1 General response

Thank you to all of the reviewers for your efforts in reviewing our paper. We have responded individually to each of your reviews, and we summarize the changes to our paper here. Changes to the submission are shown in red text, for ease of understanding the modifications.

1. We have fixed a typo in the definition of (L_0, L_1)-smoothness (pointed out by reviewer sfBG). Our definition of (L_0, L_1)-smoothness now matches that of (Zhang et al, 2020a)](https://arxiv.org/abs/2010.02519), which is standard in the literature.

2. In response to a comment by reviewer sfBG, we have added an additional result in Appendix E which removes the condition on γ in Theorem 1 for the setting of deterministic gradients, while recovering the same lower bound as in Theorem 1. Please see Appendix E for a complete description of this new result.

2 sfBG

Thank you for your helpful comments on our paper. We have responded to your individual points below.

W1: Relaxed smoothness definition Your concern actually comes from a typo: all of our proofs use the same definition of (L_0, L_1)-smoothness as previous work. We have updated the paper to fix this typo by including the condition $\|x - y\| \leq 1/L_1$. Since this was a major concern for you, we hope that you will reconsider your score.

W2: Condition for γ Thank you for pointing this out. There are two points we would like to make about this. First, we investigated whether the condition can be removed from Theorem 1, and we succeeded in removing the condition for the deterministic setting while preserving the lower bound of $\Omega(\Delta^2 L_1^2 \epsilon^{-2})$. This additional result is included in Appendix E of our revised submission; please see Appendix E for a complete discussion of this new result. Removing this condition for the stochastic setting remains open. This leads us to our second point: for the stochastic setting, we cover the practical regime where γ is chosen as a small constant (in Pytorch, the default value of the stabilization constant is 10^{-8}). We agree that requiring $\gamma \leq \tilde{O}(\Delta L_1)$ is a theoretical limitation, but we believe that our results still capture the behavior of these algorithms with practical choices of hyperparameters.

W3: Original AdaGrad As you mentioned, we did discuss this point in our limitations section. Although the lower bound for the original AdaGrad can likely be improved, we believe that our results for the decorrelated variants are an important first step towards understanding the original algorithm. This perspective was also taken by (Li and Orabona, 2019)[1].

W4: Affine Noise The reason that we focus on affine noise for Single-Step Adaptive SGD is that under the bounded noise assumption, there are single-step

[1] `https://arxiv.org/abs/1805.08114`

adaptive algorithms that are known to achieve the optimal rate. The prime example is gradient clipping, which was shown to achieve $\mathcal{O}(\Delta L_0 \sigma^2 \epsilon^{-4})$ by (Zhang et al, 2020a)[2], and this matches the lower bound for SGD (with any adaptive learning rate) from (Drori and Shamir, 2020)[3]. Note that this lower bound uses a hard instance that is smooth (and therefore relaxed smooth) and with almost surely bounded noise. Therefore the analysis of Single-Step Adaptive SGD for the bounded noise case is already tight, and relaxed smoothness does not add any difficulty compared to smoothness. Because of this, we consider the affine noise assumption for Single-Step Adaptive SGD.

3 Smgh

Thank you for your review and comments. Below we have responded to your questions and concerns.

W: Affine noise You mentioned that we only consider the affine noise setting for adaptive SGD. We want to clarify that adaptive SGD in the bounded setting is already known to achieve the optimal rate $\mathcal{O}(\Delta L_0 \sigma^2 \epsilon^{-4})$ (Zhang et al, 2020a)[4], so the setting of bounded noise is already resolved. This is our motivation for studying affine noise, which is a slightly harder setting for optimization.

T2: Comparison with (Wang et al, 2023) You said that "Table 1 provides the result of AdaGrad-Norm under affine noise, which is not equivalent to Equation 5." We should clarify that Equation 5 states the upper bound of (Wang et al, 2023)[5] in the case of bounded noise, which is a special case of their affine variance result stated in Table 1. We consider this special case in Equation 5 in order to compare against our lower bounds, which consider bounded noise.

Q1: Relaxed smoothness definition Our definition of relaxed smoothness is a slightly weaker version which does not require the function to be twice-differentiable, and this version was shown to imply the original version (Zhang et al, 2020a)[6]. Also, since all of our hard instances are twice differentiable, our results still apply for the original version of relaxed smoothness. Please let us know if we have answered your question.

