STABILITY AND GENERALIZATION OF ADVERSARIAL DIFFUSION TRAINING

*Hesam Hosseini** *Ying Cao* *Ali H. Sayed*

School of Engineering, École Polytechnique Fédérale de Lausanne

ABSTRACT

Algorithmic stability is an established tool for analyzing generalization. While adversarial training enhances model robustness, it often suffers from robust overfitting and an enlarged generalization gap. Recent work has established the convergence of adversarial training in decentralized networks, but its generalization properties remain unexplored. This work presents a stability-based generalization analysis of adversarial training under the *diffusion strategy* for *convex losses*. We derive a bound showing that the generalization error grows with both the adversarial perturbation strength and the number of training steps, a finding consistent with the single-agent case but novel for decentralized settings. Numerical experiments on logistic regression validate these theoretical predictions.

***Index Terms*—** generalization, stability, adversarial training, distributed learning

1. INTRODUCTION

Machine learning algorithms are designed to fit the training data, as well as generalize to unseen samples, which marks an important distinction between machine learning and a purely optimization-based viewpoint. The difference between empirical and population performance, known as the *generalization gap*, is a central topic in statistical learning theory [1, 2].

One framework for understanding generalization is *algorithmic stability*, which relates the sensitivity of an algorithm to perturbations in the training set to its generalization behavior [3]. Stable algorithms exhibit small changes in their outputs when a single training example is modified, and this property directly yields generalization bounds.

Adversarial training has emerged as an effective method for enhancing robustness against adversarial examples [4, 5]. However, robust training often suffers from *robust overfitting*, where the generalization gap becomes larger than in standard training [6, 7]. Stability-based analyses have shown that adversarial perturbations degrade stability and enlarge generalization bounds, with the deterioration scaling both with the perturbation radius and the number of training steps [8].

In parallel, diffusion-based algorithms have become a popular strategy for decentralized learning due to their scalability and communication efficiency [9]. Recent work has extended generalization analysis to distributed learning in the clean (non-adversarial) case [10]. While adversarial diffusion training was introduced in [11] for distributed setting, its generalization behavior remains unexplored.

In this work, we address this gap by providing a generalization analysis for adversarial training under distributed strategies. While our analysis is restricted to convex loss functions due to space limitations, it offers crucial insights into the problem. Our specific contributions are:

1. We derive a stability-based generalization bound for adversarial diffusion training under convex losses. By proving that the algorithm satisfies uniform stability, we show its generalization error grows with the perturbation strength ϵ and the number of iterations T. Our analysis provides a unifying framework, as our bounds seamlessly reduce to known results for single-agent adversarial training [8] and decentralized standard training [10] in their respective limits.

2. We combine this stability result with existing optimization guarantees [11] to characterize the trade-off between optimization and generalization, suggesting practical strategies such as early stopping.

3. We illustrate our theoretical predictions through numerical experiments on logistic regression, confirming the dependence on ϵ and T. Furthermore, our experiments provide new empirical evidence on the influence of network topology.

2. PROBLEM SETTING

In decentralized learning, multiple agents collaborate to optimize a global objective without centralizing their data. This setup provides benefits in terms of scalability and privacy. When agents are exposed to adversarial perturbations, quantifying how these perturbations affect generalization is essential and underexplored. In this work, we formalize this scenario.

We consider a network of K agents connected by a graph topology. The topology is characterized by a doubly stochas-

*This work was performed while the first author was a summer intern at the Adaptive Systems Laboratory at EPFL, Switzerland.

Emails: seyed.hosseini@epfl.ch, ying.cao@epfl.ch, ali.sayed@epfl.ch

tic combination matrix $A = [a_{\ell k}]$, where $a_{\ell k} \geq 0$, and the entries in each column sum to one. Moreover, for any two agents k and ℓ, if $a_{\ell k} > 0$, there exists a communication link from ℓ to k.

