Muonの収束と
最適なバッチサイズの解析

日本オペレーションズ・リサーチ学会 2025年秋季研究発表会
9/12(金) 15:00～15:20

明治大学　佐藤尚樹
モントリオール大学, Mila　長沼大樹
明治大学　飯塚秀明

背景：機械学習（画像分類）

入力

$\boldsymbol{z}_i \in \mathbb{R}^D$

$(i = 1, 2, \cdots, N)$

$D = 512 \times 512 \times 3$

$N = 3000000$

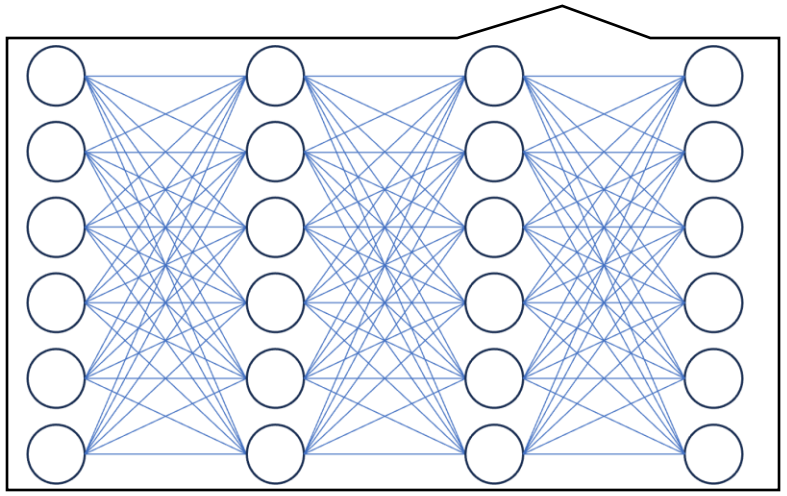
データセット

関数 $g\colon \mathbb{R}^d \to \mathbb{R}$

変数: $\boldsymbol{x} \in \mathbb{R}^d$

$\boldsymbol{z}_i \in \mathbb{R}^D$

Deep Neural Network



$d = 20$**万～**5**兆**

出力

$g(\boldsymbol{x}, \boldsymbol{z}_i) \in \mathbb{R}$

犬:3

正解

$y_i \in \mathbb{R}$

猫:2

誤差 $|g(\boldsymbol{x}, \boldsymbol{z}_i) - y_i|$

誤差の平均

$$f(\boldsymbol{x}) := \frac{1}{N} \sum_{i=1}^{N} |g(\boldsymbol{x}, \boldsymbol{z}_i) - y_i|$$

関数 $f(\boldsymbol{x})$**を最適化する**

背景：深層学習のための最適化

経験損失最小化問題

▷訓練データセット $\mathcal{S} := (\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_N)$

▷Deep Neural Network のパラメータ $\boldsymbol{x} \in \mathbb{R}^d$

$$\text{minimize } f(\boldsymbol{x}; \mathcal{S}) = \frac{1}{N} \sum_{i=1}^{N} l(\boldsymbol{x}; \boldsymbol{z}_i)$$

i 番目の訓練データ $\boldsymbol{z}_i$ に対する損失関数

▷非凸目的関数 $f\colon \mathbb{R}^d \to \mathbb{R}$ を最小化する．

背景：深層学習のための最適化

▷**確率的勾配降下法**（Stochastic Gradient Descent, SGD）

$$\boldsymbol{x}_{t+1} := \boldsymbol{x}_t - \eta_t \nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)$$

ミニバッチ確率的勾配

b個の確率的勾配の平均
（バッチサイズb=32, 1024など）

▷**慣性項付き確率的勾配降下法**（SGD with momentum）

$$\boldsymbol{m}_t := \nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t) + \beta \boldsymbol{m}_{t-1}$$

$$\boldsymbol{x}_{t+1} := \boldsymbol{x}_t - \eta_t \boldsymbol{m}_t$$

慣性係数

ただし、$\beta \in [0, 1), \boldsymbol{m}_{-1} = \boldsymbol{0}$ とする。

背景：深層学習のための最適化

▷AdamW[2017, 被引用数:31368]

$$\boldsymbol{m}_t := \beta_1 \boldsymbol{m}_{t-1} + (1-\beta_1)\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)$$

$$\boldsymbol{v}_t := \beta_2 \boldsymbol{v}_{t-1} + (1-\beta_2)\nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t) \odot \nabla f_{\mathcal{S}_t}(\boldsymbol{x}_t)$$

$$\hat{\boldsymbol{m}}_t := \boldsymbol{m}_t / (1-\beta_1^t)$$

$$\hat{\boldsymbol{v}}_t := \boldsymbol{v}_t / (1-\beta_2^t)$$

$$\boldsymbol{x}_{t+1} := \boldsymbol{x}_t - \eta_t \left\{ \mathrm{diag}\left(\frac{1}{\sqrt{\hat{v}_{t,i}}} \right) \hat{\boldsymbol{m}}_t + \lambda \boldsymbol{x}_t \right\}$$

