Proceedings of Machine Learning Research vol 313:1–16, 2026 37th International Conference on Algorithmic Learning Theory

How to Set β_1, β_2 in Adam: An Online Learning Perspective

Quan Nguyen MANHQUAN233@GMAIL.COM
University of Victoria, Canada

Editors: Matus Telgarsky and Jonathan Ullman

Abstract

While Adam is one of the most effective optimizer for training large-scale machine learning models, a theoretical understanding of how to optimally set its momentum factors, β_1 and β_2, remains largely incomplete. Prior works have shown that Adam can be seen as an instance of Follow-the-Regularized-Leader (FTRL), one of the most important class of algorithms in online learning. The prior analyses in these works required setting $\beta_1 = \sqrt{\beta_2}$, which does not cover the more practical cases with $\beta_1 \neq \sqrt{\beta_2}$. We derive novel, more general analyses that hold for both $\beta_1 \geq \sqrt{\beta_2}$ and $\beta_1 \leq \sqrt{\beta_2}$. In both cases, our results strictly generalize the existing bounds. Furthermore, we show that our bounds are tight in the worst case. We also prove that setting $\beta_1 = \sqrt{\beta_2}$ is optimal for an oblivious adversary, but sub-optimal for an non-oblivious adversary.

Keywords: Adam optimizer, hyperparameter tuning, discounted regret, online-to-nonconvex.

1. Introduction

Training a neural network is an instance of a nonsmooth nonconvex optimization problem, where the goal is to find a w^* that minimizes a function $F : \mathbb{W} \to \mathbb{R}$, where $\mathbb{W}$ is the set of possible solutions. A training algorithm usually starts from some initial solution w_0, and then iteratively update the solution $w_{t+1} = w_t + \Delta_t$. Here, Δ_t specifies the update in round t. The most efficient training algorithms are first-order methods, which computes Δ_t based on the (possibly stochastic) gradients g_t, where $\mathbb{E}[g_t] = \frac{dF}{dw_t}$. The Adam optimizer (Kingma and Ba, 2017) is among the most popular first-order methods, which computes a coordinate-wise update of the form

$$\Delta_t = -\alpha_t \frac{\sum_{s=0}^{t-1} \beta_1^{t-1-s} g_s}{\sqrt{\sum_{s=0}^{t-1} \beta_2^{t-1-s} g_s^2}}, \tag{1}$$

where $\alpha_t > 0$ is the learning rate of the optimization problem, and $0 < \beta_1, \beta_2 < 1$ are the first and second-order momentum discount factors. As justified by Ahn and Cutkosky (2025), we omit the bias-correction terms since they are coordinate-independent and can be absorbed into α_t.

Ever since Adam was introduced, numerous experimental results (see e.g. Orvieto and Gower, 2025) have shown that tuning β_1 and β_2 play a vital role in the empirical performance of Adam. Despite its practical importance, a theoretical foundation for tuning these discount factors is still lacking, and current practitioners mostly rely on either expensive grid search to tune these factors. Recent empirical findings (Orvieto and Gower, 2025) found that for each β_1, the optimal value of β_2 often satisfies $\beta_1 \leq \sqrt{\beta_2}$, and that $\beta_1 = \sqrt{\beta_2}$ is in fact optimal in a number of experiments. Figure 1 in the appendix illustrates an example.

A recently emerged framework for nonconvex optimization is the online-to-nonconvex framework (Cutkosky et al., 2023), which casts a new perspective on Δ_t. In this framework, Δ_t is the prediction in round t of an online learning algorithm in a 1-dimensional online linear optimization

© 2026 Q. Nguyen.

problem. In this framework, Ahn et al. (2024) recently showed that the Adam's update rule in Equation (1) corresponding to the output of an Follow-the-Reguralized-Leader (FTRL) algorithm (see e.g. Orabona, 2023) with the sequence of losses $\ell_t(x) = \beta_1^{-t} g_t x$ and the square function $\frac{x^2}{2}$ as the regularizer. More concretely, Equation (1) is equivalent to

$$\Delta_t = \arg\min_{x \in \mathbb{R}} \frac{1}{2\eta_t} x^2 + \sum_{s=0}^{t-1} \beta_1^{-s} g_s x = -\eta_t \sum_{s=0}^{t-1} \beta_1^{-s} g_s, \tag{2}$$

where $\eta_t = \alpha_t \frac{(\beta_1/\sqrt{\beta_2})^{t-1}}{\sqrt{\sum_{s=0}^{t-1} \beta_2^{-s} g_s^2}}$ is the learning rate of the online learning problem. The full procedure is given in Algorithm 1.

Ahn et al. (2024) considered a special version of Adam with $\beta_1 = \sqrt{\beta_2}$, and proved an upper bound on its β_1-discounted regret R_{T,β_1}. Formally, let $D > 0$ and $\mathcal{X} = [-D, D]$ be a set that contains all the possible values of Δ_t. Let $u \in \mathcal{X}$. The β_1-discounted regret with respect to u after T rounds is

$$R_{T,\beta_1}(u) = \beta_1^T \sum_{t=1}^{T} \beta_1^{-t} g_t (\Delta_t - u). \tag{3}$$

By setting $\beta_2 = \beta_1^2$, Ahn et al. (2024) and Ahn and Cutkosky (2025) showed that Adam obtains $R_{T,\beta_1}(u) \leq O(D\sqrt{\sum_{t=0}^{T} \beta_1^{T-t} g_t^2})$ for all $u \in \mathcal{X}$, which then led to optimal convergence rate on nonsmooth nonconvex optimization problems. While these results are significant, they only hold for the case $\beta_1 = \sqrt{\beta_2}$ and do not hold for more general settings. In fact, both of these works state that the analysis for the general cases that include $\beta_1 \neq \sqrt{\beta_2}$ is an important open problem. Furthermore, existing results do not explain why $\beta_1 = \sqrt{\beta_2}$ is a good choice, and how optimal or sub-optimal this choice is.

Contributions. This work takes a step towards addressing the aforementioned open problem posed by Ahn et al. (2024) and Ahn and Cutkosky (2025). More specifically, we focus on deriving regret upper bounds on $R_{T,\beta_1}(u)$ in Equation (3) for Adam with $\beta_1 \neq \sqrt{\beta_2}$. Our main contributions are as follows.

- In Section 2, under the condition $\beta_1 \leq \sqrt{\beta_2}$, Corollary 2 shows a new β_1-discounted regret bound of order $O(D\frac{\sqrt{\beta_2}}{\beta_1}\sqrt{\sum_{t=0}^{T} \beta_2^{T-t} g_t^2} + D \max_t \left|\beta_1^{T-t} g_t\right|)$. This bound strictly generalizes existing bounds in Ahn et al. (2024) and Ahn and Cutkosky (2025). Our proof of this result is not based on clipping as in Ahn et al. (2024), but relies on a generalization of the analysis for the Scale-Free FTRL algorithm (Orabona and Pal, 2015). Furthermore, Theorem 5 proves that the new generalized bound is tight.

