High-probability Convergence of Gradient Methods under Markovian Stochasticity

Polina Podzorova
MIRAI, BRAIn Lab

Nikolay Spitsyn
MIRAI, BRAIn Lab

Savelii Chezhegov
MIRAI, BRAIn Lab, Sber AI Lab

Aleksandr Beznosikov
MIRAI, BRAIn Lab, Innopolis University

Abstract

Markovian stochasticity naturally arises in many modern learning problems, including optimization with dependent data streams and reinforcement learning. Despite extensive research on convergence of stochastic gradient methods under such stochasticity, high-probability convergence remains largely unexplored. To close this gap, we establish high-probability convergence guarantees for gradient-based methods under Markovian stochasticity. We study MLMC-SGD and establish high-probability convergence results in the non-convex setting under the classical bounded stochasticity assumption, and further extend the analysis to a weaker bounded-variance assumption by incorporating gradient clipping, yielding convergence guarantees for accelerated MLMC-SGD in the convex setting. For both methods, we characterize the oracle complexity via a concentration inequality that explicitly captures the effect of Markovian dependence on the resulting convergence rates.

1 Introduction

Stochastic Gradient Descent (Robbins and Monro, 1951) is a cornerstone of machine learning (Shalev-Shwartz and Ben-David, 2014) and deep learning (Goodfellow et al., 2016) due to its efficiency, simplicity, and low computational cost. However, its behavior can vary significantly across different specialized deep learning settings, of which there exist a great number. This variability may be attributed to numerous factors; nevertheless, the most important one is the *nature of the stochasticity* inherent in the problem. As a consequence, theoretical analyses rely on a variety of assumptions on the stochastic gradient, some of which have become classical, such as unbiasedness of the gradient, bounded stochasticity, bounded variance, affine noise, and related conditions.

One such assumption concerns the Markovian nature of the stochasticity, meaning that the random variables arise as realizations of an ergodic Markov chain. This assumption is quite natural in a range of settings, including GANs (Gidel et al., 2018; Creswell et al., 2018), reinforcement learning (Sutton et al., 1998; Mnih et al., 2015; Durmus et al., 2021; Doan et al., 2020a), and the training of large language models (Devlin et al., 2019; Brown et al., 2020; Zekri et al., 2024). Indeed, the stochasticity may vary from epoch to epoch due to the trajectory induced by the model's decision graph as well as due to the sequential consumption of tokens when constructing subsequent ones. These effects suggest an inherent *sequential dependence* in the stochastic process. Moreover, as training progresses, the model increasingly approximates the true data distribution, which may be interpreted as convergence toward a stationary one. For these reasons, assuming Markovian stochasticity within the problem is a natural and practice-oriented modeling choice, even though it may not fully capture the true underlying dynamics. However, this assumption complicates theoretical analysis: the stochastic

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

gradient becomes biased with respect to the current distribution, requiring additional techniques to control the resulting bias.

On the other hand, the research community has long been concerned with the question of how strong theoretical convergence guarantees can be established for the application of specific techniques across diverse practical settings. Most existing convergence analyses are formulated in expectation and therefore characterize the behavior of an algorithm *only on average*. However, modern deep learning problems are so large-scale and data-intensive (Hestness et al., 2017; Rajbhandari et al., 2020) that running the training procedure multiple times is often infeasible. Consequently, one typically aims to train a model *only once*, and it is precisely for such a single run that convergence guarantees are most desirable. While probability boosting can recover a logarithmic dependence on failure probability from in-expectation guarantees, it requires multiple independent runs and a reliable output-selection procedure. In the Markovian setting, independent trajectories or justified restarts are not always available. In this context, high-probability convergence analyses (Gorbunov et al., 2020; Cutkosky and Mehta, 2021; Liu et al., 2023; Sadiev et al., 2023; Nguyen et al., 2023; Chezhegov et al., 2024) constitute a particularly attractive alternative, as they provide substantially stronger guarantees than expectation-based analyses and hold with a prescribed confidence level over the randomness of a single run of the stochastic process. For SGD or other gradient methods, such guarantees have been established in a number of recent works under various assumptions on the stochasticity: (i) light-tailed noise (Harvey et al., 2019; Liu et al., 2023; Madden et al., 2024), which allows one to avoid any additional algorithmic modifications; and (ii) heavy-tailed noise, which is arguably more realistic and typically necessitates the incorporation of techniques such as normalization (Cutkosky and Mehta, 2021; Hübler et al., 2024) or gradient clipping (Gorbunov et al., 2020; Sadiev et al., 2023; Nguyen et al., 2023; Chezhegov et al., 2024). However, such analyses remain relatively limited in the Markovian setting.

In light of the above discussion, a natural open question arises:

Can one establish high-probability convergence guarantees for gradient-based methods under the assumption of Markovian stochasticity under bounded-noise and bounded-second-moment assumptions?

To address this question, we formulate our contributions as follows.

Our contributions. In this work, we provide theoretical convergence guarantees for gradient-based methods under the assumption of Markovian stochasticity.

1. First, we establish high-probability convergence guarantees for MLMC-SGD in smooth non-convex optimization under bounded stochasticity.
2. We extend the theory to a specialized stochastic estimator designed to control the bias induced by Markovian stochastic gradients; in particular, we demonstrate that this estimator can be analyzed under a weaker bounded-variance assumption than in prior work. We then establish **high-probability convergence guarantees for Accelerated MLMC-SGD in the convex setting**, using gradient clipping as a mechanism for high-probability analysis.
3. We derive **oracle complexity bounds** for both algorithms, relying additionally on a suitable concentration inequality.

2 Related Works

Markovian stochasticity. The study of SGD under Markovian noise has been largely motivated by applications in reinforcement learning (Doan et al., 2020b; Sutton and Barto, 2018), where the data naturally form a Markov chain. Recent work has shown that gradients in reinforcement learning exhibit heavy-tailed behavior (Garg et al., 2021). This observation further motivates the need for robust gradient methods under more general noise assumptions, but the theoretical results themselves are of independent interest and have been extended to broader classes of problems.

Early studies date back to Duchi et al. (2012), where the authors analyzed the ergodic mirror descent with optimal rates for non-smooth optimization. Later, Dorfman and Levy (2022) introduced the technique of stochastic batching based on a truncated geometric distribution, which allowed them to adapt to the mixing time of a Markov chain, albeit under the assumption of a compact domain.

Table 1: Comparison of the results for SGD for non-convex problems.

Reference	Unbounded noise	Markovian noise	High-probability	Rate
Dorfman and Levy (2022)	✗	✓	✗	$\tilde{\mathcal{O}}\left(\frac{\sqrt{\tau_{\text{mix}}}(G+L+B)G}{\sqrt{K}}\right)$ [(1)]
Even (2023)	✗	✓✗ [(2)]	✗	$\mathcal{O}\left(\frac{\tau_{\text{mix}}(L\Delta_0+\sigma^2)}{K}+\frac{\sqrt{\tau_{\text{mix}}(L\Delta_0+\sigma^2)}\sigma}{\sqrt{K}}\right)$
Beznosikov et al. (2023)	✓✗ [(3)]	✓	✗	$\tilde{\mathcal{O}}\left(\frac{\tau_{\text{mix}}L\Delta_0}{K}+\frac{\sqrt{\tau_{\text{mix}}L\Delta_0}\sigma}{\sqrt{K}}\right)$
Liu et al. (2023)	✓✗ [(4)]	✗	✓	$\mathcal{O}\left(\frac{L\Delta_0}{K}+\frac{\sigma\sqrt{L\Delta_0}}{\sqrt{K}}+\frac{\sigma^2\log\frac{1}{\delta}}{K}\right)$
This work	✗	✓	✓	$\tilde{\mathcal{O}}\left(\frac{L\Delta_0}{K}+\frac{\tau_{\text{mix}}\sigma^2}{K}+\frac{\sqrt{\tau_{\text{mix}}}\sigma\sqrt{L\Delta_0}}{\sqrt{K}}\right)$

[(1)] Dorfman and Levy (2022) use the following notation: $G = \sup_{x,Z} \|\nabla f(x,Z)\|$, $B = \sup_x |f(x)|$ and $D = \sup_{x\in\mathcal{X}} \|x - x^*\|$, where $\mathcal{X}$ is compact.
[(2)] Even (2023) considers a Markov chain with a finite state space.
[(3)] Beznosikov et al. (2023) use the assumption $\|\nabla f(x,\xi) - \nabla f(x)\|^2 \le \sigma^2 + \beta^2 \|\nabla f(x)\|^2$, which can be reduced to Assumption 6 with $\beta = 0$ for the comparison.
[(4)] Liu et al. (2023) consider sub-Gaussian noise.

Subsequently, a line of research emerged focusing on non-accelerated SGD-type methods that provided finite-time analysis (Sun et al., 2018; Doan, 2022). However, the existing analyses are either based on strong assumptions, as in Sun et al. (2018) – where the domain was assumed to be bounded and the stochastic gradients uniformly bounded – or they exhibit an exponential dependence on the mixing time (Doan, 2022). In addition, it is worth mentioning Sun et al. (2022), who considered the distributed problem, and Wang et al. (2022), who addressed the finite-sum saddle point problem.

Accelerated algorithms have also been developed for the Markovian stochastic setting. For instance, Doan et al. (2020a) proposed an accelerated method, but the resulting bounds exhibited suboptimal dependence on the mixing time. Subsequently, Even (2023) provided the first analysis for non-convex and strongly convex settings, achieving a linear dependence on the mixing time. Later Beznosikov et al. (2023) improved convergence results for the non-convex case and provided the first accelerated method guarantees for the strongly convex case with linear mixing-time dependence. Recently, by extending these results to the non-Euclidean setting, Solodkin et al. (2024) obtained optimal rates for convex optimization.

High-probability convergence. In recent years, high-probability convergence analyses have become increasingly popular across various optimization settings. Some of the earliest assumptions under which such guarantees were established for gradient methods involved bounded stochasticity and light-tailed noise. Notable works include Nemirovski et al. (2009); Kakade and Tewari (2008); Rakhlin et al. (2011); Harvey et al. (2019); Liu et al. (2023), which study high-probability convergence of SGD. Subsequently, similar bounds were derived for accelerated variants of SGD (Ghadimi and Lan, 2012; Dvurechensky and Gasnikov, 2016). Later, these analyses were extended to the case of sub-Weibull noise (Madden et al., 2024). In addition, Gaash et al. (2025) provided convergence results under generalized smoothness assuming sub-Gaussian noise.

With respect to the bounded-variance assumption, the first high-probability analysis was conducted in Nazin et al. (2019), which studied truncated Stochastic Mirror Descent. Subsequently, Gorbunov et al. (2020) provided theoretical convergence guarantees for both SGD and the accelerated algorithm SSTM (Gasnikov and Nesterov, 2016) by incorporating *gradient clipping*, which sparked a large body of follow-up research. Later, results appeared that naturally relaxed the bounded-variance requirement to bounded α-moment with $\alpha < 2$ (Cutkosky and Mehta, 2021; Sadiev et al., 2023; Nguyen et al., 2023), which is more prevalent in practice (Zhang et al., 2020). Moreover, prior works have addressed a variety of extensions, including non-smooth optimization (Gorbunov et al., 2024a; Parletta et al., 2024), variational inequalities (Gorbunov et al., 2024b), adaptive methods (Li and Liu, 2023; Chezhegov et al., 2024), generalized smoothness (Chezhegov et al., 2025) and zero-order approaches (Kornilov et al., 2023).

Despite the widespread use of gradient clipping for high-probability convergence analyses, alternative approaches also exist. For example, normalization techniques (Cutkosky and Mehta, 2021; Hübler et al., 2024) and sign-based methods (Kornilov et al., 2025) can provide high-probability convergence guarantees; however, these approaches typically require large batch sizes. In addition, there are robust distance-based techniques (Davis et al., 2021) that can achieve similar guarantees. Furthermore, additional assumptions on the noise, such as symmetry, allow for high-probability convergence results even for nonlinear transformations of stochastic gradient (Armacki et al., 2024).

3 Preliminaries

In this section, we introduce the required notation, problem formulation, and assumptions under which we construct our theory.

Notation. The Euclidean norm in $\mathbb{R}^d$ is denoted as $\|x\| = \sqrt{\langle x, x\rangle}$. $\mathbb{E}_\xi[\cdot]$ is the expectation w.r.t. random variable ξ. (Z, d_Z) denotes a complete separable metric space endowed with its Borel σ-field $\mathcal{Z}$. $(Z^{\mathbb{N}}, \mathcal{Z}^{\otimes\mathbb{N}})$ is the corresponding canonical process. Considering the Markov kernel Q defined on $Z \times \mathcal{Z}$, we denote $\mathbb{P}_\xi$ and $\mathbb{E}_\xi$ as the corresponding probability distribution and the expected value with initial distribution ξ. We also define the corresponding canonical process $(\xi_k)_{k\in\mathbb{N}}$. If $\xi = \delta_z$, $z \in Z$, we write $\mathbb{P}_z$ and $\mathbb{E}_z$ instead of $\mathbb{P}_{\delta_z}$ and $\mathbb{E}_{\delta_z}$. We denote by $\mathcal{F}_k$ the complete history before iteration k, including the iterates, previously sampled geometric levels, and Markov-chain observations. We write $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_k]$. For a random *sequence* $\xi = \{\xi_1, \ldots, \xi_n\}$, we write $\mathbb{E}_\xi[\cdot]$ to denote the full expectation over Markov observations. The clipping operator is denoted as $\mathrm{clip}(x, \lambda) := \min\left\{1, \frac{\lambda}{\|x\|}\right\} x$.

Problem. In this work, we focus on the classical problem of stochastic optimization:

$$\min\{f(x) := \mathbb{E}_{\xi\sim\mathcal{D}}[f(x, \xi)]\}.$$

This formulation is standard and appears in many areas of machine learning. Here, f denotes a loss function, x represents the optimization variables, $\mathcal{D}$ denotes the data distribution, and ξ represents the stochasticity corresponding to sampling.

Assumptions. In this part, we introduce and briefly discuss the assumptions used in the analysis. First, we introduce the convexity assumption.

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

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

We next formalize the notion of optimality – we impose conditions to ensure that the objective value is well defined and bounded.

Assumption 2 (Optimal point)**.** The function f has an optimal point, i.e.

$$f(x^*) = \min_{x\in\mathbb{R}^d} f(x) > -\infty.$$

To control the variability of gradients and enable quantitative convergence guarantees, we additionally impose a smoothness condition on the objective function.

Assumption 3 (L-smoothness)**.** The function f is L-smooth on $\mathbb{R}^d$ with respect to the norm $\|\cdot\|$, i.e., it is differentiable and there exists $L > 0$ such that for any $x, y \in \mathbb{R}^d$ the following inequality holds:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

We assume that the gradient of f is accessible only through a noisy oracle $\nabla f(x, \xi)$. Further, we make assumptions on the noise variables.

Assumption 4 (Uniform geometric mixing)**.** The sequence $(\xi_t)_{t\geq 0}$ is a time-homogeneous Markov chain on $(Z, \mathcal{Z})$ with transition kernel Q and a unique invariant probability distribution π. The initial state need not be sampled from π. There exists a fixed integer $\tau_{\mathrm{mix}} \geq 1$ such that for every integer $t \geq 0$,

$$\sup_{z\in Z} \|Q^t(z, \cdot) - \pi\|_{\mathrm{TV}} \leq 4^{-\lfloor t/\tau_{\mathrm{mix}}\rfloor}.$$

Here $\|\mu - \nu\|_{\mathrm{TV}} := \sup_{A\in\mathcal{Z}} |\mu(A) - \nu(A)|$.

Assumption 4 is classical in the literature on optimization methods with Markovian noise (Dorfman and Levy, 2022; Doan et al., 2020a). We introduce this assumption similarly to Beznosikov et al. (2023).

The uniform geometric mixing assumption can potentially be relaxed. At the level of the Batch Lemma (see Lemma 3), summable dependence coefficients preserve the $\mathcal{O}(1/N)$ rate. For polynomial

decay of order t^{-p}, the batch bound becomes $\mathcal{O}(\log N/N)$ for $p = 1$ and $\mathcal{O}(N^{-p})$ for $0 < p < 1$, under suitable moment conditions. Non-uniform geometric mixing would additionally require control of the initial distribution through Lyapunov-type conditions. However, we leave such extensions for future work.

Given this dependence structure, we assume that the stochastic gradients are unbiased estimators of the true gradient under the invariant distribution.

Assumption 5. For all $x \in \mathbb{R}^d$ it holds that

$$\mathbb{E}_\pi[\nabla f(x, \xi)] = \nabla f(x).$$

Crucially, this assumption does not imply that $\mathbb{E}_\xi[\nabla f(x, \xi)] = \nabla f(x)$. In the Markovian setting, the stochastic gradient is generally biased conditionally, which is the primary difficulty addressed in this work.

Moreover, we formulate the assumption on the bounded stochasticity.

Assumption 6 (Uniformly bounded noise)**.** For all $\xi \in \mathcal{Z}$ and $x \in \mathbb{R}^d$ it holds that

$$\|\nabla f(x, \xi) - \nabla f(x)\|^2 \leq \sigma^2.$$

In the Markovian noise setting, existing analyses typically require a uniform bound on the noise, unlike the i.i.d. case (Agarwal et al., 2011; Bach and Perchet, 2016) where a bound on the expectation suffices. This limitation appears in numerous works (Beznosikov et al., 2023; Doan et al., 2020a; Duchi et al., 2012; Doan et al., 2020b), and removing this assumption was an open challenge. In this paper, we drop this assumption and obtain high-probability convergence under bounded variance assumptions, which are formulated below.

Assumption 7 (Conditional bounded second moment)**.** Let $(\xi_t)_{t\geq 0}$ be the Markov chain from Assumption 4 with its given initial distribution. For every $t \geq 0$, almost surely, simultaneously for all $x \in \mathbb{R}^d$, it holds that

$$\mathbb{E}\left[\|\nabla f(x, \xi_{t+1}) - \nabla f(x)\|^2 \mid \xi_t\right] \leq \sigma^2.$$

It is worth noting that when we consider finite sequences generated by the Markov chain, we impose a moment bound that holds at every transition along the considered Markov trajectory. Finally, we state a corresponding variance bound under the invariant distribution.

Assumption 8. For all $x \in \mathbb{R}^d$ the following inequality holds:

$$\mathbb{E}_\pi\|\nabla f(x, \xi) - \nabla f(x)\|^2 \leq \sigma^2.$$

It is worth noting that Assumptions 4 and 7 together imply Assumption 8; however, we will consider Assumption 8 separately for simplicity. Since we first develop the SGD-based algorithm convergence analysis under Assumption 6, it is important to clarify the following. This assumption immediately implies Assumptions 7 and 8, but the converse does not hold. Therefore, Assumption 6 is strictly stronger than the bounded-variance assumptions with respect to both the current and the invariant distributions. Moreover, if the stochasticity satisfies Assumption 6, then it is σ-sub-Gaussian. Indeed, the following result can be established straightforwardly:

$$\|\nabla f(x, \xi) - \nabla f(x)\|^2 \leq \sigma^2 \Leftrightarrow \exp\left(\frac{\|\nabla f(x, \xi) - \nabla f(x)\|^2}{\sigma^2}\right) \leq \exp(1).$$

It is enough to take the expectation of RHS. Later, we show that this observation enables us to adapt existing techniques to the Markovian stochastic setting with Assumption 6.

Example (unbounded finite-variance Markovian oracle). Let S_t be an irreducible aperiodic finite-state Markov chain and let $\eta_t \mid S_t = s \sim \mathcal{N}(0, \Sigma_s)$, conditionally independent across time, with $\sup_s \operatorname{tr}(\Sigma_s) < \infty$. Set $g(x, S_t, \eta_t) = \nabla f(x) + a_{S_t} + \eta_t$, where the vectors a_s are bounded and satisfy $\sum_s \pi_S(s) a_s = 0$. The joint process (S_t, η_t) mixes geometrically and has uniformly bounded time-marginal and invariant second moments, so Assumptions 7–8 hold (with a suitable σ), including after conditioning on any starting state. Fresh centered emissions with uniformly bounded second moments may also be used instead of Gaussian emissions, so the same construction allows finite-variance heavy tails. However, Assumption 6 fails whenever some Σ_s is nonzero, because Gaussian noise has unbounded support.

