制約付き非凸最適化問題のための確率的Frank-Wolfe法とその敵対的攻撃への応用

日本オペレーションズ・リサーチ学会 2025年春季研究発表会
3/6(木) 11:00～11:20

明治大学　佐藤尚樹
明治大学　飯塚秀明

背景：勾配法

$$\boldsymbol{\theta}_{t+1} := \boldsymbol{\theta}_t + \gamma_t \boldsymbol{d}_t$$

学習率　探索方向

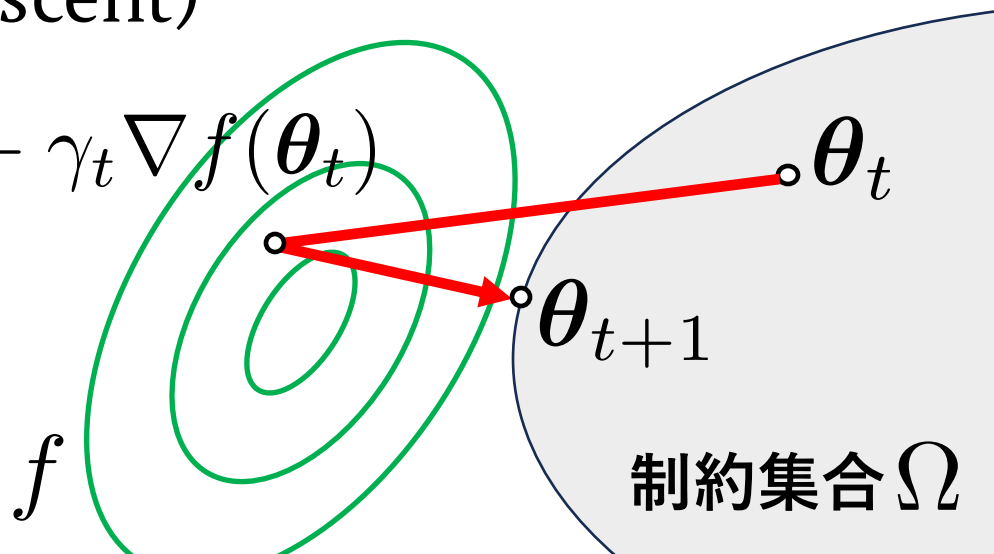
▷**最急降下法** (Gradient Descent)

$$\boldsymbol{d}_t := -\nabla f(\boldsymbol{\theta}_t)$$

全勾配

▷**射影勾配降下法** (Projected Gradient Descent)

$$\therefore \boldsymbol{\theta}_{t+1} := \Pi_{\Omega}\left(\boldsymbol{\theta}_t - \gamma_t \nabla f(\boldsymbol{\theta}_t)\right)$$



背景：勾配法

$$\boldsymbol{\theta}_{t+1} := \boldsymbol{\theta}_t + \gamma_t \boldsymbol{d}_t$$

学習率　探索方向

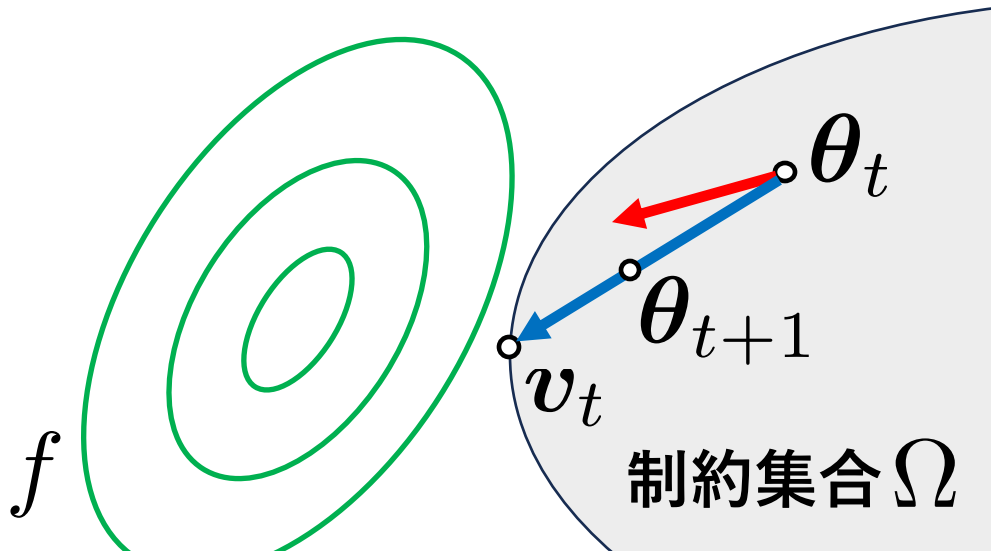
▷**Frank-Wolfe法**

$$\boldsymbol{v}_t := \underset{\boldsymbol{v}\in\Omega}{\operatorname{argmax}} \langle \boldsymbol{v}, -\nabla f(\boldsymbol{\theta}_t) \rangle$$

