STORM^{+}: Fully Adaptive SGD with Momentum for Nonconvex Optimization

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this work we investigate stochastic non-convex optimization problems where the objective is an expectation over smooth loss functions, and the goal is to find an approximate stationary point. The most popular approach to handling such problems is variance reduction techniques, which are also known to obtain tight convergence rates, matching the lower bounds in this case. Nevertheless, these techniques require a careful maintenance of anchor points in conjunction with appropriately selected "mega-batchsizes". This leads to a challenging hyperparameter tuning problem, that weakens their practicality.
Recently, [Cutkosky & Orabona, 2019] have shown that one can employ recursive momentum in order to avoid the use of anchor points and large batchsizes, and still obtain the optimal rate for this setting. Yet, their method called STORM crucially relies on the knowledge of the smoothness, as well a bound on the gradient norms.
In this work we propose STORM^{+}, a new method that is completely parameter-free, does not require large batch-sizes, and obtains the optimal $O(1/T^{1/3})$ rate for finding an approximate stationary point. Our work builds on the STORM algorithm, in conjunction with a novel approach to adaptively set the learning rate and momentum parameters.

1 Introduction

Over the past decade non-convex models have become principal tools in ML (Machine Learning), and in data-science. This predominantly includes deep models, as well as Phase Retrieval [3], Non-negative matrix factorization [11], and Matrix completion problems [7] among others.

The main workhorse for training ML models is SGD (stochastic gradient descent) and its numerous variants. One parameter that significantly affects the SGD performance is the learning rate, which often requires a careful and costly hyper-parameter tuning. Adaptive approaches to setting the learning rate like AdaGrad [5] and Adam [13] as well as non-adaptive heuristics [17, 10] are very popular in modern ML applications, yet these methods also require some tuning of hyper-parameters like momentum and the scale of the learning rate schedule.

A popular SGD heuristic that has proven to be crucial in many applications is the use of *momentum*, i.e., the use of a weighted average of past gradients instead of the current gradient [22, 13]. While adaptive approaches to setting the momentum have been investigated in the past [21, 9]. Principled and theoretically grounded approaches to doing so are less investigated. Another aspect that has not been extensively investigated is the interplay between learning rate and momentum.

In this work we explore momentum-based adaptive and parameter-free methods for stochastic non-convex optimization problems. Concretely, we focus on the setting where the objective is an *expectation over smooth losses*, and the goal is to find an approximate stationary point.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

In the general case of smooth non-convex objectives it is known that one can approach a stationary point at a rate of $O(1/T^{1/4})$, where T is the total number of samples [8]. While this rate is optimal in the general case, it is known that one can obtain an improved rate of $O(1/T^{1/3})$ if the objective is an *expectation over smooth losses*, [6, 28, 4, 24]. And this rate was recently shown to be tight [2].

Nevertheless, most of the methods for this setting rely on variance reduction techniques [12, 27, 18, 25], which require a careful maintenance of anchor points in conjunction with appropriately selected huge-batchsizes. This leads to a challenging hyper-parameter tuning problem, that weakens their practicality. One exception is the recent STORM algorithm of [4]. STORM does not require large batches nor anchor points; instead it uses a corrected momentum based gradient update that leads to implicit variance reduction, which in turn facilitates fast convergence. Unfortunately, none of the aforementioned methods (including STORM) is parameter-free. And indeed the knowledge of the smoothness together with either the noise variance or a bound on the norm the gradients is crucial to establish their guarantees.

Our contribution are as follows,

- We present STORM$^+$, a *parameter-free* momentum based method that ensures the optimal $O(1/T^{1/3})$ rate for the *expectation over smooth losses* setting. Similarly to STORM , our method does not require large-batches nor anchor points.
- STORM$^+$ implicitly adapts to the variance of the gradients. Concretely, it obtains a rate of $O(1/\sqrt{T} + \sigma^{1/3}/T^{1/3})$, which recovers the optimal $O(1/\sqrt{T})$ rate in the noiseless case. We also improve over STORM by shaving off a $(\log T)^{3/4}$ factor from the $1/T$ term.
- In STORM$^+$ we demonstrate a novel way to set the learning rate and to adaptively tie between the learning rate and momentum parameters.

2 Related Work

In the context of stochastic non-convex optimization with general smooth losses, it was shown in [8] that SGD with an appropriately selected learning rate can obtain a rate of $O(1/T^{1/4})$ for finding an approximate stationary point. And this was recently shown to be tight [2]. While the method of [8] requires a knowledge of the smoothness and variance parameters, recent works have shown that adaptive methods like AdaGrad are able to obtain this bound in a parameter free manner, as well as to adapt to the variance of the problem [16, 26, 20]. These results, in a sense, explain the success of adaptive[1] methods like AdGrad [5], Adam [13], and RMSProp [23] in handling non-convex problems.

The idea of using variance reduction techniques for non-convex problems was first suggested in the context of finite sum problems by [1, 19], showing a rate of $O(1/T^{1/4})$. This was later improved by [14] to a rate of $O(1/T^{3/10})$. The first works that have obtained the optimal $O(1/T^{1/3})$ for this setting were [6, 28]. And in [6] it is also shown how it applies more generally to the *expectation over smooth losses* setting — a setting that captures finite-sum problems as a private case.

The STORM algorithm suggested in [4], is the first algorithm to obtain the optimal $O(1/T^{1/3})$ for this setting, without the need to maintain anchor points or large batches. Instead it relies on a clever correction of the momentum term that leads to an implicit variance reduction effect. Moreover, STORM adapts to the variance of the problem by obtaining a rate of $O((\log T)^{3/4}/\sqrt{T}+\sigma^{1/3}/T^{1/3})$ without the prior knowledge of variance, but it does require the smoothness parameter and a bound on the gradient norms. Simultaneously to the work of [4], another paper [24] have obtained the same optimal bound by proposing a similar update rule. Note that [24] does calculate a single anchor point, and requires the knowledge of the smoothness and variance parameters.

