Adam with model exponential moving average is effective for nonconvex optimization

Kwangjun Ahn
Microsoft Research
Cambridge, MA
kwangjunahn@microsoft.com

Ashok Cutkosky
Boston University
Boston, MA
ashok@cutkosky.com

Abstract

In this work, we offer a theoretical analysis of two modern optimization techniques for training large and complex models: (i) adaptive optimization algorithms, such as Adam, and (ii) the model exponential moving average (EMA). Specifically, we demonstrate that a clipped version of Adam with model EMA achieves the optimal convergence rates in various nonconvex optimization settings, both smooth and nonsmooth. Moreover, when the scale varies significantly across different coordinates, we demonstrate that the coordinate-wise adaptivity of Adam is provably advantageous. Notably, unlike previous analyses of Adam, our analysis crucially relies on its core elements—momentum and discounting factors—as well as model EMA, motivating their wide applications in practice.

1 Introduction

In neural network training, the training loss $F : \mathbb{R}^d \to \mathbb{R}$ is often optimized using an iterative optimization algorithm which starts with the initial iterate $\mathbf{x}_0$ and then updates during each iteration $t = 1, 2, \dots$ as follows:

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \mathbf{z}_t \,,$$

where $\mathbf{z}_t$ denotes the increment chosen by the algorithm during the t-th iteration. One of the most popular optimization algorithms is **Adam** [Kingma and Ba, 2014]. Adam has gained significant attention due to its effectiveness in training Transformer-based language models [Zhang et al., 2020a, Kunstner et al., 2023, Jiang et al., 2023, Pan and Li, 2023, Ahn et al., 2024a, Kunstner et al., 2024, Zhang et al., 2024a].

The **model exponential moving average** (EMA) [Polyak and Juditsky, 1992, Ruppert, 1988] is an optimization technique that has gained popularity in conjunction with Adam for various recent applications. EMA maintains an exponential moving average of the model iterates, $\mathbf{x}_t$, which contributes to the stabilization of these iterates. There has been a resurgence of interest in this technique due to its effectiveness in training high-quality generative models [Yaz et al., 2018, Karras et al., 2019, Song et al., 2021b, Dhariwal and Nichol, 2021, Nichol and Dhariwal, 2021, Song et al., 2021a, Balaji et al., 2022, Karras et al., 2022, Rombach et al., 2022, Kang et al., 2023, Karras et al., 2023]. Moreover, a recent work by Block et al. [2024] demonstrates the effectiveness of EMA for both language modeling and imitation learning applications.

In this work, we theoretically study the effectiveness of these two modern optimization techniques. Our main results can be informally summarized as follows.

Theorem 1 (Informal). *CLIPPED-ADAM with the EMA on its iterates achieves the optimal convergence rate for nonconvex optimization both for smooth and nonsmooth settings (Section 5). The coordinate-wise adaptivity of Adam is particularly effective when the scale varies across different coordinates (Section 6).*

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Our main results are based on the online-to-nonconvex conversion framework of Cutkosky et al. [2023], which chooses the increment $\mathbf{z}_t$ based on an online learner of choice. In particular, our approach is quite different than the previous analyses of Adam (see Section 1.1 below). Notably, our analysis relies on the key components of Adam (momentum and adaptive learning rate) as well as EMA of the iterates, offering new, theoretical insight into their success. See Section 7 for a more detailed discussion.

At a high level, our analysis combines the main insights from the two recent works: Zhang and Cutkosky [2024] and Ahn et al. [2024b]. We first carefully modify the discounted-to-nonconvex conversion framework (Lemma 7) of Zhang and Cutkosky [2024] which converts an online learner that achieves a good discounted regret (Definition 6) into a good noncovex optimizer. We then combine it with the main insight of Ahn et al. [2024b] that an effective discounted online learner can be designed based on scale-free Follow-the-Regularized-Leader (FTRL) [Orabona and Pál, 2018]. In particular, the way we arrive at Adam is similar to Ahn et al. [2024b]: choosing a discounted version of FTRL in the discounted-to-nonconvex conversion leads to Adam.

1.1 Related work

Even though Adam is widely used in deep learning, our theoretical understanding of its inner workings, especially the importance of its core components—momentum and discounting factors—remains incomplete, as pointed out by Ahn et al. [2024b]. Most theoretical work on Adam and its variations focus on characterizing the convergence rate for convex or smooth nonconvex functions, where methods like SGD already achieve the minimax optimal convergence rate. [Reddi et al., 2018, Zhou et al., 2019, Chen et al., 2019, Zou et al., 2019, Alacaoglu et al., 2020, Guo et al., 2021, Défossez et al., 2022, Zhang et al., 2022, Li et al., 2023, Wang et al., 2023]. In fact, even the most recent results [Li et al., 2023, Wang et al., 2023] are not reflective of practice, in the sense that Adam's convergence rate worsens with momentum [Wang et al., 2023, §6] or is no better than that of SGD [Li et al., 2023, §7]. A notable exception is Crawshaw et al. [2022], which demonstrates the advantages of momentum under the generalized smoothness conditions of Zhang et al. [2020a]. However, the algorithm they analyze is signSGD, which differs significantly from the original Adam. In contrast, as we will explore in the subsequent sections, momentum and discounting factors are important in our analysis. See Section 7 for more details.

We also highlight that our analysis relies on model EMA, a technique widely used in practice as mentioned above (also see a recent work by Block et al. [2024]). It is worth noting that EMA (or model averaging in general) has shown to have generalization benefits in practice [Tarvainen and Valpola, 2017, Izmailov et al., 2018]. In this paper, we study EMA from an optimization perspective, and show that the use of EMA leads to optimal guarantees for nonconvex optimization. Interestingly, EMA naturally derives from the discounted-to-online conversion (see Algorithm 1), which, we believe, provides new theoretical insights into this practical method.

The use of EMA also represents a significant departure from most non-convex optimization analyses. While EMA is a classical technique in the *convex* setting, theoretical analyses in the non-convex setting typically randomly select an iterate as the "final output" of the optimizer, rather than using EMA. This random selection is intuitively extremely impractical (indeed, on average it actually wastes half of the computation), and is never performed in real implementations.

Our analysis follows a line of work studying convergence guarantees for non-smooth non-convex optimization. Our particular convergence criterion is similar to finding the Goldstein stationary points [Goldstein, 1977] that were first studied in the context of modern machine learning by [Zhang et al., 2020b], and has seen much subsequent interest [Tian and So, 2022, Jordan et al., 2023, Davis et al., 2020]. Other notions of convergence are also reasonable—common alternatives involve the Moreau envelope, or imposing a weak convexity condition [Davis et al., 2018, 2022a].

2 Setting for nonconvex and nonsmooth optimization

Throughout this paper, unless specified otherwise, $\|\cdot\|$ denotes the L_2 norm. Following [Cutkosky et al., 2023], we consider optimizing a loss function F that satisfies the following conditions, accessing information about F through a *stochastic gradient oracle* $\textsc{StoGrad} : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}^d$, for the set of randomness $\mathcal{Z}$.

Assumption 2. *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function with the following properties:*

- *Let $\Delta := F(\mathbf{x}_0) - \inf_{\mathbf{x}} F(\mathbf{x})$.*
- *For any two points $\mathbf{x}$ and $\mathbf{y}$, $F(\mathbf{y}) - F(\mathbf{x}) = \int_0^1 \langle \nabla F(\mathbf{x} + t(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x} \rangle \, \mathrm{d}t$.*
- **Lipschitzness.** *F is G-Lipshitz,* i.e., *for any point $\mathbf{x}$, $\|\nabla F(\mathbf{x})\| \leq G$.*
- **Stochastic gradient variance.** *For any point $\mathbf{x}$, the stochastic gradient $\mathbf{g} \leftarrow \textsc{StoGrad}(\mathbf{x}, r)$ for randomness $r \in \mathcal{Z}$ satisfies $\mathbb{E}[\mathbf{g}] = \nabla F(\mathbf{x})$ and $\mathbb{E}\,\|\mathbf{g} - \nabla F(\mathbf{x})\|^2 \leq \sigma^2$.*

The Lipschitz continuity condition is a standard assumption in nonconvex nonsmooth settings. However, as we will discuss in Section 6, one of the key insights from our results is that Adam enables us to **adapt** to the Lipschitz constants **coordinate-wise** without requiring prior knowledge of these constants. Note that we almost certainly need some form of structural assumption on the "difficulty" of the loss function; thus, relaxing the Lipschitz assumption would likely come at the cost of another assumption, such as smoothness.

The second condition, called *well-behavedess* in [Cutkosky et al., 2023, Definition 1], is a mild regularity condition. For any locally Lipschitz function F, applying an arbitrarily small perturbation to the function is sufficient to ensure this condition [Cutkosky et al., 2023, Proposition 2].

For the notion of optimality, we follow Zhang and Cutkosky [2024] and consider the following notion of stationarity for nonconvex and nonsmooth functions. This notion is a slight relaxation of the notion of a Goldstein stationarity point [Goldstein, 1977], which was further studied by recent works [Zhang et al., 2020a, Davis et al., 2022b, Tian et al., 2022, Jordan et al., 2023].

Definition 3 (**(λ, ε)-stationary point).** *Suppose $F : \mathbb{R}^d \to \mathbb{R}$ is differentiable. We say $\mathbf{x}$ is a (λ, ε)-stationary point of F if $\|\nabla F(\mathbf{x})\|^{[\lambda]} \leq \varepsilon$, where*

$$\|\nabla F(\mathbf{x})\|^{[\lambda]} := \inf_{\substack{p \in \mathcal{P}(\mathbb{R}^d), \\ \mathbb{E}_{\mathbf{y} \sim p}[\mathbf{y}] = \mathbf{x}}} \left\{ \|\mathbb{E}[\nabla F(\mathbf{y})]\| + \lambda \cdot \mathbb{E}\,\|\mathbf{y} - \mathbf{x}\|^2 \right\} .$$

To further motivate this definition, we remark that (λ, ϵ)-stationary points retain the desirable properties of Goldstein stationary points. Specifically, the following result [Zhang and Cutkosky, 2024, Lemma 2.3] demonstrates that, akin to Goldstein stationary points, (λ, ϵ)-stationary points can be reduced to first-order stationary points with appropriate choices of λ when the objective function is smooth or second-order smooth.