- In Section 3, we study Adam with $\beta_1 \geq \sqrt{\beta_2}$. Theorem 9 shows that an exponentially decaying sequence of α_t leads to an upper bound of order $O(D\sqrt{\sum_{t=0}^{T} \beta_1^{2T} \beta_2^{-t} g_t^2} + D \max_t \left|\beta_1^{T-t} g_t\right|)$ for the β_1-discounted regret. Both this bound and the analysis also recover the existing bounds and analyses in Ahn et al. (2024) and Ahn and Cutkosky (2025) when $\beta_1 = \sqrt{\beta_2}$.

- While the results in Sections 2 and 3 suggest that $\beta_1 = \sqrt{\beta_2}$ is optimal under an oblivious adversary, Theorem 11 in Section 4 shows an instance where $\beta_1 = \sqrt{\beta_2}$ is provably sub-optimal compared to $\beta_1 < \sqrt{\beta_2}$. Our construction uses an non-oblivious adversary, indicating

that optimally tuning β_1, β_2 strongly depends on the adversarial nature of the underlying environment, and that dynamically tuning the momentum factors might be beneficial.

Notations. For an integer T, we denote $[T] = \{1, 2, \ldots, T\}$. For two quantities f and g, we write $f \lesssim g, f \gtrsim g$ and $f \simeq g$ to denote $f = O(g), g = O(f)$ and $f = \Theta(g)$, respectively.

Throughout the paper, we write $p = \frac{\beta_1}{\sqrt{\beta_2}}, v_s = \beta_1^{-s} g_s$ and $D_T = \max_{t\in[T]} |\Delta_t|$. Note that $D_T \leq D$. The regularized objective function in round t of the FTRL algorithm is

$$F_t(x) = \frac{1}{2\eta_t} x^2 + \sum_{s=0}^{t-1} \beta_1^{-s} g_s x = \frac{1}{2\eta_t} x^2 + \sum_{s=0}^{t-1} v_s x. \tag{4}$$

Note that $\Delta_t = \arg\min_{x\in\mathcal{X}} F_t(x) = \mathrm{clip}_D(-\eta_t \sum_{s=0}^{t-1} v_s)$. Here, the clipping operation is $\mathrm{clip}(x) = x \min(\frac{D}{|x|}, 1)$. Also, $F_{t+1}(x) = F_t(x) - \frac{1}{2\eta_t} x^2 + v_t x + \frac{1}{2\eta_{t+1}} x^2$. The standard FTRL regret after T rounds is

$$R_T(u) := \sum_{t=1}^{T} v_t(\Delta_t - u) = \sum_{t=1}^{T} \beta_1^{-t} g_t(\Delta_t - u). \tag{5}$$

By definition, an upper bound on $R_T(u)$ implies an upper bound on the β_1-discounted regret $R_{T,\beta_1}(u)$, since $R_{T,\beta_1}(u) = \beta_1^T R_T(u)$. We can observe that $R_T(u)$ is the standard non-discounted regret of an online learning problem that predicts Δ_t and observes the loss v_t in round t. This learning procedure with non-discounted losses is in Algorithm 2, which is essentially Algorithm 1 re-written in standard online learning notations.

1.1. Related Works

Due to the vast amount of related literature on Adam and Adam-like methods, we focus on the most relevant works on the online-to-nonconvex framework and on tuning β_1, β_2. A more comprehensive review of Adam on a diverse range of topics such as implicit bias, separation from gradient descent, the SignDescent hypothesis and empirical performance, can be found in Orvieto and Gower (2025); Jin et al. (2025); Vasudeva et al. (2025) and references therein.

Online-To-Nonconvex Optimization. Viewing Δ_t as the output of an online learning algorithm was first considered by Cutkosky et al. (2023), who showed that a small shifting regret of an online learner can be converted to an optimal convergence rate for finding a stationary points. Instead of shifting regret, subsequent works (Zhang and Cutkosky, 2024; Ahn et al., 2025; Ahn and Cutkosky, 2025) considered β-discounted regret and showed that this notion of regret can also be used to bound the convergence rate of an optimization algorithm. By using online mirror descent as the online learner, Zhang and Cutkosky (2024); Ahn et al. (2025) obtained variants of gradient descent with momentum with optimal guarantee for nonconvex nonsmooth optimization. Ahn et al. (2024); Ahn and Cutkosky (2025) used FTRL as the online learner to derive Adam with $\beta_1 = \sqrt{\beta_2}$, and then proved its optimal convergence rates for nonconvex optimization. These works also provided theoretical justifications for why β_1 and β_2 should be set close, but not exactly equal, to 1.

Setting momentum factors β_1 and β_2. Adam was original proposed by Kingma and Ba (2017), who recommended several settings of the momentum factors such as $1 - \beta_1 = \sqrt{1 - \beta_2}$, $\beta_1 < \sqrt[4]{\beta_2}$ and $\beta_1 = 0.9, \beta_2 = 0.999$. The convergence issues of Adam and its variants with $\beta_1 < \sqrt{\beta_2}$ are studied in Reddi et al. (2018); Zhang et al. (2022); Alacaoglu et al. (2020). Recently, Taniguchi

Algorithm 1 Adam as Follow-the-Regularized-Leader on discounted losses

Input: $\beta_1, \beta_2 \in (0,1), (\alpha_t)_t > 0$, decision space $\mathcal{X} \subseteq \mathbb{R}$
Receive $g_0 \neq 0$
for $t = 1, \dots, T$ **do**
  Compute $\eta_t = \alpha_t \frac{(\beta_1/\sqrt{\beta_2})^{t-1}}{\sqrt{\sum_{s=0}^{t-1} \beta_2^{-s} g_s^2}}$
  Compute $\Delta_t = \arg\min_{x \in \mathcal{X}} \frac{1}{2\eta_t} x^2 + \sum_{s=0}^{t-1} \beta_1^{-s} g_s x$
  Receive $g_t \in \mathbb{R}$ and incur loss $\ell_t = \beta_1^{-t} g_t \Delta_t$
end

Algorithm 2 Follow-the-Regularized-Leader on sequence of losses $(v_t)_t$

Input: $p \in (0,1), (\alpha_t)_t > 0$ where $\alpha_{t+1} \leq \alpha_t$, decision space $\mathcal{X} \subseteq \mathbb{R}$
Receive $v_0 \neq 0$
for $t = 1, \dots, T$ **do**
  Compute $\eta_t = \alpha_t \frac{p^{t-1}}{\sqrt{\sum_{s=0}^{t-1} (p^s v_s)^2}}$
  Compute $\Delta_t = \arg\min_{x \in \mathcal{X}} \frac{1}{2\eta_t} x^2 + \sum_{s=0}^{t-1} v_s x$
  Receive $v_t \in \mathbb{R}$ and incur loss $\ell_t = v_t \Delta_t$
end

et al. (2024) showed that a modified version of Adam can converge for any β_2. Orvieto and Gower (2025) conduced some of the most comprehensive empirical study on the effects of tuning β_1, β_2 and suggested that optimal tuning would require a strong correlation between β_1 and β_2.

2. Discounted Regret Bound of Adam with $\beta_1 \leq \sqrt{\beta_2}$

In this section, we derive a bound on the β_1-discounted regret $R_{T,\beta_1}(u)$ of Adam with for $\beta_1 \leq \sqrt{\beta_2}$ and constant α. First, in Section 2.1, we show an general analysis that hold for both bounded $\mathcal{X} = [-D, D]$ and unbounded $\mathcal{X} = \mathbb{R}$ domain, as well as both oblivious and non-oblivious adversary. Then, in Section 2.3, we show that our new analysis and regret bound cannot be improved further.

