OptiML Group Meeting, 2022-2023 Winter

Adversarial training descends without descent: Finding actual descent directions based on Danskin's Theorem [Latorre et al., 2023]

Presenter: Hanseul Cho

February 1st, 2023

Overview

Published as a conference paper at ICLR 2018

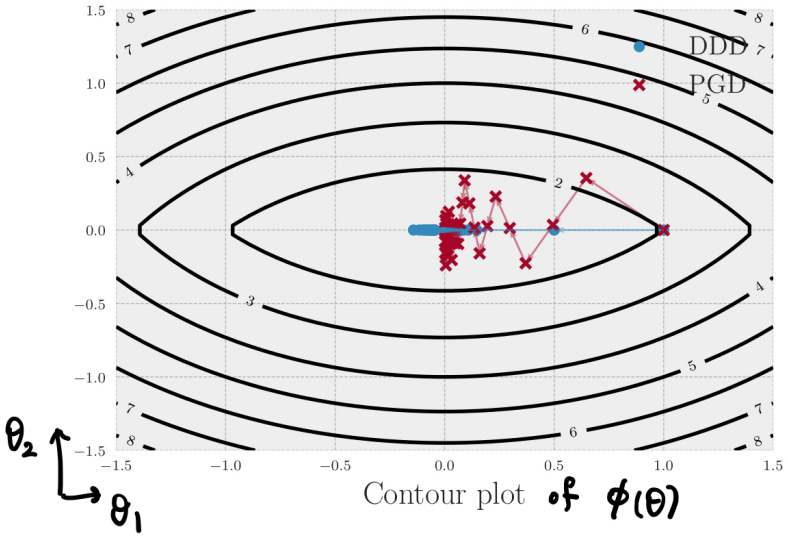
TOWARDS DEEP LEARNING MODELS RESISTANT TO ADVERSARIAL ATTACKS

Aleksander Mądry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, Adrian Vladu*
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
{madry,amakelov,ludwigs,tsipras,avladu}@mit.edu

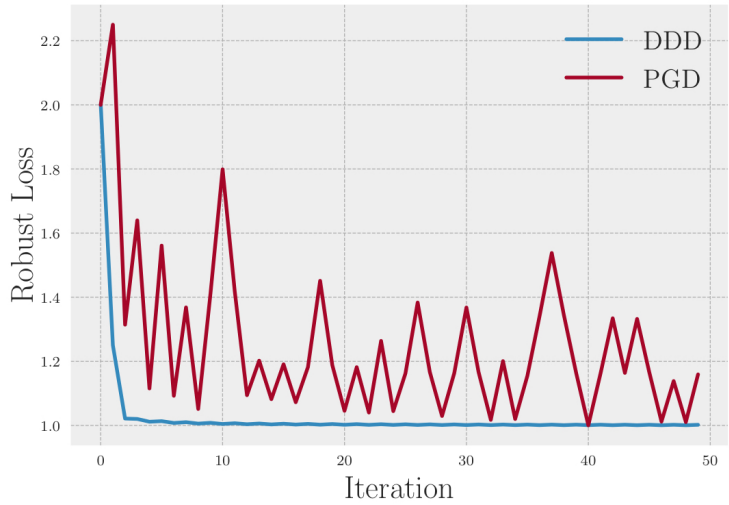
- A paper by Madry et al. [2018] is one of the most famous papers in the literature of Adversarial Training.
- Main Idea: Danskin's Theorem [Danskin, 1966, Dem'yanov, 1966, Seeger, 1988] to obtain a *descent direction* of adversarial loss.

- Latorre et al. [2023]: "Madry et al. [2018] incorrectly interpreted Danskin's theorem."
- **Danskin's Descent Direction (DDD)**: better descent directions than those obtained by Madry et al. [2018].

