Adversarial Training with Large Step Sizes: Implicit Bias and Evolution of Sharpness

author names withheld

Under Review for NExT-Game 2026

Abstract

Adversarial training (AT) can be modeled as a two-player zero-sum game. This game-theoretic characterization introduces additional challenges in analyzing its dynamics. Existing analyses of AT assume that the learner uses extremely small step sizes, which is unrealistic in practice. In this paper, we study AT dynamics under large step sizes, focusing on two aspects: the implicit bias of adversarial logistic regression and the sharpness evolution along AT trajectories. For the former, we show that adversarial logistic regression converges to a robust max-margin direction under arbitrary constant step sizes, generalizing the result of (Li et al., 2020), which requires an exponentially small step size. For the latter, we find that sharpness along AT trajectories exhibits a surprisingly regular pattern. Compared with the edge-of-stability phenomenon in standard training, this pattern contains an additional stage where sharpness oscillates while showing an overall decreasing trend.

1. Introduction

Adversarial training (AT) is an important approach for improving the robustness of machine learning models, such as neural networks [10]. AT is usually modeled as a zero-sum game between two players: the learner aims to minimize the training loss using optimization algorithms such as gradient descent, while the adversary aims to maximize the training loss by adding perturbations to the training data. This adversarial interaction makes the analysis of AT more complex, since the learner is effectively optimizing over a time-varying loss landscape that changes in an adversarial manner.

One of the challenges arising from analyzing the dynamics of AT is that, theoretically, an extremely small step size is usually needed to prove the implicit bias result of AT. For example, Li et al. [14] proved that gradient descent converge to a robust max-margin direction for adversarial logistic regression under step size $\eta = \mathcal{O}(e^{-d})$, where d is the dimensional of input. Such a requirement is unrealistic in high-dimensional settings. Similar small step sizes assumption is also needed in other works regarding to the implicit bias of adversarial logistic regression with different architectures like deep linear networks [16] and diagonal networks [17, 24]. Given the importance of understanding implicit bias of adversarial logistic regression and the huge gap between the small step sizes used in theoretical and large step size used in practice, we ask the following question:

Can we show that the implicit bias still exists in adversarial logistic regression with large step sizes?

Large step sizes are widely used in modern deep learning and often bring practical benefits [2, 5, 22, 26]. For standard training algorithms such as gradient descent, sharpness, defined as

the largest eigenvalue of the Hessian of the loss, follows a surprisingly regular pattern. This phenomenon, known as edge of stability [6], features an initial progressive sharpening phase followed by oscillations around a nearly constant value as the training loss converges to zero. Comparing with standard training, adversarial training is more complex since its trajectory follows a time-varying loss landscape induced by adversarial perturbations [15, 18]. Motivated by the edge-of-stability phenomenon in standard training, we ask:

Does adversarial training also exhibit a regular pattern of sharpness evolution?

1.1. Contributions.

In this paper, we provide affirmative answers to the above questions. In particular, our results can be summarized as follows:

Implicit Bias of ℓ_∞-adversarial Logistic Regression. We prove that, under standard separability assumptions, the gradient descent dynamics of adversarial logistic regression converge to a maximum mixed-norm margin direction between the ℓ_2 and ℓ_∞ norms for *arbitrary* constant step sizes. This result generalizes Li et al. [14] on the implicit bias of adversarial logistic regression, which requires an exponentially small step size dependent on dimension. Our approach is inspired by the recent results of Wu et al. [27] on the implicit bias of standard logistic regression with large step sizes.

Sharpness Evolution of Adversarial Training. We conduct experiments on adversarial neural network training to study how sharpness evolves along the training trajectory. We observe that, in the early stage, sharpness increases monotonically. After this phase, it begins to oscillate while exhibiting an overall decreasing trend. This behavior is markedly different from the edge-of-stability phenomenon in standard training [6], where sharpness oscillates around a constant level $2/\eta$ for step size η. To provide theoretical insight, we analyze an adversarial variant of the model in Ahn et al. [1], a canonical example for studying edge-of-stability. Our analysis shows that adversarial sharpness of this model converges to $\mathcal{O}(\eta)$, which is significantly smaller than the standard sharpness level $2/\eta$.

2. Preliminaries

2.1. Adversarial Logistic Regression

Let $L(\mathbf{w}) = \sum_{i=1}^N \ln\left(1+\exp(-y_i\mathbf{x}_i^\top\mathbf{w})\right)$ be the logistic loss, where $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ is the dataset. We consider the ℓ_∞ adversarial logistic training [1]:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \sum_{i=1}^N \max_{\|\delta_i\|_\infty\le c} \ln\left(1+\exp(-y_i(\mathbf{x}_i+\delta_i)^\top\mathbf{w})\right) \qquad \text{(Adversarial Logistic Loss)}$$

In this paper, we consider the gradient descent dynamics to solve (Adversarial Logistic Loss):

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\partial_{\mathbf{w}}\mathcal{L}_{\mathrm{Adv}}(\mathbf{w}_t) \qquad \text{(Gradient Descent)}$$

Definition 2.1 *We define the following notations:*

- $\gamma_\infty = \max_{\|\mathbf{w}\|_1=1}\min_{i\in[N]}\mathbf{x}_i^\top\mathbf{w}$, $\gamma_{2,\infty}(c) = \max_{\|\mathbf{w}\|_2=1}\min_{i\in[N]}\min_{\|\delta_i\|_\infty\le c}\langle\mathbf{x}_i+\delta_i, \mathbf{w}\rangle$

1. Our analysis can be applied to an arbitrary ℓ_p norms. Here, to be specific, we focus on the ℓ_∞ norm.

- $\mathbf{u}_{2,\infty}(c) = \operatorname{argmax}_{\|\mathbf{w}\|_2=1} \min_{i\in[N]} \min_{\|\delta_i\|_\infty \le c} \langle \mathbf{x}_i + \delta_i, \mathbf{w}\rangle$

Here, $\gamma_{2,\infty}(c)$ represents the ℓ_2 margin of the extended dataset $\{(\mathbf{x}', y) \mid \exists i,\ \|\mathbf{x}' - \mathbf{x}_i\|_\infty \le c,\ y' = y_i\}$, which is constructed by adding all adversarial examples to the original dataset within the adversarial radius c, and $\mathbf{u}_{2,\infty}(c)$ is the corresponding max margin direction.

2.2. ℓ_∞-Robust Support Vector Machine

The ℓ_∞-robust support vector machine (SVM) problem can be formulated as

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{w}\|_2^2, \qquad \text{such that } \min_{\|\delta_i\|_\infty \le c} \langle \mathbf{x}_i + \delta_i, \mathbf{w}\rangle \ge 1 \tag{ℓ_∞-Robust SVM}$$

Let $\hat{\mathbf{w}}(c)$ be the solution of (ℓ_∞-Robust SVM). Then we have $\mathbf{u}_{2,\infty} = \hat{\mathbf{w}}(c)/\|\hat{\mathbf{w}}(c)\|_2$, and $\gamma_{2,\infty}(c) = 1/\|\hat{\mathbf{w}}(c)\|_2$. We use $\mathbf{z}_i(c)$ to denote the worst-case adversarial perturbation that an adversary can construct against the classifier parameterized by $\hat{\mathbf{w}}(c)$:

$$\mathbf{z}_i(c) := \mathbf{x}_i - c\,\partial\|\hat{\mathbf{w}}(c)\|_1. \tag{1}$$

The adversarial support vectors are defined as:

$$\mathcal{S}(c) := \{i \in [N] \mid \langle \mathbf{z}_i(c), \mathbf{u}_{2,\infty}(c)\rangle = \gamma_{2,\infty}(c)\} \tag{Adversarial Support Vectors}$$

Proposition 2.2 *[Boyd and Vandenberghe [4]] We have* $\hat{\mathbf{w}}(c) = \sum_{i\in\mathcal{S}(c)} \alpha_i \mathbf{z}_i(c)$, *where* $\alpha_i \ge 0$.

2.3. General Adversarial Training

The general adversarial training problem as be modeled as a min-max problem:

$$\min_\theta L_{\text{Adv}}(\theta) := \frac{1}{N}\sum_{i=1}^N g_c(\mathbf{x}_i, \theta) \ \text{ where } \ g_c(\mathbf{x}_i, \theta) := \max_{\|\mathbf{x}_i' - \mathbf{x}_i\|_\infty \le c} g(\mathbf{x}_i', \theta). \tag{AT}$$

for some loss function g like quadratic function. In practice, it is usually difficult to explicitly solve the inner maximization problem in (AT), and several methods such as the Fast Gradient Sign Method (FGSM)[10] and the following Projected Gradient (PGD) method [18], where the inner maximization problem is approximated by iterative projected gradient ascent for $\mathcal{T}$ iterations:

$$\mathbf{x'}_t = \text{Proj}_{\mathcal{B}}\left(\mathbf{x}'_{t-1} + \alpha\,\text{Sign}(\nabla_{\mathbf{x}} g(\mathbf{x}'_{t-1}, \theta))\right),\ t \in [\mathcal{T}], \text{ and } \mathbf{x}' = \mathbf{x'}_{\mathcal{T}}, \tag{PGD}$$

where $\mathcal{B} := \{\mathbf{x}' \mid \|\mathbf{x}' - \mathbf{x}_i\|_\infty \le c\}$. Let $\{\tilde{\mathbf{x}}_i\}_{i=1}^N$ be the adversarial example founded by (PGD) under model's parameters θ. In this work, we use the terminology *sharpness* of adversarial training loss landscape at point θ as the maximal eigenvalue of the adversarial loss given $\{\tilde{\mathbf{x}}_i\}_{i=1}^N$, i.e.,

$$\text{Sharpness at } \theta := \text{Maximal eigenvalue of the Hessian } \frac{1}{N}\sum_{i=1}^N \nabla_\theta^2 g(\tilde{\mathbf{x}}_i, \theta) \tag{Sharpness}$$

When $c = 0$, our definition reduces to the standard notion of sharpness [6]. The same definition was also used by Liu et al. [15] to characterize the irregular geometry of adversarial loss landscapes.

3. Implicit Bias of Adversarial Logistic Regression

In this section, we study the implicit bias of gradient descent for linearly separable ℓ_∞-adversarial logistic regression. Previous work by Li et al. [14] shows that the trajectory of gradient descent converges in the direction of $\mathbf{u}_{2,\infty}(c)$ for $\eta = \mathcal{O}(e^{-d})$. We extend the result of Li et al. [14] to arbitrary constant step sizes, thereby bridging this gap between theory and practice. We first introduce several assumptions on the data set.