3 Setting and Preliminaries

We discuss stochastic non-convex optimization problems where the objective $f : \mathbb{R}^d \mapsto \mathbb{R}$ is of the following form,

$$f(x) := \mathbf{E}_{\xi\sim\mathcal{D}}[f(x;\xi)] \ ,$$

[1] An adaptive method is a method that updates its learning rate according to the (noisy) gradient feedback that it receives throughout the training process.

and $\mathcal{D}$ is an unknown distribution from which we may draw i.i.d. samples. Our goal is to find an approximate stationary point of f, i.e. after T draws from $\mathcal{D}$ we should output a point $\bar{x} \in \mathbb{R}^d$ such that $\mathbf{E}\|\nabla f(x)\| \leq \text{Poly}(1/T)$.

We focus on first order methods, i.e., methods that may access the gradients of $f(\cdot, \xi)$, and make the following assumptions regarding the noisy gradients and function values.
Bounded values: There exists $B > 0$ such, $\max_{x,y \in \mathbb{R}^d} |f(x) - f(y)| \leq B$.
Bounded gradients: There exists $G > 0$ such, $\|\nabla f(x;\xi)\|^2 \leq G^2$; $\forall x \in \mathbb{R}^d, \xi \in \textbf{support}\{\mathcal{D}\}$.
Bounded Variance: There exists $\sigma > 0$ such $\mathbf{E}\|\nabla f(x;\xi) - \nabla f(x)\|^2 \leq \sigma^2$; $\forall x \in \mathbb{R}^d$.
Expectation over Smooth Losses: There exists $L > 0$ such,

$$\|\nabla f(x;\xi) - \nabla f(y;\xi)\| \leq L\|x - y\| \; ; \quad \forall x, y \in \mathbb{R}^d, \xi \in \textbf{support}\{\mathcal{D}\} \; .$$

The last assumption also implies that the expected loss $f(\cdot)$ is L smooth. A property of smooth functions that we will exploit throughout the paper is the following,

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + (L/2)\|y - x\|^2 \; ; \quad \forall x, y \in \mathbb{R}^d$$

Notation: $\nabla f(x;\xi)$ relates to gradients with respect to x, i.e., $\nabla := \nabla_x$. We use $\|\cdot\|$ to denote the Euclidean norm, and we denote by x^* a global minima of $f(\cdot)$.

4 Method

In this section we present STORM^+ (STochastic Recursive Momentum $+$): a parameter-free stochastic optimization method that finds approximate stationary points at an optimal rate. We describe our method in Alg. 1 and Eq. (4), and state its guarantees in Theorem 1.

The STORM **algorithm:** The original STORM template of [4] relies on an SGD-style update with a corrected momentum. Concretely, the idea is to maintain a gradient estimate d_t which is a *corrected* weighted average of past stochastic gradients, and then update the iterates similarly to SGD,

$$x_{t+1} = x_t - \eta_t d_t \; . \tag{1}$$

Standard momentum is a weighted average of past gradients,

$$d_t = a_t \nabla f(x_t, \xi_t) + (1 - a_t) d_{t-1} \; ; \;\; \text{where } a_t \in [0, 1] \; .$$

And the above d_t is generally a biased estimate of $\nabla f(x_t)$.

In STORM it is suggested to add a correction term ,$(1 - a_t)(\nabla f(x_t, \xi_t) - \nabla f(x_{t-1}, \xi_t))$, which leads to the following update rule (again, $a_t \in [0, 1]$),

$$d_t = \nabla f(x_t, \xi_t) + (1 - a_t)(d_{t-1} - \nabla f(x_{t-1}, \xi_t)) \; ; \quad \textbf{(STORM's corrected Momentum)} \tag{2}$$

The correction term plays a crucial role here: it exploits the smoothness of $f(\cdot, \xi)$ in a way that leads to a variance reduction effect.

To see this effect one can inspect the error of the momentum d_t compared to the exact gradient at x_t,

$$\epsilon_t := d_t - \nabla f(x_t) \; .$$

The STORM update rule induces the following error dynamics,

$$\epsilon_t = (1 - a_t)\epsilon_{t-1} + a_t(\nabla f(x_t, \xi_t) - \nabla f(x_t)) + (1 - a_t) Z_t$$

where $Z_t := (\nabla f(x_t, \xi_t) - \nabla f(x_{t-1}, \xi_t)) - (\nabla f(x_t) - \nabla f(x_{t-1}))$. Due to the smoothness of the objective we have $\|Z_t\| \leq O(\|x_t - x_{t-1}\|) = O(\eta_{t-1}\|d_{t-1}\|)$. Intuitively, as we approach a stationary point (and use a small enough learning rate) then $\eta_{t-1}\|d_{t-1}\|$ decreases which in turn reduces the Z_t's. Moreover, the second term in the above dynamics, $a_t(\nabla f(x_t, \xi_t) - \nabla f(x_t))$, can be controlled by choosing a small enough momentum a_t. Thus, carefully controlling the learning rate and momentum parameters leads to a variance reduction effect which facilitates fast convergence.