最急降下方向

$$\boldsymbol{d}_t := \boldsymbol{v}_t - \boldsymbol{\theta}_t$$

$$\therefore \boldsymbol{\theta}_{t+1} := \gamma_t \boldsymbol{v}_t + (1-\gamma_t)\boldsymbol{\theta}_t$$



背景：Stochastic Frank Wolfe (SFW)

Algorithm 1 Frank Wolfe

Require: $(\gamma_t)_{t\in\mathbb{N}} \subset \mathbb{R}_{++}$

$t \leftarrow 0, \boldsymbol{\theta}_0 \in \Omega$

loop

$\boldsymbol{v}_t = \underset{\boldsymbol{v}\in\Omega}{\operatorname{argmax}} \langle \boldsymbol{v}, -\nabla f(\boldsymbol{\theta}_t) \rangle$

全勾配

$\boldsymbol{d}_t = \boldsymbol{v}_t - \boldsymbol{\theta}_t$

$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \gamma_t \boldsymbol{d}_t$

$t \leftarrow t + 1$

end loop

Algorithm 2 Stochastic Frank Wolfe (SFW)

Require: $(\gamma_t)_{t\in\mathbb{N}} \subset \mathbb{R}_{++}$

$t \leftarrow 0, \boldsymbol{\theta}_0 \in \Omega$

loop

$\boldsymbol{v}_t = \underset{\boldsymbol{v}\in\Omega}{\operatorname{argmax}} \langle \boldsymbol{v}, -\nabla f_{\mathcal{S}_t}(\boldsymbol{\theta}_t) \rangle$

ミニバッチ確率的勾配

$\boldsymbol{d}_t = \boldsymbol{v}_t - \boldsymbol{\theta}_t$

$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \gamma_t \boldsymbol{d}_t$

$t \leftarrow t + 1$

end loop

背景：SFWの活躍(応用)

▷深層学習におけるSFWの活躍の例

Pruning Ratios	50%	70%	80%	90%	95%
SFW-Pruning + SFW-Init (ours)	93.10	93.10	93.10	93.10	92.00
One-Cycle Pruning (Hubens et al., 2021)	-	-	90.87	90.72	90.67
Early Bird (You et al., 2020)	93.21	92.80	-	-	-
OTO (Chen et al., 2021)	90.35	90.35	90.35	90.35	90.35
DPF (Lin et al., 2020)	-	-	-	-	93.87
Group MDP (Deleu & Bengio, 2021)	-	-	-	89.38	-

Table 1: Comparisons to more state-of-the-art methods on VGG-16 and CIFAR-10.

▷モデルの枝刈りタスクにおいて，SGDを利用した手法よりも優れた性能をもたらすことが示された．しかも計算量も少ない．

[Miao+22] L. Miao, X. Luo, T. Chen, W. Chen, D. Liu, and Z. Wang, “Learning pruning-friendly networks via Frank-Wolfe: One-shot, any-sparsity, and no retraining”, In Proceedings of the 10th International Conference on Learning Representations, 2022.

背景：SFWの先行研究(理論)

Theorem 2. *Consider the stochastic setting of* (1) *where* f *is* G*-Lipschitz and* F *is* L*-smooth. Then, the output* x_a *of Algorithm* 2 *with parameters* $\gamma_t = \gamma = \sqrt{\frac{2(F(x_0)-F(x^*))}{TLD^2\beta}}$, $b_t = b = T$ *for all* $t \in \{0, \ldots, T-1\}$, *satisfies the following bound:*

学習率　　バッチサイズ

$$\mathbb{E}[\mathcal{G}(x_a)] \leq \frac{D}{\sqrt{T}}\left(G + \sqrt{\frac{2L(F(x_0)-F(x^*))}{\beta}}(1+\beta)\right),$$

where x^* *is an optimal solution to (stochastic) problem* (1).

[Reddi+16] S.J. Reddi, S. Sra, B. Poczos, and A. Smola, “Stochastic Frank-Wolfe Methods for Nonconvex Optimization” in *54th Annual Allerton Conference on Communication, Control, and Computing*, pp.1214-1251, 2016.

▷問題点その1. 学習率とバッチサイズの設定が現実的でない.