Synthetic example



(a)



(b)

- PGD-based method: Madry et al. [2018]
- **Danskin's Descent Direction (DDD)**: Latorre et al. [2023]

Table of Contents

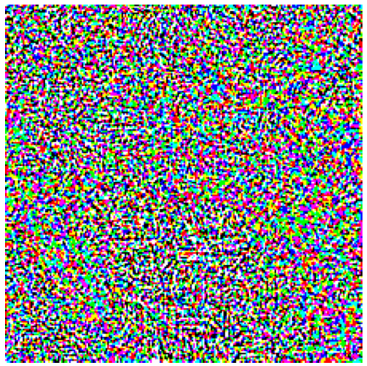
Adversarial attack

- "How to fool a model with **a small perturbation on an input**?"
- Fast Gradient Sign Method (FGSM) [Goodfellow et al., 2015]
  - $\text{sign}(\boldsymbol{v}) := [\text{sign}(v_1), \text{sign}(v_2), \text{sign}(v_3), ...]$

$\boldsymbol{x}$

"panda"
57.7% confidence

$+ .007 \times$



$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

$=$

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon" : 긴팔원숭이
99.3 % confidence

Adversarial attack

- Projected Gradient Descent (PGD) on the negative loss [Kurakin et al., 2017]
  - The multi-step variant of FGSM
  - Given a natural example x and a randomly initialized perturbation $\delta \in \mathcal{S}_0 = \{\delta : \|\delta\|_\infty = \max\{\delta_1, \delta_2, ...\} \leq \epsilon\}$,

$$x^0 := x + \delta,$$
$$x^{t+1} := \mathrm{Proj}_{x+\mathcal{S}_0}\left(x^t + \alpha\, \mathrm{sign}\left(\nabla_x \mathcal{L}(\theta, x^t, y)\right)\right)$$

Adversarial training (a.k.a. Defense)

- recall: $\mathcal{S}_0 = \{\delta : \|\delta\|_\infty \leq \epsilon\}$
- Optimization framework for Adversarial Training [Madry et al., 2018]

$$\min_\theta \rho(\theta) := \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta\in\mathcal{S}_0} \mathcal{L}(\theta, x+\delta, y) \right]$$

  - Note: optimal δ for the inner max problem clearly depends on (θ, x, y).
  - Too difficult & intractable.

- A weaker notion of robust loss for a batch $\{(x_i, y_i)\}_{i=1}^k$:

$$\min_\theta \phi(\theta) := \max_{\delta=[\delta_1,\ldots,\delta_k]\in\mathcal{S}:=\mathcal{S}_0^k} \left[g(\theta,\delta) := \frac{1}{k}\sum_{i=1}^k \mathcal{L}(\theta, x_i+\delta_i, y_i) \right]$$

  - Here, δ is a matrix with columns in $\mathcal{S}_0$
  - PGD adversary: good to approximate $\phi(\theta) = \max_{\delta\in\mathcal{S}} g(\theta,\delta)$
  - How to decrease $\phi(\theta)$? Danskin's Theorem might be useful.

Danskin's Theorem

- The theorem gives an explicit form of **directional derivative** of $\phi(\theta)$ (denoted as $D_\gamma \phi(\theta)$, for a (unit) vector γ) in general.

Theorem 1 ($\equiv$ Thm 1 by Danskin [1966])

- Let $\mathcal{S}$ be a **compact** topological space,
- $g : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}$ be a continuous function such that
  - $g(\cdot, \delta)$ is differentiable ($\forall \delta \in \mathcal{S}$) and $\nabla_\theta g(\theta, \delta)$ is continuous on $\mathbb{R}^d \times \mathcal{S}$,
- $\phi(\theta) = \max_{\delta \in \mathcal{S}} g(\theta, \delta), \quad \mathcal{S}^\star(\theta) := \arg\max_{\delta \in \mathcal{S}} g(\theta, \delta),$

Then ϕ is directionally differentiable: in the direction γ,

$$D_\gamma \phi(\theta) = \max_{\delta \in \mathcal{S}^\star(\theta)} \langle \gamma, \nabla_\theta g(\theta, \delta) \rangle .$$

In particular, if $\mathcal{S}^\star(\theta) = \{\delta_\theta^\star\}$ is a singleton set, ϕ is differentiable at θ:

$$\nabla \phi(\theta) = \nabla_\theta g(\theta, \delta_\theta^\star).$$

Corollary(?) of Danskin's Theorem

- γ is called a **descent direction** of ϕ at θ iff $D_\gamma \phi(\theta) < 0$.