4 Main Results

We study two methods: MLMC-SGD (Algorithm 1) and accelerated MLMC-SGD with clipping (Algorithm 2). Both algorithms employ a randomized multilevel batching construction related to those in Dorfman and Levy (2022); Beznosikov et al. (2023).

MLMC-SGD for non-convex case under bounded noise: theoretical guarantees and key ideas.

As stated earlier, in the non-convex setting we consider Assumption 6, which imposes bounded stochasticity. To discuss the resulting bound, we first state the corresponding convergence theorem.

Algorithm 1 MLMC-SGD

Parameters: stepsizes $\{\gamma_k\}, \forall k\ \gamma_k > 0$, number of iterations K, batch sizes B_k, sampling thresholds M_k, $T_0 = 0$.
Initialization: choose $x^0 \in \mathbb{R}^d$
for $k = 0, 1, 2, \ldots, K-1$ **do**
  Sample $J_k \sim \text{Geom}\left(\frac{1}{2}\right)$
  $g^k = g_0^k + \begin{cases} 2^{J_k}(g_{J_k}^k - g_{J_k-1}^k), & \text{if } 2^{J_k} \leq M_k \\ 0, & \text{otherwise} \end{cases}$, where
  $g_j^k = 2^{-j} B_k^{-1} \sum_{i=1}^{2^j B_k} \nabla f(x^k, \xi_{T^k+i})$
  Update $x^{k+1} = x^k - \gamma_k g^k$
  Update $T^{k+1} = T^k + 2^{J_k} B_k$
end for

Theorem 1. *Suppose that Assumptions 2, 3, 4, 5, and 6 hold. Let $K \geq 2$, $\sigma > 0$, $\delta \in (0,1)$, and $\Delta_0 = f(x^0) - f^* > 0$. Let us denote $\ell_K = \log_2 K$, $q = \log(4/\delta)$, and $U_K = \ell_K + q^2$. Then, if we choose $B_k = B = 1$, $M_k = M = K$, and*

$$\gamma_k = \gamma = \min\left\{\frac{1}{L}, \sqrt{\frac{\Delta_0}{\tau_{\text{mix}}\sigma^2 L K\, U_K}}\right\},$$

after K iterations of Algorithm 1, its iterates satisfy

$$\frac{1}{K}\sum_{k=0}^{K-1} \|\nabla f(x^k)\|^2 = \mathcal{O}\left(\max\left\{\frac{L\Delta_0}{K},\ \sigma\sqrt{\frac{\tau_{\text{mix}} L \Delta_0 U_K}{K}},\ \frac{\tau_{\text{mix}}\sigma^2(1+\ell_K q)}{K}\right\}\right) \tag{1}$$

with probability at least $1-\delta$. Here, the implicit constant is absolute. If $\sigma = 0$, deterministic GD with step $1/L$ gives the corresponding $\mathcal{O}(L\Delta_0/K)$ bound; if $\Delta_0 = 0$, a global minimizer has already been reached.

For comparison with prior work, see Table 1. As a result, we formulate the iteration complexity for Algorithm 1.

Corollary 1 (Iteration complexity)**.** *Suppose that the conditions of Theorem 1 hold. Then the required number of iterations to reach ε-accuracy, i.e.,*

$$\frac{1}{K}\sum_{k=0}^{K-1} \|\nabla f(x^k)\|^2 \leq \varepsilon^2,$$

can be upper bounded with probability at least $1-\delta$ by

$$K = \widetilde{\mathcal{O}}\left(\max\left\{1, \frac{L\Delta_0}{\varepsilon^2}, \frac{\tau_{\text{mix}}\sigma^2}{\varepsilon^2}, \frac{\tau_{\text{mix}}\sigma^2 L\Delta_0}{\varepsilon^4}\right\}\right). \tag{2}$$

The notation $\widetilde{\mathcal{O}}$ hides logarithms in ε^{-1} and δ^{-1}.

Proof sketch. The proof closely follows the framework of Liu et al. (2023), which serves as the foundation of our results. However, in contrast to the standard SGD setting, the bounded stochasticity assumption does not imply that the complete MLMC gradient estimator is bounded by the same constant. Indeed, the randomized multilevel correction may amplify the noise, which prevents us from directly applying standard sub-Gaussian concentration arguments.

To address this issue, we first establish a conditional exponential-moment bound for the complete MLMC estimator. Using the uniform mixing assumption and a bounded-differences inequality for Markov chains (Paulin, 2015), we separately control the conditional bias, the second moment, and the exponential tail of the estimator. In particular, the squared bias is of order $\mathcal{O}(\tau_{\text{mix}}\sigma^2/(BM))$, while the second-moment and exponential scales are of order $\mathcal{O}(\tau_{\text{mix}}\sigma^2 \log_2 M/B)$ and $\mathcal{O}(\tau_{\text{mix}}\sigma^2 M/B)$, respectively.

Next, we introduce an additional lemma that allows us to incorporate these estimates into the exponential-moment analysis. Following Liu et al. (2023), we apply the descent lemma and construct an exponential supermartingale. Since the corresponding moment estimate requires control of the gradient norm, we introduce a stopping argument based on the objective function gap. Applying a maximal inequality shows that the stopping threshold is not crossed with high probability. Finally, combining these bounds with an appropriate choice of the stepsize and MLMC parameters yields Theorem 1. Unlike the standard sub-Gaussian analysis, the stochastic terms explicitly depend on the mixing time. Nevertheless, we obtain a dependence on $\sqrt{\tau_{\text{mix}}}$ in the stochastic convergence term and a linear dependence on τ_{mix} in the iteration complexity, up to logarithmic factors. For more details, see Appendices B.2 and C.

Accelerated MLMC-SGD for convex case under bounded variance: theoretical guarantees and key ideas.

In the convex setting, we move away from the bounded noise assumption and instead adopt the bounded variance condition (see Assumptions 7 and 8). It is worth noting that we assume the bounded variance with respect to both the current conditional distribution at each iteration and the invariant distribution π. Nevertheless, taken together, these assumptions are noticeably weaker compared to Assumption 6 and cover a broader class of stochastic optimization problems.

Now we propose the batch lemma under Assumptions 7 and 8.

Lemma 1 (Batch Lemma)**.** *Suppose that Assumptions 4, 5, 7 and 8 hold for the random sequence $\xi := \{\xi_1, \xi_2, \ldots, \xi_n\}$. Then, for any $N \geq 1$, $\tau_{\text{mix}} \geq 1$ and $x \in \mathbb{R}^d$, the following bound holds:*

$$\mathbb{E}_\xi \left[\left\| \frac{1}{N} \sum_{t=1}^{N} \big(\nabla f(x, \xi_t) - \nabla f(x) \big) \right\|^2 \right] \leq \frac{49 \tau_{\text{mix}} \sigma^2}{N}.$$

Comparing this lemma with prior works, we in fact obtain the same type of bound as Solodkin et al. (2024). However, our assumptions are *less restrictive*, as mentioned earlier, and they allow us to provide convergence guarantees under the condition of finite variance.

One of the key mechanisms for obtaining high-probability convergence guarantees in this regime is gradient clipping. The choice of this mechanism is motivated by the following observation: without gradient clipping, convergence would likely be significantly worse. Indeed, as shown in Sadiev et al. (2023), SGD without clipping exhibits poor high-probability convergence behavior under the bounded variance assumption even if the stochastic gradient estimator is unbiased at each iteration.

Now let us provide the convergence guarantees for Algorithm 2.

Theorem 2. *Suppose that Assumptions 1, 2, 3, 4, 5, 7, and 8 hold. Then, if we choose $M_k = M = \Theta(K)$, $B_k = B = 10 \log_2(M)$, $\zeta_k = \frac{k+2}{2}$,*

$$\gamma_k = \tilde{\Theta}\left(\min\left\{ \frac{1}{L}, \frac{R_0}{\sigma \zeta_k \sqrt{K} \sqrt{\tau_{\text{mix}}}} \right\} \right), \quad \lambda_k = \tilde{\Theta}\left(\frac{R_0}{\gamma_k \zeta_k} \right),$$

after K iterations of Algorithm 2 the following bound holds with probability at least $1 - \delta$:

$$f(x_f^K) - f(x^*) = \tilde{\mathcal{O}}\left(\max\left\{ \frac{L R_0^2}{K^2}, \frac{R_0 \sqrt{\tau_{\text{mix}}} \sigma}{\sqrt{K}} \right\} \right).$$

Algorithm 2 Accelerated MLMC-SGD with clipping

Parameters: stepsizes γ_k, momentum parameters ζ_k and $\theta_k = 1 - \frac{1}{\zeta_k}$, number of iterations K, batch sizes B_k, clipping threshold λ_k, sampling thresholds M_k, $T_0 = 0$.
Initialization: choose $x^0 = x_f^0$
for $k = 0, 1, 2, \ldots, K-1$ **do**
  $x_g^k = \theta_k x_f^k + (1-\theta_k)x^k$
  Sample $J_k \sim \text{Geom}\left(\frac{1}{2}\right)$
  $g^k = g_0^k + \begin{cases} 2^{J_k}(g_{J_k}^k - g_{J_k-1}^k), & \text{if } 2^{J_k} \leq M_k \\ 0, & \text{otherwise} \end{cases}$, where
  $g_j^k = 2^{-j} B_k^{-1} \sum_{i=1}^{2^j B_k} \nabla f(x_g^k, \xi_{T^k+i})$
  $\text{clip}_{\lambda_k}(g^k) = g^k \cdot \min\left(1, \frac{\lambda_k}{\|g^k\|}\right)$
  $x_f^{k+1} = x_g^k - \gamma_k \, \text{clip}_{\lambda_k}(g^k)$
  $x^{k+1} = \zeta_k x_f^{k+1} + (1-\zeta_k)x_f^k$
  $T^{k+1} = T^k + 2^{J_k} B_k$
end for

Table 2: Comparison of the results of Accelerated MLMC-SGD for convex problem.

Reference	Bounded variance	Markovian noise	High-probability	Rate
Beznosikov et al. (2023)	✓✗[(1)]	✓	✗	$\tilde{\mathcal{O}}\left(\frac{\tau_{\text{mix}}^2 L R_0^2}{K^2} + \frac{R_0\sqrt{\tau_{\text{mix}}}\sigma}{\sqrt{K}}\right)$
Solodkin et al. (2024)	✗	✓	✗	$\tilde{\mathcal{O}}\left(\frac{L R_0^2}{K^2} + \frac{R_0\sqrt{\tau_{\text{mix}}}\sigma}{\sqrt{K}}\right)$
Gorbunov et al. (2020)	✓	✗	✓	$\tilde{\mathcal{O}}\left(\frac{L R_0^2}{K^2} + \frac{R_0\sigma}{\sqrt{K}}\right)$
This work	✓	✓	✓	$\tilde{\mathcal{O}}\left(\frac{L R_0^2}{K^2} + \frac{R_0\sqrt{\tau_{\text{mix}}}\sigma}{\sqrt{K}}\right)$

[(1)] Beznosikov et al. (2023) use the assumption $\|\nabla f(x,\xi) - \nabla f(x)\|^2 \leq \sigma^2 + \beta^2 \|\nabla f(x)\|^2$, which can be reduced to Assumption 6 with $\beta = 0$ for the comparison.

As a consequence, we can formulate the corollary on the iteration complexity.

Corollary 2 (Iteration complexity)**.** *Suppose that conditions of Theorem 2 hold. Then, the number of iterations required to reach ε-accuracy, i.e. $f(x_f^K) - f^* \leq \varepsilon$, can be upper bounded with probability at least $1-\delta$ as follows:*

$$K = \tilde{\mathcal{O}}\left(\max\left\{\sqrt{\frac{L R_0^2}{\varepsilon}}, \frac{\tau_{\text{mix}}\sigma^2 R_0^2}{\varepsilon^2}\right\}\right).$$

For more details, see Appendix D.

Proof sketch and discussion. High-probability convergence analyses of gradient methods with clipping have been extensively studied in the literature. One of the main tools for establishing such guarantees is the control of bias and variance via the so-called Clipping Lemma (Gorbunov et al., 2020; Sadiev et al., 2023; Nguyen et al., 2023). However, prior works typically assume that the stochastic estimator is *unbiased*, which does not hold in the case of Markovian noise. To address this, we provide a modified Clipping Lemma that deals with the additional bias arising from the Markovian setting (see Lemma 6).

To establish this result, it is necessary to exploit the properties of randomized batch sizes. Unfortunately, previous works (Solodkin et al., 2024) provided this result only under the assumption of bounded noise, which we *do not require*. Therefore, to make the Clipping Lemma applicable in our setting, we introduce new supporting results, including a Batch Lemma (see Appendix B, Lemma 3) and an Estimator Lemma (see Appendix B, Lemma 4), recovering the same conclusions as Solodkin et al. (2024) under weaker Assumptions 7 and 8.

The proof of Theorem 2 follows the next steps. First, the Batch Lemma (Lemma 3) provides control over the second moment of Markovian batch averages, which is then used in the Estimator Lemma (Lemma 4) to bound the bias and second moment of the MLMC estimator. Next, the Clipping Lemma (Lemma 6) controls the additional error introduced by clipping. Finally, we combine these

estimates with the accelerated descent inequality and a high-probability potential argument to establish Theorem 2.

The proof of Lemma 6 is closely related to previous works. However, due to the bias of stochastic gradients, a natural additional term arises:

$$\left\|\mathbb{E}_k g^k - \nabla f(x_g^k)\right\|.$$

The clipping-induced part of the bias is controlled as in Sadiev et al. (2023), whereas the additional Markovian bias requires the Estimator Lemma. Since the batch size is determined by a geometric level with a truncated MLMC correction, we can conclude that

$$\left\|\mathbb{E}_k \operatorname{clip}_{\lambda_k}(g^k) - \nabla f(x_g^k)\right\| \sim \frac{\sigma}{\sqrt{M_k}} + \frac{\sigma^2}{\lambda_k}.$$

Therefore, the choice of a specific M_k at each iteration is crucial, as it balances the bias introduced by clipping and the bias arising from the Markovian stochasticity. Although choosing $M_k \sim \lambda_k^2/\sigma^2$ would naturally balance these terms, it would degrade oracle complexity. Nevertheless, it turns out to be sufficient to obtain the desired convergence guarantees with appropriate M_k.

We deliberately adopt the accelerated gradient scheme, as existing work on Markovian settings uses this approach, ensuring a fair comparison of our results.

Remark 1. In fact, an alternative accelerated scheme could also have been considered, such as SSTM (Gasnikov and Nesterov, 2016), for which clipping was also incorporated for high-probability convergence guarantees (Gorbunov et al., 2020).

Our convergence bound coincides with optimal rates of corresponding guarantees up to logarithmic factors:

• Deterministic rate: $\tilde{\mathcal{O}}\left(\frac{LR_0^2}{K^2}\right)$ (Nesterov, 2013).

• Stochastic rate: $\tilde{\mathcal{O}}\left(\frac{\sqrt{\tau_{\text{mix}}}\sigma R_0}{\sqrt{K}}\right)$, matching existing bounds (Gorbunov et al., 2020; Solodkin et al., 2024) and lower bounds (Duchi et al., 2012).

Oracle complexity under randomized batching

Although we have provided results on convergence and iteration complexity, a key question remains: what is the oracle complexity of these algorithms? On the one hand, the expected number of oracle calls per iteration can be expressed as $\mathcal{O}\left(B_k \log(M_k)\right)$. On the other hand, when considering almost sure bounds, no estimate better than $\mathcal{O}\left(B_k M_k\right)$ can be achieved. For this reason, we need to concentrate the total number of oracle calls in a tighter manner. Thus, we formulate the following result:

Theorem 3 (Concentration of oracle calls)**.** *Suppose that Algorithm 1 or 2 runs for K iterations with parameters B_k and $M_k \geq 2$, $k = 0, \ldots, K-1$. Then, the total number of oracle calls N_{calls} can be upper bounded with probability at least $1-\delta$ as follows:*

$$N_{calls} \leq B\left(4K\ln(M) + \frac{4}{3}M\ln\left(\frac{1}{\delta}\right) + 2\sqrt{KM\ln\left(\frac{1}{\delta}\right)}\right)$$

with $B = \max_k B_k$ and $M = \max_k M_k$.

The key idea is that geometric level with a truncated MLMC correction with parameters M_k bounds the number of oracle calls per iteration, allowing concentration of total gradient computations around its expectation via Bernstein's inequality, yielding Theorem 3 (see Appendix E).

We are now ready to present oracle complexity results based on the convergence guarantees for Algorithms 1 and 2.

Corollary 3 (Bounded noise, non-convex case)**.** *Suppose that conditions of Theorem 1 hold. Then, the required number of oracle calls for reaching ε-accuracy, i.e., $\frac{1}{K}\sum_{k=0}^{K-1}\left\|\nabla f(x^k)\right\|^2 \leq \varepsilon^2$ with Algorithm 1, can be upper bounded with probability at least $1-2\delta$ as follows:*

$$N_{calls} = \tilde{\mathcal{O}}\left(\max\left\{\frac{L\Delta_0}{\varepsilon^2}, \frac{\tau_{\text{mix}}\sigma^2}{\varepsilon^2}, \frac{\tau_{\text{mix}}\sigma^2 L\Delta_0}{\varepsilon^4}\right\}\right).$$

Similarly, we formulate the following result.

Corollary 4 (Bounded variance, convex case)**.** *Suppose that conditions of Theorem 2 hold. Then, the required number of oracle calls to reach ε-accuracy, i.e., $f(x_f^K) - f^* \leq \varepsilon$ with Algorithm 2, can be upper bounded with probability at least $1 - 2\delta$ as follows:*

$$N_{calls} = \tilde{\mathcal{O}}\left(\max\left\{\sqrt{\frac{LR_0^2}{\varepsilon}}, \frac{\tau_{\text{mix}}\sigma^2 R_0^2}{\varepsilon^2}\right\}\right).$$

For both results, see Appendix E.

As a consequence, using Theorem 3, we obtain the following result: the bound depends only logarithmically in the high-probability regime. Moreover, to the best of our knowledge and in contrast to prior works, this is the first high-probability concentration of the number of gradient evaluations for the considered randomized MLMC batches.

5 Conclusion

In this paper, we provide high-probability convergence guarantees for gradient methods based on MLMC batching in the Markovian stochastic setting under various additional stochasticity assumptions. Moreover, we show that in several prior works the assumption of bounded noise can be relaxed to finite variance, while the results for the randomized batch-size estimator remain unchanged. We further demonstrate that employing this estimator deteriorates the oracle complexity *only logarithmically*, provided that the parameters are chosen appropriately.

Despite the results presented above, many important questions remain open. A particularly interesting challenge is to extend the proposed theory to the case of bounded α-moment with $\alpha < 2$, as this setting is the closest to our current assumptions. Moreover, our analysis is inherently limited, and extending it to more general scenarios and assumptions – such as affine noise or generalized smoothness – poses challenging yet important directions for future research. Finally, it is crucial to develop analogous analyses for distributed and federated optimization settings, which have become increasingly prominent in practice.

We hope that our results will stimulate further research along these directions and contribute to a deeper understanding of optimization theory.

Acknowledgements

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

References

Agarwal, A., Foster, D. P., Hsu, D. J., Kakade, S. M., and Rakhlin, A. (2011). Stochastic convex optimization with bandit feedback. *Advances in neural information processing systems*, 24.

Armacki, A., Yu, S., Bajovic, D., Jakovetic, D., and Kar, S. (2024). Large deviations and improved mean-squared error rates of nonlinear sgd: Heavy-tailed noise and power of symmetry. *arXiv preprint arXiv:2410.15637*.