Q2: Parameter dependence explanation Thank you for the suggestion. L_1 appears in our lower bounds because AdaGrad-type algorithms cannot operate in two stages, unlike gradient clipping. For example, Decorrelated AdaGrad-Norm must set $\eta \leq 1/L_1$ to avoid divergence on some exponential functions, and this choice of η will affect every single update, even when the algorithm has nearly converged. On the contrary, gradient clipping can avoid divergence with a proper choice of the clipping threshold, and the learning rate can be chosen independently of L_1, so that when the algorithm is close to converging, the update size is unaffected by L_1. AdaGrad lacks this ability to branch into

[2] https://arxiv.org/abs/2010.02519
[3] https://arxiv.org/abs/1910.01845
[4] https://arxiv.org/abs/2010.02519
[5] https://arxiv.org/abs/2305.18471
[6] https://arxiv.org/abs/2010.02519

two options depending on the gradient norm, and this causes the additional dependence on L_1.

Q3: Incorrect description of Theorem 4 You are correct, it should say that the term with quadratic Δ, L_1 goes to 0 instead. We have updated the paper to fix this.

Q4: Reference to (Li et al, 2023) We have referenced (Li et al, 2023)[7] in our related works, and we have added a comment to distinguish this work on Adam from other works on AdaGrad-Norm. Note that this work also exhibits a higher order polynomial dependence on L_1.

4 BkD6

Thank you for your positive review and comments. Below we have answered your question.

W1: Novelty of decorrelated results As you pointed out, we discussed this point in our limitations section. However, we believe that proving these lower bounds for the decorrelated algorithms still requires significant technical novelties. Our Lemmas 1 and 3 contain constructions of novel hard instances for these algorithms (Lemma 3 applies to both decorrelated and original AdaGrad).

Q1: Gradient noise assumptions Assumption 2 states many variations on the stochastic gradient noise assumption, and the main one considered in our paper is almost surely bounded noise. This assumption is common in the literature on relaxed smoothness (Zhang et al, 2020b)[8], (Zhang et al, 2020a)[9], (Crawshaw et al, 2022)[10]. Our last theorem also considers the assumption of affine noise, which has also been used for both smooth and relaxed smooth optimization (Bottou et al, 2016)[11], (Faw et al, 2023)[12], (Attia and Koren, 2023)[13].

5 Bg2B

Thank you for the positive review and helpful comments. We have responded to your questions and concerns below.

W1: High dimensional objectives You are correct that the hard instances in our lower bounds are high-dimensional, i.e. $d \geq T$. However, this is a common situation for lower bounds of first-order algorithms, such as (Arjevani et al, 2023)[14] and many classical lower bounds from (Nesterov, 2013)[15]. To the

[7] https://arxiv.org/abs/2304.13972
[8] https://arxiv.org/abs/1905.11881
[9] https://arxiv.org/abs/2010.02519
[10] https://arxiv.org/abs/2208.11195
[11] https://arxiv.org/abs/1606.04838
[12] https://arxiv.org/abs/2302.06570
[13] https://arxiv.org/abs/2302.08783
[14] https://arxiv.org/abs/1912.02365
[15] https://link.springer.com/book/10.1007/978-3-319-91578-4

best of our knowledge, there are no lower bounds using fixed dimension which can match the same bounds as these high-dimensional results.

W2: Generalizing for Adam It is possible that our results could generalize to Adam-type algorithms, but there are some technical difficulties. In short, some parts of our analysis can be used to analyze Adam, though it will require further work to establish a complete analysis of Adam.

To extend Theorems 2/3 for Decorrelated/Original Adam, we need to establish analogous results to Lemmas 3 and 4. Interestingly, the hard instance from Lemma 3 can be reused for Adam, and we can show that Adam will diverge on the hard instance when $\eta \geq \gamma/(L_1\sigma)\log(1 + L_1\epsilon/L_0)$ (decorrelated Adam) or $\eta \geq 1/L_1 \log(1 + L_1\epsilon/L_0)$ (original Adam). This result is enabled by the fact that the trajectory of Adam is nearly identical to that of AdaGrad for our specific hard instance, which follows from an important property of our construction: each coordinate of the input has zero stochastic gradient for every timestep except for one. Therefore, most of the gradient history is zero, so the moving averages in the numerator and denominator of Adam's will behave very similarly to AdaGrad's update.

However, Lemma 4 is not as easy to extend. Since Adam has a moving average in the denominator instead of a sum (as in AdaGrad), it's behavior on the hard instance of Lemma 4 will differ significantly from AdaGrad: the same hard instance does not yield the same complexity. So extending the analysis to Adam would require a new hard instance to replace the one in our Lemma 4.

