Understanding the Role of Adversarial Regularization in Supervised Learning

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

Abstract

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

1. Introduction

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

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

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

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

Computer Vision and Pattern Recognition (CVPR) 2021 Workshop on Adversarial Machine Learning in Real-World Computer Vision Systems and Online Challenges.

tral channel attention in regularized adversarial learning to synthesize high resolution images. Emami et al. [7] coalesce spatial attention with adversarial regularization and feature map loss to perform image-to-image translation.

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

2. Preliminaries

2.0.1 Notations

Let $X \subset \mathbb{R}^{d_x}$ and $Y \subset \mathbb{R}^{d_y}$, where d_x and d_y denote input and output dimensions, respectively. The empirical distribution of X and Y are denoted by $\mathcal{P}_X$ and $\mathcal{P}_Y$. Given an input $x \in X$, $f(\theta; x) : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_y}$ is a neural network with rectified linear unit (ReLU) activation, which is common for both supervised and adversarial learning. Here, 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 $g : \mathbb{R}^{d_y} \rightarrow \mathbb{R}$, ∇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 $\forall \theta$, $\left\|\nabla^2 f(\theta)\right\| \leq \beta$.*

2.0.2 Problem Setup

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

$$\arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{P}} \left[l\left(f(\theta; x); y\right)\right]. \tag{1}$$

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

$$\arg\min_{\theta} -\mathbb{E}_{x\sim\mathcal{P}_X} \left[g\left(\psi; f\left(\theta; x\right)\right)\right] \tag{2}$$

and the discriminator cost function is given by,

$$\begin{aligned}\arg\min_{\psi}\ & \mathbb{E}_{x\sim\mathcal{P}_X} \left[g\left(\psi; f\left(\theta; x\right)\right)\right] - \mathbb{E}_{y\sim\mathcal{P}_Y} \left[g\left(\psi; y\right)\right] \\ & + \lambda_{GP}\, \mathbb{E}_{z\sim\mathcal{P}_Z} \left[\left(\|\nabla_z g\left(\psi; z\right)\| - 1\right)^2\right].\end{aligned} \tag{3}$$

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

$$\arg\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{P}} \left[l\left(f(\theta; x); y\right) - g\left(\psi; f\left(\theta; x\right)\right)\right]. \tag{4}$$

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

3. Theoretical Analysis

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

3.1. Mitigating Vanishing Gradient

The primary assumptions are stated as following.

Assumption 1. *The 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 the near optimal region. Different from standard smoothness in optimization, it is trivial to justify **Assumption 2** by relating it to a quadratic loss function[1]

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

Proof. Refer to Appendix C.1. □

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

[1] Please refer to Appendix D for numberical experiments confirming these assumptions in practice.

the issue of gradient descent inherently resides in the near optimal region[2]. We therefore ask a fundamental question: can we attain faster convergence without having to loose any empirical risk benefits? The following sections are intended to shed light in this direction.

Lemma 2. *Suppose **Assumption 1** holds. For a differentiable discriminator $g(\psi; y)$, if $\|g - g^*\| \leq \delta$, where $g^* \triangleq g(\psi^*)$ denote optimal discriminator, then $\|-\nabla_\theta \mathbb{E}_{x\sim\mathcal{P}_X}\left[g\left(\psi; f\left(\theta; x\right)\right)\right]\| \leq L\delta$.*

Proof. Refer to Appendix C.2. □

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

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

Proof. Refer to Appendix C.3. □

To focus more on the empirical success of adversarial regularization, we study a simple convex-concave minimax optimization problem. It will certainly be interesting to borrow some ideas from the vast minimax optimization literature in various other settings [22, 24]. According to **Theorem 1**, the expected gradient of augmented objective does not vanish in the near optimal region, i.e., $\|\Delta\theta\| \to L\delta$ as $\epsilon \to 0$. In the current setting, the estimated gradients of $l(\theta)$ and $-g(\theta)$ at any instant during the optimization process are positively correlated. Thus, the gradients of augmented objective is lower bounded by $\left\|\nabla_\theta \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[l\left(f(\theta; x); y\right) - g\left(\psi; f\left(\theta; x\right)\right)\right]\right\| \geq \left\|\nabla_\theta \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[l\left(f(\theta; x); y\right)\right]\right\|$. The upper and lower bounds of the intermediate iterates justify non-vanishing gradient in the near optimal region. It is important to heed the fact that supervised learning with adversarial regularization sets a more stringent criterion, which requires convergence of both primary and secondary objectives. In a smooth-convex-concave setting, which is not necessarily true in the deep learning paradigm, $\epsilon \to 0$ promotes the reduction of δ that makes the generator close to optimal generator. Although this results in vanishing gradients, the stringent convergence criterion would have already accelerated gradient updates in the augmented objective. This will be verified in the following sections. Having mitigated the vanishing gradient issue, it seems natural to wonder whether adversarial regularization improves iteration complexity.

[2]This issue of vanishing gradient is different from the vanishing gradient pheonomenon in the initial layers of a very deep feedforward network. It exists even after residual skip connections that solves the latter.

3.2. Asymptotic Iteration Complexity

In this section, we analyze global iteration complexity of sole supervision and the augmented objective[45, 3]. The analysis is restricted to a deterministic setting. For a sequence of parameters $\{\theta_k\}_{k\in\mathbb{N}}$, the complexity of a function $l(\theta)$ is defined as

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

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 asymptotic bounds under a less restrictive setting as introduced by Zhang et al. [45]. The new condition is weaker than commonly used Lipschitz smoothness assumption. Under this condition, Zhang et al. [45] aim to resolve the mystery of why adaptive gradient methods converge faster. We use this theoretical tool to study the asymptotic convergence bounds. To circumvent tractability issues in non-convex optimization, we follow the common practice of seeking an 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 upon the assumptions made in [45]. To put more succinctly, let us recall the assumptions.

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

Assumption 4. *The function is twice differentiable.*

Assumption 5 ((L_0, L_1)-Smoothness)**.** *The function is (L_0, L_1)-smooth, i.e., there exist positive constants L_0 and L_1 such that $\left\|\nabla^2 l\left(\theta\right)\right\| \leq L_0 + L_1\left\|\nabla l\left(\theta\right)\right\|$.*

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

Proof. Refer to Appendix C.4. □

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

Proof. Refer to Appendix C.5. □

Assumption 6 (Existence of useful gradients)**.** *For arbitrarily small $\zeta > 0$, the norm of the gradients of the discriminator is lower bounded by 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 minimax optimization problems. Also, it is trivial to prove this in the inner maximization loop under concave setting. In other words, the stated assumptions are mild and derived from prior analyses for the purpose of maintaining consistency with existing literature. Next, we analyze the global iteration complexity in the adversarial setting.

Theorem 3. *Suppose the functions in $\mathscr{L}$ 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 $\mathcal{O}\left(\frac{(l(\theta_0)-l^*)\left(L_0+L_1L^2\beta\epsilon\right)}{\epsilon^2+2\epsilon\zeta-L^2\delta^2}\right)$.*

Proof. Refer to Appendix C.6. □

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

Proof. Refer to Appendix C.7. □

Since $2\epsilon\zeta - L^2\delta^2 \geq 0$, the augmented objective has a *tighter* global iteration complexity compared to sole supervision. In a simplified setup, one can easily verify this hypothesis by using first order Taylor's approximation as given by **Corollary 1** and **2**. In this case, $h\zeta\epsilon > 0$ 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., $\|g - g^*\| \leq \delta \leq \frac{\sqrt{2\epsilon\zeta}}{L}$, adversarial acceleration is guaranteed. Notably, the empirical risk and iteration complexity benefit from this provided the discriminator and the generator are trained alternatively as typically followed in practice.

3.3. Sub-Optimality Gap

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

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

Proof. Refer to Appendix C.8. □

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

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

Proof. Refer to Appendix C.9. □

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

4. Concluding Remarks

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

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017. 20

[2] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018. 1, 7

[3] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, pages 1–50, 2019. 3

[4] Emily L Denton, Soumith Chintala, Rob Fergus, et al. Deep generative image models using laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pages 1486–1494, 2015. 1

[5] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 2, 17

[6] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011. 1, 7

[7] Hajar Emami, Majid Moradi Aliabadi, Ming Dong, and Ratna Babu Chinnam. Spa-gan: Spatial attention gan for image-to-image translation. *arXiv preprint arXiv:1908.06616*, 2019. 2

[8] Brendan J Frey and Delbert Dueck. Mixture modeling by affinity propagation. In *Advances in neural information processing systems*, pages 379–386, 2006. 22

[9] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007. 22

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 7

[11] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017. 20

[12] Mikael Henaff, Alfredo Canziani, and Yann LeCun. Model-predictive policy learning with uncertainty regularization for driving in dense traffic. *arXiv preprint arXiv:1901.02705*, 2019. 2, 7, 17

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1

[14] Levent Karacan, Zeynep Akata, Aykut Erdem, and Erkut Erdem. Learning to generate images of outdoor scenes from attributes and semantic layouts. *arXiv preprint arXiv:1612.00215*, 2016. 1

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 1

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[17] Teuvo Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990. 23

[18] Akira Kudo, Yoshiro Kitamura, Yuanzhong Li, Satoshi Iizuka, and Edgar Simo-Serra. Virtual thin slice: 3d conditional gan-based super-resolution for ct slice interval. In *International Workshop on Machine Learning for Medical Image Reconstruction*, pages 91–100. Springer, 2019. 1

[19] Simon Lacoste-Julien, Mark Schmidt, and Francis Bach. A simpler approach to obtaining an o (1/t) convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012. 7, 8

[20] Charis Lanaras, José Bioucas-Dias, Silvano Galliani, Emmanuel Baltsavias, and Konrad Schindler. Super-resolution of sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 146:305–319, 2018. 7

[21] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 1, 7

[22] Tianyi Lin, Chi Jin, Michael Jordan, et al. Near-optimal algorithms for minimax optimization. *arXiv preprint arXiv:2002.02417*, 2020. 3, 7

[23] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 22

[24] Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2703–2717. SIAM, 2018. 3, 7

[25] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 1

[26] Vaishnavh Nagarajan and J Zico Kolter. Generalization in deep networks: The role of distance from initialization. In *Neural Information Processing Systems (NeurIPS) Workshop, Deep Learning: Bridging Theory and Practice*, 2017. 1, 8

[27] Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012. 7

[28] J v Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928. 7

[29] Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. The role of over-parametrization in generalization of neural networks. In *Proceedings of Intenational Conference on Learning Represenations (ICLR)*, 2019. 1, 8, 9

[30] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 1

[31] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 20

[32] Litu Rout. Alert: Adversarial learning with expert regularization using tikhonov operator for missing band reconstruction. *IEEE Transactions on Geoscience and Remote Sensing*, 2020. 2, 7

[33] Litu Rout, Indranil Misra, S Manthira Moorthi, and Debajyoti Dhar. S2a: Wasserstein gan with spatio-spectral laplacian attention for multi-spectral band synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshop*, 2020. 1

[34] Muhammad Sarmad, Hyunjoo Jenny Lee, and Young Min Kim. Rl-gan-net: A reinforcement learning agent controlled gan network for real-time point cloud shape completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5898–5907, 2019. 2, 7, 17

[35] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017. 7

[36] Yibing Song, Chao Ma, Xiaohe Wu, Lijun Gong, Linchao Bao, Wangmeng Zuo, Chunhua Shen, Rynson WH Lau, and Ming-Hsuan Yang. Vital: Visual tracking via adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8990–8999, 2018. 1

[37] Matthew Staib, Sashank J Reddi, Satyen Kale, Sanjiv Kumar, and Suvrit Sra. Escaping saddle points with adaptive gradient methods. *arXiv preprint arXiv:1901.09149*, 2019. 7

[38] Tijmen Tieleman and Geoffrey Hinton. 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, 2012. 1

[39] A.M. Turing. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 237(641):37–72, 1952.

[40] Xiaolong Wang and Abhinav Gupta. Generative image modeling using style and structure adversarial networks. In *European Conference on Computer Vision*, pages 318–335. Springer, 2016. 1

[41] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 1

[42] Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems*, pages 4148–4158, 2017. 1

[43] Wenqi Xian, Patsorn Sangkloy, Varun Agrawal, Amit Raj, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Texturegan: Controlling deep image synthesis with texture patches. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8456–8465, 2018. 1

[44] Yuan Xue, Tao Xu, Han Zhang, L Rodney Long, and Xiaolei Huang. Segan: Adversarial network with multi-scale l-1 loss for medical image segmentation. *Neuroinformatics*, 16(3-4):383–392, 2018. 1, 2, 3, 7, 17

[45] Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2019. 3

[46] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank J Reddi, Sanjiv Kumar, and Suvrit Sra. Why adam beats sgd for attention models. *arXiv preprint arXiv:1912.03194*, 2019. 7, 13

[47] Dongruo Zhou, Yiqi Tang, Ziyan Yang, Yuan Cao, and Quanquan Gu. On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*, 2018. 7

[48] Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic nested variance reduction for nonconvex optimization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 3925–3936. Curran Associates Inc., 2018. 7

[49] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1

A. More Related Works

A.1. Adversarial Regularization

Although the improvement in empirical risk is minimal, recent studies on conditional generative models have shown a significant gain in iteration complexity. The spectral and spatial super resolution based on adversarial regularization is proven to achieve *faster convergence* and *better empirical risk* compared to purely supervised learning [20, 32]. Further, Ledig et al. [21] show improvement in perceptual quality of high resolution images in the adversarial setting. Despite superior 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 [44]. This inhibits erroneous gradient estimation by the discriminator that apparently improves perceptual quality. However, these benign properties of loss surface do not fully explain the practical observations. The present account is intended to provide further insights to this problem.

