# CMA-ESを用いたニューラルネットワークの重み行列の最適化における頑健性の検証

Robustness verification of CMA-ES optimization for neural networks

清水 洸希 *1　小宮山 純平 *2　豊田 正史 *2
Shimizu Hiroki　Komiyama Junpei　Toyoda Masashi

*1東京大学大学院 情報理工学系研究科
Graduate School of Information Science and Technology, the University of Tokyo

*2東京大学 生産技術研究所
Institute of Industrial Science, the University of Tokyo

This paper aims to verify robustness of covariance matrix adaptation evolution strategy (CMA-ES) optimization for neural networks (NN). We added label noise to the training dataset. Unlike stochastic gradient descent (SGD), which is the state-of-the-art optimizer of NN, a CMA-ES based optimization was robust against label noise.

## 1. はじめに

ニューラルネットワーク (NN) は近年，画像認識や機械翻訳の分野において人間の精度を超える，あるいはそれに迫る性能を発揮しており [1][2]，今後，実世界のさまざまな分野での応用が期待されている．

NN のモデルの評価には，画像分類においては，MNIST や CIFAR-10 といったデータセットが一般的に用いられる．これらのデータは，学習データに対して正しく教師ラベルが付与されており，非常にラベルノイズの小さいデータセットであると言える．一方で，実世界における応用を考えた場合，誤った教師ラベルが付与されることが低くない確率で起きることが予想され，NN の最適化には強い頑健性が求められる．

NN の最適化には，確率的勾配降下法 (SGD) と呼ばれるアルゴリズムが性能や実行速度の観点から広く使われている．最適化の手法としては，SGD 以外にもいくつかのアルゴリズムが存在しており，清水らの研究 [3] により，2 層の非常に浅い NN においては，Covariance matrix adaptation evolution strategy (CMA-ES)[4] が SGD と同等の精度が得られることが報告されている．

CMA-ES は進化戦略と呼ばれるアルゴリズムの一種であり，多変量正規分布により複数の NN を生成し，評価値の高い NN の重み行列により多変量正規分布のパラメタを更新することで NN の最適化を行う．進化戦略による最適化の特徴として，一般に頑健性が高いと言われているが [5]，それについて検証が行われることは非常に少なく，特に NN の最適化にあたっては先行研究がない．

そこで，本稿では，学習データのラベルに対して人為的にノイズを付与し，それを用いた最適化を行うことで，CMA-ES による最適化の頑健性を検証し，SGD により最適化を行った場合との比較を行う．

## 2. 関連研究

ラベルノイズ下での NN の最適化の頑健性向上に関する研究は近年盛んに行われている．

損失関数に対して，ラベルノイズに関する正則化を与えることで，性能の向上を図った研究として，Patrini らによる研究 [6] と Ghoshr らによる研究 [7] が挙げられる．Patrini らは，ノイズ発生確率の情報を損失関数に与えることで，ラベルノイズ下での性能を向上させた．ノイズの発生確率が未知の場合でも，それを推定する手法を提案しているが，その場合はネットワークの構造に工夫をしなければノイズを過学習するという課題もある．Patrini らの手法は損失関数をクロスエントロピー関数に限定しているという課題があったが，Ghoshr らは平均二乗誤差や平均絶対誤差に対してもラベルノイズ化での最適化を可能とした．

Goldberger ら [8] は，NN の構造に対して，ラベルノイズに対する適応層を追加することで性能の向上を図った．

以上の研究は，損失関数や層に対しての提案であり，いずれも本稿の CMA-ES と組み合わせて用いることができる．一方で，CMA-ES と組み合わせることのできないような，最適化アルゴリズム自体への提案をする研究は知りうる限り存在せず，今後の研究が期待される．

本稿では，CMA-ES による最適化の頑強性を明確に検証するため，これらの既存研究との組み合わせは行わずに実験を行った．

## 3. CMA-ES

### 3.1 概要

CMA-ES は Hansen らが考案した最適化アルゴリズムである．SGD が 1 つの NN を定義し，その重み行列を更新していくのに対して，CMA-ES では N 個のネットワークを定義し，その重み行列を多変量正規分布を用いて生成し，評価値の高い NN の重み行列を用いて，多変量正規分布のパラメタを更新することによって NN を間接的に最適化する．SGD が最適化の過程で，目的関数や活性化関数の連続性や微分可能性を必要とするのに対して，CMA-ES はそれらを必要としないという特徴がある．CMA-ES における課題としては，各ステップで N 個のネットワークの重み行列のパラメタを保持する必要があるため，莫大なメモリを消費するという点が挙げられる．