The original STORM paper [4] makes the following choices,

$$\eta_t = \theta / \left(w + \sum_{i=1}^{t} \|g_i\|^2 \right)^{1/3} , \quad \& \quad a_t = c L^2 \eta_{t-1}^2 \; , \tag{3}$$

Algorithm 1 STORM$^+$

Input: #iterations T , $x_1 \in \mathbb{R}^d$

Sample ξ_1 and set $d_1 = g_1 = \nabla f(x_1, \xi_1)$

for $t = 1, ..., T$ **do**

$\quad a_{t+1} \leftarrow 1/\left(1 + \sum_{i=1}^{t} \|g_i\|^2\right)^{2/3} \quad \& \quad \eta_t \leftarrow 1/\left(\sum_{i=1}^{t} \|d_i\|^2/a_{i+1}\right)^{1/3}$

$\quad x_{t+1} \leftarrow x_t - \eta_t d_t$

$\quad$ Sample ξ_{t+1} and set $g_{t+1} := \nabla f(x_{t+1}; \xi_{t+1})$, and $\tilde{g}_t := \nabla f(x_t; \xi_{t+1})$

$\quad d_{t+1} \leftarrow g_{t+1} + (1 - a_{t+1})(d_t - \tilde{g}_t)$

end for

Choose $\bar{x}_T$ uniformly at random from $\{x_1, \ldots, x_T\}$

return $\bar{x}_T$

where we denote $g_t := \nabla f(x_t, \xi_t)$. The above choice of learning rate is inspired by AdaGrad [5], which also sets the learning rate inversely proportional to the cumulative square norms of past gradients. Note that θ and w are constants that depend on the smoothness of the objective L, as well as on the bound on the gradients G, and c is an absolute constant independent of the problem's characteristics. These choices of the constants and especially the choice of $a_t \propto L^2 \eta_{t-1}^2$ is crucial for the analysis of the original STORM . And actually the convergence proof for STORM breaks down unless we encode this prior knowledge into η_t and a_t.

Next we describe a parameter free choice.

Our STORM$^+$ **algorithm:** STORM$^+$ relies on the original STORM template described in Equations (1),(2), with the following parameter-free choices of learning rate and momentum parameter,

$$\eta_t = 1/\left(\sum_{i=1}^{t} \|d_i\|^2/a_{i+1}\right)^{1/3}, \quad \& \quad a_t = 1/\left(1 + \sum_{i=1}^{t-1} \|g_i\|^2\right)^{2/3}, \tag{4}$$

where again we denote $g_t := \nabla f(x_t, \xi_t)$. Note that in contrast to the original STORM our adaptive learning rate builds on past momentums $\{d_1, \ldots, d_t\}$'s as well as on the momentum parameters $\{a_1, \ldots, a_{t+1}\}$. Our momentum term is similar to the adaptive choice of STORM , yet it does not require a bound on the gradients nor on the smoothness parameter, which is crucial for the original analysis. Finally, note that the above choice ensures $a_t \in [0, 1]$.

For completeness we present our method in Alg. 1, where it can be seen that STORM$^+$ is a combination of the original STORM template (Equations (1),(2)) together with the specific choices of η_t, a_t appearing in Eq. (4). Note that the solution that STORM$^+$ outputs is a point chosen uniformly at random among all iterates. This latter is quite standard in non-convex optimization.

Notation: In Alg. 1 and throughout the rest of the paper we will employ the following notation,

$$g_t := \nabla f(x_t, \xi_t)\ ; \ \ \tilde{g}_t := \nabla f(x_t, \xi_{t+1})\ ; \ \ \bar{g}_t := \nabla f(x_t)\ .$$

Next we present our main theorem regarding STORM$^+$ (Alg. 1):

Theorem 1. *Under the conditions detailed in Section 3,* STORM$^+$ *ensures,*

$$\mathbf{E}\|\nabla f(\bar{x}_T)\| \leq O\left(\frac{M}{\sqrt{T}} + \frac{\kappa\sigma^{1/3}}{T^{1/3}}\right)\ ,$$

where $\kappa = O(B^{3/4} + L^{3/2})$*;* $M = O(1 + L^{9/4} + B^{9/8} + G^5 + (LG^4)^{3/2})$*, and the expectation is with respect to the randomization of the samples as well as the algorithm's.*

Theorem 1 demosntrates that in the stochastic case STORM$^+$ achieves the optimal $O(1/T^{1/3})$ rate for our setting. Moreover, it can be seen that STORM$^+$ implicitly adapts to the variance of the noise. Concretely, in the noiseless case where $\sigma = 0$, STORM$^+$ recovers the optimal $O(1/\sqrt{T})$ rate.

We note that by scaling the learning rate by a constant factor may enable to obtain better dependence on L and B.

5 Analysis

In this section we provide an analysis of the STORM^+ algorithm. In Section 5.1 we provide an analysis for the offline case where $\sigma = 0$, and establish a convergence rate of $O(1/\sqrt{T})$. For simplicity and due to space limitations we do not provide the full proof of the stochastic case. Instead, in Section 5.2 we fully analyze a simplified version of STORM^+ where we use $a_{t+1} := 1/t^{2/3}$ instead of the adaptive momentum term that appears in Alg. 1. We show that this version achieves a convergence rate of $O(1/T^{1/3})$ in the stochastic case (though it does not adapt to the variance). Then in 5.1 we provide a proof sketch for STORM^+ that establishes Theorem 1.

5.1 Offline Case

Here we analyze STORM^+ in the case where $\sigma = 0$, and demonstrate a rate of $O(1/\sqrt{T})$ for finding an approximate stationary point.