Apart from supervised learning, the notion of adversarial regularization has also been studied in Reinforcement Learning (RL). Henaff et al. [12] use adversarial learning with expert regularization to learn a predictive policy that allows to drive in a simulated dense traffic. Sarmad et al. [34] use RL agent controlled GAN and L_2-loss to convert noisy, partial point cloud into high-fidelity data.

A.2. Accelerated Gradients

The idea of accelerated training has long been an interesting area of research. An elegant line of work focuses on variance reduction that aims to address stochastic and finite sum problems by averaging the stochastic noise [35, 48]. Among momentum based acceleration, much theoretical progress has been made to accelerate any smooth convex optimization [27, 2]. Further, many efforts have been made towards changing the step size across iterations based on estimated gradient norm [6, 37, 47]. Adversarial regularization is similar to these methods in a sense that it offers acceleration in the near optimal region.

A.3. Minimax Optimization

The seminal work of Neumann [28] 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 [10]. 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 [22, 24] that may encourage further investigation from an algorithmic point of view. It will be useful to borrow some ideas from the vast literature of minimax optimization under less restrictive setting.

B. Omitted Theoretical Analysis

B.1. 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 [19]. 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 the rates for global clipping. One may wish to analyze this for coordinate-wise clipping [46].

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

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.10. □

Observe that by eliminating the discriminator and setting $\alpha = 2$, we recover exactly the SGD rate, i.e., $\mathcal{O}\left(\frac{G^2}{\mu T}\right)$ [19]. Thus, the augmented objective converges in strongly-convex-strongly-concave setting. It is determined by the convergence of the inner maximization loop.

Theorem 7 (Nonconvex-nonconcave convergence)**.** *Suppose **Assumption 7** holds. Let* $l(\theta_k) \triangleq l(\theta_k) - g(\theta_k)$ *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* $\tau_k = G(\eta_k L)^{\frac{-1}{\alpha}}$ *and constant step size* $\eta_k = \left(\frac{R_0^\alpha L^{2-2\alpha}}{G^2 T^\alpha}\right)^{\frac{1}{3\alpha-2}}$*, where* $R_0 = l(\theta_0) - l(\theta^*)$*, the sequence* $\{\theta_k\}$ *satisfies*

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

Proof. Refer to Appendix C.11. □

By setting $\alpha = 2$ and discarding adversarial acceleration, we obtain the standard SGD rate, $\mathcal{O}\left(\frac{G}{\sqrt{T}}\right)$. It is important that adversarial regularization converges even under nonconvex-nonconcave setting. 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.

B.2. Generalization Error

Motivated by the role of over-parametrization in generalization [26, 29], we study generalization error in the augmented objective. We use Rademacher complexity to get a bound on the generalization error. Since it depends on hypothesis class, we use a set of restricted parameters of trained networks to get a tighter bound. 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. The restricted hypothesis class then becomes

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

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 $1 - \delta$.

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

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

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

Relative Generalization Error: We define relative generalization error as

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

To be consistent with [29] 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 [29], one can easily prove that the Rademacher complexity of $l\ o\ \mathcal{F}_{\mathcal{W}}$ is bounded by

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

Adapted to current setting, the generalization error becomes

$$\mathcal{O}\left(\left\|U^0\right\|_2\left\|V\right\|_F + \left\|U-U^0\right\|_F\left\|V\right\|_F + \sqrt{h}\right). \tag{5}$$

C. Technical Proofs

C.1. Proof of Lemma 1

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. □

C.2. Proof of Lemma 2

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. □

C.3. Proof of Theorem 1

By applying triangle inequality after simplification,

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

which completes the statement of the theorem. □

C.4. Proof of Theorem 2

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

$$\gamma(t) = t\theta_{k+1} + (1-t)\theta_k \forall t \in [0,1]. \tag{6}$$

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

$$\begin{aligned}l\left(\theta_{k+1}\right) &= l\left(\theta_k\right) + \nabla l\left(\theta_k\right)\left(\theta_{k+1}-\theta_k\right) + \frac{1}{2}\left(\theta_{k+1}-\theta_k\right)^T \nabla^2 l\left(\theta_k\right)\left(\theta_{k+1}-\theta_k\right) \\ &= l\left(\theta_k\right) - h_k\left\|\nabla l\left(\theta_k\right)\right\|^2 + \frac{1}{2}\left(\theta_{k+1}-\theta_k\right)^T \nabla^2 l\left(\theta_k\right)\left(\theta_{k+1}-\theta_k\right).\end{aligned} \tag{7}$$

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

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

We know by **Assumption 5**

$$\left\|\nabla^2 l\left(\theta\right)\right\| \leq L_0 + L_1\left\|\nabla l\left(\theta\right)\right\|. \tag{9}$$

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

$$\begin{aligned}l\left(\theta_{k+1}\right) &\leq l\left(\theta_k\right) - h_k\left\|\nabla l\left(\theta_k\right)\right\|^2 \\ &\qquad + \frac{h_k^2\left\|\nabla l\left(\theta_k\right)\right\|^2}{2}\int_0^1 \left(L_0 + L_1\left\|\nabla l\left(\gamma(t)\right)\right\|\right) dt \\ &\leq l\left(\theta_k\right) - h_k\left\|\nabla l\left(\theta_k\right)\right\|^2 \\ &\qquad + \frac{h_k^2\left\|\nabla l\left(\theta_k\right)\right\|^2}{2}\int_0^1 \left(L_0 + L_1 L^2\beta\epsilon\right) dt \\ &\leq l\left(\theta_k\right) - h_k\left\|\nabla l\left(\theta_k\right)\right\|^2 + \frac{h_k^2\left\|\nabla l\left(\theta_k\right)\right\|^2\left(L_0 + L_1L^2\beta\epsilon\right)}{2}.\end{aligned} \tag{10}$$

Let us choose $h_k = \frac{1}{L_0 + L_1L^2\beta\epsilon}$. Now,

$$\begin{aligned}l\left(\theta_{k+1}\right) &\leq l\left(\theta_k\right) - \frac{h_k\left\|\nabla l\left(\theta_k\right)\right\|^2}{2} \\ &\leq l\left(\theta_k\right) - \frac{\left\|\nabla l\left(\theta_k\right)\right\|^2}{2\left(L_0 + L_1L^2\beta\epsilon\right)}.\end{aligned} \tag{11}$$

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

$$\begin{aligned}\sum_{k=0}^{T-1} l\left(\theta_{k+1}\right) - l\left(\theta_k\right) &\leq \frac{-T\epsilon^2}{2\left(L_0 + L_1L^2\beta\epsilon\right)} \\ \implies T &\leq \frac{2\left(l(\theta_0) - l^*\right)\left(L_0 + L_1L^2\beta\epsilon\right)}{\epsilon^2}.\end{aligned} \tag{12}$$

Therefore, we get

$$\sup_{\theta_0 \in \left\{\mathbb{R}^{h\times d_x}, \mathbb{R}^{d_y \times h}\right\}, l \in \mathscr{L}} \mathcal{T}_\epsilon \left(A_h \left[l, \theta_0\right], l\right) = \mathcal{O}\left(\frac{\left(l(\theta_0) - l^*\right)\left(L_0 + L_1 L^2 \beta \epsilon\right)}{\epsilon^2}\right) \tag{13}$$

which finishes the proof. □

C.5. Proof of Corollary 1

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

$$l\left(\theta_{k+1}\right) = l\left(\theta_k\right) - h_k \left\|\nabla l\left(\theta_k\right)\right\|^2 \leq l\left(\theta_k\right) - h_k \epsilon^2. \tag{14}$$

By telescopic sum, $\sum_{k=0}^{T-1} l\left(\theta_{k+1}\right) - l\left(\theta_k\right) \leq -T h_k \epsilon^2$. So

$$\sup_{\theta_0 \in \left\{\mathbb{R}^{h\times d_x}, \mathbb{R}^{d_y \times h}\right\}, l \in \mathscr{L}} \mathcal{T}_\epsilon \left(A_h \left[l, \theta_0\right], l\right) = \mathcal{O}\left(\frac{\left(l\left(\theta_0\right) - l^*\right)}{h\epsilon^2}\right) \tag{15}$$

which finishes the proof. □

C.6. Proof of Theorem 3

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

$$\theta_{k+1} = \theta_k - h_k \nabla \left(l\left(\theta_k\right) - g\left(\psi; f\left(\theta_k; x\right)\right)\right). \tag{16}$$

Using Taylor's expansion, triangle inequality, and Cauchy-Schwarz inequality as in **Theorem 2**, we obtain

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

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

$$\begin{aligned} l\left(\theta_{k+1}\right) \leq\ & l\left(\theta_k\right) - h_k \left\|\nabla l\left(\theta_k\right)\right\|^2 - h_k \left\|\nabla l\left(\theta_k\right)\right\| \zeta \\ & + \frac{h_k^2 \left\|\nabla l\left(\theta_k\right) - \nabla g\left(\psi; f\left(\theta_k; x\right)\right)\right\|^2}{2} \int_0^1 \left(L_0 + L_1 \left\|\nabla l(\gamma(t))\right\|\right) dt. \end{aligned} \tag{18}$$

Upon simplification using arguments of **Theorem 2** and applying Minkowski's inequality,

$$\begin{aligned} l\left(\theta_{k+1}\right) \leq\ & l\left(\theta_k\right) - h_k \left\|\nabla l\left(\theta_k\right)\right\|^2 - h_k \left\|\nabla l\left(\theta_k\right)\right\| \zeta \\ & + \frac{h_k^2 \left(\left\|\nabla l\left(\theta_k\right)\right\|^2 + \left\|\nabla g\left(\psi; f\left(\theta_k; x\right)\right)\right\|^2\right)}{2} \left(L_0 + L_1 L^2 \beta \epsilon\right). \end{aligned} \tag{19}$$

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

$$\begin{aligned} l\left(\theta_{k+1}\right) &\leq l\left(\theta_k\right) - \frac{h_k \left\|\nabla l\left(\theta_k\right)\right\|^2}{2} - h_k \left\|\nabla l\left(\theta_k\right)\right\| \zeta + \frac{h_k \left\|\nabla g\left(\psi; f\left(\theta_k; x\right)\right)\right\|^2}{2} \\ &\leq l\left(\theta_k\right) - \frac{h_k \left\|\nabla l\left(\theta_k\right)\right\|^2}{2} - h_k \left\|\nabla l\left(\theta_k\right)\right\| \zeta + \frac{h_k L^2 \delta^2}{2}, \text{ (\textbf{Lemma 2}).} \end{aligned} \tag{20}$$

Assuming T iterations to find an 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,

$$\begin{aligned} \sum_{k=0}^{T-1} l\left(\theta_{k+1}\right) - l\left(\theta_k\right) &\leq \frac{-T\left(\epsilon^2 + 2\epsilon\zeta - L^2\delta^2\right)}{2\left(L_0 + L_1 L^2 \beta \epsilon\right)} \\ \implies T &\leq \frac{2\left(l(\theta_0) - l^*\right)\left(L_0 + L_1 L^2 \beta \epsilon\right)}{\epsilon^2 + 2\epsilon\zeta - L^2\delta^2}. \end{aligned} \tag{21}$$

Therefore, we obtain

$$\sup_{\theta_0 \in \left\{\mathbb{R}^{h \times d_x}, \mathbb{R}^{d_y \times h}\right\}, l \in \mathscr{L}} \mathcal{T}_\epsilon \left(A_h \left[l, \theta_0\right], l\right) = \mathcal{O}\left(\frac{\left(l(\theta_0) - l^*\right)\left(L_0 + L_1 L^2 \beta \epsilon\right)}{\epsilon^2 + 2\epsilon\zeta - L^2\delta^2}\right) \tag{22}$$

which finishes the proof. □

C.7. Proof of Corollary 2

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

$$\begin{aligned} l\left(\theta_{k+1}\right) &= l\left(\theta_k\right) - h_k \left\|\nabla l\left(\theta_k\right)\right\|^2 - h_k \left\|\nabla l\left(\theta_k\right)\right\| \left\|\nabla g\left(\psi; f\left(\theta_k; x\right)\right)\right\| \\ &\leq l\left(\theta_k\right) - h_k \epsilon^2 - h_k \epsilon \zeta. \end{aligned} \tag{23}$$

By telescopic sum, $\sum_{k=0}^{T-1} l\left(\theta_{k+1}\right) - l\left(\theta_k\right) \leq -T h_k \epsilon^2 - T h_k \epsilon \zeta$. 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. □

