Preprint: 1–13, 2026

Finite-Time Nonsmooth Convergence of Signum and Muon

Dingzhi Yu[1,2] YUDZ@LAMDA.NJU.EDU.CN
Yuxing Liu[2] YUXING6@ILLINOIS.EDU
Rui Pan[2] RUIP4@ILLINOIS.EDU
Lijun Zhang[1] ZHANGLJ@LAMDA.NJU.EDU.CN
Yang You[3] YOUY@COMP.NUS.EDU.SG
Difan Zou[4] DZOU@CS.HKU.HK
[1]*Nanjing University* [2]*University of Illinois Urbana-Champaign*
[3]*National University of Singapore* [4]*The University of Hong Kong*

Abstract

We study whether momentum alone can give finite-time guarantees for sign-based methods on nonsmooth objectives. We formulate Signum as a constrained method via Euclidean projection onto a closed convex domain $\mathcal{X}$, with $\mathcal{X} = \mathbb{R}^d$ representing unconstrained optimization. The answer depends sharply on this domain and on the oracle model. In constrained online convex optimization over arbitrary compact convex sets, Signum has linear regret for every stepsize and momentum schedule. The same construction transfers to constrained diagonal Muon. In unconstrained deterministic optimization, every fixed momentum coefficient admits convex Lipschitz counterexamples, and the same diagonal mechanism covers the standard fixed-coefficient Lion regime (Chen et al., 2023). However, this mechanism has a finite-time residual scale proportional to $1 - \beta$, so it does not rule out horizon-dependent or coupled choices with $\beta \uparrow 1$. In stochastic optimization, conditional unbiasedness alone is insufficient. For every fixed $\beta < 1$, an adapted unbiased oracle can freeze the true progress coordinate while maintaining a positive Goldstein gap. This stochastic shield does not settle i.i.d. data or the coupled regime $1 - \beta_t = \Theta(\eta_t)$ under uniform moment bounds. The remaining core question is therefore the unconstrained, coupled, finite-time behavior of Signum and related sign-momentum methods.

Keywords: SignSGD, Signum, Lion, Muon, momentum, nonsmooth optimization

1. Introduction

Sign-based methods, most notably SignSGD (Bernstein et al., 2018), update an iterate using only the sign of a gradient estimate. This makes them attractive in distributed learning and low-precision training, where communication and arithmetic are major bottlenecks (Bernstein et al., 2019; Yu et al., 2026a). They are also useful abstractions for adaptive optimizers such as Adam (Kingma and Ba, 2014), where both theory and experiments have revealed a close connection between Adam-like methods and sign descent (Balles and Hennig, 2018; Kunstner et al., 2023). A more recent source of attention is Muon (Jordan et al., 2024), a matrix sign optimizer used for neural-network hidden layers. On diagonal matrices, Muon's polar factor reduces to an entrywise sign, so failures of signed momentum on diagonal objectives can be converted into failures of diagonal Muon (Parshakova et al., 2026). Thus nonsmooth pathologies of Signum are not only toy examples for a scalar sign method. They also probe the robustness of a modern matrix sign optimizer that has been used effectively in LLM training (Liu et al., 2025; Kimi Team, 2025; Zeng et al., 2026; DeepSeek-AI, 2026).

In smooth optimization, the role of momentum in sign-based methods has been studied extensively, and it can provably reduce the bias of signed stochastic gradients and improve the guarantees of SignSGD-type methods (Sun et al., 2023; Jiang et al., 2025; Yu et al., 2026b; Jiang et al., 2026; Tao et al., 2026). However, the nonsmooth case is less understood. Nonsmoothness appears naturally through ReLU activations, max-pooling, mixture-of-experts routing, clipping, and quantization (Nair and Hinton, 2010; Krizhevsky et al., 2012; Shazeer et al., 2017; Pascanu et al., 2013; Gupta et al., 2015). Smooth descent inequalities no longer apply, and the coordinatewise sign map is intrinsically biased. Indeed, deterministic SignSGD can fail even on simple convex Lipschitz functions (Karimireddy et al., 2019). Existing finite-time remedies use mechanisms beyond plain momentum, such as error feedback (Karimireddy et al., 2019) or structural stochasticity (Yu et al., 2026a). This raises the question studied here.

Can momentum by itself restore finite-time nonsmooth convergence for sign-based methods?

