

Muon with Finite Newton–Schulz: The Smoothing Benefit in Nonsmooth Nonconvex Optimization

arXiv:2608.26288

Mingyi Li¹ Taira Tsuchiya^{1,2}

¹The University of Tokyo ²RIKEN

Multi-agent seminar @ UTokyo · October 5, 2026

Muon (Jordan et al., 2024)：行列パラメータ向けの最適化手法
直交化したモーメントの方向に更新

Muon (Jordan et al., 2024)：行列パラメータ向けの最適化手法
直交化したモメンタムの方向に更新

LLM の事前学習で AdamW に代わる有力な手法

- スケーリング実験で効率を大幅に改善 (Liu et al., 2025)
- 派生版が Kimi K2 や GLM-4.5 などの最先端モデルの学習に採用 (Kimi Team, 2025; GLM-4.5 Team, 2025)
- 事前学習の体系的なベンチマークでトップクラスの性能 (Wen et al., 2026)

Muon (Jordan et al., 2024)：行列パラメータ向けの最適化手法
直交化したモメンタムの方向に更新

LLM の事前学習で AdamW に代わる有力な手法

- スケーリング実験で効率を大幅に改善 (Liu et al., 2025)
- 派生版が Kimi K2 や GLM-4.5 などの最先端モデルの学習に採用 (Kimi Team, 2025; GLM-4.5 Team, 2025)
- 事前学習の体系的なベンチマークでトップクラスの性能 (Wen et al., 2026)

実用性は、直交化を安価に計算できるかどうか依存

Muon の更新則

行列パラメータ $W_t \in \mathbb{R}^{m \times n}$ と確率的勾配 G_t に対し：

$$M_t = \beta M_{t-1} + (1 - \beta) G_t \quad (\text{モメンタム, } M_0 = 0)$$

$$W_{t+1} = W_t - D \operatorname{polar}(M_t) \quad (D > 0)$$

Muon の更新則

行列パラメータ $W_t \in \mathbb{R}^{m \times n}$ と確率的勾配 G_t に対し：

$$M_t = \beta M_{t-1} + (1 - \beta) G_t \quad (\text{モメンタム, } M_0 = 0)$$

$$W_{t+1} = W_t - D \text{ polar}(M_t) \quad (D > 0)$$

thin SVD $M = U \Sigma V^\top$ ($\Sigma \succ 0$) に対し, **polar factor** (直交化した M) は

$$\text{polar}(M) = UV^\top$$

写像 polar は特異ベクトルを保ち, 正の特異値をすべて 1 に変換
(0 の特異値は 0 のまま)

Muon の更新則

行列パラメータ $W_t \in \mathbb{R}^{m \times n}$ と確率的勾配 G_t に対し：

$$M_t = \beta M_{t-1} + (1 - \beta) G_t \quad (\text{モメンタム, } M_0 = 0)$$

$$W_{t+1} = W_t - D \text{ polar}(M_t) \quad (D > 0)$$

thin SVD $M = U \Sigma V^\top$ ($\Sigma \succ 0$) に対し, **polar factor** (直交化した M) は

$$\text{polar}(M) = UV^\top$$

写像 polar は特異ベクトルを保ち, 正の特異値をすべて 1 に変換
(0 の特異値は 0 のまま)

LLM の規模では, SVD による $\text{polar}(M)$ の計算コストが過大

polar factor の計算 : Newton–Schulz 反復

Newton–Schulz 反復 (Kovarik, 1970; Björck and Bowie, 1971) :

行列積だけで $\text{polar}(Y)$ に収束する反復 ($\|Y\|_{\text{op}} \leq 1$ となるように正規化しておく)

$$Y \leftarrow \frac{15}{8} Y - \frac{5}{4} (YY^T) Y + \frac{3}{8} (YY^T)^2 Y$$

$\text{polar}(Y) = (YY^T)^{-1/2} Y$ の $\lambda^{-1/2}$ を, $\lambda = 1$ まわりの 2 次 Taylor 多項式
 $p(\lambda) = \frac{15}{8} - \frac{5}{4}\lambda + \frac{3}{8}\lambda^2$ で置き換えた $p(YY^T) Y$

polar factor の計算 : Newton-Schulz 反復

Newton-Schulz 反復 (Kovarik, 1970; Björck and Bowie, 1971) :

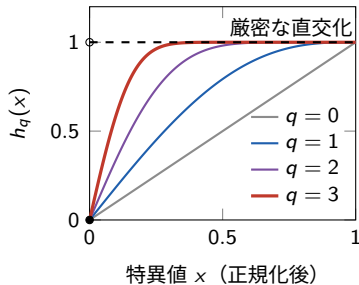
行列積だけで $\text{polar}(Y)$ に収束する反復 ($\|Y\|_{\text{op}} \leq 1$ となるように正規化しておく)

$$Y \leftarrow \frac{15}{8}Y - \frac{5}{4}(YY^T)Y + \frac{3}{8}(YY^T)^2Y$$

$\text{polar}(Y) = (YY^T)^{-1/2}Y$ の $\lambda^{-1/2}$ を, $\lambda = 1$ まわりの 2 次 Taylor 多項式 $p(\lambda) = \frac{15}{8} - \frac{5}{4}\lambda + \frac{3}{8}\lambda^2$ で置き換えた $p(YY^T)Y$

特異値への作用 : $Y = U\Sigma V^T$ とすると

- ① $(YY^T)^k Y = U\Sigma^{2k+1}V^T$
- ② 1 回の反復 $Y \mapsto p(YY^T)Y$ は, 特異ベクトルを保ち, 各特異値 x を $f(x) := x p(x^2) = \frac{15}{8}x - \frac{5}{4}x^3 + \frac{3}{8}x^5$ に変換
- ③ q 回の反復後は $Y_q = U h_q(\Sigma) V^T$, $h_q := f \circ q$



polar factor の計算：Newton–Schulz 反復

Newton–Schulz 反復 (Kovarik, 1970; Björck and Bowie, 1971) :

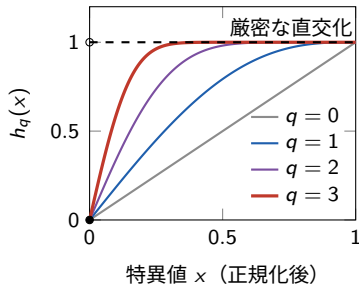
行列積だけで $\text{polar}(Y)$ に収束する反復 ($\|Y\|_{\text{op}} \leq 1$ となるように正規化しておく)

$$Y \leftarrow \frac{15}{8} Y - \frac{5}{4} (YY^T) Y + \frac{3}{8} (YY^T)^2 Y$$

$\text{polar}(Y) = (YY^T)^{-1/2} Y$ の $\lambda^{-1/2}$ を, $\lambda = 1$ まわりの 2 次 Taylor 多項式 $p(\lambda) = \frac{15}{8} - \frac{5}{4}\lambda + \frac{3}{8}\lambda^2$ で置き換えた $p(YY^T) Y$

特異値への作用 : $Y = U\Sigma V^T$ とすると

- ① $(YY^T)^k Y = U\Sigma^{2k+1} V^T$
- ② 1 回の反復 $Y \mapsto p(YY^T) Y$ は, 特異ベクトルを保ち, 各特異値 x を $f(x) := x p(x^2) = \frac{15}{8}x - \frac{5}{4}x^3 + \frac{3}{8}x^5$ に変換
- ③ q 回の反復後は $Y_q = U h_q(\Sigma) V^T$, $h_q := f \circ q$



Muon が実行するのは数ステップだけ：**有限回の Newton–Schulz 反復**

既存の Muon の収束解析

ほとんど全ての収束解析が目的関数の平滑性を仮定

- 多くの解析は Newton–Schulz 反復を厳密な polar factor で置き換える
(Li and Hong, 2025; Kovalev, 2025; Pethick et al., 2025; Shen et al., 2026)
- 有限回の Newton–Schulz 反復を考慮した解析も、これを近似誤差として評価
(Kim and Oh, 2026; Shulgin et al., 2026; Choudhury et al., 2026)
→ **厳密な polar factor に近いほど保証が改善**

既存の Muon の収束解析 (cont'd)

非平滑な関数ではどうか？ ReLU, max-pooling, ℓ_1 正則化, ヒンジ損失, 絶対値損失

既存の Muon の収束解析 (cont'd)

非平滑な関数ではどうか？ ReLU, max-pooling, ℓ_1 正則化, ヒンジ損失, 絶対値損失

最適化アルゴリズムを更新方向 $X_t = W_t - W_{t-1}$ をオンライン学習器で決定するとみなす (後ほど解説, online-to-nonconvex-conversion)

モメンタム M_t の厳密直交化を用いる **exact-polar Muon** の更新は,

$$X_{t+1} = -D \operatorname{polar}(M_t) \in \arg \min_{\|X\|_{\text{op}} \leq D} \sum_{s=1}^t \beta^{t-s} \langle G_s, X \rangle$$

since $-D \operatorname{polar}(S) \in \arg \min_{\|X\|_{\text{op}} \leq D} \langle S, X \rangle$ and $M_t = (1 - \beta) \sum_{s \leq t} \beta^{t-s} G_s$

既存の Muon の収束解析 (cont'd)

非平滑な関数ではどうか？ ReLU, max-pooling, ℓ_1 正則化, ヒンジ損失, 絶対値損失

最適化アルゴリズムを更新方向 $X_t = W_t - W_{t-1}$ をオンライン学習器で決定するとみなす (後ほど解説, online-to-nonconvex-conversion)

モメンタム M_t の厳密直交化を用いる **exact-polar Muon** の更新は,

$$X_{t+1} = -D \operatorname{polar}(M_t) \in \arg \min_{\|X\|_{\text{op}} \leq D} \sum_{s=1}^t \beta^{t-s} \langle G_s, X \rangle$$

since $-D \operatorname{polar}(S) \in \arg \min_{\|X\|_{\text{op}} \leq D} \langle S, X \rangle$ and $M_t = (1 - \beta) \sum_{s \leq t} \beta^{t-s} G_s$

exact-polar Muon = follow-the-leader (FTL)

FTL のリグレットは線形になりうるため, 停留点の保証なし

exact-polar Muon は, 凸で Lipschitz な関数でも収束しない例が存在 (Parshakova et al., 2026)

既存の Muon の収束解析 (cont'd)

非平滑な関数ではどうか？ ReLU, max-pooling, ℓ_1 正則化, ヒンジ損失, 絶対値損失

最適化アルゴリズムを更新方向 $X_t = W_t - W_{t-1}$ をオンライン学習器で決定するとみなす (後ほど解説, online-to-nonconvex-conversion)

モメンタム M_t の厳密直交化を用いる **exact-polar Muon** の更新は,

$$X_{t+1} = -D \operatorname{polar}(M_t) \in \arg \min_{\|X\|_{\text{op}} \leq D} \sum_{s=1}^t \beta^{t-s} \langle G_s, X \rangle$$

since $-D \operatorname{polar}(S) \in \arg \min_{\|X\|_{\text{op}} \leq D} \langle S, X \rangle$ and $M_t = (1 - \beta) \sum_{s \leq t} \beta^{t-s} G_s$

exact-polar Muon = follow-the-leader (FTL)

FTL のリグレットは線形になりうるため, 停留点の保証なし

exact-polar Muon は, 凸で Lipschitz な関数でも収束しない例が存在 (Parshakova et al., 2026)

Jiang et al. (2026): 学習器を置き換えた Pion と Leon を提案

→ Muon にない要素を追加し, 有限回の Newton-Schulz 反復は対象外

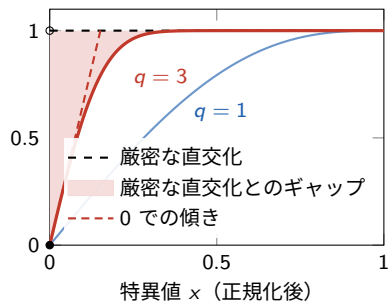
既存の解析：一方は有限回の Newton–Schulz 反復を誤差としか扱えず，平滑性が必要
他方は非平滑な目的関数を扱えるが，Muon が実際に行う更新を置き換える

Muon が実際に使うのは，厳密な polar factor ではなく
有限回の Newton–Schulz 反復

有限回の Newton–Schulz 反復は最適化にどのような影響を与えるか？
有限回の Newton–Schulz 反復が誤差ではなく利点となる場合はあるか？
Muon は非平滑非凸な関数の停留点を見つけられるか？