Corollary(?) 1 [Madry et al., 2018]

Choose a $\delta^\star \in \mathcal{S}^\star(\theta)$. If θ is not a local minimizer of $g(\cdot, \delta^\star)$, under the same assumptions in Danskin's Theorem, $-\nabla_\theta g(\theta, \delta^\star)$ is a descent direction for ϕ at θ.

- True (and can be easily deduced) when $\mathcal{S}^\star(\theta)$ is a singleton set.
- Base on this, Madry et al. [2018] propose a framework of Adversarial Training (AT):
  1. Given a parameter θ and a batch $\{(x_i, y_i)\}_{i=1}^k$, find a set of near-optimal adversarial examples $\{(x_i + \delta_i)\}_{i=1}^k$ with PGD
  2. Update θ along a *stochastic descent direction*:

$$\theta \leftarrow \theta - \beta \cdot \frac{1}{k} \sum_{i=1}^{k} \nabla_\theta \mathcal{L}(\theta, x_i + \delta_i, y_i)$$

Table of Contents

Counterexample 1 (well-known)

Corollary(?) 1 [Madry et al., 2018] in short

Choose a $\delta^\star \in \mathcal{S}^\star(\theta)$. If $-\nabla_\theta g(\theta, \delta^\star) \neq 0$, it is a descent direction for ϕ at θ.

- Let $\mathcal{S} := [-1, 1]$ and $g(\theta, \delta) = \theta\delta$ $(\theta \in \mathbb{R}, \delta \in \mathcal{S})$. Then

$$\phi(\theta) = \max_{\delta \in [-1,1]} \theta\delta = |\theta|.$$

  Note that, at $\theta = 0$, $\mathcal{S}^\star(0) = \{-1, 1\}$.
- Choosing $\delta = 1 \in \mathcal{S}^\star(0)$, we have $g(\theta, 1) = \theta$ and thus

$$-\nabla_\theta g(0, 1) = -1 \neq 0.$$

- Likewise, choosing $\delta = -1 \in \mathcal{S}^\star(0)$, we have $g(\theta, -1) = -\theta$ and thus

$$-\nabla_\theta g(0, -1) = 1 \neq 0.$$

- However, $\theta = 0$ is a global minimizer of $\phi(\theta) = |\theta|$, which means there exists no descent direction. $(\Rightarrow\!\!\!\!\!\!\not\leftarrow)$

Counterexample 2 (non-local-optimum)

Corollary(?) 1 [Madry et al., 2018] in short

Choose a $\delta^\star \in \mathcal{S}^\star(\theta)$. If $-\nabla_\theta g(\theta, \delta^\star) \neq 0$, it is a descent direction for ϕ at θ.

- Let $\mathcal{S} := [0, 1]$; let u, v be unit vectors s.t. $-1 < \langle u, v \rangle < 0$ (obtuse); let $g(\theta, \delta) = \delta\langle \theta, u \rangle + (1 - \delta)\langle \theta, v \rangle + \delta(\delta - 1)$. $(\theta \in \mathbb{R}^2, \delta \in \mathcal{S})$
- Claim 1: at $\theta = 0$, $-\nabla_\theta g(0, \delta^\star)$ are *ascent* directions $\forall \delta^\star \in \mathcal{S}^\star(0)$.
  - $\mathcal{S}^\star(0) = \arg\max_{\delta \in [0,1]} \delta(\delta - 1) = \{0, 1\}$.
  - $\nabla_\theta g(0, 0) = v$ and $\nabla_\theta g(0, 1) = u$.
  - However, applying Danskin's Theorem, we can have $D_{-v}\phi(0) = D_{-u}\phi(0) = -\langle u, v \rangle > 0$, which disproves Corollary 1.
- Claim 2: However, $\theta = 0$ is not a local minima of $\phi(\cdot)$ (unlike c.e.1).
  - If $\gamma = -(u + v)$, applying Danskin's Theorem,

$$D_\gamma \phi(0) = -1 - \langle u, v \rangle < 0 : \quad \text{descent direction.}$$