Each agent k observes independent samples of random variables $(\boldsymbol{x}, \boldsymbol{y})$ drawn from a local distribution $\mathcal{D}_k$. Here $\boldsymbol{x}$ plays the role of a feature vector and $\boldsymbol{y}$ plays the role of a label. For a convex loss function $Q_k(w; \boldsymbol{x}, \boldsymbol{y})$, the robust local risk is defined as

$$R_k(w) \triangleq \mathbb{E}_{(x,y)\sim\mathcal{D}_k}\left[\max_{\delta\in\Delta_\epsilon} Q_k(w; \boldsymbol{x}+\delta, \boldsymbol{y})\right], \qquad (1)$$

where $\Delta_\epsilon = \{\delta : \|\delta\| \leq \epsilon\}$ represents the set of admissible adversarial perturbations. The inner maximization models worst-case perturbations within a radius ϵ, enabling each agent to train robustly against adversarial examples. The global robust risk aggregates the local risks as

$$R(w) \triangleq \sum_{k=1}^{K} \pi_k R_k(w), \qquad w^\star \triangleq \arg\min_{w\in\mathbb{R}^M} R(w). \qquad (2)$$

using weight coefficients $\pi_k \geq 0$ that satisfy $\sum_{k=1}^{K} \pi_k = 1$.

In practice, the distributions $\{\mathcal{D}_k\}$ are unknown, and each agent only has access to a finite dataset

$$S_k \triangleq \{(x_{k,i}, y_{k,i})\}_{i=1}^{N}, \qquad S \triangleq (S_1, \ldots, S_K). \qquad (3)$$

Each agent is then associated with the following local empirical robust risk

$$\widehat{R}_k(w) \triangleq \frac{1}{N}\sum_{i=1}^{N} \max_{\delta\in\Delta_\epsilon} Q_k(w; x_{k,i}+\delta, y_{k,i}), \qquad (4)$$

leading to the global empirical objective

$$R_S(w) \triangleq \sum_{k=1}^{K} \pi_k \widehat{R}_k(w), \qquad \widehat{w} \triangleq \arg\min_{w} R_S(w). \qquad (5)$$

In this work, we focus on using the adversarial diffusion strategy [11] to solve (5). For simplicity we assume that the inner maximization in $\widehat{R}_k(w)$ has a unique solution, ensuring the adversarial loss

$$g_k(w; x, y) \triangleq \max_{\delta\in\Delta_\epsilon} Q_k(w; x+\delta, y) \qquad (6)$$

is differentiable. This property holds for common loss functions, such as logistic regression under typical perturbation sets, as illustrated in [11]. Agents then update via the adapt-then-combine (ATC) diffusion recursion:

$$\phi_{k,n} = w_{k,n-1} - \mu_n \nabla g_k(w_{k,n-1}; x_{k,n}, y_{k,n}), \qquad (7)$$

$$w_{k,n} = \sum_{\ell\in\mathcal{N}_k} a_{\ell k}\, \phi_{\ell,n}, \qquad (8)$$

with step-size $\mu_n > 0$. This recursion incorporates neighbor information while adapting to local observations. For each agent k after T iterations, the algorithm outputs

$$F_k(S) \triangleq w_{k,T} \qquad (9)$$

The quality of the learned solution is measured by the expected excess risk

$$\mathbb{E}_{F,S}[R(F_k(S)) - R(w^\star)] \leq \underbrace{\mathbb{E}_{F,S}[R(F_k(S)) - R_S(F_k(S))]}_{\epsilon_{\text{gen}}} + \underbrace{\mathbb{E}_{F,S}[R_S(F_k(S)) - R_S(\widehat{w})]}_{\epsilon_{\text{opt}}}, \qquad (10)$$

where the expectation is taken over both the randomness of the data and the algorithm's internal stochasticity. The term ϵ_{gen} quantifies the generalization gap between empirical and population performance, and ϵ_{opt} captures suboptimality due to decentralized optimization. While ϵ_{opt} has been studied extensively in prior work [11], our focus is on bounding ϵ_{gen} via algorithmic stability, which we address next.