We focus on the Signum update of Bernstein et al. (2018), written with a Euclidean projection,

$$m_{t+1} = \beta_t m_t + (1-\beta_t) g_t, \qquad x_{t+1} = \Pi_{\mathcal{X}}(x_t - \eta_t \operatorname{sign}(m_{t+1})), \tag{1}$$

where $\mathcal{X} \subseteq \mathbb{R}^d$ is closed and convex, $\Pi_{\mathcal{X}}$ is Euclidean projection, $\operatorname{sign}(\cdot)$ is applied coordinatewise, and $\operatorname{sign}(0) = 0$. The unconstrained case is $\mathcal{X} = \mathbb{R}^d$, for which $\Pi_{\mathcal{X}}$ is the identity. When $\beta_t \equiv \beta$, Equation (1) is sign-equivalent to the common implementation $m_{t+1} = g_t + \beta m_t$, $x_{t+1} = \Pi_{\mathcal{X}}(x_t - \eta_t \operatorname{sign}(m_{t+1}))$, after a positive rescaling of the momentum (Karimireddy et al., 2019). This equivalence need not hold for arbitrary time-varying β_t.

In online convex optimization (OCO) (Zinkevich, 2003), an algorithm chooses $x_t \in \mathcal{X}$ before observing a convex loss ℓ_t. Its regret after T rounds is

$$\operatorname{Regret}_T := \sum_{t=0}^{T-1} \ell_t(x_t) - \min_{x \in \mathcal{X}} \sum_{t=0}^{T-1} \ell_t(x),$$

and the target is $\operatorname{Regret}_T = o(T)$. For unconstrained nonsmooth nonconvex stationarity, we use the Goldstein gap

$$G_\delta(x) := \operatorname{dist}\left(0, \operatorname{conv} \bigcup\nolimits_{y \in B_\infty(x,\delta)} \partial f(y)\right), \tag{2}$$

a standard finite-time stationarity measure in nonsmooth optimization (Cutkosky et al., 2023).

Open Problem 1 (Finite-Time Nonsmooth Signum) *Determine which schedules (η_t, β_t) make constrained Signum satisfy $\operatorname{Regret}_T = o(T)$ in OCO, or make unconstrained Signum satisfy an explicit finite-time bound on $\min_{t<T} \mathbb{E}\left[G_\delta(x_t)\right]$ under exact or stochastic nonsmooth oracles. If no such schedules exist in a given model, prove a matching impossibility result, and determine which conclusions transfer to diagonal Muon.*

Table 1 summarizes the finite-time picture across constrained and unconstrained models.

The constrained result is a complete negative answer for arbitrary compact convex sets. The unconstrained results are more nuanced. Fixed momentum is not enough, but the known deterministic diagonal obstruction weakens as β approaches one. This leaves the coupled regime $1 - \beta_t = \Theta(\eta_t)$ as the most plausible route by which momentum alone might still help.

Table 1: Finite-time status by oracle type, domain, stepsize schedule, momentum schedule, and result reference. OCO means online convex optimization, det. means deterministic, stoch. means stochastic, const. means constant, summ. means summable, and decr. means strictly decreasing nonsummable.

Oracle	**Domain $\mathcal{X}$**	η_t	β_t	**Reference**
OCO	$\mathcal{X} \subset \mathbb{R}^2$	any	any in $[0, 1]$	Theorem 2
det.	$\mathbb{R}^2$	const., summ., decr.	fixed $\beta < 1$	Proposition 5
det.	$\mathbb{R}^2$	const.	fixed $\beta < 1$	Proposition 6
stoch.	$\mathbb{R}^2$	any	fixed $\beta < 1$	Theorem 9
det./stoch.	$\mathbb{R}^d$	diminishing	$1 - \beta_t = \Theta(\eta_t)$	Section 5

2. Constrained Problems

The constrained OCO version of Open Problem 1 is false for arbitrary compact convex feasible sets. The obstruction is geometric: a coordinatewise sign step can be orthogonal to the objective's progress coordinate after projection onto a rotated set.

Theorem 2 (Constrained OCO failure) *For every schedule $\eta_t \geq 0$, every momentum schedule $\beta_t \in [0, 1]$, and $m_0 = 0$, there exist a two-dimensional compact convex full-dimensional set $\mathcal{X}$, a repeated linear loss $\ell_t \equiv \ell$, and an initialization $x_0 \in \mathcal{X}$ such that constrained Signum has linear regret. More precisely, for every $a > 1$ there is an instance with*

$$\mathrm{Regret}_T \geq 2(a - 1)T.$$

The construction uses a rotated rectangle. The repeated linear gradient makes the signed direction move only along the thin coordinate of the rectangle; Euclidean projection then keeps the coordinate that matters for minimizing the loss fixed. The full proof is in Appendix A.

