Normalized SGD in the Convex Regime: High-Probability Guarantees, Momentum and Lower Bounds

Savelii Chezhegov
MIRAI, BRAIn Lab, Sber AI Lab
savachezhegov2017@gmail.com

Eduard Gorbunov
MBZUAI
eduard.gorbunov@mbzuai.ac.ae

Abstract

Normalized stochastic gradient descent (Normalized SGD) is a classical scale-invariant method that uses only the gradient direction. While Normalized SGD has recently been studied in non-convex stochastic optimization, its convex theory remains much less developed. In this paper, we derive deterministic convergence guarantees for inner and outer momentum and high-probability best-iterate function-value bounds for mini-batch Normalized SGD, with and without outer momentum. Our stochastic results hold for smooth convex objectives under a bounded α-central moment of the heavy-tailed noise, without symmetry or strong convexity assumptions. The resulting oracle complexity $\tilde{\mathcal{O}}\left(\frac{LR_0^2}{\varepsilon}\left(1+\left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right)\right)$, which is worse than the corresponding guarantees for Clip-SGD, raises the question of whether the gap is intrinsic to normalization. We answer this through a series of lower bounds for constant-step normalized methods. First, for a constant batch size, we prove the necessary condition $\Omega\left(\max\left\{1, \left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right)$, matching the choice of batch size in the upper bounds. Second, for arbitrary predetermined batch-size schedules, we establish the oracle lower bound $\Omega\left(\max\left\{\frac{LR_0^2}{\varepsilon}, \sqrt{\frac{LR_0^2}{\varepsilon}}\left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right)$, showing that Normalized SGD cannot attain the optimal complexity of stochastic first-order methods in general. Finally, for non-decreasing batch sizes, we prove the matching oracle lower bound $\Omega\left(\frac{LR_0^2}{\varepsilon}\left(1+\left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right)\right)$ up to logarithmic factors, establishing optimality of our upper bound within this class.

1 Introduction

Machine learning (Shalev-Shwartz and Ben-David, 2014) and deep learning (Goodfellow et al., 2016) have become central tools across modern science and technology. Their recent progress has been driven not only by larger models and datasets, but also by the design of optimization algorithms that remain stable and efficient at scale. While Stochastic Gradient Descent (Robbins and Monro, 1951) and adaptive methods (Duchi et al., 2011; Tieleman, 2012; Kingma and Ba, 2014; Loshchilov and Hutter, 2017; Reddi et al., 2019) remain the standard choices in many applications, the rapid growth of large-scale training has motivated a new wave of optimizers that modify the geometry of the update direction rather than only its coordinate-wise scale.

One particularly active line of work is built around normalizing or orthogonalizing the updates. Recent methods such as Muon (Jordan et al., 2024) and related orthogonalized optimizers (Pethick et al., 2025; Liu et al., 2025) apply momentum to the stochastic gradients and then pass the resulting momentum through a normalization- or orthogonalization-type map (more generally, a linear minimization oracle, or LMO) before using it to update the parameters. These methods have attracted considerable

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

attention in large-scale language-model training (Liu et al., 2025), where changing the geometry of the update can yield strong empirical performance. Conceptually, the simplest method in this family is Normalized Stochastic Gradient Descent (Normalized SGD): rather than stepping along the stochastic gradient itself, it discards the gradient's magnitude and moves along its Euclidean-normalized direction, possibly after momentum averaging. Normalized SGD can therefore serve as a basic model for understanding the behavior of modern normalized and orthogonalized update rules.

Beyond its role as a conceptual model, the normalization in Normalized SGD also makes it possible to establish strong convergence guarantees under assumptions that better reflect the behavior observed in real machine learning and deep learning tasks. One such assumption is heavy-tailed noise: the stochastic gradients may have unbounded variance, and only their α-th central moment is assumed to be bounded for some $\alpha \in (1, 2]$. Heavy-tailed noise has been observed in many problems, including the training of attention-based language models (Zhang et al., 2020c) and policy optimization in reinforcement learning (Garg et al., 2021). Several works have shown that normalization enables convergence under this type of noise (Cutkosky and Mehta, 2020, 2021; Liu and Zhou, 2023; Hübler et al., 2024a; Sun et al., 2025), and some of them further establish high-probability convergence guarantees (Cutkosky and Mehta, 2021; Liu and Zhou, 2023; Hübler et al., 2024a). Such guarantees have attracted growing interest in recent years because they describe the behavior of individual training runs more faithfully than bounds in expectation.

Although Normalized SGD has been studied extensively, most existing analyses focus on the non-convex setting. In particular, to the best of our knowledge, (i) the role of momentum in Normalized SGD remains poorly understood for convex problems, and (ii) no high-probability convergence guarantees are known for the method in the convex setting under heavy-tailed noise. Motivated by these gaps, we make the following contributions.

Our contribution. First, we derive convergence guarantees for deterministic Normalized GD with inner and outer momentum. Next, we establish high-probability best-iterate bounds for mini-batch Normalized SGD, with and without outer momentum, on smooth convex objectives. Under a bounded α-central moment of the noise, for a fixed momentum parameter, the number of oracle calls sufficient to reach ε-accuracy with probability at least $1 - \delta$ is

$$\tilde{\mathcal{O}}\left(\frac{LR_0^2}{\varepsilon}\left(1 + \left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right)\right).$$

Finally, we prove negative results for constant-step methods with predetermined batch sizes. They show that the constant batch size is chosen optimally for the considered approaches. Moreover, we also demonstrate that the provided upper bound on oracle complexity is tight, up to logarithmic factors, for Normalized SGD with *non-decreasing batch sizes*. For arbitrary predetermined schedules, we obtain a weaker oracle lower bound that still separates this method from the achievable bounds for first-order methods. For more details, see Table 1.

2 Preliminaries

Notation. The Euclidean norm in $\mathbb{R}^d$ is denoted as $\|x\| = \sqrt{\langle x, x\rangle}$. The $\mathbb{E}_\xi[\cdot]$ denotes the expectation w.r.t. random variable ξ. We also denote the conditional expectation w.r.t. stochasticity at iteration t as $\mathbb{E}_t[\cdot]$. $R_0 = \|x_0 - x^*\|$ denotes the initial distance to the solution of the problem. Normalization operator is defined to be zero at zero vector, i.e. $0/\|0\| = 0$. $\mathcal{O}(\cdot)$ and $\Omega(\cdot)$ denote the standard complexity upper and lower bounds, respectively. $\tilde{\mathcal{O}}$ hides the polylogarithmic dependence.

Problem Setup. We consider classical stochastic optimization problem, which can be written as

$$\min_{x\in\mathbb{R}^d} \{f(x) := \mathbb{E}_{\xi\sim\mathcal{D}}[f(x,\xi)]\}.$$

This formulation is a cornerstone in machine learning (Shalev-Shwartz and Ben-David, 2014) and appears in an enormous number of practical tasks. Here, f denotes the loss function (or empirical risk), x denotes the parameters of the model, $\mathcal{D}$ is the data distribution, and ξ denotes the stochasticity introduced by sampling the data.

Assumptions. We formulate the considered assumptions below. Let us start with the convexity.

Assumption 1. *The function f is convex, i.e. for all $x, y \in \mathbb{R}^d$ the following inequality holds:*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle .$$

We also use L-smoothness assumption, which is considered as a classical one in the existing literature.

Assumption 2. *The function f is L-smooth, i.e. for all $x, y \in \mathbb{R}^d$ the following inequality holds:*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2.$$

The next assumption describes the stochastic oracle: we assume that it is unbiased and has a bounded α-central moment.

Assumption 3. *For every $x \in \mathbb{R}^d$, the stochastic oracle $\nabla f(x, \xi)$ is unbiased and has a bounded α-central moment with $\alpha \in (1, 2]$, i.e.*

$$\mathbb{E}_\xi[\nabla f(x, \xi)] = \nabla f(x), \qquad \mathbb{E}_\xi[\|\nabla f(x, \xi) - \nabla f(x)\|^\alpha] \leq \sigma^\alpha.$$

Each oracle call uses a fresh draw from distribution, independent from previous draws.

In recent years, this assumption has become increasingly natural: several works have shown that heavy-tailed noise arises in modern machine learning and deep learning, with prominent examples including image classification (Simsekli et al., 2019), the training of attention-based language models (Zhang et al., 2020c), and reinforcement learning (Garg et al., 2021). Motivated by these observations, many recent works adopt Assumption 3 (Zhang et al., 2020c; Cutkosky and Mehta, 2021; Sadiev et al., 2023; Gorbunov et al., 2024b; Hübler et al., 2024a; El Bakkali El Kadi et al., 2025). As a result, it has become a standard assumption in the literature, and it is both more general and more practically relevant than the classical bounded-variance assumption, which corresponds to $\alpha = 2$ (Nemirovski et al., 2009; Takáč et al., 2013; Ghadimi and Lan, 2013).

3 Related Works

Normalized methods in the convex case. Deterministic normalized methods have been studied under convexity and related assumptions (Nesterov, 1984; Hazan et al., 2015; Grimmer, 2019). Building on Levy (2017), Orabona (2023) gives an online-learning reduction that adapts to local Hölder smoothness. Further results cover generalized smoothness (Vankov et al., 2025; Lobanov et al., 2024). Gorbunov et al. (2025) study related clipped and adaptive methods.

For stochastic methods, Hazan et al. (2015) derive high-probability function-value bounds under bounded mini-batch objectives and strict local quasi-convexity, selecting the output using objective values. Levy (2017) studies projected adaptive normalization under bounded stochastic gradients; Turan et al. (2021) assume random adversarial corruption, an acute-angle condition, and a bounded domain. Kovalev (2025) analyzes inner-momentum trust-region methods with weight decay, including Normalized SGD with weight decay. Nonlinear SGD includes normalization (Jakovetić et al., 2023; Armacki et al., 2025); the latter work gives high-probability last-iterate bounds under strong convexity and additive symmetric noise with a density positive near zero, without moment assumptions. Our best-iterate guarantees require neither strong convexity nor noise symmetry.

Normalized methods in the non-convex case. Inner momentum avoids large mini-batches under bounded variance (Cutkosky and Mehta, 2020). Further results cover generalized smoothness with bounded noise (Zhang et al., 2020a), parameter-agnostic guarantees (Hübler et al., 2024b), and algorithm-specific lower bounds (Yang et al., 2023). Under heavy-tailed noise, normalized momentum has been analyzed with clipping (Cutkosky and Mehta, 2021; Liu et al., 2023) and without clipping (Hübler et al., 2024a; Liu and Zhou, 2025; Sun et al., 2025). Hübler et al. (2024a) obtain high-probability mini-batch guarantees matching oracle lower bounds up to logarithmic factors. Under additional stochastic-oracle smoothness, Fradin et al. (2026) match lower bounds in expectation using variance-reduced normalized momentum and give high-probability bounds for a double-clipped variant. These results concern stationarity rather than function-value accuracy.

Robust convex optimization and gradient clipping. Clipping yields high-probability guarantees without light-tail assumptions (Nazin et al., 2019; Gorbunov et al., 2020). Under bounded central moments, Sadiev et al. (2023); Nguyen et al. (2023) analyze clipped gradient methods allowing infinite variance; see also Gorbunov et al. (2024b). Mai and Johansson (2021) study clipping stability for nonsmooth convex objectives with quadratic growth and rapidly growing subgradients. High-probability guarantees also cover nonsmooth objectives (Liu and Zhou, 2023; Parletta et al., 2024; Gorbunov et al., 2024a) and last iterates (Chezhegov et al., 2026). Tsai et al. (2022) and Das et al. (2024) analyze clipped SGD for streaming estimation under finite second-moment conditions. For online convex problems with Lipschitz losses and unbounded domains, Zhang and Cutkosky (2022) combine clipping and regularization to obtain high-probability regret bounds. Shestakov et al. (2026) connect convex finite-sum clipping guarantees under heavy-tailed data with generalization.

Zhang et al. (2020b) motivate clipping through generalized smoothness in non-convex optimization; convex heavy-tailed analyses cover clipped SGD (El Bakkali El Kadi et al., 2025). Further results concern adaptive methods (Chezhegov et al., 2025), median-based estimators under structured noise (Puchkin et al., 2024), and variational inequalities under bounded variance (Gorbunov et al., 2022).

Clipping bias is studied by Chen et al. (2020) and Koloskova et al. (2023). In private federated learning, Andrew et al. (2021) adapt client-update clipping using privately estimated norm quantiles. With fixed clipping and heavy-tailed noise, Khah et al. (2026) obtain high-probability convergence to a neighborhood. Distributed applications include deterministic error feedback (Khirirat et al., 2023), Byzantine robustness with partial participation (Malinovsky et al., 2024), and private double-momentum methods under sub-Gaussian noise (Islamov et al., 2025, 2026).

Without clipping, stochastic mirror descent (Vural et al., 2022) and projected SGD (Fatkhullin et al., 2025) attain optimal expected convex rates under bounded-domain and raw-gradient-moment assumptions. Our clipping comparison concerns high-probability guarantees under the central-moment model in Assumption 3; our lower bounds concern constant-step mini-batch normalized SGD with predetermined batch sizes, including a constant-batch outer-momentum result.

4 Main Results

In this section, we present theoretical guarantees of the considered algorithms. It is natural to start with deterministic approaches.

4.1 Deterministic case

We first recall the deterministic convex guarantee for Normalized GD. This result is known: it follows, for instance, from Hazan et al. (2015, Theorem 4.2), which holds even for quasi-convex, locally smooth functions. We include a short proof for completeness, as it serves as a template for the stochastic analysis below.

Algorithm 1 Normalized GD

Input: Starting point x_0, learning rate γ
for $k = 0, \ldots, K-1$ **do**
  $x_{k+1} = x_k - \gamma \frac{\nabla f(x_k)}{\|\nabla f(x_k)\|}$
end for

Theorem 1. *Suppose that Assumptions 1 and 2 hold. Then, after K iterations of Algorithm 1 with $\gamma = \frac{R_0}{\sqrt{K}}$, we have*

$$\min_{k \in 0,\ldots,K-1} (f(x_k) - f^*) = \mathcal{O}\left(\frac{LR_0^2}{K}\right).$$

Next, we propose theoretical results for the approaches with momentum. But before we proceed to the convergence theorems, let us briefly discuss the use of momentum in normalization-based approaches. In general, the underlying estimator is either the full gradient or its approximation in the form of a stochastic gradient. Nevertheless, momentum technique can be applied in different ways.

The most common approach is to apply momentum before normalization (Cutkosky and Mehta, 2020): the parameters are updated along the normalized direction $m_t / \| m_t \|$, where

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{1}$$

and g_t is the gradient estimate at iteration t. This form of momentum is rooted in the heavy-ball method and can reduce the variance of the stochastic gradients when β_1 is close to 1; in practice, β_1 is typically set between 0.9 and 0.95. Since (1) is not the only way to incorporate momentum, we refer to this approach as *inner momentum* to distinguish it from the alternatives discussed below.

Alternatively, momentum can be applied after normalization:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) \frac{g_t}{\|g_t\|}. \tag{2}$$

We refer to this scheme as *outer momentum*. It is conceptually different from (1): inner momentum averages the raw stochastic gradients, so their magnitudes influence the resulting direction, whereas outer momentum averages unit-norm directions and thus discards magnitude information. Outer momentum is used considerably less often, but it still appears in several recent works. Zhao et al. (2024) analyze a related update, with a different scaling and without bias correction, in expectation in the non-convex setting. Tuddenham et al. (2022) propose Orthogonal-SGDM, which combines orthogonalization with outer momentum to accelerate neural network training. Chezhegov et al. (2025) combine outer momentum with gradient clipping in their analysis of adaptive methods. Finally, Pethick et al. (2025, Appendix D.3) propose ALMOND, which applies outer momentum to the outputs of an LMO rather than to normalized gradients, generalizing the method of Zhao et al. (2024) to arbitrary norms.

Since inner and outer momentum behave differently and both are used in practice, we analyze both; to the best of our knowledge, outer momentum has not been analyzed in the convex setting.

Algorithm 2 Normalized GD with inner momentum

Input: Starting point x_0, learning rate γ, momentum parameter $\beta_1 \in [0, 1)$, $m_{-1} = 0$
for $k = 0, \dots, K-1$ **do**
  $m_k = \beta_1 m_{k-1} + (1 - \beta_1) \nabla f(x_k)$
  $x_{k+1} = x_k - \gamma \frac{m_k}{\|m_k\|}$
end for

Theorem 2. *Suppose that Assumptions 1 and 2 hold. Then, after K iterations of Algorithm 2 with $\gamma = \frac{R_0 \sqrt{1-\sqrt{\beta_1}}}{\sqrt{2K}}$, we have*

$$\min_{k \in 0, \dots, K-1} (f(x_k) - f^*) = \mathcal{O}\left(\frac{L R_0^2}{(1 - \sqrt{\beta_1}) K} \right).$$

The proof follows a conceptually different route from that of Theorem 1. There, i.e., for $\beta_1 = 0$, we derive a one-iteration bound and sum it over the iterations. For inner momentum, we instead combine the descent lemma with a *proof by contradiction*. Specifically, at each iteration k, the descent lemma yields a bound involving the weighted average of past suboptimality gaps

$$\sum_{t=0}^{k} \frac{(1 - \beta_1)\beta_1^{k-t}}{1 - \beta_1^{k+1}}, \big(f(x_t) - f^*\big),$$

and we show that, for at least one $k \in \{0, \dots, K-1\}$, this average is smaller than $\frac{4 L R_0^2}{(1-\sqrt{\beta_1})K}$, which is the desired bound. Assuming the contrary leads to a contradiction. Such arguments are uncommon in the analysis of optimization methods, particularly deterministic ones. Nevertheless, this approach yields, to the best of our knowledge, the first convergence guarantees for normalization combined with classical (inner) momentum without weight decay in the convex setting. The full proof is deferred to Section B.

Now we provide an upper bound for outer momentum.

Algorithm 3 Normalized GD with outer momentum

Input: Starting point x_0, learning rate γ, momentum parameter $\beta_1 \in [0, 1)$, $m_{-1} = 0$
for $k = 0, \ldots, K-1$ **do**
  $m_k = \beta_1 m_{k-1} + (1-\beta_1)\frac{\nabla f(x_k)}{\|\nabla f(x_k)\|}$
  $\tilde{m}_k = \frac{m_k}{1-\beta_1^{k+1}}$
  $x_{k+1} = x_k - \gamma \tilde{m}_k$
end for

Theorem 3. *Suppose that Assumptions 1 and 2 hold. Then, after K iterations of Algorithm 3 with $\gamma = \frac{R_0\sqrt{1-\beta_1}}{\sqrt{2K}}$, we have*

$$\min_{k\in 0,\ldots,K-1}(f(x_k) - f^*) = \mathcal{O}\left(\frac{LR_0^2}{(1-\beta_1)K}\right).$$

As expected, the bound in Theorem 2 recovers that of Theorem 1 up to an additional factor $\frac{1}{1-\sqrt{\beta_1}}$, while the bound in Theorem 3 does so up to a factor $\frac{1}{1-\beta_1}$. These two factors are equivalent up to a numerical constant, since

$$\frac{1-\beta_1}{1-\sqrt{\beta_1}} = 1 + \sqrt{\beta_1} \in [1, 2].$$

Hence, the upper bounds in Theorems 2 and 3 differ *only by a numerical constant*.

Remark 1. *In fact, convexity is not required to establish the results of Theorems 1, 2, and 3; it suffices to assume the **star-convexity** of f, i.e., for all $x \in \mathbb{R}^d$ the following inequality holds:*

$$f^* \geq f(x) + \langle \nabla f(x), x^* - x\rangle.$$

The proofs of the mentioned theorems remain unchanged.

4.2 Stochastic case

In this section, we present our results for the stochastic setting. We begin with high-probability convergence guarantees for Normalized SGD and its variant with outer momentum.

High-probability convergence. Before presenting these guarantees, we briefly outline the proof methodology and the main technical tools it relies on.

Key proof highlights. To prove such results, we use the idea from Sadiev et al. (2023): we consider certain probabilistic events whose probabilities can be estimated below by induction. Generally speaking, for algorithms with normalization (Cutkosky and Mehta, 2021; Hübler et al., 2024a) and other similar structures, such as sign-based methods Kornilov et al. (2025), the proof is constructed differently. At the same time, (Sadiev et al., 2023) use this approach to study gradient clipping. We decided to follow the route of (Sadiev et al., 2023) for the following clear reason: the proposed inductive structure allows one to control R_t, which is critical for our analysis due to the descent lemma in the convex case.

However, while several works on gradient clipping (e.g., (Sadiev et al., 2023; Gorbunov et al., 2024b; Chezhegov et al., 2025)) use a certain good event E_k (for more details, see Section C), we also consider an additional event S_k:

$$S_k = \left\{\min_{t=0,\ldots,k-1}(f(x_t) - f^*) \leq \frac{Const(\beta_1)LR_0^2\ln^2\left(\frac{2K}{\delta}\right)}{K}\right\},$$

which is needed to construct the induction step. That is, whereas in all existing results E_k implies convergence, this does not hold in our setting – thus, we need to control the event of convergence additionally. After establishing the bound $\mathbb{P}\{S_K \cup E_K\} \geq 1-\delta$ by inductive step, we do not obtain a convergence bound directly; rather, we show that $\overline{S_K} \cap E_K = \varnothing$, and this already leads to the desired result.

We would also like to emphasize separately the new lemma for the bias part. Based on the existing literature, in the non-convex setting one typically uses the so-called expected angle bound lemma (Zhao et al., 2021; Hübler et al., 2024a). However, in the convex case, such a structure does not appear; instead, we provide Lemma 4 to give an upper bound for

$$\left\| \mathbb{E}\left[\frac{g_t}{\|g_t\|}\right] - \frac{\nabla f(x_t)}{\|\nabla f(x_t)\|} \right\|.$$

According to the existing literature, this result was not provided before. For more details, see Appendix A and C. □

We are now ready to state our high-probability guarantees for Normalized SGD in the convex case.

Algorithm 4 Normalized SGD

Input: Starting point x_0, learning rate γ, batch size schedule $\{B_k\}_{k=0}$
for $k = 0, \ldots, K-1$ **do**
  Compute batched stochastic gradient $\nabla f(x_k, \xi_k) = \frac{1}{B_k}\sum_{i=1}^{B_k} \nabla f(x_k, \xi_k^i)$
  $x_{k+1} = x_k - \gamma \frac{\nabla f(x_k,\xi_k)}{\|\nabla f(x_k,\xi_k)\|}$
end for

Theorem 4. *Suppose that Assumptions 1, 2 and 3 hold. Then, after K iterations with $\ln\left(\frac{2K}{\delta}\right) \geq 1$ of Algorithm 4 with $\gamma = \frac{R_0}{32\sqrt{K}\ln\left(\frac{2K}{\delta}\right)}$ and batch size $B_k \equiv B = \left\lceil \max\left\{1, \left(\frac{\sigma K}{128 L R_0 \ln^2\left(\frac{2K}{\delta}\right)}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right\rceil$, we have*

$$\min_{k=0,\ldots,K-1}(f(x_k) - f^*) = \mathcal{O}\left(\frac{L R_0^2 \ln^2\left(\frac{K}{\delta}\right)}{K}\right)$$

with probability at least $1-\delta$. As a result, the number of oracle calls required to reach ε-accuracy (i.e. $\min_{k=0,\ldots,K-1}(f(x_k) - f^*) \leq \varepsilon$) with probability $1-\delta$, is upper bounded as follows:

$$\#\textit{oracle calls} = \tilde{\mathcal{O}}\left(\frac{L R_0^2}{\varepsilon}\left(1 + \left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right)\right).$$

Next, let us formulate our result for the Normalized SGD with outer momentum.

