructed and original image (Ledig et al. 2017; Wang et al. 2018). In medical image segmentation, multi-scale L_1-loss with adversarial regularization is shown to outperform sole supervision (Xue et al. 2018). In medical image analysis, a 3d conditional GAN along with L_1-distance is used to super resolve CT scan imagery (Kudo et al. 2019).

Furthermore, Isola et al. 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 that would otherwise require hand-engineered loss. Xian et al. 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 (Mirza and Osindero 2014; Denton et al. 2015; Wang and Gupta 2016). Karacan et al. use this technique to efficiently generate images of outdoor scenes. Rout et al.

*Under Review

combine spatial and Laplacian spectral channel attention in regularized adversarial learning to synthesize high resolution images. Emami et al. coalesce spatial attention with adversarial regularization and feature map loss to perform image-to-image translation.

As per these prior and concurrent works (Rangnekar et al. 2017; Xue et al. 2018; Rout 2020; Dong et al. 2015; Henaff, Canziani, and LeCun 2019; Sarmad, Lee, and Kim 2019), it is understandable that supervised learning with adversarial regularization boosts empirical performance. More importantly, this behavior is consistent across a wide variety of problems and network configurations. As much beneficial as this has been so far, to our knowledge, the theoretical understanding still remains relatively less explored. Aiming to bridge this gap, we provide theoretical and empirical evidence of better performance due to adversarial regularization when compared with sole supervision.

2 Related Works

Adversarial Regularization The spectral and spatial super resolution based on adversarial regularization (Rangnekar et al. 2017; Rout 2020) is proven to achieve *faster convergence* and *better empirical risk* compared to purely supervised learning (Lanaras et al. 2018). Further, Ledig et al. showed improvement in perceptual quality of high resolution images in adversarial setting. Despite superior empirical performance, the theoretical understanding of such phenomena remains elusive. To this end, the theoretical analysis suggests that there is a constant that bounds the total empirical risk above (Xue et al. 2018). As a result, this inhibits erroneous gradient estimation by the discriminator that apparently improves perceptual quality. However, these benign properties of loss surface do not fully explain this phenomenon at a fundamental level. The present account in this paper is intended to provide further insights to this problem.

Apart from supervised and adversarial learning, the notion of adversarial regularization has also been studied in Reinforcement Learning (RL). Henaff, Canziani, and LeCun use adversarial learning with expert regularization to learn a predictive policy that allows to drive in simulated dense traffic. Sarmad, Lee, and Kim use RL agent controlled GAN and L_2-distance between global feature vectors to convert noisy, partial point cloud into high-fidelity data.

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 (Schmidt, Le Roux, and Bach 2017; Zhou, Xu, and Gu 2018). Among momentum based acceleration, much theoretical progress has been made to accelerate any smooth convex optimization (Nesterov 2012; Carmon et al. 2018). Further, many efforts have been made towards changing the step size across iterations based on estimated gradient norm (Duchi, Hazan, and Singer 2011; Staib et al. 2019; Zhou et al. 2018). Adversarial regularization is similar to these methods in a sense that it offers acceleration in the near optimal region.

Minimax Optimization The seminal work of Neumann in solving the problem of minimax optimization has been a central part of game theory. Recently, a rapid increase in interest is seen to study the intrinsic properties of minimax problems. The increasing popularity owes in part to the discovery of generative adversarial networks (Goodfellow et al. 2014). In this paper, to focus more on the empirical success of adversarial regularization, we study a simple minimax optimization problem. However, we wish to allude some interesting line of work by Lin, Jin, and Jordan; Lin et al.; Jin, Netrapalli, and Jordan; Mertikopoulos, Papadimitriou, and Piliouras in this direction that may encourage further investigation from algorithmic point of view. It will certainly be useful to borrow some ideas from the vast literature of minimax optimization under less restrictive setting. Though it is beyond the scope of this discussion, the definition of local optimality by Jin, Netrapalli, and Jordan is likely to pave the way for better understanding of minimax optimization, and consequently adversarial regularization.

3 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 $x \in X$, $f(\theta; x) : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$ is a neural network with rectified linear unit (ReLU) activation, common for both supervised and adversarial learning. Here, θ denotes the trainable parameters of the generator, $f(\theta; .)$. On the other hand, the discriminator, $g(\psi; .)$ has trainable parameters collected by ψ. The optimal values of these parameters are represented by θ^* and 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 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. (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,

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

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

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

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

4 Theoretical Analysis

This section states the assumptions and their justifications in the context of adversarial regularization. The theoretical findings are intended to provide convincing reasons to 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 study. 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 adversarial regularization helps tighten the existing bounds in supervised learning literature. Thus, the emphasis is on providing a theoretical justification to the practial success of supervised learning with adversarial regularization.

Warm-Up: Mitigating Vanishing Gradient in Near Optimal Region

The primary assumptions are stated as following.

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.

Lemma 1. *Let **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. This is a crucial result. So we sketch the proof as following. 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

$$
\begin{aligned}
\left\|\nabla_p l\left(p; y\right)\right\| - \left\|\nabla_q l\left(q; y\right)\right\| &\leq \left\|\nabla_p l\left(p; y\right) - \nabla_q l\left(q; y\right)\right\| \\
&\leq \beta \left\|p - q\right\| \\
&\leq \beta L \left\|\theta - \theta^*\right\|.
\end{aligned}
$$

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

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

Upon substituting optimality condition, i.e., $\|\nabla_q l\left(q; y\right)\| = 0$, 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. □

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 having to loose 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. Using similar arguments from **Lemma 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, $\|-\nabla_\theta \mathbb{E}_{x\sim\mathcal{P}_X} \left[g\left(\psi; f\left(\theta; x\right)\right)\right]\| \leq L\delta$, which finishes the proof. □

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 (Xue et al. 2018).

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. By applying triangle inequality after simplification,

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

which completes the statement of the theorem. □

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 (Lin, Jin, and Jordan 2019; Lin et al. 2020; Jin, Netrapalli, and Jordan 2019; Mertikopoulos, Papadimitriou, and Piliouras 2018). 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.

Main Results: Asymptotic Iteration Complexity

In this section, we analyze global iteration complexity of sole supervision and adversarial regularization (Zhang et al. 2019a; Carmon et al. 2019). 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

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

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

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

We derive the asymptotic bounds under a less restrictive setting as introduced in (Zhang et al. 2019a). The new condition is weaker than commonly used Lipschitz smoothness assumption. Under this condition, the authors of (Zhang et al. 2019a) 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 (Zhang et al. 2019a). 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* $\mathcal{O}\left(\frac{\left(l(\theta_0)-l^*\right)\left(L_0+L_1L^2\beta\epsilon\right)}{\epsilon^2}\right)$.

Proof. Refer to Appendix C.

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 C.

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 the discriminator to provide useful gradients until convergence. It is a valid assumption in convex-concave 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 sole purspose of maintaining consistency with existing literature. 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* $\mathcal{O}\left(\frac{\left(l(\theta_0)-l^*\right)\left(L_0+L_1L^2\beta\epsilon\right)}{\epsilon^2+2\epsilon\zeta-L^2\delta^2}\right)$.

Proof. Refer to Appendix C.

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 C. 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 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 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.

Main Results: Sub-Optimality Gap

Here, we analyze the continuous time gradient flow in both approaches. The sub-optimality gap of generator and discriminator are defined by $\kappa(t) = \kappa(\theta(t)) \coloneqq l\left(\theta(t)\right) - l\left(\theta^*\right)$ and $\pi(t) = \pi(\theta(t)) \coloneqq g\left(\theta^*\right) - g\left(\theta(t)\right)$, 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 C.

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 C.

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. It is worth mentioning that the sub-optimality gap in adversarial regularization is at least as good as sole supervision which justifies the emprical gain in practice. 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 empirical risk benefits in adversarial regularization.

Main Results: Provable Convergence

This section covers 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 (Lacoste-Julien, Schmidt, and Bach 2012). 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. One may wish to analyze this for coordinate-wise clipping (Zhang et al. 2019b).

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 zero momentum. 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 C.

Observe that by eliminating adversary and setting $\alpha=2$, we recover exactly the SGD rate, i.e., $\mathcal{O}\left(\frac{G^2}{\mu T}\right)$ (Lacoste-Julien, Schmidt, and Bach 2012). Thus, adversarial regularization converges in strongly-convex-strongly-concave setting. 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 zero momentum. Given constant clipping $\tau_k=G\left(\eta_k L\right)^{\frac{-1}{\alpha}}$ and constant step size $\eta_k=\left(\frac{R_0^\alpha L^{2-2\alpha}}{G^2T^\alpha}\right)^{\frac{1}{3\alpha-2}}$, where $R_0=l(\theta_0)-l(\theta^*)$, the sequence $\{\theta_k\}$ satisfies*

$$\frac{1}{T}\sum_{k=1}^T\mathbb{E}\left[\|\nabla l\left(\theta_{k-1}\right)\|^2\right]\leq\mathcal{O}\left(G^{\frac{2\alpha}{3\alpha-2}}\left(\frac{R_0L}{T}\right)^{\frac{2\alpha-2}{3\alpha-2}}-\frac{1}{T}\sum_{k=1}^T\mathbb{E}\left[\|\nabla g(\theta_{k-1})\|^2\right]\right).$$

Proof. Refer to Appendix C.

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 that 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 guarantees provide more insights to our understanding of adversarial training in practice. While this paper studies minimax optimization under nonconvex-smooth settings, it will be interesting to derive convergence guarantees under nonconvex-nonsmooth setting.

Main Results: Generalization Error

Motivated by the role of over-parametrization in generalization (Neyshabur et al. 2017; Nagarajan and Kolter 2017; Neyshabur et al. 2019), 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 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 $\mathcal{S}$.

$$\mathcal{R}_{\mathcal{S}}\left(\mathcal{H}\right)=\frac{1}{m}\mathbb{E}_{\xi_i\in\{\pm1\}^m}\left[\sup_{f\in\mathcal{H}}\sum_{i=1}^m\xi_i f(x_i)\right].$$

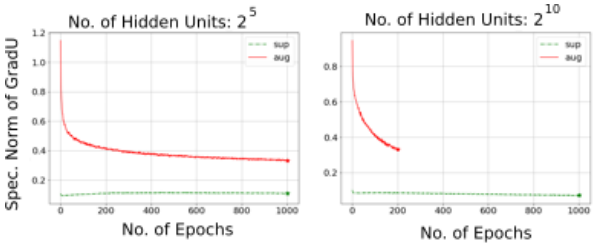

Figure 1: Comparison of gradients — supervised (sup) and augmented (aug) — in the *hidden layer* on MNIST. Adversarial regularization mitigates vanishing gradient issue.

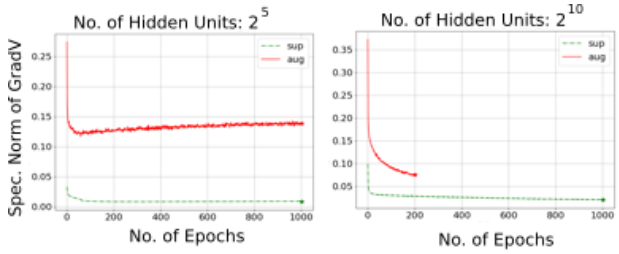

Figure 2: Comparison of gradients — supervised (sup) and augmented (aug) — in the *top layer* on MNIST. Adversarial regularization mitigates vanishing gradient issue.

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 Neyshabur et al. 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 Neyshabur et al., 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.

5 Experiments

Our experiments aim to answer the following questions[1].

[1]While the preliminary observations are reported in the main paper, additional experimental results are supplied in the appendix.

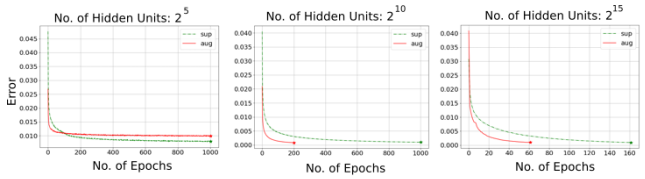

Figure 3: Comparison of optimal empirical risk on MNIST. Adversarial regularization converges faster.

- How does adversarial regularization mitigate vanishing gradients in the near optimal region?
- How does adversarial regularization accelerate training?
- How does adversarial regularization achieve tighter sub-optimality gap?
- How does adversarial regularization converge under practical settings?

Results on MNIST

Figure 1 and 2 provide 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 in purely supervised objective is smaller than adversarial learning. This is consistent with the theoretical analyses in Section 4. The main reason for such non-vanishing gradient is the feedback signal from discriminator. Further, the rate of convergence is at least as good as sole supervision, as marked by $\star$ in Figure 1 and 2.

Figure 3 offers experimental support to better empirical risk in adversarial setting. Here, we observe the significance of near optimal region, i.e., ϵ with 32 hidden units. Since the expressive power of such a network is very small in both approaches, evidently neither of those 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 optimum and thereby, it satisfies the assumptions of **Theorem 1**. It is to be noted that the tightness of the reported bounds is asserted in the near optimal region. This is evident from the stability of the Lipschitz constant L over iterations as shown in Figure 1 and 2. Under this circumstance, the optimal empirical risk in augmented objective can be provably better than sole supervision as predicted by the proposed theorems. Figure 3 supports this theory as augmented objective consistently achieves better performance either by risk or by rate of convergence for networks with sufficient expressive power.

Furthermore, we compare the optimal empirical risk and iteration complexity with different number of hidden units in Figure 4. To better interpret the theorems, one can infer from Figure 4 (a) that the value of ϵ in **Theorem 1** is approximately equal to 0.005. The number of epochs required to find a first order stationary point in adversarial learning is always less than or equal to supervised learning, which validates our theorems. The value of ϵ is more relevant to the present body of analysis as it is a major part of the inverse mapping in practical scenarios. Moreover, it is not hard to

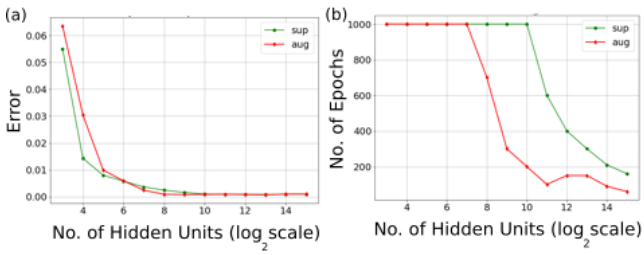

Figure 4: Comparison on MNIST. (a) Optimal empirical risk. (b) Iteration Complexity. Adversarial regularization attains tighter ϵ-stationary point at an optimal rate.

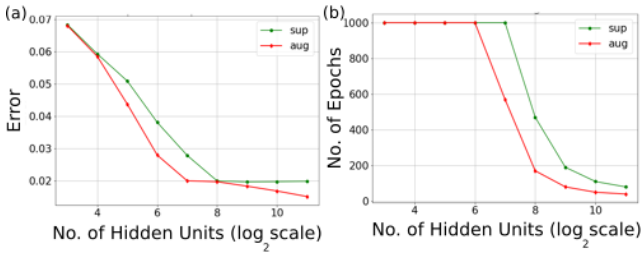

Figure 5: Comparison on CIFAR10. (a) Optimal empirical risk. (b) Iteration Complexity. Adversarial regularization attains tighter ϵ-stationary point at an optimal rate.

estimate δ in some rare occurences where the mapping function is approximated by the discriminator.

