ON THE CONVERGENCE OF ADAM, REVISITED

STEVEN HEILMAN AND SAMPAD MOHANTY

Abstract. We show that projected Adam for online optimization with arbitrary moment decay parameters $\beta_1, \beta_2 \in [0,1)$ can have average regret bounded away from zero. A similar result of Reddi-Kale-Kumar from 2018 required $\beta_1 < \sqrt{\beta_2}$. Similar to their result, we use a three-periodic sequence of linear functions on $[-1,1]$ with slopes $c, -1, -1$, though we use c slightly larger than 2. This nonzero average regret result extends to Adam variants such as AdamW, RMSProp, NAdam, Adan, AdaMax, Muon, and to an i.i.d. variant of the three-periodic sequence of slopes for Adam.

1. Introduction

In online minimization on $[-1,1]$, we are presented with a sequence of functions $f_1, f_2, \ldots$ where $f_t \colon [-1,1] \to \mathbb{R}$ for all $t \geq 1$. At time $t \geq 1$, we know $f_1(x_1), \ldots, f_{t-1}(x_{t-1})$ and $f_1'(x_1), \ldots, f_{t-1}'(x_{t-1})$, and we produce $x_t \in [-1,1]$. For a fixed time horizon $T \geq 1$, the goal is to minimize the regret R_T at time T against the best fixed comparator in $[-1,1]$, where

$$R_T := \sum_{t=1}^{T} f_t(x_t) - \min_{x \in [-1,1]} \sum_{t=1}^{T} f_t(x). \tag{1}$$

In contemporary applications, f_t often depends on the t-th portion (or batch) of a large dataset. The most popular optimization method for these applications is Adam [KB15] (Adaptive Moment Estimation). Adam and its variants perform online optimization to train neural networks, transformers, large language models, etc. With nearly 250,000 citations, [KB15] is currently one of the all time most highly cited scientific papers.

Under additional assumptions such as varying its parameters, Adam is known to converge, in the sense that $R_T/T \to 0$ as $T \to \infty$ [RKK18]. However, it is also known that Adam might not converge, i.e. there are examples of sequences of fairly reasonable functions $f_1, f_2, \ldots$ where projected Adam produces $x_1, x_2, \ldots$ with R_T/T not converging to 0 as $T \to \infty$. However, these results only apply with restrictions on Adam's parameters [RKK18]. In order to understand these parameters, let us define projected Adam.

Definition 1.1 (Adam Optimization Method [KB15]). *Fix*

$$b := \beta_1 \in [0,1), \qquad q := \beta_2 \in [0,1), \qquad \varepsilon \geq 0, \qquad \alpha_t > 0, \quad \forall\, t \geq 1.$$

Date: July 3, 2026.
Email: stevenmheilman@gmail.com, sbmohant@usc.edu
S.H. is supported by NSF Grant CCF AF 2448108.
2020 Mathematics Subject Classification: 68W27, 65K10, 68Q32
Keywords: Adam, online optimization, regret
Department of Mathematics, University of Southern California, Los Angeles, CA 90089.

Let $x_1 \in [-1,1]$ be arbitrary. Define $x_2, x_3, \ldots \in [-1,1]$ as follows.

$$m_t := bm_{t-1} + (1-b)g_t, \qquad v_t := qv_{t-1} + (1-q)g_t^2, \qquad g_t := f_t'(x_t), \qquad \forall\, t \geq 1 \tag{2}$$

with the standard initialization $m_0 = v_0 = 0$. The projected update with step sizes $\alpha_t > 0$ is

$$x_{t+1} := \Pi_{[-1,1]}\left(x_t - \alpha_t h_t\right), \qquad h_t := \frac{m_t}{\sqrt{v_t} + \varepsilon}, \qquad \forall\, t \geq 1, \tag{3}$$

where $\Pi_{[-1,1]}(x) := -1_{\{x<-1\}} + x1_{\{-1\leq x\leq 1\}} + 1_{\{x>1\}}$ is projection of $x \in \mathbb{R}$ to the nearest element of $[-1,1]$. Here α_t is called the learning rate or step size. For example, one could use $\alpha_t := \alpha/\sqrt{t}$ for some $\alpha > 0$. Also, if $\varepsilon = 0$, then h_t is only defined when $v_t \neq 0$.

Some authors may refer to Adam as the above optimization method, but with no projection term $\Pi_{[-1,1]}$ appearing (3). We will not do that. Unless otherwise stated, we only refer to Adam as the method defined in Definition 1.1.

Remark 1.2. *We define RMSProp to be the Adam optimization method with $\beta_1 = 0$. Other implementations called RMSProp may include momentum, centering, different epsilon placement, or bias corrections; those variants require separate notation, although the same short-memory denominator mechanism often persists.*

Remark 1.3. *We briefly contrast Adam with other optimization methods:*

- $x_{t+1} := x_t - \alpha_t g_t$ *(Gradient Descent)*
- $m_t := bm_{t-1} + g_t$, $x_{t+1} := x_t - \alpha_t m_t$ *(Heavy Ball)*
- $m_t := bm_{t-1} + f_t'(x_t - \alpha_t m_{t-1})$, $x_{t+1} := x_t - \alpha_t m_t$ *(Nesterov Accelerated Gradient)*
- $m_t := bm_{t-1} + (1-b)g_t$, $v_t := \max(qv_{t-1}, |g_t|)$, $h_t := \frac{m_t}{v_t+\varepsilon}$, x_{t+1} *as in* (3) *(AdaMax)*
- *Same as Adam with* $x_{t+1} := \Pi_{[-1,1]}\left((1-\lambda\alpha_t)x_t - \alpha_t h_t\right)$ *for some* $\lambda \geq 0$ *(AdamW)*
- *Same as Adam, but with* $h_t := \frac{\beta_1 m_t + (1-\beta_1)g_t}{\sqrt{v_t}+\varepsilon}$ *(NAdam)*
- *Same as Adam, with q changing over time (NosAdam)*

The main parameters that can adjust the behavior of Adam are β_1 and β_2. From the recursion (2), we see that β_1 quantifies the amount of exponentially decaying "memory" of past derivatives of f_t (where $b = \beta_1$ close to 1 is a "larger" amount of such memory), since m_t is approximately a function of $1/\log(1/b)$ previous time steps. Likewise, $q = \beta_2$ quantifies the amount of "memory" of past squared gradients of f_t.

Here are some cited examples of Adam used to train large language models, together with their parameter descriptions.

- BERT was trained with Adam "with learning rate of 10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, L2 weight decay of 0.01, learning rate warmup over the first 10,000 steps, and linear decay of the learning rate." [DCK+19].
- GPT-3 was trained with Adam "with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and $\varepsilon = 10^{-8}$, we clip the global norm of the gradient at 1.0, and we use cosine decay for learning rate down to 10% of its value" [BMR+20].
- Llama 2 was trained "using the AdamW optimizer [LH19], with $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\varepsilon = 10^{-5}$. We use a cosine learning rate schedule, with warmup of 2000 steps, and decay final learning rate down to 10% of the peak learning rate. We use a weight decay of 0.1 and gradient clipping of 1.0." [TMS+23].
- DeepSeek-V3 "employ[s] the AdamW optimizer [LH19] with hyper-parameters set to $\beta_1 = 0.9$, $\beta_2 = 0.95$ and weight decay 0.1." [D24]

Despite the empirical success of Adam, it is known that it might not converge to its optimum. The main result of Reddi, Kale, and Kumar [RKK18] showed that Adam might not converge to its optimum for a sequence of linear functions on $[-1,1]$.

Theorem 1.4 ([RKK18, Theorem 2]). *Let $\beta_1 < \sqrt{\beta_2}$, let $\alpha > 0$ and let $\alpha_t := \alpha/\sqrt{t}$, for all $t \geq 1$. Then there exists a sequence of functions $f_1, f_2, \ldots : [-1,1] \to \mathbb{R}$ such that the Adam optimization method has regret satisfying: R_T/T does not converge to zero as $T \to \infty$.*

The example used was $f_t(x) = -x$ for all $t \geq 1$ except $t \bmod c = 1$, in which case $f_t(x) = cx$, for all $x \in [-1,1]$. That is, the slope of f_t is c-periodic, where c is chosen to be a sufficiently large number, as a function of β_1, β_2. The idea is that the large positive slope that appears once is sufficient to offset the other smaller negative slopes.

Theorem 1.4 was also extended [RKK18, Theorem 3] to the setting where the f_t have random dependence on t. That is, $f_t(x) = -x$ with probability $1-p$, and $f_t(x) = cx$ with probability p for some appropriate $0 < p < 1$, with $f_1, f_2, \ldots$ i.i.d. random functions. In [RKK18, Theorem 5], it is also shown that Adam can converge to its optimum if the parameters β_1, β_2 change over time.

As pointed out in [RKK18], the paper that introduced Adam [KB15, Corollary 4.2] mistakenly claimed that Adam does converge, i.e. it has R_T/T converging to zero as $T \to \infty$. Investigating this issue then led to Theorem 1.4.

Note that in the above four examples of BERT, GPT-3, Llama 2 and DeepSeek-V3, they already choose $\beta_1 < \sqrt{\beta_2}$, i.e. they choose parameters where Theorem 1.4 applies.

Nevertheless, the results of [RKK18] left open the question of the existence of similar counterexamples for $\beta_1 \geq \sqrt{\beta_2}$. Moreover, the choice of slope c can be arbitrarily large when β_1 or β_2 are close to 1, i.e. $c \approx \max(1/\log(1/\beta_1), 1/\log(1/\beta_2))$ is required in [RKK18]. So, it was not clear if an example for Adam with nonzero average regret could be constructed with uniformly bounded slopes, even when $\beta_1 < \sqrt{\beta_2}$.

1.1. **Our Contribution.** In this work we provide such a family of examples with nonzero average regret for Adam for all parameters $\beta_1, \beta_2 \in [0,1)$ and with uniformly bounded gradients.

Theorem 1.5 (Main). *Let $\alpha_t > 0$ with $\lim_{t\to\infty} \alpha_t = 0$, $\lim_{t\to\infty} \frac{\alpha_{t+1}}{\alpha_t} = 1$, $\sum_{t=1}^{\infty} \alpha_t = \infty$. $\forall$ $\beta_1, \beta_2 \in [0,1)$, $\varepsilon \geq 0$, $\exists$ $f_1, f_2, \ldots : [-1,1] \to \mathbb{R}$ with $1 \leq |f_t'(x)| \leq 3$, $\forall$ $t \geq 1, x \in [-1,1]$ such that Adam has regret satisfying: R_T/T does not converge to zero as $T \to \infty$.*

The example we use is simply $f_t(x) = -x$ for all $t \geq 1$ except $t \bmod 3 = 1$, in which case $f_t(x) = (2+\delta)x$, for all $x \in [-1,1]$, where $\delta > 0$ is chosen to be sufficiently small, depending on $\beta_1, \beta_2, \varepsilon$. That is, the slope of f_t is 3-periodic.

Since $1 \leq |f_t'(x)| \leq 3$ for all $t \geq 1$, $x \in [-1,1]$, the derivatives of the functions are uniformly bounded above and below, for all β_1, β_2.

This same example showed nonzero regret of the $\beta_1 = 0$ case (known as RMSProp) of Adam in [RKK18, Theorem 6] and [HWD19, Lemma 1], inspiring Theorem 1.5.

Despite the similarity of our example to the one from [RKK18], our analysis is different and arguably simpler.