▷ $\lambda = 0$ のとき，Adam[2014, 被引用数:224371]と一致する．

背景：深層学習のための最適化

▷Muon (Momentum Orthogonalized by Newton-Shulz)

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

▷探索方向は，勾配の行列を直交化した行列

$$O_t := \operatorname*{argmin}_{O \in \{O \in \mathbb{R}^{m \times n} : O^\top O = I_n\}} \|O - M_t\|_{\mathrm{F}}$$

▷特異値分解は計算コストがかかるので，Newton-Shultz反復5回で近似する．

▷Newton-Shultz反復[Higham2008]

$$X_0 := M_t / \|M_t\|_{\mathrm{F}}$$

$$X_{k+1} := aX_k + b(X_k X_k^\top)X_k + c(X_k X_k^\top)^2 X_k$$

$$(a = 3.4445, b = -4.7750, c = 2.0315)$$

[Higham2008] Nicholas J. Higham “Functions of matrices – theory and computation” SIAM 2008.

背景：深層学習のための最適化

▷Muon (Momentum Orthogonalized by Newton-Shulz)

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

▷重み減衰 (weight decay)

$$W_{t+1} := W_t - \eta(O_t + \lambda W_t)$$
$$= (1-\eta\lambda)W_t - \eta O_t$$

ただし，$\lambda > 0$ は重み減衰係数で，0.06など．

▷実験で性能が良くなったので採用された．

なぜ良くなるのか分かっていない．

動機-その1

▷**追加の慣性項**と**重み減衰**を有する「よく使われるMuon」の優位性を理論的に調査したい.

準備：仮定

(A1)関数 $f\colon \mathbb{R}^{m\times n} \to \mathbb{R}$ **は** L **-平滑とする.**

$$\forall A, B \in \mathbb{R}^{m\times n}\colon \|\nabla f(A) - \nabla f(B)\|_{\mathrm{F}} \leq L\|A - B\|_{\mathrm{F}}$$

(A2) $(W_t)_{t\in\mathbb{N}} \subset \mathbb{R}^{m\times n}$ **をMuonによって生成された点列とするとき,**

(i)任意の $t \in \mathbb{N}$ **と** $i \in [N]$ **に対して，次の式が成り立つとする.**

$$\mathbb{E}_{\xi_{t,i}}\left[\mathsf{G}_{\xi_{t,i}}(W_t)\right] = \nabla f(W_t)$$

(ii)次の式を満たす非負定数 σ^2 **が存在するとする.**

$$\mathbb{E}_{\xi_{t,i}}\left[\|\mathsf{G}_{\xi_{t,i}}(W_t) - \nabla f(W_t)\|_{\mathrm{F}}^2\right] \leq \sigma^2$$

(A3)NewtonShultzで得られた近似解が O_t **の定義を満たすとする.**

重み減衰付きMuonの望ましい性質-その1

▷**命題3.1. Muonを$\eta \leq 1/\lambda$の下で実行すると，次が成り立つ.**

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

証明)

$$\begin{aligned} \|W_t\|_{\mathrm{F}} &= \|(1-\eta\lambda)W_{t-1} - \eta O_t\|_{\mathrm{F}} \\ &\leq (1-\eta\lambda)\|W_{t-1}\|_{\mathrm{F}} + \eta\sqrt{n} \\ &\leq (1-\eta\lambda)^t \|W_0\|_{\mathrm{F}} + \eta\sqrt{n}\sum_{k=0}^{t-1}(1-\eta\lambda)^k \\ &\leq (1-\eta\lambda)^t \|W_0\|_{\mathrm{F}} + \frac{\sqrt{n}}{\lambda} \\ &\leq (1-\eta\lambda)^t \|W_0\|_{\mathrm{F}} + \frac{\sqrt{n}}{\lambda}. \end{aligned}$$

▷**Muonの探索方向$O_t \in \mathbb{R}^{m\times n}$は直交化された行列なので，$O_t^\top O_t = I_n$が満たされる.**
したがって，任意の時刻で次が成り立つ.

$$\|O_t\|_{\mathrm{F}} = \sqrt{n}$$

これはSGDやAdamでは成り立たない.

▷**命題3.1によると，パラメータのノルムの上界は単調減少する．特に，$\eta = 1/\lambda$ のとき，上界は最小となる.**

重み減衰付きMuonの望ましい性質-その2

▷**命題3.2. Muonを$\eta \leq 1/\lambda$の下で実行すると，次が成り立つ.**

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

▷命題3.2によると，全勾配のノルムの上界も単調減少する.
特に, $\eta = 1/\lambda$ のとき，上界は最小となる.