Results on CIFAR10

These theorems also justify the experiments conducted on CIFAR10 dataset. As shown in Figure 5, supervised learning with adversarial regularization performs better than sole supervision both in terms of optimal empirical risk and iteration complexity. Here, ϵ is approximately equal to 0.06.

Results on Generalization Error

The generalization trend in sole supervision is shown in Figure 6(a) and 6(c). As per equation (4), 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). The generalization bound, therefore, is expected to explain the generalization error in adversarial learning with expert regularization. However, as shown in Figure 6(b) and 6(d), this bound does not fully explain the generalization error observed in adversarial learning.

In Figure 7, 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.

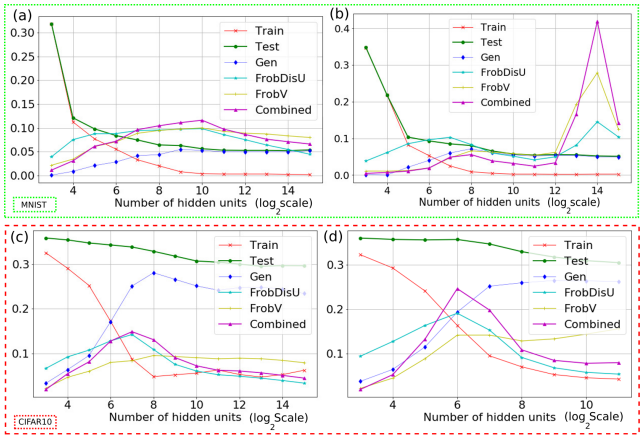

Figure 6: Generalization error on MNIST and CIFAR10. Adversarial training requires new generalization bound.

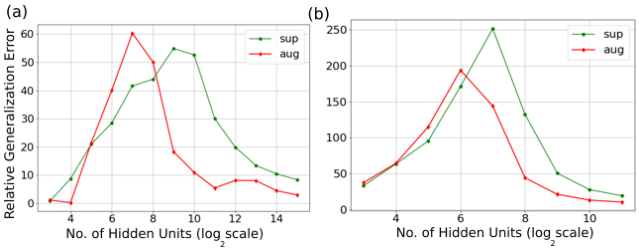

Figure 7: Relative generalization. (a) MNIST. (b) CIFAR10. Augmented objective has better relative generalization error.

6 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. From theoretical perspective, the iteration complexity, gradient flow, provable convergence guarantee, and the analysis of generalization error provided further insights to the empirical findings of adversarial regularization as independently reported in previous works.

While these theoretical analyses provided several key insights to better understand the practical success of adversarial regularization, it is far from being conclusive. Moreover, it paves the way for several open questions: (i) What explains the generalization behavior in adversarial learning? (ii) Does adversarial regularization improve sample complexity? In this paper, we do not explain generalization gap and sample complexity. Nevertheless, it will be interesting to understand the effect of implicit gradient estimation by an adversary on these theoretic puzzles.

7 Broader Impact

In this paper, we primarily focused on understanding the role of adversarial regularization in supervised learning. At a fundamental level, we provided a theoretical justification supported by empirical evidence to corroborate commonly observed phenomena in practice. We believe this work does not present any foreseable societal consequence.

References

Carmon, Y.; Duchi, J. C.; Hinder, O.; and Sidford, A. 2018. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization* 28(2): 1751–1772.

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

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

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2015. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence* 38(2): 295–307.

Dozat, T. 2016. Incorporating nesterov momentum into adam .

Duchi, J.; Hazan, E.; and Singer, Y. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* 12(Jul): 2121–2159.

Emami, H.; Aliabadi, M. M.; Dong, M.; and Chinnam, R. B. 2019. SPA-GAN: Spatial Attention GAN for Image-to-Image Translation. *arXiv preprint arXiv:1908.06616* .

Frey, B. J.; and Dueck, D. 2006. Mixture modeling by affinity propagation. In *Advances in neural information processing systems*, 379–386.

Frey, B. J.; and Dueck, D. 2007. Clustering by passing messages between data points. *science* 315(5814): 972–976.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

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

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

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

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

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4401–4410.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Kohonen, T. 1990. The self-organizing map. *Proceedings of the IEEE* 78(9): 1464–1480.

Kudo, A.; Kitamura, Y.; Li, Y.; Iizuka, S.; and Simo-Serra, E. 2019. Virtual thin slice: 3D conditional GAN-based Super-resolution for CT slice interval. In *International Workshop on Machine Learning for Medical Image Reconstruction*, 91–100. Springer.

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

Lanaras, C.; Bioucas-Dias, J.; Galliani, S.; Baltsavias, E.; and Schindler, K. 2018. Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS Journal of Photogrammetry and Remote Sensing* 146: 305–319.

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

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

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

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9(Nov): 2579–2605.

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

Mirza, M.; and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* .

Nagarajan, V.; and Kolter, J. Z. 2017. Generalization in Deep Networks: The Role of Distance from Initialization. In *Neural Information Processing Systems (NeurIPS) Workshop, Deep Learning: Bridging Theory and Practice*.

Nesterov, Y. 2012. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization* 22(2): 341–362.

Neumann, J. v. 1928. Zur theorie der gesellschaftsspiele. *Mathematische annalen* 100(1): 295–320.

Neyshabur, B.; Bhojanapalli, S.; McAllester, D.; and Srebro, N. 2017. Exploring generalization in deep learning. In

Advances in Neural Information Processing Systems, 5947–5956.

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

Park, T.; Liu, M.-Y.; Wang, T.-C.; and Zhu, J.-Y. 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2337–2346.

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

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

Rout, L.; Misra, I.; Moorthi, S. M.; and Dhar, D. 2020. 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*.

Sarmad, M.; Lee, H. J.; and Kim, Y. M. 2019. 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*, 5898–5907.

Schmidt, M.; Le Roux, N.; and Bach, F. 2017. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming* 162(1-2): 83–112.

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

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

Tieleman, T.; and Hinton, G. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning* 4(2): 26–31.

Turing, A. 1952. The Chemical Basis of Morphogenesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 237(641): 37–72.

Wang, X.; and Gupta, A. 2016. Generative image modeling using style and structure adversarial networks. In *European Conference on Computer Vision*, 318–335. Springer.

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

Wilson, A. C.; Roelofs, R.; Stern, M.; Srebro, N.; and Recht, B. 2017. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems*, 4148–4158.

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

Xue, Y.; Xu, T.; Zhang, H.; Long, L. R.; and Huang, X. 2018. Segan: Adversarial network with multi-scale L-1 loss for medical image segmentation. *Neuroinformatics* 16(3-4): 383–392.

Zeiler, M. D. 2012. ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701* .

Zhang, J.; He, T.; Sra, S.; and Jadbabaie, A. 2019a. Why Gradient Clipping Accelerates Training: A Theoretical Justification for Adaptivity. In *International Conference on Learning Representations*.

Zhang, J.; Karimireddy, S. P.; Veit, A.; Kim, S.; Reddi, S. J.; Kumar, S.; and Sra, S. 2019b. Why ADAM Beats SGD for Attention Models. *arXiv preprint arXiv:1912.03194* .

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

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

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

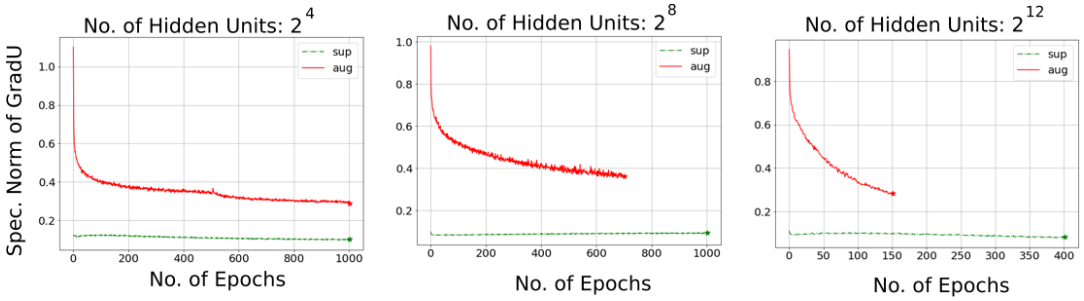

Figure 8: Comparison of gradient updates between supervised and augmented objective as observed in the *hidden layer* on MNIST.

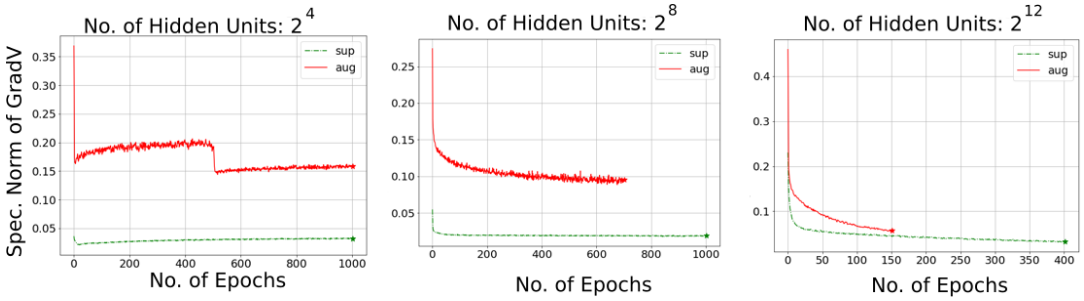

Figure 9: Comparison of gradient updates between supervised and augmented objective as observed in the *top layer* on MNIST.

Appendix

A More Experiments

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

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. Experiments are conducted on 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. The convergence criterion is set to be mean square error of 0.001 for MNIST and 0.02 for CIFAR10. We train on both datasets for a maximum of 1000 epochs, or until convergence. In these settings, 13 architectures with the number of hidden units (h) ranging from 2^3 to 2^{15} are trained on both datasets. All parameters are initialized from uniform distribution. The experiments are conducted on a Linux system with 64GB RAM and 2 x V100 gpus using PyTorch library.

Experimental Results

Results on MNIST As shown in Figure 8 and 9, 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 10 where we observe this particular phenomenon across a wide variety of architectures. One may argue that the difference in empirical risk is minimal. However, it is always better to discover a first order stationary point relatively faster without having to loose any risk benefits. From another perspective, the notion of multiple critical points in deep neural networks acts in favor of adversarial learning that allows faster convergence. It seems to us as an interesting line of future work.

Results on CIFAR10 Similar to MNIST, we also observe vanishing gradient issue on CIFAR10, which is shown in Figure 11 and 12. Figure 13 illustrates how model capacity correlates with empirical risk and thereby, satisfies the assumption of **Theorem 1**. Across a wide variety of architectures, observe 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 13, though both methods start with almost same initial empirical risk, augmented objective traverses through a shorter path and attains minimal risk upon convergence. The slight difference in error at the begining is mainly due to adversarial acceleration in the first step itself.

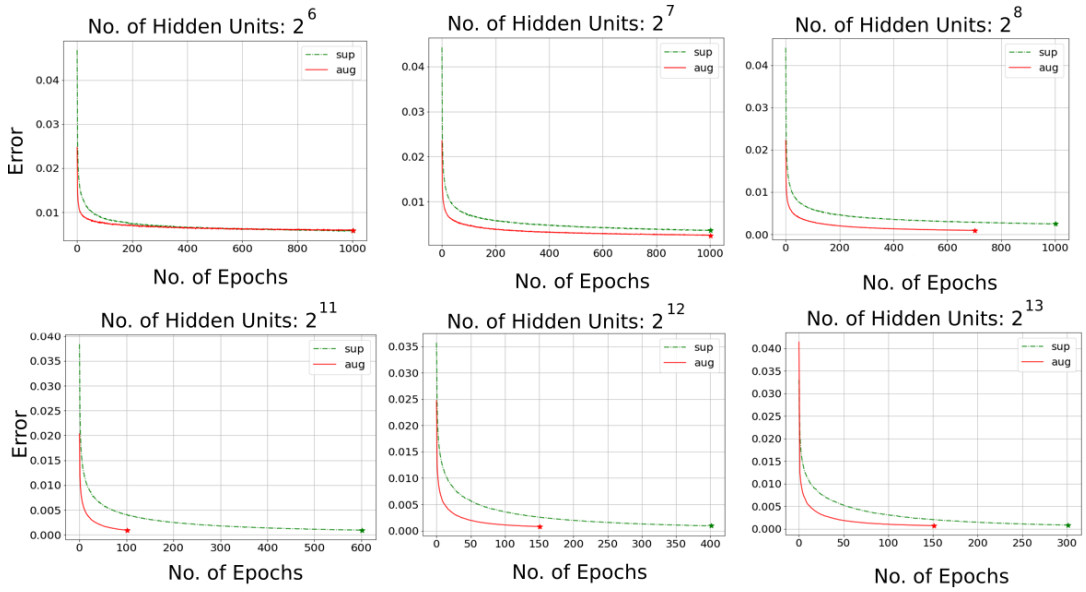

Figure 10: Comparison of optimal empirical risk on MNIST.

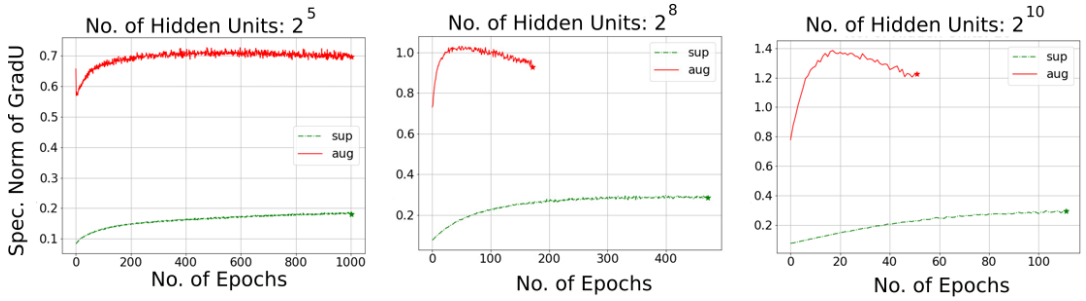

Figure 11: Comparison of gradient updates between supervised and augmented objective as observed in the *hidden layer* on CIFAR10.