Lemma 4. *If F is L-smooth, then an $(L^2\varepsilon^{-1}, \varepsilon)$-stationary point $\mathbf{x}$ of F satisfies $\|\nabla F(\mathbf{x})\| \leq 2\varepsilon$. Moreover, if F is H-second-order-smooth, then an $(H/2, \varepsilon)$-stationary point $\mathbf{x}$ of F satisfies $\|\nabla F(\mathbf{x})\| \leq 2\varepsilon$.*

Moreover, as shown by [Zhang and Cutkosky, 2024, Lemma 2.4], (λ, ε)-stationary points can also be reduced to Goldstein stationary points when F is Lipschitz.

Lemma 5. *Suppose F is G-Lipschitz. For any $\lambda, \varepsilon, \delta > 0$, a (λ, ε)-stationary point is a (δ, ε')-Goldstein stationary point, where $\varepsilon' = (1 + \frac{2G}{\lambda\delta^2}) \cdot \varepsilon$.*

Now we design algorithms that find (λ, ε)-stationary points efficiently.

3 Discounted-to-nonconvex conversion: online learning of increments

Our main results are built on the online-to-nonconvex conversion framework of Cutkosky et al. [2023]. At its core, this framework involves selecting the increment $\mathbf{z}_t$ using an online learner, as discussed by Ahn et al. [2024b]. Specifically, we follow a variant developed by Zhang and Cutkosky [2024], which carefully incorporates the discounting factor in the conversion process. Note that we make slight modifications to the version proposed by Zhang and Cutkosky [2024] as follows. Here $\mathrm{Exp}(1)$ denotes the exponential random variable with mean 1.

Given Algorithm 1, it turns out we need to design an online learner that minimizes the discounted regret, formally defined below. It is worth noting that discounted regret has been recently studied with the goal of better adapting online learners to dynamic environments [Ahn et al., 2024b, Zhang et al., 2024b, Jacobsen and Cutkosky, 2024].

Algorithm 1 Discounted-to-nonconvex conversion (choosing increments via online learning)

```
Input: Initial point $\mathbf{x}_0$, $T \in \mathbb{N}$, online learning algorithm $\mathcal{A}$, and discounting factor $\beta \in (0,1)$
for $t = 1, 2 \ldots, T$ do
    Receive $\mathbf{z}_t$ from $\mathcal{A}$ // choose the increment using an online learner
    Update $\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} + \alpha_t \mathbf{z}_t$, where $\alpha_t \sim \mathrm{Exp}(1)$ i.i.d.
    Compute $\mathbf{g}_t \leftarrow \textsc{StoGrad}(\mathbf{x}_t, r_t)$ with freshly sampled randomness $r_t$
    Send $\ell_t^{[\beta]}(\mathbf{z}) := \langle \beta^{-t}\mathbf{g}_t, \mathbf{z} \rangle$ to $\mathcal{A}$
    // Maintain exponential moving average (for output only):
    Update $\overline{\mathbf{x}}_t \leftarrow \frac{\beta-\beta^t}{1-\beta^t}\overline{\mathbf{x}}_{t-1} + \frac{1-\beta}{1-\beta^t}\mathbf{x}_t$    (Equivalently, $\overline{\mathbf{x}}_t \leftarrow \frac{1-\beta}{1-\beta^t}\sum_{s=1}^t \beta^{t-s}\mathbf{x}_s$)
end for
```

Definition 6 (Discounted regret). *For a comparator* $\mathbf{u}$, *the* β*-discounted regret is defined as*

$$\mathrm{Regret}_t^{[\beta]}(\mathbf{u}) := \beta^t \cdot \sum_{s=1}^{t} (\ell_s^{[\beta]}(\mathbf{z}_s) - \ell_s^{[\beta]}(\mathbf{u})) = \sum_{s=1}^{t} \beta^{t-s} \langle \mathbf{g}_s, \mathbf{z}_s - \mathbf{u} \rangle \ .$$

The discounted regret of an online learner $\mathcal{A}$ can be used to upper bound the norm of averaged gradients, as shown in the following result.

Lemma 7 (Discounted-to-nonconvex conversion). *Suppose that* F *satisfies Assumption 2. Then for the comparator sequence chosen as* $\mathbf{u}_t := -D\frac{\sum_{s=1}^t \beta^{-s}\nabla F(\mathbf{x}_s)}{\|\sum_{s=1}^t \beta^{-s}\nabla F(\mathbf{x}_s)\|}$, *Algorithm 1 gives*

$$\mathop{\mathbb{E}}_{t\sim[T]} \mathbb{E}\left\|\mathop{\mathbb{E}}_{\mathbf{y}_t} \nabla F(\mathbf{y}_t)\right\| \le \frac{\Delta}{DT} + \frac{2G+\sigma}{(1-\beta)T} + \sigma\sqrt{1-\beta}$$
$$+ \frac{1}{DT}\left[\beta \cdot \mathbb{E}\left[\mathrm{Regret}_T^{[\beta]}(\mathbf{u}_T)\right] + (1-\beta)\cdot\sum_{t=1}^{T}\mathbb{E}\left[\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t)\right]\right] ,$$

where $\mathbf{y}_t$ *is distributed over* $\{\mathbf{x}_s\}_{s=1}^t$ *as* $\mathbb{P}(\mathbf{y}_t = \mathbf{x}_s) = \beta^{t-s}\cdot\frac{1-\beta}{1-\beta^t}$ *for* $s = 1, 2, \ldots, t$.

The proof combines the techniques of [Cutkosky et al., 2023, Theorem 7] and [Zhang and Cutkosky, 2024, Theorem 3.3]. See Appendix A for details.

We briefly explain how Lemma 7 can be used to find a (λ, ϵ)-stationary point (Definition 3). Recall that (λ, ϵ)-stationarity essentially requires producing a point $\mathbf{x} = \mathbb{E}[\mathbf{y}]$ such that both $\|\mathbb{E}[\nabla F(\mathbf{y})]\|$ and $\mathbb{E}\|\mathbf{y} - \mathbf{x}\|^2$ are small.

Given this context, Lemma 7 states that as long as the discounted regret of the online learner $\mathcal{A}$ is small, we can ensure that the EMA iterates $\overline{\mathbf{x}}_t = \mathbb{E}[\mathbf{y}_t]$ serve as good candidates for (λ, ϵ)-stationarity, since the term $\mathbb{E}\|\mathbb{E}_{\mathbf{y}_t} \nabla F(\mathbf{y}_t)\|$ can be kept small. The remaining task is to bound the variance term, $\mathbb{E}\|\mathbf{y}_t - \overline{\mathbf{x}}_t\|^2$, which will be addressed later in Lemma 10.

Moreover, the comparator $\mathbf{u}_t$ roughly models the *update direction that an oracle algorithm with perfect knowledge of the loss would select*. In the proof of Lemma 7, we demonstrate that moving along the $\mathbf{u}_t$ direction effectively decreases the loss value, which forms the basis for establishing our convergence guarantee.

Thanks to the discounted-to-nonconvex conversion, the task now reduces to designing an online learner that achieves low discounted regret.

4 Scale-free Follow-the-Regularized-Leader (FTRL)

In this section, we introduce an algorithmic component, called the Followed-The-Regularized-Leader (FTRL), a powerful online learning technique with various applications [Gordon, 1999, Kalai and Vempala, 2005, Shalev-Shwartz and Singer, 2006, Abernethy et al., 2008, Nesterov, 2009, Hazan and Kale, 2010].

For the setting, consider the online linear optimization (OLO) setting, where during each round $t = 1, \dots, T$, and online learner chooses $\mathbf{z}_t$, and then the linear loss $\ell_t(\cdot) = \langle \mathbf{v}_t, \cdot \rangle$ is revealed by the environment. Here the goal of the online learner is to minimize the regret defined as $\sum_t \langle \mathbf{v}_t, \mathbf{z}_t - \mathbf{u} \rangle$, where $\mathbf{u}$ is the comparator in hindsight. For this setting, FTRL is presented in Algorithm 2.

Algorithm 2 Follow-the-Regularized-Leader (FTRL)

```
Require: Regularizers $\{\psi_t\} : \mathbb{R}^d \to \mathbb{R}$, the domain $\mathcal{D} \subseteq \mathbb{R}^d$
for $t = 1, 2, \dots, T$ do
    Update $\mathbf{z}_t \leftarrow \mathrm{argmin}_{\mathbf{z} \in \mathcal{D}} \left[ \psi_t(\mathbf{z}) + \sum_{s=1}^{t-1} \ell_s(\mathbf{z}) \right]$
    Receive the next loss $\ell_t(\cdot) = \langle \mathbf{v}_t, \cdot \rangle$
end for
```

The key insight of Ahn et al. [2024b] and Zhang et al. [2024b] is that in order to design an online learner for discounted regret, it is important that the online learner is *scale-free* as described below. In particular, following Ahn et al. [2024b], we consider a gradient adaptive scale-free FTRL algorithm called *scale-free FTRL* [Orabona and Pál, 2018].

We will focus on the case where $\mathcal{D} = \mathbb{B}_D$, the d-dimensional L_2-ball of radius $D > 0$. Scale-free FTRL is given by Algorithm 2 with the following choie:

$$\psi_t(\cdot) = \frac{1}{\eta_t} \left\| \cdot \right\|^2 \quad \text{and} \quad \eta_t = \frac{D}{\sqrt{\sum_{s=1}^{t-1} \left\| \mathbf{v}_s \right\|^2}}.$$

Then using the clipping operator $\mathrm{clip}_D(\mathbf{x}) := \mathbf{x} \min(D/\|\mathbf{x}\|, 1)$, we can write down the update rule more explicitly as follows:

$$\mathbf{z}_t \leftarrow \operatorname*{argmin}_{\mathbf{z} \in \mathbb{B}_D} \left[\frac{1}{\eta_t} \left\| \mathbf{z} \right\|^2 + \sum_{s=1}^{t-1} \langle \mathbf{v}_s, \mathbf{z} \rangle \right] = -\mathrm{clip}_D \left(D \frac{\sum_{s=1}^{t-1} \mathbf{v}_s}{\sqrt{\sum_{s=1}^{t-1} \left\| \mathbf{v}_s \right\|^2}} \right) . \quad (\text{SCALE-FREE FTRL})$$

Here, if the denominator is zero, *i.e.*, $\mathbf{v}_1 = \cdots = \mathbf{v}_{t-1} = \mathbf{0}$, then we set $\mathbf{z}_t \leftarrow 0$. Note that this algorithm is scale-free in the sense that when the loss sequence is scaled by a scalar $c > 0$, the updates remain the same.