Assumption 3.1 *We assume the following assumptions hold for the data set* $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$*:*

(1). The dataset is linearly separable, i.e., $\gamma_\infty > 0$*, and the adversarial radius* $c < \gamma_\infty$*.*

(2). $\mathrm{Rank}\{\mathbf{z}_i(c), i \in \mathcal{S}(c)\} = \mathrm{Rank}\{\mathbf{z}_i(c), i \in [N]\} = d$*, where* $\mathcal{S}(c)$ *is defined in Section 2.2.*

(3). The coefficients α_i *in Proposition 2.2 satisfy* $\alpha_i > 0$*.*

Assumption *(1)* implies that the extended data set is also linearly separable. This assumption is also used in Li et al. [14]. Assumption *(2)* requires that the adversarial support vectors do not collapse to a low dimensional space. Similar assumptions as *(2)* is also needed in the standard logistic regression setting, e.g., [23, Theorem 4] and [27, Assumption 3]. Assumption *(3)* is the robust analogue of the non-degeneracy condition used in Soudry et al. [23, Appendix B].

Theorem 3.2 *Let* $v_t = \langle \mathbf{w}_t, \mathbf{u}_{2,\infty}(c) \rangle$ *and* $\eta > 0$*. Then for the* (Gradient Descent)*, we have*

(1). $v_{t+1} - v_t \geq \eta \gamma_{2,\infty}(c) \frac{1}{2} \exp\left(-v_t \gamma_{2,\infty}(c)\right)$*, which implies* $v_t \geq \frac{1}{\gamma_{2,\infty}(c)} \log\left(1 + \frac{\eta \gamma_{2,\infty}^2(c)}{2} t\right)$*.*

(2). $\|\mathrm{Proj}_{\mathbf{u}_{2,\infty}^\perp(c)}(\mathbf{w}_t)\|_2 \leq C$*, where* C *depends on* N*,* c*,* d*,* η *but is independent of* t*.*

In particular, we have $\lim_{t\to\infty} \mathbf{w}_t / \|\mathbf{w}_t\|_2 = \mathbf{u}_{2,\infty}(c)$ *with a* $\mathcal{O}(1/\log(\eta\, \gamma_{2,\infty}^2(c)\, t))$ *rate.*

The proof of Theorem 3.2 is inspired by Wu et al. [27], who studied the implicit bias of standard logistic regression under large step sizes. We generalize their approach to adversarial training by decomposing the parameter space into the robust max-margin direction spanned by $\mathbf{u}_{2,\infty}(c)$ and its orthogonal complement, and then tracking the projections of the dynamics onto these two subspaces. Detailed proofs are provided in Appendix B.

4. Evolution of Sharpness

The classical descent lemma for deterministic gradient descent in convex optimization problems $\min_\theta f(\theta)$ states that if the sharpness at θ is smaller than $2/\eta$, where η is the step size, then the loss decreases monotonically. However, in modern deep learning practice, people usually employ large step sizes in the regime where the descent lemma fails, and the loss exhibits non-monotonic behavior. Recent work of Cohen et al. [6] finds that an interesting pattern can arise if we tracing the evolution of the sharpness along the process, which can be characterized as a two-stages dynamics:

- *Progressive Sharpening*: If the sharpness is less than $2/\eta$, it tends to increase monotonically.
- *Edge of Stability Stage*: The sharpness oscillate around $2/\eta$, while the train loss behaves non-monotonically, yet consistently decreases over long timescales.

For adversarial training, since the (Sharpness) can be seen as a perturbation of the clean sharpness with an $\mathcal{O}(c)$ term, thus one could expect the evolution of sharpness for AT is also similar to standard training. *Surprisingly, our experiments suggest that such intuition is incorrect.*

4.1. Experimental Results

In Figure 1, we present our numerical experiments on the sharpness evolution of AT under ResNet and Vision Transformer on CIFAR-10 data set. The training algorithm is the deterministic gradient descent. Our architectures follow the implementation in Cohen et al. [7]. We also use the MSE loss as it can most align with the edge of stability phenomena [6]. Our main observation is that the sharpness evolution follows a very different two-stage dynamics as with standard training:

- *Stage* I: In this stage, the sharpness tends to increase monotonically, which is similar to the progressive sharpening stage in standard training.
- *Stage* II: After a certain point in training, the sharpness begins to decrease oscillatory. Meanwhile, the adversarial training loss decreases non-monotonically. Moreover, the sharpness at the end of training is far smaller than $2/\eta$.

The phenomena reported above are consistent across different architectures and adversarial radii. Importantly, even for a very small adversarial radius, e.g., $c = 0.5/288$, which is much smaller than the commonly used value $8/255$ in practice [21], it can lead to the behaviors of sharpness of AT being very different from that of standard training. In the following section, we provide a theoretical analysis of a minimalist example to verify this phenomenon.

4.2. Theoretical Investigation on a Minimalist Example

In this section, we study an adversarial variant of a two-dimensional example originally introduced by Ahn et al. [1] to investigate the edge-of-stability (EoS) phenomenon in standard training.

Let $\ell(s)$ be defined as the following Huber loss: $\ell(s) := \frac{s^2}{2}\mathbf{1}_{\{|s|\le 1\}} + \left(|s| - \frac{1}{2}\right)\mathbf{1}_{\{|s|>1\}}$. Let the step size $\eta > 0$, and consider the following EoS initialization according to Ahn et al. [1]:

$$(x_0, y_0) = \sqrt{(2+\delta)/\eta}\,(\tilde{x}, \tilde{y}), \quad \delta > 0, \tag{EoS Regime}$$

where $\tilde{y} > \tilde{x} > 0$ and $\tilde{x}^2 + \tilde{y}^2 = 1$. Given an adversarial radius $\rho \ge 0$, we consider the objective:

$$\min_{(x,y)\in\mathbb{R}^2} \mathcal{L}_\rho(x, y) = \min_{(x,y)\in\mathbb{R}^2} \max_{|\delta|\le\rho} \ell\left((x+\delta)y\right) \tag{ρ-adversarial Objective}$$

When $\rho = 0$, equation above reduces to the standard training objective studied in Section 2 of Ahn et al. [1]. We consider the following gradient descent dynamics

$$(x_{t+1}, y_{t+1}) = (x_t, y_t) - \eta\,\partial\mathcal{L}_\rho(x_t, y_t). \tag{2}$$

We define the sharpness of (ρ-adversarial Objective) as the maximal eigenvalue of the Hessian:

$$\lambda_t^\rho := \lambda_{\max}\left(\nabla^2\mathcal{L}_\rho(x_t, y_t)\right)$$

Proposition 4.1 *Let (x_0, y_0) be initialized according to* (EoS Regime)*. Under certain boundary-avoidance assumptions as described in Assumption C.3 and for sufficiently small η, we have:*

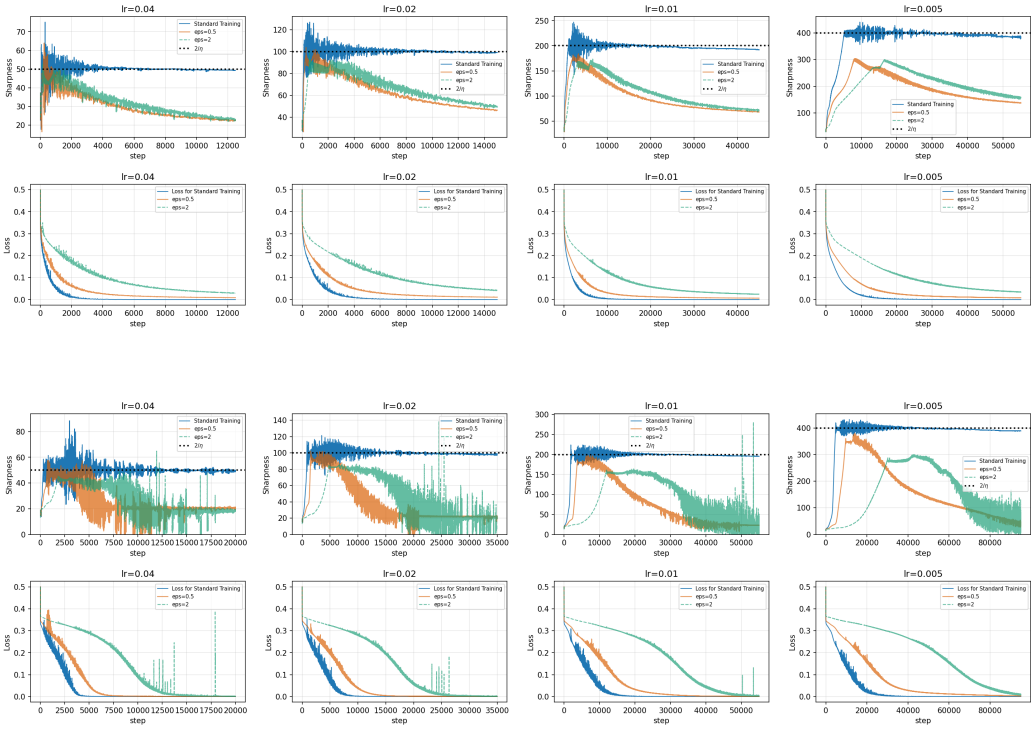

Figure 1: Sharpness evolution of adversarial training. Top: *ResNet*. Bottom: *Vision Transformer*. Dataset: CIFAR-10. Unlike the edge-of-stability phenomenon [6], adversarial training exhibits a similar initial progressive sharpening stage, followed by an oscillatory decline in sharpness. The adversarial radius is $\epsilon/255$. Details are presented in Appendix 4.1.

- *(Ahn et al. [1, Theorem 2]) For* $\rho = 0$*, it holds that* $\lim_{t\to+\infty} \lambda_t^0 \approx \frac{2}{\eta}$.
- *For* $\rho > 0$ *while* $\rho = \mathcal{O}(\sqrt{\eta})$*, it holds that* $\lim_{t\to+\infty} \lambda_t^\rho = \mathcal{O}(\eta)$.