Bach, F. and Perchet, V. (2016). Highly-smooth zero-th order online optimization. In *Conference on Learning Theory*, pages 257–283. PMLR.

Bennett, G. (1962). Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45.

Beznosikov, A., Samsonov, S., Sheshukova, M., Gasnikov, A., Naumov, A., and Moulines, E. (2023). First order methods with markovian noise: from acceleration to variational inequalities. *Advances in Neural Information Processing Systems*, 36:44820–44835.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chezhegov, S., Beznosikov, A., Horváth, S., and Gorbunov, E. (2025). Convergence of clipped-sgd for convex (l_0, l_1)-smooth optimization with heavy-tailed noise. *arXiv preprint arXiv:2505.20817*.

Chezhegov, S., Klyukin, Y., Semenov, A., Beznosikov, A., Gasnikov, A., Horváth, S., Takáč, M., and Gorbunov, E. (2024). Clipping improves adam-norm and adagrad-norm when the noise is heavy-tailed. *arXiv preprint arXiv:2406.04443*.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65.

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.

Davis, D., Drusvyatskiy, D., Xiao, L., and Zhang, J. (2021). From low probability to high confidence in stochastic convex optimization. *Journal of machine learning research*, 22(49):1–38.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Doan, T. T. (2022). Finite-time analysis of markov gradient descent. *IEEE Transactions on Automatic Control*, 68(4):2140–2153.

Doan, T. T., Nguyen, L. M., Pham, N. H., and Romberg, J. (2020a). Convergence rates of accelerated markov gradient descent with applications in reinforcement learning. *arXiv preprint arXiv:2002.02873*.

Doan, T. T., Nguyen, L. M., Pham, N. H., and Romberg, J. (2020b). Finite-time analysis of stochastic gradient descent under markov randomness. *arXiv preprint arXiv:2003.10973*.

Dorfman, R. and Levy, K. Y. (2022). Adapting to mixing time in stochastic optimization with markovian data. In *International Conference on Machine Learning*, pages 5429–5446. PMLR.

Duchi, J. C., Agarwal, A., Johansson, M., and Jordan, M. I. (2012). Ergodic mirror descent. *SIAM Journal on Optimization*, 22(4):1549–1578.

Durmus, A., Moulines, E., Naumov, A., Samsonov, S., and Wai, H.-T. (2021). On the stability of random matrix product with markovian noise: Application to linear stochastic approximation and td learning. In *Conference on Learning Theory*, pages 1711–1752. PMLR.

Dvurechensky, P. and Gasnikov, A. (2016). Stochastic intermediate gradient method for convex problems with stochastic inexact oracle. *Journal of Optimization Theory and Applications*, 171:121–145.

Dzhaparidze, K. and Van Zanten, J. (2001). On bernstein-type inequalities for martingales. *Stochastic processes and their applications*, 93(1):109–117.

Even, M. (2023). Stochastic gradient descent under markovian sampling schemes. In *International Conference on Machine Learning*, pages 9412–9439. PMLR.

Freedman, D. A. (1975). On tail probabilities for martingales. *the Annals of Probability*, pages 100–118.

Gaash, O., Levy, K. Y., and Carmon, Y. (2025). Convergence of clipped sgd on convex (l_0, l_1)-smooth functions. *arXiv preprint arXiv:2502.16492*.

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.

Gasnikov, A. and Nesterov, Y. (2016). Universal fast gradient method for stochastic composit optimization problems. *arXiv preprint arXiv:1604.05275*.

Ghadimi, S. and Lan, G. (2012). Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492.

Gidel, G., Berard, H., Vignoud, G., Vincent, P., and Lacoste-Julien, S. (2018). A variational inequality perspective on generative adversarial networks. *arXiv preprint arXiv:1802.10551*.

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

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.

Harvey, N. J., Liaw, C., and Randhawa, S. (2019). Simple and optimal high-probability bounds for strongly-convex stochastic gradient descent. *arXiv preprint arXiv:1909.00843*.

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M. M. A., Yang, Y., and Zhou, Y. (2017). Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*.

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

Kakade, S. M. and Tewari, A. (2008). On the generalization ability of online strongly convex programming algorithms. *Advances in neural information processing systems*, 21.

Kornilov, N., Shamir, O., Lobanov, A., Dvinskikh, D., Gasnikov, A., Shibaev, I., Gorbunov, E., and Horváth, S. (2023). Accelerated zeroth-order method for non-smooth stochastic convex optimization problem with infinite variance. *Advances in Neural Information Processing Systems*, 36:64083–64102.

Kornilov, N., Zmushko, P., Semenov, A., Gasnikov, A., and Beznosikov, A. (2025). Sign operator for coping with heavy-tailed noise: High probability convergence bounds with extensions to distributed optimization and comparison oracle. *arXiv preprint arXiv:2502.07923*.

Li, S. and Liu, Y. (2023). High probability analysis for non-convex stochastic optimization with clipping. *arXiv preprint arXiv:2307.13680*.

Liu, Z., Nguyen, T. D., Nguyen, T. H., Ene, A., and Nguyen, H. (2023). High probability convergence of stochastic gradient methods. In *International Conference on Machine Learning*, pages 21884–21914. PMLR.

Madden, L., Dall'Anese, E., and Becker, S. (2024). High probability convergence bounds for non-convex stochastic gradient descent with sub-weibull noise. *Journal of Machine Learning Research*, 25(241):1–36.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533.

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*, 80:1607–1627.

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. (2013). *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media.

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.

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.

Paulin, D. (2015). Concentration inequalities for markov chains by marton couplings and spectral methods.

Rajbhandari, S., Rasley, J., Ruwase, O., and He, Y. (2020). Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Rakhlin, A., Shamir, O., and Sridharan, K. (2011). Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*.

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 Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29563–29648. PMLR.

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

Solodkin, V., Veprikov, A., and Beznosikov, A. (2024). Methods for optimization problems with markovian stochasticity and non-euclidean geometry. *arXiv preprint arXiv:2408.01848*.

Sun, T., Li, D., and Wang, B. (2022). Adaptive random walk gradient descent for decentralized optimization. In *International conference on machine learning*, pages 20790–20809. PMLR.

Sun, T., Sun, Y., and Yin, W. (2018). On markov chain gradient descent. *Advances in neural information processing systems*, 31.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: an introduction*. Adaptive computation and machine learning series. The MIT Press, Cambridge, Massachusetts, second edition.

Sutton, R. S., Barto, A. G., et al. (1998). *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Wang, P., Lei, Y., Ying, Y., and Zhou, D.-X. (2022). Stability and generalization for markov chain stochastic gradient methods. *Advances in Neural Information Processing Systems*, 35:37735–37748.

Zekri, O., Odonnat, A., Benechehab, A., Bleistein, L., Boullé, N., and Redko, I. (2024). Large language models as markov chains. *arXiv preprint arXiv:2410.02724*.

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

Contents

1 Introduction **1**

2 Related Works **2**

3 Preliminaries **4**

4 Main Results **6**

5 Conclusion **10**

A Additional Notation **15**

B Technical lemmas **15**

B.1 Bias estimation lemmas 15

B.2 Non-convex case lemmas 21

B.3 Convex case lemmas 23

C MLMC-SGD with Bounded Noise, Non-Convex Case **24**

D Accelerated MLMC-SGD with Bounded Variance and Clipping, Convex Case **27**

D.1 Descent Lemma 27

D.2 Main theorem, convex case 30

E Oracle complexity **35**

A Additional Notation

Here we introduce additional notation and facts.

Table 3: Auxiliary notation used in the proofs.

Symbol	Definition
Ω_k	$\mathrm{clip}_{\lambda_k}(g^k) - \nabla f(x_g^k)$
Ω_k^u	$\mathrm{clip}_{\lambda_k}(g^k) - \mathbb{E}_k\,\mathrm{clip}_{\lambda_k}(g^k)$
Ω_k^b	$\mathbb{E}_k\,\mathrm{clip}_{\lambda_k}(g^k) - \nabla f(x_g^k)$
R_k	$\left\|x^k - x^*\right\|$

Moreover, our analysis involves sums of martingale-difference sequences, to which Bernstein's inequality can be applied (Bennett, 1962; Dzhaparidze and Van Zanten, 2001; Freedman, 1975). For completeness, we state the version used in this paper as Lemma 2 below.

Lemma 2 (Bernstein's inequality)**.** *Let the sequence of random variables $\{X_i\}_{i\geq 1}$ form a martingale difference sequence, i.e.*

$$\mathbb{E}[X_i \mid X_{i-1}, \ldots, X_1] = 0 \qquad \text{for all } i \geq 1.$$

Assume that conditional variances $\sigma_i^2 \overset{def}{=} \mathbb{E}[X_i^2 \mid X_{i-1}, \ldots, X_1]$ exist and are bounded, and also that there exists a 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 $n \geq 1$,

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

and

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

For clarity, we summarize the role of the technical lemmas and their assumptions in Table 4.

Table 4: Technical lemmas and their role in the convergence analysis.

Result	Assumptions	Role in the analysis
Lemma 2	Martingale conditions	Concentration inequalities
Lemma 3	4, 5, 7, 8	Markovian batch error
Lemma 4	4, 5, 7, 8	MLMC bias and second moment
Lemma 5	Moment condition	Exponential-moment control
Lemma 6	4, 5, 7, 8	Clipping bias and variance
Lemma 7	2, 3	Non-convex descent inequality
Lemmas 8, 9	1, 2, 3	Convex descent inequality

B Technical lemmas

B.1 Bias estimation lemmas

We begin with the Batch Lemma, which serves as the foundational tool in our analysis. Building upon it, we prove the Estimator Lemma (Lemma 4) below, which directly provides control over the bias induced by the Markovian stochasticity. This bias gives rise to additional terms in the main theorems, and their estimation is delegated to technical lemmas. Similar technical apparatus has been employed in prior work on convergence under Markovian sampling; see, e.g., Beznosikov et al. (2023); Solodkin et al. (2024).

Lemma 3 (Lemma 1). *Suppose that Assumptions 4, 5, 7 and 8 hold for the random sequence $\xi = \{\xi_1, \xi_2, \ldots, \xi_N\}$. Then, for any $N \geq 1$ and $\tau_{\mathrm{mix}} \geq 1$, the following bound holds:*

$$\mathbb{E}_\xi\left[\left\|\frac{1}{N}\sum_{k=1}^N \big(\nabla f(x, \xi_k) - \nabla f(x)\big)\right\|^2\right] \leq \frac{49\tau_{\mathrm{mix}}\sigma^2}{N}.$$

Proof. We begin by expanding the squared norm.

$$\mathbb{E}_\xi\left[\left\|\sum_{k=1}^{N}(\nabla f(x,\xi_k)-\nabla f(x))\right\|^2\right] = \mathbb{E}_\xi\left[\sum_{k=1}^{N}\left\|\nabla f(x,\xi_k)-\nabla f(x)\right\|^2\right] + 2\,\mathbb{E}_\xi\left[\sum_{1\le i<j\le N}\left\langle\nabla f(x,\xi_i)-\nabla f(x),\nabla f(x,\xi_j)-\nabla f(x)\right\rangle\right].$$

Next, we decompose the last term in the following way: in both parts of the scalar product, we add and subtract the expectation of the corresponding stochastic gradient with respect to the trajectory distribution to divide it into four parts. Hence, we get

$$\begin{aligned}
&\mathbb{E}_\xi\left[\left\|\sum_{k=1}^{N}(\nabla f(x,\xi_k)-\nabla f(x))\right\|^2\right] = \mathbb{E}_\xi\left[\sum_{k=1}^{N}\left\|\nabla f(x,\xi_k)-\nabla f(x)\right\|^2\right]\\
&+2\,\mathbb{E}_\xi\left[\sum_{1\le i<j\le N}\langle\nabla f(x,\xi_i)-\nabla f(x),\nabla f(x,\xi_j)-\nabla f(x)\rangle\right]\\
&=\mathbb{E}_\xi\left[\sum_{k=1}^{N}\left\|\nabla f(x,\xi_k)-\nabla f(x)\right\|^2\right]\\
&+2\,\mathbb{E}_\xi\left[\sum_{1\le i<j\le N}\langle\nabla f(x,\xi_i)-\mathbb{E}_\xi[\nabla f(x,\xi_i)],\nabla f(x,\xi_j)-\mathbb{E}_\xi[\nabla f(x,\xi_j)]\rangle\right]\\
&+2\,\mathbb{E}_\xi\left[\sum_{1\le i<j\le N}\langle\nabla f(x,\xi_i)-\mathbb{E}_\xi[\nabla f(x,\xi_i)],\mathbb{E}_\xi[\nabla f(x,\xi_j)]-\nabla f(x)\rangle\right] \qquad (3)\\
&+2\,\mathbb{E}_\xi\left[\sum_{1\le i<j\le N}\langle\mathbb{E}_\xi[\nabla f(x,\xi_i)]-\nabla f(x),\nabla f(x,\xi_j)-\mathbb{E}_\xi[\nabla f(x,\xi_j)]\rangle\right] \qquad (4)\\
&+2\,\mathbb{E}_\xi\left[\sum_{1\le i<j\le N}\langle\mathbb{E}_\xi[\nabla f(x,\xi_i)]-\nabla f(x),\mathbb{E}_\xi[\nabla f(x,\xi_j)]-\nabla f(x)\rangle\right].
\end{aligned}$$

Noting that the terms (3) and (4) are equal to zero, we derive the following:

$$\mathbb{E}_\xi\left[\left\|\sum_{k=1}^{N}(\nabla f(x,\xi_k)-\nabla f(x))\right\|^2\right] = \mathbb{E}_\xi\left[\sum_{k=1}^{N}\left\|(\nabla f(x,\xi_k)-\nabla f(x))\right\|^2\right] \qquad (5)$$

$$+2\,\mathbb{E}_\xi\left[\sum_{1\le i<j\le N}\langle\nabla f(x,\xi_i)-\mathbb{E}_\xi[\nabla f(x,\xi_i)],\nabla f(x,\xi_j)-\mathbb{E}_\xi[\nabla f(x,\xi_j)]\rangle\right] \qquad (6)$$

$$+2\,\mathbb{E}_\xi\left[\sum_{1\le i<j\le N}\langle\mathbb{E}_\xi[\nabla f(x,\xi_i)]-\nabla f(x),\mathbb{E}_\xi[\nabla f(x,\xi_j)]-\nabla f(x)\rangle\right]. \qquad (7)$$

Step 1: Bounding the variance term (5). By Assumption 7 and the tower property, we have

$$\mathbb{E}_\xi\left[\sum_{k=1}^{N}\left\|(\nabla f(x,\xi_k)-\nabla f(x))\right\|^2\right] \le \sigma^2 N.$$

Step 2: Bounding the bias term (7). Applying the Cauchy-Schwarz inequality yields

$$2\,\mathbb{E}_\xi\left[\sum_{1\le i<j\le N}\langle \mathbb{E}_\xi[\nabla f(x,\xi_i)]-\nabla f(x), \mathbb{E}_\xi[\nabla f(x,\xi_j)]-\nabla f(x)\rangle\right]$$

$$\le 2\,\mathbb{E}_\xi\left[\sum_{1\le i<j\le N}\left\|\mathbb{E}_\xi[\nabla f(x,\xi_i)]-\nabla f(x)\right\|\cdot\left\|\mathbb{E}_\xi[\nabla f(x,\xi_j)]-\nabla f(x)\right\|\right]. \tag{8}$$

Step 2.1: Estimating the conditional expectation. Now let us bound $\left\|\mathbb{E}_\xi\nabla f(x,\xi_j)-\nabla f(x)\right\|$. Since $i<j$, one can use the tower property to obtain the following:

$$\begin{aligned}\left\|\mathbb{E}_\xi[\nabla f(x,\xi_j)]-\nabla f(x)\right\| &= \left\|\mathbb{E}_\xi\left[\mathbb{E}_\xi\left(\nabla f(x,\xi_j)-\nabla f(x)\mid \xi_i\right)\right]\right\| \\ &\le \mathbb{E}_\xi\left\|\mathbb{E}_\xi\left(\nabla f(x,\xi_j)-\nabla f(x)\mid \xi_i\right)\right\|. \end{aligned}\tag{9}$$

Therefore, we need to analyze the conditional expectation $\left\|\mathbb{E}\left[\nabla f(x,\xi_j)-\nabla f(x)\mid \xi_i\right]\right\|$. To bound it, we use a standard technique based on the Radon-Nikodym derivative, which allows us to express the difference between the conditional distribution and the invariant distribution in terms of the total variation distance. We now introduce several definitions. Let $\mu_i(\cdot):=\mathbb{P}\left(\xi_j\in\cdot\mid\xi_i\right)$ and let $h\left(z\right):=\nabla f(x,z)-\nabla f(x)$. Now, using the Assumption 5, we write the corresponding integrals with respect to measures defined on the same probability space.

$$\left\|\mathbb{E}_\xi\left(\nabla f(x,\xi_j)-\nabla f(x)\mid\xi_i\right)\right\| = \left\|\int_Z h(z)d\mu_i(z)-\int_Z h(z)d\pi(z)\right\|. \tag{10}$$

Let $\psi_i:=\mu_i+\pi$, $p_i:=\frac{d\mu_i}{d\psi_i}$, $q_i:=\frac{d\pi}{d\psi_i}$. From this, we obtain that $d\mu_i(z)=p_i(z)d\psi_i(z)$ and $d\pi(z)=q_i(z)d\psi_i(z)$. Then we have:

$$\begin{aligned}\left\|\mathbb{E}_\xi\left(\nabla f(x,\xi_j)-\nabla f(x)\mid\xi_i\right)\right\| &= \left\|\int_Z h(z)(p_i(z)-q_i(z))d\psi_i(z)\right\| \\ &\le \int_Z \|h(z)\|\,|p_i(z)-q_i(z)|d\psi_i(z).\end{aligned}$$

Applying the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}\left\|\mathbb{E}_\xi\left(\nabla f(x,\xi_j)-\nabla f(x)\mid\xi_i\right)\right\| &\le \int_Z \|h(z)\|\sqrt{p_i(z)+q_i(z)}\cdot\frac{|p_i(z)-q_i(z)|}{\sqrt{p_i(z)+q_i(z)}}d\psi_i(z) \\ &\le \left(\int_Z \|h(z)\|^2\left(p_i(z)+q_i(z)\right)d\psi_i(z)\right)^{\frac12}\cdot\left(\int_Z\frac{\left(p_i(z)-q_i(z)\right)^2}{\left(p_i(z)+q_i(z)\right)}d\psi_i(z)\right)^{\frac12}.\end{aligned}$$

Since $p_i+q_i=1$, $\left(\int_Z \|h(z)\|^2\left(p_i(z)+q_i(z)\right)d\psi_i(z)\right)^{\frac12}$ is equal to $\left(\int_Z \|h(z)\|^2 d\psi_i(z)\right)^{\frac12}$. Applying Assumptions 8 and 7 and taking additional expectation, we obtain the bound

$$\mathbb{E}_\xi\left(\int_Z \|h(z)\|^2\left(p_i(z)+q_i(z)\right)d\psi_i(z)\right)^{\frac12} = \mathbb{E}_\xi\left(\mathbb{E}_\xi\left[\|h(\xi_j)\|^2\mid\xi_i\right]+\mathbb{E}_\pi\left[\|h(\xi_j)\|^2\right]\right)^{\frac12}\le\sqrt{2}\sigma.$$

To justify the final step, we note that the function $\sqrt{\cdot}$ is concave, and Jensen's inequality gives

$$\begin{aligned}\mathbb{E}_\xi\left(\mathbb{E}_\xi\left[\|h(\xi_j)\|^2\mid\xi_i\right]+\mathbb{E}_\pi\left[\|h(\xi_j)\|^2\right]\right)^{\frac12} &\le \left(\mathbb{E}_\xi\left[\mathbb{E}_\xi\left[\|h(\xi_j)\|^2\mid\xi_i\right]\right]+\mathbb{E}_\pi\left[\|h(\xi_j)\|^2\right]\right)^{\frac12} \\ &= \left(\mathbb{E}_\xi\left[\|h(\xi_j)\|^2\right]+\mathbb{E}_\pi\left[\|h(\xi_j)\|^2\right]\right)^{\frac12}.\end{aligned}$$

