ニューラルネットワークモデルにおける 加速の効果と学習率に制限を加えた適応的な学習アルゴリズム

岡部 駿,* 伏木 忠義

1 はじめに

近年，機械学習の技術が著しい発達を遂げている．機械学習は人間が行う学習能力と同等の機能をコンピュータ内で実現しようとする技術であり，その発達の背景には深層ニューラルネットワークモデルの成功が大きく影響している．物体の認識率を競う ILSVRC2012 では，深層ニューラルネットワークモデルが他のネットワークモデルをはるかに上回る高精度機能を実現した．その後，画像認識にとどまらず様々な分野で深層ニューラルネットワークモデルによって高い性能が得られることが報告されている．実際に深層ニューラルネットワークモデルが利用される場面としては，自動車の自動運転システム，国立がん研究センターによる胃がん領域の検出，音声認識機能の Siri，囲碁の世界トップクラスの棋士に勝利した囲碁ソフト AlphaGo[9] などが挙げられる．

深層ニューラルネットワークモデルの有用性は広く認知されるようになっているが，大規模なネットワークを用いるためその学習に大きな時間が必要になることが課題の一つとして存在する．深層ニューラルネットワークモデルの学習アルゴリズムとしては，確率的勾配降下法，モーメンタム法，Adam などがよく利用されている．Adam [4] は適応的な学習率を用いたアルゴリズムであり，バイアス補正を行った 1 次モーメントと 2 次モーメントの推定値を利用している．Adam に Nesterov の加速 [7] の効果を加えた Nadam [1] やクリッピングの効果を加えた AdaBound [5] など，Adam にはいくつかの派生アルゴリズムが存在する．Adam では適応的な学習率を用いているため，学習率が極端な値をとることがありそれによって性能の低下が起こりえる．AdaBound では学習率に上限と下限を設けることで学習をスムーズにしている．本研究では，AdaBound に Nesterov の加速の効果を加えた学習アルゴリズムを提案しその性能を評価する．

本論文の構成は以下のとおりである．第 2 節では，提案アルゴリズムを示し，Adam [4] や AdaBound [5] と同様の設定で凸最適化問題における性能を評価する．第 3 節では，実データを用いて提案アルゴリズムの画像認識における精度を評価する．第 4 節ではまとめを行う．

2 提案アルゴリズム

本論文で提案する手法は AdaBound[5] に Nesterov の加速勾配降下法 [7] のアイデアを組み合わせた適応的学習アルゴリズムとなっている．まず提案手法のベースとなる AdaBound を Algorithm1 に記す．Algorithm1 において，E_t は大きさ M のミニバッチにおける誤差関数を表す．また，$\mathbf{x} \odot \mathbf{y}$ はベクトル $\mathbf{x}$ とベクトル $\mathbf{y}$ の成分ごとの積をとることで得られるベクトルを指し，$\mathrm{Clip}(\mathbf{x}, a, b)$ はベクトル $\mathbf{x}$ の各成分を $[a, b]$ 内の値にする関数で (つまり，$\mathbf{x}$ の各成分が b 以上の場合は b に a 以下の場合は a にする関数)，$\prod_{\mathcal{F},Q} \mathbf{x} = \underset{\mathbf{y}\in\mathcal{F}}{\mathrm{argmin}} (\mathbf{x}-\mathbf{y})^{\mathrm{T}} Q (\mathbf{x}-\mathbf{y})$ としている．

提案手法である NadaBound を Algorithm2 に記す．

Algorithm 2 は Nadam [1] に AdaBound と同様に学習率にクリッピング手法による動的制限を加えている．また $0 \le \mu < 1$，$0 \le \nu < 1$，$0 \le \mu_t < 1$ と仮定し，ε はごく小さい定数を用いるようにとる．

2020.5.25 受理

* 鶴岡市立藤島中学校

Algorithm 1 AdaBound(ミニバッチ学習)

Require: 時刻の初期化 : $t = 1$，学習率 :α，時刻 1 でのパラメータ : $\mathbf{w}_1$，誤差分の定数 :ε
Require: 指数減衰率 :$\beta_{1,t}$およびβ_2，一次と二次のモーメントの変数を初期化 : $\mathbf{m}_0 = \mathbf{0}, \mathbf{v}_0 = \mathbf{0}$
Require: 学習率を制限する下界の関数η_l, 上界の関数η_u

while 終了条件を満たさない **do**
　訓練データの集合 $\{(\mathbf{x}_{(1)}, \mathbf{y}_{(1)}), (\mathbf{x}_{(2)}, \mathbf{y}_{(2)}), \cdots, (\mathbf{x}_{(N)}, \mathbf{y}_{(N)})\}$ から M 個のデータをランダムに抽出してミニバッチを作成する
　勾配の推定値を計算する : $\mathbf{g}_t \leftarrow \nabla E_t(\mathbf{w}_t)$
　バイアス付きの一次モーメントの推定を計算する : $\mathbf{m}_t \leftarrow \beta_{1,t}\mathbf{m}_{t-1} + (1-\beta_{1,t})\mathbf{g}_t$
　バイアス付きの二次モーメントの推定を計算する : $\mathbf{v}_t \leftarrow \beta_2\mathbf{v}_{t-1} + (1-\beta_2)\mathbf{g}_t \odot \mathbf{g}_t$
　二次モーメントを対角行列に変換する : $V_t = \mathrm{diag}\,(\mathbf{v}_t)$
　学習率の動的な制限を与える :$\hat{\boldsymbol{\eta}}_t = \mathrm{Clip}(\alpha/(\sqrt{V_t} + \varepsilon), \eta_l(t), \eta_u(t))$
　$\boldsymbol{\eta}_t = \hat{\boldsymbol{\eta}}_t/\sqrt{t}$
　更新を適用する : $\mathbf{w}_{t+1} \leftarrow \prod_{\mathcal{F},\mathrm{diag}(\boldsymbol{\eta}_t^{-1})}(\mathbf{w}_t - \boldsymbol{\eta}_t \odot \mathbf{m}_t)$
　$t \leftarrow t + 1$
end while

補題 1. ([6]) $Q \in \mathcal{S}_+^d$, $\mathcal{F} \subset \mathbb{R}^d$ を凸集合とする．ここで，$\mathcal{S}_+^d$ は正定値行列の集合である．$\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^d$ に対して，$\mathbf{u}_1 = \arg\min_{\mathbf{x}\in\mathcal{F}} \|Q^{1/2}(\mathbf{x} - \mathbf{z}_1)\|$,$\mathbf{u}_2 = \arg\min_{\mathbf{x}\in\mathcal{F}} \|Q^{1/2}(\mathbf{x} - \mathbf{z}_2)\|$ とする．このとき，次式が成り立つ．

$$\|Q^{1/2}(\mathbf{u}_1 - \mathbf{u}_2)\| \le \|Q^{1/2}(\mathbf{z}_1 - \mathbf{z}_2)\|.$$

補題 2. 関数 $E : \mathbb{R}^d \to \mathbb{R}$ は微分可能な凸関数であると仮定する．このとき，任意の $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ において次の不等式が成り立つ．

$$E(\mathbf{y}) \ge E(\mathbf{x}) + \langle \nabla E(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle.$$

