CONVERGENCE ANALYSIS OF STORM UNDER DIFFERENT GEOMETRIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Stochastic recursive momentum (STORM) achieves fast convergence for nonconvex optimization via the variance reduction effect, but existing analyses rely on the strong average smoothness assumption. In this paper, we study the convergence of STORM for different objectives without average smoothness. We first revisit the results under average smoothness, obtaining the $O(T^{-1/3})$ bound for nonconvex objectives and the $O(\sigma^2/(\mu T))$ bound for last-iterate output under the μ-Polyak–Łojasiewicz (PL) condition. Without average smoothness, we design an auxiliary sequence and compare the STORM update with it in the analysis. With the help of this sequence, we prove that STORM still attains an $O(T^{-1/4})$ rate for nonconvex objectives, which is optimal under standard smoothness. For convex and λ-strongly convex objectives, we further prove averaged and last-iterate bounds with optimal rates of $O(\sigma R/\sqrt{T})$ and $O(\sigma^2/(\lambda T))$, respectively. All the obtained results use the same STORM recursion with different hyperparameter choices.

1 INTRODUCTION

In this paper, we consider the stochastic optimization problem

$$\min_{x\in\mathbb{R}^d} F(x), \qquad F(x)=\mathbb{E}_{\xi\sim\mathcal{D}}[f(x;\xi)], \tag{1}$$

where function $F(\cdot)$ is differentiable and L-smooth. We have access to an unbiased stochastic gradient $g(x,\xi)$ with bounded variance such that $\mathbb{E}_\xi g(x,\xi)=\nabla F(x)$ and $\mathbb{E}_\xi \left\|g(x,\xi)-\nabla F(x)\right\|^2 \le \sigma^2$. We measure the convergence rate by the expected gradient norm $\mathbb{E}\left\|\nabla F(x)\right\|$ for the general nonconvex functions and by the expected function gap $\mathbb{E}[F(x)]-F^*$ for the other cases.

For nonconvex objectives, stochastic gradient descent (SGD) attains an optimal $O(T^{-1/4})$ convergence rate under the smoothness assumption (Ghadimi & Lan, 2013). Variance reduction can further improve this rate to $O(T^{-1/3})$ under the additional average smoothness condition (Fang et al., 2018; Arjevani et al., 2023). Among such methods, STORM (Cutkosky & Orabona, 2019) is attractive because it uses a constant batch size per iteration and does not require periodically computing gradients with large batches. Starting from $d_0=g(x_0,\xi_0)$, it performs

$$x_{t+1}=x_t-\eta_t d_t, \tag{2}$$

$$d_{t+1}=g(x_{t+1},\xi_{t+1})+(1-a_{t+1})\big(d_t-g(x_t,\xi_{t+1})\big). \tag{3}$$

Here η_t is the stepsize and a_{t+1} is the momentum parameter.

Although STORM achieves the optimal rate for nonconvex objectives under average smoothness, many issues remain to be investigated. First, existing analysis relies on the stronger *average smoothness* condition, which assumes that $\mathbb{E}_\xi \left\|g(x,\xi)-g(y,\xi)\right\|^2 \le \mathcal{L}^2\left\|x-y\right\|^2$. It remains unknown whether STORM can achieve the same $O(T^{-1/4})$ rate as SGD under standard smoothness alone. Second, existing guarantees do not establish optimal convergence rates for convex or strongly convex objectives without the average smoothness assumption.

In this paper, we analyze these questions for the original STORM framework. First, we revisit the previous analysis under average smoothness and obtain the following results through a simpler and more transparent analysis. We do not regard the average smoothness results as the main contribution.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Convergence of STORM under different settings. All rows assume an L-smooth objective and an unbiased oracle with variance at most σ^2. AS denotes average smoothness. The nonconvex rows measure the expected gradient norm, while the other rows measure the expected function gap. Only the dominant term is presented.

Objective	Assumption	Output	Convergence rate
Nonconvex	AS	Uniform random	$T^{-1/3}$
Nonconvex	–	Uniform random	$T^{-1/4}$
μ-PL	AS	Last iterate	$\sigma^2/(\mu T)$
Convex	–	Weighted mean; last iterate	$\sigma R/\sqrt{T}$
λ-strongly convex	–	Weighted mean; last iterate	$\sigma^2/(\lambda T)$

- With average smoothness, we first improve the previous analysis of STORM, achieving the $O(T^{-1/3})$ rate for nonconvex functions. Compared with previous analysis, we remove the logarithmic factor in the bound and avoid using large batches in the first iteration.
- With average smoothness and the μ-PL condition, we obtain a *last-iterate* bound with the convergence rate of $O(\sigma^2/(\mu T))$ via a two-phase stepsize design, allowing exponential decay of the initialization term.

Then, we investigate the convergence of STORM without the average smoothness condition, achieving the following results, which are the main contributions of this paper.

- With standard smoothness, we prove the $O(T^{-1/4})$ convergence rate for STORM, which matches the classical SGD rate and is optimal under standard assumptions.
- For smooth and convex objectives, we obtain the weighted-mean and last-iterate bounds of order $\mathcal{O}(LR^2/T + \sigma R/\sqrt{T})$, where R is the initial distance to a minimizer.
- For smooth and λ-strongly convex objectives, we demonstrate that the weighted-mean and last-iterate bounds can be further improved to $\mathcal{O}(\sigma^2/(\lambda T))$.

Table 1 summarizes the above convergence guarantees, which are all optimal in terms of the iteration number T under the corresponding settings.

2 Related work

Variance reduction for nonconvex objectives. SARAH (Nguyen et al., 2017) first introduced a recursive estimator based on successive stochastic gradient differences. SPIDER (Fang et al., 2018) then uses this principle to achieve $O(\varepsilon^{-3})$ stochastic gradient complexity and $O(n + \sqrt{n}\,\varepsilon^{-2})$ finite-sum complexity for ensuring $\mathbb{E}\,\|\nabla F(x)\| \le \varepsilon$, where n is the number of component functions. These guarantees assume average smoothness. SPIDER periodically refreshes its estimator using a large batch or the full gradient. STORM (Cutkosky & Orabona, 2019) instead incorporates a fresh-gradient difference term at every step, removing these periodic refreshes. However, the original STORM analysis contains logarithmic factors in the rate. In particular, its guarantee is $O(\log(T)T^{-2/3})$ in expected squared gradient norm, hence $O(\sqrt{\log(T)}T^{-1/3})$ in expected norm.[1] Other variants of STORM have also been investigated in the literature. STORM+ (Levy et al., 2021) obtains $O(T^{-1/3})$ with adaptive parameters under the extra assumption of bounded gradients and bounded objective value. Ada-STORM (Jiang et al., 2024) attains the same rates without additional assumptions but with a large initial batchsize on the order $\mathcal{O}(T^{1/3})$.

STORM-based methods for convex optimization. The STORM estimator has also been studied beyond nonconvex optimization. Zhang et al. (2020) combine the STORM estimator with Frank–Wolfe updates and obtain an $O(T^{-1/2})$ expected function-gap bound for convex objectives. However,

[1] In Cutkosky & Orabona (2019), the last simplification in Theorem 1's proof drops a factor $\sqrt{M}$ uncarefully, where $M = O(\log(T+2))$. Furthermore, their Appendix B, Theorem 2, explicitly retains the logarithmic factor for the nonadaptive guarantee.

their analysis requires a bounded feasible set, samplewise smoothness, and bounded stochastic gradients. More recently, Dahan & Levy (2025) combine STORM with anytime averaging in μ^2-SGD and attain the $O(T^{-1/2})$ rate. Their method evaluates gradients at running averages of a separate projected sequence, and its analysis uses samplewise smoothness and a bounded feasible set.

Variance reduction under the μ-PL condition. The variance reduction technique PAGE (Li et al., 2021) obtains an expected function gap of at most ε with $O((B + \sqrt{B}\,\mathcal{L}/\mu)\log(2\Delta_0/\varepsilon))$ stochastic gradient evaluations, where $B = \lceil 2\sigma^2/(\mu\varepsilon)\rceil$. This result assumes the average smoothness and bounded variance, and its refresh batch depends on the target accuracy. For compositional objectives, RECOVER (Qi et al., 2021) obtains the sample complexity $O(\max\{\mu^{-3/2}\varepsilon^{-1/2}, \mu^{-1}\varepsilon^{-1}\})$. For stochastic multi-level compositional optimization, Jiang et al. (2022) develop stage-wise SMVR using STORM estimators. For objectives satisfying the μ-PL condition, their stage-wise analysis gives complexity $O(\mu^{-1}\varepsilon^{-1} + \mu^{-3/2}\varepsilon^{-1/2})$, which reduces to $O((\mu\varepsilon)^{-1})$ when $\varepsilon \le \mu$.

3 Assumptions

We now state the assumptions for problem (1). The first two assumptions are used throughout the paper and the remaining assumptions specify the setting of each theorem.

Assumption 1 (Standard smoothness)**.** The differentiable objective $F : \mathbb{R}^d \to \mathbb{R}$ has finite infimum $F^* = \inf_x F(x)$ and satisfies

$$\|\nabla F(x) - \nabla F(y)\| \le L\,\|x - y\| \quad \text{for all } x, y \in \mathbb{R}^d. \tag{4}$$

Assumption 2 (Stochastic oracle)**.** The stochastic oracle g is unbiased and has bounded variance:

$$\mathbb{E}_\xi g(x,\xi) = \nabla F(x), \qquad \mathbb{E}_\xi\,\|g(x,\xi) - \nabla F(x)\|^2 \le \sigma^2 \quad \text{for all } x \in \mathbb{R}^d. \tag{5}$$

Assumption 3 (Average smoothness)**.** Stochastic gradient differences satisfy

$$\mathbb{E}_\xi\,\|g(x,\xi) - g(y,\xi)\|^2 \le \mathcal{L}^2\,\|x - y\|^2 \quad \text{for all } x, y \in \mathbb{R}^d. \tag{6}$$

Note that this assumption is also called mean-square smoothness. By Jensen's inequality, Assumptions 2 and 3 imply that F is also $\mathcal{L}$-smooth, so we assume $L \le \mathcal{L}$ without loss of generality.

Assumption 4 (PL condition)**.** For some $0 < \mu \le L$, the objective satisfies

$$\tfrac{1}{2}\,\|\nabla F(x)\|^2 \ge \mu(F(x) - F^*) \quad \text{for all } x \in \mathbb{R}^d. \tag{7}$$

Assumption 5 (Convexity)**.** The objective F has a minimizer x^* and satisfies

$$F(y) \ge F(x) + \langle \nabla F(x), y - x\rangle \quad \text{for all } x, y \in \mathbb{R}^d. \tag{8}$$

Assumption 6 (Strong convexity)**.** The objective F has a minimizer x^* and satisfies

$$F(y) \ge F(x) + \langle \nabla F(x), y - x\rangle + \tfrac{\lambda}{2}\,\|y - x\|^2 \quad \text{for all } x, y \in \mathbb{R}^d. \tag{9}$$

Strong convexity also implies the PL condition with $\mu = \lambda$, and the converse does not hold.

We initialize at x_0, and define $\Delta_0 = F(x_0) - F^*$, $R = \|x_0 - x^*\|$ and $\kappa = L/\lambda$ when applicable.

4 Convergence under Average Smoothness

We first revisit the convergence results under average smoothness (Assumption 3). We begin with the general nonconvex objectives, then use the PL condition to obtain the function-gap guarantee.

4.1 Nonconvex Objectives

For the STORM algorithm, the estimator d_t tracks the true function gradient $\nabla F(x_t)$, and we define $q_t = \mathbb{E}\,\|d_t - \nabla F(x_t)\|^2$ for its mean squared error. Average smoothness bounds the movement error by a term proportional to $\mathcal{L}^2\eta_t^2\mathbb{E}\,\|d_t\|^2$ via $\mathbb{E}\,\|g(x_{t+1},\xi_{t+1}) - g(x_t,\xi_{t+1})\|^2 \le \mathcal{L}^2\mathbb{E}\,\|x_{t+1} - x_t\|^2$. The descent inequality also contains a negative term involving $\mathbb{E}\,\|d_t\|^2$. To cancel them, we use the potential $\Psi_t = \mathbb{E}[F(x_t) - F^*] + \eta q_t/(2a_t)$ to obtain

$$\Psi_{t+1} - \Psi_t \le -\frac{\eta}{2}\mathbb{E}\,\|\nabla F(x_t)\|^2 + \eta a_{t+1}\sigma^2.$$

The sum of $a_{t+1} = (t+3)^{-2/3}$ is $O(T^{1/3})$, which leads to the following bound.