定理 (informal) 目的関数 $\mathcal{L}$ は Lipschitz 連続（非凸・非平滑でもよい）で、不偏かつ有界な確率的勾配が得られるとする．このとき，Newton–Schulz 反復を $q = O(\log(1/\varepsilon))$ 回用いるモメンタム付き Muon を実行すると， $O(\rho^{-1}\varepsilon^{-3} + \varepsilon^{-2})$ 個の確率的勾配で， (ρ, ε) -停留点を得られる．

定理 (informal) 目的関数 $\mathcal{L}$ は Lipschitz 連続（非凸・非平滑でもよい）で、不偏かつ有界な確率的勾配が得られるとする．このとき，Newton-Schulz 反復を $q = O(\log(1/\varepsilon))$ 回用いるモメンタム付き Muon を実行すると， $O(\rho^{-1}\varepsilon^{-3} + \varepsilon^{-2})$ 個の確率的勾配で， (ρ, ε) -停留点を得られる．



有限回の NS 反復は直交化の写像を平滑化

- 厳密な直交化 ($q = \infty$) = FTL
- 反復なし ($q = 0$)
= FTRL with squared Frobenius norm
- 両者が釣り合うのは $q = O(\log(1/\varepsilon))$
- 平滑な $\mathcal{L}$ でも（ほぼ）最適レートを達成

発表の流れ

1. はじめに
2. **問題設定**
3. Online-to-nonconvex conversion (O2NC)
4. 一般のスペクトル写像に対する discounted regret
5. 有限回の Newton-Schulz 反復を用いる Muon
6. Newton-Schulz 反復以外の写像
7. まとめ

問題設定と仮定

$$\min_{W \in \mathbb{R}^{m \times n}} \mathcal{L}(W) := \mathbb{E}_{\zeta \sim \mathcal{D}}[\ell(W; \zeta)], \quad m \leq n$$

ただし $\ell(W; \zeta)$ はデータ分布 $\mathcal{D}$ からのサンプル ζ での損失

仮定 目的関数 $\mathcal{L}$ は微分可能で, $\Delta_{\mathcal{L}} := \mathcal{L}(W_0) - \inf_W \mathcal{L}(W) < \infty$ とする.
 オラクルは各勾配評価点 W で, 過去と独立な乱数を用いて次を満たす G を返す:

$$\mathbb{E}[G \mid W] = \nabla \mathcal{L}(W), \quad \mathbb{E}[\|G\|_{\text{F}}^2 \mid W] \leq \Gamma^2, \quad \mathbb{E}[\|G - \nabla \mathcal{L}(W)\|_{\text{F}}^2 \mid W] \leq \sigma^2$$

微分不可能な点がある場合は, ランダムな摂動で微分可能な代理関数に帰着 (後述)

- このとき $\|\nabla \mathcal{L}(W)\|_{\text{F}} \leq \Gamma$ で $\mathcal{L}$ は Γ -Lipschitz
- **平滑性** (勾配の Lipschitz 連続性) の仮定なし

微分不可能な点がある場合

仮定では $\mathcal{L}$ は微分可能だが、ReLU ネットワークなどでは微分不可能な点がありうる
局所 Lipschitz な $\mathcal{L}$ なら、**ランダムな摂動**で微分可能な問題に帰着 (Cutkosky et al., 2023) :

$$\hat{\mathcal{L}}_\delta(W) := \mathbb{E}_Z[\mathcal{L}(W + \delta Z)], \quad Z \text{ は Frobenius 単位球内の一様分布}$$

- $\hat{\mathcal{L}}_\delta$ は微分可能で Lipschitz : 仮定を満たす (値を計算する必要はない)
- $\hat{\mathcal{L}}_\delta$ の確率的勾配 : Z を引き, $W + \delta Z$ で元のオラクルを呼んだ出力 G でよい

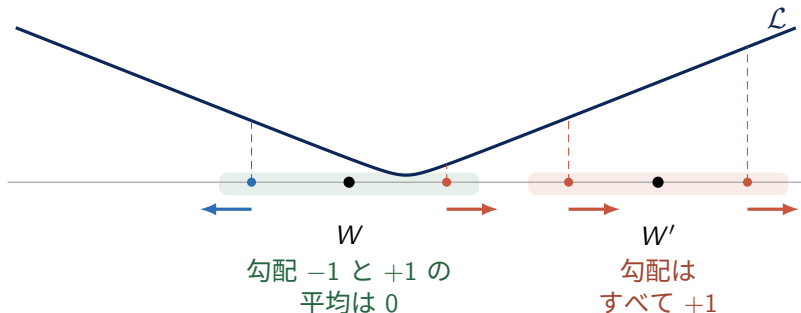
$$\mathbb{E}[G \mid W] = \mathbb{E}_Z[\nabla \mathcal{L}(W + \delta Z)] = \nabla \hat{\mathcal{L}}_\delta(W) \quad W + \delta Z \text{ では確率 } 1 \text{ で } \mathcal{L} \text{ が微分可能}$$

- $\hat{\mathcal{L}}_\delta$ に対する半径 $\rho \geq \delta$ の停留性 (後述の (ρ, ε) -停留性) の保証は,
Cutkosky et al. (2023) と同様に, $\mathcal{L}$ に対する半径 2ρ の保証になる

Cutkosky et al. (2023) の注記 : “In practice we suspect that such perturbation arguments are unnecessary: intuitively an algorithm is unlikely to query a point of non-differentiability.”

平滑性がない場合の停留性

平滑な場合の目標は $\|\nabla \mathcal{L}(W)\|_* \leq \varepsilon$ だが、平滑性がないとこの目標は厳しすぎる：
折れ曲がりの近くでは、最小点のごく近くでも勾配ノルムは ≈ 1 のまま ($\|\cdot\|_*$: 核ノルム)



その点での勾配ノルム： W でも W' でも ≈ 1
 図示した近傍の勾配の平均： W では ≈ 0 , W' では ≈ 1

(ρ, ε) -停留点

点 W が (ρ, ε) -停留点：期待値が W で W から平均的に距離が ρ 程度しか離れていないような分布をうまく選ぶことで、勾配の平均を ε 以下にできる

定義 (Jiang et al., 2026)

$$\mathcal{P}(W; \rho) := \left\{ p : \mathbb{R}^{m \times n} \text{ 上の有限台の分布 } \mid \mathbb{E}_{Y \sim p}[Y] = W, \mathbb{E}_{Y \sim p} \|Y - W\|_{\text{op}} \leq \rho \right\}$$

点 W が (ρ, ε) -**停留点**であるとは、次が成り立つことをいう：

$$\|\nabla \mathcal{L}(W)\|_*^{[\rho]} := \inf_{p \in \mathcal{P}(W; \rho)} \|\mathbb{E}_{Y \sim p} \nabla \mathcal{L}(Y)\|_* \leq \varepsilon$$

- $\mathcal{P}(W; \rho)$ ：平均 W ， W からの平均距離 ρ 以下の分布（前の図の 2 点分布など）
- Goldstein 型の緩和 (Goldstein, 1977)：一般の Lipschitz 関数では不可避 (Zhang et al., 2020; Kornowski and Shamir, 2022)

有限台の制限は Jiang らの定義どおり（有限個のモーメント条件を保って有限台に置き換えられるので本質的ではない）

平滑な場合の停留点との比較

平滑な場合の目標は ε -**停留点** : $\|\nabla \mathcal{L}(W)\|_* \leq \varepsilon$

常に $\|\nabla \mathcal{L}(W)\|_*^{[\rho]} \leq \|\nabla \mathcal{L}(W)\|_*$ なので, ε -停留点は (ρ, ε) -停留点 (W に全質量をもつ分布)

平滑な場合の停留点との比較

平滑な場合の目標は ε -**停留点** : $\|\nabla \mathcal{L}(W)\|_* \leq \varepsilon$

常に $\|\nabla \mathcal{L}(W)\|_*^{[\rho]} \leq \|\nabla \mathcal{L}(W)\|_*$ なので, ε -停留点は (ρ, ε) -停留点 (W に全質量をもつ分布)

補題 $\|\nabla \mathcal{L}(W) - \nabla \mathcal{L}(W')\|_* \leq L_{\text{op}} \|W - W'\|_{\text{op}}$ ($\forall W, W'$) ならば,

$$\|\nabla \mathcal{L}(W)\|_* \leq \|\nabla \mathcal{L}(W)\|_*^{[\rho]} + L_{\text{op}} \rho.$$

特に $\rho = \varepsilon / L_{\text{op}}$ とすれば, (ρ, ε) -停留点は 2ε -停留点.

証明: 任意の $p \in \mathcal{P}(W; \rho)$ と $Y \sim p$ に対し

$$\|\nabla \mathcal{L}(W)\|_* \leq \|\mathbb{E} \nabla \mathcal{L}(Y)\|_* + \|\mathbb{E} [\nabla \mathcal{L}(W) - \nabla \mathcal{L}(Y)]\|_*$$

三角不等式

$$\leq \|\mathbb{E} \nabla \mathcal{L}(Y)\|_* + \mathbb{E} \|\nabla \mathcal{L}(W) - \nabla \mathcal{L}(Y)\|_*$$

Jensen の不等式

$$\leq \|\mathbb{E} \nabla \mathcal{L}(Y)\|_* + L_{\text{op}} \mathbb{E} \|W - Y\|_{\text{op}} \leq \|\mathbb{E} \nabla \mathcal{L}(Y)\|_* + L_{\text{op}} \rho$$

平滑性と $\mathbb{E} \|W - Y\|_{\text{op}} \leq \rho$

右辺の ρ についての下限を取ればよい

この事実は, 後で平滑な目的関数を扱うときに使用

Input: momentum $\beta \in (0, 1)$, radius $D > 0$, operator-norm bound $G_{\text{op}} > 0$, depth q

$M_0 = 0, X_1 = 0$

for $t = 1, 2, \dots$:

 receive G_t

$M_t \leftarrow \beta M_{t-1} + (1 - \beta)G_t$

momentum

$Y \leftarrow M_t / G_{\text{op}}$

normalization

repeat q times:

$Y \leftarrow \frac{15}{8}Y - \frac{5}{4}(YY^\top)Y + \frac{3}{8}(YY^\top)^2Y$

Newton–Schulz step

$X_{t+1} \leftarrow -D Y$ ($W_{t+1} \leftarrow W_t - D Y$)

next update

Input: momentum $\beta \in (0, 1)$, radius $D > 0$, operator-norm bound $G_{\text{op}} > 0$, depth q

$M_0 = 0, X_1 = 0$

for $t = 1, 2, \dots$:

 receive G_t

$M_t \leftarrow \beta M_{t-1} + (1 - \beta) G_t$

momentum

$Y \leftarrow M_t / G_{\text{op}}$

normalization

repeat q times:

$Y \leftarrow \frac{15}{8} Y - \frac{5}{4} (YY^\top) Y + \frac{3}{8} (YY^\top)^2 Y$

Newton-Schulz step

$X_{t+1} \leftarrow -D Y$ ($W_{t+1} \leftarrow W_t - D Y$)

next update

確率 1 で $\|G_t\|_{\text{op}} \leq G_{\text{op}}$ と仮定: $\|M_t\|_{\text{op}} \leq G_{\text{op}}$ なので M_t / G_{op} の特異値は $[0, 1]$

$$X_{t+1} = -D \mathcal{H}_{h_q}(M_t / G_{\text{op}}), \quad \mathcal{H}_h(U \Sigma V^\top) := U \text{diag}(h(\sigma_1), \dots, h(\sigma_r)) V^\top$$

($U \Sigma V^\top$ は thin SVD) h_q は f の q 回合成 ($[1, \infty)$ 上で 1). 厳密な直交化は $h(x) = \mathbf{1}\{x > 0\}$ に対応

発表の流れ

1. はじめに
2. 問題設定
3. **Online-to-nonconvex conversion (O2NC)**
4. 一般のスペクトル写像に対する discounted regret
5. 有限回の Newton-Schulz 反復を用いる Muon
6. Newton-Schulz 反復以外の写像
7. まとめ