Proof. In the case where $\sigma = 0$ one can directly show by induction that $d_t = \bar{g}_t := \nabla f(x_t)$. So the update rule becomes $x_{t+1} = x_t - \eta_t \bar{g}_t$. Now, using the smoothness of the objective implies,

$$\Delta_{t+1} - \Delta_t = f(x_{t+1}) - f(x_t) \leq -\eta_t \|\bar{g}_t\|^2 + L\eta_t^2 \|\bar{g}_t\|^2/2 \,,$$

here we denoted $\Delta_t := f(x_t) - f(x^*)$, where $x^* \in \arg\min f(x)$. Dividing by η_t, re-arranging and summing gives,

$$\begin{aligned}
\sum_{t=1}^{T} \|\bar{g}_t\|^2 &\leq -\frac{1}{\eta_T}\Delta_{t+1} + \sum_{t=1}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\Delta_t + \frac{L}{2}\sum_{t=1}^{T}\eta_t\|\bar{g}_t\|^2 \\
&\leq B\sum_{t=1}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + \frac{L}{2}\sum_{t=1}^{T}\frac{\|\bar{g}_t\|^2}{\left(\sum_{i=1}^{t}\|\bar{g}_i\|^2\right)^{1/3}} \\
&\leq \frac{B}{\eta_T} + L\left(\sum_{i=1}^{t}\|\bar{g}_i\|^2\right)^{2/3} \leq B\left(\sum_{t=1}^{T}\|\bar{g}_t\|^2/a_{t+1}\right)^{1/3} + L\left(\sum_{i=1}^{t}\|\bar{g}_i\|^2\right)^{2/3} \\
&\leq B\left(1 + \sum_{t=1}^{T}\|\bar{g}_t\|^2\right)^{2/9}\left(\sum_{t=1}^{T}\|\bar{g}_t\|^2\right)^{1/3} + L\left(\sum_{i=1}^{t}\|\bar{g}_i\|^2\right)^{2/3} \qquad (5)
\end{aligned}$$

where the second inequality uses $\eta_t = \left(\sum_{i=1}^{t}\|\bar{g}_i\|^2/a_{i+1}\right)^{-1/3} \leq \left(\sum_{i=1}^{t}\|\bar{g}_i\|^2\right)^{-1/3}$ which holds since $d_t = \bar{g}_t$ and $a_t \leq 1$, we also use a $\Delta_t \in [0, B]$ together with $\eta_t^{-1} - \eta_{t-1}^{-1} \geq 0$, and finally we use the notation $1/\eta_0 := 0$. The third inequality uses Lemma 2 below; and the last inequality uses $1/a_{t+1} \leq (1/a_{T+1}) = \left(1 + \sum_{t=1}^{T}\|\bar{g}_t\|^2\right)^{2/3}$, which holds since a_t is monotonically non-increasing.

From Eq. (5) one can derive the following bound, $\sum_{t=1}^{T}\|\bar{g}_t\|^2 \leq O(1 + L^3 + B^{9/4})$. Using the definition of $\bar{x}_T$ as well as Jensen's inequality implies,

$$\mathbf{E}\|\nabla f(\bar{x}_T)\| := \mathbf{E}\|\bar{g}(\bar{x}_T)\| \leq \sqrt{\mathbf{E}\|\bar{g}(\bar{x}_T)\|^2} = \sqrt{\sum_{t=1}^{T}\|\bar{g}_t\|^2/T} \leq O(\sqrt{1 + L^3 + B^{9/4}}/\sqrt{T}) \,.$$

which establishes the bound. In the proof we have used the technical lemma below,

Lemma 2. *Let $b_1, ..., b_n \geq 0$, and $p \in (0, 1)$ be a rational number. Then,*

$$\sum_{i=1}^{n}\frac{b_i}{\left(\sum_{j=1}^{i} b_j\right)^p} \leq \frac{1}{1-p}\left(\sum_{i=1}^{n} b_i\right)^{1-p}$$

□

5.2 Stochastic Case Analysis of Simplified STORM^+

Here we analyze a simplified version of STORM^+ in the stochatic setting. While this version does not adapt to the noise variance, it exhibits the optimal rate of $O(1/T^{1/3})$ in the stochastic case, and its analysis illustrates some of the main ideas that we employ in the proof of the fully adaptive STORM^+ (which is more involved).

The version that we analyze here differs from STORM^+ in the choice of the momentum parameters. Here we choose $a_1 = 1$ and $a_{t+1} = 1/t^{2/3}$; $\forall t \geq 1$, in contracts to the adaptive choice that we make in Alg. 1. Note that we keep picking $\eta_t = 1/\left(\sum_{i=1}^{t} \|d_i\|^2/a_{i+1}\right)^{1/3}$.

Proof. The proof is composed of two parts. In the first we bound the cumulative expectation of errors $\mathbf{E}\sum_{t=1}^{T}\|\epsilon_t\|^2$, where ϵ_t is the difference between the corrected momentum d_t and the exact gradient $\bar{g}_t$, i.e. $\epsilon_t = d_t - \bar{g}_t$. Thus, in the first part we relate the above sum to the sum of exact gradients $\mathbf{E}\sum_{t=1}^{T}\|\bar{g}_t\|^2$. Then, in the second part we divide into two sub-cases the first where $\mathbf{E}\sum_{t=1}^{T}\|\epsilon_t\|^2 \leq (1/2)\mathbf{E}\sum_{t=1}^{T}\|\bar{g}_t\|^2$ and its complement. In one of these sub-cases we also use the smoothness of the objective together with the update rule, similarly to what we do in Eq. (5).

First Part: Bounding $\mathbf{E}\sum_{t=1}^{T}\|\epsilon_t\|^2$. The update rule for d_t induces the following error dynamics,

$$\epsilon_t = (1-a_t)\epsilon_{t-1} + a_t(g_t - \bar{g}_t) + (1-a_t)Z_t \tag{6}$$

where $Z_t := (g_t - \tilde{g}_{t-1}) - (\bar{g}_t - \bar{g}_{t-1})$. Letting $\mathcal{H}_t$ be the history to time t, i.e., $\mathcal{H}_t := \{x_1, \xi_1, \xi_2, \xi_3 \ldots, \xi_t\}$ and recalling that $a_t = a_t(x_1, \xi_1, \ldots, \xi_{t-1})$, and $x_t = x_t(x_1, \xi_1, \ldots, \xi_{t-1})$ immediately implies that $\mathbf{E}[a_t(g_t - \bar{g}_t)|\mathcal{H}_{t-1}] = \mathbf{E}[(1-a_t)Z_t|\mathcal{H}_{t-1}] = 0$, as well as $\mathbf{E}[(1-a_t)\epsilon_{t-1}|\mathcal{H}_{t-1}] = (1-a_t)\epsilon_{t-1}$.