- Lesson: a 'single' optimal δ is not enough.

A Mistake in the proof

- Directional derivative $D_\gamma \phi(\theta)$ is defined as the one-sided limit.
  - In general, $D_\gamma \phi(\theta)$ and $D_{-\gamma} \phi(\theta)$ are independent items.
- However, Madry et al. [2018] mistakenly assumes the two-sided limit.
  - They correctly showed the fact that if $\delta^\star \in \mathcal{S}^\star(\theta)$ then $\gamma = \nabla_\theta g(\theta, \delta^\star) \neq 0$ satisfies $D_\gamma \phi(\theta) > 0$:

$$D_\gamma \phi(\theta) = \max_{\delta \in \mathcal{S}^\star(\theta)} \langle \gamma, \nabla_\theta g(\theta, \delta) \rangle \geq \|\nabla_\theta g(\theta, \delta^\star)\|^2 > 0.$$

  - This fact does not implies $D_{-\gamma} g(\theta, \delta^\star) < 0$.

Table of Contents

Obtaining a better descent direction

- Adversarial Training relying on a "single" optimal adversary does not always lead to a descent direction
- Solution: use multiple adversarial perturbations per data sample.

- Goal: to obtain steepest descent direction $\gamma^\star$ for the robust loss on a batch $\{(x_i, y_i)\}_{i=1}^k$:

$$\gamma^\star \in \underset{\gamma:\|\gamma\|_2=1}{\arg\min} \max_{\delta \in \mathcal{S}^\star(\theta)} \langle \gamma, \nabla_\theta g(\theta, \delta) \rangle, \quad g(\theta, \delta) = \frac{1}{k} \sum_{i=1}^{k} \mathcal{L}(\theta, x_i + \delta_i, y_i)$$

Obtaining a better descent direction

- Assume: $\mathcal{S}^\star(\theta) = \mathcal{S}^\star_m(\theta) := \{\delta^{(1)}, \ldots, \delta^{(m)}\}$ $(m < \infty)$

Theorem 2 [Latorre et al., 2023]

Denote by $\nabla_\theta g(\theta, \mathcal{S}^\star_m(\theta))$ the matrix with columns $\nabla_\theta g(\theta, \delta^{(i)})$ $(i \in [m])$. As long as θ is not a local minimizer of the robust loss $\phi(\theta)$, the steepest descent direction of ϕ at θ can be computed as

$$\gamma^\star := -\frac{\nabla_\theta g(\theta, \mathcal{S}^\star_m(\theta))\alpha^\star}{\|\nabla_\theta g(\theta, \mathcal{S}^\star_m(\theta))\alpha^\star\|_2}, \quad \alpha^\star \in \underset{\alpha \in \Delta_m}{\arg\min} \|\nabla_\theta g(\theta, \mathcal{S}^\star_m(\theta))\alpha\|_2^2,$$

where $\Delta_m := \{\alpha \in \mathbb{R}^m : \sum_{i=1}^m \alpha_i = 1, \alpha \geq 0\}$.

- In practice, the finiteness assumption might not hold.
  - If $\mathcal{S}^\star_m(\theta)$ can ϵ-approximates $\mathcal{S}^\star(\theta)$ and $\nabla_\theta g(\theta, \delta)$ is Lipschitz conti. in δ, then $\gamma^\star$ is still the best choice of direction to move (Theorem 3).