2.1. A More General Regret Analysis for both Bounded and Unbounded Domain

Recall that $p = \frac{\beta_1}{\sqrt{\beta_2}}$ and $D_T = \max_{t \in [T]} |\Delta_t|$. We have $p \leq 1$ throughout this section. The following theorem states the regret bound of Algorithm 1 with $\beta_1 \leq \sqrt{\beta_2}$.

Theorem 1 *For any $T \geq 2, \beta_1 \leq \sqrt{\beta_2}$, any sequence $(\alpha_t)_t$ where $\alpha_{t+1} \leq \alpha_t$ and any sequence of $(g_t)_{t=0,\dots,T}$, Algorithm 1 guarantees*

$$R_T(u) \leq \frac{u^2}{\alpha_{T+1}\beta_1^T} \sqrt{\sum_{t=0}^{T} \beta_2^{T-t} g_t^2} + \frac{\sqrt{6\beta_2}}{2\beta_1} \left(\max_{t \in [T]} \frac{\alpha_t \sqrt{\beta_2^t}}{\beta_1^t} \right) \sqrt{\sum_{t=0}^{T} \beta_2^{-t} g_t^2} + 7 D_T \max_{0 \leq t \leq T} \left| \beta_1^{-t} g_t \right|. \tag{6}$$

Before showing the proof of Theorem 1, we discuss the generality of the bound in Equation (6) in comparison to existing bound in Ahn et al. (2024); Ahn and Cutkosky (2025). Letting $(\alpha_t)_t = \alpha$ be constant and noting that $\max_{t\in[T]} \frac{\sqrt{\beta_2^t}}{\beta_1^t} = \max_{t\in[T]} \frac{1}{p^t} = \frac{1}{p^T}$ for $p \le 1$, we obtain the following result.

Corollary 2 *Let $\alpha > 0$ be a constant. For any $T \ge 2$, $\beta_1 \le \sqrt{\beta_2}$ and any sequence of $(g_t)_{t=0,\dots,T}$, Algorithm 1 with $\alpha_t = \alpha$ guarantees*

$$R_T(u) \le \left(\frac{u^2}{\alpha} + \frac{\alpha\sqrt{6\beta_2}}{2\beta_1}\right)\frac{1}{\beta_1^T}\sqrt{\sum_{t=0}^{T}\beta_2^{T-t}g_t^2} + 7D_T \max_{0\le t\le T}\left|\beta_1^{-t}g_t\right|. \tag{7}$$

Remark 3 *Multiplying β_1^T on both sides of Equation (7) results in the following bound for the β_1-discounted regret:*

$$R_{T,\beta_1}(u) = \beta_1^T R_T(u) \le \left(\frac{u^2}{\alpha} + \frac{\alpha\sqrt{6\beta_2}}{2\beta_1}\right)\sqrt{\sum_{t=0}^{T}\beta_2^{T-t}g_t^2} + 7D_T \max_{0\le t\le T}\left|\beta_1^{T-t}g_t\right|. \tag{8}$$

Ignoring the (small) constant factors, setting $\beta_2 = \beta_1^2$ leads to the bound in Ahn et al. (2024, Theorem B.2). In addition, on the bounded domain $\mathcal{X} = [-D, D]$, setting $\beta_2 = \beta_1^2$, $|u| = D$ and $\alpha = \Theta(D)$ recovers the $O(D\sqrt{\sum_{t=1}^{T}\beta_1^{T-t}g_t^2})$ bound in Ahn and Cutkosky (2025, Theorem 9).

Remark 4 *Under an oblivious adversary, i.e. when the sequence $(g_t)_t$ is fixed regardless of the choice of β_2, choosing $\beta_2 = \beta_1^2$ is optimal for the regret bound in Equation (7). This follows from the fact that $T - t \ge 0$ and thus, $\beta_2^{T-t} \ge \beta_1^{2(T-t)}$ for all $t \in [T]$.*

2.2. Proof of Theorem 1

As mentioned above, Algorithm 2 is exactly Algorithm 1 written in standard online learning notations. Hence, we will focus on analyzing the regret $R_T(u) = \sum_{t=1}^{T} v_t(\Delta_t - u)$ of the FTRL with time-varying learning rate in Algorithm 2. Our proof is essentially a more general version of the proof for the Scale-Free FTRL algorithm in Orabona and Pal (2015). We first give a proof sketch that highlights the main steps of the analysis, and then give the full proof.

Proof Sketch. Our first step is ensure that the sequence of learning rates $(\eta_t)_t$ is non-increasing, which indeed follows from the fact that $p \le 1$ and $\eta_{t+1} \le \eta_t$. Having a non-increasing sequence of learning rates, we can employ the standard FTRL analysis (Orabona, 2023) and obtain

$$R_T(u) \le \frac{u^2\sqrt{\sum_{t=1}^{T}\beta_2^{-t}g_t^2}}{\alpha_{T+1}p^T} + \sum_{t=1}^{T}\min\left\{\underbrace{\frac{\alpha_t}{2}\frac{p^{t-1}v_t^2}{\sqrt{\sum_{s=0}^{t-1}(p^s v_s)^2}}}_{(a)}, \underbrace{2D_T|v_t|}_{(b)}\right\}.$$

Observe that if $p = 1$, then (a) would recover the well-known quantity $\sum_{t=1}^{T}\frac{v_t^2}{\sqrt{\sum_{s=0}^{t-1}v_s^2}}$ that frequently appears in the analysis of FTRL with adaptive learning rates (e.g. Ito et al., 2024)

and Scale-Free FTRL in particular (Orabona and Pal, 2015; Ahn et al., 2024). Existing works proceed by considering two different cases on the magnitude of $|v_t|$, thereby creating two different telescopic sums. If $|v_t|$ is large, then the first term (a) can be very large and we resort to bounding $(b) \lesssim O(D_T \max_{s\le t}|v_s| - \max_{s\le t-1}|v_s|)$. On the other hand, if $|v_t|$ is small then we can bound $(a) \le O(\sqrt{\sum_{s\le t} v_s^2} - \sqrt{\sum_{s\le t-1} v_s^2})$.

In our more general case of $p \le 1$, we need a more fine-grained analysis. Instead of $|v_t|$, we examine $p^{t-1}|v_t|$. In the first case where $p^{t-1}|v_t|$ is large, it turns out that we still have $(b) \lesssim O(D_T \max_{s\le t}|v_s| - \max_{s\le t-1}|v_s|)$ for $p \le 1$. On the contrary, when $p^{t-1}|v_t|$ is small then $(p^t v_t)^2$ is small, which allows us to bound $(a) \le O(\sqrt{\sum_{s\le t}(p^s v_s)^2} - \sqrt{\sum_{s\le t-1}(p^s v_s)^2})$.