Corollary 3 (Constrained Muon) *Constrained diagonal Muon has the same linear-regret counterexample for every momentum and stepsize schedule.*

This follows because, on diagonal matrices, the Muon polar factor reduces to the entrywise sign on diagonal coordinates (Parshakova et al., 2026). The proof is also given in Appendix A.

Remark 4 *Theorem 2 is a constrained-geometry counterexample. It does not settle unconstrained Signum or unconstrained Muon, where an invariant of this kind must come from the objective and the dynamics rather than from the feasible set.*

Having isolated the obstruction created by projection, we now remove the feasible-set geometry and ask what remains. The deterministic exact-oracle model is the cleanest next case, since any failure must come from the nonsmooth objective and the momentum recursion.

3. Unconstrained Deterministic Momentum

We therefore set $\mathcal{X} = \mathbb{R}^2$ and use exact subgradients. Fixed momentum still cannot guarantee finite-time nonsmooth convergence. The central construction is the convex family

$$f_{a,b,C}(u,v) = |au + v| + b|u + C|, \qquad a, b, C > 0. \tag{3}$$

Along the diagonal $u = v = \xi$ and inside the strip $u > -C$, if $r = \operatorname{sign}(\xi)$ then a natural subgradient is

$$g(\xi, \xi) = (ar + b, r). \tag{4}$$

The second coordinate alternates with the side of the kink, while the first coordinate has a persistent positive bias. For fixed β, choosing a large enough makes the momentum sign follow the wrong diagonal direction.

Proposition 5 (Fixed-β deterministic failure) *Fix $\beta \in [0, 1)$ and choose $a > (1 + \beta)/(1 - \beta)$. For Equation (1) with $\beta_t \equiv \beta$ and $f_{a,1,C}$, each of the following stepsize regimes admits choices of C, x_0, and $m_0 = 0$ for which the stated failure occurs:*

1. *If $\eta_t \equiv \eta > 0$, then $x_t = (-1)^t(\eta/2, \eta/2)$ is an exact period-2 orbit.*
2. *If $\sum_t \eta_t < \infty$, then x_t converges to a nonstationary point.*
3. *If η_t is strictly decreasing to 0 and $\sum_t \eta_t = \infty$, then $x_t \to (0, 0)$, which is nonstationary.*

In each case, there exists $\delta > 0$ such that $G_\delta(x_t) \geq (1 + a^2)^{-1/2}$ for every t.

Proposition 5 rules out fixed-β finite-time guarantees under standard deterministic nonsmooth assumptions. It also gives linear regret on repeated losses over any bounded domain containing the constructed trajectory and the minimizer.

The high-momentum finite-time interpretation requires a second observation. The same diagonal mechanism does not give a β-uniform Goldstein lower bound when $\beta \uparrow 1$ under a fixed Lipschitz normalization.

Proposition 6 (Residual scale of the diagonal obstruction) *Fix $\beta \in [0, 1)$ and consider the family (3). If*

$$\frac{b}{a} < \frac{1 - \beta}{1 + \beta}, \tag{5}$$

then, for every constant stepsize $\eta > 0$, there exist C and x_0 such that Signum with $\beta_t \equiv \beta$ follows the exact two-cycle $x_t = (-1)^t(\eta/2, \eta/2)$. If also $b \leq a$, then along this cycle there is a $\delta > 0$ such that

$$G_\delta(x_t) \geq \frac{b}{\sqrt{a^2 + 1}} \quad \text{for every } t.$$

In particular, choosing $a = 1$ and $b = (1 - \beta)/4$ gives an $O(1)$-Lipschitz convex example with a Goldstein floor $\Omega(1 - \beta)$. Conversely, condition (5) forces this local diagonal Goldstein gap to be $O(1 - \beta)$.

Thus fixed momentum fails, but the diagonal lower bound is compatible with horizon-dependent choices such as $\beta = 1 - \Theta(1/T)$. This is why the unconstrained, high-momentum, schedule-universal problem remains open. Recent work of Parshakova et al. (2026) proves a schedule-adaptive counterexample for diagonal Muon/sign-momentum in the low-momentum range $\beta < 1/2$. Schedule-universal deterministic failures in the high-momentum range $\beta \geq 1/2$ are not settled by the diagonal oscillator alone.

Corollary 7 (Deterministic diagonal Muon) *The deterministic Signum failures in Proposition 5 and the residual-scale construction in Proposition 6 transfer to diagonal Muon.*

The same exact-gradient construction also applies to standard fixed-coefficient Lion.

