F-018

自然勾配法を用いたニューラルネットワークの学習のスキップ接続による影響

長瀬 准平[1,a)]　長沼 大樹[2,3,b)]
Nagase Jumpei[1,a)]　Naganuma Hiroki[2,3,b)]

1. はじめに

自然勾配法 [1] はその高い収束性にもかかわらず、実用上用いられることは極めて稀である。その原因として、深層学習の設定において、損失関数の非凸性により局所解に陥りやすいことや、学習が不安定になる問題が挙げられる。特に、大きなバッチサイズにおいてはこれらの問題が顕在化し汎化性能が劣化することが知られている [6]。一方で、モデルアーキテクチャの一つであるスキップ接続 [3] は、汎化と大域的最適解の保証においての問題点である損失関数の非凸性を改善することが知られている [2], [5], [9]。本研究では、この凸性の観点から、自然勾配法を用いたニューラルネットワークの学習において、スキップ接続がもたらす影響を考察すると共に数値実験の結果を比較する。

2. 導入

2.1 残差スキップ接続

残差スキップ接続 [3] は近年の深層学習において広く用いられている重要なモデルアーキテクチャの一つである。残差スキップ接続を持たないニューラルネットワークを g としたとき、入力 $\mathbf{x} \in \mathbb{R}^n$ に対して、残差スキップ接続を付与したモデルの出力 $\mathbf{y} \in \mathbb{R}^n$ は次式で表される：

$$\mathbf{y} = g(\mathbf{x}) + \mathbf{x} \tag{1}$$

このとき、元のモデル g は残差である $\mathbf{y} - \mathbf{x}$ を学習するとが期待される。この残差学習の構造が深い層の学習に重要であると考えられているが、スキップ接続が学習に及ぼす影響については未だ明らかでないことも多い。

2.2 自然勾配法

自然勾配法 [1] はパラメータ θ についてリーマン計量がフィッシャー情報行列 $F_\theta \in \mathbb{R}^{d\times d}$ で定まるリーマン空間の座標系として勾配を計算する。初期値 $\theta^{(0)}$ から t 回更新を行って得られるパラメータを $\theta^{(t)} \in \mathbb{R}^d$、$\eta \in \mathbb{R}$ を学習係数、L を損失関数とすると、自然勾配法による更新後のパラメータ $\theta^{(t+1)} \in \mathbb{R}^d$ は、

$$\theta^{(t+1)} = \theta^{(t)} - \eta F^{-1}_{\theta^{(t)}} \nabla L(\theta^{(t)}) \tag{2}$$

と計算される。本研究では、自然勾配法の近似手法である K-FAC[7] を用いて実験・評価を行っている。K-FAC はクロネッカー因子分解を用いることでメモリ消費量と計算量を抑えているが、その収束レートは、十分な幅を持つ NN において、厳密な自然勾配法と同様の大域的最適解への高速な収束性が知られている [4]。

3. スキップ接続を持つモデルでの自然勾配法

非凸性を緩和するとされるスキップ接続と、非凸性の影響を受けやすいとされる自然勾配法の両者の関係を調べるため、線型のニューラルネットがスキップ接続を持たない場合のモデル f と持つ場合のモデル f_s のそれぞれについての自然勾配法の更新式を考察する。入力を $\mathbf{x} \in \mathbb{R}^n$ に対して、f および f_s を次のように定める。

$$f(\mathbf{x}; \mathbf{W}_1, \mathbf{W}_2) = \mathbf{W}_2\mathbf{W}_1\mathbf{x} \tag{3}$$

$$f_s(\mathbf{x}; \mathbf{W}'_1, \mathbf{W}'_2) = \mathbf{W}'_2\mathbf{W}'_1\mathbf{x} + \mathbf{W}'_1\mathbf{x} \tag{4}$$

自然勾配法の更新 (2) に用いられるフィッシャー情報行列は次のように計算される。

$$F_\theta := \mathbb{E}\left[\nabla_\theta f(\mathbf{x};\theta) \nabla_\theta f(\mathbf{x};\theta)^T\right] + \lambda I. \tag{5}$$

ここで $\lambda \in \mathbb{R}$ はハイパーパラメータであり、λI は F_θ の退化を防ぐ他、二次最適化の学習を安定化させる効果を持つダンピング項である [8]。モデル f_s について、$\mathbf{W}'_2$ に関する勾配はスキップ接続の影響を受けない。すなわち、$\nabla_{\mathbf{W}_2} f = \nabla_{\mathbf{W}'_2} f_s$ である。一方で、モデル f の $\mathbf{W}_1$ と、モデル f_s の $\mathbf{W}'_1$ に関する勾配を考えると、

$$\begin{aligned}
\nabla_{\mathbf{W}_1} f(\mathbf{x};\theta) \nabla_{\mathbf{W}_1} f(\mathbf{x};\theta)^T &= \mathbf{W}_2\mathbf{x}\mathbf{x}^T\mathbf{W}_2^T \\
\nabla_{\mathbf{W}'_1} f_s(\mathbf{x};\theta) \nabla_{\mathbf{W}'_1} f_s(\mathbf{x};\theta)^T &= (\mathbf{W}'_2 + I)\mathbf{x}\left((\mathbf{W}'_2 + I)\mathbf{x}\right)^T \\
&= \mathbf{W}'_2\mathbf{x}\mathbf{x}^T\mathbf{W}_2'^T + \mathbf{x}\mathbf{x}^T
\end{aligned}$$