Let us now present the regret bound of SCALE-FREE FTRL.

Lemma 8 (Gradient-adaptive regret bound)**.** *For any $T > 0$, loss sequence $\mathbf{v}_{1:T}$ and comparator $\mathbf{u} \in \mathbb{R}^d$ s.t. $\|\mathbf{u}\| \leq D$,* SCALE-FREE FTRL *guarantees the following regret bound:*

$$\sum_{t=1}^{T} \langle \mathbf{v}_t, \mathbf{z}_t - \mathbf{u} \rangle \leq 4D \sqrt{\sum_{t=1}^{T} \left\| \mathbf{v}_t \right\|^2} .$$

We note that Lemma 8 follows (with a slightly worse constant) from [Orabona and Pál, 2018, Theorem 1], and the version we invoke is here due to [Ahn et al., 2024b, Theorem A.1].

Recall from Lemma 7 that an online learner for the discounted-to-nonconvex conversion (Algorithm 1) needs to have a low discounted regret. To achieve this, following Ahn et al. [2024b] and Zhang et al. [2024b], we simply substitute $\mathbf{v}_t \leftarrow \beta^{-t} \mathbf{g}_t$ into SCALE-FREE FTRL, resulting in the update

$$\mathbf{z}_t \leftarrow -\mathrm{clip}_D \left(D \frac{\sum_{s=1}^{t-1} \beta^{-s} \mathbf{g}_s}{\sqrt{\sum_{s=1}^{t-1} \beta^{-2s} \left\| \mathbf{g}_s \right\|^2}} \right) . \quad (\beta\text{-FTRL})$$

Here again, if the denominator is zero, *i.e.*, $\mathbf{g}_1 = \cdots = \mathbf{g}_{t-1} = \mathbf{0}$, then we set $\mathbf{z}_t \leftarrow 0$. Then, the following result characterizes the discounted regret guarantee of β-FTRL.

Theorem 9 (Discounted regret bound). *Let $\beta \in (0, 1]$. For any $T > 0$, loss sequence $\mathbf{g}_{1:T}$ and comparator $\mathbf{u} \in \mathbb{R}^d$ s.t. $\|\mathbf{u}\| \leq D$, β-FTRL guarantees the following static regret bound*

$$\mathrm{Regret}_T^{[\beta]}(\mathbf{u}) \leq 4D \sqrt{\sum_{t=1}^{T} \beta^{2(T-t)} \left\| \mathbf{g}_t \right\|^2} .$$

We next use this result to design an algorithm for nonconvex optimization.

5 Discounted-FTRL leads to adaptive nonconvex optimization

In this section, as a warm-up, let us see the implications of choosing $\mathcal{A} = \beta$-FTRL in Algorithm 1. First, let us obtain a bound on the expected discounted regret. By Theorem 9 together with Jensen's inequality, we have the following regret bound for any $t = 1, 2, \dots, T$:

$$\mathbb{E}\left[\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t)\right] \le 4D\,\mathbb{E}\sqrt{\sum_{t=1}^{T}\beta^{2(T-t)}\left\|\mathbf{g}_t\right\|^2} \le 4D\sqrt{\sum_{t=1}^{T}\beta^{2(T-t)}\,\mathbb{E}\left\|\mathbf{g}_t\right\|^2}\,.$$

Since $\mathbb{E}\left\|\mathbf{g}_t\right\|^2 \le G^2 + \sigma^2$ and $\frac{1}{\sqrt{1-\beta^2}} \le \frac{1}{\sqrt{1-\beta}}$, it follows that

$$\mathbb{E}\left[\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t)\right] \le \frac{4D(G+\sigma)}{\sqrt{1-\beta^2}} \le \frac{4D(G+\sigma)}{\sqrt{1-\beta}}\,. \tag{1}$$

5.1 From gradient-adaptive regret to nonconvex optimization

In order to obtain nonconvex optimization guarantees in terms of the (λ, ε)-stationarity (Definition 3), we need to handle the variance term. Following [Zhang and Cutkosky, 2024, Lemma 3.2], the variance term can be bounded as follows.

Lemma 10 (Variance bound)**.** *Using the notations of Lemma 7, for any $t = 1, 2, \dots, T$, β-FTRL satisfies*

$$\mathop{\mathbb{E}}_{t\sim[T]}\mathbb{E}\left\|\mathbf{y}_t - \overline{\mathbf{x}}_t\right\|^2 \le 12\frac{D^2}{(1-\beta)^2}\,.$$

Proof. From [Zhang and Cutkosky, 2024, Lemma 3.2], it follows that $\mathbb{E}\sum_{t=1}^{T}\left\|\mathbf{y}_t - \overline{\mathbf{x}}_t\right\|^2 \le \frac{12}{(1-\beta)^2}\,\mathbb{E}\sum_{t=1}^{T}\left\|\mathbf{z}_t\right\|^2$. Now since $\left\|\mathbf{z}_t\right\| \le D$ for all $t = 1, 2, \dots, T$, after dividing each side by T, we get the desired inequality. □

Plugging the regret bound (1) into Lemma 7 and combining it with Lemma 10, we arrive at the following optimization guarantee in terms of the (λ, ε)-stationarity. See Section B.1 for a proof.

Theorem 11. *Suppose that F satisfies Assumption 2 and consider any $\lambda > 0$. For $C > 0$, choose $\mathcal{A} = \beta$-FTRL in Algorithm 1 with the following parameters:*

$$\beta = 1 - \left(\frac{\varepsilon}{10C}\right)^2,\quad D = \frac{(1-\beta)\varepsilon^{1/2}}{4\lambda^{1/2}},\quad \text{and}\quad T = \frac{1}{1-\beta}\cdot\max\left\{\frac{4\Delta\lambda^{1/2}}{\varepsilon^{3/2}}, \frac{12C}{\varepsilon}\right\}.$$

Then we have $\mathbb{E}_{t\sim[T]}\,\mathbb{E}\left\|\nabla F(\overline{\mathbf{x}}_t)\right\|^{[\lambda]} \le (1 + \frac{G+\sigma}{C})\varepsilon$. In other words, a randomly chosen ***model EMA*** *$\overline{\mathbf{x}}_t$ is a $(\lambda, (1 + \frac{G+\sigma}{C})\varepsilon)$-stationary point, in expectation.*

5.2 Optimality and gradient adaptivity

Here, we discuss several notable aspects of the guarantee provided in Theorem 11.

5.2.1 Optimality

As shown in [Zhang and Cutkosky, 2024, Corollary 5.1], the lower bound on the iteration complexity for finding a (λ, ε)-stationary point is $\Omega((G+\sigma)^2\Delta\lambda^{1/2}\varepsilon^{-7/2})$, provided that $\lambda \le \frac{G^4}{\Delta^2}\varepsilon^{-1}$. Theorem 11 implies that setting $C = G + \sigma$ achieves this optimal iteration complexity.

Corollary 12. *In Theorem 11, choosing $C = G + \sigma$ leads to the following iteration complexity for finding a (λ, ε)-stationary point:*

$$O\left(\max\left\{\frac{(G+\sigma)^2\Delta\lambda^{1/2}}{\varepsilon^{7/2}}, \frac{(G+\sigma)^3}{\varepsilon^3}\right\}\right).$$

In particular, treating G, σ, and Δ as constants, as long as $\lambda \gtrsim \varepsilon$, this leads to the optimal complexity of $O((G+\sigma)^2\Delta\lambda^{1/2}\varepsilon^{-7/2})$.

In light of Lemma 4, the above optimal complexity can be converted into the optimal complexities for smooth settings.

Corollary 13 (Smooth settings). *Corollary 12 implies the following optimal iteration complexity for smooth settings. Choosing $\lambda = O(\varepsilon^{-1})$, it implies the optimal complexity of $O(\varepsilon^{-4})$ for smooth loss functions [Arjevani et al., 2023]. Similarly, with $\lambda = O(1)$, it achieves the optimal iteration complexitiy of $O(\varepsilon^{-7/2})$ for second-order smooth loss functions [Arjevani et al., 2020].*

We next discuss the benefits of using the gradient-adaptive regret bound (Theorem 9) by considering the case where we do not have knowledge of G, σ.

5.2.2 Gradient adaptivity

A remarkable consequence of Theorem 11 is that, due to the gradient-adaptive regret bound of Theorem 9, the final convergence guarantee has a better dependence on G, σ in the case when we do not have knowledge of them. For concreteness, in the following discussion, we treat G, σ, Δ as constants, and focus on the regime $\lambda \gtrsim \varepsilon$.

First, our Theorem 11 with $C = 1$ (since we do not know G, σ) leads to the following iteration complexity for finding a (λ, ε)-stationary point:

$$O\left((G+\sigma)^{7/2}\Delta\lambda^{1/2}\varepsilon^{-7/2}\right)$$

The price we pay for not knowing G, σ relative to the lower bound is a multiplicative factor of $(G+\sigma)^{3/2}$. To see the benefit of this adaptive regret approach, let us consider the guarantees given by Zhang and Cutkosky [2024]. Their approach is based on choosing online gradient descent for $\mathcal{A}$ in Algorithm 1, when the learning rate is not properly tuned with the knowledge of G and σ, it would lead to the following (suboptimal) discounted regret bound:

$$\mathbb{E}\left[\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t)\right] \le O\left(\frac{D(G+\sigma)^2}{\sqrt{1-\beta}}\right).$$

Then, the resulting iteration complexity becomes $O(\Delta\lambda^{1/2}(\frac{\varepsilon}{(G+\sigma)^2})^{-7/2})$, which is equal to $O\left((G+\sigma)^7\Delta\lambda^{1/2}\varepsilon^{-7/2}\right)$. This is larger than the complexity due to our adaptive approach by a multiplicative factor of $(G+\sigma)^{7/2}$.