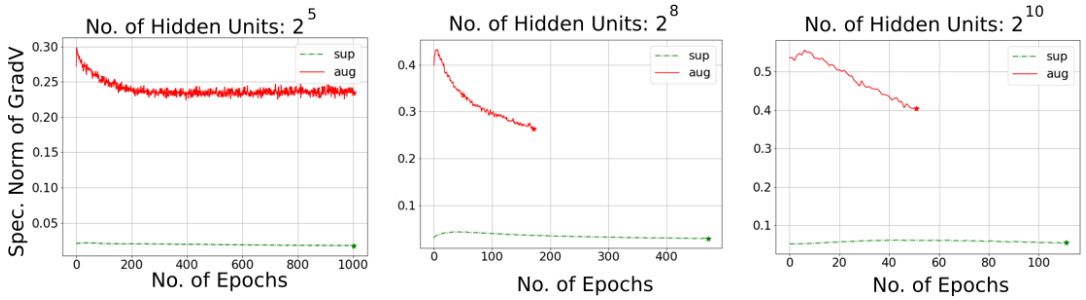

Figure 12: Comparison of gradient updates between supervised and augmented objective as observed in the *top layer* on CIFAR10.

Results on Various Networks To study the impact of these findings on more realistic scenarios, we experiment on various network configurations. As shown in Figure 14 and 15, the issue of vanishing gradient is persistent across these experimented configurations. Furthermore, the discussion on adversarial acceleration is also supported by Figure 16. 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

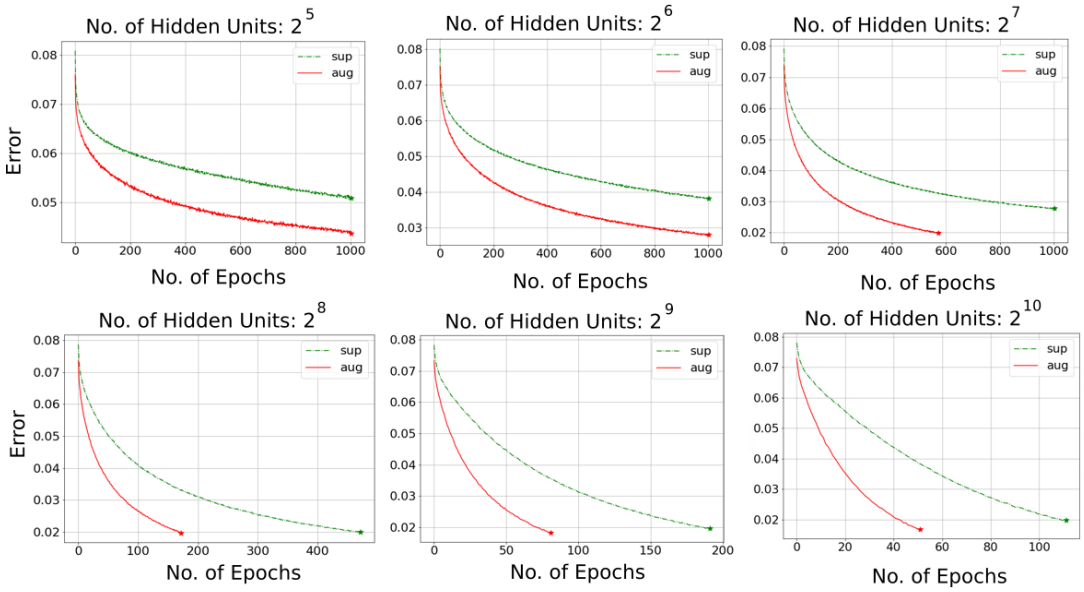

Figure 13: Comparison of optimal empirical risk on CIFAR10.

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	✓

fundamental level.

B Omitted Main Results: Neural Topology

Implementation Details

In neural topology, we analyze the geometry of neurons present in the hidden and the top layer. Here, three different architectures with 2^{13}, 2^{14} and 2^{15} hidden units are used to ensure sufficient expressive power. The core of our visualization is neural interaction which is modelled by Affinity Propagation (AP) (Frey and Dueck 2006, 2007). Since each model has large number of neurons in the hidden layer, we restrict our topological analysis to a fixed subset of 2048 neurons. Due to extreme time and space complexity in AP, we first reduce the dimension of neurons in the hidden layer from $\mathbb{R}^{d_x}$ (here, $d_x = 784$) to $\mathbb{R}^{10}$ using PCA and thereafter, to $\mathbb{R}^2$ using t-SNE (Maaten and Hinton 2008). In case of top layer, we directly apply t-SNE to map neurons in $\mathbb{R}^{d_y}$ to $\mathbb{R}^2$ (here, $d_y = 10$). Note that the absolute units of x and y axes are not important in these neural topology diagrams.

NTA on MNIST

In the experiments with 2^{14} hidden units, we observe emergence of evolutionary patterns in adversarial framework. As shown in Figure 17, even though both systems are initialized with similar topology in weight space, the final topology in regularized adversarial learning changes drastically. It is quite apparent from Figure 17(d), both in the hidden and the top layers, that adversarially learned weights lie on a different geometrical surface compared to sole supervision. Particularly intriguing is the self-organization tendency of these artificial neurons in a topological sense (Kohonen 1990). We observe this sparse self-organization behavior on a wide variety of architectures. In all these configurations, adversarial learning tries to exploit sparsity in data to reorganize neurons in neural embedded vector space.

In Figure 18 and 19, we also observe this drastic change in neural topology from initiation. The arguments are still supported in another architecture with 2^{15} hidden units on MNIST (Figure 18 and 19). Adversarial regularization exploits sparsity in data distribution and neural embedded topological vector space, provided it exists. This suggests participation of a smaller subset

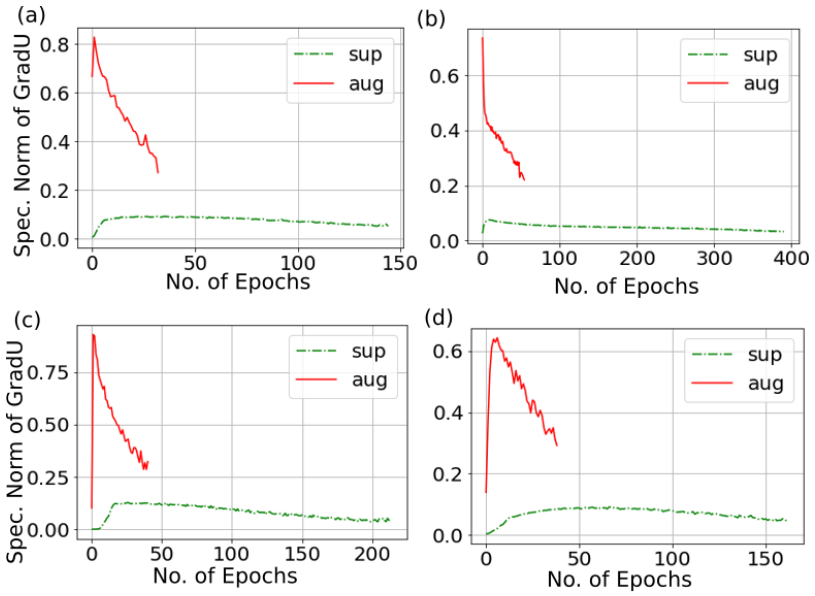

Figure 14: Comparison of gradient updates between supervised and augmented objective as observed in the *first layer* on MNIST. (a) Multi-Layer Perceptron. (b) Exponential Activation. (c) Residual Network. (d) Dense Network.

of neurons in achieving desired task. Vanishing gradient phenomenon, which we observed in both the layers, adds on to the explanation of lacking evolutionary patterns in sole supervision. Figure 20 and 21 illustrate similar observations with 2^{13} hidden units on MNIST.

Further, we study the neural topology of other fixed subsets of neurons in a network with 2^{13} hidden units as shown in Figure 22 and 23. In this analysis, we focus on 4 subsets of 2048 neurons each sequentially. Since we repeatedly observe new patterns even with random seeds, it ensures that the geometry of neural embedded vector space has indeed changed drastically. Also, we analyze the topology of a randomly selected subset of 2048 neurons with 2^{13} hidden units. As shown in Figure 24, the final topology in adversarial learning does lie on a different manifold as compared to sole supervision. In addition, Figure 25 shows emergence of *global pattern* in adversarial learning due to more local interaction (Turing 1952).

Perturbation Sensitivity In Figure 26 and 27, we investigate the sensitivity of the topological diagrams to local perturbation. The perturbation model considered here follows Gaussian distribution with mean and standard deviation same as that of the fully trained weights. Here, the percentage perturbation corresponds to the fraction of the total energy in the weight vectors. For conciseness, we study sensitivity in the top layer on MNIST with 2^{13} hidden nodes. As shown in Figure 26 and 27, the final topology retains sparse representation with low and moderate level Gaussian perturbation. However, we observe slight reduction of sparsity with extreme perturbation as shown in Figure 27. These experimental results indicate that the sparse nature of neural anatomy in augmented objective is not due to minor deviations from the neural anatomy of sole supervision. Thus, there is a significant difference between the final topology of adversarial regularization and sole supervision in the neural embedded vector space.

NTA on Over-Parameterization

It is well known that highly over-parameterized deep neural networks sprinkle the corresponding parametric space with lots of good solutions. However, it is not fully understood how to reduce this dependency on over-parameterization while still achieve required performance. In this paper, we illustrate this phenomenon using topological diagrams of fully trained networks. The fact that all neurons in the weight space do not contribute equally to the main task highlights the existence of redundant neurons in over-parameterized networks — though in a positive sense.

In Figure 28, we study the neural topology of a network which is trained on randomly labelled pairs of MNIST dataset. With 2^{13} nodes in hidden layer, the augmented objective converges to 0.004 MSE after 1000 epochs. It is interesting to observe these patterns even when trained on a randomly labelled dataset. This purportedly implies that adversarial training is the predominant source that constitutes the basis of such pattern formation.

NTA on FashionMNIST

Additionally, the experiments on FashionMNIST demonstrate similar pattern formation on three different subsets as shown in Figure 29 and 30.

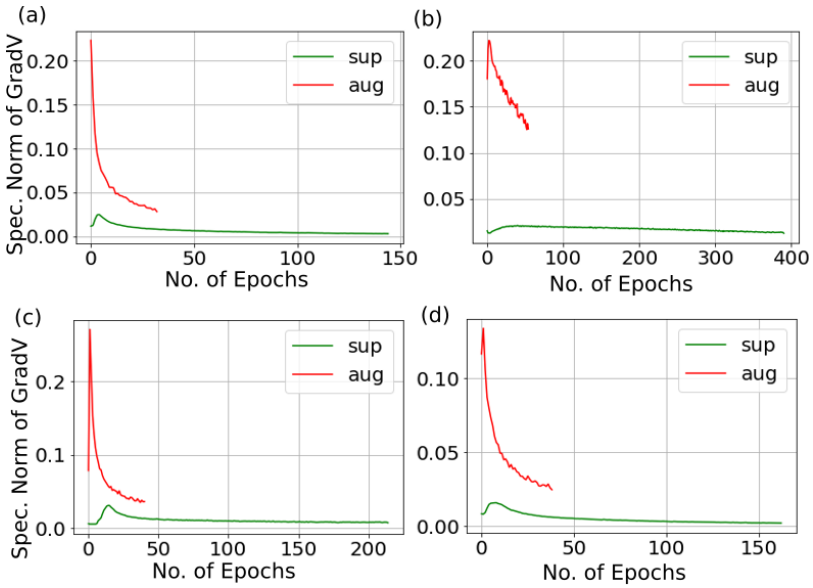

Figure 15: Comparison of gradient updates between supervised and augmented objective as observed in the *last layer* on MNIST. (a) Multi-Layer Perceptron. (b) Exponential Activation. (c) Residual Network. (d) Dense Network.

Neural Anatomy

Since we believe all neurons lie on a common manifold due to single channel representation of input data, it makes more sense to study their topology on single channel datasets, such as MNIST and FashionMNIST. However, there are several potential improvements and extensions to the present account of neural topology analysis. A particularly interesting research direction would be to design an experiment for multi-channel dataset, such as SVHN and CIFAR10. We believe that studying channel specific topology might give insights to design better architectures. Also, it is worth unveiling whether there exist such patterns in convolutional neural networks.

An interesting observation in most of these diagrams is the emergence of animal shaped patterns with central and assistant nervous systems. Turing's theory predicted that the emergence of patterns on the skin of an animal is due to chemical substances, called morphogens reacting together and diffusing through tissues (Turing 1952). In the context of morhphogenesis, while one reaction favors the growth of patterns, another tries to prohibit it. In the nascent state of understanding, this forms the chemical basis of morphogenesis. To our surprise, the adversarial game between generator and discriminator also forms a similar basis for the evolutionary pattern formation in neural topology, suggesting further research in this direction might prove beneficial.

The neural topology in adversarial regularization has essentially two components: a central nervous system/dense branch and an assistant nervous system/narrow branch. The resemblance of dense branch with neural topology diagram in sole supervision suggests that adversarial learning somehow exploits sparsity in over-parameterized neural networks. Furthermore, it provides accelerated gradients in the optimization process. As it turns out, adversarial learning depends upon a very few primary processing elements to efficiently perform the same task. It is however unclear at the moment the exact role of each of these individual branches. It makes one wonder whether local neural interaction, which is believed to be the primary cause of such evolutionary patterns, can help in reasoning, interpretability and designing efficient architectures upon further investigation.

To our knowledge, one can not at present hope to make progress in understanding the electrical, chemical and mechanical properties of neurons in the fabric of space and time that influence the emergence of evolutionary patterns. It is hoped, however, that the simplified architectures retained for discussion are those of greatest importance at this juncture. Thus, the present account of the problem is vastly a simplification and an idealization of actual neural anatomy. It is intended to bridge the gap between chemical basis of morphogenesis and an equivalent mathematical basis of neural topology.

C Technical Proofs

Proof of Theorem 2

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

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

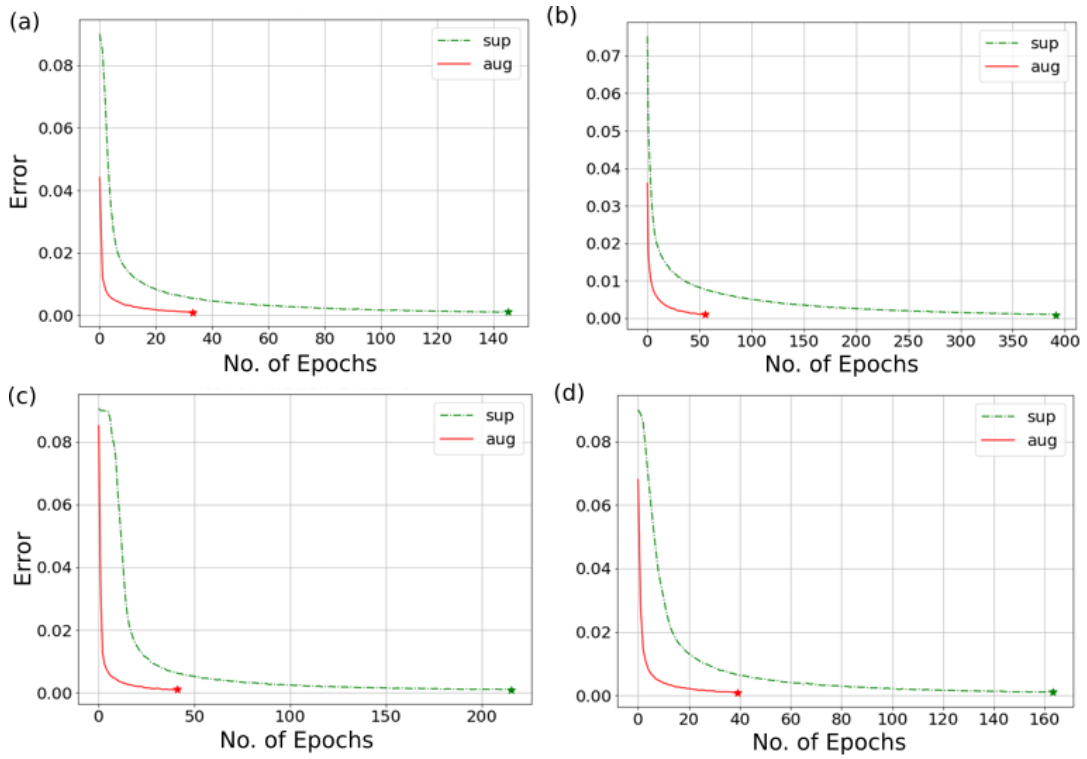

Figure 16: Comparison of optimal empirical risk on MNIST. (a) Multi-Layer Perceptron. (b) Exponential Activation. (c) Residual Network. (d) Dense Network.