Thus, taking the square of the above equation and then taking the expectation gives,

$$\begin{aligned}
\mathbf{E}\|\epsilon_t\|^2 &\leq (1-a_t)^2\mathbf{E}\|\epsilon_{t-1}\|^2 + \|(1-a_t)Z_t + a_t(g_t - \bar{g}_t)\|^2 \\
&\leq (1-a_t)^2\mathbf{E}\|\epsilon_{t-1}\|^2 + 2(1-a_t)^2\|Z_t\|^2 + 2a_t^2\mathbf{E}\|g_t - \bar{g}_t\|^2 \\
&\leq (1-a_t)\mathbf{E}\|\epsilon_{t-1}\|^2 + 8L^2\mathbf{E}\eta_{t-1}^2\|d_{t-1}\|^2 + 2a_t^2\sigma^2 \ ,
\end{aligned} \tag{7}$$

where the second line uses $\|b+c\|^2 \leq 2\|b\|^2 + 2\|c\|^2$, and the last line uses $\mathbf{E}\|g_t - \bar{g}_t\|^2 \leq \sigma^2$, and $(1-a_t) \in [0,1]$, as well as the smoothness assumption that implies $\|Z_t\| \leq \|g_t - \tilde{g}_{t-1}\| + \|\bar{g}_t - \bar{g}_{t-1}\| \leq 2L\|x_t - x_{t-1}\| = 2L\eta_{t-1}\|d_{t-1}\|$.

Dividing Eq. (7) by a_t and re-arranging implies,

$$\mathbf{E}\|\epsilon_{t-1}\|^2 \leq \frac{1}{a_t}(\mathbf{E}\|\epsilon_{t-1}\|^2 - \mathbf{E}\|\epsilon_t\|^2) + 8L^2\mathbf{E}[\eta_{t-1}^2\|d_{t-1}\|^2/a_t] + 2a_t\sigma^2 \ .$$

Summing the above, and using $\epsilon_0 := 0$ gives,

$$\mathbf{E}\sum_{t=1}^{T}\|\epsilon_{t-1}\|^2 \leq \underbrace{-\frac{\mathbf{E}\|\epsilon_T\|^2}{a_T}}_{\text{(A)}} + \underbrace{\sum_{t=1}^{T-1}\left(\frac{1}{a_{t+1}} - \frac{1}{a_t}\right)\mathbf{E}\|\epsilon_t\|^2}_{\text{(B)}} + 8L^2\underbrace{\mathbf{E}[\sum_{t=1}^{T}\eta_{t-1}^2\|d_{t-1}\|^2/a_t]}_{\text{(C)}} + 2\sigma^2\underbrace{\sum_{t=1}^{T}a_t}_{\text{(D)}} \tag{8}$$

Next we bound all the term on the RHS of the above equation:

Bounding (A)**:** Since $a_T \leq 1$ we can bound $-\mathbf{E}\|\epsilon_T\|^2/a_T \leq -\mathbf{E}\|\epsilon_T\|^2$

Bounding (B)**:** Note that $G(z) = z^{2/3}$ is a concave function in $\mathbb{R}_+$. Thus applying the gradient inequality implies that $\forall z_1, z_2 \geq 0$ we have $(z_1+z_2)^{2/3} - z_1^{2/3} \leq \frac{2}{3}z_1^{-1/3}z_2$. Therefore $\forall t \geq 2$,

$$1/a_{t+1} - 1/a_t = t^{2/3} - (t-1)^{2/3} \leq 2(t-1)^{-1/3}/3 \leq 2/3 \ .$$

Moreover, $1/a_2 - 1/a_1 = 0$. These imply that (B) $\leq (2/3)\mathbf{E}\sum_{t=1}^{T}\|\epsilon_t\|^2$.

Bounding (C)**:** By definition of η_t we have,

$$(\mathrm{C}) = \mathbf{E}\sum_{t=1}^{T} \frac{\|d_{t-1}\|^2/a_t}{\left(\sum_{i=1}^{t-1}\|d_i\|^2/a_{i+1}\right)^{2/3}} \leq 3\mathbf{E}\left(\sum_{t=1}^{T-1}\|d_t\|^2/a_{t+1}\right)^{1/3} \leq 3T^{2/9}\left(\mathbf{E}\sum_{t=1}^{T}\|d_t\|^2\right)^{1/3} .$$

where the first inequality uses Lemma 2, and the second inequality uses $1/a_t \leq 1/a_{T+1} \leq T^{2/3}$ as well as Jensen's inequality with respect to the concave function $U(z) = z^{1/3}$, defined over $\mathbb{R}_+$.

Bounding (D)**:** Lemma 2 immediately imply that $(\mathrm{D}) = 1 + \sum_{t=1}^{T-1} 1/t^{2/3} \leq 1 + 3T^{1/3} \leq 4T^{1/3}$. Plugging these bounds into Eq. (8) and re-arranging yields,

$$\mathbf{E}\sum_{t=1}^{T}\|\epsilon_t\|^2 \leq 72L^2T^{2/9}\left(\mathbf{E}\sum_{t=1}^{T}\|d_t\|^2\right)^{1/3} + 24\sigma^2T^{1/3} . \tag{9}$$

Second Part: Bounding $\mathbf{E}\sum_{t=1}^{T}\|\bar{g}_t\|^2$. Here we use the bound of Eq. (9) in order to bound the sum of square gradients. Let us divide into two sub-cases.