Concerning the second factor $\left(\int_Z \frac{(p_i(z)-q_i(z))^2}{p_i(z)+q_i(z)} d\psi_i(z)\right)^{\frac{1}{2}}$, we note that $\int_Z \frac{(p_i-q_i)^2}{p_i+q_i} d\psi_i(z) \leq \int_Z |p_i - q_i| d\psi_i(z)$, since $|p_i - q_i| \leq p_i + q_i \leq 1$. By the definition of the total variation distance, it follows that

$$\left(\int_Z \frac{(p_i(z)-q_i(z))^2}{p_i(z)+q_i(z)} d\psi_i(z)\right)^{\frac{1}{2}} \leq \sqrt{2}\, \|\mu_i - \pi\|_{TV}^{\frac{1}{2}}.$$

The final step is to apply Assumption 4. Set $x = (j-i)/\tau_{\text{mix}}$. Since $\lfloor x \rfloor \geq x - 1$,

$$\|\mu_i - \pi\|_{TV} \leq \left(\frac{1}{4}\right)^{\lfloor x \rfloor} \leq \left(\frac{1}{4}\right)^{x-1} = 4\left(\frac{1}{4}\right)^{x},$$
$$\|\mu_i - \pi\|_{TV}^{\frac{1}{2}} \leq 2 \cdot \left(\frac{1}{2}\right)^{\frac{j-i}{\tau_{\text{mix}}}}.$$

As a result, we get the following:

$$\mathbb{E}_\xi \|\mathbb{E}_\xi\left(\nabla f(x,\xi_j) - \nabla f(x) \mid \xi_i\right)\| \leq 4\sigma \left(\frac{1}{2}\right)^{\frac{j-i}{\tau_{\text{mix}}}}. \tag{11}$$

Consequently, returning to the term in (9), we obtain the following bound:

$$\|\mathbb{E}_\xi[\nabla f(x,\xi_j)] - \nabla f(x)\| \leq 4\sigma \left(\frac{1}{2}\right)^{\frac{j-i}{\tau_{\text{mix}}}}. \tag{12}$$

From Assumption 7, it directly follows that $\left\|\mathbb{E}_\xi[\nabla f(x,\xi_i)] - \nabla f(x)\right\| \leq \mathbb{E}_\xi\left\|\nabla f(x,\xi_i) - \nabla f(x)\right\| \leq \sigma$. Combining this result with the established bound (12), we can then derive an explicit upper bound for the term (8):

$$\begin{aligned} &2\,\mathbb{E}_\xi\left[\sum_{1\leq i<j\leq N} \left\|\mathbb{E}_\xi[\nabla f(x,\xi_i)] - \nabla f(x)\right\| \cdot \left\|\mathbb{E}_\xi[\nabla f(x,\xi_j)] - \nabla f(x)\right\|\right] \\ &\leq 2\,\mathbb{E}_\xi\left[\sum_{1\leq i<j\leq N} \left\|\mathbb{E}_\xi[\nabla f(x,\xi_i)] - \nabla f(x)\right\| \cdot 4\sigma\left(\frac{1}{2}\right)^{\frac{j-i}{\tau_{\text{mix}}}}\right] \leq 8\,\sigma^2 \sum_{1\leq i<j\leq N}\left(\frac{1}{2}\right)^{\frac{j-i}{\tau_{\text{mix}}}}. \end{aligned} \tag{13}$$

Step 2.2: Summing the geometric series. Let us define $r := \left(\frac{1}{2}\right)^{\frac{1}{\tau_{\text{mix}}}}$. Using this notation, we can rewrite and bound the sum:

$$\begin{aligned} \sum_{1\leq i<j\leq N}\left(\frac{1}{2}\right)^{\frac{j-i}{\tau_{\text{mix}}}} &= \sum_{j=2}^{N}\sum_{i=1}^{j-1}\left(\frac{1}{2}\right)^{\frac{j-i}{\tau_{\text{mix}}}} = \sum_{j=2}^{N}\sum_{k=1}^{j-1}\left(\frac{1}{2}\right)^{\frac{k}{\tau_{\text{mix}}}} = \sum_{j=2}^{N}\sum_{k=1}^{j-1} r^k \\ &= \frac{(N-1)r}{1-r} - \frac{r^2(1-r^{N-1})}{(1-r)^2} \leq \frac{(N-1)r}{1-r} \leq (N-1)\left(\frac{\tau_{\text{mix}}}{\ln 2} + 1\right). \end{aligned}$$

The last inequality follows from the fact that $\frac{\left(\frac{1}{2}\right)^{1/x}}{1-\left(\frac{1}{2}\right)^{1/x}} \leq \frac{x}{\ln 2} + 1$. Furthermore, the following bound holds:

$$(N-1)\left(\frac{\tau_{\text{mix}}}{\ln 2} + 1\right) \leq (N-1)\left(2\,\tau_{\text{mix}} + 1\right) \leq 3\,N\,\tau_{\text{mix}}.$$

Step 2.3: Final bound for (7). According to the above, we get the following.

$$\begin{aligned} 2\,\mathbb{E}_\xi\left[\sum_{1\leq i<j\leq N}\langle \mathbb{E}_\xi[\nabla f(x,\xi_i)] - \nabla f(x), \mathbb{E}_\xi[\nabla f(x,\xi_j)] - \nabla f(x)\rangle\right] &\leq 8\sigma^2 \sum_{1\leq i<j\leq N}\left(\frac{1}{2}\right)^{\frac{j-i}{\tau_{\text{mix}}}} \\ &\leq 24 N \tau_{\text{mix}} \sigma^2. \end{aligned}$$

Step 3: Bounding the covariance term (6). We start with decomposing the second part of the scalar product:

$$
\begin{aligned}
&2\,\mathbb{E}_\xi\left[\sum_{1\le i<j\le N}\langle \nabla f(x,\xi_i)-\mathbb{E}_\xi[\nabla f(x,\xi_i)],\nabla f(x,\xi_j)-\mathbb{E}_\xi[\nabla f(x,\xi_j)]\rangle\right]\\
&=\ 2\,\mathbb{E}_\xi\left[\sum_{1\le i<j\le N}\langle \nabla f(x,\xi_i)-\mathbb{E}_\xi[\nabla f(x,\xi_i)],\nabla f(x,\xi_j)\rangle\right]\\
&\quad-2\,\mathbb{E}_\xi\left[\sum_{1\le i<j\le N}\langle \nabla f(x,\xi_i)-\mathbb{E}_\xi[\nabla f(x,\xi_i)],\,\mathbb{E}_\xi[\nabla f(x,\xi_j)]\rangle\right] \qquad (14)\\
&=\ 2\,\mathbb{E}_\xi\left[\sum_{1\le i<j\le N}\langle \nabla f(x,\xi_i)-\mathbb{E}_\xi[\nabla f(x,\xi_i)],\nabla f(x,\xi_j)\rangle\right]\\
&\quad-2\,\mathbb{E}_\xi\left[\sum_{1\le i<j\le N}\langle \nabla f(x,\xi_i)-\mathbb{E}_\xi[\nabla f(x,\xi_i)],\nabla f(x)\rangle\right], \qquad (15)
\end{aligned}
$$

where the last equality holds because in (14) and (15) we get that $\mathbb{E}_\xi\left[\nabla f(x,\xi_i)-\mathbb{E}_\xi[\nabla f(x,\xi_i)]\right]=0$. Therefore, we employ the tower property of conditional expectation:

$$
\begin{aligned}
&2\mathbb{E}_\xi\left[\sum_{1\le i<j\le N}\langle \nabla f(x,\xi_i)-\mathbb{E}_\xi\nabla f(x,\xi_i),\nabla f(x,\xi_j)-\mathbb{E}_\xi\nabla f(x,\xi_j)\rangle\right]\\
&\le 2\mathbb{E}_\xi\left[\sum_{1\le i<j\le N}\langle \nabla f(x,\xi_i)-\mathbb{E}_\xi\nabla f(x,\xi_i),\nabla f(x,\xi_j)-\nabla f(x)\rangle\right]\\
&=2\mathbb{E}_\xi\left[\sum_{1\le i<j\le N}\mathbb{E}_\xi\Big(\langle \nabla f(x,\xi_i)-\mathbb{E}_\xi\nabla f(x,\xi_i),\nabla f(x,\xi_j)-\nabla f(x)\rangle\mid\xi_i\Big)\right]\\
&=2\mathbb{E}_\xi\left[\sum_{1\le i<j\le N}\langle \nabla f(x,\xi_i)-\mathbb{E}_\xi\nabla f(x,\xi_i),\,\mathbb{E}_\xi\left(\nabla f(x,\xi_j)-\nabla f(x)\mid\xi_i\right)\rangle\right]\\
&\le 2\mathbb{E}_\xi\left[\sum_{1\le i<j\le N}\left\|\nabla f(x,\xi_i)-\mathbb{E}_\xi\nabla f(x,\xi_i)\right\|\cdot\left\|\mathbb{E}_\xi\left(\nabla f(x,\xi_j)-\nabla f(x)\mid\xi_i\right)\right\|\right],\\
&\le 2\left[\sum_{1\le i<j\le N}\left(\mathbb{E}_\xi\left[\left\|\nabla f(x,\xi_i)-\mathbb{E}_\xi\nabla f(x,\xi_i)\right\|^2\right]\right)^{\frac12}\cdot\left(\mathbb{E}_\xi\left[\left\|\mathbb{E}_\xi\left(\nabla f(x,\xi_j)-\nabla f(x)\mid\xi_i\right)\right\|^2\right]\right)^{\frac12}\right],
\end{aligned}
$$

where in the last line we apply Cauchy-Schwarz inequality.

The same decomposition as in (10), together with $\|\mu_i-\pi\|_{TV}^{1/2}\le 2\cdot(1/2)^{(j-i)/\tau_{\text{mix}}}$, yields

$$
\left\|\mathbb{E}_\xi\big(\nabla f(x,\xi_j)-\nabla f(x)\mid\xi_i\big)\right\|\le\left(\mathbb{E}_\xi\left[\|h(\xi_j)\|^2\mid\xi_i\right]+\mathbb{E}_\pi\left[\|h(\xi_j)\|^2\right]\right)^{\frac12}\cdot 2\sqrt{2}\left(\frac12\right)^{\frac{j-i}{\tau_{\text{mix}}}}.
$$

Squaring and taking the expectation gives

$$
\mathbb{E}_\xi\left[\left\|\mathbb{E}_\xi\big(\nabla f(x,\xi_j)-\nabla f(x)\mid\xi_i\big)\right\|^2\right]\le 16\sigma^2\left(\frac12\right)^{\frac{2(j-i)}{\tau_{\text{mix}}}},
$$

and therefore

$$
\left(\mathbb{E}_\xi\left[\left\|\mathbb{E}_\xi\big(\nabla f(x,\xi_j)-\nabla f(x)\mid\xi_i\big)\right\|^2\right]\right)^{\frac12}\le 4\sigma\left(\frac12\right)^{\frac{j-i}{\tau_{\text{mix}}}}.
$$

Since $\left(\mathbb{E}_\xi\left[\|\nabla f(x,\xi_i)-\mathbb{E}_\xi\nabla f(x,\xi_i)\|^2\right]\right)^{1/2}\leq\sigma$, we obtain

$$2\,\mathbb{E}_\xi\left[\sum_{1\leq i<j\leq N}\langle\nabla f(x,\xi_i)-\mathbb{E}_\xi\nabla f(x,\xi_i),\nabla f(x,\xi_j)-\mathbb{E}_\xi\nabla f(x,\xi_j)\rangle\right]$$
$$\leq 8\sigma^2\sum_{1\leq i<j\leq N}\left(\frac{1}{2}\right)^{\frac{j-i}{\tau_{\text{mix}}}}\leq 24N\tau_{\text{mix}}\sigma^2.$$

Combining (5), (6) and (7),

$$\mathbb{E}_\xi\left[\left\|\sum_{k=1}^N\big(\nabla f(x,\xi_k)-\nabla f(x)\big)\right\|^2\right]\leq\sigma^2N+24N\tau_{\text{mix}}\sigma^2+24N\tau_{\text{mix}}\sigma^2$$
$$=N\sigma^2\big(1+48\tau_{\text{mix}}\big)\leq 49N\tau_{\text{mix}}\sigma^2.$$

Dividing by N^2 completes the proof. □

Now we are ready to formulate the Estimator Lemma. This lemma is similar to the one used in Solodkin et al. (2024). However, in contrast to the latter, we prove it under weaker assumptions and with a different formulation of Assumption 4. Nevertheless, the lemma remains valid in the setting of *light-tailed noise*, and we also use it in the bounded-noise non-convex analysis.

Lemma 4 (Estimator lemma)**.** *Suppose that Assumptions 4, 5, 7 and 8 hold for ξ^k, where ξ^k is the sequence comprising Markovian random variables obtained on iteration k, $J_k\sim$ Geom $\left(\frac{1}{2}\right)$ and*

$$g^k=g_0^k+\begin{cases}2^{J_k}(g^k_{J_k}-g^k_{J_k-1}), & \text{if } 2^{J_k}\leq M_k\\ 0, & \text{otherwise}\end{cases},\quad \text{with } g_j^k=2^{-j}B_k^{-1}\sum_{i=1}^{2^jB_k}\nabla f(x_k,\xi_{T_k+i}).$$

Then, one can obtain the following bounds:

$$\mathbb{E}_k\|\nabla f(x_k)-g^k\|^2\leq\frac{490\tau_{\text{mix}}\,\sigma^2\log_2 M_k}{B_k};\tag{16}$$

$$\|\nabla f(x_k)-\mathbb{E}_k g^k\|^2\leq\frac{98\tau_{\text{mix}}\sigma^2}{B_kM_k}.\tag{17}$$

Proof. We start by proving (16). The initial part of the proof parallels that of Lemma 2 in Solodkin et al. (2024). To show that $\mathbb{E}_k[g^k]=\mathbb{E}_k\left[g^k_{[\log_2 M_k]}\right]$, we simply compute conditional expectation w.r.t. J_k:

$$\mathbb{E}_k[g^k]=\mathbb{E}_k\left[\mathbb{E}_{J_k}[g^k]\right]=\mathbb{E}_k[g_0^k]+\sum_{i=1}^{[\log_2 M_k]}\mathbb{P}\{J_k=i\}\cdot 2^i\mathbb{E}_k[g_i^k-g_{i-1}^k]=\mathbb{E}_k[g_0^k]$$
$$+\sum_{i=1}^{[\log_2 M_k]}\mathbb{E}_k[g_i^k-g_{i-1}^k]=\mathbb{E}_k\left[g^k_{[\log_2 M_k]}\right].$$

Next, we take the conditional expectation with respect to J_k and obtain

$$\mathbb{E}_k\|\nabla f(x^k)-g^k\|^2\leq 2\mathbb{E}_k\|\nabla f(x^k)-g_0^k\|^2+2\mathbb{E}_k\|g^k-g_0^k\|^2$$
$$=2\mathbb{E}_k\|\nabla f(x^k)-g_0^k\|^2+2\sum_{j=1}^{\lfloor\log_2 M_k\rfloor}2^j\mathbb{E}_k\|g_j^k-g_{j-1}^k\|^2.$$

To bound the last term, we use the fact that g_j^k is the average of two consecutive batches of size $2^{j-1}B_k$. Therefore,

$$g_j^k-g_{j-1}^k=\frac{1}{2}\left(\frac{1}{2^{j-1}B_k}\sum_{i=2^{j-1}B_k+1}^{2^jB_k}\nabla f(x^k,\xi_{T^k+i})-g_{j-1}^k\right).$$

Applying the inequality $\|a-b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ and Lemma 3 to both halves of the batch, we obtain

$$\mathbb{E}_k\|g_j^k - g_{j-1}^k\|^2 \leq \frac{1}{2}\mathbb{E}_k\left\|\frac{1}{2^{j-1}B_k}\sum_{i=2^{j-1}B_k+1}^{2^jB_k}\nabla f(x^k,\xi_{T^k+i}) - \nabla f(x^k)\right\|^2 + \frac{1}{2}\mathbb{E}_k\|g_{j-1}^k - \nabla f(x^k)\|^2$$
$$\leq \frac{49\tau_{\text{mix}}\sigma^2}{2^{j-1}B_k} = \frac{98\tau_{\text{mix}}\sigma^2}{2^jB_k}.$$

Substituting this bound into the previous inequality and applying Lemma 3 once again, we get

$$\mathbb{E}_k\|\nabla f(x^k) - g^k\|^2 \leq \frac{98\tau_{\text{mix}}\sigma^2}{B_k} + 2\sum_{j=1}^{\lfloor\log_2 M_k\rfloor} 2^j\frac{98\tau_{\text{mix}}\sigma^2}{2^jB_k} = \frac{(98+196\lfloor\log_2 M_k\rfloor)\tau_{\text{mix}}\sigma^2}{B_k}$$
$$\leq \frac{490\tau_{\text{mix}}\sigma^2\log_2 M_k}{B_k}.$$

To prove (17), it is enough to substitute $\mathbb{E}_k[g^k] = \mathbb{E}_k\left[g^k_{\lfloor\log_2 M_k\rfloor}\right]$, use $2^{\lfloor\log_2 M_k\rfloor} \geq \frac{M_k}{2}$, and apply Lemma 3. This finishes the proof. □

B.2 Non-convex case lemmas

Next, we formulate additional results which we use in the non-convex case.

Lemma 5 (Conditional exponential moment bound)**.** *Suppose that X is a random vector and, for some $H, V > 0$,*

$$\mathbb{E}_k\exp\left(\frac{\|X\|^2}{H}\right) \leq 1 + \frac{V}{H}. \tag{18}$$

Then, for all constants (with respect to X) $a \in \mathbb{R}^d$ and $b \geq 0$ such that

$$\sqrt{H}\|a\| \leq \frac{1}{4}, \qquad b^2 \leq \frac{1}{2H},$$

the following inequality holds for any $p > 0$:

$$\mathbb{E}_k\exp\left(\langle a, X\rangle + b^2\|X\|^2\right) \leq \exp\left[\left(2V + \frac{p}{2}\right)\|a\|^2 + b^2V + \frac{\|\mathbb{E}_kX\|^2}{2p}\right]. \tag{19}$$

Proof. Throughout this proof, all expectations are conditional on $\mathcal{F}_k$. First, for $\sqrt{H}\|a\| \leq 1/2$, using $e^u - 1 - u \leq u^2e^{|u|}/2$ and $\sqrt{H}\|a\|\,\|X\|/\sqrt{H} \leq (\|X\|^2/H + H\|a\|^2)/2$, we obtain

$$\mathbb{E}_k\left[e^{\langle a,X\rangle} - 1 - \langle a, X\rangle\right] \leq \frac{\|a\|^2}{2}\mathbb{E}_k\left[\|X\|^2e^{\|a\|\|X\|}\right] \leq \frac{e^{1/8}\|a\|^2H}{2}\mathbb{E}_k\left[e^{\|X\|^2/H} - 1\right] \leq V\|a\|^2,$$

where in the second inequality we used $ue^{u/2} \leq e^u - 1$ for $u \geq 0$. Since $\log t \leq t - 1$, this implies

$$\mathbb{E}_k\exp\left(\langle a, X - \mathbb{E}_kX\rangle\right) \leq \exp\left(V\|a\|^2\right), \qquad \sqrt{H}\|a\| \leq \frac{1}{2}. \tag{20}$$

Furthermore, conditional Jensen's inequality gives, for $0 \leq s \leq 1/H$,

$$\mathbb{E}_ke^{s\|X\|^2} \leq \left(1 + \frac{V}{H}\right)^{sH} \leq e^{sV}. \tag{21}$$

Applying Cauchy-Schwarz inequality, (20) with $2a$, and (21) with $s = 2b^2$, we have

$$\mathbb{E}_ke^{\langle a,X\rangle + b^2\|X\|^2} \leq e^{\langle a,\mathbb{E}_kX\rangle}\left(\mathbb{E}_ke^{2\langle a, X-\mathbb{E}_kX\rangle}\right)^{1/2}\left(\mathbb{E}_ke^{2b^2\|X\|^2}\right)^{1/2}$$
$$\leq \exp\left(\langle a, \mathbb{E}_kX\rangle + 2V\|a\|^2 + b^2V\right).$$

Finally, using Young's inequality $\langle a, \mathbb{E}_kX\rangle \leq p\|a\|^2/2 + \|\mathbb{E}_kX\|^2/(2p)$ completes the proof. □

Conditional exponential-moment bound under bounded noise. For Algorithm 1, additionally impose Assumption 6, and let $B_k \equiv B \geq 1$, $M_k \equiv M \geq 2$, and $\sigma > 0$. We assume that the geometric levels are mutually independent and independent of the Markov chain, with $\mathbb{P}(J_k = j) = 2^{-j}$ for $j \geq 1$. In this paragraph, $\mathcal{F}_k$ denotes the full history before iteration k, including the previously sampled levels and the Markov-chain states up to T^k, and $\mathbb{E}_k = \mathbb{E}[\,\cdot \mid \mathcal{F}_k]$. The random index T^k is a stopping time for the joint filtration generated by the chain and the levels sampled before the corresponding batches; the strong Markov property therefore applies after conditioning on $\mathcal{F}_k$. Put $\eta_k = g^k - \nabla f(x^k)$ and define

$$D_M = \frac{26\tau_{\mathrm{mix}}\sigma^2}{BM}, \qquad V_M = \frac{260\tau_{\mathrm{mix}}\sigma^2 \log_2 M}{B}, \qquad H_M = \frac{2^{15}\tau_{\mathrm{mix}}\sigma^2 M}{B}. \tag{22}$$

Then the following estimates hold almost surely:

$$\|\mathbb{E}_k \eta_k\|^2 \leq D_M, \qquad \mathbb{E}_k \exp\left(\frac{\|\eta_k\|^2}{H_M}\right) \leq 1 + \frac{V_M}{H_M}. \tag{23}$$

In particular, since $e^u \geq 1 + u$, this yields a second-moment bound with constant 260. The same bounds hold when conditioning on the coarser filtration generated by the iterates.