Algorithm 5 Normalized SGD with outer momentum

Input: Starting point x_0, learning rate γ, momentum parameter $\beta_1 \in [0, 1)$, batch size schedule B_k, $m_{-1} = 0$
for $k = 0, \ldots, K-1$ **do**
  Compute batched stochastic gradient $\nabla f(x_k, \xi_k) = \frac{1}{B_k}\sum_{i=1}^{B_k} \nabla f(x_k, \xi_k^i)$
  $m_k = \beta_1 m_{k-1} + (1-\beta_1)\frac{\nabla f(x_k,\xi_k)}{\|\nabla f(x_k,\xi_k)\|}$
  $\tilde{m}_k = \frac{m_k}{1-\beta_1^{k+1}}$
  $x_{k+1} = x_k - \gamma \tilde{m}_k$
end for

Theorem 5. *Suppose that Assumptions 1, 2 and 3 hold. Then, after K iterations with $\ln\left(\frac{2K}{\delta}\right) \geq 1$ of Algorithm 5 with $\gamma = \frac{R_0\sqrt{1-\beta_1}}{32 C(\beta_1)\sqrt{K}\ln\left(\frac{2K}{\delta}\right)}$, batch size $B_k \equiv B = \left\lceil \max\left\{1, \left(\frac{\sigma K(1-\beta_1)}{128 L R_0 \ln^2\left(\frac{2K}{\delta}\right)}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right\rceil$ and $C(\beta_1) = 1 + \ln\left(\frac{1}{1-\beta_1}\right)$, we have*

$$\min_{k=0,\ldots,K-1}(f(x_k) - f^*) = \mathcal{O}\left(\frac{L R_0^2 C(\beta_1)^2 \ln^2\left(\frac{K}{\delta}\right)}{(1-\beta_1)K}\right)$$

with probability at least $1-\delta$*. As a result, the number of oracle calls required to reach* ε*-accuracy (i.e.* $\min_{k=0,\ldots,K-1}(f(x_k)-f^*)\leq\varepsilon$*) with probability* $1-\delta$*, is upper bounded as follows:*

$$\#\textit{oracle calls} = \tilde{\mathcal{O}}\left(\frac{LR_0^2}{(1-\beta_1)\varepsilon}\left(1+\left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right)\right).$$

Remark 2. *Similarly to Remark 1, it is enough to consider the* ***star-convexity*** *of* f *to derive the same result as in Theorems 4 and 5.*

Outer momentum does not improve the oracle complexity in our upper bounds. It can reduce the variability of normalized directions, but it does not generally remove their normalization bias. This distinction explains why averaging after normalization need not have the same effect as averaging the gradients before normalization

Inner momentum – the open question. One of the most important questions concerns convergence guarantees for Normalized SGD with inner momentum, as this is the form most commonly encountered in practice. At present, based on the existing literature, there are no high-probability convergence analyses for Algorithm 6, even in the non-convex case. Two works are worth mentioning here:

- Cutkosky and Mehta (2021): high-probability guarantees are provided for Algorithm 6 in non-convex scenario, but it requires the gradient clipping technique.
- Hübler et al. (2024a): the authors propose the convergence of Algorithm 6 for non-convex setup only in-expectation manner. They also emphasize that extension on high-probability guarantees is a strong challenge: it might be impossible according to their empirical evaluation.

Algorithm 6 Normalized SGD with inner momentum

Input: Starting point x_0, learning rate γ, momentum parameter $\beta_1\in[0,1)$, $m_{-1}=0$
for $k=0,\ldots,K-1$ **do**
  Sample stochastic gradient $\nabla f(x_k,\xi_k)$
  $m_k=\beta_1 m_{k-1}+(1-\beta_1)\nabla f(x_k,\xi_k)$
  $x_{k+1}=x_k-\gamma\frac{m_k}{\|m_k\|}$
end for

Consequently, without using additional tools to control stochasticity, such as gradient clipping, it remains *unclear whether high-probability convergence guarantees can be established even in the non-convex setup for* Algorithm 6.

4.3 Negative results

A natural question arises: can we improve the bounds from Theorems 4 and 5? Indeed, it is known that the lower bound on oracle calls for stochastic optimization under the assumption on bounded α-central moment is

$$\#\text{oracle calls} = \Omega\left(\left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right);$$

hence, there is a gap between normalization approaches and first-order methods. To better understand its behavior (whether it stems from the tightness of our analysis or from a fundamental difference in the effect of normalization between the convex and nonconvex settings), we present the following negative results for Normalized SGD. We start with a result for the constant batch size.

Theorem 6 (Lower bound on constant batch)**.** *Suppose that* $\varepsilon\leq\frac{LR_0^2}{16}$*,* $K\geq 1$*,* $\gamma>0$ *and* $\beta_1\in[0,1)$*. Then there exist a one-dimensional function* f *satisfying Assumptions 1, 2 and stochastic oracles* $\{g_k\}_{k\geq 0}$ *satisfying Assumption 3 such that after run of* Normalized SGD with outer momentum

for K iterations with a constant batch size B, the following bound holds:

$$\mathbb{P}\left\{\min_{k=0,\ldots,K}(f(x_k)-f^*)>\varepsilon\right\}<\delta$$
$$\Rightarrow B=\Omega\left(\max\left\{1,\ln\left(1+(1-\delta)^{\frac{1}{C(\gamma,\beta_1)}}\right)\left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right),$$

where $C(\gamma,\beta_1)=\left\lceil\frac{5R_0}{8\gamma\left(2+\ln\left(\frac{2}{1-\beta_1}\right)\right)}\right\rceil\geq 1$. *In particular, if* $\delta\leq\frac{1}{2}$, *we get*

$$\mathbb{P}\left\{\min_{k=0,\ldots,K}(f(x_k)-f^*)>\varepsilon\right\}<\delta\Rightarrow B=\Omega\left(\max\left\{1,\left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right).$$

Remark 3. *This result also holds for* Normalized SGD *– it is enough to substitute* $\beta_1=0$.

Theorem 6 shows that the choice of constant batch size in Theorems 4 and 5 is *optimal* in heavy-tailed regime with respect to σ, R_0 and ε for Algorithms 4 and 5. To verify this, it suffices to set the convergence criterion $\min_{k=0,\ldots,K}(f(x_k)-f^*)$ equal to ε in the upper bounds and rewrite the resulting expression for the batch size.

However, this result does not constitute a lower bound on the oracle complexity of Normalized SGD. It is entirely possible that a time-varying batch size allows one to obtain a bound matching the optimal complexity bound for first-order methods. To investigate whether this is indeed the case, we establish the following result.

Theorem 7 (Lower bound on oracle complexity)**.** *Suppose that* $d\geq 2$, $\varepsilon\leq\frac{LR_0^2}{64}$, $K\geq 1$, $\gamma>0$ *and* $\{B_k\}_{k\geq 0}$ *is arbitrary predetermined batch size schedule. Also suppose that* Normalized SGD *satisfies*

$$\mathbb{P}\left\{\min_{k=0,\ldots,K}(f(x_k)-f^*)>\varepsilon\right\}<\delta$$

for every objective and stochastic oracles satisfying Assumptions 1–3. Then, we have

$$N_\varepsilon=\Omega\left(\max\left\{\frac{LR_0^2}{\varepsilon},(1-\delta)^{1/M_\gamma}\sqrt{\frac{LR_0^2}{\varepsilon}}\left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right),$$

where $N_\varepsilon=\sum\limits_{k=0}^{K-1}B_k$ *is the total number of oracle calls and* $M_\gamma=\left\lceil\frac{13R_0}{16\gamma}\right\rceil\geq 1$. *In particular, if* $\delta\leq\frac{1}{2}$, *we get*

$$N_\varepsilon=\Omega\left(\max\left\{\frac{LR_0^2}{\varepsilon},\sqrt{\frac{LR_0^2}{\varepsilon}}\left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right).$$

Surprisingly, it turns out that constant step Normalized SGD with arbitrary predetermined batch size schedule cannot attain the stochastic term from optimal bound in the class of first-order methods. More importantly, in such a case Normalized SGD is optimal in the nonconvex setting; it was shown by Hübler et al. (2024a). Thus, Theorem 7 provides direct evidence that normalization *has different effects in convex and nonconvex settings*.

Theorem 7 leaves a gap between our upper (Theorem 4) and lower bounds: indeed, the difference in the stochastic term is $\varepsilon^{-1/2}$. In fact, it is not immediately clear whether this gap arises from the upper or the lower bounds. To shed light on this question, we consider a particular class of batch size schedules. Specifically, we focus on *non-decreasing* batch size schedules for two reasons: (i) they are a natural and well-motivated choice from a practical perspective (Smith et al., 2017), and (ii) they admit convenient constructions that facilitate theoretical analysis. In this setting, we present the following negative result.

Theorem 8 (Lower bound on oracle complexity for non-decreasing batch sizes)**.** *Suppose that* $d \geq 2$, $\varepsilon \leq \frac{LR_0^2}{64}$, $K \geq 1$, $\gamma > 0$, $\delta \leq \frac{1}{2}$ *and* $\{B_k\}_{k\geq 0}$ *is predetermined non-decreasing batch size schedule. Also suppose that* Normalized SGD *satisfies*

$$\mathbb{P}\left\{\min_{k=0,\ldots,K}(f(x_k) - f^*) > \varepsilon\right\} < \delta$$

for every objective and stochastic oracles satisfying Assumptions 1–3. Then, we have

$$N_\varepsilon = \Omega\left(\frac{LR_0^2}{\varepsilon}\max\left\{1, \left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right)$$

where $N_\varepsilon = \sum_{k=0}^{K-1} B_k$ *is the total number of oracle calls after* K *iterations.*

Thus, Theorem 4 is tight up to logarithmic factors within the class of constant-step methods with predetermined non-decreasing batch sizes. This follows from the fact that a constant batch size across all iterations is a special case of a non-decreasing batch size schedule.

Remark 4. *It is worth noting that results provided in Theorems 7 and 8 do not cover adaptive batch sizes or changing steps.*

Table 1: Stochastic guarantees for constant-step methods. All simplified lower bounds assume $\delta \leq \frac{1}{2}$. For more details, see Sections 4.2 and 4.3.

Method	Type	Oracle Complexity	Batch Size Schedule
Algorithms 4 and 5	Negative result	$\Omega\left(\max\left\{1, \left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right)$ per step[(1)]	Constant
Algorithm 4	Upper bound	$\tilde{\mathcal{O}}\left(\frac{LR_0^2}{\varepsilon}\left(1 + \left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right)\right)$	Constant
	Lower bound I	$\Omega\left(\frac{LR_0^2}{\varepsilon} + \sqrt{\frac{LR_0^2}{\varepsilon}\left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}}\right)$	Arbitrary
	Lower bound II	$\Omega\left(\frac{LR_0^2}{\varepsilon}\left(1 + \left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right)\right)$[(2)]	Non-decreasing
Algorithm 5	Upper bound	$\tilde{\mathcal{O}}\left(\frac{LR_0^2}{(1-\beta_1)\varepsilon}\left(1 + \left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right)\right)$	Constant

[(1)] This is the lower bound on $B_k \equiv B$ from Theorem 6 for Normalized SGD with/without outer momentum.
[(2)] Matching lower bound for predetermined non-decreasing batch sizes.

5 Conclusion

In this paper, we studied normalized methods in smooth convex optimization. We derived convergence guarantees for deterministic Normalized GD with inner and outer momentum, and high-probability best-iterate bounds for mini-batch Normalized SGD with and without outer momentum under a bounded α-central moment of the noise.

Our lower bounds show that constant batch size is chosen optimally and cannot be reduced. In parallel, we demonstrate that Normalized SGD with constant step size cannot attain the optimal bound on oracle complexity in the class of first-order methods with additional factor $\Omega(\varepsilon^{-1/2})$. Moreover, the provided upper bound on oracle complexity is tight up to logarithmic factors for constant-step Normalized SGD with predetermined *non-decreasing* batch sizes. For arbitrary predetermined batch sizes, a gap remains between the upper and lower bounds. Other open questions concern changing learning rates, adaptive batch sizes, selecting an accurate output using only stochastic gradients, and high-probability convergence of unclipped inner momentum with polylogarithmic confidence dependence.

We hope that our results serve an important foundation for further research of normalized approaches and understanding of behavior of modern methods in deep learning.

Acknowledgments

The work of S. Chezhegov was supported by the Ministry of Economic Development of the Russian Federation (agreement No. 139-15-2025-013, dated June 20, 2025, IGK 000000C313925P4B0002).

Use of generative AI. We used generative AI tools to polish the text and assist in checking the proofs. The upper bounds were derived by the authors, who also developed the key ideas behind the lower bounds. Generative AI helped formalize the lower-bound arguments. We take full responsibility for the results and the entire content of the paper.

References

Andrew, G., Thakkar, O., McMahan, B., and Ramaswamy, S. (2021). Differentially private learning with adaptive clipping. *Advances in neural information processing systems*, 34:17455–17466.

Armacki, A., Yu, S., Sharma, P., Joshi, G., Bajovic, D., Jakovetic, D., and Kar, S. (2025). High-probability convergence bounds for online nonlinear stochastic gradient descent under heavy-tailed noise. In *AISTATS*, pages 1774–1782.

Chen, X., Wu, S. Z., and Hong, M. (2020). Understanding gradient clipping in private SGD: A geometric perspective. *Advances in neural information processing systems*, 33:13773–13782.

Chezhegov, S., Parletta, D. A., Paudice, A., and Gorbunov, E. (2026). High-probability bounds for the last iterate of clipped SGD. In *The Fourteenth International Conference on Learning Representations*.

Chezhegov, S., Yaroslav, K., Semenov, A., Beznosikov, A., Gasnikov, A., Horváth, S., Takáč, M., and Gorbunov, E. (2025). Clipping improves adam-norm and adagrad-norm when the noise is heavy-tailed. In *Forty-second International Conference on Machine Learning*.

Cutkosky, A. and Mehta, H. (2020). Momentum improves normalized sgd. In *International conference on machine learning*, pages 2260–2268. PMLR.

Cutkosky, A. and Mehta, H. (2021). High-probability bounds for non-convex stochastic optimization with heavy tails. *Advances in Neural Information Processing Systems*, 34:4883–4895.

Das, A., Nagaraj, D., Pal, S., Suggala, A. S., and Varshney, P. (2024). Near-optimal streaming heavy-tailed statistical estimation with clipped SGD. *Advances in Neural Information Processing Systems*, 37:8834–8900.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7).

El Bakkali El Kadi, T., Chezhegov, S., Beznosikov, A., Horváth, S., and Gorbunov, E. (2025). High-probability convergence of clipped sgd under heavy-tailed noise and (L_0, L_1)-smoothness. *arXiv e-prints*, pages arXiv–2505.

Fatkhullin, I., Hübler, F., and Lan, G. (2025). Can SGD handle heavy-tailed noise? *arXiv preprint arXiv:2508.04860*.

Fradin, A., Sadiev, A., Condat, L., and Richtárik, P. (2026). Tight lower bounds and optimal algorithms for stochastic nonconvex optimization with heavy-tailed noise. In Khan, E., Li, Y., Solin, A., and Ramdas, A., editors, *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pages 2548–2556. PMLR.

Garg, S., Zhanson, J., Parisotto, E., Prasad, A., Kolter, Z., Lipton, Z., Balakrishnan, S., Salakhutdinov, R., and Ravikumar, P. (2021). On proximal policy optimization's heavy-tailed gradients. In *International Conference on Machine Learning*, pages 3610–3619. PMLR.

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

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Gorbunov, E., Danilova, M., Dobre, D., Dvurechenskii, P., Gasnikov, A., and Gidel, G. (2022). Clipped stochastic methods for variational inequalities with heavy-tailed noise. *Advances in Neural Information Processing Systems*, 35:31319–31332.

Gorbunov, E., Danilova, M., and Gasnikov, A. (2020). Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *Advances in Neural Information Processing Systems*, 33:15042–15053.

Gorbunov, E., Danilova, M., Shibaev, I., Dvurechensky, P., and Gasnikov, A. (2024a). High-probability complexity bounds for non-smooth stochastic convex optimization with heavy-tailed noise. *Journal of Optimization Theory and Applications*, pages 1–60.

Gorbunov, E., Sadiev, A., Danilova, M., Horváth, S., Gidel, G., Dvurechensky, P., Gasnikov, A., and Richtárik, P. (2024b). High-probability convergence for composite and distributed stochastic minimization and variational inequalities with heavy-tailed noise. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F., editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 15951–16070. PMLR.

Gorbunov, E., Tupitsa, N., Choudhury, S., Aliev, A., Richtárik, P., Horváth, S., and Takáč, M. (2025). Methods for convex (L_0, L_1)-smooth optimization: Clipping, acceleration, and adaptivity. In *International Conference on Learning Representations*, volume 2025, pages 97261–97305.

Grimmer, B. (2019). Convergence rates for deterministic and stochastic subgradient methods without Lipschitz continuity. *SIAM Journal on Optimization*, 29(2):1350–1365.

Hazan, E., Levy, K., and Shalev-Shwartz, S. (2015). Beyond convexity: Stochastic quasi-convex optimization. *Advances in neural information processing systems*, 28.

Hübler, F., Fatkhullin, I., and He, N. (2024a). From gradient clipping to normalization for heavy tailed sgd. *arXiv preprint arXiv:2410.13849*.

Hübler, F., Yang, J., Li, X., and He, N. (2024b). Parameter-agnostic optimization under relaxed smoothness. In *International Conference on Artificial Intelligence and Statistics*, pages 4861–4869. PMLR.

Islamov, R., Horvath, S., Lucchi, A., Richtarik, P., and Gorbunov, E. (2025). Double momentum and error feedback for clipping with fast rates and differential privacy. *arXiv preprint arXiv:2502.11682*.

Islamov, R., Malinovsky, G., Gaponov, A., Lucchi, A., Richtárik, P., and Gorbunov, E. (2026). Byzantine-robust and differentially private federated optimization under weaker assumptions. In Perković, E. and Malinsky, D., editors, *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pages 2378–2441. PMLR.

Jakovetić, D., Bajović, D., Sahu, A. K., Kar, S., Milos˘ević, N., and Stamenković, D. (2023). Nonlinear gradient mappings and stochastic optimization: A general framework with applications to heavy-tail noise. *SIAM Journal on Optimization*, 33(2):394–423.

Jordan, K., Jin, Y., Boza, V., Jiacheng, Y., Cesista, F., Newhouse, L., and Bernstein, J. (2024). Muon: An optimizer for hidden layers in neural networks, 2024. *URL https://kellerjordan. github. io/posts/muon*, 6(3):4.

Khah, S. V., Chezhegov, S., Farahmand, S., Horváth, S., and Gorbunov, E. (2026). Differentially private Clipped-SGD: High-probability convergence with arbitrary clipping level. In Khan, E., Li, Y., Solin, A., and Ramdas, A., editors, *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pages 4762–4770. PMLR.

Khirirat, S., Gorbunov, E., Horváth, S., Islamov, R., Karray, F., and Richtárik, P. (2023). Clip21: Error feedback for gradient clipping. *arXiv preprint arXiv:2305.18929*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koloskova, A., Hendrikx, H., and Stich, S. U. (2023). Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *International Conference on Machine Learning*, pages 17343–17363. PMLR.

Kornilov, N., Zmushko, P., Semenov, A., Ikonnikov, M., Gasnikov, A., and Beznosikov, A. (2025). Sign operator for coping with heavy-tailed noise in non-convex optimization: High probability bounds under (L_0, L_1)-smoothness. *arXiv preprint arXiv:2502.07923*.

Kovalev, D. (2025). Understanding gradient orthogonalization for deep learning via non-euclidean trust-region optimization. *arXiv preprint arXiv:2503.12645*.

Levy, K. (2017). Online to offline conversions, universality and adaptive minibatch sizes. *Advances in Neural Information Processing Systems*, 30.

Liu, J., Su, J., Yao, X., Jiang, Z., Lai, G., Du, Y., Qin, Y., Xu, W., Lu, E., Yan, J., et al. (2025). Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*.

Liu, Z., Zhang, J., and Zhou, Z. (2023). Breaking the lower bound with (little) structure: Acceleration in non-convex stochastic optimization with heavy-tailed noise. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2266–2290. PMLR.

Liu, Z. and Zhou, Z. (2023). Stochastic nonsmooth convex optimization with heavy-tailed noises: High-probability bound, in-expectation rate and initial distance adaptation. *arXiv preprint arXiv:2303.12277*.

Liu, Z. and Zhou, Z. (2025). Nonconvex stochastic optimization under heavy-tailed noises: Optimal convergence without gradient clipping. In *International Conference on Learning Representations*, volume 2025, pages 92529–92554.

Lobanov, A., Gasnikov, A., Gorbunov, E., and Takáč, M. (2024). Linear convergence rate in convex setup is possible! gradient descent method variants under (L_0, L_1)-smoothness. *arXiv preprint arXiv:2412.17050*.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Mai, V. V. and Johansson, M. (2021). Stability and convergence of stochastic gradient clipping: Beyond Lipschitz continuity and smoothness. In *International Conference on Machine Learning*, pages 7325–7335. PMLR.

Malinovsky, G., Richtárik, P., Horváth, S., and Gorbunov, E. (2024). Byzantine robustness and partial participation can be achieved at once: Just clip gradient differences. *Advances in Neural Information Processing Systems*, 37:34900–34979.

Nazin, A. V., Nemirovsky, A. S., Tsybakov, A. B., and Juditsky, A. B. (2019). Algorithms of robust stochastic optimization based on mirror descent method. *Automation and Remote Control*, 80(9):1607–1627.

Nemirovski, A. S., Juditsky, A. B., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609.

Nesterov, Y. E. (1984). Minimization methods for nonsmooth convex and quasiconvex functions. *Matekon*, 29(3):519–531.

Nguyen, T. D., Nguyen, T. H., Ene, A., and Nguyen, H. (2023). Improved convergence in high probability of clipped gradient methods with heavy tailed noise.

Orabona, F. (2023). Normalized gradients for all. *arXiv preprint arXiv:2308.05621*.

Parletta, D. A., Paudice, A., Pontil, M., and Salzo, S. (2024). High probability bounds for stochastic subgradient schemes with heavy tailed noise. *SIAM Journal on Mathematics of Data Science*, 6(4):953–977.

Pethick, T., Xie, W., Antonakopoulos, K., Zhu, Z., Silveti-Falls, A., and Cevher, V. (2025). Training deep learning models with norm-constrained lmos. *arXiv preprint arXiv:2502.07529*.

Puchkin, N., Gorbunov, E., Kutuzov, N., and Gasnikov, A. (2024). Breaking the heavy-tailed noise barrier in stochastic optimization problems. In *International Conference on Artificial Intelligence and Statistics*, pages 856–864. PMLR.

Reddi, S. J., Kale, S., and Kumar, S. (2019). On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Sadiev, A., Danilova, M., Gorbunov, E., Horváth, S., Gidel, G., Dvurechensky, P., Gasnikov, A., and Richtárik, P. (2023). High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. In *International conference on machine learning*, pages 29563–29648. PMLR.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Shestakov, A., Takáč, M., and Gorbunov, E. (2026). From optimization to generalization under heavy-tailed data: The role of gradient clipping. In *Forty-third International Conference on Machine Learning*.

Simsekli, U., Sagun, L., and Gurbuzbalaban, M. (2019). A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837. PMLR.

Smith, S. L., Kindermans, P.-J., Ying, C., and Le, Q. V. (2017). Don't decay the learning rate, increase the batch size. *arXiv preprint arXiv:1711.00489*.

Sun, T., Liu, X., and Yuan, K. (2025). Revisiting gradient normalization and clipping for nonconvex sgd under heavy-tailed noise: Necessity, sufficiency, and acceleration. *Journal of Machine Learning Research*, 26(237):1–42.

Takáč, M., Bijral, A., Richtárik, P., and Srebro, N. (2013). Mini-batch primal and dual methods for svms. In *In 30th International Conference on Machine Learning, ICML 2013*.

Tieleman, T. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26.

Tsai, C.-P., Prasad, A., Balakrishnan, S., and Ravikumar, P. (2022). Heavy-tailed streaming statistical estimation. In *International Conference on Artificial Intelligence and Statistics*, pages 1251–1282. PMLR.

Tuddenham, M., Prügel-Bennett, A., and Hare, J. (2022). Orthogonalising gradients to speed up neural network optimisation. *arXiv preprint arXiv:2202.07052*.

Turan, B., Uribe, C. A., Wai, H.-T., and Alizadeh, M. (2021). On robustness of the normalized subgradient method with randomly corrupted subgradients. In *2021 American Control Conference (ACC)*, pages 965–971. IEEE.

Vankov, D., Rodomanov, A., Nedich, A., Sankar, L., and Stich, S. (2025). Optimizing (L_0, L_1)-smooth functions by gradient methods. In *International Conference on Learning Representations*, volume 2025, pages 15953–15979.