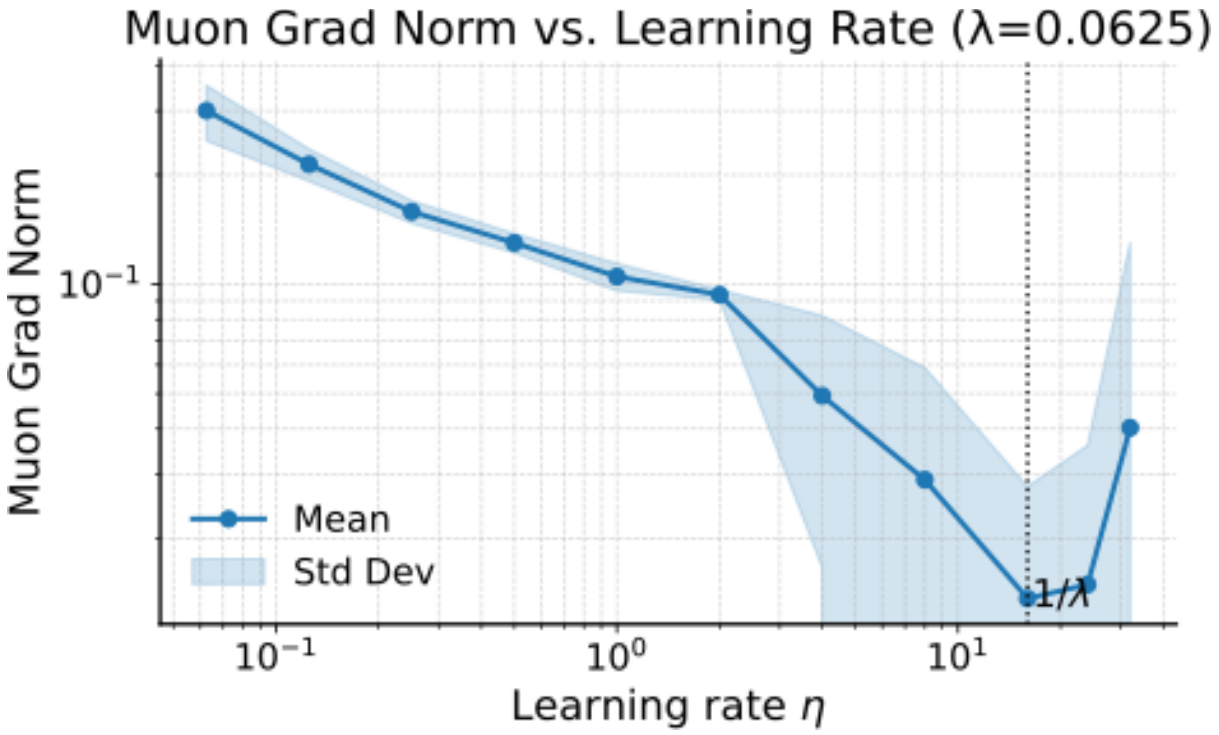
▷勾配の有界性は，MomentumやAdamの解析では仮定されるもので，一般に強い.

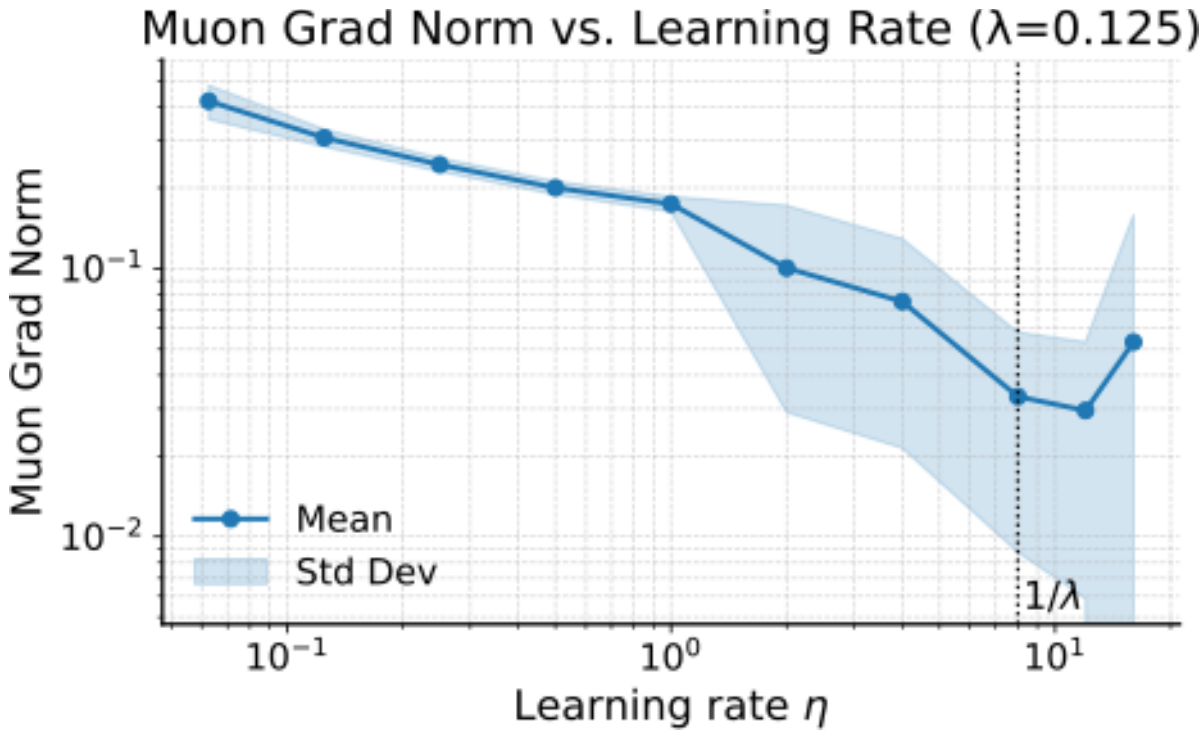
▷**系3.1. Muonを$\eta \leq 1/\lambda$の下で実行すると，次が成り立つ.**