C.8. Proof of Theorem 4

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

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

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

C.9. Proof of Theorem 5

In supervised learning with adversarial regularization, the parameters are updated by $\frac{d\theta(t)}{dt} = -\nabla l(\theta(t)) + \nabla g(\theta(t))$. Using arguments of **Theorem 4**, we obtain

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

Since $l(.)$ is convex and $g(.)$ is concave, 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. □

C.10. Proof of Theorem 6

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

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

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:*

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

One can easily prove this using **Lemma 2** of [46]. 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

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

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

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

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

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

which finishes the proof. □

C.11. Proof of Theorem 7

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

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

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

$$\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\left\|\nabla l(\theta_{k-1})\right\|^2 + \left\|\nabla g(\theta_{k-1})\right\|^2 + 2\left\|\nabla l(\theta_{k-1})\right\|\left\|\nabla g(\theta_{k-1})\right\|\right] \leq 4G^{\frac{2\alpha}{3\alpha-2}}\left(\frac{R_0 L}{T}\right)^{\frac{2\alpha-2}{3\alpha-2}}. \tag{48}$$

Therefore,

$$\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\left\|\nabla l(\theta_{k-1})\right\|^2\right] + \frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\left\|\nabla g(\theta_{k-1})\right\|^2\right] \leq 4G^{\frac{2\alpha}{3\alpha-2}}\left(\frac{R_0 L}{T}\right)^{\frac{2\alpha-2}{3\alpha-2}}. \tag{49}$$

Upon simplification,

$$\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\left\|\nabla l(\theta_{k-1})\right\|^2\right] \leq \mathcal{O}\left(G^{\frac{2\alpha}{3\alpha-2}}\left(\frac{R_0 L}{T}\right)^{\frac{2\alpha-2}{3\alpha-2}} - \frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\left\|\nabla g(\theta_{k-1})\right\|^2\right]\right) \tag{50}$$

which finishes the proof. □

D. Experiments

The experimental section aims to answer the following questions. How does adversarial regularization (a) mitigate vanishing gradients in the near optimal region? (b) accelerate training? (c) achieve tighter sub-optimality gap? (d) converge under practical setting? and (e) cast doubts on the generalization measure?

D.1. Implementation Details

Unless specified otherwise, the experiments are conducted on a two layer neural network with ReLU activation function. For completeness, we also experiment with practical neural network architectures. We do not use weight decay, dropout, or normalization. In these settings, 13 architectures are trained with the number of hidden units ranging from 2^3 to 2^{15}. We use SGD with momentum for training the networks. 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.

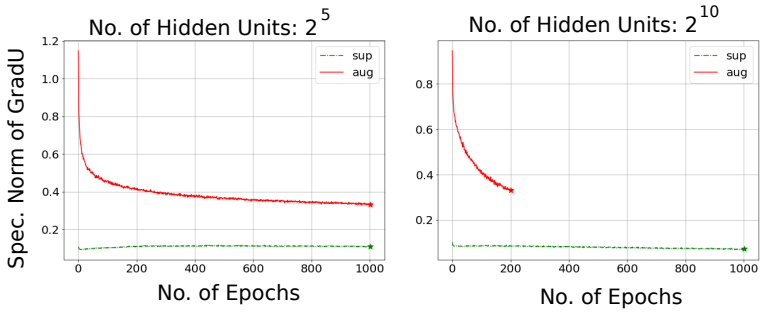

Figure 1. Comparison of gradients between supervised (sup) and augmented (aug) objective in the *hidden layer* (GradU) on MNIST. Adversarial regularization mitigates vanishing gradient issue.

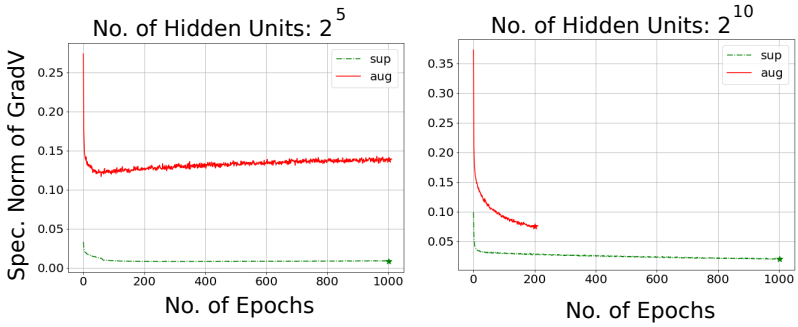

Figure 2. Comparison of gradients between supervised (sup) and augmented (aug) objective in the *top layer* (GradV) on MNIST. Adversarial regularization mitigates vanishing gradient issue.

D.1.1 MNIST Dataset

We use SGD with momentum 0.9, batch size 64, and a fixed learning rate of 0.01 on MNIST dataset. Here, $n = 60000$ samples of size $[28 \times 28]$ are used in training and 10000 samples are used in testing. The convergence criterion is set to be the mean square error of 0.001 or a maximum of 1000 epochs.

D.1.2 CIFAR10 Dataset

On CIFAR10 dataset, we use 50000 samples of size $[32 \times 32 \times 3]$ in training and remaining 10000 samples in testing. All the hyperparameters are same as on MNIST except for the convergence error of 0.02.

D.1.3 Tiny Imagenet Dataset

Different from MNIST, the learning rate is set to be 0.001 and convergence error is chosen as 0.02. Out of 100000 samples of size $[64 \times 64 \times 3]$, we use $n = 90000$ in training and 10000 samples in testing.

D.2. Experimental Results

D.2.1 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 architectures, the spectral norm of the gradients estimated in the purely supervised objective is smaller than the augmented objective. This is consistent with the theoretical analyses presented in Section 3. The main reason for such non-vanishing gradient is the feedback from discriminator. As marked by $\star$ in Figure 1 and 2, the adversarial regularization is at least as good as sole supervision in terms of the number iterations required to attain convergence. In other words, it ensures faster convergence in an over-parameterized setting with potential improvement in accuracy.

Figure 3 offers empirical support to sub-optimality gap in the 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 the approaches, evidently neither of those meets the convergence criteria. However, as the capacity increases, the supervised cost which is

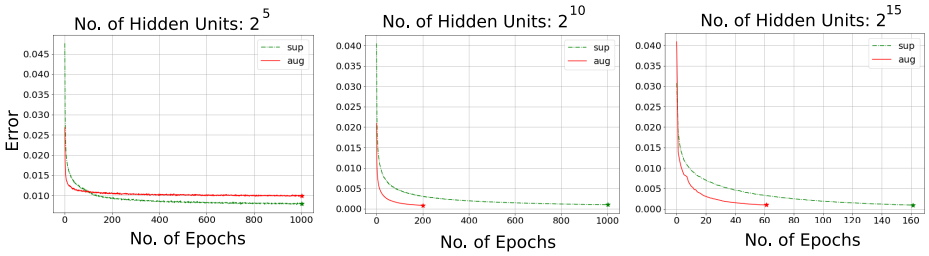

Figure 3. Comparison of optimal empirical risk on MNIST. Adversarial regularization converges faster.

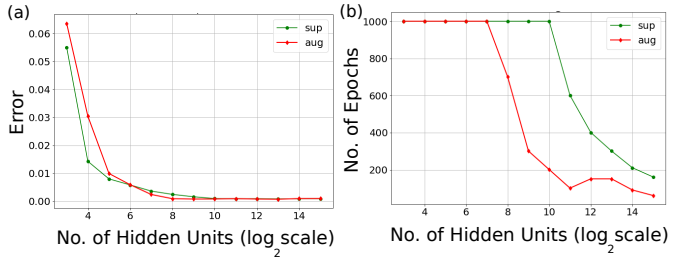

Figure 4. Comparison on MNIST. (a) Optimal empirical risk. (b) Iteration Complexity. Adversarial regularization attains tighter ϵ-stationary point at an optimal rate.

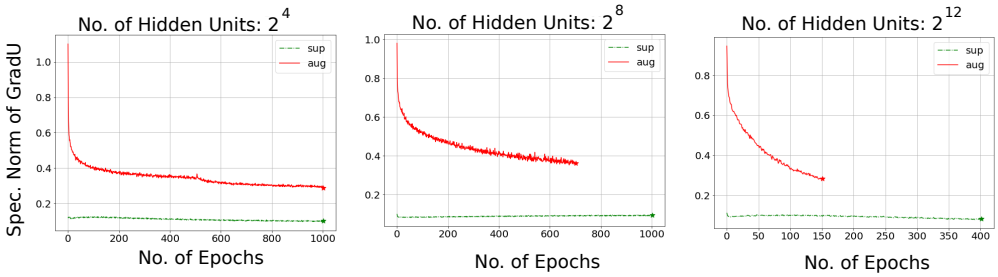

Figure 5. Comparison of gradient updates between supervised and augmented objective as observed in the *hidden layer* on MNIST.

common in both the 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 condition, the optimal empirical risk in the augmented objective can be provably better than sole supervision as predicted by the proposed theorems. Figure 3 supports this theory as the augmented objective consistently achieves better performance either by risk or by the 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 typically sought after in practice. Moreover, it is not hard to estimate δ in some rare occurences where the mapping function is approximated by the discriminator.

To this end, it is quite clear that the augmented objective achieves faster convergence as compared to the purely supervised objective. However, it is necessary to verify this hypothesis in other architectures. As shown in Figure 5 and 6, the estimated gradient 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 7 where we observe this particular phenomenon across a wide variety of architectures. Although the difference in empirical risk is minimal, 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. This provides a reasonable justification to the practical success of regularized adversarial learning [44, 5, 12, 34].

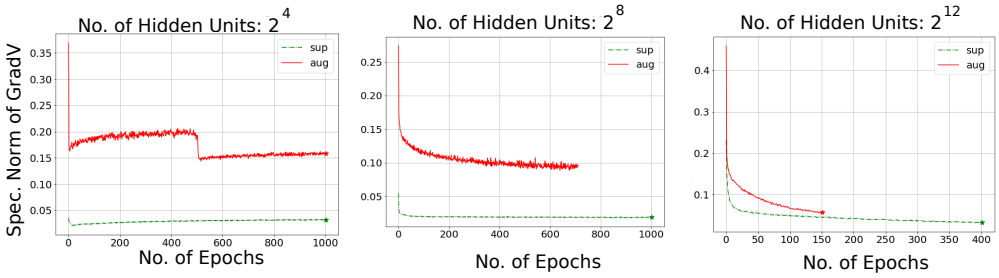

Figure 6. Comparison of gradient updates between supervised and augmented objective as observed in the *top layer* on MNIST.

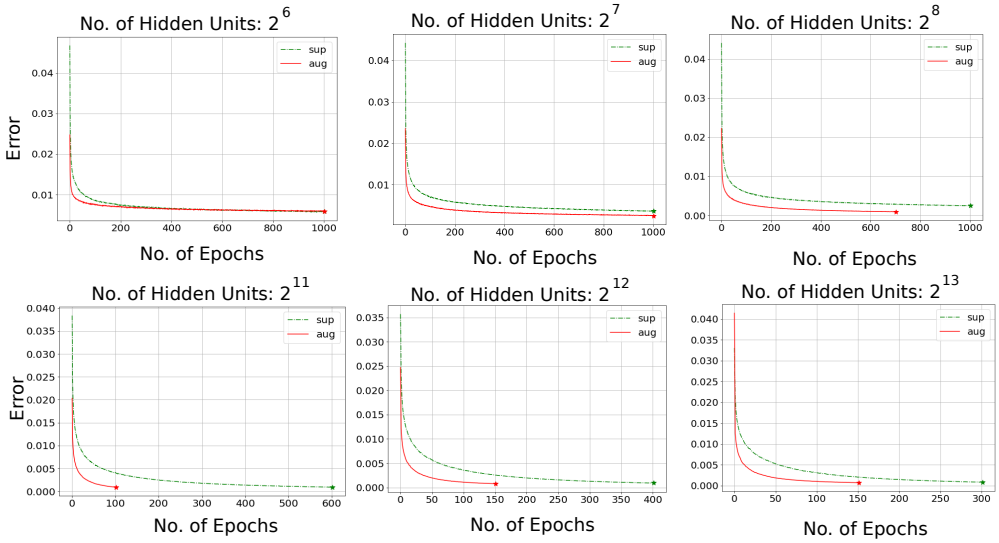

Figure 7. Comparison of optimal empirical risk on MNIST.

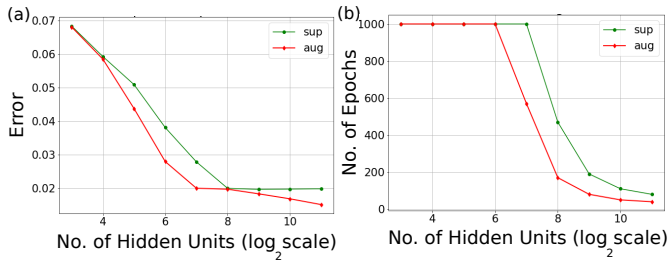

Figure 8. Comparison on CIFAR10. (a) Optimal empirical risk. (b) Iteration Complexity. Adversarial regularization attains tighter ϵ-stationary point at an optimal rate.