Proposition 8 (Fixed-coefficient Lion) *Consider Lion (Chen et al., 2023) with fixed coefficients*

$$d_t = \beta_1 m_t + (1-\beta_1) g_t, \qquad x_{t+1} = x_t - \eta_t \operatorname{sign}(d_t), \qquad m_{t+1} = \beta_2 m_t + (1-\beta_2) g_t. \tag{6}$$

If $\beta_1, \beta_2 \in [0,1)$ *satisfy*

$$\beta_1 < \frac{1}{2-\beta_2}, \tag{7}$$

then a suitable member of the family (3) gives the same constant, summable, and strictly decreasing nonsummable failures as in Proposition 5. In particular, this covers the standard regime $\beta_1 \leq \beta_2 < 1$.

4. Unconstrained Stochastic Momentum

The deterministic results in Section 3 use exact subgradients. Any convergence proof for Signum must first overcome this deterministic obstruction, and a stochastic oracle then adds a second difficulty. The current sample both estimates a subgradient and selects the signed direction through m_{t+1}. This feedback makes finite-time convergence harder to justify. Even under a conditionally unbiased oracle whose distribution may adapt to the algorithmic filtration, fixed momentum fails for every $\beta < 1$, including the high-momentum range.

The construction uses the orthogonal directions

$$e = (1,1), \qquad h = (1,-1),$$

and the convex objective

$$f(x) = c|\langle e, x\rangle - P|, \qquad 0 < c < b. \tag{8}$$

The oracle adds zero-mean noise in the flat h direction, but chooses its conditional distribution so that the momentum sign is determined by that flat direction. The true progress coordinate $\langle e, x_t\rangle$ never changes.

Theorem 9 (Stochastic shield) *Fix any* $\beta \in [0,1)$*, any nonnegative stepsize sequence* (η_t)*, any* $0 < c < b$*, and* $m_0 = 0$*. There exist a choice of* P*, an initialization with* $\langle e, x_0\rangle < P$*, and a conditionally unbiased stochastic oracle for (8) such that*

$$\langle e, x_t\rangle = \langle e, x_0\rangle \quad \textit{for all } t \geq 0$$

almost surely. If $P - \langle e, x_0 \rangle > 2\delta$*, then*

$$G_\delta(x_t) = \sqrt{2}\, c \quad \textit{for all } t \geq 0$$

almost surely. Moreover, the oracle has bounded moments for each fixed $\beta < 1$*, and*

$$\|x_t - x_0\|_2 \leq \sqrt{2} \sum_{i<t} \eta_i.$$

The moment bound in Theorem 9 deteriorates as $(1-\beta)^{-1}$. Hence the shield is an all-time bounded-iterate counterexample for summable stepsizes, and a finite-horizon lower bound for arbitrary stepsizes. It does not refute i.i.d. data or the coupled regime $1 - \beta_t = \Theta(\eta_t)$ under a uniform moment bound.

The proof also pinpoints a stochastic pitfall. If $g_t = h_t + \zeta_t$ and $\mathbb{E}[\zeta_t \mid \mathcal{F}_t] = 0$, one cannot treat $\langle \zeta_t, \operatorname{sign}(m_{t+1}) \rangle$ as a martingale difference, because m_{t+1} contains g_t. In the shield construction,

$$\mathbb{E}[\langle \zeta_t, \operatorname{sign}(m_{t+1}) \rangle \mid \mathcal{F}_t] = 2(1+\beta)b > 0$$

after the first step. Conditional unbiasedness alone therefore gives no sign-alignment guarantee.

Corollary 10 (Stochastic diagonal Muon) *The stochastic shield transfers to diagonal Muon. In particular, there is a conditionally unbiased stochastic diagonal oracle for which diagonal Muon freezes the same progress coordinate and keeps the same positive Goldstein gap.*

5. What Remains Open

The results above give a fairly sharp finite-time map. Constrained OCO over arbitrary compact sets is impossible. Unconstrained deterministic fixed momentum is impossible, but the standard diagonal obstruction weakens as β approaches one. Adapted stochastic oracles can defeat fixed momentum, but the construction uses noise whose bound grows like $(1-\beta)^{-1}$ and is not an i.i.d. data model.

The main unresolved positive direction is the coupled regime

$$\beta_t = 1 - \tau\eta_t, \qquad m_{t+1} = m_t + \tau\eta_t(g_t - m_t), \tag{9}$$

where $\tau > 0$. This keeps the spatial memory length

$$\frac{\eta_t}{1-\beta_t} = \frac{1}{\tau}$$

fixed while sending the fixed-β residual scale to zero. Xiao et al. (2023) prove asymptotic convergence for a broad inertial stochastic subgradient framework that includes Signum in this coupled form, under additional stability assumptions and sufficiently small stepsize constants. Their result is qualitative: it does not provide a finite-time Goldstein rate, an OCO regret bound, or an iteration complexity.