Danskin's Descent Direction

- Based on Theorem 2&3, the authors propose an algorithm:

Algorithm: Danskin's Descent Direction [Latorre et al., 2023]

for $t = 0, \ldots, T-1$ do:

- Draw k natural examples: $(x_1, y_1), \ldots, (x_k, y_k) \sim \mathcal{D}$
- Obtain m near-optimal perturbation matrices $\delta^{(1)}, \cdots, \delta^{(m)} \in \mathcal{S}_0^k$
- Obtain steepest descent direction $\gamma^\star$ (Based on Theorem 2)
- Update $\theta_{t+1} \leftarrow \theta_t + \beta_t \gamma^\star$

Danskin's Descent Direction

- In more detail:

for $t = 0, \ldots, T-1$ do:

- Draw k natural examples: $(x_1, y_1), \ldots, (x_k, y_k) \sim \mathcal{D}$
- Obtain m different perturbation matrices (w/ accelerated proximal PGD):

$$\delta^{(1)}, \ldots, \delta^{(m)} \leftarrow \underset{\delta \in \mathcal{S}_0^k}{\text{MAXIMIZE}}\ g(\theta, \delta) = \frac{1}{k} \sum_{i=1}^{k} \mathcal{L}(\theta, x_i + \delta_i, y_i)$$

- $M \leftarrow [\nabla_\theta g(\theta, \delta^{(i)})]_{i=1}^m \in \mathbb{R}^{d \times m}$
- Obtain $\alpha^\star$ (w/ simplex projection algorithm [Duchi et al., 2008]):

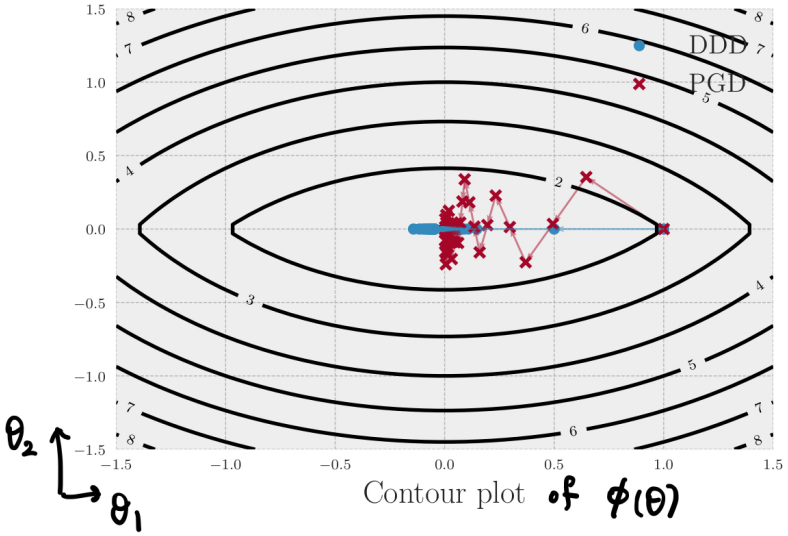
$$\alpha^\star \leftarrow \underset{\alpha \in \Delta_m}{\text{MINIMIZE}}\ \|M\alpha\|_2^2$$

- Steepest descent direction $\gamma^\star \leftarrow -\frac{M\alpha^\star}{\|M\alpha^\star\|_2}$
- Update $\theta_{t+1} \leftarrow \theta_t + \beta_t \gamma^\star$