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

$$\begin{aligned} l\left(\theta_{k+1}\right) &= l\left(\theta_k\right) + \nabla l\left(\theta_k\right)\left(\theta_{k+1} - \theta_k\right) + \frac{1}{2}\left(\theta_{k+1} - \theta_k\right)^T \nabla^2 l\left(\theta_k\right)\left(\theta_{k+1} - \theta_k\right) \\ &= l\left(\theta_k\right) - h_k \left\|\nabla l\left(\theta_k\right)\right\|^2 + \frac{1}{2}\left(\theta_{k+1} - \theta_k\right)^T \nabla^2 l\left(\theta_k\right)\left(\theta_{k+1} - \theta_k\right), \ \left(\because \theta_{k+1} - \theta_k = -h_k \nabla l(\theta_k)\right). \end{aligned} \tag{5}$$

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

$$\begin{aligned} l\left(\theta_{k+1}\right) &\leq l\left(\theta_k\right) - h_k \left\|\nabla l\left(\theta_k\right)\right\|^2 + \frac{1}{2}\left\|\left(\theta_{k+1} - \theta_k\right)\right\| \left\|\nabla^2 l\left(\theta_k\right)\left(\theta_{k+1} - \theta_k\right)\right\| \\ &\leq l\left(\theta_k\right) - h_k \left\|\nabla l\left(\theta_k\right)\right\|^2 + \frac{1}{2}\left\|\left(\theta_{k+1} - \theta_k\right)\right\|^2 \int_0^1 \left\|\nabla^2 l\left(\gamma(t)\right)\right\| dt. \end{aligned} \tag{6}$$

We know by **Assumption 5**

$$\left\|\nabla^2 l\left(\theta\right)\right\| \leq L_0 + L_1 \left\|\nabla l\left(\theta\right)\right\|. \tag{7}$$

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

$$\begin{aligned} l\left(\theta_{k+1}\right) &\leq l\left(\theta_k\right) - h_k \left\|\nabla l\left(\theta_k\right)\right\|^2 + \frac{h_k^2 \left\|\nabla l\left(\theta_k\right)\right\|^2}{2} \int_0^1 \left(L_0 + L_1 \left\|\nabla l\left(\gamma(t)\right)\right\|\right) dt \\ &\leq l\left(\theta_k\right) - h_k \left\|\nabla l\left(\theta_k\right)\right\|^2 + \frac{h_k^2 \left\|\nabla l\left(\theta_k\right)\right\|^2}{2} \int_0^1 \left(L_0 + L_1 L^2 \beta \epsilon\right) dt \\ &\leq l\left(\theta_k\right) - h_k \left\|\nabla l\left(\theta_k\right)\right\|^2 + \frac{h_k^2 \left\|\nabla l\left(\theta_k\right)\right\|^2 \left(L_0 + L_1 L^2 \beta \epsilon\right)}{2}. \end{aligned} \tag{8}$$

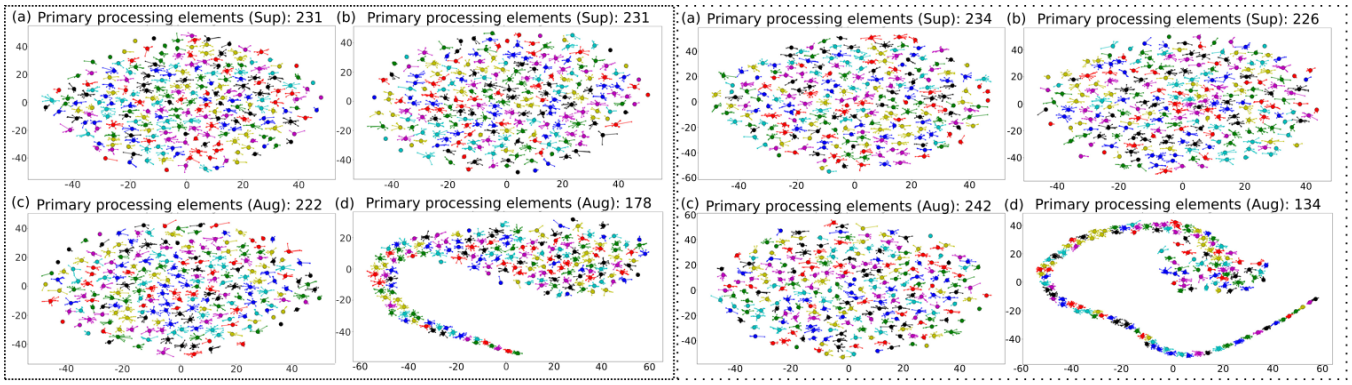

Figure 17: NTA in *hidden layer* (left) and *top layer* (right). (a) Initial and (b) final topology in supervised learning. (c) Initial and (d) final topology in adversarial learning.

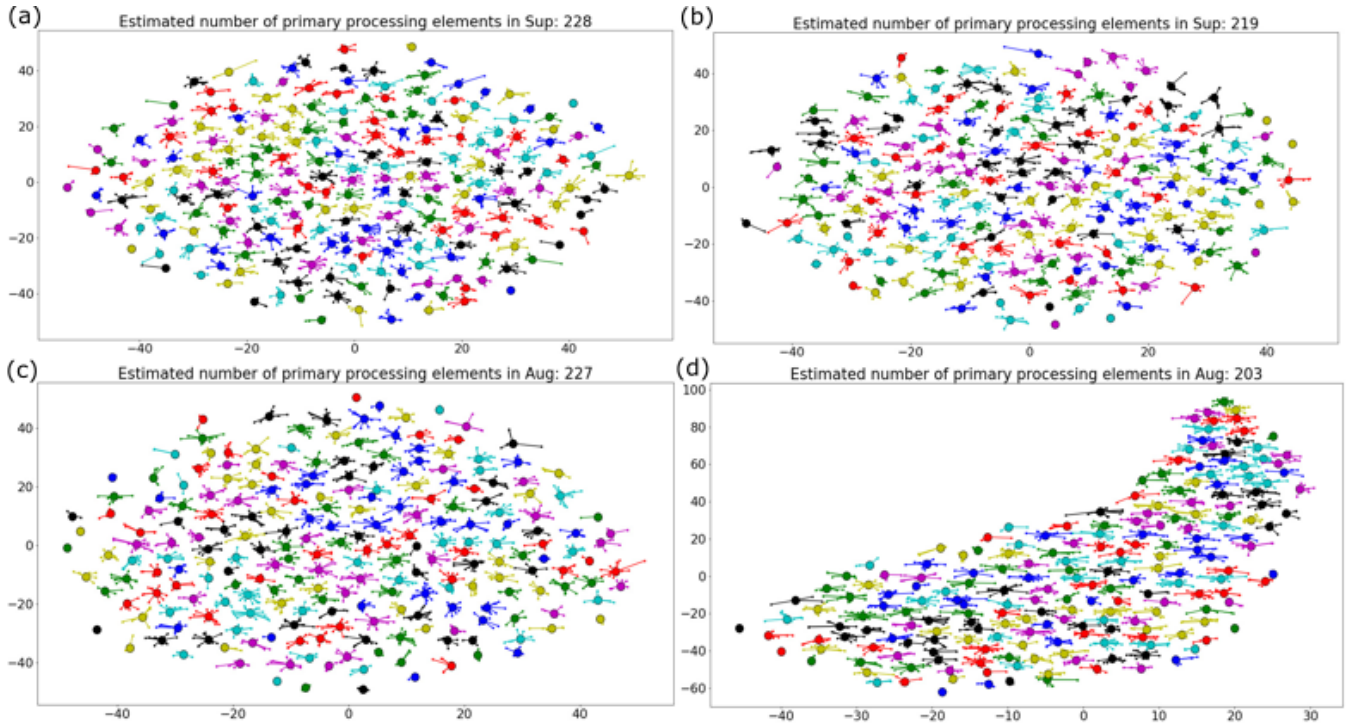

Figure 18: NTA in the *hidden layer* with 2^{15} hidden units. (a) Initial and (b) final topology in supervised learning. (c) Initial and (d) final topology in adversarial learning.

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

$$\begin{aligned} l\left(\theta_{k+1}\right) &\leq l\left(\theta_k\right) - \frac{h_k \left\|\nabla l\left(\theta_k\right)\right\|^2}{2} \\ &\leq l\left(\theta_k\right) - \frac{\left\|\nabla l\left(\theta_k\right)\right\|^2}{2\left(L_0 + L_1 \lambda M\right)}. \end{aligned} \tag{9}$$

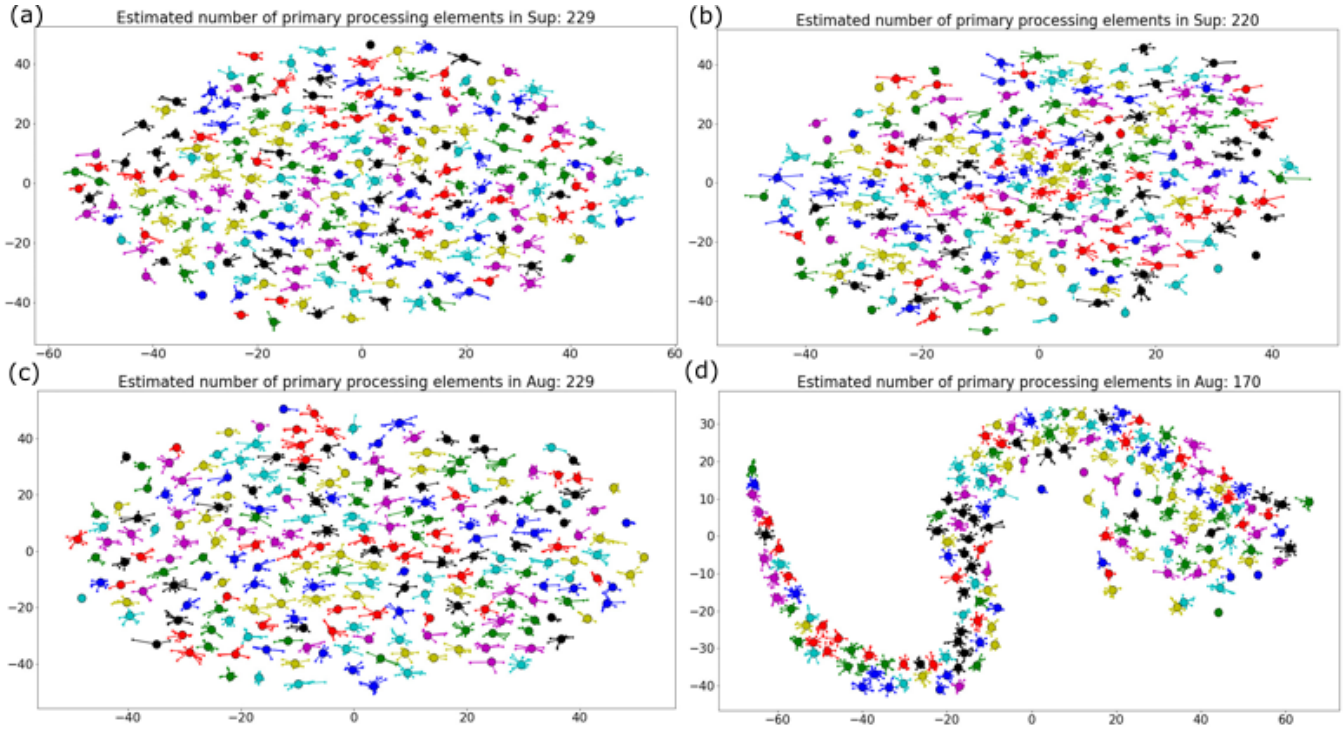

Figure 19: NTA in the *top layer* with 2^{15} hidden units. (a) Initial and (b) final topology in supervised learning. (c) Initial and (d) final topology in adversarial learning.

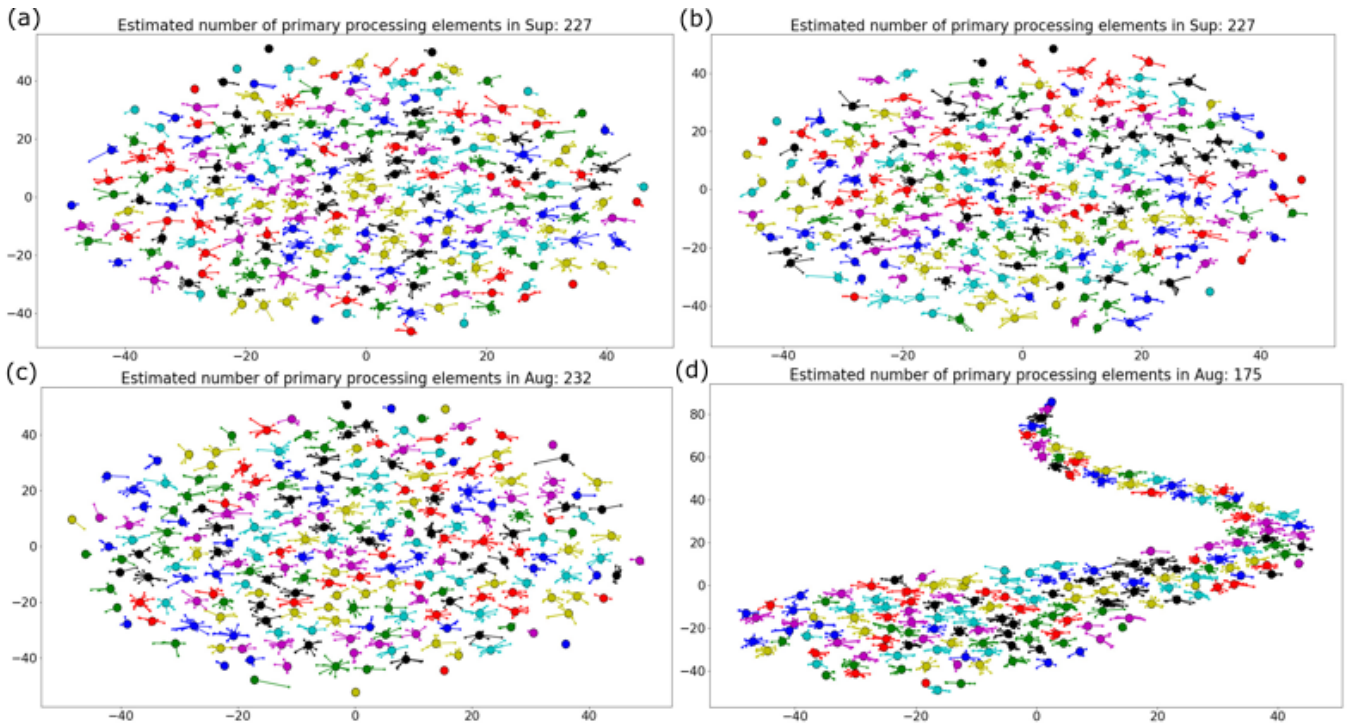

Figure 20: NTA in the *hidden layer* with 2^{13} hidden units. (a) Initial and (b) final topology in supervised learning. (c) Initial and (d) final topology in adversarial learning.