最適化手法をオンライン学習とみなす

最適化手法

更新方向 X_t を選ぶ

$W_t \leftarrow W_{t-1} + X_t$

確率的勾配 G_t を得る

=

オンライン学習器

行動（更新方向） X_t を出す

損失 $X \mapsto \langle G_t, X \rangle$ を受け取る

損失 $\langle G_t, X_t \rangle$ を被る

オンライン学習器の出力は実行可能領域上の点 W_t ではなく更新方向 X_t

online-to-batch ではオンライン学習器が実行可能領域上の点を出力

損失が $\mathcal{L}$ の変化に対応するのは 1 次近似 $\mathcal{L}(W_t) - \mathcal{L}(W_{t-1}) \approx \langle \nabla \mathcal{L}(W_{t-1}), X_t \rangle$ から

(O2NC で厳密な等式に)

最適化手法をオンライン学習とみなす

最適化手法

更新方向 X_t を選ぶ

$W_t \leftarrow W_{t-1} + X_t$

確率的勾配 G_t を得る

=

オンライン学習器

行動（更新方向） X_t を出す

損失 $X \mapsto \langle G_t, X \rangle$ を受け取る

損失 $\langle G_t, X_t \rangle$ を被る

オンライン学習器の出力は実行可能領域上の点 W_t ではなく更新方向 X_t

online-to-batch ではオンライン学習器が実行可能領域上の点を出力

損失が $\mathcal{L}$ の変化に対応するのは 1 次近似 $\mathcal{L}(W_t) - \mathcal{L}(W_{t-1}) \approx \langle \nabla \mathcal{L}(W_{t-1}), X_t \rangle$ から
(O2NC で厳密な等式に)

既存の最適化手法は、既知のオンライン学習器に対応：

モメンタム付き SGD はオンライン勾配降下法, Adam は discounted FTRL,

Muon（厳密な直交化）は follow-the-leader

いずれも小さな変更を含めた対応 (Zhang and Cutkosky, 2024; Ahn et al., 2024; Jiang et al., 2026)

Online-to-nonconvex conversion (O2NC)

O2NC (Cutkosky et al., 2023) : オンライン学習器に更新方向を選ばせ、リグレットの上界を停留性の保証に変換する枠組み

for $t = 1, 2, \dots$:

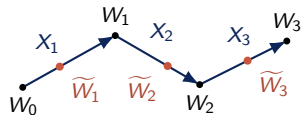
the learner chooses an update X_t , $\|X_t\|_{\text{op}} \leq D$

update $W_t \leftarrow W_{t-1} + X_t$

query a stochastic gradient G_t at $\widetilde{W}_t \leftarrow W_{t-1} + u_t X_t$,

$u_t \sim \text{Unif}[0, 1]$

feed the linear loss $X \mapsto \langle G_t, X \rangle$ to the learner



勾配評価点 $\widetilde{W}_t$:
線分上の一様ランダムな点

Online-to-nonconvex conversion (O2NC)

O2NC (Cutkosky et al., 2023) : オンライン学習器に更新方向を選ばせ、リグレットの上界を停留性の保証に変換する枠組み

for $t = 1, 2, \dots$:

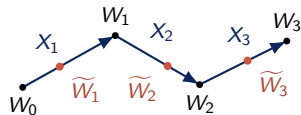
the learner chooses an update X_t , $\|X_t\|_{\text{op}} \leq D$

update $W_t \leftarrow W_{t-1} + X_t$

query a stochastic gradient G_t at $\widetilde{W}_t \leftarrow W_{t-1} + u_t X_t$,

$u_t \sim \text{Unif}[0, 1]$

feed the linear loss $X \mapsto \langle G_t, X \rangle$ to the learner



勾配評価点 $\widetilde{W}_t$:
線分上の一様ランダムな点

固定点 U ($\|U\|_{\text{op}} \leq D$) に対するリグレット

$$\text{Reg}_T(U) = \sum_{t=1}^T \langle G_t, X_t - U \rangle$$

O2NC (Cutkosky et al., 2023) : オンライン学習器に更新方向を選ばせ、リグレットの上界を停留性の保証に変換する枠組み

for $t = 1, 2, \dots$:

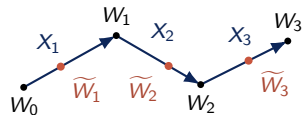
the learner chooses an update X_t , $\|X_t\|_{\text{op}} \leq D$

update $W_t \leftarrow W_{t-1} + X_t$

query a stochastic gradient G_t at $\widetilde{W}_t \leftarrow W_{t-1} + u_t X_t$,

$u_t \sim \text{Unif}[0, 1]$

feed the linear loss $X \mapsto \langle G_t, X \rangle$ to the learner



勾配評価点 $\widetilde{W}_t$:
線分上の一様ランダムな点

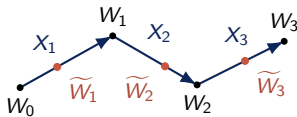
固定点 U ($\|U\|_{\text{op}} \leq D$) に対するリグレット

$$\text{Reg}_T(U) = \sum_{t=1}^T \langle G_t, X_t - U \rangle$$

学習器から見ると、 $\mathcal{L}$ の非平滑非凸性は予測不能な勾配列 $(G_t)_t$ として表れる
平滑な場合は勾配の予測可能性を活用して良いレートが得られる

1 ラウンドの厳密な等式

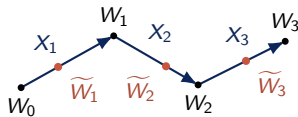
学習器が選んだ更新方向 $X_t = W_t - W_{t-1}$ に対し、
勾配評価点 $\widetilde{W}_t = W_{t-1} + u_t X_t$ ($u_t \sim \text{Unif}[0, 1]$) で
確率的勾配 G_t を取得



$$\begin{aligned}
 \text{線分に沿った基本定理 } \mathcal{L}(W_t) - \mathcal{L}(W_{t-1}) &= \int_0^1 \langle \nabla \mathcal{L}(W_{t-1} + u(W_t - W_{t-1})), W_t - W_{t-1} \rangle du \\
 X_t \text{ の定義} &= \int_0^1 \langle \nabla \mathcal{L}(W_{t-1} + uX_t), X_t \rangle du \\
 u_t \text{ は一様} &= \mathbb{E}_{u_t} \langle \nabla \mathcal{L}(\widetilde{W}_t), X_t \rangle \\
 G_t \text{ は不偏} &= \mathbb{E}[\langle G_t, X_t \rangle \mid W_{t-1}, X_t]
 \end{aligned}$$

1 ラウンドの厳密な等式

学習器が選んだ更新方向 $X_t = W_t - W_{t-1}$ に対し、
勾配評価点 $\widetilde{W}_t = W_{t-1} + u_t X_t$ ($u_t \sim \text{Unif}[0, 1]$) で
確率的勾配 G_t を取得



$$\begin{aligned}
 \text{線分に沿った基本定理 } \mathcal{L}(W_t) - \mathcal{L}(W_{t-1}) &= \int_0^1 \langle \nabla \mathcal{L}(W_{t-1} + u(W_t - W_{t-1})), W_t - W_{t-1} \rangle du \\
 X_t \text{ の定義} &= \int_0^1 \langle \nabla \mathcal{L}(W_{t-1} + uX_t), X_t \rangle du \\
 u_t \text{ は一様} &= \mathbb{E}_{u_t} \langle \nabla \mathcal{L}(\widetilde{W}_t), X_t \rangle \\
 G_t \text{ は不偏} &= \mathbb{E}[\langle G_t, X_t \rangle \mid W_{t-1}, X_t]
 \end{aligned}$$

平滑性を仮定しなくても、損失 $\langle G_t, X_t \rangle$ は $\mathcal{L}$ の変化量の不偏推定量：

$$\mathbb{E}[\mathcal{L}(W_T)] - \mathcal{L}(W_0) = \sum_{t=1}^T \mathbb{E} \langle G_t, X_t \rangle$$

リグレットの上界から勾配の平均の上界へ

$U := -D \operatorname{polar}(\sum_{t=1}^T G_t)$ (事後的に最良の固定方向, $\|U\|_{\text{op}} \leq D$) に対し, 定義より

$$\sum_{t=1}^T \langle G_t, X_t \rangle = \operatorname{Reg}_T(U) + \langle \sum_{t=1}^T G_t, U \rangle = \operatorname{Reg}_T(U) - D \left\| \sum_{t=1}^T G_t \right\|_*$$

左辺の期待値は $\mathbb{E}[\mathcal{L}(W_T)] - \mathcal{L}(W_0) \geq -\Delta_{\mathcal{L}}$ なので

$$\mathbb{E} \left\| \frac{1}{T} \sum_{t=1}^T G_t \right\|_* \leq \frac{\mathbb{E}[\operatorname{Reg}_T(U)] + \Delta_{\mathcal{L}}}{DT} : \text{勾配の平均はリグレットで抑えられる}$$

OGD なら $\operatorname{Reg}_T(U) = O(\sqrt{T})$ で, 右辺は $T \rightarrow \infty$ で 0 へ (この上界は U の選び方によらない)

リグレットの上界から勾配の平均の上界へ

$U := -D \operatorname{polar}(\sum_{t=1}^T G_t)$ (事後的に最良の固定方向, $\|U\|_{\text{op}} \leq D$) に対し, 定義より

$$\sum_{t=1}^T \langle G_t, X_t \rangle = \operatorname{Reg}_T(U) + \langle \sum_{t=1}^T G_t, U \rangle = \operatorname{Reg}_T(U) - D \left\| \sum_{t=1}^T G_t \right\|_*$$

左辺の期待値は $\mathbb{E}[\mathcal{L}(W_T)] - \mathcal{L}(W_0) \geq -\Delta_{\mathcal{L}}$ なので

$$\mathbb{E} \left\| \frac{1}{T} \sum_{t=1}^T G_t \right\|_* \leq \frac{\mathbb{E}[\operatorname{Reg}_T(U)] + \Delta_{\mathcal{L}}}{DT} : \text{勾配の平均はリグレットで抑えられる}$$

OGD なら $\operatorname{Reg}_T(U) = O(\sqrt{T})$ で, 右辺は $T \rightarrow \infty$ で 0 へ (この上界は U の選び方によらない)

評価点 $\widetilde{W}_t$ は軌道全体に分布 (互いの距離は最大 DT)

一点での停留性には, **直近のラウンド**の勾配だけを平均

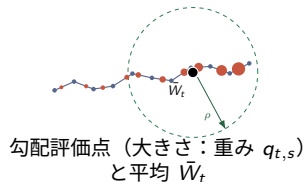
- ブロックに分けて平均 (Cutkosky et al., 2023) (付録のスライド)
- **モメンタムの重み β^{t-s} で平均** **次のスライドから解説**

モメンタムが小さければ停留点

重み $q_{t,s} := (1 - \beta)\beta^{t-s}/(1 - \beta^t)$: 和は 1, 直近ほど重い
($s = t$ で最大, 過去ほど β 倍ずつ減る)

$$\frac{M_t}{1 - \beta^t} = \sum_{s \leq t} q_{t,s} G_s \quad \text{直近の勾配の平均}$$

$$\bar{W}_t := \sum_{s \leq t} q_{t,s} \widetilde{W}_s \quad \text{直近の勾配評価点の平均}$$

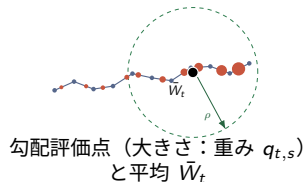


モメンタムが小さければ停留点

重み $q_{t,s} := (1 - \beta)\beta^{t-s}/(1 - \beta^t)$: 和は 1, 直近ほど重い
($s = t$ で最大, 過去ほど β 倍ずつ減る)

$$\frac{M_t}{1 - \beta^t} = \sum_{s \leq t} q_{t,s} G_s \quad \text{直近の勾配の平均}$$

$$\bar{W}_t := \sum_{s \leq t} q_{t,s} \bar{W}_s \quad \text{直近の勾配評価点の平均}$$



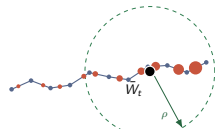
停留性の分布: $\bar{W}_s$ を確率 $q_{t,s}$ で選ぶ分布 p は, 平均が $\bar{W}_t$, $\bar{W}_t$ からの平均距離が $\mathbb{E}_{Y \sim p} \|Y - \bar{W}_t\|_{\text{op}} \leq 2D \mathbb{E}_{s,s' \sim q_{t,\cdot}} |s - s'| \leq 4D\beta/(1 - \beta)$ (隣接する評価点の距離 $\leq 2D$)
 $D = (1 - \beta)\rho/(4\beta)$ とすれば平均距離 $\leq \rho$, すなわち $p \in \mathcal{P}(\bar{W}_t; \rho)$