Next, we build on the results from this section and consider a better approach to design an adaptive nonconvex optimizer.

6 Coordinate-wise adaptivity via (clipped-)Adam

In this section, we consider the setting where the Lipschitzness constants vary across different coordinates, which is empirically observed to be reflective of practical neural network training (see, *e.g.* [Crawshaw et al., 2022, Zhuang et al., 2022]). Formally, we consider the following setting.

Assumption 14. *Under the same setting as Assumption 2, we replace the last two conditions with the following coordinate-wise version:*

- *For each coordinate $i = 1, 2, \ldots, d$, there is a Lipschitzness constant $G_i > 0$ and a variance constant $\sigma_i > 0$ such that $\forall \mathbf{x}$, $|\partial_i F(\mathbf{x})| \le G_i$ and the stochastic gradient $\mathbf{g} \leftarrow \textsc{StoGrad}(\mathbf{x}, r)$ satisfies $\mathbb{E}[\mathbf{g}[i]] = \partial_i F(\mathbf{x}_i)$ and $\mathbb{E}\left|\mathbf{g}[i] - \partial_i F(\mathbf{x})\right|^2 \le \sigma_i^2$. (Here, $\partial_i F$ denotes the partial derivative of F w.r.t. the i-th coordinate.)*

Let $\boldsymbol{G} := (G_1, \ldots, G_d)$ and $\boldsymbol{\sigma} := (\sigma_1, \ldots, \sigma_d)$. Then, the above condition implies the last two conditions in Assumption 2 with $G = \|\boldsymbol{G}\|_2$ and $\sigma = \|\boldsymbol{\sigma}\|_2$.

As we mentioned before, the previous approaches [Cutkosky et al., 2023, Zhang and Cutkosky, 2024] choose the online learner $\mathcal{A}$ to be online gradient descent, and hence choosing the learning rate requires the knowledge of G_i, σ_i for all i. However, for neural network training, d is equal to the number of parameters in the network, so tuning them individually is computationally infeasible. We instead consider running β-FTRL **coordinate-wise** in Algorithm 1, which will automatically adapt to each coordinate. We begin with an important observation that such an approach in fact leads to a popular optimizer widely used in practice.

6.1 Coordinate-wise discounted FTRL corresponds to (clipped-)Adam

For notational simplicity, fix a coordinate among $i = 1, 2, \dots, d$, and let us denote the iterate by x_t, the stochastic gradient by g_t, and the update by z_t. Then the resulting optimizer becomes:

$$\boxed{z_{t+1} = -\mathrm{clip}_D\left(D \frac{\sum_{s=1}^t \beta_1^{t-s} g_s}{\sqrt{\sum_{s=1}^t \beta_2^{t-s} g_s^2}} \right),} \tag{CLIPPED-ADAM}$$

where $\beta_1 = \beta$ and $\beta_2 = \beta^2$. Here, again if the denominator is zero, *i.e.*, if $g_1 = \cdots = g_t = 0$, then we set the update to be zero, *i.e.*, $z_{t+1} = 0$. Note that **this is almost exactly the Adam optimizer** [Kingma and Ba, 2014], except that now we add clipping to control the variance of the iterates relative to their EMA. Notably, CLIPPED-ADAM retains one of the most important properties of Adam: it is *scale-invariant*. The scale invariance causes the optimizer to make updates of the same magnitude on each coordinate even when the scale differs across different coordinates.

In practice, we expect that the clipping operation will effectively be a no-op. This is because, when the algorithm is converging (even if the convergence is somewhat slow), the gradients are likely to behave as approximately mean-zero random variables (due to factors such as stochastic noise, unstable training trajectories, etc.). In such cases, standard concentration inequalities imply that $\sum_{s=1}^t \beta^{t-s} g_s \lesssim \sqrt{\sum_{s=1}^t (\beta^{t-s} g_s)^2}$, and hence, the clipping has no effect.

We also remark that CLIPPED-ADAM does not consider the "bias correction" terms in the original updates of Adam [Kingma and Ba, 2014]. However, note that the bias correction terms are coordinate-independent, and they can be merged into the scalar D.

6.2 Nonconvex optimization guarantees of CLIPPED-ADAM

We next discuss the theoretical guarantees of CLIPPED-ADAM for nonconvex and nonsmooth optimizaton. Inspired by [Duchi et al., 2010, McMahan and Streeter, 2010], where the coordinate-wise online learners lead to regret bounds with respect to the L_1 norms of stochastic gradients, we consider the following variant of Definition 3, in the same vein as [Cutkosky et al., 2023, Section 4].

Definition 15 **((λ, ε)-L_1-stationary point).** *Suppose $F : \mathbb{R}^d \to \mathbb{R}$ is differentiable. We say $\mathbf{x}$ is a (λ, ε)-L_1-stationary point of F if $\|\nabla F(\mathbf{x})\|_1^{[\lambda]} \le \varepsilon$, where*

$$\|\nabla F(\mathbf{x})\|_1^{[\lambda]} := \inf_{\substack{p \in \mathcal{P}(\mathbb{R}^d), \\ \mathbb{E}_{\mathbf{y}\sim p}[\mathbf{y}] = \mathbf{x}}} \left\{ \|\mathbb{E}[\nabla F(\mathbf{y})]\|_1 + \lambda \cdot \mathbb{E}\|\mathbf{y} - \mathbf{x}\|_2^2 \right\}.$$

Using the fact $\|\cdot\|_1 \le \sqrt{d}\,\|\cdot\|_2$, one can connect the two notions of (λ, ε)-stationary points.

Lemma 16. *A $(\lambda/\sqrt{d}, \varepsilon/\sqrt{d})$-stationary point is a (λ, ε)-L_1-stationary point.*

In order to obtain the guarantee in terms of L_1-norm, we consider the coordinate-wise version of discounted-to-online conversion, in the same vein as [Cutkosky et al., 2023, Appendix G]. See Section A.1 for details.

Lemma 17 **(L_1-variant of Lemma 7).** *Suppose that F satisfies Assumption 14. Consider the comparator sequence chosen as $\mathbf{u}_t$ defined as $\mathbf{u}_t[i] := -D\frac{\sum_{s=1}^t \beta^{-s}\partial_i F(\mathbf{x}_s)}{\left|\sum_{s=1}^t \beta^{-s}\partial_i F(\mathbf{x}_s)\right|}$ for $i = 1, 2, \dots, d$. Then, Algorithm 1 gives*

$$\mathop{\mathbb{E}}_{t\sim[T]} \mathbb{E}\left\| \mathop{\mathbb{E}}_{\mathbf{y}_t} \nabla F(\mathbf{y}_t) \right\|_1 \le \frac{\Delta}{DT} + \frac{2\|\boldsymbol{G} + \boldsymbol{\sigma}\|_1}{(1-\beta)T} + \|\boldsymbol{\sigma}\|_1 \sqrt{1-\beta}$$
$$+ \frac{1}{DT}\left[\beta \cdot \mathbb{E}\left[\mathrm{Regret}_T^{[\beta]}(\mathbf{u}_T)\right] + (1-\beta) \cdot \sum_{t=1}^T \mathbb{E}\left[\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t)\right] \right],$$

where $\mathbf{y}_t$ is distributed over $\{\mathbf{x}_s\}_{s=1}^t$ as $\mathbb{P}(\mathbf{y}_t = \mathbf{x}_s) = \beta^{t-s} \cdot \frac{1-\beta}{1-\beta^t}$ for $s = 1, 2, \dots, t$.

Next, let us consider the (expected) regret bound. Fix a coordinate $i = 1, \dots, d$. Then, by the one-dimensional version of Theorem 9 together with Jensen's inequality, we have the following regret bound for any $t = 1, 2, \dots, T$:

$$\mathbb{E}\left[\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t[i])\right] \le 4D\sqrt{\sum_{t=1}^{T} \beta^{2(T-t)}\,\mathbb{E}\left|\mathbf{g}_t[i]\right|^2} \le \frac{4D(G_i+\sigma_i)}{\sqrt{1-\beta}}\,.$$

Hence, taking the sum over all coordinates $i = 1, \dots, d$, we obtain

$$\mathbb{E}\left[\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t)\right] \le \frac{4D\left\|\boldsymbol{G}+\boldsymbol{\sigma}\right\|_1}{\sqrt{1-\beta}}\,. \tag{2}$$

Combining these together, one get the following guarantee in terms of the L_1 norm. See Section B.2 for a proof.

Theorem 18. *Suppose that F satisfies Assumption 14 and consider any $\lambda > 0$. For $C > 0$, choose the coordinate-wise optimizer* CLIPPED-ADAM *in Algorithm 1 with the following parameters:*

$$\beta = 1 - \left(\frac{\varepsilon}{10C}\right)^2,\quad D = \frac{(1-\beta)\varepsilon^{1/2}}{4d^{1/2}\lambda^{1/2}},\quad \text{and}\quad T = \frac{1}{1-\beta}\cdot\max\left\{\frac{4\Delta d^{1/2}\lambda^{1/2}}{\varepsilon^{3/2}},\ \frac{12C}{\varepsilon}\right\}.$$

*Then we have $\mathbb{E}_{t\sim[T]}\,\mathbb{E}\left\|\nabla F(\overline{\mathbf{x}}_t)\right\|_1^{[\lambda]} \le (1+\frac{\|\boldsymbol{G}+\boldsymbol{\sigma}\|_1}{C})\varepsilon$. In other words, a randomly chosen **model EMA** $\overline{\mathbf{x}}_t$ is a $(\lambda, (1+\frac{\|\boldsymbol{G}+\boldsymbol{\sigma}\|_1}{C})\varepsilon)$-$L_1$-stationary point, in expectation.*

6.3 Benefits of coordinate-wise adaptivity of CLIPPED-ADAM

In this section, we discuss the benefits of coordinate-wise adaptivity by examining the guarantee from Theorem 18 and compare it with that of Theorem 11. We begin with the (λ, ε)-L_1-stationary point guarantee due to Theorem 18. We consider the scenario where β is carefully tuned by making the optimal choice of C.