Theorem 1. *Under Assumptions 1, 2, and 3, setting* $\eta_t = \eta = \frac{1}{2\mathcal{L}(T+2)^{1/3}}$, $a_t = (t+2)^{-2/3}$, *we have the following guarantee*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla F(x_t)\right\|^2 \le \frac{4\mathcal{L}\Delta_0(T+2)^{1/3}}{T} + \frac{2^{2/3}\sigma^2}{T} + \frac{6\sigma^2}{T}\big((T+2)^{1/3} - 2^{1/3}\big). \tag{10}$$

For I uniform on $\{0,\dots,T-1\}$ and independent of the trajectory,

$$\mathbb{E}\left\|\nabla F(x_I)\right\| = O(\sqrt{\mathcal{L}\Delta_0 + \sigma^2}\, T^{-1/3}).$$

When $\sigma = 0$, the alternative choice $\eta = 1/(4\mathcal{L})$ and $a_t = 1/4$ gives

$$\mathbb{E}\left\|\nabla F(x_I)\right\| = O(\sqrt{\mathcal{L}\Delta_0}/\sqrt{T}).$$

Remark. Theorem 1 gives an $O(T^{-1/3})$ rate and the corresponding gradient complexity is $O(\varepsilon^{-3})$, matching the lower bound for nonconvex objectives with average smoothness (Arjevani et al., 2023). The theorem removes the logarithmic factor in the original STORM guarantee (Cutkosky & Orabona, 2019) and avoids using a large initial batch on the order of $\mathcal{O}(T^{1/3})$ (Jiang et al., 2024).

4.2 Objectives Satisfying the PL Condition

The PL inequality converts the negative gradient-norm term in the descent bound into a decrease proportional to the current function gap, i.e., $-\left\|\nabla F(x_t)\right\|^2 \le -2\mu(F(x_t) - F^*)$. In this case, we set the potential $V_t = \mathbb{E}[F(x_t) - F^*] + q_t/\mu$ and couple the parameters through $a_{t+1} = \mu\eta_t$. The descent and tracking inequalities will then give

$$V_{t+1} \le (1-\mu\eta_t)V_t + 2\mu\eta_t^2\sigma^2 \quad \text{when } \eta_t \le \mu/(8\mathcal{L}^2). \tag{11}$$

Then, we can prove the following theorem under the PL condition.

Theorem 2. ***(Harmonic schedule)*** *Under Assumptions 1–4, for $s \ge 16(\mathcal{L}/\mu)^2$, set $\eta_t = \frac{2}{\mu(t+s)}$, $a_{t+1} = \frac{2}{t+s}$. For every integer $T \ge 0$, the last iterate satisfies*

$$\mathbb{E}[F(x_T) - F^*] \le \left(\frac{s}{T+s}\right)^2\left(\Delta_0 + \frac{\sigma^2}{\mu}\right) + \frac{8\sigma^2}{\mu(T+s)}. \tag{12}$$

(Two-phase schedule) *Alternatively, let $m = \lfloor T/2 \rfloor$ and use*

$$\mu\eta_t = a_{t+1} = \begin{cases} 2/s, & 0 \le t < m, \\ 2/(s+t-m), & m \le t < T. \end{cases} \tag{13}$$

Without resetting either state, this gives

$$\mathbb{E}[F(x_T) - F^*] \le e^{-2m/s}\left(\Delta_0 + \frac{\sigma^2}{\mu}\right) + \frac{24\sigma^2}{\mu(T+2s)}. \tag{14}$$

Remark. The first hyperparameter schedule gives quadratic decay of the initialization term and the stochastic order $\sigma^2/(\mu T)$. Then, the alternative two-phase schedule first reduces the initial error geometrically and then decreases the stepsize to reduce the stochastic error. It carries (x_t, d_t) between phases without restarting and introduces no extra non-constant factor in the leading stochastic term.

The dominant order $O(1/(\mu T))$ is unimprovable, since lower bounds for strongly convex functions also hold for PL objectives (Agarwal et al., 2012). One important feature of our results is that the same estimator runs through both phases with one sample per step, and the rate is for the last iterate.

Core of the analysis. We have already proved $V_{t+1} \le (1-\mu\eta_t)V_t + 2\mu\eta_t^2\sigma^2$ in the previous analysis. When using $\mu\eta_t = 2/(s+t)$, we have:

$$\prod_{j=k+1}^{T-1}\left(1 - \frac{2}{j+s}\right) = \frac{(k+s-1)(k+s)}{(T+s-2)(T+s-1)}.$$

The initial error is thus multiplied by $O(s^2/(T+s)^2)$, while each noise term's $(k+s)^{-2}$ factor is canceled, up to constants, by its product weight $(k+s-1)(k+s)$. Their sum gives $O(\sigma^2/[\mu(T+s)])$. In the first phase of the alternative schedule, by using a constant learning rate $\mu\eta_t = 2/s$, the inequality $(1-2/s)^m \le e^{-2m/s}$ gives exponential decay.

5 CONVERGENCE WITHOUT AVERAGE SMOOTHNESS

Without the average smoothness, nearby query points may not have nearby stochastic gradients. The main difficulty is the conditional bias of the estimator d_t. Denote $e_t = d_t - \nabla F(x_t)$. We separate this inherited error from a centered increment ν_t via $e_t = \beta e_{t-1} + \nu_t$. Bounded variance can control $\mathbb{E}\left\|e_t\right\|^2$, but this alone would leave a nonvanishing term in the usual descent inequality.

We handle the inherited error in two ways. For nonconvex objectives, we shift the iterate by a multiple of e_{t-1}, so that the resulting recursion contains only the centered noise. Standard smoothness then controls the difference between the two gradient evaluation points. For convex objectives, we compare STORM with an auxiliary sequence driven by the same ν_t. The noise then cancels in their difference, which can be controlled by nonexpansiveness, or by contraction under strong convexity.

5.1 NONCONVEX OBJECTIVES

The faster convergence guarantee in Theorem 1 uses the average smoothness condition. Without it, we show that STORM can still attain the optimal $O(T^{-1/4})$ rate with standard smoothness.

Theorem 3. *Under Assumptions 1 and 2, fixing any $\beta \in (0,1)$ and setting $a_{t+1} = 1-\beta$ and $0 < \eta_t = \eta \le 1/(2L)$, we have the following guarantee*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla F(x_t)\right\|^2 \le \frac{2\Delta_0}{\eta T} + \frac{1+\beta}{(1-\beta)^3}L\eta\sigma^2. \tag{15}$$

When $\Delta_0, \sigma > 0$, choose $\eta = \min\{1/(2L), \sqrt{\Delta_0/(6L\sigma^2 T)}\}$ and $0 < \beta \le 1/2$. An independent uniform index $I \in \{0,\dots,T-1\}$ satisfies

$$\mathbb{E}\left\|\nabla F(x_I)\right\| = \mathcal{O}\left(\sqrt{L\Delta_0}\,T^{-1/2} + (L\Delta_0\sigma^2)^{1/4}\,T^{-1/4}\right). \tag{16}$$

When $\sigma = 0$, the alternative choice $\eta = 1/(2L)$ and $0 < \beta \le 1/2$ gives

$$\mathbb{E}\left\|\nabla F(x_I)\right\| = O(\sqrt{L\Delta_0}/\sqrt{T}).$$

Remark. When $\sigma > 0$, the $O(T^{-1/4})$ rate matches SGD (Ghadimi & Lan, 2013) and is optimal under standard smoothness (Arjevani et al., 2023). The theorem also allows any fixed $\beta \in (0, 1/2]$.

Core of the analysis. Define the estimation error $e_t = d_t - \nabla F(x_t)$ and $e_{-1} = 0$. The error obeys $e_t = \beta e_{t-1} + \nu_t$, with $\nu_t = g(x_t,\xi_t) - \nabla F(x_t) - \beta\big(g(x_{t-1},\xi_t) - \nabla F(x_{t-1})\big)$ and $\nu_0 = e_0$. Note that ν_t has conditional mean zero and second moment at most $(1+\beta)^2\sigma^2$. Although e_t itself is conditionally biased, we can define the auxiliary iterations $z_t = x_t - \eta\beta e_{t-1}/(1-\beta)$ which satisfy

$$z_{t+1} = z_t - \eta\nabla F(x_t) - \frac{\eta}{1-\beta}\nu_t.$$

This removes the inherited error from the update. Standard smoothness controls the difference between $\nabla F(x_t)$ and $\nabla F(z_t)$, such that $\left\|\nabla F(x_t) - \nabla F(z_t)\right\|^2 \le L^2\left\|x_t - z_t\right\|^2 \le \frac{L^2\eta^2\beta^2}{(1-\beta)^2}\left\|e_{t-1}\right\|^2$. The shift thus has expected squared size $O(\eta^2\sigma^2)$ for fixed β. After deriving a descent argument, we can establish the bound (15). This argument uses the smoothness of F only, and it never bounds a difference of sampled gradients by the distance between the query points.

5.2 CONVEX OBJECTIVES

We now use convexity to obtain function-gap guarantees without the average smoothness assumption. The key observation is that standard smoothness and convexity already provide useful stability properties for the objective gradient map. These properties allow us to compare STORM with a carefully designed auxiliary sequence, even when average smoothness is not assumed.

Lemma 4. *Under Assumptions 1 and 5, for every $0 \le \eta \le 1/L$ and all $x, y \in \mathbb{R}^d$,*

$$\left\|x - \eta\nabla F(x) - y + \eta\nabla F(y)\right\| \le \left\|x - y\right\|. \tag{17}$$

If strong convexity (Assumption 6) also holds, then

$$\|x - \eta\nabla F(x) - y + \eta\nabla F(y)\| \le (1 - \lambda\eta)\|x - y\|. \tag{18}$$

Under Assumption 1 alone, even without convexity, we have

$$\|\nabla F(y)\|^2 \le 2L(F(y) - F^*), \tag{19}$$

$$F(x) - F^* \le 2(F(y) - F^*) + L\|x - y\|^2. \tag{20}$$

Remark. Inequality (17) shows that a gradient step would not increase the distance between two points. Under strong convexity, (18) can further give a contraction. Both statements concern the exact gradient ∇F and therefore require no regularity of individual stochastic gradients, avoiding the requirement of average smoothness. Once the distance between two trajectories is controlled, the inequality (20) transfers a function-gap guarantee from one trajectory to the other.

Comparison with an auxiliary sequence. Let $\beta = 1 - a \in [0, 1/2]$. We compare STORM with an auxiliary sequence that uses exact objective gradients and the centered increments. This sequence is only used for analysis and requires no real computation.

Lemma 5. *Under Assumptions 1, 2, and 5, let $0 \le \beta \le 1/2$ and choose $0 < \eta_t \le 1/L$. Write $e_t = d_t - \nabla F(x_t)$ and $q_t = \mathbb{E}\|e_t\|^2$. Set $e_{-1} = 0$ and $\nu_0 = e_0$, and define*

$$\nu_t = \zeta(x_t, \xi_t) - \beta\zeta(x_{t-1}, \xi_t), \qquad e_t = \beta e_{t-1} + \nu_t, \qquad \zeta(x, \xi) = g(x, \xi) - \nabla F(x). \tag{21}$$

For $\mathcal{F}_t = \sigma(\xi_0, \dots, \xi_{t-1})$, we have

$$\mathbb{E}[\nu_t \mid \mathcal{F}_t] = 0, \qquad \mathbb{E}[\|\nu_t\|^2 \mid \mathcal{F}_t] \le S^2 := (1 + \beta)^2\sigma^2, \qquad q_t \le 3\sigma^2. \tag{22}$$

The auxiliary sequence $y_0 = x_0$,

$$y_{t+1} = y_t - \eta_t\big(\nabla F(y_t) + \nu_t\big)$$

satisfies

$$\sqrt{\mathbb{E}\|x_t - y_t\|^2} \le \sqrt{3}\beta\sigma\sum_{j<t}\eta_j. \tag{23}$$

Under Assumption 6, this improves to

$$\sqrt{\mathbb{E}\|x_t - y_t\|^2} \le \sqrt{3}\beta\sigma/\lambda. \tag{24}$$

Core of the analysis. To see why the comparison works, subtract the two updates:

$$x_{t+1} - y_{t+1} = \big(x_t - \eta_t\nabla F(x_t)\big) - \big(y_t - \eta_t\nabla F(y_t)\big) - \eta_t\beta e_{t-1}.$$

The shared increment ν_t cancels. Applying (17) and then the triangle inequality in L^2 gives

$$\sqrt{\mathbb{E}\|x_{t+1} - y_{t+1}\|^2} \le \sqrt{\mathbb{E}\|x_t - y_t\|^2} + \eta_t\beta\sqrt{q_{t-1}}.$$

Since $x_0 = y_0$ and $q_{t-1} \le 3\sigma^2$, summation gives (23). Under strong convexity, inequality (18) instead multiplies the first term on the right by $1 - \lambda\eta_t$. With $\mathbb{E}\|e_t\|^2 \le 3\sigma^2$, this gives

$$\sqrt{\mathbb{E}\|x_{t+1} - y_{t+1}\|^2} \le (1 - \lambda\eta_t)\sqrt{\mathbb{E}\|x_t - y_t\|^2} + \sqrt{3}\,\beta\sigma\eta_t,$$

which further yields (24).

