Muonの収束と最適なバッチサイズの解析

会員　明治大学
Université de Montréal, Mila
会員　明治大学
*佐藤 尚樹　SATO Naoki
長沼 大樹　NAGANUMA Hiroki
飯塚 秀明　IIDUKA Hideaki

1. はじめに

画像分類や言語モデリングなどの深層学習のための最適化においては，長らくAdamをはじめとする適応的手法が事実上の標準手法であったが，近年提案された新しい最適化アルゴリズム：Muonは，収束速度と汎化性能の両方でそれらよりも優れていることが，多くの研究で実験的に観察されている．本発表は，Muonのアルゴリズムを紹介し，その収束解析と最適なバッチサイズの解析を行う．

2. Muon

Moment**U**m **O**rthogonalized by **N**ewton-Shulz (Muon) [1] は行列構造を持つパラメータに特化した最適化アルゴリズムで，次のように定義される．

アルゴリズム 1 最も基本的なMuon

Require: $\eta > 0, \beta \in [0,1), M_{-1} := \mathbf{0}, W_0 \in \mathbb{R}^{m\times n}$
  for $t = 0$ to $T-1$ **do**
    $M_t := \beta M_{t-1} + (1-\beta)\nabla f_{\mathcal{S}_t}(W_t)$
    $O_t := \mathrm{NewtonSchulz}(M_t)$
    $W_{t+1} := W_t - \eta O_t$
  end for
  return W_T

Muonの探索方向 $O_t \in \mathbb{R}^{m\times n}$ は，次のように勾配情報の行列を直交化して得られる．

$$O_t := \operatorname*{argmin}_{O:\mathbb{R}^{m\times n}:\|O\|_2\leq 1} \|O - M_t\|_F.$$

このとき M_t の特異値分解が $U_tS_tV_t^\top$ であるならば，$O_t = U_tV_t^\top$ が成り立つ．ただし，$U_t \in \mathbb{R}^{m\times r}, S_t \in \mathbb{R}^{r\times r}, V_t \in \mathbb{R}^{n\times r}$ で，$r > 0$ は行列 M_t のランクである．しかし特異値分解の計算は非常に高価であるため，MuonはNewton-Schulz反復法 [2, 3] を使って O_t を高速に近似している．

確率的勾配降下法やAdamが，深層学習モデルのパラメータ行列をベクトル化して取り扱うのに対して，Muonは行列のまま取り扱うため，パラメータ行列本来の構造的特性を損なわない．

より実用的には，次のように追加の慣性項と重み減衰を加えたMuonがよく利用されている．主

アルゴリズム 2 よく利用されるMuon

Require: $\eta, \lambda > 0, \beta \in [0,1), M_{-1} := \mathbf{0}, W_0 \in \mathbb{R}^{m\times n}$
  for $t = 0$ to $T-1$ **do**
    $M_t := \beta M_{t-1} + (1-\beta)\nabla f_{\mathcal{S}_t}(W_t)$
    $C_t := \beta M_t + (1-\beta)\nabla f_{\mathcal{S}_t}(W_t)$
    $O_t := \mathrm{NewtonSchulz}(C_t)$
    $W_{t+1} := (1-\eta\lambda)W_t - \eta O_t$
  end for
  return W_T

な実装では慣性係数 $\beta \in [0,1)$ は2つの慣性項で共有されている．また，$\lambda > 0$ は重み減衰係数である．これらはユーザーが自由に設定できるハイパーパラメータであり，$\beta = 0.95, \lambda = 0.06$ などがよく使われる．本発表はよく利用されるMuonの解析を行う．

3. 収束解析

本稿では特に重み減衰ありのMuonの優れた理論的性質を紹介する．以下の解析では，経験損失関数 $f\colon \mathbb{R}^{m\times n} \to \mathbb{R}$ が L-平滑であることを仮定している．

命題 1（点列の有界性）$\eta \leq \frac{1}{\lambda}$ を仮定する．点列 $(W_t)_{t\in\mathbb{N}}$ が *Muon* で生成されたとすると，

$$\|W_t\|_F \leq \begin{cases} (1-\eta\lambda)^t\|W_0\|_F + \frac{\sqrt{n}}{\lambda} & \text{if } \eta < \frac{1}{\lambda}, \\ \frac{\sqrt{n}}{\lambda} & \text{if } \eta = \frac{1}{\lambda}. \end{cases}$$