定理 3. Algorithm2 において，$1 > \mu_1 \ge \mu_2 \ge \cdots \ge \mu_T \ge 0$，$\eta_l(T) \ge \cdots \ge \eta_l(2) \ge \eta_l(1) = L > 0$, $\eta_u(T) \le \cdots \le \eta_u(2) \le \eta_u(1) = R < \infty$ とし，$\mathscr{F} \subset \mathbb{R}^d$ は閉かつ凸な集合とする．任意の $\mathbf{w}, \mathbf{x} \in \mathscr{F}$ において $\|\mathbf{w} - \mathbf{x}\|_\infty \le D_\infty$，また任意の $\mathbf{w} \in \mathscr{F}$，$t \in \{1, 2, \ldots, T\}$ において $\|\mathbf{g}_t\| \le G_2$ であり，関数 $E_t(\mathbf{w})$ は微分可能な凸関数とする．ここで，$\mathbf{v}^* = \arg\min_{\mathbf{w}} \sum_{t=1}^T E_t(\mathbf{w})$，$R_T = \sum_{t=1}^T (E_t(\mathbf{w}_t) - E_t(\mathbf{v}^*))$ とするとき，次式が成り立つ．

$$\begin{aligned} R_T \le{}& \frac{D_\infty^2 d}{2L} + \frac{D_\infty^2}{2(1-\mu_1)} \sum_{t=2}^T \sum_{i=1}^d (\eta_{t,i}^{-1} - \eta_{t-1,i}^{-1}) + \frac{RG_2^2\sqrt{d}}{2} \sum_{t=1}^T \frac{1-\mu_t}{\sqrt{t}(1-\prod_{k=1}^t \mu_k)^2} \\ &+ RG_2^2\sqrt{d} \sum_{t=1}^T \frac{\mu_{t+1}}{\sqrt{t}(1-\prod_{k=1}^t \mu_k)(1-\prod_{k=1}^{t+1} \mu_k)} + \frac{RG_2^2\sqrt{d}}{2} \sum_{t=1}^T \frac{\mu_{t+1}^2 + \mu_{t+1}}{\sqrt{t}(1-\mu_t)(1-\prod_{k=1}^{t+1} \mu_k)^2} \\ &+ \frac{D_\infty^2 d}{2(1-\mu_1)L} \sum_{t=1}^T \mu_{t+1}\sqrt{t} + \frac{D_\infty^2 d}{2(1-\mu_1)L} \sum_{t=1}^T \mu_t \sqrt{t}. \end{aligned}$$

Algorithm 2 NadaBound(ミニバッチ学習)

Require: 時刻の初期化 : $t = 1$, 学習率 : α, 誤差分の定数 : ε, 位置パラメータ : $\mathbf{w}_t$
Require: モーメンタム : μ_t, 指数減衰率 : μとν, 一次と二次のモーメントの初期化 : $\mathbf{m}_0 = \mathbf{0}, \mathbf{v}_0 = \mathbf{0}$
Require: 学習率を制限する下界の関数η_l, 上界の関数$\eta_u, \eta_{0,i} = \eta_u(1)$
while 終了条件を満たさない **do**
  訓練データの集合 $\{(\mathbf{x}_{(1)}, \mathbf{y}_{(1)}), (\mathbf{x}_{(2)}, \mathbf{y}_{(2)}), \cdots, (\mathbf{x}_{(N)}, \mathbf{y}_{(N)})\}$ から M 個のデータをランダムに抽出してミニバッチを作成する
  勾配の推定値を計算する : $\mathbf{g}_t \leftarrow \nabla E_t(\mathbf{w}_t)$
  勾配の推定に補正を行う : $\hat{\mathbf{g}}_t \leftarrow \frac{\mathbf{g}_t}{1-\prod_{k=1}^{t}\mu_k}$
  一次モーメントの推定を計算する : $\mathbf{m}_t \leftarrow \mu\mathbf{m}_{t-1} + (1-\mu)\mathbf{g}_t$
  二次モーメントの推定を計算する : $\mathbf{v}_t \leftarrow \nu\mathbf{v}_{t-1} + (1-\nu)\mathbf{g}_t \odot \mathbf{g}_t$
  一次モーメントの推定に補正を行う : $\hat{\mathbf{m}}_t \leftarrow \frac{\mathbf{m}_t}{1-\prod_{k=1}^{t+1}\mu_k}$
  二次モーメントの推定に補正を行う : $\hat{\mathbf{v}}_t \leftarrow \frac{\mathbf{v}_t}{1-\nu^t}$
  加速の効果を与える : $\bar{\mathbf{m}}_t \leftarrow (1-\mu_t)\hat{\mathbf{g}}_t + \mu_{t+1}\hat{\mathbf{m}}_t$
  学習率の動的な制限を与える : $\hat{\boldsymbol{\eta}}_t \leftarrow \mathrm{Clip}(\alpha/(\sqrt{\hat{\mathbf{v}}_t} + \varepsilon), \eta_l(t), \eta_u(t))$
  学習率のより小さい値を選択して更新する : $\boldsymbol{\eta}_t \leftarrow (\hat{\boldsymbol{\eta}}_t/\sqrt{t}) \wedge \boldsymbol{\eta}_{t-1}$
  更新を適用する : $\mathbf{w}_{t+1} \leftarrow \prod_{\mathcal{F},\mathrm{diag}(\boldsymbol{\eta}_t^{-1})}(\mathbf{w}_t - \boldsymbol{\eta}_t \odot \bar{\mathbf{m}}_t)$
  $t \leftarrow t+1$
end while

証明. まず $\mathcal{F}$ は閉かつ凸な集合なので，$\mathbf{v}^*$ は存在する．

$$\begin{aligned}
\mathbf{w}_{t+1} &= \prod_{\mathcal{F},\mathrm{diag}(\boldsymbol{\eta}_t^{-1})} (\mathbf{w}_t - \boldsymbol{\eta}_t \odot \bar{\mathbf{m}}_t) \\
&= \arg\min_{\mathbf{w}\in\mathcal{F}} (\mathbf{w} - (\mathbf{w}_t - \boldsymbol{\eta}_t \odot \bar{\mathbf{m}}_t))\mathrm{diag}(\boldsymbol{\eta}_t^{-1})(\mathbf{w} - (\mathbf{w}_t - \boldsymbol{\eta}_t \odot \bar{\mathbf{m}}_t)) \\
&= \arg\min_{\mathbf{w}\in\mathcal{F}} \{\boldsymbol{\eta}_t^{-1/2} \odot (\mathbf{w} - (\mathbf{w}_t - \boldsymbol{\eta}_t \odot \bar{\mathbf{m}}_t))\}^{\mathrm{T}}\{\boldsymbol{\eta}_t^{-1/2} \odot (\mathbf{w} - (\mathbf{w}_t - \boldsymbol{\eta}_t \odot \bar{\mathbf{m}}_t))\} \\
&= \arg\min_{\mathbf{w}\in\mathcal{F}} \|\boldsymbol{\eta}_t^{-1/2} \odot (\mathbf{w} - (\mathbf{w}_t - \boldsymbol{\eta}_t \odot \bar{\mathbf{m}}_t))\|. \qquad (1)
\end{aligned}$$

補題 1 に $\mathbf{u}_t = \mathbf{w}_{t+1}$，$\mathbf{u}_2 = \mathbf{v}^*$ とすることで次のように表せる．

$$\begin{aligned}
\|\boldsymbol{\eta}_t^{-1/2} \odot (\mathbf{w}_{t+1} - \mathbf{v}^*)\|^2 &\le \|\boldsymbol{\eta}_t^{-1/2} \odot (\mathbf{w}_t - \boldsymbol{\eta}_t \odot \bar{\mathbf{m}}_t - \mathbf{v}^*)\|^2 \\
&= \|\boldsymbol{\eta}_t^{-1/2} \odot (\mathbf{w}_t - \mathbf{v}^*)\|^2 + \|\boldsymbol{\eta}_t^{1/2} \odot \bar{\mathbf{m}}_t\|^2 - 2\langle\mu_{t+1}\hat{\mathbf{m}}_t + (1-\mu_t)\hat{\mathbf{g}}_t, \mathbf{w}_t - \mathbf{v}^*\rangle. \qquad (2)
\end{aligned}$$