Remark. Thus the inherited estimator error affects the trajectory comparison through the smaller quantity βe_{t-1}. Combining (23) with (20) gives

$$\mathbb{E}[F(x_t) - F^*] \le 2\mathbb{E}[F(y_t) - F^*] + 3L\beta^2\sigma^2\left(\sum_{j<t}\eta_j\right)^2.$$

Choosing $\beta = O(1/(T + 1))$ makes this additional error compatible with the desired rates.

Stepsize blocks. Note that we also want to derive the rates for the last iterate output, and an average guarantee does not by itself control $F(x_T)$. We therefore arrange the stepsizes so that a good value in an early block controls values in the following shorter block. The blocks shrink geometrically toward T, and the stepsize is halved at each boundary. Larger early steps reduce the initial error and smaller late steps limit the fresh noise injected near the returned point. For any integer $T \geq 1$, define

$$m = \lfloor \log_2(T+1) \rfloor, \qquad U_i = T + 1 - \left\lfloor (T+1)/2^i \right\rfloor \ (0 \leq i \leq m), \qquad U_{m+1} = T+1. \quad (25)$$

Each block $\{U_i, \dots, U_{i+1} - 1\}$ is nonempty, and the last is $\{T\}$. For example, when $T = 12$, their lengths are $7, 3, 2, 1$. The construction partitions the $T+1$ output, and it does not restart the estimator. To see how the blocks are compared, write $h_t = \nabla F(y_t) + \nu_t$ and fix indices $p \leq q$. Expanding $\|y_{t+1} - y_p\|^2$, taking expectations, and then using $\langle \nabla F(y_t), y_t - y_p \rangle \geq F(y_t) - F(y_p)$ gives

$$2\sum_{t=p}^{q} \eta_t \mathbb{E}[F(y_t) - F(y_p)] \leq \sum_{t=p}^{q} \eta_t^2 \mathbb{E}\,\|h_t\|^2 . \quad (26)$$

Lemma 4 is useful again here. Conditional centering and (19) imply

$$\mathbb{E}\,\|h_t\|^2 = \mathbb{E}\,\|\nabla F(y_t)\|^2 + \mathbb{E}\,\|\nu_t\|^2 \leq 2L\,\mathbb{E}[F(y_t) - F^*] + S^2.$$

Thus the right-hand side of (26) can be controlled by function errors and the noise variance. The stepsize restriction allows the function-error terms to be absorbed into the resulting estimates, without assuming globally bounded stochastic gradients.

We then compare successive blocks using a reference point from the preceding block. Halving the stepsizes makes the accumulated comparison errors summable, and the final singleton block gives a last-iterate guarantee.

Theorem 6. *Under Assumptions 1, 2, and 5, for every integer $T \geq 1$, choose*

$$\beta = \frac{1}{T+2}, \qquad \eta_t = 2^{-i}\eta \quad (U_i \leq t < U_{i+1}), \qquad 0 < \eta \leq \frac{1}{32L}. \quad (27)$$

Define $\bar{x}_T^\eta = \sum_{t=0}^{T} \eta_t x_t / \sum_{t=0}^{T} \eta_t$. Then

$$\mathbb{E}[F(\bar{x}_T^\eta) - F^*] \leq \frac{4R^2}{\eta(T+1)} + 5\eta\sigma^2. \quad (28)$$

Both $z_T = x_T$ and $z_T = (T+1)^{-1}\sum_{t=0}^{T} x_t$ obey

$$\mathbb{E}[F(z_T) - F^*] \leq \frac{8\mathrm{e}R^2}{\eta(T+1)} + 82\mathrm{e}\eta\sigma^2. \quad (29)$$

For $R, \sigma > 0$, take $\eta = \min\{1/(32L),\ 2R/(\sqrt{41}\sigma\sqrt{T+1})\}$. All three outputs $\widehat{x}_T$ satisfy

$$\mathbb{E}[F(\widehat{x}_T) - F^*] \leq \mathcal{O}\left(\frac{LR^2}{T} + \frac{\sigma R}{\sqrt{T}}\right). \quad (30)$$

If $\sigma = 0$, use $\eta = 1/(32L)$ to obtain

$$\mathbb{E}[F(\widehat{x}_T) - F^*] \leq \mathcal{O}\left(\frac{LR^2}{T}\right). \quad (31)$$

Remark. The dominant term $\sigma R/\sqrt{T}$ matches the SGD guarantees and the lower bound for convex objectives (Agarwal et al., 2012; Stich, 2019; Liu & Zhou, 2024). Compared with previous analysis of STORM, our rates remove the additional average smoothness assumption. Also note that we also establish a last-iterate bound in our analysis.

5.3 Strongly Convex Objectives

Under λ-strong convexity, the companion distance contracts. With $\beta = 1/(T+s+1)$, the function-gap comparison takes the form

$$\mathbb{E}[F(x_t) - F^*] \leq 2\mathbb{E}[F(y_t) - F^*] + \frac{3\kappa\sigma^2}{\lambda(T+s+1)^2}.$$

Thus the error introduced by transferring the companion guarantee back to STORM can be absorbed into the desired $O(\sigma^2/(\lambda T))$ stochastic term. For a carefully designed weighted mean, this permits exact cancellation of consecutive distance terms. For the last point, we combine this contraction with the rounded blocks in (25). The two results use different schedules.

Theorem 7 (Weighted mean)**.** *Under Assumptions 1, 2, and 6, for any $s \geq 4\kappa$, use*

$$\beta = \frac{1}{T+s+1}, \qquad \eta_t = \frac{2}{\lambda(t+s+1)}, \qquad \bar{x}_T = \frac{2\sum_{t=0}^{T}(t+s)x_t}{(T+1)(T+2s)}. \tag{32}$$

Then, we can ensure that

$$\mathbb{E}[F(\bar{x}_T) - F^*] \leq \frac{2\lambda s(s-1)R^2}{(T+1)(T+2s)} + \frac{19\sigma^2}{\lambda(T+s+1)}. \tag{33}$$

The same bound holds for x_I, where I has probabilities proportional to $t+s$.

Next, we present the guarantee for the last iterate below.

Theorem 8 (Last iterate)**.** *Under Assumptions 1, 2, and 6, let $T \geq 1$ be any integer, $s \geq 4\kappa^2$, and $\beta = 1/(T+s+1)$. For the rounded blocks* (25)*, use*

$$\eta_t = 2^{-i}\frac{2}{\lambda(t+s+1)} \quad (U_i \leq t < U_{i+1}). \tag{34}$$

After T primal updates, the last iterate satisfies

$$\mathbb{E}[F(x_T) - F^*] \leq \frac{288\lambda(s+1)^2R^2}{(T+1)^2} + \frac{2451\sigma^2}{\lambda(T+1)}. \tag{35}$$

Remark. The dominant term $O(\sigma^2/(\lambda T))$ of the last iterate bound has the same order as the mean, which is optimal for strongly convex functions (Agarwal et al., 2012). Compared with previous results, we remove the average smoothness assumption and establish a last-iterate bound.

6 Numerical Experiments

We study three questions suggested by the theory: how the objective geometry affects convergence, whether STORM converges when average smoothness fails, and how the prescribed schedules behave for averaged and last-iterate outputs. Our contribution is an analysis of the original STORM algorithm and its empirical performance has already been studied extensively in previous literature, so we focus on experiments that isolate these questions instead of repeating a broad benchmark evaluation.

Experimental setup. We use the following four objectives with known minima and exact gradients. An orthogonal change of coordinates $z = Qx$ gives

$$F_{\mathrm{nc}}(x) = \sum_{i=1}^{10}\left(1 - \cos z_i + 0.05z_i^2\right), \qquad F_{\mathrm{PL}}(x) = \frac{1}{2}\sum_{i=1}^{10}(z_i + 2\tanh z_i)^2,$$

$$F_{\mathrm{cv}}(x) = \frac{1}{2}\sum_{i=1}^{10}\left(z_i^2 - \log(1+z_i^2)\right), \qquad F_{\mathrm{sc}}(x) = \frac{1}{2}\sum_{i=1}^{10}h_i z_i^2,$$

where h_i are equally spaced in $[1,4]$. For the average-smooth, we add independent standard Gaussian noise to the coordinates of the gradient. For the experiments without average smoothness, we use

$$g(x;J,\phi) = \nabla F(x) + Q^\top\zeta(Qx;J,\phi), \qquad \zeta_i(z;J,\phi) = \sqrt{2}\sin(2^{J_i}z_i + \phi_i), \tag{36}$$

where $\Pr(J_i = j) = 2^{-j-1}$ and ϕ_i is uniform on $[0, 2\pi]$. The oracle is unbiased and has $\sigma^2 = 10$, but admits no finite average-smoothness constant. We run 256 independent trials. All parameters are chosen from the stated analytical constants, without tuning hyperparameters.

Nonconvex and PL objectives under average smoothness. For F_{nc}, we use the exact schedule in Theorem 1 and measure $T^{-1}\sum_{t=0}^{T-1}\|\nabla F(x_t)\|^2$, whose reference order is $T^{-2/3}$. For F_{PL}, we compare the harmonic and two-phase schedules of Theorem 2 and report $F(x_T) - F^*$. Figure 1(a) is roughly consistent with the $T^{-2/3}$ reference. On the PL problem, the two-phase schedule gives a smaller gap at smaller budgets, whereas the harmonic schedule becomes better at the larger budgets.

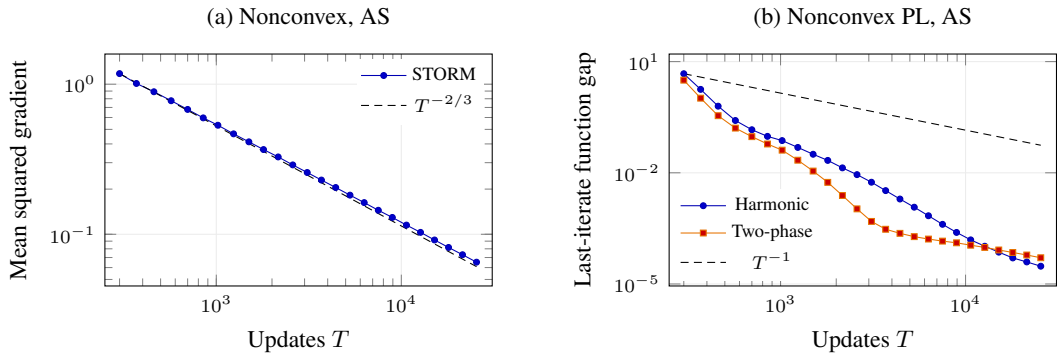

Figure 1: Convergence under average smoothness (log–log scales).

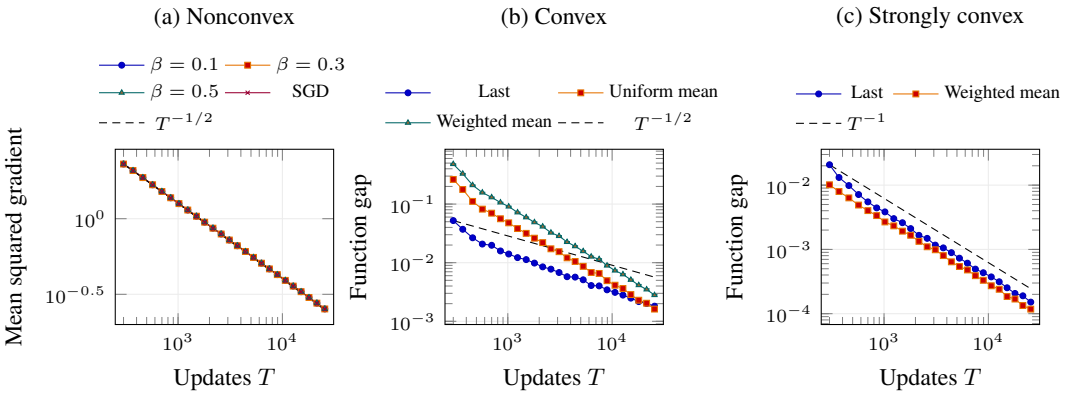

Figure 2: Convergence without average smoothness (log–log scales).