Corollary 19. *In Theorem 18, choosing $C = \left\|\boldsymbol{G}+\boldsymbol{\sigma}\right\|_1$ leads to the following iteration complexity for finding a (λ, ε)-L_1-stationary point:*

$$O\left(\max\left\{\frac{\left\|\boldsymbol{G}+\boldsymbol{\sigma}\right\|_1^2\Delta d^{1/2}\lambda^{1/2}}{\varepsilon^{7/2}},\ \frac{\left\|\boldsymbol{G}+\boldsymbol{\sigma}\right\|_1^3}{\varepsilon^3}\right\}\right).$$

In order to better appreciate the benefits of coordinate-wise adaptivity, let us compare the above iteration complexity with that of Theorem 11.

For concreteness, we treat $G = \left\|\boldsymbol{G}\right\|_2$ and $\sigma = \left\|\boldsymbol{\sigma}\right\|_2$ as constants throughout, and more importantly, we assume that the **coordinates are heterogeneous** in the sense that

$$\left\|\boldsymbol{G}+\boldsymbol{\sigma}\right\|_1 \approx \left\|\boldsymbol{G}+\boldsymbol{\sigma}\right\|_2\,. \tag{3}$$

The assumption (3) roughly says that a few coordinates of $\boldsymbol{G}+\boldsymbol{\sigma}$ take much larger values than the rest; if all the coordinates of $\boldsymbol{G}+\boldsymbol{\sigma}$ have similar magnitudes, then $\left\|\boldsymbol{G}+\boldsymbol{\sigma}\right\|_1 \approx \sqrt{d}\left\|\boldsymbol{G}+\boldsymbol{\sigma}\right\|_2$. In the case $\lambda \gtrsim \varepsilon$, Corollary 19 implies that the iteration complexity is

$$O(\left\|\boldsymbol{G}+\boldsymbol{\sigma}\right\|_1^2\Delta d^{1/2}\lambda^{1/2}\varepsilon^{-7/2})\,. \tag{4}$$

Next, let us consider the counterpart that does not adapt to each coordinate separately. In this case, we apply Lemma 16, which tells us that to find a (λ, ε)-L_1 stationary point it suffices to find a $(\lambda/\sqrt{d}, \varepsilon/\sqrt{d})$-stationary point. Then, from Corollary 12, the iteration complexity is $O(\left\|\boldsymbol{G}+\boldsymbol{\sigma}\right\|_2^2\Delta(\lambda/\sqrt{d})^{1/2}(\varepsilon/\sqrt{d})^{-7/2})$, *i.e.*,

$$O(\left\|\boldsymbol{G}+\boldsymbol{\sigma}\right\|_2^2\Delta d^{3/2}\lambda^{1/2}\varepsilon^{-7/2})\,. \tag{5}$$

Hence, when (3) holds, (4) can be lower than (5) by a multiplicative factor of d, showing the benefits of coordinate-wise adaptivity.

7 Discussion

Our analyses of Adam based on the discounted-to-online conversion is quite different than the previous ones. As discussed in Section 1.1, the previous analyses often result in guarantees that are not quite reflective of practice—*e.g.*, the rates get better without momentum and the rates are no better than that of non-adaptive methods. In contrast, our analyses and results highlight the role of the practical components as highlighted below.

- **Momentum.** In order to obtain a low discounted regret, any sensible online learner should integrate the past history of stochastic gradients $\mathbf{g}_{1:t}$ to make the decision $\mathbf{z}_{t+1}$. Such online learners under the discounted-to-online conversion lead to momentum methods that integrate $\mathbf{g}_{1:t}$ to obtain the next increment $\mathbf{z}_{t+1}$. In particular, non-momentum methods would correspond to aggressive online learners that only use the last gradient $\mathbf{g}_t$ to make the decision $\mathbf{z}_{t+1}$. This perspective provides new insights into understanding the role of momentum, as echoed by Ahn et al. [2024b].
- **Adaptive learning rates.** The adaptive learning rate due to β-FTRL leads to a gradient-adaptive regret bound (Theorem 9), which is important to obtain a better Lipshitzness dependence (Section 5.2) as well as the coordinate-wise adaptivity for high-dimension settings (Section 6.3). Our analysis offers theoretical benefits of adaptive learning rate from a discounted regret perspective.
- **Model EMA.** Lastly, the discounted-to-nonconvex conversion (Algorithm 1) naturally leads to guarantees in terms of the model EMA, $\overline{\mathbf{x}}_t$. At a high level (see Appendix A precise details), this is because for a dynamic environment, it is important to discount the losses such that online learners adapt to changing environments. The appearance of model EMA in the discounted-to-nonconvex conversion provides a new perspective on its role.

Our analyses and results have several limitations and raise several interesting questions. Firstly, CLIPPED-ADAM does not precisely match the original Adam algorithm, warranting further investigation into the original Adam update. Specifically, our analysis suggests choosing $\beta_1 = \beta$ and $\beta_2 = \beta^2$, which does not align with the commonly used practical choices. Understanding the exact roles of these practical choices for β_1 and β_2 would be valuable.

In Section 5.2, we observed that our iteration complexity for finding a (λ, ε)-stationary point is $O(\Delta(G+\sigma)^{7/2}\lambda^{1/2}\varepsilon^{-7/2})$ when G and σ are unknown. Investigating whether this complexity is optimal presents another intriguing direction for future research.

Lastly, from a practical standpoint, developing a more advanced online learner for discounted regret and designing an algorithm that surpasses Adam in practicality would have significant practical implications.

Funding Acknowledgments

AC is supported by NSF grant number CCF-2211718.

References

Jacob D. Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Conference on Learning Theory (COLT)*, pages 263–274. Omnipress, 2008.

Kwangjun Ahn, Xiang Cheng, Minhak Song, Chulhee Yun, Ali Jadbabaie, and Suvrit Sra. Linear attention is (maybe) all you need (to understand transformer optimization). In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=0uI5415ry7.

Kwangjun Ahn, Zhiyu Zhang, Yunbum Kook, and Yan Dai. Understanding Adam optimizer via online learning of updates: Adam is FTRL in disguise. In *International Conference on Machine Learning*. PMLR, 2024b.

Ahmet Alacaoglu, Yura Malitsky, Panayotis Mertikopoulos, and Volkan Cevher. A new regret analysis for Adam-type algorithms. In *International conference on machine learning (ICML)*, pages 202–210. PMLR, 2020.

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Ayush Sekhari, and Karthik Sridharan. Second-order information in non-convex stochastic optimization: Power and limitations. In *Conference on Learning Theory*, pages 242–299, 2020.

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1): 165–214, 2023.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. eDiff-I: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

Adam Block, Dylan J Foster, Akshay Krishnamurthy, Max Simchowitz, and Cyril Zhang. Butterfly effects of SGD noise: Error amplification in behavior cloning and autoregression. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=CgPs04l9TO.

Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of A class of Adam-type algorithms for non-convex optimization. In *International Conference on Learning Representations (ICLR)*, 2019.

Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to unbounded smoothness of generalized signsgd. *Advances in neural information processing systems*, 35:9955–9968, 2022.

Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. In *International Conference on Machine Learning (ICML)*, volume 202, pages 6643–6670. PMLR, 2023.

Damek Davis, Dmitriy Drusvyatskiy, Kellie J MacPhee, and Courtney Paquette. Subgradient methods for sharp weakly convex functions. *Journal of Optimization Theory and Applications*, 179:962–982, 2018.

Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D Lee. Stochastic subgradient method converges on tame functions. *Foundations of computational mathematics*, 20(1):119–154, 2020.

Damek Davis, Mateo Díaz, and Dmitriy Drusvyatskiy. Escaping strict saddle points of the moreau envelope in nonsmooth optimization. *SIAM Journal on Optimization*, 32(3):1958–1983, 2022a.

Damek Davis, Dmitriy Drusvyatskiy, Yin Tat Lee, Swati Padmanabhan, and Guanghao Ye. A gradient sampling method with complexity guarantees for lipschitz functions in high and low dimensions. In *Advances in Neural Information Processing Systems*, volume 35, pages 6692–6703. Curran Associates, Inc., 2022b.

Alexandre Défossez, Leon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of Adam and AdaGrad. *Transactions on Machine Learning Research (TMLR)*, 2022.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. In *Conference on Learning Theory (COLT)*, pages 257–269, 2010.

AA Goldstein. Optimization of lipschitz continuous functions. *Mathematical Programming*, 13: 14–22, 1977.

Geoffrey J. Gordon. Regret bounds for prediction problems. In *Conference on Learning Theory (COLT)*, pages 29–40. ACM, 1999.

Zhishuai Guo, Yi Xu, Wotao Yin, Rong Jin, and Tianbao Yang. A novel convergence analysis for algorithms of the Adam family. *arXiv preprint arXiv:2112.03459*, 2021.

Elad Hazan and Satyen Kale. Extracting certainty from uncertainty: Regret bounded by variation in costs. *Machine Learning*, 80:165–188, 2010.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, pages 876–885. Association For Uncertainty in Artificial Intelligence (AUAI), 2018.

Andrew Jacobsen and Ashok Cutkosky. Online linear regression in dynamic environments via discounting. In *International Conference on Machine Learning*. PMLR, 2024.

Kaiqi Jiang, Dhruv Malik, and Yuanzhi Li. How does adaptive optimization impact local neural network geometry? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://openreview.net/forum?id=gIG8LvTLuc.

Michael Jordan, Guy Kornowski, Tianyi Lin, Ohad Shamir, and Manolis Zampetakis. Deterministic nonsmooth nonconvex optimization. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 4570–4597. PMLR, 2023.

Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences (JCSS)*, 71(3):291–307, 2005.