Case 1 : Assume that $\mathbf{E}\sum_{t=1}^{T}\|\epsilon_t\|^2 \geq \frac{1}{2}\mathbf{E}\sum_{t=1}^{T}\|\bar{g}_t\|^2$. Combining the condition of Case 1 with $\|d_t\|^2 \leq 2\|\bar{g}_t\|^2 + 2\|\epsilon_t\|^2$ (due to $d_t = \bar{g}_t + \epsilon_t$), implies that $\mathbf{E}\sum_{t=1}^{T}\|d_t\|^2 \leq 6\mathbf{E}\sum_{t=1}^{T}\|\epsilon_t\|^2$. Plugging this inside Eq. (9) yields,

$$\mathbf{E}\sum_{t=1}^{T}\|\epsilon_t\|^2 \leq 72L^2T^{2/9}\left(6\mathbf{E}\sum_{t=1}^{T}\|\epsilon_t\|^2\right)^{1/3} + 24\sigma^2T^{1/3} .$$

The above immediately implies that $\mathbf{E}\sum_{t=1}^{T}\|\epsilon_t\|^2 \leq O((L^3+\sigma^2)T^{1/3})$, and due to the condition of Case 1 we therefore have, $\mathbf{E}\sum_{t=1}^{T}\|\bar{g}_t\|^2 \leq O((L^3+\sigma^2)T^{1/3})$. This concludes the first case.

Case 2 : Assume that $\mathbf{E}\sum_{t=1}^{T}\|\epsilon_t\|^2 \leq \frac{1}{2}\mathbf{E}\sum_{t=1}^{T}\|\bar{g}_t\|^2$. Combining the condition of Case 2 with $\|d_t\|^2 \leq 2\|\bar{g}_t\|^2 + 2\|\epsilon_t\|^2$ (due to $d_t = \bar{g}_t + \epsilon_t$), implies that $\mathbf{E}\sum_{t=1}^{T}\|d_t\|^2 \leq 3\mathbf{E}\sum_{t=1}^{T}\|\bar{g}_t\|^2$. Now using the update rule $x_{t+1} = x_t - \eta_t d_t$ together with smoothness of $f(\cdot)$, one can show in a similar manner to our derivation of Eq. (5) the following bound (we defer this to the appendix),

$$\sum_{t=1}^{T}\|\bar{g}_t\|^2 \leq \sum_{t=1}^{T}\|\epsilon_t\|^2 + 2BT^{2/9}\left(\sum_{t=1}^{T}\|d_t\|^2\right)^{1/3} + \frac{3}{2}L\left(\sum_{t=1}^{T}\|d_t\|^2\right)^{2/3} \tag{10}$$

Taking the expectation of the above equation and plugging in $\mathbf{E}\sum_{t=1}^{T}\|d_t\|^2 \leq 3\mathbf{E}\sum_{t=1}^{T}\|\bar{g}_t\|^2$ as well as $\mathbf{E}\sum_{t=1}^{T}\|\epsilon_t\|^2 \leq \frac{1}{2}\mathbf{E}\sum_{t=1}^{T}\|\bar{g}_t\|^2$ gives,

$$\mathbf{E}\sum_{t=1}^{T}\|\bar{g}_t\|^2 \leq \frac{1}{2}\mathbf{E}\sum_{t=1}^{T}\|\bar{g}_t\|^2 + 2BT^{2/9}\left(3\mathbf{E}\sum_{t=1}^{T}\|\bar{g}_t\|^2\right)^{1/3} + \frac{3}{2}L\left(3\mathbf{E}\sum_{t=1}^{T}\|\bar{g}_t\|^2\right)^{2/3} \tag{11}$$

where we also used Jensen's inequality with respect to he concave functions $z^{1/3}$ and $z^{2/3}$ defined over $\mathbb{R}_+$. The above immediately implies, $\mathbf{E}\sum_{t=1}^{T}\|\bar{g}_t\|^2 \leq O(L^3 + B^{3/2}T^{1/3})$. This concludes the second case.

Summary. We have shown that $\mathbf{E}\sum_{t=1}^{T}\|\bar{g}_t\|^2 \leq O((L^3+\sigma^2+B^{3/2})T^{1/3})$, combining this with the definition of $\bar{x}_T$ and using Jensen's inequality similarly to what we did in the offline analysis provides,

$$\mathbf{E}\|\nabla f(\bar{x}_T)\| = O(\sqrt{L^3+\sigma^2+B^{3/2}}/T^{1/3}) ,$$

which concludes the proof. □

5.3 Stochastic Case Analysis of STORM^+

Here we provide a sketch of the proof for the STORM^+ algorithm appearing in Alg. 1. In a high level the analysis follows similar lines to the analysis of simplified STORM^+ appearing in Section 5.2.

There are two extra challenges compared to the analysis of simplified STORM^+ : **(i)** Now a_t is a random variable that depends on the noisy samples. **(ii)** The differences $1/a_{t+1} - 1/a_t$ are not necessarily smaller than 1. Recall that in the analysis appearing in Section 5.2 we used $1/a_{t+1} - 1/a_t \leq 2/3$, which was crucial to bounding term (B).

Among the tools that we use to address the first challenge is a version of Young's inequality, that we mention in the apppendix. To cope with the second challenge, when we bound the expectation of $\sum_{t=1}^{T} \|\epsilon_t\|^2$, it is split into two,

$$\sum_{t=1}^{T} \|\epsilon_t\|^2 = \sum_{t=1}^{\tau^*} \|\epsilon_t\|^2 + \sum_{t=\tau^*+1}^{T} \|\epsilon_t\|^2$$

where τ^* is a time-step after which we can ensure that $1/a_{t+1} - 1/a_t \leq 2/3$.

Next we proceed with the proof sketch.