背景：SFWの先行研究(理論)

目的関数のL-平滑性から，

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{gap} \leq \frac{A}{T\gamma} + B\gamma$$

が成り立つ．ただし，AとBは正の定数とする．
このとき $\gamma = \frac{1}{\sqrt{T}}$ **を使えば，**

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{gap} \leq \frac{A}{\sqrt{T}} + \frac{B}{\sqrt{T}} = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$$

が成り立ち，あたかもgapが0に収束するかのように見える．

▷問題点その2. 実は収束は保証されていない．

背景：SFWの先行研究(理論)

Algorithm	Learning Rate	Gradient (batch size)	Momentum
SFW (Reddi et al., 2016)	$\gamma_t = \mathcal{O}\left(\sqrt{\frac{1}{TL}}\right)$	mini-batch ($b = T$)	No
AsySFW (Gu et al., 2019)	$\gamma_t = \mathcal{O}\left(\sqrt{\frac{1}{TL}}\right)$	mini-batch ($b = T$)	No
FW-SD (Grigas et al., 2019)	$\gamma_t = \mathcal{O}\left(\sqrt{\frac{1}{L}}\right)$	mini-batch ($b_t \leq n$)	No
SFW (Négiar et al., 2020)	$\gamma_t = \frac{2}{t+2}$	mini-batch ($b \leq n$)	No
SCG (Mokhtari et al., 2020)	$\gamma_t = \frac{1}{T}$	Noise	Yes (Increasing)
AdaSFW (Combettes et al., 2021)	$\gamma_t = \frac{2}{t+2}$	mini-batch $\left(b_t = \mathcal{O}\left(\frac{t^2}{L^2}\right)\right)$	No
AdaCSFW (Combettes et al., 2021)	$\gamma_t = \frac{2}{t+2}$	mini-batch ($b \leq n$)	No
any SFW methods (Nazykov et al., 2024)	$\gamma_t = \sqrt{\frac{1}{T}}$	Noise	No

動機

- **現実的な学習率とバッチサイズの設定の下での収束解析を提供したい.**

貢献

▷実用的な定数/減少学習率，定数/増加バッチサイズを有するSFWの収束解析を提供した．

	定数バッチサイズ $b_t := b$	増加バッチサイズ
定数学習率 $\gamma_t := \gamma$	$\mathcal{O}\left(\frac{1}{T} + \frac{1}{\sqrt{b}} + \gamma\right)$ (Theorem 3.1)	$\mathcal{O}\left(\frac{1}{T} + \gamma\right)$ (Theorem 3.2)
減少 (I) $\gamma_t := \frac{1}{t+1}$	$\mathcal{O}\left(\frac{1}{\sqrt{b}}\right)$ (Theorem 3.3)	-
減少 (II) $\gamma_t := \frac{1}{(t+1)^a} (a \in [0.5, 1))$	$\mathcal{O}\left(\frac{1}{T^{\min\{1-a,a\}}} + \frac{1}{\sqrt{b}}\right)$ (Theorerm 3.4(i))	$\mathcal{O}\left(\frac{1}{T^{\min\{1-a,a\}}}\right)$ (Theorem 3.5(i))
減少 (III)	$\mathcal{O}\left(\frac{1}{T} + \frac{1}{\sqrt{b}}\right)$ (Theorem 3.4(ii))	$\mathcal{O}\left(\frac{1}{T}\right)$ (Theorem 3.5(ii))

$$\gamma_t := (\underbrace{\gamma, \gamma, \cdots, \gamma}_{K}, \underbrace{\eta\gamma, \eta\gamma, \cdots, \eta\gamma}_{K}, \cdots, \underbrace{\eta^{P-1}\gamma, \eta^{P-1}\gamma, \cdots, \eta^{P-1}\gamma}_{K}), \gamma > 0, \eta \in (0,1), KP = T$$

$$b_t := (\underbrace{b, b, \cdots, b}_{E}, \underbrace{\lambda b, \lambda b, \cdots, \lambda b}_{E}, \cdots, \underbrace{\lambda^{Q-1} b, \lambda^{Q-1} b, \cdots, \lambda^{Q-1} b}_{E}), b > 0, \lambda > 1, EQ = T$$

準備：最適化問題

経験損失最小化問題

▷訓練データセット $S := (z_1, z_2, \cdots, z_n)$

▷Deep Neural Network のパラメータ $x \in \mathbb{R}^d$

$$\min_{\theta \in \Omega} \left\{ f(\theta) := \frac{1}{n} \sum_{i=0}^{n-1} f_i(\theta) \right\}$$

i 番目の訓練データ z_i に対する損失関数

▷非凸目的関数 $f : \mathbb{R}^d \to \mathbb{R}$ を最適化する。

準備：仮定

仮定

(A1)制約集合 $\Omega \subset \mathbb{R}^d$ は凸集合で，直径 $D \in \mathbb{R}$ を有する，i.e.,

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \Omega \colon \|\boldsymbol{x} - \boldsymbol{y}\| \leq D$$