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(W_t)\|_{\mathrm{F}} \leq \begin{cases} \frac{L\|W_0\|_{\mathrm{F}}}{\eta\lambda T} + \frac{L}{\lambda} + L\|W^\star\|_{\mathrm{F}} & \text{if } \eta < \frac{1}{\lambda}, \\ \frac{L}{\lambda} + L\|W^\star\|_{\mathrm{F}} & \text{if } \eta = \frac{1}{\lambda}. \end{cases}$$

▷全勾配ノルムの1乗の上界は，期待値なしで，およそ1/Tのオーダーで減少する.

数値実験-重み減衰付きMuonの望ましい性質





▷**理論の主張：**

Muonの学習率と重み減衰係数は，$\eta = 1/\lambda$ を満たすとき，勾配ノルム(の上界)は最小となる．

▷**実験の観察：**

ややずれはあるが，おおむね$\eta = 1/\lambda$ 付近で最小となっている．

Muonの収束解析–追加慣性項なし/重み減衰あり

▷**定理3.2(i) Muonを$\eta \leq 1/\lambda$の下で実行すると，次が成り立つ．**

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(W_t)\|_{\mathrm{F}}\right] \leq \frac{f(W_0)+\rho\|W_0\|_{\mathrm{F}}^2}{\eta T} + \frac{\Delta^2+2\sqrt{2n}\Delta}{(1-\beta)T} + \left(1-\beta+\frac{\lambda}{2}\right)\frac{\sigma^2}{b}$$

$$+\nu\sqrt{\frac{\sigma^2}{b}} + 2\gamma n(1+\gamma) + (1+L\eta)n + \frac{\rho n}{\lambda} + \frac{\lambda D_0^2}{2}$$

$$= \mathcal{O}\left(\frac{1}{T} + \left(1-\beta+\frac{\lambda}{2}\right)\frac{1}{b} + n\right)$$

ただし，$\Delta := \|M_0 - \nabla f(W_0)\|_{\mathrm{F}}^2, \bar{\beta} := \dfrac{(2\beta+1)(1-\beta)}{2}, \nu := \sqrt{2(1-\beta)n},$

$\gamma := \dfrac{L\eta}{1-\beta}, \rho := \dfrac{1+2(1+L\eta)\lambda}{2}, D_0 := L\left(\|W_0\|_{\mathrm{F}} + \dfrac{\sqrt{n}}{\lambda} + \|W^\star\|_{\mathrm{F}}\right)$ **とする．**