これより，

$$\begin{aligned}
\langle\hat{\mathbf{g}}_t, \mathbf{w}_t - \mathbf{v}^*\rangle &\le \frac{1}{2(1-\mu_t)}\left\{\|\boldsymbol{\eta}_t^{-1/2} \odot (\mathbf{w}_t - \mathbf{v}^*)\|^2 - \|\boldsymbol{\eta}_t^{-1/2} \odot (\mathbf{w}_{t+1} - \mathbf{v}^*)\|^2\right\} \\
&\quad + \frac{1}{2(1-\mu_t)}\|\boldsymbol{\eta}_t^{1/2} \odot \bar{\mathbf{m}}_t\|^2 - \frac{\mu_{t+1}}{1-\mu_t}\langle\hat{\mathbf{m}}_t, \mathbf{w}_t - \mathbf{v}^*\rangle \\
&\le \frac{1}{2(1-\mu_t)}\left\{\|\boldsymbol{\eta}_t^{-1/2} \odot (\mathbf{w}_t - \mathbf{v}^*)\|^2 - \|\boldsymbol{\eta}_t^{-1/2} \odot (\mathbf{w}_{t+1} - \mathbf{v}^*)\|^2\right\} \\
&\quad + \frac{1}{2(1-\mu_t)}\|\boldsymbol{\eta}_t^{1/2} \odot \bar{\mathbf{m}}_t\|^2 + \frac{\mu_{t+1}}{2(1-\mu_t)}\|\boldsymbol{\eta}_t^{1/2} \odot \hat{\mathbf{m}}_t\|^2 + \frac{\mu_{t+1}}{2(1-\mu_t)}\|\boldsymbol{\eta}_t^{-1/2} \odot (\mathbf{w}_t - \mathbf{v}^*)\|^2 \qquad (3)
\end{aligned}$$

となる．最後の不等号は右辺第 4 項にヤングの不等式を用いた．そして，

$$\begin{aligned}\sum_{t=1}^{T}(E_t(\mathbf{w}_t)-E_t(\mathbf{v}^*)) &\le \sum_{t=1}^{T}\langle \mathbf{g}_t, \mathbf{w}_t-\mathbf{v}^*\rangle \\ &= \sum_{t=1}^{T}\langle\Big(1-\prod_{k=1}^{t}\mu_k\Big)\hat{\mathbf{g}}_t, \mathbf{w}_t-\mathbf{v}^*\rangle \\ &\le \sum_{t=1}^{T}\Bigg[\frac{1}{2(1-\mu_t)}\Big\{\|\boldsymbol{\eta}_t^{-1/2}\odot(\mathbf{w}_t-\mathbf{v}^*)\|^2-\|\boldsymbol{\eta}_t^{-1/2}\odot(\mathbf{w}_{t+1}-\mathbf{v}^*)\|^2\Big\}+\frac{1}{2(1-\mu_t)}\|\boldsymbol{\eta}_t^{1/2}\odot\bar{\mathbf{m}}_t\|^2 \\ &\quad+\frac{\mu_{t+1}}{2(1-\mu_t)}\|\boldsymbol{\eta}_t^{1/2}\odot\hat{\mathbf{m}}_t\|^2+\frac{\mu_{t+1}}{2(1-\mu_t)}\|\boldsymbol{\eta}_t^{-1/2}\odot(\mathbf{w}_t-\mathbf{v}^*)\|^2+\frac{\prod_{k=1}^{t}\mu_k}{2(1-\mu_t)}\|\boldsymbol{\eta}_t^{-1/2}\odot(\mathbf{w}_{t+1}-\mathbf{v}^*)\|^2\Bigg]\end{aligned} \tag{4}$$

となる．まず式（4）の右辺第 1 項を評価する．η_t の定義と帰納法を用いることで $L \le \sqrt{t}\|\boldsymbol{\eta}_t\|_\infty \le R$ となる．また $0 \le \mu_t \le \mu_1 < 1$ と $\eta_{t,i}^{-1} \ge \eta_{t-1,i}^{-1}$ に注意して変形を行うと，

$$\begin{aligned}&\sum_{t=1}^{T}\frac{1}{2(1-\mu_t)}\Big\{\|\boldsymbol{\eta}_t^{-1/2}\odot(\mathbf{w}_t-\mathbf{v}^*)\|^2-\|\boldsymbol{\eta}_t^{-1/2}\odot(\mathbf{w}_{t+1}-\mathbf{v}^*)\|^2\Big\} \\ &=\frac{1-\mu_1}{2(1-\mu_1)}\|\boldsymbol{\eta}_1^{-1/2}\odot(\mathbf{w}_1-\mathbf{v}^*)\|^2+\sum_{t=2}^{T}\Big\{\frac{1}{2(1-\mu_t)}\|\boldsymbol{\eta}_t^{-1/2}\odot(\mathbf{w}_t-\mathbf{v}^*)\|^2-\frac{1}{2(1-\mu_{t-1})}\|\boldsymbol{\eta}_{t-1}^{-1/2}\odot(\mathbf{w}_t-\mathbf{v}^*)\|^2\Big\} \\ &\quad-\frac{1}{2(1-\mu_T)}\|\boldsymbol{\eta}_T^{-1/2}\odot(\mathbf{w}_{T+1}-\mathbf{v}^*)\|^2 \\ &\le\frac{1}{2}\|\boldsymbol{\eta}_1^{-1/2}\odot(\mathbf{w}_1-\mathbf{v}^*)\|^2+\sum_{t=2}^{T}\Big\{\frac{1}{2(1-\mu_t)}\|\boldsymbol{\eta}_t^{-1/2}\odot(\mathbf{w}_t-\mathbf{v}^*)\|^2-\frac{1}{2(1-\mu_{t-1})}\|\boldsymbol{\eta}_{t-1}^{-1/2}\odot(\mathbf{w}_t-\mathbf{v}^*)\|^2\Big\} \\ &\le\frac{1}{2}\|\boldsymbol{\eta}_1^{-1/2}\odot(\mathbf{w}_1-\mathbf{v}^*)\|^2+\sum_{t=2}^{T}\Big\{\frac{1}{2(1-\mu_t)}\|\boldsymbol{\eta}_t^{-1/2}\odot(\mathbf{w}_t-\mathbf{v}^*)\|^2-\frac{1}{2(1-\mu_t)}\|\boldsymbol{\eta}_{t-1}^{-1/2}\odot(\mathbf{w}_t-\mathbf{v}^*)\|^2\Big\} \\ &\le\frac{1}{2}\|\boldsymbol{\eta}_1^{-1/2}\odot(\mathbf{w}_1-\mathbf{v}^*)\|^2+\sum_{t=2}^{T}\frac{1}{2(1-\mu_t)}\Big\{\|\boldsymbol{\eta}_t^{-1/2}\odot(\mathbf{w}_t-\mathbf{v}^*)\|^2-\|\boldsymbol{\eta}_{t-1}^{-1/2}\odot(\mathbf{w}_t-\mathbf{v}^*)\|^2\Big\} \\ &\le\frac{1}{2}\sum_{i=1}^{d}\eta_{1,i}^{-1}(w_{1,i}-v_i^*)^2+\sum_{t=2}^{T}\sum_{i=1}^{d}\frac{1}{2(1-\mu_t)}(\eta_{t,i}^{-1}-\eta_{t-1,i}^{-1})(w_{t,i}-v_i^*)^2 \\ &\le\frac{D_\infty^2 d}{2L}+\frac{D_\infty^2}{2(1-\mu_1)}\sum_{t=2}^{T}\sum_{i=1}^{d}(\eta_{t,i}^{-1}-\eta_{t-1,i}^{-1})\end{aligned} \tag{5}$$