From Proposition 4.1, we observe a sharp transition depending on whether $\rho = 0$, i.e., standard training and adversarial training. For small η, the adversarial dynamics converge to points with sharpness $\mathcal{O}(\eta)$, which is much smaller than the standard edge-of-stability level $2/\eta$, even for very small positive ρ. This coincides with the experimental results reported in Section 4.1. This difference comes from the dynamics: when $\rho = 0$, x_t becomes small and y_t changes slowly, so the trajectory converges to $(0, y_\infty)$ with $y_\infty \neq 0$ and sharpness close to $2/\eta$ as proved by Ahn et al. [1]; when $\rho > 0$, y_t keeps decreasing after x_t becomes small and eventually converges to 0, leading to much lower sharpness. The detailed proof is given in Appendix C.

5. Conclusion

In this paper, we studied adversarial training dynamics under large step sizes, focusing on two aspects: the implicit bias of adversarial logistic regression with separable data and the sharpness evolution of adversarial deep neural network training. A theoretical explanation for the different sharpness behaviors between adversarial and standard training remains an interesting future direction.

References

[1] Kwangjun Ahn, Sébastien Bubeck, Sinho Chewi, Yin Tat Lee, Felipe Suarez, and Yi Zhang. Learning threshold neurons via edge of stability. *Advances in Neural Information Processing Systems*, 36:19540–19569, 2023.

[2] Maksym Andriushchenko, Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Sgd with large step sizes learns sparse features. In *International Conference on Machine Learning*, pages 903–925. PMLR, 2023.

[3] Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *International Conference on Machine Learning*, pages 948–1024. PMLR, 2022.

[4] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[5] Yuhang Cai, Jingfeng Wu, Song Mei, Michael Lindsey, and Peter Bartlett. Large stepsize gradient descent for non-homogeneous two-layer networks: Margin improvement and fast optimization. *Advances in Neural Information Processing Systems*, 37:71306–71351, 2024.

[6] Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.

[7] Jeremy Cohen, Alex Damian, Ameet Talwalkar, J Zico Kolter, and Jason D. Lee. Understanding optimization in deep learning with central flows. In *The Thirteenth International Conference on Learning Representations*, 2025.

[8] Alex Damian, Eshaan Nichani, and Jason D Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. *arXiv preprint arXiv:2209.15594*, 2022.

[9] Chen Fan, Mark Schmidt, and Christos Thrampoulidis. Implicit bias of spectral descent and muon on multiclass separable data. *Advances in Neural Information Processing Systems*, 38: 39622–39669, 2026.

[10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[11] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR, 2018.

[12] Ziwei Ji and Matus Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018.

[13] Ziwei Ji, Nathan Srebro, and Matus Telgarsky. Fast margin maximization via dual acceleration. In *International Conference on Machine Learning*, pages 4860–4869. PMLR, 2021.

[14] Yan Li, Ethan X.Fang, Huan Xu, and Tuo Zhao. Implicit bias of gradient descent based adversarial training on separable data. In *International Conference on Learning Representations*, 2020.

[15] Chen Liu, Mathieu Salzmann, Tao Lin, Ryota Tomioka, and Sabine Süsstrunk. On the loss landscape of adversarial training: Identifying challenges and how to overcome them. *Advances in Neural Information Processing Systems*, 33:21476–21487, 2020.

[16] Bochen Lv and Zhanxing Zhu. Implicit bias of adversarial training for deep neural networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=l8It-0lE5e7.

[17] Bochen Lyu and Zhanxing Zhu. Analyzing the implicit bias of adversarial training from a generalized margin perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[19] Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3051–3059. PMLR, 2019.

[20] Mor Shpigel Nacson, Kavya Ravichandran, Nathan Srebro, and Daniel Soudry. Implicit bias of the step size in linear diagonal neural networks. In *International Conference on Machine Learning*, pages 16270–16295. PMLR, 2022.

[21] Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu. Bag of tricks for adversarial training. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Xb8xvrtB8Ce.

[22] Dan Qiao, Kaiqi Zhang, Esha Singh, Daniel Soudry, and Yu-Xiang Wang. Stable minima cannot overfit in univariate relu networks: Generalization by large step sizes. *Advances in Neural Information Processing Systems*, 37:94163–94208, 2024.

[23] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19 (70):1–57, 2018.

[24] Nikolaos Tsilivis, Natalie Frank, Nati Srebro, and Julia Kempe. The price of implicit bias in adversarially robust generalization. *Advances in Neural Information Processing Systems*, 37: 58023–58057, 2024.

[25] Bohan Wang, Qi Meng, Huishuai Zhang, Ruoyu Sun, Wei Chen, Zhi-Ming Ma, and Tie-Yan Liu. Does momentum change the implicit regularization on separable data? *Advances in Neural Information Processing Systems*, 35:26764–26776, 2022.

[26] Yuqing Wang, Zhenghao Xu, Tuo Zhao, and Molei Tao. Good regularity creates large learning rate implicit biases: edge of stability, balancing, and catapult. *Journal of Machine Learning Research*, 26(273):1–68, 2025.

[27] Jingfeng Wu, Vladimir Braverman, and Jason D Lee. Implicit bias of gradient descent for logistic regression at the edge of stability. *Advances in Neural Information Processing Systems*, 36:74229–74256, 2023.

[28] Jingfeng Wu, Peter L Bartlett, Matus Telgarsky, and Bin Yu. Large stepsize gradient descent for logistic loss: Non-monotonicity of the loss improves optimization efficiency. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 5019–5073. PMLR, 2024.

[29] Jingfeng Wu, Pierre Marion, and Peter Bartlett. Large stepsizes accelerate gradient descent for regularized logistic regression. *Advances in Neural Information Processing Systems*, 38: 104485–104525, 2026.

[30] Chenyang Zhang, Difan Zou, and Yuan Cao. The implicit bias of adam on separable data. *Advances in Neural Information Processing Systems*, 37:23988–24021, 2024.

[31] Ruiqi Zhang, Jingfeng Wu, Licong Lin, and Peter L Bartlett. Minimax optimal convergence of gradient descent in logistic regression via large and adaptive stepsizes. *arXiv preprint arXiv:2504.04105*, 2025.

Appendix A. Related Works

Implicit Bias of (Adversarial) Logistic Regression. The implicit bias of gradient-based methods for logistic regression problem is often understood through margin maximization. In standard training, Soudry et al. [23] show that gradient descent converges along a max margin direction for linear model under linearly separable data, which was generalized by Ji and Telgarsky [12] to non-separable data. Subsequent works have extended the results of Ji and Telgarsky [12], Soudry et al. [23] to other algorithms such as steepest descent [11], stochastic gradient descent [19], momentum-based method [13, 25, 30] and spectral descent [9].

Comparing with standard training, implicit bias of adversarial training algorithms are less understood. Li et al. [14] showed that gradient descent for separable adversarial logistic regression converges to a robust max-margin direction, and their analysis requires a very small step size that depends exponentially on the dimension. Later works studied related robust implicit-bias questions from generalized-margin and robust-generalization perspectives [17, 24].

Large Step Sizes and Edge-of-Stability Phenomena. Large step sizes can produce non-monotone trajectories while still preserving useful long-time behavior. For logistic regression, Wu et al. [27, 28, 29] showed that large step size gradient descent can still minimize the loss and exhibit max-margin implicit bias despite oscillations. Zhang et al. [31] proved that GD with large, adaptive step sizes is minimax optimal among first-order batch methods. Cai et al. [5] shows that gradient descent under large step size can improve the convergence of non-Homogeneous two-layer neural networks. Besides the benefits of convergence, Nacson et al. [20] showed that large step sizes can alter the implicit bias of linear diagonal neural networks.

In neural network training, Cohen et al. [6] identified the edge-of-stability (EoS) phenomenon, in which the sharpness stays close to $2/\eta$. Moreover, Ahn et al. [1] show that EoS can improve the generalization of neural networks. Several theoretical works have also been proposed to explain the EoS phenomenon; see, e.g., [3, 7, 8] and the references therein. In adversarial training, however, the loss landscape may exhibit different curvature and gradient structures [15]. Our experiments and minimalist model suggest that the corresponding sharpness dynamics do not necessarily settle near $2/\eta$, but may instead enter an oscillatory decreasing phase.

Appendix B. Proofs of Theorem 3.2

In the following, we assume $\mathbf{w}_0 = \mathbf{0}$. Without loss of generality, we also assume $y_i = 1$ for all $i \in [N]$ in what follows since if $y_i = -1$, we can equivalently take the negative of $\mathbf{x}_i$ as input. We introduce the following decomposition of $\mathbb{R}^d$, which was motivated by [27]:

$$\begin{aligned} \mathcal{F} : \mathbb{R}^d &\to \mathbb{R}^d \times \mathbb{R}^d \\ \mathbf{x} &\to \left(\langle \mathbf{x}, \mathbf{u}_{2,\infty}(c) \rangle \mathbf{u}_{2,\infty}(c), \mathrm{Proj}_{\mathbf{u}_{2,\infty}^{\perp}(c)}(\mathbf{x}) \right). \end{aligned} \tag{3}$$

We write

$$\mathbf{w}_t = v_t \mathbf{u}_{2,\infty}(c) + \mathbf{r}_t, \tag{4}$$

where $\langle \mathbf{r}_t, \mathbf{u}_{2,\infty}(c) \rangle = 0$. We have

$$\begin{aligned} -\nabla_{\mathbf{w}} \mathcal{L}_{\mathrm{Adv}}(\mathbf{w}_t) &= -\sum_{i=1}^{N} \nabla_{\mathbf{w}} \ln \left(1 + \exp(-\mathbf{x}_i^\top \mathbf{w}_t + c\|\mathbf{w}_t\|_1) \right) \\ &= \sum_{i=1}^{N} \lambda_{i,t} \left(\mathbf{x}_i - c\partial \|\mathbf{w}_t\|_1 \right), \end{aligned}$$

where

$$\lambda_{i,t} = \frac{\exp(-\mathbf{x}_i^\top \mathbf{w}_t + c\|\mathbf{w}_t\|_1)}{1 + \exp(-\mathbf{x}_i^\top \mathbf{w}_t + c\|\mathbf{w}_t\|_1)} = \frac{1}{1 + \exp(\mathbf{x}_i^\top \mathbf{w}_t - c\|\mathbf{w}_t\|_1)} > 0. \tag{5}$$

Thus the dynamics can be written as

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \eta \sum_{i=1}^{N} \lambda_{i,t} \left(\mathbf{x}_i - c\partial \|\mathbf{w}_t\|_1 \right) \tag{6}$$

B.1. Some Useful Lemmas

Lemma B.1 *Assume the original dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ is linearly separable, i.e., $\exists\, \mathbf{w} \in \mathbb{R}^d$ such that $y_i \mathbf{w}^\top \mathbf{x}_i > 0$ for all $i \in [N]$. If $c < \gamma_\infty$, we have $\gamma_{2,\infty}(c) > 0$.*

Proof Let $\mathbf{u}_\infty = \mathrm{argmax}_{\|\mathbf{w}\|_1 = 1} \min_{i \in [n]} \mathbf{x}_i^\top \mathbf{w}$. By definition, we have $\gamma_\infty = \min_{i \in [N]} \mathbf{x}_i^\top \mathbf{u}_\infty$.