As in [RKK18, Theorem 6] in the $\beta_1 = 0$ case of Adam, we show that every three iterations of Adam leads to a net positive movement of $x_1, x_2, \ldots$ towards the point $x = 1$, whereas the regret minimizer is $x = -1$. However, we depart from [RKK18] by using an elementary fixed point argument via the contractive mapping theorem. A related perspective was used in [BW19], albeit for quadratic functions.

This example also shows nonzero average regret for AdamW, RMSProp, NAdam, Adan, AdaMax, and Muon.

One might naturally ask if Theorem 1.5 holds when the highly structured periodic $f_1, f_2, \ldots$ is changed to a less structured i.i.d. variant of the above example, e.g. if for any $t \geq 1$, $f_t(x) = ax$ with probability $1/3$, and $f_t(x) = -x$ with probability $2/3$, where $f_1, f_2, \ldots$ are all i.i.d. We show the same nonzero average regret conclusion does hold in this case. We present this result in the Appendix, Section A. Consequently, the 3-periodicity of the example used in Theorem 1.5 is not required to obtain the theorem's conclusion.

The proof of Theorem 1.5 is written for the uncorrected moments. The same projected update with standard bias-corrected moments, with

$$\widetilde{m}_t = \frac{m_t}{1-b^t}, \qquad \widetilde{v}_t = \frac{v_t}{1-q^t}, \qquad h_t := \frac{\widetilde{m}_t}{\sqrt{\widetilde{v}_t} + \varepsilon},$$

has the same asymptotic properties, since $(1-b^t)^{-1}$ and $(1-q^t)^{-1}$ tend to one as $t \to \infty$. Therefore the same asymptotic argument applies to the bias-corrected case.

1.2. **Outline of Proof of Theorem 1.5.**

- Let $a \geq 2$. Let $f_{3k+1}(x) = ax$ and $f_{3k+2}(x) = f_{3k+3}(x) = -x$ for all $k \geq 0$, $x \in \mathbb{R}$.
- A contractive mapping argument shows $(m_{3k+1}, m_{3k+2}, m_{3k+3})$ and $(v_{3k+1}, v_{3k+2}, v_{3k+3})$ from (2) converge to $(M_1(a), M_2(a), M_3(a))$ and $(V_1(a), V_2(a), V_3(a))$, as $k \to \infty$.
- Verify that the negative mean drift $S(a) := \sum_{i=1}^{3} \frac{M_i(a)}{\sqrt{V_i(a)+\varepsilon}}$ of $x_1, x_2, \ldots$ from three iterations of Adam, is negative when $a = 2$.
- A continuity argument shows, for $\delta > 0$ small enough, $S(a) = S(2+\delta) < 0$, so the negative mean drift is still negative for such a.
- Conclude then that $\lim_{t\to\infty} x_t = 1$.
- Since $\sum_{t=1}^{3} f_t(x) = \delta x$, $\sum_{t=1}^{T} f_t(x)$ is minimized at $x = -1$ for T large, so $\lim_{T\to\infty} R_T/T = 2\delta/3 > 0$, thereby completing the proof.

This argument is flexible enough to extend to other variants of Adam.

Theorem 1.6. *Theorem 1.5 also holds for: AdamW, NAdam, Adan, AdaMax and Muon*

Theorem 1.7. *Let $\alpha > 0$. Then Theorem 1.5 holds almost surely for Adam with i.i.d. selection of the functions $f_1, f_2, \ldots$. and with step size $\alpha_t = \alpha/\sqrt{t}$ for all $t \geq 1$*

1.3. **Organization.** Theorem 1.5 will be proven in Section 5, by combining the previous Sections 2, 3 and 4.

Theorem 1.6 will be stated more formally as separate versions of Theorem 1.5, spread across the following sections: 6 for AdamW; 7 for NAdam; 8 for Adan; 9 for AdaMax; and 10 for Muon. Theorem 1.7 is proven in Section A.

1.4. **Further Discussion and Related Work.**

1.4.1. *Adam Alternatives such as AMSGrad.* Due to the convergence issues they found for Adam, Reddi et al. [RKK18] proposed AMSGrad, which adds an additional parameter $\widehat{v}_t$ to

Definition 1.1, and then changes (3) to

$$x_{t+1} := \Pi_{[-1,1]}(x_t - \alpha_t h_t), \qquad h_t := \frac{m_t}{\sqrt{\widehat{v}_t} + \varepsilon}, \qquad \widehat{v}_t := \max(v_t, \widehat{v}_{t-1}) \qquad \forall\, t \geq 1,$$

With this change, the previous periodicity issues for the squared gradient are removed. AMSGrad then has provable regret bounds of the form $R_T = O(T^{1/2})$, so in particular $R_T/T \to 0$ as $T \to \infty$ [AMM+20], assuming $\beta_1 < \sqrt{\beta_2}$. ([RKK18] also proved a regret bound of this form, but it needed to assume that β_1 decreased over time.)

Despite the superior theoretical guarantees of AMSGrad when compared to Adam, it appears that Adam is still more widely used in practice.

Subsequent variants, including Yogi [ZRS+18] and AdaBound/AMSBound [LXL+19], were partly motivated by overcoming Adam's convergence issues found in [RKK18].

1.4.2. *Adam divergence with unbounded gradients.* In this work, we fix the parameters $\beta_1, \beta_2 \in [0, 1)$, and then produce an example of nonzero average regret for Adam with derivatives uniformly bounded above and below. One might make these choices in the opposite order, i.e. fixing a function sequence (with possibly large derivatives) and then choosing β_1, β_2 to obtain a convergent method. The latter perspective is taken in [ZCS+22, ZLC+26]. They show it is possible to choose β_1, β_2 (after the functions being optimized are fixed) such that Adam converges.

They also show that, for any $0 \leq \beta_1, \beta_2 < 1$, there are functions such that Adam on the real line (without projection) diverges. Their example [ZLC+26, Equation (3.1)] is the following quadratic modification of [RKK18]: for any $x \in \mathbb{R}$, $a > 0$, $1 \leq i \leq n-1$, $n \geq 4$,

$$f_0(x) := \begin{cases} (1+(n-1)a)x & , \text{ if } x \geq -1 \\ \frac{(1+(n-1)a)}{2}(x+2)^2 - \frac{3n}{2} & , \text{ if } x < -1. \end{cases} \qquad f_i(x) := \begin{cases} -ax & , \text{ if } x \geq -1 \\ -\frac{a}{2}(x+2)^2 + \frac{3}{2} & , \text{ if } x < -1. \end{cases}$$

There are, however, some issues with this example, namely these functions are discontinuous unless $a = 1$, and the proof of [ZLC+26, Theorem 3.5] is only provided when $a = 1$ and when the step size is constant in each training epoch. These issues are fixable, but more importantly condition C1 [ZLC+26, Equation (8.3)] seems to require $\beta_1 < \sqrt{1 - \beta_2}$, i.e. not all $\beta_1, \beta_2 \in [0, 1)$ are covered by their proof for the $a = 1$ case; similarly, the suggested choice of $a = (n-1)^{-2}$ does not seem to allow all β_1, β_2 values in condition C1. Also, condition C3 [ZLC+26, Equation (8.5)] requires choosing a suitably small step size. In any case, [ZLC+26, Theorem 3.5] is incomparable to our Theorem 1.5 since their functions have quadratic components with unbounded gradients on an unbounded domain, whereas our functions have gradients bounded above and below on the bounded domain $[-1, 1]$ with projection onto that domain. Despite the above issues, the following modification should reproduce the result of [ZLC+26, Theorem 3.5]: $f_i(x) = -x^2$ for $0 < i < 2n/3$ and $f_i(x) = 16x^2$ for $2n/3 \leq i \leq n-1$, where n is chosen sufficiently large depending on β_1, β_2, since on the set $[0, \infty)$ we have $m_i/\sqrt{v_i} \approx -1$ for most $0 < i < 2n/3$ and $m_i/\sqrt{v_i} \approx 1$ for most $2n/3 \leq i \leq n-1$, so that $x_1, x_2, \ldots$ tends toward $+\infty$ while the true minimum occurs at 0.

A different perspective for Adam is taken in Ahn, Zhang, Kook, and Dai [AZK+24] where they interpret Adam as a discounted Follow-the-Regularized-Leader method.

1.4.3. *Dynamical Systems Approach.* Da Silva and Gazeau [BG20] derive a continuous-time ODE system for adaptive first-order methods and analyze the convergence and stability of the limiting dynamics.

Bai, Zhao, Zhou, Xu, and Zhang [BZZ+26] study Adam on highly degenerate polynomials and give a hyperparameter phase diagram containing stable convergence, spikes, and SignGD-like oscillation regimes. These papers concern related adaptive optimizers and stability phenomena, but not the bounded online-linear regret setting of Theorem 1.5.

1.4.4. *Nonconvergence in traditional stochastic optimization frameworks.* The results below concern traditional stochastic optimization, instead of online optimization.

Wang and Klabjan [WK22] give stochastic divergence examples for Adam in unconstrained strongly convex optimization, including examples that diverge in expectation or with high probability and examples that persist for large mini-batches. They also propose a variance-reduced Adam-type method and prove convergence under a variance-reduction assumption.

Dereich, Graeber, and Jentzen [DGJ24] prove a nonconvergence result for Adam and other adaptive stochastic-gradient methods when the learning rates are asymptotically bounded away from zero.

Dereich, Do, Jentzen, and von Wurstemberger [DDJ+25] prove an Adam symmetry theorem for stochastic strongly convex quadratic problems. In their formulation, Adam converges to the true minimizer if and only if the data distribution is symmetric.

Jentzen and Riekert [JR25] prove that Adam and SGD-type methods can fail with high probability to converge to global minimizers in shallow ReLU-network training landscapes. Do, Hannibal, and Jentzen [DHJ24] prove analogous high-probability nonconvergence to global minimizers for a broad class of SGD methods, including Adam, in data-driven supervised deep learning with ReLU activations. Do, Jentzen, and Riekert [DJR25] show nonconvergence of the true risk to the optimal risk for a large class of SGD-type methods, again including Adam.

Toint [Toi23] gives a very simple deterministic one-dimensional example showing that fixed-stepsize Adam can diverge on a smooth function with Lipschitz continuous gradient, without gradient noise, irrespective of the method parameters.

1.4.5. *Contrast with NosAdam, AMSGrad, AdaGrad.* The one-dimensional counterexample we presented for the nonzero average regret of Adam and its relatives does not extend in a straightforward way to Adam variants with "longer long-term memory" such as AdaGrad, AMSGrad, NosAdam, etc. For example, instead of using the iteration for v_t from (2), AMSGrad keeps track of the maximum of v_t with the additional parameter $\widehat{v}_t := \max(v_t, \widehat{v}_{t-1})$, and it then uses $h_t := \frac{m_t}{\sqrt{\widehat{v}_t}+\varepsilon}$ in (3). This eliminates the periodicity issue of v_t that occurs for these counterexamples. And indeed, these other methods often have better provable regret bounds than Adam.

2. Steady-state Moments via Contraction

We now prepare to prove Theorem 1.5. We first show the promised convergence of m_{3k+i} and v_{3k+1} as $k \to \infty$ using the contractive mapping theorem.

Throughout this paper, we assume the gradients $g_t = f_t'(x_t)$ from (2) satisfy

$$g_{3k+1} = a = 2 + \delta, \qquad g_{3k+2} = -1, \qquad g_{3k+3} = -1, \qquad \forall\, k \geq 0 \tag{4}$$

where $\delta > 0$ will be chosen sufficiently small.