と表せる．次に式 (4) の右辺第 4 項を評価する．

$$\begin{aligned}\sum_{t=1}^{T}\frac{\mu_{t+1}}{2(1-\mu_t)}\|\boldsymbol{\eta}_t^{-1/2}\odot(\mathbf{w}_t-\mathbf{v}^*)\|^2 &= \sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\mu_{t+1}}{2(1-\mu_t)}\eta_{t,i}^{-1}(w_{t,i}-v_i^*)^2 \\ &\le \frac{D_\infty^2 d}{2(1-\mu_1)L}\sum_{t=1}^{T}\mu_{t+1}\sqrt{t}.\end{aligned} \tag{6}$$

次に式（4）の右辺第 2 項を評価する．

$$\begin{aligned}
\sum_{t=1}^{T}\frac{1}{2(1-\mu_t)}\|\boldsymbol{\eta}_t^{1/2}\odot\bar{\mathbf{m}}_t\|^2 &= \sum_{t=1}^{T}\frac{1}{2(1-\mu_t)}\|\boldsymbol{\eta}_t^{1/2}\odot\{(1-\mu_t)\hat{\mathbf{g}}_t+\mu_{t+1}\hat{\mathbf{m}}_t\}\|^2\\
&= \sum_{t=1}^{T}\frac{1}{2(1-\mu_t)}\|(1-\mu_t)(\boldsymbol{\eta}_t^{1/2}\odot\hat{\mathbf{g}}_t)+\mu_{t+1}(\boldsymbol{\eta}_t^{1/2}\odot\hat{\mathbf{m}}_t)\|^2\\
&= \sum_{t=1}^{T}\frac{1}{2(1-\mu_t)}\Big\{(1-\mu_t)^2\|\boldsymbol{\eta}_t^{1/2}\odot\hat{\mathbf{g}}_t\|^2+2(1-\mu_t)\mu_{t+1}\langle\boldsymbol{\eta}_t^{1/2}\odot\hat{\mathbf{g}}_t,\boldsymbol{\eta}_t^{1/2}\odot\hat{\mathbf{m}}_t\rangle\\
&\quad+\mu_{t+1}^2\|\boldsymbol{\eta}_t^{1/2}\odot\hat{\mathbf{m}}_t\|^2\Big\}\\
&= \sum_{t=1}^{T}\frac{1-\mu_t}{2}\|\boldsymbol{\eta}_t^{1/2}\odot\hat{\mathbf{g}}_t\|^2+\sum_{t=1}^{T}\mu_{t+1}\langle\boldsymbol{\eta}_t,\hat{\mathbf{g}}_t\odot\hat{\mathbf{m}}_t\rangle\\
&\quad+\sum_{t=1}^{T}\frac{\mu_{t+1}^2}{2(1-\mu_t)}\|\boldsymbol{\eta}_t^{1/2}\odot\hat{\mathbf{m}}_t\|^2. \qquad (7)
\end{aligned}$$

式 (7) と，式 (4) の右辺第 3 項をまとめて，

$$\begin{aligned}
&\sum_{t=1}^{T}\Big\{\frac{1}{2(1-\mu_t)}\|\boldsymbol{\eta}_t^{1/2}\odot\bar{\mathbf{m}}_t\|^2+\frac{\mu_{t+1}}{2(1-\mu_t)}\|\boldsymbol{\eta}_t^{1/2}\odot\hat{\mathbf{m}}_t\|^2\Big\}\\
\le&\sum_{t=1}^{T}\frac{1-\mu_t}{2}\|\boldsymbol{\eta}_t^{1/2}\odot\hat{\mathbf{g}}_t\|^2+\sum_{t=1}^{T}\mu_{t+1}\|\boldsymbol{\eta}_t\|\cdot\|\hat{\mathbf{g}}_t\odot\hat{\mathbf{m}}_t\|\\
&+\sum_{t=1}^{T}\frac{\mu_{t+1}^2+\mu_{t+1}}{2(1-\mu_t)}\|\boldsymbol{\eta}_t^{1/2}\odot\hat{\mathbf{m}}_t\|^2. \qquad (8)
\end{aligned}$$

ここで仮定と，$\|\mathbf{m}_t\|=\|\beta_{1t}\mathbf{m}_{t-1}+(1-\beta_{1t})\mathbf{g}_t\|\le\beta_{1t}\|\mathbf{m}_{t-1}\|+(1-\beta_{1t})\|\mathbf{g}_t\|$ より帰納法を用いると，$t\in\{1,2,\ldots,T\}$ において $\|\mathbf{m}_t\|\le G_2$ が成り立つことがわかる．実際に $t=1$ のとき，

$$\|\mathbf{m}_1\|\le\beta_1\|\mathbf{m}_0\|+(1-\beta_1)\|\mathbf{g}_1\|=\|\mathbf{g}_1\|\le G_2 \qquad (9)$$

となる．$t=n$ のとき $\|\mathbf{m}_n\|\le G_2$ が成り立つとすると，$t=n+1$ のときは

$$\begin{aligned}
\|\mathbf{m}_{n+1}\|&\le\beta_{1(n+1)}\|\mathbf{m}_n\|+(1-\beta_{1(n+1)})\|\mathbf{g}_{n+1}\|\\
&\le\beta_{1(n+1)}G_2+(1-\beta_{1(n+1)})G_2=G_2 \qquad (10)
\end{aligned}$$

である．よって $t\in\{1,2,\ldots,T\}$ において $\|\mathbf{m}_t\|\le G_2$. これより $\|\hat{\mathbf{m}}_t\|\le\frac{G_2}{1-\prod_{k=1}^{t+1}\mu_k}$，$\|\hat{\mathbf{g}}_t\|\le\frac{G_2}{1-\prod_{k=1}^{t}\mu_k}$ である．また $\|\boldsymbol{\eta}_t\|$ の定義から $\|\boldsymbol{\eta}_t\|\le\frac{R\sqrt{d}}{\sqrt{t}}$，そして $\forall\mathbf{x},\mathbf{y}\in\mathbb{R}^d$ において

$$\|\mathbf{x}\odot\mathbf{y}\|^2=\sum_{i=1}^{d}x_i^2y_i^2\le\Big(\max y_i^2\Big)\sum_{i=1}^{d}x_i^2\le\sqrt{\sum_{i=1}^{d}y_i^4}\cdot\|\mathbf{x}\|^2=\|\mathbf{y}\odot\mathbf{y}\|\cdot\|\mathbf{x}\|^2 \qquad (11)$$

となる．また $\|\hat{\mathbf{g}}_t \odot \hat{\mathbf{m}}_t\|^2 \le \|\hat{\mathbf{g}}_t\|^2 \|\hat{\mathbf{m}}_t\|^2$ である．よって式 (8) を次のように変形する．

$$\begin{aligned}
(\text{与式}) &\le \sum_{t=1}^{T} \frac{1-\mu_t}{2} \|\boldsymbol{\eta}_t\| \cdot \|\hat{\mathbf{g}}_t\|^2 + \sum_{t=1}^{T} \mu_{t+1} \|\boldsymbol{\eta}_t\| \cdot \|\hat{\mathbf{g}}_t\| \cdot \|\hat{\mathbf{m}}_t\| \\
&\quad + \sum_{t=1}^{T} \frac{\mu_{t+1}^2 + \mu_{t+1}}{2(1-\mu_t)} \|\boldsymbol{\eta}_t\| \cdot \|\hat{\mathbf{m}}_t\|^2 \\
&\le \frac{RG_2^2\sqrt{d}}{2} \sum_{t=1}^{T} \frac{1-\mu_t}{\sqrt{t}(1-\prod_{k=1}^{t}\mu_k)^2} + RG_2^2\sqrt{d} \sum_{t=1}^{T} \frac{\mu_{t+1}}{\sqrt{t}(1-\prod_{k=1}^{t}\mu_k)(1-\prod_{k=1}^{t+1}\mu_k)} \\
&\quad + \frac{RG_2^2\sqrt{d}}{2} \sum_{t=1}^{T} \frac{\mu_{t+1}^2 + \mu_{t+1}}{\sqrt{t}(1-\mu_t)(1-\prod_{k=1}^{t+1}\mu_k)^2}. \qquad (12)
\end{aligned}$$

