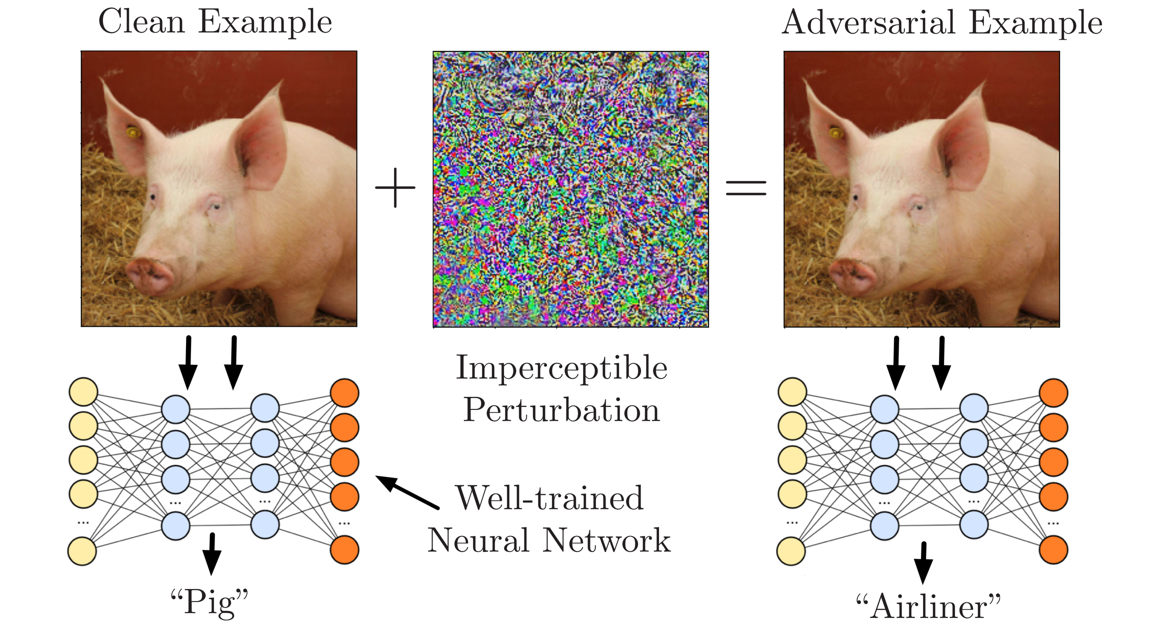
Implicit Bias of Gradient Descent Based Adversarial Training on Separable Data

Yan Li*, Ethan X.Fang◇, Huan Xu*, Tuo Zhao*
*Georgia Tech ◇Pennsylvania State University

Background

▷ **Adversarial Examples**: "Flying Pig" [2].



All current deep neural network (DNN) models are subject to adversarial examples.

▷ **Practical Implications**:
- Autonomous driving system.
- Biometric authentication system.

How could we obtain a robust model? Unfortunately, no definite answer with theoretical justification yet.

Popular practice: **Adversarial training**. Directly minimize the worst-case loss for a given perturbation set Δ:

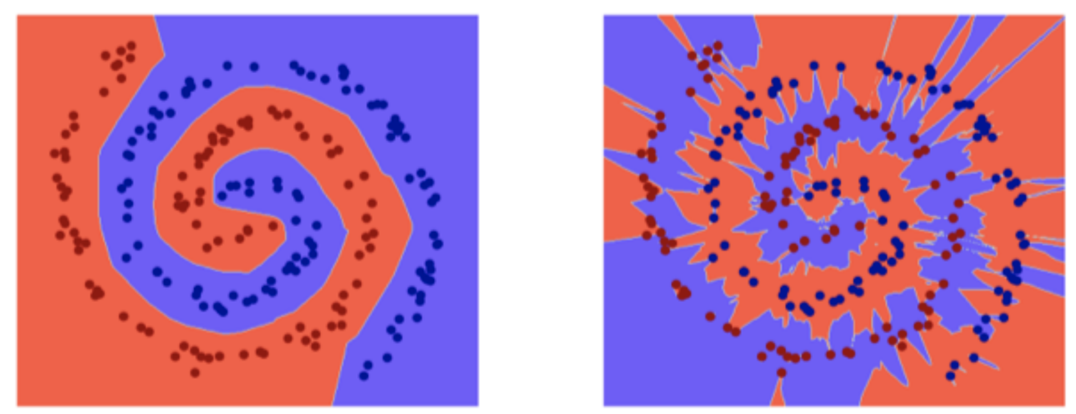
$$\theta_{\text{robust}} = \underset{\theta \in \mathbb{R}^d}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} \max_{\delta_i \in \Delta} \ell(x_i + \delta_i, y_i, \theta).$$

Question: How does adversarial training promote robustness?
Existing Results: Robust VC-dimension, Distributionally robust optimization, Adversarial risk via function transformation, etc.