Moreover, we have

$$\begin{aligned} \gamma_{2,\infty}(c) &\geq \min_{i \in [N]} \min_{\|\delta_i\|_\infty \leq c} (\mathbf{x}_i + \delta_i)^\top \mathbf{u}_\infty \\ &= \min_{i \in [N]} \mathbf{x}_i^\top \mathbf{u}_\infty - c\|\mathbf{u}_\infty\|_1 \\ &= \gamma_\infty - c \qquad (\text{Since } \min_{i \in [N]} \mathbf{x}_i^\top \mathbf{u}_\infty = \gamma_\infty \text{ and } \|\mathbf{u}_\infty\|_1 = 1) \\ &> 0. \end{aligned}$$

This completes the proof. ■

Proposition B.2 *Under Assumption 3.1, if* $\mathbf{v} \perp \mathbf{u}_{2,\infty}(c)$*, then* $\mathcal{S}(c)$ *is non-separable by* $\mathbf{v}$*. That is,* $\exists\, i, j \in \mathcal{S}(c)$ *such that*

$$\langle \mathbf{z}_i(c), \mathbf{v} \rangle > 0 \qquad \text{and} \qquad \langle \mathbf{z}_j(c), \mathbf{v} \rangle < 0.$$

Proof We have

$$0 = \langle \mathbf{v}, \mathbf{u}_{2,\infty}(c) \rangle = \sum_{i \in \mathcal{S}(c)} \alpha_i \langle \mathbf{v}, \mathbf{z}_i(c) \rangle, \qquad \text{where } \alpha_i > 0 \text{ from (2) of Assumption 3.1.}$$

Moreover, by (3) of Assumption 3.1, $\langle \mathbf{v}, \mathbf{z}_i(c) \rangle$ cannot all be zero for $i \in \mathcal{S}(c)$. Thus there must exist some $i, j \in \mathcal{S}(c)$ such that $\langle \mathbf{z}_i(c), \mathbf{v} \rangle > 0$ and $\langle \mathbf{z}_i(c), \mathbf{v} \rangle < 0$. ■

In the following we will prove Theorem 3.2. Recall from (Gradient Descent), we have:

$$\begin{aligned} \mathbf{w}_{t+1} &= \mathbf{w}_t - \eta \nabla_{\mathbf{w}} \mathcal{L}_{\text{Adv}}(\mathbf{w}_t) \\ &= \mathbf{w}_t - \eta \sum_{i=1}^{N} \nabla_{\mathbf{w}} \ln \Big(1 + \exp(-\mathbf{x}_i^\top \mathbf{w}_t + c \|\mathbf{w}_t\|_1) \Big) \end{aligned}$$

B.2. Proof of (1) in Theorem 3.2

Taking the inner product on both sides of (6) with $\mathbf{u}_{2,\infty}(c)$, we get

$$v_{t+1} = v_t + \eta \sum_{i=1}^{N} \lambda_{i,t} \langle \mathbf{x}_i - c\partial \|\mathbf{w}_t\|_1, \mathbf{u}_{2,\infty}(c) \rangle. \tag{7}$$

Lemma B.3 *We have* $\langle \mathbf{x}_i - c\partial \|\mathbf{w}_t\|_1, \mathbf{u}_{2,\infty}(c) \rangle \geq \gamma_{2,\infty}(c)$*. Thus* v_t *as defined in (7) is an increasing function of* t*. In particular, if we choose* $\mathbf{w}_0 = \mathbf{0}$*, then* $v_t > 0$ *for all* $t > 1$*.*

Proof By definition, for any $i \in [N]$, we have

$$\begin{aligned} \gamma_{2,\infty}(c) &\leq \min_{\|\delta_i\|_\infty \leq c} \langle \mathbf{x}_i + \delta_i, \mathbf{u}_{2,\infty}(c) \rangle \\ &= \langle \mathbf{x}_i, \mathbf{u}_{2,\infty}(c) \rangle - c \|\mathbf{u}_{2,\infty}(c)\|_1 \end{aligned}$$

Moreover, since $\|\partial \|\mathbf{w}_t\|_1\|_\infty = 1$, we have $\|\mathbf{u}_{2,\infty}(c)\|_1 \geq \langle \mathbf{u}_{2,\infty}(c), \partial \|\mathbf{w}_t\|_1 \rangle$. Combining these together, we get

$$\gamma_{2,\infty}(c) \leq \langle \mathbf{x}_i, \mathbf{u}_{2,\infty}(c) \rangle - c \|\mathbf{u}_{2,\infty}(c)\|_1 \leq \langle \mathbf{x}_i, \mathbf{u}_{2,\infty}(c) \rangle - c \langle \mathbf{u}_{2,\infty}(c), \partial \|\mathbf{w}_t\|_1 \rangle = \langle \mathbf{x}_i - c\partial \|\mathbf{w}_t\|_1, \mathbf{u}_{2,\infty}(c) \rangle.$$

■

Now we are ready to prove (1) in Theorem 3.2.

Proof [Proof of (1) in Theorem 3.2] By Proposition B.2, there exists some $j \in \mathcal{S}(c)$ such that

$$\langle \mathbf{z}_j(c), \mathbf{r}_t \rangle \leq 0. \tag{8}$$

For the j as defined in (8), we have

$$-\mathbf{x}_j^\top \mathbf{w}_t + c\|\mathbf{w}_t\|_1 \geq -\mathbf{x}_j^\top \mathbf{w}_t + c\langle \partial\|\mathbf{u}_{2,\infty}(c)\|_1, \mathbf{w}_t\rangle = -\langle \mathbf{z}_j(c), \mathbf{w}_t\rangle \tag{9}$$

Taking the decomposition $\mathbf{w}_t = v_t\mathbf{u}_{2,\infty}(c) + \mathbf{r}_t$ as stated in (4) into (9), we have

$$-\langle \mathbf{z}_j(c), \mathbf{w}_t\rangle = -v_t\langle \mathbf{z}_j(c), \mathbf{u}_{2,\infty}(c)\rangle - \langle \mathbf{z}_j(c), \mathbf{r}_t\rangle.$$

Since $j \in \mathcal{S}(c)$, we have $\langle \mathbf{z}_j(c), \mathbf{u}_{2,\infty}(c)\rangle = \gamma_{2,\infty}(c)$. Moreover, from (8), we get

$$-\langle \mathbf{z}_j(c), \mathbf{w}_t\rangle \geq -v_t\gamma_{2,\infty}(c) \tag{10}$$

Combining (10) and (9), we get

$$-\mathbf{x}_j^\top \mathbf{w}_t + c\|\mathbf{w}_t\|_1 \geq -v_t\gamma_{2,\infty}(c). \tag{11}$$

Now we can provide a lower bound on $\lambda_{j,t}$:

$$\begin{aligned}
\lambda_{j,t} &= \frac{\exp(-\mathbf{x}_i^\top \mathbf{w}_t + c\|\mathbf{w}_t\|_1)}{1+\exp(-\mathbf{x}_i^\top \mathbf{w}_t + c\|\mathbf{w}_t\|_1)} \\
&\geq \frac{1}{2}\min\{1, \exp(-\mathbf{x}_i^\top \mathbf{w}_t + c\|\mathbf{w}_t\|_1)\} \qquad (\text{Since } \frac{e^a}{1+e^a} \geq \frac{1}{2}\min\{1, e^a\}) \\
&\geq \frac{1}{2}\min\{1, \exp\left(-v_t\gamma_{2,\infty}(c)\right)\} \qquad (\text{From (11)}) \\
&\geq \frac{1}{2}\exp\left(-v_t\gamma_{2,\infty}(c)\right) \qquad (\text{Since Lemma } B.3)
\end{aligned} \tag{12}$$

Now we return to the recursion formula (7):

$$\begin{aligned}
v_{t+1} &= v_t + \eta\sum_{i=1}^N \lambda_{i,t}\langle \mathbf{x}_i - c\partial\|\mathbf{w}_t\|_1, \mathbf{u}_{2,\infty}(c)\rangle \\
&\geq v_t + \eta\gamma_{2,\infty}(c)\sum_{i=1}^N \lambda_{i,t} \qquad (\text{Since Lemma } B.3) \\
&\geq v_t + \eta\gamma_{2,\infty}(c)\lambda_{j,t} \\
&\geq v_t + \eta\gamma_{2,\infty}(c)\frac{1}{2}\exp\left(-v_t\gamma_{2,\infty}(c)\right) \qquad (\text{Since (12)})
\end{aligned} \tag{13}$$

Solving the recurrence of (13) completes the proof of part (1). ■

B.3. Proof of (2) in Theorem 3.2

Lemma B.4 *[Monotonicity of ℓ_1 norm] We have*

$$\langle \mathbf{x} - \mathbf{y}, \partial\|\mathbf{x}\|_1 - \partial\|\mathbf{y}\|_1 \rangle \geq 0$$

Proof In fact, this holds for any convex function f; here we use the special case $f(\mathbf{x}) = \|\mathbf{x}\|_1$. ■

Moreover, we have

$$\begin{aligned}
\mathbf{x}_i^\top \mathbf{w}_t - c\|\mathbf{w}_t\|_1 &= \langle \mathbf{x}_i, \mathbf{w}_t \rangle - c\langle \mathbf{w}_t, \partial\|\mathbf{w}_t\|_1 \rangle \\
&= \langle \mathbf{w}_t, \mathbf{x}_i - c\partial\|\mathbf{w}_t\|_1 \rangle \\
&= \langle v_t \mathbf{u}_{2,\infty}(c) + \mathbf{r}_t, \mathbf{x}_i - c\partial\|\mathbf{w}_t\|_1 \rangle \\
&= v_t \langle \mathbf{u}_{2,\infty}(c), \mathbf{x}_i - c\partial\|\mathbf{w}_t\|_1 \rangle + \langle \mathbf{r}_t, \mathbf{x}_i - c\partial\|\mathbf{w}_t\|_1 \rangle
\end{aligned} \tag{14}$$

We denote $\mathbf{g}_t = \mathrm{Proj}_{\mathbf{u}_{2,\infty}^\perp(c)} \left(\sum_{i=1}^N \lambda_{i,t}(\mathbf{x}_i - c\partial\|\mathbf{w}_t\|_1) \right)$, that is, the projection of negative gradient into the non-separable space. Thus we have

$$\mathbf{r}_{t+1} = \mathbf{r}_t + \eta \mathbf{g}_t. \tag{15}$$

Our aim is to show $\{\mathbf{r}_t\}$ is uniformly bounded for all $t > 0$. From (15), we have

$$\|\mathbf{r}_{t+1}\|_2^2 = \|\mathbf{r}_t\|_2^2 + 2\eta\langle \mathbf{r}_t, \mathbf{g}_t \rangle + \eta^2 \|\mathbf{g}_t\|_2^2 \tag{16}$$

In the following we will provide bounds on $\langle \mathbf{r}_t, \mathbf{g}_t \rangle$ and $\|\mathbf{g}_t\|_2^2$ to prove the result.