3. STABILITY ANALYSIS

We now analyze the stability of the adversarial diffusion strategy (7)–(8) under convex losses to establish a stability-based generalization bound. Our approach follows a four-step recipe: (i) assume smoothness of local losses Q_k, (ii) derive approximate Lipschitz properties of the adversarial objective g_k, (iii) relate stability to generalization via average model stability, and (iv) show that the adversarial diffusion recursion satisfies stability condition.

We adopt standard smoothness assumptions, which are commonly used in the context of decentralized learning and adversarial training [3, 8–13].

Assumption 1 (Smooth loss functions). *For any data point (x, y) and any two model parameters w, w', the loss function $Q_k(w; x, y)$ satisfies:*

$$\|Q_k(w; x, y) - Q_k(w'; x, y)\| \leq L_w \|w - w'\|, \qquad (11)$$

$$\|\nabla_w Q_k(w; x, y) - \nabla_w Q_k(w'; x, y)\| \leq L_{ww}\|w - w'\|, \qquad (12)$$

$$\|\nabla_w Q_k(w; x, y) - \nabla_w Q_k(w; x', y)\| \leq L_{wx}\|x - x'\|. \qquad (13)$$

□

These conditions make the loss well-behaved; they hold for many convex models [8, 11].

Consider the adversarial loss defined in (6), then g_k inherits the convexity and smoothness properties from Q_k up to a perturbation-dependent term [8, 11].

Lemma 1 (Convexity of adversarial loss). *If $Q_k(w; x, y)$ is convex in w, then $g_k(w; x, y)$ is also convex in w.* □

Lemma 2 (Lipschitz properties of adversarial loss). *Let Q_k satisfy Assumption 1. Then, for all w_1, w_2 and any (x, y), it holds that:*

$$\|g_k(w_1; x, y) - g_k(w_2; x, y)\| \leq L_w \|w_1 - w_2\|. \tag{14}$$

$$\|\nabla g_k(w_1; x, y) - \nabla g_k(w_2; x, y)\| \leq L_{ww} \|w_1 - w_2\| + 2L_{wx}\epsilon. \tag{15}$$

□

The additive term proportional to ϵ in (15) captures the sensitivity introduced by adversarial perturbations. We adopt the notion of on-average model stability [14] to connect algorithmic stability to generalization.

Definition 1 (On-average model stability). *Let S and $\tilde{S}$ be two independent datasets as defined in (3), and let $S^{(ij)}$ denote S with the i-th sample of agent j replaced by its counterpart in $\tilde{S}$. For agent k a randomized algorithm F_k is* on-average η-stable *if*

$$\mathbb{E}_{S,\tilde{S},F}\left[\frac{1}{KN}\sum_{j=1}^{K}\sum_{i=1}^{N}\|F_k(S) - F_k(S^{(ij)})\|\right] \leq \eta. \tag{16}$$

□

Intuitively, this notion captures the idea that a stable algorithm should not change its output significantly when a single sample is modified. The smaller the parameter η, the less sensitive the algorithm is to individual data points, which in turn according to below lemma implies better generalization.

Lemma 3 (Generalization via stability [14]**).** *If F_k is on-average η-stable and (11) is satisfied, then*

$$\left|\mathbb{E}_{F,S}[R(F_k(S)) - R_S(F_k(S))]\right| \leq L_w \eta. \tag{17}$$

□

Hence, to bound generalization, it suffices to bound the on-average stability of the adversarial diffusion algorithm.