Lemma 2.1. *Assume (4) holds. Then there exist unique triples* (M_1, M_2, M_3) *and* (V_1, V_2, V_3) *that are fixed points of three iterations of (2). Moreover,* $|m_{3k+i} - M_i| \le O(b^{3k})$ *and* $|v_{3k+i} - V_i| \le O(q^{3k})$ *for all* $k \ge 0$, $1 \le i \le 3$.

Note that, by (4), the iteration (2) does not depend on x_t.

Remark 2.2. *We will show using elementary algebra that*

$$M_1(a) = \frac{a - b - b^2}{1 + b + b^2}, \qquad M_2(a) = \frac{ab - b^2 - 1}{1 + b + b^2}, \qquad M_3(a) = \frac{ab^2 - b - 1}{1 + b + b^2}. \tag{5}$$

$$V_1(a) = \frac{a^2 + q + q^2}{1 + q + q^2}, \qquad V_2(a) = \frac{a^2 q + q^2 + 1}{1 + q + q^2}, \qquad V_3(a) = \frac{a^2 q^2 + q + 1}{1 + q + q^2}. \tag{6}$$

Here we added the parameter a *to our notation to emphasize the dependence of* M_i, V_i *on* a.

Proof. Let $M_1 \in \mathbb{R}$. Recall $g_1 = a$, $g_2 = g_3 = -1$, by (4), so two iterations of (2) give

$$M_2 \overset{(2)}{=} bM_1 - (1 - b), \qquad t = 2 \tag{7}$$

$$M_3 \overset{(2)}{=} bM_2 - (1 - b) \overset{(7)}{=} b^2 M_1 - (1 - b)(1 + b), \qquad t = 3. \tag{8}$$

If we have a fixed point (M_1, M_2, M_3), then (2) for $t = 4$ should return to M_1, i.e. M_1 would be equal to (using $g_4 = a$)

$$bM_3 + (1 - b)a \overset{(8)}{=} b^3 M_1 + (1 - b)(a - b - b^2).$$

Thus the one-period return map for M_1 is the affine contraction $\Phi_b \colon \mathbb{R} \to \mathbb{R}$ defined by

$$\Phi_b(M) := b^3 M + (1 - b)(a - b - b^2), \qquad \forall\, M \in \mathbb{R}. \tag{9}$$

Since $b \in [0, 1)$, we have $b^3 < 1$, so Φ_b has a unique fixed point by the contractive mapping theorem ($|\Phi_b(M) - \Phi_b(M')| \le b^3 |M - M'| < |M - M'|$ for all $M, M' \in \mathbb{R}$). Similarly, M_2, M_3 are each the unique fixed point of a contraction, each of the form $M \mapsto b^3 M + \text{constant}$. Solving for M_1 in $\Phi_b(M_1) = M_1$ gives the first part of (5), then (7) and (8) yield the last part of (5). The contraction property for Φ_b implies $|m_{3k+i} - M_i| \le O(b^{3k})$, $\forall\, k \ge 0$, $1 \le i \le 3$.

The argument for (V_1, V_2, V_3) is analogous. Since g_t does not depend on x_t, three iterations of (2) for v_t results in a contractive mapping of the form $\Psi_q(V) := q^3 V +$ constant, i.e.

$$\Psi_q(V) := q^3 V + (1 - q)(a^2 + q + q^2), \qquad \forall\, V \in \mathbb{R}. \tag{10}$$

Since $q \in [0, 1)$, this map has a unique fixed point V_1, by the contractive mapping theorem. Solving for $\Psi_q(V_1) = V_1$ produces the first equation in (6), and then (2) yields the last two parts of (6). The contraction property implies $|v_{3k+i} - V_i| \le O(q^{3k})$, $\forall\, k \ge 0$, $1 \le i \le 3$. □

3. Drift Away from the Minimizer

In Lemma 2.1, we found exponential convergence of the m_{3k+i} and v_{3k+i} terms from (2) to their limiting values as $k \to \infty$, $\forall\, 1 \le i \le 3$. In this section, we then deduce the "drift" of the iterates x_t themselves to the right endpoint $x = 1$. This "drift" will be quantified by

$$S(a) := \sum_{i=1}^{3} \frac{M_i(a)}{\sqrt{V_i(a)} + \varepsilon}. \tag{11}$$

If $S(a) < 0$, then the Adam update $x \mapsto x - \alpha_t h_t$ has positive net drift toward $x = 1$.

Lemma 3.1. *There exists $d \in (0,1)$ such that, for all $0 \le \delta \le d$, $a := 2+\delta$ satisfies*

$$S(a) < 0, \qquad M_1(a) > 0, \qquad M_2(a) < 0, \qquad M_3(a) < 0. \tag{12}$$

Proof. Let $a = 2$. Since $b \in [0,1)$, we have by (5) that

$$M_1(2) = \frac{(1-b)(2+b)}{1+b+b^2} > 0, \quad M_2(2) = -\frac{(1-b)^2}{1+b+b^2} < 0, \quad M_3(2) = -\frac{(1-b)(1+2b)}{1+b+b^2} < 0. \tag{13}$$

Moreover,

$$M_1(2) + M_2(2) + M_3(2) = 0. \tag{14}$$

For the second moment, we have by (6)

$$V_1(2) = \frac{4+q+q^2}{1+q+q^2}, \qquad V_2(2) = \frac{1+4q+q^2}{1+q+q^2}, \qquad V_3(2) = \frac{1+q+4q^2}{1+q+q^2}. \tag{15}$$

Since $q < 1$,

$$V_1(2) - V_2(2) \overset{(15)}{=} \frac{3(1-q)}{1+q+q^2} > 0, \qquad V_1(2) - V_3(2) \overset{(15)}{=} \frac{3(1-q^2)}{1+q+q^2} > 0.$$

Therefore

$$\sqrt{V_1(2)} + \varepsilon > \sqrt{V_2(2)} + \varepsilon, \qquad \sqrt{V_1(2)} + \varepsilon > \sqrt{V_3(2)} + \varepsilon. \tag{16}$$

Let

$$A := -M_2(2) \overset{(13)}{>} 0, \qquad B := -M_3(2) \overset{(13)}{>} 0. \tag{17}$$

Since $M_1(2) = A + B$ by (14), we obtain

$$\begin{aligned} S(2) &\overset{(11)}{=} \frac{A+B}{\sqrt{V_1(2)}+\varepsilon} - \frac{A}{\sqrt{V_2(2)}+\varepsilon} - \frac{B}{\sqrt{V_3(2)}+\varepsilon} \\ &= A\left(\frac{1}{\sqrt{V_1(2)}+\varepsilon} - \frac{1}{\sqrt{V_2(2)}+\varepsilon}\right) + B\left(\frac{1}{\sqrt{V_1(2)}+\varepsilon} - \frac{1}{\sqrt{V_3(2)}+\varepsilon}\right) \overset{(16)\wedge(17)}{<} 0. \end{aligned} \tag{18}$$

This strict negativity holds for every $b, q \in [0,1)$ and every $\varepsilon \ge 0$.

By continuity of $S(a)$ via (11), (5) and (6), there exists $d > 0$ such that (12) holds for all $0 \le \delta \le d$. Replacing d by $\min(d, .9)$ completes the proof. $\square$

4. A projection lemma

The net drift result of Section 3 does not immediately apply to Adam, due to the projection term $\Pi_{[-1,1]}$ in (3). In this section, we therefore analyze this projection term applied thrice.

Lemma 4.1. *Let $P = \Pi_{[-1,1]}$ so $P(x) = -1_{\{x<-1\}} + x1_{\{-1\le x\le 1\}} + 1_{\{x>1\}}$ for all $x \in \mathbb{R}$. If $u_1 \le 0$, $u_2 \ge 0$, $u_3 \ge 0$, and $U := u_1 + u_2 + u_3$, then for every $x \in [-1,1]$,*

$$P\big(P(P(x+u_1)+u_2)+u_3\big) \ge P(x+U). \tag{19}$$

Consequently, if $U \ge c > 0$, then

$$P\big(P(P(x+u_1)+u_2)+u_3\big) \ge \min(1, x+c). \tag{20}$$

Proof. Let $x \in [-1,1]$. Since $x \le 1$ and $u_1 \le 0$, we have $x + u_1 \le 1$, hence

$$P(x+u_1) = \max(-1, x+u_1) \ge x + u_1.$$

Also, for any $z \in \mathbb{R}$ and any $w \ge 0$,

$$P(P(z)+w) \ge P(z+w). \tag{21}$$

Indeed, if $z \le 1$, then $P(z) \ge z$, and the claim follows from monotonicity of P; if $z > 1$, both sides are equal to 1 since $P(z) = 1$ and $w \ge 0$. Applying (21) twice (using $u_2, u_3 \ge 0$)

$$P\big(P(P(x+u_1)+u_2)+u_3\big) \ge P(x+u_1+u_2+u_3) = P(x+U).$$

So (19) holds. Now, assume $U \ge c > 0$. Then monotonicity of P gives

$$P(x+U) \ge P(x+c) = \min(1, x+c),$$

where the last equality uses $x \in [-1,1]$ so that $x + c \ge -1$. □

5. Main theorem

We now prove Theorem 1.5, restated as Theorem 5.1 below.

Theorem 5.1. *Fix $b, q \in [0,1)$, $\varepsilon \ge 0$, and α_t satisfying $\lim_{t\to\infty} \alpha_t = 0$, $\lim_{t\to\infty} \alpha_{t+1}/\alpha_t = 1$ and $\sum_{t=1}^{\infty} \alpha_t = \infty$. There exists $\delta > 0$, depending only on b, q and ε, such that for every initial point $x_1 \in [-1,1]$, projected Adam on $[-1,1]$ applied to*

$$f_{3k+1}(x) = (2+\delta)x, \qquad f_{3k+2}(x) = -x, \qquad f_{3k+3}(x) = -x, \tag{22}$$

satisfies

$$\lim_{t\to\infty} x_t = 1. \tag{23}$$

For every horizon $T \ge 1$, the best fixed comparator in $[-1,1]$ is $x_T^* = -1$, and the average regret satisfies

$$\lim_{T\to\infty} \frac{R_T}{T} = \frac{2\delta}{3} > 0. \tag{24}$$

Proof. Choose $\delta > 0$ so that (12) holds. Let $a := 2 + \delta$. Since the m_{3k+i} and v_{3k+i} terms converge as $k \to \infty$ by Lemma 2.1 $\forall\, 1 \le i \le 3$, (3) implies that the h_{3k+i} terms also converge:

$$\lim_{k\to\infty} h_{3k+i} = H_i := \frac{M_i(a)}{\sqrt{V_i(a)} + \varepsilon}, \qquad \forall\, i = 1,2,3. \tag{25}$$

By (12),

$$H_1 > 0, \qquad H_2 < 0, \qquad H_3 < 0, \qquad H_1 + H_2 + H_3 = S(a) < 0. \tag{26}$$

Let $\eta := -S(a) > 0$. Define the three unprojected increments

$$u_{k,i} := -\alpha_{3k+i} h_{3k+i}, \qquad \forall\, k \ge 0, \qquad i = 1,2,3. \tag{27}$$

Then for all k sufficiently large, (26) and (25) imply

$$u_{k,1} < 0, \qquad u_{k,2} > 0, \qquad u_{k,3} > 0. \tag{28}$$

Furthermore, using $\lim_{t\to\infty} \alpha_{t+1}/\alpha_t = 1$,

$$U_k := u_{k,1} + u_{k,2} + u_{k,3} \overset{(27)}{=} -\sum_{i=1}^{3} \alpha_{3k+i} h_{3k+i} \overset{(25)}{=} -\alpha_{3k+1} \cdot \left(\sum_{i=1}^{3} H_i + o_k(1) \right) = \alpha_{3k+1}(\eta + o_k(1)).$$

Consequently, there are constants $c > 0$ and $k_0 \geq 1$ such that

$$U_k \geq c \cdot \alpha_{3k+1} \qquad \text{for all } k \geq k_0. \tag{29}$$

Let $z_k := x_{3k+1}$ be the iterate at the start of a period, for all $k \geq 0$. Applying Lemma 4.1 with $u_i = u_{k,i}$ for each $1 \leq i \leq 3$ which is valid by (28) and (29),

$$z_{k+1} \overset{(3)\wedge(27)}{\geq} \min\Big(1, z_k + c \cdot \alpha_{3k+1}\Big), \qquad \forall\, k \geq k_0. \tag{30}$$

Since $z_k \leq 1$ for all $k \geq 1$ and $\sum_{k \geq k_0} \alpha_k = \infty$, we have $\sum_{k \geq k_0} \alpha_{3k+1} = \infty$, which follows since $\lim_{k\to\infty} \alpha_{k+1}/\alpha_k = 1$. Then iterating (30) forces z_k to equal 1 after finitely many periods. Indeed, once $\sum_{k=k_0}^{k_1} c \cdot \alpha_{3k+1} > 1 - z_{k_0}$ for some $k_1 > k_0$, (30) gives $z_{k+1} \geq 1$, while projection onto $[-1, 1]$ from (3) gives $z_{k+1} \leq 1$. Then $\forall\ k > k_1$, (30) gives $z_{k+1} = 1$ whenever $z_k = 1$.