Lemma B.5 *Let $\beta(c) = -\max_{\|\mathbf{q}\|_2=1, \mathbf{q} \perp \mathbf{u}_{2,\infty}(c)} \min_{i \in \mathcal{S}(c)} \langle \mathbf{z}_i(c), \mathbf{q} \rangle$, then we have $\beta(c) > 0$.*

Proof From Proposition B.2, we have

$$\exists i \in \mathcal{S}(c),\ \langle \mathbf{z}_i(c), \mathbf{q} \rangle < 0, \forall \mathbf{q} \perp \mathbf{u}_{2,\infty}(c).$$

Moreover, since the set $\{\mathbf{q} \mid \|\mathbf{q}\|_2 = 1, \mathbf{q} \perp \mathbf{u}_{2,\infty}(c)\}$ is a compact set, thus the maximal is also strictly positive, which implies $\beta(c) > 0$. ■

Lemma B.6 *We have*

$$\langle \mathbf{r}_t, \partial\|\mathbf{w}_t\|_1 - \partial\|\hat{\mathbf{w}}(c)\|_1 \rangle = \langle \mathbf{w}_t - v_t \mathbf{u}_{2,\infty}(c), \partial\|\mathbf{w}_t\|_1 - \partial\|\hat{\mathbf{w}}(c)\|_1 \rangle \geq 0.$$

Proof Recall that we have $v(t) > 0$ from Lemma B.3, thus $\partial\|\hat{\mathbf{w}}(c)\|_1 = \partial\|\mathbf{u}_{2,\infty}(c)\|_1 = \partial\|v_t \mathbf{u}_{2,\infty}(c)\|_1$. Then we apply Lemma B.4 to get

$$\langle \mathbf{w}_t - v_t \mathbf{u}_{2,\infty}(c), \partial\|\mathbf{w}_t\|_1 - \partial\|\hat{\mathbf{w}}(c)\|_1 \rangle = \langle \mathbf{w}_t - v_t \mathbf{u}_{2,\infty}(c), \partial\|\mathbf{w}_t\|_1 - \partial\|v_t \mathbf{u}_{2,\infty}(c)\|_1 \rangle \geq 0.$$

■

We define $p_{i,t} = \langle \mathbf{r}_t, \mathbf{x}_i - c\partial\|\mathbf{w}_t\|_1 \rangle$, which is an important term appearing in (14). Then we have the following upper bound for $p_{i,t}$:

Lemma B.7 *We have*

$$p_{i,t} \leq \langle \mathbf{z}_i(c), \mathbf{r}_t \rangle$$

Proof

$$\begin{aligned} p_{i,t} &= \langle \mathbf{r}_t, \mathbf{x}_i - c\partial\|\mathbf{w}_t\|_1 \rangle \\ &= \langle \mathbf{r}_t, \mathbf{x}_i - c\partial\|\hat{\mathbf{w}}(c)\|_1 + c\partial\|\hat{\mathbf{w}}(c)\|_1 - c\partial\|\mathbf{w}_t\|_1 \rangle \\ &= \langle \mathbf{r}_t, \mathbf{x}_i - c\partial\|\hat{\mathbf{w}}(c)\|_1 \rangle + \langle \mathbf{r}_t, c\partial\|\hat{\mathbf{w}}(c)\|_1 - c\partial\|\mathbf{w}_t\|_1 \rangle \\ &\leq \langle \mathbf{r}_t, \mathbf{x}_i - c\partial\|\hat{\mathbf{w}}(c)\|_1 \rangle \qquad \text{(By Lemma } B.6\text{)} \\ &= \langle \mathbf{z}_i(c), \mathbf{r}_t \rangle \end{aligned}$$

■

We choose two particular elements in $\mathcal{S}(c)$. First, from Lemma B.5, there exists some $k \in \mathcal{S}(c)$, such that

$$\langle \mathbf{z}_k(c), \mathbf{r}(t) \rangle \leq -\beta(c)\|\mathbf{r}_t\|_2 \tag{17}$$

which implies

$$p_{k,t} \leq -\beta(c)\|\mathbf{r}_t\|_2 \tag{18}$$

from Lemma B.7. Second, we choose $j \in [N]$ such that

$$p_{j,t} = \min_{i \in [N]} p_{i,t}, \tag{19}$$

thus we have

$$p_{j,t} \leq p_{k,t} \leq -\beta(c)\|\mathbf{r}_t\|_2. \tag{20}$$

Lemma B.8 (Bounds on $\lambda_{i,t}$) *For any $i \in [N]$, the term $\lambda_{i,t}$ as defined in (5) satisfies*

$$\lambda_{i,t} \leq \frac{1}{1 + \exp(\gamma_{2,\infty}(c)v_t + p_{j,t})}, \ \forall i \in [N] \tag{21}$$

where the j-term is defined in (19). Moreover, we have a lower bound on $\lambda_{k,t}$:

$$\lambda_{k,t} \geq \frac{1}{1 + \exp(\gamma_{2,\infty}(c)v_t)} \geq \frac{1}{2}\exp(-\gamma_{2,\infty}(c)v_t), \tag{22}$$

where the k-term is defined in (18).

Proof We first prove (21). We have

$$\begin{aligned} \langle \mathbf{x}_i, \mathbf{w}_t \rangle - c\|\mathbf{w}_t\|_1 &= \langle \mathbf{x}_i, \mathbf{w}_t \rangle - c\langle \partial\|\mathbf{w}_t\|_1, \mathbf{w}_t \rangle \\ &= \langle \mathbf{x}_i - c\partial\|\mathbf{w}_t\|_1, \mathbf{w}_t \rangle \\ &= \langle v_t\mathbf{u}_{2,\infty}(c) + \mathbf{r}_t, \mathbf{x}_i - c\partial\|\mathbf{w}_t\|_1 \rangle \\ &= v_t\langle \mathbf{u}_{2,\infty}(c), \mathbf{x}_i - c\partial\|\mathbf{w}_t\|_1 \rangle + p_{i,t} \end{aligned} \tag{23}$$

From Lemma B.3, we have $\langle \mathbf{x}_i - c\partial\|\mathbf{w}_t\|_1, \mathbf{u}_{2,\infty}(c)\rangle \geq \gamma_{2,\infty}(c)$. Take this into (23), we get

$$(23) \geq v_t\gamma_{2,\infty}(c) + p_{i,t} \geq v_t\gamma_{2,\infty}(c) + p_{j,t} \tag{24}$$

Take (24) into (5), we get (21).

Now we prove (22). We have

$$\begin{aligned}\langle \mathbf{x}_k, \mathbf{w}_t\rangle - c\|\mathbf{w}_t\|_1 &\leq \langle \mathbf{z}_k(c), \mathbf{w}_t\rangle \\ &= \gamma_{2,\infty}(c)v_t + \langle \mathbf{z}_k(c), \mathbf{r}_t\rangle \\ &\leq \gamma_{2,\infty}(c)v_t \qquad (\text{From (17)},\ \langle \mathbf{z}_k(c), \mathbf{r}_t\rangle < 0)\end{aligned} \tag{25}$$

Take (25) into into (5), we get (22). ■

Proposition B.9 (Upper bound of $\mathbf{g}_t$) *We have* $\|\mathbf{g}_t\|_2^2 \leq \frac{N^2M^2}{1+\exp(\gamma_{2,\infty}(c)+p_{i,t})}$, *where* $M = 1 + c\sqrt{d}$.