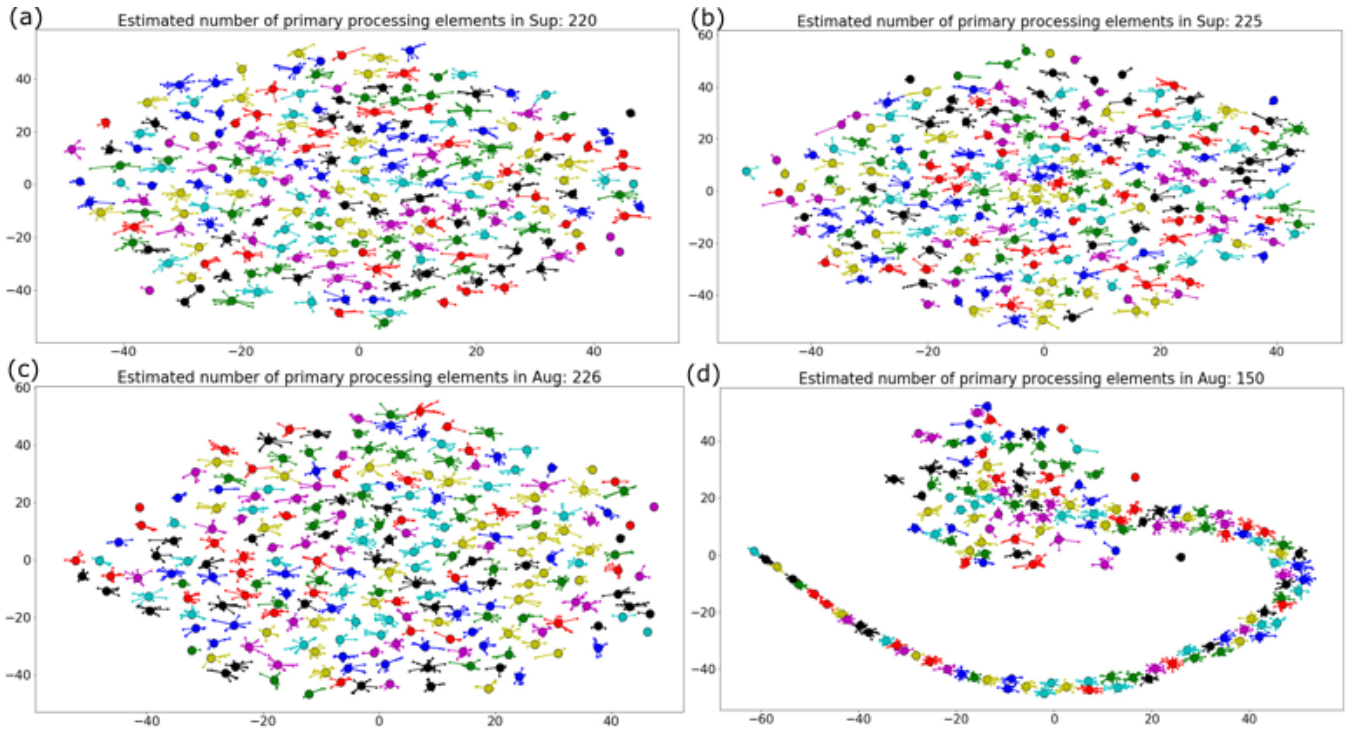

Figure 21: NTA in the *top layer* with 2^{13} hidden units. (a) Initial and (b) final topology in supervised learning. (c) Initial and (d) final topology in adversarial learning.

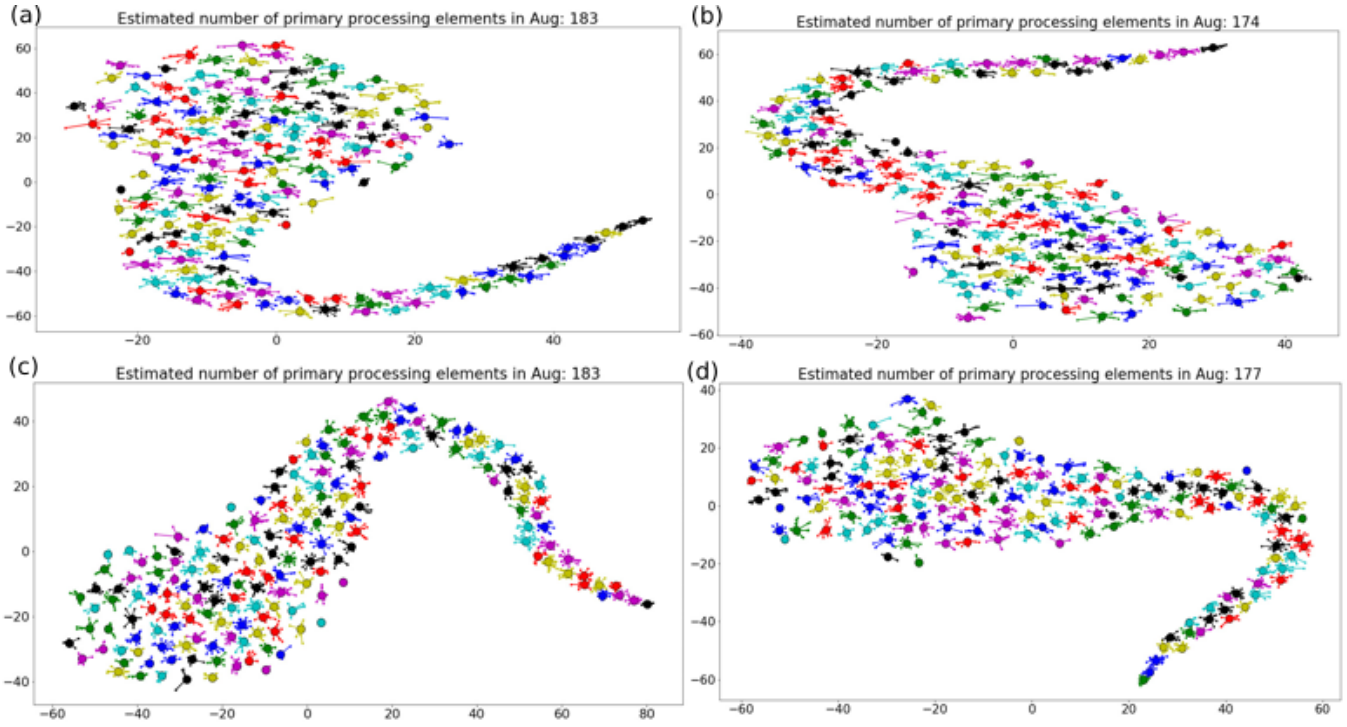

Figure 22: NTA in the *hidden layer* with 2^{13} hidden units. (a) First subset (0-2048) (b) Second subset (2048-4096) (c) Third subset (4096-6144) (d) Fourth subset (6144-8192) final topology in adversarial learning.

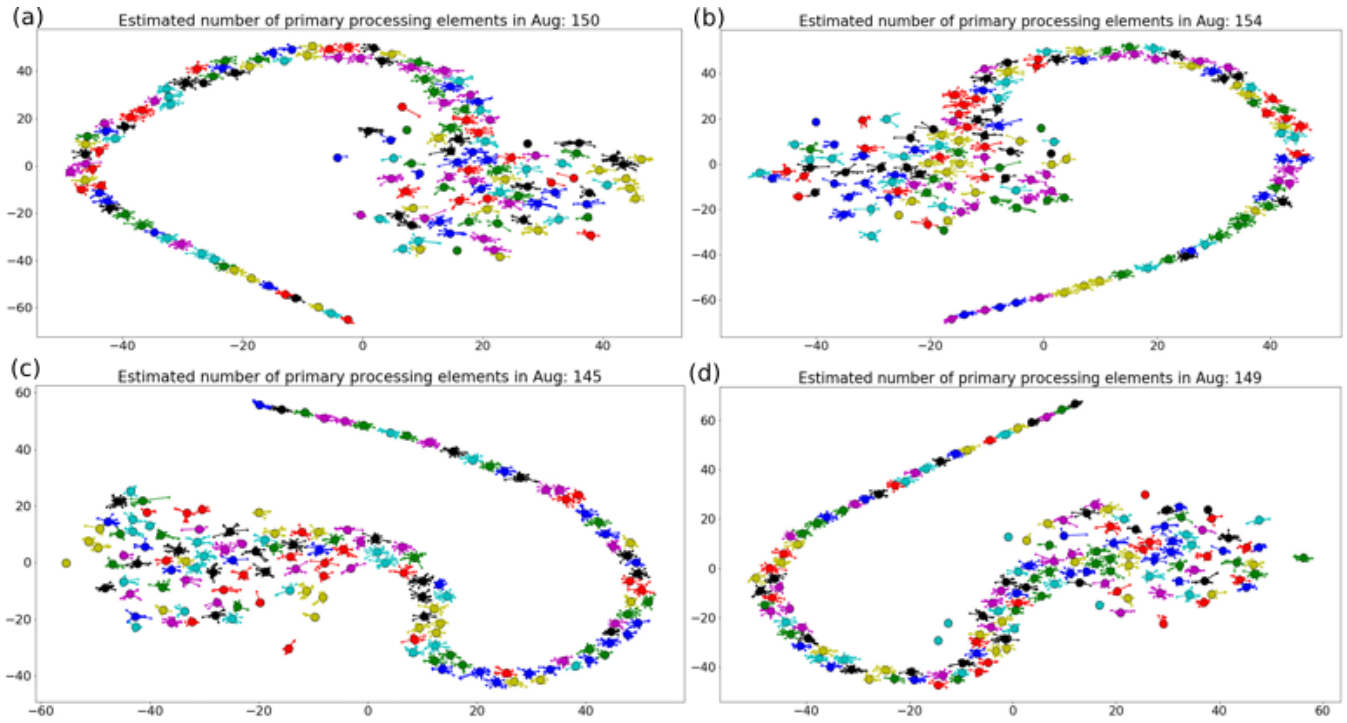

Figure 23: NTA in the *top layer* with 2^{13} hidden units. (a) First subset (0-2048) (b) Second subset (2048-4096) (c) Third subset (4096-6144) (d) Fourth subset (6144-8192) final topology in adversarial learning.

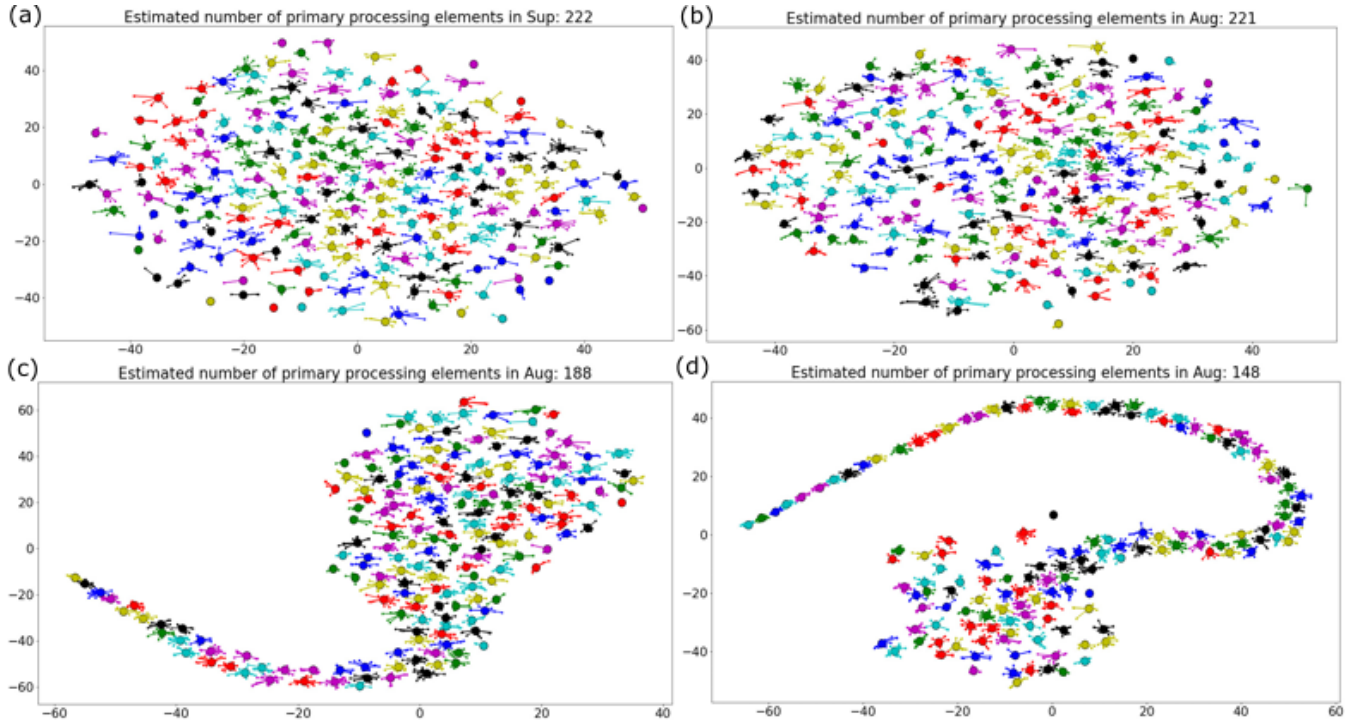

Figure 24: NTA of a random subset of 2048 neurons with 2^{13} hidden units.(a) Hidden and (b) top layer topology in supervised learning. (c) Hidden and (d) top layer topology in adversarial learning.