Overall, it seems promising to reuse both the overall proof structure and Lemma 3 to analyze Adam, but finishing the proof will require a new construction to fill the hole left by Lemma 4. We leave this question to future work.

W3: Explanation of Theorem 4 constants Thank you for the feedback, and you are correct that the length put some limits on the amount of exposition that we can put in the main body. These constants are error terms arising from the probability of divergence of the biased random walk from Section C.2, and their order depends on σ_2. To provide a sense for these constants without specifying all technical details, we describe their order for different regimes of σ_2 and δ on page 9.

Q1: Difficulties of unified analysis This is a good observation. There are many technical details that create difficulty in simultaneously analyzing AdaGrad-Norm (shared learning rate) and AdaGrad (coordinate-wise learning rate), so let us touch on one. The source of the difficulty is that with coordinate-wise learning rates, each coordinate is affected only by the history of gradients for that particular coordinate, but not for other coordinates.

In Lemma 3, the coordinates of the objective correspond to time steps in the trajectory, and each coordinate sees a nonzero gradient for exactly one timestep. With coordinate-wise learning rates, each coordinate is unaffected by previous history, so Decorrelated AdaGrad and AdaGrad can be analyzed together. Back to the original question, Decorrelated AdaGrad-Norm uses a shared learning rate for each coordinate, so we cannot separate the behavior of each coordinate into separate timesteps, and the history becomes an important factor in the analysis. This is one reason why the objective from Lemma 1 is so different from that of

Lemma 3, and it is not clear whether all three of these algorithms could have a unified analysis.

Q2: γ requirement of Theorem 3 In the case that $\gamma > \sigma$, our current constructions cannot "force divergence" of AdaGrad in the case that $\eta \geq 1/L_1$, which is a key component of our analysis. The reason is that our construction relies on noise in the stochastic gradient to force the algorithm along a trajectory where $\|\nabla F(x_t)\|$ never decreases. If $\gamma > \sigma$, then the noise in the stochastic gradient is dominated by the γ in the adaptive learning rate denominator, and the step size is too small to follow this trajectory. This is not to say that the choice $\gamma > \sigma$ is impossible to handle, but doing so will likely require some new construction.

Also, with AdaGrad there is no fear of the algorithm "exploding", since even when $\gamma = 0$ the denominator of the adaptive learning rate is always larger than the magnitude of the stochastic gradient. This means that the update size of AdaGrad is bounded by η, no matter the choice of γ.

Lastly, can you elaborate what you meant by "The 'better noise-dependency' argument of AdaGrad over DAG is also true only when $\gamma \leq \sigma$"? We believe that this better noise-dependence of AdaGrad over DAG still holds when $\gamma > \sigma$, since the update size of AdaGrad is bounded by η, but the update size of DAG can grow with σ.

Q3: High probability analysis Theorem 4 uses a high-probability analysis because it relies on the probability of divergence of a biased random walk (as opposed to the expectation of a random walk variable after a given time). You are correct that the structure of the arguments of Theorems 1-3 is similar in spirit to that of Theorem 4, but these two settings use very different technical tools, and this leads to the two different types of guarantees.

Q4: Affine noise in Theorem 4 Actually, the complexity of Single-Step Adaptive SGD is already completely characterized by existing work, since the upper bound of gradient clipping (Zhang et al, 2020a)[16] matches the lower bound of adaptive SGD (Drori and Shamir, 2020)[17], so that the best complexity of Single-Step Adaptive SGD is $\mathcal{O}(\Delta L_0 \sigma^2 \epsilon^{-4})$, which recovers the optimal rate from the smooth setting. Note that the lower bound of (Drori and Shamir, 2020)[18] was introduced for the smooth setting, so their hard instance is consequently relaxed smooth. It also has almost surely bounded gradient noise. Since the complexity in this setting is already completely characterized, we focus on the slightly harder setting of affine noise.

Q5: Comparison with (Gorbunov et al, 2023) Can you please provide the specific reference of (Gorbunov et al, 2023)? We are not sure to which paper you refer.

[16] https://arxiv.org/abs/2010.02519
[17] https://arxiv.org/abs/1910.01845
[18] https://arxiv.org/abs/1910.01845

6 AVzC

Thank you for the positive review. Please let us know if you think of any questions we can answer.