Proof We have

$$\|\mathbf{g}_t\|_2 \leq \sum_{i=1}^{N} \lambda_{i,t}\|\mathrm{Proj}_{\mathbf{u}_{2,\infty}^{\perp}(c)}(\mathbf{x}_i - c\partial\|\mathbf{w}_t\|_1)\|_2$$

Moreover, since we assume $\|\mathbf{x}_i\|_2 \leq 1$ and $\|\partial\|\mathbf{w}_t\|_1\|_2 \leq \sqrt{d}$, we have

$$\|\mathrm{Proj}_{\mathbf{u}_{2,\infty}^{\perp}(c)}(\mathbf{x}_i - c\partial\|\mathbf{w}_t\|_1)\|_2 \leq 1 + c\sqrt{d} = M.$$

Combining the above, we get $\|\mathbf{g}_t\|_2 \leq \sum_{i=1}^{N} \lambda_{i,t}(1 + c\sqrt{d}) \leq \frac{MN}{1+\exp(\gamma_{2,\infty}(c)+p_{j,t})}$, where the last inequality comes from Lemma B.8. ■

Proposition B.10 (Upper bounds on $\langle \mathbf{r}_t, \mathbf{g}_t\rangle$) *We have*

$$\langle \mathbf{r}_t, \mathbf{g}_t\rangle \leq \exp\left(-\gamma_{2,\infty}(c)v_t\right)\left(N - \frac{1}{4}\beta(c)\|\mathbf{r}_t\|_2\right) - \frac{1}{2}\beta(c)\frac{\|\mathbf{r}_t\|_2}{1+\exp(\gamma_{2,\infty}(c)+p_{j,t})}$$

Proof We have

$$\begin{aligned}\langle \mathbf{r}_t, \mathbf{g}_t\rangle &= \langle \mathbf{r}_t, \mathrm{Proj}_{\mathbf{u}_{2,\infty}^{\perp}(c)}\left(\sum_{i=1}^{N}\lambda_{i,t}(\mathbf{x}_i - c\partial\|\mathbf{w}_t\|_1)\right)\rangle \\ &\leq \sum_{i=1}^{N}\lambda_{i,t}\langle \mathbf{r}_t, \mathrm{Proj}_{\mathbf{u}_{2,\infty}^{\perp}(c)}\left(\mathbf{x}_i - c\partial\|\mathbf{w}_t\|_1\right)\rangle \\ &= \sum_{i=1}^{N}\lambda_{i,t}p_{i,t} \qquad (\text{Since } \mathbf{u}_{2,\infty}(c) \perp \mathbf{r}_t) \\ &\leq \sum_{i,p_{i,t}>0}\lambda_{i,t}p_{i,t} + \frac{1}{2}\lambda_{j,t}p_{j,t} + \lambda_{k,t}p_{k,t} \qquad (j,k \text{ are defined in (19) and (18)})\end{aligned}$$

For terms in $\sum_{i,p_{i,t}>0} \lambda_{i,t} p_{i,t}$, we have

$$\begin{aligned}\lambda_{i,t} &\leq \exp(-\mathbf{x}^\top \mathbf{w}_t + c\|\mathbf{w}_t\|_1) \\ &= \exp\left(-v_t \langle \mathbf{u}_{2,\infty}(c), \mathbf{x}_i - c\partial\|\mathbf{w}_t\|_1\rangle - \langle \mathbf{r}_t, \mathbf{x}_i - c\partial\|\mathbf{w}_t\|_1\rangle\right) \qquad \text{(Since (14))} \\ &= \exp\left(-v_t \langle \mathbf{u}_{2,\infty}(c), \mathbf{x}_i - c\partial\|\mathbf{w}_t\|_1\rangle - p_{i,t}\right),\end{aligned}$$

thus

$$\begin{aligned}\sum_{i,p_{i,t}>0} \lambda_{i,t} p_{i,t} &= \sum_{i,p_{i,t}>0} \exp\left(-v_t \langle \mathbf{u}_{2,\infty}(c), \mathbf{x}_i - c\partial\|\mathbf{w}_t\|_1\rangle - p_{i,t}\right) p_{i,t} \\ &\leq \sum_{i,p_{i,t}>0} \exp\left(-v_t \langle \mathbf{u}_{2,\infty}(c), \mathbf{x}_i - c\partial\|\mathbf{w}_t\|_1\rangle\right) \\ &\leq N \exp(-\gamma_{2,\infty}(c) v_t) \qquad \text{(Lemma } B.3\text{)} \end{aligned} \tag{26}$$

For the term $\frac{1}{2}\lambda_{j,t} p_{j,t}$, we have

$$\begin{aligned}\frac{1}{2}\lambda_{j,t} p_{j,t} &\leq \frac{1}{2}\lambda_{j,t}\left(-\beta(c)\|\mathbf{r}_t\|_2\right) \qquad \text{(From (20))} \\ &= -\frac{1}{2}\frac{\beta(c)\|\mathbf{r}_t\|_2}{1+\exp(\gamma_{2,\infty}(c) v_t + p_{j,t})}.\end{aligned} \tag{27}$$

For the term $\frac{1}{2}\lambda_{k,t} p_{k,t}$, we have

$$\frac{1}{2}\lambda_{k,t} p_{k,t} \leq -\frac{1}{4}\beta(c)\|\mathbf{r}_t\|_2 \exp(-\gamma_{2,\infty} v_t). \qquad \text{(From (18))} \tag{28}$$

Combine (26), (27), and (28) together, we get

$$\langle \mathbf{r}_t, \mathbf{g}_t\rangle \leq \exp\left(-\gamma_{2,\infty}(c) v_t\right)\left(N - \frac{1}{4}\beta(c)\|\mathbf{r}_t\|_2\right) - \frac{1}{2}\beta(c)\frac{\|\mathbf{r}_t\|_2}{1+\exp(\gamma_{2,\infty}(c) + p_{j,t})},$$

this completes the proof. ■

Now we return to the recursion formula

$$\|\mathbf{r}_{t+1}\|_2^2 = \|\mathbf{r}_t\|_2^2 + 2\eta\langle \mathbf{r}_t, \mathbf{g}_t\rangle + \eta^2\|\mathbf{g}_t\|_2^2. \tag{29}$$

We have obtained an upper bound on $2\eta\langle \mathbf{r}_t, \mathbf{g}_t\rangle$ in Proposition B.10 and an upper bound on $\eta^2\|\mathbf{g}_t\|_2^2$ in Proposition B.9. Combining these bounds, we get

$$\begin{aligned}\|\mathbf{r}_{t+1}\|_2^2 \leq \|\mathbf{r}_t\|_2^2 &+ 2\eta\exp(-\gamma_{2,\infty}(c) v_t)\left(N - \frac{\beta(c)}{4}\|\mathbf{r}_t\|_2\right) \\ &+ \frac{\eta}{1+\exp(\gamma_{2,\infty}(c) v_t + p_{j,t})}(-\beta(c)\|\mathbf{r}_t\|_2 + \eta M^2 N^2)\end{aligned} \tag{30}$$

Thus once $\|\mathbf{r}_t\|_2 \geq \max\{\frac{4N}{\beta(c)}, \frac{\eta M^2 N^2}{\beta(c)}\}$, we have $\|\mathbf{r}_{t+1}\|_2 \leq \|\mathbf{r}_t\|_2$.

Otherwise, we have $\|\mathbf{r}_t\|_2 < \max\{\frac{4N}{\beta(c)}, \frac{\eta M^2 N^2}{\beta(c)}\}$, and then

$$\|\mathbf{r}_{t+1}\|_2 \leq \|\mathbf{r}_t\|_2 + \eta\|\mathbf{g}_t\|_2$$

We have a simple bound $\|\mathbf{g}_t\|_2 \leq MN$, which simply uses $\lambda_{i,t} < 1$. This implies $\|\mathbf{r}_{t+1}\|_2 \leq \|\mathbf{r}_t\|_2 + \eta MN$.

Combine all together, we get

$$\|\mathbf{r}_t\|_2 \leq \max\{\frac{4N}{\beta(c)}, \frac{\eta M^2 N^2}{\beta(c)}\} + \eta MN,$$

This finishes the proof of (2) in Theorem 3.2.

Appendix C. Appendix of Section 4.1

C.1. Experimental Details of Section 4.1

Experiments in Fig. 1 were conducted on a four-class subset of CIFAR-10 with 400 training examples. Images were normalized using the CIFAR-10 channel-wise mean and standard deviation, and models were trained with the mean-squared error loss on one-hot class labels. We compare standard full-batch gradient descent with adversarial training. For standard training, parameters are updated by

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta L(\theta_t),$$

with learning rates

$$\eta \in \{0.04,\ 0.02,\ 0.01,\ 0.005\}.$$

For adversarial training, at every optimization step we generate ℓ_∞-PGD adversarial examples and compute the training loss and gradient on these adversarial inputs. PGD uses 15 attack steps, with perturbation budgets $\epsilon \in \{0.5, 2.0\}$ in pixel-value units and step sizes $\alpha = 0.1\epsilon$, i.e. $\alpha = 0.05$ for $\epsilon = 0.5$ and $\alpha = 0.2$ for $\epsilon = 2.0$.

We evaluate two architectures. The first is a small ResNet with base width 16, GELU activations, and three residual stages of sizes $(3, 3, 3)$, using group normalization and a final linear classifier. The second is a ViT-style Transformer using image patches of size 4×4, embedding dimension 64, depth 4, 8 attention heads, MLP hidden dimension 256, and a linear classification head. These architectures are the same as those in the recent benchmark of the edge-of-stability literature [7]. In all runs, the top eigenvalue of the effective Hessian is estimated every 100 steps using LOBPCG. The sharpness curves report the top Hessian eigenvalue; the dotted horizontal line in each panel denotes the edge-of-stability threshold $2/\eta$.

C.2. Proof of Proposition 4.1

We first note that as in (Adversarial Logistic Loss), the inner max problem of (ρ-adversarial Objective) has an explicit solution:

$$\max_{|\delta| \le \rho} \ell\left((x+\delta)y\right) = \ell\left((x + \rho\,\mathrm{Sign}(x))\,y\right). \tag{31}$$

Thus (Adversarial Logistic Loss) has an explicit formula:

$$\max_{|\delta| \le \rho} \ell\big((x+\delta)y\big) = \begin{cases} \dfrac{\big((x+\rho\,\mathrm{Sign}(x))\,y\big)^2}{2}, & \big|(x+\rho\,\mathrm{Sign}(x))y\big| \le 1, \\ \big|(x+\rho\,\mathrm{Sign}(x))y\big| - \dfrac{1}{2}, & \big|(x+\rho\,\mathrm{Sign}(x))\,y\big| > 1. \end{cases}$$

In the following, we define the region where $\big|(x+\rho\,\mathrm{Sign}(x))y\big| \le 1$ holds as the *quadratic region*, and the region where $\big|(x+\rho\,\mathrm{Sign}(x))y\big| > 1$ holds as the *linear region*.