Proof Sketch of Theorem 1. The proof is composed of three parts: **(a)** In the first part we bound the cumulative expectation of errors $\mathbf{E}\sum_{t=1}^{\tau^*} \|\epsilon_t\|^2$, where $\epsilon_t := d_t - \bar{g}_t$, and τ^* is a stopping time after which we can ensure that $1/a_{t+1} - 1/a_t \leq 2/3$. **(b)** In the second part we use our bound on $\mathbf{E}\sum_{t=1}^{\tau^*} \|\epsilon_t\|^2$ in order to bound the total sum of square errors, $\mathbf{E}\sum_{t=1}^{T} \|\epsilon_t\|^2$. **(c)** Then, in the last part we divide into two sub-cases the first where $\mathbf{E}\sum_{t=1}^{T} \|\epsilon_t\|^2 \leq (1/2)\mathbf{E}\sum_{t=1}^{T} \|\bar{g}_t\|^2$ and its complement. In one of these sub-cases we also use the smoothness of the objective together with the update rule, similarly to what we do in Eq. (5).

First Part: Bounding $\mathbf{E}\sum_{t=1}^{\tau^*} \|\epsilon_t\|^2$. Recall the error dynamics of STORM^+ appearing in Eq. (6). Taking the square and summing up to some $\tau^* \in [T]$ enables to bound,

$$\sum_{t=1}^{\tau^*} \|\epsilon_t\|^2 \leq \sum_{t=1}^{\tau^*} (1-a_t)\|\epsilon_{t-1}\|^2 + 2\sum_{t=1}^{\tau^*} \|Z_t\|^2 + 2\sum_{t=1}^{\tau^*} a_t^2 \|g_t - \bar{g}_t\|^2 + \sum_{t=1}^{\tau^*} M_t \ ,$$

where $\{M_t\}_{t\in[T]}$ is a martingale difference sequence such $\mathbf{E}[M_t|\mathcal{H}_{t-1}] = 0$, where $\mathcal{H}_t$ is the history to time t, i.e., $\mathcal{H}_t := \{x_1, \xi_1, \xi_2, \xi_3 \ldots, \xi_t\}$; also recall that $Z_t := (g_t - \tilde{g}_{t-1}) - (\bar{g}_t - \bar{g}_{t-1})$.

Now let us define $\beta := \min\{1, 1/G^4\}$, and $\tau^* = \max\{t \in [T] : a_t \geq \beta\}$. Recalling that a_{t+1} is measurable with respect to $\mathcal{H}_t$ implies that $\tau^* \in [T]$ is a stopping time.

Re-arranging the above and using the definition of τ^* implies,

$$\beta \sum_{t=1}^{\tau^*} \|\epsilon_t\|^2 \leq \|\epsilon_{\tau^*}\|^2 + \sum_{t=1}^{\tau^*-1} a_{t+1}\|\epsilon_t\|^2 \leq 2\underbrace{\sum_{t=1}^{T} \|Z_t\|^2}_{\text{(i)}} + 2\underbrace{\sum_{t=1}^{T} a_t^2 \|g_t - \bar{g}_t\|^2}_{\text{(ii)}} + \underbrace{\sum_{t=1}^{\tau^*} M_t}_{\text{(iii)}}$$

where we used $\tau^* \leq T$, as well as $\beta \leq 1$. Next we bound the expected value of the above terms.
Bounding (i)**.** As in the previous section, the smoothness property implies that $\|Z_t\|^2 \leq 4L^2\eta_{t-1}^2\|d_{t-1}\|^2$. Using the expression for η_{t-1} together with Lemma 2 enables to show,

$$\text{(i)} \leq 4L^2 \sum_{t=1}^{T} \frac{\|d_{t-1}\|^2}{(\sum_{i=1}^{t-1} \|d_i\|^2)^{2/3}} \leq 12L^2 (\sum_{t=1}^{T} \|d_t\|^2)^{1/3} \ .$$

Bounding (ii)**.** One can directly show that $\mathbf{E}[a_t^2\|g_t - \bar{g}_t\|^2] \leq \mathbf{E}[a_t^2\|g_t\|^2]$. Using this together with the expression for a_t, it is possible to show that,

$$\mathbf{E}(\text{ii}) \leq \mathbf{E}\sum_{t=1}^{T} \frac{\|g_t\|^2}{(1+\sum_{i=1}^{t-1} \|g_i\|^2)^{4/3}} \leq C_1 \ .$$

where C_1 is a constant, and the second inequality is due to a lemma that we describe in the appendix.
Bounding (iii)**.** Since $\tau^* \in [T]$ is a bounded stopping time, and M_t is a martingale difference sequence, then Doob's optional stopping theorem [15] implies $\mathbf{E}(\text{iii}) = \mathbf{E}\sum_{t=1}^{\tau^*} M_t = 0$.

Conclusion. The above together with Jensen's inequality for $U(z) = z^{1/3}$ defined over $\mathbb{R}_+$, yields,

$$\mathbf{E}\sum_{t=1}^{\tau^*}\|\epsilon_t\|^2 \leq 2C_1/\beta + 24(L^2/\beta)(\mathbf{E}\sum_{t=1}^{T}\|d_t\|^2)^{1/3}\,. \tag{12}$$

Second Part: Bounding $\mathbf{E}\sum_{t=1}^{T}\|\epsilon_t\|^2$. Recall the error dynamics of STORM$^+$ appearing in Eq. (6). Dividing by $\sqrt{a_t}$, taking the square and summing up to some T enables to bound,

$$\frac{1}{a_t}\|\epsilon_t\|^2 \leq (\frac{1}{a_t} - 1)\|\epsilon_{t-1}\|^2 + 2\frac{\|Z_t\|^2}{a_t} + 2a_t\|g_t - \bar{g}_t\|^2 + Y_t$$

where $\{Y_t\}_{t\in[T]}$ is a martingale difference sequence such $\mathbf{E}[Y_t|\mathcal{H}_{t-1}] = 0$.