All existing results hold uniformly within a function class, while SGD only explores a limited subset of the class.

We propose to study from a computational perspective which has implications on learning theory.

Implicit Bias: Neural network can easily overfit training data. Training algorithm biases toward a certain kind of solutions.



(a). (b).

Implicit Bias of Algorithms: network (a) is learnt by SGD (Smooth Boundary). Both networks overfits training data. Only network (a) generalizes well.

Implicit Bias of Gradient Descent

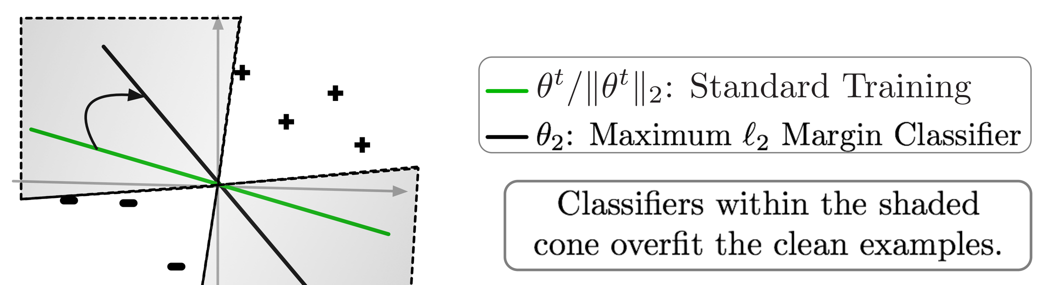
Directly analyzing DNNs is beyond current technical limit.

▷ A simplified yet non-trivial example, training a linear classifier on linearly separable data $\{(x_i, y_i)\}_{i=1}^n$. We aim to solve

$$\min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i x_i^\top \theta), \ell \text{ exponential/logistic loss.} \quad (1)$$

- Only the **direction** of the linear classifier is important.
- There is **no finite minimizer** of $\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(y_i x_i^\top \theta)$.

But there exists infinite amount of solutions at infinity.



- **Implicit bias** [1, 3] of gradient descent to solve (1) :

$$1 - \langle \theta^t / \|\theta^t\|_2, \theta_2 \rangle = \mathcal{O}(\log n / \log t),$$

where θ_q (here $q = 2$) and the optimal value γ_q is defined by:

$$\theta_q = \underset{\|\theta\|_p = 1}{\text{argmax}} \min_{i=1,\ldots,n} y_i x_i^\top \theta, \quad \text{with } 1/p + 1/q = 1.$$

GDAT on Separable Data

GDAT on Separable Data with ℓ_q Perturbation

Input: Data points $\{(x_i, y_i)\}_{i=1}^n$, perturbation level $c < \gamma_q$ and step sizes $\{\eta^t\}_{t=0}^{T-1}$.
Init: Set $\theta^0 = 0$.
For $t = 0 \ldots T-1$:
  For $i = 1 \ldots n$, solve $\widehat{\delta}_i = \text{argmax}_{\|\delta_i\|_q \leq c} \ell(y_i x_i^\top \theta^t)$.
  Set $\widetilde{x}_i = x_i + \widehat{\delta}_i$, for $i = 1 \ldots n$.
  Update $\theta^{t+1} = \theta^t - (\eta^t/n) \cdot \sum_{i=1}^n \nabla \ell(y_i \widetilde{x}_i \theta^t)$.

Question: When can GDAT possess implicit bias?

Theorem 1. *When perturbation level $c < \gamma_q$, no finite stationary point exists for $\mathcal{L}_{\text{adv}}(\theta)$. For $c > \gamma_q$, $\mathcal{L}_{\text{adv}}(\theta)$ admits a unique finite minimizer.*

Remarks:

- No finite minimizer ⇒ investigate implicit bias.
- Minimization with no finite solution is rarely studied in the optimization literature.
- For non-separable data, adversarial training is equivalent to regularization [4].

Questions: Can we characterize the implicit bias of GDAT on separable data? How is it related to adversary geometry?