▷**全勾配ノルムの1乗の上界は，期待値ありでも，およそ1/Tのオーダーで減少する．**

Muonの収束解析–追加慣性項あり/重み減衰あり

▷**定理3.2(ii) Muonを$\eta \le 1/\lambda$の下で実行すると，次が成り立つ．**

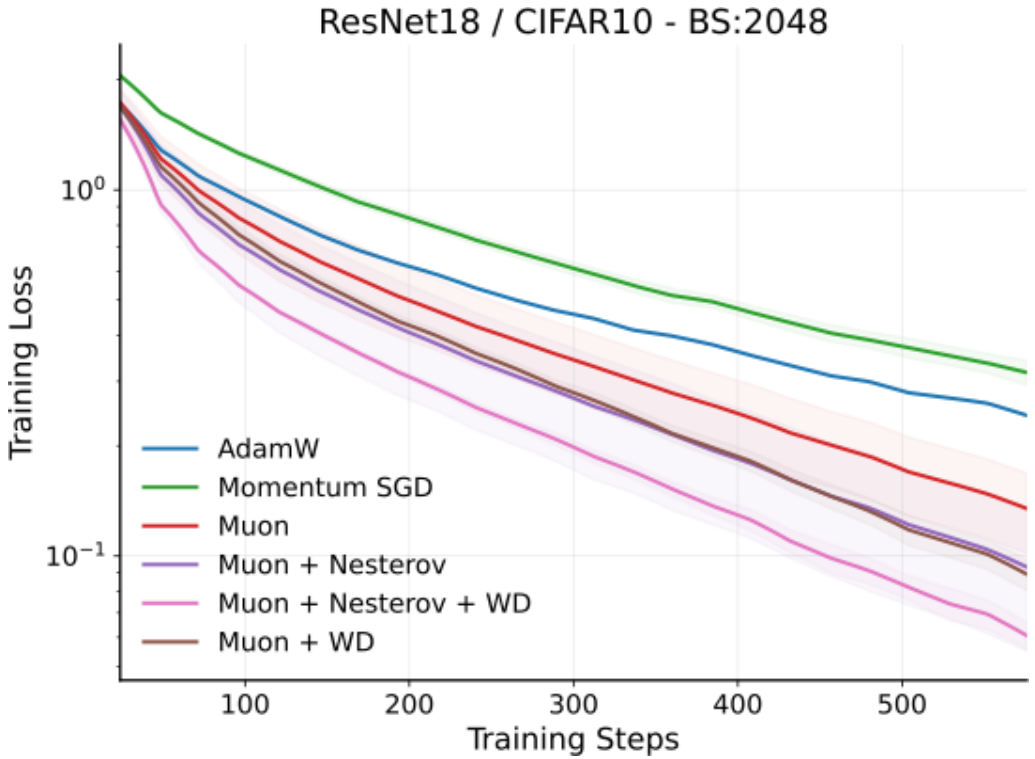
$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(W_t)\|_{\mathrm{F}}\right] \le \frac{f(W_0)+\rho\|W_0\|_{\mathrm{F}}^2}{\eta T} + \frac{\beta(\Delta^2+2\sqrt{2n}\Delta)}{(1-\beta)T} + \left(\bar{\beta}+\frac{\lambda}{2}\right)\frac{\sigma^2}{b}$$
$$+\beta\nu\sqrt{\frac{\sigma^2}{b}} + 2\gamma\beta n(1+\gamma) + (1+L\eta)n + \frac{\rho n}{\lambda} + \frac{\lambda D_0^2}{2}$$
$$= \mathcal{O}\left(\frac{1}{T} + \frac{(2\beta+1)(1-\beta)+\lambda}{2}\cdot\frac{1}{b} + n\right),$$

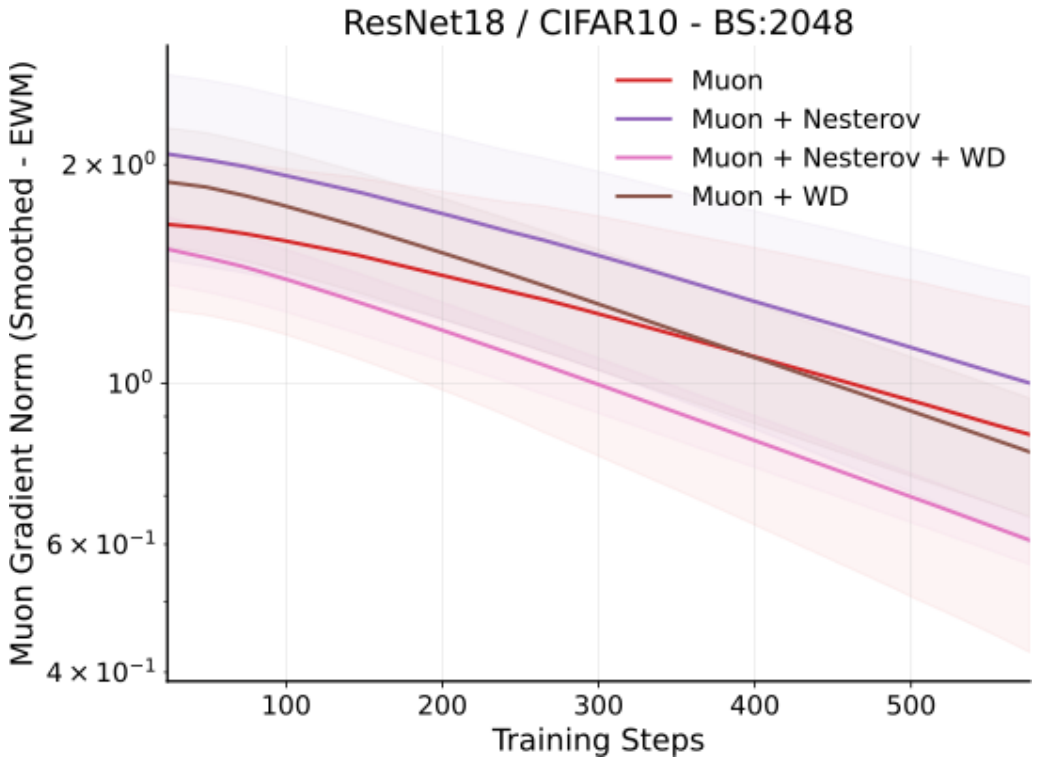
ただし，$\Delta := \|M_0 - \nabla f(W_0)\|_{\mathrm{F}}^2, \bar{\beta} := \dfrac{(2\beta+1)(1-\beta)}{2}, \nu := \sqrt{2(1-\beta)n},$