Vural, N. M., Yu, L., Balasubramanian, K., Volgushev, S., and Erdogdu, M. A. (2022). Mirror descent strikes again: Optimal stochastic convex optimization under infinite noise variance. In *Conference on Learning Theory*, pages 65–102. PMLR.

Yang, J., Li, X., Fatkhullin, I., and He, N. (2023). Two sides of one coin: the limits of untuned sgd and the power of adaptive methods. *Advances in Neural Information Processing Systems*, 36:74257–74288.

Zhang, B., Jin, J., Fang, C., and Wang, L. (2020a). Improved analysis of clipping algorithms for non-convex optimization. *Advances in Neural Information Processing Systems*, 33:15511–15521.

Zhang, J. and Cutkosky, A. (2022). Parameter-free regret in high probability with heavy tails. *Advances in Neural Information Processing Systems*, 35:8000–8012.

Zhang, J., He, T., Sra, S., and Jadbabaie, A. (2020b). Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*.

Zhang, J., Karimireddy, S. P., Veit, A., Kim, S., Reddi, S., Kumar, S., and Sra, S. (2020c). Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393.

Zhao, S.-Y., Shi, C.-W., Xie, Y.-P., and Li, W.-J. (2024). Stochastic normalized gradient descent with momentum for large-batch training. *Science China Information Sciences*, 67(11):212101.

Zhao, S.-Y., Xie, Y.-P., and Li, W.-J. (2021). On the convergence and improvement of stochastic normalized gradient descent. *Science China Information Sciences*, 64(3):132103.

Contents

1 Introduction **1**

2 Preliminaries **2**

3 Related Works **3**

4 Main Results **4**

4.1 Deterministic case 4

4.2 Stochastic case 6

4.3 Negative results 8

5 Conclusion **10**

A Auxiliary facts and results **17**

B Deterministic Case **20**

B.1 Upper bound without momentum 20

B.2 Upper bounds with inner and outer momentum 20

C Stochastic Case **26**

C.1 Upper bound without momentum 26

C.2 Upper bound with outer momentum 30

D Negative Results and Lower Bounds **37**

A Auxiliary facts and results

Here we introduce the additional notation and auxiliary results.

- $g_t = \frac{1}{B}\sum\limits_{i=1}^{B} \nabla f(x_t, \xi_t^i)$ is the batched stochastic gradient at iteration t;
- $\theta_t = \frac{g_t}{\|g_t\|} - \frac{\nabla f(x_t)}{\|\nabla f(x_t)\|}$ denotes the difference between the normalized stochastic gradient and the normalized true gradient at iteration t;
- $\theta_t^u = \frac{g_t}{\|g_t\|} - \mathbb{E}_t\left[\frac{g_t}{\|g_t\|}\right]$ denotes the unbiased part of θ_t;
- $\theta_t^b = \mathbb{E}_t\left[\frac{g_t}{\|g_t\|}\right] - \frac{\nabla f(x_t)}{\|\nabla f(x_t)\|}$ is the bias part of θ_t.
- $R_t = \|x_t - x^*\|$ – the distance to the optimal point at iteration t.

We first recall a standard consequence of L-smoothness.

Proposition 1. *If the function f is L-smooth, i.e. satisfies Assumption 2, then for all $x \in \mathbb{R}^d$ the next inequality holds:*

$$\|\nabla f(x)\|^2 \leq 2L(f(x) - f^*).$$

The following lemma gives a bound on the momentum estimator.

Lemma 1 (Bound for the momentum sequence)**.** *Suppose that the sequence $\{m_t\}_{t\geq 0}$ is generated by the rule*

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$$

with $m_{-1} = 0$ and $\beta_1 \in [0,1]$. Then, $\|m_t\|^2$ can be upper bounded as

$$\|m_t\|^2 \leq (1-\beta_1^{t+1}) \sum_{j=0}^{t} \beta_1^{t-j}(1-\beta_1)\|g_j\|^2.$$

Proof. This is Lemma C.2 from Chezhegov et al. (2025). □

The following lemma bounds a weighted sum that appears in the analysis of outer momentum.

Lemma 2. *Suppose that $x \in [0,1)$. Let us consider the sum*

$$S_k = \sum_{i=0}^{k-1} \frac{x^i}{\sum_{j=0}^{i} x^j}.$$

Then, we have that for all $k \geq 1$ the following inequality holds:

$$S_k \leq 1 + \ln\left(\frac{1}{1-x}\right).$$

Proof. Let us denote $A_i = \sum\limits_{j=0}^{i} x^j$ with $A_{-1} = 0$. Then, we have

$$S_k = \sum_{i=0}^{k-1} \frac{x^i}{\sum_{j=0}^{i} x^j} = \sum_{i=0}^{k-1} \frac{A_i - A_{i-1}}{A_i} = 1 + \sum_{i=1}^{k-1}\left(1 - \frac{A_{i-1}}{A_i}\right).$$

Moreover, we get

$$1 - \frac{A_{i-1}}{A_i} \leq \ln\left(\frac{A_i}{A_{i-1}}\right)$$

for all $i \geq 1$. Thus, we derive

$$S_k = \sum_{i=0}^{k-1} \frac{x^i}{\sum_{j=0}^{i} x^j} \leq 1 + \sum_{i=1}^{k-1} \ln\left(\frac{A_i}{A_{i-1}}\right) = 1 + \ln(A_{k-1})$$

since $A_0 = 1$. Noting that A_k is uniformly bounded as $\frac{1}{1-x}$, we conclude the proof. □

The next lemma allows us to control the error of mini-batching under heavy-tailed noise.

Lemma 3. *Let $\alpha \in (1, 2]$, and $X_1, \ldots, X_n \in \mathbb{R}^d$ be a martingale difference sequence, i.e.,*

$$\mathbb{E}\left[X_j \mid X_{j-1}, \ldots, X_1\right] = 0 \quad \text{a.s. for all } j = 1, \ldots, n,$$

satisfying

$$\mathbb{E}\left[\|X_j\|^\alpha\right] < \infty \quad \text{for all } j = 1, \ldots, n.$$

Define

$$S_n := \sum_{j=1}^{n} X_j,$$

then

$$\mathbb{E}\left[\|S_n\|^\alpha\right] \leq 2 \sum_{j=1}^{n} \mathbb{E}\left[\|X_j\|^\alpha\right].$$

Proof. This is Lemma 10 from Hübler et al. (2024a). □

As a consequence, the next corollary holds.

Corollary 1. *Let the conditions of Lemma 3 hold. Then, if we take the batch of size B of stochastic variables $X_1, \ldots, X_B$, which satisfies*

$$\mathbb{E}\left[X_j \mid X_{j-1}, \ldots, X_1\right] = 0 \quad \text{a.s. for all } j = 1, \ldots, B,$$

and

$$\mathbb{E}\left[\|X_j\|^\alpha\right] \leq \sigma^\alpha \quad \text{for all } j = 1, \ldots, B,$$

then

$$\mathbb{E}\left[\left\|\frac{1}{B}\sum_{i=1}^{B} X_i\right\|\right] \leq \frac{2\sigma}{B^{\frac{\alpha-1}{\alpha}}}.$$

Proof. The result can be directly obtained from Lemma 3 with Jensen's inequality. □

We also present a new lemma for structures related to normalization; specifically, we provide a bound on the bias part of the normalization operator.

Lemma 4 (Bound on the bias part of normalization)**.** *Let X is a stochastic vector such that*

$$\mathbb{E}[X] = x, \quad x \neq 0 \qquad \text{and} \qquad \mathbb{E}\|X - x\|^\alpha \leq \sigma^\alpha.$$

Then, we have

$$\left\|\mathbb{E}\left[\frac{X}{\|X\|}\right] - \frac{x}{\|x\|}\right\| \leq \frac{8\sigma}{\|x\|}.$$

Proof. For simplicity, let us denote $s := X - x$. Now, let us introduce the following indicators:

$$\mathbb{I}_1 := \mathbb{I}\left\{\|s\| \leq \frac{\|x\|}{2}\right\}, \qquad \mathbb{I}_2 := \mathbb{I}\left\{\|s\| > \frac{\|x\|}{2}\right\}.$$

Therefore, one can decompose the target expression as follows:

$$\mathbb{E}\left[\frac{X}{\|X\|}\right] - \frac{x}{\|x\|} = \mathbb{E}\left[\left(\frac{X}{\|X\|} - \frac{x}{\|x\|}\right)\mathbb{I}_1\right] + \mathbb{E}\left[\left(\frac{X}{\|X\|} - \frac{x}{\|x\|}\right)\mathbb{I}_2\right]$$

since the events in the definition of $\mathbb{I}_1$ and $\mathbb{I}_2$ are *complement*. Thus, due to the triangle inequality, we get

$$\left\|\mathbb{E}\left[\frac{X}{\|X\|}\right] - \frac{x}{\|x\|}\right\| \leq \left\|\mathbb{E}\left[\left(\frac{X}{\|X\|} - \frac{x}{\|x\|}\right)\mathbb{I}_1\right]\right\| + \left\|\mathbb{E}\left[\left(\frac{X}{\|X\|} - \frac{x}{\|x\|}\right)\mathbb{I}_2\right]\right\|. \tag{3}$$

Term related to $\mathbb{I}_1$. Bounding the first term from (3), we have

$$\left\|\mathbb{E}\left[\left(\frac{X}{\|X\|}-\frac{x}{\|x\|}\right)\mathbb{I}_1\right]\right\| \leq \mathbb{E}\left\|\left(\frac{X}{\|X\|}-\frac{x}{\|x\|}\right)\mathbb{I}_1\right\|.$$

Additionally, one can obtain

$$\frac{X}{\|X\|}-\frac{x}{\|x\|} = x\left(\frac{1}{\|x+s\|}-\frac{1}{\|x\|}\right)+\frac{s}{\|x+s\|} = x\left(\frac{\|x\|-\|x+s\|}{\|x\|\,\|x+s\|}\right)+\frac{s}{\|x+s\|}.$$

Consequently, we derive

$$\left\|\frac{X}{\|X\|}-\frac{x}{\|x\|}\right\| \leq \frac{\big|\,\|x\|-\|x+s\|\,\big|}{\|x+s\|}+\frac{\|s\|}{\|x+s\|} \leq \frac{2\,\|s\|}{\|x+s\|},$$

where in the last inequality we also apply a triangle inequality. Moreover, into the event of indicator $\mathbb{I}_1$, we get

$$\|x+s\| \geq \frac{\|x\|}{2}.$$

Therefore, we conclude

$$\left\|\mathbb{E}\left[\left(\frac{X}{\|X\|}-\frac{x}{\|x\|}\right)\mathbb{I}_1\right]\right\| \leq \mathbb{E}\left\|\left(\frac{X}{\|X\|}-\frac{x}{\|x\|}\right)\mathbb{I}_1\right\| \leq \mathbb{E}\left[\frac{4\,\|s\|}{\|x\|}\right] \leq \frac{4\sigma}{\|x\|}. \tag{4}$$

Term related to $\mathbb{I}_2$. Bounding the second term from (3), one can get

$$\left\|\mathbb{E}\left[\left(\frac{X}{\|X\|}-\frac{x}{\|x\|}\right)\mathbb{I}_2\right]\right\| \leq \mathbb{E}\left\|\left(\frac{X}{\|X\|}-\frac{x}{\|x\|}\right)\mathbb{I}_2\right\| = \mathbb{E}\left[\left\|\frac{X}{\|X\|}-\frac{x}{\|x\|}\right\|\mathbb{I}_2\right] \leq 2\mathbb{E}[\mathbb{I}_2]$$

since the random variable $\left\|\frac{X}{\|X\|}-\frac{x}{\|x\|}\right\|$ is uniformly bounded by 2. By the definition of $\mathbb{I}_2$, we obtain the next result using Markov's inequality

$$\mathbb{E}[\mathbb{I}_2] = \mathbb{P}\left\{\|s\| > \frac{\|x\|}{2}\right\} \leq \frac{2\mathbb{E}[\|s\|]}{\|x\|} \leq \frac{2\sigma}{\|x\|},$$

where in the last transition we apply the Jensen's inequality to the convex function x^α: $\mathbb{E}[\|s\|] \leq \sqrt[\alpha]{\mathbb{E}[\|s\|^\alpha]}$. Consequently, we have

$$\left\|\mathbb{E}\left[\left(\frac{X}{\|X\|}-\frac{x}{\|x\|}\right)\mathbb{I}_2\right]\right\| \leq 2\mathbb{E}[\mathbb{I}_2] \leq \frac{4\sigma}{\|x\|}, \tag{5}$$

Substituting (4) and (5) into (3), we finish the proof. □

To obtain high-probability guarantees, we use Bernstein's inequality.

Lemma 5 (Bernstein's inequality)**.** *Let the sequence of random variables* $\{X_i\}_{i\geq 1}$ *form a martingale difference sequence, i.e.,* $\mathbb{E}\left[X_i \mid X_{i-1},\ldots,X_1\right]=0$ *for all* $i \geq 1$. *Assume that conditional variances* $\sigma_i^2 = \mathbb{E}\left[X_i^2 \mid X_{i-1},\ldots,X_1\right]$ *exist and are bounded and also assume that there exists deterministic constant* $c>0$ *such that* $|X_i| \leq c$ *almost surely for all* $i \geq 1$. *Then for all* $b>0$, $G>0$ *and* $n \geq 1$

$$\mathbb{P}\left\{\left|\sum_{i=1}^n X_i\right| > b \text{ and } \sum_{i=1}^n \sigma_i^2 \leq G\right\} \leq 2\exp\left(-\frac{b^2}{2G+\frac{2cb}{3}}\right).$$

B Deterministic Case

In this section, we present proofs of convergence for deterministic methods (Algorithms 1, 2 and 3).

B.1 Upper bound without momentum

Theorem 9 (Convergence theorem for Algorithm 1)**.** *Suppose that Assumptions 1 and 2 hold. Then, after K iterations of Algorithm 1 with $\gamma = \frac{R_0}{\sqrt{K}}$, we have*

$$\min_{k\in 0,\dots,K-1}(f(x_k) - f^*) \leq \frac{2LR_0^2}{K}.$$

Proof. Let us start with the update rule:

$$\|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 - \frac{2\gamma \langle \nabla f(x_k), x_k - x^*\rangle}{\|\nabla f(x_k)\|} + \gamma^2,$$

where we also suppose that $\nabla f(x_k) \neq 0$; otherwise, the final bound holds since $f(x_k) - f^* = 0$. Applying convexity of f (Assumption 1) with L-smoothness (Proposition 1), we derive

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^*\|^2 - \frac{2\gamma \langle \nabla f(x_k), x_k - x^*\rangle}{\|\nabla f(x_k)\|} + \gamma^2 \\ &\leq \|x_k - x^*\|^2 - \frac{2\gamma(f(x_k) - f^*)}{\sqrt{2L(f(x_k) - f^*)}} + \gamma^2 \\ &= \|x_k - x^*\|^2 - \frac{\sqrt{2}\gamma}{\sqrt{L}}\sqrt{f(x_k) - f^*} + \gamma^2,\end{aligned}$$

Rearranging terms gives the following inequality:

$$\frac{\sqrt{2}\gamma}{\sqrt{L}}\sqrt{f(x_k) - f^*} \leq \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \gamma^2.$$

Summing up, we obtain

$$\sum_{k=0}^{K-1}\frac{\sqrt{2}\gamma}{\sqrt{L}}\sqrt{f(x_k) - f^*} \leq \|x_0 - x^*\|^2 - \|x_K - x^*\|^2 + K\gamma^2 \leq \|x_0 - x^*\|^2 + K\gamma^2.$$

Consequently, dividing both parts by $\frac{\sqrt{2}K\gamma}{\sqrt{L}}$, and mentioning that LHS can be bounded below as a minimal value of the weighted sum, we get

$$\min_{k\in 0,\dots,K-1}\sqrt{f(x_k) - f^*} \leq \frac{\sqrt{L}R_0^2}{\sqrt{2}K\gamma} + \frac{\sqrt{L}\gamma}{\sqrt{2}}.$$

Therefore, squaring both sides, we derive

$$\min_{k\in 0,\dots,K-1} f(x_k) - f^* \leq \left(\frac{\sqrt{L}R_0^2}{\sqrt{2}K\gamma} + \frac{\sqrt{L}\gamma}{\sqrt{2}}\right)^2.$$

Substituting $\gamma = \frac{R_0}{\sqrt{K}}$, we conclude the proof. □

B.2 Upper bounds with inner and outer momentum

Theorem 10 (Convergence theorem for Algorithm 2 – inner momentum, Theorem 2)**.** *Suppose that Assumptions 1 and 2 hold. Then, after K iterations of Algorithm 2 with $\gamma = \frac{R_0\sqrt{1-\sqrt{\beta_1}}}{\sqrt{2K}}$, we have*

$$\min_{k\in 0,\dots,K-1}(f(x_k) - f^*) \leq \frac{4LR_0^2}{(1-\sqrt{\beta_1})K}.$$

Proof. Starting with the update rule:

$$\|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 - \frac{2\gamma \langle m_k, x_k - x^* \rangle}{\|m_k\|} + \gamma^2. \tag{6}$$

Let us decompose the scalar product:

$$\begin{aligned} -\langle m_k, x_k - x^* \rangle &= -\beta_1 \langle m_{k-1}, x_k - x^* \rangle - (1-\beta_1) \langle \nabla f(x_k), x_k - x^* \rangle \\ &= -\beta_1 \langle m_{k-1}, x_{k-1} - x^* \rangle + \beta_1 \gamma \|m_{k-1}\| - (1-\beta_1) \langle \nabla f(x_k), x_k - x^* \rangle \\ &= -\sum_{t=0}^{k} (1-\beta_1)\beta_1^{k-t} \langle \nabla f(x_t), x_t - x^* \rangle + \sum_{t=0}^{k-1} \beta_1^{k-t} \gamma \|m_t\| \end{aligned} \tag{7}$$

due to the recursive decomposition of $\langle m_i, x_i - x^* \rangle$. Now, we have two cases. First, if $m_k = 0$, from (7) we have

$$\sum_{t=0}^{k} (1-\beta_1)\beta_1^{k-t} \langle \nabla f(x_t), x_t - x^* \rangle = \sum_{t=0}^{k-1} \beta_1^{k-t} \gamma \|m_t\| = \sum_{t=0}^{k} \beta_1^{k-t} \gamma \|m_t\|. \tag{8}$$

We can upper bound $\|m_t\|$ using Lemma 1 as follows:

$$\|m_t\| \leq \sqrt{(1-\beta_1^{t+1}) \sum_{j=0}^{t} (1-\beta_1)\beta_1^{t-j} \|\nabla f(x_j)\|^2}. \tag{9}$$

Thus, one can get

$$\begin{aligned} \sum_{t=0}^{k} \beta_1^{k-t} \gamma \|m_t\| &\leq \sum_{t=0}^{k} \beta_1^{k-t} \gamma \sqrt{(1-\beta_1^{t+1}) \sum_{j=0}^{t} (1-\beta_1)\beta_1^{t-j} \|\nabla f(x_j)\|^2} \\ &\leq \sum_{t=0}^{k} \beta_1^{k-t} \gamma \sqrt{(1-\beta_1^{k+1}) \sum_{j=0}^{t} (1-\beta_1)\beta_1^{t-j} \|\nabla f(x_j)\|^2} \\ &= \sum_{t=0}^{k} \beta_1^{\frac{k-t}{2}} \gamma \sqrt{(1-\beta_1^{k+1}) \sum_{j=0}^{t} (1-\beta_1)\beta_1^{k-j} \|\nabla f(x_j)\|^2}, \end{aligned}$$

where we use $\beta_1^{t+1} \geq \beta_1^{k+1}$ and put $\beta_1^{\frac{k-t}{2}}$ under the square root. Next, we apply Jensen's inequality to the concave function $\sqrt{x}$ to derive

$$\begin{aligned} &\sum_{t=0}^{k} \frac{\beta_1^{\frac{k-t}{2}}}{C(\beta_1, k)} \gamma \sqrt{(1-\beta_1^{k+1}) \sum_{j=0}^{t} (1-\beta_1)\beta_1^{k-j} \|\nabla f(x_j)\|^2} \\ &\leq \gamma \sqrt{(1-\beta_1^{k+1}) \sum_{t=0}^{k} \frac{\beta_1^{\frac{k-t}{2}}}{C(\beta_1, k)} \sum_{j=0}^{t} (1-\beta_1)\beta_1^{k-j} \|\nabla f(x_j)\|^2}, \end{aligned}$$

where $C(\beta_1, k) = \sum\limits_{i=0}^{k} \beta_1^{\frac{i}{2}} = \frac{1-\sqrt{\beta_1}^{k+1}}{1-\sqrt{\beta_1}} \leq \frac{1}{1-\sqrt{\beta_1}}$ is the normalization parameter for Jensen's inequality. Thus, we have

$$\sum_{t=0}^{k} \beta_1^{k-t} \gamma \|m_t\| \leq \gamma \sqrt{(1-\beta_1^{k+1}) C(\beta_1, k) \sum_{t=0}^{k} \beta_1^{\frac{k-t}{2}} \sum_{j=0}^{t} (1-\beta_1)\beta_1^{k-j} \|\nabla f(x_j)\|^2}.$$

Next, we can estimate the coefficient for $\|\nabla f(x_r)\|^2$ for all $r = 0, \ldots, k$ to obtain

$$\begin{aligned}\sum_{t=0}^{k} \beta_1^{k-t}\gamma \|m_t\| &\leq \gamma \sqrt{(1-\beta_1^{k+1})C(\beta_1,k)\sum_{t=0}^{k}\beta_1^{\frac{k-t}{2}}\sum_{j=0}^{t}(1-\beta_1)\beta_1^{k-j}\|\nabla f(x_j)\|^2} \\ &= \gamma \sqrt{(1-\beta_1^{k+1})C(\beta_1,k)\sum_{r=0}^{k}\sum_{t=r}^{k}(1-\beta_1)\beta_1^{k-r}\beta_1^{\frac{k-t}{2}}\|\nabla f(x_r)\|^2} \\ &\leq \frac{\gamma}{1-\sqrt{\beta_1}} \sqrt{(1-\beta_1^{k+1})\sum_{r=0}^{k}(1-\beta_1)\beta_1^{k-r}\|\nabla f(x_r)\|^2}, \end{aligned} \tag{10}$$

where in the last inequality we use $C(\beta_1, k) \leq \frac{1}{1-\sqrt{\beta_1}}$ and $\sum\limits_{t=r}^{k} \beta_1^{\frac{k-t}{2}} \leq \frac{1}{1-\sqrt{\beta_1}}$. Consequently, substituting (10) into (8) and applying the convexity to LHS, we have

$$\sum_{t=0}^{k}(1-\beta_1)\beta_1^{k-t}(f(x_t)-f^*) \leq \frac{\gamma}{1-\sqrt{\beta_1}} \sqrt{(1-\beta_1^{k+1})\sum_{t=0}^{k}(1-\beta_1)\beta_1^{k-t}\|\nabla f(x_t)\|^2}.$$

Applying Proposition 1, one can get

$$\sum_{t=0}^{k}(1-\beta_1)\beta_1^{k-t}(f(x_t)-f^*) \leq \frac{\gamma}{1-\sqrt{\beta_1}} \sqrt{2L(1-\beta_1^{k+1})\sum_{t=0}^{k}(1-\beta_1)\beta_1^{k-t}(f(x_t)-f^*)}.$$

Solving the inequality, we derive

$$\sum_{t=0}^{k}(1-\beta_1)\beta_1^{k-t}(f(x_t)-f^*) \leq \frac{2L\gamma^2(1-\beta_1^{k+1})}{(1-\sqrt{\beta_1})^2}.$$

Dividing both sides by $(1-\beta_1^{k+1})$ and noting that LHS is the convex combination of nonnegative terms, we conclude

$$\min_{k=0,\ldots,K-1}(f(x_k)-f^*) \leq \min_{t=0,\ldots,k}(f(x_t)-f^*) \leq \frac{2L\gamma^2}{(1-\sqrt{\beta_1})^2} = \frac{LR_0^2}{(1-\sqrt{\beta_1})K},$$

where in the last equation we substitute the choice of γ. As a result, if for some $k = 0, \ldots, K-1$ $m_k = 0$, then we obtain the desired result. Otherwise, for all $k = 0, \ldots, K-1$ $m_k \neq 0$; substitution of (7) into (6) gives

$$\begin{aligned}\|x_{k+1}-x^*\|^2 &= \|x_k-x^*\|^2 - \frac{2\gamma}{\|m_k\|}\sum_{t=0}^{k}(1-\beta_1)\beta_1^{k-t}\langle\nabla f(x_t), x_t-x^*\rangle \\ &\quad + \sum_{t=0}^{k-1}\frac{2\beta_1^{k-t}\gamma^2\|m_t\|}{\|m_k\|} + \gamma^2 \\ &\leq \|x_k-x^*\|^2 - \frac{2\gamma}{\|m_k\|}\left[\sum_{t=0}^{k}(1-\beta_1)\beta_1^{k-t}\langle\nabla f(x_t), x_t-x^*\rangle - \sum_{t=0}^{k}\beta_1^{k-t}\gamma\|m_t\|\right],\end{aligned}$$

where we use $\gamma^2 \leq 2\gamma^2\frac{\|m_k\|}{\|m_k\|}$. Next, we can upper bound the scalar product by convexity (Assumption 1) to obtain

$$\|x_{k+1}-x^*\|^2 \leq \|x_k-x^*\|^2 - \frac{2\gamma}{\|m_k\|}\left[\sum_{t=0}^{k}(1-\beta_1)\beta_1^{k-t}(f(x_t)-f^*) - \sum_{t=0}^{k}\beta_1^{k-t}\gamma\|m_t\|\right].$$

To bound the RHS term related to $\|m_t\|$, we apply (10) to get

$$\begin{aligned}\|x_{k+1}-x^*\|^2 &\leq \|x_k-x^*\|^2 - \frac{2\gamma}{\|m_k\|}\sum_{t=0}^{k}(1-\beta_1)\beta_1^{k-t}(f(x_t)-f^*) \\ &\quad + \frac{2\gamma^2}{(1-\sqrt{\beta_1})\|m_k\|}\sqrt{(1-\beta_1^{k+1})\sum_{t=0}^{k}(1-\beta_1)\beta_1^{k-t}\|\nabla f(x_t)\|^2}.\end{aligned}$$

Moreover, due to Proposition 1, one can have

$$\|x_{k+1}-x^*\|^2 \leq \|x_k-x^*\|^2 - \frac{2\gamma}{\|m_k\|}\sum_{t=0}^{k}(1-\beta_1)\beta_1^{k-t}(f(x_t)-f^*) + \frac{2\gamma^2}{(1-\sqrt{\beta_1})\|m_k\|}\sqrt{2L(1-\beta_1^{k+1})\sum_{t=0}^{k}(1-\beta_1)\beta_1^{k-t}(f(x_t)-f^*)}.$$

Now, let us denote $A_k = (1-\beta_1^{k+1})\sum\limits_{t=0}^{k}(1-\beta_1)\beta_1^{k-t}(f(x_t)-f^*)$. Then, our expression can be rewritten in the following way:

$$\begin{aligned}\|x_{k+1}-x^*\|^2 &\leq \|x_k-x^*\|^2 - \frac{2\gamma}{(1-\beta_1^{k+1})\|m_k\|}A_k + \frac{2\gamma^2\sqrt{2L}}{(1-\sqrt{\beta_1})\|m_k\|}\sqrt{A_k}\\ &= \|x_k-x^*\|^2 - \frac{2\gamma}{\|m_k\|}\left(\frac{A_k}{1-\beta_1^{k+1}} - \frac{\sqrt{2L}\gamma\sqrt{A_k}}{(1-\sqrt{\beta_1})}\right). \end{aligned} \tag{11}$$

Now, we are ready to show via contradiction that (11) allows us to obtain the convergence guarantees. First, let us consider two cases.