This implies convergence of the full sequence (x_t) to 1, since the within-period moves have magnitude $O(\alpha_t)$. To see this, note by (2) that (m_t) is bounded since (g_t) is bounded, i.e. $|m_t| \leq \max(|m_{t-1}|, |g_t|) \leq 3$ for all $t \geq 0$. Also, since $|g_t| \geq 1$ for every t,

$$v_t \overset{(2)}{=} q v_{t-1} + (1-q) g_t^2 \overset{(2)}{\geq} 1 - q > 0.$$

Hence

$$|h_t| \overset{(3)}{=} \left| \frac{m_t}{\sqrt{v_t} + \varepsilon} \right| \leq \frac{3}{\sqrt{1-q}}, \qquad \forall\, t \geq 1.$$

Therefore $|x_{t+1} - x_t| \leq |\alpha_t| \frac{3}{\sqrt{1-q}}$, $\forall\ t \geq 1$. Since $x_{3k+1} \to 1$ as $k \to \infty$ and $\lim_{t\to\infty} \alpha_t = 0$, this implies that $x_t \to 1$ as $t \to \infty$. That is, (23) holds.

It remains to prove (24). For any $T \geq 1$, let

$$G_T := \sum_{t=1}^{T} g_t. \tag{31}$$

Writing $T = 3k + r$, $r \in \{0, 1, 2\}$, gives (recalling $a = 2 + \delta$)

$$G_T \overset{(4)}{=} \begin{cases} k\delta, & r = 0, \\ k\delta + a, & r = 1, \\ k\delta + a - 1, & r = 2. \end{cases} \tag{32}$$

All three quantities are positive since $a = 2 + \delta$ and $\delta > 0$. Therefore the best fixed comparator is always $x_T^* = -1$, and

$$\min_{x\in[-1,1]} \sum_{t=1}^{T} f_t(x) \overset{(22)\wedge(31)}{=} G_T \cdot \min_{x\in[-1,1]} x = -G_T. \tag{33}$$

The regret is then

$$R_T \overset{(1)\wedge(22)\wedge(33)}{=} \sum_{t=1}^{T} g_t x_t + G_T \overset{(31)}{=} \sum_{t=1}^{T} g_t \cdot (x_t + 1). \tag{34}$$

Since $\lim_{t\to\infty} x_t = 1$ by (23) and (g_t) is bounded by (4),

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} g_t \cdot (x_t - 1) = 0.$$

Also $\lim_{T\to\infty} G_T/T = \delta/3$ by (32). Hence

$$\frac{R_T}{T} \overset{(34)\wedge(31)}{=} \frac{2G_T}{T} + \frac{1}{T}\sum_{t=1}^{T} g_t(x_t - 1) \to \frac{2\delta}{3}, \qquad \text{as } T \to \infty.$$

This proves (24) and completes the proof. ☐

6. AdamW

We now define AdamW and extend Theorem 1.5 to AdamW.

Definition 6.1 (**AdamW Optimization Method**)**.** *Let $x_1 \in [-4,-2]$ be arbitrary. Define $x_2, x_3, \ldots \in [-4,-2]$ as follows. The first and second moment recursions of AdamW* [LH19] *with parameters $b = \beta_1$ and $q = \beta_2$ are the same as Adam, i.e.* (2) *holds, with the standard initialization $m_0 = v_0 = 0$. The projected update then adds a single extra term $\lambda \geq 0$ to* (3) *as follows*

$$x_{t+1} := \Pi_{[-4,-2]}\left((1-\lambda\alpha_t)x_t - \alpha_t h_t\right), \qquad h_t := \frac{m_t}{\sqrt{v_t} + \varepsilon}, \qquad \forall\, t \geq 1. \tag{35}$$

Theorem 6.2 (AdamW Counterexample)**.** *Let $\alpha_t > 0$ satisfy $\lim_{t\to\infty} \alpha_t = 0$, $\lim_{t\to\infty} \frac{\alpha_{t+1}}{\alpha_t} = 1$ and $\sum_{t=1}^{\infty} \alpha_t = \infty$. Fix $\beta_1, \beta_2 \in [0,1)$, $\varepsilon \geq 0$ and $\lambda \geq 0$. Consider projected AdamW on the domain $\mathcal{F} = [-4,-2]$. Then there exists $\delta > 0$, depending only on b, q, ε, such that for the linear functions* (22)*, the iterates of AdamW satisfy*

$$\lim_{t\to\infty} x_t = -2.$$

However, the best fixed comparator is $x^* = -4$, and

$$\lim_{T\to\infty} \frac{R_T}{T} = \frac{2\delta}{3}.$$

Proof. Define $u_{k,i}$ from (27). The corresponding increments for AdamW are then

$$u_{k,i} - \lambda\alpha_{3k+i}x_{3k+i}.$$

On $\mathcal{F} = [-4,-2]$, we have $x_{3k+i} \leq -2$, and hence

$$-\lambda\alpha_{3k+i}x_{3k+i} \geq 0.$$

We now adapt the remaining parts of Theorem 5.1 to AdamW.

Put $\mathcal{F} = [-4,-2]$, and for any $t \geq 1$, let

$$u_t := -\alpha_t h_t.$$

For fixed t, g_t is a constant that does not depend on x_t, so (h_t) also does not depend on x_t.

Since $\lim_{t\to\infty} \alpha_t = 0$, we may choose $T = 3K + 1$ sufficiently large that

$$0 \leq \lambda\alpha_t \leq 1, \qquad \forall\, t \geq T.$$

For any $t \geq T$, define the one-step maps

$$\mathcal{A}_t(x) := \Pi_{\mathcal{F}}(x + u_t), \qquad \mathcal{W}_t(x) := \Pi_{\mathcal{F}}\big((1-\lambda\alpha_t)x + u_t\big), \qquad \forall\, x \in [-4,-2].$$

The map $\mathcal{W}_t$ is nondecreasing since $1 - \lambda\alpha_t \geq 0$ and Π_F is nondecreasing. Furthermore, for every $x \in F$,

$$(1-\lambda\alpha_t)x + u_t = x + u_t - \lambda\alpha_t x \geq x + u_t,$$

since $x \leq -2 < 0$. Therefore

$$\mathcal{W}_t(x) \geq \mathcal{A}_t(x), \qquad \forall\, x \in \mathcal{F}. \tag{36}$$

Let $(y_t)_{t \geq T}$ be the auxiliary projected Adam sequence without weight decay, initialized by

$$y_T := x_T, \qquad y_{t+1} := \mathcal{A}_t(y_t), \qquad \forall\, t \geq T.$$

The period-three argument from Theorem 5.1, translated from $[-1,1]$ to $\mathcal{F} = [-4,-2]$, gives $\lim_{t\to\infty} y_t = -2$. We claim that

$$x_t \geq y_t, \qquad \forall\, t \geq T.$$

This holds at time T by definition of y_T. If it holds at any $t \geq T$, then monotonicity of $\mathcal{W}_t$ and (36) give

$$x_{t+1} \overset{(35)}{=} \mathcal{W}_t(x_t) \geq \mathcal{W}_t(y_t) \geq \mathcal{A}_t(y_t) = y_{t+1}.$$

Thus the claim follows by induction. Since both sequences lie in $[-4,-2]$, $-2 \geq x_t \geq y_t$ and $\lim_{t\to\infty} y_t = -2$ imply that $\lim_{t\to\infty} x_t = -2$.

The cumulative gradient over each period is still $\delta > 0$, so the best fixed comparator on $[-4,-2]$ is the left endpoint -4. Since the iterates converge to the right endpoint -2 and the interval length is 2, the same regret computation from (24) gives $\lim_{T\to\infty} \frac{R_T}{T} = \frac{2\delta}{3}$. $\square$

7. NAdam

We now define NAdam and extend Theorem 1.5 to NAdam. Recall that projected NAdam is defined exactly as in Definition 1.1, but instead of the h_t from (3) we have

$$h_t := \frac{bm_t + (1-b)g_t}{\sqrt{v_t} + \varepsilon}, \qquad \forall\, t \geq 1.$$

Theorem 7.1 (NAdam Counterexample)**.** *Let $\alpha_t > 0$ satisfy $\lim_{t\to\infty} \alpha_t = 0$, $\lim_{t\to\infty} \frac{\alpha_{t+1}}{\alpha_t} = 1$ and $\sum_{t=1}^{\infty} \alpha_t = \infty$. Fix $b, q \in [0,1)$, $\varepsilon \geq 0$. Consider projected NAdam on the domain $\mathcal{F} = [-1,1]$. Then there exists $\delta > 0$, depending only on b, q, ε, such that for the linear functions (22), the iterates of NAdam satisfy*

$$\lim_{t\to\infty} x_t = 1.$$

However, the best fixed comparator is $x^* = -1$, and

$$\lim_{T\to\infty} \frac{R_T}{T} = \frac{2\delta}{3}.$$

Proof. The recursions for m_t and v_t are identical to Adam, so Lemma 2.1 applies, and the limiting values M_i, V_i from Lemma 2.1 are unchanged. For NAdam with $a = 2$, the limiting numerators of the h terms are U_1, U_2, U_3 where

$$U_i = bM_i(2) + (1-b)g_i, \qquad \forall\, i \in \{1,2,3\}.$$

Since $\sum_{i=1}^{3} M_i(2) = 0$, by (5) and $\sum_{i=1}^{3} g_i = 0$, when $a = 2$, we have $U_1 + U_2 + U_3 = 0$. Moreover $U_1 > 0$ and $U_2, U_3 < 0$, since both $M_i(2)$ and g_i have these signs by (12) and (4). The remaining details follow those of Adam in Theorem 5.1 with $S_{\text{NAdam}}(a) := \sum_{i=1}^{3} \frac{U_i(a)}{\sqrt{V_i(a)}+\varepsilon}$, $U_i(a) := bM_i(a) + (1-b)g_i$ for all $i \in \{1,2,3\}$. For example, $S_{\text{NAdam}}(2) < 0$ by repeating the proof of (18) mutatis mutandis, so $S_{\text{NAdam}}(a) < 0$ for all $a \in \mathbb{R}$ near 2, and so on. $\square$

8. Adan

We define the Adan optimization method [XZL+24] and extend Theorem 1.5 to it. Let

$$b := \beta_1, \qquad q := \beta_2, \qquad r := \beta_3, \qquad b, q, r \in [0, 1).$$

Let $g_0 := -1$, $m_0 = d_0 = n_0 := 0$. For any $t \geq 1$, define

$$m_t := b m_{t-1} + (1-b) g_t,$$
$$d_t := q d_{t-1} + (1-q)(g_t - g_{t-1}),$$
$$n_t := r n_{t-1} + (1-r)\big(g_t + q(g_t - g_{t-1})\big)^2.$$
$$x_{t+1} := \Pi_{[-1,1]}\left(x_t - \alpha_t h_t\right), \qquad h_t := \frac{m_t + q d_t}{\sqrt{n_t} + \varepsilon}.$$

This definition in terms of decay coefficients b, q, r may differ from other definitions.