(A2) $(\boldsymbol{\theta}_t)_{t \in \mathbb{N}} \subset \mathbb{R}^d$ を最適化手法によって生成された点列とするとき，

(i)任意の $t \in \mathbb{N}$ に対して，次の式が成り立つとする．

$$\mathbb{E}_{\xi_t}\left[\mathsf{G}_{\xi_t}(\boldsymbol{\theta}_t)\right] = \nabla f(\boldsymbol{\theta}_t)$$

(ii)次の式を満たす非負定数 σ^2 が存在するとする．

$$\mathbb{E}_{\xi_t}\left[\|\mathsf{G}_{\xi_t}(\boldsymbol{\theta}_t) - \nabla f(\boldsymbol{\theta}_t)\|^2\right] \leq \sigma^2$$

準備：仮定

仮定続き

(A3) 関数$f\colon \mathbb{R}^d \to \mathbb{R}$は制約集合$\Omega$上で連続的微分可能で，
L-平滑とする，i.e.,

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \Omega\colon \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq \|\boldsymbol{x} - \boldsymbol{y}\|$$

(A4) 時刻$t \in \mathbb{N}$で，全勾配∇fはミニバッチ$\mathcal{S}_t$で次のように近似されるとする．

$$\nabla f_{\mathcal{S}_t}(\boldsymbol{\theta}_t) := \frac{1}{b}\sum_{i\in[b]} \mathsf{G}_{\xi_{t,i}}(\boldsymbol{\theta}_t) = \frac{1}{b}\sum_{\{i\colon \boldsymbol{z}_i \in \mathcal{S}_t\}} \nabla l_i(\boldsymbol{\theta}_t)$$

ミニバッチ
確率的勾配

確率的勾配

準備：Frank-Wolfe gap

▷制約付き最適化においては，$\|\nabla f(\boldsymbol{\theta}_t)\|$を収束の指標として使うことはできない．

▷Frank-Wolfe法の解析では，収束の指標として伝統的に **Frank-Wolfe gap** $\mathcal{G}(\boldsymbol{\theta}_t)$が利用されてきた．

$$\mathcal{G}(\boldsymbol{\theta}_t) := \max_{\boldsymbol{v}\in\Omega} \langle \boldsymbol{v} - \boldsymbol{\theta}, -\nabla f(\boldsymbol{\theta}) \rangle$$

▷$\boldsymbol{\theta}^\star \in \mathbb{R}^d$を$\Omega$上での局所的最適解であるとすると，

$$\forall \boldsymbol{v} \in \Omega\colon \langle \boldsymbol{v} - \boldsymbol{\theta}^\star, \nabla f(\boldsymbol{\theta}^\star) \rangle \geq 0 \quad \text{i.e.,} \quad \mathcal{G}(\boldsymbol{\theta}^\star) \leq 0$$

が成り立つことから，gapを小さくすることを目指す．

収束解析ー定数バッチサイズ $b_t := b$

▷**定数学習率** $\gamma_t := \gamma$

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\mathcal{G}(\boldsymbol{\theta}_t)\right] \leq \frac{\mathbb{E}\left[f(\boldsymbol{\theta}_0)\right] - \mathbb{E}\left[f(\boldsymbol{\theta}^\star)\right]}{\gamma T} + \frac{D\sigma}{\sqrt{b}} + \frac{LD^2\gamma}{2} = \mathcal{O}\left(\frac{1}{T} + \frac{1}{\sqrt{b}} + \gamma\right)$$

▷**減少学習率** $\gamma_t := \dfrac{1}{(t+1)^a}(a \in [0.5, 1))$

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\mathcal{G}(\boldsymbol{\theta}_t)\right] \leq \frac{2\max\{\bar{f}, |f(\boldsymbol{\theta}^\star)|\}}{T^{1-a}} + \frac{D\sigma}{\sqrt{b}} + \frac{LD^2}{2(1-a)T^a} = \mathcal{O}\left(\frac{1}{T^{\min\{1-a,a\}}} + \frac{1}{\sqrt{b}}\right)$$

▷**減少学習率** $\gamma_t := (\underbrace{\gamma, \gamma, \cdots, \gamma}_{K}, \underbrace{\eta\gamma, \eta\gamma, \cdots, \eta\gamma}_{K}, \cdots, \underbrace{\eta^{P-1}\gamma, \eta^{P-1}\gamma, \cdots, \eta^{P-1}\gamma}_{K}), \gamma > 0, \eta \in (0,1), KP = T$