モメンタムが小さければ停留点

重み $q_{t,s} := (1 - \beta)\beta^{t-s}/(1 - \beta^t)$: 和は 1, 直近ほど重い
($s = t$ で最大, 過去ほど β 倍ずつ減る)

$$\frac{M_t}{1 - \beta^t} = \sum_{s \leq t} q_{t,s} G_s \quad \text{直近の勾配の平均}$$

$$\bar{W}_t := \sum_{s \leq t} q_{t,s} \bar{W}_s \quad \text{直近の勾配評価点の平均}$$



勾配評価点 (大きさ: 重み $q_{t,s}$)
と平均 $\bar{W}_t$

停留性の分布: $\bar{W}_s$ を確率 $q_{t,s}$ で選ぶ分布 p は, 平均が $\bar{W}_t$, $\bar{W}_t$ からの平均距離が $\mathbb{E}_{Y \sim p} \|Y - \bar{W}_t\|_{\text{op}} \leq 2D \mathbb{E}_{s,s' \sim q_{t,\cdot}} |s - s'| \leq 4D\beta/(1 - \beta)$ (隣接する評価点の距離 $\leq 2D$)
 $D = (1 - \beta)\rho/(4\beta)$ とすれば平均距離 $\leq \rho$, すなわち $p \in \mathcal{P}(\bar{W}_t; \rho)$

$$\text{雑音がない場合 } (G_s = \nabla \mathcal{L}(\bar{W}_s)) : \quad \|\nabla \mathcal{L}(\bar{W}_t)\|_*^{[\rho]} \leq \left\| \sum_{s \leq t} q_{t,s} \nabla \mathcal{L}(\bar{W}_s) \right\|_* = \frac{\|M_t\|_*}{1 - \beta^t}$$

雑音 $E_s := G_s - \nabla \mathcal{L}(\bar{W}_s)$ がある場合は右辺に $(1 - \beta) \left\| \sum_{s \leq t} \beta^{t-s} E_s \right\|_* / (1 - \beta^t)$ が加わる (補題で評価)

あとは, ほとんどの時刻 t で $\|M_t\|_*$ が小さいことを示せば十分

Discounted regret とモメンタムの恒等式

目的： $\|M_t\|_*$ を学習器のリグレットで抑える

定義 (Zhang and Cutkosky, 2024; Jiang et al., 2026) 時刻 t での **discounted regret** :

$$\text{Reg}_t^{[\beta]}(D) := \max_{\|U\|_{\text{op}} \leq D} \sum_{s=1}^t \beta^{t-s} \langle G_s, X_s - U \rangle \quad \text{直近 } \approx 1/(1-\beta) \text{ ラウンドでのリグレット}$$

Discounted regret とモメンタムの恒等式

目的： $\|M_t\|_*$ を学習器のリグレットで抑える

定義 (Zhang and Cutkosky, 2024; Jiang et al., 2026) 時刻 t での **discounted regret** :

$$\text{Reg}_t^{[\beta]}(D) := \max_{\|U\|_{\text{op}} \leq D} \sum_{s=1}^t \beta^{t-s} \langle G_s, X_s - U \rangle \quad \text{直近 } \approx 1/(1-\beta) \text{ ラウンドでのリグレット}$$

$$\begin{aligned} \text{Reg}_t^{[\beta]}(D) &= \sum_{s=1}^t \beta^{t-s} \langle G_s, X_s \rangle + \max_{\|U\|_{\text{op}} \leq D} \langle \sum_{s=1}^t \beta^{t-s} G_s, -U \rangle \\ &= \sum_{s=1}^t \beta^{t-s} \langle G_s, X_s \rangle + \frac{D}{1-\beta} \|M_t\|_* \quad \text{双対性と } M_t = (1-\beta) \sum_{s \leq t} \beta^{t-s} G_s \end{aligned}$$

$$\text{どの学習器でも} \quad \frac{D}{1-\beta} \|M_t\|_* = \text{Reg}_t^{[\beta]}(D) + \underbrace{\left(- \sum_{s=1}^t \beta^{t-s} \langle G_s, X_s \rangle \right)}_{\text{期待値は直近の } \mathcal{L} \text{ の減少量}}$$

モメンタムが大きいのは、直近で $\mathcal{L}$ が減少したか、discounted regret が大きいときだけ $\mathcal{L}$ の減少は合計 $\Delta_{\mathcal{L}}$ まで：リグレットが小さければ、モメンタムが大きい時刻は少ない

時刻についての和

目的：ほとんどの時刻で $\|M_t\|_*$ が小さいことを，リグレットと $\Delta_{\mathcal{L}}$ で示す

① 時刻 t の恒等式に重み $a_t := 1 - \beta$ ($t < T$), $a_T := 1$ を掛けて足す：

$$\frac{D}{1 - \beta} \sum_{t=1}^T a_t \|M_t\|_* = \sum_{t=1}^T a_t \text{Reg}_t^{[\beta]}(D) - \sum_{s=1}^T \langle G_s, X_s \rangle$$

ラウンド s の損失は時刻 $t \geq s$ の恒等式に重み β^{t-s} で現れ，合計は $\sum_{t \geq s} a_t \beta^{t-s} = (1 - \beta^{T-s}) + \beta^{T-s} = 1$

② 最後の項の期待値は $\mathcal{L}(W_0) - \mathbb{E}[\mathcal{L}(W_T)] \leq \Delta_{\mathcal{L}}$

リグレットが小さければ $\sum_t a_t \|M_t\|_*$ も小さい

時刻についての和

目的：ほとんどの時刻で $\|M_t\|_*$ が小さいことを，リグレットと $\Delta_{\mathcal{L}}$ で示す

① 時刻 t の恒等式に重み $a_t := 1 - \beta$ ($t < T$), $a_T := 1$ を掛けて足す：

$$\frac{D}{1-\beta} \sum_{t=1}^T a_t \|M_t\|_* = \sum_{t=1}^T a_t \text{Reg}_t^{[\beta]}(D) - \sum_{s=1}^T \langle G_s, X_s \rangle$$

ラウンド s の損失は時刻 $t \geq s$ の恒等式に重み β^{t-s} で現れ，合計は $\sum_{t \geq s} a_t \beta^{t-s} = (1 - \beta^{T-s}) + \beta^{T-s} = 1$

② 最後の項の期待値は $\mathcal{L}(W_0) - \mathbb{E}[\mathcal{L}(W_T)] \leq \Delta_{\mathcal{L}}$

リグレットが小さければ $\sum_t a_t \|M_t\|_*$ も小さい

③ 出力：確率 $\mathbb{P}(\tau = t) = \frac{a_t(1-\beta^t)}{(1-\beta)^T}$ で τ を選び $\bar{W}_\tau$ を返す (**discounted O2NC**)

重み a_t で ② を使い，因子 $1 - \beta^t$ は停留性の不等式の分母と相殺

$$\mathbb{E} \|\nabla \mathcal{L}(\bar{W}_\tau)\|_*^{[\rho]} \leq \frac{1}{DT} \left(\sum_{t=1}^T a_t \mathbb{E}[\text{Reg}_t^{[\beta]}(D)] + \Delta_{\mathcal{L}} \right) + \text{雑音項}$$

補題 (Jiang et al., 2026) ステップ幅 $D = (1 - \beta)\rho/(4\beta)$ の discounted O2NC は次を満たす：

$$\mathbb{E}\|\nabla\mathcal{L}(\bar{W}_T)\|_*^{[\rho]} \leq \underbrace{\frac{4\Delta_{\mathcal{L}}}{(1-\beta)\rho T}}_{\mathcal{L}(W_0) \text{ と下限の差}} + \underbrace{\frac{1}{T}\mathbb{E}\left[\text{Reg}_T^{[\beta]}(1) + (1-\beta)\sum_{t<T}\text{Reg}_t^{[\beta]}(1)\right]}_{\text{学習器の discounted regret}} \\ + \underbrace{\left(1-\beta + \frac{\beta}{T}\right)\frac{\sqrt{r}\sigma}{\sqrt{1-\beta^2}}}_{\text{雑音}}$$

$\text{Reg}_t^{[\beta]}(1) : \|U\|_{\text{op}} \leq 1$ に対する X_s/D の discounted regret で, $\text{Reg}_t^{[\beta]}(D) = D \cdot \text{Reg}_t^{[\beta]}(1)$.

証明：前のスライドまでの議論と，マルチンゲールによる雑音の評価（付録）

補題 (Jiang et al., 2026) ステップ幅 $D = (1 - \beta)\rho/(4\beta)$ の discounted O2NC は次を満たす：

$$\begin{aligned} \mathbb{E} \|\nabla \mathcal{L}(\bar{W}_T)\|_*^{[\rho]} &\leq \underbrace{\frac{4\Delta_{\mathcal{L}}}{(1-\beta)\rho T}}_{\mathcal{L}(W_0) \text{ と下限の差}} + \underbrace{\frac{1}{T} \mathbb{E} \left[\text{Reg}_T^{[\beta]}(1) + (1-\beta) \sum_{t < T} \text{Reg}_t^{[\beta]}(1) \right]}_{\text{学習器の discounted regret}} \\ &\quad + \underbrace{\left(1 - \beta + \frac{\beta}{T}\right) \frac{\sqrt{r}\sigma}{\sqrt{1-\beta^2}}}_{\text{雑音}} \end{aligned}$$

$\text{Reg}_t^{[\beta]}(1) : \|U\|_{\text{op}} \leq 1$ に対する X_s/D の discounted regret で, $\text{Reg}_t^{[\beta]}(D) = D \cdot \text{Reg}_t^{[\beta]}(1)$.

証明：前のスライドまでの議論と，マルチンゲールによる雑音の評価（付録）

- 学習器に依存するのは discounted regret だけ
- **すべての t で $\mathbb{E}[\text{Reg}_t^{[\beta]}(1)] \lesssim B/\sqrt{1-\beta}$** ($B \asymp \sqrt{r}(\Gamma + \sigma)$) なら，
 $1 - \beta \asymp (\varepsilon/B)^2$, $\boxed{T \asymp \Delta_{\mathcal{L}} B^2 / (\rho \varepsilon^3) + B^2 / \varepsilon^2}$ で上界は $\lesssim \varepsilon$

補題 (Jiang et al., 2026) ステップ幅 $D = (1 - \beta)\rho/(4\beta)$ の discounted O2NC は次を満たす：

$$\begin{aligned} \mathbb{E} \|\nabla \mathcal{L}(\bar{W}_T)\|_*^{[\rho]} &\leq \underbrace{\frac{4\Delta_{\mathcal{L}}}{(1-\beta)\rho T}}_{\mathcal{L}(W_0) \text{ と下限の差}} + \underbrace{\frac{1}{T} \mathbb{E} \left[\text{Reg}_T^{[\beta]}(1) + (1-\beta) \sum_{t < T} \text{Reg}_t^{[\beta]}(1) \right]}_{\text{学習器の discounted regret}} \\ &\quad + \underbrace{\left(1 - \beta + \frac{\beta}{T}\right) \frac{\sqrt{r}\sigma}{\sqrt{1-\beta^2}}}_{\text{雑音}} \end{aligned}$$

$\text{Reg}_t^{[\beta]}(1) : \|U\|_{\text{op}} \leq 1$ に対する X_s/D の discounted regret で, $\text{Reg}_t^{[\beta]}(D) = D \cdot \text{Reg}_t^{[\beta]}(1)$.