The remaining questions are therefore the following.

Open Problem 11 (High-momentum deterministic Signum) *For fixed* $\beta \in [1/2, 1)$*, does there exist one unconstrained convex Lipschitz objective and one initialization on which Signum, or diagonal Muon through the Signum reduction, fails for every admissible stepsize schedule?*

Open Problem 12 (Coupled finite-time Signum) *For Signum with $1 - \beta_t = \Theta(\eta_t)$, prove or refute a finite-time nonsmooth Goldstein guarantee under exact subgradients, or under a stochastic model with i.i.d. samples or uniform bounded moments.*

Open Problem 13 (Stochastic sign alignment) *Identify stochastic assumptions that prevent the current noise from being positively aligned with the current signed momentum direction. Candidate mechanisms include i.i.d. sampling assumptions, sample splitting, delayed signs using $\operatorname{sign}(m_t)$, structural stochasticity, or error feedback.*

Open Problem 14 (Time-varying Lion) *Determine whether time-varying Lion admits finite-time nonsmooth convergence, and identify the relations among $\eta_t/(1 - \beta_{1,t})$ and $\eta_t/(1 - \beta_{2,t})$ that separate convergence from failure.*

Taken together, these results suggest that fixed momentum cannot fix nonsmooth sign descent. The only remaining plausible momentum-only regime is one where the momentum coefficient approaches one in a controlled way, and where the stochastic model prevents current noise from selecting an adversarial sign direction.

References

Lukas Balles and Philipp Hennig. Dissecting Adam: The sign, magnitude and variance of stochastic gradients. In *Proceedings of the 35th International Conference on Machine Learning*, pages 404–413, 2018.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *Proceedings of the 35th International Conference on Machine Learning*, pages 560–569, 2018.

Jeremy Bernstein, Jiawei Zhao, Kamyar Azizzadenesheli, and Anima Anandkumar. SignSGD with majority vote is communication efficient and fault tolerant. In *International Conference on Learning Representations*, 2019.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Yao Liu, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V. Le. Symbolic discovery of optimization algorithms. In *Advances in Neural Information Processing Systems*, pages 49205–49233, 2023.

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*, pages 6643–6670, 2023.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence, 2026.

Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. Deep learning with limited numerical precision. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1737–1746, 2015.

Wei Jiang, Dingzhi Yu, Sifan Yang, Wenhao Yang, and Lijun Zhang. Improved analysis for sign-based methods with momentum updates. *arXiv preprint arXiv:2507.12091*, 2025.

Wei Jiang, Mao Xu, Wenhao Yang, Yibo Wang, Zechao Li, and Lijun Zhang. Convergence analysis of the Lion optimizer in centralized and distributed settings. In *Proceedings of the 43rd International Conference on Machine Learning*, page to appear, 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/.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3252–3261, 2019.

Kimi Team. Kimi K2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, page 84–90, 2012.

Frederik Kunstner, Jacques Chen, Jonathan Wilder Lavington, and Mark Schmidt. Noise Is Not the Main Factor Behind the Gap Between SGD and Adam on Transformers, But Sign Descent Might Be. In *International Conference on Learning Representations*, 2023.

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.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*, pages 807–814, 2010.

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.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1310–1318, 2013.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Tao Sun, Qingsong Wang, Dongsheng Li, and Bao Wang. Momentum ensures convergence of SIGNSGD under weaker assumptions. In *Proceedings of the 40th International Conference on Machine Learning*, pages 33077–33099, 2023.

Hongyi Tao, Dingzhi Yu, and Lijun Zhang. When and Why SignSGD Outperforms SGD: A Theoretical Study Based on ℓ_1-norm Lower Bounds. *arXiv preprint arXiv:2605.06615*, 2026.

Nachuan Xiao, Xiaoyin Hu, and Kim-Chuan Toh. Stochastic subgradient methods with guaranteed global stability in nonsmooth nonconvex optimization. *arXiv preprint arXiv:2307.10053*, 2023.

Dingzhi Yu, Rui Pan, Yuxing Liu, and Tong Zhang. StoSignSGD: Unbiased Structural Stochasticity Fixes SignSGD for Training Large Language Models. *arXiv preprint arXiv:2604.15416*, 2026a.