D.2.2 Results on CIFAR10

These theorems also justify the experiments conducted on CIFAR10 dataset. As shown in Figure 8, supervised learning with adversarial regularization performs better than sole supervision both in terms of optimal empirical risk and iteration complexity. Here, ϵ is found to be approximately equal to 0.06. Similar to MNIST, we observe the vanishing gradient issue on CIFAR10, which is shown in Figure 9 and 10. Figure 11 illustrates how model capacity correlates with empirical risk, and thereby satisfies the assumption of **Theorem 1**. Across a wide variety of architectures, the 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. Although both the methods start with almost identical initial empirical risk, as shown in Figure 11, the augmented objective allows to traverse through a shorter path and attain minimal risk upon convergence. It is to be noted that the slight difference in error at the begining is because of adversarial acceleration in the first step itself.

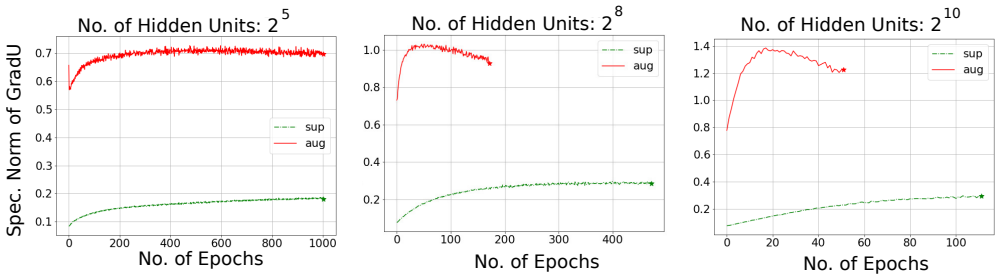

Figure 9. Comparison of gradient updates between supervised and augmented objective as observed in the *hidden layer* on CIFAR10.

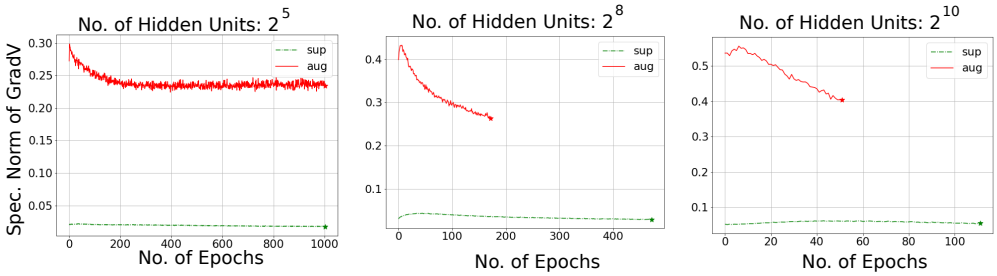

Figure 10. Comparison of gradient updates between supervised and augmented objective as observed in the *top layer* on CIFAR10.

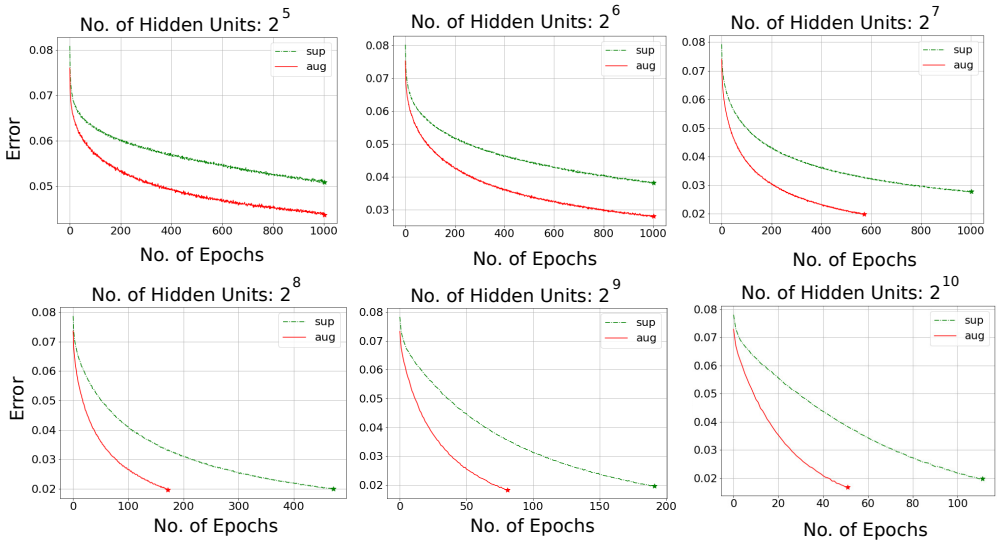

Figure 11. Comparison of optimal empirical risk on CIFAR10.

D.2.3 Results on Tiny ImageNet

To disentangle the effect of adversarial acceleration on a large-scale dataset, we experiment with several variants of adversarial networks. In this setting, the primary function approximator, $f(\theta; x)$ consists of 6 `Conv2d` layers with stride $(2, 2)$. The output is taken from a `Linear` layer with 200 classes on top of the final `Conv2d` layer. In practice, we observe that a shallow discriminator is usually sufficient to offer adversarial acceleration. We therefore choose a two layer fully connected network with 1024 hidden nodes. For optimization, ADAM is used with a learning rate of 0.001. Here, the discriminator is updated once for every single update of the generator.

Since the augmented objective requires training of the discriminator in addition to the generator, it takes more training time to reach convergence as given by WGAN in Table 1. However, when we introduce Gradient Penalty (GP) and Weight Clipping (WC), the ablation study suggests a significant acceleration in training to achieve similar performance. Interestingly, the proposed hypothesis of adversarial acceleration holds on several variants of adversarial training objectives and activation functions. With the results on three different datasets, we have empirically verified the robustness of the proposed theorems.

Architecture	No. Layers	Activation	No. Epochs (Runtime) Sup	No. Epochs (Runtime) Aug	Hypothesis
WGAN	7	ReLU	1000 (159 m)	1000 (366 m)	✓
WGAN + GP [11]	7	ReLU	1000 (159 m)	110 (64 m)	✓
WGAN + WC [1]	7	ReLU	1000 (159 m)	84 (32 m)	✓
DCGAN [31]	7	Sigmoid	1000 (159 m)	80 (29 m)	✓

Table 1. Hypothesis Testing on Various Adversarial Training Configurations.

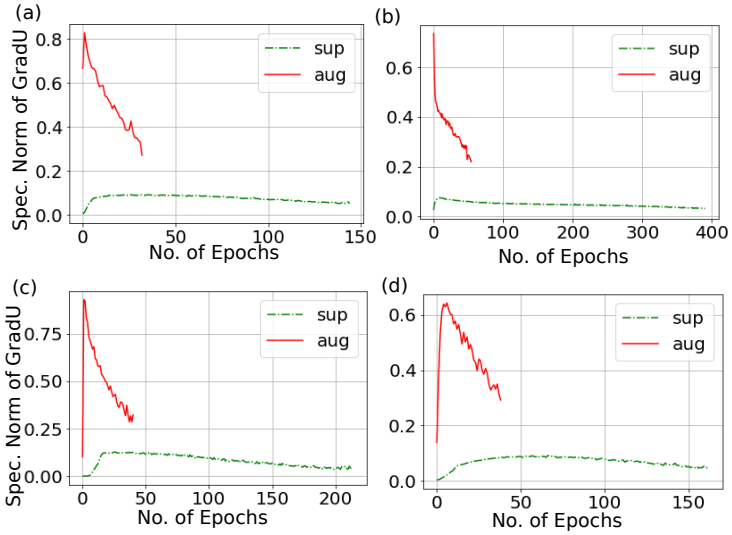

Figure 12. Comparison of gradient updates between supervised and augmented objective as observed in the *first layer* on MNIST. (a) Multi-Layer Perceptron. (b) Exponential Activation. (c) Residual Network. (d) Dense Network.

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	✓

Table 2. Hypothesis Testing on Various Generator Network Configurations.

Next, we verify this hypothesis on more practical generator networks.

D.3. Practical Architectures

To study the impact of these findings in realistic situations, we experiment on various generator network configurations. As shown in Figure 12 and 13, the issue of vanishing gradient is persistent across these experimented configurations[3]. Furthermore, the discussion on adversarial acceleration is also supported by Figure 14. In addition, Table 2 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 MLPs with Exponential Linear Unit (ELU) activations, convolution layers, skip connections, dense connections, L_1 regularized networks, and L_2 regularized networks. Thus, the augmented objective owes its benefits to adversarial learning at a fundamental level.

[3]The number of layers is reasonably high given the complexity of the MNIST dataset. Note that one can efficiently classify the MNIST digits by a two layer fully connected network.

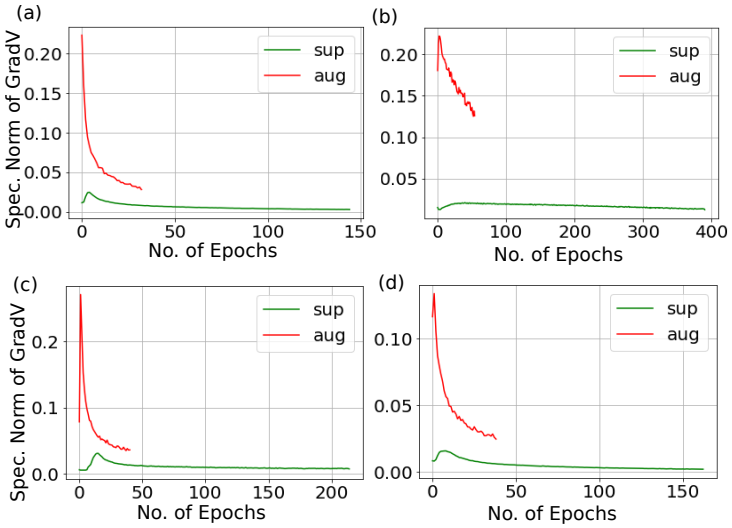

Figure 13. Comparison of gradient updates between supervised and augmented objective as observed in the *last layer* on MNIST. (a) Multi-Layer Perceptron. (b) Exponential Activation. (c) Residual Network. (d) Dense Network.

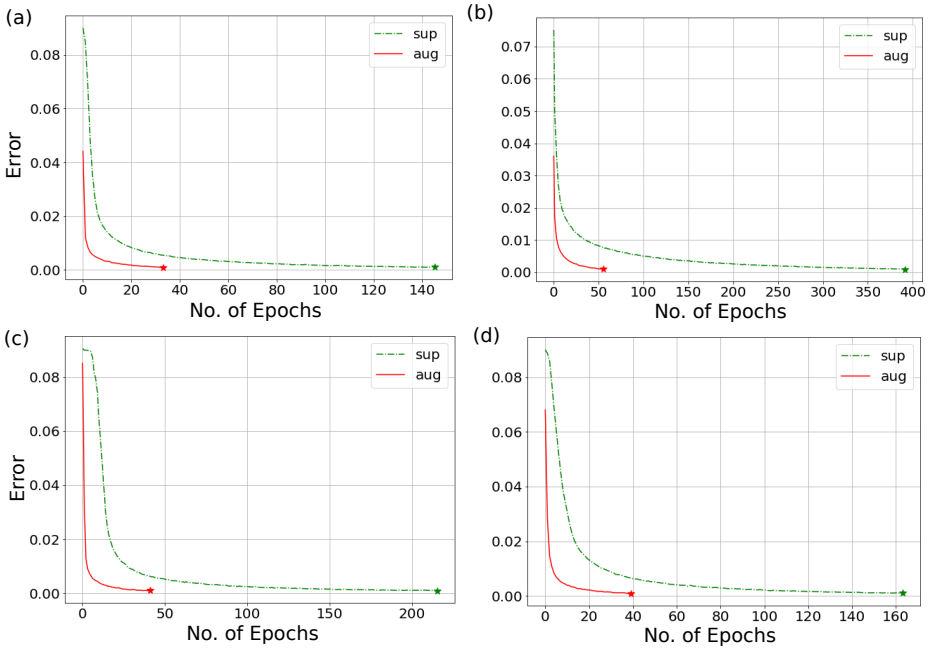

Figure 14. Comparison of optimal empirical risk on MNIST. (a) Multi-Layer Perceptron. (b) Exponential Activation. (c) Residual Network. (d) Dense Network.