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10134, 2023.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. *arXiv preprint arXiv:2312.02696*, 2023.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2014.

Frederik Kunstner, Jacques Chen, Jonathan Wilder Lavington, and Mark Schmidt. Noise is not the main factor behind the gap between SGD and Adam on Transformers, but sign descent might be. In *International Conference on Learning Representations (ICLR)*, 2023.

Frederik Kunstner, Robin Yadav, Alan Milligan, Mark Schmidt, and Alberto Bietti. Heavy-tailed class imbalance and why adam outperforms gradient descent on language models. *arXiv preprint arXiv:2402.19449*, 2024.

Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of Adam under relaxed assumptions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://openreview.net/forum?id=yEewbkBNzi.

H. Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, pages 244–256, 2010.

Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

Francesco Orabona and Dávid Pál. Scale-free online learning. *Theoretical Computer Science*, 716: 50–69, 2018.

Yan Pan and Yuanzhi Li. Toward understanding why Adam converges faster than SGD for Transformers. *arXiv preprint arXiv:2306.00204*, 2023.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

Shai Shalev-Shwartz and Yoram Singer. Online learning meets optimization in the dual. In *Conference on Learning Theory (COLT)*, volume 4005, pages 423–437. Springer, 2006.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. URL `https://openreview.net/forum?id=St1giarCHLP`.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL `https://openreview.net/forum?id=PxTIG12RRHS`.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Lai Tian and Anthony Man-Cho So. No dimension-free deterministic algorithm computes approximate stationarities of lipschitzians. *arXiv preprint arXiv:2210.06907*, 2022.

Lai Tian, Kaiwen Zhou, and Anthony Man-Cho So. On the finite-time complexity and practical computation of approximate stationarity concepts of lipschitz functions. In *International Conference on Machine Learning*, pages 21360–21379. PMLR, 2022.

Bohan Wang, Jingwen Fu, Huishuai Zhang, Nanning Zheng, and Wei Chen. Closing the gap between the upper bound and lower bound of Adam's iteration complexity. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yasin Yaz, Chuan-Sheng Foo, Stefan Winkler, Kim-Hui Yap, Georgios Piliouras, Vijay Chandrasekhar, et al. The unusual effectiveness of averaging in GAN training. In *International Conference on Learning Representations*, 2018.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations (ICLR)*, 2020a.

Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *International Conference on Machine Learning*, 2020b.

Qinzi Zhang and Ashok Cutkosky. Random scaling and momentum for non-smooth non-convex optimization. In *International Conference on Machine Learning*. PMLR, 2024.

Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, and Zhi-Quan Luo. Adam can converge without any modification on update rules. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 28386–28399, 2022.

Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhi-Quan Luo. Why transformers need adam: A hessian perspective. *arXiv preprint arXiv:2402.16788*, 2024a.

Zhiyu Zhang, David Bombara, and Heng Yang. Discounted adaptive online prediction. In *International Conference on Machine Learning*. PMLR, 2024b.

Zhiming Zhou, Qingru Zhang, Guansong Lu, Hongwei Wang, Weinan Zhang, and Yong Yu. Adashift: Decorrelation and convergence of adaptive learning rate methods. In *International Conference on Learning Representations (ICLR)*, 2019.

Zhenxun Zhuang, Mingrui Liu, Ashok Cutkosky, and Francesco Orabona. Understanding AdamW through proximal methods and scale-freeness. *Transactions on Machine Learning Research*, 2022.

Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of Adam and RMSProp. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11127–11135, 2019.

Appendix

A Proof of discounted-to-nonconvex conversion (Lemma 7)

Note first that via a change of summation, we get

$$\begin{aligned}
\sum_{n=1}^{T}\sum_{t=1}^{n}\beta^{n-t}(1-\beta)(F(\mathbf{x}_t)-F(\mathbf{x}_{t-1})) &= \sum_{t=1}^{T}\sum_{n=t}^{T}\beta^{n-t}(1-\beta)(F(\mathbf{x}_t)-F(\mathbf{x}_{t-1})) \\
&= \sum_{t=1}^{T}(1-\beta^{T-t+1})(F(\mathbf{x}_t)-F(\mathbf{x}_{t-1})) \\
&= F(\mathbf{x}_T)-F(\mathbf{x}_0)-\sum_{t=1}^{T}\beta^{T-t+1}(F(\mathbf{x}_t)-F(\mathbf{x}_{t-1}))\,.
\end{aligned}$$

Rearranging the above together with the fact $F(\mathbf{x}_0)-F(\mathbf{x}_T)\le F(\mathbf{x}_0)-\inf_{\mathbf{x}}F(\mathbf{x})=:\Delta$, we get

$$-\Delta \le \underbrace{\mathbb{E}\left[\sum_{n=1}^{T}\sum_{t=1}^{n}\beta^{n-t}(1-\beta)(F(\mathbf{x}_t)-F(\mathbf{x}_{t-1}))\right]}_{\textcircled{A}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\beta^{T-t+1}(F(\mathbf{x}_t)-F(\mathbf{x}_{t-1}))\right]}_{\textcircled{B}}\,.$$

We also recall the following fact about exponential random variable due to [Zhang and Cutkosky, 2024, Lemma 3.1].

Lemma 20. *Let $\alpha\sim Exp(\lambda)$ for some $\lambda>0$, then*

$$\underset{\alpha}{\mathbb{E}}[F(\mathbf{x}+\alpha\mathbf{z})-F(\mathbf{x})]=\underset{\alpha}{\mathbb{E}}[\langle\nabla F(\mathbf{x}+\alpha\mathbf{z}),\mathbf{z}\rangle]/\lambda\,.$$

By Lemma 20 with $\lambda=1$, together with $\mathbf{x}_t=\mathbf{x}_{t-1}+\alpha_t\mathbf{z}_t$, it follows that

$$\mathbb{E}[F(\mathbf{x}_t)-F(\mathbf{x}_{t-1})]=\mathbb{E}\left[\langle\mathbf{g}_t,\mathbf{z}_t\rangle\right].$$

This identity indicates that the function gap is exactly equal to the linearization of the function gap. In this sense, the randomization renders the first-order Taylor approximation perfectly accurate.

Now, with this result, we will address each term separately.

Analysis of $\textcircled{A}$

Note that for each $t\le n$, since $\mathbf{x}_t=\mathbf{x}_{t-1}+\alpha_t\mathbf{z}_t$ for α_t~Exp(1), Lemma 20 yields

$$\begin{aligned}
\mathbb{E}[F(\mathbf{x}_t)-F(\mathbf{x}_{t-1})] &= \mathbb{E}\left\langle\nabla F(\mathbf{x}_t),\mathbf{z}_t\right\rangle = \mathbb{E}\left\langle\nabla F(\mathbf{x}_t),\mathbf{u}_n\right\rangle + \mathbb{E}\left\langle\nabla F(\mathbf{x}_t),\mathbf{z}_t-\mathbf{u}_n\right\rangle \\
&= \mathbb{E}\left\langle\nabla F(\mathbf{x}_t),\mathbf{u}_n\right\rangle + \mathbb{E}\left\langle\nabla F(\mathbf{x}_t)-\mathbf{g}_t,\mathbf{z}_t-\mathbf{u}_n\right\rangle + \mathbb{E}\left\langle\mathbf{g}_t,\mathbf{z}_t-\mathbf{u}_n\right\rangle \\
&= \underbrace{\mathbb{E}\left\langle\nabla F(\mathbf{x}_t),\mathbf{u}_n\right\rangle}_{\textcircled{1}} + \underbrace{\mathbb{E}\left\langle\nabla F(\mathbf{x}_t)-\mathbf{g}_t,-\mathbf{u}_n\right\rangle}_{\textcircled{2}} + \underbrace{\mathbb{E}\left\langle\mathbf{g}_t,\mathbf{z}_t-\mathbf{u}_n\right\rangle}_{\textcircled{3}},
\end{aligned}$$

where the last line follows from the fact $\mathbb{E}[\langle \nabla F(\mathbf{x}_t) - \mathbf{g}_t, \mathbf{z}_t \rangle] = 0$. More specifically, note that the randomness in the stochastic gradient oracle is independent of the randomness due to α_t. Since $\mathbb{E}[\nabla F(\mathbf{x}_t) - \mathbf{g}_t] = 0$, it follows that

$$\mathbb{E}[\langle \nabla F(\mathbf{x}_t) - \mathbf{g_t}, \mathbf{z_t} \rangle] = \mathbb{E}[\langle \mathbb{E}[\nabla F(\mathbf{x}_t) - \mathbf{g_t}], \mathbf{z_t} \rangle] = 0,$$

where the inner expectation is with respect to the randomness in the stochastic gradient oracle and the outer is with respect to all other quantities.

Now let us handle each term.

