1S-03

自然勾配法におけるDampingによるFisher情報行列の正定値への影響と鞍点回避

The Effect of Damping on Positive Definiteness of Fisher Information Matrix in Natural Gradient Descent and Escaping Saddle Point

藤森 岳 *1　長瀬准平♯*2　長沼大樹♯*3
Gaku Fujimori　Jumpei Nagase　Hiroki Naganuma
*1東京理科大学　*2芝浦工業大学　*3モントリオール大学, Mila
Tokyo University of Science　Shibaura Institute of Technology　Université de Montréal, Mila

近年の深層学習の最適化手法において, 収束性の観点から二次最適化手法が見直されている．特に自然勾配法は計算コストの低い近似手法である K-FAC の提案により注目を集めている．これらの手法ではフィッシャー情報行列 F の逆行列を求める必要があるが, 深層学習の設定では F が退化する問題がある．そこで, F の正定値性を保つため一般に Damping と呼ばれるヒューリスティックな調整手法が用いられている．本研究では Damping が学習に与える影響を調査した．結果として，Damping 係数の大きさにより学習前半と後半の振る舞いが異なること，特に，係数が大きい場合は学習初期の鞍点回避効果があることが示された．

1. はじめに

深層学習における最適化手法は, 計算コストとその実装の手軽さの観点から一次の最適化手法に分類される勾配降下法が用いられることが一般的である．収束性の観点で優れている二次最適化手法は, パラメータ数 p の統計的モデルに対し $O(p^3)$ の計算複雑度を持つため, 深層学習のような p 自体が巨大な問題設定には不向きである．しかしながら, 自然勾配法 [Amari 98] の計算コストを削減した近似手法である K-FAC[Martens 15] などの提案により, 深層学習への適用が注目されている．自然勾配法は, Fisher 情報行列 F の逆行列を用いる最適化手法であり, データ数 n に対し $p >> n$ となることが一般的な深層学習の問題設定では, F は退化するため, 逆行列計算ができない．そのため実用上は, 逆行列の計算時に, F の対角成分に定数を加える Damping によって正定値性を保つ工夫がなされている．最適化手法の研究において, 未チューニングなハイパーパラメータは, 最適化手法の特性を大きく左右する [Choi 19] にもかかわらず, Damping の値はヒューリスティックに依存しており, その変化による影響は明らかでない．本研究では, K-FAC における Damping の値のグリッドサーチを行い, 学習への影響を鞍点回避・正定値性への影響の側面から調査した．

2. 自然勾配法と K-FAC 法

確率的勾配降下法を始めとする勾配法は, パラメータ空間をユークリッド空間と仮定した場合の最急方向にパラメータの更新を行う．対して, 自然勾配法 [Amari 98] は KL 距離を最小化する方向へのパラメータ更新を行う二次最適化手法である．リーマン計量がフィッシャー情報行列 $F_\theta \in \mathbb{R}^{p\times p}$ で定まるリーマン空間の座標系として勾配を計算する．t は更新回数として, パラメータを $\theta^{(t)} \in \mathbb{R}^p$ とし, $\eta \in \mathbb{R}$ を学習率, L を損失関数とすると, パラメータの更新は

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

と計算される．本研究では, 自然勾配法の近似手法の中でも最も一般に用いられている K-FAC[Martens 15] を採用した．K-FAC はクロネッカー因子分解により F_θ のブロック対角行列を更に近似し, メモリ消費量と計算量を抑えつつ高速な収束性を持つ．

連絡先: 1418098@ed.tus.ac.jp , ♯ equal contribution

3. Damping の学習効果

一般的な深層学習における自然勾配法では, F_θ の正定値性は担保されておらず, 自然勾配法以外の二次最適化手法であるニュートン法や, その近似手法である Adam においても同様である．逆行列を計算するための工夫として, F_θ の対角成分に定数を加える Damping が行われることが一般的である；

$$\theta^{(t+1)} = \theta^{(t)} - \eta \left(F_{\theta^{(t)}} + \lambda \mathbf{I}\right)^{-1} \nabla L(\theta^{(t)}) \tag{2}$$