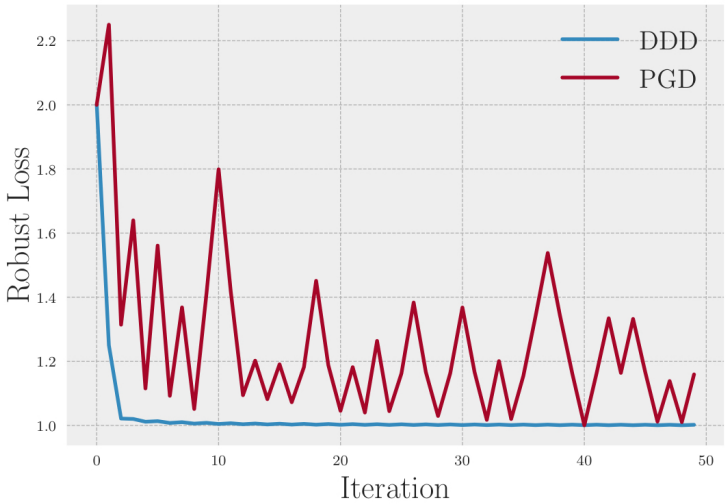
Table of Contents

Synthetic example

- $g(\theta, \delta) = \delta(\theta_1^2 + (\theta_2 + 1)^2) + (1 - \delta)(\theta_1^2 + (\theta_2 - 1)^2)$ where $\delta \in [0, 1]$.



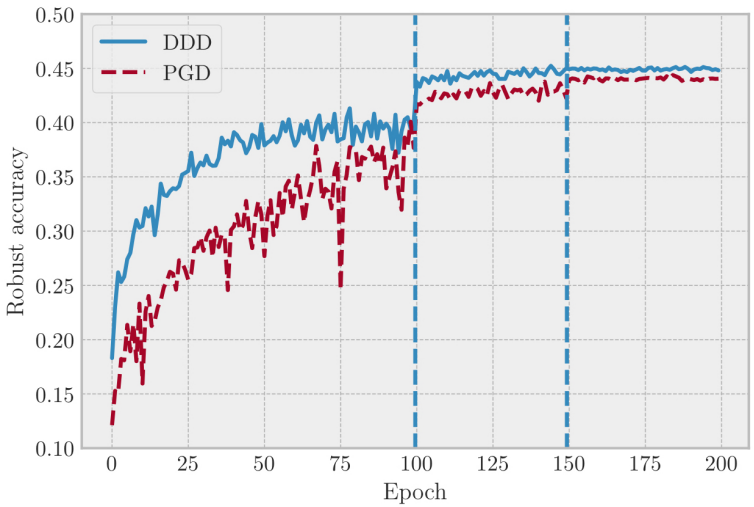
(a)



(b)