GDAT Adapts to Adversary Examples

Consider the following large margin classifier:

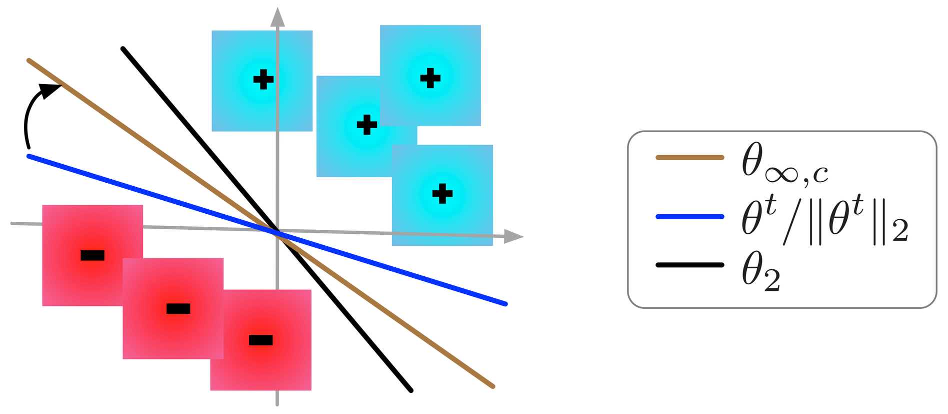
$$\theta_{q,c} = \underset{\|\theta\|_2 = 1}{\text{argmax}} \min_{i=1,\ldots,n} \min_{\|\delta_i\|_q \leq c} y_i (x_i + \delta_i)^\top \theta.$$

Robustness: $\theta_{q,c}$ is in the same direction to the solution of

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_2 \text{ s.t. } y_i \widetilde{x}_i^\top \theta \geq 1 \text{ for all } \|\widetilde{x}_i - x_i\|_q \leq c, \forall i = 1 \ldots n.$$

Minimum mix-norm: $\theta_{q,c}$ is in the same direction to the solution of (here $1/p + 1/q = 1$)

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_2 + \eta(c) \|\theta\|_p \text{ s.t. } y_i x_i^\top \theta \geq 1, \forall i = 1 \ldots n.$$



Theorem 2. *Let $c < \gamma_q$, $\eta^0 = 1$ and $\eta^t = \eta \leq \min\{1/M_p, 1\}$ for $t \geq 1$, where $M_p = \left[(1 + c\sqrt{d})^2 + \frac{c(p-1)}{\gamma_{2,q}} d^{\frac{3p-2}{2p-2}}\right] \exp(c\sqrt{d})$. Then*

$$1 - \langle \theta^t / \|\theta^t\|_2, \theta_{q,c} \rangle = \mathcal{O}(\log n / \log t).$$

GDAT Accelerates Convergence ($q = 2$)

Theorem 3. *Let c and total number of iterations T satisfy $\gamma_2 - c = \left(\frac{n^{1+1/\alpha} \log T}{\eta T}\right)^{1/2}$, set $\eta^0 = 1$ and $\eta^t = \eta$ for $t = 1 \ldots T-1$. We have $\theta_{2,c} = \theta_2$, and*

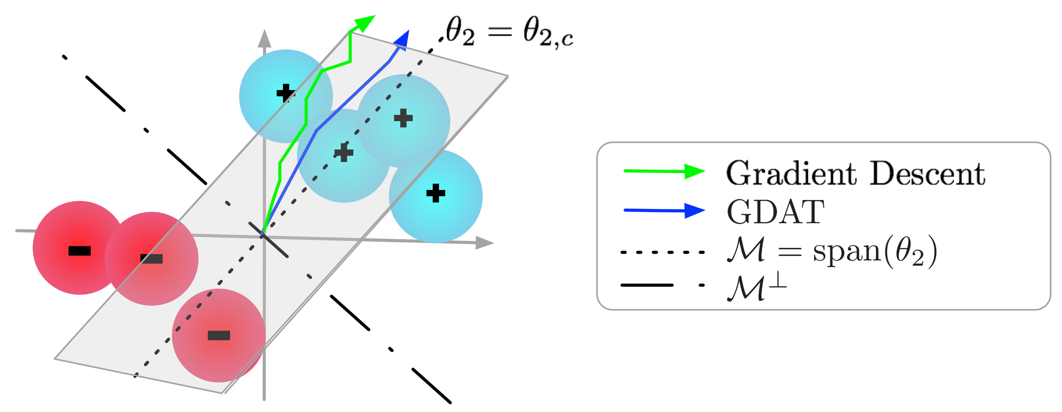
$$1 - \langle \theta^T / \|\theta^T\|_2, \theta_2 \rangle = \mathcal{O}\left(\frac{n^{(1+1/\alpha)/2} K \log T}{\sqrt{\eta T}}\right).$$

Exponential Acceleration by GDAT!