最後に式（4）の右辺第 5 項を評価する．

$$\begin{aligned}
\sum_{t=1}^{T} \frac{\prod_{k=1}^{t}\mu_k}{2(1-\mu_t)} \|\boldsymbol{\eta}_t^{-1/2} \odot (\mathbf{w}_{t+1} - \mathbf{v}^*)\|^2 &\le \frac{1}{2(1-\mu_1)} \sum_{t=1}^{T} \prod_{k=1}^{t} \mu_k \|\boldsymbol{\eta}_t^{-1/2} \odot (\mathbf{w}_{t+1} - \mathbf{v}^*)\|^2 \\
&= \frac{1}{2(1-\mu_1)} \sum_{t=1}^{T} \sum_{i=1}^{d} \prod_{k=1}^{t} \mu_k \eta_{t,i}^{-1} (w_{t+1,i} - v_i^*)^2 \\
&\le \frac{D_\infty^2}{2(1-\mu_1)} \sum_{t=1}^{T} \sum_{i=1}^{d} \mu_t \eta_{t,i}^{-1} \\
&\le \frac{D_\infty^2 d}{2(1-\mu_1)L} \sum_{t=1}^{T} \mu_t \sqrt{t}. \qquad (13)
\end{aligned}$$

以上の結果をまとめて，次の最終的な不等式が得られる．

$$\begin{aligned}
R_T &\le \frac{D_\infty^2 d}{2L} + \frac{D_\infty^2}{2(1-\mu_1)} \sum_{t=2}^{T} \sum_{i=1}^{d} (\eta_{t,i}^{-1} - \eta_{t-1,i}^{-1}) + \frac{RG_2^2\sqrt{d}}{2} \sum_{t=1}^{T} \frac{1-\mu_t}{\sqrt{t}(1-\prod_{k=1}^{t}\mu_k)^2} \\
&\quad + RG_2^2\sqrt{d} \sum_{t=1}^{T} \frac{\mu_{t+1}}{\sqrt{t}(1-\prod_{k=1}^{t}\mu_k)(1-\prod_{k=1}^{t+1}\mu_k)} + \frac{RG_2^2\sqrt{d}}{2} \sum_{t=1}^{T} \frac{\mu_{t+1}^2 + \mu_{t+1}}{\sqrt{t}(1-\mu_t)(1-\prod_{k=1}^{t+1}\mu_k)^2} \\
&\quad + \frac{D_\infty^2 d}{2(1-\mu_1)L} \sum_{t=1}^{T} \mu_{t+1}\sqrt{t} + \frac{D_\infty^2 d}{2(1-\mu_1)L} \sum_{t=1}^{T} \mu_t \sqrt{t}. \qquad (14)
\end{aligned}$$

□

系 4. 定理 3 において $\mu_t = \beta\lambda^{t-1}$ とする ($0 \le \beta < 1$, $0 \le \lambda < 1$)．このとき次の不等式が成立する．

$$\begin{aligned}
R_T &\le \frac{D_\infty^2 d}{2L} + \frac{D_\infty^2 d\sqrt{T}}{2(1-\beta)L} + \frac{RG_2^2\sqrt{d}}{2(1-\beta)^2}(2\sqrt{T} - 1) + \frac{\beta\lambda RG_2^2\sqrt{d}}{(1-\beta)^2(1-\lambda)} \\
&\quad + \frac{\beta\lambda RG_2^2\sqrt{d}}{(1-\beta)^3(1-\lambda)} + \frac{\beta(1+\lambda)D_\infty^2 d}{2(1-\beta)(1-\lambda)^2 L}.
\end{aligned}$$

証明. まず，式 (14) の右辺第 1 項，第 2 項を整理する．

$$\begin{aligned}
&\frac{D_\infty^2 d}{2L} + \frac{D_\infty^2}{2(1-\mu_1)} \sum_{t=2}^{T} \sum_{i=1}^{d} (\eta_{t,i}^{-1} - \eta_{t-1,i}^{-1}) \\
&= \frac{D_\infty^2 d}{2L} + \frac{D_\infty^2}{2(1-\beta)} \sum_{i=1}^{d} (\eta_{T,i}^{-1} - \eta_{1,i}^{-1}) \\
&\leq \frac{D_\infty^2 d}{2L} + \frac{D_\infty^2}{2(1-\beta)} \sum_{i=1}^{d} \eta_{T,i}^{-1} \\
&\leq \frac{D_\infty^2 d}{2L} + \frac{D_\infty^2 d\sqrt{T}}{2(1-\beta)L}.
\end{aligned} \tag{15}$$

次に式 (14) の右辺第 3 項について，

$$\begin{aligned}
\frac{RG_2^2\sqrt{d}}{2} \sum_{t=1}^{T} \frac{1-\mu_t}{\sqrt{t}(1-\prod_{k=1}^{t}\mu_k)^2} &\leq \frac{RG_2^2\sqrt{d}}{2} \sum_{t=1}^{T} \frac{1}{\sqrt{t}(1-\prod_{k=1}^{t}\mu_k)^2} \\
&\leq \frac{RG_2^2\sqrt{d}}{2} \sum_{t=1}^{T} \frac{1}{\sqrt{t}(1-\mu_1)^2} \quad \left(\because \frac{1}{1-\prod_{k=1}^{t}\mu_k} \leq \frac{1}{1-\mu_1}\right) \\
&= \frac{RG_2^2\sqrt{d}}{2(1-\beta)^2} \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \\
&\leq \frac{RG_2^2\sqrt{d}}{2(1-\beta)^2}(2\sqrt{T}-1).
\end{aligned} \tag{16}$$

次に式 (14) の右辺第 4 項について，

$$\begin{aligned}
RG_2^2\sqrt{d} \sum_{t=1}^{T} \frac{\mu_{t+1}}{\sqrt{t}(1-\prod_{k=1}^{t}\mu_k)(1-\prod_{k=1}^{t+1}\mu_k)} &\leq RG_2^2\sqrt{d} \sum_{t=1}^{T} \frac{\mu_{t+1}}{\sqrt{t}(1-\mu_1)^2} \quad \left(\because \frac{1}{1-\prod_{k=1}^{t+1}\mu_k} \leq \frac{1}{1-\mu_1}\right) \\
&= \frac{\beta RG_2^2\sqrt{d}}{(1-\beta)^2} \sum_{t=1}^{T} \frac{\lambda^t}{\sqrt{t}} \\
&\leq \frac{\beta RG_2^2\sqrt{d}}{(1-\beta)^2} \sum_{t=1}^{T} \lambda^t \\
&\leq \frac{\beta\lambda RG_2^2\sqrt{d}}{(1-\beta)^2(1-\lambda)}.
\end{aligned} \tag{17}$$