Recall that the adaptation step is

$$G_{\mu,g}(w) = w - \mu \nabla_w g(w; x, y). \tag{18}$$

A key property to guarantee stability is the *non-expansiveness* of the adaptation step: the distance between two iterates w_1, w_2 remains controlled under $G_{\mu,g}$. It is well-known that SGD is non-expansive under the assumed Lipschitz conditions [3] and similarly adversarial updates are approximately non-expansive up to a perturbation-dependent term [8]:

Lemma 4 (Adversarial adaptation update). *Under Lemmas 1, 2, if $\mu < 1/L_{ww}$, then for any w_1, w_2, it holds that,*

$$\|G_{\mu,g}(w_1) - G_{\mu,g}(w_2)\| \leq \|w_1 - w_2\| + 2\mu L_{wx}\epsilon. \tag{19}$$

□

This lemma implies that the adversarial adaptation step approximately preserves stability up to a controlled additive error proportional to ϵ.

Theorem 1 (Generalization bound for adversarial diffusion training). *Consider K agents, each with N training samples, running the adversarial diffusion algorithm in* (7)–(8) *with step-sizes $\mu_n \leq 1/L_{ww}$. let $F_k(S) = w_{k,T}$ be the model parameter of agent k after T iterations, then under assumption 1 and Lemmas 3 and 4, it holds for every agent k that*

$$\left|\mathbb{E}[R(F_k(S)) - R_S(F_k(S))]\right| \;\leq\; 2L_w\left(L_{wx}\epsilon + \frac{L_w}{KN}\right)\sum_{n=1}^{T}\mu_n. \tag{20}$$

□

This bound highlights two sources of generalization error: (i) the adversarial perturbation bounded by ϵ, and (ii) the cumulative effect of training iterations in $\sum_{n=1}^{T}\mu_n$. In the special case of a fixed step size $\mu_n = \mu$, the bound simplifies to

$$\left|\mathbb{E}[R(F_k(S)) - R_S(F_k(S))]\right| \;\leq\; 2L_w\mu T\left(L_{wx}\epsilon + \frac{L_w}{KN}\right), \tag{21}$$

making the dependence on training duration T explicit. This result generalizes prior findings in single-agent adversarial training [8] and distributed training without perturbations [10] to the adversarial diffusion setting.

4. OPTIMIZATION-GENERALIZATION TRADE-OFF

As shown in (10), the performance of a learning algorithm is captured by the expected excess risk, which reflects both optimization and generalization errors. We analyze this trade-off for decentralized adversarial training by combining our generalization results with optimization guarantees. Theorem 1 implies that for a constant step-size μ

$$\epsilon_{\text{gen}} = O\left(\mu T\left(\epsilon + \frac{1}{KN}\right)\right) \tag{22}$$

while for the optimization error, the convergence analysis in [11] yields:

$$\epsilon_{\text{opt}} = O\left(\frac{1}{\mu T}\right) + O\left(\mu\right) \tag{23}$$

Combining these bounds gives the overall excess risk:

$$\mathbb{E}[R(F_k(S))) - R(w^\star)] = O\left(\mu T\left(\epsilon + \frac{1}{KN}\right)\right) + O\left(\frac{1}{\mu T}\right) + O\left(\mu\right). \tag{24}$$

This expression encapsulates the core trade-off in decentralized adversarial training. The $O(\epsilon\mu T)$ term represents