Case 1: $-\frac{A_k}{1-\beta_1^{k+1}} + \frac{\sqrt{2L}\gamma\sqrt{A_k}}{(1-\sqrt{\beta_1})} \geq -\frac{A_k}{2(1-\beta_1^{k+1})}$. For this case, one can obtain

$$A_k \leq \frac{8L\gamma^2(1-\beta_1^{k+1})^2}{(1-\sqrt{\beta_1})^2} = \frac{4LR_0^2(1-\beta_1^{k+1})^2}{(1-\sqrt{\beta_1})K}.$$

by solving the quadratic inequality and substituting the choice of γ. By the definition of A_k, we derive

$$\sum_{t=0}^{k}\frac{(1-\beta_1)\beta_1^{k-t}}{1-\beta_1^{k+1}}(f(x_t)-f^*) \leq \frac{4LR_0^2}{(1-\sqrt{\beta_1})K},$$

which implies

$$\min_{k=0,\ldots,K-1} f(x_k)-f^* \leq \min_{t=0,\ldots,k} f(x_t)-f^* \leq \frac{4LR_0^2}{(1-\sqrt{\beta_1})K}, \tag{12}$$

where we use the property of convex combination.

Case 2: $-\frac{A_k}{1-\beta_1^{k+1}} + \frac{\sqrt{2L}\gamma\sqrt{A_k}}{(1-\sqrt{\beta_1})} < -\frac{A_k}{2(1-\beta_1^{k+1})}$. This case implies

$$\|x_{k+1}-x^*\|^2 \leq \|x_k-x^*\|^2 - \frac{\gamma}{(1-\beta_1^{k+1})\|m_k\|}A_k.$$

Moreover, from (9) we have

$$\begin{aligned}\|m_k\| &\leq \sqrt{(1-\beta_1^{k+1})\sum_{t=0}^{k}(1-\beta_1)\beta_1^{k-t}\|\nabla f(x_t)\|^2}\\ &\leq \sqrt{2L(1-\beta_1^{k+1})\sum_{t=0}^{k}(1-\beta_1)\beta_1^{k-t}(f(x_t)-f^*)} = \sqrt{2LA_k},\end{aligned}$$

where we apply Proposition 1. As a consequence, *Case 2* implies

$$\|x_{k+1}-x^*\|^2 \leq \|x_k-x^*\|^2 - \frac{\gamma\sqrt{A_k}}{\sqrt{2L}(1-\beta_1^{k+1})}.$$

Next, *only two options* are possible. The first one: there exists $k_0 \in \overline{0, K-1}$ such that A_{k_0} satisfies *Case 1*. This immediately leads to a final result

$$\min_{k=0,\ldots,K-1} f(x_k)-f^* \leq \frac{8L\gamma^2}{(1-\sqrt{\beta_1})^2} = \frac{4LR_0^2}{(1-\sqrt{\beta_1})K}$$

due to (12). Otherwise, let us assume that for all $k = 0, \ldots, K-1$ the *Case 2* holds. Thus, we have

$$\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - \frac{\gamma\sqrt{A_k}}{\sqrt{2L}(1-\beta_1^{k+1})},$$

and, after summing up and rearranging terms, we get

$$\sum_{k=0}^{K-1} \frac{\gamma\sqrt{A_k}}{\sqrt{2L}(1-\beta_1^{k+1})} \leq \|x_0 - x^*\|^2 - \|x_K - x^*\|^2 \leq R_0^2.$$

Dividing both sides by $\frac{\gamma K}{\sqrt{2L}}$, one can obtain

$$\frac{1}{K}\sum_{k=0}^{K-1} \frac{\sqrt{A_k}}{(1-\beta_1^{k+1})} \leq \frac{\sqrt{2L}R_0^2}{\gamma K}.$$

Therefore, the following inequality holds:

$$\min_{k=0,\ldots,K-1} \frac{\sqrt{A_k}}{(1-\beta_1^{k+1})} \leq \frac{\sqrt{2L}R_0^2}{\gamma K}.$$

Squaring both sides, we have

$$\min_{k=0,\ldots,K-1} \frac{A_k}{(1-\beta_1^{k+1})^2} \leq \frac{2LR_0^4}{\gamma^2 K^2} = \frac{4LR_0^2}{(1-\sqrt{\beta_1})K},$$

where we substitute the choice of γ. Therefore, there exists $\tilde{k} \in \overline{0, K-1}$ such that $A_{\tilde{k}} \leq \frac{4LR_0^2(1-\beta_1^{\tilde{k}+1})^2}{(1-\sqrt{\beta_1})K}$, which means that $A_{\tilde{k}}$ satisfies *Case 1* and arises a contradiction. Consequently, after K iterations of Algorithm 2 with $\gamma = \frac{R_0\sqrt{1-\sqrt{\beta_1}}}{\sqrt{2K}}$ there always exists $k_0 \in \overline{0, K-1}$ such that A_{k_0} satisfies *Case 1*. Applying (12), we conclude the proof. □

Theorem 11 (Convergence theorem for Algorithm 3 – outer momentum, Theorem 3)**.** *Suppose that Assumptions 1 and 2 hold. Then, after K iterations of Algorithm 3 with $\gamma = \frac{R_0\sqrt{1-\beta_1}}{\sqrt{2K}}$, we have*

$$\min_{k\in 0,\ldots,K-1} (f(x_k) - f^*) \leq \frac{4LR_0^2}{(1-\beta_1)K}.$$

Proof. Starting with the update rule:

$$\|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 - 2\gamma \langle \tilde{m}_k, x_k - x^* \rangle + \gamma^2 \|\tilde{m}_k\|^2. \tag{13}$$

Decomposing the scalar product recursively with $m_{-1} = 0$ and assuming that $\nabla f(x_k) \neq 0$ for all $k = 0, \ldots, K-1$ (otherwise, the final bound obviously holds), we derive

$$\begin{aligned}
-\langle \tilde{m}_k, x_k - x^* \rangle &= -\frac{1}{1-\beta_1^{k+1}} \langle m_k, x_k - x^* \rangle \\
&= -\frac{1}{1-\beta_1^{k+1}} \left(\beta_1 \langle m_{k-1}, x_k - x^* \rangle + (1-\beta_1) \frac{\langle \nabla f(x_k), x_k - x^* \rangle}{\|\nabla f(x_k)\|} \right) \\
&= -\frac{\beta_1}{1-\beta_1^{k+1}} \langle m_{k-1}, x_{k-1} - x^* \rangle + \frac{\beta_1}{1-\beta_1^{k+1}} \gamma \langle m_{k-1}, \tilde{m}_{k-1} \rangle \\
&\quad - \frac{1-\beta_1}{1-\beta_1^{k+1}} \frac{\langle \nabla f(x_k), x_k - x^* \rangle}{\|\nabla f(x_k)\|} \\
&= -\frac{\beta_1}{1-\beta_1^{k+1}} \langle m_{k-1}, x_{k-1} - x^* \rangle + \frac{\beta_1}{1-\beta_1^{k+1}} \gamma \|\tilde{m}_{k-1}\|^2 (1-\beta_1^k) \\
&\quad - \frac{1-\beta_1}{1-\beta_1^{k+1}} \frac{\langle \nabla f(x_k), x_k - x^* \rangle}{\|\nabla f(x_k)\|} \\
&= \sum_{t=0}^{k} -\frac{(1-\beta_1)\beta_1^{k-t}}{1-\beta_1^{k+1}} \frac{\langle \nabla f(x_t), x_t - x^* \rangle}{\|\nabla f(x_t)\|} + \sum_{t=0}^{k-1} \frac{\beta_1^{k-t}}{1-\beta_1^{k+1}} \gamma \|\tilde{m}_t\|^2 (1-\beta_1^{t+1}).
\end{aligned}$$

Substituting this result into (13) with the bound $\gamma^2\|\tilde{m}_k\|^2 \leq 2\gamma^2\|\tilde{m}_k\|^2$, one can get

$$\|x_{k+1}-x^*\|^2 \leq \|x_k-x^*\|^2 - 2\gamma\sum_{t=0}^{k}\frac{(1-\beta_1)\beta_1^{k-t}}{1-\beta_1^{k+1}}\frac{\langle\nabla f(x_t), x_t-x^*\rangle}{\|\nabla f(x_t)\|} + \sum_{t=0}^{k}\frac{2\beta_1^{k-t}(1-\beta_1^{t+1})}{1-\beta_1^{k+1}}\gamma^2\|\tilde{m}_t\|^2.$$

Applying the convexity (Assumption 1) and L-smoothness (Proposition 1), we have

$$\|x_{k+1}-x^*\|^2 \leq \|x_k-x^*\|^2 - \frac{\sqrt{2}\gamma}{\sqrt{L}}\sum_{t=0}^{k}\frac{(1-\beta_1)\beta_1^{k-t}}{1-\beta_1^{k+1}}\sqrt{f(x_t)-f^*} + \gamma^2\sum_{t=0}^{k}\frac{2\beta_1^{k-t}(1-\beta_1^{t+1})}{1-\beta_1^{k+1}}$$

since $\|\tilde{m}_t\| \leq 1$ due to the fact that $\tilde{m}_t$ is the convex combination of normalized vectors. Moreover, one can obtain

$$\sum_{t=0}^{k}\frac{2\beta_1^{k-t}(1-\beta_1^{t+1})}{1-\beta_1^{k+1}} \leq \frac{2}{1-\beta_1^{k+1}}\sum_{t=0}^{k}\beta_1^{k-t} \leq \frac{2}{1-\beta_1}. \tag{14}$$

Consequently, rearranging terms, we get

$$\frac{\sqrt{2}\gamma}{\sqrt{L}}\sum_{t=0}^{k}\frac{(1-\beta_1)\beta_1^{k-t}}{1-\beta_1^{k+1}}\sqrt{f(x_t)-f^*} \leq \|x_k-x^*\|^2 - \|x_{k+1}-x^*\|^2 + \frac{2\gamma^2}{1-\beta_1}.$$

Noting that the LHS is the convex combination of $\sqrt{f(x_t)-f^*}$, we have

$$\min_{t=0,\ldots,k}\frac{\sqrt{2}\gamma}{\sqrt{L}}\sqrt{f(x_t)-f^*} \leq \|x_k-x^*\|^2 - \|x_{k+1}-x^*\|^2 + \frac{2\gamma^2}{1-\beta_1}.$$

Summing up, one can derive

$$\sum_{k=0}^{K-1}\min_{t=0,\ldots,k}\frac{\sqrt{2}\gamma}{\sqrt{L}}\sqrt{f(x_t)-f^*} \leq \|x_0-x^*\|^2 + \frac{2K\gamma^2}{1-\beta_1}. \tag{15}$$

Moreover, the LHS can be lower bounded as

$$K\min_{k=0,\ldots,K-1}\sqrt{f(x_k)-f^*} \leq \sum_{k=0}^{K-1}\min_{t=0,\ldots,k}\sqrt{f(x_t)-f^*}$$

since $\min_{k=0,\ldots,K-1}\sqrt{f(x_k)-f^*} \leq \min_{t=0,\ldots,k}\sqrt{f(x_t)-f^*}$. Therefore, substituting the lower bound on LHS of (15) and dividing both sides by $\frac{\sqrt{2}\gamma K}{\sqrt{L}}$, we obtain

$$\min_{k=0,\ldots,K-1}\sqrt{f(x_k)-f^*} \leq \frac{\sqrt{L}R_0^2}{\sqrt{2}\gamma K} + \frac{\sqrt{2}\sqrt{L}\gamma}{1-\beta_1}.$$

Choosing $\gamma = \frac{R_0\sqrt{1-\beta_1}}{\sqrt{2K}}$, we conclude

$$\min_{k=0,\ldots,K-1}\sqrt{f(x_k)-f^*} \leq \frac{2\sqrt{L}R_0}{\sqrt{1-\beta_1}\sqrt{K}}.$$

Squaring both sides, one can get the final bound as follows:

$$\min_{k=0,\ldots,K-1}(f(x_k)-f^*) \leq \frac{4LR_0^2}{(1-\beta_1)K}.$$

This finishes the proof. □

C Stochastic Case

C.1 Upper bound without momentum

First, let us provide a descent lemma.

Lemma 6 (Descent lemma)**.** *Suppose that Assumptions 1 and 2 hold. Then, if we assume that* $\nabla f(x_k) \neq 0$ *for all* $k = 0, \ldots, K-1$*, after* K *iterations of Algorithm 4, we have*

$$\sum_{k=0}^{K-1} \frac{2\gamma(f(x_k) - f^*)}{\|\nabla f(x_k)\|} \leq \|x_0 - x^*\|^2 - \|x_K - x^*\|^2 + K\gamma^2 - \sum_{k=0}^{K-1} 2\gamma \left\langle \theta_k^u, x_k - x^* \right\rangle - \sum_{k=0}^{K-1} 2\gamma \left\langle \theta_k^b, x_k - x^* \right\rangle$$

Proof. This result is closely related to Theorem 9. Starting with the update rule and using the notation, we get

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &\leq \|x_k - x^*\|^2 - 2\gamma \left\langle \frac{g_k}{\|g_k\|}, x_k - x^* \right\rangle + \gamma^2 \\
&= \|x_k - x^*\|^2 - 2\gamma \left\langle \frac{\nabla f(x_k)}{\|\nabla f(x_k)\|}, x_k - x^* \right\rangle + \gamma^2 \\
&\quad - 2\gamma \left\langle \frac{g_k}{\|g_k\|} - \frac{\nabla f(x_k)}{\|\nabla f(x_k)\|}, x_k - x^* \right\rangle.
\end{aligned}$$

Therefore, one can obtain

$$\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - \frac{2\gamma(f(x_k) - f^*)}{\|\nabla f(x_k)\|} - 2\gamma \left\langle \frac{g_k}{\|g_k\|} - \frac{\nabla f(x_k)}{\|\nabla f(x_k)\|}, x_k - x^* \right\rangle + \gamma^2,$$

where we use $-\langle \nabla f(x_k), x_k - x^* \rangle \leq -(f(x_k) - f^*)$ due to convexity. By the definition of θ_k, we derive

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &\leq \|x_k - x^*\|^2 - \frac{2\gamma(f(x_k) - f^*)}{\|\nabla f(x_k)\|} - 2\gamma \left\langle \theta_k, x_k - x^* \right\rangle + \gamma^2 \\
&= \|x_k - x^*\|^2 - \frac{2\gamma(f(x_k) - f^*)}{\|\nabla f(x_k)\|} + \gamma^2 \\
&\quad - 2\gamma \left\langle \theta_k^u, x_k - x^* \right\rangle - 2\gamma \left\langle \theta_k^b, x_k - x^* \right\rangle,
\end{aligned}$$

where in the last equation we decompose θ_k according to the notation. Summing up and rearranging terms lead to the following inequality:

$$\sum_{k=0}^{K-1} \frac{2\gamma(f(x_k) - f^*)}{\|\nabla f(x_k)\|} \leq \|x_0 - x^*\|^2 - \|x_K - x^*\|^2 + K\gamma^2 - \sum_{k=0}^{K-1} 2\gamma \left\langle \theta_k^u, x_k - x^* \right\rangle - \sum_{k=0}^{K-1} 2\gamma \left\langle \theta_k^b, x_k - x^* \right\rangle.$$

This concludes the proof. □

Theorem 12. *Suppose that Assumptions 1, 2 and 3 hold. Then, after* K *iterations with* $\ln\left(\frac{2K}{\delta}\right) \geq 1$ *of Algorithm 4 with* $\gamma = \frac{R_0}{32\sqrt{K}\ln\left(\frac{2K}{\delta}\right)}$ *and batch size* $B_k \equiv B = \left\lceil \max\left\{1, \left(\frac{\sigma K}{128 L R_0 \ln^2\left(\frac{2K}{\delta}\right)}\right)^{\frac{\alpha}{\alpha-1}}\right\} \right\rceil$*, we have*

$$\min_{k=0,\ldots,K-1} (f(x_k) - f^*) \leq \frac{8 \cdot 32^2 L R_0^2 \ln^2\left(\frac{2K}{\delta}\right)}{K}$$

with probability at least $1 - \delta$*.*

Proof. The proof controls the union of a success event and an event on which the iterates remain in a bounded neighborhood of x^*. This approach is similar to the ideas used in a number of papers (Sadiev et al., 2023; Gorbunov et al., 2024b). First, let us introduce the probabilistic good events E_k :

$$
\begin{aligned}
-\sum_{l=0}^{t-1} 2\gamma \left\langle \theta_l^u, x_l - x^* \right\rangle &\leq \frac{R_0^2}{2}; \\
-\sum_{l=0}^{t-1} 2\gamma \left\langle \theta_l^b, x_l - x^* \right\rangle &\leq \sum_{l=0}^{t-1} \frac{64\gamma R_0 \sigma}{B^{\frac{\alpha-1}{\alpha}} \|\nabla f(x_l)\|}; \\
R_t &\leq \sqrt{2} R_0
\end{aligned}
$$

for all $t = 0, \ldots, k$, and success events S_k:

$$
\min_{t=0,\ldots,k-1} (f(x_t) - f^*) \leq \frac{8 \cdot 32^2 L R_0^2 \ln^2\left(\frac{2K}{\delta}\right)}{K}.
$$

Here $S_0 = \varnothing$. In the definition of E_k, the first two inequalities are required only for those t for which $\nabla f(x_j) \neq 0$ for every $j < t$; the radius bound is required for every $t \leq k$. Thus, all displayed ratios are well defined whenever they are used. We want to show via induction that $\mathbb{P}\{S_k \cup E_k\} \geq 1 - \frac{k\delta}{K}$. The main idea can be expressed as follows: the event E_k **does not imply a convergence bound itself**; thus, we also need to control the convergence events. Now we are ready to start the proof. For $k = 0$ it is obvious that $\mathbb{P}\{S_0 \cup E_0\} \geq 1$. Now, let us assume that $\mathbb{P}\{S_{T-1} \cup E_{T-1}\} \geq 1 - \frac{(T-1)\delta}{K}$. Since $S_{T-1} \subseteq S_T$, we have $(S_{T-1} \cup E_{T-1}) \cap S_T \cup (\overline{S_T} \cap E_{T-1}) = S_{T-1} \cup E_{T-1}$. Moreover, $(S_{T-1} \cup E_{T-1}) \cap S_T \subseteq S_T \cup E_T$. Thus, it is enough to consider the event $\overline{S_T} \cap E_{T-1}$. This event implies $\nabla f(x_t) \neq 0$ for all $t = 0, \ldots, T-1$. Therefore, we can apply Lemma 6 to get

$$
\sum_{l=0}^{t-1} \frac{2\gamma(f(x_l) - f^*)}{\|\nabla f(x_l)\|} \leq R_0^2 - R_t^2 - \sum_{l=0}^{t-1} 2\gamma \left\langle \theta_l^u, x_l - x^* \right\rangle - \sum_{l=0}^{t-1} 2\gamma \left\langle \theta_l^b, x_l - x^* \right\rangle + t\gamma^2
$$

for all $t = 0, \ldots, T$. Moreover, the event $\overline{S_T} \cap E_{T-1}$ implies

$$
\begin{aligned}
\sum_{l=0}^{t-1} \frac{2\gamma(f(x_l) - f^*)}{\|\nabla f(x_l)\|} &\leq R_0^2 - R_t^2 - \sum_{l=0}^{t-1} 2\gamma \left\langle \theta_l^u, x_l - x^* \right\rangle - \sum_{l=0}^{t-1} 2\gamma \left\langle \theta_l^b, x_l - x^* \right\rangle + t\gamma^2 \\
&\leq R_0^2 - R_t^2 + \frac{R_0^2}{2} + \sum_{l=0}^{t-1} \frac{64\gamma R_0 \sigma}{B^{\frac{\alpha-1}{\alpha}} \|\nabla f(x_l)\|} + \frac{t R_0^2}{32^2 K} \\
&\leq 2R_0^2 - R_t^2 + \sum_{l=0}^{t-1} \frac{8 \cdot 32^2 \gamma L R_0^2 \ln^2\left(\frac{2K}{\delta}\right)}{K \|\nabla f(x_l)\|}
\end{aligned}
$$

for $t = 0, \ldots, T-1$, since $B \geq \left(\frac{\sigma K}{128 L R_0 \ln^2\left(\frac{2K}{\delta}\right)}\right)^{\frac{\alpha}{\alpha-1}}$. Since we consider the event $\overline{S_T} \cap E_{T-1}$, then we have

$$
f(x_t) - f^* > \frac{8 \cdot 32^2 L R_0^2 \ln^2\left(\frac{2K}{\delta}\right)}{K}
$$

for all $t = 0, \ldots, T-2$, which leads to

$$
\sum_{l=0}^{t-1} \frac{\gamma(f(x_l) - f^*)}{\|\nabla f(x_l)\|} \leq 2R_0^2 - R_t^2
$$

on the event $\overline{S_T} \cap E_{T-1}$ for all $t = 0, \ldots, T-1$. Using that the LHS is non-negative, we derive that $\overline{S_T} \cap E_{T-1}$ implies

$$
\sum_{l=0}^{t-1} \frac{\gamma(f(x_l) - f^*)}{\|\nabla f(x_l)\|} \leq 2R_0^2
$$

and

$$R_t^2 \leq 2R_0^2$$

for all $t = 0, \ldots, T-1$. Next, let us define random vector η_t for all $t = 0, \ldots, T-1$ as follows:

$$\eta_t = \begin{cases} x_t - x^*, & \|x_t - x^*\| \leq \sqrt{2}R_0; \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

Therefore, $\|\eta_t\| \leq \sqrt{2}R_0$ with probability 1. Moreover, the event $\overline{S_T} \cap E_{T-1}$ implies $R_t \leq \sqrt{2}R_0$, which allows to use truncated vectors η_t to obtain

$$\sum_{t=0}^{T-1} \frac{2\gamma(f(x_t) - f^*)}{\|\nabla f(x_t)\|} \leq R_0^2 - R_T^2 - \underbrace{\sum_{t=0}^{T-1} 2\gamma \left\langle \theta_t^u, \eta_t \right\rangle}_{①} - \underbrace{\sum_{t=0}^{T-1} 2\gamma \left\langle \theta_t^b, \eta_t \right\rangle}_{②} + T\gamma^2.$$

Now, we are ready to provide upper bounds for ① and ②.