ここで, λ は正の定数, $\mathbf{I}$ は恒等行列である．

Damping による学習の安定化の効果は, 正定値性への影響以外にも様々な観点で議論されている．例として数値解析の分野では, 学習初期における Damping の効果 [Nocedal 06] や, 信頼領域法との関連 [Martens 12] から, 問題に応じて Damping 項を設定することが多い．また Levenberg-Marquardt 法 [Moré 78] のように動的に Damping 項を調整する手法も知られている．しかしながら深層学習においては，Damping パラメータ λ の探索や Levenberg-Marquardt 法の計算コストが非常に高いことから, 問題に依らず固定された Damping 項を用いることが一般的である．

4. 実験

機械学習フレームワークとして PyTorch *1 を, データセットとして MNIST *2 を, 学習には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．KFAC 内の Damping 項を 1e-5 から 1e+2 まで 10 倍ごとの間隔でグリッドサーチを行った．各 Damping 項において学習率, 学習率における線形減衰係数, 慣性項をベイズ最適化でハイパーパラメータ探索, 各実験 300 試行を行った後, 訓練損失の低い順に 100 件の試行についてその性能の平均を比較した．

5. 結果・考察

異なる Damping 係数を用いた場合の学習曲線の結果を図 2 に示す．デフォルトの Damping 係数は 1e-3 であるが, 最

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

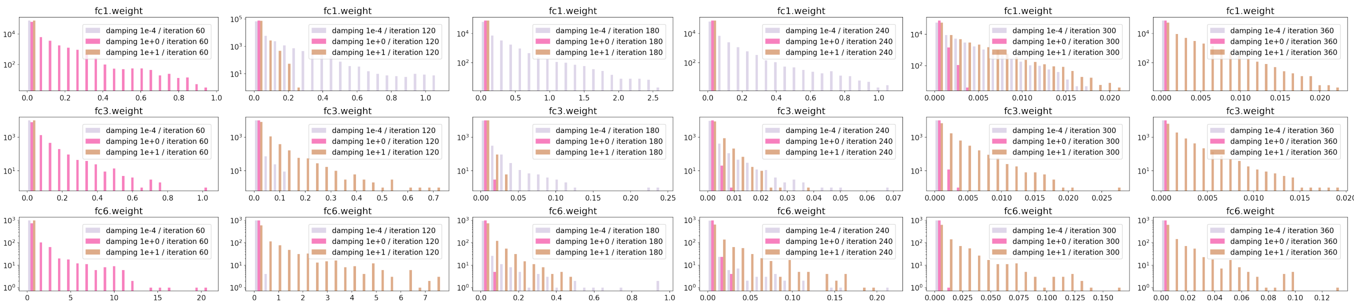

図 1: 損失関数の勾配 $(\nabla L(\theta))^2$ のヒストグラムの結果．水平方向は 60 iteration から 360 iteration まで 60 iteration ごとの推移を示す．垂直方向はレイヤの変化を示す．学習が最も良く進んだ Damping = 1e+0（凡例:ピンク色）の場合は，学習初期に勾配が大きいことから θ の更新が行われていて Plateau を回避している状態が見て取れる．対して Damping 係数が適切でない場合，学習初期の勾配の値が 0 に張り付く状況が確認された．

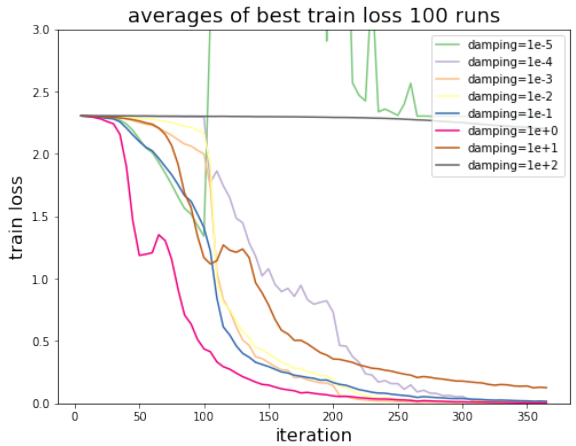

図 2: 異なる Damping 係数を用いた場合の訓練損失下位 100 件の試行の学習曲線の平均．損失関数の最小化において，適切な Damping 係数が存在することが確認できた．また，緑の学習曲線で示す通り，1e-5 などの小さすぎる Damping 係数 を用いた場合には学習が発散している．

