第22回情報論的学習理論ワークショップ (IBIS2019)
The 22nd Information-Based Induction Sciences Workshop
2019.11.20-23, ウインクあいち（名古屋市）

早期終了タイミングを予測する: 深層学習における確率勾配の分布の変化点検出

八島慶汰[1] 石川康太[2] 佐藤育郎[2] 野村哲弘[1] 横田理央[1] 松岡聡[3,1]
1: 東京工業大学, 2: デンソーアイティーラボラトリ, 3: 理化学研究所

I. 貢献

DNNのSGD学習の振る舞いと汎化性能の関係を実験的に調べるため，学習データ標本ごとの勾配ノルムの分布を毎イテレーション追跡し

- 確率勾配のノルムの分布は
  学習初期では正規分布に近く,
  学習後期 (=過学習時) ではべき分布に近い
- 正規分布・べき分布の尤度が拮抗する
  変化点の近くで汎化が最良となる

ことを実験的に確認した.

II. 背景 (SGDの確率微分方程式による解釈)

ミニバッチSGD (確率的勾配降下) の学習則：

$$W_{k+1} = W_k - \eta \left(\frac{1}{|B|} \sum_{i \in B} \nabla f(x_i; W_k) \right)$$

(W_k: 重み, x_i: サンプル, f: ロス関数, η: 学習率, B: ミニバッチ)

ミニバッチ勾配 = 真の勾配 + 確率的ノイズ：

$$\Delta W = -\eta \left(\underbrace{\frac{1}{|N|} \sum_{i \in N} \nabla f(x_i; W)}_{\text{真の勾配（フルバッチ）}} + \underbrace{\left(\frac{1}{|N|} \sum_{i \in N} \nabla f(x_i; W) - \frac{1}{|B|} \sum_{i \in B} \nabla f(x_i; W) \right)}_{\text{ノイズ項（分散 } \propto 1/|B|\text{）}} \right)$$

⇩ 連続時間による近似 (η→0)

- Brownian motion [1, 2]

$$dW_t = -\nabla f(W_t) dt + \sqrt{\beta^{-1} D(W_t)} dB_t$$

$$\beta := (\eta / |B|)^{-1} \text{ : 逆温度}$$

→ 解の分布はエントロピーの高い状態をとる[3]

- Levy motion [4]

$$dW_t = -\nabla f(W_t) dt + \beta^{-(\alpha-1)/\alpha} \sigma dL_t^{\alpha}$$

→ ヘビーテールな勾配ノイズによってNarrow minimaから抜け出しやすく，Wide minimaに留まりやすくなる[4]

ところが,

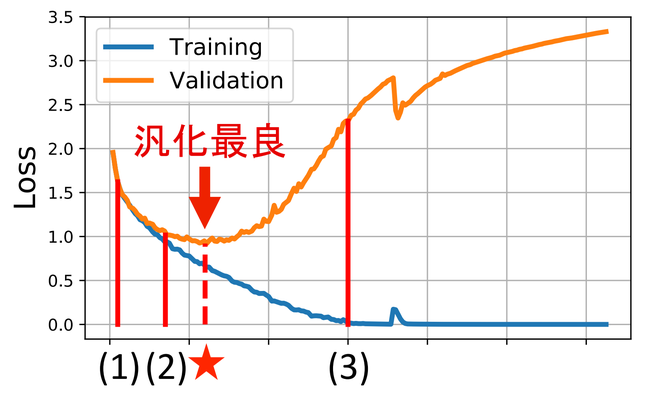
- 実際は，過学習を避けるためvalidation setを用いた早期終了が必要
- 早期終了時の確率勾配の性質は？？？