D.4. Generalization Error

The generalization trend in sole supervision is shown in Figure 15(a) and 15(c). As per equation (5), the combined measure of the Frobenius norm of top layer (FrobV), i.e., $\|V\|_F$ and distance from initialization of hidden layer (FrobDisU), i.e., $\left\|U - U^0\right\|_F$ explains the generalization gap (Gen) on MNIST and CIFAR10. We verify this measure in our experimental setting and study whether it can explain generalization in adversarial learning. Recall 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 15(b) and 15(d), this bound does not fully explain the generalization error observed in the adversarial setting. In Figure 16, 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

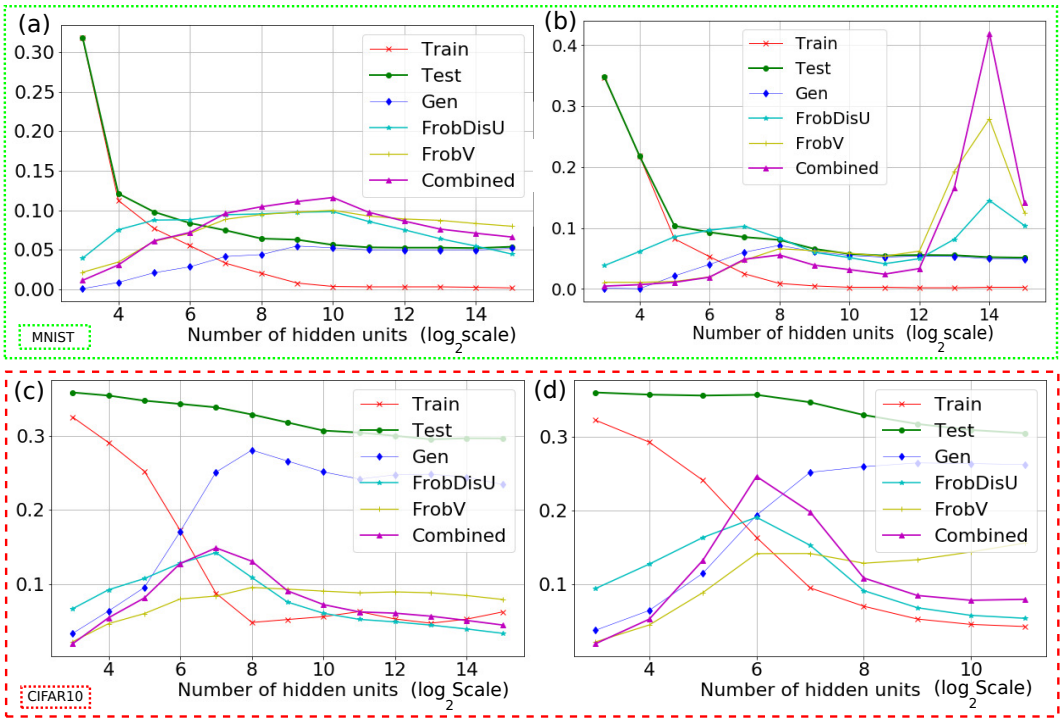

Figure 15. Generalization error on MNIST and CIFAR10. Adversarial training requires new generalization bound.

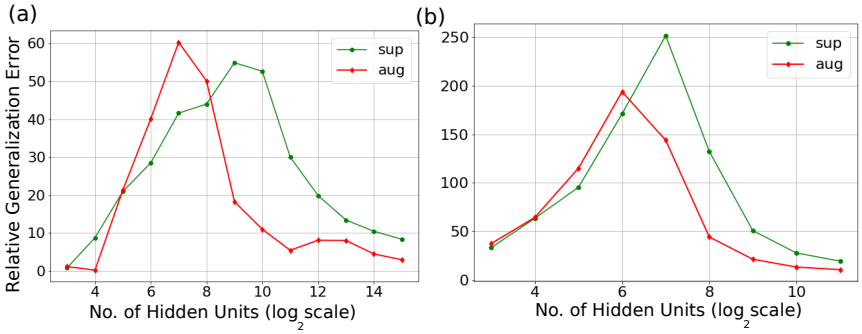

Figure 16. Relative generalization. (a) MNIST. (b) CIFAR10. Augmented objective has better relative generalization error.

convergence. We believe that the contribution of the discriminator needs to be efficiently characterized in the formulation of generalization gap, which we wish to explore as a part of the future work.

E. Discussion on Neural Topology Analysis

E.1. Implementation Details

In Neural Topology Analysis (NTA), we analyze the geometry of neurons present in the hidden layer 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) [8, 9]. The number of cluster centers represents total number of primary processing elements. 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 [23]. In the 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.

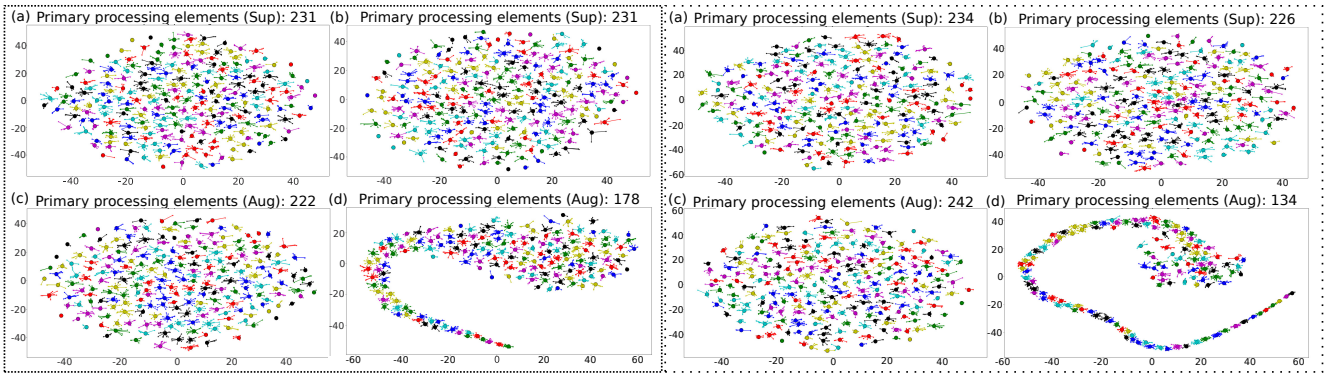

Figure 17. NTA in *hidden layer* (left) and *top layer* (right). (a) Initial and (b) final topology in supervised learning. (c) Initial and (d) final topology in adversarial learning.

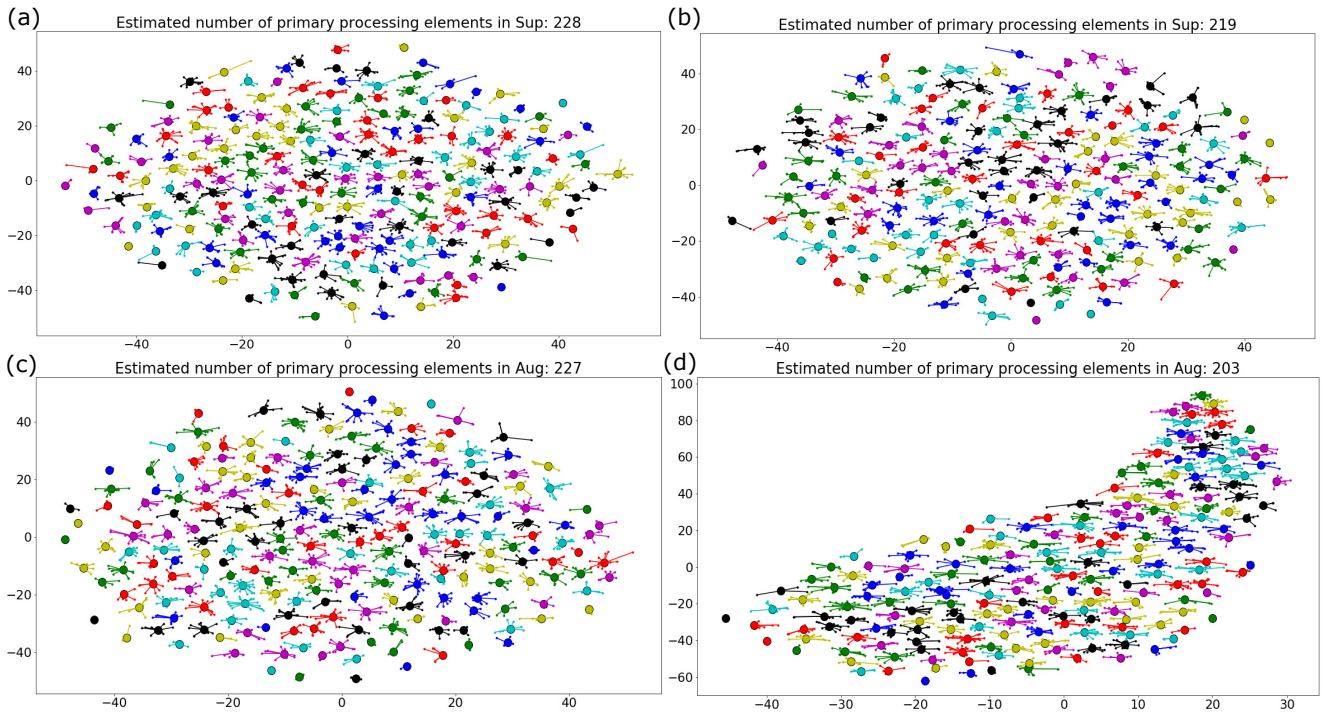

Figure 18. NTA in the *hidden layer* with 2^{15} hidden units. (a) Initial and (b) final topology in supervised learning. (c) Initial and (d) final topology in adversarial learning.

E.2. NTA on MNIST

In the experiments with 2^{14} hidden units, we observe emergence of evolutionary patterns in the adversarial framework. Figure 17 shows that despite similar topology at initialization, 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 [17]. We observe this sparse self-organization behavior on a wide variety of architectures, as shown in Figure 18, 19, 20, and 21. In all these configurations, adversarial learning tries to exploit sparsity in data to reorganize neurons.

Further, we study the neural topology of other fixed subsets of neurons in Figure 22 and 23. In this analysis, we focus on 4 subsets sequentially, each consisting of 2048 neurons. Since we repeatedly observe new patterns even with random seeds, it ensures that the organization of neurons has indeed changed drastically. Also, we analyze the topology of a randomly selected subset of 2048 neurons in Figure 24. In addition, Figure 25 shows emergence of *global pattern* in adversarial learning.

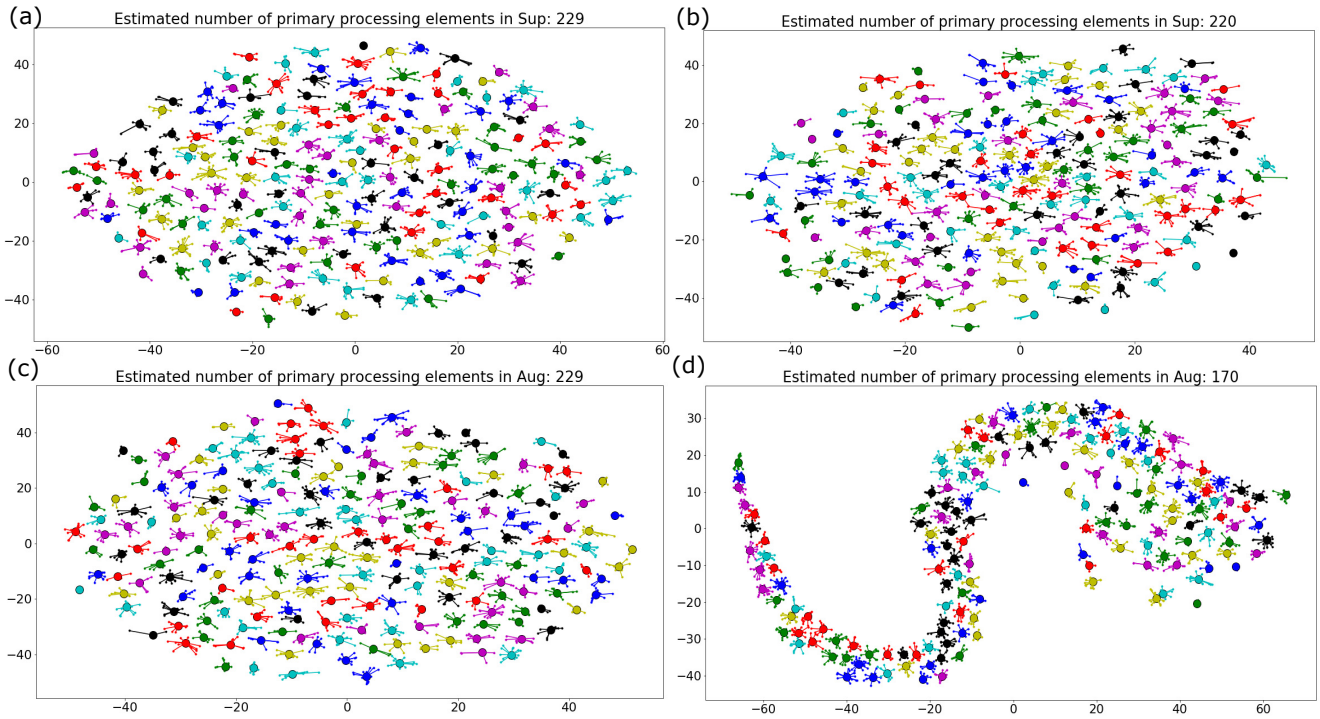

Figure 19. NTA in the *top layer* with 2^{15} hidden units. (a) Initial and (b) final topology in supervised learning. (c) Initial and (d) final topology in adversarial learning.