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\mathcal{G}(\boldsymbol{\theta}_t)\right] \leq \frac{2\max\{\bar{f}, |f(\boldsymbol{\theta}^\star)|\}}{T\underline{\gamma}} + \frac{D\sigma}{\sqrt{b}} + \frac{\gamma_\infty LD^2}{2T} = \mathcal{O}\left(\frac{1}{T} + \frac{1}{\sqrt{b}}\right)$$

収束解析ー増加バッチサイズ

$$b_t := (\underbrace{b, b, \cdots, b}_{E}, \underbrace{\lambda b, \lambda b, \cdots, \lambda b}_{E}, \cdots, \underbrace{\lambda^{Q-1} b, \lambda^{Q-1} b, \cdots, \lambda^{Q-1} b}_{E}), b > 0, \lambda > 1, EQ = T$$

▷**定数学習率** $\gamma_t := \gamma$

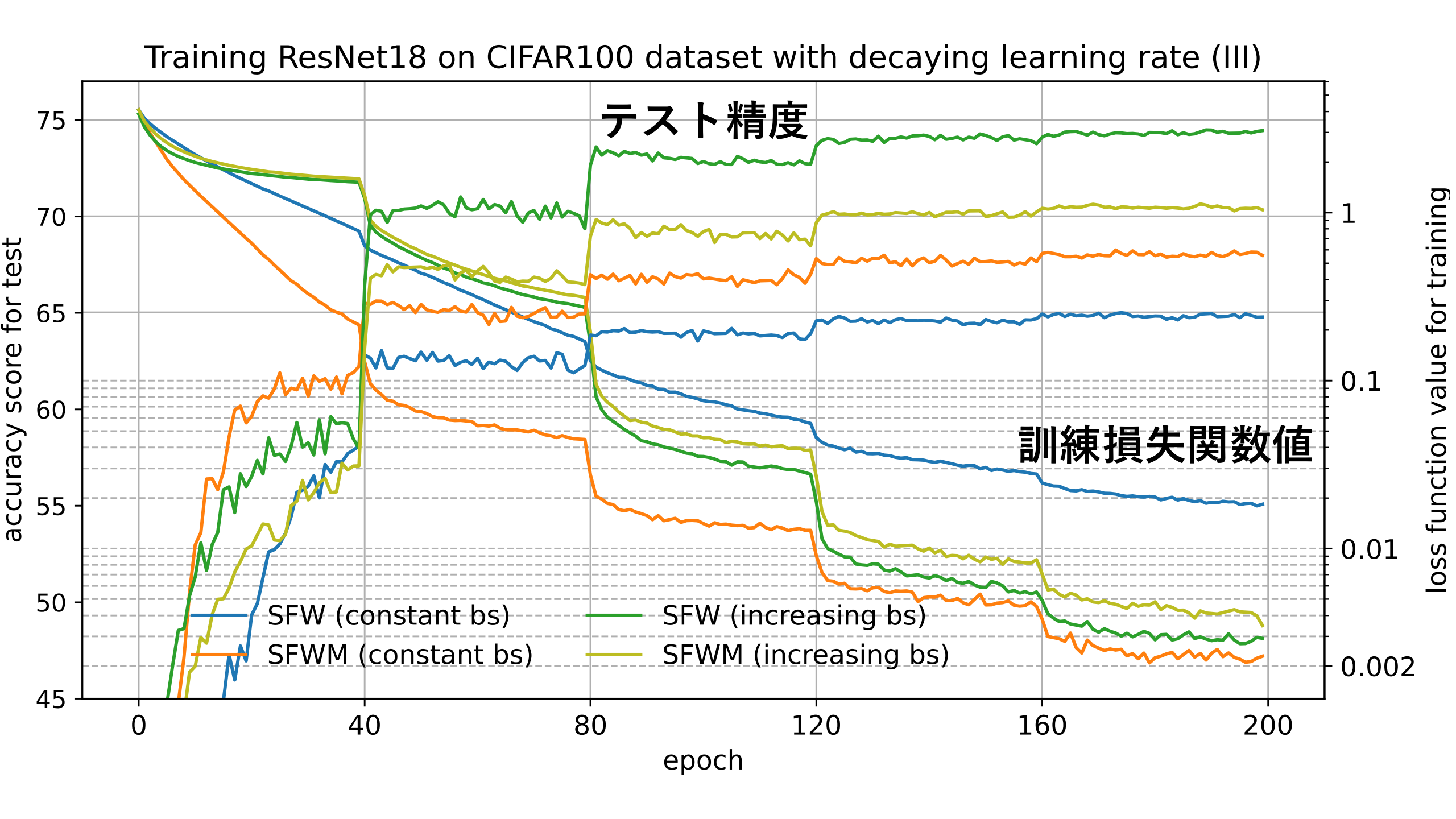
$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\mathcal{G}(\boldsymbol{\theta}_t)\right] \leq \frac{\mathbb{E}\left[f(\boldsymbol{\theta}_0)\right] - \mathbb{E}\left[f(\boldsymbol{\theta}^\star)\right]}{\gamma T} + \frac{D\sigma}{T}\frac{E\sqrt{\lambda}}{\sqrt{b}(\sqrt{\lambda}-1)} + \frac{LD^2\gamma}{2} = \mathcal{O}\left(\frac{1}{T} + \gamma\right)$$