Nonconvex objectives without average smoothness. For F_{nc}, we replace the Gaussian oracle by (36). We test $\beta \in \{0.1, 0.3, 0.5\}$ with the common stepsize prescribed in Theorem 3. SGD uses the same stepsize as a reference. We measure the average squared gradient. Here, the theoretical stochastic order is $T^{-1/2}$, corresponding to $T^{-1/4}$ in gradient norm. In Figure 2(a), increasing T by a factor of 85 reduces the squared-gradient criterion by about 9.1 for each tested correction coefficient, close to $\sqrt{85} \approx 9.2$. The three fixed-β curves and SGD are nearly indistinguishable at this scale.

Averaged and last iterates without average smoothness. For F_{cv}, we use the rounded block schedule in Theorem 6 and evaluate the last iterate, the uniform mean, and the stepsize-weighted mean of the same trajectory. For F_{sc}, we run the two schedules in Theorems 7 and 8 separately. In Figure 2(b), the uniform mean is slightly better than the last iterate at the largest tested budget. For strong convexity, both curves in Figure 2(c) follow the T^{-1} reference closely at larger budgets.

Also note that our theoretical guarantees are worst-case upper bounds. A particular objective can converge faster than the worst-case rate in practice.

7 Conclusion

In this paper, we studied the convergence of STORM under different objective geometries and smoothness assumptions. Under average smoothness, we provided a direct analysis of the logarithm-free nonconvex rate and established a last-iterate guarantee with an optimal stochastic term under the PL condition. Under standard smoothness alone, STORM matches the nonconvex SGD rate and attains optimal stochastic orders for both averaged and last iterates in the convex and strongly convex settings. These guarantees follow from different parameter choices within the same STORM recursion, with a single initial sample and a constant number of gradient evaluations per iteration.

AI USE STATEMENT

In this work, we used AI tools mainly for polishing writing, assisting in the writing of proofs, and designing experiments. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Alekh Agarwal, Peter L. Bartlett, Pradeep Ravikumar, and Martin J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

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

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems*, 2019.

Tehila Dahan and Kfir Y. Levy. Do stochastic, feel noiseless: Stable stochastic optimization via a double momentum mechanism. In *Proceedings of the 13th International Conference on Learning Representations*, 2025.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, 2018.

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

Wei Jiang, Bokun Wang, Yibo Wang, Lijun Zhang, and Tianbao Yang. Optimal algorithms for stochastic multi-level compositional optimization. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 10195–10216, 2022.

Wei Jiang, Sifan Yang, Yibo Wang, and Lijun Zhang. Adaptive variance reduction for stochastic optimization under weaker assumptions. In *Advances in Neural Information Processing Systems*, 2024.

Kfir Levy, Ali Kavis, and Volkan Cevher. STORM+: Fully adaptive SGD with recursive momentum for nonconvex optimization. In *Advances in Neural Information Processing Systems*, 2021.

Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtárik. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 6286–6295, 2021.

Zijian Liu and Zhengyuan Zhou. Revisiting the last-iterate convergence of stochastic gradient methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=xxaEhwC1I4`.

Lam M. Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 2613–2621, 2017.

Qi Qi, Zhishuai Guo, Yi Xu, Rong Jin, and Tianbao Yang. An online method for a class of distributionally robust optimization with non-convex objectives. In *Advances in Neural Information Processing Systems*, 2021.

Sebastian U. Stich. Unified optimal analysis of the (Stochastic) gradient method. *arXiv preprint arXiv:1907.04232*, 2019.

Mingrui Zhang, Zebang Shen, Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. One sample stochastic Frank–Wolfe. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pp. 4012–4023, 2020.

A PROOF OF THEOREM 1

We prove Theorem 1 by combining the descent inequality with a bound on the gradient estimation error. Let $\mathcal{F}_t = \sigma(\xi_0, \dots, \xi_{t-1})$. Define

$$e_t = d_t - \nabla F(x_t), \qquad q_t = \mathbb{E}\left\|e_t\right\|^2, \qquad \Delta_t = \mathbb{E}[F(x_t) - F^*].$$

Thus $\Delta_0 = F(x_0) - F^*$ and $q_0 \le \sigma^2$.

Lemma 9. *Under Assumptions 1 and 2, for every deterministic $\eta_t \ge 0$, the primal update satisfies*

$$\Delta_{t+1} - \Delta_t \le -\frac{\eta_t}{2}\mathbb{E}\left\|\nabla F(x_t)\right\|^2 - \frac{\eta_t(1 - L\eta_t)}{2}\mathbb{E}\left\|d_t\right\|^2 + \frac{\eta_t}{2}q_t. \tag{37}$$

If Assumption 3 also holds and $0 \le a_{t+1} \le 1$, then

$$q_{t+1} \le (1 - a_{t+1})^2 q_t + 2a_{t+1}^2\sigma^2 + 2\mathcal{L}^2\eta_t^2\mathbb{E}\left\|d_t\right\|^2. \tag{38}$$

Proof. Smoothness and $x_{t+1} = x_t - \eta_t d_t$ give

$$F(x_{t+1}) \le F(x_t) - \eta_t \left\langle \nabla F(x_t), d_t \right\rangle + \frac{L\eta_t^2}{2}\left\|d_t\right\|^2.$$

By the definition of e_t,

$$\left\|e_t\right\|^2 = \left\|d_t\right\|^2 + \left\|\nabla F(x_t)\right\|^2 - 2\left\langle \nabla F(x_t), d_t \right\rangle.$$

Solving this identity for the inner product and substituting yields

$$\begin{aligned} F(x_{t+1}) - F(x_t) &\le -\frac{\eta_t}{2}\left\|\nabla F(x_t)\right\|^2 - \frac{\eta_t}{2}\left\|d_t\right\|^2 + \frac{\eta_t}{2}\left\|e_t\right\|^2 + \frac{L\eta_t^2}{2}\left\|d_t\right\|^2 \\ &= -\frac{\eta_t}{2}\left\|\nabla F(x_t)\right\|^2 - \frac{\eta_t(1 - L\eta_t)}{2}\left\|d_t\right\|^2 + \frac{\eta_t}{2}\left\|e_t\right\|^2. \end{aligned}$$

Taking expectations proves (37).

To prove the error recursion, abbreviate $a = a_{t+1}$ and write $\zeta(x, \xi) = g(x, \xi) - \nabla F(x)$. Subtracting $\nabla F(x_{t+1})$ from the estimator update gives the exact identity

$$\begin{aligned} e_{t+1} &= (1 - a)e_t + \zeta(x_{t+1}, \xi_{t+1}) - (1 - a)\zeta(x_t, \xi_{t+1}) \\ &= (1 - a)e_t + a\zeta(x_{t+1}, \xi_{t+1}) + (1 - a)\big(\zeta(x_{t+1}, \xi_{t+1}) - \zeta(x_t, \xi_{t+1})\big). \end{aligned}$$

Conditional on $\mathcal{F}_{t+1}$, the fresh-sample terms together have mean zero, so their cross term with the fixed e_t vanishes. We then use $\left\|u + v\right\|^2 \le 2\left\|u\right\|^2 + 2\left\|v\right\|^2$ to obtain

$$\begin{aligned} \mathbb{E}[\left\|e_{t+1}\right\|^2 \mid \mathcal{F}_{t+1}] \le (1 - a)^2\left\|e_t\right\|^2 &+ 2a^2\mathbb{E}[\left\|\zeta(x_{t+1}, \xi_{t+1})\right\|^2 \mid \mathcal{F}_{t+1}] \\ &+ 2(1 - a)^2\mathbb{E}\big[\left\|\zeta(x_{t+1}, \xi_{t+1}) - \zeta(x_t, \xi_{t+1})\right\|^2 \mid \mathcal{F}_{t+1}\big]. \end{aligned}$$

The first conditional noise moment is at most σ^2 by (5) and fresh-sample independence. For the second one, centering a random vector subtracts the squared norm of its mean:

$$\begin{aligned} &\mathbb{E}\big[\left\|\zeta(x_{t+1}, \xi_{t+1}) - \zeta(x_t, \xi_{t+1})\right\|^2 \mid \mathcal{F}_{t+1}\big] \\ &\quad = \mathbb{E}\big[\left\|g(x_{t+1}, \xi_{t+1}) - g(x_t, \xi_{t+1})\right\|^2 \mid \mathcal{F}_{t+1}\big] - \left\|\nabla F(x_{t+1}) - \nabla F(x_t)\right\|^2 \\ &\quad \le \mathcal{L}^2\left\|x_{t+1} - x_t\right\|^2 = \mathcal{L}^2\eta_t^2\left\|d_t\right\|^2. \end{aligned}$$

The inequality uses (6). Substitution gives

$$\mathbb{E}[\left\|e_{t+1}\right\|^2 \mid \mathcal{F}_{t+1}] \le (1 - a)^2\left\|e_t\right\|^2 + 2a^2\sigma^2 + 2(1 - a)^2\mathcal{L}^2\eta_t^2\left\|d_t\right\|^2.$$

Finally, $(1 - a)^2 \le 1$ gives (38). □

Lemma 10. *Under Assumptions 1, 2, and 3, let $\eta_t = \eta > 0$ and $L\eta \le 1/2$. Choose an auxiliary weight $a_0 > 0$ and deterministic coefficients satisfying*

$$0 < a_{t+1} \le \frac{1}{2}, \qquad a_{t+1} \ge 4\mathcal{L}^2\eta^2, \qquad a_{t+1}^{-1} - a_t^{-1} \le \frac{1}{2}.$$

With

$$\Psi_t = \Delta_t + \frac{\eta}{2a_t}q_t,$$

one has

$$\Psi_{t+1} \le \Psi_t - \frac{\eta}{2}\mathbb{E}\left\|\nabla F(x_t)\right\|^2 + \eta a_{t+1}\sigma^2. \tag{39}$$

Proof. Multiply (38) by $\eta/(2a_{t+1})$ and add it to (37). This gives

$$\Psi_{t+1} \le \Delta_t - \frac{\eta}{2}\mathbb{E}\left\|\nabla F(x_t)\right\|^2 + \left(\frac{\eta}{2} + \frac{\eta(1-a_{t+1})^2}{2a_{t+1}}\right) q_t + \left(-\frac{\eta(1-L\eta)}{2} + \frac{\mathcal{L}^2\eta^3}{a_{t+1}}\right)\mathbb{E}\left\|d_t\right\|^2 + \eta a_{t+1}\sigma^2.$$

Subtract $\Psi_t = \Delta_t + \eta q_t/(2a_t)$. The remaining coefficient of q_t is

$$\begin{aligned}\frac{\eta}{2} + \frac{\eta(1-a_{t+1})^2}{2a_{t+1}} - \frac{\eta}{2a_t} &= \frac{\eta}{2}\left(1 + a_{t+1}^{-1} - 2 + a_{t+1} - a_t^{-1}\right) \\ &= \frac{\eta}{2}\left(a_{t+1}^{-1} - a_t^{-1} - 1 + a_{t+1}\right) \\ &\le \frac{\eta}{2}\left(\frac{1}{2} - 1 + \frac{1}{2}\right) = 0.\end{aligned}$$

The coefficient of $\mathbb{E}\left\|d_t\right\|^2$ is also nonpositive:

$$-\frac{\eta(1-L\eta)}{2} + \frac{\mathcal{L}^2\eta^3}{a_{t+1}} \le -\frac{\eta}{4} + \frac{\mathcal{L}^2\eta^3}{4\mathcal{L}^2\eta^2} = 0.$$

Dropping these two nonpositive terms proves (39). □

Proof of Theorem 1. Fix an arbitrary integer $T \ge 1$ and use:

$$\eta = \frac{1}{2\mathcal{L}(T+2)^{1/3}}, \qquad a_t = (t+2)^{-2/3}.$$

We first verify the conditions of Lemma 10 for $t = 0, \dots, T-1$. Since $L \le \mathcal{L}$,

$$L\eta \le \frac{1}{2(T+2)^{1/3}} \le \frac{1}{2}.$$

The mixing sequence decreases, and hence

$$a_{t+1} \le a_1 = 3^{-2/3} < \frac{1}{2}, \qquad a_{t+1} \ge a_T = (T+2)^{-2/3} = 4\mathcal{L}^2\eta^2.$$

Because $u \mapsto u^{2/3}$ is concave, its unit increments decrease as u increases. Therefore

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

Thus, all conditions of Lemma 10 hold. Rearrange (39) and sum over the T steps:

$$\begin{aligned}\frac{\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla F(x_t)\right\|^2 &\le \sum_{t=0}^{T-1}(\Psi_t - \Psi_{t+1}) + \eta\sigma^2\sum_{t=0}^{T-1} a_{t+1} \\ &= \Psi_0 - \Psi_T + \eta\sigma^2\sum_{t=0}^{T-1} a_{t+1} \\ &\le \Delta_0 + \frac{\eta\sigma^2}{2a_0} + \eta\sigma^2\sum_{t=0}^{T-1} a_{t+1}.\end{aligned}$$

Here $\Psi_T \ge 0$ because $F(x_T) \ge F^*$ and $q_T \ge 0$, and the initial variance bound gives $q_0 \le \sigma^2$. Dividing by $\eta T/2$ yields

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla F(x_t)\right\|^2 \le \frac{2\Delta_0}{\eta T} + \frac{\sigma^2}{a_0 T} + \frac{2\sigma^2}{T}\sum_{t=0}^{T-1}(t+3)^{-2/3}.$$

Since $u^{-2/3}$ is decreasing, integrating over each interval $[t+2, t+3]$ gives

$$\begin{aligned}\sum_{t=0}^{T-1}(t+3)^{-2/3} &\le \sum_{t=0}^{T-1}\int_{t+2}^{t+3} u^{-2/3}\,du \\ &= \int_{2}^{T+2} u^{-2/3}\,du \\ &= 3\big((T+2)^{1/3}-2^{1/3}\big).\end{aligned}$$

Substitute this bound, $\eta^{-1}=2\mathcal{L}(T+2)^{1/3}$, and $a_0^{-1}=2^{2/3}$ to obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla F(x_t)\right\|^2 \le \frac{4\mathcal{L}\Delta_0(T+2)^{1/3}}{T}+\frac{2^{2/3}\sigma^2}{T}+\frac{6\sigma^2}{T}\big((T+2)^{1/3}-2^{1/3}\big),$$

which is (10).