To simplify the notation, we will also denote

$$q_t := |x + \rho\,\mathrm{Sign}(x)| = |x_t| + \rho.$$

Lemma C.1 *If $y_t > 0$, then in the linear region, the gradient descent dynamic (2) can be written as:*

$$\begin{aligned} x_{t+1} &= x_t - \eta \operatorname{Sign}(x_t)\, y_t, \\ y_{t+1} &= y_t - \eta q_t. \end{aligned} \tag{32}$$

While in the quadratic region, the gradient descent dynamic (2) can be written as:

$$\begin{aligned} x_{t+1} &= (1 - \eta\, y_t^2) x_t - \eta \operatorname{Sign}(x_t)\, \rho\, y_t^2, \\ y_{t+1} &= (1 - \eta\, q_t^2) y_t. \end{aligned} \tag{33}$$

Proof Since we assume $y_t > 0$, we have

$$\big|(x_t + \rho \operatorname{Sign}(x_t)) y_t\big| = \big|x_t + \rho \operatorname{Sign}(x_t)\big| y_t = q_t y_t.$$

We first consider the linear region $\big|(x_t + \rho \operatorname{Sign}(x_t)) y_t\big| > 1$. In this case we have

$$\max_{|\delta| \le \rho} \ell\left((x + \delta) y\right) = \big|(x + \rho \operatorname{Sign}(x)) y\big| - \frac{1}{2} = q_t y_t - \frac{1}{2},$$

and

$$y_{t+1} = y_t - \eta\, \partial_y \max_{|\delta| \le \rho} \ell\left((x_t + \delta) y_t\right) = y_t - \eta q_t,$$

this finished the y-part in (32). For the x-part, we have

$$\begin{aligned} x_{t+1} &= x_t - \eta\, \partial_x \max_{|\delta| \le \rho} \ell\left((x_t + \delta) y_t\right) \\ &= x_t - \eta \left(\partial_x q_t\right) y_t \\ &= x_t - \eta \operatorname{Sign}(x_t)\, y_t, \end{aligned}$$

and this finished the x-part of the (32).

Next we consider the quadratic region $\big|(x_t + \rho \operatorname{Sign}(x)) y_t\big| \le 1$. In this case, we have

$$\max_{|\delta| \le \rho} \ell\left((x_t + \delta) y_t\right) = \frac{\left((x_t + \rho \operatorname{Sign}(x_t))\, y_t\right)^2}{2} = \frac{q_t^2 y_t^2}{2},$$

and

$$y_{t+1} = y_t - \eta\, \partial_y \max_{|\delta| \le \rho} \ell\left((x_t + \delta) y_t\right) = \left(1 - \eta\, q_t^2\right) y_t,$$

this finished the y-part in (33). For the x-part, we have

$$\begin{aligned} x_{t+1} &= x_t - \eta\, \partial_x \max_{|\delta| \le \rho} \ell\left((x_t + \delta) y_t\right) \\ &= x_t - \eta \left(\partial_x q_t^2\right) y_t^2 / 2. \end{aligned}$$

Note that

$$\partial_x q_t^2/2 = q_t \partial_x q_t = \operatorname{Sign}(x_t)\, q_t = \operatorname{Sign}(x_t)\ (|x_t| + \rho) = x_t + \operatorname{Sign}(x_t)\, \rho,$$

thus

$$\begin{aligned} x_{t+1} &= x_t - \eta\, (\partial_x q_t^2)\, y_t/2 \\ &= x_t - \eta\ (x_t + \operatorname{Sign}(x_t)\, \rho)\ y_t^2 \\ &= \left(1 - \eta\, y_t^2\right) x_t - \eta \operatorname{Sign}(x_t)\, \rho\, y_t^2, \end{aligned}$$

this finished the the x-part of the (33). ■

Lemma C.2 *If (x_y, y_t) lies in the quadratic region, we have*

$$\lambda_t^\rho \le (|y_t| + q_t)^2 .$$

Proof We first calculate the Hessian matrix $\nabla^2 \mathcal{L}_\rho(x_t, y_t) = \nabla^2 \left(\max_{|\delta| \le \rho} \ell\left((x_t + \delta) y_t\right)\right)$ in the quadratic region. We have

$$\begin{aligned} \nabla^2 \left(\max_{|\delta| \le \rho} \ell\left((x_t + \delta) y_t\right) \right) &= \nabla^2 \left(\frac{\left((x_t + \rho \operatorname{Sign}(x_t))\, y_t\right)^2}{2} \right) \\ &= \nabla^2 \left(\frac{q_t^2\, y_t^2}{2} \right) \end{aligned}$$

Note that we have

- $\nabla_{xx} \left(\dfrac{q_t^2\, y_t^2}{2} \right) = \nabla_{xx} (q_t^2/2) \cdot y_t^2 = y_t^2$
- $\nabla_{yy} \left(\dfrac{q_t^2\, y_t^2}{2} \right) = q_t^2/2 \cdot \nabla_{yy} (y_t^2) = q_t^2$
- $\nabla_{xy} \left(\dfrac{q_t^2\, y_t^2}{2} \right) = \partial_x \left(q_t^2 y_t \right) = 2 \operatorname{Sign}(x_t)\, q_t\, y_t$

Combining above, we get

$$H_t := \nabla^2 \left(\max_{|\delta| \le \rho} \ell\left((x_t + \delta) y_t\right) \right) = \begin{bmatrix} y_t^2 & 2 \operatorname{Sign}(x_t)\, q_t\, y_t \\ 2 \operatorname{Sign}(x_t)\, q_t\, y_t & q_t^2 \end{bmatrix}.$$

Moreover, since

$$\lambda_t^\rho \le \|H_t\|_\infty := \max\{y_t^2 + 2\, q_t\, y_t,\ q_t^2 + 2\, q_t\, y_t\} \le (|y_t| + q_t)^2,$$

this finished the proof. ■

High-Level Overview of the Proof. In the following, we will analysis the dynamics given by Lemma (C.1). We first provide a high-Level overview of the proof. To study the dynamics, we first introduce the following time horizon, which is motivated by Lemma 6 in [1]:

$$s := \inf\{t \geq 0 \mid x_t < 0 \text{ or } q_t y_t \leq 1\} \tag{34}$$

In other worlds, for all $t \leq s$, the trajectories (x_t, y_t) lies in the linear region, and $x_t \geq 0$. Moreover, for small η, the initialization given by (EoS Regime) must lies in the linear region, thus $s > 0$ and is well-defined.

It turns out that in the region $t \leq s$, the following value

$$D_t := y_t^2 - q_t^2 \tag{35}$$

is nearly conversed. Since the initialization given by (EoS Regime) makes $D_0 = \Omega(1/\eta)$, thus for $t < s$, y_t keeps a magnitude of $\mathcal{O}(1/\sqrt{\eta})$, which result in x_t significantly decreases in each iteration according to (32). In Lemma C.4, we will prove $s = \mathcal{O}(1/\eta)$. This results in the training trajectories enter a small tube with $q_s = \mathcal{O}(\eta)$ as shown in Lemma C.5. Moreover, in Lemma C.6, we prove that once the trajectory enters this tube, it will remain in this tube for the rest of the training time. After the trajectory enter the tube, we will show the y-coordinate of the trajectory gradually be compressed to zero and the trajectory future enter the quadratic region, then we use Lemma C.2 to prove the upper bound on the sharpness when t tends to infinity.

Assumption C.3 *We assume the following assumptions hold in the training process:*

- *(Boundary-avoidance) The training trajectories do not hit the non-smooth set*

$$x_t = 0,\ (|x_t| + \rho)\, y_t = 1$$

- *(Magnitudes of Adversarial Radius) We assume $\rho = \mathcal{O}(\sqrt{\eta})$ and $\rho \neq 0$.*

Note that the boundary-avoidance part of Assumption C.3 is to avoid the cases where the solution of inner-max problem of (ρ-adversarial Objective) is not unique when $x = 0$, and the Hessian on the boundary points $(|x_t| + \rho)\, y_t = 1$ is not well-defined. Similar assumptions are also needed in [1].

Lemma C.4 *Under Assumption C.3, let*

$$s := \inf\{t \geq 0 \mid x_t < 0 \text{ or } q_t y_t \leq 1\}, \qquad \text{(Entrance Time)}$$

then for sufficient small step size η, we have $s \in \mathcal{O}(1/\eta)$.

Proof For any $k \leq s - 1$, according to the definition of (Entrance Time), we have

$$x_k > 0, \quad q_k = x_k + \rho, \quad q_k\, y_k > 1,$$

thus the trajectory lies in the linear region, and the dynamics as defined in (32) can be written as

$$\begin{aligned} q_{k+1} &= q_k - \eta\, y_k \\ y_{k+1} &= y_k - \eta\, q_k. \end{aligned} \tag{36}$$

Define

$$D_k := y_k^2 - q_k^2.$$

We first compute D_0, where (x_0, y_0) is initialize according to (EoS Regime). Since in this case, $q_0 = x_0 + \rho$, we have

$$D_0 = y_0^2 - q_0^2 = (y_0^2 - x_0^2) - 2\rho\, x_0 - \rho^2 = (2+\delta)/\eta - 2\rho\, x_0 - \rho^2.$$

Then using $x_0 = \mathcal{O}(\eta^{-1/2})$ and $\rho = \mathcal{O}(\sqrt{\eta})$ as stated in Assumption C.3, we have

$$D_0 = \Omega(1/\eta). \tag{37}$$

Moreover, in time region $k \le s$, the trajectory lies in the linear region, we have

$$\begin{aligned} D_k = y_k^2 - q_k^2 &= (y_{k-1} - \eta\, q_{k-1})^2 - (q_{k-1} - \eta\, y_{k-1})^2 \\ &= (1-\eta^2)\,(y_{k-1}^2 - q_{k-1}^2) \\ &= (1-\eta^2) D_{k-1}. \end{aligned} \tag{38}$$

Using above recurrence relation, we get

$$y_k \ge \sqrt{D_k} = (1-\eta^2)^{k/2}\sqrt{D_0}. \tag{39}$$

Moreover, we have

$$0 < x_{s-1} = x_0 - \eta \sum_{k=1}^{s-2} y_k \le x_0 - \eta\sqrt{D_0} \sum_{k=0}^{s-2} \left(1-\eta^2\right)^{k/2},$$

where the inequality in the right hand side used (39). Thus we have

$$\sum_{k=0}^{s-2} \left(1-\eta^2\right)^{k/2} \le x_0 / \left(\eta\sqrt{D_0}\right)$$

According to (EoS Regime), we have

$$x_0 = \mathcal{O}(\eta^{-\frac{1}{2}}), \text{ and } \sqrt{D_0} = \Omega(\eta^{-\frac{1}{2}}) \quad \text{from (37)},$$

these gives

$$\sum_{k=0}^{s-2} \left(1-\eta^2\right)^{k/2} = \mathcal{O}(1/\eta).$$

The left hand side is a geometric progression, with an explicit solution given by

$$\sum_{k=0}^{s-2} \left(1-\eta^2\right)^{k/2} = \frac{1 - \left(1-\eta^2\right)^{\frac{s-1}{2}}}{1 - \sqrt{1-\eta^2}}. \tag{40}$$