▷**減少学習率** $\gamma_t := \dfrac{1}{(t+1)^a} (a \in [0.5, 1))$

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\mathcal{G}(\boldsymbol{\theta}_t)\right] \leq \frac{2\max\{\bar{f}, |f(\boldsymbol{\theta}^\star)|\}}{T^{1-a}} + \frac{D\sigma}{T}\frac{E\sqrt{\lambda}}{\sqrt{b}(\sqrt{\lambda}-1)} + \frac{LD^2}{2(1-a)T^a} = \mathcal{O}\left(\frac{1}{T^{\min\{1-a,a\}}}\right)$$

▷**減少学習率** $\gamma_t := (\underbrace{\gamma, \gamma, \cdots, \gamma}_{K}, \underbrace{\eta\gamma, \eta\gamma, \cdots, \eta\gamma}_{K}, \cdots, \underbrace{\eta^{P-1}\gamma, \eta^{P-1}\gamma, \cdots, \eta^{P-1}\gamma}_{K}), \gamma > 0, \eta \in (0,1), KP = T$

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\mathcal{G}(\boldsymbol{\theta}_t)\right] \leq \frac{2\max\{\bar{f}, |f(\boldsymbol{\theta}^\star)|\}}{T\underline{\gamma}} + \frac{D\sigma}{T}\frac{E\sqrt{\lambda}}{\sqrt{b}(\sqrt{\lambda}-1)} + \frac{\gamma_\infty LD^2}{2T} = \mathcal{O}\left(\frac{1}{T}\right)$$

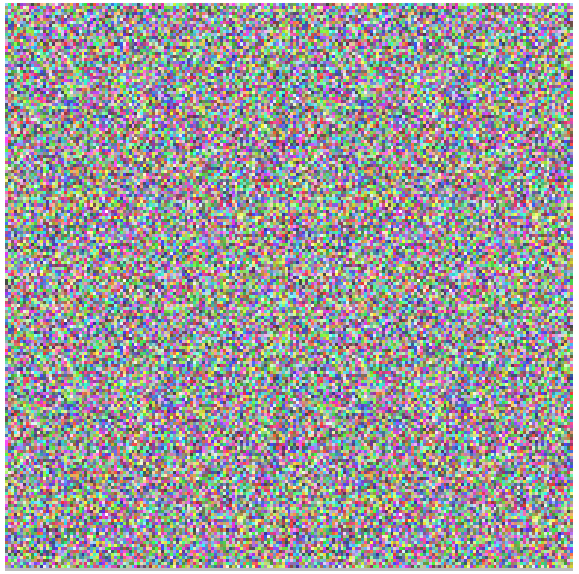

Training ResNet18 on CIFAR100 dataset with decaying learning rate (III)
テスト精度
訓練損失関数値
accuracy score for test
loss function value for training
epoch
75
70
65
60
55
50
45
1
0.1
0.01
0.002
0
40
80
120
160
200
SFW (constant bs)
SFWM (constant bs)
SFW (increasing bs)
SFWM (increasing bs)

敵対的攻撃＝敵対的サンプルを作る

“cat”

$\boldsymbol{x}$

＋ 0.007 ×



noise

＝

“dog”

Adversarial Examples

$\boldsymbol{x}_{\mathrm{adv}}$

ロバストモデル

- ロバストモデルとは，敵対的攻撃に備えて，防御のための特別な細工を施された訓練済みモデルのこと．
- いくつかの防御手法の中で，ロバストモデルを構成するための最も標準的な手法は敵対的訓練．
- 敵対的訓練とは，モデルの訓練時に，訓練データに敵対的サンプルを混ぜる手法．
- 敵対的攻撃を研究するモチベーションは，
  - 敵対的訓練に敵対的サンプルが必要だから．
  - ロバストモデルの性能評価に敵対的サンプルが必要だから．

敵対的攻撃は制約付き非凸最適化問題

経験損失関数

正解ラベル

何らかの
距離関数

許容できる
ノイズの大きさ

$$\text{maximize } f(g(\boldsymbol{x}_{\text{adv}}), \boldsymbol{y}) \text{ s.t. } d(\boldsymbol{x}_{\text{adv}}, \boldsymbol{x}) \leq \epsilon$$