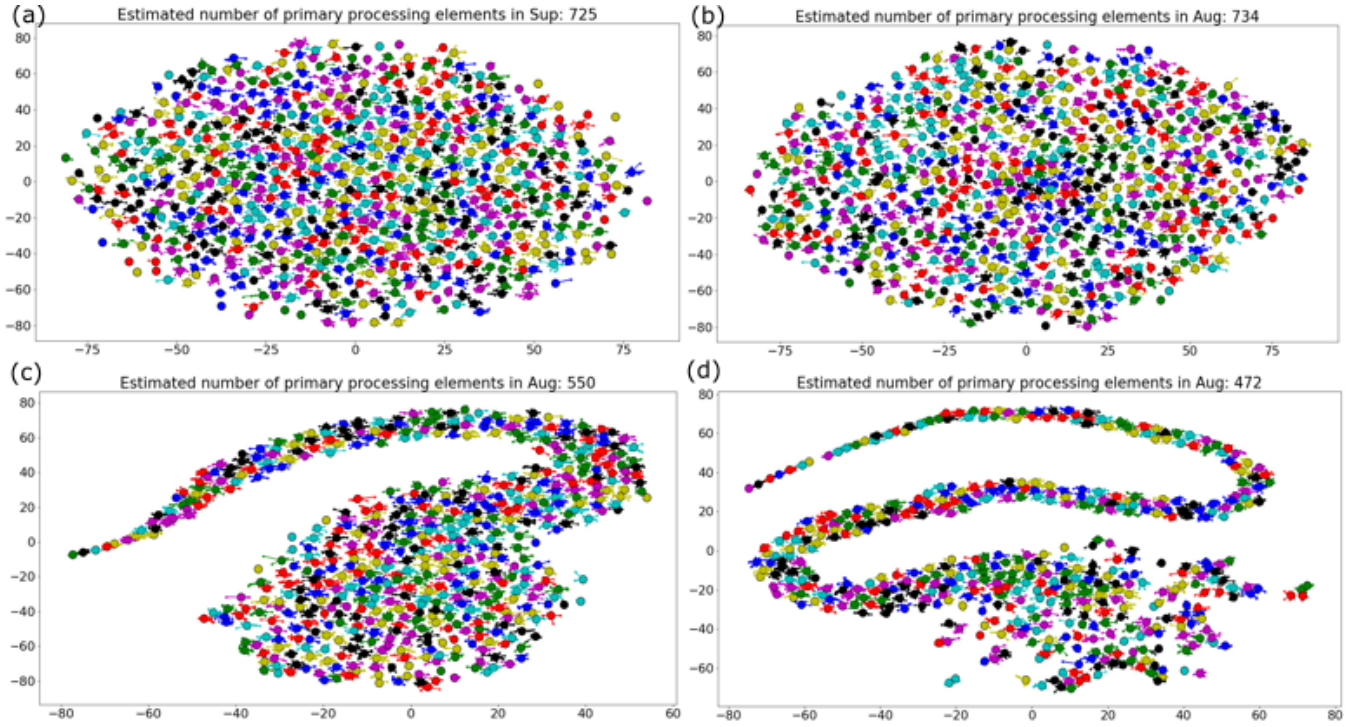

Figure 25: NTA of all 2^{13} hidden units.(a) Hidden and (b) top layer final topology in supervised learning. (c) Hidden and (d) top layer final topology in adversarial learning.

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,

$$\begin{aligned} \sum_{k=0}^{T-1} l(\theta_{k+1}) - l(\theta_k) &\leq \frac{-T\epsilon^2}{2(L_0 + L_1\lambda M)} \\ \implies T &\leq \frac{2(l(\theta_0) - l^*)\left(L_0 + L_1 L^2 \beta\epsilon\right)}{\epsilon^2}. \end{aligned} \tag{10}$$

Therefore, we get

$$\sup_{\theta_0 \in \left\{\mathbb{R}^{h\times d_x}, \mathbb{R}^{d_y \times h}\right\}, l \in \mathscr{L}} \mathcal{T}_\epsilon(A_h[l, \theta_0], l) = \mathcal{O}\left(\frac{(l(\theta_0) - l^*)\left(L_0 + L_1 L^2\beta\epsilon\right)}{\epsilon^2}\right) \tag{11}$$

which finishes the proof. □

Proof of Corollary 1

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

$$\begin{aligned} l(\theta_{k+1}) &= l(\theta_k) - h_k \|\nabla l(\theta_k)\|^2 \\ &\leq l(\theta_k) - h_k\epsilon^2. \end{aligned} \tag{12}$$

By telescopic sum,

$$\begin{aligned} \sum_{k=0}^{T-1} l(\theta_{k+1}) - l(\theta_k) &\leq -T h_k \epsilon^2 \\ \implies T &\leq \frac{(l(\theta_0) - l^*)}{h_k\epsilon^2}. \end{aligned} \tag{13}$$

So,

$$\sup_{\theta_0 \in \left\{\mathbb{R}^{h\times d_x}, \mathbb{R}^{d_y \times h}\right\}, l \in \mathscr{L}} \mathcal{T}_\epsilon(A_h[l, \theta_0], l) = \mathcal{O}\left(\frac{(l(\theta_0) - l^*)}{h\epsilon^2}\right) \tag{14}$$

which finishes the proof. □

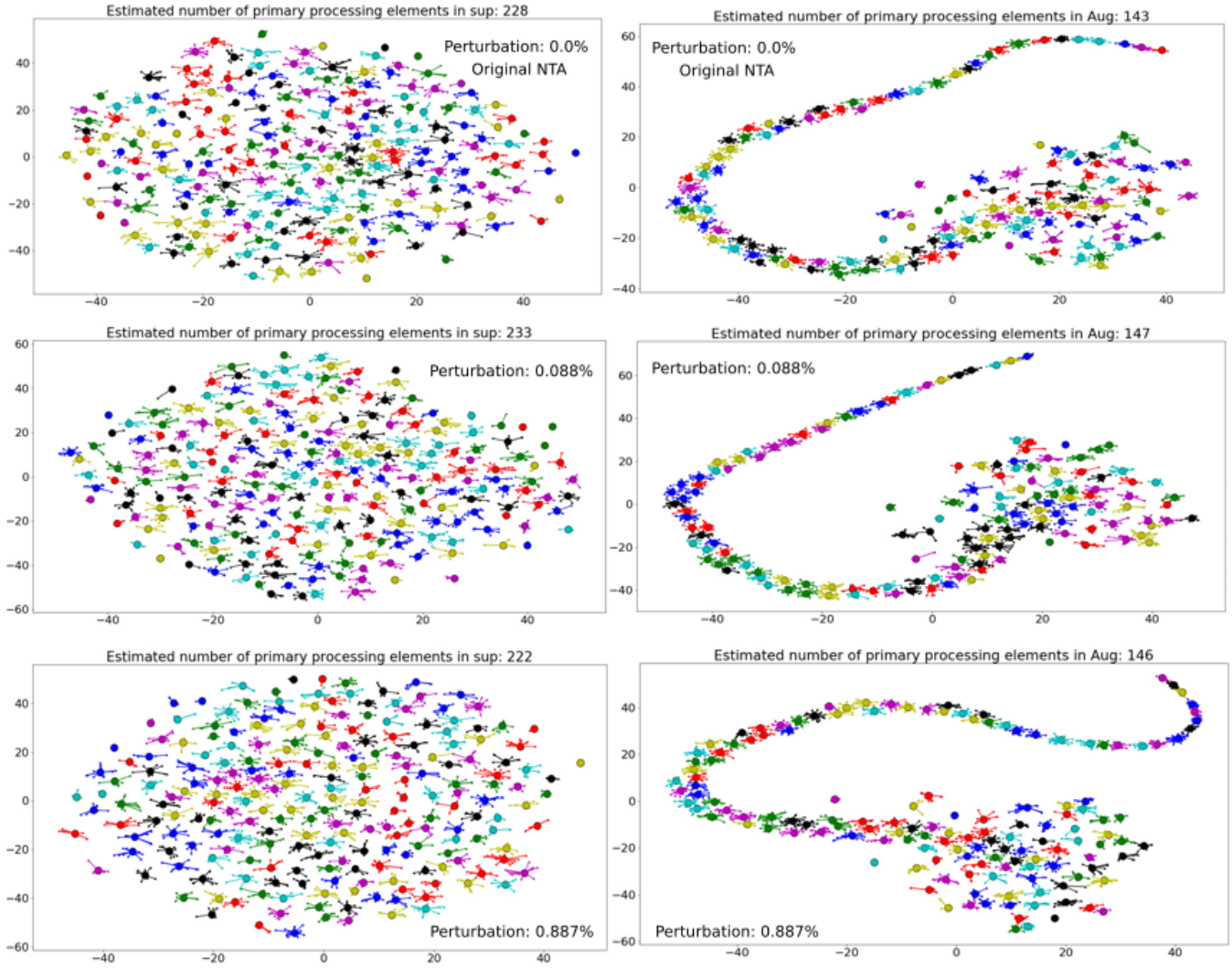

Figure 26: NTA in the *top layer* with 2^{13} hidden units. Comparison of sensitivity to low level Gaussian perturbation. Final topology in supervised learning (left) and adversarial learning (right).

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

$$\theta_{k+1} = \theta_k - h_k \nabla \left(l\left(\theta_k\right) - g\left(\psi; f\left(\theta_k; x\right)\right)\right). \tag{15}$$

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

$$l\left(\theta_{k+1}\right) \leq l\left(\theta_k\right) - h_k \left\|\nabla l\left(\theta_k\right)\right\|^2 - h_k \left\|\nabla l\left(\theta_k\right)\right\| \left\|\nabla g\left(\psi; f\left(\theta_k; x\right)\right)\right\| + \frac{h_k^2 \left\|\nabla\left(l\left(\theta_k\right) - g\left(\psi; f\left(\theta_k; x\right)\right)\right)\right\|^2}{2} \int_0^1 \left\|\nabla^2 l(\gamma(t))\right\| dt. \tag{16}$$

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

$$l\left(\theta_{k+1}\right) \leq l\left(\theta_k\right) - h_k \left\|\nabla l\left(\theta_k\right)\right\|^2 - h_k \left\|\nabla l\left(\theta_k\right)\right\| \zeta + \frac{h_k^2 \left\|\nabla l\left(\theta_k\right) - \nabla g\left(\psi; f\left(\theta_k; x\right)\right)\right\|^2}{2} \int_0^1 \left(L_0 + L_1 \left\|\nabla l(\gamma(t))\right\|\right) dt. \tag{17}$$

Upon simplification using arguments of Appendix C and applying Minkowski's inequality,

$$l\left(\theta_{k+1}\right) \leq l\left(\theta_k\right) - h_k \left\|\nabla l\left(\theta_k\right)\right\|^2 - h_k \left\|\nabla l\left(\theta_k\right)\right\| \zeta + \frac{h_k^2 \left(\left\|\nabla l\left(\theta_k\right)\right\|^2 + \left\|\nabla g\left(\psi; f\left(\theta_k; x\right)\right)\right\|^2\right)}{2} \left(L_0 + L_1 \lambda M\right). \tag{18}$$

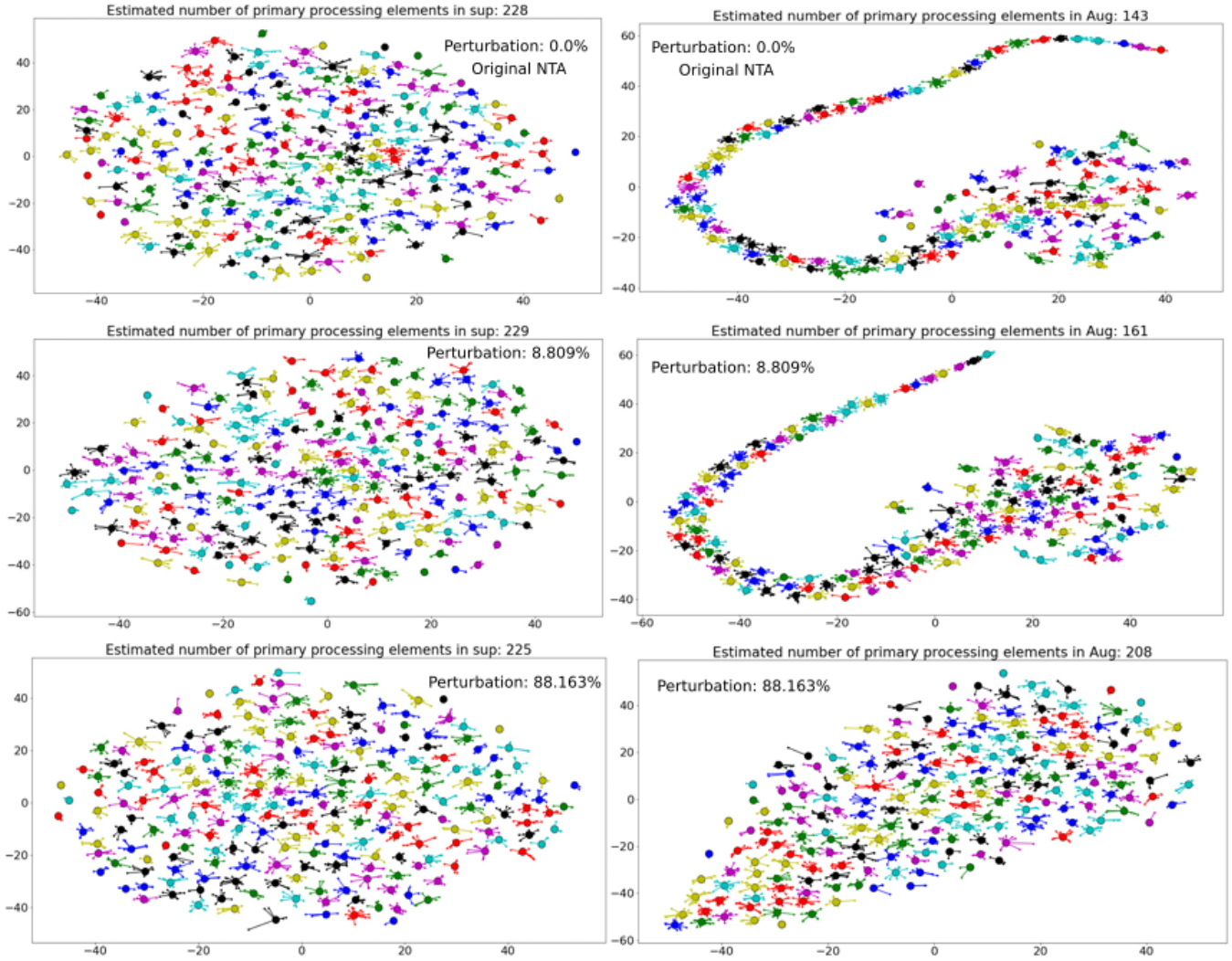

Figure 27: NTA in the *top layer* with 2^{13} hidden units. Comparison of sensitivity to moderate and extreme level Gaussian perturbation. Final topology in supervised learning (left) and adversarial learning (right).