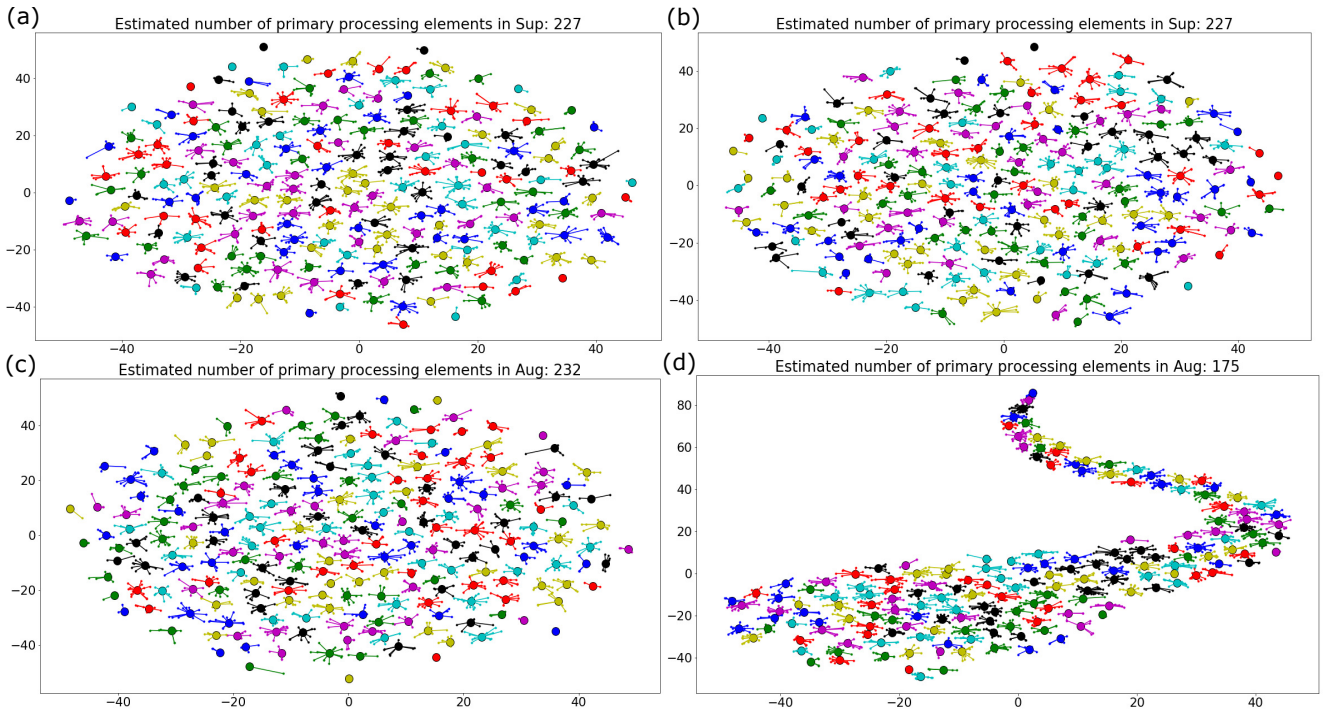

Figure 20. NTA in the *hidden layer* with 2^{13} hidden units. (a) Initial and (b) final topology in supervised learning. (c) Initial and (d) final topology in adversarial learning.

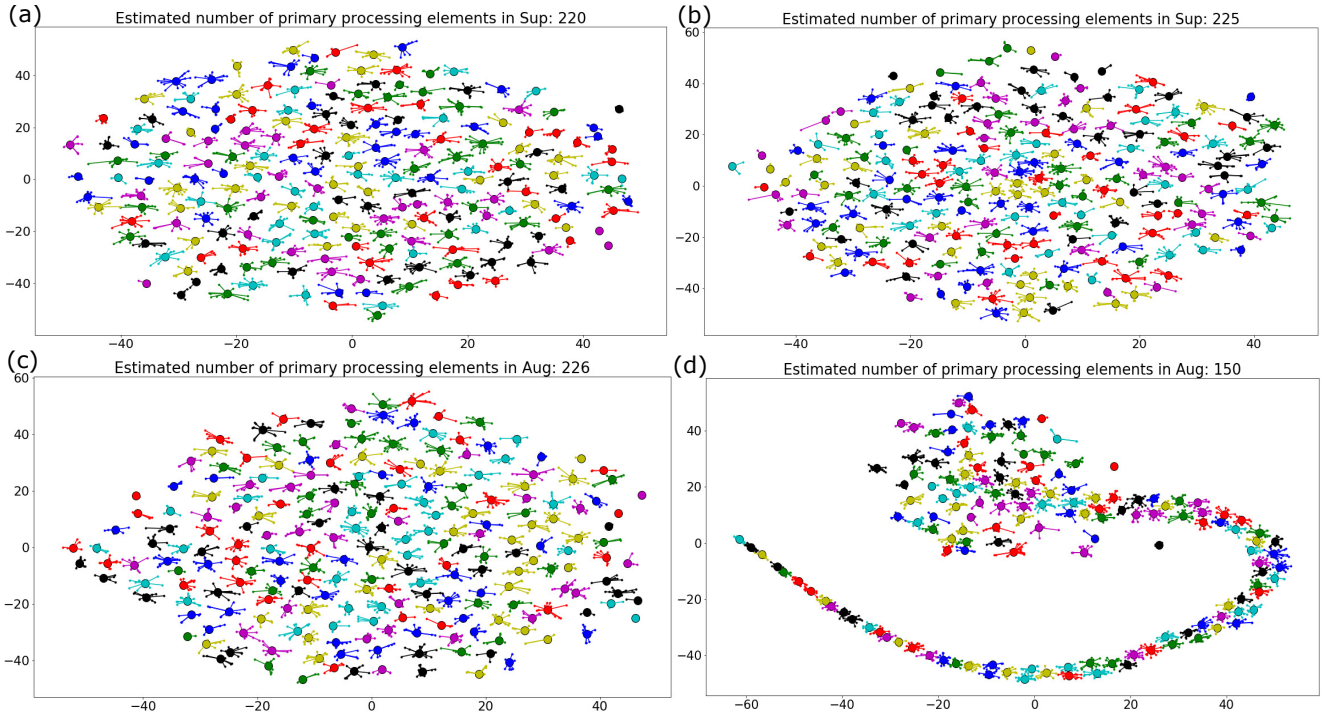

Figure 21. NTA in the *top layer* with 2^{13} hidden units. (a) Initial and (b) final topology in supervised learning. (c) Initial and (d) final topology in adversarial learning.

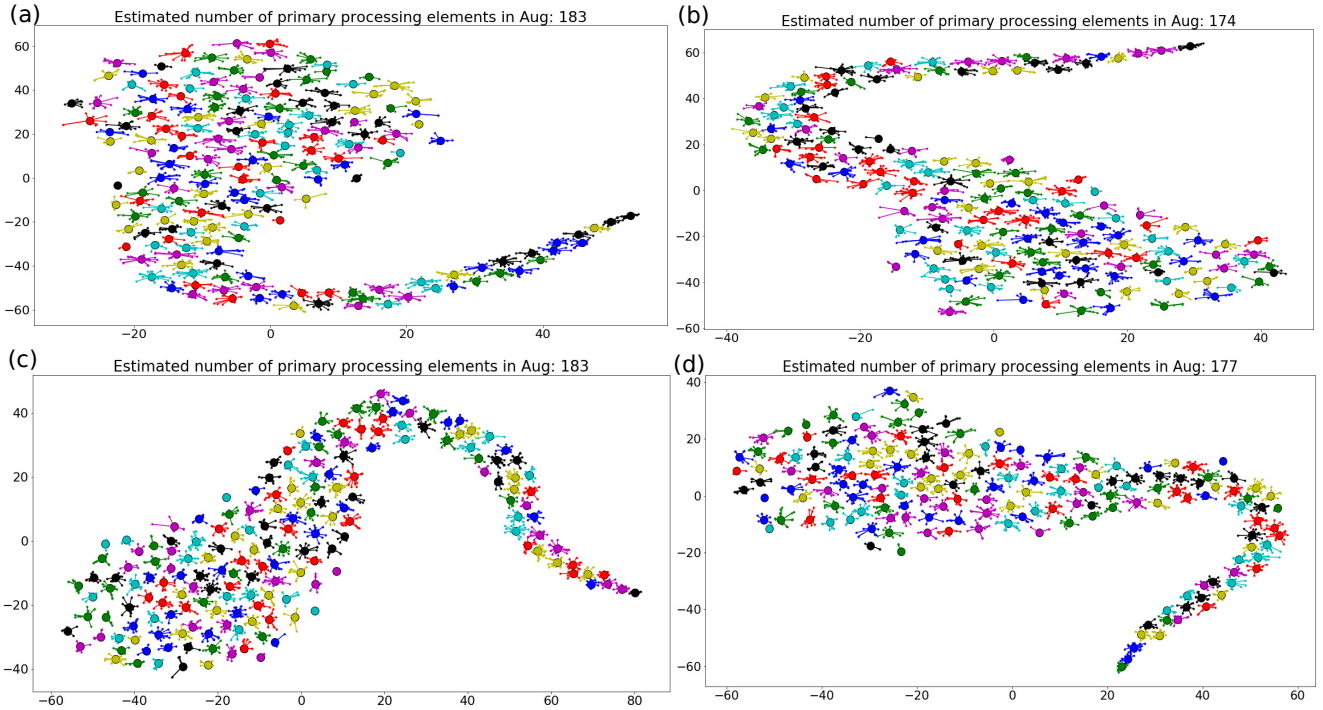

Figure 22. NTA in the *hidden layer* with 2^{13} hidden units. (a) First subset (0-2048) (b) Second subset (2048-4096) (c) Third subset (4096-6144) (d) Fourth subset (6144-8192) final topology in adversarial learning.

E.3. Perturbation Sensitivity

In Figure 26 and 27, we investigate the sensitivity of the topological diagrams to local perturbation. The perturbation model follows Gaussian distribution with mean and standard deviation same as that of the fully trained weights. Here, the

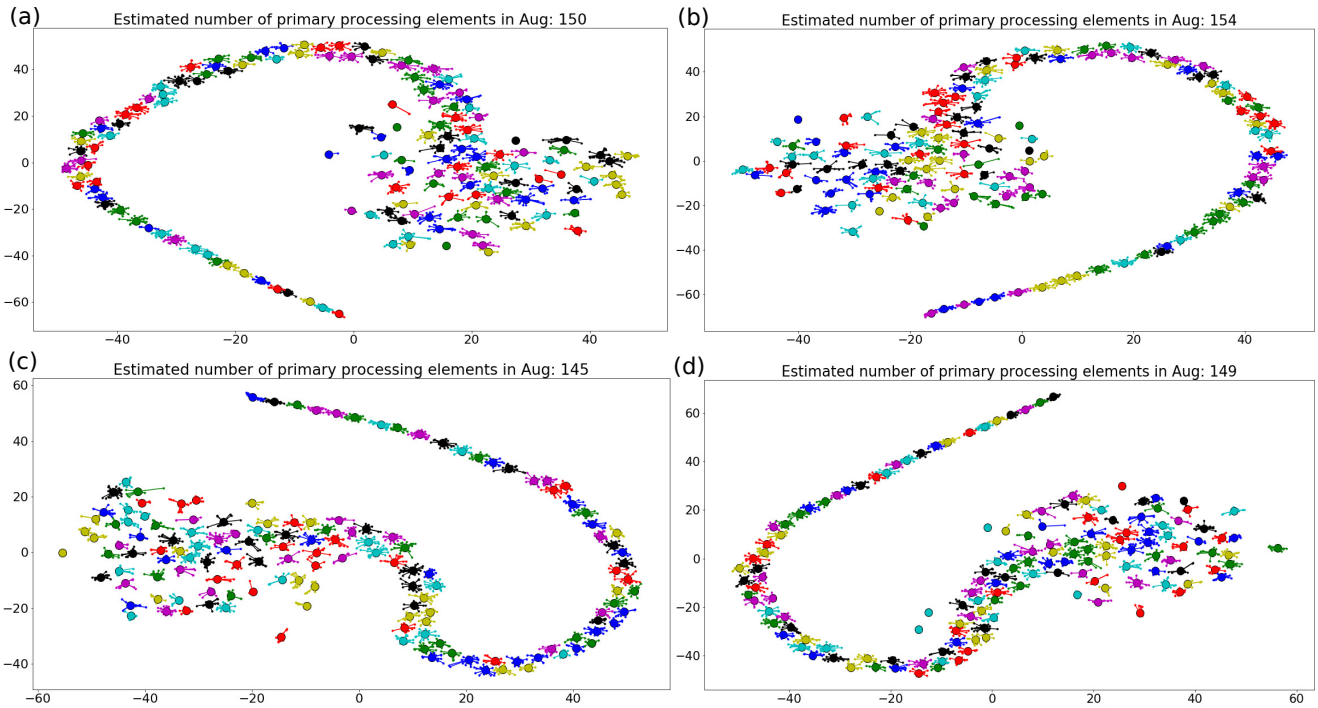

Figure 23. NTA in the *top layer* with 2^{13} hidden units. (a) First subset (0-2048) (b) Second subset (2048-4096) (c) Third subset (4096-6144) (d) Fourth subset (6144-8192) final topology in adversarial learning.