Finally, let I be uniform on $\{0,\dots,T-1\}$ and independent of the trajectory. Then

$$\mathbb{E}\left\|\nabla F(x_I)\right\|^2 = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla F(x_t)\right\|^2.$$

Cauchy–Schwarz gives $\mathbb{E}\left\|\nabla F(x_I)\right\| \le (\mathbb{E}\left\|\nabla F(x_I)\right\|^2)^{1/2}$. Since $T+2\le 3T$ for $T\ge 1$, the displayed second-moment bound is $O((\mathcal{L}\Delta_0+\sigma^2)T^{-2/3})$, proving the stated expected-norm rate. □

The noiseless case.

If $\sigma=0$, take $\eta=1/(4\mathcal{L})$ and $a_t=1/4$ for every $t\ge 0$, including the potential weight a_0. Then $L\eta\le 1/4$, $a_{t+1}=4\mathcal{L}^2\eta^2=1/4$, and $a_{t+1}^{-1}-a_t^{-1}=0$, so the certificate applies. Since $q_0=0$ and the noise term vanishes, summing it gives

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla F(x_t)\right\|^2 \le \frac{2\Delta_0}{\eta T}=\frac{8\mathcal{L}\Delta_0}{T}.$$

B Proof of Theorem 2

Proof. Retain the notation $e_t=d_t-\nabla F(x_t)$, $q_t=\mathbb{E}\left\|e_t\right\|^2$, and $\Delta_t=\mathbb{E}[F(x_t)-F^*]$. The potential in this proof has a constant estimator-error weight:

$$V_t=\Delta_t+\frac{q_t}{\mu}.$$

Also note that $\Delta_t\le V_t$ and $V_0\le\Delta_0+\sigma^2/\mu$.

First, consider any one step satisfying

$$a_{t+1}=\mu\eta_t,\qquad 0<\mu\eta_t\le\frac{1}{2},\qquad \eta_t\le\min\left\{\frac{1}{2L},\frac{\mu}{8\mathcal{L}^2}\right\}.$$

Add (38) divided by μ to (37). The resulting inequality is

$$\begin{aligned}V_{t+1}\le\Delta_t-\frac{\eta_t}{2}\mathbb{E}\left\|\nabla F(x_t)\right\|^2+\left(\frac{(1-\mu\eta_t)^2}{\mu}+\frac{\eta_t}{2}\right)q_t \\ +\left(-\frac{\eta_t(1-L\eta_t)}{2}+\frac{2\mathcal{L}^2\eta_t^2}{\mu}\right)\mathbb{E}\left\|d_t\right\|^2+2\mu\eta_t^2\sigma^2.\end{aligned}$$

By the PL inequality,

$$\Delta_t-\frac{\eta_t}{2}\mathbb{E}\left\|\nabla F(x_t)\right\|^2\le(1-\mu\eta_t)\Delta_t.$$

For the coefficient of q_t/μ, direct expansion gives

$$\begin{aligned}(1-\mu\eta_t)^2+\frac{\mu\eta_t}{2} &= 1-\frac{3}{2}\mu\eta_t+(\mu\eta_t)^2\\ &= 1-\mu\eta_t+\mu\eta_t\left(\mu\eta_t-\frac{1}{2}\right)\\ &\le 1-\mu\eta_t.\end{aligned}$$

Finally, $L\eta_t \le 1/2$ and $\eta_t \le \mu/(8\mathcal{L}^2)$ imply

$$-\frac{\eta_t(1-L\eta_t)}{2}+\frac{2\mathcal{L}^2\eta_t^2}{\mu} \le -\frac{\eta_t}{4}+\eta_t\frac{2\mathcal{L}^2}{\mu}\frac{\mu}{8\mathcal{L}^2}=0.$$

Since $\mathbb{E}\left\|d_t\right\|^2 \ge 0$, that term can be discarded. Combining the other two coefficients gives

$$V_{t+1} \le (1-\mu\eta_t)V_t + 2\mu\eta_t^2\sigma^2. \tag{40}$$

For either schedule in the theorem, the largest stepsize is $2/(\mu s)$. Since $s \ge 16(\mathcal{L}/\mu)^2$,

$$\frac{2}{\mu s} \le \frac{\mu}{8\mathcal{L}^2}, \qquad \frac{2}{s} \le \frac{\mu^2}{8\mathcal{L}^2} \le \frac{1}{8}, \qquad L\frac{2}{\mu s} \le \frac{L\mu}{8\mathcal{L}^2} \le \frac{1}{8}.$$

Thus (40) holds at every step of either schedule. In particular, $s \ge 16$, so the scalar factors used below are positive.

The first schedule. Substituting $\eta_t = 2/[\mu(t+s)]$ into (40) gives

$$V_{t+1} \le \frac{t+s-2}{t+s}V_t + \frac{8\sigma^2}{\mu(t+s)^2}. \tag{41}$$

The first two substitutions are

$$\begin{aligned}V_1 &\le \frac{s-2}{s}V_0+\frac{8\sigma^2}{\mu s^2},\\ V_2 &\le \frac{s-1}{s+1}\frac{s-2}{s}V_0+\frac{8\sigma^2}{\mu}\left(\frac{s-1}{(s+1)s^2}+\frac{1}{(s+1)^2}\right).\end{aligned}$$

Induction therefore gives

$$V_T \le V_0\prod_{k=0}^{T-1}\frac{k+s-2}{k+s}+\frac{8\sigma^2}{\mu}\sum_{j=0}^{T-1}\frac{1}{(j+s)^2}\prod_{k=j+1}^{T-1}\frac{k+s-2}{k+s}. \tag{42}$$

For every $0 \le u \le T$,

$$\prod_{k=u}^{T-1}\frac{k+s-2}{k+s}=\frac{(u+s-2)(u+s-1)}{(T+s-2)(T+s-1)}. \tag{43}$$

Taking $u=0$ gives the initialization product. Taking $u=j+1$ in each noise term gives

$$\begin{aligned}\frac{1}{(j+s)^2}\prod_{k=j+1}^{T-1}\frac{k+s-2}{k+s} &= \frac{1}{(j+s)^2}\frac{(j+s-1)(j+s)}{(T+s-2)(T+s-1)}\\ &= \frac{(j+s-1)/(j+s)}{(T+s-2)(T+s-1)}.\end{aligned}$$

Consequently, (42) becomes the explicit bound

$$V_T \le \frac{(s-2)(s-1)V_0+\frac{8\sigma^2}{\mu}\sum_{j=0}^{T-1}\frac{j+s-1}{j+s}}{(T+s-2)(T+s-1)}. \tag{44}$$

We now bound the initialization and noise terms separately.

For $r = 1, 2$, positivity of the denominators and $rT \geq 0$ imply

$$\frac{s-r}{T+s-r} \leq \frac{s}{T+s}.$$

Indeed, after cross multiplication this is precisely $(s-r)(T+s) \leq s(T+s-r)$. Multiplying the two inequalities gives

$$\frac{(s-2)(s-1)}{(T+s-2)(T+s-1)} \leq \left(\frac{s}{T+s}\right)^2.$$

Every summand $(j+s-1)/(j+s)$ in the noise numerator is at most one. Furthermore,

$$(T+s-2)(T+s-1) - T(T+s) = (s-3)(T+s) + 2 \geq 0,$$

because $s \geq 16$. Hence

$$\frac{\sum_{j=0}^{T-1}(j+s-1)/(j+s)}{(T+s-2)(T+s-1)} \leq \frac{T}{(T+s-2)(T+s-1)} \leq \frac{1}{T+s}.$$

Combining the two bounds yields

$$V_T \leq \left(\frac{s}{T+s}\right)^2 V_0 + \frac{8\sigma^2}{\mu(T+s)}. \tag{45}$$

Using $\Delta_T \leq V_T$ and $V_0 \leq \Delta_0 + \sigma^2/\mu$ proves (12).

The two-phase schedule.

Now fix any integer $T \geq 1$, put $m = \lfloor T/2 \rfloor$, and use (13). During the first m updates, (40) gives

$$V_{t+1} \leq \left(1 - \frac{2}{s}\right) V_t + \frac{8\sigma^2}{\mu s^2}.$$

Repeated substitution gives

$$\begin{aligned} V_m &\leq \left(1 - \frac{2}{s}\right)^m V_0 + \frac{8\sigma^2}{\mu s^2} \sum_{j=0}^{m-1} \left(1 - \frac{2}{s}\right)^j \\ &= \left(1 - \frac{2}{s}\right)^m V_0 + \frac{4\sigma^2}{\mu s} \left[1 - \left(1 - \frac{2}{s}\right)^m\right] \\ &\leq e^{-2m/s} V_0 + \frac{4\sigma^2}{\mu s}. \end{aligned}$$

The final line uses $1 - u \leq e^{-u}$ and drops a factor between zero and one from the noise term.

There remain $n = T - m$ updates. Their parameters are $\eta_{m+j} = 2/[\mu(s+j)]$ for $j = 0, \ldots, n-1$. Applying the bound (45) with initial value V_m gives

$$V_T \leq \left(\frac{s}{s+n}\right)^2 V_m + \frac{8\sigma^2}{\mu(s+n)}.$$

The potential V_m already includes the estimation error q_m. Hence this argument applies directly to the remaining updates, without restarting x_m or d_m. Substitute the preceding bound for V_m to obtain

$$\begin{aligned} V_T &\leq \left(\frac{s}{s+n}\right)^2 e^{-2m/s} V_0 + \frac{4\sigma^2 s}{\mu(s+n)^2} + \frac{8\sigma^2}{\mu(s+n)} \\ &\leq e^{-2m/s} V_0 + \frac{12\sigma^2}{\mu(s+n)} \\ &\leq e^{-2m/s} V_0 + \frac{24\sigma^2}{\mu(T+2s)}. \end{aligned}$$

The second line uses $s/(s+n) \leq 1$ and the third uses $n = \lceil T/2 \rceil \geq T/2$. Finally, $\Delta_T \leq V_T$ and the initial bound on V_0 give (14). □

C Proof of Theorem 3

Throughout this section, $\eta_t = \eta$ and $a_{t+1} = 1-\beta$, where $0 < \eta \le 1/(2L)$ and $0 < \beta < 1$. The main step of the proof is to absorb the inherited estimation error into an auxiliary iterate. The remaining noise then has conditional mean zero and can be treated by the usual smoothness descent argument.