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

$$\begin{aligned} l\left(\theta_{k+1}\right) &\leq l\left(\theta_k\right) - \frac{h_k \left\|\nabla l\left(\theta_k\right)\right\|^2}{2} - h_k \left\|\nabla l\left(\theta_k\right)\right\| \zeta + \frac{h_k \left\|\nabla g\left(\psi; f\left(\theta_k; x\right)\right)\right\|^2}{2} \\ &\leq l\left(\theta_k\right) - \frac{h_k \left\|\nabla l\left(\theta_k\right)\right\|^2}{2} - h_k \left\|\nabla l\left(\theta_k\right)\right\| \zeta + \frac{h_k L^2 \delta^2}{2}, \text{ (from } \textbf{Lemma 2}\text{)}. \end{aligned} \tag{19}$$

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,

$$\begin{aligned} \sum_{k=0}^{T-1} l\left(\theta_{k+1}\right) - l\left(\theta_k\right) &\leq \frac{-T\left(\epsilon^2 + 2\epsilon\zeta - L^2\delta^2\right)}{2\left(L_0 + L_1 L^2 \beta \epsilon\right)} \\ \implies T &\leq \frac{2\left(l(\theta_0) - l^*\right)\left(L_0 + L_1 L^2 \beta \epsilon\right)}{\epsilon^2 + 2\epsilon\zeta - L^2\delta^2}. \end{aligned} \tag{20}$$

Therefore, we obtain

$$\sup_{\theta_0 \in \left\{\mathbb{R}^{h \times d_x}, \mathbb{R}^{d_y \times h}\right\}, l \in \mathscr{L}} \mathcal{T}_\epsilon\left(A_h\left[l, \theta_0\right], l\right) = \mathcal{O}\left(\frac{\left(l(\theta_0) - l^*\right)\left(L_0 + L_1 \lambda M\right)}{\epsilon^2 + 2\epsilon\zeta - \delta^2 M^2}\right) \tag{21}$$

which finishes the proof. □

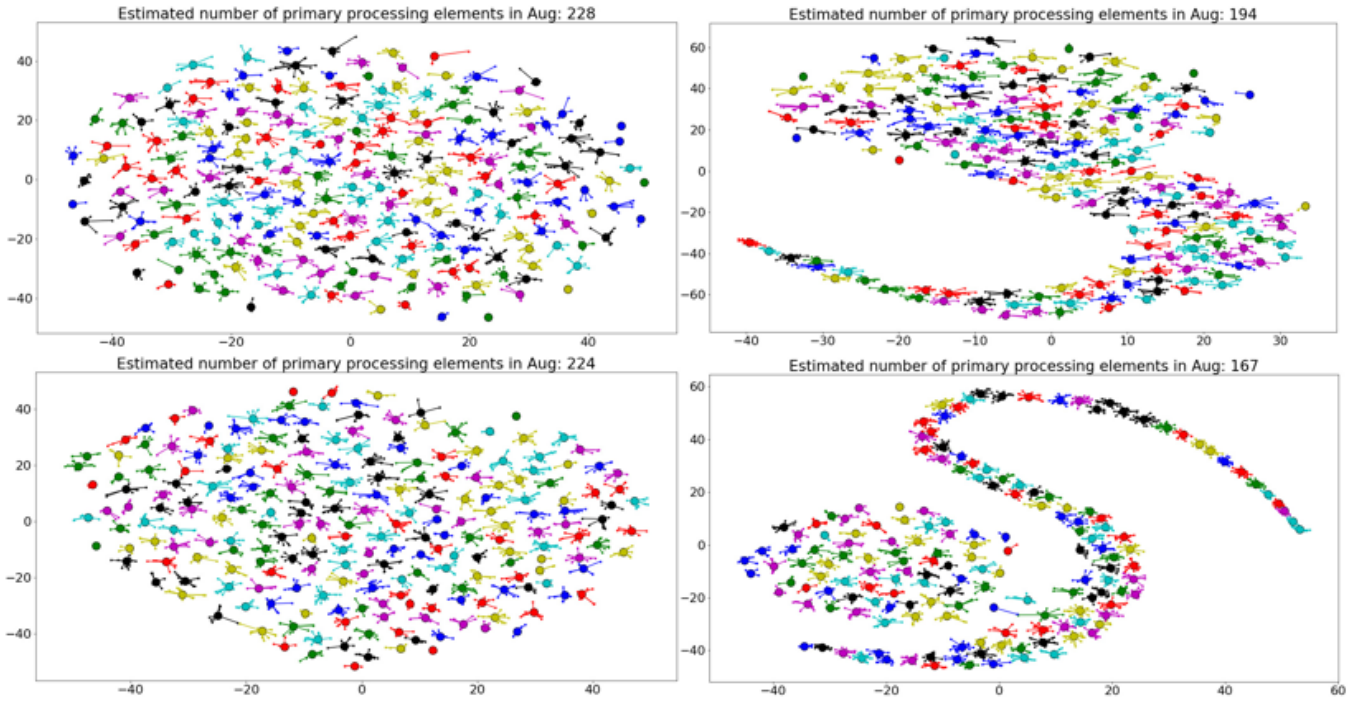

Figure 28: NTA in *adversarial learning* with 2^{13} hidden units. Initial and final topology in *hidden layer* (first row) and *top layer* (second row).

Proof of Corollary 2

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

$$\begin{aligned} l\left(\theta_{k+1}\right) &= l\left(\theta_k\right) - h_k \left\|\nabla l\left(\theta_k\right)\right\|^2 - h_k \left\|\nabla l\left(\theta_k\right)\right\| \left\|\nabla g\left(\psi; f\left(\theta_k; x\right)\right)\right\| \\ &\leq l\left(\theta_k\right) - h_k \epsilon^2 - h_k \epsilon \zeta. \end{aligned} \tag{22}$$

By telescopic sum,

$$\begin{aligned} \sum_{k=0}^{T-1} l\left(\theta_{k+1}\right) - l\left(\theta_k\right) &\leq -T h_k \epsilon^2 - T h_k \epsilon \zeta \\ \implies T &\leq \frac{\left(l\left(\theta_0\right) - l^*\right)}{h_k \epsilon^2 + h_k \epsilon \zeta}. \end{aligned} \tag{23}$$

Therefore,

$$\sup_{\theta_0 \in \left\{\mathbb{R}^{h \times d_x}, \mathbb{R}^{d_y \times h}\right\}, l \in \mathscr{L}} \mathcal{T}_\epsilon\left(A_h\left[l, \theta_0\right], l\right) = \mathcal{O}\left(\frac{\left(l\left(\theta_0\right) - l^*\right)}{h\epsilon^2 + h\epsilon\zeta}\right) \tag{24}$$

which finishes the proof. □

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 $r^2(t) = \frac{1}{2}\left\|\theta(t) - \theta^*\right\|^2$. Now differentiating both sides, we get

$$\begin{aligned} \frac{dr^2(t)}{dt} &= \left\langle \frac{d\theta(t)}{dt}, \theta(t) - \theta^* \right\rangle \\ &= \left\langle -\nabla l(\theta(t)), \theta(t) - \theta^* \right\rangle. \end{aligned} \tag{25}$$

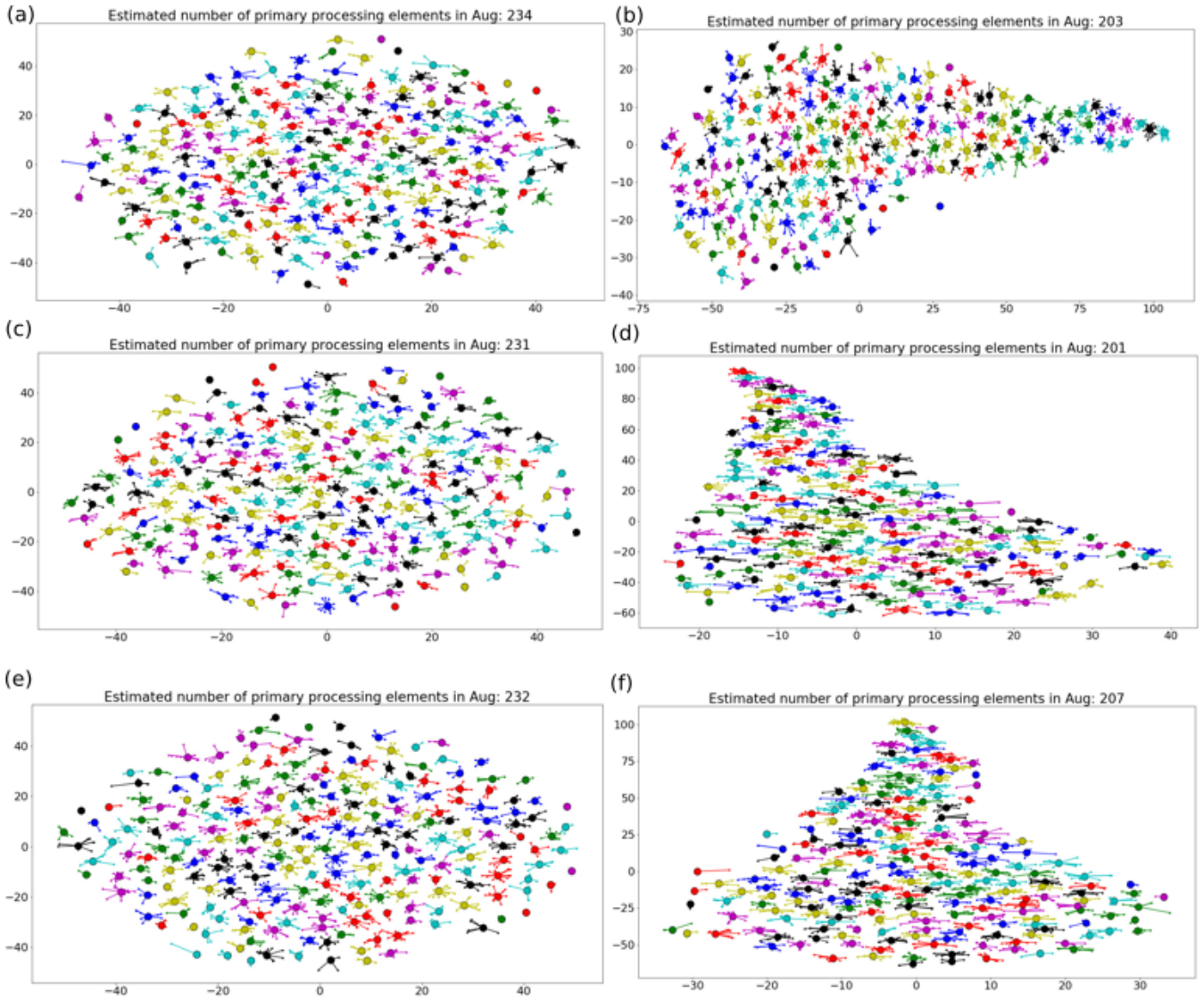

Figure 29: NTA in the *hidden layer* with 2^{13} hidden units on FashionMNIST. (a) First subset (0-2048) (b) Second subset (2048-4096) (c) Third subset (4096-6144) initial (left) and final (right) topology in adversarial learning.

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

$$\begin{aligned} \frac{1}{T}\int_0^T \frac{dr^2(t)}{dt}dt &\leq \frac{1}{T}\int_0^T -\kappa(t)dt \\ \implies \frac{1}{T}\left(r^2(T) - r^2(0)\right) &\leq -\frac{1}{T}\int_0^T \kappa(t)dt \\ \implies \frac{1}{T}\int_0^T \kappa(\theta(t))dt &\leq \frac{r^2(0)}{T}. \end{aligned} \tag{26}$$

By Jensen's inequality,

$$\kappa\left(\frac{1}{T}\int_0^T \theta(t)dt\right) \leq \frac{1}{T}\int_0^T \kappa(\theta(t))dt. \tag{27}$$

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. □

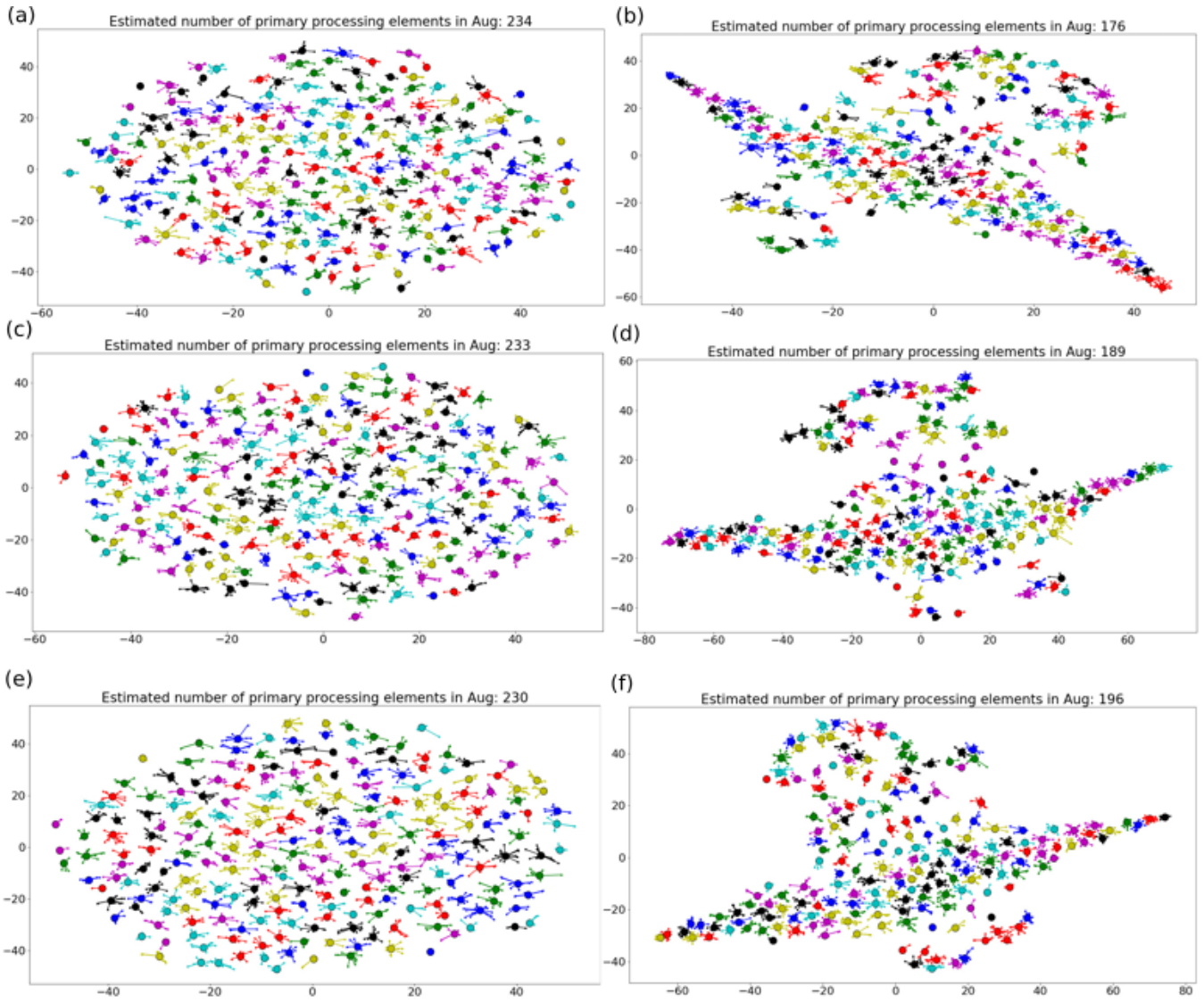

Figure 30: NTA in the *top layer* with 2^{13} hidden units on FashionMNIST. (a) First subset (0-2048) (b) Second subset (2048-4096) (c) Third subset (4096-6144) initial (left) and final (right) topology in adversarial learning.