Upper bound for ①. At first, we get

$$\mathbb{E}_t[-2\gamma \left\langle \theta_t^u, \eta_t \right\rangle] = 0$$

due to the definition of θ_t^u. Moreover, we have

$$|-2\gamma \left\langle \theta_t^u, \eta_t \right\rangle| \leq 8\gamma R_0 \leq \frac{9R_0^2}{16 \ln\left(\frac{2K}{\delta}\right)}$$

due to the choice of γ. Additionally, we have bounded conditional variance for each term:

$$\sigma_t^2 = \mathbb{E}_t[4\gamma^2 \left\langle \theta_t^u, \eta_t \right\rangle^2] \leq 8\gamma^2 R_0^2 \mathbb{E}_t[\|\theta_t^u\|^2].$$

Therefore, one can apply Bernstein's inequality with $b = \frac{R_0^2}{2}$ and $G = \frac{R_0^4}{32 \ln\left(\frac{2K}{\delta}\right)}$ to derive

$$\mathbb{P}\left\{ |①| > b \text{ and } \sum_{t=0}^{T-1} \sigma_t^2 \leq G \right\} \leq 2 \exp\left(-\frac{b^2}{2G + {}^{2cb}/_3} \right) = \frac{\delta}{K}.$$

Thus, we get

$$\mathbb{P}\left\{ |①| \leq b \text{ or } \sum_{t=0}^{T-1} \sigma_t^2 > G \right\} \geq 1 - \frac{\delta}{K}.$$

Moreover, we have

$$\sum_{t=0}^{T-1} \sigma_t^2 \leq \sum_{t=0}^{T-1} 8\gamma^2 R_0^2 \mathbb{E}_t[\|\theta_t^u\|^2] \leq 32\gamma^2 R_0^2 T \leq \frac{R_0^4}{32 \ln\left(\frac{2K}{\delta}\right)} = G.$$

Upper bound for ②. The event $\overline{S_T} \cap E_{T-1}$ implies

$$-\sum_{t=0}^{T-1} 2\gamma \left\langle \theta_t^b, \eta_t \right\rangle \leq \sum_{t=0}^{T-1} 2\gamma \left\|\theta_t^b\right\| \|\eta_t\| \overset{\text{Lemma 4, Corollary 1}}{\leq} \sum_{t=0}^{T-1} \frac{64\gamma R_0 \sigma}{B^{\frac{\alpha-1}{\alpha}} \|\nabla f(x_t)\|},$$

where in the last inequality we apply Lemma 4 to bound $\left\|\theta_t^b\right\|$ (it is possible due to the fact that $\overline{S}_T$ implies $\nabla f(x_t) \neq 0$ for all $t = 0, \ldots, T-1$) and Corollary 1 to derive $\sigma \to \frac{2\sigma}{B^{\frac{\alpha-1}{\alpha}}}$. Consequently, the event $\overline{S_T} \cap E_{T-1} \cap E_{①}$, where

$$E_{①} = \left\{ |①| \leq \frac{R_0^2}{2} \text{ or } \sum_{t=0}^{T-1} \sigma_t^2 > \frac{R_0^4}{32 \ln\left(\frac{2K}{\delta}\right)} \right\},$$

implies

$$\sum_{t=0}^{T-1} \frac{2\gamma(f(x_t) - f^*)}{\|\nabla f(x_t)\|} \leq R_0^2 - R_T^2 + \frac{R_0^2}{2} + \sum_{t=0}^{T-1} \frac{64\gamma R_0 \sigma}{B^{\frac{\alpha-1}{\alpha}} \|\nabla f(x_t)\|} + \frac{TR_0^2}{32^2 K}. \tag{17}$$

Moreover, we can derive the following inclusion:

$$(S_{T-1} \cup E_{T-1}) \cap S_T \cup (\overline{S}_T \cap E_{T-1} \cap E_{①}) \subseteq S_T \cup E_T.$$

Indeed, we have $(S_{T-1} \cup E_{T-1}) \cap S_T \subseteq S_T$; otherwise, into the event $\overline{S_T} \cap E_{T-1} \cap E_{①}$ we get for all $t = 0, \ldots, T-1$:

$$f(x_t) - f^* > \frac{8 \cdot 32^2 L R_0^2 \ln^2\left(\frac{2K}{\delta}\right)}{K}. \tag{18}$$

Thus, applying $B \geq \left(\frac{\sigma K}{128 L R_0 \ln^2\left(\frac{2K}{\delta}\right)}\right)^{\frac{\alpha}{\alpha-1}}$ and rearranging terms, one can obtain

$$\sum_{t=0}^{T-1} \frac{\gamma(f(x_t) - f^*)}{\|\nabla f(x_t)\|} \leq R_0^2 - R_T^2 + \frac{R_0^2}{2} + \frac{T R_0^2}{32^2 K}.$$

Noting that LHS is non-negative, $\overline{S_T} \cap E_{T-1} \cap E_{①}$ under (18) implies

$$R_T^2 \leq R_0^2 + \frac{R_0^2}{2} + \frac{T R_0^2}{32^2 K} \leq 2R_0^2.$$

Therefore, $R_T \leq \sqrt{2} R_0$, and, as a consequence, we have an inclusion of $\overline{S_T} \cap E_{T-1} \cap E_{①}$ into E_T. As a result, we have $\overline{S_{T-1}} \cap E_{T-1} \cap E_{①} \subseteq S_T \cup E_T$, and, subsequently,

$$(S_{T-1} \cup E_{T-1}) \cap S_T \cup (\overline{S}_T \cap E_{T-1} \cap E_{①}) \subseteq S_T \cup E_T.$$

Due to the fact that

$$(S_{T-1} \cup E_{T-1}) \cap S_T \cup (\overline{S}_T \cap E_{T-1}) = S_{T-1} \cup E_{T-1}$$

and

$$(S_{T-1} \cup E_{T-1}) \cap S_T \cap (\overline{S}_T \cap E_{T-1}) = \varnothing,$$

we derive

$$\begin{aligned}\mathbb{P}\{S_T \cup E_T\} &\geq \mathbb{P}\{(S_{T-1} \cup E_{T-1})\} - \mathbb{P}(\overline{S}_T \cap E_{T-1} \cap \overline{E}_{①})\} \\ &\geq \mathbb{P}\{(S_{T-1} \cup E_{T-1})\} - \mathbb{P}\{\overline{E}_{①}\} \geq 1 - \frac{T\delta}{K}.\end{aligned}$$

This concludes the inductive step. Particularly, we have

$$\mathbb{P}\{S_K \cup E_K\} \geq 1 - \delta.$$

To obtain the probability guarantee on the convergence bound, it is enough to prove that

$$\overline{S_K} \cap E_K = \varnothing.$$

Indeed, the event $\overline{S_K} \cap E_K$ implies

$$\begin{aligned}\sum_{k=0}^{K-1} \frac{2\gamma(f(x_k) - f^*)}{\|\nabla f(x_k)\|} &\leq R_0^2 - R_K^2 + \frac{R_0^2}{2} + \sum_{k=0}^{K-1} \frac{64\gamma R_0 \sigma}{B^{\frac{\alpha-1}{\alpha}} \|\nabla f(x_k)\|} + \frac{R_0^2}{32^2} \\ &\leq 2R_0^2 + \sum_{k=0}^{K-1} \frac{8 \cdot 32^2 \gamma L R_0^2 \ln^2\left(\frac{2K}{\delta}\right)}{K \|\nabla f(x_k)\|}.\end{aligned}$$

Moreover, this event implies

$$f(x_k) - f^* > \frac{8 \cdot 32^2 L R_0^2 \ln^2\left(\frac{2K}{\delta}\right)}{K}$$

for all $k = 0, \ldots, K-1$. Consequently, the event $\overline{S_K} \cap E_K$ implies

$$\sum_{k=0}^{K-1} \frac{\gamma(f(x_k) - f^*)}{\|\nabla f(x_k)\|} \leq 2R_0^2.$$

Applying Proposition 1 to the denominator of LHS, one can obtain

$$\sum_{k=0}^{K-1} \frac{\gamma\sqrt{f(x_k)-f^*}}{\sqrt{2L}} \leq 2R_0^2.$$

Dividing both sides by $\frac{\gamma K}{\sqrt{2L}}$, we get

$$\frac{1}{K}\sum_{k=0}^{K-1} \sqrt{f(x_k)-f^*} \leq \frac{2\sqrt{2L}R_0^2}{\gamma K}.$$

Therefore, the next result follows: the event $\overline{S_K} \cap E_K$ implies

$$\min_{k=0,\ldots,K-1}(f(x_k)-f^*) \leq \frac{8\cdot 32^2 L R_0^2 \ln^2\left(\frac{2K}{\delta}\right)}{K}.$$

This is a direct contradiction with the fact that we are in the event $\overline{S_K} \cap E_K$; thus, $\overline{S_K} \cap E_K = \varnothing$. To conclude, we derive

$$\mathbb{P}\{S_K \cup E_K\} = \mathbb{P}\{S_K\} = \mathbb{P}\left\{\min_{k=0,\ldots,K-1}(f(x_k)-f^*) \leq \frac{8\cdot 32^2 L R_0^2 \ln^2\left(\frac{2K}{\delta}\right)}{K}\right\} \geq 1-\delta.$$

To achieve $\min_{k=0,\ldots,K-1}(f(x_k)-f^*) \leq \varepsilon$ with probability at least $1-\delta$, it is enough to make

$$K = \mathcal{O}\left(\frac{LR_0^2}{\varepsilon}\ln^2\left(\frac{LR_0^2}{\varepsilon\delta}\right)\right) \quad \text{iterations;}$$

consequently, the number of oracle calls is bounded as follows:

$$N_{\text{oracle}} = \mathcal{O}\left(\frac{LR_0^2}{\varepsilon}\ln^2\left(\frac{LR_0^2}{\varepsilon\delta}\right)\max\left\{1, \left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right) = \tilde{\mathcal{O}}\left(\frac{LR_0^2}{\varepsilon}\left(1+\left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right)\right).$$

This finishes the proof. □

C.2 Upper bound with outer momentum

We first prove a descent lemma.

Lemma 7 (Descent lemma). *Suppose that Assumptions 1 and 2 hold. Then, if we assume that $\nabla f(x_k) \neq 0$ for all $k = 0, \ldots, K-1$, after K iterations of Algorithm 5, we have*

$$\begin{aligned}\sum_{k=0}^{K-1}\sum_{t=0}^{k} \frac{(1-\beta_1)\beta_1^{k-t}}{1-\beta_1^{k+1}} \frac{2\gamma(f(x_t)-f^*)}{\|\nabla f(x_t)\|} &\leq \|x_0-x^*\|^2 - \|x_K-x^*\|^2 + \frac{2K\gamma^2}{1-\beta_1} \\ &\quad - \sum_{k=0}^{K-1}\sum_{t=0}^{k} \frac{2\gamma(1-\beta_1)\beta_1^{k-t}}{1-\beta_1^{k+1}} \langle \theta_t^u, x_t - x^*\rangle \\ &\quad - \sum_{k=0}^{K-1}\sum_{t=0}^{k} \frac{2\gamma(1-\beta_1)\beta_1^{k-t}}{1-\beta_1^{k+1}} \left\langle \theta_t^b, x_t - x^*\right\rangle.\end{aligned}$$

Proof. This result is closely related to Theorem 11. Starting with the update rule, we get

$$\|x_{k+1}-x^*\|^2 = \|x_k-x^*\|^2 - 2\gamma\langle \tilde{m}_k, x_k - x^*\rangle + \gamma^2\|\tilde{m}_k\|^2. \tag{19}$$

Decomposing the scalar product recursively with $m_{-1}=0$ in the same manner as in Theorem 11, we obtain

$$-\langle \tilde{m}_k, x_k - x^*\rangle = -\sum_{t=0}^{k} \frac{(1-\beta_1)\beta_1^{k-t}}{1-\beta_1^{k+1}} \left\langle \frac{g_t}{\|g_t\|}, x_t - x^*\right\rangle + \sum_{t=0}^{k-1} \frac{\beta_1^{k-t}}{1-\beta_1^{k+1}}\gamma\|\tilde{m}_t\|^2(1-\beta_1^{t+1}).$$

Substituting this result into (19) with the bound $\gamma^2\|\tilde{m}_k\|^2 \leq 2\gamma^2\|\tilde{m}_k\|^2$, we have

$$\|x_{k+1}-x^*\|^2 \leq \|x_k-x^*\|^2 - \sum_{t=0}^{k}\frac{2\gamma(1-\beta_1)\beta_1^{k-t}}{1-\beta_1^{k+1}}\left\langle\frac{g_t}{\|g_t\|}, x_t-x^*\right\rangle + \sum_{t=0}^{k}\frac{2\beta_1^{k-t}}{1-\beta_1^{k+1}}\gamma^2\|\tilde{m}_t\|^2(1-\beta_1^{t+1}).$$

Using the notation of θ_t, we derive

$$\|x_{k+1}-x^*\|^2 \leq \|x_k-x^*\|^2 - \sum_{t=0}^{k}\frac{2\gamma(1-\beta_1)\beta_1^{k-t}}{1-\beta_1^{k+1}}\left\langle\frac{\nabla f(x_t)}{\|\nabla f(x_t)\|}, x_t-x^*\right\rangle - \sum_{t=0}^{k}\frac{2\gamma(1-\beta_1)\beta_1^{k-t}}{1-\beta_1^{k+1}}\langle\theta_t, x_t-x^*\rangle + \sum_{t=0}^{k}\frac{2\beta_1^{k-t}}{1-\beta_1^{k+1}}\gamma^2\|\tilde{m}_t\|^2(1-\beta_1^{t+1}).$$

Applying (14) from Theorem 11 with $\|\tilde{m}_t\| \leq 1$, rearranging terms and using that $\langle\nabla f(x_t), x_t-x^*\rangle \geq f(x_t)-f^*$ due to convexity of f, we obtain the next inequality:

$$\sum_{t=0}^{k}\frac{(1-\beta_1)\beta_1^{k-t}}{1-\beta_1^{k+1}}\frac{2\gamma(f(x_t)-f^*)}{\|\nabla f(x_t)\|} \leq \|x_k-x^*\|^2 - \|x_{k+1}-x^*\|^2 + \frac{2\gamma^2}{1-\beta_1} - \sum_{t=0}^{k}\frac{2\gamma(1-\beta_1)\beta_1^{k-t}}{1-\beta_1^{k+1}}\langle\theta_t, x_t-x^*\rangle.$$

Decomposing $\theta_t = \theta_t^u + \theta_t^b$ and summing up, we conclude

$$\sum_{k=0}^{K-1}\sum_{t=0}^{k}\frac{(1-\beta_1)\beta_1^{k-t}}{1-\beta_1^{k+1}}\frac{2\gamma(f(x_t)-f^*)}{\|\nabla f(x_t)\|} \leq \|x_0-x^*\|^2 - \|x_K-x^*\|^2 + \frac{2K\gamma^2}{1-\beta_1} - \sum_{k=0}^{K-1}\sum_{t=0}^{k}\frac{2\gamma(1-\beta_1)\beta_1^{k-t}}{1-\beta_1^{k+1}}\langle\theta_t^u, x_t-x^*\rangle - \sum_{k=0}^{K-1}\sum_{t=0}^{k}\frac{2\gamma(1-\beta_1)\beta_1^{k-t}}{1-\beta_1^{k+1}}\langle\theta_t^b, x_t-x^*\rangle.$$

This finishes the proof. □

Now we are ready to prove the high-probability convergence of Algorithm 5.

Theorem 13. *Suppose that Assumptions 1, 2 and 3 hold. Then, after K iterations with $\ln\left(\frac{2K}{\delta}\right) \geq 1$ of Algorithm 5 with $\gamma = \frac{R_0\sqrt{1-\beta_1}}{32C(\beta_1)\sqrt{K}\ln\left(\frac{2K}{\delta}\right)}$, batch size $B_k \equiv B = \left\lceil\max\left\{1, \left(\frac{\sigma K(1-\beta_1)}{128LR_0\ln^2\left(\frac{2K}{\delta}\right)}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right\rceil$ and $C(\beta_1) = 1+\ln\left(\frac{1}{1-\beta_1}\right)$, we have*

$$\min_{k=0,\ldots,K-1}(f(x_k)-f^*) \leq \frac{8\cdot 32^2 LR_0^2C(\beta_1)^2\ln^2\left(\frac{2K}{\delta}\right)}{K(1-\beta_1)}$$

with probability at least $1-\delta$.

Proof. The proof is constructed similarly to Theorem 12. First, let us introduce the probabilistic good events E_k :

$$-\sum_{l=0}^{t-1}\sum_{j=0}^{l}\frac{2\gamma(1-\beta_1)\beta_1^{l-j}}{1-\beta_1^{l+1}}\left\langle\theta_j^u, x_j-x^*\right\rangle \le \frac{R_0^2}{2};$$
$$-\sum_{l=0}^{t-1}\sum_{j=0}^{l}\frac{2\gamma(1-\beta_1)\beta_1^{l-j}}{1-\beta_1^{l+1}}\left\langle\theta_j^b, x_j-x^*\right\rangle \le \sum_{l=0}^{t-1}\sum_{j=0}^{l}\frac{(1-\beta_1)\beta_1^{l-j}}{1-\beta_1^{l+1}}\frac{64\gamma R_0\sigma}{B^{\frac{\alpha-1}{\alpha}}\left\|\nabla f(x_j)\right\|};$$
$$R_t \le \sqrt{2}R_0$$

for all $t=0,\ldots,k$, and success events S_k:

$$\min_{t=0,\ldots,k-1}(f(x_t)-f^*) \le \frac{8\cdot 32^2 LR_0^2C(\beta_1)^2\ln^2\left(\frac{2K}{\delta}\right)}{K(1-\beta_1)}.$$

Here $S_0=\varnothing$. In the definition of E_k, the first two inequalities are required only for those t for which $\nabla f(x_j)\neq 0$ for every $j<t$. Thus, all displayed ratios are well defined whenever they are used. We want to show via induction that $\mathbb{P}\{S_k\cup E_k\}\ge 1-\frac{k\delta}{K}$. For $k=0$ it is obvious that $\mathbb{P}\{S_0\cup E_0\}\ge 1$. Now, let us assume that $\mathbb{P}\{S_{T-1}\cup E_{T-1}\}\ge 1-\frac{(T-1)\delta}{K}$. Since $S_{T-1}\subseteq S_T$, we have $(S_{T-1}\cup E_{T-1})\cap S_T\cup(\overline{S_T}\cap E_{T-1}) = S_{T-1}\cup E_{T-1}$, and $(S_{T-1}\cup E_{T-1})\cap S_T\subseteq S_T\cup E_T$. Hence, it is enough to consider the event $\overline{S_T}\cap E_{T-1}$. This event implies $\nabla f(x_t)\neq 0$ for all $t=0,\ldots,T-1$. Thus, the event $\overline{S}_T\cap E_{T-1}$ allows us to apply Lemma 7:

$$\begin{aligned}\sum_{l=0}^{t-1}\sum_{j=0}^{l}\frac{(1-\beta_1)\beta_1^{l-j}}{1-\beta_1^{l+1}}\frac{2\gamma(f(x_j)-f^*)}{\left\|\nabla f(x_j)\right\|} &\le R_0^2-R_t^2+\frac{2t\gamma^2}{1-\beta_1}\\ &\quad-\sum_{l=0}^{t-1}\sum_{j=0}^{l}\frac{2\gamma(1-\beta_1)\beta_1^{l-j}}{1-\beta_1^{l+1}}\left\langle\theta_j^u, x_j-x^*\right\rangle\\ &\quad-\sum_{l=0}^{t-1}\sum_{j=0}^{l}\frac{2\gamma(1-\beta_1)\beta_1^{l-j}}{1-\beta_1^{l+1}}\left\langle\theta_j^b, x_j-x^*\right\rangle\end{aligned}$$

for all $t=0,\ldots,T$. What is more, the event $\overline{S_T}\cap E_{T-1}$ implies

$$\begin{aligned}\sum_{l=0}^{t-1}\sum_{j=0}^{l}\frac{(1-\beta_1)\beta_1^{l-j}}{1-\beta_1^{l+1}}\frac{2\gamma(f(x_j)-f^*)}{\left\|\nabla f(x_j)\right\|} &\le R_0^2-R_t^2+\frac{2t\gamma^2}{1-\beta_1}\\ &\quad-\sum_{l=0}^{t-1}\sum_{j=0}^{l}\frac{2\gamma(1-\beta_1)\beta_1^{l-j}}{1-\beta_1^{l+1}}\left\langle\theta_j^u, x_j-x^*\right\rangle\\ &\quad-\sum_{l=0}^{t-1}\sum_{j=0}^{l}\frac{2\gamma(1-\beta_1)\beta_1^{l-j}}{1-\beta_1^{l+1}}\left\langle\theta_j^b, x_j-x^*\right\rangle\\ &\le R_0^2-R_t^2+\frac{R_0^2}{2}+\frac{2tR_0^2}{32^2K}\\ &\quad+\sum_{l=0}^{t-1}\sum_{j=0}^{l}\frac{(1-\beta_1)\beta_1^{l-j}}{1-\beta_1^{l+1}}\frac{64\gamma R_0\sigma}{B^{\frac{\alpha-1}{\alpha}}\left\|\nabla f(x_j)\right\|}\\ &\le 2R_0^2-R_t^2\\ &\quad+\sum_{l=0}^{t-1}\sum_{j=0}^{l}\frac{(1-\beta_1)\beta_1^{l-j}}{1-\beta_1^{l+1}}\frac{8\cdot 32^2\gamma LR_0^2C(\beta_1)^2\ln^2\left(\frac{2K}{\delta}\right)}{K(1-\beta_1)\left\|\nabla f(x_j)\right\|}\end{aligned}$$

for all $t = 0, \ldots, T-1$, since $B \geq \left(\frac{\sigma K(1-\beta_1)}{128 L R_0 C(\beta_1)^2 \ln^2\left(\frac{2K}{\delta}\right)} \right)^{\frac{\alpha}{\alpha-1}}$. Additionally, the event $\overline{S_T} \cap E_{T-1}$ implies

$$f(x_t) - f^* > \frac{8 \cdot 32^2 L R_0^2 C(\beta_1)^2 \ln^2\left(\frac{2K}{\delta}\right)}{K(1-\beta_1)}$$

for all $t = 0, \ldots, T-2$; thus, we have

$$\sum_{l=0}^{t-1} \sum_{j=0}^{l} \frac{(1-\beta_1)\beta_1^{l-j}}{1-\beta_1^{l+1}} \frac{\gamma(f(x_j) - f^*)}{\|\nabla f(x_j)\|} \leq 2R_0^2 - R_t^2$$

into the event $\overline{S_T} \cap E_{T-1}$ for all $t = 0, \ldots, T-1$. Consequently, one can obtain that $\overline{S}_T \cap E_{T-1}$ implies

$$\sum_{l=0}^{t-1} \sum_{j=0}^{l} \frac{(1-\beta_1)\beta_1^{l-j}}{1-\beta_1^{l+1}} \frac{\gamma(f(x_j) - f^*)}{\|\nabla f(x_j)\|} \leq 2R_0^2$$

and

$$R_t^2 \leq 2R_0^2$$

for all $t = 0, \ldots, T-1$. Next, similarly to Theorem 12, we define random vector η_t for all $t = 0, \ldots, T-1$ as follows:

$$\eta_t = \begin{cases} x_t - x^*, & \|x_t - x^*\| \leq \sqrt{2} R_0; \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

Therefore, $\|\eta_t\| \leq \sqrt{2} R_0$ with probability 1. Moreover, the event $\overline{S_T} \cap E_{T-1}$ implies $R_t \leq \sqrt{2} R_0$, which allows to use truncated vectors η_t to reformulate descent lemma using the new notation:

$$\begin{aligned} \sum_{t=0}^{T-1} \sum_{j=0}^{t} \frac{(1-\beta_1)\beta_1^{t-j}}{1-\beta_1^{t+1}} \frac{2\gamma(f(x_j) - f^*)}{\|\nabla f(x_j)\|} &\leq R_0^2 - R_T^2 + \frac{2T\gamma^2}{1-\beta_1} \\ &\quad \underbrace{- \sum_{t=0}^{T-1} \sum_{j=0}^{t} \frac{2\gamma(1-\beta_1)\beta_1^{t-j}}{1-\beta_1^{t+1}} \left\langle \theta_j^u, \eta_j \right\rangle}_{①} \\ &\quad \underbrace{- \sum_{t=0}^{T-1} \sum_{j=0}^{t} \frac{2\gamma(1-\beta_1)\beta_1^{t-j}}{1-\beta_1^{t+1}} \left\langle \theta_j^b, \eta_j \right\rangle}_{②}. \end{aligned}$$

Next, we provide upper bounds for ① and ②.