Theorem 8.1 (Adan Counterexample)**.** *Let $\alpha_t > 0$ satisfy $\lim_{t\to\infty} \alpha_t = 0$, $\lim_{t\to\infty} \frac{\alpha_{t+1}}{\alpha_t} = 1$ and $\sum_{t=1}^{\infty} \alpha_t = \infty$. Fix $b, q, r \in [0, 1)$, $\varepsilon \geq 0$. Consider projected Adan on the domain $\mathcal{F} = [-1, 1]$. Then there exists $\delta > 0$, depending only on b, q, r, ε, such that for the linear functions (22), the iterates of Adan satisfy*

$$\lim_{t\to\infty} x_t = 1.$$

However, the best fixed comparator is $x^* = -1$, and

$$\lim_{T\to\infty} \frac{R_T}{T} = \frac{2\delta}{3}.$$

Proof. Since g_t does not depend on x_t, three Adan iterations for m_t or d_t or n_t results in a contractive mapping. For example, $m_{3k+1} \mapsto m_{3(k+1)+1}$ corresponds to a map $M \mapsto b^3 M + \text{constant}$. Since this map is a contraction from $\mathbb{R}$ to $\mathbb{R}$, it has a unique fixed point, by the contractive mapping theorem. Thus, $\lim_{k\to\infty}(m_{3k+1}, m_{3k+2}, m_{3k+3}) =: (M_1, M_2, M_3)$, $\lim_{k\to\infty}(n_{3k+1}, n_{3k+2}, n_{3k+3}) =: (N_1, N_2, N_3)$, $\lim_{k\to\infty}(d_{3k+1}, d_{3k+2}, d_{3k+3}) =: (D_1, D_2, D_3)$.

The first-moment values from (13) are

$$M_1 = \frac{2 - b - b^2}{1 + b + b^2}, \quad M_2 = \frac{2b - b^2 - 1}{1 + b + b^2}, \quad M_3 = \frac{2b^2 - b - 1}{1 + b + b^2}. \tag{37}$$

The gradient-difference values are

$$D_1 = \frac{3(1 - q^2)}{1 + q + q^2}, \quad D_2 = -\frac{3(1 - q)}{1 + q + q^2}, \quad D_3 = -\frac{3q(1 - q)}{1 + q + q^2}. \tag{38}$$

Thus the limiting Adan numerators are

$$U_i := M_i + q D_i, \qquad \forall\, i \in \{1, 2, 3\}. \tag{39}$$

From (37), (38) and $q \geq 0$, they satisfy

$$U_1 > 0, \qquad U_2 < 0, \qquad U_3 < 0, \qquad U_1 + U_2 + U_3 = 0.$$

Write

$$A := -U_2 > 0, \qquad B := -U_3 > 0,$$

so that $U_1 = A + B$. Set

$$s_1 := (2 + 3q)^2, \qquad s_2 := (1 + 3q)^2, \qquad s_3 := 1.$$

Then

$$N_1 = \frac{s_1 + r^2 s_2 + r s_3}{1 + r + r^2}, \qquad N_2 = \frac{r s_1 + s_2 + r^2 s_3}{1 + r + r^2}, \qquad N_3 = \frac{r^2 s_1 + r s_2 + s_3}{1 + r + r^2}.$$

Define

$$w_i := \frac{1}{\sqrt{N_i} + \varepsilon}, \qquad \forall\, i \in \{1, 2, 3\}. \tag{40}$$

The limiting drift of three Adan iterations at $a = 2$ is

$$S_{\text{Adan}}(2) = \sum_{i=1}^{3} U_i w_i = A(w_1 - w_2) + B(w_1 - w_3). \tag{41}$$

We always have $N_1 > N_3$ since $N_1 - N_3 = (1 - r)[(s_1 - s_3) + r(s_1 - s_2)]/(1 + r + r^2) > 0$. If $N_1 \geq N_2$, then $w_1 \leq w_2$ and $w_1 < w_3$, so $S_{\text{Adan}}(2) < 0$.

It remains to consider the case $N_2 > N_1 > N_3$. (In the case $q = 0$, we have $s_1 = 4, s_2 = 1$, and $N_1 - N_2 = 3(1 - r)/(1 + r + r^2) > 0$, i.e. this case cannot occur when $q = 0$, so we may assume $q > 0$.) The function $\phi\colon [0, \infty) \to \mathbb{R}$ defined by

$$\phi(x) = \frac{1}{\sqrt{x} + \varepsilon}, \qquad \forall\, x \geq 0$$

has $-\phi'(x)$ positive and decreasing. Hence using $N_2 > N_1 > N_3$

$$\frac{w_3 - w_1}{w_1 - w_2} \overset{(40)}{=} \frac{\int_{N_3}^{N_1} -\phi'(x)dx}{\int_{N_1}^{N_2} -\phi'(x)dx} \geq \frac{N_1 - N_3}{N_2 - N_1}. \tag{42}$$

A direct calculation gives

$$\frac{N_1 - N_3}{N_2 - N_1} > \frac{1}{q}. \tag{43}$$

Indeed, after canceling the common $(1 + r + r^2)/(1 - r)$ factor, this is equivalent to

$$q\big((1 + r)s_1 - r s_2 - s_3\big) > (1 + r)s_2 - s_1 - r s_3,$$

and the left side minus the right side is equal to

$$3\Big(3q^3 + 4q^2 + 3q + 1 - r(q^2 + q)\Big) > 0.$$

On the other hand,

$$\frac{A}{B} \leq \frac{1}{q}.$$

This follows since, using $b, q \in [0, 1)$ and $q \neq 0$

$$\frac{-M_2}{-M_3} \overset{(37)}{=} \frac{1 - b}{1 + 2b} \leq 1 < \frac{1}{q}, \tag{44}$$

while

$$\frac{-qD_2}{-qD_3} \overset{(38)}{=} \frac{1}{q}. \tag{45}$$

Therefore

$$\frac{A}{B} = \frac{U_2}{U_3} \overset{(39)}{=} \frac{-M_2 - qD_2}{-M_3 - qD_3} \overset{(44)\wedge(45)}{\leq} \frac{1}{q} \overset{(42)\wedge(43)}{<} \frac{w_3 - w_1}{w_1 - w_2}.$$

Rearranging this inequality gives

$$A(w_1 - w_2) - B(w_3 - w_1) < 0,$$

so by (41) we get (in all cases) that

$$S_{\text{Adan}}(2) < 0.$$

By continuity, there exists $\delta > 0$ such that, with $a = 2 + \delta$, the limiting Adan period sum $\sum_{i=1}^{3} U_i(a) w_i(a)$ remains negative and the numerator signs remain

$$U_1(a) > 0, \qquad U_2(a) < 0, \qquad U_3(a) < 0.$$

The projection and regret argument from Theorem 5.1 then applies verbatim. □

9. AdaMax

We define the AdaMax optimization method and extend Theorem 1.5 to it. The AdaMax optimization method is defined by the following recursions:

$m_t := bm_{t-1} + (1-b)g_t$, $v_t := \max(qv_{t-1}, |g_t|)$, $h_t := \frac{m_t}{v_t + \varepsilon}$, x_{t+1} as in (3), for all $t \geq 1$.

As usual, $m_0 = v_0 = 0$.

Theorem 9.1 (AdaMax Counterexample)**.** *Let $\alpha_t > 0$ satisfy $\lim_{t\to\infty} \alpha_t = 0$, $\lim_{t\to\infty} \frac{\alpha_{t+1}}{\alpha_t} = 1$ and $\sum_{t=1}^{\infty} \alpha_t = \infty$. Fix $b, q \in [0, 1)$, $\varepsilon \geq 0$. Consider projected AdaMax on the domain $\mathcal{F} = [-1, 1]$. Then there exists $\delta > 0$, depending only on b, q, ε, such that for the linear functions (22), the iterates of AdaMax satisfy*

$$\lim_{t\to\infty} x_t = 1.$$

However, the best fixed comparator is $x^* = -1$, and

$$\lim_{T\to\infty} \frac{R_T}{T} = \frac{2\delta}{3}.$$

Proof. The steady-state values (M_1, M_2, M_3) from Lemma 2.1 apply here as well, since the m_t recursion for AdaMax is the same as Adam. To find the steady-state (V_1, V_2, V_3) we plug them into their recursion as follows:

$$V_2 = \max(qV_1, 1)$$
$$V_3 = \max(qV_2, 1) = \max(q^2 V_1, 1)$$

One more iteration gives

$$\max(q^3 V_1, a)$$

So the fixed point of $V \mapsto \max(q^3 V, a)$ must satisfy $V = a$. (Note that $\Phi(V) := \max(q^3 V, a)$ satisfies $|\Phi(V) - \Phi(V')| \leq q^3 |V - V'|$ for all $V, V' \in \mathbb{R}$ and $0 \leq q < 1$ implies that Φ is a contraction, so Φ has a unique fixed point.) The corresponding maps for V_2, V_3 are also contractions. Then solving for V_2, V_3 gives

$$V_1(a) = a, \quad V_2(a) = \max(qa, 1), \quad V_3(a) = \max(q^2 a, 1), \tag{46}$$

and when $a = 2$ we have

$$V_1 = 2, \quad V_2 = \max(2q, 1), \quad V_3 = \max(2q^2, 1).$$

Since $q < 1$ we therefore have

$$V_1 > V_2, \qquad V_1 > V_3, \tag{47}$$

For any $a \geq 2$, define $S_{\text{AdaMax}}(a) := \sum_{i=1}^{3} \frac{M_i(a)}{V_i(a)+\varepsilon}$. We then have

$$S_{\text{AdaMax}}(2) = \sum_{i=1}^{3} \frac{M_i(2)}{V_i(2)+\varepsilon}.$$

Let $A, B > 0$ as in (17). Since $M_1(2) = A + B$ by (14), we obtain

$$\begin{aligned} S_{\text{AdaMax}}(2) &= \frac{A+B}{V_1(2)+\varepsilon} - \frac{A}{V_2(2)+\varepsilon} - \frac{B}{V_3(2)+\varepsilon} \\ &= A\left(\frac{1}{V_1(2)+\varepsilon} - \frac{1}{V_2(2)+\varepsilon}\right) + B\left(\frac{1}{V_1(2)+\varepsilon} - \frac{1}{V_3(2)+\varepsilon}\right) \stackrel{(47)}{<} 0. \end{aligned}$$

This strict negativity holds for every $b, q \in [0,1)$ and every $\varepsilon \geq 0$. The remaining details follow those of Theorem 5.1, e.g. observing that $V_1(a), V_2(a), V_3(a)$ are continuous functions of a by (46), so $S_{\text{AdaMax}}(a) < 0$ for a near 2, and so on. $\square$

10. Muon

We define the Muon optimization method and extend Theorem 1.5 to it. The Muon optimization method is defined for functions of matrices $f_t \colon \mathcal{K} \to \mathbb{R}\ \forall\, t \geq 1$ where $\mathcal{K} \subset \mathbb{R}^{n \times n}$. The iterations satisfy

$$m_t = bm_{t-1} + (1-b)\nabla f_t(x_t), \qquad x_{t+1} = \Pi_{\mathcal{K}}\left(x_t - \alpha_t \operatorname{Polar}(m_t)\right), \qquad \forall\, t \geq 1,$$

with $m_0 = 0$, and $\Pi_{\mathcal{K}}$ denoting the projection to the nearest point in $\mathcal{K}$, with respect to the Euclidean (Frobenius) metric on $\mathbb{R}^{n \times n}$. Here $m_t, x_t \in \mathbb{R}^{n \times n}$ for all $t \geq 1$ and

$$\operatorname{Polar}(A) := A(A^*A)^{-1/2}$$

is defined for any invertible $n \times n$ matrix A [PKC+26]. More generally, $\operatorname{Polar}(A) := UV$ when A is an $n \times n$ matrix with reduced singular value decomposition $A = UDV$ (noting that the product UV is well-defined even though U, V are not uniquely determined by A). Also $\operatorname{Polar}(0) := 0$. In practice, $\operatorname{Polar}(A)$ can be approximated by Newton-Schulz iterations.