Proof. The estimation error has bounded second moment. Let $\mathcal{F}_t = \sigma(\xi_0, \dots, \xi_{t-1})$ denote the information available before sample ξ_t is drawn. Write $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t]$ and define

$$e_t = d_t - \nabla F(x_t), \qquad e_{-1} = 0.$$

For $t \ge 1$, subtract $\nabla F(x_t)$ from the estimator update:

$$\begin{aligned} e_t &= \beta e_{t-1} + \nu_t, \\ \nu_t &= g(x_t, \xi_t) - \nabla F(x_t) - \beta\big(g(x_{t-1}, \xi_t) - \nabla F(x_{t-1})\big). \end{aligned}$$

Set $\nu_0 = e_0$. The error recursion therefore holds also at $t = 0$:

$$e_t = \beta e_{t-1} + \nu_t, \qquad t \ge 0. \tag{46}$$

Both query points in ν_t are $\mathcal{F}_t$-measurable. Since ξ_t is independent of $\mathcal{F}_t$, oracle unbiasedness gives

$$\mathbb{E}_t \nu_t = 0. \tag{47}$$

The two oracle errors in ν_t use the same sample and need not be independent. We bound their sum by the conditional L^2 triangle inequality and then use (5):

$$\begin{aligned} \big(\mathbb{E}_t \|\nu_t\|^2\big)^{1/2} &\le \big(\mathbb{E}_t \|g(x_t, \xi_t) - \nabla F(x_t)\|^2\big)^{1/2} + \beta\big(\mathbb{E}_t \|g(x_{t-1}, \xi_t) - \nabla F(x_{t-1})\|^2\big)^{1/2} \\ &\le (1+\beta)\sigma. \end{aligned}$$

For $t = 0$, initialization gives the stronger bound $\mathbb{E}\|\nu_0\|^2 \le \sigma^2$. Thus, for every $t \ge 0$,

$$\mathbb{E}_t \|\nu_t\|^2 \le (1+\beta)^2 \sigma^2. \tag{48}$$

Squaring (46) and conditioning on $\mathcal{F}_t$ eliminates its cross term, because e_{t-1} is $\mathcal{F}_t$-measurable and (47) holds. Consequently,

$$\mathbb{E}\|e_t\|^2 \le \beta^2 \mathbb{E}\|e_{t-1}\|^2 + (1+\beta)^2\sigma^2.$$

Starting from $e_{-1} = 0$, repeated substitution gives

$$\mathbb{E}\|e_t\|^2 \le (1+\beta)^2\sigma^2 \sum_{j=0}^{t} \beta^{2j} \le \frac{(1+\beta)^2}{1-\beta^2}\sigma^2 = \frac{1+\beta}{1-\beta}\sigma^2. \tag{49}$$

An auxiliary iterate removes the inherited error. Define, only for the analysis,

$$z_t = x_t - \frac{\eta\beta}{1-\beta} e_{t-1}. \tag{50}$$

Then $z_0 = x_0$. The factor $\beta/(1-\beta)$ is chosen so that the error cancels from the following update:

$$\begin{aligned} z_{t+1} &= x_{t+1} - \frac{\eta\beta}{1-\beta} e_t \\ &= x_t - \eta\nabla F(x_t) - \eta e_t - \frac{\eta\beta}{1-\beta} e_t \\ &= x_t - \eta\nabla F(x_t) - \frac{\eta}{1-\beta} e_t \\ &= x_t - \frac{\eta\beta}{1-\beta} e_{t-1} - \eta\nabla F(x_t) - \frac{\eta}{1-\beta}\nu_t. \end{aligned}$$

The last equality uses (46). Hence

$$z_{t+1} = z_t - \eta\nabla F(x_t) - \frac{\eta}{1-\beta}\nu_t. \tag{51}$$

The update has centered noise, but evaluates the gradient at x_t rather than z_t. Standard smoothness will control this difference. Indeed, (50) gives the exact identity

$$\|x_t - z_t\|^2 = \frac{\eta^2\beta^2}{(1-\beta)^2}\|e_{t-1}\|^2.$$

No auxiliary sequence is computed by STORM.

Descent along the auxiliary sequence. Apply standard smoothness to (51):

$$F(z_{t+1}) \le F(z_t) - \eta\langle \nabla F(z_t), \nabla F(x_t) + \nu_t/(1-\beta)\rangle + \frac{L\eta^2}{2}\|\nabla F(x_t) + \nu_t/(1-\beta)\|^2.$$

Conditional on $\mathcal{F}_t$, $\mathbb{E}_t\nu_t = 0$. Expanding the squared norm before taking conditional expectations therefore yields

$$\mathbb{E}_t F(z_{t+1}) \le F(z_t) - \eta\langle \nabla F(z_t), \nabla F(x_t)\rangle + \frac{L\eta^2}{2}\|\nabla F(x_t)\|^2 + \frac{L\eta^2}{2(1-\beta)^2}\mathbb{E}_t\|\nu_t\|^2.$$

To express the decrease in terms of the gradient at the actual iterate, use $\langle u, v\rangle \le \|u\|^2 + \|v\|^2/4$:

$$\begin{aligned}
-\langle \nabla F(z_t), \nabla F(x_t)\rangle &= -\|\nabla F(x_t)\|^2 + \langle \nabla F(x_t) - \nabla F(z_t), \nabla F(x_t)\rangle \\
&\le -\frac{3}{4}\|\nabla F(x_t)\|^2 + \|\nabla F(x_t) - \nabla F(z_t)\|^2 \\
&\le -\frac{3}{4}\|\nabla F(x_t)\|^2 + \frac{L^2\eta^2\beta^2}{(1-\beta)^2}\|e_{t-1}\|^2.
\end{aligned}$$

Since $L\eta \le 1/2$, the coefficient of the squared gradient is at most $-\eta(3/4 - L\eta/2) \le -\eta/2$. We obtain

$$\mathbb{E}_t F(z_{t+1}) \le F(z_t) - \frac{\eta}{2}\|\nabla F(x_t)\|^2 + \frac{L^2\eta^3\beta^2}{(1-\beta)^2}\|e_{t-1}\|^2 + \frac{L\eta^2(1+\beta)^2}{2(1-\beta)^2}\sigma^2.$$

Taking full expectations and using (49), which also bounds $\mathbb{E}\|e_{-1}\|^2 = 0$, gives

$$\mathbb{E}F(z_{t+1}) \le \mathbb{E}F(z_t) - \frac{\eta}{2}\mathbb{E}\|\nabla F(x_t)\|^2 + \frac{L^2\eta^3\beta^2(1+\beta)}{(1-\beta)^3}\sigma^2 + \frac{L\eta^2(1+\beta)^2}{2(1-\beta)^2}\sigma^2. \tag{52}$$

Summation and the choice of parameters. Rearrange (52) and sum over $t = 0, \dots, T-1$. The function values telescope, so

$$\frac{\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(x_t)\|^2 \le F(z_0) - \mathbb{E}F(z_T) + \frac{L^2\eta^3T\beta^2(1+\beta)}{(1-\beta)^3}\sigma^2 + \frac{L\eta^2T(1+\beta)^2}{2(1-\beta)^2}\sigma^2.$$

Since $z_0 = x_0$ and $F(z_T) \ge F^*$, division by $\eta T/2$ gives the more precise bound

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(x_t)\|^2 \le \frac{2\Delta_0}{\eta T} + \frac{(1+\beta)^2}{(1-\beta)^2}L\eta\sigma^2 + \frac{2\beta^2(1+\beta)}{(1-\beta)^3}L^2\eta^2\sigma^2. \tag{53}$$

The condition $2L\eta \le 1$ bounds the last term by $\beta^2(1+\beta)L\eta\sigma^2/(1-\beta)^3$. Combining it with the preceding noise term uses only the identity

$$\frac{(1+\beta)^2}{(1-\beta)^2} + \frac{\beta^2(1+\beta)}{(1-\beta)^3} = \frac{(1+\beta)((1+\beta)(1-\beta)+\beta^2)}{(1-\beta)^3} = \frac{1+\beta}{(1-\beta)^3}.$$

This proves (15).

For $\eta = 1/(2L\sqrt{T})$, substitution gives

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(x_t)\|^2 \le \frac{4L\Delta_0 + (1+\beta)\sigma^2/(2(1-\beta)^3)}{\sqrt{T}}.$$

If $0 < \beta \le 1/2$, then $(1+\beta)/(2(1-\beta)^3) \le (3/2)/(2(1/2)^3) = 6$, which proves the uniform bound. For an independent uniform index $I \in \{0, \dots, T-1\}$,

$$\mathbb{E}\|\nabla F(x_I)\| \le \sqrt{\mathbb{E}\|\nabla F(x_I)\|^2} = \sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(x_t)\|^2}.$$

□

Dependence on the variance. For $0 < \beta \le 1/2$, (15) also gives

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla F(x_t)\right\|^2 \le \frac{2\Delta_0}{\eta T} + 12L\eta\sigma^2.$$

When $\Delta_0, \sigma > 0$, choose $\eta = \min\{1/(2L), \sqrt{\Delta_0/(6L\sigma^2 T)}\}$. The elementary inequalities

$$\eta^{-1} \le 2L + \sqrt{\frac{6L\sigma^2 T}{\Delta_0}}, \qquad \eta \le \sqrt{\frac{\Delta_0}{6L\sigma^2 T}}$$

then yield

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla F(x_t)\right\|^2 \le \frac{4L\Delta_0}{T} + 4\sqrt{6}\,\sigma\sqrt{\frac{L\Delta_0}{T}}.$$

Thus the bound recovers the usual initialization and variance orders of nonconvex SGD.

If $\sigma = 0$, taking $\eta = 1/(2L)$ directly gives $4L\Delta_0/T$.

D Proof of Lemma 4

Proof. For fixed y, let $\phi(z) = F(z) - \langle \nabla F(y), z\rangle$. Convexity and $\nabla\phi(y) = 0$ imply that y minimizes ϕ. The smoothness upper model at x, evaluated at $x - L^{-1}\nabla\phi(x)$, therefore gives

$$\phi(y) \le \phi\big(x - L^{-1}\nabla\phi(x)\big) \le \phi(x) - \frac{\left\|\nabla\phi(x)\right\|^2}{2L}.$$

Substituting the definition of ϕ yields

$$F(x) - F(y) - \langle \nabla F(y), x - y\rangle \ge \frac{\left\|\nabla F(x) - \nabla F(y)\right\|^2}{2L}.$$

Interchanging x, y and adding cancels the function values. With $u = x - y$ and $v = \nabla F(x) - \nabla F(y)$, the result is $\langle v, u\rangle \ge \left\|v\right\|^2/L$. Hence

$$\left\|u - \eta v\right\|^2 = \left\|u\right\|^2 - 2\eta\,\langle u, v\rangle + \eta^2\left\|v\right\|^2 \le \left\|u\right\|^2 - \eta(2/L - \eta)\left\|v\right\|^2 \le \left\|u\right\|^2.$$

Taking square roots proves (17).

Contraction under strong convexity. Assume first that $L > \lambda$ and set $H(z) = F(z) - \lambda\left\|z\right\|^2/2$. Subtracting the quadratic term from the strong-convexity lower model and the smoothness upper model gives

$$0 \le H(z) - H(w) - \langle \nabla H(w), z - w\rangle \le \frac{L - \lambda}{2}\left\|z - w\right\|^2.$$

Thus H is convex and has an upper quadratic model with coefficient $L - \lambda$. The preceding proof of nonexpansiveness uses only these two properties, so it applies to H. Since $\eta \le 1/L$ and $L > \lambda$, we have $1 - \lambda\eta > 0$ and

$$\frac{\eta}{1 - \lambda\eta} \le \frac{1}{L - \lambda};$$

after multiplying by the positive denominators, this is just $L\eta \le 1$. Using $\nabla F(z) = \nabla H(z) + \lambda z$,

$$\begin{aligned}
&\left\|x - \eta\nabla F(x) - y + \eta\nabla F(y)\right\| \\
&\quad = (1 - \lambda\eta)\left\|x - \frac{\eta}{1 - \lambda\eta}\nabla H(x) - y + \frac{\eta}{1 - \lambda\eta}\nabla H(y)\right\| \\
&\quad \le (1 - \lambda\eta)\left\|x - y\right\|.
\end{aligned}$$

If $L = \lambda$, the lower and upper models for H coincide, so H is affine. Hence $\nabla F(x) - \nabla F(y) = \lambda(x - y)$ and (18) holds with equality.