For the denominator term $1-\sqrt{1-\eta^2} = \frac{\eta^2}{1+\sqrt{1-\eta^2}}$, since $\eta < 1$, we have

$$1 \le 1+\sqrt{1-\eta^2} \le 2,$$

thus

$$\eta^2/2 < 1-\sqrt{1-\eta^2} < \eta^2,$$

which gives

$$\frac{1-\left(1-\eta^2\right)^{\frac{s-1}{2}}}{\eta^2} < \frac{1-\left(1-\eta^2\right)^{\frac{s-1}{2}}}{1-\sqrt{1-\eta^2}} = \sum_{k=0}^{s-2}\left(1-\eta^2\right)^{k/2} = \mathcal{O}(1/\eta).$$

This gives

$$1-\left(1-\eta^2\right)^{\frac{s-1}{2}} = \mathcal{O}(\eta), \quad \text{i.e.,} \quad \exists\, M>0,\ \left(1-\eta^2\right)^{\frac{s-1}{2}} \ge 1-M\eta.$$

Now we take $\log$ to the both side, which gives

$$\frac{s-1}{2} < \frac{\log(1-M\eta)}{\log(1-\eta^2)} = \frac{M}{\eta}\cdot\frac{1+\mathcal{O}(\eta)}{1+\mathcal{O}(\eta^2)} = \mathcal{O}(1/\eta),$$

this finished the proof. ■

Lemma C.5 *Under Assumption C.3, at entrance time s as defined in (Entrance Time), we have $q_s \in \mathcal{O}(\sqrt{\eta})$, i.e., there exists some constant C_q, such that $q_s \le C_q\sqrt{\eta}$. Moreover, we have $y_s = \mathcal{O}(\eta^{-\frac{1}{2}})$ and $y_s > 0$.*

Proof $y_s = \mathcal{O}(\eta^{-\frac{1}{2}})$ is easy to prove since in the linear region $k \le s-1$, we have

$$y_{k+1} = y_k - \eta\, q_k,\ q_k > 0 \Rightarrow y_s \le y_0 = \mathcal{O}(\eta^{-\frac{1}{2}}). \tag{41}$$

According to the definition of s, we have the following possibilities:

- (Case 1): $q_s\, y_s < 1$, and $x_s > 0$
- (Case 2): $x_s < 0$, and $q_s\, y_s > 1$

In the following, we will prove the statement holds in both cases.

We first consider Case 1. Recall that case 1 implies $\forall t \le s$, the trajectory lies in the linear region, where the following recurrence holds according to (38):

$$D_s = D_0\ \left(1-\eta^2\right)^s.$$

Since $s = \mathcal{O}(1/\eta)$ according to Lemma C.4, we have $(1-\eta^2)^s = \Omega(1)$. Moreover, we have $D_0 = \Omega(1/\eta)$ according to (37), these together gives

$$D_s = \Omega(1/\eta).$$

Since $D_s = y_s^2 - q_s^2$, above gives

$$y_s \geq \sqrt{y_s^2 - q_s^2} = \sqrt{D_s} = \Omega(1/\sqrt{\eta}).$$

Moreover, since $q_s\, y_s < 1$ according to the assumption of Case 1, we get

$$q_s = \mathcal{O}(\sqrt{\eta}),$$

and this finished the proof.

Now we consider Case 2. In this case, we have $x_s < 0$ and $x_{k-1} > 0$ for all $k \leq s-1$. This gives

$$x_s = x_{s-1} - \eta\, y_{s-1} < 0 \Rightarrow |x_s| = \eta\, y_{s-1} - x_{s-1} < \eta\, y_{s-1}. \tag{42}$$

Since $y_{s-1} = \mathcal{O}(\eta^{-\frac{1}{2}})$ according to (41), this gives $|x_s| = \mathcal{O}(\eta^{\frac{1}{2}})$. Moreover, under Assumption C.3 with $\rho = \mathcal{O}(\eta^{\frac{1}{2}})$, we get

$$q_s = |x_s| + \rho = \mathcal{O}(\sqrt{\eta}),$$

this finished the proof of q_s part.

Finally, since $D_{s-1} > 0$, we have $y_{s-1} > q_{s-1}$, and $y_s = y_{s-1} - \eta\, q_{s-1} \geq (1-\eta) y_{s-1} > 0$. This finished the y-part of the proof. ■

From Lemma C.5, there exist constant $Q,\ Y$ which are independent of η, such that

$$q_s \leq Q\,\sqrt{\eta},\ 0 < y_s < Y \eta^{-\frac{1}{2}}. \tag{43}$$

Moreover, since in Assumption C.3, we assume $\rho = \mathcal{O}(\eta)$, thus there exists some constant R such that

$$\rho \leq R\,\sqrt{\eta}.$$

Now we fix some constant M such that

$$M \geq \max\{Q,\ Y,\ R,\ Y^2(1+R)\}, \tag{44}$$

and our aim is to prove the following lemma about the invariance of the tube.

Lemma C.6 *Under Assumption C.3 and assume η satisfies $(M+R)\,\eta < 1$, we have for all $t \geq s$, the following holds:*

$$|x_t| \leq M\,\sqrt{\eta},\ 0 < y_t \leq Y\,\eta^{-\frac{1}{2}}. \tag{45}$$

Moreover, y_t is a non-increasing function after $t > s$.

Proof We prove by using induction. The base case $t = s$ is true from (43) and the definition of M.

Assume that for a time point $t \geq s$, the inequalities in (45) hold. Then we need to prove that at time $t + 1$, (45) still holds.

We first note that the assumption $(M + R)\eta < 1$ implies

$$\eta\, q_t^2 \leq \eta\, ((M + R)\sqrt{\eta})^2 = \eta^2\, (M + R)^2 < 1. \tag{46}$$

We dived the discussion into two cases:

- (Case 1): $y_t\, q_t > 1$
- (Case 2): $y_t\, q_t < 1$

where we do not include the case $y_t\, q_t = 1$ for Assumption C.3.

We first consider Case 1. In this region, according to (32), we have

$$y_{t+1} = y_t - \eta\, q_t,\ y_t > 1/q_t. \tag{47}$$

It is easy to see y_t is a non-increasing function. Moreover, since $\eta\, q_t^2 < 1$, we have

$$\eta\, q_t < 1/q_t < y_t,$$

therefore $0 < y_{t+1} < y_t \leq Y\, \eta^{-\frac{1}{2}}$, and this finished y_{t+1} part of the induction. For the x_{t+1} part, we have

$$|x_{t+1}| = |x_t - \eta\, \mathrm{Sign}(x_t)\, y_t| = |\, |x_t| - \eta\, y_t\, | \leq \max\{|x_t|,\ \eta\, y_t\} \leq \max\{M, Y\}\sqrt{\eta},$$

and this finished the proof of Case 1.

Next we consider Case 2. If $y_t\, q_t < 1$, then according to (33), we have

$$y_{t+1} = (1 - \eta\, q_t^2) y_t. \tag{48}$$

Moreover, since $1 - \eta\, q_t^2 < 1$ according to (46), we have

$$y_{t+1} = (1 - \eta\, q_t^2) y_t \leq y_t, \tag{49}$$

and this finished y_{t+1} part of the induction. Next we consider x_{t+1}, in this case, we have

$$\begin{aligned} x_{t+1} &= x_t - \eta\, \mathrm{Sign}(x_t)\, q_t\, y_t^2 \\ &= \mathrm{Sign}(x_t)\, \left((1 - \eta\, y_t^2)|x_t| - \eta\, \rho\, y_t^2\right), \end{aligned}$$

thus

$$|x_{t+1}| = |\, (1 - \eta\, y_t^2)|x_t| - \rho\, \eta\, y_t^2\, |.$$

Recall that we have $\eta\, y_t^2 = \mathcal{O}(1)$ according to the base case. We divide into cases $\eta\, y_t > 1$ and $\eta\, y_t \leq 1$.

If $\eta\, y_t^2 \leq 1$, we have

$$|x_{t+1}| \leq (1 - \eta\, y_t^2)|x_t| + \rho\, \eta\, y_t^2 \leq \max\{|x_t|, \rho\} \leq M\sqrt{\eta}.$$

If $\eta\, y_t^2 > 1$, we have $y_t > 1/\sqrt{\eta}$. Moreover, since $y_t\, q_t < 1$, we have

$$|x_t| < |x_t| + \rho = q_t < 1/y_t < \sqrt{\eta},$$

thus

$$\begin{aligned} |x_{t+1}| &\leq (\eta\, y_t^2 - 1)|x_t| + \rho\, \eta\, y_t^2 \\ &\leq \eta\, y_t^2\, (|x_t| + \rho) \\ &\leq Y^2(1 + R)\sqrt{\eta} \\ &\leq M\sqrt{\eta}, \end{aligned}$$

and this finished the proof. ■

In the following, we combine Lemma C.4-C.6 to provide a proof of Proposition 4.1.

Proof [Proof of Proposition 4.1] From Lemma C.6, we know $y_t > 0$ and it is a non-increasing function of t, thus $\lim_{t\to\infty} y_t$ exists, we denote

$$y_\infty := \lim_{t\to\infty} y_t \geq 0.$$

We will show in fact $y_\infty = 0$. If $y_\infty > 0$, note that in the linear region $q_t\, y_t > 1$, after each iteration y_t will decrease $\eta\, q_t > \eta\, \rho$. Thus there must exists some $T > 0$, such that for $t > T$, the trajectory (x_t, y_t) always lies in the quadratic region where $q_t\, y_t < 1$, and we have

$$y_{t+1} = (1 - \eta\, q_t^2)\, y_t.$$

Since $q_t > \rho$, we have $y_{t+1} \leq (1 - \eta\, \rho^2) y_t$, and this enforces $\lim_{t\to\infty} y_t = 0$.

Finally, according to Lemma C.2 and Lemma C.6, we have

$$\lambda_t^\rho \leq (y_t + q_t)^2,\ q_t = \mathcal{O}(\sqrt{\eta}).$$

Through taking the limit in above as $t \to \infty$, we get

$$\lim_{t\to\infty} \lambda_t^\rho \leq \limsup_{t\to\infty} q_t^2 = \mathcal{O}(\eta),$$

and this finished the proof.

■