次に式 (14) の右辺第 5 項について，

$$\begin{aligned}
\frac{RG_2^2\sqrt{d}}{2} \sum_{t=1}^{T} \frac{\mu_{t+1}^2+\mu_{t+1}}{\sqrt{t}(1-\mu_t)(1-\prod_{k=1}^{t+1}\mu_k)^2} &\leq \frac{RG_2^2\sqrt{d}}{2} \sum_{t=1}^{T} \frac{\mu_{t+1}^2+\mu_{t+1}}{\sqrt{t}(1-\mu_1)^3} \quad \left(\because \frac{1}{1-\mu_t} \leq \frac{1}{1-\mu_1}\right) \\
&\leq \frac{RG_2^2\sqrt{d}}{2(1-\beta)^3} \sum_{t=1}^{T} \frac{2\mu_{t+1}}{\sqrt{t}} \\
&= \frac{RG_2^2\sqrt{d}}{(1-\beta)^3} \sum_{t=1}^{T} \frac{\beta\lambda^t}{\sqrt{t}} \\
&\leq \frac{\beta\lambda RG_2^2\sqrt{d}}{(1-\beta)^3(1-\lambda)}.
\end{aligned} \tag{18}$$

次に式 (14) の右辺第 6 項，第 7 項について，

$$\begin{aligned}\frac{D_\infty^2 d}{2(1-\mu_1)L}\sum_{t=1}^{T}\mu_{t+1}\sqrt{t}+\frac{D_\infty^2 d}{2(1-\mu_1)L}\sum_{t=1}^{T}\mu_t\sqrt{t} &\le \frac{D_\infty^2 d}{2(1-\mu_1)L}\Big(\sum_{t=1}^{T}\mu_{t+1}\cdot t+\sum_{t=1}^{T}\mu_t\cdot t\Big)\\ &= \frac{\beta D_\infty^2 d}{2(1-\beta)L}\Big(\sum_{t=1}^{T}\lambda^t\cdot t+\sum_{t=1}^{T}\lambda^{t-1}\cdot t\Big)\\ &\le \frac{\beta(1+\lambda)D_\infty^2 d}{2(1-\beta)(1-\lambda)^2 L}. \end{aligned} \tag{19}$$

よって式 (15) から式 (19) を式 (14) に適用すると，次の不等式が得られる．

$$\begin{aligned}R_T \le{}& \frac{D_\infty^2 d}{2L}+\frac{D_\infty^2 d\sqrt{T}}{2(1-\beta)L}+\frac{RG_2^2\sqrt{d}}{2(1-\beta)^2}(2\sqrt{T}-1)+\frac{\beta\lambda RG_2^2\sqrt{d}}{(1-\beta)^2(1-\lambda)}\\ &+\frac{\beta\lambda RG_2^2\sqrt{d}}{(1-\beta)^3(1-\lambda)}+\frac{\beta(1+\lambda)D_\infty^2 d}{2(1-\beta)(1-\lambda)^2 L}. \end{aligned} \tag{20}$$

□

系 4 より，μ_t が適当な値をとるとき R_T は $O(\sqrt{T})$ で抑えられることがわかる．ゆえに $T\to\infty$ のとき，$R_T/T\to 0$ に近づく．また，$\mu_t=\beta/t(0\le\beta<1)$ のときも $R_T/T\to 0(T\to\infty)$ である．

3　数値実験

本節では，画像データを用いて本論文で提案する学習アルゴリズムと他の学習アルゴリズムの性能を比較する．実験の目的は，提案するアルゴリズムは学習の初期エポックにおいて画像に写る物体を正しく認識する精度がより高く，そして学習を加速させることができるかである．ここで精度 (単位 : %) は全てのテスト画像に対して，テスト画像に写る物体が正しく認識される数の割合を表す．

本研究では AdaBound [5] と同様の設定で数値実験を行う．実験に用いるネットワークモデルとデータセットの対応は次の表の通りである．MNIST データに対しては，恒等関数を活性化関数とした 2

データセット	モデル	層構造
MNIST	Feedforward	2 層順伝播
CIFAR-10	ResNet34	34 層畳み込み
CIFAR-10	DenseNet121	121 層畳み込み

層順伝播型ニューラルネットワークに用いる．MNIST データは分類難易度が他のデータを比較しても難しくないため，シンプルなネットワークで比較を行う．CIFAR-10 データに対しては，深層畳み込みニューラルネットワークの ResNet34[2] と DenseNet121[3] を用いる．

実験に用いる学習アルゴリズムは，比較的よく用いられるモーメンタム法と Adam，NadaBound のもととなった AdaBound と Nadam, そして提案手法の NadaBound である．

(1) モーメンタム法

学習率を 0.01，モーメンタムパラメータを 0.9 とする．

(2)Adam

学習率を 0.001，誤差定数を 10^{-8}，2 種類の指数減衰率の値をそれぞれ 0.9，0.999 とする．

(3)AdaBound

学習率を $\eta_t = 0.001$，誤差定数を $\varepsilon = 10^{-8}$，2 種類の指数減衰率の値をそれぞれ $\rho_1 = 0.9$，$\rho_2 = 0.999$，$\alpha = 0.5$，時刻 t で学習率の値がとり得る範囲を決める関数を $\eta_l(t) = 1 - \frac{1}{(1-\rho_2)t+1}$，$\eta_u(t) = 1 + \frac{1}{(1-\rho_2)t}$ とする．

(4)Nadam

MNIST データセットの実験は学習率を $\eta_t = 0.002$，CIFAR-10 の実験は学習率を $\eta_t = 0.001$ とする．また誤差定数を $\varepsilon = 10^{-8}$，2 種類の指数減衰率の値をそれぞれ $\mu = 0.99$，$\beta = 0.999$，モーメンタムパラメータを $\mu_t = \mu(1 - 0.5 \times 0.96^{(t/250)})$ とする．

(5)NadaBound

MNIST データセットの実験は学習率を $\eta_t = 0.002$，CIFAR-10 の実験は学習率を $\eta_t = 0.001$ とする．また誤差定数を $\varepsilon = 10^{-8}$，2 種類の指数減衰率の値をそれぞれ $\mu = 0.99$，$\beta = 0.999$，モーメンタムパラメータを $\mu_t = \mu(1 - 0.5 \times 0.96^{(t/250)})$，$\alpha = 0.5$，時刻 t で学習率の値がとり得る範囲を決める関数を $\eta_l(t) = 1 - \frac{1}{(1-\rho_2)t+1}$，$\eta_u(t) = 1 + \frac{1}{(1-\rho_2)t}$ とする．

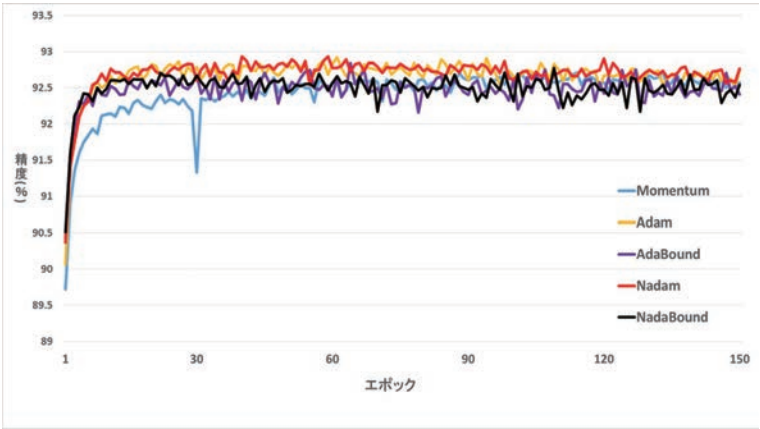
各表とも平均値と標準偏差は有効数字 6 桁で表す．

最初に 2 層順伝播型ニューラルネットワークによる MNIST データセットの比較結果をみる．学習の最初の 9 エポックで精度を比較したデータが表 1 である．また，150 エポック時の精度の比較した結果については精度を表すグラフは図 1a，テストデータの誤差のグラフは図 1b にまとめている．初期 9 エポックの精度の平均値と標準偏差については 100 回の試行から算出している．