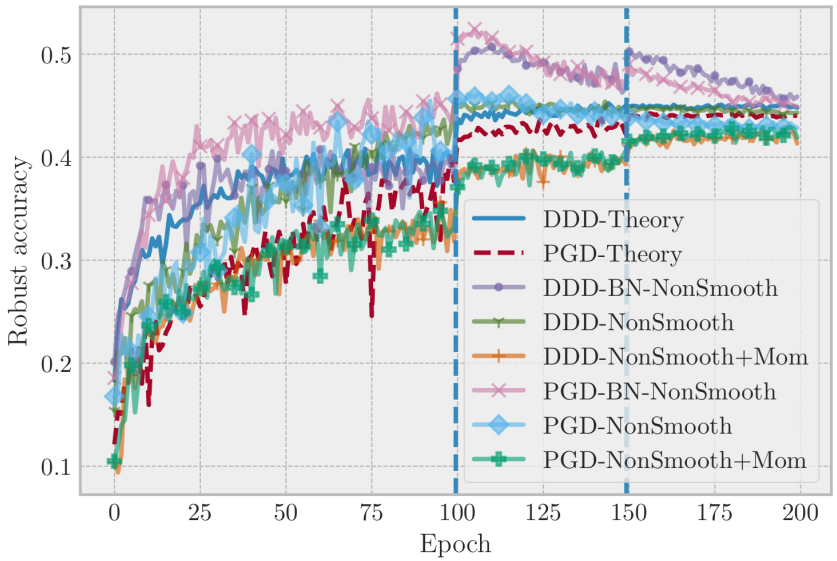
Experiments on CIFAR10

- Mainly compare with ℓ_∞-PGD ($\epsilon = 8/255 \approx 0.03$, $\alpha = 2/255$, $n_{\text{inner}} = 7$)
  - ResNet18 with SGD, using the setting from Pang et al. [2021] ("Bag of Tricks for Adversarial Training") with some modifications
- To meet theoretical assumptions:
  - **ReLU** $\rightarrow$ **CELU**: to ensure (continuously) differentiability
  - **BN** $\rightarrow$ **GroupNorm**: to remove intra-batch dependencies
  - **momentum 0.9** $\rightarrow$ **0.0** : to remove dependency on previous descent directions
  - $m = 10$ perturbations per sample



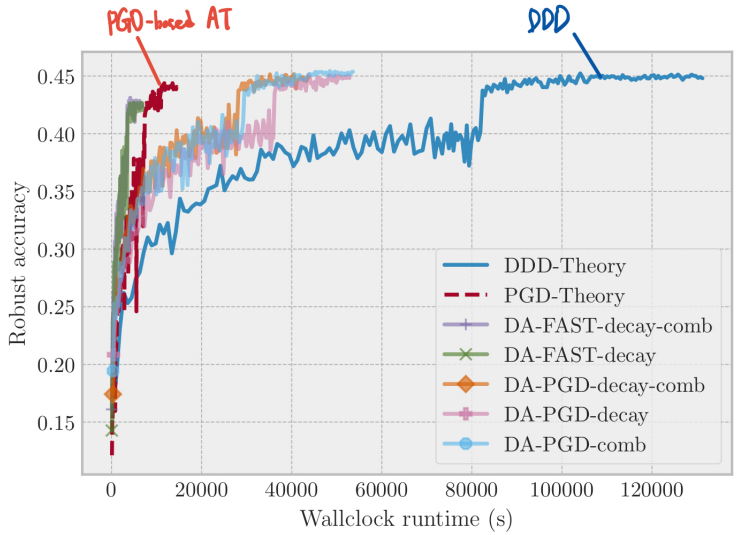
- robust accuracy: accuracy on validation set + PGD-7 adversary

Experiments on CIFAR10 - ablation study



- Momentum hurts robust accuracy
- With ReLU + BN + Early stopping, PGD-based AT seems better than DDD
  - "Outside the scope of existing theory but might achieve better performance."

Remark on wall-clock runtime



- Naive DDD has $\sim 10-12\times$ overhead than PGD :
  1. Generating m adversarial examples - m-times overhead
  2. m separate gradient samples - m forward-backward passes
  3. $\star$ Additional inner optimization problem (simplex projection algorithm)

- Authors also provide two heuristic approaches to speed up ($12\times \rightarrow 3\times$ overhead) while maintaining the benefits of DDD:
  - `decay` (decay m linearly 10 to 1)
  - `comb` (efficient combinatorial batch construction method; Appendix B.4.1)

References I

J. M. Danskin. The theory of max-min, with applications. *SIAM Journal on Applied Mathematics*, 14(4):641–664, 1966.

V. Dem'yanov. On the solution of several minimax problems. i. *Cybernetics*, 2(6):47–53, 1966.

J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 272–279, 2008.

I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. URL `http://arxiv.org/abs/1412.6572`.

A. Kurakin, I. J. Goodfellow, and S. Bengio. Adversarial machine learning at scale. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=BJm4T4Kgx`.

F. Latorre, I. Krawczuk, L. T. Dadi, T. Pethick, and V. Cevher. Adversarial training descends without descent: Finding actual descent directions based on danskin's theorem. In *Submitted to The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=I3HCE7Ro78H`. under review.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=rJzIBfZAb`.

T. Pang, X. Yang, Y. Dong, H. Su, and J. Zhu. Bag of tricks for adversarial training. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=Xb8xvrtB8Ce`.

A. Seeger. Second order directional derivatives in parametric optimization problems. *Mathematics of Operations Research*, 13 (1):124–139, 1988.