も学習が上手く進んだのは Damping 係数が 1e+0 のときである．また，Damping 係数が 1e-5 のように極端に小さい場合は学習が不安定になり，逆に 1e+2 のようにさらに大きくなると学習速度が低下することが確認された．特に，Damping 項が十分大きいとき，式 (2) における $(F_{\theta^{(t)}} + \lambda\mathbf{I})^{-1}$ は対角成分が非常に小さい単位行列として近似でき，学習率が非常に小さい確率的勾配降下法と見做せるため，学習速度が低下する．

代表的な Damping 係数として 1e-4, 1e+0, 1e+1 を選び，Damping 係数と損失関数の勾配の挙動の関係を示したのが図１である．図２では学習曲線を確認したが，損失関数の勾配は学習の更新幅に関わるため，より詳細な学習の挙動や速度を図１の結果から考察する．学習初期においては Damping 項が大きいほうが勾配の値が大きく学習が速く進むが，学習後半においては Damping 項が小さいほうが勾配の値が大きくなっており，Damping 項の大きさによって学習の前半と後半で異なる挙動になることがわかる．これは図１に記載されていない他の係数の場合にも 1e-4 から 1e+1 の範囲でも同様である．学習初期はモデルのパラメータが安定せず F も安定しないため，Damping 項を大きくした SGD に近いパラメータ更新を行う場合の Plateau 回避を行うことを確認した．対して学習後半では F が安定するため Damping 項 $\lambda \to 0$ の時，更新方向は $(F_{\theta^{(t)}} + \lambda\mathbf{I})^{-1}\nabla L(\theta^{(t)}) \to F^{-1}_{\theta^{(t)}}\nabla L(\theta^{(t)})$ となり，二次最適化の恩恵を受け高速に収束することを示唆する結果となった．

6.　おわりに

自然勾配法における Damping 項の定数について，学習初期の Plateau 回避と学習後半の学習速度に関するトレードオフの関係を確認した．適応的な Damping 係数 の決定手法として，数値解析における Levenberg-Marquardt 法 [Moré 78] などが知られているが，深層学習などの問題設定では計算量に対するメリットが十分でないため一般には用いられていない．今後の課題として，本論文の結果を踏まえ適応的かつ計算コストの低い Damping 決定手法の開発が考えられる．

謝辞

本研究は，JSPS 科研費 JP21J12812 の支援を受けたものである．

参考文献

[Amari 98] Amari, S.-I.: Natural gradient works efficiently in learning, *Neural computation*, Vol. 10, No. 2, pp. 251–276 (1998)

[Choi 19] Choi, D., Shallue, C. J., Nado, Z., Lee, J., Maddison, C. J., and Dahl, G. E.: On empirical comparisons of optimizers for deep learning, *arXiv preprint arXiv:1910.05446* (2019)

[Martens 12] Martens, J. and Sutskever, I.: Training deep and recurrent networks with hessian-free optimization, in *Neural networks: Tricks of the trade*, pp. 479–535, Springer (2012)

[Martens 15] Martens, J. and Grosse, R.: Optimizing neural networks with kronecker-factored approximate curvature, in *International conference on machine learning*, pp. 2408–2417PMLR (2015)

[Moré 78] Moré, J. J.: The Levenberg-Marquardt algorithm: implementation and theory, in *Numerical analysis*, pp. 105–116, Springer (1978)

[Nocedal 06] Nocedal, J. and Wright, S. J.: Numerical Optimization, Springer series in operations research, *Siam J Optimization* (2006)