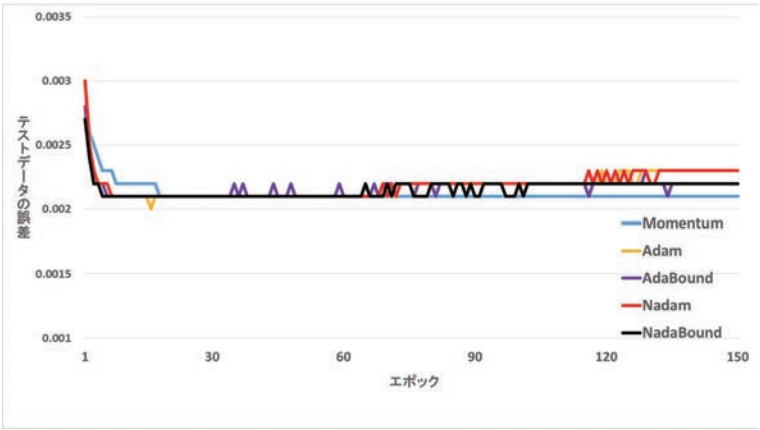
表 1: 順伝播型ニューラルネットワークによる MNIST データセットの学習結果 (初期の 9 エポックにおける平均値と標準偏差)

項目	アルゴリズム	エポック 1	エポック 2	エポック 3	エポック 4	エポック 5	エポック 6	エポック 7	エポック 8	エポック 9
平均値	Momentum	89.9402	90.883	91.2982	91.6782	91.6784	91.8318	91.9044	91.9694	92.0288
	Adam	90.6762	91.317	91.7828	92.1018	92.2732	92.4006	92.4478	92.4828	92.5536
	AdaBound	90.5398	91.6074	91.978	92.1748	92.2732	92.3082	92.3598	92.391	92.4386
	Nadam	91.0898	91.897	92.2132	92.3864	92.499	92.554	92.6112	92.6398	92.659
	NadaBound	91.179	92.011	92.2816	92.427	92.4866	92.513	92.5324	92.5664	92.5786
標準偏差	Momentum	0.11076	0.08193	0.08255	0.71409	0.08516	0.15097	0.07422	0.07836	0.1641
	Adam	0.11534	0.08552	0.08099	0.07143	0.0646	0.07732	0.0808	0.0718	0.0737
	AdaBound	0.18317	0.13452	0.17953	0.13432	0.11453	0.10693	0.10747	0.10397	0.09755
	Nadam	0.33841	0.23789	0.20563	0.14069	0.12007	0.11562	0.11613	0.12633	0.10253
	NadaBound	0.1108	0.19036	0.08474	0.07242	0.08817	0.07627	0.08769	0.09061	0.08117

次に ResNet34 で CIFAR-10 データセットの分類を行った結果をみる．学習の最初の 8 エポックで精度を比較した結果が表 2 である．また，150 エポック時の精度を比較したグラフは図 2a，テストデータの誤差のグラフは図 2b にまとめている．初期 8 エポックの精度の平均値と標準偏差については 100 回の試行から算出している．



(a) 縦軸は精度 (%)，横軸はエポックの経過をみる．



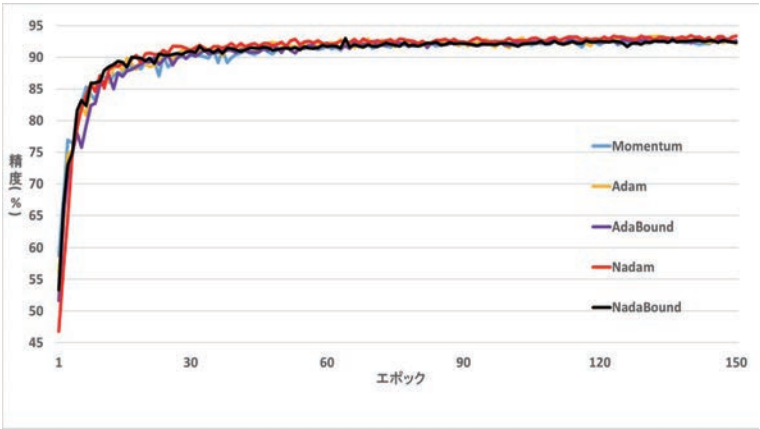
(b) 縦軸は誤差の値，横軸はエポックの経過をみる．

図 1: 順伝播型ニューラルネットで MNIST を 150 エポック学習したときの推移．

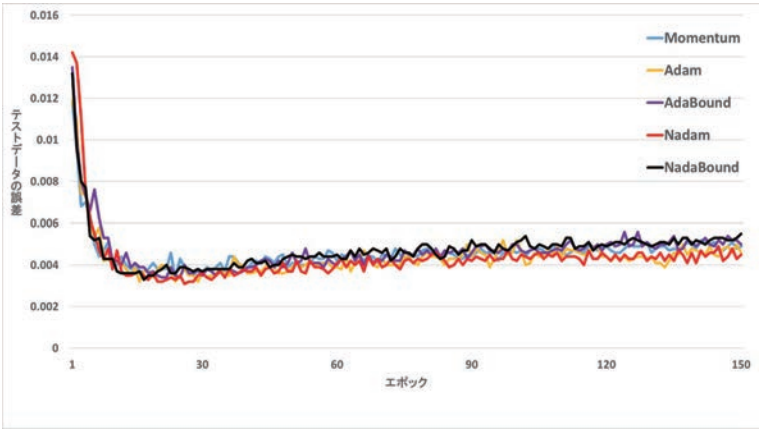
表 2: ResNet34 による CIFAR-10 データセットの学習結果 (初期の 8 エポックにおける平均値と標準偏差)

項目	アルゴリズム	エポック 1	エポック 2	エポック 3	エポック 4	エポック 5	エポック 6	エポック 7	エポック 8
平均値	Momentum	54.2002	67.7486	73.4902	76.798	79.986	81.1634	82.2946	83.583
	Adam	51.746	62.6834	70.7036	75.5018	78.9872	81.3542	82.777	83.6548
	AdaBound	49.9154	61.695	69.239	74.5325	76.54	79.4396	80.8868	82.4528
	Nadam	45.9746	58.259	68.4632	73.9398	77.4436	79.8898	81.9932	83.4622
	NadaBound	53.5402	65.8406	73.0188	77.4676	80.4962	82.6564	84.3186	85.4458
標準偏差	Momentum	3.015063	2.357754	2.490147	2.345702	1.479258	1.614838	1.445465	1.141825
	Adam	3.64085	3.30318	2.60535	2.45696	2.09091	2.28848	1.83517	1.51996
	AdaBound	3.988265	3.686561	2.339456	2.093196	1.857805	1.689745	1.588919	1.36749
	Nadam	3.914669	5.030281	3.537677	3.088551	2.156569	2.153107	1.668854	1.715514
	NadaBound	2.8803	2.68067	2.14737	1.77233	1.63726	1.53185	0.9862	0.85448

最後に DenseNet121 で CIFAR-10 データセットの分類を行った結果をみる．学習の最初の 6 エポックで精度を比較した結果が表 3 である．また，120 エポック時の精度を比較したグラフは図 3a，テストデータの誤差のグラフが図 3b にまとめている．初期 6 エポックの精度の平均値と標準偏差については 100 回の試行から算出している．