$$\gamma := \frac{L\eta}{1-\beta}, \rho := \frac{1+2(1+L\eta)\lambda}{2}, D_0 := L\left(\|W_0\|_{\mathrm{F}} + \frac{\sqrt{n}}{\lambda} + \|W^\star\|_{\mathrm{F}}\right)$$ **とする．**

▷**追加の慣性項がある方が，僅かに上界が小さくなる．**

数値実験-Muonの収束解析





▷理論の主張：

追加の慣性項を加えると，やや勾配ノルム(の上界)が小さくなる(紫よりピンクが小さいはず).

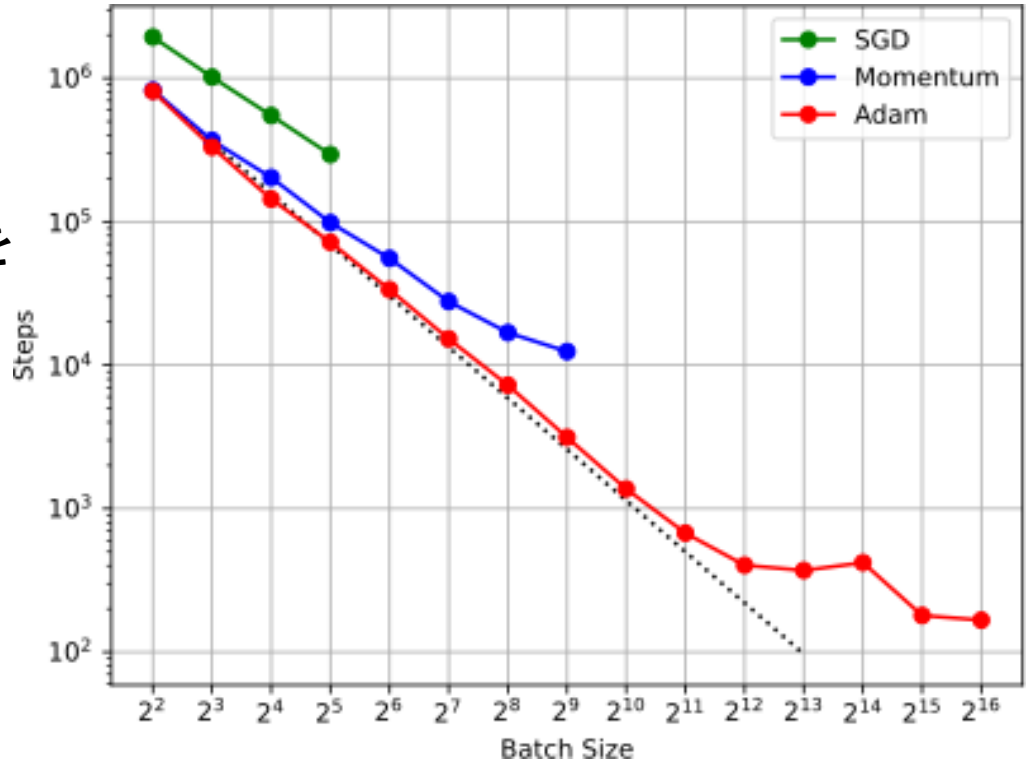
▷実験の観察：

重み減衰もある場合にはその傾向が見られる．訓練損失にもその傾向がある.