Key Technical Ingredients:

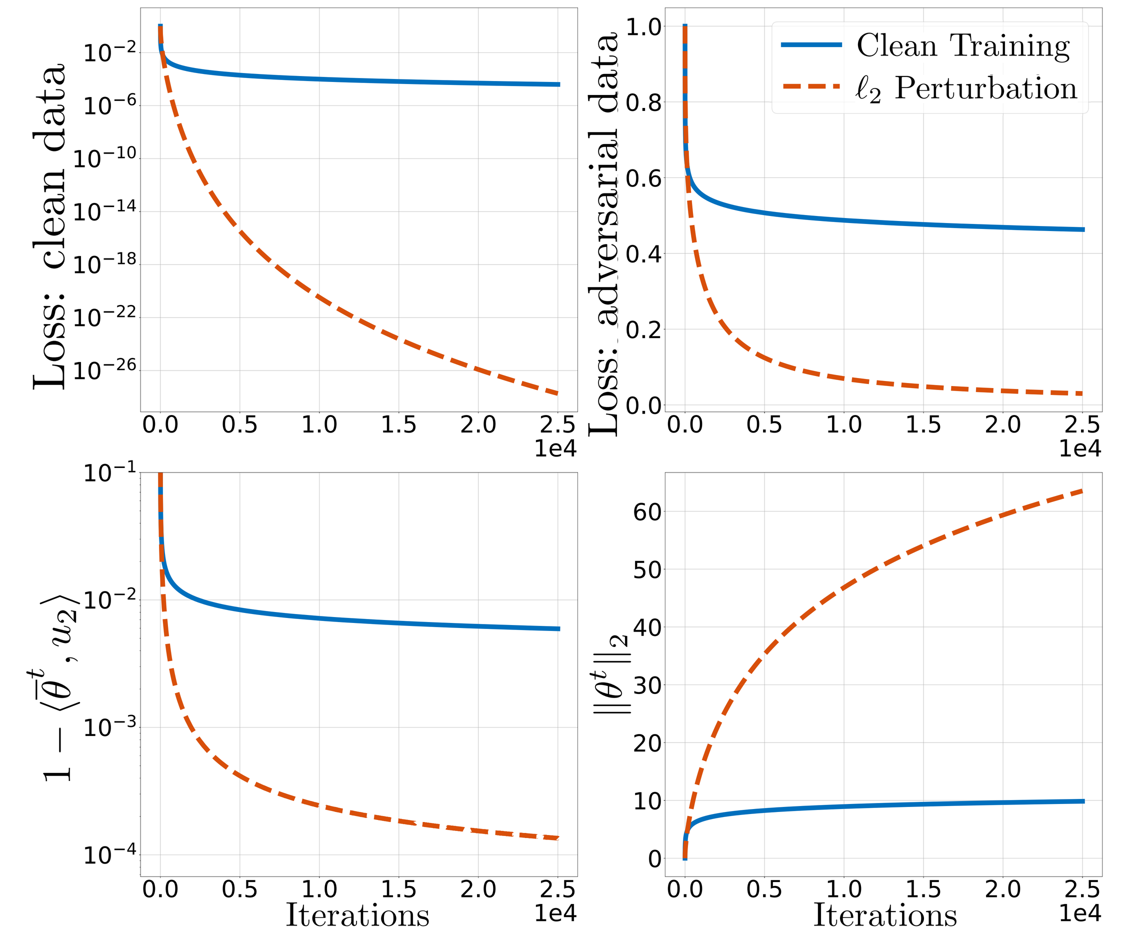
- Projection of θ^t onto the orthogonal space $\mathcal{M}^\perp = \{\theta : \langle \theta, \theta_2 \rangle = 0\}$ is bounded for all $t \geq 0$.
- For projection of θ^t onto the space $\mathcal{M} = \text{span}(\theta_2)$, its increment satisfies **Generalized Perceptron Lemma**:

$$\langle \theta^{t+1} - \theta^t, \theta_2 \rangle \geq \eta^t \mathcal{L}_{\text{adv}}(\theta^t)(\gamma_2 - c).$$