証明：前のスライドまでの議論と，マルチンゲールによる雑音の評価（付録）

- 学習器に依存するのは discounted regret だけ
- **すべての t で $\mathbb{E}[\text{Reg}_t^{[\beta]}(1)] \lesssim B/\sqrt{1-\beta}$** ($B \asymp \sqrt{r}(\Gamma + \sigma)$) なら，
 $1 - \beta \asymp (\varepsilon/B)^2$, $\boxed{T \asymp \Delta_{\mathcal{L}} B^2 / (\rho \varepsilon^3) + B^2 / \varepsilon^2}$ で上界は $\lesssim \varepsilon$

次：有限回の Newton-Schulz 反復を用いる Muon の discounted regret

発表の流れ

1. はじめに
2. 問題設定
3. Online-to-nonconvex conversion (O2NC)
4. **一般のスペクトル写像に対する discounted regret**
5. 有限回の Newton-Schulz 反復を用いる Muon
6. Newton-Schulz 反復以外の写像
7. まとめ

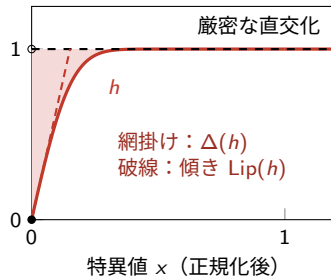
一般のスペクトル写像を用いる Muon

写像 h_q をスペクトル写像 h に置き換え，
正規化定数 $C > 0$ を固定 (Muon は $C = G_{\text{op}}$) :

$$X_{t+1} = -D \mathcal{H}_h(M_t/C) \quad \mathcal{H}_h(U\Sigma V^\top) = U \text{diag}(h(\sigma_i)) V^\top$$

仮定 写像 $h: [0, \infty) \rightarrow [0, 1]$ は
連続かつ非減少で，次を満たす：

- (i) $h(0) = 0$
- (ii) $\text{Lip}(h) := \sup_{x \neq y} |h(x) - h(y)| / |x - y| < \infty$
- (iii) $\Delta(h) := \int_0^\infty (1 - h(x)) \, dx < \infty$



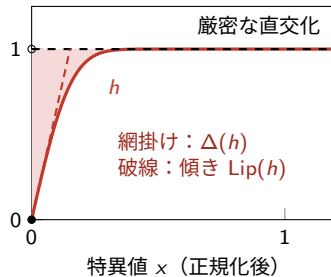
一般のスペクトル写像を用いる Muon

写像 h_q をスペクトル写像 h に置き換え，
正規化定数 $C > 0$ を固定 (Muon は $C = G_{\text{op}}$) :

$$X_{t+1} = -D \mathcal{H}_h(M_t/C) \quad \mathcal{H}_h(U\Sigma V^\top) = U \text{diag}(h(\sigma_i)) V^\top$$

仮定 写像 $h: [0, \infty) \rightarrow [0, 1]$ は
連続かつ非減少で，次を満たす：

- (i) $h(0) = 0$
- (ii) $\text{Lip}(h) := \sup_{x \neq y} |h(x) - h(y)| / |x - y| < \infty$
- (iii) $\Delta(h) := \int_0^\infty (1 - h(x)) \, dx < \infty$



$\Delta(h)$ ：厳密な直交化に対する h の近似誤差

厳密な直交化： $\Delta = 0$ だが $\text{Lip} = \infty$

有限回の Newton-Schulz 反復 h_q ($[1, \infty)$ 上で $h_q = 1$)：どちらも有限

定理

$$\mathbb{E}\left[\text{Reg}_t^{[\beta]}(D)\right] \leq D \left(\underbrace{\frac{rC \Delta(h)}{1-\beta}}_{\text{ペナルティ項}} + \underbrace{\frac{2\Gamma^2 \text{Lip}(h)}{C}}_{\text{安定性項}} \right)$$

$$\text{Reg}_t^{[\beta]}(D) = \max_{\|U\|_{\text{op}} \leq D} \sum_{s \leq t} \beta^{t-s} \langle G_s, X_s - U \rangle$$

定理

$$\mathbb{E}\left[\text{Reg}_t^{[\beta]}(D)\right] \leq D \left(\underbrace{\frac{rC \Delta(h)}{1-\beta}}_{\text{ペナルティ項}} + \underbrace{\frac{2\Gamma^2 \text{Lip}(h)}{C}}_{\text{安定性項}} \right)$$

$$\text{Reg}_t^{[\beta]}(D) = \max_{\|U\|_{\text{op}} \leq D} \sum_{s \leq t} \beta^{t-s} \langle G_s, X_s - U \rangle$$

follow-the-regularized-leader (FTRL) やオンライン鏡像降下法の解析と同様の、ペナルティ項と安定性項のトレードオフ (Shalev-Shwartz, 2012; McMahan, 2017)

定理

$$\mathbb{E}\left[\text{Reg}_t^{[\beta]}(D)\right] \leq D \left(\underbrace{\frac{rC \Delta(h)}{1-\beta}}_{\text{ペナルティ項}} + \underbrace{\frac{2\Gamma^2 \text{Lip}(h)}{C}}_{\text{安定性項}} \right)$$

$$\text{Reg}_t^{[\beta]}(D) = \max_{\|U\|_{\text{op}} \leq D} \sum_{s \leq t} \beta^{t-s} \langle G_s, X_s - U \rangle$$

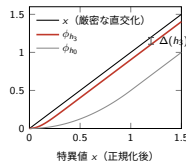
follow-the-regularized-leader (FTRL) やオンライン鏡像降下法の解析と同様の、ペナルティ項と安定性項のトレードオフ (Shalev-Shwartz, 2012; McMahan, 2017)

有限回の Newton-Schulz 反復： q を増やすと $\Delta(h_q)$ は減少， $\text{Lip}(h_q)$ は増加
反復回数 q でペナルティ項と安定性項の釣り合いを調整
有限回の Newton-Schulz 反復を用いる Muon は，それ自体が FTRL (後述)

証明の方針：更新はポテンシャルの勾配

平滑化したポテンシャル

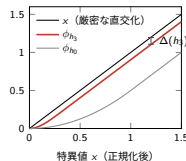
$$\tilde{\Phi}(M) := C \sum_{i=1}^r \phi_h(\sigma_i(M)/C), \quad \phi_h(x) := \int_0^x h(u) du$$



証明の方針：更新はポテンシャルの勾配

平滑化したポテンシャル

$$\tilde{\Phi}(M) := C \sum_{i=1}^r \phi_h(\sigma_i(M)/C), \quad \phi_h(x) := \int_0^x h(u) du$$



補題 $\tilde{\Phi}$ は凸かつ連続微分可能で $\nabla \tilde{\Phi} = \mathcal{H}_h(\cdot/C)$. よって更新は $X_{t+1} = -D \mathcal{H}_h(M_t/C) = -D \nabla \tilde{\Phi}(M_t)$: モメンタムでのポテンシャルの勾配

証明. 一般に ϕ が凸・微分可能で $\phi'(0) = 0$ なら, $F(S) := \sum_i \phi(\sigma_i(S))$ も凸・微分可能で,

$$\nabla F(U\Sigma V^\top) = U \text{diag}(\phi'(\sigma_i(S))) V^\top.$$

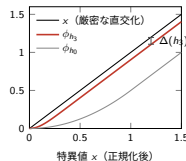
$\phi = \phi_h$ にこれを適用すれば良い.

ポテンシャルの勾配を選ぶ学習器: gradient-based prediction algorithm (Abernethy et al., 2014)

証明の方針：更新はポテンシャルの勾配

平滑化したポテンシャル

$$\tilde{\Phi}(M) := C \sum_{i=1}^r \phi_h(\sigma_i(M)/C), \quad \phi_h(x) := \int_0^x h(u) du$$



補題 $\tilde{\Phi}$ は凸かつ連続微分可能で $\nabla \tilde{\Phi} = \mathcal{H}_h(\cdot/C)$. よって更新は $X_{t+1} = -D \mathcal{H}_h(M_t/C) = -D \nabla \tilde{\Phi}(M_t)$: モメンタムでのポテンシャルの勾配

証明. 一般に ϕ が凸・微分可能で $\phi'(0) = 0$ なら, $F(S) := \sum_i \phi(\sigma_i(S))$ も凸・微分可能で,

$$\nabla F(U\Sigma V^\top) = U \text{diag}(\phi'(\sigma_i(S))) V^\top.$$

$\phi = \phi_h$ にこれを適用すれば良い.

ポテンシャルの勾配を選ぶ学習器: gradient-based prediction algorithm (Abernethy et al., 2014)

exact-polar Muon: $\text{polar}(M) \in \partial \|M\|_*$ なので, $M \mapsto \|M\|_*$ が対応するポテンシャル
 $(h = \mathbf{1}\{x > 0\})$ では $\phi_h(x) = x$, $\tilde{\Phi} = \|\cdot\|_*$

証明：モメンタムによるリグレット分解

補題 ポテンシャル $\tilde{\Phi}$ の Bregman ダイバージェンスを $B_{\tilde{\Phi}}$ とすると,

$$\text{Reg}_t^{[\beta]}(D) = \frac{D}{1-\beta} \left(\underbrace{\|M_t\|_* - \tilde{\Phi}(M_t)}_{\text{ペナルティ項}} + \underbrace{\sum_{s=1}^t \beta^{t-s} B_{\tilde{\Phi}}(M_s \| M_{s-1})}_{\text{安定性項}} + \underbrace{\sum_{s=1}^{t-1} \beta^{t-s-1} V_s}_{\leq 0} \right)$$

ただし $V_s := (1-\beta)(\tilde{\Phi}(M_s) - \langle M_s, \nabla \tilde{\Phi}(M_s) \rangle)$ とする.

証明：モメンタムによるリグレット分解

補題 ポテンシャル $\tilde{\Phi}$ の Bregman ダイバージェンスを $B_{\tilde{\Phi}}$ とすると,

$$\text{Reg}_t^{[\beta]}(D) = \frac{D}{1-\beta} \left(\underbrace{\|M_t\|_* - \tilde{\Phi}(M_t)}_{\text{ペナルティ項}} + \underbrace{\sum_{s=1}^t \beta^{t-s} B_{\tilde{\Phi}}(M_s \| M_{s-1})}_{\text{安定性項}} + \underbrace{\sum_{s=1}^{t-1} \beta^{t-s-1} V_s}_{\leq 0} \right)$$

ただし $V_s := (1-\beta)(\tilde{\Phi}(M_s) - \langle M_s, \nabla \tilde{\Phi}(M_s) \rangle)$ とする.

方針：

1. 双対性より $\text{Reg}_t^{[\beta]}(D) = \sum_{s \leq t} \beta^{t-s} \langle G_s, X_s \rangle + \frac{D}{1-\beta} \|M_t\|_*$ (モメンタムの恒等式)
2. 勾配を $(1-\beta)G_s = (M_s - M_{s-1}) + (1-\beta)M_{s-1}$ と分解：Bregman ダイバージェンスと $\tilde{\Phi}(M_s) - \tilde{\Phi}(M_{s-1})$ が現れ、後者の割引和を整理した残りが V_s

証明：モメンタムによるリグレット分解

補題 ポテンシャル $\tilde{\Phi}$ の Bregman ダイバージェンスを $B_{\tilde{\Phi}}$ とすると,

$$\text{Reg}_t^{[\beta]}(D) = \frac{D}{1-\beta} \left(\underbrace{\|M_t\|_* - \tilde{\Phi}(M_t)}_{\text{ペナルティ項}} + \underbrace{\sum_{s=1}^t \beta^{t-s} B_{\tilde{\Phi}}(M_s \| M_{s-1})}_{\text{安定性項}} + \underbrace{\sum_{s=1}^{t-1} \beta^{t-s-1} V_s}_{\leq 0} \right)$$

ただし $V_s := (1-\beta)(\tilde{\Phi}(M_s) - \langle M_s, \nabla \tilde{\Phi}(M_s) \rangle)$ とする.