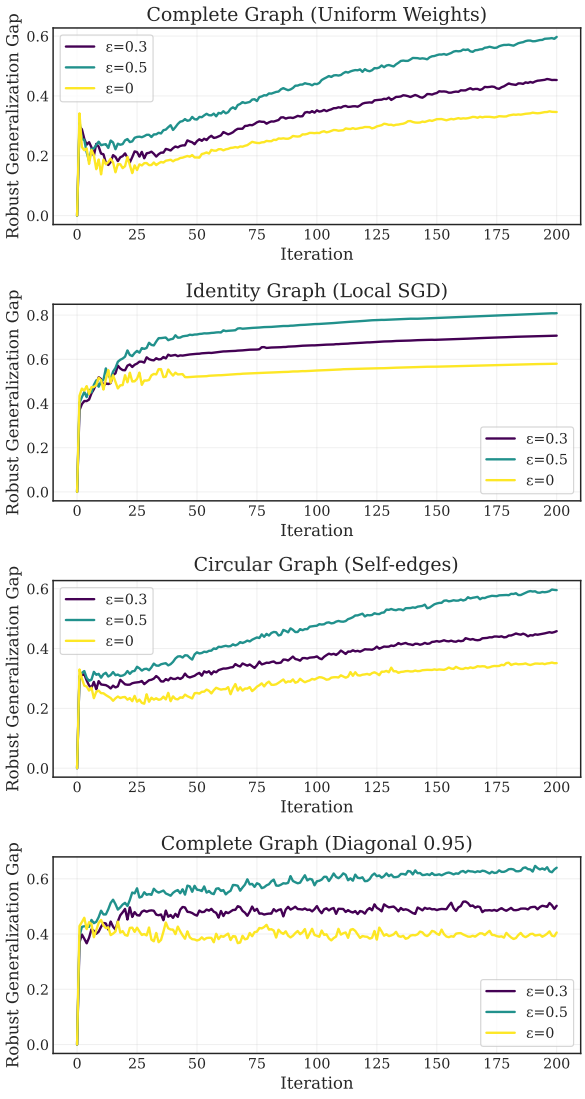

Fig. 1: Robust generalization gap for different communication graphs.

the price of robustness, which grows with both the step size and the number of training iterations. The competing $O(\mu T/KN)$ and $O(1/\mu T)$ terms demonstrate the fundamental optimization-generalization trade-off: longer training improves optimization but harms generalization, an effect mitigated by the total data size KN. The $O(\mu)$ term reflects the asymptotic noise controlled by the step size.

This analysis yields design guidelines: early stopping is necessary to halt training before generalization error dominates; a decaying step-size is useful to balance convergence speed with final error; and collaboration across the network which implicitly increases N, directly improves generalization by allowing more training without overfitting.

5. COMPUTER SIMULATIONS

We illustrate the theoretical predictions of Theorem 1 through logistic regression experiments, a convex setting aligned with our assumptions yet challenging under adversarial training.

To illustrate the generalization error, we consider a network of $K = 10$ agents, each with a local dataset S_k of $N = 10$ samples. The robust local risk at agent k is

$$R_k(w) = \mathbb{E}\Big[\max_{\|\delta\|\leq\epsilon} \ln(1 + e^{-y(\boldsymbol{x}+\delta)^\top w})\Big], \tag{25}$$

whose inner maximization under an ℓ_2-norm constraint admits the closed-form solution via the fast gradient method (FGM) [2, 15]:

$$\delta^\star = -\epsilon\, y\, \frac{w}{\|w\|}. \tag{26}$$

Synthetic data is generated following [10]. Each sample has a label $y \in \{-1, 1\}$ (uniformly assigned) and a feature vector $\boldsymbol{x} \in \mathbb{R}^{200}$, drawn from $\mathcal{N}(\mathbf{1}, I)$ for $y = 1$ and $\mathcal{N}(-\mathbf{1}, I)$ for $y = -1$, with label flips at probability 0.1 to avoid linear separability.

We implement the adversarial diffusion algorithm (7)–(8) with step-size $\mu = 0.03$ and evaluate the robust generalization gap over 15 trials. To study connectivity effects, we test four topologies: (1) Complete graph ($a_{\ell k} = 1/K$), (2) Isolated agents ($A = I$), (3) Circular graph ($a_{\ell k} = 1/3$ for self and neighbors), and (4) Star-like graph ($a_{kk} = 0.95, 0.05$ for neighbors).

Figure 1 shows the robust generalization gap across topologies as the perturbation radius ϵ and the number of training iterations T vary. Results confirm Theorem 1: the gap increases with ϵ and with T.