Proof of (23). Let $h_x(z) = \nabla f(x, z) - \nabla f(x)$. From the uniform TV mixing in Assumption 4 and Assumptions 5, 6,

$$\|h_x(z)\| \leq \sigma, \qquad \pi h_x = 0, \qquad \|Q^r h_x(z)\| \leq 2\sigma \|Q^r(z, \cdot) - \pi\|_{\mathrm{TV}} \leq 2\sigma 4^{-\lfloor r/\tau_{\mathrm{mix}} \rfloor}. \tag{24}$$

For deterministic $a_i \in \{-1, 1\}$ and an arbitrary initial distribution, expanding the squared norm and conditioning the later observation on the earlier one yields

$$\mathbb{E}\left\| \frac{1}{B} \sum_{i=1}^{N} a_i h_x(\xi_i) \right\|^2 \leq \frac{\sigma^2}{B^2} \left(N + 4 \sum_{r=1}^{N-1} (N - r) 4^{-\lfloor r/\tau_{\mathrm{mix}} \rfloor} \right) \leq \frac{6\tau_{\mathrm{mix}}\sigma^2 N}{B^2}. \tag{25}$$

Here we use $\sum_{r \geq 1} 4^{-\lfloor r/\tau_{\mathrm{mix}} \rfloor} = 4\tau_{\mathrm{mix}}/3 - 1$. All these estimates are uniform in x and the initial state.

Let $m = \lfloor \log_2 M \rfloor$. Conditional on $J_k = j \leq m$, cancellation of the first base batch gives

$$\eta_{k,j} = \frac{1}{B} \left(- \sum_{i=B+1}^{2^{j-1}B} h_{x^k}(\xi_{T^k+i}) + \sum_{i=2^{j-1}B+1}^{2^j B} h_{x^k}(\xi_{T^k+i}) \right). \tag{26}$$

For $j = 1$, the first sum is empty. The nonzero terms constitute a contiguous Markov segment of length $N_j = (2^j - 1)B$. Let

$$\alpha_j = \frac{2\tau_{\mathrm{mix}}\sigma^2(2^j - 1)}{B}, \qquad F_j = \|\eta_{k,j}\|.$$

By (25), $\mathbb{E}_k F_j^2 \leq 3\alpha_j$. Replacing one state in the segment changes F_j by at most $2\sigma/B$, so the sum of squared bounded-difference constants is at most $4\sigma^2 N_j / B^2$. The Markov-chain bounded-differences inequality (Paulin, 2015), implies, conditionally on $\mathcal{F}_k$,

$$\mathbb{P}(F_j - \mathbb{E}_k F_j \geq t \mid \mathcal{F}_k) \leq 2 \exp\left(-\frac{t^2}{18\alpha_j}\right), \qquad t \geq 0. \tag{27}$$

Indeed, Assumption 4 makes the usual mixing time at TV distance $1/4$ no larger than τ_{mix}, while the dependence parameter of the cited inequality is at most nine times this mixing time. This applies to arbitrary initial distributions.

Set $Z_j = (F_j - \mathbb{E}_k F_j)_+$. Since $\mathbb{E}_k F_j \leq \sqrt{3\alpha_j}$, we have $F_j^2 \leq 2Z_j^2 + 6\alpha_j$. Integrating (27),

$$\mathbb{E}_k \exp\left(\frac{F_j^2}{128\alpha_j}\right) \leq e^{6/128} \left(1 + \frac{72}{128 - 36}\right) < e. \tag{28}$$

On $J_k > m$, only the base batch is used; the same proof applies to $F_0 = \|g_0^k - \nabla f(x^k)\|$ with $\alpha_0 = 2\tau_{\mathrm{mix}}\sigma^2/B$.

Put $S_j^2 = 128\alpha_j$ for $j = 0, \dots, m$. Our choice of H_M ensures $r_j = S_j^2/H_M \le 1/128$. Conditional Jensen's inequality and (28) give

$$\mathbb{E}_k e^{F_j^2/H_M} \le e^{r_j} \le 1 + e^{1/128} r_j.$$

Averaging with respect to the independent geometric level and using

$$\sum_{j=1}^{m} 2^{-j} S_j^2 + 2^{-m} S_0^2 = \frac{256\tau_{\text{mix}}\sigma^2}{B}\left(m - 1 + 2^{1-m}\right) \le \frac{256\tau_{\text{mix}}\sigma^2 m}{B},$$

we obtain the second bound in (23), since $256e^{1/128} < 260$ and $m \le \log_2 M$.

Finally, telescoping the levels gives $\mathbb{E}_k g^k = \mathbb{E}_k g_m^k$. Applying (25) with positive weights,

$$\|\mathbb{E}_k \eta_k\|^2 \le \mathbb{E}_k \|g_m^k - \nabla f(x^k)\|^2 \le \frac{6\tau_{\text{mix}}\sigma^2}{2^m B} \le \frac{12\tau_{\text{mix}}\sigma^2}{BM} \le D_M.$$

The conditional bounds hold because the chain after the history has the same transition kernel with an arbitrary starting state, and all the estimates above are uniform in that state and in the fixed iterate x^k. □

B.3 Convex case lemmas

We now establish several useful bounds for the clipped gradient estimator. These bounds control the centered error and the bias of the clipped gradient, and are essential for applying Bernstein's inequality in the proof of Theorem 2.

Lemma 6 (Clipping lemma)**.** *Suppose that Assumptions 4, 5, 7, and 8 hold. Additionally, let us assume that* $\left\|\nabla f(x_g^k)\right\| \le \frac{\lambda_k}{2}$ *at the considered iteration* k *of Algorithm 2. Then, the following bounds hold:*

$$\left\| \operatorname{clip}_{\lambda_k}(g^k) - \mathbb{E}_k[\operatorname{clip}_{\lambda_k}(g^k)] \right\| \le 2\lambda_k; \tag{29}$$

$$\left\|\mathbb{E}_k[\operatorname{clip}_{\lambda_k}(g^k)] - \nabla f(x_g^k)\right\| \le \frac{10\sqrt{\tau_{\text{mix}}}\,\sigma}{\sqrt{M_k B_k}} + \frac{1960\tau_{\text{mix}}\,\sigma^2 \log_2 M_k}{B_k \lambda_k} \tag{30}$$

$$\mathbb{E}_k\left\|\operatorname{clip}_{\lambda_k}(g^k) - \nabla f(x_g^k)\right\|^2 \le \frac{4900\tau_{\text{mix}}\sigma^2 \log_2 M_k}{B_k} \tag{31}$$

$$\mathbb{E}_k\left\|\operatorname{clip}_{\lambda_k}(g^k) - \mathbb{E}_k \operatorname{clip}_{\lambda_k}(g^k)\right\|^2 \le \frac{4900\tau_{\text{mix}}\,\sigma^2 \log_2 M_k}{B_k} \tag{32}$$

Proof. The outline of the proof is similar to that of Lemma 5.1 in Sadiev et al. (2023). Furthermore, the proof of (29) proceeds identically, as the latter constitutes a general assertion. To prove (30), we introduce two random indicator variables:

$$\chi = \mathbb{I}_{\{g^k : \|g^k\| > \lambda_k\}} = \begin{cases} 1, & \text{if } \|g^k\| > \lambda_k, \\ 0, & \text{otherwise} \end{cases}, \tag{33}$$

$$\eta = \mathbb{I}_{\{g^k : \left\|g^k - \nabla f(x_g^k)\right\| > \frac{\lambda_k}{2}\}} = \begin{cases} 1, & \text{if } \left\|g^k - \nabla f(x_g^k)\right\| > \frac{\lambda_k}{2}, \\ 0, & \text{otherwise} \end{cases}. \tag{34}$$

Applying (33), (34), the triangle inequality, Lemma 4, and the Cauchy-Schwarz inequality, we obtain the following:

$$\begin{aligned}
\left\|\mathbb{E}_k \operatorname{clip}_{\lambda_k}(g^k) - \nabla f(x_g^k)\right\| &= \left\|\mathbb{E}_k\left[g^k + \chi\left(\frac{\lambda_k}{\|g^k\|} - 1\right) g^k\right] - \nabla f(x_g^k)\right\| \\
&\le \left\|\mathbb{E}_k g^k - \nabla f(x_g^k)\right\| + \left\|\mathbb{E}_k\left[\chi\left(\frac{\lambda_k}{\|g^k\|} - 1\right) g^k\right]\right\| \\
&\le \frac{10\sqrt{\tau_{\text{mix}}}\,\sigma}{\sqrt{M_k B_k}} + \left\|\mathbb{E}_k\left[\chi\left(\frac{\lambda_k}{\|g^k\|} - 1\right) g^k\right]\right\| \\
&\le \frac{10\sqrt{\tau_{\text{mix}}}\,\sigma}{\sqrt{M_k B_k}} + \mathbb{E}_k\left[\eta\left\|g^k - \nabla f(x_g^k)\right\| + \eta\|\nabla f(x_g^k)\|\right] \\
&\le \frac{10\sqrt{\tau_{\text{mix}}}\,\sigma}{\sqrt{M_k B_k}} + \left(\mathbb{E}_k\left\|g^k - \nabla f(x_g^k)\right\|^2\right)^{\frac{1}{2}} \left(\mathbb{E}_k \eta\right)^{\frac{1}{2}} + \frac{\lambda_k}{2}\mathbb{E}_k \eta.
\end{aligned} \tag{35}$$

Noting that $\mathbb{E}_k[\eta] = \mathbb{P}_k\big(\|g^k - \nabla f(x_g^k)\| > \frac{\lambda_k}{2}\big) \le \frac{4}{\lambda_k^2}\,\mathbb{E}_k\|g^k - \nabla f(x_g^k)\|^2 \le \frac{4}{\lambda_k^2}\cdot\frac{490\tau_{\text{mix}}\,\sigma^2\log_2 M_k}{B_k}$ due to Markov's inequality and Lemma 4, we can rewrite (35):

$$\begin{aligned}\big\|\mathbb{E}_k\ \text{clip}_{\lambda_k}(g^k) - \nabla f(x_g^k)\big\| &\le \frac{10\sqrt{\tau_{\text{mix}}}\,\sigma}{\sqrt{M_k B_k}} + \frac{2}{\lambda_k}\cdot\frac{490\tau_{\text{mix}}\,\sigma^2\log_2 M_k}{B_k}\\ &\quad + \frac{\lambda_k}{2}\cdot\frac{4}{\lambda_k^2}\cdot\frac{490\tau_{\text{mix}}\,\sigma^2\log_2 M_k}{B_k}\\ &= \frac{10\sqrt{\tau_{\text{mix}}}\,\sigma}{\sqrt{M_k B_k}} + \frac{4}{\lambda_k}\cdot\frac{490\tau_{\text{mix}}\,\sigma^2\log_2 M_k}{B_k}\\ &= \frac{10\sqrt{\tau_{\text{mix}}}\,\sigma}{\sqrt{M_k B_k}} + \frac{1960\tau_{\text{mix}}\,\sigma^2\log_2 M_k}{B_k\lambda_k}.\end{aligned}$$

To prove (31), we use the same random variables χ and η as defined above. Using $\big\|\text{clip}_{\lambda_k}(g^k) - \nabla f(x_g^k)\big\| \le \big\|\text{clip}_{\lambda_k}(g^k)\big\| + \big\|\nabla f(x_g^k)\big\| \le \frac{3\lambda_k}{2}$ and the triangle inequality, we obtain the following:

$$\begin{aligned}\mathbb{E}_k\big\|\text{clip}_{\lambda_k}(g^k) - \nabla f(x_g^k)\big\|^2 &= \mathbb{E}_k\Big[\chi\big\|\tfrac{\lambda_k}{\|g^k\|}g^k - \nabla f(x_g^k)\big\|^2 + (1-\chi)\big\|g^k - \nabla f(x_g^k)\big\|^2\Big]\\ &\le \mathbb{E}_k\Big[\Big(\tfrac{3\lambda_k}{2}\Big)^2\chi + (1-\chi)\big\|g^k - \nabla f(x_g^k)\big\|^2\Big]\\ &\le \mathbb{E}_k\eta\cdot\Big(\frac{3\lambda_k}{2}\Big)^2 + \frac{490\tau_{\text{mix}}\log_2 M_k\,\sigma^2}{B_k}\\ &\le \frac{4}{\lambda_k^2}\cdot\frac{490\tau_{\text{mix}}\sigma^2\log_2 M_k}{B_k}\cdot\frac{9\lambda_k^2}{4} + \frac{490\tau_{\text{mix}}\sigma^2\log_2 M_k}{B_k}\\ &= \frac{9\cdot 490\,\tau_{\text{mix}}\sigma^2\log_2 M_k}{B_k} + \frac{490\tau_{\text{mix}}\sigma^2\log_2 M_k}{B_k}\\ &= \frac{4900\tau_{\text{mix}}\sigma^2\log_2 M_k}{B_k}.\end{aligned}$$

Using variance decomposition and (31), we conclude that (32) holds. □

C MLMC-SGD with Bounded Noise, Non-Convex Case

In this section, we establish high-probability convergence guarantees for Algorithm 1 in the non-convex setting under Assumption 6. The proof retains the descent and exponential-moment framework of Liu et al. (2023). In contrast to the standard sub-Gaussian analysis, we control the full MLMC error through two distinct scales: its second moment and its exponential tail. A stopping argument allows us to apply Lemma 5 without assuming globally bounded gradients.

Throughout this section, $\mathcal{F}_k$ denotes the full history prior to iteration k, and $\mathbb{E}_k = \mathbb{E}[\cdot \mid \mathcal{F}_k]$. The geometric levels are independent of the Markov chain and of one another, with $\mathbb{P}(J_k = j) = 2^{-j}$ for $j \ge 1$. The proof is organized as follows:

1. Lemma 7 gives the descent inequality for one iteration.
2. Theorem 4 constructs a stopped exponential supermartingale using Lemma 5.
3. Corollary 5 provides a high-probability estimate and proves that the stopping threshold is not crossed.
4. Theorem 5 chooses the parameters and establishes the convergence rate.
5. Corollary 6 gives the iteration complexity.

Lemma 7 (Descent lemma)**.** *Suppose that Assumptions 2 and 3 hold, and Algorithm 1 is applied. Let $\eta_k = g^k - \nabla f(x^k)$. Then, for any $k \ge 0$,*

$$C_k := \gamma_k\left(1 - \frac{L\gamma_k}{2}\right)\|\nabla f(x^k)\|^2 + f(x^{k+1}) - f(x^k) \le (L\gamma_k^2 - \gamma_k)\langle\nabla f(x^k), \eta_k\rangle + \frac{L\gamma_k^2}{2}\|\eta_k\|^2.$$

Proof. The claim follows directly from the L-smoothness inequality after substituting $x^{k+1} = x^k - \gamma_k g^k$, and coincides with the standard descent calculation of Liu et al. (2023). □

Theorem 4 (Exponential-moment estimate)**.** *Suppose that Assumptions 2, 3, 4, 5, and 6 hold. Let $B_k \equiv B \geq 1$, $M_k \equiv M \geq 2$ and $0 < \gamma_k = \gamma \leq 1/L$. Put $\eta_k = g^k - \nabla f(x^k)$, and let D_M, V_M, H_M be given by (22). For $R \geq f(x^0) - f^*$, define*

$$T_R = \inf\{0 \leq k \leq K : f(x^k) - f^* > R\}, \quad w = \frac{1}{\gamma(8V_M + 4\sqrt{2LRH_M} + L\gamma H_M)}, \tag{36}$$

where $\inf \varnothing = \infty$, and set

$$v_k = w\gamma\left(\frac{1 - L\gamma}{4} + 2V_M w\gamma(1 - L\gamma)^2\right).$$

Define C_k as in Lemma 7 and

$$Z_k = \mathbb{I}_{\{k<T_R\}}\left(wC_k - v_k\|\nabla f(x^k)\|^2 - w\gamma D_M - \frac{wL\gamma^2 V_M}{2}\right), \qquad 0 \leq k < K,$$

$$S_k = \sum_{i=k}^{K-1} Z_i, \qquad 0 \leq k \leq K.$$

Then, for every $0 \leq k \leq K$,

$$\mathbb{E}_k[\exp(S_k)] \leq 1,$$

and $\exp(\sum_{i=0}^{n-1} Z_i)$, $0 \leq n \leq K$, is a nonnegative supermartingale starting from 1.

Proof. The proof follows the same exponential-moment induction as Theorem 4.3 of Liu et al. (2023). Since f is L-smooth and bounded below by f^*,

$$\|\nabla f(x)\|^2 \leq 2L(f(x) - f^*). \tag{37}$$

Thus, for $k < T_R$ we have $\|\nabla f(x^k)\|^2 \leq 2LR$. From Lemma 7,

$$C_k \leq -\gamma(1 - L\gamma)\langle \nabla f(x^k), \eta_k\rangle + \frac{L\gamma^2}{2}\|\eta_k\|^2. \tag{38}$$

We apply Lemma 5 with

$$a = w\gamma(L\gamma - 1)\nabla f(x^k), \qquad b^2 = \frac{wL\gamma^2}{2}.$$

By the choice of w and (37), on $\{k < T_R\}$,

$$\sqrt{H_M}\|a\| \leq w\gamma\sqrt{2LRH_M} \leq \frac{1}{4}, \qquad b^2 \leq \frac{1}{2H_M}, \qquad 2V_M w\gamma \leq \frac{1}{4}.$$

Using (23) and Lemma 5, we get, on this event,

$$\begin{aligned}
&\mathbb{E}_k \exp(wC_k) \\
&\leq \exp\left(-w\gamma(1 - L\gamma)\langle \nabla f(x^k), \mathbb{E}_k\eta_k\rangle + 2V_M w^2\gamma^2(1 - L\gamma)^2\|\nabla f(x^k)\|^2 + \frac{wL\gamma^2 V_M}{2}\right) \\
&\leq \exp\left(v_k\|\nabla f(x^k)\|^2 + w\gamma D_M + \frac{wL\gamma^2 V_M}{2}\right),
\end{aligned}$$

where the last step follows from

$$-(1-L\gamma)\langle \nabla f(x^k), \mathbb{E}_k\eta_k\rangle \leq \frac{1 - L\gamma}{4}\|\nabla f(x^k)\|^2 + (1-L\gamma)\|\mathbb{E}_k\eta_k\|^2 \leq \frac{1 - L\gamma}{4}\|\nabla f(x^k)\|^2 + D_M.$$

Consequently, $\mathbb{E}_k e^{Z_k} \leq 1$ on $\{k < T_R\}$. On $\{k \geq T_R\}$, $Z_k = 0$, so the same bound holds.