(a) 縦軸は精度 (%)，横軸はエポックの経過をみる．



(b) 縦軸は誤差の値，横軸はエポックの経過をみる．

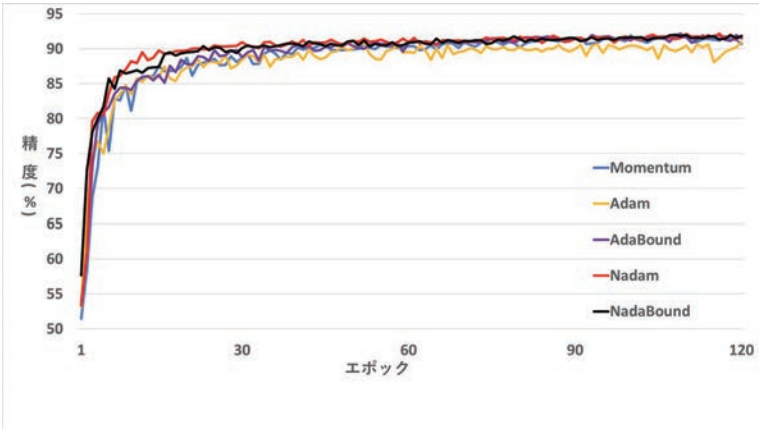
図 2: ResNet34 で CIFAR-10 を 150 エポック学習したときの推移．

表 3: DenseNet121 による CIFAR-10 データセットの学習結果 (初期の 6 エポックにおける平均値と標準偏差)

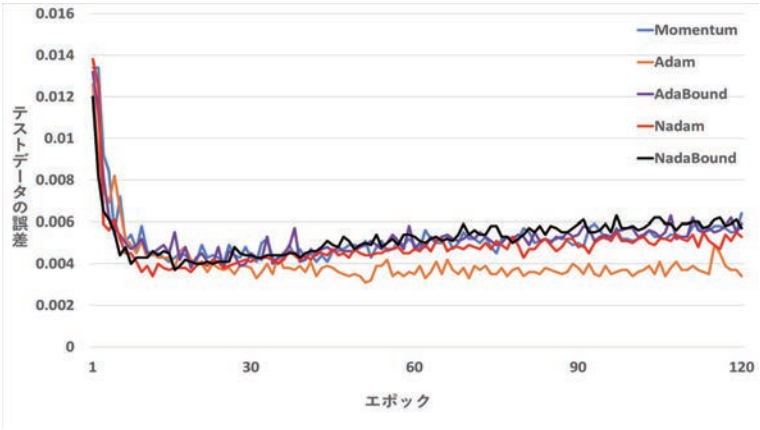
項目	アルゴリズム	エポック 1	エポック 2	エポック 3	エポック 4	エポック 5	エポック 6
平均値	Momentum	49.7852	63.441	69.8922	74.7414	77.6538	79.9406
	Adam	52.6759	65.4319	72.8889	76.9925	79.1053	80.4247
	AdaBound	52.1158	65.2138	71.5288	75.6066	78.8854	80.5911
	Nadam	53.6842	66.1861	74.8537	79.296	81.7516	83.4151
	NadaBound	55.5983	68.6095	75.8373	79.4268	82.2439	83.4347
標準偏差	Momentum	2.53328	2.3279	3.064	2.24995	1.76853	1.69707
	Adam	2.69627	2.83518	3.31527	1.71659	1.86123	1.86675
	AdaBound	3.05121	2.14528	2.30914	1.84658	1.65697	1.25202
	Nadam	3.85412	4.46478	3.3341	2.30415	2.20031	1.79243
	NadaBound	2.27815	1.96479	1.62435	1.61664	1.11524	0.9882

各表から，NadaBound は他の手法に劣らない精度を出し，かつ標準偏差が徐々に小さくなることが分かる．各ネットワークモデルについて結果を分析する．

2 層順伝播型ネットワークでは各エポックにおける精度の平均値が他の手法と同等かそれ以上の数値を出している．また，各エポックの標準偏差も振れ幅が小さく，NadaBound は学習の初期から安定して高い精度を実現している．ResNet34 では，エポックによって他の手法より見劣りする数値もある



(a) 縦軸は精度 (%)，横軸はエポックの経過をみる．



(b) 縦軸は誤差の値，横軸はエポックの経過をみる．

図 3: DenseNet121 で CIFAR-10 を 120 エポック学習したときの推移．

が，十分な加速の効果を発揮してエポック 7，エポック 8 では平均値が他より高いことがわかる．標準偏差も徐々に小さくなっている．DenseNet121 では各エポックにおける精度の平均値が高く標準偏差も小さいことから，いずれの試行も安定して良い精度を高めていると考えられる．以上から，学習の初期段階では各ネットワークモデルにおおよそ順応して，加速による精度の高さや安定性を保持していると推測される．

次に，各図から NadaBound と他の手法の分析を行う．2 層順伝播ニューラルネットワークでは MNIST の学習自体は難しくなかったため，各手法で精度の高さには違いが少ない．Nadam と NadaBound は，学習の加速を行ったため，テスト誤差は学習途中から過学習を引き起こしたと考えられる．ResNet34 では，NadaBound は学習の初期から高い精度を維持しつつ最後まで大きくぶれることなく学習を終えている．DenseNet121 も ResNet34 と同様の結果となっている．

NadaBound は，各ネットワークモデルにおいて学習後半のエポックにおける精度の平均値は他の手法より特別高い精度を実現していない．しかし NadaBound は学習を加速するアルゴリズム，Nadam と同様の加速を行っている．また，学習の初期エポックの精度に関する標準偏差が Nadam より小さいことから，学習の初めから Nadam よりも安定していることが今回の検証から考えることができる．安定して学習を行い，そして学習を加速させることは，DenseNet [3] や YOLO [8] 等の学習に時間がかかるような複雑なネットワークでも充分に効果があると考える．

また，今回の検証で学習の初期エポックにおける精度の標準偏差が小さい結果が得られた．その要因となったのが，学習率にクリッピング手法を適用したことではないかと考える．これは Adam と AdaBound，Nadam と NadaBound について精度の標準偏差を比較したことからである．学習率を動的に変化させ，かつ制限を加えることによって，学習率が大きすぎる値や小さすぎる値となることを防ぐため，学習が安定して進行したのではないかと考える．

4 まとめ

本論文では新たな学習アルゴリズムの NadaBound を提案し，凸最適化問題における収束性を示した．また，既存の学習アルゴリズムと NadaBound の性能を比較する計算機実験を行い，学習の初期エポックにおいて既存の学習アルゴリズムよりも良い精度を実現するという結果が得られた．数値実験の結果から，学習率にクリッピング手法を適用することでネットワークの学習が安定して進むと考えられる．

今回の数値実験で見られた学習率にクリッピング手法を適用して得られた学習の安定化に関しての数理的な解析が今後の課題である．

参考文献

[1] T. Dozat. Incorporating nesterov momentum into adam. *International Conference for Learning Representations*, 2016.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[3] G. Huang, Z. Liu, K. Weinberger, and L. Maaten. Densely connected convolutional networks. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[4] D. Kingma and J. Ba. Adam:A method for stochastic optimization. *In Proceedings of the 3rd International Conference on Learning Representations*, 2015.

[5] L. Lio, Y. Xiong, Y. Liu, and X. Sun. Adaptive gradient methods with dynamic bound of learning rate. *In Proceedings of the 7rd International Conference on Learning Representations*, 2019.

[6] H. Mcmahan and M. Streeter. Adaptive bound optimization for online convex optimization. *Proceedings of the 23rd Annual Conference on Learning Theory*, 2010.

[7] Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathmatics Doklady*, 27:372-376, 1983.

[8] J. Redmon, S. Divvala, R. Girshick and A. Farhadi. You only look once: Unified,real-time object detection. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[9] D. Silver and D. Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529:484-489, 2016.