となり、モデル f_s の勾配には $\mathbf{x}\mathbf{x}^T$ が付加されていることがわかる。$\mathbf{x}\mathbf{x}^T$ は半正定値行列であり固有値が非負なので、モデル f と比べて f_s は F_θ の退化が起こりにくくなる。これはダンピング項 λI に類する効果であり、スキップ接続にも学習を安定化させる効果があると考えられる。

1 芝浦工業大学
Shibaura Institute of Technology
2 モントリオール大学
Université de Montréal
3 モントリオール学習アルゴリズム研究所
Montreal Institute for Learning Algorithms (MILA)
a) nb20106@shibaura-it.ac.jp
b) naganuma.hiroki@mila.quebec

4. 実験

機械学習フレームワークとしては PyTorch[*1] を、データセットとして MNIST[*2]を基本のモデルアーキテクチャとして中間層の層毎の素子数を 100 とした 6 層の線形 NN モデルを使用した。最適化手法としては SGD と K-FAC を用いて比較実験を行った。ただし、正則化は加えていない。ハイパーパラメータは学習率、学習率における線形減衰係数、慣性項をベイズ最適化によりチューニングし、各実験 300 試行を行った後、訓練精度の高い上位 100 件の試行についてその性能の平均を比較した。

4.1 小規模なバッチサイズにおける数値実験結果

表 1 Average Training Accuracy (BS=256)

	SGD	K-FAC
残差スキップ接続なし	92.16%	92.86%
残差スキップ接続あり	91.40%	93.64%
残差スキップ接続の有無による差	-0.7%	0.8%

表 2 Average Validation Accuracy (BS=256)

	SGD	K-FAC
残差スキップ接続なし	91.13%	91.32%
残差スキップ接続あり	90.32%	91.53%
残差スキップ接続の有無による差	-0.8%	0.2%

小規模なバッチサイズを用いた場合、K-FAC は Train・Validation の両者において SGD を超える性能を示したほか、残差スキップ接続をモデルに加えることにより、更に高い性能を示した。

4.2 大規模なバッチサイズにおける数値実験結果

表 3 Average Training Accuracy (BS=8192)

	SGD	K-FAC
残差スキップ接続なし	**92.87**%	92.56%
残差スキップ接続あり	93.24%	**93.57**%
残差スキップ接続の有無による差	0.3%	**1.0**%

表 4 Average Validation Accuracy (BS=8192)

	SGD	K-FAC
残差スキップ接続なし	**91.88**%	90.92%
残差スキップ接続あり	92.00%	**92.12**%
残差スキップ接続の有無による差	0.1%	**1.2**%

大規模なバッチサイズを用いた場合、残差スキップ接続がないモデルに対して K-FAC は Validation Accuracy が著しく低下し SGD よりも低い性能を示した。また、Training Accuracy にも同様の低下が見られた。残差スキップ接続を加えた場合、この性能劣化が改善し、Training・Validation のいずれも SGD を超える性能を示した。

[*1] https://pytorch.org/
[*2] http://yann.lecun.com/exdb/mnist/

5. おわりに

本研究は、近年広く用いられるアーキテクチャであるスキップ接続と、高い収束性を持つ自然勾配法の関係性について考察したものである。自然勾配法は非凸性の影響が大きな設定では有効でないとされており、実用上用いられていなかったが、本研究の結果として、スキップ接続を持つモデルの学習においては自然勾配法の性能劣化が防げることを実験により確認した。また、自然勾配法の学習を安定化させるためのダンピング法とスキップ接続の関連性が示された。以上のことから、近年広く用いられているスキップ接続を持つモデルにおいては自然勾配法が有効に使える可能性がある。今後の展望として、自然勾配法で有効に学習できるモデルアーキテクチャを提案することや、自然勾配法を始めとした学習アルゴリズムとアーキテクチャの関係についての理解を深めることが挙げられる。

謝辞 本研究は、JSPS 科研費 JP21J12812 の支援を受けたものである。本研究の K-FAC の実装に協力頂いた筑波大学の本川哲哉氏に感謝する。

参考文献

[1] Amari, S.-I.: Natural gradient works efficiently in learning, *Neural computation*, Vol. 10, No. 2, pp. 251–276 (1998).
[2] Hardt, M. and Ma, T.: Identity Matters in Deep Learning, *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, OpenReview.net (2017).
[3] He, K., Zhang, X., Ren, S. and Sun, J.: Deep residual learning for image recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778 (2016).
[4] Karakida, R. and Osawa, K.: Understanding Approximate Fisher Information for Fast Convergence of Natural Gradient Descent in Wide Neural Networks, *arXiv preprint arXiv:2010.00879* (2020).
[5] Li, H., Xu, Z., Taylor, G., Studer, C. and Goldstein, T.: Visualizing the Loss Landscape of Neural Nets, *Advances in Neural Information Processing Systems*, Vol. 31, Curran Associates, Inc. (2018).
[6] Ma, L., Montague, G., Ye, J., Yao, Z., Gholami, A., Keutzer, K. and Mahoney, M.: Inefficiency of k-fac for large batch size training, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, No. 04, pp. 5053–5060 (2020).
[7] Martens, J. and Grosse, R.: Optimizing neural networks with kronecker-factored approximate curvature, *International conference on machine learning*, PMLR, pp. 2408–2417 (2015).
[8] Martens, J. and Sutskever, I.: Training deep and recurrent networks with hessian-free optimization, *Neural networks: Tricks of the trade*, Springer, pp. 479–535 (2012).
[9] Yun, C., Sra, S. and Jadbabaie, A.: Are deep ResNets provably better than linear predictors?, *Advances in Neural Information Processing Systems*, Vol. 32, Curran Associates, Inc. (2019).