①: Note that using the definition of $\mathbf{y}_n$, we have

$$\mathbb{E} \sum_{t=1}^{n} \beta^{n-t}(1-\beta) \langle \nabla F(\mathbf{x}_t), \mathbf{u}_n \rangle = (1-\beta)\, \mathbb{E} \left\langle \sum_{t=1}^{n} \beta^{n-t} \nabla F(\mathbf{x}_t), -D \frac{\sum_{t=1}^{n} \beta^{n-t} \nabla F(\mathbf{x}_t)}{\|\sum_{t=1}^{n} \beta^{n-t} \nabla F(\mathbf{x}_t)\|} \right\rangle$$

$$= (1-\beta^n)\, \mathbb{E} \left\langle \sum_{t=1}^{n} \frac{1-\beta}{1-\beta^n} \beta^{n-t} \nabla F(\mathbf{x}_t), -D \frac{\sum_{t=1}^{n} \frac{1-\beta}{1-\beta^n} \beta^{n-t} \nabla F(\mathbf{x}_t)}{\left\|\sum_{t=1}^{n} \frac{1-\beta}{1-\beta^n} \beta^{n-t} \nabla F(\mathbf{x}_t)\right\|} \right\rangle$$

$$= -D(1-\beta^n)\, \mathbb{E} \left\| \mathop{\mathbb{E}}_{\mathbf{y}_n} \nabla F(\mathbf{y}_n) \right\| \le -D\, \mathbb{E} \left\| \mathop{\mathbb{E}}_{\mathbf{y}_n} \nabla F(\mathbf{y}_n) \right\| + DG\beta^n \,.$$

Therefore, summing over $n = 1, \ldots, T$, we obtain:

$$\mathbb{E} \sum_{n=1}^{T} \sum_{t=1}^{n} \beta^{n-t}(1-\beta) \langle \nabla F(\mathbf{x}_t), \mathbf{u}_n \rangle \le -D\, \mathbb{E} \sum_{t=1}^{T} \left\| \mathop{\mathbb{E}}_{\mathbf{y}_t} \nabla F(\mathbf{y}_t) \right\| + \frac{DG}{1-\beta} \,.$$

②: For the second term, using Cauchy-Schwartz inequality, we have

$$\mathbb{E} \sum_{t=1}^{n} \beta^{n-t} \langle \nabla F(\mathbf{x}_t) - \mathbf{g}_t, -\mathbf{u}_n \rangle \le \sqrt{\mathbb{E} \left\| \sum_{t=1}^{n} \beta^{n-t} (\nabla F(\mathbf{x}_t) - \mathbf{g}_t) \right\|^2 \mathbb{E} \left\| \mathbf{u}_n \right\|^2} \,.$$

Using the bounded variance assumption on the stochastic gradient oracle, we have

$$\mathbb{E} \left\| \sum_{t=1}^{n} \beta^{n-t} (\nabla F(\mathbf{x}_t) - \mathbf{g}_t) \right\|^2 = \mathbb{E} \sum_{t=1}^{n} \beta^{2(n-t)} \left\| \nabla F(\mathbf{x}_t) - \mathbf{g}_t \right\|^2 \le \frac{\sigma^2}{1-\beta^2} \,.$$

Therefore, summing over $n = 1, \ldots, T$, and using the fact that $\frac{1}{1-\beta^2} \le \frac{1}{1-\beta}$, we get the following bound on the second term:

$$\mathbb{E} \sum_{n=1}^{T} \sum_{t=1}^{n} \beta^{n-t}(1-\beta) \langle \nabla F(\mathbf{x}_t) - \mathbf{g}_t, -\mathbf{u}_n \rangle \le \sum_{n=1}^{T} (1-\beta) \cdot \frac{\sigma D}{\sqrt{1-\beta^2}} \le \sigma D T \sqrt{1-\beta} \,.$$

③: Lastly, for the third term, we have

$$\begin{aligned} \mathbb{E} \sum_{n=1}^{T} \sum_{t=1}^{n} \beta^{n-t}(1-\beta) \langle \mathbf{g}_t, \mathbf{z}_t - \mathbf{u}_n \rangle &= (1-\beta)\, \mathbb{E} \sum_{n=1}^{T} \left[\sum_{t=1}^{n} \mathbb{E} \left\langle \beta^{n-t} \mathbf{g}_t, \mathbf{z}_t - \mathbf{u}_n \right\rangle \right] \\ &= (1-\beta)\, \mathbb{E} \sum_{t=1}^{T} \mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t) \,. \end{aligned}$$

Analysis of Ⓑ

Note that for each t, since $\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} + \alpha_t \mathbf{z}_t$ for $\alpha_t \sim \mathrm{Exp}(1)$, Lemma 20 yields

$$\begin{aligned} \mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}_{t-1})] &= \mathbb{E} \langle \nabla F(\mathbf{x}_t), \mathbf{z}_t \rangle = \mathbb{E} \langle \mathbf{g}_t, \mathbf{z}_t \rangle \\ &= \mathbb{E} \langle \mathbf{g}_t, \mathbf{z}_t - \mathbf{u}_T \rangle + \mathbb{E} \langle \mathbf{g}_t, \mathbf{u}_T \rangle \le \mathbb{E} \langle \mathbf{g}_t, \mathbf{z}_t - \mathbf{u}_T \rangle + D(G+\sigma) \,. \end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\beta^{T-t+1}(F(\mathbf{x}_t)-F(\mathbf{x}_{t-1}))\right] &= \beta\,\mathbb{E}\sum_{t=1}^{T}\left[\left\langle\beta^{T-t}\mathbf{g}_t,\mathbf{z}_t-\mathbf{u}_T\right\rangle+\beta^{T-t}D(G+\sigma)\right]\\ &\le \beta\,\mathbb{E}[\mathrm{Regret}_T^{[\beta]}(\mathbf{u}_T)]+\frac{D(G+\sigma)}{1-\beta}\,.\end{aligned}$$

Combining $\textcircled{A}$ and $\textcircled{B}$

Combining the above analyses and rearranging, it follows that

$$\begin{aligned}D\,\mathbb{E}\sum_{t=1}^{T}\left\|\mathop{\mathbb{E}}_{\mathbf{y}_t}\nabla F(\mathbf{y}_t)\right\| &\le \Delta+\frac{DG}{1-\beta}+\sigma DT\sqrt{1-\beta}+(1-\beta)\,\mathbb{E}\sum_{t=1}^{T}\left[\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t)\right]\\ &\quad+\beta\,\mathbb{E}[\mathrm{Regret}_T^{[\beta]}(\mathbf{u}_T)]+\frac{D(G+\sigma)}{1-\beta}\,.\end{aligned}$$

Dividing both sides by DT, we get the desired result.

A.1 Proof of the coordinate-wise version (Lemma 17)

The proof closely follows that of Lemma 7. In particular, with $\Delta \coloneqq F(\mathbf{x}_0)-\inf_{\mathbf{x}}F(\mathbf{x})$, we have

$$-\Delta \le \underbrace{\mathbb{E}\left[\sum_{n=1}^{T}\sum_{t=1}^{n}\beta^{n-t}(1-\beta)(F(\mathbf{x}_t)-F(\mathbf{x}_{t-1}))\right]}_{\textcircled{A}}+\underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\beta^{T-t+1}(F(\mathbf{x}_t)-F(\mathbf{x}_{t-1}))\right]}_{\textcircled{B}}\,.$$

We begin with the term $\textcircled{B}$. Using the same decomposition as before, we have

$$\begin{aligned}\mathbb{E}[F(\mathbf{x}_t)-F(\mathbf{x}_{t-1})] &= \mathbb{E}\left\langle\mathbf{g}_t,\mathbf{z}_t-\mathbf{u}_T\right\rangle+\mathbb{E}\left\langle\mathbf{g}_t,\mathbf{u}_T\right\rangle=\mathbb{E}\left\langle\mathbf{g}_t,\mathbf{z}_t-\mathbf{u}_T\right\rangle+\sum_{i=1}^{d}\mathbb{E}\,\mathbf{g}_t[i]\mathbf{u}_T[i]\\ &\le \mathbb{E}\left\langle\mathbf{g}_t,\mathbf{z}_t-\mathbf{u}_T\right\rangle+D\sum_{i=1}^{d}(G_i+\sigma_i)\,.\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\beta^{T-t+1}(F(\mathbf{x}_t)-F(\mathbf{x}_{t-1}))\right] &= \beta\,\mathbb{E}\sum_{t=1}^{T}\left[\left\langle\beta^{T-t}\mathbf{g}_t,\mathbf{z}_t-\mathbf{u}_T\right\rangle+\beta^{T-t}\sum_{i=1}^{d}D(G_i+\sigma_i)\right]\\ &\le \beta\,\mathbb{E}[\mathrm{Regret}_T^{[\beta]}(\mathbf{u}_T)]+\frac{D\sum_{i=1}^{d}(G_i+\sigma_i)}{1-\beta}\,.\end{aligned}$$

Moving onto the term $\textcircled{A}$, we again use the same decomposition:

$$\mathbb{E}[F(\mathbf{x}_t)-F(\mathbf{x}_{t-1})]=\underbrace{\mathbb{E}\left\langle\nabla F(\mathbf{x}_t),\mathbf{u}_n\right\rangle}_{\textcircled{1}}+\underbrace{\mathbb{E}\left\langle\nabla F(\mathbf{x}_t)-\mathbf{g}_t,-\mathbf{u}_n\right\rangle}_{\textcircled{2}}+\underbrace{\mathbb{E}\left\langle\mathbf{g}_t,\mathbf{z}_t-\mathbf{u}_n\right\rangle}_{\textcircled{3}}\,.$$

As before, let us handle each term one by one separately.

$\textcircled{1}$: Note that using the definition of $\mathbf{y}_n$, for each coordinate $i = 1, \dots, d$, we have

$$\mathbb{E}\sum_{t=1}^{n} \beta^{n-t}(1-\beta)\partial_i F(\mathbf{x}_t)\mathbf{u}_n[i] = (1-\beta)\,\mathbb{E}\left[\left(\sum_{t=1}^{n} \beta^{n-t}\nabla F(\mathbf{x}_t)\right)\left(-D\frac{\sum_{t=1}^{n}\beta^{n-t}\partial_i F(\mathbf{x}_t)}{|\sum_{t=1}^{n}\beta^{s-t}\partial_i F(\mathbf{x}_t)|}\right)\right]$$

$$= (1-\beta^n)\,\mathbb{E}\left[\left(\sum_{t=1}^{n}\frac{1-\beta}{1-\beta^n}\beta^{n-t}\nabla F(\mathbf{x}_t)\right)\left(-D\frac{\sum_{t=1}^{n}\frac{1-\beta}{1-\beta^n}\beta^{n-t}\nabla F(\mathbf{x}_t)}{\left|\sum_{t=1}^{n}\frac{1-\beta}{1-\beta^n}\beta^{n-t}\nabla F(\mathbf{x}_t)\right|}\right)\right]$$

$$= -D(1-\beta^n)\,\mathbb{E}\left|\mathop{\mathbb{E}}_{\mathbf{y}_n}\partial_i F(\mathbf{y}_n)\right| \le -D\,\mathbb{E}\left|\mathop{\mathbb{E}}_{\mathbf{y}_n}\partial_i F(\mathbf{y}_n)\right| + DG_i\beta^n\,.$$