Upper bound for ①. At first, we combine terms as follows:

$$- \sum_{t=0}^{T-1} \sum_{j=0}^{t} \frac{2\gamma(1-\beta_1)\beta_1^{t-j}}{1-\beta_1^{t+1}} \left\langle \theta_j^u, \eta_j \right\rangle = - \sum_{j=0}^{T-1} \left(\sum_{t=j}^{T-1} \frac{2\gamma(1-\beta_1)\beta_1^{t-j}}{1-\beta_1^{t+1}} \right) \left\langle \theta_j^u, \eta_j \right\rangle.$$

For simplicity, we denote the sum above in another indices to derive

$$① = - \sum_{t=0}^{T-1} \left(\sum_{p=t}^{T-1} \frac{2\gamma(1-\beta_1)\beta_1^{p-t}}{1-\beta_1^{p+1}} \right) \left\langle \theta_t^u, \eta_t \right\rangle.$$

Then, we get

$$\mathbb{E}_t \left[- \left(\sum_{p=t}^{T-1} \frac{2\gamma(1-\beta_1)\beta_1^{p-t}}{1-\beta_1^{p+1}} \right) \left\langle \theta_t^u, \eta_t \right\rangle \right] = 0$$

due to the definition of θ_t^u. Moreover, we can have

$$\sum_{p=t}^{T-1} \frac{(1-\beta_1)\beta_1^{p-t}}{1-\beta_1^{p+1}} \leq 1 + \ln\left(\frac{1}{1-\beta_1}\right) = C(\beta_1).$$

Indeed, we can apply Lemma 2 in the following way

$$\sum_{p=t}^{T-1} \frac{2\gamma(1-\beta_1)\beta_1^{p-t}}{1-\beta_1^{p+1}} = 2\gamma \sum_{p=t}^{T-1} \frac{\beta_1^{p-t}}{\sum_{i=0}^{p}\beta_1^i} \leq 2\gamma \sum_{p=t}^{T-1} \frac{\beta_1^{p-t}}{\sum_{i=0}^{p-t}\beta_1^i} = 2\gamma \sum_{p=0}^{T-t-1} \frac{\beta_1^{p}}{\sum_{i=0}^{p}\beta_1^i}.$$

The RHS of the inequality above is bounded by a direct result of Lemma 2. Therefore, we can bound each term of ① almost surely:

$$\left| -\left(\sum_{p=t}^{T-1} \frac{2\gamma(1-\beta_1)\beta_1^{p-t}}{1-\beta_1^{p+1}}\right) \langle \theta_t^u, \eta_t \rangle \right| \leq 8\gamma C(\beta_1) R_0 \leq \frac{9R_0^2}{16\ln\left(\frac{2K}{\delta}\right)}$$

due to the choice of γ. Additionally, we have bounded conditional variance for each term:

$$\sigma_t^2 = \mathbb{E}_t\left[\left(\sum_{p=t}^{T-1} \frac{2\gamma(1-\beta_1)\beta_1^{p-t}}{1-\beta_1^{p+1}}\right)^2 \langle \theta_t^u, \eta_t \rangle^2\right] \leq 8\gamma^2 C(\beta_1)^2 R_0^2 \mathbb{E}_t[\|\theta_t^u\|^2].$$

Therefore, one can apply Bernstein's inequality with $b = \frac{R_0^2}{2}$ and $G = \frac{R_0^4}{32\ln\left(\frac{2K}{\delta}\right)}$ to derive

$$\mathbb{P}\left\{|①| > b \text{ and } \sum_{t=0}^{T-1} \sigma_t^2 \leq G\right\} \leq 2\exp\left(-\frac{b^2}{2G + 2cb/3}\right) = \frac{\delta}{K}.$$

Thus, we get

$$\mathbb{P}\left\{|①| \leq b \text{ or } \sum_{t=0}^{T-1} \sigma_t^2 > G\right\} \geq 1 - \frac{\delta}{K}.$$

Moreover, we have

$$\sum_{t=0}^{T-1} \sigma_t^2 \leq \sum_{t=0}^{T-1} 8\gamma^2 C(\beta_1)^2 R_0^2 \mathbb{E}_t[\|\theta_t^u\|^2] \leq 32\gamma^2 C(\beta_1)^2 R_0^2 T \leq \frac{R_0^4}{32\ln\left(\frac{2K}{\delta}\right)} = G.$$

Upper bound for ②. The event $\overline{S_T} \cap E_{T-1}$ implies

$$-\sum_{t=0}^{T-1}\sum_{j=0}^{t} \frac{2\gamma(1-\beta_1)\beta_1^{t-j}}{1-\beta_1^{t+1}} \langle \theta_j^b, \eta_j \rangle \leq \sum_{t=0}^{T-1}\sum_{j=0}^{t} \frac{2\gamma(1-\beta_1)\beta_1^{t-j}}{1-\beta_1^{t+1}} \left\|\theta_j^b\right\| \|\eta_j\|$$

$$\overset{\text{Lemma 4, Cor. 1}}{\leq} \sum_{t=0}^{T-1}\sum_{j=0}^{t} \frac{(1-\beta_1)\beta_1^{t-j}}{1-\beta_1^{t+1}} \frac{64\gamma R_0 \sigma}{B^{\frac{\alpha-1}{\alpha}} \|\nabla f(x_j)\|},$$

where in the last inequality we apply Lemma 4 to bound $\left\|\theta_t^b\right\|$ (it is possible due to the fact that the event $\overline{S_T}$ implies $\nabla f(x_t) \neq 0$ for all $t = 0, \ldots, T-1$) and Corollary 1 to derive $\sigma \to \frac{2\sigma}{B^{\frac{\alpha-1}{\alpha}}}$. Consequently, the event $\overline{S_T} \cap E_{T-1} \cap E_①$, where

$$E_① = \left\{|①| \leq \frac{R_0^2}{2} \text{ or } \sum_{t=0}^{T-1} \sigma_t^2 > \frac{R_0^4}{32\ln\left(\frac{2K}{\delta}\right)}\right\},$$

implies

$$\sum_{t=0}^{T-1}\sum_{j=0}^{t} \frac{(1-\beta_1)\beta_1^{t-j}}{1-\beta_1^{t+1}} \frac{2\gamma(f(x_j)-f^*)}{\|\nabla f(x_j)\|} \leq R_0^2 - R_T^2 + \frac{2TR_0^2}{32^2 K} + \frac{R_0^2}{2}$$

$$+ \sum_{t=0}^{T-1}\sum_{j=0}^{t} \frac{(1-\beta_1)\beta_1^{t-j}}{1-\beta_1^{t+1}} \frac{64\gamma R_0 \sigma}{B^{\frac{\alpha-1}{\alpha}} \|\nabla f(x_j)\|}. \tag{21}$$

We can also derive the next inclusion:

$$(S_{T-1} \cup E_{T-1}) \cap S_T \cup (\overline{S}_T \cap E_{T-1} \cap E_{①}) \subseteq S_T \cup E_T.$$

Similarly to Theorem 12, we consider following cases: we have $S_{T-1} \subseteq S_T$; moreover, into the event $\overline{S_T} \cap E_{T-1} \cap E_{①}$ we get

$$f(x_t) - f^* > \frac{8 \cdot 32^2 L R_0^2 C(\beta_1)^2 \ln^2\left(\frac{2K}{\delta}\right)}{K(1-\beta_1)}. \tag{22}$$

Consequently, using that $B \geq \left(\frac{\sigma K(1-\beta_1)}{128 L R_0 C(\beta_1)^2 \ln^2\left(\frac{2K}{\delta}\right)}\right)^{\frac{\alpha}{\alpha-1}}$ and rearranging terms, one can obtain

$$\sum_{t=0}^{T-1} \sum_{j=0}^{t} \frac{(1-\beta_1)\beta_1^{t-j}}{1-\beta_1^{t+1}} \frac{\gamma(f(x_j) - f^*)}{\|\nabla f(x_j)\|} \leq R_0^2 - R_T^2 + \frac{R_0^2}{2} + \frac{2TR_0^2}{32^2 K}.$$

Since the LHS of the inequality above is non-negative, the event $\overline{S_T} \cap E_{T-1} \cap E_{①}$ under (22) implies

$$R_T^2 \leq R_0^2 + \frac{R_0^2}{2} + \frac{2TR_0^2}{32^2 K} \leq 2R_0^2.$$

Therefore, $R_T \leq \sqrt{2} R_0$, and, as a consequence, we have an inclusion of $\overline{S_T} \cap E_{T-1} \cap E_{①}$ into E_T. Thus, similarly to Theorem 12, we obtain

$$\mathbb{P}\{S_T \cup E_T\} \geq \mathbb{P}\{S_{T-1} \cup E_{T-1}\} - \mathbb{P}\{\overline{E}_{①}\} \geq 1 - \frac{T\delta}{K}.$$

This concludes the inductive step. Particularly, we have

$$\mathbb{P}\{S_K \cup E_K\} \geq 1 - \delta.$$

To obtain the probability guarantee on the convergence bound, we may prove that

$$\overline{S_K} \cap E_K = \varnothing.$$

Indeed, the event $\overline{S_K} \cap E_K$ implies

$$\begin{aligned}
&\sum_{k=0}^{K-1} \sum_{j=0}^{k} \frac{(1-\beta_1)\beta_1^{k-j}}{1-\beta_1^{k+1}} \frac{2\gamma(f(x_j) - f^*)}{\|\nabla f(x_j)\|} \\
&\quad \leq 2R_0^2 - R_K^2 + \sum_{k=0}^{K-1} \sum_{j=0}^{k} \frac{(1-\beta_1)\beta_1^{k-j}}{1-\beta_1^{k+1}} \frac{64\gamma R_0 \sigma}{B^{\frac{\alpha-1}{\alpha}} \|\nabla f(x_j)\|} \\
&\quad \leq 2R_0^2 + \sum_{k=0}^{K-1} \sum_{j=0}^{k} \frac{(1-\beta_1)\beta_1^{k-j}}{1-\beta_1^{k+1}} \frac{8 \cdot 32^2 \gamma L R_0^2 C(\beta_1)^2 \ln^2\left(\frac{2K}{\delta}\right)}{K(1-\beta_1)\|\nabla f(x_j)\|}.
\end{aligned}$$

Moreover, this event implies

$$f(x_k) - f^* > \frac{8 \cdot 32^2 C(\beta_1)^2 L R_0^2 \ln^2\left(\frac{2K}{\delta}\right)}{K(1-\beta_1)}$$

for all $k = 0, \ldots, K-1$. Consequently, the event $\overline{S_K} \cap E_K$ implies

$$\sum_{k=0}^{K-1} \sum_{j=0}^{k} \frac{(1-\beta_1)\beta_1^{k-j}}{1-\beta_1^{k+1}} \frac{\gamma(f(x_j) - f^*)}{\|\nabla f(x_j)\|} \leq 2R_0^2.$$

Applying Proposition 1 to denominators of LHS, one can get

$$\sum_{k=0}^{K-1} \sum_{j=0}^{k} \frac{(1-\beta_1)\beta_1^{k-j}}{1-\beta_1^{k+1}} \frac{\gamma\sqrt{f(x_j) - f^*}}{\sqrt{2L}} \leq 2R_0^2.$$

Noting that $\sum\limits_{j=0}^{k} \frac{(1-\beta_1)\beta_1^{k-j}}{1-\beta_1^{k+1}} \sqrt{f(x_j) - f^*}$ is a convex combination of $\sqrt{f(x_j) - f^*}$, which can be bounded below as $\min_{j=0,\ldots,k} \sqrt{f(x_j) - f^*}$, and dividing both sides by $\frac{\gamma K}{\sqrt{2L}}$, we get

$$\min_{k=0,\ldots,K-1} \sqrt{f(x_k) - f^*} \leq \frac{2\sqrt{2L}R_0^2}{\gamma K}.$$

Therefore, the next result follows: the event $\overline{S_K} \cap E_K$ implies

$$\min_{k=0,\ldots,K-1} (f(x_k) - f^*) \leq \frac{8 \cdot 32^2 L R_0^2 C(\beta_1)^2 \ln^2\left(\frac{2K}{\delta}\right)}{K(1-\beta_1)}.$$

This is a direct contradiction with the fact that we are in the event $\overline{S_K} \cap E_K$; thus, $\overline{S_K} \cap E_K = \varnothing$. To conclude, we derive

$$\begin{aligned}\mathbb{P}\{S_K \cup E_K\} = \mathbb{P}\{S_K\} &= \mathbb{P}\left\{\min_{k=0,\ldots,K-1} (f(x_k) - f^*) \leq \frac{8 \cdot 32^2 L R_0^2 C(\beta_1)^2 \ln^2\left(\frac{2K}{\delta}\right)}{K(1-\beta_1)}\right\} \\ &\geq 1 - \delta.\end{aligned}$$

To achieve $\min_{k=0,\ldots,K-1}(f(x_k) - f^*) \leq \varepsilon$ with probability at least $1-\delta$, it is enough to make

$$K = \tilde{\mathcal{O}}\left(\frac{LR_0^2}{(1-\beta_1)\varepsilon} \ln^2\left(\frac{LR_0^2}{(1-\beta_1)\varepsilon\delta}\right)\right) \quad \text{iterations,}$$

where $\tilde{\mathcal{O}}$ hides $\ln\left(\frac{1}{1-\beta_1}\right)$. Consequently, the number of oracle calls is bounded as follows:

$$N_{\text{oracle}} = \tilde{\mathcal{O}}\left(\frac{LR_0^2}{(1-\beta_1)\varepsilon}\left(1 + \left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right)\right),$$

where $\tilde{\mathcal{O}}$ also hides $\ln\left(\frac{1}{1-\beta_1}\right)$ and $\ln\left(\frac{1}{\delta}\right)$. This finishes the proof. □

D Negative Results and Lower Bounds

In this section, we provide negative results for Normalized SGD and its variants. First, we formulate a lower bound on constant batch size for Normalized SGD with outer momentum.

Theorem 14 (Lower bound on constant batch)**.** *Suppose that $\varepsilon \leq \frac{LR_0^2}{16}$, $K \geq 1$, $\gamma > 0$ and $\beta_1 \in [0, 1)$. Then there exist a one-dimensional function f satisfying Assumptions 1, 2 and stochastic oracles $\{g_k\}_{k=0}$ satisfying Assumption 3 such that after run of Normalized SGD with outer momentum for K iterations with a constant batch size B, the following bound holds:*

$$\mathbb{P}\left\{\min_{k=0,\ldots,K}(f(x_k) - f^*) > \varepsilon\right\} < \delta$$
$$\Rightarrow B = \Omega\left(\max\left\{1, \ln\left(1 + (1-\delta)^{\frac{1}{C(\gamma,\beta_1)}}\right)\left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right),$$

where $C(\gamma, \beta_1) = \left\lceil \frac{5R_0}{8\gamma\left(2+\ln\left(\frac{2}{1-\beta_1}\right)\right)} \right\rceil \geq 1$. In particular, if $\delta \leq \frac{1}{2}$, we get

$$\mathbb{P}\left\{\min_{k=0,\ldots,K}(f(x_k) - f^*) > \varepsilon\right\} < \delta \Rightarrow B = \Omega\left(\max\left\{1, \left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right).$$

Proof. **Part 1: function.** Let us consider the one-dimensional function f, based on the Huber loss;

$$f(x) = \begin{cases} 0, & x < 0; \\ \frac{Lx^2}{2}, & 0 \leq x \leq \frac{\mu}{L}; \\ \mu x - \frac{\mu^2}{2L}, & x > \frac{\mu}{L}, \end{cases} \tag{23}$$

where $\mu = \frac{4\varepsilon}{R_0}$ for convenience. We take $x^* = 0$ and initialize $x_0 = R_0$. The function from (23) is convex and L-smooth; therefore, f satisfies Assumptions 1 and 2. Moreover, since $\varepsilon \leq \frac{LR_0^2}{16}$, we derive

$$f\left(\frac{\mu}{L}\right) = \frac{\mu^2}{2L} = \frac{8\varepsilon^2}{LR_0^2} \leq \frac{\varepsilon}{2}.$$

Thus, if $f(x) - f^* = \varepsilon$, then $x > \frac{\mu}{L}$. Consequently,

$$\mu x - \frac{\mu^2}{2L} = \varepsilon \Rightarrow x = \frac{\varepsilon}{\mu} + \frac{\mu}{2L} = \frac{R_0}{4} + \frac{2\varepsilon}{LR_0} \leq \frac{3R_0}{8}.$$

Let us denote $x_b = \frac{R_0}{4} + \frac{2\varepsilon}{LR_0}$. Therefore,

$$\mathbb{P}\left\{\min_{t=0,\ldots,K} f(x_t) - f^* > \varepsilon\right\} = \mathbb{P}\left\{\forall t = 0, \ldots, K \quad x_t > x_b\right\} \tag{24}$$

Part 2: stochastic oracle. To continue the proof, we consider the following two-point oracle:

$$\hat{g}_t = \begin{cases} -\nabla f(x_t), & \text{with prob. } 1-p \\ \frac{2-p}{p}\nabla f(x_t), & \text{with prob. } p, \end{cases}$$

where $p = \min\left\{1, 2^{\frac{\alpha+1}{\alpha-1}}\left(\frac{\mu}{\sigma}\right)^{\frac{\alpha}{\alpha-1}}\right\}$. This stochastic gradient is unbiased; indeed,

$$\mathbb{E}\hat{g}_t = -(1-p)\nabla f(x_t) + (2-p)\nabla f(x_t) = \nabla f(x_t).$$

Moreover, the α-central moment is bounded: in the case $p = 1$ it is obvious, and for the case $p < 1$ we get

$$\begin{aligned}\mathbb{E}\left\|\hat{g}_t - \nabla f(x_t)\right\|^\alpha &= (1-p)2^\alpha \left\|\nabla f(x_t)\right\|^\alpha + p\left(\frac{2-2p}{p}\right)^\alpha \left\|\nabla f(x_t)\right\|^\alpha \\ &= 2^\alpha((1-p) + (1-p)^\alpha p^{1-\alpha})\left\|\nabla f(x_t)\right\|^\alpha \\ &\leq 2^{\alpha+1}p^{1-\alpha}\left\|\nabla f(x_t)\right\|^\alpha,\end{aligned}$$

where in the last inequality we apply $1 - p \leq 1 \leq p^{1-\alpha}$. According to (23), the gradient norm is bounded: $\|\nabla f(x)\| \leq \mu$; thus, using the definition of p, we derive

$$\mathbb{E}\,\|\hat{g}_t - \nabla f(x_t)\|^\alpha \leq 2^{\alpha+1} p^{1-\alpha}\,\|\nabla f(x_t)\|^\alpha \leq \sigma^\alpha.$$

Consequently, the stochastic oracle $\hat{g}_t$ satisfies Assumption 3. It is worth noting that we draw its scalar multiplier independently on every call.