Comparison of function gaps. The smoothness upper model evaluated at one gradient step gives

$$F^* \le F\big(y - L^{-1}\nabla F(y)\big) \le F(y) - \frac{\left\|\nabla F(y)\right\|^2}{2L}.$$

Rearrangement proves (19). Apply the upper model once more to $F(x)$, and use Young's inequality:

$$\begin{aligned} F(x) - F^* &\le F(y) - F^* + \langle \nabla F(y), x - y \rangle + \frac{L}{2} \|x - y\|^2 \\ &\le F(y) - F^* + \frac{\|\nabla F(y)\|^2}{2L} + L \|x - y\|^2 \\ &\le 2(F(y) - F^*) + L \|x - y\|^2 . \end{aligned}$$

The two function-gap inequalities require smoothness only. □

E Proof of Lemma 5

Proof. Let $\mathcal{F}_t = \sigma(\xi_0, \dots, \xi_{t-1})$ be the information before drawing ξ_t. Write $\zeta(x, \xi) = g(x, \xi) - \nabla F(x)$ and $e_t = d_t - \nabla F(x_t)$. Set $e_{-1} = 0$, $\nu_0 = e_0$, and $\nu_t = \zeta(x_t, \xi_t) - \beta \zeta(x_{t-1}, \xi_t)$ for $t \ge 1$. Subtracting $\nabla F(x_t)$ from (3) yields $e_t = \beta e_{t-1} + \nu_t$. Fresh-sample independence and unbiasedness give $\mathbb{E}[\nu_t \mid \mathcal{F}_t] = 0$. The conditional L^2 triangle inequality gives

$$\big(\mathbb{E}[\|\nu_t\|^2 \mid \mathcal{F}_t]\big)^{1/2} \le \big(\mathbb{E}[\|\zeta(x_t, \xi_t)\|^2 \mid \mathcal{F}_t]\big)^{1/2} + \beta \big(\mathbb{E}[\|\zeta(x_{t-1}, \xi_t)\|^2 \mid \mathcal{F}_t]\big)^{1/2} \le (1 + \beta)\sigma.$$

At $t = 0$ the same bound follows directly from (5). Thus, with $S^2 = (1 + \beta)^2 \sigma^2$,

$$\mathbb{E}[\nu_t \mid \mathcal{F}_t] = 0, \qquad \mathbb{E}[\|\nu_t\|^2 \mid \mathcal{F}_t] \le S^2 \le 9\sigma^2/4. \tag{54}$$

The estimation error. The previous error is $\mathcal{F}_t$-measurable, so $\mathbb{E} \langle e_{t-1}, \nu_t \rangle = 0$. Squaring the error recursion and substituting recursively from $e_{-1} = 0$ gives

$$\mathbb{E} \|e_t\|^2 \le \beta^2 \mathbb{E} \|e_{t-1}\|^2 + S^2 \le S^2 \sum_{j=0}^{t} \beta^{2j} \le \frac{(1 + \beta)^2 \sigma^2}{1 - \beta^2} = \frac{1 + \beta}{1 - \beta} \sigma^2 \le 3\sigma^2.$$

Define the mathematical companion by

$$y_0 = x_0, \qquad y_{t+1} = y_t - \eta_t (\nabla F(y_t) + \nu_t). \tag{55}$$

Subtracting (55) from (2) cancels the shared innovation:

$$x_{t+1} - y_{t+1} = x_t - \eta_t \nabla F(x_t) - y_t + \eta_t \nabla F(y_t) - \eta_t \beta e_{t-1}.$$

For $r_t = (\mathbb{E} \|x_t - y_t\|^2)^{1/2}$, Minkowski's inequality, (17), and the estimation-error bound above yield $r_{t+1} \le r_t + \sqrt{3} \beta \sigma \eta_t$. Summing $r_{j+1} - r_j \le \sqrt{3} \beta \sigma \eta_j$ from $j = 0$ to $t - 1$, with $r_0 = 0$, gives $r_t \le \sqrt{3} \beta \sigma \sum_{j<t} \eta_j$. Squaring gives

$$\mathbb{E} \|x_t - y_t\|^2 \le 3\beta^2 \sigma^2 \left(\sum_{j=0}^{t-1} \eta_j \right)^2 . \tag{56}$$

Under λ-strong convexity, (18) instead gives $r_{t+1} \le (1 - \lambda \eta_t) r_t + \sqrt{3} \beta \sigma \eta_t$. Starting from $r_0 = 0$, the induction

$$r_{t+1} \le (1 - \lambda \eta_t) \frac{\sqrt{3} \beta \sigma}{\lambda} + \sqrt{3} \beta \sigma \eta_t = \frac{\sqrt{3} \beta \sigma}{\lambda}$$

proves

$$\mathbb{E} \|x_t - y_t\|^2 \le 3\beta^2 \sigma^2 / \lambda^2. \tag{57}$$

Under strong convexity, this gives

$$\mathbb{E}[F(x_t) - F^*] \le 2\mathbb{E}[F(y_t) - F^*] + 3L\beta^2 \sigma^2 / \lambda^2. \tag{58}$$

Combining the same pointwise inequality with (56) gives the convex assertion explicitly:

$$\mathbb{E}[F(x_t) - F^*] \le 2\mathbb{E}[F(y_t) - F^*] + 3L\beta^2 \sigma^2 \left(\sum_{j=0}^{t-1} \eta_j \right)^2 . \tag{59}$$

□

F Other useful lemmas

F.1 A shared one-step inequality

Lemma 11. *Let y_t be an adapted recursion of the form* (55) *with* (54)*, where y_t is $\mathcal{F}_t$-measurable. Under Assumptions 1 and 5, suppose $y_0 = x_0$. Define $D_t = \mathbb{E}\,\|y_t - x^*\|^2$, $\delta_t = \mathbb{E}[F(y_t) - F^*]$, and $h_t = \nabla F(y_t) + \nu_t$. For deterministic $0 < \eta_t \le 1/(2L)$,*

$$\mathbb{E}\,\|h_t\|^2 \le 2L\delta_t + S^2, \tag{60}$$

$$D_{t+1} \le D_t - \eta_t\delta_t + \eta_t^2 S^2. \tag{61}$$

Proof. Note that $\|\nabla F(y_t)\| \le L\,\|y_t - x^*\|$ and conditional centering justifies $\mathbb{E}\,\langle \nabla F(y_t), \nu_t\rangle = 0$. Expanding $\|h_t\|^2$ and using (19) proves (60). For distance, the same centering and convexity give

$$\begin{aligned} D_{t+1} &= D_t - 2\eta_t\mathbb{E}\,\langle h_t, y_t - x^*\rangle + \eta_t^2\mathbb{E}\,\|h_t\|^2 \\ &\le D_t - 2\eta_t\mathbb{E}\,\langle \nabla F(y_t), y_t - x^*\rangle + \eta_t^2(2L\delta_t + S^2) \\ &\le D_t - 2\eta_t(1 - L\eta_t)\delta_t + \eta_t^2 S^2 \\ &\le D_t - \eta_t\delta_t + \eta_t^2 S^2. \end{aligned}$$

The third line uses $\langle \nabla F(y_t), y_t - x^*\rangle \ge F(y_t) - F^*$. The last uses $2(1 - L\eta_t) \ge 1$ and $\delta_t \ge 0$. □

F.2 Rounded-block geometry for arbitrary integer horizons

Lemma 12. *For the partition* (25) *and $T \ge 1$, put $M = \lceil (T+1)/4 \rceil$, $K = \lceil (T+1)/2 \rceil$. Define*

$$B_0 = \{M - 1, \dots, K - 1\}, \qquad B_i = \{U_i, \dots, U_{i+1} - 1\} \quad (1 \le i \le m). \tag{62}$$

Then $U_1 = K$, $B_m = \{T\}$, and

$$M \ge (T+1)/4, \quad K \ge (T+1)/2, \quad K \in \{2M - 1, 2M\},$$

$$n := K - M + 1 \ge (T+1)/4, \quad K/n \le 2, \tag{63}$$

$$|B_0| \le 2|B_1|, \qquad |B_i| \le 3|B_{i+1}| \quad (1 \le i < m). \tag{64}$$

In particular, a suffix of B_i together with all of B_{i+1} contains at most $4|B_{i+1}|$ indices.

Proof. The identities $U_1 = \lceil (T+1)/2 \rceil$ and $\lfloor (T+1)/2^m \rfloor = 1$ prove the boundary assertions. Writing $T+1 = 4q+r$ with $r \in \{0,1,2,3\}$ verifies $K \in \{2M-1, 2M\}$. Hence $n \ge M \ge (T+1)/4$ and $K \le 2n$. The same four cases give

r	0	1	2	3
$\lvert B_0\rvert$	$q+1$	$q+1$	$q+1$	$q+2$
$\lvert B_1\rvert$	q	q	$q+1$	$q+1$.

For $r = 0, 1$, $T \ge 1$ implies $q \ge 1$; for $r = 2, 3$, $q \ge 0$ is enough. Thus $|B_0| \le 2|B_1|$ in every case, including $T = 1, 2$.

For the remaining ratios, let $r_i = \lfloor (T+1)/2^i \rfloor$. For $1 \le i < m$, $r_{i+1} = \lfloor r_i/2 \rfloor$ and $|B_i| = r_i - r_{i+1} = \lceil r_i/2 \rceil$. For $1 \le i < m-1$, write $r_i = 2r_{i+1} + \varepsilon$ with $\varepsilon \in \{0, 1\}$. It follows that

$$|B_i| = r_{i+1} + \varepsilon \le 2\lceil r_{i+1}/2 \rceil + 1 = 2|B_{i+1}| + 1 \le 3|B_{i+1}|.$$

For the final pair, $r_m = 1$ and $r_{m-1} \in \{2, 3\}$, so $|B_{m-1}| \le 2$ and $|B_m| = 1$. □

G Proof of Theorem 6

G.1 General stepsize-weighted and constant-step means

Theorem 13. *Under Assumptions 1, 2, and 5, let $T \ge 1$ be any integer, take $\beta = 1/(T+2)$, and choose deterministic stepsizes $0 < \eta_t \le 1/(2L)$ for $0 \le t \le T$. Define*

$$\Lambda_T = \sum_{t=0}^{T} \eta_t, \qquad \bar{x}_T^{\eta} = \frac{1}{\Lambda_T}\sum_{t=0}^{T} \eta_t x_t.$$

Then

$$\mathbb{E}[F(\bar{x}_T^\eta) - F^*] \le \frac{2R^2}{\Lambda_T} + \frac{9\sigma^2}{2\Lambda_T}\sum_{t=0}^{T}\eta_t^2 + \frac{3L\sigma^2\Lambda_T^2}{(T+2)^2}. \tag{65}$$

Proof. Apply Lemma 5, and use the notation $D_t = \mathbb{E}\left\|y_t - x^*\right\|^2$ and $\delta_t = \mathbb{E}[F(y_t) - F^*]$. Rearrange (61) as $\eta_t\delta_t \le D_t - D_{t+1} + \eta_t^2 S^2$. Summing over $t = 0, \dots, T$ cancels each interior distance:

$$\sum_{t=0}^{T}\eta_t\delta_t \le \sum_{t=0}^{T}(D_t - D_{t+1}) + S^2\sum_{t=0}^{T}\eta_t^2 = D_0 - D_{T+1} + S^2\sum_{t=0}^{T}\eta_t^2 \le R^2 + S^2\sum_{t=0}^{T}\eta_t^2.$$

The final distance is nonnegative and can be discarded.