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 C, we obtain

$$\frac{dr^2(t)}{dt} = \langle -\nabla l(\theta(t)), \theta(t) - \theta^* \rangle + \langle \nabla g(\theta(t)), \theta(t) - \theta^* \rangle . \tag{28}$$

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

$$\begin{aligned} \frac{1}{T}\int_0^T \frac{dr^2(t)}{dt} dt &\leq -\frac{1}{T}\int_0^T \kappa(t)dt - \frac{1}{T}\int_0^T \pi(t)dt \\ \implies \frac{1}{T}\left(r^2(T) - r^2(0)\right) &\leq -\frac{1}{T}\int_0^T \kappa(t)dt - \frac{1}{T}\int_0^T \pi(t)dt \\ \implies \frac{1}{T}\int_0^T \kappa(\theta(t))dt &\leq \frac{r^2(0)}{T} - \frac{1}{T}\int_0^T \pi(\theta(t))dt. \end{aligned} \tag{29}$$

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

$$\kappa\left(\frac{1}{T}\int_0^T \theta(t)dt\right) = \mathcal{O}\left(\frac{\|\theta(0)-\theta^*\|^2}{2T} - \pi\left(\frac{1}{T}\int_0^T \theta(t)dt\right)\right) \tag{30}$$

which finishes the proof. □

Proof of Theorem 6

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

$$\begin{aligned}
\|\theta_k - \theta^*\|^2 &= \|\theta_{k-1} - \eta_k\hat{\mathfrak{g}}_{k-1} - \theta^*\|^2 \\
&= \|\theta_{k-1}-\theta^*\|^2 - 2\eta_k\langle\theta_{k-1}-\theta^*, \hat{\mathfrak{g}}_{k-1}\rangle + \eta_k^2\|\hat{\mathfrak{g}}_{k-1}\|^2 \\
&= \|\theta_{k-1}-\theta^*\|^2 - 2\eta_k\langle\theta_{k-1}-\theta^*, \nabla\mathfrak{l}(\theta_{k-1})\rangle - 2\eta_k\langle\theta_{k-1}-\theta^*, b_{k-1}\rangle + \eta_k^2\|\hat{\mathfrak{g}}_{k-1}\|^2 \\
&\leq \|\theta_{k-1}-\theta^*\|^2 - 2\eta_k\langle\theta_{k-1}-\theta^*, \nabla\mathfrak{l}(\theta_{k-1})\rangle + \underbrace{2\eta_k\|\theta_{k-1}-\theta^*\|\|b_{k-1}\|}_{\text{By Cauchy-Schwarz inequality}} + \eta_k^2\|\hat{\mathfrak{g}}_{k-1}\|^2 \\
&\leq \|\theta_{k-1}-\theta^*\|^2 - 2\eta_k\langle\theta_{k-1}-\theta^*, \nabla\mathfrak{l}(\theta_{k-1})\rangle + \underbrace{\eta_k\left(\|\theta_{k-1}-\theta^*\|^2 + \|b_{k-1}\|^2\right)}_{\text{By AM-GM inequality}} + \eta_k^2\|\hat{\mathfrak{g}}_{k-1}\|^2
\end{aligned} \tag{31}$$

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

$$\begin{aligned}
\|\theta_k-\theta^*\|^2 &\leq \|\theta_{k-1}-\theta^*\|^2 - 2\eta_k\left(\mathfrak{l}(\theta_{k-1}) - \mathfrak{l}(\theta^*)\right) - \eta_k\mu\|\theta_{k-1}-\theta^*\|^2 + \eta_k\left(\|\theta_{k-1}-\theta^*\|^2 + \|b_{k-1}\|^2\right) + \eta_k^2\|\hat{\mathfrak{g}}_{k-1}\|^2 \\
&\leq \|\theta_{k-1}-\theta^*\|^2(1-\eta_k\mu+\eta_k) - 2\eta_k\left(\mathfrak{l}(\theta_{k-1}) - \mathfrak{l}(\theta^*)\right) + \eta_k\|b_{k-1}\|^2 + \eta_k^2\|\hat{\mathfrak{g}}_{k-1}\|^2.
\end{aligned} \tag{32}$$

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:*

$$\mathbb{E}\left[\|\hat{\mathfrak{g}}(\theta)\|^2\right] \leq G^\alpha\tau^{2-\alpha} \text{ and } \|\mathbb{E}\left[\hat{\mathfrak{g}}(\theta)\right] - \nabla l(\theta) + \nabla g(\theta)\|^2 \leq G^{2\alpha}\tau^{2-2\alpha}. \tag{33}$$

One can easily prove this using **Lemma 2** of (Zhang et al. 2019b). Upon rearranging, taking expectation of both sides, and using **Lemma 3**,

$$\mathbb{E}\left[\mathfrak{l}(\theta_{k-1})\right] - \mathfrak{l}(\theta^*) \leq \mathbb{E}\left[\left(\frac{\eta_k^{-1}-\mu+1}{2}\right)\|\theta_{k-1}-\theta^*\|^2 - \frac{\eta_k^{-1}}{2}\|\theta_k-\theta^*\|^2\right] + \frac{1}{2}G^{2\alpha}\tau^{2-2\alpha} + \frac{\eta_k}{2}G^\alpha\tau^{2-\alpha}. \tag{34}$$

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 $\tau_k = Gk^{\frac{1}{\alpha}}\mu^{\frac{1}{\alpha}}$, $\eta_k = \frac{5}{2\mu(k+1)}$ and multiply k both sides. Thus,

$$k\mathbb{E}\left[\mathfrak{l}(\theta_{k-1})\right] - k\mathfrak{l}(\theta^*) \leq \mathbb{E}\left[k(k-1)\|\theta_{k-1}-\theta^*\|^2 - k(k+1)\|\theta_k-\theta^*\|^2\right] + \frac{G^2k^{\frac{2-\alpha}{\alpha}}\mu^{\frac{2-2\alpha}{\alpha}}}{2}\left[\frac{5}{2}\left(\frac{k}{k+1}\right)+1\right]. \tag{35}$$

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

$$k\mathbb{E}\left[\mathfrak{l}(\theta_{k-1})\right] - k\mathfrak{l}(\theta^*) \leq \mathbb{E}\left[k(k-1)\|\theta_{k-1}-\theta^*\|^2 - k(k+1)\|\theta_k-\theta^*\|^2\right] + \frac{7G^2k^{\frac{2-\alpha}{\alpha}}\mu^{\frac{2-2\alpha}{\alpha}}}{4}. \tag{36}$$

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

$$\sum_{k=1}^T k\mathbb{E}\left[\mathfrak{l}(\theta_{k-1})\right] - \mathfrak{l}(\theta^*)\sum_{k=1}^T k \leq \mathbb{E}\left[-T(T+1)\|\theta_T-\theta^*\|^2\right] + \frac{7G^2\mu^{\frac{2-2\alpha}{\alpha}}}{4}\sum_{k=1}^T k^{\frac{2-\alpha}{\alpha}}. \tag{37}$$

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

$$\sum_{k=1}^T k\mathbb{E}\left[\mathfrak{l}(\theta_{k-1})\right] - \mathfrak{l}(\theta^*)\frac{T(T+1)}{2} \leq \frac{7G^2\mu^{\frac{2-2\alpha}{\alpha}}}{4}(T+1)^{\frac{2}{\alpha}}. \tag{38}$$

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

$$\frac{\sum_{k=1}^{T} k\mathbb{E}\left[\mathfrak{l}(\theta_{k-1})\right]}{\sum_{k=1}^{T} k} - \mathfrak{l}(\theta^*) \leq 7G^2\mu^{\frac{2-2\alpha}{\alpha}}\left(T+1\right)^{\frac{2-2\alpha}{\alpha}}. \tag{39}$$

By Jensen's inequality,

$$\mathbb{E}\left[\mathfrak{l}\left(\frac{\sum_{k=1}^{T} k\theta_{k-1}}{\sum_{k=1}^{T} k}\right)\right] - \mathfrak{l}(\theta^*) \leq \mathcal{O}\left(G^2\left(\mu(T+1)\right)^{\frac{2-2\alpha}{\alpha}}\right) \tag{40}$$

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

$$\mathbb{E}\left[l\left(\bar{\theta}\right)\right] - l(\theta^*) \leq \mathcal{O}\left(G^2\left(\mu(T+1)\right)^{\frac{2-2\alpha}{\alpha}} - \left(g\left(\theta^*\right) - \mathbb{E}\left[g\left(\bar{\theta}\right)\right]\right)\right), \tag{41}$$

which finishes the proof. □

Proof of Theorem 7

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

$$\begin{aligned}
\mathfrak{l}\left(\theta_k\right) &\leq \mathfrak{l}\left(\theta_{k-1}\right) + \left\langle \nabla\mathfrak{l}\left(\theta_{k-1}\right), \theta_k - \theta_{k-1}\right\rangle + \frac{L}{2}\left\|\theta_k - \theta_{k-1}\right\|^2 \\
&\leq \mathfrak{l}\left(\theta_{k-1}\right) + \left\langle \nabla\mathfrak{l}\left(\theta_{k-1}\right), -\eta_k\hat{\mathfrak{g}}_{k-1}\right\rangle + \frac{\eta_k^2 L}{2}\left\|\hat{\mathfrak{g}}_{k-1}\right\|^2 \\
&\leq \mathfrak{l}\left(\theta_{k-1}\right) - \eta_k\left|\left\|\nabla\mathfrak{l}(\theta_{k-1})\right\|^2 - \eta_k\langle\nabla\mathfrak{l}(\theta_{k-1}), b_{k-1}\rangle + \frac{\eta_k^2 L}{2}\left\|\hat{\mathfrak{g}}_{k-1}\right\|^2\right. \\
&\leq \mathfrak{l}\left(\theta_{k-1}\right) - \eta_k\left|\left\|\nabla\mathfrak{l}(\theta_{k-1})\right\|^2 + \underbrace{\eta_k\left\|\nabla\mathfrak{l}(\theta_{k-1})\right\|\left\|b_{k-1}\right\|}_{\text{By Cauchy-Schwarz inequality}} + \frac{\eta_k^2 L}{2}\left\|\hat{\mathfrak{g}}_{k-1}\right\|^2\right. \\
&\leq \mathfrak{l}\left(\theta_{k-1}\right) - \eta_k\left|\left\|\nabla\mathfrak{l}(\theta_{k-1})\right\|^2 + \underbrace{\frac{\eta_k}{2}\left(\left\|\nabla\mathfrak{l}(\theta_{k-1})\right\|^2 + \left\|\ b_{k-1}\right\|^2\right)}_{\text{By AM-GM inequality}} + \frac{\eta_k^2 L}{2}\left\|\hat{\mathfrak{g}}_{k-1}\right\|^2\right.
\end{aligned} \tag{42}$$

Taking expectation of both sides,

$$\mathbb{E}\left[\mathfrak{l}(\theta_k) - \mathfrak{l}(\theta_{k-1})\right] \leq \mathbb{E}\left[\frac{-\eta_k}{2}\left\|\nabla\mathfrak{l}(\theta_{k-1})\right\|^2\right] + \frac{\eta_k}{2}G^{2\alpha}\tau^{2-2\alpha} + \frac{\eta_k^2 L}{2}G^{\alpha}\tau^{2-\alpha}. \tag{43}$$

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

$$\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\left\|\nabla\mathfrak{l}(\theta_{k-1})\right\|^2\right] \leq \frac{2\eta_k^{-1}}{2}\left(\mathfrak{l}(\theta_0) - \mathfrak{l}(\theta^*)\right) + G^{2\alpha}\tau^{2-2\alpha} + \eta_k L G^{\alpha}\tau^{2-\alpha}. \tag{44}$$

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

$$\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\left\|\nabla\mathfrak{l}(\theta_{k-1})\right\|^2\right] \leq \frac{2\eta_k^{-1}R_0}{T} + 2G^2\left(\eta_k L\right)^{\frac{2\alpha-2}{\alpha}}. \tag{45}$$

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

$$\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\left\|\nabla\mathfrak{l}(\theta_{k-1})\right\|^2\right] \leq 4G^{\frac{2\alpha}{3\alpha-2}}\left(\frac{R_0 L}{T}\right)^{\frac{2\alpha-2}{3\alpha-2}} \tag{46}$$

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

$$\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\left\|\nabla l(\theta_{k-1})\right\|^2 + \left\|\nabla g(\theta_{k-1})\right\|^2 - 2\langle\nabla l(\theta_{k-1}), \nabla g(\theta_{k-1})\rangle\right] \leq 4G^{\frac{2\alpha}{3\alpha-2}}\left(\frac{R_0 L}{T}\right)^{\frac{2\alpha-2}{3\alpha-2}}. \tag{47}$$

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

$$\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\left\|\nabla l(\theta_{k-1})\right\|^2+\left\|\nabla g(\theta_{k-1})\right\|^2+2\left\|\nabla l(\theta_{k-1})\right\|\left\|\nabla g(\theta_{k-1})\right\|\right]\leq 4G^{\frac{2\alpha}{3\alpha-2}}\left(\frac{R_0L}{T}\right)^{\frac{2\alpha-2}{3\alpha-2}}. \tag{48}$$

Therefore,

$$\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\left\|\nabla l(\theta_{k-1})\right\|^2\right]+\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\left\|\nabla g(\theta_{k-1})\right\|^2\right]\leq 4G^{\frac{2\alpha}{3\alpha-2}}\left(\frac{R_0L}{T}\right)^{\frac{2\alpha-2}{3\alpha-2}}. \tag{49}$$

Upon simplification,

$$\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\left\|\nabla l(\theta_{k-1})\right\|^2\right]\leq \mathcal{O}\left(G^{\frac{2\alpha}{3\alpha-2}}\left(\frac{R_0L}{T}\right)^{\frac{2\alpha-2}{3\alpha-2}}-\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\left\|\nabla g(\theta_{k-1})\right\|^2\right]\right) \tag{50}$$

which finishes the proof. □