Dingzhi Yu, Hongyi Tao, Yuanyu Wan, Luo Luo, and Lijun Zhang. Sign-based optimizers are effective under heavy-tailed noise. *arXiv preprint arXiv:2602.07425*, 2026b.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chengxing Xie, Cunxiang Wang, et al. GLM-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, pages 928–936, 2003.

Appendix A. Proofs for Constrained Problems

Proof of Theorem 2. Use coordinates

$$z = \frac{x_1 + x_2}{2}, \qquad w = \frac{x_1 - x_2}{2}, \qquad x = (z + w, z - w).$$

Fix $0 < \rho < 1$ and define the rotated rectangle

$$\mathcal{X}_\rho = \{(z + w, z - w) : z \in [-1, 1],\ w \in [-\rho, \rho]\}.$$

Let $\ell(x) = \langle c, x\rangle$ with $c = (1, -a)$ and $a > 1$. In (z, w) coordinates,

$$\ell(z, w) = (1 - a)z + (1 + a)w.$$

A minimizer over $\mathcal{X}_\rho$ is $(z_\star, w_\star) = (1, -\rho)$, or $x_\star = (1 - \rho, 1 + \rho)$. Initialize at $x_0 = (-1, -1)$, equivalently $(z_0, w_0) = (-1, 0)$.

Since the loss is repeated and linear, $g_t = c$ for all t. The momentum has the form $m_{t+1} = \lambda_{t+1} c$ with $\lambda_{t+1} \geq 0$. Whenever $\lambda_{t+1} > 0$, $\operatorname{sign}(m_{t+1}) = (1, -1)$; if $\lambda_{t+1} = 0$, the update is zero and the lower bound only becomes easier. The raw step by $-\eta_t(1, -1)$ keeps z fixed and sends w to $w - \eta_t$. Moreover,

$$\big\|(z + w, z - w) - (z' + w', z' - w')\big\|_2^2 = 2\{(z - z')^2 + (w - w')^2\}.$$

Thus Euclidean projection onto $\mathcal{X}_\rho$ clips z and w separately. Hence $z_t = -1$ for every t, while $w_t \in [-\rho, 0]$. Therefore

$$\begin{aligned}\ell(x_t) - \ell(x_\star) &= (1 - a)(z_t - 1) + (1 + a)(w_t + \rho) \\ &= 2(a - 1) + (1 + a)(w_t + \rho) \geq 2(a - 1).\end{aligned}$$

Summing over $t = 0, \ldots, T - 1$ proves the theorem. □

Proof of Corollary 3. On diagonal matrices, the Muon polar factor reduces to the entrywise sign on the diagonal coordinates (Parshakova et al., 2026). Apply Theorem 2 to (W_{11}, W_{22}), take the linear loss $F(W) = W_{11} - aW_{22}$, and constrain these two diagonal entries to $\mathcal{X}_\rho$. Frobenius projection reduces to Euclidean projection on these two coordinates, so the proof is unchanged. □

Appendix B. Proofs for Deterministic Signum

Proof of Proposition 5. Set $b = 1$ in (3). For the constant and strictly decreasing nonsummable constructions, let $r_t = (-1)^t$ and suppose that x_t lies on the diagonal side with sign r_t. Then Equation (4) gives $g_t = (1 + ar_t, r_t)$. Writing $m_t = (p_t, q_t)$ and taking $m_0 = 0$,

$$q_{t+1} = (1-\beta)\sum_{i=0}^{t} \beta^{t-i} r_i = \frac{1-\beta}{1+\beta} r_t\big(1 - (-\beta)^{t+1}\big),$$

so $\operatorname{sign}(q_{t+1}) = r_t$. Also,

$$p_{t+1} = (1-\beta)\sum_{i=0}^{t} \beta^{t-i}(1 + ar_i) = 1 - \beta^{t+1} + aq_{t+1}.$$

When t is even, $p_{t+1} > 0$. When t is odd,

$$p_{t+1} = (1-\beta^{t+1})\left(1 - \frac{a(1-\beta)}{1+\beta}\right) < 0,$$

by the choice $a > (1+\beta)/(1-\beta)$. Hence $\operatorname{sign}(m_{t+1}) = (r_t, r_t)$.

For constant $\eta_t \equiv \eta$, choose $C > \eta/2$ and $x_0 = (\eta/2, \eta/2)$. Then $x_{t+1} = x_t - \eta(r_t, r_t) = r_{t+1}(\eta/2, \eta/2)$.