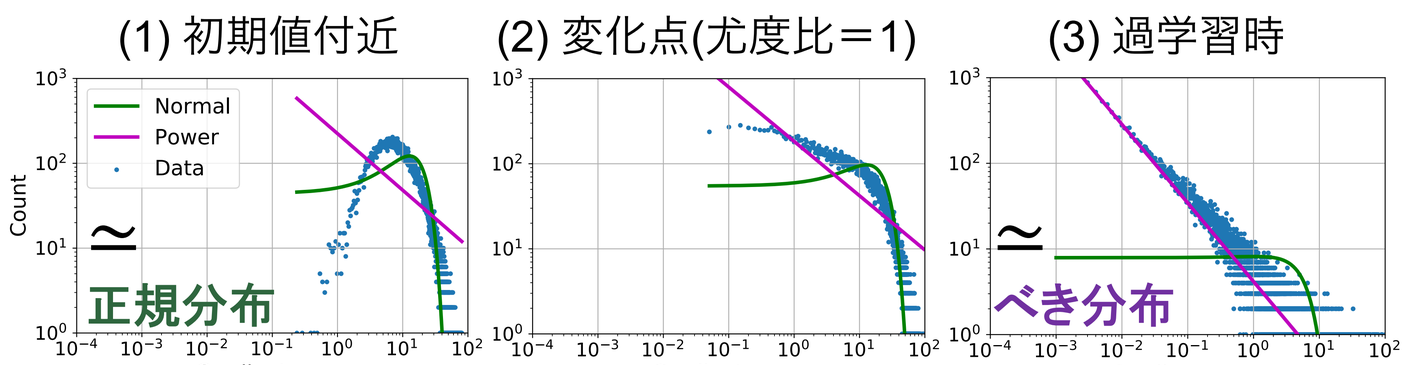
III.実験: 確率勾配のノルム分布の切り替わり

目的 初期値付近と過学習時での標本ごとの勾配ノルムの分布形状の違いを調べる

要領
- Conv4層 + FC2層, 第2Conv層の勾配を解析
- CIFAR10データセット
- 正規分布，べき分布の尤度を測定

結果 確率勾配のノルムは,
初期値付近で正規分布の尤度が高く，過学習時でべき分布の尤度が高い





IV.実験: 変化点と早期終了タイミングの関係

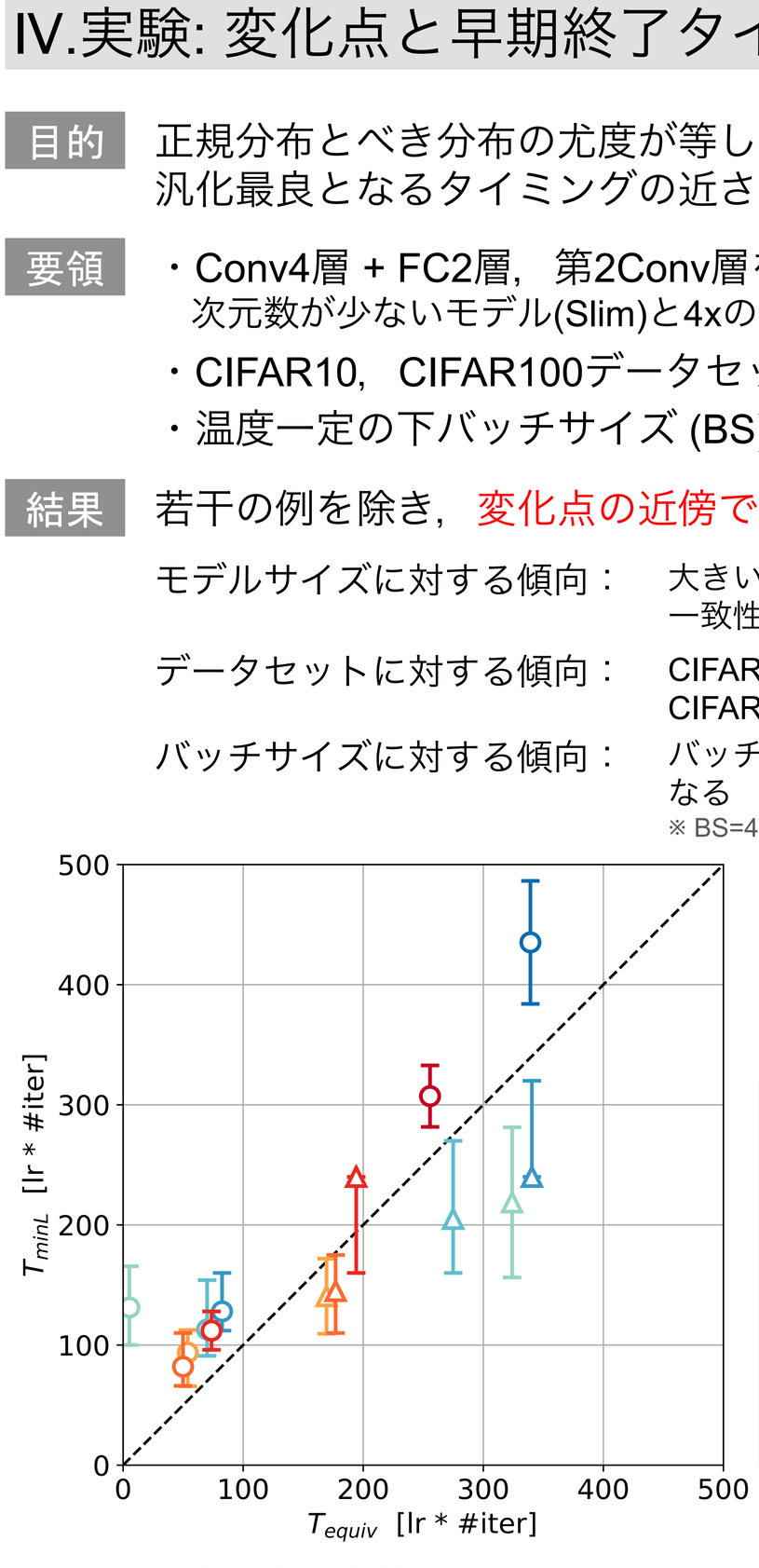
目的 正規分布とべき分布の尤度が等しくなる変化点と汎化最良となるタイミングの近さを評価する