We prove the claim by backward induction. The case $k = K$ is immediate, since $S_K = 0$. Suppose $\mathbb{E}_{k+1} e^{S_{k+1}} \leq 1$. Then, since Z_k is $\mathcal{F}_{k+1}$-measurable,

$$\mathbb{E}_k e^{S_k} = \mathbb{E}_k\left[e^{Z_k}\mathbb{E}_{k+1} e^{S_{k+1}}\right] \leq \mathbb{E}_k e^{Z_k} \leq 1.$$

Finally, $\mathbb{E}_k e^{Z_k} \leq 1$ implies that $\exp(\sum_{i=0}^{n-1} Z_i)$ is a nonnegative supermartingale. This concludes the proof. □

Corollary 5 (High-probability estimate)**.** *Suppose that the conditions of Theorem 4 hold. Then, with probability at least* $1-\delta$*, for* $\delta \in (0,1)$*,*

$$\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla f(x^k)\|^2 \le \frac{4\Delta_0}{\gamma K} + 4D_M + 2L\gamma V_M + \frac{32V_M\log(4/\delta)}{K} + \frac{132L\gamma H_M\log^2(4/\delta)}{K}, \tag{39}$$

where $\Delta_0 = f(x^0) - f^*$*.*

Proof. Choose the deterministic stopping threshold

$$R = 2\left(\Delta_0 + \gamma K D_M + \frac{L\gamma^2 K V_M}{2} + 8\gamma V_M\log\frac{4}{\delta} + L\gamma^2 H_M\log\frac{4}{\delta}\right) + 64L\gamma^2 H_M\log^2\frac{4}{\delta}. \tag{40}$$

By Theorem 4 and the maximal inequality for a nonnegative supermartingale, with probability at least $1-\delta/4$ we have simultaneously for all $n \le K$

$$f(x^{n\wedge T_R}) - f^* + \frac{\gamma}{2}\sum_{k<n\wedge T_R}\|\nabla f(x^k)\|^2 \le \Delta_0 + \gamma K D_M + \frac{L\gamma^2 K V_M}{2} + \frac{\log(4/\delta)}{w}. \tag{41}$$

Here we used the elementary estimate, valid for $0 \le L\gamma \le 1$,

$$1 - \frac{L\gamma}{2} - \frac{v_k}{w\gamma} \ge 1 - \frac{L\gamma}{2} - \frac{1-L\gamma}{4} - \frac{(1-L\gamma)^2}{4} \ge \frac{1}{2}.$$

Substituting the choice of w into (41), we obtain

$$f(x^{n\wedge T_R}) - f^* + \frac{\gamma}{2}\sum_{k<n\wedge T_R}\|\nabla f(x^k)\|^2 \le \Delta_0 + \gamma K D_M + \frac{L\gamma^2 K V_M}{2} + 8\gamma V_M\log\frac{4}{\delta} + L\gamma^2 H_M\log\frac{4}{\delta} + 4\gamma\sqrt{2LRH_M}\log\frac{4}{\delta}.$$

By Young's inequality,

$$4\gamma\sqrt{2LRH_M}\log\frac{4}{\delta} \le \frac{R}{2} + 16L\gamma^2 H_M\log^2\frac{4}{\delta}.$$

Together with (40), this makes the last right-hand side strictly smaller than R. If $T_R \le K$, the left-hand side at $n = T_R$ is larger than R, a contradiction. Therefore $T_R > K$ on this event. Since $f(x^K) \ge f^*$, we obtain

$$\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla f(x^k)\|^2 \le \frac{2R}{\gamma K}.$$

Expanding R and using $\log(4/\delta) > 1$ yields (39). The claimed probability $1-\delta$ follows from $1-\delta/4 \ge 1-\delta$. □

Theorem 5 (Theorem 1: non-convex convergence)**.** *Suppose that Assumptions 2, 3, 4, 5, and 6 hold. Let* $K \ge 2$*,* $\sigma > 0$*,* $\Delta_0 = f(x^0) - f^* > 0$*,* $\delta \in (0,1)$*. Choose* $B = 1$*,* $M_k = M = K$ *and*

$$\gamma_k = \gamma = \min\left\{\frac{1}{L}, \sqrt{\frac{\Delta_0}{\tau_{\text{mix}}\sigma^2 LK(\log_2 K + \log^2(4/\delta))}}\right\}.$$

Then, after K *iterations of Algorithm 1, with probability at least* $1-\delta$*,*

$$\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla f(x^k)\|^2 \le \frac{4L\Delta_0}{K} + 4325900\,\sigma\sqrt{\frac{\tau_{\text{mix}}L\Delta_0(\log_2 K + \log^2(4/\delta))}{K}} + \frac{\tau_{\text{mix}}\sigma^2}{K}\left(104 + 8320\log_2 K\log\frac{4}{\delta}\right). \tag{42}$$

Proof. Substitute $B = 1, M = K$ into (22):

$$D_K = \frac{26\tau_{\text{mix}}\sigma^2}{K}, \qquad V_K = 260\tau_{\text{mix}}\sigma^2 \log_2 K, \qquad H_K = 2^{15}\tau_{\text{mix}}\sigma^2 K.$$

Applying Corollary 5, we obtain

$$\frac{1}{K}\sum_{k=0}^{K-1} \|\nabla f(x^k)\|^2 \le \frac{4\Delta_0}{\gamma K} + \frac{104\tau_{\text{mix}}\sigma^2}{K} + 520\tau_{\text{mix}} L\gamma\sigma^2 \log_2 K + \frac{8320\tau_{\text{mix}}\sigma^2 \log_2 K \log(4/\delta)}{K} + 4325376\tau_{\text{mix}} L\gamma\sigma^2 \log^2(4/\delta).$$

From the choice of γ,

$$\frac{\Delta_0}{\gamma K} \le \frac{L\Delta_0}{K} + \sigma\sqrt{\frac{\tau_{\text{mix}} L\Delta_0(\log_2 K + \log^2(4/\delta))}{K}},$$
$$\tau_{\text{mix}} L\gamma\sigma^2 \le \sigma\sqrt{\frac{\tau_{\text{mix}} L\Delta_0}{K(\log_2 K + \log^2(4/\delta))}}.$$

Since each of $\log_2 K$ and $\log^2(4/\delta)$ is bounded by their sum, the last three displays yield (42). This concludes the proof. □

Corollary 6 (Iteration complexity)**.** *Under the conditions of Theorem 5, the number of iterations sufficient for $K^{-1}\sum_{k=0}^{K-1} \|\nabla f(x^k)\|^2 \le \varepsilon^2$ with probability at least $1-\delta$ satisfies*

$$K = \widetilde{\mathcal{O}}\left(\max\left\{\frac{L\Delta_0}{\varepsilon^2}, \frac{\tau_{\text{mix}}\sigma^2}{\varepsilon^2}, \frac{\tau_{\text{mix}}\sigma^2 L\Delta_0}{\varepsilon^4}\right\}\right).$$

Proof. Put

$$A = \max\left\{2, \frac{L\Delta_0}{\varepsilon^2}, \frac{\tau_{\text{mix}}\sigma^2}{\varepsilon^2}, \frac{\tau_{\text{mix}}\sigma^2 L\Delta_0}{\varepsilon^4}\right\}.$$

Choose

$$K = \left\lceil CA\big(1 + \log(2+A) + \log(4/\delta)\big)^4 \right\rceil$$

with a sufficiently large universal constant C. Then $\log K = O(\log(2+A) + \log(4/\delta) + \log C)$, and inserting this choice into (42) shows that each term is at most $\varepsilon^2/3$ for all sufficiently large universal C. This proves the claim. □

D Accelerated MLMC-SGD with Bounded Variance and Clipping, Convex Case

We analyze accelerated MLMC-SGD with clipping. Here we use the *bounded variance* assumption, which is less restrictive.

D.1 Descent Lemma

First, we propose the auxiliary lemma for Algorithm 2. This is an important step that enables the key transition in the *descent lemma*. We state it separately in order to avoid overburdening the technical proof.

Lemma 8 (Auxiliary lemma)**.** *Suppose that Assumptions 1, 2 and 3 hold. Then, for any iteration of Algorithm 2 and for any $u \in \mathbb{R}^d$, the following inequality holds:*

$$f(x_f^{k+1}) \le f(u) - \langle \nabla f(x_g^k), u - x_g^k\rangle - \gamma_k \langle \nabla f(x_g^k), \text{clip}_{\lambda_k}(g^k)\rangle + \frac{L\gamma_k^2}{2}\|\text{clip}_{\lambda_k}(g^k)\|^2.$$

Proof. Considering the L-smoothness property:

$$f(x_f^{k+1}) \leq f(x_g^k) + \langle \nabla f(x_g^k), x_f^{k+1} - x_g^k \rangle + \frac{L}{2}\|x_f^{k+1} - x_g^k\|^2.$$

Applying line 8 of Algorithm 2, we obtain:

$$\begin{aligned} f(x_f^{k+1}) &\leq f(x_g^k) + \langle \nabla f(x_g^k), x_f^{k+1} - x_g^k \rangle + \frac{L}{2}\|x_f^{k+1} - x_g^k\|^2 \\ &= f(x_g^k) - \gamma_k \langle \nabla f(x_g^k), \operatorname{clip}_{\lambda_k}(g^k) \rangle + \frac{L\gamma_k^2}{2}\|\operatorname{clip}_{\lambda_k}(g^k)\|^2. \end{aligned}$$

By Assumption 1, we have:

$$\begin{aligned} f(x_f^{k+1}) &\leq f(x_g^k) - \gamma_k \langle \nabla f(x_g^k), \operatorname{clip}_{\lambda_k}(g^k) \rangle + \frac{L\gamma_k^2}{2}\|\operatorname{clip}_{\lambda_k}(g^k)\|^2 \\ &\leq f(u) - \langle \nabla f(x_g^k), u - x_g^k \rangle - \gamma_k \langle \nabla f(x_g^k), \operatorname{clip}_{\lambda_k}(g^k) \rangle + \frac{L\gamma_k^2}{2}\|\operatorname{clip}_{\lambda_k}(g^k)\|^2. \end{aligned}$$

This concludes the proof. □

This is a standard derivation for accelerated methods adapted to our setting with clipping. The key steps are to expand the squared distance and use the inequality from Lemma 8.

Lemma 9 (Descent lemma)**.** *Suppose that conditions of Lemma 8 hold. Then, if $\gamma_k \leq \frac{1}{2L}$, the sequence of γ_k is non-increasing and $\theta_k = 1 - \frac{1}{\zeta_k}$, where $\zeta_k = \frac{k+2}{2}$, the next inequality holds:*

$$\begin{aligned} \|x^{k+1} - x^*\|^2 + 2\gamma_k\zeta_k^2\left(f(x_f^{k+1}) - f(x^*)\right) + \frac{\gamma_k^2\zeta_k^2}{2}\|\nabla f(x_g^k)\|^2 \\ \leq \|x^k - x^*\|^2 + 2\gamma_k\zeta_k(\zeta_k - 1)\left(f(x_f^k) - f(x^*)\right) \\ + 2\gamma_k^3\zeta_k^2 L\langle \nabla f(x_g^k), \Omega_k \rangle - 2\gamma_k\zeta_k\langle \Omega_k, x^k - x^* \rangle \\ + 2\gamma_k^2\zeta_k^2\|\Omega_k\|^2, \end{aligned}$$

Moreover, if all stepsizes satisfy $\gamma_k \leqslant \frac{1}{2L}$, after K iterations of Algorithm 2, the following inequality holds:

$$\begin{aligned} 2\gamma_{K-1}\zeta_{K-1}^2\left(f(x_f^K) - f(x^*)\right) + \frac{\gamma_{K-1}^2\zeta_{K-1}^2}{2}\|\nabla f(x_g^{K-1})\|^2 \\ \leq \|x^0 - x^*\|^2 - \|x^K - x^*\|^2 \\ + \sum_{k=0}^{K-1} 2\gamma_k^2\zeta_k^2\|\Omega_k\|^2 + \sum_{k=0}^{K-1} 2L\gamma_k^3\zeta_k^2\langle \nabla f(x_g^k), \Omega_k \rangle \\ - \sum_{k=0}^{K-1} 2\gamma_k\zeta_k\langle \Omega_k, x^k - x^* \rangle. \end{aligned}$$

Proof. We start with the squared distance to the solution at the $(k+1)^{th}$ iteration, substituting the update rule from Algorithm 2:

$$\begin{aligned} \|x^{k+1} - x^*\|^2 &= \|\zeta_k x_f^{k+1} + (1 - \zeta_k)x_f^k - x^*\|^2 \\ &= \|\zeta_k(x_g^k - \gamma_k \operatorname{clip}_{\lambda_k}(g^k)) + (1 - \zeta_k)x_f^k - x^*\|^2 \\ &= \|\zeta_k x_g^k + (1 - \zeta_k)x_f^k - x^*\|^2 + \gamma_k^2\zeta_k^2\|\operatorname{clip}_{\lambda_k}(g^k)\|^2 \\ &\quad - 2\gamma_k\zeta_k\langle \operatorname{clip}_{\lambda_k}(g^k), \zeta_k x_g^k + (1 - \zeta_k)x_f^k - x^* \rangle \end{aligned} \tag{43}$$

It is worth noting that in the case $k = 0$ it is obvious due to the initialization. Otherwise, applying line 4 of Algorithm 2 yields:

$$\zeta_k x_g^k + (1-\zeta_k)x_f^k = \zeta_k x_g^k + \frac{1-\zeta_k}{\theta_k}\theta_k x_f^k = \zeta_k x_g^k + \frac{1-\zeta_k}{\theta_k}\big(x_g^k - (1-\theta_k)x^k\big)$$
$$= \Big(\zeta_k + \frac{1-\zeta_k}{\theta_k}\Big)x_g^k - \frac{(1-\theta_k)(1-\zeta_k)}{\theta_k}x^k.$$

Moreover,

$$\frac{(1-\theta_k)(1-\zeta_k)}{\theta_k} = \frac{1-\theta_k-\zeta_k+\theta_k\zeta_k}{\theta_k} = \frac{1-\zeta_k+\theta_k\zeta_k}{\theta_k} - 1.$$

As a consequence, we get

$$\zeta_k x_g^k + (1-\zeta_k)x_f^k = x^k - \frac{1-\zeta_k+\theta_k\zeta_k}{\theta_k}x^k + \frac{1-\zeta_k+\theta_k\zeta_k}{\theta_k}x_g^k$$
$$= x^k - \Big(\frac{1-\zeta_k+\theta_k\zeta_k}{\theta_k}\Big)(x^k - x_g^k). \tag{44}$$

Next, we substitute the choice of $\theta_k = 1 - 1/\zeta_k$ into $\frac{1-\zeta_k+\theta_k\zeta_k}{\theta_k}$ to obtain

$$\frac{1-\zeta_k+\theta_k\zeta_k}{\theta_k} = \zeta_k + \frac{1-\zeta_k}{\theta_k} = 0. \tag{45}$$

Therefore, applying (45) to (44) gives

$$\zeta_k x_g^k + (1-\zeta_k)x_f^k = x^k. \tag{46}$$

Thus, we use (46) to reformulate (43):

$$\|x^{k+1} - x^*\|^2 = \|\zeta_k x_g^k + (1-\zeta_k)x_f^k - x^*\|^2 + \gamma_k^2\zeta_k^2\|\operatorname{clip}_{\lambda_k}(g^k)\|^2$$
$$- 2\gamma_k\zeta_k\langle \operatorname{clip}_{\lambda_k}(g^k), \zeta_k x_g^k + (1-\zeta_k)x_f^k - x^*\rangle$$
$$= \|x^k - x^*\|^2 + \gamma_k^2\zeta_k^2\|\operatorname{clip}_{\lambda_k}(g^k)\|^2 - 2\gamma_k\zeta_k\langle \operatorname{clip}_{\lambda_k}(g^k), x^k - x^*\rangle \tag{47}$$

Applying Lemma 8 with $u = x_f^k$ and $u = x^*$, one can obtain the following inequalities:

$$f(x_f^{k+1}) \le f(x_f^k) - \langle \nabla f(x_g^k), x_f^k - x_g^k\rangle - \gamma_k\langle \nabla f(x_g^k), \operatorname{clip}_{\lambda_k}(g^k)\rangle + \frac{L\gamma_k^2}{2}\|\operatorname{clip}_{\lambda_k}(g^k)\|^2; \tag{48}$$

$$f(x_f^{k+1}) \le f(x^*) - \langle \nabla f(x_g^k), x^* - x_g^k\rangle - \gamma_k\langle \nabla f(x_g^k), \operatorname{clip}_{\lambda_k}(g^k)\rangle + \frac{L\gamma_k^2}{2}\|\operatorname{clip}_{\lambda_k}(g^k)\|^2. \tag{49}$$