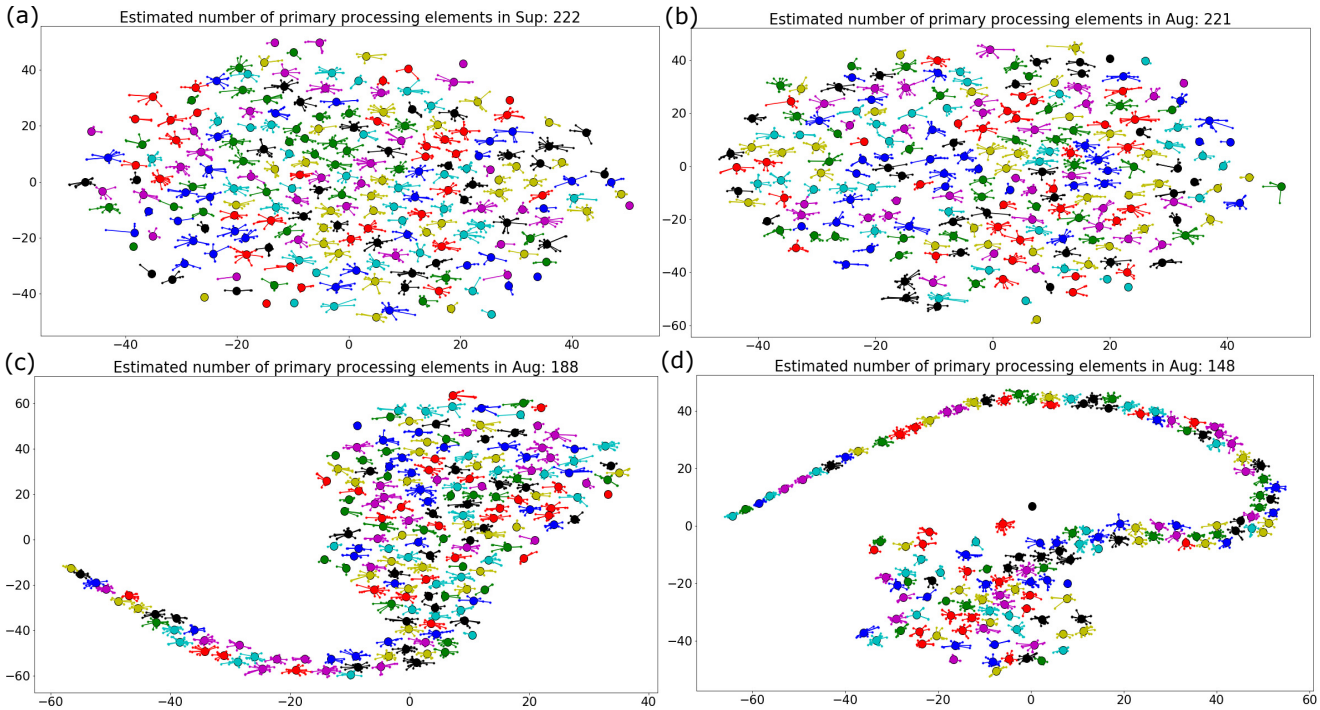

Figure 24. NTA of a random subset of 2048 neurons with 2^{13} hidden units.(a) Hidden and (b) top layer topology in supervised learning. (c) Hidden and (d) top layer topology in adversarial learning.

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. As shown in Figure 26 and 27, the final topology retains sparse representation with

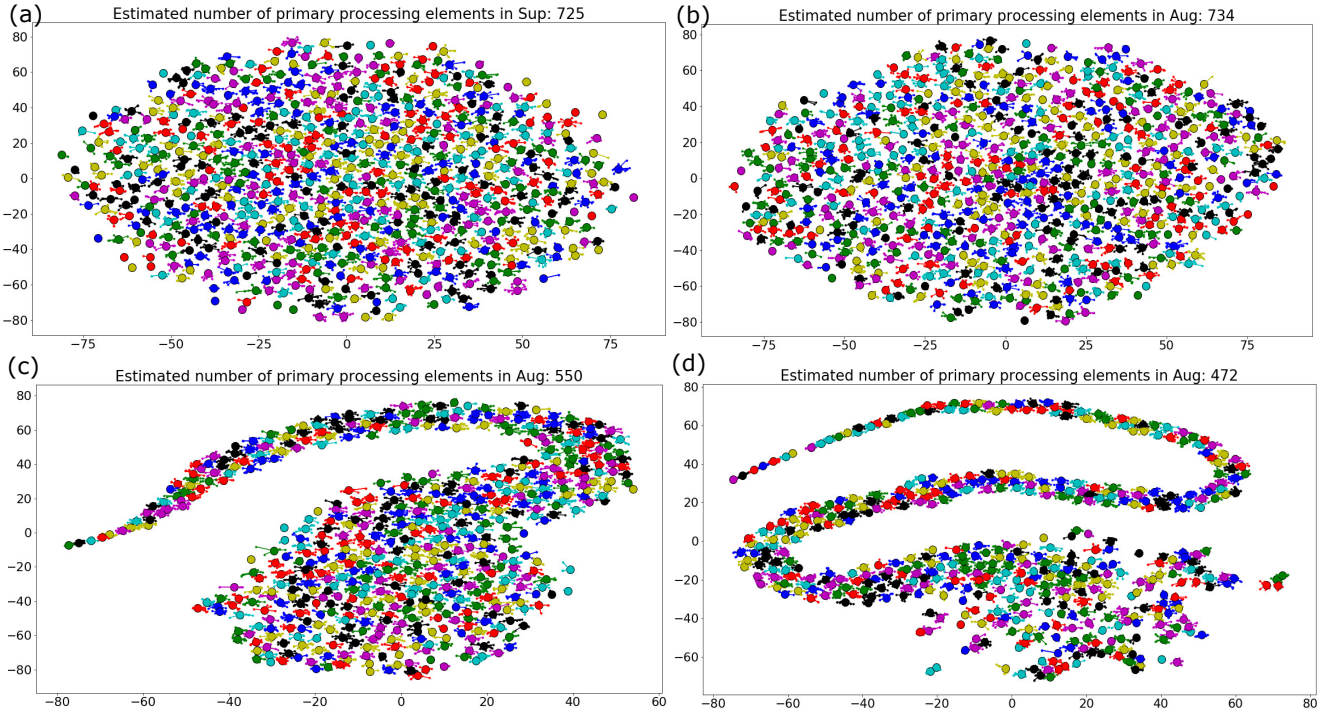

Figure 25. NTA of all 2^{13} hidden units.(a) Hidden and (b) top layer final topology in supervised learning. (c) Hidden and (d) top layer final topology in adversarial learning.

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 topology in augmented objective is not due to minor deviations from the neural topology of sole supervision. Thus, there is a significant difference between the final topology of adversarial regularization and sole supervision.

E.4. NTA on Over-Parameterization

In Figure 28, we study the neural topology of a network that is trained on randomly labelled pairs of MNIST dataset. With 2^{13} nodes in the 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.

E.5. NTA on FashionMNIST

Figure 29 and 30 demonstrate similar pattern formation on three different subsets of the hidden nodes. It is interesting to note that these patterns are quite different from the patterns observed in MNIST experiments. Nevertheless, these results favor the arguments on pattern formation due to adversarial interaction.

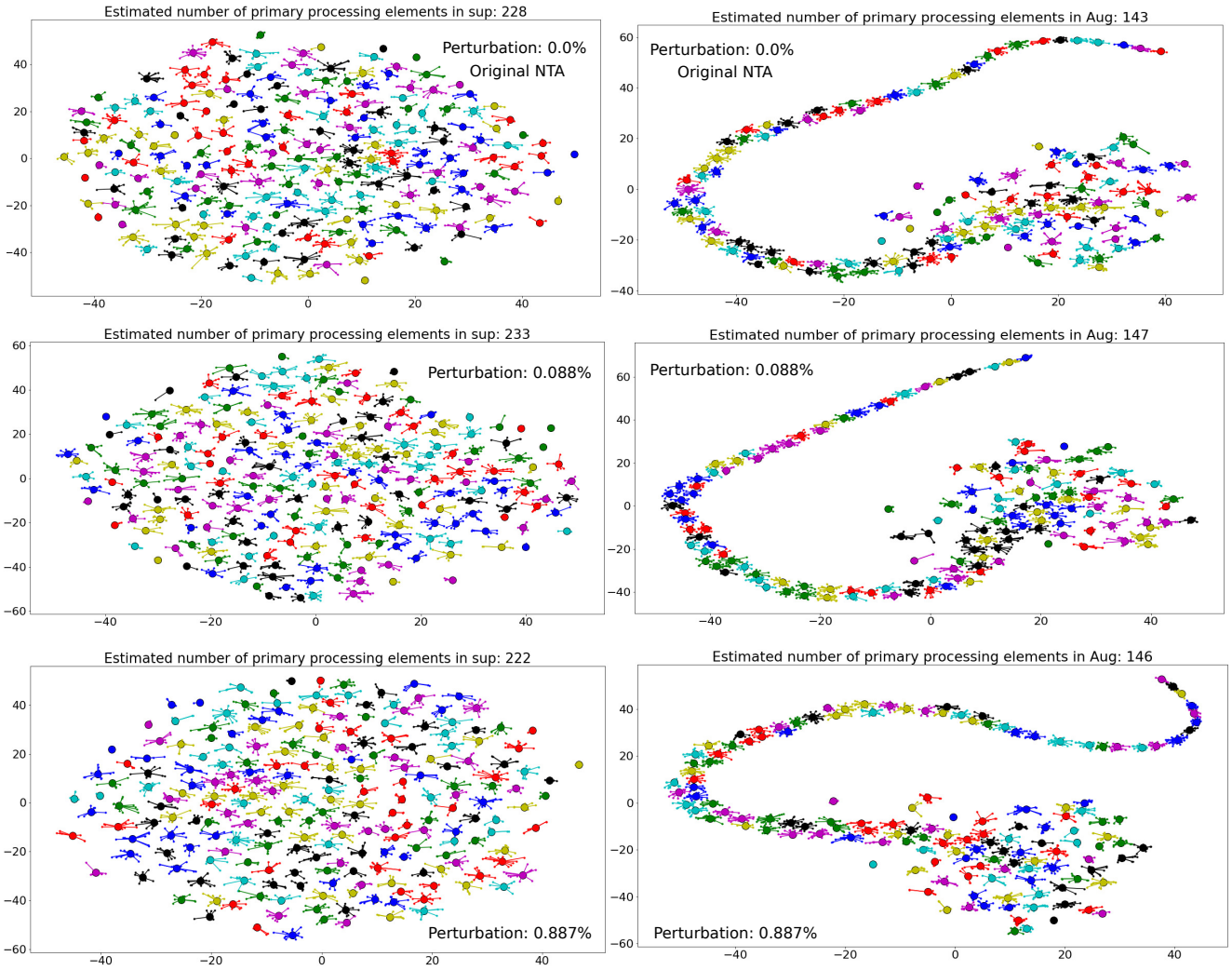

Figure 26. NTA in the *top layer* with 2^{13} hidden units. Comparison of sensitivity to low level Gaussian perturbation. Final topology in supervised learning (left) and adversarial learning (right).

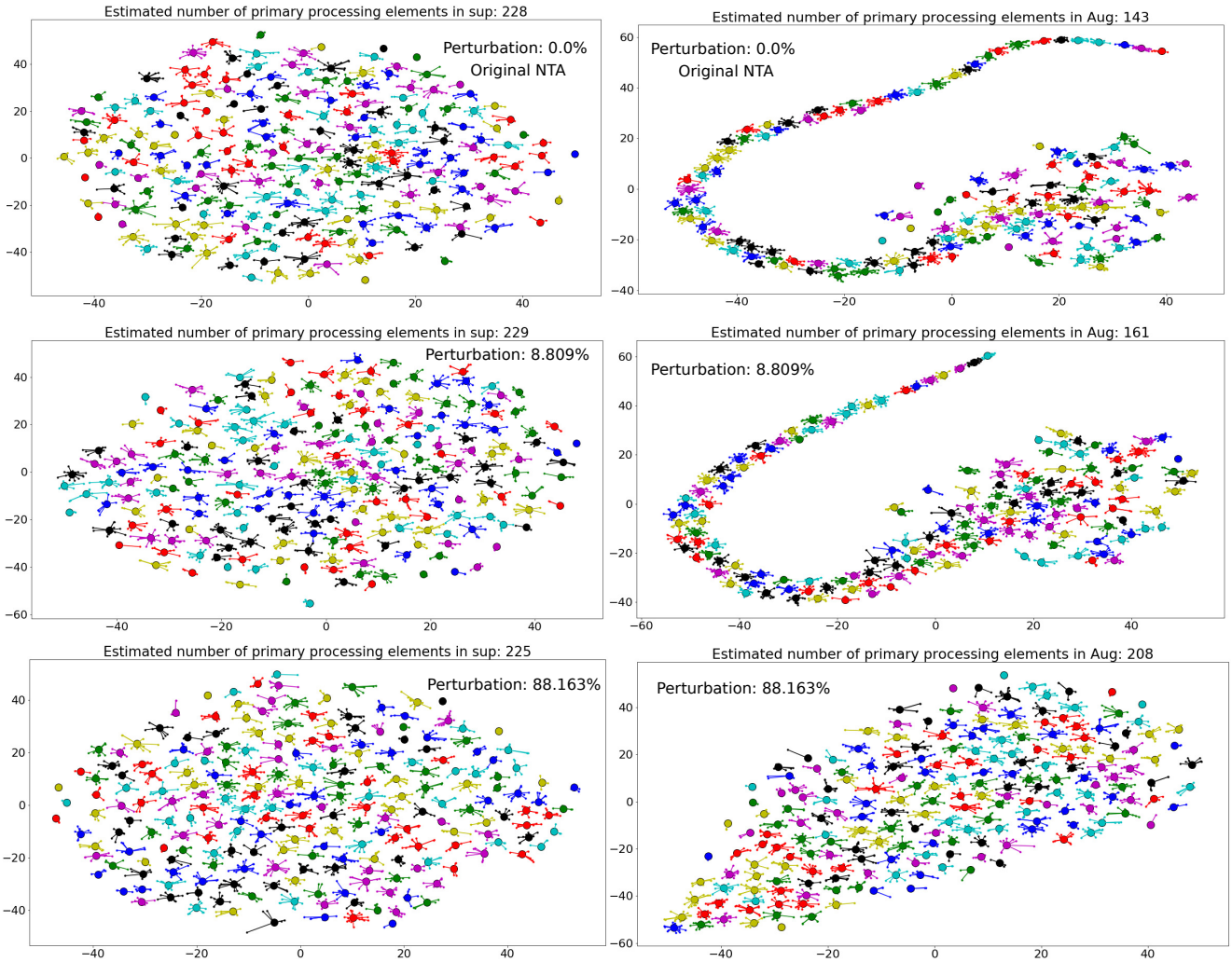

Figure 27. NTA in the *top layer* with 2^{13} hidden units. Comparison of sensitivity to moderate and extreme level Gaussian perturbation. Final topology in supervised learning (left) and adversarial learning (right).