In the case $n = 1$ with $\mathcal{K} := [-1, 1] \subset \mathbb{R}$, Muon becomes a signed momentum method:

$$m_t = bm_{t-1} + (1-b)f_t'(x_t), \qquad x_{t+1} = \Pi_{[-1,1]}\left(x_t - \alpha_t \operatorname{sign}(m_t)\right), \qquad \forall\, t \geq 1,$$

(Here $\operatorname{sign}(0) := 0$.) We will demonstrate this method has nonzero average regret. The $n = 1$ example can then be extended to the $n > 1$ case by choosing each f_t to be a function of one diagonal entry of its input matrix, e.g. $\mathcal{K} = \{\operatorname{diag}(x, 0, \dots, 0) \colon x \in [-1, 1]\}$.

Theorem 10.1 (Muon Counterexample)**.** *Let $\alpha_t > 0$ satisfy $\lim_{t\to\infty} \alpha_t = 0$, $\lim_{t\to\infty} \frac{\alpha_{t+1}}{\alpha_t} = 1$ and $\sum_{t=1}^{\infty} \alpha_t = \infty$. Fix $b \in [0, 1)$. Consider Muon on the domain $\mathcal{F} = [-1, 1]$. Then there exists $\delta > 0$, depending only on b, such that for the linear functions (22), the iterates of Muon satisfy*

$$\lim_{t\to\infty} x_t = 1.$$

However, the best fixed comparator is $x^* = -1$, and

$$\lim_{T\to\infty} \frac{R_T}{T} = \frac{2\delta}{3}.$$

Proof. By Lemma 2.1, the steady-state momentum values are

$$M_1(a) = \frac{a-b-b^2}{1+b+b^2}, \qquad M_2(a) = \frac{ab-b^2-1}{1+b+b^2}, \qquad M_3(a) = \frac{ab^2-b-1}{1+b+b^2}.$$

At $a = 2$,

$$M_1(2) > 0, \qquad M_2(2) < 0, \qquad M_3(2) < 0.$$

Therefore, by continuity, there exists $\delta > 0$ such that, with $a = 2 + \delta$,

$$M_1(a) > 0, \qquad M_2(a) < 0, \qquad M_3(a) < 0.$$

By Lemma 2.1, $(m_{3k+1}, m_{3k+2}, m_{3k+3})$ converges exponentially to this period-three steady-state as $k \to \infty$. Hence, there is some $k_0 > 0$ such that, for all $k > k_0$,

$$\operatorname{sign}(m_{3k+1}) = +1, \qquad \operatorname{sign}(m_{3k+2}) = -1, \qquad \operatorname{sign}(m_{3k+3}) = -1.$$

Thus the three unprojected scalar increments in the kth period are

$$u_{k,1} = -\alpha_{3k+1}, \qquad u_{k,2} = \alpha_{3k+2}, \qquad u_{k,3} = \alpha_{3k+3}.$$

Since $\lim_{t\to\infty} \alpha_{t+1}/\alpha_t = 1$, their sum satisfies

$$U_k := -\alpha_{3k+1} + \alpha_{3k+2} + \alpha_{3k+3} = \alpha_{3k+1}(1 + o(1)).$$

Denote $z_k := x_{3k+1}$. Lemma 4.1 as used for Adam then gives

$$z_{k+1} \geq \min(1, z_k + \alpha_{3k+1}/2), \qquad \forall\, k \geq k_0.$$

Since $z_k \leq 1$ for all $k \geq 1$ and $\sum_{k\geq k_0} \alpha_k = \infty$, we have $\sum_{k\geq k_0} \alpha_{3k+1} = \infty$, which follows since $\lim_{k\to\infty} \alpha_{k+1}/\alpha_k = 1$. So, the sequence (z_k) reaches 1 at some finite value of k. The within-period moves have size $O(\alpha_{3k}) = o(1)$ by assumption, so $x_t \to 1$ as $t \to \infty$.

Finally, every period has cumulative gradient

$$(2 + \delta) - 1 - 1 = \delta > 0.$$

Hence the best fixed comparator in $[-1, 1]$ is -1. Since $\lim_{t\to\infty} x_t = 1$, the same scalar regret computation as for Adam gives

$$\lim_{T\to\infty} \frac{R_T}{T} = \frac{2\delta}{3}.$$

$\square$

Appendix A. Adam with i.i.d. Slopes

We now give a stochastic version of the Adam counterexample from Theorem 1.5. For simplicity, we only consider the step size $\alpha_t := \alpha/\sqrt{t}$ for all $t \geq 1$. Instead of presenting the gradients in the deterministic period-three order $a, -1, -1$, we draw them independently at each time t. The slope a appears with probability $1/3$, and the slope -1 appears with probability $2/3$.

More formally, let $(X_t)_{t\geq 1}$ be i.i.d. Bernoulli random variables with

$$\mathbb{P}(X_t = 1) = \frac{1}{3}, \qquad \mathbb{P}(X_t = 0) = \frac{2}{3}, \qquad \forall\, t \geq 1.$$

For any $a > 0$, define

$$g_t(a) := -1 + (a + 1)X_t, \qquad \forall\, t \geq 1.$$

Thus, for any $t \geq 1$,

$$g_t(a) = a \quad \text{with probability } 1/3, \qquad g_t(a) = -1 \quad \text{with probability } 2/3.$$

The optimized functions are again

$$f_t(x) = g_t(a)x, \qquad \forall\, x \in [-1, 1],\, t \geq 1.$$

Theorem A.1 (Adam i.i.d. random-slope counterexample)**.** *Fix $b, q \in [0, 1)$, $\varepsilon \geq 0$, and $\alpha > 0$. There exists $\delta > 0$, depending only on b, q, ε, such that, with*

$$a = 2 + \delta,$$

projected Adam on $[-1, 1]$, driven by the i.i.d. slopes

$$g_t(a) = \begin{cases} a, & \textit{with probability } 1/3, \\ -1, & \textit{with probability } 2/3, \end{cases}$$

satisfies

$$\lim_{t\to\infty} x_t = 1 \qquad \textit{almost surely.}$$

Moreover, the best fixed comparator is eventually -1, and the average regret satisfies

$$\lim_{T\to\infty} \frac{R_T}{T} = \frac{2\delta}{3} \qquad \textit{almost surely.}$$

Proof. We first analyze the balanced value $a = 2$. It is convenient to work with a two-sided i.i.d. extension $(X_t)_{t\in\mathbb{Z}}$. For $\lambda \in [0, 1)$, define the stationary weighted average

$$A_{\lambda,t} := (1 - \lambda) \sum_{\ell=0}^{\infty} \lambda^\ell X_{t-\ell}. \tag{48}$$

Then $\mathbb{E}[A_{\lambda,t}] = \frac{1}{3}$. For $a = 2$, we have, for all $t \in \mathbb{Z}$

$$g_t(2) = -1 + 3X_t, \qquad g_t(2)^2 = 1 + 3X_t.$$

The stationary versions of the first and second Adam moments are therefore

$$m_t^*(2) = (1 - b) \sum_{\ell=0}^{\infty} b^\ell g_{t-\ell}(2) = -1 + 3A_{b,t}, \qquad v_t^*(2) = (1 - q) \sum_{\ell=0}^{\infty} q^\ell g_{t-\ell}(2)^2 = 1 + 3A_{q,t}.$$

Define the stationary normalized Adam direction

$$h_t^*(2) := \frac{m_t^*(2)}{\sqrt{v_t^*(2)} + \varepsilon} = \frac{-1 + 3A_{b,t}}{\sqrt{1 + 3A_{q,t}} + \varepsilon}.$$

Step 1. Proving a positive drift. In the proof of Theorem 5.1, we used (12) to show that $S(2) < 0$ and $S(a) < 0$ for a near 2. In the current proof, the analogous statement is that $h_t^*(2)$ has negative mean. That is, we claim that the following expression does not depend on $t \in \mathbb{Z}$ and

$$\mu(2) := \mathbb{E}[h_t^*(2)] < 0. \tag{49}$$

Let $\phi(y) := \frac{1}{\sqrt{1+3y}+\varepsilon}$, $\forall\, y \geq 0$. Then ϕ is strictly decreasing on $[0, 1]$. Since $\mathbb{E}[A_{\lambda,t}] = 1/3$,

$$\mu(2) = \mathbb{E}\left[(-1 + 3A_{b,t})\phi(A_{q,t})\right] = 3\,\mathrm{Cov}(A_{b,t}, \phi(A_{q,t}))\,.$$

The random variable $A_{b,t}$ is an increasing function of the coordinates $(X_t, X_{t-1}, X_{t-2}, \ldots)$. The random variable $A_{q,t}$ is also an increasing function of these same coordinates, and therefore $\phi(A_{q,t})$ is a decreasing function of them. By the Harris correlation inequality for product measures,

$$\operatorname{Cov}(A_{b,t}, \phi(A_{q,t})) \leq 0.$$

The inequality is in fact strict. For any $t \geq 1$ define the σ-algebra

$$\mathcal{G}_t := \sigma(X_{t-1}, X_{t-2}, \ldots)$$

and define B_t, Q_t so that

$$A_{b,t} = B_t + (1-b)X_t, \qquad A_{q,t} = Q_t + (1-q)X_t,$$

and such that B_t and Q_t are $\mathcal{G}_t$-measurable. Set $p := \mathbb{P}(X_t = 1) = 1/3$. Conditional on $\mathcal{G}_t$,

$$\mathbb{E}[A_{b,t} \mid \mathcal{G}_t] = B_t + p(1-b), \qquad \mathbb{E}[\phi(A_{q,t}) \mid \mathcal{G}_t] = (1-p)\phi(Q_t) + p\phi(Q_t + 1 - q),$$

$$\operatorname{Cov}\big(A_{b,t}, \phi(A_{q,t}) \mid \mathcal{G}_t\big) = p(1-p)(1-b)\big[\phi(Q_t + 1 - q) - \phi(Q_t)\big] < 0,$$

since $b, q < 1$ and ϕ is strictly decreasing. Moreover, $\mathbb{E}[A_{b,t} \mid \mathcal{G}_t]$ is an increasing function of the coordinates $(X_{t-1}, X_{t-2}, \ldots)$, whereas $\mathbb{E}[\phi(A_{q,t}) \mid \mathcal{G}_t]$ is a decreasing function of those coordinates. The Harris correlation inequality, now applied to the product measure of the past coordinates, therefore gives

$$\operatorname{Cov}(\mathbb{E}[A_{b,t} \mid \mathcal{G}_t], \mathbb{E}[\phi(A_{q,t}) \mid \mathcal{G}_t]) \leq 0.$$