Part 3: probability of failure. Let us consider the estimator g_t based on the batch size B on each iteration:

$$g_t = \frac{1}{B}\sum_{i=1}^{B} \hat{g}_t^i,$$

where $\hat{g}_t^i$ is the ith sample. Since our goal is to bound a failure probability, we first need to find a probability to take a step away from the optimum. In the case $x_k > \frac{\mu}{L}$ we get

$$g_k = \frac{1}{B}\sum_{i=1}^{B} \hat{g}_k^i = \frac{1}{B}\left((B-j)(-\mu) + \frac{2-p}{p} j\mu\right),$$

where $j \sim Bin(B, p)$. We define these independent counts for all $k \geq 0$, including after the method reaches the target set. Thus, while $x_k > x_b$, the sign of g_k depends on j:

$$\text{sign}(g_k) = -1 \Leftrightarrow -B\mu + \frac{2}{p} j\mu < 0 \Leftrightarrow j < \frac{Bp}{2}.$$

Now let us consider two cases. The first option is $Bp \geq \frac{1}{2}$ – this case directly leads to the desired bound on B after substitution of p and μ:

$$B \geq \frac{1}{2p} = \Omega\left(\max\left\{1, \left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right).$$

Otherwise, if $Bp < \frac{1}{2}$, we derive

$$j < \frac{Bp}{2} < \frac{1}{4};$$

it is possible iff $j = 0$. Therefore, we get

$$\mathbb{P}\{\text{sign}(g_k) = -1\} = \mathbb{P}\{j = 0\} = (1-p)^B. \tag{25}$$

Now let us look at the update rule

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1-\beta_1)\frac{g_t}{\|g_t\|}; \\ \tilde{m}_t &= \frac{m_t}{1-\beta_1^{t+1}}; \\ x_{t+1} &= x_t - \gamma \tilde{m}_t. \end{aligned}$$

Therefore, the total step from the initial point x_0 can be expressed as

$$x_t = x_0 - \gamma \sum_{k=0}^{t-1} \tilde{m}_k.$$

For simplicity, let us denote $P_t = \sum\limits_{k=0}^{t-1} \tilde{m}_k$. Substituting this form of x_t in (24), we have

$$\begin{aligned} \mathbb{P}\left\{\min_{t=0\ldots K} f(x_t) - f^* > \varepsilon\right\} &= \mathbb{P}\left\{\forall t = 0, \ldots, K \quad x_t > x_b\right\} \\ &\geq \mathbb{P}\left\{\forall t = 0, \ldots, K \quad P_t < \frac{5R_0}{8\gamma}\right\}, \end{aligned} \tag{26}$$

where we use the upper bound on x_b and $R_0 = |x_0 - x^*| = x_0$. Moreover, P_t can be rewritten as follows:

$$P_t = \sum_{k=0}^{t-1} \tilde{m}_k = \sum_{k=0}^{t-1} \sum_{i=0}^{k} \frac{(1-\beta_1)\beta_1^{k-i}}{1-\beta_1^{k+1}} \operatorname{sign}(g_i)$$

due to the definition of $\tilde{m}_k$. Rearranging terms gives

$$P_t = \sum_{k=0}^{t-1} \sum_{i=0}^{k} \frac{(1-\beta_1)\beta_1^{k-i}}{1-\beta_1^{k+1}} \operatorname{sign}(g_i) = \sum_{k=0}^{t-1} \left(\sum_{j=k}^{t-1} \frac{(1-\beta_1)\beta_1^{j-k}}{1-\beta_1^{j+1}} \right) \operatorname{sign}(g_k) = \sum_{k=0}^{t-1} a_k \operatorname{sign}(g_k), \tag{27}$$

where $a_k = \sum\limits_{j=k}^{t-1} \frac{(1-\beta_1)\beta_1^{j-k}}{1-\beta_1^{j+1}} \leq \sum\limits_{j=0}^{\infty} \frac{(1-\beta_1)\beta_1^{j}}{1-\beta_1^{j+1}}$. Thus, one can derive

$$\begin{aligned}
a_k &\leq \sum_{j=0}^{\infty} \frac{(1-\beta_1)\beta_1^j}{1-\beta_1^{j+1}} = \sum_{j=0}^{\infty} \frac{\beta_1^j}{1+\beta_1+\ldots+\beta_1^j} \\
&= \sum_{j=0}^{\left\lceil \frac{1}{1-\beta_1} \right\rceil - 1} \frac{\beta_1^j}{1+\beta_1+\ldots+\beta_1^j} + \sum_{j=\left\lceil \frac{1}{1-\beta_1} \right\rceil}^{\infty} \frac{\beta_1^j}{1+\beta_1+\ldots+\beta_1^j} \\
&\leq \sum_{j=0}^{\left\lceil \frac{1}{1-\beta_1} \right\rceil - 1} \frac{1}{j+1} + \sum_{j=\left\lceil \frac{1}{1-\beta_1} \right\rceil}^{\infty} \frac{\beta_1^j}{1+\beta_1+\ldots+\beta_1^{\left\lceil \frac{1}{1-\beta_1} \right\rceil - 1}},
\end{aligned}$$

where we divided the series into two terms, for the first term we use $1+\beta_1+\ldots+\beta_1^j \geq (j+1)\beta_1^j$, and for the second one we apply $1+\beta_1+\ldots+\beta_1^j \geq 1+\beta_1+\ldots+\beta_1^{\left\lceil \frac{1}{1-\beta_1} \right\rceil - 1}$ since $j \geq \left\lceil \frac{1}{1-\beta_1} \right\rceil - 1$. Moreover, we get

$$\sum_{j=\left\lceil \frac{1}{1-\beta_1} \right\rceil}^{\infty} \beta_1^j = \beta_1^{\left\lceil \frac{1}{1-\beta_1} \right\rceil} \sum_{j=0}^{\infty} \beta_1^j = \frac{\beta_1^{\left\lceil \frac{1}{1-\beta_1} \right\rceil}}{1-\beta_1}.$$

Therefore, one can have

$$a_k \leq \sum_{j=0}^{\left\lceil \frac{1}{1-\beta_1} \right\rceil - 1} \frac{1}{j+1} + \frac{\beta_1^{\left\lceil \frac{1}{1-\beta_1} \right\rceil}}{1-\beta_1^{\left\lceil \frac{1}{1-\beta_1} \right\rceil}}, \tag{28}$$

where we apply $(1-\beta_1)\left(1+\beta_1+\ldots+\beta_1^{\left\lceil \frac{1}{1-\beta_1} \right\rceil - 1}\right) = 1-\beta_1^{\left\lceil \frac{1}{1-\beta_1} \right\rceil}$. The first term from (28) can be bounded by logarithm:

$$\sum_{j=0}^{\left\lceil \frac{1}{1-\beta_1} \right\rceil - 1} \frac{1}{j+1} \leq 1 + \ln\left(\left\lceil \frac{1}{1-\beta_1} \right\rceil \right).$$

To bound the he second term, it can be shown that

$$\beta_1^{\left\lceil \frac{1}{1-\beta_1} \right\rceil} = (1-(1-\beta_1))^{\left\lceil \frac{1}{1-\beta_1} \right\rceil} \leq \exp\left(-(1-\beta_1) \left\lceil \frac{1}{1-\beta_1} \right\rceil \right) \leq \exp(-1).$$

Consequently, we obtain

$$\frac{\beta_1^{\left\lceil \frac{1}{1-\beta_1} \right\rceil}}{1-\beta_1^{\left\lceil \frac{1}{1-\beta_1} \right\rceil}} \leq \frac{\exp(-1)}{1-\exp(-1)} \leq 1.$$

Applying the bounds above to (28) and using that $\left\lceil \frac{1}{1-\beta_1} \right\rceil \leq \frac{2}{1-\beta_1}$, we derive

$$a_k \leq 2 + \ln\left(\frac{2}{1-\beta_1}\right)$$

for all $k \geq 0$. Additionally, one can easily verify that

$$a_0 \geq a_1 \geq \ldots \geq a_{t-1}.$$

Indeed,

$$\begin{aligned} a_k &= \sum_{j=k}^{t-1} \frac{(1-\beta_1)\beta_1^{j-k}}{1-\beta_1^{j+1}} = \sum_{j=k}^{t-2} \frac{(1-\beta_1)\beta_1^{j-k}}{1-\beta_1^{j+1}} + \frac{(1-\beta_1)\beta_1^{t-1-k}}{1-\beta_1^t} \geq \sum_{j=k}^{t-2} \frac{(1-\beta_1)\beta_1^{j-k}}{1-\beta_1^{j+1}} \\ &= \sum_{j=k+1}^{t-1} \frac{(1-\beta_1)\beta_1^{j-k-1}}{1-\beta_1^j} \geq \sum_{j=k+1}^{t-1} \frac{(1-\beta_1)\beta_1^{j-k-1}}{1-\beta_1^{j+1}} = a_{k+1}. \end{aligned}$$

Then, if we denote $S_r = \sum\limits_{k=0}^{r-1} \operatorname{sign}(g_k)$, one can derive

$$\begin{aligned} P_t &= \sum_{k=0}^{t-1} a_k (S_{k+1} - S_k) = a_{t-1} S_t + \sum_{k=0}^{t-2} (a_k - a_{k+1}) S_{k+1} \\ &\leq \max_{k=0,\ldots,t} S_k \left(a_{t-1} + \sum_{k=0}^{t-2} (a_k - a_{k+1})\right) = a_0 \max_{k=0,\ldots,t} S_k \\ &\leq \left(2 + \ln\left(\frac{2}{1-\beta_1}\right)\right) \max_{k=0,\ldots,t} S_k, \end{aligned} \tag{29}$$

where we use $S_0 = 0$, $(a_k - a_{k+1}) \geq 0$ and $a_k \leq 2 + \ln\left(\frac{2}{1-\beta_1}\right)$. Therefore, applying (29) to (26), one can have

$$\begin{aligned} \mathbb{P}\left\{\min_{t=0,\ldots,K} f(x_t) - f^* > \varepsilon\right\} &\geq \mathbb{P}\left\{\forall t = 0,\ldots,K \quad P_t < \frac{5R_0}{8\gamma}\right\} \\ &\geq \mathbb{P}\left\{\forall t = 0,\ldots,K \quad \left(2 + \ln\left(\frac{2}{1-\beta_1}\right)\right) \max_{k=0,\ldots,t} S_k < \frac{5R_0}{8\gamma}\right\} \\ &= \mathbb{P}\left\{\left(2 + \ln\left(\frac{2}{1-\beta_1}\right)\right) \max_{k=0,\ldots,K} S_k < \frac{5R_0}{8\gamma}\right\} \\ &\geq \mathbb{P}\left\{\forall k \geq 0 \quad \left(2 + \ln\left(\frac{2}{1-\beta_1}\right)\right) S_k < \frac{5R_0}{8\gamma}\right\}. \end{aligned}$$

The probability of RHS is equal to the probability of never achieving the upper bound $\frac{5R_0}{8\gamma\left(2+\ln\left(\frac{2}{1-\beta_1}\right)\right)}$ in the random walk; thus, one can get

$$\mathbb{P}\left\{\forall k \geq 0 \quad \left(2 + \ln\left(\frac{2}{1-\beta_1}\right)\right) S_k < \frac{5R_0}{8\gamma}\right\} = \begin{cases} 1 - \left(\frac{p_{\text{step}}}{1-p_{\text{step}}}\right)^{\left\lceil \frac{5R_0}{8\gamma\left(2+\ln\left(\frac{2}{1-\beta_1}\right)\right)} \right\rceil}, & p_B < \frac{1}{2}; \\ 0, & p_B \geq \frac{1}{2}, \end{cases}$$

where p_{step} denotes the probability of $\operatorname{sign}(g_k) = 1$; according to (25), it is equal to $1 - (1-p)^B$. Since the case $Bp < \frac{1}{2}$ is considered, we have

$$1 - (1-p)^B \leq Bp < \frac{1}{2}.$$

Consequently, we derive

$$\mathbb{P}\left\{\min_{t=0,\ldots,K} f(x_t) - f^* > \varepsilon\right\} \geq 1 - \left(\frac{1-(1-p)^B}{(1-p)^B}\right)^{\left\lceil \frac{5R_0}{8\gamma\left(2+\ln\left(\frac{2}{1-\beta_1}\right)\right)} \right\rceil}.$$

Now let us find the constraints on B. Comparing the probability above with failure probability δ, we get

$$1 - \left(\frac{1-(1-p)^B}{(1-p)^B}\right)^{C(\gamma,\beta_1)} \geq \delta.$$

where $C(\gamma,\beta_1) = \left\lceil \frac{5R_0}{8\gamma\left(2+\ln\left(\frac{2}{1-\beta_1}\right)\right)} \right\rceil$. Then, we have

$$(1-p)^B \geq \frac{1}{1+(1-\delta)^{\frac{1}{C(\beta_1)}}}$$

It is equivalent to

$$B \leq \frac{\ln\left(1+(1-\delta)^{\frac{1}{C(\gamma,\beta_1)}}\right)}{-\ln(1-p)}.$$

Consequently, using that $\frac{-1}{\ln(1-p)} \geq \frac{1}{2p}$, one can obtain

$$B \leq \frac{\ln\left(1+(1-\delta)^{\frac{1}{C(\gamma,\beta_1)}}\right)}{2p} \Rightarrow \mathbb{P}\left\{\min_{k=0,\ldots,K}(f(x_k)-f^*) > \varepsilon\right\} \geq \delta,$$

Substituting p (in the case $Bp \leq \frac{1}{2}$ only the second choice of p is possible due to $B \geq 1$) and μ, we get

$$\mathbb{P}\left\{\min_{k=0,\ldots,K}(f(x_k)-f^*) > \varepsilon\right\} \leq \delta \Rightarrow B \geq \ln\left(1+(1-\delta)^{\frac{1}{C(\gamma,\beta_1)}}\right) \cdot 2^{-\frac{4\alpha}{\alpha-1}} \left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}.$$

Combining the above with the lower bound from $Bp > \frac{1}{2}$ and using that $B \geq 1$, we conclude

$$\mathbb{P}\left\{\min_{k=0,\ldots,K}(f(x_k)-f^*) > \varepsilon\right\} < \delta \Rightarrow B = \Omega\left(\max\left\{1, \ln\left(1+(1-\delta)^{\frac{1}{C(\gamma,\beta_1)}}\right)\left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right).$$

If $\delta \leq \frac{1}{2}$, $1+(1-\delta)^{\frac{1}{C(\gamma,\beta_1)}} \geq \frac{3}{2}$, and

$$\mathbb{P}\left\{\min_{k=0,\ldots,K}(f(x_k)-f^*) > \varepsilon\right\} < \delta \Rightarrow B = \Omega\left(\max\left\{1, \left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right).$$

This finishes the proof. □

Theorem 15 (Lower bound on oracle complexity)**.** *Suppose that $\varepsilon \leq \frac{LR_0^2}{64}$, $K \geq 1$, $\gamma > 0$ and $\{B_k\}_{k=0}$ is arbitrary predetermined batch size schedule. Also suppose that* Normalized SGD *satisfies*

$$\mathbb{P}\left\{\min_{k=0,\ldots,K}(f(x_k)-f^*) > \varepsilon\right\} < \delta$$

for every objective and stochastic oracles satisfying Assumptions 1–3. Then, we have

$$N_\varepsilon = \Omega\left(\max\left\{\frac{LR_0^2}{\varepsilon}, (1-\delta)^{1/M_\gamma}\sqrt{\frac{LR_0^2}{\varepsilon}}\left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right),$$

where $N_\varepsilon = \sum\limits_{k=0}^{K-1} B_k$ is the total number of oracle calls and $M_\gamma = \left\lceil \frac{13R_0}{16\gamma} \right\rceil \geq 1$. In particular, if $\delta \leq \frac{1}{2}$, we get

$$N_\varepsilon = \Omega\left(\max\left\{\frac{LR_0^2}{\varepsilon}, \sqrt{\frac{LR_0^2}{\varepsilon}}\left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right).$$

Proof. Before we start, let us denote

$$q = \frac{\alpha}{\alpha-1}, \qquad \kappa = \frac{LR_0^2}{\varepsilon}.$$

Part 1: deterministic lower bound. We first prove that a normalized-gradient method with constant step size must use a horizon of order κ to guarantee success uniformly over a stated class. In this part, the oracle is exact; hence the stochastic central moment is zero, so the oracle is admissible for every prescribed noise level σ. Let

$$\rho = \frac{4\varepsilon}{LR_0} = \frac{4R_0}{\kappa}.$$

Define the one-dimensional Huber function

$$f_\tau(x) = \begin{cases} \frac{L}{2}x^2, & |x| \le \tau, \\ L\tau|x| - \frac{L}{2}\tau^2, & |x| \ge \tau. \end{cases}$$

It is easy to verify that $f_\tau(x)$ is convex and L-smooth; thus, Assumptions 1 and 2 hold. This function will be useful for some cases in the proof below. Now we are ready to split the argument into three regimes for the constant step size γ.

Regime I: large steps fail for every horizon. In this part, we want to show that the constraint $\gamma \le \sqrt{\frac{8\varepsilon}{L}}$ is necessary for a uniform guarantee. Let $r_\varepsilon = \sqrt{\frac{2\varepsilon}{L}}$, and suppose that

$$\gamma > \sqrt{\frac{8\varepsilon}{L}} = 2r_\varepsilon.$$

If $\gamma \ge 2R_0$, we can consider the one-dimensional quadratic function

$$f(x) = \frac{L}{2}x^2.$$

Its ε-sublevel set is

$$\left\{x : \frac{L}{2}x^2 \le \varepsilon\right\} = [-r_\varepsilon, r_\varepsilon].$$

We initialize at $x_0 = R_0$. Then, $|x_0| = R_0$, the first step gives

$$x_1 = R_0 - \gamma \le -R_0,$$

and the next step gives

$$x_2 = x_1 + \gamma = R_0.$$

Thus the trajectory alternates between R_0 and $R_0 - \gamma$. Moreover,

$$|R_0 - \gamma| = \gamma - R_0 \ge R_0.$$

Additionally, using $\varepsilon \le \frac{LR_0^2}{64}$, we have

$$r_\varepsilon = \sqrt{\frac{2\varepsilon}{L}} \le \sqrt{\frac{2}{L} \cdot \frac{LR_0^2}{64}} = \frac{R_0}{4\sqrt{2}} < R_0.$$

Therefore, Normalized GD cannot achieve ε-accuracy. Otherwise, in the case $\gamma < 2R_0$ we consider two-dimensional function

$$F(x, y) = \frac{L}{2}x^2 + f_\tau(y)$$

We choose the initial point as follows. Since $\gamma < 2R_0$, let us choose odd n such that $n > \frac{2R_0}{\gamma}$. Then, $n \ge 3$. Next, we define $\tau = \sqrt{\frac{R_0^2 - \frac{\gamma^2}{4}}{n^2 - 1}}$. Therefore, the start point is $z_0 = (s, n\tau)$, where $s = \sqrt{\frac{\gamma^2}{4} - \tau^2}$ (it is correct due to $\tau < \frac{\gamma}{2}$ in such case). Consequently, $\|z_0\|^2 = s^2 + n^2\tau^2 = R_0^2$ due to the fact that $(0, 0)$ is the optimal point of $F(x, y)$. Now, we want to prove that the following holds for the trajectory:

$$x_k \in \{-s, s\}; \qquad |y_k| \ge \tau.$$

Indeed, the conditions above hold for the initial point. Let us assume that this holds for the point (x_k, y_k). Then, the update rule is

$$x_{k+1} = x_k - \gamma \frac{Lx_k}{L\gamma/2} = -x_k;$$
$$y_{k+1} = y_k - \gamma \frac{L\tau}{L\gamma/2} \operatorname{sign}(y_k) = y_k - 2\tau \operatorname{sign}(y_k),$$

where we use that $\nabla F(x_k, y_k) = L(x_k, \tau \operatorname{sign}(y_k))$. Since start point is $(s, n\tau)$ with odd n, all points from the trajectory can be expressed as $(\pm s, m\tau)$, where m is odd. Consequently, we derive

$$F(x_k, y_k) \geq \frac{L}{2}s^2 + \frac{L}{2}\tau^2 = \frac{L\gamma^2}{8} > \varepsilon.$$

As a result, ε-accuracy is also non-achievable. Thus, a method that succeeds uniformly over the class must satisfy

$$\gamma \leq \sqrt{\frac{8\varepsilon}{L}}. \tag{30}$$

Regime II: very small steps require a linear-in-κ number of iterations. Suppose that $\gamma < 2\rho$. It is possible due to $2\rho \leq \sqrt{\frac{8\varepsilon}{L}}$. In that case, let us use the one-dimensional objective $f(x) = f_\rho(x)$ and initialize at $x_0 = R_0$. Let x_ε be the positive endpoint of the ε-sublevel set, i.e., $f_\rho(x_\varepsilon) = \varepsilon$. This case is achievable in the linear part of function: it is enough to substitute $x = \rho$ to show that

$$f_\rho(\rho) = \frac{L}{2}\rho^2 = \frac{8\varepsilon^2}{LR_0^2} \leq \frac{\varepsilon}{8}.$$

Thus, we get

$$L\rho x_\varepsilon - \frac{L}{2}\rho^2 = \varepsilon,$$

and hence

$$x_\varepsilon = \frac{\varepsilon}{L\rho} + \frac{\rho}{2}.$$

Using $\rho = \frac{4\varepsilon}{LR_0}$, we have

$$x_\varepsilon = \frac{R_0}{4} + \frac{2\varepsilon}{LR_0}.$$

Since $\varepsilon \leq \frac{LR_0^2}{64}$, one can derive

$$x_\varepsilon \leq \frac{R_0}{4} + \frac{R_0}{32} = \frac{9R_0}{32}.$$

Consequently, to reach the ε-sublevel set, the method must decrease the variable at least

$$R_0 - x_\varepsilon \geq \frac{23R_0}{32}.$$

Each normalized-gradient step has length at most γ. Consequently, the convergence requires

$$K\gamma \geq \frac{23R_0}{32} \Rightarrow K \geq \frac{23R_0}{32\gamma}.$$

However, $\gamma < 2\rho$, which leads to

$$\gamma < 2\rho = \frac{8\varepsilon}{LR_0} = \frac{8R_0}{\kappa} \Rightarrow \frac{1}{\gamma} > \frac{\kappa}{8R_0}.$$

Therefore, we can provide the lower bound

$$K > \frac{23R_0}{32}\frac{\kappa}{8R_0} = \frac{23}{256}\kappa = \frac{23LR_0^2}{256\varepsilon}.$$

Regime III: intermediate steps produce an exact zigzag. The last option is

$$2\rho \le \gamma \le \sqrt{\frac{8\varepsilon}{L}}.$$

Let us define the two-dimensional objective

$$F(x, y) = \frac{L}{2}x^2 + f_\rho(y).$$

This function is convex and L-smooth, i.e. satisfies Assumptions 1 and 2. Next, we denote the start point as $z_0 = (s, r_0)$, where

$$s = \sqrt{\frac{\gamma^2}{4} - \rho^2}, \qquad r_0 = \sqrt{R_0^2 - s^2}.$$

Then, one can have

$$\|z_0\|^2 = s^2 + r_0^2 = R_0^2.$$

As long as $y_k > \rho$, we are in the linear branch of f_ρ, and hence

$$\nabla F(x_k, y_k) = L(x_k, \rho).$$

At either value $x_k \in \{s, -s\}$, the gradient norm is

$$\|\nabla F(x_k, y_k)\| = L\sqrt{s^2 + \rho^2}.$$

By the definition of s, $s^2 + \rho^2 = \gamma^2/4$. Thus,

$$\|\nabla F(x_k, y_k)\| = \frac{L\gamma}{2}.$$

Consequently, one normalized-gradient step gives

$$x_{k+1} = x_k - \gamma\frac{Lx_k}{L\gamma/2} = -x_k,$$
$$y_{k+1} = y_k - \gamma\frac{L\rho}{L\gamma/2} = y_k - 2\rho.$$

Hence, as long as $y_i > \rho$ for all $0 \le i \le k-1$, the trajectory behaves as:

$$x_k = (-1)^k s, \qquad y_k = y_0 - 2k\rho.$$

We now verify that the trajectory formula is self-consistent for a constant fraction of κ iterations. The bound on s gives

$$s^2 = \frac{\gamma^2}{4} - \rho^2 \le \frac{\gamma^2}{4} \le \frac{2\varepsilon}{L} \le \frac{R_0^2}{32},$$

where we use $\varepsilon \le LR_0^2/64$. Therefore, we get

$$r_0 = \sqrt{R_0^2 - s^2} \ge \sqrt{\frac{31}{32}}R_0 > \frac{3R_0}{4}.$$

In parallel, let us denote

$$T = \left\lfloor \frac{\kappa}{32} \right\rfloor.$$

For every $0 \le k \le T$, we have

$$2k\rho \le \frac{\kappa}{16} \cdot \frac{4R_0}{\kappa} = \frac{R_0}{4}.$$

Therefore,

$$y_k = r_0 - 2k\rho \ge \frac{3R_0}{4} - \frac{R_0}{4} = \frac{R_0}{2}.$$

Since

$$\rho = \frac{4\varepsilon}{LR_0} \leq \frac{4}{LR_0} \cdot \frac{LR_0^2}{64} = \frac{R_0}{16},$$

we have $y_k \geq \frac{R_0}{2} > \rho$. Thus, the trajectory formula is self-consistent for all $0 \leq k \leq T$. Moreover, for every such k,

$$F(x_k, y_k) \geq f_\rho(y_k).$$

Since $y_k > \rho$ for all $0 \leq k \leq T$, the linear formula for f_ρ applies:

$$f_\rho(y_k) = L\rho y_k - \frac{L}{2}\rho^2.$$

Using $y_k \geq \frac{R_0}{2}$, we get

$$f_\rho(y_k) \geq L\rho\frac{R_0}{2} - \frac{L}{2}\rho^2.$$

Substituting $\rho = \frac{4\varepsilon}{LR_0}$ gives

$$F(x_k, y_k) \geq 2\varepsilon - \frac{8\varepsilon^2}{LR_0^2}.$$

Since $\varepsilon \leq \frac{LR_0^2}{64}$, we have

$$F(x_k, y_k) \geq 2\varepsilon - \frac{\varepsilon}{8} = \frac{15}{8}\varepsilon > \varepsilon.$$

Hence none of the iterates $0 \leq k \leq T$ is ε-optimal. Using that $\kappa = \frac{LR_0^2}{\varepsilon} \geq 64$, we have

$$T = \left\lfloor \frac{\kappa}{32} \right\rfloor \geq \frac{\kappa}{64}.$$

Therefore, in considered regime, one can get

$$K \geq \frac{\kappa}{64}.$$

Deterministic lower bound. Combining the three regimes, Normalized GD with constant step size must satisfy

$$K \geq \frac{\kappa}{64}, \tag{31}$$

Equivalently,

$$K = \Omega\left(\frac{LR_0^2}{\varepsilon}\right).$$

Since every batch contains at least one oracle call, $B_k \geq 1$, we also have

$$N_\varepsilon = \sum_{k=0}^{K-1} B_k \geq K.$$

Therefore, the deterministic part alone already implies

$$N_\varepsilon = \Omega\left(\frac{LR_0^2}{\varepsilon}\right). \tag{32}$$

Moreover, the constant step size must satisfy (30); otherwise, Normalized GD cannot guarantee ε-accuracy uniformly over this class.

Part 2: stochastic lower bound. We now consider a second hard instance. In this part we assume the admissible step size condition (30). If (30) or (31) fails, the deterministic construction above (see **Regime I**) violates the required guarantee. Let

$$\mu = \frac{8\varepsilon}{R_0}.$$

Next, we define the one-sided Huber function

$$f(x) = \begin{cases} 0, & x \le 0, \\ \frac{L}{2}x^2, & 0 < x \le \frac{\mu}{L}, \\ \mu x - \frac{\mu^2}{2L}, & x > \frac{\mu}{L}. \end{cases}$$

The function $f(x)$ is convex and L-smooth, i.e. Assumptions 1 and 2 hold. Moreover, its minimizer is $x^* = 0$, and $F^* = 0$. Also we choose the initial point $x_0 = R_0$.