Proof of Theorem 1 The standard FTRL analysis (e.g. Orabona, 2023, Lemma 7.1) shows that

$$R_T(u) = \frac{u^2}{\eta_{T+1}} + F_{T+1}(\Delta_{T+1}) - F_{T+1}(u) + \sum_{t=1}^{T} F_t(\Delta_t) - F_{t+1}(\Delta_{t+1}) + v_t\Delta_t \tag{9}$$

$$\le \frac{u^2}{\eta_{T+1}} + \sum_{t=1}^{T} F_t(\Delta_t) - F_{t+1}(\Delta_{t+1}) + v_t\Delta_t, \tag{10}$$

where we used $F_{T+1}(\Delta_{T+1}) - F_{T+1}(u) \le 0$ in the inequality. From here, we will bound $F_t(\Delta_t) - F_{t+1}(\Delta_{t+1}) + v_t\Delta_t$ by two different ways and then take their minimum. First, we have

$$\begin{aligned}
&F_t(\Delta_t) - F_{t+1}(\Delta_{t+1}) + v_t\Delta_t \\
&= F_t(\Delta_t) - (F_t(\Delta_{t+1}) - \frac{1}{2\eta_t}\Delta_{t+1}^2 + v_t\Delta_{t+1} + \frac{1}{2\eta_{t+1}}\Delta_{t+1}^2) + v_t\Delta_t \\
&= F_t(\Delta_t) - F_t(\Delta_{t+1}) + (\frac{1}{\eta_t} - \frac{1}{\eta_{t+1}})\Delta_{t+1}^2 + v_t(\Delta_t - \Delta_{t+1}) \\
&\le v_t(\Delta_t - \Delta_{t+1}) \le 2D_T|v_t|,
\end{aligned} \tag{11}$$

where the first inequality is due to the following facts:

- For $p \le 1$ and $\alpha_{t+1} \le \alpha_t$, we have $\eta_{t+1} = \alpha_{t+1}\frac{p^t}{\sqrt{\sum_{s=0}^{t}\beta_2^{-s}g_s^2}} \le \alpha_t\frac{p^{t-1}}{\sqrt{\sum_{s=0}^{t-1}\beta_2^{-s}g_s^2}} = \eta_t$. In other words, the sequence of learning rates $(\eta_t)_t$ is non-increasing.
- $F_t(\Delta_t) \le F_t(\Delta_{t+1})$ by the definition of Δ_t.

Next, using either a local-norm analysis as in Orabona (2023) or the strong convexity of F_t, we obtain

$$F_t(\Delta_t) - F_t(\Delta_{t+1}) + v_t(\Delta_t - \Delta_{t+1}) \le \frac{\eta_t v_t^2}{2}. \tag{12}$$

Combining Equation (11) and Equation (12) leads to

$$F_t(\Delta_t) - F_{t+1}(\Delta_{t+1}) + v_t\Delta_t \le \min\left\{\frac{\eta_t v_t^2}{2}, 2D_T|v_t|\right\}.$$

Plugging this into Equation (10) and expanding the definition of η_t, we obtain

$$R_T(u) \leq \frac{u^2}{\eta_{T+1}} + \sum_{t=1}^{T} \min\left\{\frac{\eta_t v_t^2}{2}, 2D_T|v_t|\right\} = \frac{u^2\sqrt{\sum_{t=1}^{T}\beta_2^{-t}g_t^2}}{\alpha_{T+1}p^T} + \sum_{t=1}^{T}\min\left\{\frac{\eta_t v_t^2}{2}, 2D_T|v_t|\right\}$$
$$= \frac{u^2\sqrt{\sum_{t=0}^{T}\beta_2^{-t}g_t^2}}{\alpha_{T+1}p^T} + \sum_{t=1}^{T}\min\left\{\frac{\alpha_t}{2}\frac{p^{t-1}v_t^2}{\sqrt{\sum_{s=0}^{t-1}(p^s v_s)^2}}, 2D_T|v_t|\right\}. \tag{13}$$

We consider two cases. First, if $p^{t-1}|v_t| \geq \sqrt{2}\sqrt{\sum_{s=0}^{t-1}(p^s v_s)^2}$, then

$$2D_T|v_t| = 2D_T\frac{\sqrt{2}|v_t| - |v_t|}{\sqrt{2}-1} \leq 2D_T\frac{\sqrt{2}(|v_t| - \frac{\sqrt{\sum_{s=1}^{t-1}(p^s v_s)^2}}{p^{t-1}})}{\sqrt{2}-1} \tag{14}$$
$$\leq 7D_T(|v_t| - \sqrt{\sum_{s=0}^{t-1} v_s^2}) \leq 7D_T(\max_{s\leq t}|v_s| - \max_{s\leq t-1}|v_s|), \tag{15}$$

where the the second inequality is due to $\sum_{s=1}^{t-1}(p^s v_s)^2 \geq (p^{t-1})^2\sum_{s=1}^{t-1} v_s^2$ for $p \leq 1$.

On the other hand, if $p^{t-1}|v_t| < \sqrt{2}\sqrt{\sum_{s=0}^{t-1}(p^s v_s)^2}$, then we have

$$\frac{p^{t-1}v_t^2}{\sqrt{\sum_{s=0}^{t-1}(p^s v_s)^2}} = \frac{\sqrt{3}}{p^{t+1}}\frac{(p^t v_t)^2}{\sqrt{3\sum_{s=0}^{t-1}(p^s v_s)^2}} \leq \frac{\sqrt{3}}{p^{t+1}}\frac{(p^t v_t)^2}{\sqrt{(p^t v_t)^2 + \sum_{s=1}^{t-1}(p^s v_s)^2}} \tag{16}$$
$$\leq \frac{\sqrt{6}}{p^{t+1}}\left(\sqrt{\sum_{s=0}^{t}(p^s v_s)^2} - \sqrt{\sum_{s=0}^{t-1}(p^s v_s)^2}\right), \tag{17}$$

where we used $p^t|v_t| \leq p^{t-1}|v_t|$ in the first inequality and $\frac{x^2}{\sqrt{x^2+y}} \leq \sqrt{2}(\sqrt{x^2+y} - \sqrt{y})$ for all $x \in \mathbb{R}, y \geq 0$ in the second inequality. Consequently,

$$\alpha_t\frac{p^{t-1}v_t^2}{\sqrt{\sum_{s=0}^{t-1}(p^s v_s)^2}} \leq \frac{\alpha_t\sqrt{6}}{p^{t+1}}\left(\sqrt{\sum_{s=0}^{t}(p^s v_s)^2} - \sqrt{\sum_{s=0}^{t-1}(p^s v_s)^2}\right) \tag{18}$$
$$\leq \frac{\sqrt{6}}{p}\left(\max_{t\in[T]}\frac{\alpha_t}{p^t}\right)\left(\sqrt{\sum_{s=0}^{t}(p^s v_s)^2} - \sqrt{\sum_{s=0}^{t-1}(p^s v_s)^2}\right) \tag{19}$$

Plugging Equations (15) and (19) into Equation (13) and summing over T rounds, we obtain

$$R_T(u) \leq \left(\frac{u^2}{\alpha_{T+1}p^T} + \frac{\sqrt{6}}{2p}\left(\max_{t\in[T]}\frac{\alpha_t}{p^t}\right)\right)\sqrt{\sum_{t=0}^{T}(p^t v_t)^2} + 7D_T\max_{0\leq t\leq T}|v_t|. \tag{20}$$

Using $p = \frac{\beta_1}{\sqrt{\beta_2}}$ and $v_t = \beta_1^{-t}g_t$ leads to the desired statement in Equation (6). ■

2.3. The generalized regret bound is tight

Consider the bounded domain $\mathcal{X} = [-D, D]$. We rewrite the order of the generalized regret bound in Corollary 2 as a function of p and $(v_t)_t$:

$$R_T(u) \lesssim B_{T,\alpha,p,(v_t)_t}(u) := \left(\frac{u^2}{\alpha} + \frac{\alpha}{p}\right)\frac{1}{p^T}\sqrt{\sum_{t=0}^{T}(p^t g_t)^2} + D\max_{0\le t\le T}|v_t|. \tag{21}$$

Recall that this bound holds for *any* choices of $0 < p \le 1, u \in \mathcal{X}, \alpha > 0$ and $(v_t)_t$. The following theorem states that this bound is tight and cannot be significantly improved further.