The law of total covariance consequently yields $\operatorname{Cov}(A_{b,t}, \phi(A_{q,t})) < 0$. Thus

$$\mu(2) = 3 \operatorname{Cov}(A_{b,t}, \phi(A_{q,t})) < 0.$$

Now consider general a near 2. Since for all $t \in \mathbb{Z}$

$$g_t(a) = -1 + (a+1)X_t, \qquad g_t(a)^2 = 1 + (a^2 - 1)X_t,$$

the stationary moments are

$$m_t^*(a) = -1 + (a+1)A_{b,t}, \qquad v_t^*(a) = 1 + (a^2-1)A_{q,t}. \tag{50}$$

Define

$$h_t^*(a) := \frac{m_t^*(a)}{\sqrt{v_t^*(a)} + \varepsilon}, \qquad \mu(a) := \mathbb{E}[h_t^*(a)].$$

The denominator is bounded away from zero and the integrand is bounded and continuous in a in a neighborhood of 2. Hence, by dominated convergence, $a \mapsto \mu(a)$ is continuous. Since $\mu(2) < 0$, there exists $\delta > 0$ such that, with $a = 2 + \delta$, we still have

$$\mu(a) < 0. \tag{51}$$

Decreasing δ if necessary, assume also $\delta \leq 1$.

Set $\gamma := -\mu(a) > 0$. Thus the stationary Adam direction has negative mean: $\mathbb{E}[h_t^*(a)] = -\gamma$. Equivalently, the mean update direction $-h_t^*(a)$ is positive.

We next transfer this stationary drift to the actual Adam process initialized at $m_0 = v_0 = 0$. The actual moments m_t satisfy

$$m_t = (1-b)\sum_{\ell=0}^{t-1} b^\ell g_{t-\ell}(a), \qquad \forall\, t \geq 1,$$

while the stationary version is

$$m_t^*(a) = (1-b)\sum_{\ell=0}^{\infty} b^\ell g_{t-\ell}(a), \qquad \forall\, t \in \mathbb{Z}.$$

Since the gradients g_t satisfy $|g_t| \le 3$ for all t,

$$|m_t - m_t^*(a)| \le 3b^t, \qquad |v_t - v_t^*(a)| \le 9q^t.$$

Also $v_t^*(a) \ge 1$ for all $t \in \mathbb{Z}$ and, for the actual process,

$$v_t \overset{(2)}{=} qv_{t-1} + (1-q)g_t^2 \overset{(2)}{\ge} 1-q > 0, \qquad \forall\, t \ge 1.$$

Therefore $h_t := \frac{m_t}{\sqrt{v_t}+\varepsilon}$ satisfies

$$|h_t - h_t^*(a)| \le C\rho^t, \qquad \forall\, t \ge 1, \tag{52}$$

for some $C < \infty$ and $\rho \in (0,1)$.

Step 2. We now prove the following tail excursion estimate. This technical Lemma has no determinstic analogue, i.e. it was not needed in the proof of Theorem 5.1.

Lemma A.2 (Weighted positive-drift estimate)**.** *Let $(X_t)_{t\in\mathbb{Z}}$ be i.i.d. random variables, and let $(Z_t)_{t\in\mathbb{Z}}$ be a bounded stationary Bernoulli shift of the form*

$$Z_t := F(X_t, X_{t-1}, X_{t-2}, \ldots), \qquad \forall\, t \in \mathbb{Z},$$

where F is real-valued. Assume that there are constants $C_0 < \infty$ and $\rho \in (0,1)$ such that, for every $\ell \ge 0$, changing only the coordinate $X_{t-\ell}$ can change Z_t by at most $C_0\rho^\ell$. Suppose

$$\bar{z} := \mathbb{E}Z_t > 0. \tag{53}$$

Let $\alpha > 0$. Let $\alpha_t := \alpha/\sqrt{t}$ for all $t \ge 1$. Then as $T \to \infty$,

$$\sum_{t=1}^{T} \alpha_t Z_t \to +\infty \qquad \textit{almost surely,}$$

and the uniform tail adverse excursion satisfies: as $n \to \infty$,

$$\Delta_n := \sup_{n \le r \le s < \infty} \max\Big(-\sum_{t=r}^{s} \alpha_t Z_t,\, 0\Big) \longrightarrow 0 \qquad \textit{almost surely.} \tag{54}$$

Proof. The weighted strong law follows from the ergodic theorem by Abel summation. Let

$$S_n := \sum_{t=1}^{n} Z_t, \qquad c_n := \frac{S_n}{n}.$$

By the ergodic theorem, $c_n \to \bar{z}$ almost surely as $n \to \infty$. Write $w_t = t^{-1/2}$ for all $t \ge 1$ and $W_T = \sum_{t=1}^{T} w_t$ for all $T \ge 1$. Summation by parts gives

$$\sum_{t=1}^{T} w_t Z_t = w_T S_T + \sum_{t=1}^{T-1} (w_t - w_{t+1}) S_t.$$

Since $S_t = tc_t$,

$$\frac{\sum_{t=1}^{T} w_t Z_t}{W_T} = \frac{T w_T}{W_T} c_T + \sum_{t=1}^{T-1} \frac{t(w_t - w_{t+1})}{W_T} c_t.$$

The coefficients on the right are nonnegative and sum to one since

$$Tw_T + \sum_{t=1}^{T-1} t(w_t - w_{t+1}) = \sum_{t=1}^{T} w_t = W_T.$$

So, the Toeplitz lemma implies that, as $T \to \infty$,

$$\frac{\sum_{t=1}^{T} w_t Z_t}{\sum_{t=1}^{T} w_t} \to \bar{z} \qquad \text{almost surely.} \tag{55}$$

Since $\sum_{t=1}^{T} \alpha_t \sim 2\alpha\sqrt{T}$, this implies as $T \to \infty$

$$\sum_{t=1}^{T} \alpha_t Z_t \to +\infty \qquad \text{almost surely.} \tag{56}$$

It remains to prove (54). Fix $\eta > 0$, and for $N \geq 1$ let

$$I_N := \{N, N+1, \ldots, 2N-1\}.$$

For an interval $[r, s] \subseteq I_N$, set

$$m := s - r + 1, \qquad Y_{r,s} := \sum_{t=r}^{s} \alpha_t Z_t. \tag{57}$$

We will apply McDiarmid's inequality to $Y_{r,s}$, regarded as a function of the independent coordinates $(X_j)_{j \leq s}$. By assumption, changing only X_j can change $Y_{r,s}$ by at most

$$d_j := C_0 \sum_{t=\max\{r,j\}}^{s} \alpha_t \rho^{t-j}, \qquad j \leq s,$$

where an empty sum is interpreted as zero. Since $\alpha_t \leq \alpha/\sqrt{N}$ for $t \in I_N$, we have

$$d_j \leq \frac{C_0 \alpha}{(1-\rho)\sqrt{N}}, \qquad \forall\, r \leq j \leq s,$$

$$d_j \leq \frac{C_0 \alpha}{(1-\rho)\sqrt{N}} \rho^{r-j}, \qquad \forall\, j < r.$$

It follows that

$$\sum_{j=-\infty}^{s} d_j^2 \leq \frac{C_1(m+1)}{N}$$

for a constant $C_1 < \infty$ depending only on C_0, ρ, α.

McDiarmid's inequality therefore gives

$$\mathbb{P}\big(Y_{r,s} - \mathbb{E}Y_{r,s} \leq -u\big) \leq \exp\left(-\frac{2u^2}{\sum_{j=-\infty}^{s} d_j^2}\right), \qquad \forall\, u > 0.$$

One may justify its use for the countable family $(X_j)_{j \leq s}$ by first fixing all coordinates before a finite time, applying the finite-dimensional inequality, and then sending that time to

$-\infty$. The assumed summable coordinate sensitivities make the corresponding approximation uniform. Since $t \le 2N$ on I_N and $[r,s] \subseteq I_N$,

$$\mathbb{E}Y_{r,s} \overset{(57)\wedge(53)}{=} \bar{z}\sum_{t=r}^{s}\alpha_t \overset{(57)}{\ge} \frac{\alpha\bar{z}}{\sqrt{2N}}\,m =: c_0\frac{m}{\sqrt{N}} \overset{(53)}{>} 0.$$

Consequently, using $-\eta := -u + \mathbb{E}Y_{r,s}$, for N sufficiently large,

$$\mathbb{P}(Y_{r,s} \le -\eta) \le \exp\Big(-c_1 N\big(\eta + c_0 m/\sqrt{N}\big)^2/(m+1)\Big) \le \exp(-c_\eta\sqrt{N}).$$

For the last inequality, use that $\eta > 0$ for N sufficiently large by definition of η,

$$\big(\eta + c_0 m/\sqrt{N}\big)^2 \ge 2\eta c_0 m/\sqrt{N},$$

and $m/(m+1) \ge 1/2$ by (57). Taking a union bound over the at most N^2 intervals with integer endpoints in I_N gives, for all N sufficiently large

$$\mathbb{P}\Big(\exists\, r,s \in I_N,\ r \le s\colon \sum_{t=r}^{s}\alpha_t Z_t \le -\eta\Big) \le N^2\exp(-c_\eta\sqrt{N}).$$

This bound is summable along the dyadic sequence $N = 2^j$. Therefore, by the Borel–Cantelli lemma, almost surely there exists $j_0(\eta)$ such that, for every $j \ge j_0(\eta)$, every subinterval of the dyadic block

$$B_j := \{2^j, \dots, 2^{j+1}-1\}$$

has weighted sum greater than $-\eta$.

We also claim that every sufficiently late dyadic block has positive weighted sum. Define

$$A_T := \sum_{t=1}^{T}\alpha_t Z_t.$$

The weighted strong law (55) proven above gives

$$A_T = \bar{z}\sum_{t=1}^{T}\alpha_t + o(\sqrt{T}) \qquad \text{almost surely.}$$

Hence

$$A_{2^{j+1}-1} - A_{2^j-1} = \sum_{t=2^j}^{2^{j+1}-1}\alpha_t Z_t = \bar{z}\sum_{t=2^j}^{2^{j+1}-1}\alpha_t + o(2^{j/2}) > 0$$

for all sufficiently large j, since $\sum_{t=2^j}^{2^{j+1}-1}\alpha_t \asymp 2^{j/2}$.

Now consider an interval $[r,s]$ lying sufficiently far in the tail. If it is contained in one dyadic block, its weighted sum is greater than $-\eta$. Otherwise, decompose it into a terminal piece of its first dyadic block, a collection of complete dyadic blocks, and an initial piece of its final dyadic block. The complete dyadic blocks have nonnegative sum, and each of the two boundary pieces has sum greater than $-\eta$ (almost surely, for sufficiently large j). Thus

$$\sum_{t=r}^{s}\alpha_t Z_t > -2\eta$$

for every sufficiently late interval $[r,s]$.