For strictly decreasing nonsummable stepsizes, define the alternating tail

$$\varepsilon_t := \sum_{j=t}^{\infty} (-1)^j \eta_j.$$

The alternating-series theorem gives $0 < (-1)^t \varepsilon_t < \eta_t$. Choosing $C > \eta_0$ and $x_0 = (\varepsilon_0, \varepsilon_0)$ yields

$$x_{t+1} = x_t - \eta_t((-1)^t, (-1)^t) = (\varepsilon_{t+1}, \varepsilon_{t+1}),$$

so $x_t \to (0,0)$.

For an arbitrary summable schedule, let $S = \sum_{t\geq 0} \eta_t$, choose $C > S + 1$, and initialize at $x_0 = (S+1, S+1)$. The iterates remain on the positive diagonal, so $g_t = (a+1, 1)$ and $\operatorname{sign}(m_{t+1}) = (1,1)$ for all t. Therefore

$$x_t = \left(S + 1 - \sum_{i=0}^{t-1} \eta_i,\ S + 1 - \sum_{i=0}^{t-1} \eta_i\right) \to (1,1).$$

In each construction the trajectory has a positive distance from $u = -C$. Choose $\delta > 0$ so that the ℓ_∞ neighborhoods of the trajectory remain inside $u > -C$. The local convexified subdifferential is then contained in

$$\{(1 + as, s) : s \in [-1, 1]\}.$$

The distance from 0 to this segment is $1/\sqrt{1+a^2}$, which proves the Goldstein lower bound. □

Proof of Proposition 6. Assume $x_t = r_t(\eta/2, \eta/2)$ with $r_t = (-1)^t$. Then $g_t = (ar_t + b, r_t)$. Writing $m_{t+1} = (p_{t+1}, q_{t+1})$ and using $m_0 = 0$,

$$q_{t+1} = (1-\beta)\sum_{i=0}^{t} \beta^{t-i} r_i = \frac{1-\beta}{1+\beta} r_t \left(1 - (-\beta)^{t+1}\right),$$

and the bias part in the first coordinate is

$$h_{t+1} := (1-\beta)\sum_{i=0}^{t} \beta^{t-i} = 1 - \beta^{t+1}, \qquad p_{t+1} = bh_{t+1} + aq_{t+1}.$$

For even t, $p_{t+1} > 0$. For odd t,

$$\frac{|q_{t+1}|}{h_{t+1}} = \frac{1-\beta}{1+\beta},$$

so condition (5) gives $p_{t+1} < 0$. Hence $\operatorname{sign}(m_{t+1}) = (r_t, r_t)$ for every t, and the update maps $r_t(\eta/2, \eta/2)$ to $r_{t+1}(\eta/2, \eta/2)$.

Inside the strip $u > -C$, the local convexified subdifferential is contained in

$$\{(as + b, s) : s \in [-1, 1]\}.$$

If $b \le a$, the distance from the origin to this segment is $b/\sqrt{a^2+1}$. This gives the stated lower bound for a sufficiently small δ. Finally, condition (5) implies

$$\frac{b}{\sqrt{a^2+1}} < \frac{1-\beta}{1+\beta}\frac{a}{\sqrt{a^2+1}} \le \frac{1-\beta}{1+\beta},$$

which proves the residual-scale upper limitation of this diagonal mechanism. □

Proof of Corollary 7. Work with diagonal matrices $W \in \mathbb{R}^{2\times 2}$ and identify $x = (W_{11}, W_{22})$. For the objectives in Propositions 5 and 6, set

$$F(W) = f_{a,b,C}(W_{11}, W_{22}).$$

If W_0 and the momentum are diagonal, every chosen subgradient is diagonal and the momentum remains diagonal. The Muon polar factor of a diagonal matrix is the diagonal matrix of entrywise signs on its nonzero diagonal entries. Hence the two diagonal entries of Muon follow exactly the Signum recursion for $x = (W_{11}, W_{22})$, and the claimed failures and residual-scale lower bounds transfer unchanged. □

Appendix C. Proof of Fixed-Coefficient Lion

Proof of Proposition 8. Along the alternating diagonal trajectory used in Proposition 5, the second coordinate of m_t is

$$q_t = (1-\beta_2)\sum_{i=0}^{t-1} \beta_2^{t-1-i} r_i = -\frac{1-\beta_2}{1+\beta_2} r_t \big(1 - (-\beta_2)^t\big).$$