Theorem 5 *Let $D > 0$ and $0.4 \le p \le 0.6$ be arbitrary universal constants. On the bounded domain $\mathcal{X} = [-D, D]$, there exists a sequence $(v_t)_t$ where $v_t > 0$ for all $t \in [T]$, such that with $\alpha_t = \alpha = \frac{D}{4}$ and $u = -D$, the sequence of $(\Delta_t)_t$ produced by Algorithm 2 satisfies $R_T(u) = \Omega(B_{T,\alpha,p,(v_t)_t}(u))$.*

Proof Let $\kappa > 0$ be a constant. We select a constant $v_0 > 0$ arbitrarily and define a sequence $v_t = \kappa^t v_0$ for $t = 1, 2, \ldots, T$. Under this sequence $(v_t)_t$, the learning rates are

$$\eta_t = \alpha\frac{p^{t-1}}{\sqrt{\sum_{s=0}^{t-1}(p^s v_s)^2}} = \alpha\frac{p^{t-1}}{\sqrt{\sum_{s=0}^{t-1}(p^s\kappa^s v_0)^2}} = \alpha\frac{p^{t-1}}{v_0\sqrt{\sum_{s=0}^{t-1}(p^2\kappa^2)^s}} = \alpha\frac{p^{t-1}}{v_0}\sqrt{\frac{p^2\kappa^2-1}{(p^2\kappa^2)^t-1}}. \tag{22}$$

The pre-clipping update $\bar{\Delta}_t$ in each round is

$$\bar{\Delta}_t = -\eta_t\sum_{s=0}^{t-1}v_s = -\eta_t\sum_{s=0}^{t-1}\kappa^s v_0 = -v_0\eta_t\frac{\kappa^t-1}{\kappa-1} = -\alpha p^{t-1}\frac{\kappa^t-1}{\kappa-1}\sqrt{\frac{p^2\kappa^2-1}{(p^2\kappa^2)^t-1}}. \tag{23}$$

Note that this formula of $\bar{\Delta}_t$ holds for any $(\alpha_t)_t = \alpha$ and $\kappa > 0$.

Next, we further require $\kappa \ge \frac{1}{p^2}$. The following lemma shows that $\left|\bar{\Delta}_t\right| < D$, and thus the clipping operation is never used and can be safely ignored during our analysis.

Lemma 6 *For any $p \in (0, 0.6]$ and $\kappa \ge \frac{1}{p^2}$, we have $\left|\bar{\Delta}_t\right| \le \frac{D}{2}$ holds for all $t \in [T]$.*

Proof We have $\kappa > 1$ in this case. From Equation (23), we re-write the pre-clipping update as

$$\bar{\Delta}_t = -\alpha\frac{\sqrt{p^2\kappa^2-1}}{p(\kappa-1)}\frac{p^t(\kappa^t-1)}{\sqrt{(p^2\kappa^2)^t-1}}.$$

Our proof makes uses of two technical lemmas, Lemma 13 and Lemma 14. By Lemma 13, we have $\frac{p^t(\kappa^t-1)}{\sqrt{(p^2\kappa^2)^t-1}} \le 1$ for all $\kappa \ge \frac{1}{p^2}$. By Lemma 14, we have $\frac{\sqrt{p^2\kappa^2-1}}{p(\kappa-1)} \le 2$. Therefore, $\left|\bar{\Delta}_t\right| = \alpha\frac{\sqrt{p^2\kappa^2-1}}{p(\kappa-1)}\frac{p^t(\kappa^t-1)}{\sqrt{(p^2\kappa^2)^t-1}} \le 2\alpha = \frac{D}{2}$. ■

Lemma 6 implies that for every round t, we have $\Delta_t = \bar{\Delta}_t$. Moreover, $\Delta_t \geq -\frac{D}{2}$. We can then compute a lower bound for the regret $R_T(-D)$. For $u = -D$, the regret is lower bounded by

$$R_T(-D) = \sum_{t=1}^{T} v_t(\Delta_t + D) \geq \sum_{t=1}^{T} v_t(D - \frac{D}{2}) = \frac{v_0 D}{2} \sum_{t=1}^{T} \kappa^t = v_0 D \kappa \frac{\kappa^T - 1}{\kappa - 1}, \tag{24}$$

where the first inequality used $v_t > 0$ and $\Delta_t \geq \frac{-D}{2}$.

Finally, we compute (the order of) the upper bound $B_{T,\alpha,p,(v_t)_t}(-D)$. Note that $\max_{0 \leq tT} |v_t| = \max_{0 \leq tT} v_0 \kappa^t = v_0 \kappa^T$. With $p \geq 0.4$, Equation (21) implies that the regret bound is of order

$$\begin{aligned} B_{T,\alpha,p,(v_t)_t}(-D) &\simeq (\frac{u^2}{\alpha} + \frac{\alpha}{p}) \frac{1}{p^T} \sqrt{\sum_{t=0}^{T} (p^t v_t)^2} + D \max_{t \in [0,T]} |v_t| \\ &\simeq D v_0 \left(\kappa^T + \frac{1}{p^T} \sqrt{\sum_{t=0}^{T} (p^2 \kappa^2)^t} \right) = D v_0 \left(\kappa^T + \frac{1}{p^T} \sqrt{\frac{(p^2\kappa^2)^T - 1}{p^2 \kappa^2 - 1}} \right). \end{aligned} \tag{25}$$

By Equation (24), $R_T(-D) = \Omega(v_0 D(\kappa^T - 1))$. From the fact that $\kappa^T - 1 = \Omega(\kappa^T)$ and $\kappa^T = \Omega(\frac{1}{p^T}\sqrt{(p^2\kappa^2)^T - 1})$ for $T \geq 2$ and $\kappa > 1$, we conclude that $R_T(-D) = \Omega(B_{T,\alpha,p,(v_t)_t}(-D))$. ■