方針：

1. 双対性より $\text{Reg}_t^{[\beta]}(D) = \sum_{s \leq t} \beta^{t-s} \langle G_s, X_s \rangle + \frac{D}{1-\beta} \|M_t\|_*$ (モメンタムの恒等式)
2. 勾配を $(1-\beta)G_s = (M_s - M_{s-1}) + (1-\beta)M_{s-1}$ と分解：Bregman ダイバージェンスと $\tilde{\Phi}(M_s) - \tilde{\Phi}(M_{s-1})$ が現れ、後者の割引和を整理した残りが V_s

累積勾配による分解 (Abernethy et al., 2014; Jiang et al., 2026) と異なり、**モメンタム**で分解
これにより、雑音のない平滑な目的関数で最適な ε^{-2} の依存性を達成

証明：ペナルティ項

$$\tilde{\Phi}(M) = C \sum_{i=1}^r \phi_h(\sigma_i(M)/C), \quad \phi_h(x) = \int_0^x h(u) \, du$$

核ノルムは $\|M\|_* = \sum_{i=1}^r \sigma_i(M)$ なので

$$\begin{aligned} \|M\|_* - \tilde{\Phi}(M) &= \sum_{i=1}^r \left(\sigma_i(M) - C \phi_h(\sigma_i(M)/C) \right) = C \sum_{i=1}^r \int_0^{\sigma_i(M)/C} (1 - h(u)) \, du \\ &\leq C \sum_{i=1}^r \int_0^\infty (1 - h(u)) \, du = rC \Delta(h) \end{aligned}$$

また $h \leq 1$ なので、この差は非負

証明：ペナルティ項

$$\tilde{\Phi}(M) = C \sum_{i=1}^r \phi_h(\sigma_i(M)/C), \quad \phi_h(x) = \int_0^x h(u) du$$

核ノルムは $\|M\|_* = \sum_{i=1}^r \sigma_i(M)$ なので

$$\begin{aligned} \|M\|_* - \tilde{\Phi}(M) &= \sum_{i=1}^r \left(\sigma_i(M) - C \phi_h(\sigma_i(M)/C) \right) = C \sum_{i=1}^r \int_0^{\sigma_i(M)/C} (1 - h(u)) du \\ &\leq C \sum_{i=1}^r \int_0^\infty (1 - h(u)) du = rC \Delta(h) \end{aligned}$$

また $h \leq 1$ なので、この差は非負

ペナルティ項：
$$\frac{D}{1-\beta} (\|M_t\|_* - \tilde{\Phi}(M_t)) \leq \frac{D rC \Delta(h)}{1-\beta}$$

証明：安定性項

事実. 写像 h が Lipschitz で $h(0) = 0$ なら $\|\mathcal{H}_h(X) - \mathcal{H}_h(Y)\|_F \leq \text{Lip}(h) \|X - Y\|_F$.
 よって $\nabla\tilde{\Phi} = \mathcal{H}_h(\cdot/C)$ は $\|\cdot\|_F$ に関して $(\text{Lip}(h)/C)$ -Lipschitz.

安定性項.

$$\begin{aligned} \mathbb{E} B_{\tilde{\Phi}}(M_s \| M_{s-1}) &\leq \frac{\text{Lip}(h)}{2C} \mathbb{E} \|M_s - M_{s-1}\|_F^2 \\ &\leq \frac{2 \text{Lip}(h)(1 - \beta)^2 \Gamma^2}{C} \end{aligned}$$

$\nabla\tilde{\Phi}$ が L -Lipschitz なら
 $B_{\tilde{\Phi}}(X \| Y) \leq \frac{L}{2} \|X - Y\|_F^2$

$\mathbb{E} \|M_s - M_{s-1}\|_F^2 \leq 4(1 - \beta)^2 \Gamma^2$

証明：安定性項

事実. 写像 h が Lipschitz で $h(0) = 0$ なら $\|\mathcal{H}_h(X) - \mathcal{H}_h(Y)\|_F \leq \text{Lip}(h) \|X - Y\|_F$.
 よって $\nabla\tilde{\Phi} = \mathcal{H}_h(\cdot/C)$ は $\|\cdot\|_F$ に関して $(\text{Lip}(h)/C)$ -Lipschitz.

安定性項. $\nabla\tilde{\Phi}$ が L -Lipschitz なら $B_{\tilde{\Phi}}(X\|Y) \leq \frac{L}{2} \|X - Y\|_F^2$

$$\begin{aligned} \mathbb{E}B_{\tilde{\Phi}}(M_s\|M_{s-1}) &\leq \frac{\text{Lip}(h)}{2C} \mathbb{E}\|M_s - M_{s-1}\|_F^2 \\ &\leq \frac{2\text{Lip}(h)(1-\beta)^2\Gamma^2}{C} \end{aligned}$$

重み β^{t-s} で足し合わせると $((1-\beta)\sum_{s=1}^t \beta^{t-s} \leq 1)$

$$\text{安定性項： } \frac{D}{1-\beta} \sum_{s=1}^t \beta^{t-s} \mathbb{E}B_{\tilde{\Phi}}(M_s\|M_{s-1}) \leq D \frac{2\Gamma^2 \text{Lip}(h)}{C}$$

証明：安定性項

事実. 写像 h が Lipschitz で $h(0) = 0$ なら $\|\mathcal{H}_h(X) - \mathcal{H}_h(Y)\|_F \leq \text{Lip}(h) \|X - Y\|_F$.
 よって $\nabla\tilde{\Phi} = \mathcal{H}_h(\cdot/C)$ は $\|\cdot\|_F$ に関して $(\text{Lip}(h)/C)$ -Lipschitz.

安定性項. $\nabla\tilde{\Phi}$ が L -Lipschitz なら $B_{\tilde{\Phi}}(X\|Y) \leq \frac{L}{2} \|X - Y\|_F^2$

$$\begin{aligned} \mathbb{E}B_{\tilde{\Phi}}(M_s\|M_{s-1}) &\leq \frac{\text{Lip}(h)}{2C} \mathbb{E}\|M_s - M_{s-1}\|_F^2 \\ &\leq \frac{2\text{Lip}(h)(1-\beta)^2\Gamma^2}{C} \end{aligned}$$

重み β^{t-s} で足し合わせると $((1-\beta)\sum_{s=1}^t \beta^{t-s} \leq 1)$

$$\text{安定性項: } \frac{D}{1-\beta} \sum_{s=1}^t \beta^{t-s} \mathbb{E}B_{\tilde{\Phi}}(M_s\|M_{s-1}) \leq D \frac{2\Gamma^2 \text{Lip}(h)}{C}$$

最後の項. $\tilde{\Phi}$ は凸で $\tilde{\Phi}(0) = 0$ なので,

$$0 = \tilde{\Phi}(0) \geq \tilde{\Phi}(M_s) - \langle M_s, \nabla\tilde{\Phi}(M_s) \rangle = \frac{V_s}{1-\beta}.$$

発表の流れ

1. はじめに
2. 問題設定
3. Online-to-nonconvex conversion (O2NC)
4. 一般のスペクトル写像に対する discounted regret
5. **有限回の Newton–Schulz 反復を用いる Muon**
6. Newton–Schulz 反復以外の写像
7. まとめ

スペクトル写像としての有限回 Newton–Schulz 反復

写像 $f(x) = \frac{15}{8}x - \frac{5}{4}x^3 + \frac{3}{8}x^5$ に対し,
 $[0, 1]$ 上で $h_q = f \circ q$, $[1, \infty)$ 上で $h_q = 1$:
 連続かつ非減少で, $h_q(0) = 0$

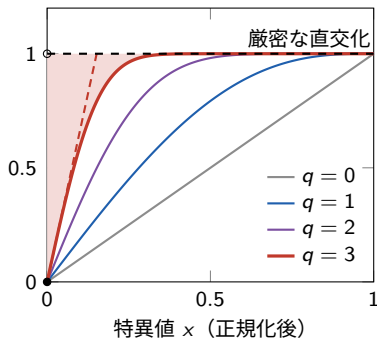
補題 定数 $A := f'(0) = 15/8$ に対し,

$$\text{Lip}(h_q) = A^q, \quad \frac{1}{2A^q} \leq \Delta(h_q) \leq \frac{1}{A^q + 1}$$

$$\Delta(h) = \int_0^\infty (1 - h(x)) \, dx$$

反復回数を増やすと, ギャップは指数的に縮小,
 傾きは指数的に増大
 一様な誤差は縮まない:

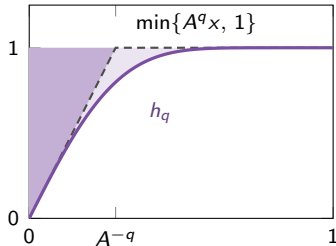
任意の q で $\sup_{x>0} (1 - h_q(x)) = 1$



網掛け: $\Delta(h_3)$, 破線: h_3 の 0 での傾き A^3

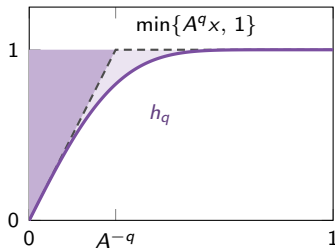
下界. $\text{Lip}(h_q) = A^q$ より各 $x \geq 0$ で
 $h_q(x) \leq \min\{A^q x, 1\}$ なので,
ギャップは三角形を含む:

$$\Delta(h_q) \geq \int_0^{A^{-q}} (1 - A^q x) dx = \frac{1}{2A^q}$$



下界. $\text{Lip}(h_q) = A^q$ より各 $x \geq 0$ で
 $h_q(x) \leq \min\{A^q x, 1\}$ なので,
 ギャップは三角形を含む:

$$\Delta(h_q) \geq \int_0^{A^{-q}} (1 - A^q x) dx = \frac{1}{2A^q}$$



上界 (方針). 関数 $g_L(x) := 1 - (1 - x)^L$ は $f \geq g_A$ かつ $g_L \circ g_{L'} = g_{LL'}$ を満たすので,
 f の単調性より $[0, 1]$ 上で $h_q = f^{\circ q} \geq g_A^{\circ q} = g_{A^q}$ となり

$$\Delta(h_q) \leq \int_0^1 (1 - g_{A^q}(x)) dx = \int_0^1 (1 - x)^{A^q} dx = \frac{1}{A^q + 1}$$

有限回 Newton–Schulz 反復での discounted regret

一般の h に対する定理 ($C = G_{\text{op}}$) と補題 ($\Delta(h_q) \leq A^{-q}$, $\text{Lip}(h_q) = A^q$) より

$$\begin{aligned} \mathbb{E}[\text{Reg}_t^{[\beta]}(D)] &\leq D \left(\underbrace{\frac{rG_{\text{op}} \Delta(h_q)}{1 - \beta}}_{\text{ペナルティ項}} + \underbrace{\frac{2\Gamma^2 \text{Lip}(h_q)}{G_{\text{op}}}}_{\text{安定性項}} \right) \\ &\leq D \left(\frac{rG_{\text{op}}}{(1 - \beta) A^q} + \frac{2 A^q \Gamma^2}{G_{\text{op}}} \right) \end{aligned}$$

反復回数 q を増やすと、ペナルティ項は減少、安定性項は増加
 $q \rightarrow \infty$ で安定性項は発散：厳密な直交化が 0 で不連続であることの反映

有限回 Newton–Schulz 反復での discounted regret

一般の h に対する定理 ($C = G_{\text{op}}$) と補題 ($\Delta(h_q) \leq A^{-q}$, $\text{Lip}(h_q) = A^q$) より

$$\begin{aligned} \mathbb{E}[\text{Reg}_t^{[\beta]}(D)] &\leq D \left(\underbrace{\frac{rG_{\text{op}} \Delta(h_q)}{1-\beta}}_{\text{ペナルティ項}} + \underbrace{\frac{2\Gamma^2 \text{Lip}(h_q)}{G_{\text{op}}}}_{\text{安定性項}} \right) \\ &\leq D \left(\frac{rG_{\text{op}}}{(1-\beta) A^q} + \frac{2 A^q \Gamma^2}{G_{\text{op}}} \right) \end{aligned}$$

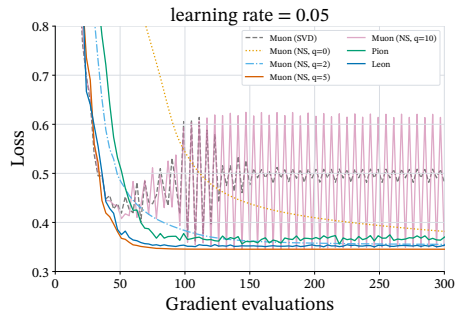
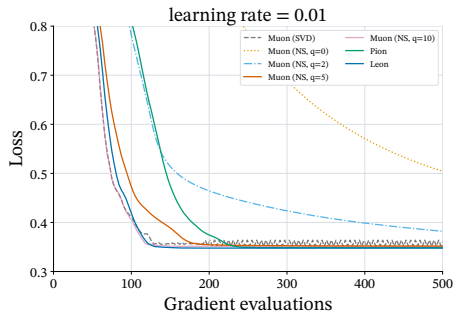
反復回数 q を増やすと、ペナルティ項は減少、安定性項は増加
 $q \rightarrow \infty$ で安定性項は発散：厳密な直交化が 0 で不連続であることの反映
 2 項を釣り合わせると：