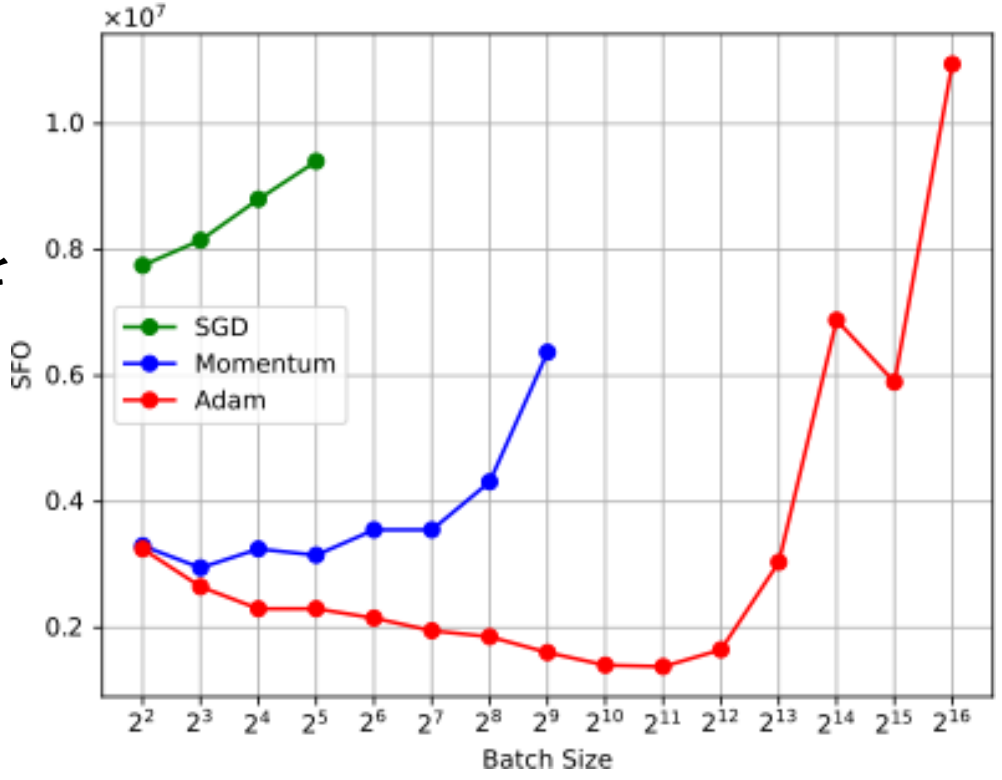
背景：クリティカルバッチサイズ

- バッチサイズを2倍にしていくと，訓練に必要な反復回数は半減していくが，あるバッチサイズ $b^\star$ からはその傾向が見られなくなる(収穫逓減).
- T 回の反復での総計算量は Tb となり，$b^\star$ は訓練に必要な計算量を最小にするバッチサイズである.

ある閾値を達成するための反復回数 T



ある閾値を達成するための計算量 Tb



動機-その2

- **先行研究によると，**
  - **クリティカルバッチサイズは最適化手法の種類に依存する．**

    Guodong Zhang et al., “Which algorithmic choices matter at which batch sizes? insights from a noisy quadratic model,” NeurIPS2019.

  - **クリティカルバッチサイズはハイパーパラメータに依存する．**

    Naoki Sato and Hideaki Iiduka, “Existence and Estimation of Critical Batch Size for Training Generative Adversarial Networks with Two Time-Scale Update Rule,” ICML2023.

  - **クリティカルバッチサイズはデータセットの大きさに依存する．**

    Hanlin Zhang et al., “How Does Critical Batch Size Scale in Pre-training?,” ICLR2025.

- **Muonのクリティカルバッチサイズがどのように定まるのか調査したい．**

Muonのクリティカルバッチサイズ-その1

▷収束解析より，

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

▷訓練が十分に完了したとき，ある閾値 $\epsilon > 0$ に対して，

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

▷この式を満たすための反復回数 T_{Muon} は，

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

▷この訓練にかかる計算量は，

$$T_{\mathrm{Muon}} b = \frac{Xb^2}{\epsilon b - Y}$$

Muonのクリティカルバッチサイズ-その2

▷この訓練にかかる計算量は,

$$T_{\mathrm{Muon}} b = \frac{X b^2}{\epsilon b - Y}$$

▷この計算量を最小にするbは,

$$b^{\star}_{\mathrm{Muon}} = \frac{2Y}{\epsilon}$$

▷収束解析を思い出すと,

$$=: Y$$

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(W_t)\|_{\mathrm{F}}\right] \leq \frac{f(W_0)+\rho\|W_0\|_{\mathrm{F}}^2}{\eta T} + \frac{\beta(\Delta^2+2\sqrt{2n}\Delta)}{(1-\beta)T} + \left(\frac{(2\beta+1)(1-\beta)+\lambda}{2}\right)\frac{\sigma^2}{b}$$