Stochastic oracle. Let us define the stochastic oracle

$$\hat{g}_t = \begin{cases} 0, & \text{with probability } 1 - p_\varepsilon, \\ \frac{\nabla f(x_t)}{p_\varepsilon}, & \text{with probability } p_\varepsilon. \end{cases}$$

We first consider the rare-signal regime $p_\varepsilon < 1$, where

$$p_\varepsilon = \min\left\{1, 2^{\frac{1}{\alpha-1}}\left(\frac{8\varepsilon}{\sigma R_0}\right)^q\right\}. \tag{33}$$

It is easy to verify that the stochastic oracle is unbiased, i.e.

$$\mathbb{E}[\hat{g}_t] = \nabla f(x_t).$$

Moreover, one can derive

$$\begin{aligned} \mathbb{E}\left\|\hat{g}_t - \nabla f(x_t)\right\|^\alpha &= (1-p_\varepsilon)\left\|\nabla f(x_t)\right\|^\alpha + p_\varepsilon\left\|\nabla f(x_t)\right\|^\alpha\left(\frac{1-p_\varepsilon}{p_\varepsilon}\right)^\alpha \\ &\le (1-p_\varepsilon)\left(\frac{8\varepsilon}{R_0}\right)^\alpha + p_\varepsilon\left(\frac{8\varepsilon}{R_0}\right)^\alpha\left(\frac{1-p_\varepsilon}{p_\varepsilon}\right)^\alpha \end{aligned}$$

where we use $\|\nabla f(x)\| \le \mu$ for all x due to the definition of f. Moreover, we get

$$1 - p_\varepsilon + (1-p_\varepsilon)^\alpha p_\varepsilon^{1-\alpha} \le 2p_\varepsilon^{1-\alpha}.$$

Applying the inequality above and substituting p_ε, one can obtain

$$\mathbb{E}\left\|\hat{g}_t - \nabla f(x_t)\right\|^\alpha \le \left(\frac{8\varepsilon}{R_0}\right)^\alpha\left(\frac{8\varepsilon}{\sigma R_0}\right)^{-\alpha} = \sigma^\alpha.$$

Thus, the oracle satisfies Assumption 3. We use an independent Bernoulli draw on each call.

Active batches. In this step of the proof, we consider the mini-batch estimator at iteration k, which we denote as g_k:

$$g_k = \frac{1}{B_k}\sum_{i=1}^{B_k}\hat{g}_k^i.$$

We call a batch *active*, if it contains at least one Huber signal, i.e., at least one oracle call within this batch. If a batch is not active, then all its oracle calls are equal to 0, and $x_{k+1} = x_k$.

Let x_ε be the point where $f(x_\varepsilon) = \varepsilon$. This point lies in the linear branch of the Huber function. Consequently,

$$f(x) = \mu x - \frac{\mu^2}{2L}.$$

Then, we get

$$\mu x_\varepsilon - \frac{\mu^2}{2L} = \varepsilon \Rightarrow x_\varepsilon = \frac{\varepsilon}{\mu} + \frac{\mu}{2L}.$$

Substituting $\mu = \frac{8\varepsilon}{R_0}$ gives

$$x_\varepsilon = \frac{R_0}{8} + \frac{4\varepsilon}{LR_0}.$$

Using that $\varepsilon \leq \frac{LR_0^2}{64}$, we get Hence

$$x_\varepsilon \leq \frac{R_0}{8} + \frac{R_0}{16} = \frac{3R_0}{16}.$$

The initial Huber coordinate is $x_0 = R_0$. Therefore, to make $f(x) \leq \varepsilon$, the method must decrease the coordinate by at least

$$\Delta_x = x_0 - x_\varepsilon \geq R_0 - \frac{3R_0}{16} = \frac{13R_0}{16}. \tag{34}$$

Since each active batch can decrease the coordinate by at most γ, success requires at least

$$M_\gamma = \left\lceil \frac{13R_0}{16\gamma} \right\rceil$$

active batches. Using the admissible stepsize condition (30), we obtain

$$M_\gamma \geq \frac{13R_0}{16}\sqrt{\frac{L}{8\varepsilon}}. \tag{35}$$

Let S_N be the number of Huber signals among all $N = N_\varepsilon$ oracle calls. Since each oracle call is a Huber signal independently with probability p_ε, we have

$$S_N \sim \mathrm{Bin}(N, p_\varepsilon).$$

If the method succeeds, then at least M_γ active batches must occur. Hence, we have

$$\{\text{success}\} \subseteq \{S_N \geq M_\gamma\}.$$

Therefore,

$$\mathbb{P}(\text{success}) \leq \mathbb{P}(S_N \geq M_\gamma).$$

If the success probability is at least $1 - \delta$, then

$$\mathbb{P}(S_N \geq M_\gamma) \geq 1 - \delta.$$

Let us upper bound $\mathbb{P}(S_N \geq M_\gamma)$ using the factorial moment. First,

$$\mathbf{1}\{S_N \geq M_\gamma\} \leq \binom{S_N}{M_\gamma}.$$

Taking expectations,

$$\mathbb{P}(S_N \geq M_\gamma) \leq \mathbb{E}\binom{S_N}{M_\gamma}.$$

For a binomial random variable $S_N \sim \mathrm{Bin}(N, p_\varepsilon)$,

$$\mathbb{E}\binom{S_N}{M_\gamma} = \binom{N}{M_\gamma} p_\varepsilon^{M_\gamma}.$$

Moreover, we can use the following inequality:

$$\binom{N}{M_\gamma} \leq \frac{N^{M_\gamma}}{M_\gamma!}.$$

Thus, we derive

$$\mathbb{P}(S_N \geq M_\gamma) \leq \frac{(Np_\varepsilon)^{M_\gamma}}{M_\gamma!}.$$

Since $\mathbb{P}(S_N \geq M_\gamma) \geq 1-\delta$, we get

$$1-\delta \leq \frac{(Np_\varepsilon)^{M_\gamma}}{M_\gamma!}.$$

Thus,

$$N \geq \frac{\left((1-\delta)M_\gamma!\right)^{1/M_\gamma}}{p_\varepsilon} \Rightarrow N \geq \frac{(1-\delta)^{1/M_\gamma}}{e}\frac{M_\gamma}{p_\varepsilon},$$

where we use $M_\gamma! \geq (M_\gamma/e)^{M_\gamma}$. Substituting (35) and applying the expression for p_ε, we derive

$$\begin{aligned} N &\geq \frac{(1-\delta)^{1/M_\gamma}}{e}\cdot\frac{13R_0}{16}\sqrt{\frac{L}{8\varepsilon}}\cdot\frac{1}{p_\varepsilon} = \frac{13(1-\delta)^{1/M_\gamma}}{16e}\sqrt{\frac{LR_0^2}{8\varepsilon}}\frac{1}{p_\varepsilon} \\ &= \frac{13(1-\delta)^{1/M_\gamma}}{16e}\sqrt{\frac{LR_0^2}{8\varepsilon}}2^{-\frac{1}{\alpha-1}}\left(\frac{\sigma R_0}{8\varepsilon}\right)^q = \frac{13(1-\delta)^{1/M_\gamma}}{32\sqrt{2}e}\cdot 2^{-\frac{1}{\alpha-1}}\sqrt{\frac{LR_0^2}{\varepsilon}}\left(\frac{\sigma R_0}{8\varepsilon}\right)^q. \end{aligned}$$

Finally, since $q = \frac{\alpha}{\alpha-1}$, one can have

$$N = \Omega\left((1-\delta)^{1/M_\gamma}\sqrt{\frac{LR_0^2}{\varepsilon}}\left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right). \tag{36}$$

Complementary regime: $p_\varepsilon = 1$**.** In this case, the definition of p_ε implies

$$\left(\frac{\sigma R_0}{\varepsilon}\right)^q \leq 2^{\frac{1}{\alpha-1}}8^q.$$

The right-hand side depends only on α. Since $\kappa \geq 64$, we have $\sqrt{\kappa} \leq \kappa$, and hence the deterministic lower bound (32) also implies the claimed stochastic term in this regime.

Final conclusion. The assumed uniform guarantee must hold on both the deterministic and stochastic hard instances. Therefore, combining (32) with (36) and the complementary-regime argument gives

$$N_\varepsilon = \Omega\left(\max\left\{\frac{LR_0^2}{\varepsilon}, (1-\delta)^{1/M_\gamma}\sqrt{\frac{LR_0^2}{\varepsilon}}\left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right).$$

In particular, if $\delta \leq \frac{1}{2}$, one can use $M_\gamma \geq 1$ to derive

$$N_\varepsilon = \Omega\left(\max\left\{\frac{LR_0^2}{\varepsilon}, \sqrt{\frac{LR_0^2}{\varepsilon}}\left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right)$$

This finishes the proof. □

Theorem 16 (Lower bound on oracle complexity for non-decreasing batch sizes)**.** *Suppose that $\varepsilon \leq \frac{LR_0^2}{64}$, $K \geq 1$, $\gamma > 0$, $\delta \leq \frac{1}{2}$ and $\{B_k\}_{k=0}$ is predetermined non-decreasing batch size schedule. Also suppose that* Normalized SGD *satisfies*

$$\mathbb{P}\left\{\min_{k=0,\ldots,K}(f(x_k)-f^*) > \varepsilon\right\} < \delta$$

for every objective and stochastic oracles satisfying Assumptions 1–3. Then, we have

$$N_\varepsilon = \Omega\left(\frac{LR_0^2}{\varepsilon}\max\left\{1, \left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right)$$

where $N_\varepsilon = \sum_{k=0}^{K-1} B_k$ is the total number of oracle calls after K iterations.

Proof. As in the proof of Theorem 15, let us denote

$$q = \frac{\alpha}{\alpha-1}, \qquad \kappa = \frac{LR_0^2}{\varepsilon}.$$

Part 1: deterministic lower bound. We first recall the deterministic part of the proof of Theorem 15. It shows that constant step size Normalized GD which reaches an ε-accuracy function-value gap uniformly over the class of convex L-smooth functions with initial distance R_0 must use

$$K \geq c_0 \kappa, \qquad c_0 = \frac{1}{64}. \tag{37}$$

iterations (and oracle calls).

Part 2: stochastic hard instance. Now we prove the stochastic part of the lower bound. We assume that the step size satisfies

$$\gamma \leq \sqrt{\frac{8\varepsilon}{L}}. \tag{38}$$

If this restriction or (37) is violated, a deterministic instance from the proof of Theorem 15 already violates the required guarantee. Otherwise, both restrictions hold, and we apply the uniform guarantee to a new stochastic instance. Let

$$\mu = \frac{2\varepsilon}{R_0}.$$

Consider the one-sided Huber function

$$f(x) = \begin{cases} 0, & x \leq 0, \\ \frac{L}{2}x^2, & 0 < x \leq \frac{\mu}{L}, \\ \mu x - \frac{\mu^2}{2L}, & x > \frac{\mu}{L}. \end{cases}$$

This function is convex and L-smooth; therefore, Assumptions 1 and 2 hold. Moreover, $x^* = 0$ and $f^* = 0$. We initialize the method at $x_0 = R_0$. Let x_ε be the positive point such that $f(x_\varepsilon) = \varepsilon$. Since

$$f\left(\frac{\mu}{L}\right) = \frac{\mu^2}{2L} = \frac{2\varepsilon^2}{LR_0^2} \leq \frac{\varepsilon}{32} < \varepsilon,$$

the point x_ε lies in the linear branch of f. Thus,

$$\mu x_\varepsilon - \frac{\mu^2}{2L} = \varepsilon \Rightarrow x_\varepsilon = \frac{\varepsilon}{\mu} + \frac{\mu}{2L} = \frac{R_0}{2} + \frac{\varepsilon}{LR_0}$$

Using $\varepsilon \leq \frac{LR_0^2}{64}$, we get

$$x_\varepsilon \leq \frac{R_0}{2} + \frac{R_0}{64} = \frac{33R_0}{64}.$$

Therefore, to reach the ε-sublevel set, the method must decrease the coordinate by at least $\frac{31R_0}{64}$.

Stochastic oracle. Let us define the following two-point stochastic oracle:

$$\hat{g}_t = \begin{cases} -\nabla f(x_t), & \text{with probability } 1-p, \\ \frac{2-p}{p}\nabla f(x_t), & \text{with probability } p, \end{cases}$$

where

$$p = 2^{\frac{\alpha+1}{\alpha-1}} \left(\frac{\mu}{\sigma}\right)^q = 2^{\frac{2\alpha+1}{\alpha-1}} \left(\frac{\varepsilon}{\sigma R_0}\right)^q. \tag{39}$$

First, we consider the rare-signal regime

$$p \leq \frac{1}{16}. \tag{40}$$

The oracle is unbiased, since

$$\mathbb{E}[\hat{g}_t] = -(1-p)\nabla f(x_t) + (2-p)\nabla f(x_t) = \nabla f(x_t).$$

Moreover, the α-central moment is bounded. Indeed,

$$\mathbb{E}\left[\|\hat{g}_t - \nabla f(x_t)\|^\alpha\right] = 2^\alpha \left((1-p) + (1-p)^\alpha p^{1-\alpha}\right) \|\nabla f(x_t)\|^\alpha \leq 2^{\alpha+1} p^{1-\alpha} \|\nabla f(x_t)\|^\alpha.$$

Since $\|\nabla f(x)\| \leq \mu$ for all x, the definition of p gives

$$\mathbb{E}\left[\|\hat{g}_t - \nabla f(x_t)\|^\alpha\right] \leq 2^{\alpha+1} p^{1-\alpha} \mu^\alpha = \sigma^\alpha.$$

Thus, Assumption 3 holds.

Behavior of the mini-batch estimator. Before reaching the ε-sublevel set, we have $x_k > x_\varepsilon > \frac{\mu}{L}$, and hence $\nabla f(x_k) = \mu$. Let J_k be the number of rare positive samples in the k-th mini-batch. Then, we have

$$J_k \sim \mathrm{Bin}(B_k, p).$$

The batched estimator can be written as

$$g_k = \frac{1}{B_k}\left((B_k - J_k)(-\mu) + J_k \frac{2-p}{p}\mu\right) = \mu\left(-1 + \frac{2J_k}{B_k p}\right).$$

If

$$B_k p \le \frac{1}{16}, \tag{41}$$

then $g_k < 0$ whenever $J_k = 0$, and $g_k > 0$ whenever $J_k \ge 1$. Therefore, in this case the probability of a step toward the optimum is

$$\vartheta_k := \mathbb{P}\{g_k > 0\} = \mathbb{P}\{J_k \ge 1\} = 1 - (1-p)^{B_k} \le B_k p \le \frac{1}{16}. \tag{42}$$

Random-walk estimate. Before reaching the ε-sublevel set, the iterate is in the linear branch of the Huber function. Therefore, the update direction is determined only by the sign of g_k. Let us denote

$$Z_k = \mathrm{sign}(g_k), \qquad S_r = \sum_{k=0}^{r-1} Z_k.$$

Here $Z_k = 1$ corresponds to a step toward the optimum, while $Z_k = -1$ corresponds to a step away from the optimum. Consequently, before success we have

$$x_r = x_0 - \gamma S_r.$$

Thus, in order to reach the ε-sublevel set, it is necessary to have

$$S_r \ge \frac{\Delta}{\gamma} \ge \frac{31 R_0}{64\gamma}$$

for some r. Let

$$m = \left\lceil \frac{31 R_0}{64\gamma} \right\rceil, \qquad \tau_{\mathrm{hit}} = \inf\{0 \le r \le K : S_r \ge m\}. \tag{43}$$

Then, for any $0 \le n \le K$, we have

$$\{\tau_{\mathrm{succ}} \le n\} \subseteq \{\tau_{\mathrm{hit}} \le n\}, \tag{44}$$

where

$$\tau_{\mathrm{succ}} = \inf\{0 \le r \le K : f(x_r) - f^* \le \varepsilon\}$$

Assume now that all batches under consideration satisfy

$$B_k p \le \frac{1}{16}.$$

Then, according to (42), the conditional probability of the event $Z_k = 1$ is bounded as

$$\mathbb{P}\{Z_k = 1 \mid \mathcal{F}_k\} \le \theta, \qquad \theta = \frac{1}{16}.$$

Let $a = \frac{\theta}{1-\theta} = \frac{1}{15}$. We claim that a^{-S_r} is a supermartingale over this prefix of small batches. Indeed,

$$\begin{aligned}\mathbb{E}\left[a^{-S_{r+1}} \mid \mathcal{F}_r\right] &= a^{-S_r}\left(\mathbb{P}\{Z_r = 1 \mid \mathcal{F}_r\}a^{-1} + \mathbb{P}\{Z_r = -1 \mid \mathcal{F}_r\}a\right) \\ &\le a^{-S_r}\left(\theta a^{-1} + (1-\theta)a\right) = a^{-S_r}.\end{aligned}$$

Therefore, by optional stopping,

$$1 = a^{-S_0} \ge \mathbb{E}\left[a^{-S_{\min\{\tau_{\mathrm{hit}}, n\}}}\right] \ge \mathbb{P}\{\tau_{\mathrm{hit}} \le n\} a^{-m}.$$

Thus,

$$\mathbb{P}\{\tau_{\mathrm{hit}} \le n\} \le a^m \le \frac{1}{15}, \tag{45}$$

where in the last step we used $m \ge 1$.

Using the non-decreasing schedule. Let $n = \lceil \frac{4K}{5} \rceil$. Now, we consider two cases.

Case 1: a large batch appears before time n. Assume that there exists $t < n$ such that

$$B_t p > \frac{1}{16}.$$

Then, one can get

$$B_t > \frac{1}{16p}.$$

Since the batch-size schedule is non-decreasing, for all $k \geq t$ we have

$$B_k \geq B_t > \frac{1}{16p}.$$

Moreover, since $t < n = \lceil \frac{4K}{5} \rceil$, the suffix length satisfies

$$K - t \geq \frac{K}{5}.$$

Thus, we derive

$$N_\varepsilon = \sum_{k=0}^{K-1} B_k \geq \sum_{k=t}^{K-1} B_k > \frac{K}{5} \cdot \frac{1}{16p} = \frac{K}{80p}. \tag{46}$$

Case 2: all first n batches are small. Assume now that

$$B_k p \leq \frac{1}{16}, \qquad k = 0, \ldots, n-1.$$

Let T_n be the number of steps toward the optimum among the first n steps before success:

$$T_n = \sum_{k=0}^{n-1} \mathbf{1}\{k < \tau_{\text{succ}}\} \mathbf{1}\{\text{the } k\text{-th step is toward the optimum}\}.$$

Using (42), we get

$$\mathbb{E}[T_n] \leq \frac{n}{16}.$$

Therefore, by Markov's inequality,

$$\mathbb{P}\left\{T_n \geq \frac{n}{4}\right\} \leq \frac{1}{4}. \tag{47}$$

At the same time, by (45),

$$\mathbb{P}\{\tau_{\text{succ}} \leq n\} \leq \mathbb{P}\{\tau_{\text{hit}} \leq n\} \leq \frac{1}{15}. \tag{48}$$

On the event $\{\tau_{\text{succ}} > n\} \cap \left\{T_n < \frac{n}{4}\right\}$ the random walk satisfies

$$S_n = T_n - (n - T_n) = 2T_n - n < -\frac{n}{2}.$$

Since $n \geq \frac{4K}{5}$, we have

$$S_n < -\frac{2K}{5}.$$

After time n, there are at most $K - n \leq \frac{K}{5}$ iterations left. Even if all these remaining steps are toward the optimum, the random walk can increase by at most $\frac{K}{5}$. Therefore, for all $r \in \{n, \ldots, K\}$, we have

$$S_r < -\frac{2K}{5} + \frac{K}{5} = -\frac{K}{5} < 0.$$

In particular, the random walk cannot hit the positive level $m \geq 1$. Hence, success is impossible on this event. Consequently,

$$\begin{aligned}\mathbb{P}\{\text{success}\} &\leq \mathbb{P}\{\tau_{\text{succ}} \leq n\} + \mathbb{P}\left\{T_n \geq \frac{n}{4}\right\} \\ &\leq \frac{1}{15} + \frac{1}{4} = \frac{19}{60} < \frac{1}{2}.\end{aligned}$$

This contradicts the required guarantee $1 - \delta \geq \frac{1}{2}$. Thus, Case 2 is impossible. Therefore, Case 1 must hold, and by (46),

$$N_\varepsilon \geq \frac{K}{80p}. \tag{49}$$

Rare-signal regime. The fixed horizon already satisfies (37), since success is required on every admissible instance. Combining this restriction with (49), obtained on the stochastic instance, we have

$$N_\varepsilon \geq \frac{1}{80p} \cdot \frac{LR_0^2}{64\varepsilon} = \frac{1}{5120p} \frac{LR_0^2}{\varepsilon}.$$

Substituting the value of p from (39), we get

$$N_\varepsilon \geq \frac{2^{-\frac{2\alpha+1}{\alpha-1}}}{5120} \frac{LR_0^2}{\varepsilon} \left(\frac{\sigma R_0}{\varepsilon}\right)^q. \tag{50}$$

Complementary regime. It remains to consider the case

$$p > \frac{1}{16}.$$

According to (39), this implies

$$2^{\frac{2\alpha+1}{\alpha-1}} \left(\frac{\varepsilon}{\sigma R_0}\right)^q > \frac{1}{16}.$$

Equivalently,

$$\left(\frac{\sigma R_0}{\varepsilon}\right)^q < 16 \cdot 2^{\frac{2\alpha+1}{\alpha-1}}.$$

Thus, the stochastic multiplier is bounded by a constant depending on α. Therefore, the deterministic lower bound implies

$$N_\varepsilon = \Omega\left(\frac{LR_0^2}{\varepsilon} \max\left\{1, \left(\frac{\sigma R_0}{\varepsilon}\right)^q\right\}\right) \tag{51}$$

in the complementary regime.

Final conclusion. Combining (32), (50), and (51), we conclude

$$N_\varepsilon = \Omega\left(\frac{LR_0^2}{\varepsilon} \max\left\{1, \left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right).$$

This finishes 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 state the deterministic and high-probability best-iterate guarantees and the restrictions on the algorithms covered by the lower bounds; see Sections 1 and 4.

Guidelines:

- The answer [N/A] 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 [N/A] 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: We discuss the assumptions, the best-iterate criterion, the restrictions to constant stepsizes and predetermined batch sizes in the lower bounds, and open questions on adaptive batches and inner momentum in Section 4 and the Conclusion.

Guidelines:

- The answer [N/A] 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 are stated in Section 2 and in the theorem statements. Complete proofs are provided in Appendices A–D.

Guidelines:

- The answer [N/A] 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: [N/A]

   Justification: The paper is theoretical and contains no numerical experiments.

   Guidelines:

   - The answer [N/A] 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: [N/A]

   Justification: The paper is theoretical and contains no numerical experiments.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper is theoretical and contains no numerical experiments.

   Guidelines:

   - The answer [N/A] 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: [N/A]

   Justification: The paper is theoretical and contains no numerical experiments.

   Guidelines:

   - The answer [N/A] 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: [N/A]

   Justification: The paper is theoretical and contains no numerical experiments.

   Guidelines:

   - The answer [N/A] 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 conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] 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: [N/A]

    Justification: This work establishes mathematical results for stochastic optimization and does not introduce a particular application or deployment with direct societal impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This theoretical paper does not release data or models that require safeguards.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: This theoretical paper does not reuse or redistribute third-party datasets, code, or model weights. Generative-AI assistance is disclosed separately in the Acknowledgments.

    Guidelines:

    - The answer [N/A] 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: [N/A]

Justification: The paper introduces theoretical results and algorithm descriptions, but no datasets, model checkpoints, or research software.

Guidelines:

- The answer [N/A] 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: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] 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: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] 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.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The "Use of generative AI" paragraph in the Acknowledgments describes assistance with text polishing, proof checking, and formalizing the lower-bound arguments. It attributes the upper bounds and key lower-bound ideas to the authors and states their responsibility for the results and the paper's content.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.