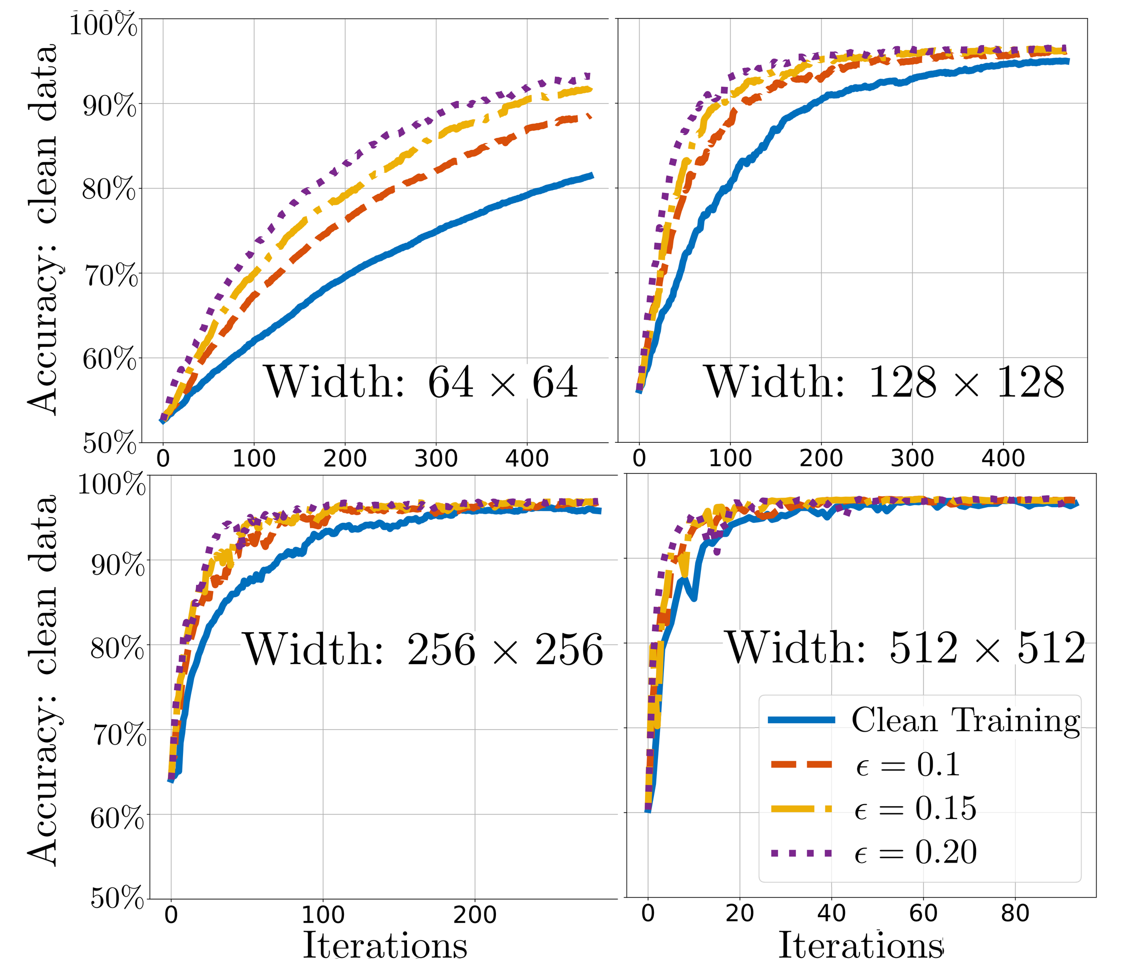
Experiments

Linear Classifiers: We generate data with $\gamma_2 = 1$. We set $c = 0.95$. $\eta = 0.1$ for GDAT and $\eta = 1$ for standard training.



Clean Training v.s. GDAT (ℓ_2 perturbation)

Neural Networks: We use MNIST dataset. Network consists of one hidden layer. The width of hidden layer varies in $\{64 \times 64, 128 \times 128, 256 \times 256, 512 \times 512\}$. We use ℓ_∞ perturbation with perturbation level $\epsilon \in \{0.1, 0.15, 0.20\}$.



References

[1] Ji, Z. and Telgarsky, M. (2019). The implicit bias of gradient descent on nonseparable data. In *Proceedings of the Thirty-Second Conference on Learning Theory*.

[2] Madry, A., Makelov, A., Schmidt, L., Tsipras, D. and Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*.

[3] Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S. and Srebro, N. (2018). The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research* **19** 2822–2878.

[4] Xu, H., Caramanis, C. and Mannor, S. (2009). Robustness and regularization of support vector machines. *Journal of Machine Learning Research* **10** 1485–1510.