Remark 7 *An important point of clarification is that, our upper and lower bounds do* not *suggest $\beta_1 = \sqrt{\beta_2}$ is uniformly better than on every problem instance. Under an oblivious adversary, the optimality of setting $\beta_1 = \sqrt{\beta_2}$ comes from minimizing* the regret bound *in Remark 3, not from minimizing* the regret itself. *In other words, if the problem admits a loss sequence specified in our Theorem 5, then choosing $\beta_1 = \sqrt{\beta_2}$ is always going to be better than $\beta_1 < \sqrt{\beta_2}$. This, however, does not exclude other loss sequences (still from an oblivious adversary) where setting $\beta_1 < \sqrt{\beta_2}$ is actually better, because the regret upper bound may be loose in that case, and the actual regret might be smaller.*

Remark 8 *The worst-case lower bound in our Theorem 5 is algorithm-dependent. This leaves the question of whether there are other algorithms that may achieve better worst-case or problem-dependent regret bounds. For example, an algorithm whose regret depends on the total variation of the sequence of losses may significantly outperform Adam when the loss sequences (i.e. the gradients) are not changing much (i.e. in lazy-training regime where the parameters of a model do not deviate much from the initialization). We leave this as a future work.*

3. An Analysis for Adam with $\beta_1 \geq \sqrt{\beta_2}$

In this section, we extend the technique in the proof of Theorem 1 to derive a β_1-discounted regret bound for Adam with $\beta_1 \geq \sqrt{\beta_2}$. With a different choice of choice of $(\alpha_t)_t$, we will show a bound that, surprisingly, reduces to the bound in Ahn et al. (2024, Theorem B.1) when $\beta_1 = \sqrt{\beta_2}$. Furthermore, under an oblivious adversary, this bound also suggests that choosing $\beta_2 = \beta_1^2$ is optimal.

With $\beta_1 \geq \sqrt{\beta_2}$, we have $p \geq 1$. To ensure that the sequence of learning rates $(\eta_t)_t$ is non-increasing, we use $\alpha_t = \frac{\alpha}{p^{t-1}}$, where $\alpha > 0$ is a constant. In other words, α_t decay exponentially. The learning rate η_t in Algorithms 1 and 2 becomes

$$\eta_t = \alpha_t \frac{p^{t-1}}{\sqrt{\sum_{s=1}^{t-1} (p^s v_s)^2}} = \frac{\alpha}{\sqrt{\sum_{s=1}^{t-1} (p^s v_s)^2}}, \tag{26}$$

which is non-increasing. The pre-clipping update in round t is $\bar{\Delta}_t = -\eta \sum_{s=1}^{t-1} v_s = -\frac{\alpha \sum_{s=1}^{t-1} v_s}{\sum_{s=1}^{t-1} (p^s v_s)^2}$, which could be understood as RMSProp with momentum. The following theorem states the regret bound of this algorithm.

Theorem 9 *For any $\alpha > 0, T \geq 2$, Algorithm 1 with $\beta_1 \geq \sqrt{\beta_2}, \alpha_t = \frac{\alpha}{p^{t-1}}$ guarantees*

$$R_T(u) \leq \left(\frac{u^2}{\alpha} + \frac{\alpha\sqrt{6}}{2} \right) \sqrt{\sum_{t=1}^{T} \beta_2^{-t} g_t^2} + 7D_T \max_{0 \leq t \leq T} \left| \beta_1^{-t} g_t \right| \tag{27}$$

Before showing the proof of Theorem 9, we discuss its implication.

Remark 10 *Multiplying β_1^T on both sides of Equation (27) leads to*

$$R_{T,\beta_1}(u) = \beta_1^T R_T(u) \leq (\frac{u^2}{\alpha} + \frac{\alpha\sqrt{6}}{2}) \sqrt{\sum_{t=0}^{T} (\beta_1^2)^T \beta_2^{-t} g_t^2} + 7D_T \max_{t \in [0,T]} \left| \beta_1^{T-t} g_t \right|. \tag{28}$$

Under an oblivious adversary, the β_1-discounted regret bound in Equation (28) is decreasing in β_2. Therefore, setting $\beta_2 = \beta_1^2$ is optimal for an oblivious adversary. Furthermore, setting $\beta_2 = \beta_1^2$ recovers both Ahn et al. (2024, Theorem B.1) and Ahn and Cutkosky (2025, Theorem 9).

Proof of Theorem 9 Since the learning rates $(\eta_t)_t$ in Equation (26) is non-increasing, we can proceed as in the proof of Theorem 1 and obtain

$$R_T(u) \leq \frac{u^2}{\eta_{T+1}} + \sum_{t=1}^{T} \min \left\{ \frac{\eta_t v_t^2}{2}, 2D_T |v_t| \right\} \tag{29}$$

$$= \frac{u^2}{\alpha} \sqrt{\sum_{t=0}^{T} (p^t v_t)^2} + \sum_{t=1}^{T} \min \left\{ \frac{\alpha v_t^2}{2\sqrt{\sum_{s=0}^{t-1} (p^s v_s)^2}}, 2D_T |v_t| \right\}. \tag{30}$$

We consider two cases. First, if $|v_t| \geq \sqrt{2}\sqrt{\sum_{s=0}^{t-1} (p^s v_s)^2}$, then

$$2D_T |v_t| = 2D_T \frac{\sqrt{2}|v_t| - |v_t|}{\sqrt{2} - 1} \leq 2D_T \frac{\sqrt{2}}{\sqrt{2} - 1} \left(|v_t| - \sqrt{\sum_{s=0}^{t-1} (p^s v_s)^2} \right) \tag{31}$$

$$\leq 7D_T \left(|v_t| - \sqrt{\sum_{s=0}^{t-1} v_s^2} \right) \leq 7D_T (\max_{0 \leq s \leq t} |v_s| - \max_{0 \leq s \leq t-1} |v_s|), \tag{32}$$

where the the second inequality is due to $\sum_{s=1}^{t-1} (p^s v_s)^2 \geq \sum_{s=1}^{t-1} v_s^2$ for $p \geq 1$.

On the other hand, if $|v_t| < \sqrt{2}\sqrt{\sum_{s=0}^{t-1}(p^s v_s)^2}$, then we have

$$\frac{v_t^2}{\sqrt{\sum_{s=0}^{t-1}(p^s v_s)^2}} = \sqrt{3}\frac{v_t^2}{\sqrt{3\sum_{s=0}^{t-1}(p^s v_s)^2}} \leq \sqrt{3}\frac{v_t^2}{\sqrt{v_t^2 + \sum_{s=1}^{t-1}(p^s v_s)^2}} \tag{33}$$

$$\leq \sqrt{6}\left(\sqrt{v_t^2 + \sum_{s=0}^{t-1}(p^s v_s)^2} - \sqrt{\sum_{s=0}^{t-1}(p^s v_s)^2}\right) \tag{34}$$

$$\leq \sqrt{6}\left(\sqrt{\sum_{s=0}^{t}(p^s v_s)^2} - \sqrt{\sum_{s=0}^{t-1}(p^s v_s)^2}\right), \tag{35}$$

where we used $p^{2t}v_t^2 \geq v_t^2$ for $p \geq 1$ in the last inequality. Plugging Equations (32) and (35) into Equation (30) and summing over T rounds, we obtain, we obtain

$$R_T(u) \leq \left(\frac{u^2}{\alpha} + \frac{\alpha\sqrt{6}}{2}\right)\sqrt{\sum_{t=0}^{T}(p^t v_t)^2} + 7D_T \max_{t\in[0,T]}|v_t|.. \tag{36}$$

Using $p = \frac{\beta_1}{\sqrt{\beta_2}}$ and $v_t = \beta_1^{-t} g_t$ leads to the desired statement in Theorem 9. ■

4. When is setting $\beta_1 = \sqrt{\beta_2}$ not optimal? An Example with Non-Oblivious Adversary

When using Adam for optimizing a target function in the online-to-nonconvex objective $F(\cdot)$, for a fixed β_1, different values of β_2 inevitably leads to different sequence of updates $(\Delta_t)_t$, which in turn leads to different sequence of w_t. Consequently, the sequence of gradients $(g_t)_t$ also varies with different values of β_2. It follows that the adversary is non-oblivious, since g_{t+1} depends on the outputs $(\Delta_s)_{s=1,2,\ldots,t}$ of the algorithm in past t rounds.