が成り立つ．

命題1は，重み減衰ありのMuonで生成された点列のノルムは確率1で有界で，時刻 t に対して

単調減少し，その上界は $t \to \infty$ で $\frac{\sqrt{n}}{\lambda}$ に収束することを意味している．特に，$\eta = \frac{1}{\lambda}$ が成り立つときその上界は最小となる．

命題 2（勾配の有界性）$\eta \leq \frac{1}{\lambda}$ を仮定する．点列 $(W_t)_{t\in\mathbb{N}}$ が *Muon* で生成されたとすると，

$$\|\nabla f(W_t)\|_F \leq \begin{cases} L(1-\eta\lambda)^t\|W_0\|_F + D & \text{if } \eta < \frac{1}{\lambda}, \\ D & \text{if } \eta = \frac{1}{\lambda}. \end{cases}$$

が成り立つ．ただし，$D := \frac{L}{\lambda} + L\|W^\star\|_F$ とする．

命題 2 は命題 1 から得られる結果で，勾配のノルムが確率 1 で有界であり，時刻 t に対して単調減少することを意味している．また同様に，特に $\eta = \frac{1}{\lambda}$ が成り立つときその上界が最小となる．これらの結果は，Muon の探索方向が直交化演算のおかげで常に $\|O_t\|_F = \sqrt{n}$ を満たすために成り立つもので，Muon の優れた理論的特性の一つであると言える．正確には探索方向のノルムが有界であれば任意の最適化アルゴリズムに対して同様の結果を得られるが，確率的勾配降下法や Adam に対しては成り立たない結果である．これらの命題を利用すると，次の定理を得られる．

定理 1（よく利用される *Muon* の収束解析）$\eta \leq \frac{1}{\lambda}$ を仮定する．点列 $(W_t)_{t\in\mathbb{N}}$ が *Muon* で生成されたとすると，

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(W_t)\|_F\right] \leq \mathcal{O}\left(\frac{1}{T} + \frac{(2\beta+1)(1-\beta)+\lambda}{2}\cdot\frac{1}{b} + n\right).$$

が成り立つ．ただし，b はバッチサイズである．

確率的勾配降下法や Adam などの標準的な収束解析では，$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(W_t)\|_F^2\right]$ がおよそ $1/T$ のオーダーで減少するが，定理 1 のように Muon は $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(W_t)\|_F\right]$ がおよそ $1/T$ のオーダーで減少する．この結果は Muon の高速な収束を保証するものである．

4. 最適なバッチサイズの解析

収束解析で得られた式を利用すると，訓練に必要な計算量を最小にする最適なバッチサイズの式を得られる．定理 1 から，正数 $X, Y, Z > 0$ に対して

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(W_t)\|_F^2\right] \leq \frac{X}{T} + \frac{Y}{b} + Z$$

が成り立つ．訓練が十分に完了したとき，ある閾値 $\epsilon > 0$ に対して

$$\frac{X}{T} + \frac{Y}{b} < \epsilon$$

が成り立つと仮定すると，この式を満たす総反復回数 T_{Muon} は

$$T_{\text{Muon}} = \frac{Xb}{\epsilon b - Y}$$

と表せる．Muon は 1 回の反復で b 個の確率的勾配を計算するので，T_{Muon} 回の反復では $T_{\text{Muon}}b$ 個の確率的勾配を計算する．これが訓練にかかる総計算量であり，$T_{\text{Muon}}b$ を最小にする b が計算量の意味で最適なバッチサイズとなる．以上より Muon の最適なバッチサイズは

$$b^\star_{\text{Muon}} = \frac{Y}{2} = \frac{\{(2\beta+1)(1-\beta)+\lambda\}\,\sigma^2}{\epsilon}$$

となり，適切なハイパーパラメータの設定のためにいくつかの知見を得られる．定理等の詳細な証明は [4] を参照されたい．

参考文献

[1] K. Jordan, Y. Jin, V. Boza, Y. Jiacheng, F. Cesista, L. Newhouse, and J. Bernstein, "Muon: An optimizer for hidden layers in neural networks," `https://kellerjordan.github.io/posts/muon/`, 2024.

[2] J. Bernstein and L. Newhouse, "Old optimizer, new norm: An anthology," `https://arxiv.org/abs/2409.20325`, 2024.

[3] Z. Kovarik, "Some iterative methods for improving orthonormality," SIAM Journal on Numerical Analysis, vol. 7, no. 3, pp. 386 – 389, 1970.

[4] N. Sato, H. Naganuma, and H. Iiduka, "Analysis of Muon's convergence and critical batch size," `https://arxiv.org/abs/2507.01598`, 2025.