Therefore, summing over $i = 1, \dots, d$ and then $n = 1, \dots, T$, we obtain:

$$\mathbb{E}\sum_{n=1}^{T}\sum_{t=1}^{n}\beta^{n-t}(1-\beta)\left\langle\nabla F(\mathbf{x}_t), \mathbf{u}_n\right\rangle \le -D\,\mathbb{E}\sum_{t=1}^{T}\left\|\mathop{\mathbb{E}}_{\mathbf{y}_t}\nabla F(\mathbf{y}_t)\right\|_1 + \frac{D\sum_{i=1}G_i}{1-\beta}\,.$$

$\textcircled{2}$: For each coordinate $i = 1, \dots, d$, we have

$$\mathbb{E}\sum_{t=1}^{n}\beta^{n-t}(\partial_i F(\mathbf{x}_t) - \mathbf{g}_t[i])(-\mathbf{u}_n[i]) \le \sqrt{\mathbb{E}\left|\sum_{t=1}^{n}\beta^{n-t}(\partial_i F(\mathbf{x}_t) - \mathbf{g}_t[i])\right|^2\mathbb{E}\left|\mathbf{u}_n[i]\right|^2}\,.$$

Using the coordinate-wise bounded variance assumption on the stochastic gradient oracle,

$$\mathbb{E}\left|\sum_{t=1}^{n}\beta^{n-t}(\partial_i F(\mathbf{x}_t) - \mathbf{g}_t[i])\right|^2 = \mathbb{E}\sum_{t=1}^{n}\beta^{2(n-t)}\left|\partial_i F(\mathbf{x}_t) - \mathbf{g}_t[i]\right|^2 \le \frac{\sigma_i^2}{1-\beta^2}\,.$$

Therefore, summing over $n = 1, \dots, T$, and using the fact that $\frac{1}{1-\beta^2} \le \frac{1}{1-\beta}$, we get the following bound on the second term:

$$\mathbb{E}\sum_{n=1}^{T}\sum_{t=1}^{n}\beta^{n-t}(1-\beta)\left\langle\nabla F(\mathbf{x}_t) - \mathbf{g}_t, -\mathbf{u}_n\right\rangle \le \sum_{n=1}^{T}(1-\beta)\cdot\frac{D\sum_{i=1}^{d}\sigma_i}{\sqrt{1-\beta^2}}$$

$$\le DT\left(\sum_{i=1}^{d}\sigma_i\right)\sqrt{1-\beta}\,.$$

$\textcircled{3}$: We use the same manipulation as before:

$$\mathbb{E}\sum_{n=1}^{T}\sum_{t=1}^{n}\beta^{n-t}(1-\beta)\left\langle\mathbf{g}_t, \mathbf{z}_t - \mathbf{u}_n\right\rangle = (1-\beta)\,\mathbb{E}\sum_{t=1}^{T}\mathrm{Regret}_t^{[\beta]}(\mathbf{u}_t)\,.$$

Combining the above, we get the desired result in Lemma 17.

B Proof of main theorems

B.1 Proof of Theorem 11

By Definition 3, since $\mathbb{E}[\mathbf{y}_t] = \overline{\mathbf{x}}_t$, it holds that

$$\mathop{\mathbb{E}}_{t\sim[T]}\left\|\nabla F(\overline{\mathbf{x}}_t)\right\|^{[\lambda]} \le \mathop{\mathbb{E}}_{t\sim[T]}\left[\left\|\mathop{\mathbb{E}}_{\mathbf{y}_t}\nabla F(\mathbf{y}_t)\right\| + \lambda\mathop{\mathbb{E}}_{\mathbf{y}_t}\left\|\mathbf{y}_t - \overline{\mathbf{x}}_t\right\|^2\right]\,.$$

We begin with the second term (the variance term). By Lemma 10, we have

$$\lambda\mathop{\mathbb{E}}_{t\sim[T]}\mathop{\mathbb{E}}_{\mathbf{y}_t}\left\|\mathbf{y}_t - \overline{\mathbf{x}}_t\right\|^2 \le 12\frac{\lambda D^2}{(1-\beta)^2}\,.$$

Hence, by choosing $D = \frac{(1-\beta)\varepsilon^{1/2}}{4\lambda^{1/2}}$, it follows that $\lambda \cdot \mathbb{E}_{t\sim[T]}\, \mathbb{E}_{\mathbf{y}_t} \left\|\mathbf{y}_t - \overline{\mathbf{x}}_t\right\|^2 \le \varepsilon$.

Next consider the first term (the norm of the averaged gradients). Plugging the regret bound (1) into Lemma 7, we get

$$\begin{aligned}\mathop{\mathbb{E}}_{t\sim[T]} \left\|\mathop{\mathbb{E}}_{\mathbf{y}_t} \nabla F(\mathbf{y}_t)\right\| &\le \frac{\Delta}{DT} + \frac{2G+\sigma}{(1-\beta)T} + \sigma\sqrt{1-\beta} + \frac{4(G+\sigma)}{T\sqrt{1-\beta}} + 4(G+\sigma)\sqrt{1-\beta} \\ &\le \frac{4\Delta\lambda^{1/2}}{(1-\beta)\varepsilon^{1/2}T} + \frac{6G+5\sigma}{(1-\beta)T} + (4G+5\sigma)\sqrt{1-\beta}\,,\end{aligned}$$

where the last line follows since $\frac{1}{\sqrt{1-\beta}} \le \frac{1}{1-\beta}$ and $D = \frac{(1-\beta)\varepsilon^{1/2}}{4\lambda^{1/2}}$. Choosing $\beta = 1 - (\frac{\varepsilon}{10C})^2$, the last term is bounded by $\frac{G+\sigma}{2C}\varepsilon$. Moreover, choosing $T = (1-\beta)^{-1} \cdot \max\left\{4\Delta\lambda^{1/2}\varepsilon^{-3/2},\ 12C\varepsilon^{-1}\right\}$, the first and second terms are bounded by ε and $\frac{G+\sigma}{2C}\varepsilon$, respectively. This concludes the proof.

B.2 Proof of Theorem 18

By Definition 15, since $\mathbb{E}[\mathbf{y}_t] = \overline{\mathbf{x}}_t$, it holds that

$$\mathop{\mathbb{E}}_{t\sim[T]} \|\nabla F(\overline{\mathbf{x}}_t)\|_1^{[\lambda]} \le \mathop{\mathbb{E}}_{t\sim[T]} \left[\left\|\mathop{\mathbb{E}}_{\mathbf{y}_t} \nabla F(\mathbf{y}_t)\right\|_1 + \lambda \mathop{\mathbb{E}}_{\mathbf{y}_t} \|\mathbf{y}_t - \overline{\mathbf{x}}_t\|_2^2\right].$$

We begin with the second term (the variance term). This time, given that now each coordinate of update $\mathbf{z}_t$ is bounded by D, *i.e.*, $|\mathbf{z}_t[i]| \le D$, applying the variance bound due to Lemma 10 coordinate-wise implies:

$$\lambda \mathop{\mathbb{E}}_{t\sim[T]} \mathop{\mathbb{E}}_{\mathbf{y}_t} \|\mathbf{y}_t - \overline{\mathbf{x}}_t\|_2^2 \le 12\frac{\lambda d D^2}{(1-\beta)^2}.$$

Hence, by choosing $D = \frac{(1-\beta)\varepsilon^{1/2}}{4d^{1/2}\lambda^{1/2}}$, it follows that $\lambda\, \mathbb{E}_{t\sim[T]}\, \mathbb{E}_{\mathbf{y}_t} \left\|\mathbf{y}_t - \overline{\mathbf{x}}_t\right\|^2 \le \varepsilon$.

Next consider the first term (the L_1-norm of the averaged gradients). Plugging the regret bound (2) into Lemma 17, and doing similar manipulations as the proof of Theorem 11, we get

$$\begin{aligned}\mathop{\mathbb{E}}_{t\sim[T]} \left\|\mathop{\mathbb{E}}_{\mathbf{y}_t} \nabla F(\mathbf{y}_t)\right\|_1 &\le \frac{\Delta}{DT} + \frac{\|6\boldsymbol{G} + 5\boldsymbol{\sigma}\|_1}{(1-\beta)T} + \|4\boldsymbol{G} + 5\boldsymbol{\sigma}\|_1 \sqrt{1-\beta} \\ &= \frac{4\Delta d^{1/2}\lambda^{1/2}}{(1-\beta)\varepsilon^{1/2}T} + \frac{\|6\boldsymbol{G} + 5\boldsymbol{\sigma}\|_1}{(1-\beta)T} + \|4\boldsymbol{G} + 5\boldsymbol{\sigma}\|_1 \sqrt{1-\beta}\,,\end{aligned}$$

where the last line follows since $D = \frac{(1-\beta)\varepsilon^{1/2}}{4d^{1/2}\lambda^{1/2}}$. Choosing $\beta = 1 - (\frac{\varepsilon}{10C})^2$, the last term is bounded by $\frac{\|\boldsymbol{G}+\boldsymbol{\sigma}\|_1}{2C} \cdot \varepsilon$. Moreover, choosing $T = (1-\beta)^{-1} \cdot \max\left\{4\Delta d^{1/2}\lambda^{1/2}\varepsilon^{-3/2},\ 12C\varepsilon^{-1}\right\}$, the first and second terms are bounded by ε and $\frac{\|\boldsymbol{G}+\boldsymbol{\sigma}\|_1}{2C} \cdot \varepsilon$, respectively. This concludes the proof.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions regarding the analysis of Adam with EMA for nonconvex and nonsmooth optimization.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discuss the limitations of the work in Section 7.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The assumptions and the proofs are provided both in the main text and the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: This paper is a theory paper and does not have any experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: This paper is a theory paper and does not have any experiments.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.

- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: This paper is a theory paper and does not have any experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: This paper is a theory paper and does not have any experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: This paper is a theory paper and does not have any experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research adheres to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This paper is a theory paper, and we do not forsee any direct societal implications arising from the research.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper is a theory paper and does not have any experiments.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper accurately credits the existing results by citing the papers that inspired the techniques used in this work.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not introduce any new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.