Thus the second coordinate of the Lion direction d_t equals

$$\beta_1 q_t + (1-\beta_1) r_t = B_t r_t,$$

where

$$B_t = (1-\beta_1) - \beta_1 \frac{1-\beta_2}{1+\beta_2}\big(1-(-\beta_2)^t\big) \geq 1 - 2\beta_1 + \beta_1\beta_2 > 0$$

under (7). The first coordinate of d_t is a positive bounded bias plus $aB_t r_t$. Choosing a larger than the reciprocal of the lower margin forces $\operatorname{sign}(d_t) = (r_t, r_t)$ for the alternating constructions. The constant, summable, and strictly decreasing nonsummable failures then follow exactly as in Appendix B. □

Appendix D. Proofs for the Stochastic Shield

Proof of Theorem 9. Choose P so that $\langle e, x_0\rangle < P$. While the process stays on this side, $\partial f(x_t) = \{-ce\}$. Use stochastic gradients

$$g_t = -ce + \xi_t h, \qquad \mathbb{E}[\xi_t \mid \mathcal{F}_t] = 0.$$

Decompose $m_t = A_t e + B_t h$. Then

$$A_{t+1} = \beta A_t - (1-\beta)c, \qquad B_{t+1} = \beta B_t + (1-\beta)\xi_t.$$

With $m_0 = 0$,

$$A_t = -c(1-\beta^t), \qquad |A_t| < c.$$

At $t = 0$, choose $\xi_0 = \pm b/(1-\beta)$ with probabilities $1/2, 1/2$, so $B_1 = \pm b$ and $\mathbb{E}[\xi_0] = 0$. For $t \geq 1$, suppose $B_t = sb$ with $s \in \{-1, +1\}$. Define

$$\xi_t = \begin{cases} sb, & \text{with probability } (1+\beta)/2, \\ -s\frac{1+\beta}{1-\beta} b, & \text{with probability } (1-\beta)/2. \end{cases}$$

This distribution has zero conditional mean. It also gives

$$B_{t+1} = \begin{cases} sb, & \text{with probability } (1+\beta)/2, \\ -sb, & \text{with probability } (1-\beta)/2. \end{cases}$$

Thus $|B_t| = b > c > |A_t|$ for all $t \geq 1$. Therefore

$$\operatorname{sign}(m_{t+1}) = \operatorname{sign}(B_{t+1}) h,$$

and since $\langle e, h\rangle = 0$,

$$\langle e, x_{t+1}\rangle = \langle e, x_t\rangle - \eta_t \langle e, \operatorname{sign}(m_{t+1})\rangle = \langle e, x_t\rangle .$$

This proves the progress-coordinate invariant.

If $P - \langle e, x_0\rangle > 2\delta$, then every $y \in B_\infty(x_t, \delta)$ remains on the side $\langle e, y\rangle < P$, because changing two coordinates by at most δ changes $\langle e, y\rangle$ by at most 2δ. Hence the Goldstein hull is the singleton $\{-ce\}$ and $G_\delta(x_t) = \|ce\|_2 = \sqrt{2}c$.

For fixed $\beta < 1$,

$$|\xi_t| \le \frac{1+\beta}{1-\beta} b, \qquad \|g_t\|_2 \le \sqrt{2}\left(c + \frac{1+\beta}{1-\beta} b\right).$$

Since every update direction is $\pm h$, the displacement bound

$$\|x_t - x_0\|_2 \le \sqrt{2} \sum_{i<t} \eta_i$$

also follows. □

Noise-sign correlation in the shield. Let $\zeta_t = \xi_t h$. For $t \ge 1$ and $B_t = sb$,

$$\mathbb{E}[\xi_t \operatorname{sign}(B_{t+1}) \mid \mathcal{F}_t] = \frac{1+\beta}{2} b + \frac{1-\beta}{2} \frac{1+\beta}{1-\beta} b = (1+\beta) b.$$

Therefore

$$\mathbb{E}[\langle \zeta_t, \operatorname{sign}(m_{t+1}) \rangle \mid \mathcal{F}_t] = 2(1+\beta) b > 0.$$

Proof of Corollary 10. Let $W \in \mathbb{R}^{m \times n}$ with $m, n \ge 2$ and define

$$f(W) = c|W_{11} + W_{22} - P|.$$

Use stochastic diagonal subgradients

$$G_t = \operatorname{Diag}(-c + \xi_t,\ -c - \xi_t,\ 0, \ldots).$$

The conditional expectation is $\operatorname{Diag}(-c, -c, 0, \ldots) \in \partial f(W_t)$ whenever $W_{11} + W_{22} < P$. On diagonal matrices, the Muon polar factor is the entrywise sign on diagonal coordinates, so the two-coordinate proof of Theorem 9 is unchanged. □