The experimental results also reveal the influence of network topology. While trends with ϵ and T are almost consistent, the magnitude of the generalization gap depends on the communication graph: well-connected (complete) graphs achieve lower gaps than sparsely isolated agents. Incorporating network connectivity into theoretical bounds is a promising direction for future work.

6. CONCLUSION

We analyzed the generalization performance of decentralized adversarial training via algorithmic stability, deriving a bound for convex loss functions over arbitrary graphs. Our results reveal a fundamental trade-off between optimization and generalization. Experiments with logistic regression corroborate the bound, showing that the generalization error increases with perturbation radius and training time, varying by graph structure. Although we focused on the convex setting to maintain theoretical rigor, this work establishes a foundation for future inquiry into non-convex objectives and the role of graph design in mitigating robust overfitting.

7. REFERENCES

[1] Olivier Bousquet and André Elisseeff, "Stability and generalization," *Journal of Machine Learning Research*, vol. 2, pp. 499–526, 2002.

[2] A. H. Sayed, *Inference and Learning from Data: Learning*, Cambridge University Press, 2022.

[3] M. Hardt, B. Recht, and Y. Singer, "Train faster, generalize better: Stability of stochastic gradient descent," in *Proceedings of International Conference on Machine Learning*. PMLR, 2016, pp. 1225–1234.

[4] T. Bai, J. Luo, J. Zhao, B. Wen, and Q. Wang, "Recent advances in adversarial training for adversarial robustness," in *Proceedings of International Joint Conference on Artificial Intelligence*, 2021, pp. 4312–4321.

[5] W. Zhao, S. Alwidian, and Q. H. Mahmoud, "Adversarial training methods for deep learning: A systematic review," *Algorithms*, vol. 15, no. 8, 2022.

[6] Leslie Rice, Eric Wong, and J. Zico Kolter, "Overfitting in adversarially robust deep learning," in *International Conference on Machine Learning (ICML)*, 2020, pp. 8093–8104.

[7] C. Yu, B. Han, L. Shen, J. Yu, C. Gong, M. Gong, and T. Liu, "Understanding robust overfitting of adversarial training and beyond," in *International Conference on Machine Learning*. PMLR, 2022, pp. 25595–25610.

[8] J. Xiao, Y. Fan, R. Sun, J. Wang, and Z.-Q. Luo, "Stability analysis and generalization bounds of adversarial training," *Advances in Neural Information Processing Systems*, vol. 35, pp. 15446–15459, 2022.

[9] A. H. Sayed, "Adaptation, learning, and optimization over networks," *Foundations and Trends in Machine Learning*, vol. 7, pp. 311–801, 2014.

[10] B. Le Bars, A. Bellet, M. Tommasi, K. Scaman, and G. Neglia, "Improved stability and generalization guarantees of the decentralized SGD algorithm," in *Proceedings of International Conference on Machine Learning*, 2024.

[11] Y. Cao, E. Rizk, S. Vlaski, and A. H. Sayed, "Decentralized adversarial training over graphs," *IEEE Transactions on Information Theory*, vol. 71, no. 7, pp. 5570–5600, 2025.

[12] A. Sinha, H. Namkoong, R. Volpi, and J. Duchi, "Certifying some distributional robustness with principled adversarial training," in *Proc. International Conference on Learning Representations*, 2018, pp. 1–34.

[13] S. Vlaski and A. H. Sayed, "Distributed learning in non-convex environments—part i: Agreement at a linear rate," *IEEE Transactions on Signal Processing*, vol. 69, pp. 1242–1256, 2021.

[14] Y. Lei and Y. Ying, "Fine-grained analysis of stability and generalization for stochastic gradient descent," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5809–5819.

[15] T. Miyato, A. M. Dai, and I. Goodfellow, "Adversarial training methods for semi-supervised text classification," in *Proc. International Conference on Learning Representations*, 2017, pp. 1–11.