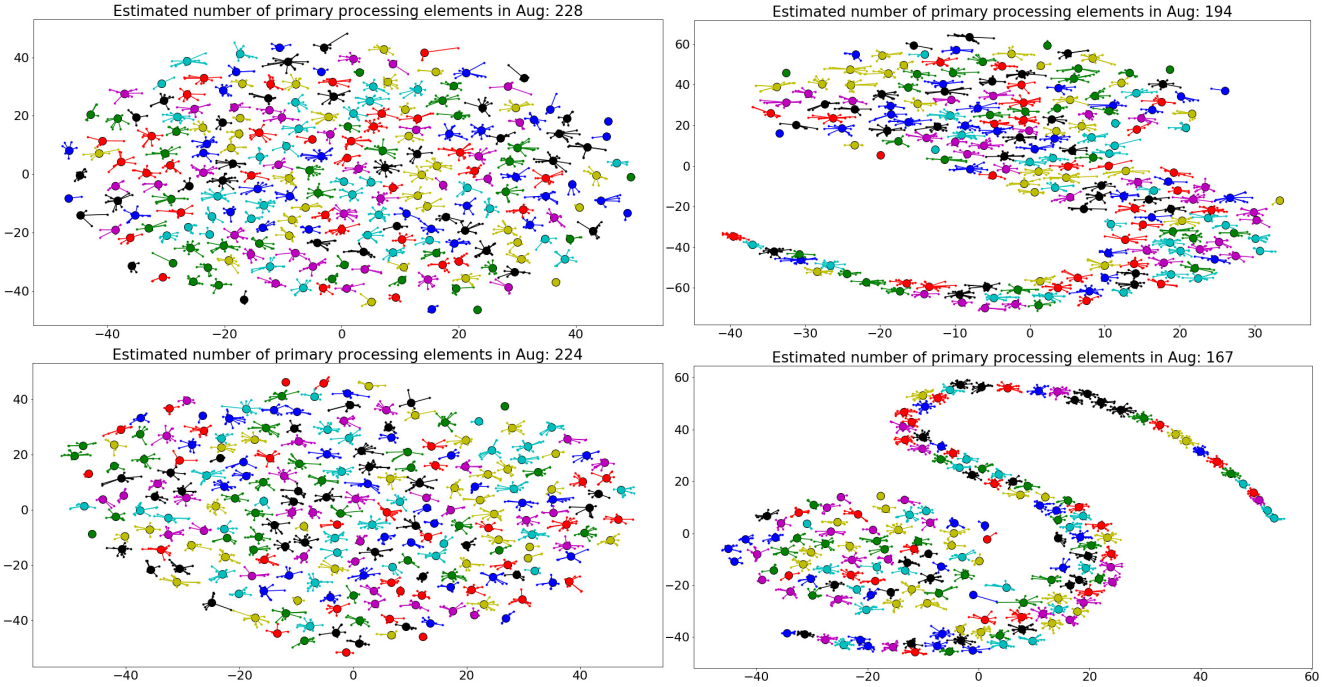

Figure 28. NTA in *adversarial learning* with 2^{13} hidden units. Initial and final topology in *hidden layer* (first row) and *top layer* (second row), respectively.

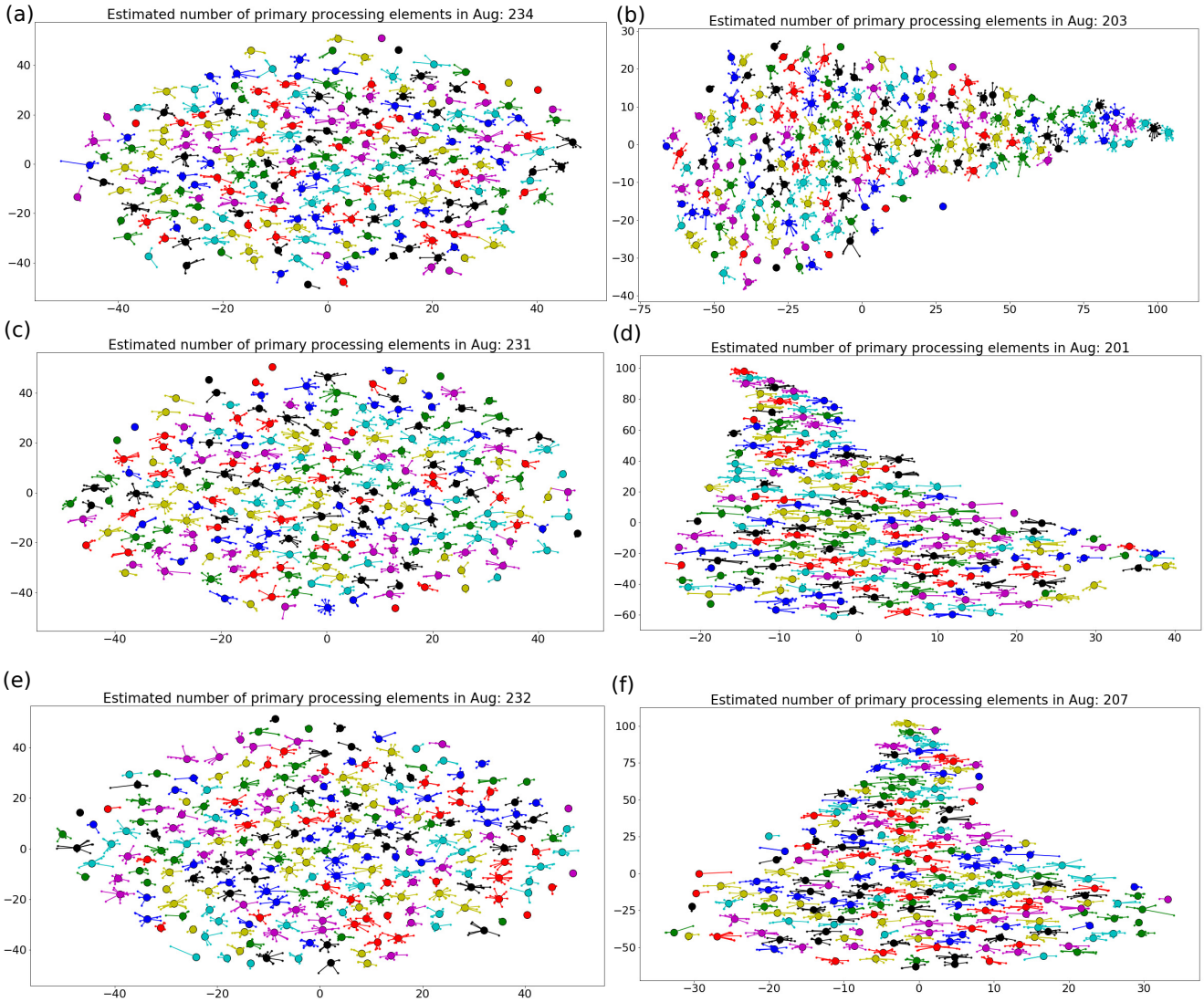

Figure 29. NTA in the *hidden layer* with 2^{13} hidden units on FashionMNIST. (a) First subset (0-2048) (b) Second subset (2048-4096) (c) Third subset (4096-6144) initial (left) and final (right) topology in adversarial learning.

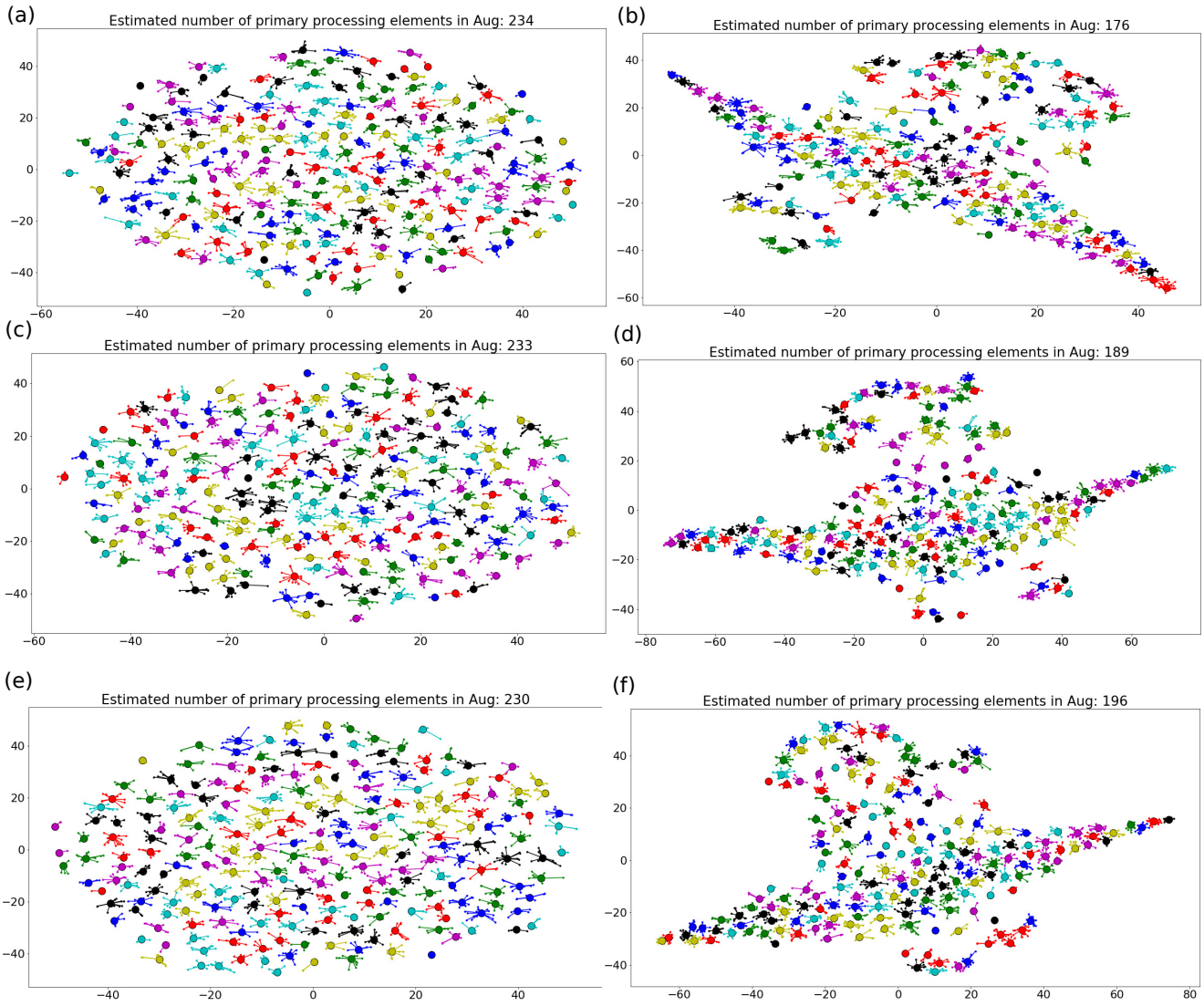

Figure 30. NTA in the *top layer* with 2^{13} hidden units on FashionMNIST. (a) First subset (0-2048) (b) Second subset (2048-4096) (c) Third subset (4096-6144) initial (left) and final (right) topology in adversarial learning.