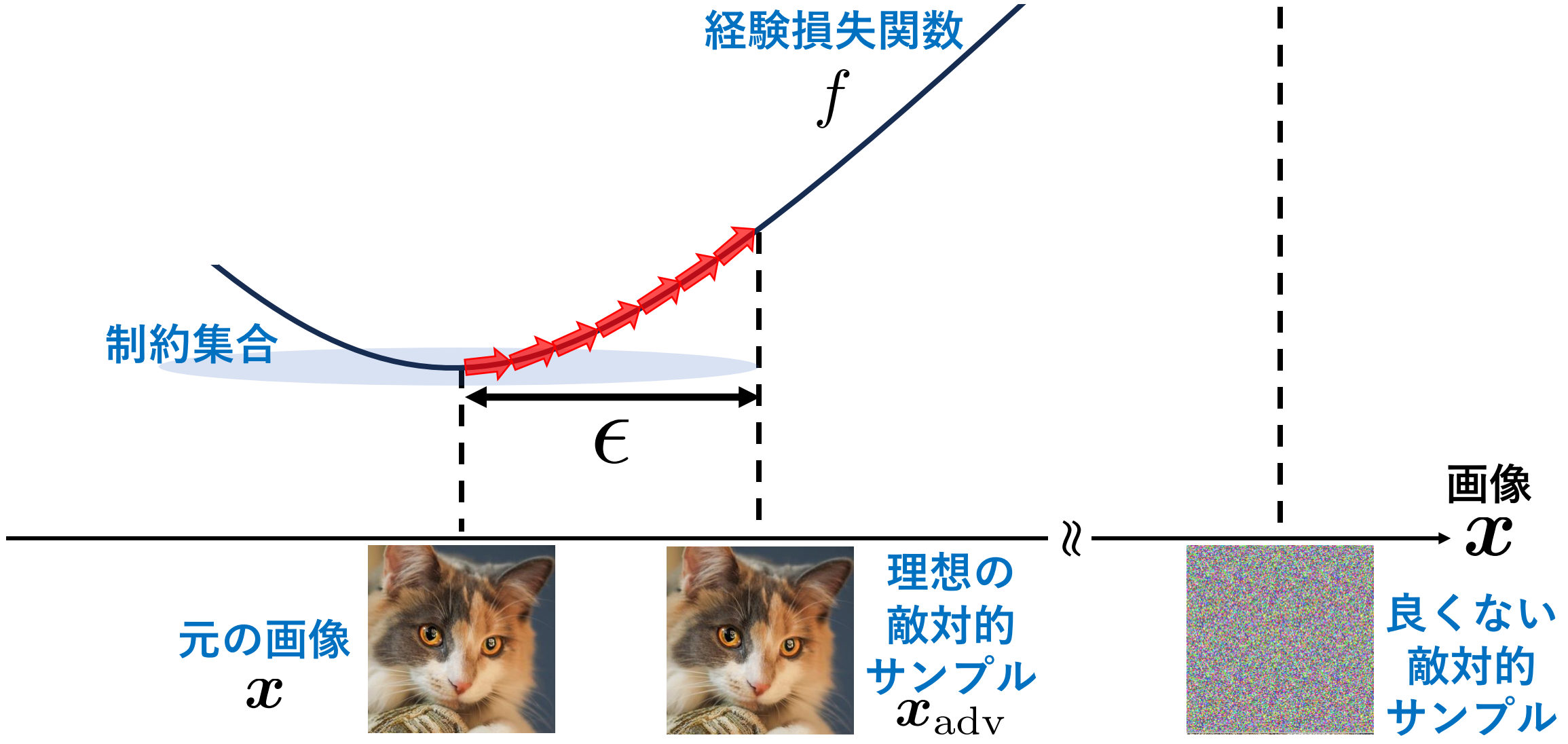
分類モデル(DNN)
の出力

攻撃後の画像

元の画像

▷攻撃後の画像と元の画像の差異を最大 ε に抑えながら，なるべく損失関数値を悪化させ，分類モデルの出力を歪ませる.