Combining inequalities (47), (48), and (49) with multiplicative factors 1, $2\gamma_k\zeta_k(\zeta_k-1)$, and $2\gamma_k\zeta_k$, respectively, we get the following:

$$\|x^{k+1} - x^*\|^2 + 2\gamma_k\zeta_k^2 f(x_f^{k+1})$$
$$\le \|x^k - x^*\|^2 + \gamma_k^2\zeta_k^2\|\operatorname{clip}_{\lambda_k}(g^k)\|^2 - 2\gamma_k\zeta_k\langle \operatorname{clip}_{\lambda_k}(g^k), x^k - x^*\rangle$$
$$+ 2\gamma_k\zeta_k(\zeta_k-1)f(x_f^k) - 2\gamma_k\zeta_k(\zeta_k-1)\langle \nabla f(x_g^k), x_f^k - x_g^k\rangle$$
$$- 2\gamma_k^2\zeta_k(\zeta_k-1)\langle \nabla f(x_g^k), \operatorname{clip}_{\lambda_k}(g^k)\rangle + \gamma_k^3\zeta_k(\zeta_k-1)L\|\operatorname{clip}_{\lambda_k}(g^k)\|^2$$
$$+ 2\gamma_k\zeta_k f(x^*) - 2\gamma_k\zeta_k\langle \nabla f(x_g^k), x^* - x_g^k\rangle - 2\gamma_k^2\zeta_k\langle \nabla f(x_g^k), \operatorname{clip}_{\lambda_k}(g^k)\rangle$$
$$+ \gamma_k^3\zeta_k L\|\operatorname{clip}_{\lambda_k}(g^k)\|^2$$
$$= \|x^k - x^*\|^2 + 2\gamma_k\zeta_k(\zeta_k-1)f(x_f^k) + 2\gamma_k\zeta_k f(x^*)$$
$$+ \gamma_k^2\zeta_k^2(1+L\gamma_k)\|\operatorname{clip}_{\lambda_k}(g^k)\|^2 - 2\gamma_k\zeta_k\langle \operatorname{clip}_{\lambda_k}(g^k), x^k - x^*\rangle$$
$$- 2\gamma_k^2\zeta_k^2\langle \nabla f(x_g^k), \operatorname{clip}_{\lambda_k}(g^k)\rangle - 2\gamma_k\zeta_k\langle \nabla f(x_g^k), (\zeta_k-1)(x_f^k - x_g^k) + x^* - x_g^k\rangle$$
$$= \|x^k - x^*\|^2 + 2\gamma_k\zeta_k(\zeta_k-1)f(x_f^k) + 2\gamma_k\zeta_k f(x^*)$$
$$+ \gamma_k^2\zeta_k^2(1+L\gamma_k)\|\operatorname{clip}_{\lambda_k}(g^k)\|^2 - 2\gamma_k\zeta_k\langle \operatorname{clip}_{\lambda_k}(g^k) - \nabla f(x_g^k), x^k - x^*\rangle$$
$$- 2\gamma_k^2\zeta_k^2\langle \nabla f(x_g^k), \operatorname{clip}_{\lambda_k}(g^k)\rangle,$$

where in the second equation we apply (46): $(\zeta_k - 1)x_f^k - \zeta_k x_g^k = -x^k$. Next, we use the notation $\Omega_k = \text{clip}_{\lambda_k}(g^k) - \nabla f(x_g^k)$ to have

$$\begin{aligned}\|x^{k+1} - x^*\|^2 + 2\gamma_k\zeta_k^2 f(x_f^{k+1}) &\leq \|x^k - x^*\|^2 + 2\gamma_k\zeta_k(\zeta_k - 1)f(x_f^k) + 2\gamma_k\zeta_k f(x^*) \\ &+ \gamma_k^2\zeta_k^2(1 + L\gamma_k)\|\nabla f(x_g^k) + \Omega_k\|^2 - 2\gamma_k\zeta_k\langle\Omega_k, x^k - x^*\rangle \\ &- 2\gamma_k^2\zeta_k^2\langle\nabla f(x_g^k), \nabla f(x_g^k) + \Omega_k\rangle \\ &= \|x^k - x^*\|^2 + 2\gamma_k\zeta_k(\zeta_k - 1)f(x_f^k) + 2\gamma_k\zeta_k f(x^*) \\ &- \gamma_k^2\zeta_k^2(1 - L\gamma_k)\|\nabla f(x_g^k)\|^2 - 2\gamma_k\zeta_k\langle\Omega_k, x^k - x^*\rangle \\ &+ 2L\gamma_k^3\zeta_k^2\langle\nabla f(x_g^k), \Omega_k\rangle + 2\gamma_k^2\zeta_k^2\|\Omega_k\|^2,\end{aligned}$$

where we also use $\gamma_k \leq \frac{1}{L}$. Therefore, subtracting $2\gamma_k\zeta_k^2 f(x^*)$ gives the following:

$$\begin{aligned}\|x^{k+1} - x^*\|^2 + 2\gamma_k\zeta_k^2\left(f(x_f^{k+1}) - f(x^*)\right) &\leq \|x^k - x^*\|^2 + 2\gamma_k\zeta_k(\zeta_k - 1)\left(f(x_f^k) - f(x^*)\right) \\ &+ \gamma_k^2\zeta_k^2(1 + L\gamma_k)\|\nabla f(x_g^k) + \Omega_k\|^2 - 2\gamma_k\zeta_k\langle\Omega_k, x^k - x^*\rangle \\ &- 2\gamma_k^2\zeta_k^2\langle\nabla f(x_g^k), \nabla f(x_g^k) + \Omega_k\rangle \qquad (50) \\ &\leq \|x^k - x^*\|^2 + 2\gamma_k\zeta_k(\zeta_k - 1)\left(f(x_f^k) - f(x^*)\right) \\ &- \gamma_k^2\zeta_k^2(1 - L\gamma_k)\|\nabla f(x_g^k)\|^2 - 2\gamma_k\zeta_k\langle\Omega_k, x^k - x^*\rangle \\ &+ 2L\gamma_k^3\zeta_k^2\langle\nabla f(x_g^k), \Omega_k\rangle + 2\gamma_k^2\zeta_k^2\|\Omega_k\|^2. \qquad (51)\end{aligned}$$

Since the stepsize satisfies $\gamma_k \leq \frac{1}{2L}$, we have $(\gamma_k L - 1) \leq -\frac{1}{2}$. Therefore, the inequality (51) takes the final form:

$$\begin{aligned}\|x^{k+1} - x^*\|^2 + 2\gamma_k\zeta_k^2\left(f(x_f^{k+1}) - f(x^*)\right) + \frac{\gamma_k^2\zeta_k^2}{2}\|\nabla f(x_g^k)\|^2 &\leq \|x^k - x^*\|^2 + 2\gamma_k\zeta_k(\zeta_k - 1)\left(f(x_f^k) - f(x^*)\right) \\ &+ 2\gamma_k^3\zeta_k^2 L\langle\nabla f(x_g^k), \Omega_k\rangle - 2\gamma_k\zeta_k\langle\Omega_k, x^k - x^*\rangle \\ &+ 2\gamma_k^2\zeta_k^2\|\Omega_k\|^2,\end{aligned}$$

which concludes the first part of the lemma. Additionally, summing up, and using that the sequence of stepsizes γ_k is non-increasing and $\zeta_k(\zeta_k - 1) \leq \zeta_{k-1}^2$ with the fact that $\zeta_0(\zeta_0 - 1) \leq 0$, we claim that the second part of the lemma holds. □

D.2 Main theorem, convex case

Theorem 6 (Theorem 2 and Corollary 2). *Suppose that Assumptions 1, 2, 3, 4, 5, 7, and 8 hold. Then, if we choose parameters in the following way:*

$$\zeta_k = \frac{k+2}{2}, \quad \theta_k = 1 - \frac{1}{\zeta_k}, \quad M_k \equiv M = \frac{43 \cdot 175^2 K}{93^2}, \quad B_k \equiv B = 10\log_2(M); \qquad (52)$$

$$\gamma_k = \min\left\{\frac{1}{360000L\ln^2\left(\frac{6K}{\delta}\right)}, \frac{R_0}{\sqrt{43}\cdot 175\sqrt{\tau_{\text{mix}}}\sigma\zeta_k\sqrt{K}\ln\left(\frac{6K}{\delta}\right)}\right\}; \qquad (53)$$

$$\lambda_k = \frac{R_0}{93\gamma_k\,\zeta_k\ln\left(\frac{6K}{\delta}\right)}, \qquad (54)$$

then, after K iterations of Algorithm 2, the output x_f^K satisfies the following inequality with probability at least $1 - \delta$:

$$f(x_f^K) - f(x^*) \leq \mathcal{O}\left(\max\left\{\frac{LR_0^2\ln^2\left(\frac{K}{\delta}\right)}{K^2}, \frac{\sqrt{\tau_{\text{mix}}}\sigma R_0\ln\left(\frac{6K}{\delta}\right)}{\sqrt{K}}\right\}\right).$$

As a consequence, to obtain an ε-accuracy, i.e. $f(x_f^K) - f(x^*) \leq \varepsilon$ with probability at least $1-\delta$, it suffices to run the algorithm for

$$K = \tilde{\mathcal{O}}\left(\max\left\{\sqrt{\frac{LR_0^2}{\varepsilon}}, \frac{\tau_{\text{mix}}\sigma^2 R_0^2}{\varepsilon^2}\right\}\right)$$

iterations.

Proof. Similarly to the proof of Sadiev et al. (2023), we introduce the probabilistic events E_k in the following way:

$$\sum_{l=0}^{t-1} 2L\gamma_l^3\zeta_l^2\langle\nabla f(x_g^l), \Omega_l\rangle - \sum_{l=0}^{t-1} 2\gamma_l\zeta_l\langle\Omega_l, x^l - x^*\rangle + \sum_{l=0}^{t-1} 2\gamma_l^2\zeta_l^2\|\Omega_l\|^2 \leq R_0^2; \tag{55}$$

$$R_t \leq \sqrt{2}R_0 \tag{56}$$

for all $t = 0, \ldots, k$, where $R_t := \|x^t - x^*\|$. We want to prove via induction that the event E_k holds with probability at least $1 - \frac{k\delta}{K}$. For the case $k = 0$ it is obvious. Next, let us assume that the event E_{T-1} holds for some $T \leq K$ with probability at least $1 - \frac{(T-1)\delta}{K}$. At the same time, we can apply Lemma 9:

$$\begin{aligned}2\gamma_{t-1}\zeta_{t-1}^2\left(f(x_f^t) - f(x^*)\right) \leq{}& \|x^0 - x^*\|^2 - \|x^t - x^*\|^2 - \frac{\gamma_{t-1}^2\zeta_{t-1}^2}{2}\|\nabla f(x_g^{t-1})\|^2 \\ &+ \sum_{l=0}^{t-1} 2L\gamma_l^3\zeta_l^2\langle\nabla f(x_g^l), \Omega_l\rangle - \sum_{l=0}^{t-1} 2\gamma_l\zeta_l\langle\Omega_l, x^l - x^*\rangle \\ &+ \sum_{l=0}^{t-1} 2\gamma_l^2\zeta_l^2\|\Omega_l\|^2.\end{aligned}$$

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

$$\begin{aligned}2\gamma_{t-1}\zeta_{t-1}^2\left(f(x_f^t) - f(x^*)\right) \leq{}& \|x^0 - x^*\|^2 - \|x^t - x^*\|^2 \\ &+ \sum_{l=0}^{t-1} 2L\gamma_l^3\zeta_l^2\langle\nabla f(x_g^l), \Omega_l\rangle - \sum_{l=0}^{t-1} 2\gamma_l\zeta_l\langle\Omega_l, x^l - x^*\rangle \\ &+ \sum_{l=0}^{t-1} 2\gamma_l^2\zeta_l^2\|\Omega_l\|^2 \leq 2R_0^2\end{aligned} \tag{57}$$

for all $t = 0, \ldots, T-1$. Next, let us denote η_t and ν_t for $t = 0, \ldots, T-1$ in the following way:

$$\nu_t = \begin{cases}\nabla f(x_g^t), & \|\nabla f(x_g^t)\| \leq \frac{2R_0}{\gamma_t\zeta_t}; \\ 0, & \text{otherwise},\end{cases} \tag{58}$$

and

$$\eta_t = \begin{cases}x_t - x^*, & \|x^t - x^*\| \leq 2R_0; \\ 0, & \text{otherwise}.\end{cases} \tag{59}$$

Thus, η_t and ν_t can be bounded with probability 1. We can also observe that the event E_{T-1} leads to inequality

$$f(x_f^t) - f(x^*) \leq \frac{R_0^2}{\gamma_{t-1}\zeta_{t-1}^2}$$

for all $1 \leq t \leq T-1$. Moreover, according to the update rule of Algorithm 2, for $t \geq 1$,

$$x_f^t = \frac{1}{\zeta_{t-1}}x^t + \left(1 - \frac{1}{\zeta_{t-1}}\right)x_f^{t-1}.$$

Since $\zeta_{t-1} \geq 1$, $x_f^0 = x^0$, and $\|x^t - x^*\| \leq \sqrt{2}R_0$ on E_{T-1}, we conclude by induction that

$$\|x_f^t - x^*\| \leq \sqrt{2}R_0$$

for all $0 \leq t \leq T-1$. Hence, for $1 \leq t \leq T-1$, using L-smoothness and the inequality $\|\nabla f(x)\|^2 \leq 2L(f(x) - f(x^*))$, we obtain

$$\begin{aligned}
\|\nabla f(x_g^t)\| &\leq \|\nabla f(x_g^t) - \nabla f(x_f^t)\| + \|\nabla f(x_f^t)\| \\
&\leq \frac{L}{\zeta_t}\|x^t - x_f^t\| + \sqrt{2L(f(x_f^t) - f(x^*))} \\
&\leq \frac{L}{\zeta_t}\|x^t - x^*\| + \frac{L}{\zeta_t}\|x_f^t - x^*\| + \sqrt{\frac{2LR_0^2}{\gamma_{t-1}\zeta_{t-1}^2}} \\
&\leq \frac{2\sqrt{2}LR_0}{\zeta_t} + \sqrt{\frac{2LR_0^2}{\gamma_{t-1}\zeta_{t-1}^2}} \overset{(54)}{\leq} \frac{\lambda_t}{2}.
\end{aligned}$$

For $t = 0$, we have $x_g^0 = x^0$. Since $\nabla f(x^*) = 0$, by L-smoothness,

$$\|\nabla f(x_g^0)\| \leq L\|x^0 - x^*\| = LR_0 \leq \frac{\lambda_0}{2}.$$

Thus, on the event E_{T-1},

$$\|\nabla f(x_g^t)\| \leq \frac{\lambda_t}{2}$$

for all $t = 0, \ldots, T-1$. Consequently, bounds from Lemma 6 are applicable at each iteration. For this purpose, we adopt the following notation:

$$\Omega_t^u \overset{def}{:=} \operatorname{clip}_{\lambda_t}(g^t) - \mathbb{E}_t \operatorname{clip}_{\lambda_t}(g^t), \quad \Omega_t^b \overset{def}{:=} \mathbb{E}_t \operatorname{clip}_{\lambda_t}(g^t) - \nabla f(x_g^t). \tag{60}$$

Lemma 6 and the definitions of (52) provide us with bounds necessary to proceed with the analysis:

$$\|\Omega_t^u\| \leq 2\lambda_t; \tag{61}$$

$$\|\Omega_t^b\| \leq \frac{4\sqrt{\tau_{\text{mix}}}\,\sigma}{\sqrt{M}} + \frac{196\tau_{\text{mix}}\,\sigma^2}{\lambda_t}; \tag{62}$$

$$\mathbb{E}_t\|\Omega_t^u\|^2 \leq 490\tau_{\text{mix}}\sigma^2 \tag{63}$$

due to the definition of $B = 10\log_2(M)$. Note that in our notation we have: $\Omega_t = \Omega_t^u + \Omega_t^b$. Thus, it follows from event E_{T-1} that

$$\begin{aligned}
R_t^2 \leq R_0^2 &+ \underbrace{\sum_{l=0}^{T-1} 4\gamma_l^2\zeta_l^2(\|\Omega_l^u\|^2 - \mathbb{E}_l\|\Omega_l^u\|^2)}_{①} + \underbrace{\sum_{l=0}^{T-1} 4\gamma_l^2\zeta_l^2\mathbb{E}_l\|\Omega_l^u\|^2}_{②} + \underbrace{\sum_{l=0}^{T-1} 4\gamma_l^2\zeta_l^2\|\Omega_l^b\|^2}_{③} \\
&\underbrace{- \sum_{l=0}^{T-1} 2\gamma_l\zeta_l\langle\Omega_l^b, \eta_l\rangle}_{④} \underbrace{- \sum_{l=0}^{T-1} 2\gamma_l\zeta_l\langle\Omega_l^u, \eta_l\rangle}_{⑤} + \underbrace{\sum_{l=0}^{T-1} 2L\gamma_l^3\zeta_l^2\langle\nu_l, \Omega_l^b\rangle}_{⑥} + \underbrace{\sum_{l=0}^{T-1} 2L\gamma_l^3\zeta_l^2\langle\nu_l, \Omega_l^u\rangle}_{⑦}.
\end{aligned}$$

We need to show that with high probability, the following inequality holds: ① + ② + ③ + ④ + ⑤ + ⑥ + ⑦ $\leq R_0^2$. Subsequently, we can bound these terms separately.

Upper bound for ①. We begin by noting that the sequence $4\gamma_l^2\zeta_l^2(\|\Omega_l^u\|^2 - \mathbb{E}_l\|\Omega_l^u\|^2)$ is a martingale difference sequence, i.e. $\mathbb{E}_l[4\gamma_l^2\zeta_l^2(\|\Omega_l^u\|^2 - \mathbb{E}_l\|\Omega_l^u\|^2)] = 0$. This allows us to use Bernstein's inequality (Lemma 2) to obtain an estimate for this term. Indeed, we have:

$$\begin{aligned}
|4\gamma_l^2\zeta_l^2(\|\Omega_l^u\|^2 - \mathbb{E}_l\|\Omega_l^u\|^2)| &\leq 4\gamma_l^2\zeta_l^2(\|\Omega_l^u\|^2 + \mathbb{E}_l\|\Omega_l^u\|^2) \overset{(61)}{\leq} 4 \cdot 8\gamma_l^2\zeta_l^2\lambda_l^2 \qquad (64) \\
&\overset{(54)}{=} \frac{32 \cdot \gamma_l^2\zeta_l^2 R_0^2}{93^2 \cdot \gamma_l^2\zeta_l^2 \ln^2\left(\frac{6K}{\delta}\right)} \\
&\leq \frac{R_0^2}{270\ln^2\left(\frac{6K}{\delta}\right)} \overset{def}{=:} c. \qquad (65)
\end{aligned}$$

The next step is to compute the variances $\hat{\sigma_l}^2 = \mathbb{E}_l\left[4\gamma_l^2\zeta_l^2(\|\Omega_l^u\|^2 - \mathbb{E}_l\|\Omega_l^u\|^2)\right]^2 \forall l = 0, \ldots, T-1$. From the definition of variance, it follows that under event E_{T-1} we have

$$\begin{aligned}\hat{\sigma}_l^2 &\leq \frac{R_0^2}{270\ln^2\left(\frac{6K}{\delta}\right)}\mathbb{E}_l[4\gamma_l^2\zeta_l^2|\|\Omega_l^u\|^2 - \mathbb{E}_l\|\Omega_l^u\|^2|] \leq \frac{8\gamma_l^2\zeta_l^2R_0^2}{270\ln^2\left(\frac{6K}{\delta}\right)}\mathbb{E}_l[\|\Omega_l^u\|^2] \\ &\overset{(63)}{\leq} \frac{490\cdot 8\tau_{\text{mix}}\sigma^2\gamma_l^2\zeta_l^2R_0^2}{270\ln^2\left(\frac{6K}{\delta}\right)} \overset{(53)}{\leq} \frac{490\cdot 8\tau_{\text{mix}}\sigma^2\zeta_l^2R_0^4}{270\cdot 43\cdot 175^2K\zeta_l^2\sigma^2\tau_{\text{mix}}\ln^4\left(\frac{6K}{\delta}\right)} \\ &\leq \frac{490\cdot 8R_0^4}{270\cdot 43\cdot 175^2K\ln\left(\frac{6K}{\delta}\right)} \leq \frac{R_0^4}{100K\ln\left(\frac{6K}{\delta}\right)}.\end{aligned}$$

Summing the conditional variances over all $l = 0, \ldots, T-1$, we obtain that event E_{T-1} implies:

$$\sum_{l=0}^{T-1}\hat{\sigma}_l^2 \leq \frac{R_0^4}{100\ln\left(\frac{6K}{\delta}\right)} \overset{def}{=}: G. \tag{66}$$

Defining $b \overset{def}{=} \frac{R_0^2}{7}$ and using Bernstein's inequality, we get

$$\mathbb{P}\left(|①| > \frac{R_0^2}{7} \text{ and } \sum_{l=0}^{T-1}\hat{\sigma}_l^2 \leq \frac{R_0^4}{100\ln\left(\frac{6K}{\delta}\right)}\right) \leq 2\exp\left(\frac{-b^2}{2G + \frac{2bc}{3}}\right) \leq \frac{\delta}{3K}.$$

Upper bound for ②. The event E_{T-1} implies:

$$\begin{aligned}② = \sum_{l=0}^{T-1}4\gamma_l^2\zeta_l^2\mathbb{E}_l\|\Omega_l^u\|^2 &\overset{(63)}{\leq} 490\tau_{\text{mix}}\sigma^2T\cdot 4\gamma_l^2\zeta_l^2 \overset{(53)}{\leq} \frac{490\tau_{\text{mix}}\sigma^2T\cdot 4R_0^2}{43\cdot 175^2\ln^2\left(\frac{6K}{\delta}\right)K\sigma^2\tau_{\text{mix}}} \\ &\leq \frac{490\cdot 4R_0^2}{43\cdot 175^2\ln\left(\frac{6K}{\delta}\right)} \leq \frac{R_0^2}{7}.\end{aligned}$$