Setup. While our results in the previous sections indicate that setting $\beta_1 = \sqrt{\beta_2}$ is optimal under an oblivious adversary, they do not have any implication for an non-oblivious adversary. In this section, we present a result showing that $\beta_1 = \sqrt{\beta_2}$ may no longer be optimal when the adversary is non-oblivious. To this end, we consider the bounded domain $\mathcal{X} = [-1, 1]$, i.e. $D = 1$. Let $a, b \in (0, 1)$ be two universal constants, $a \neq b$. Let $K = \max\{\frac{1}{1-a}, \frac{1}{1-b}\}$, and $\alpha = \frac{1}{K}$.

Fix $\beta_1 \in (0, 1)$. We will compare the regret of two instances of Algorithm 2, denoted by $\mathcal{A}$ and $\mathcal{A}'$. Algorithm $\mathcal{A}$ uses $\beta_1 = p\sqrt{\beta_2}$, where $p < 1$. Algorithm $\mathcal{A}'$ uses $\beta_1 = \sqrt{\beta_2}$. Both algorithms use $\alpha_t = \alpha = \frac{1}{K}$.

Non-oblivious Adversary. Fix an arbitrary $v > 0$ and define two sequences $v_t = a^t v$ and $v'_t = b^t v$ for $t = 0, 1, 2, \ldots$. Note that $v_0 = v'_0 = v$, which resembles practical scenarios where the very first gradients evaluated when the model has just been initialized. For algorithms $\mathcal{A}$ and $\mathcal{A}'$, the adversary will use the losses $(v_t)_t$ and $(v'_t)_t$, respectively.

Regret Analysis. Let Δ_t and Δ'_t be the updates in round t of $\mathcal{A}$ and $\mathcal{A}'$, respectively. Because all of losses $(v_t)_t$ and $(v'_t)_t$ are positive, the optimal updates for both $\mathcal{A}$ and $\mathcal{A}'$ is $u = -D$. The following theorem shows that the regret of $\mathcal{A}$ is strictly smaller than that of $\mathcal{A}'$, which indicates that setting $\beta_1 = \sqrt{\beta_2}$ is not optimal.

Theorem 11 *Let $u = -D$. For any $T \geq 2$, the sequence of updates from $\mathcal{A}$ and $\mathcal{A}'$ satisfy*

$$\sum_{t=1}^{T} v_t(\Delta_t - u) < \sum_{t=1}^{T} v'_t(\Delta'_t - u). \tag{37}$$

Proof First, we show that no clipping is required in both $\mathcal{A}$ and $\mathcal{A}'$. Since both algorithms use the same constant $\alpha = \frac{1}{K}$ for the sequence $(\alpha_t)_t$, we can apply Equation (23) twice, once for $\kappa = a$ and once for $\kappa = b$, to obtain

$$\bar{\Delta}_t = -\frac{p^{t-1}\sqrt{1-(pa)^2}}{K\sqrt{1-(pa)^{2t}}}\frac{1-a^t}{1-a}, \tag{38}$$

$$\bar{\Delta}'_t = -\frac{p^{t-1}\sqrt{1-(pb)^2}}{K\sqrt{1-(pb)^{2t}}}\frac{1-b^t}{1-b}, \tag{39}$$

where $\bar{\Delta}_t$ and $\bar{\Delta}'_t$ are the pre-clipping updates in round t of $\mathcal{A}$ and $\mathcal{A}'$, respectively. Since $D = 1$, it suffices to show that $\left|\bar{\Delta}_t\right| \leq 1$ and $\left|\bar{\Delta}'_t\right| \leq 1$. We have $\left|\bar{\Delta}_t\right| \leq 1$ follow from $K(1-a) \geq 1$ and

$$p^{t-1}(1-a^t)\sqrt{1-(pa)^2} \leq \sqrt{1-(pa)^2} \leq \sqrt{1-(pa)^{2t}} \tag{40}$$

for any $t \geq 1$ and $pa \leq 1$. The same argument applies for $\bar{\Delta}'_t$. We conclude that no clipping happens, which implies that $\Delta_t = \bar{\Delta}_t$ and $\Delta'_t = \bar{\Delta}'_t$

Let $u = -D$ be the comparator for both $\mathcal{A}$ and $\mathcal{A}'$. The regret in each round t of the algorithm $\mathcal{A}$ and $\mathcal{A}'$ are, respectively,

$$f_{\mathcal{A}}(t,u) := v_t(\Delta_t - u) = va^t\left(1 - \frac{p^{t-1}\sqrt{1-(pa)^2}}{K\sqrt{1-(pa)^{2t}}}\frac{1-a^t}{1-a}\right), \tag{41}$$

$$f_{\mathcal{A}'}(t,u) := v'_t(\Delta'_t - u) = vb^t\left(1 - \frac{\sqrt{1-b^2}}{K\sqrt{1-b^{2t}}}\frac{1-b^t}{1-b}\right). \tag{42}$$

Finally, we specify a sufficient condition for a and b such that the per-round regret of $\mathcal{A}$ is smaller than that of $\mathcal{A}'$.

Lemma 12 *For any $a, b \in (0,1)$ such that $a < b^2$, we have $f_{\mathcal{A}}(t,u) < f_{\mathcal{A}'}(t,u)$ for all $t \geq 1$.*

Proof Since $a < b^2 < b$, we have $1-a > 1-b$, thus $K = \max\{\frac{1}{1-a}, \frac{1}{1-b}\} = \frac{1}{1-b}$. Hence,

$$f_{\mathcal{A}}(t,u) = a^t\left(1 - \frac{p^t\sqrt{1-(pa)^2}}{K\sqrt{1-(pa)^{2t}}}\frac{1-a^t}{1-a}\right) \leq a^t < (b^2)^t = b^t(1-(1-b^t)) \tag{43}$$

$$= b^t\left(1 - \frac{1-b^t}{K(1-b)}\right) \leq b^t\left(1 - \frac{\sqrt{1-b^2}}{K\sqrt{1-b^{2t}}}\frac{1-b^t}{1-b}\right) = f_{\mathcal{A}'}(t,u), \tag{44}$$

where the last inequality is due to $1-b^2 \leq 1-b^{2t}$ for any $b \in (0,1)$. ■

The desired statement in Theorem 11 immediately follows from Lemma 12 and summing up over $t \in [T]$. ■

5. Conclusion and Future Works

In this work, we studied the Adam optimizer from an online learning perspective. By considering Adam as an instance of FTRL, we derived more general discounted regret bounds for Adam that hold beyond the restrictive setting of $\beta_1 = \sqrt{\beta_2}$ often required in existing works. For both cases $\beta_1 \geq \sqrt{\beta_2}$ and $\beta_1 \leq \sqrt{\beta_2}$, our new analyses and their bounds strictly generalize existing results. Moreover, we show that our bounds are worst-case tight and cannot be significantly improved furthers. Our results imply that when using Adam in the online-to-nonconvex framework, a rigorous approach towards tuning the two momentum factors β_1 and β_2 would require an exact modelling of the adversary. Future works include characterizing the adversary, i.e. how the sequence of gradients changes according to different values of β_1, β_2, on popular convex and non-convex objectives.