要領
- Conv4層 + FC2層，第2Conv層を解析
  次元数が少ないモデル(Slim)と4xのモデル(Fat)を使用
- CIFAR10，CIFAR100データセット
- 温度一定の下バッチサイズ (BS) を変化 (1〜4096)

結果 若干の例を除き，変化点の近傍で汎化が最良となる

モデルサイズに対する傾向：	大きいモデルではより両タイミングの一致性が高い
データセットに対する傾向：	CIFAR10では相対的に変化点が早く、CIFAR100では遅い
バッチサイズに対する傾向：	バッチが大きいほど両タイミングが遅くなる

※ BS=4096, Slim, C100 のみ外れ値 $T_{equiv} \simeq 4,000$



エラーバー：
Validation lossが最小値+5%のタイミング

	Model	Data	BS	LR	Lmin
○	Slim	cifar10	1	6.250e-05	0.935
△	Slim	cifar100	1	3.125e-04	2.588
○	Slim	cifar10	16	1.000e-03	0.934
△	Slim	cifar100	16	5.000e-03	2.516
○	Slim	cifar10	256	1.600e-02	0.956
△	Slim	cifar100	256	8.000e-02	2.827
○	Slim	cifar10	4096	2.560e-01	1.031
○	Fat	cifar10	1	6.250e-05	0.940
△	Fat	cifar100	1	3.125e-04	2.535
○	Fat	cifar10	16	1.000e-03	0.907
△	Fat	cifar100	16	5.000e-03	2.577
○	Fat	cifar10	256	1.600e-02	0.943
△	Fat	cifar100	256	8.000e-02	3.044
○	Fat	cifar10	4096	2.560e-01	0.960

参考文献:
1. Li, Qianxiao+, “Stochastic modified equations and adaptive stochastic gradient algorithms.”, ICML 2017
2. Smith, Samuel L.+, "A bayesian perspective on generalization and stochastic gradient descent.”, ICLR 2018
3. Chaudhari, Pratik+, "Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks.”, ICLR 2018
4. Simsekli, Umut+. "A tail-index analysis of stochastic gradient noise in deep neural networks.”, ICML 2019