Upper bound for ③. The event E_{T-1} implies:

$$\begin{aligned}③ = \sum_{l=0}^{T-1}4\gamma_l^2\zeta_l^2\|\Omega_l^b\|^2 &\overset{(62)}{\leq} \sum_{l=0}^{T-1}4\gamma_l^2\zeta_l^2\left(\frac{4\sqrt{\tau_{\text{mix}}}\sigma}{\sqrt{M}} + \frac{196\tau_{\text{mix}}\sigma^2}{\lambda_l}\right)^2 \\ &\overset{(52)(53)(54)}{\leq} \sum_{l=0}^{T-1}\frac{8\cdot 16\cdot 93^2\tau_{\text{mix}}\sigma^2R_0^2}{175^4\cdot 43^2K\ln^2\left(\frac{6K}{\delta}\right)\tau_{\text{mix}}\sigma^2} \\ &+ \sum_{l=0}^{T-1}\frac{8\cdot 196^2\cdot 93^2\tau_{\text{mix}}^2\sigma^4\cdot\ln^2\left(\frac{6K}{\delta}\right)R_0^4}{175^4\cdot 43^2R_0^2K^2\ln^4\left(\frac{6K}{\delta}\right)\tau_{\text{mix}}^2\sigma^4} \leq \frac{8\cdot 197^2\cdot 93^2\cdot R_0^2}{175^4\cdot 43^2} \leq \frac{R_0^2}{7}.\end{aligned}$$

Upper bound for ④. The event E_{T-1} implies that

$$\begin{aligned}④ = \sum_{l=0}^{T-1}2\gamma_l\zeta_l\langle\Omega_l^b, \eta_l\rangle &\leq \sum_{l=0}^{T-1}2\gamma_l\zeta_l\|\eta_l\|\cdot\|\Omega_l^b\| \\ &\overset{(59)(62)}{\leq} \sum_{l=0}^{T-1}4\cdot 196\frac{R_0\gamma_l\zeta_l\tau_{\text{mix}}\sigma^2}{\lambda_l} + \sum_{l=0}^{T-1}\frac{4\cdot 4R_0\gamma_l\zeta_l\sqrt{\tau_{\text{mix}}}\sigma}{\sqrt{M}} \\ &\overset{(53)(54)}{\leq} \sum_{l=0}^{T-1}\frac{4\cdot 196\cdot 93R_0\zeta_l^2\tau_{\text{mix}}\sigma^2R_0^2\ln\left(\frac{6K}{\delta}\right)}{175^2\cdot 43\ln^2\left(\frac{6K}{\delta}\right)K\sigma^2\tau_{\text{mix}}R_0} \\ &+ \sum_{l=0}^{T-1}\frac{4\cdot 196\cdot 93\sqrt{\tau_{\text{mix}}}\sigma R_0^2}{175^2\cdot 43\ln\left(\frac{6K}{\delta}\right)K\sigma\sqrt{\tau_{\text{mix}}}} \\ &\leq \frac{8\cdot 196\cdot 93R_0^2}{175^2\cdot 43} \leq \frac{R_0^2}{7}.\end{aligned}$$

Upper bound for ⑤. In order to bound this term, we note that according to definition (60) $\mathbb{E}_l[2\gamma_l\zeta_l\langle\Omega_l^u,\eta_l\rangle]=0$, which allows us to apply Bernstein's theorem. Indeed, we have:

$$|2\gamma_l\zeta_l\langle\Omega_l^u,\eta_l\rangle|\leq 2\gamma_l\zeta_l\|\Omega_l^u\|\cdot\|\eta_l\| \overset{(59)(61)}{\leq} 8\gamma_l\zeta_l R_0\lambda_l \overset{(54)}{\leq} \frac{8R_0^2}{93\ln\left(\frac{6K}{\delta}\right)} \overset{def}{=:} c.$$

The next step is to compute the variances $\hat{\sigma}_l^2=\mathbb{E}_l[2\gamma_l\zeta_l\langle\Omega_l^u,\eta_l\rangle]^2\ \forall l=0,\ldots,T-1$. From the definition of variance, it follows that under event E_{T-1} we have

$$\hat{\sigma}_l^2\leq 4\gamma_l^2\zeta_l^2\mathbb{E}_l[\|\Omega_l^u\|\cdot\|\eta_l\|]^2 \overset{(59)(63)}{\leq} 4\gamma_l^2\zeta_l^2\cdot 4R_0^2\cdot 490\tau_{\text{mix}}\sigma^2 \overset{(53)}{\leq} \frac{16\cdot 490R_0^4\tau_{\text{mix}}\sigma^2}{43\cdot 175^2K\ln^2\left(\frac{6K}{\delta}\right)\tau_{\text{mix}}\sigma^2}$$

$$\leq \frac{16\cdot 490R_0^4}{43\cdot 175^2\ln\left(\frac{6K}{\delta}\right)K}.$$

Summing over all $l=0,\ldots,T-1$, we obtain that event E_{T-1} implies:

$$\sum_{l=0}^{T-1}\hat{\sigma}_l^2\leq\frac{16\cdot 490R_0^4}{43\cdot 175^2\ln\left(\frac{6K}{\delta}\right)}\overset{def}{=:}G. \tag{67}$$

Defining $b\overset{def}{=}\frac{R_0^2}{7}$ and using Bernstein's inequality, we get that

$$\mathbb{P}\left(|⑤|>\frac{R_0^2}{7}\ \text{ and }\ \sum_{l=0}^{T-1}\hat{\sigma}_l^2\leq\frac{16\cdot 490R_0^4}{43\cdot 175^2\ln\left(\frac{6K}{\delta}\right)}\right)\leq 2\exp\left(\frac{-b^2}{2G+\frac{2bc}{3}}\right)\leq\frac{\delta}{3K}.$$

Upper bound for ⑥. The event E_{T-1} implies:

$$\begin{aligned}⑥=\sum_{l=0}^{T-1}2L\gamma_l^3\zeta_l^2\langle\nu_l,\Omega_l^b\rangle &\overset{(58)}{\leq}\sum_{l=0}^{T-1}2L\gamma_l^3\zeta_l^2\cdot\frac{2R_0}{\gamma_l\zeta_l}\|\Omega_l^b\| \overset{L\gamma_l\leq 1}{\leq}\sum_{l=0}^{T-1}4\gamma_l\zeta_l R_0\|\Omega_l^b\|\\ &\overset{(62)}{\leq}\sum_{l=0}^{T-1}4\gamma_l\zeta_l\cdot\left(\frac{4\sqrt{\tau_{\text{mix}}}\sigma}{\sqrt{M}}+\frac{196\tau_{\text{mix}}\sigma^2}{\lambda_l}\right)\\ &\overset{(52)(53)(54)}{\leq}\sum_{l=0}^{T-1}4\frac{200\cdot 93R_0^2}{43\cdot 175^2K\ln\left(\frac{6K}{\delta}\right)}\leq\frac{R_0^2}{7}.\end{aligned}$$

Upper bound for ⑦. The analysis here closely follows the one presented for bounding term ⑤. To bound ⑦, we note that according to definition (60) $\mathbb{E}_l[2L\gamma_l^3\zeta_l^2\langle\nu_l,\Omega_l^u\rangle]=0$, which allows us to apply Bernstein's theorem. Indeed, we have:

$$|2L\gamma_l^3\zeta_l^2\langle\nu_l,\Omega_l^u\rangle|\overset{L\gamma_l\leq 1}{\leq}2\gamma_l^2\zeta_l^2\|\Omega_l^u\|\cdot\|\nu_l\|\overset{(61)(58)}{\leq}8\gamma_l\zeta_l R_0\lambda_l\overset{(54)}{\leq}\frac{8R_0^2}{93\ln\left(\frac{6K}{\delta}\right)}\overset{def}{=:}c.$$

The next step is to compute the variances $\hat{\sigma}_l^2=\mathbb{E}_l[2L\gamma_l^3\zeta_l^2\langle\Omega_l^u,\nu_l\rangle]^2\ \forall l=0,\ldots,T-1$. From the definition of variance, it follows that under event E_{T-1} we have

$$\hat{\sigma}_l^2\overset{L\gamma_l\leq 1}{\leq}4\gamma_l^4\zeta_l^4\mathbb{E}_l[\|\Omega_l^u\|\cdot\|\nu_l\|]^2\overset{(58)(63)}{\leq}16\gamma_l^2\zeta_l^2\cdot R_0^2\cdot 490\tau_{\text{mix}}\sigma^2\overset{(53)}{\leq}\frac{16\cdot 490R_0^4\tau_{\text{mix}}\sigma^2}{43\cdot 175^2K\ln^2\left(\frac{6K}{\delta}\right)\tau_{\text{mix}}\sigma^2}$$

$$\leq\frac{16\cdot 490R_0^4}{43\cdot 175^2\ln\left(\frac{6K}{\delta}\right)K}.$$

Summing over all $l=0,\ldots,T-1$, we obtain that event E_{T-1} implies:

$$\sum_{l=0}^{T-1}\hat{\sigma}_l^2\leq\frac{16\cdot 490R_0^4}{43\cdot 175^2\ln\left(\frac{6K}{\delta}\right)}\overset{def}{=:}G. \tag{68}$$

Defining $b\overset{def}{=}\frac{R_0^2}{7}$ and using Bernstein's inequality, we get that

$$\mathbb{P}\left(|⑦|>\frac{R_0^2}{7}\ \text{ and }\ \sum_{l=0}^{T-1}\hat{\sigma}_l^2\leq\frac{16\cdot 490R_0^4}{43\cdot 175^2\ln\left(\frac{6K}{\delta}\right)}\right)\leq 2\exp\left(\frac{-b^2}{2G+\frac{2bc}{3}}\right)\leq\frac{\delta}{3K}.$$

Now, we have the upper bounds for ① - ⑦. If we denote

$$E_{①} = \left\{ |①| \leq \frac{R_0^2}{7} \text{ or } \sum_{l=0}^{T-1} \hat{\sigma}_l^2 > \frac{R_0^4}{100 \ln\left(\frac{6K}{\delta}\right)} \right\};$$
$$E_{⑤} = \left\{ |⑤| \leq \frac{R_0^2}{7} \text{ or } \sum_{l=0}^{T-1} \hat{\sigma}_l^2 > \frac{16 \cdot 490 R_0^4}{43 \cdot 175^2 \ln\left(\frac{6K}{\delta}\right)} \right\};$$
$$E_{⑦} = \left\{ |⑦| \leq \frac{R_0^2}{7} \text{ or } \sum_{l=0}^{T-1} \hat{\sigma}_l^2 > \frac{16 \cdot 490 R_0^4}{43 \cdot 175^2 \ln\left(\frac{6K}{\delta}\right)} \right\},$$

then the event $E_{T-1} \cap E_{①} \cap E_{⑤} \cap E_{⑦}$ implies:

$$R_T^2 \leq R_0^2 + ① + ② + ③ + ④ + ⑤ + ⑥ + ⑦ \leq 2R_0^2$$
$$② \leq \frac{R_0^2}{7}, \quad ③ \leq \frac{R_0^2}{7}, \quad ④ \leq \frac{R_0^2}{7}, \quad ⑥ \leq \frac{R_0^2}{7}.$$

To complete the induction step, observe that the probability of the event E_T can be bounded as follows:

$$\mathbb{P}(E_T) \geq \mathbb{P}(E_{T-1} \cap E_{①} \cap E_{⑤} \cap E_{⑦}) \geq 1 - \left(\frac{\delta(T-1)}{K} + \frac{\delta}{3K} + \frac{\delta}{3K} + \frac{\delta}{3K} \right) = 1 - \frac{T\delta}{K}$$

. Thus, we have shown that inequalities (55), (56) hold for $t = T$. Consequently, they are true for all $t = 0, ..., T$. Hence, the event E_K holds with probability at least $1 - \delta$. More specifically, inequality (57) is satisfied with probability at least $1 - \delta$. This enables us to derive a bound for $f(x_f^K) - f(x^*)$:

$$f(x_f^K) - f(x^*) \leq \frac{R_0^2}{\gamma_{K-1} \zeta_{K-1}^2}. \tag{69}$$

Substituting γ_{K-1} and ζ_{K-1} into (69), we obtain:

$$f(x_f^K) - f(x^*) = \mathcal{O}\left(\max\left\{ \frac{LR_0^2 \ln^2\left(\frac{K}{\delta}\right)}{K^2}, \frac{\sqrt{\tau_{\text{mix}}} \sigma R_0 \ln\left(\frac{K}{\delta}\right)}{\sqrt{K}} \right\} \right).$$

To obtain $f(x_f^K) - f(x^*) \leq \varepsilon$ with probability at least $1 - \delta$, it suffices to choose K such that both terms inside the maximum are $O(\varepsilon)$. This yields

$$K = \tilde{\mathcal{O}}\left(\max\left\{ \sqrt{\frac{LR_0^2}{\varepsilon}}, \frac{\tau_{\text{mix}} \sigma^2 R_0^2}{\varepsilon^2} \right\} \right).$$

This concludes the proof. □

E Oracle complexity

In this section, we provide the theory for the oracle complexity.

Theorem 7 (Theorem 3). *Suppose that Algorithm 1 or 2 runs K iterations with parameters B_k and $M_k \geq 2$, $k = 0, \ldots, K-1$. Then, the total number of oracle calls N_{calls} can be upper bounded with probability at least $1 - \delta$ as follows:*

$$N_{calls} \leq B\left(4K \ln(M) + \frac{4}{3} M \ln\left(\frac{1}{\delta}\right) + 2\sqrt{KM \ln\left(\frac{1}{\delta}\right)} \right),$$

where $B = \max_k B_k$, $M = \max_k M_k$.

Proof. The oracle complexity of the algorithm on iteration k equals the batch size $\mathcal{C}_k := \mathbb{I}\{2^{J_k} \leq M_k\} 2^{J_k} B_k + \mathbb{I}\{2^{J_k} > M_k\} B_k$. To obtain a high-probability guarantee on the total oracle cost, we apply Bernstein's inequality.

The first step is to bound $\mathbb{E}[\mathcal{C}_k]$ and $Var[\mathcal{C}_k]$ $\forall k = 0, \ldots, K-1$. For this purpose, we define the parameter $m_k := \lfloor \log_2 (M_k) \rfloor$. Then, we get

$$\mathbb{P}(\mathcal{C}_k = 2^t B_k) = 2^{-t} \quad \forall t = 1, \ldots, m_k \quad \Longrightarrow \quad \mathbb{P}(\mathcal{C}_k = B_k) = 1 - \frac{1}{2} - \cdots - \frac{1}{2^{m_k}} = \frac{1}{2^{m_k}}.$$

Therefore,

$$\begin{aligned}\mathbb{E}[\mathcal{C}_k] &= 2^{-m_k} B_k + B_k \underbrace{(1 + \cdots + 1)}_{m_k} \\ &= B_k (m_k + 2^{-m_k}) \leq 2 B_k m_k = 2 B_k \log_2 (M_k) \leq 4 B_k \ln (M_k),\end{aligned}$$

and

$$Var[\mathcal{C}_k] = \mathbb{E}[\mathcal{C}_k^2] - \mathbb{E}[\mathcal{C}_k]^2 = B_k^2 \left(2^{-m_k} + \sum_{t=1}^{m} \frac{4^t}{2^t} - (2^{-m_k} + m_k)^2 \right) \leq 2^{m_k+1} B_k^2 \leq 2 B_k^2 M_k.$$

Let $X_k := \mathcal{C}_k - \mathbb{E}[\mathcal{C}_k]$. Therefore, we can apply Bernstein's inequality (see Lemma 2) to the martingale difference sequence X_k, since $\mathbb{E}[X_k | X_{k-1}, \ldots, X_0] = 0$ due to the independence. Taking into account that $M_k \leq M$ and $B_k \leq B$ for all k with the definition of $b := B \left(\frac{4}{3} M \ln \left(\frac{1}{\delta} \right) + 2 \sqrt{K M \ln \left(\frac{1}{\delta} \right)} \right)$ and the fact that $|X_k| = |C_k - \mathbb{E}[C_k]| \leq 2 B_k M_k \leq 2BM$, we obtain that

$$\mathbb{P} \left(\sum_{k=0}^{K-1} X_k > b \right) \leq \exp \left(\frac{-b^2}{4KB^2M + \frac{2}{3} \cdot 2BM \cdot b} \right) \leq \exp \left(- \ln \left(\frac{1}{\delta} \right) \right) = \delta$$

$$\begin{aligned}&\mathbb{P} \left(\sum_{k=0}^{K-1} X_k > b \right) \leq \delta \\ &\Longrightarrow \mathbb{P} \left(\sum_{k=0}^{K-1} \mathcal{C}_k \leq B \left(4K \ln (M) + \frac{4}{3} M \ln \left(\frac{1}{\delta} \right) + 2 \sqrt{K M \ln \left(\frac{1}{\delta} \right)} \right) \right) \geq 1 - \delta.\end{aligned}$$

This finishes the proof. □

Corollary 7 (Corollary 3). *Suppose that conditions of Theorem 1 hold. Then, the required number of oracle calls for reaching ε-accuracy, i.e. $\frac{1}{K} \sum_{k=0}^{K-1} \left\| \nabla f(x_k) \right\|^2 \leq \varepsilon^2$, can be upper bounded with probability at least $1 - 2\delta$ as follows:*

$$N_{calls} = \tilde{\mathcal{O}} \left(\max \left\{ \frac{L \Delta_0}{\varepsilon^2}, \frac{\tau_{\text{mix}} \sigma^2}{\varepsilon^2}, \frac{\tau_{\text{mix}} \sigma^2 L \Delta_0}{\varepsilon^4} \right\} \right).$$

Proof. Using Corollary 6 and Theorem 7 with $B = 1$, $M = K$, we derive

$$N_{calls} = \tilde{\mathcal{O}} \left(\max \left\{ \frac{L \Delta_0}{\varepsilon^2}, \frac{\tau_{\text{mix}} \sigma^2}{\varepsilon^2}, \frac{\tau_{\text{mix}} \sigma^2 L \Delta_0}{\varepsilon^4} \right\} \right),$$

where $\tilde{\mathcal{O}}$ hides the polylogarithmic dependency on ε and δ. This concludes the proof. □

Corollary 8 (Corollary 4). *Suppose that conditions of Theorem 2 hold. Then, the required number of oracle calls to reach ε-accuracy, i.e., $f(x_f^K) - f^* \leq \varepsilon$ with Algorithm 2, can be upper bounded with probability at least $1 - 2\delta$ as follows:*

$$N_{calls} = \tilde{\mathcal{O}} \left(\max \left\{ \sqrt{\frac{L R_0^2}{\varepsilon}}, \frac{\tau_{\text{mix}} \sigma^2 R_0^2}{\varepsilon^2} \right\} \right).$$

Proof. Using Theorem 7 with $B_k = 10 \log_2 (M_k)$, $M_k = \frac{43 \cdot 175^2 K}{93^2}$ we have that if Algorithm 2 runs K iterations, the number of oracle calls can be bounded as

$$N_{calls} = \tilde{\mathcal{O}} (K)$$

with probability at least $1-\delta$. Indeed, according to the conditions of Theorem 2 (equal to Theorem 6), we get

$$M_k \equiv M = \Theta(K) \quad \text{and } B_k \equiv B = \Theta\left(\ln\left(K\right)\right).$$

Therefore, Theorem 7 gives

$$N_{calls} = \tilde{\mathcal{O}}\left(K\right),$$

with probability at least $1-\delta$, where $\tilde{\mathcal{O}}$ hides the polylogarithmic dependency on K and δ. Applying Theorem 6 for the boundary on K with probability at least $1-\delta$, we derive the final result with probability at least $1-2\delta$. □

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: see Section 1 and contribution paragraph.

   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 discussed limitations of our results in 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: All assumptions and completed proofs are provided in Section 3 and Appendix.

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 does not include 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]

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 paper follows 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 paper is theoretical 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: The paper poses no such risks.

    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: The paper does not use such assets.

    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 does not release new assets.

    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 nor 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 nor 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: [N/A]

Justification: the core method development in this paper does not involve LLMs.

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.