Acknowledgments

We thank Ali Mortazavi and Nishant Mehta for useful discussion on the proof and interpretation of Theorem 1. We thank the helpful suggestions from all reviewers. In particular, our Remarks 7 and 8 on clarifying the significance of our lower bound come from our rebuttal answer to reviewer 2J4Y.

References

Kwangjun Ahn and Ashok Cutkosky. Adam with model exponential moving average is effective for nonconvex optimization. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NeurIPS '24, Red Hook, NY, USA, 2025. Curran Associates Inc. ISBN 9798331314385.

Kwangjun Ahn, Zhiyu Zhang, Yunbum Kook, and Yan Dai. Understanding Adam optimizer via online learning of updates: Adam is FTRL in disguise. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Kwangjun Ahn, Gagik Magakyan, and Ashok Cutkosky. General framework for online-to-nonconvex conversion: Schedule-free SGD is also effective for nonconvex optimization. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=etxseIT47b.

Ahmet Alacaoglu, Yura Malitsky, Panayotis Mertikopoulos, and Volkan Cevher. A new regret analysis for Adam-type algorithms. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 202–210. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/alacaoglu20b.html.

Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 6643–6670. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/cutkosky23a.html.

Shinji Ito, Taira Tsuchiya, and Junya Honda. Adaptive learning rate for follow-the-regularized-leader: Competitive analysis and best-of-both-worlds. In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 2522–2563. PMLR, 30 Jun–03 Jul 2024. URL https://proceedings.mlr.press/v247/ito24a.html.

Ruinan Jin, Xiao Li, Yaoliang Yu, and Baoxiang Wang. A comprehensive framework for analyzing the convergence of adam: Bridging the gap with SGD. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=CiKWAofp7n.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

Francesco Orabona. A modern introduction to online learning. *CoRR*, abs/1912.13213, 2023. URL http://arxiv.org/abs/1912.13213.

Francesco Orabona and David Pal. Scale-free algorithms for online linear optimization, 2015. URL https://arxiv.org/abs/1502.05744.

Antonio Orvieto and Robert M. Gower. In search of adam's secret sauce. In *High-dimensional Learning Dynamics 2025*, 2025. URL https://openreview.net/forum?id=kjSBaukyRT.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ryQu7f-RZ.

Shohei Taniguchi, Keno Harada, Gouki Minegishi, Yuta Oshima, Seong Cheol Jeong, Go Nagahara, Tomoshi Iiyama, Masahiro Suzuki, Yusuke Iwasawa, and Yutaka Matsuo. ADOPT: Modified adam can converge with any β_2 with the optimal rate. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=rzvVm0LsyK.

Bhavya Vasudeva, Jung Whan Lee, Vatsal Sharan, and Mahdi Soltanolkotabi. The rich and the simple: On the implicit bias of adam and sgd, 2025. URL https://arxiv.org/abs/2505.24022.

Qinzi Zhang and Ashok Cutkosky. Random scaling and momentum for non-smooth non-convex optimization. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, and Zhi-Quan Luo. Adam can converge without any modification on update rules. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Appendix A. Technical Lemmas

Lemma 13 *For any* $x \in (0, 1], y \geq \frac{1}{x^2}$, *we have* $\frac{x^t(y^t-1)}{\sqrt{(x^2y^2)^t-1}} \leq 1$ *for all* $t \geq 1$.

Proof The inequality is equivalent to

$$\begin{aligned} & x^{2t}(y^{2t} - 2y^t + 1) \leq x^{2t}y^{2t} - 1 \\ \Leftrightarrow \quad & x^{2t}(2y^t - 1) \geq 1 \\ \Leftrightarrow \quad & 2y^t \geq 1 + \frac{1}{x^{2t}}. \end{aligned}$$

The last inequality is true due to $2y^t \geq \frac{2}{x^{2t}} \geq 1 + \frac{1}{x^{2t}}$ holds for all $x \in (0, 1]$. ■

Lemma 14 *For any* $x \in (0, 0.6], y \geq \frac{1}{x^2}$, *we have* $\frac{\sqrt{x^2y^2-1}}{x(y-1)} \leq 2$.

Proof The inequality is equivalent to

$$\begin{aligned} & 4x^2(y^2 - 2y + 1) \geq x^2y^2 - 1 \\ \Leftrightarrow \quad & x^2(3y^2 - 8y + 4) \geq -1. \end{aligned}$$

The last inequality is true due to $y \geq \frac{1}{x^2} > \frac{8}{3}$ holds for all $x \in (0, 0.6]$. ■

Appendix B. A Recent Empirical Finding From Orvieto and Gower (2025)

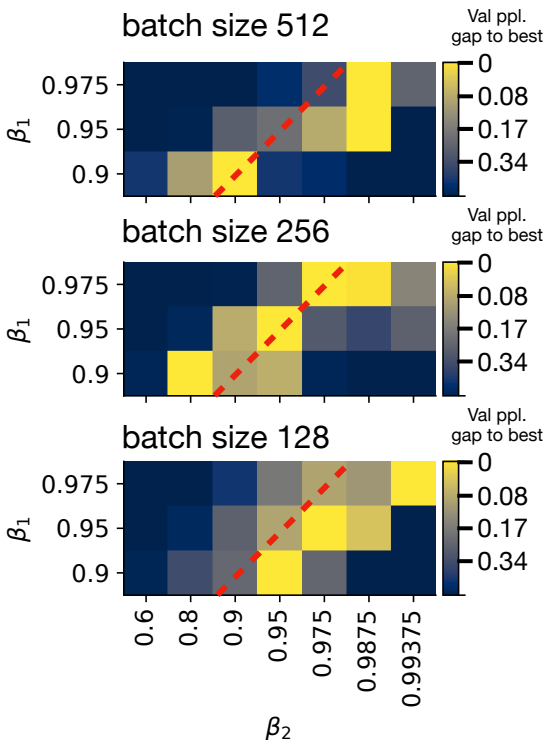

Figure 1: This is Figure 3 in Orvieto and Gower (2025), demonstrating the empirical results of tuning β_1, β_2 across three batch sizes for training $160M$-parameter transformers. Yellow indicates optimal performances, while dark blue indicates sub-optimal performances. The smallest $\frac{\beta_1}{\sqrt{\beta_2}}$ ratio of a yellow box is approximately 1, achieved at batch size 256, $\beta_1 = 0.9$ and $\beta_2 = 0.8$.