Finally, apply the preceding argument simultaneously to the countable sequence $\eta = 1/k$, $k \geq 1$. On the resulting probability-one event, for every k we have

$$\Delta_n \leq \frac{2}{k}$$

for all sufficiently large n. Therefore as $n \to \infty$,

$$\Delta_n \longrightarrow 0 \qquad \text{almost surely.}$$

$\square$

Step 3. Applying Lemma A.2. Returning to the proof of Theorem A.1, we will apply Lemma A.2, so we verify its coordinate-sensitivity assumption. Let

$$Z_t^* := -h_t^*(a) = -\frac{m_t^*(a)}{\sqrt{v_t^*(a)} + \varepsilon}. \tag{58}$$

From (50) and (48), changing only $X_{t-\ell}$ changes the stationary moment $m_t^*(a)$ by at most

$$|\Delta m_t^*(a)| = (a+1)(1-b)b^\ell$$

and similarly $v_t^*(a)$ changes by at most

$$|\Delta v_t^*(a)| = (a^2-1)(1-q)q^\ell.$$

On the region $\{(m,v) \in \mathbb{R}^2 \colon |m| \leq a,\ 1 \leq v \leq a^2\}$, the function $F(m,v) := -\frac{m}{\sqrt{v}+\varepsilon}$ satisfies

$$\left|\frac{\partial F}{\partial m}\right| \leq 1, \qquad \left|\frac{\partial F}{\partial v}\right| = \frac{|m|}{2\sqrt{v}(\sqrt{v}+\varepsilon)^2} \leq \frac{a}{2}.$$

It then follows from (58) that the change in Z_t^* from changing $X_{t-\ell}$ is at most

$$|\Delta Z_t^*| \leq (a+1)(1-b)b^\ell + \frac{a}{2}(a^2-1)(1-q)q^\ell.$$

Choose any $\rho_0 \in \big(\max(b,q), 1\big)$. Then there is a constant $C_0 < \infty$ such that

$$|\Delta Z_t^*| \leq C_0 \rho_0^\ell, \qquad \ell \geq 0.$$

Moreover, Z_t^* is bounded, since for all $2 \leq a \leq 3$

$$|m_t^*(a)| \leq a, \qquad v_t^*(a) \geq 1.$$

Thus Z_t^* satisfies all the hypotheses of Lemma A.2.

We apply Lemma A.2 first to the stationary process Z_t^* from (58). It has positive mean by (51). The actual process $Z_t := -h_t$ differs from Z_t^* by an exponentially decaying error by (52): $|Z_t - Z_t^*| \leq C\rho^t$. Consequently,

$$\sum_{t=1}^{\infty} \alpha_t |Z_t - Z_t^*| < \infty, \qquad \lim_{n\to\infty} \sup_{n \leq r \leq s < \infty} \sum_{t=r}^{s} \alpha_t |Z_t - Z_t^*| = 0.$$

Thus the two conclusions of Lemma A.2 also hold for $Z_t = -h_t$.

Step 4. Proving $\lim_{t\to\infty} x_t = 1$. We now prove that $\lim_{t\to\infty} x_t = 1$. The update is

$$x_{t+1} = \Pi_{[-1,1]}(x_t + \alpha_t Z_t), \qquad \forall\, t \geq 1.$$

Suppose there exists m such that $x_t < 1$ for all $t > m$. Then for any $n > m$ we have

$$1 > x_{n+1} \geq x_m + \sum_{t=m}^{n} \alpha_t Z_t,$$

since projection on $[-1, 1]$ can only increase $x_t < 1$ when $t > m$. But this contradicts (56). We therefore conclude that $x_t = 1$ for infinitely many times t.

Fix n, and let $\tau \geq n$ be a time such that $x_\tau = 1$. For all $m \geq \tau$, let

$$r_m := \max\{r \in [\tau, m] \colon x_r = 1\}.$$

If $r_m = m$, then $1 - x_m = 0$. Otherwise, none of the iterates

$$x_{r_m+1}, \ldots, x_m$$

equals 1. Hence projection to 1 does not occur for the indices from r_m through $m-1$. Projection to -1 can only increase the iterate, so an induction gives

$$x_m \geq 1 + \sum_{t=r_m}^{m-1} \alpha_t Z_t.$$

Since $x_m \leq 1$, we have $-\sum_{t=r_m}^{m-1} \alpha_t Z_t \geq 0$, so for all $m \geq \tau \geq n$,

$$1 - x_m \leq -\sum_{t=r_m}^{m-1} \alpha_t Z_t = \max\Big(-\sum_{t=r_m}^{m-1} \alpha_t Z_t,\, 0\Big) \overset{(54)}{\leq} \Delta_n.$$

Since $\Delta_n \to 0$ almost surely by (54), for any $\eta > 0$ we may first choose n so large that $\Delta_n < \eta$, and then choose a hitting time $\tau \geq n$. The preceding bound gives

$$0 \leq 1 - x_m < \eta, \qquad \forall\, m \geq \tau.$$

Therefore

$$\lim_{t\to\infty} x_t = 1 \qquad \text{almost surely.}$$

Step 5. Regret bound. It remains to compute regret. By the strong law of large numbers and since $a = 2 + \delta$,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} g_t(a) = \mathbb{E}[g_t(a)] = \frac{a}{3} - \frac{2}{3} = \frac{\delta}{3} \qquad \text{almost surely.}$$

Let $G_T := \sum_{t=1}^{T} g_t(a)$. Then $G_T/T \to \delta/3 > 0$, and hence $G_T > 0$ for all sufficiently large T, almost surely. Therefore the best fixed comparator is eventually $x_T^* = -1$, and for all sufficiently large T, $R_T = \sum_{t=1}^{T} g_t(a) x_t + G_T$. Equivalently,

$$R_T = 2G_T + \sum_{t=1}^{T} g_t(a)(x_t - 1).$$

Since $x_t \to 1$ almost surely and the gradients g_t are uniformly bounded,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} g_t(a)(x_t - 1) = 0 \qquad \text{almost surely.}$$

Combining this with $\lim_{T\to\infty} G_T/T = \delta/3$, we obtain

$$\lim_{T\to\infty} \frac{R_T}{T} = \frac{2\delta}{3} \qquad \text{almost surely.}$$

□

Acknowledgement. ChatGPT 5.5 assisted in the preparation of this manuscript.

References

[AMM+20] Ahmet Alacaoglu, Yura Malitsky, Panayotis Mertikopoulos, and Volkan Cevher. *A new regret analysis for Adam-type algorithms*. International Conference on Machine Learning (2020), 119, pp. 202–210.

[AZK+24] Kwangjun Ahn, Zhiyu Zhang, Yunbum Kook, and Yan Dai. *Understanding Adam optimizer via online learning of updates: Adam is FTRL in disguise*. International Conference on Machine Learning (2024).

[BG20] André Belotto da Silva and Maxime Gazeau. *A general system of differential equations to model first order adaptive algorithms*. Journal of Machine Learning Research (2020), 21 (129), pp. 1–42.

[BMR+20] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. *Language models are few-shot learners*. Advances in Neural Information Processing Systems (2020) 159, pp. 1877–1901.

[BW19] Sebastian Bock and Martin Georg Weiß. *Non-convergence and limit cycles in the Adam optimizer*. International Conference on Artificial Neural Networks (2019), vol 11728.

[BZZ+26] Zhiwei Bai, Jiajie Zhao, Zhangchen Zhou, Zhi-Qin John Xu, and Yaoyu Zhang. *Towards understanding Adam convergence on highly degenerate polynomials*. International Conference on Machine Learning (2026), to appear.

[D24] DeepSeek-AI. *DeepSeek-V3 Technical Report*. (2024), Preprint, arXiv:2412.19437.

[DCK+19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. North American Chapter of the Association for Computational Linguistics (2019), pp. 4171–4186.

[DDJ+25] Steffen Dereich, Thang Do, Arnulf Jentzen, and Philippe von Wurstemberger. *Adam symmetry theorem: characterization of the convergence of the stochastic Adam optimizer*. (2025), Preprint, arXiv:2511.06675.

[DGJ24] Steffen Dereich, Robin Graeber, and Arnulf Jentzen. *Non-convergence of Adam and other adaptive stochastic gradient descent optimization methods for non-vanishing learning rates*. (2024), Preprint, arXiv:2407.08100.

[DHJ24] Thang Do, Sonja Hannibal, and Arnulf Jentzen. *Non-convergence to global minimizers in data driven supervised deep learning: Adam and stochastic gradient descent optimization provably fail to converge to global minimizers in the training of deep neural networks with ReLU activation*. Journal of Mathematical Analysis and Applications (2026), 130724.

[DJR25] Thang Do, Arnulf Jentzen, and Adrian Riekert. *Non-convergence to the optimal risk for Adam and stochastic gradient descent optimization in the training of deep neural networks*. (2025), Preprint, arXiv:2503.01660.

[HWD19] Haiwen Huang, Chang Wang, and Bin Dong. *Nostalgic Adam: Weighting More of the Past Gradients When Designing the Adaptive Learning Rate*. International Joint Conference on Artificial Intelligence (2019), pp. 2556–2562.

[JR25] Arnulf Jentzen and Adrian Riekert. *Non-convergence to global minimizers for Adam and stochastic gradient descent optimization and constructions of local minimizers in the training of artificial neural networks*. SIAM/ASA Journal on Uncertainty Quantification (2025), 13 (3), pp. 1294–1333.

[KB15] Diederik P. Kingma and Jimmy Ba. *Adam: A method for stochastic optimization*. International Conference on Learning Representations (2015). (Poster)

[LH19] Ilya Loshchilov and Frank Hutter. *Decoupled Weight Decay Regularization*. International Conference on Learning Representations (2019) (Poster).

[LXL+19] Liangchen Luo, Yuanhao Xiong, Yan Liu, and Xu Sun. *Adaptive gradient methods with dynamic bound of learning rate*. International Conference on Learning Representations (2019).

[PKC+26] Tetiana Parshakova, Ahmed Khaled, Michael Crawshaw, Guillaume Garrigos, and Robert M. Gower. *Muon Does Not Converge on Convex Lipschitz Functions*. (2026), Preprint, arXiv:2605.08980.

[RKK18] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. *On the convergence of Adam and beyond*. International Conference on Learning Representations (2018).

[Toi23] Philippe L. Toint. *Divergence of the ADAM algorithm with fixed-stepsize: a (very) simple example*. (2023), Preprint, arXiv:2308.00720.

[TMS+23] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. (2023), Preprint, arXiv:2307.09288.

[WK22] Ruiqi Wang and Diego Klabjan. *Divergence results and convergence of a variance reduced version of Adam*. (2022), Preprint, arXiv:2210.05607.

[XZL+24] Xingyu Xie, Pan Zhou, Huan Li, Zhouchen Lin, and Shuicheng Yan. *Adan: Adaptive Nesterov Momentum Algorithm for Faster Optimizing Deep Models*. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024), 46 (12), pp. 9508–9520.

[ZCS+22] Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, and Zhi-Quan Luo. *Adam can converge without any modification on update rules*. Advances in Neural Information Processing Systems (2022).

[ZLC+26] Yushun Zhang, Bingran Li, Congliang Chen, Zhi-Quan Luo, and Ruoyu Sun. *Adam converges without any modification on update rules*. (2026), Preprint, arXiv:2603.02092.

[ZRS+18] Manzil Zaheer, Sashank J. Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. *Adaptive methods for nonconvex optimization*. Advances in Neural Information Processing Systems (2018).