敵対的攻撃


経験損失関数
f
制約集合
ϵ
画像
x
元の画像
x
理想の
敵対的
サンプル
x_{adv}
良くない
敵対的
サンプル

Auto-PGD

Algorithm 1 APGD

Input: f, S, $x^{(0)}$, η, N_{iter}, $W = \{w_0, \ldots, w_n\}$
Output: $x_{\max}$, $f_{\max}$
$x^{(1)} \leftarrow P_S\left(x^{(0)} + \eta \nabla f(x^{(0)})\right)$
$f_{\max} \leftarrow \max\{f(x^{(0)}), f(x^{(1)})\}$
$x_{\max} \leftarrow x^{(0)}$ **if** $f_{\max} \equiv f(x^{(0)})$ **else** $x_{\max} \leftarrow x^{(1)}$
for $k = 1$ **to** $N_{\text{iter}} - 1$ **do**
  $z^{(k+1)} \leftarrow P_S\left(x^{(k)} + \eta \nabla f(x^{(k)})\right)$
  $x^{(k+1)} \leftarrow P_S\Big(x^{(k)} + \alpha(z^{(k+1)} - x^{(k)}) + (1-\alpha)(x^{(k)} - x^{(k-1)})\Big)$
  if $f(x^{(k+1)}) > f_{\max}$ **then**
    $x_{\max} \leftarrow x^{(k+1)}$ and $f_{\max} \leftarrow f(x^{(k+1)})$
  end if
  if $k \in W$ **then**
    if Condition 1 **or** Condition 2 **then**
      $\eta \leftarrow \eta/2$ and $x^{(k+1)} \leftarrow x_{\max}$
    end if
  end if
end for

PGDによる元の画像の更新

ある細工

Auto-FW

Algorithm 4 Auto-Frank Wolfe

Require: $f, \Omega, \boldsymbol{x}_0, \gamma_0, N_{iter}, W = \{w_0, \cdots, w_n\}$

$\boldsymbol{x}_{\text{adv}} \leftarrow \boldsymbol{x}_0, \boldsymbol{m}_{-1} \leftarrow \boldsymbol{0}$

for $t = 0$ to $N_{iter} - 1$ **do**

$\boldsymbol{v}_t = \underset{\boldsymbol{v} \in \Omega}{\text{argmax}} \langle \boldsymbol{v}, \nabla f(\boldsymbol{x}_t) \rangle$

$\boldsymbol{d}_t = \boldsymbol{v}_t - \boldsymbol{x}_t$

$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \gamma_t \boldsymbol{d}_t$

if $f(\boldsymbol{x}_{t+1}) > f(\boldsymbol{x}_{\text{adv}})$ **then**

$\boldsymbol{x}_{\text{adv}} \leftarrow \boldsymbol{x}_{t+1}$

end if

if $t \in W$ **then**

if Condition (i) or (ii) is satisfied **then**

$\gamma_{t+1} \leftarrow \gamma_t / 2$

$\boldsymbol{x}_{t+1} \leftarrow \boldsymbol{x}_{\text{adv}}$

end if

end if

end for

FWによる元の画像の更新

現在最も関数値を悪化させる画像を敵対的サンプルに設定

攻撃がうまくいっていないならば，学習率を半分にし，画像は良かったところまで引き戻す．

数値実験(CIFAR100, L_∞ 距離, $\epsilon = 8/255$)

paper	Architecture	clean accuracy	APGD	AFW
(Wang et al., 2023)	WideResNet-70-16	75.22	48.11	48.16
(Wang et al., 2023)	WideResNet-28-10	72.58	44.05	44.12
(Debenedetti et al., 2023)	XCiT-L12	70.76	38.97	39.02
(Rebuffi et al., 2021)	WideResNet-70-16	63.56	38.28	38.31
(Debenedetti et al., 2023)	XCiT-M12	69.21	38.69	38.77
(Pang et al., 2022)	WideResNet-70-16	65.56	36.66	36.71
(Debenedetti et al., 2023)	XCiT-S12	67.34	37.08	37.06
(Rebuffi et al., 2021)	WideResNet-28-16	62.41	35.73	35.8
(Jia et al., 2022)	WideResNet-34-20	67.31	36.46	36.62
(Addepalli et al., 2022a)	WideResNet-34-10	68.75	36.84	36.92
(Cui et al., 2021)	WideResNet-34-10	62.97	37.05	37.17
(Sehwag et al., 2022)	WideResNet-34-10	65.93	35.77	35.89
(Pang et al., 2022)	WideResNet-28-10	63.66	35.25	35.26
(Jia et al., 2022)	WideResNet-34-10	64.89	35.28	35.53
(Addepalli et al., 2022b)	WideResNet-34-10	65.73	35.64	36.07
(Cui et al., 2021)	WideResNet-34-20	62.55	34.08	34.27
(Cui et al., 2021)	WideResNet-34-10	60.64	33.99	34.10

まとめ

- SFWの現実的な学習率とバッチサイズの設定の下での収束解析を提供した.
- 特に，Frank-Wolfe gapが0に収束するのは，減少学習率と増加バッチサイズを同時に使うときだけであることを示した.
- 新しい敵対的攻撃手法Auto-FWを提案した.
- 最新のロバストモデルへの攻撃においてAuto-FWとAuto-PGDの性能を比較し，その限界を示した.