Re-arranging the above and summing one can show,

$$\sum_{t=1}^{T}\|\epsilon_{t-1}\|^2 \leq \underbrace{-\frac{1}{a_T}\|\epsilon_T\|^2}_{\text{(A)}} + \underbrace{\sum_{t=1}^{T}(\frac{1}{a_{t+1}} - \frac{1}{a_t})\|\epsilon_t\|^2}_{\text{(B)}} + \underbrace{2\sum_{t=1}^{T}\frac{\|Z_t\|^2}{a_t}}_{\text{(C)}} + \underbrace{2\sum_{t=1}^{T}a_t\|g_t - \bar{g}_t\|^2}_{\text{(D)}} + \underbrace{\sum_{t=1}^{T}Y_t}_{\text{(E)}}$$

Now, due to the martingale property $\mathbf{E}(\text{E}) = 0$. Next, we focus on bounding term (B),
Bounding (B)**.** Using the definition of τ^* one can show that $1/a_{t+1} \leq 1/\tilde{\beta}\,; \forall t \leq \tau^*$, where $1/\tilde{\beta} := (1/\beta^{3/2} + G^2)^{2/3}$. Moreover, we can show,

$$1/a_{t+1} - 1/a_t \leq 2/3\,; \quad \forall t \geq \tau^* + 1$$

This enables to decompose and bound (B) according to τ^*,

$$\begin{aligned}\sum_{t=1}^{T}(\frac{1}{a_{t+1}} - \frac{1}{a_t})\|\epsilon_t\|^2 &= \sum_{t=1}^{\tau^*}(\frac{1}{a_{t+1}} - \frac{1}{a_t})\|\epsilon_t\|^2 + \sum_{t=\tau^*+1}^{T}(\frac{1}{a_{t+1}} - \frac{1}{a_t})\|\epsilon_t\|^2 \\ &\leq \frac{1}{\tilde{\beta}}\sum_{t=1}^{\tau^*}\|\epsilon_t\|^2 + \frac{2}{3}\sum_{t=\tau^*+1}^{T}\|\epsilon_t\|^2 \leq \frac{1}{\tilde{\beta}}\sum_{t=1}^{\tau^*}\|\epsilon_t\|^2 + \frac{2}{3}\sum_{t=1}^{T}\|\epsilon_t\|^2\,.\end{aligned} \tag{13}$$

This enables to use Eq. (12) to bound the expected value of term (B).

From here the analysis of the other terms and bounding $\mathbf{E}\sum_{t=1}^{T}\|\epsilon_{t-1}\|^2$ is done similarly to our analysis of simplified STORM$^+$.

Third Part: Bounding $\mathbf{E}\sum_{t=1}^{T}\|\bar{g}_t\|^2$. In this part we divide into two sub-cases depending whether $\mathbf{E}\sum_{t=1}^{T}\|\epsilon_t\|^2 \geq (1/2)\mathbf{E}\sum_{t=1}^{T}\|\bar{g}_t\|^2$ or not. And continue similarly to our analysis of simplified STORM$^+$. The rest of the details appear in the appendix. □

6 Conclusion

We have presented a novel parameter-free and adaptive algorithm for non-convex optimization that obtains the optimal rate in the setting of expectation over smooth losses. Our approach suggests a new way to set the learning rate and momentum jointly and adaptively throughout the learning process, which might open up new avenues to both practical and theoretical developments in this direction.

References

[1] Zeyuan Allen-Zhu and Elad Hazan. Variance reduction for faster non-convex optimization. In *International conference on machine learning*, pages 699–707. PMLR, 2016.

[2] Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*, 2019.

[3] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[4] Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019.

[5] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[6] C Fang, CJ Li, Z Lin, and T Zhang. Near-optimal non-convex optimization via stochastic path integrated differential estimator. *Advances in Neural Information Processing Systems*, 31:689, 2018.

[7] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. *Advances in Neural Information Processing Systems*, pages 2981–2989, 2016.

[8] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[9] Alaa Ali Hameed, Bekir Karlik, and Mohammad Shukri Salman. Back-propagation algorithm with variable adaptive momentum. *Knowledge-Based Systems*, 114:79–87, 2016.

[10] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 558–567, 2019.

[11] Patrik O Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, 5(9), 2004.

[12] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26:315–323, 2013.

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

[14] Lihua Lei, Cheng Ju, Jianbo Chen, and Michael I Jordan. Non-convex finite-sum optimization via scsg methods. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 2345–2355, 2017.

[15] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

[16] Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 983–992. PMLR, 2019.

[17] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

[18] Mehrdad Mahdavi, Lijun Zhang, and Rong Jin. Mixed optimization for smooth functions. *Advances in neural information processing systems*, 26:674–682, 2013.

[19] Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pages 314–323. PMLR, 2016.

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

[21] Vishwak Srinivasan, Adepu Ravi Sankar, and Vineeth N Balasubramanian. Adine: An adaptive momentum method for stochastic gradient descent. In *Proceedings of the ACM india joint international conference on data science and management of data*, pages 249–256, 2018.

[22] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2013.

[23] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

[24] Quoc Tran-Dinh, Nhan H Pham, Dzung T Phan, and Lam M Nguyen. Hybrid stochastic gradient descent algorithms for stochastic nonconvex optimization. *arXiv preprint arXiv:1905.05920*, 2019.

[25] Chong Wang, Xi Chen, Alex Smola, and Eric P Xing. Variance reduction for stochastic gradient optimization. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 1*, pages 181–189, 2013.

[26] Rachel Ward, Xiaoxia Wu, and Leon Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. In *International Conference on Machine Learning*, pages 6677–6686. PMLR, 2019.

[27] Lijun Zhang, Mehrdad Mahdavi, and Rong Jin. Linear convergence with condition number independent access of full gradients. *Advances in Neural Information Processing Systems*, 26:980–988, 2013.

[28] Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic nested variance reduction for nonconvex optimization. *Advances in Neural Information Processing Systems*, 31:3921–3932, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [N/A]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]