定理

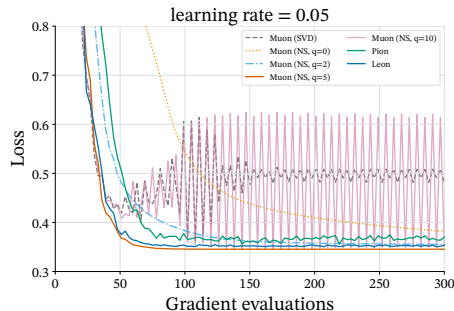
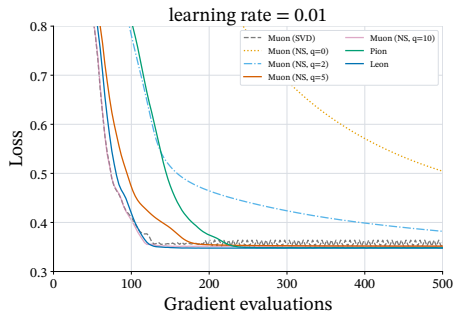
$$q = \left\lceil \log_A \frac{G_{\text{op}} \sqrt{r}}{\Gamma \sqrt{1-\beta}} \right\rceil \text{ とすると } \mathbb{E}[\text{Reg}_t^{[\beta]}(D)] \leq (1+2A) D \Gamma \sqrt{\frac{r}{1-\beta}}$$

D で割ると、すべての t で $\mathbb{E}[\text{Reg}_t^{[\beta]}(1)] \lesssim B/\sqrt{1-\beta}$: O2NC の保証に必要な条件

非凸・非平滑な問題 (Jiang et al., 2026 の変種), $W \in \mathbb{R}^{20 \times 20}$, $\beta = 0.9$, 学習率 = D



非凸・非平滑な問題（Jiang et al., 2026 の変種）， $W \in \mathbb{R}^{20 \times 20}$ ， $\beta = 0.9$ ，学習率 = D



- 学習率 0.01（左）：反復回数が少ない（ $q = 0, 2$ ）と低速
- 学習率 0.05（右）：exact-polar Muon（SVD）と $q = 10$ は振動， $q = 0$ は低速
高速かつ安定なのは $q = 2, 5$

ペナルティ項と安定性項のトレードオフと整合
（反復点での勾配を使用，ベンチマークではない）

主結果：非凸・非平滑な目的関数

定理 定数を $B := \sqrt{r}((1+2A)\Gamma + \sigma) \asymp \sqrt{r}(\Gamma + \sigma)$ とし, $\varepsilon \leq B$ とする. 有限回の Newton–Schulz 反復を用いる Muon を学習器として, discounted O2NC を

$$1 - \beta = \left(\frac{\varepsilon}{3B}\right)^2, \quad D = \frac{(1 - \beta)\rho}{4\beta}, \quad q = \left\lceil \log_A \frac{3BG_{\text{op}}\sqrt{r}}{\Gamma\varepsilon} \right\rceil$$

で実行する. このとき

$$T = O\left(\frac{\Delta_{\mathcal{L}} r(\Gamma + \sigma)^2}{\rho\varepsilon^3} + \frac{r(\Gamma + \sigma)^2}{\varepsilon^2}\right)$$

個の確率的勾配で, 出力 $\bar{W}_T$ は期待値の意味で (ρ, ε) -停留点となる.

$\bar{W}_T$: discounted O2NC の出力 (勾配評価点の重み付き平均)

主結果：非凸・非平滑な目的関数

定理 定数を $B := \sqrt{r}((1+2A)\Gamma + \sigma) \asymp \sqrt{r}(\Gamma + \sigma)$ とし, $\varepsilon \leq B$ とする. 有限回の Newton-Schulz 反復を用いる Muon を学習器として, discounted O2NC を

$$1 - \beta = \left(\frac{\varepsilon}{3B}\right)^2, \quad D = \frac{(1 - \beta)\rho}{4\beta}, \quad q = \left\lceil \log_A \frac{3BG_{\text{op}}\sqrt{r}}{\Gamma\varepsilon} \right\rceil$$

で実行する. このとき

$$T = O\left(\frac{\Delta_{\mathcal{L}} r(\Gamma + \sigma)^2}{\rho\varepsilon^3} + \frac{r(\Gamma + \sigma)^2}{\varepsilon^2}\right)$$

個の確率的勾配で, 出力 $\bar{W}_T$ は期待値の意味で (ρ, ε) -停留点となる.

$\bar{W}_T$: discounted O2NC の出力 (勾配評価点の重み付き平均)

証明: discounted regret $O(\Gamma\sqrt{r/(1-\beta)}) + \text{O2NC の保証 } (1 - \beta \asymp (\varepsilon/B)^2)$

比較：非凸・非平滑な目的関数

各手法が (ρ, ε) -停留点を得るのに必要な確率的勾配の数（NS：Newton–Schulz 反復）：

手法	平滑化	直交化	確率的勾配の数
exact-polar Muon	なし	厳密	discounted O2NC の保証の対象外
Pion (Jiang et al., 2026)	ガウス摂動	厳密	$O\left(\max\left\{\frac{\nu^2 \ Q\ _*^2 \Delta_{\mathcal{L}}}{\rho \varepsilon^3}, \frac{\nu^2 \ Q\ _*^2}{\varepsilon^2}, \frac{\nu r G}{\varepsilon}\right\}\right)$
Leon (Jiang et al., 2026)	hyperbolic smoothing	厳密	$O\left(\max\left\{\frac{\ Q\ _*^2 \Delta_{\mathcal{L}}}{\rho \varepsilon^3}, \frac{\ Q\ _*^2}{\varepsilon^2}, \frac{r G}{\varepsilon}\right\}\right)$
本研究	有限回 NS	NS 反復の回数 $q = O(\log(1/\varepsilon))$	$O\left(\max\left\{\frac{r(\Gamma + \sigma)^2 \Delta_{\mathcal{L}}}{\rho \varepsilon^3}, \frac{r(\Gamma + \sigma)^2}{\varepsilon^2}\right\}\right)$

- 主要項 $\rho^{-1}\varepsilon^{-3}$ は共通, Pion と Leon の仮定は $\mathbb{E}[G_t G_t^\top] \preceq Q^2$ と $\|G_t\|_{\text{op}} \leq G$ (確率 1 で) 定数 $\|Q\|_*^2$ と $r(\Gamma + \sigma)^2$ は異なる仮定に由来し, 直接比較は不可能
- Pion の上界は, polar factor の期待値を厳密に計算できることを仮定
Monte Carlo 版でこの上界を保つには, 各ラウンドで T 個のサンプルが必要
- Pion と Leon の上界は ε が十分小さいときに成立 ($\nu = (n/(n-r-1))^{1/4}$)

平滑な非凸目的関数

仮定 さらに, 任意の W, W' に対し $\|\nabla\mathcal{L}(W) - \nabla\mathcal{L}(W')\|_* \leq L_{\text{op}}\|W - W'\|_{\text{op}}$.

定理 目標精度を $\varepsilon \leq 12rG_{\text{op}}$ とする. Muon を学習器として, $q = \lceil \log_A(12rG_{\text{op}}/\varepsilon) \rceil$ と適切な β, D で discounted O2NC を実行すると,

$$T = O\left(\frac{L_{\text{op}}\Delta_{\mathcal{L}}}{\varepsilon^2} + \frac{r\sigma^2}{\varepsilon^2} + \frac{r\sigma^2 L_{\text{op}}\Delta_{\mathcal{L}}}{\varepsilon^4}\right)$$

個の確率的勾配で $\mathbb{E}\|\nabla\mathcal{L}(\bar{W}_T)\|_* \leq \varepsilon$ となる.

精度 ε について最適: 雑音なし ε^{-2} (Carmon et al., 2020), 雑音あり ε^{-4} (Arjevani et al., 2023)

平滑な非凸目的関数

仮定 さらに, 任意の W, W' に対し $\|\nabla\mathcal{L}(W) - \nabla\mathcal{L}(W')\|_* \leq L_{\text{op}}\|W - W'\|_{\text{op}}$.

定理 目標精度を $\varepsilon \leq 12rG_{\text{op}}$ とする. Muon を学習器として, $q = \lceil \log_A(12rG_{\text{op}}/\varepsilon) \rceil$ と適切な β, D で discounted O2NC を実行すると,

$$T = O\left(\frac{L_{\text{op}}\Delta_{\mathcal{L}}}{\varepsilon^2} + \frac{r\sigma^2}{\varepsilon^2} + \frac{r\sigma^2 L_{\text{op}}\Delta_{\mathcal{L}}}{\varepsilon^4}\right)$$

個の確率的勾配で $\mathbb{E}\|\nabla\mathcal{L}(\bar{W}_T)\|_* \leq \varepsilon$ となる.

精度 ε について最適: 雑音なし ε^{-2} (Carmon et al., 2020), 雑音あり ε^{-4} (Arjevani et al., 2023)

証明の考え方:

- (ρ, ε) から ε へ: $\|\nabla\mathcal{L}(W)\|_* \leq \|\nabla\mathcal{L}(W)\|_*^{[\rho]} + L_{\text{op}}\rho$
ただし, $\rho = \varepsilon/L_{\text{op}}$ として非平滑な場合の上界を使うだけでは, 雑音がなくても ε^{-4}
- 安定性項: $M_s - M_{s-1}$ を信号と雑音に分解し, 信号は $\text{Lip}(h_q)$ によらずに評価
モメンタムが勾配に追従するため信号は小さく (平滑性), A^q が掛かるのは σ^2 だけ

比較：平滑な非凸目的関数

文献	決定論的	確率的	直交化
Kovalev (2025)	$O\left(\frac{L_{\text{op}}\Delta_{\mathcal{L}}}{\varepsilon^2}\right)$	$O\left(\max\left\{\frac{L_{\text{op}}\Delta_{\mathcal{L}}}{\varepsilon^2}, \frac{\sqrt{r}\sigma}{\varepsilon}, \frac{r^{3/2}\sigma^3}{\varepsilon^3}, \frac{r\sigma^2 L_{\text{op}}\Delta_{\mathcal{L}}}{\varepsilon^4}\right\}\right)$	厳密
Shen et al. (2026)	$O\left(\frac{L_{\text{op}}\Delta_{\mathcal{L}}}{\varepsilon^2}\right)$	$O\left(\max\left\{\frac{L_{\text{op}}\Delta_{\mathcal{L}}}{\varepsilon^2}, \frac{r^2\sigma^4}{L_{\text{op}}\Delta_{\mathcal{L}}\varepsilon^2}, \frac{r\sigma^2 L_{\text{op}}\Delta_{\mathcal{L}}}{\varepsilon^4}\right\}\right)$	厳密
Shulgin et al. (2026)	$O\left(\frac{(1+\delta)^2 L_{\text{op}}\Delta_{\mathcal{L}}}{(1-\delta)^2 \varepsilon^2}\right)$	$O\left(\frac{(1+\delta)r\sigma^2 L_{\text{op}}\Delta_{\mathcal{L}}}{(1-\delta)^4 \varepsilon^4}\right)$	近似 LMO
Kim and Oh (2026)	$O\left(\frac{\chi_q^2 L_{\text{op}}\Delta_{\mathcal{L}}}{\varepsilon^2}\right)$	$O\left(\max\left\{\frac{\chi_q^2 L_{\text{op}}\Delta_{\mathcal{L}}}{\varepsilon^2}, \frac{\chi_q^2 r^2 \sigma^2}{\varepsilon^2}, \frac{\chi_q^4 r\sigma^2 L_{\text{op}}\Delta_{\mathcal{L}}}{\varepsilon^4}\right\}\right)$	有限回 NS
本研究	$O\left(\frac{L_{\text{op}}\Delta_{\mathcal{L}}}{\varepsilon^2}\right)$	$O\left(\max\left\{\frac{L_{\text{op}}\Delta_{\mathcal{L}}}{\varepsilon^2}, \frac{r\sigma^2}{\varepsilon^2}, \frac{r\sigma^2 L_{\text{op}}\Delta_{\mathcal{L}}}{\varepsilon^4}\right\}\right)$	有限回 NS