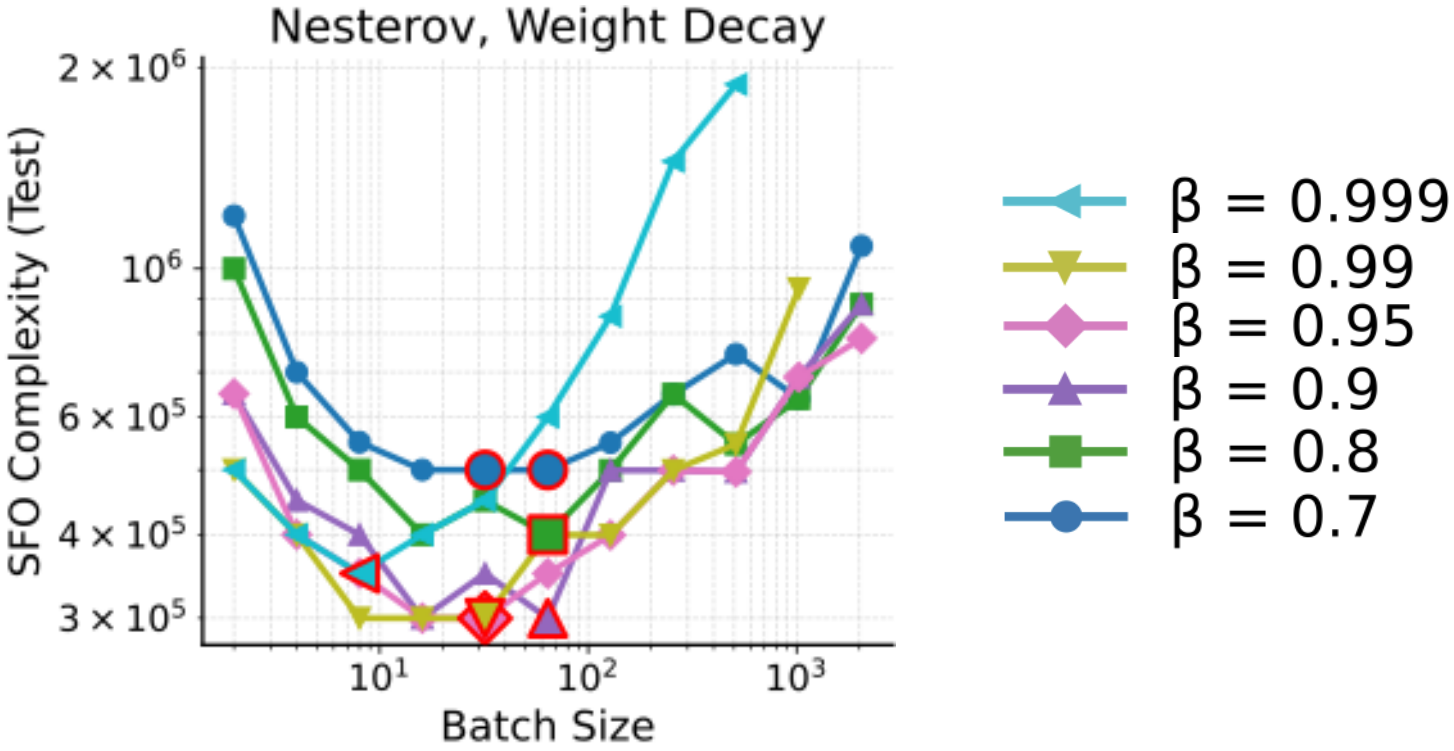
$$+\beta\nu\sqrt{\frac{\sigma^2}{b}} + 2\gamma\beta n(1+\gamma) + (1+L\eta)n + \frac{\rho n}{\lambda} + \frac{\lambda D_0^2}{2}$$

▷したがって，Muonのクリティカルバッチサイズは,

$$b^{\star}_{\mathrm{Muon}} = \frac{\{(2\beta+1)(1-\beta)+\lambda\}\sigma^2}{\epsilon}$$

▷β →大のとき $b^{\star}_{\mathrm{Muon}}$ →小

数値実験-クリティカルバッチサイズ



▷**理論の主張：**

慣性係数β→大のとき，クリティカルバッチサイズ$b^{\star}_{\mathrm{Muon}}$→小 (水色が一番左にあるはず)

▷**実験の観察：**

おおむねその傾向が見られる．

まとめ

- よく利用されるMuonの望ましい性質と収束解析を紹介した.
  - 探索方向が有界ならば，重み減衰を加えることでパラメータと全勾配のノルムの上界は単調減少する.
  - 全勾配ノルムの1乗の上界がおよそ1/Tのオーダーで減少する.
- よく利用されるMuonの最適なバッチサイズの推定を紹介した.
  - 収束解析の結果から，クリティカルバッチサイズの推定式を導出できる.
  - Muonのクリティカルバッチサイズは，慣性係数と重み減衰係数に依存していた.