### 3.2 CMA-ES の最適化ステップ

CMA-ES は多変量正規分布により NN の重み行列を生成するが，多変量正規分布 $N(m, C)$ はパラメタとして平均 m と分散共分散行列 C を持つ．ここで平均 m および分散共分散行列の次元数は NN の重み行列の要素数である．CMA-ES の最適

連絡先: {shimizu, jkomiyama, toyoda}@tkl.iis.u-tokyo.ac.jp

化ステップではまず，$N(m, C)$ から N 個の重み行列 $\{x_i\}_{i=1}^N$ をサンプリングする．これらの N 個の NN に対して，SGD と同様に学習データを順伝播させ，各 $x$ に対してコストを計算する．そのコストに基づき，各 $x$ に対して 1 から N までの順位をつけ，以下の式 (1)〜(5) よりパラメタ m 及び C の更新を行う．ここで $t$ は最適化のステップ数，$\sigma$ は収束に応じて値が大きくなる変数，$r$ は閾値である．

コストの計算については，一般的な進化戦略においては，全ての学習データのコストを計算して順位をつけるが，Morse らの研究 [9] により，全てのデータではなく，ミニバッチに区切ったデータを用いた方が性能が向上することが報告されており，本稿においても，このミニバッチを採用した．

$$m^{t+1} = m^t + \eta_m \sum_{i=1}^{N} \tilde{w}_i (x_i - m^t) \quad (1)$$

$$C^{t+1} = C^t + \eta_c \sum_{i=1}^{N} \tilde{w}_i \left( \frac{(x_i - m^t)(x_i - m^t)^T}{\sigma^t} - C^t \right) \quad (2)$$

$$p_\sigma^{t+1} = p_\sigma^t + C^{t-\frac{1}{2}} \frac{m^{t+1} - m^t}{\sigma^t} \quad (3)$$

$$\sigma^{t+1} = \sigma^t \exp\left( \frac{\|p_\sigma\|}{E\|N(0, I)\|} - 1 \right) \quad (4)$$

$$\tilde{w}_i = \frac{1}{N} \mathbb{I}_{(rank(x_i) \geq r)} \quad (5)$$

## 4. 実験

実世界におけるデータにはラベルノイズが発生している可能性が高いと考えられるが，どの程度のノイズがあるかを知ることが難しいため，検証においては，ベンチマーク用のデータセットである MNIST を用いた．MNIST のデータを 49500 個の学習データと 5500 個のテストデータに分け，学習データのラベルを，一定の割合でランダムに書き換えることで，ラベルノイズを人為的に発生させた．発生させるノイズは 0%から 20%まで 5%刻みで変化させた．

検証には入力層と出力層が直結した 2 層の NN を用いた．非常に浅い NN であるが，CMA-ES のメモリ制約上の都合からこれを使用した．比較対象として，CMA-ES と SGD の他に，SGD の学習率を自動的に調整するアルゴリズムである ADAM[10] も用いた．各手法のパラメタ設定として，バッチサイズは全て 256 を用い，学習率は ADAM は初期設定値，SGD については 0.01 とした．評価には F 値を用いた．

実験の結果を表 1 及び図 1 に示した．表 1 を見ると，ノイズが 0%のときは ADAM 及び SGD が優れているが，ノイズが大きくなるに従って ADAM 及び SGD が大きく性能を下げたのに対して，CMA-ES は小さい下げ幅に止まった．ラベルノイズが 20%の場合においては，他の手法が F 値で 0.9 を下回ったのに対して，CMA-ES は 0.9 以上を維持しており，頑健性の高さを示した．

表 1: CMA-ES と SGD との F 値による比較 (MNIST)

| ラベルノイズ [%] | CMA-ES | ADAM | SGD |
|---|---|---|---|
| 0 | 0.918 | 0.924 | 0.922 |
| 5 | 0.918 | 0.915 | 0.913 |
| 10 | 0.918 | 0.910 | 0.910 |
| 15 | 0.917 | 0.906 | 0.903 |
| 20 | 0.916 | 0.899 | 0.900 |