- 主要項は exact-polar Muon の上界と一致し, Kim and Oh (2026) と比べて因子 χ_q がなく, $r^2\sigma^2/\varepsilon^2$ が $r\sigma^2/\varepsilon^2$ に改善 (アルゴリズムは少し異なる)
- $\delta < 1$: 近似的な線形最小化オラクル (LMO) の一様誤差 $\chi_q = (1 - \varepsilon_q)^{-1}$, ε_q : q 回の NS 反復の一様誤差
- 本研究は $\|G_t\|_{\text{op}} \leq G_{\text{op}}$ (確率 1 で) と, Frobenius ノルムでの二次モーメントの上界も仮定

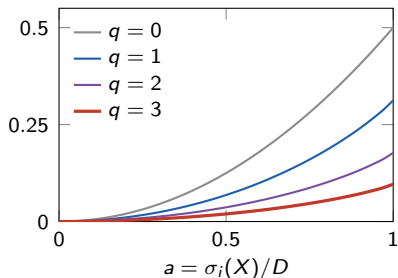
定理 関数 ϕ_h ($\phi_h(x) = \int_0^x h(u) du$) の Fenchel 共役 ϕ_h^* を用いて, 正則化関数を $R_{h,c}(X) := DC \sum_i \phi_h^*(\sigma_i(X)/D)$ とすると,

$$X_{t+1} \in \arg \min_{\|X\|_{\text{op}} \leq D} \left\{ \sum_{s=1}^t \beta^{t-s} \langle G_s, X \rangle + \frac{R_{h,c}(X)}{1 - \beta} \right\}$$

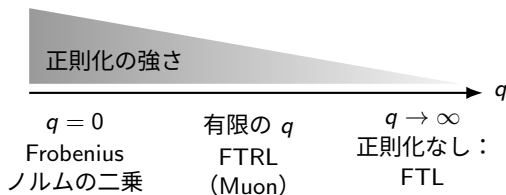
有限回 Newton-Schulz 反復を用いる Muon は FTRL

定理 関数 ϕ_h ($\phi_h(x) = \int_0^x h(u) du$) の Fenchel 共役 ϕ_h^* を用いて, 正則化関数を $R_{h,C}(X) := DC \sum_i \phi_h^*(\sigma_i(X)/D)$ とすると,

$$X_{t+1} \in \arg \min_{\|X\|_{\text{op}} \leq D} \left\{ \sum_{s=1}^t \beta^{t-s} \langle G_s, X \rangle + \frac{R_{h,C}(X)}{1-\beta} \right\}$$



特異値ごとの正則化関数 $\phi_{h_q}^*(a)$



反復回数で正則化の強さを調整

発表の流れ

1. はじめに
2. 問題設定
3. Online-to-nonconvex conversion (O2NC)
4. 一般のスペクトル写像に対する discounted regret
5. 有限回の Newton-Schulz 反復を用いる Muon
6. **Newton-Schulz 反復以外の写像**
7. まとめ

ギャップと傾きのトレードオフは避けられない

任意の h ($0 \leq h \leq 1$, $h(0) = 0$) は
 $h(x) \leq \min\{\text{Lip}(h)x, 1\}$ を満たすので、
 ギャップは三角形を含む：

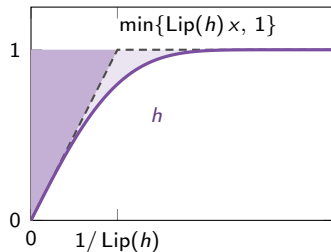
$$\Delta(h) \geq \frac{1}{2 \text{Lip}(h)}$$

ギャップを小さくするには、大きい傾きが必要

ギャップ $\Delta(h) = \int_0^\infty (1 - h(x)) dx$, 傾き $\text{Lip}(h)$

有限回の反復は $\Delta(h_q) \leq 1/\text{Lip}(h_q)$:
 高々 2 倍の差で最適

任意の次数 $\kappa \geq 1$ の Taylor 多項式による
 Newton-Schulz 反復でも同様 (Muon は $\kappa = 2$) :
 A を, 次数 κ の多項式 f_κ の 0 での傾き
 $A_\kappa = f'_\kappa(0)$ に置き換えるだけ



薄い部分：ギャップ $\Delta(h)$, 濃い部分：三角形

別のスペクトル写像：SoftMuon

SoftMuon (Feoktistov et al., 2026)：同じモメンタム $M_t = U \text{diag}(\sigma_i) V^\top$ に対し

$$X_{t+1} = -D M_t (M_t^\top M_t + a^{-2} I)^{-1/2} = -D U \text{diag}(\tilde{h}_a(\sigma_i)) V^\top, \quad \tilde{h}_a(x) = \frac{ax}{\sqrt{1 + a^2 x^2}}$$

ただし $a > 0$ は平滑化のパラメータ

Mustafi et al. (2026) の regularized Muon も同じ写像

別のスペクトル写像：SoftMuon

SoftMuon (Feoktistov et al., 2026)：同じモメンタム $M_t = U \text{diag}(\sigma_i) V^\top$ に対し

$$X_{t+1} = -D M_t (M_t^\top M_t + a^{-2} I)^{-1/2} = -D U \text{diag}(\tilde{h}_a(\sigma_i)) V^\top, \quad \tilde{h}_a(x) = \frac{ax}{\sqrt{1 + a^2 x^2}}$$

ただし $a > 0$ は平滑化のパラメータ

Mustafi et al. (2026) の regularized Muon も同じ写像

本研究の仮定を満たし、傾きとギャップは

$$\text{Lip}(\tilde{h}_a) = a, \quad \Delta(\tilde{h}_a) = \frac{1}{a}$$

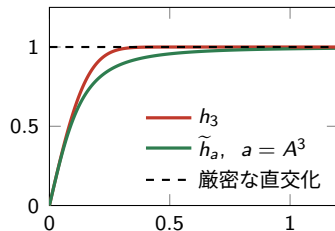
よって同じ保証が成立：

反復回数 q の代わりに a で調整

SoftMuon は学習中に a を変更し、反復点での勾配を使用

本研究の保証は、 a を固定して

discounted O2NC の中で使う場合

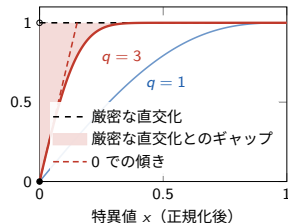


特異値 x (正規化後)

0 での傾きは同じ
ギャップが小さいのは h_3

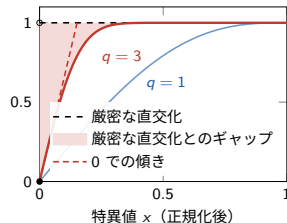
有限回の Newton-Schulz 反復は、誤差ではなく平滑化として働く：
 Muon を follow-the-leader から follow-the-regularized-leader へ

1. **平滑化.** 0 での跳びを有限の傾きに緩和
 exact-polar Muon ($q = \infty$) は FTL
2. **トレードオフ.** discounted regret $\leq$
 ペナルティ項 $\propto \Delta(h)$ + 安定性項 $\propto \text{Lip}(h)$
 反復回数 q で両者の釣り合いを調整
3. **保証.** O2NC の中で $q = O(\log(1/\varepsilon))$:
 $O(\rho^{-1}\varepsilon^{-3} + \varepsilon^{-2})$ 個の確率的勾配で (ρ, ε) -停留点
 平滑な目的関数では ε について最適



有限回の Newton-Schulz 反復は、誤差ではなく平滑化として働く：
Muon を follow-the-leader から follow-the-regularized-leader へ

1. **平滑化.** 0 での跳びを有限の傾きに緩和
exact-polar Muon ($q = \infty$) は FTL
2. **トレードオフ.** discounted regret $\leq$
ペナルティ項 $\propto \Delta(h)$ + 安定性項 $\propto \text{Lip}(h)$
反復回数 q で両者の釣り合いを調整
3. **保証.** O2NC の中で $q = O(\log(1/\varepsilon))$ ：
 $O(\rho^{-1}\varepsilon^{-3} + \varepsilon^{-2})$ 個の確率的勾配で (ρ, ε) -停留点
平滑な目的関数では ε について最適



実際の Muon との違い (今後の課題)

- 正規化：定数 G_{op} vs. データ依存の正規化 (例： $\|M_t\|_F$)
- 係数：Taylor 係数 vs. 経験的に調整した 5 次多項式
- ランダムな勾配評価点と平均化した出力 (O2NC) vs. 反復点での勾配

参考文献 I

- Jacob Abernethy, Chansoo Lee, Abhinav Sinha, and Ambuj Tewari. Online linear optimization via smoothing. In *Proceedings of the 27th Conference on Learning Theory*, volume 35, pages 807–823, 2014.
- Kwangjun Ahn, Zhiyu Zhang, Yunbum Kook, and Yan Dai. Understanding Adam optimizer via online learning of updates: Adam is FTRL in disguise. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 619–640, 2024.
- Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1–2):165–214, 2023.
- Åke Björck and C. Bowie. An iterative algorithm for computing the best estimate of an orthogonal matrix. *SIAM Journal on Numerical Analysis*, 8(2):358–364, 1971.
- Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points I. *Mathematical Programming*, 184:71–120, 2020.
- Sayantan Choudhury, Xiaoran Cheng, Martin Takáč, Sen Na, and Mladen Kolar. Muon with Nesterov momentum: Heavy-tailed noise and (randomized) inexact polar decomposition. *arXiv preprint arXiv:2605.06884*, 2026.

参考文献 II

- Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 6643–6670, 2023.
- Dmitrii Feoktistov, Timofey Belinsky, Andrey Veprikov, Amir Zainullin, and Aleksandr Beznosikov. Softsign: Smooth sign in your optimizer for better parameter heterogeneity handling. In *International Conference on Machine Learning*, 2026.
- GLM-4.5 Team. GLM-4.5: Agentic, reasoning, and coding (ARC) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.
- A. A. Goldstein. Optimization of Lipschitz continuous functions. *Mathematical Programming*, 13(1): 14–22, 1977.
- Ruichen Jiang, Zakaria Mhammedi, Mehryar Mohri, and Aryan Mokhtari. Adaptive matrix online learning through smoothing with guarantees for nonsmooth nonconvex optimization. In *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336, pages 3782–3824, 2026.
- Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL <https://kellerjordan.github.io/posts/muon/>.

参考文献 III

- Gyu Yeol Kim and Min-hwan Oh. Convergence of Muon with Newton–Schulz. In *International Conference on Learning Representations*, 2026.
- Kimi Team. Kimi K2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.
- Guy Kornowski and Ohad Shamir. Oracle complexity in nonsmooth nonconvex optimization. *Journal of Machine Learning Research*, 23(314):1–44, 2022.
- Dmitry Kovalev. Understanding gradient orthogonalization for deep learning via non-Euclidean trust-region optimization. *arXiv preprint arXiv:2503.12645*, 2025.
- Zdislav Kovarik. Some iterative methods for improving orthonormality. *SIAM Journal on Numerical Analysis*, 7(3):386–389, 1970.
- Jiaxiang Li and Mingyi Hong. A note on the convergence of Muon. *arXiv preprint arXiv:2502.02900*, 2025.
- Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025.
- H. Brendan McMahan. A survey of algorithms and analysis for adaptive online learning. *Journal of Machine Learning Research*, 18(90):1–50, 2017.
- Aratrika Mustafi, Soumya Mukherjee, and Bharath K. Sriperumbudur. Move on Muon: A Hamiltonian probability gradient flow perspective of Muon optimizer. *arXiv preprint arXiv:2605.23871*, 2026.

参考文献 IV

- Tetiana Parshakova, Ahmed Khaled, Michael Crawshaw, Guillaume Garrigos, and Robert M. Gower. Muon does not converge on convex Lipschitz functions. *arXiv preprint arXiv:2605.08980*, 2026.
- Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 49069–49104, 2025.
- Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.
- Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the convergence analysis of Muon. *Transactions on Machine Learning Research*, 2026.
- Egor Shulgin, Sultan AlRashed, Peter Richtárik, and Francesco Orabona. Beyond the ideal: Analyzing the inexact Muon update. In *International Conference on Artificial Intelligence and Statistics*, volume 300, pages 5167–5175, 2026.
- Kaiyue Wen, David Hall, Tengyu Ma, and Percy Liang. Fantastic pretraining optimizers and where to find them. In *International Conference on Learning Representations*, 2026.

参考文献 V

Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 11173–11182, 2020.

Qinzi Zhang and Ashok Cutkosky. Random scaling and momentum for non-smooth non-convex optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 58780–58799, 2024.