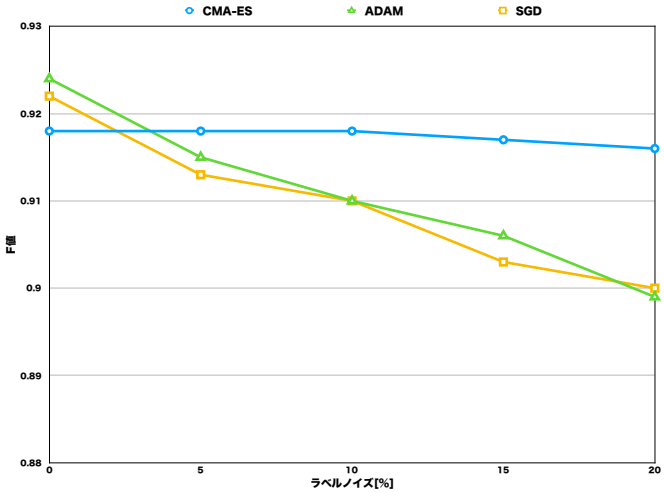


図 1: CMA-ES と SGD との比較 (MNIST)

## 5. 考察

SGD が，ラベルノイズの割合が上がるに連れて急激に性能が低下したことについて，SGD が各学習データのコストに対して勾配を計算し，重み行列の値を更新していることに原因があると考えられる．ミニバッチ学習法を用いることで，1 つの学習データが与える影響を小さくすることができるが，その影響が必ず勾配と重み行列の更新値へ直接及ぶからである．

一方で，CMA-ES においては，各学習データが更新に直接与える影響は，各 NN の順位を決定する際のコストの算出のみである．確率分布のパラメタの更新は，その順位に基づいて，各 NN の重み行列の値から行われるため，ラベルノイズを含んだ学習データが順位の上下に影響を与えない限りは，最適化への影響は生じないと言える．これにより，CMA-ES はラベルノイズが上昇しても F 値が大きく下がらなかったと考えられる．

## 6. 終わりに

本稿では，NN の最適化における CMA-ES の頑健性の高さを示した．一方で，検証実験において，メモリの制約上から 2 層という非常に浅い NN しか用いることができないという問題が生じた．

今後の課題としては，(1)MNIST 以外のデータセットでの検証，(2) より深い NN での検証，(3) 既存の損失関数や層への正則化手法との組み合わせの検証，が挙げられる．

## 謝辞

本研究は JSPS 科研費 16H02905, 17K12736 の助成を受けたものです．

## 参考文献

[1] Stanford Vision Lab. ImageNet, 2015.

[2] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff

Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. pp. 1–23, 2016.

[3] 清水洸希, 小宮山純平, 豊田正史. 進化戦略を併用した neural network の重み最適化. 第 10 回データ工学と情報マネジメントに関するフォーラム (DEIM2018), 2018.

[4] Nikolaus Hansen. The CMA Evolution Strategy: A Tutorial. Vol. 102, No. 2006, pp. 75–102, 2016.

[5] Hitoshi Iba. 進化計算と深層学習. *Ohmsha*, 2017.

[6] Giorgio Patrini, Alessandro Rozza, Aditya Menon, Richard Nock, and Lizhen Qu. Making Deep Neural Networks Robust to Label Noise: a Loss Correction Approach. sep 2016.

[7] Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. Robust Loss Functions under Label Noise for Deep Neural Networks. pp. 1919–1925, 2017.

[8] Jacob Goldberger and Ehud Ben-Reuven. Training Deep Neural Networks using a Noise Adaptation Layer. *Iclr 2017*, No. 2014, pp. 1–9, 2017.

[9] Gregory Morse and Kenneth O. Stanley. Simple Evolutionary Optimization Can Rival Stochastic Gradient Descent in Neural Networks. *Proceedings of the 2016 on Genetic and Evolutionary Computation Conference - GECCO '16*, No. Gecco, pp. 477–484, 2016.

[10] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. dec 2014.