The weights η_t/Λ_T are positive and sum to one. Apply convexity to the weighted mean of STORM iterates, then apply (59) to each summand:

$$\begin{aligned}
\mathbb{E}[F(\bar{x}_T^\eta) - F^*] &\le \frac{1}{\Lambda_T}\sum_{t=0}^{T}\eta_t\mathbb{E}[F(x_t) - F^*] \\
&\le \frac{2}{\Lambda_T}\sum_{t=0}^{T}\eta_t\delta_t + \frac{3L\beta^2\sigma^2}{\Lambda_T}\sum_{t=0}^{T}\eta_t\left(\sum_{j=0}^{t-1}\eta_j\right)^2 \\
&\le \frac{2R^2}{\Lambda_T} + \frac{2S^2}{\Lambda_T}\sum_{t=0}^{T}\eta_t^2 + 3L\beta^2\sigma^2\Lambda_T^2.
\end{aligned}$$

For the last term, every prefix sum is at most Λ_T, and $\sum_t \eta_t = \Lambda_T$. Substituting $\beta = 1/(T+2)$ and $S^2 \le 9\sigma^2/4$ proves the theorem. □

Corollary 14. *With $\eta_t = \eta \le 1/(2L)$, the weighted mean becomes the uniform mean such that $\bar{x}_T = (T+1)^{-1}\sum_{t=0}^{T} x_t$, and*

$$\mathbb{E}[F(\bar{x}_T) - F^*] \le \frac{2R^2}{\eta(T+1)} + 6\eta\sigma^2. \tag{66}$$

If $R > 0$ and $\sigma > 0$, choose $\eta = \min\{1/(2L),\ R/(\sqrt{3}\sigma\sqrt{T+1})\}$ to obtain

$$\mathbb{E}[F(\bar{x}_T) - F^*] \le \frac{4LR^2}{T+1} + \frac{4\sqrt{3}\sigma R}{\sqrt{T+1}}. \tag{67}$$

Proof. Here $\Lambda_T = (T+1)\eta$ and $\sum_t \eta_t^2 = (T+1)\eta^2$. Since $(T+1)/(T+2) \le 1$ and $L\eta \le 1/2$, the last two terms of (65) are at most $(9/2)\eta\sigma^2 + 3L\eta^2\sigma^2 \le 6\eta\sigma^2$. For the tuned result, use

$$\frac{1}{\eta} = \max\{2L, \sqrt{3}\sigma\sqrt{T+1}/R\} \le 2L + \sqrt{3}\sigma\sqrt{T+1}/R, \qquad \eta \le \frac{R}{\sqrt{3}\sigma\sqrt{T+1}}.$$

The first inequality bounds the initialization term by $4LR^2/(T+1) + 2\sqrt{3}\sigma R/\sqrt{T+1}$; the second bounds the noise term by $2\sqrt{3}\sigma R/\sqrt{T+1}$, giving the stated bound. If $\sigma = 0$, take $\eta = 1/(2L)$. □

G.2 The companion's last-iterate bound

Lemma 15. *Under Assumptions 1 and 5, let $y_{t+1} = y_t - \eta_t(\nabla F(y_t) + \nu_t)$, with $y_0 = x_0$, y_t measurable with respect to the filtration $\mathcal{F}_t$, $\mathbb{E}[\nu_t \mid \mathcal{F}_t] = 0$, and $\mathbb{E}[\|\nu_t\|^2 \mid \mathcal{F}_t] \le S^2$. For $T \ge 1$ and the stepsizes in* (27)*,*

$$\max\left\{\delta_T,\ \frac{1}{T+1}\sum_{t=0}^{T}\delta_t\right\} \le \frac{4\mathrm{e}R^2}{\eta(T+1)} + 18\mathrm{e}\eta S^2, \qquad \delta_t = \mathbb{E}[F(y_t) - F^*]. \tag{68}$$

Proof. Use $D_t = \mathbb{E}\,\|y_t - x^*\|^2$ and $h_t = \nabla F(y_t) + \nu_t$. Note that (60)–(61) apply. Put $M = \lceil (T+1)/4 \rceil$ and $K = \lceil (T+1)/2 \rceil$. The first schedule block is $0, \dots, K-1$, with $\eta_t = \eta$.

A reference value in the first half. Discard the nonpositive term $-\eta\delta_t$ in (61) for $0 \le t < M-1$ and sum. This yields

$$D_{M-1} \le D_0 + (M-1)\eta^2 S^2 = R^2 + (M-1)\eta^2 S^2.$$

Now keep that decrease and sum from $t = M-1$ to $K-1$:

$$\begin{aligned}\eta \sum_{t=M-1}^{K-1} \delta_t &\le D_{M-1} - D_K + (K-M+1)\eta^2 S^2 \\ &\le R^2 + \big((M-1) + (K-M+1)\big)\eta^2 S^2 \\ &= R^2 + K\eta^2 S^2.\end{aligned}$$

Let A_0 be the minimum of δ_t over these $K-M+1$ indices. A minimum is at most the mean, so

$$A_0 \le \frac{R^2}{(K-M+1)\eta} + \frac{K}{K-M+1}\eta S^2 \le \frac{4R^2}{\eta(T+1)} + 2\eta S^2. \tag{69}$$

The last line uses $K - M + 1 \ge (T+1)/4$ and $K/(K-M+1) \le 2$ from Lemma 12. We will also need the mean over the entire first block. Summing directly from 0 to $K-1$ gives

$$\frac{1}{K}\sum_{t=0}^{K-1} \delta_t \le \frac{R^2}{K\eta} + \eta S^2 \le \frac{2R^2}{\eta(T+1)} + \eta S^2, \tag{70}$$

where $K \ge (T+1)/2$.

Comparison with an earlier point. Fix deterministic indices $p \le q$. For every $p \le t \le q$, expand the squared distance from y_p:

$$\|y_{t+1} - y_p\|^2 = \|y_t - y_p\|^2 - 2\eta_t \langle h_t, y_t - y_p\rangle + \eta_t^2 \|h_t\|^2.$$

Although y_p is random, it is $\mathcal{F}_t$-measurable when $t \ge p$. Conditional centering therefore gives

$$\begin{aligned}\mathbb{E}\,\langle h_t, y_t - y_p\rangle &= \mathbb{E}\,\langle \nabla F(y_t), y_t - y_p\rangle \\ &\ge \mathbb{E}[F(y_t) - F(y_p)] = \delta_t - \delta_p.\end{aligned}$$

The inequality follows by rearranging $F(y_p) \ge F(y_t) + \langle \nabla F(y_t), y_p - y_t\rangle$. After taking expectations and summing the distance identity, the initial distance is zero and the terminal distance is nonnegative:

$$\begin{aligned}2\sum_{t=p}^{q} \eta_t(\delta_t - \delta_p) &\le -\mathbb{E}\,\|y_{q+1} - y_p\|^2 + \sum_{t=p}^{q} \eta_t^2 \mathbb{E}\,\|h_t\|^2 \\ &\le 2L\sum_{t=p}^{q} \eta_t^2 \delta_t + S^2 \sum_{t=p}^{q} \eta_t^2.\end{aligned}$$

The right side still depends on the function errors. Write $\delta_t = (\delta_t - \delta_p) + \delta_p$ there, and move the first part to the left. This is an exact rearrangement, giving

$$2\sum_{t=p}^{q} \eta_t(1 - L\eta_t)(\delta_t - \delta_p) \le (2L\delta_p + S^2)\sum_{t=p}^{q} \eta_t^2. \tag{71}$$

The differences $\delta_t - \delta_p$ need not be nonnegative.

Comparison of consecutive blocks. Use the comparison blocks (62). For each i, choose a deterministic minimizer τ_i of δ_t on B_i and set $A_i = \delta_{\tau_i}$. These are indices minimizing expected errors, not data-dependent choices made by the algorithm. Write $\bar{\delta}_i = |B_i|^{-1}\sum_{t\in B_i} \delta_t$ for a block mean. In particular, $A_i \le \bar{\delta}_i$ and $A_m = \bar{\delta}_m = \delta_T$ because $B_m = \{T\}$.

In (71), take $p = \tau_i$ and $q = \max B_{i+1}$ for $0 \le i < m$. On the remaining part of B_i, each difference $\delta_t - A_i$ is nonnegative. Its coefficient $2\eta_t(1 - L\eta_t)$ is positive, so discarding this source part can only

decrease the left side. Throughout the target block B_{i+1} the stepsize equals $\eta 2^{-(i+1)}$. The remaining inequality is therefore

$$2\eta 2^{-(i+1)}(1 - L\eta 2^{-(i+1)})|B_{i+1}|(\bar{\delta}_{i+1} - A_i) \le (2LA_i + S^2) \sum_{t=\tau_i}^{\max B_{i+1}} \eta_t^2.$$

Every comparison span has at most $4|B_{i+1}|$ indices by (64). Its largest stepsize is at most $\eta 2^{-i}$. Moreover, $1 - L\eta 2^{-(i+1)} \ge 1/2$. Thus the ratio of the sum of squared stepsizes to the coefficient of $\bar{\delta}_{i+1} - A_i$ is at most

$$\frac{4|B_{i+1}|\eta^2 2^{-2i}}{|B_{i+1}|\eta 2^{-(i+1)}} = 8\eta 2^{-i} =: c_i.$$

Since $2LA_i + S^2 \ge 0$, it follows that

$$A_{i+1} \le \bar{\delta}_{i+1} \le (1 + 2Lc_i)A_i + c_i S^2, \qquad c_i = 8\eta 2^{-i}. \tag{72}$$

The mean difference $\bar{\delta}_{i+1} - A_i$ itself may be negative; division by its positive coefficient remains valid in that case. The length bounds used here include $T = 1$ and $T = 2$, whose blocks are respectively $B_0 = \{0\}, B_1 = \{1\}$ and $B_0 = \{0, 1\}, B_1 = \{2\}$.

Expansion of the scalar recursion. The first two substitutions in (72) (the second when $m \ge 2$) are

$$A_1 \le (1 + 2Lc_0)A_0 + c_0 S^2,$$
$$A_2 \le (1 + 2Lc_1)(1 + 2Lc_0)A_0 + \big(c_0(1 + 2Lc_1) + c_1\big)S^2.$$

More generally, induction gives, for $0 \le i < m$,

$$\bar{\delta}_{i+1} \le A_0 \prod_{j=0}^{i}(1 + 2Lc_j) + S^2 \sum_{j=0}^{i} c_j \prod_{\ell=j+1}^{i} (1 + 2Lc_\ell). \tag{73}$$

To see the induction step, use the same formula with index $i - 1$ to bound $A_i \le \bar{\delta}_i$, multiply it by $1 + 2Lc_i$, and add $c_i S^2$. This extends every existing product by its last factor and adds the $j = i$ summand, whose empty product equals one.

The coefficients have a geometric sum:

$$\sum_{j=0}^{m-1} c_j = 8\eta \sum_{j=0}^{m-1} 2^{-j} \le 16\eta.$$

For any product occurring in (73), use $1 + u \le \exp(u)$ for $u \ge 0$ to obtain

$$\prod(1 + 2Lc_j) \le \exp\left(2L \sum_{j=0}^{m-1} c_j\right) \le \exp(32L\eta) \le \mathrm{e}.$$

Hence every tail-block mean satisfies

$$\begin{aligned}\bar{\delta}_{i+1} &\le \mathrm{e}A_0 + 16\mathrm{e}\eta S^2 \\ &\le \frac{4\mathrm{e}R^2}{\eta(T+1)} + 18\mathrm{e}\eta S^2.\end{aligned} \tag{74}$$

The second line inserts (69). The coefficients c_j form a geometric sequence, so neither term incurs a factor proportional to the number of blocks.

The last iterate and the uniform mean. The last block is a singleton, so (74) with $i + 1 = m$ is already the desired last-point inequality. For the full mean, partition $0, \dots, T$ into the entire first schedule block and $B_1, \dots, B_m$:

$$\frac{1}{T+1}\sum_{t=0}^{T}\delta_t = \frac{K}{T+1}\left(\frac{1}{K}\sum_{t=0}^{K-1}\delta_t\right) + \sum_{i=1}^{m}\frac{|B_i|}{T+1}\bar{\delta}_i.$$

The coefficients on the right sum to one. The first mean is bounded by (70), which is no larger than the right side of (74); every remaining mean is bounded by (74) itself. Their convex combination obeys the same bound. This completes the proof of (68). □

G.3 Proof of the convex theorem

Proof of Theorem 6. Lemma 5 gives $S^2 \le 9\sigma^2/4$ and a companion to which Lemma 15 applies. Because $\eta_t \le \eta$ and $\beta = 1/(T+2)$, for every $0 \le t \le T$,

$$\mathbb{E}\left\|x_t - y_t\right\|^2 \le 3\beta^2\sigma^2(t\eta)^2 \le 3\sigma^2\eta^2. \tag{75}$$

Transfer the companion's last-point guarantee using (59) and (75):

$$\begin{aligned}\mathbb{E}[F(x_T) - F^*] &\le 2\delta_T + 3L\sigma^2\eta^2 \\ &\le \frac{8\mathrm{e}R^2}{\eta(T+1)} + 36\mathrm{e}\eta S^2 + 3L\sigma^2\eta^2 \\ &\le \frac{8\mathrm{e}R^2}{\eta(T+1)} + (81\mathrm{e} + 3/32)\eta\sigma^2 \\ &\le \frac{8\mathrm{e}R^2}{\eta(T+1)} + 82\mathrm{e}\eta\sigma^2.\end{aligned}$$

Here $36(9/4) = 81$ and $L\eta \le 1/32$. For the uniform mean, convexity gives

$$\mathbb{E}[F(\bar{x}_T) - F^*] \le \frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[F(x_t) - F^*] \le \frac{2}{T+1}\sum_{t=0}^{T}\delta_t + 3L\sigma^2\eta^2.$$

Use the mean part of (68) and the same numerical calculation. This proves the last-point and uniform-mean bound.

The stepsize-weighted mean. Use Theorem 13 directly, and put $K = \lceil (T+1)/2 \rceil$. The first K weights equal η, every weight is at most η, and $K \ge (T+1)/2$. Therefore

$$\Lambda_T \ge K\eta \ge (T+1)\eta/2, \qquad \sum_{t=0}^{T}\eta_t^2 \le \eta\Lambda_T, \qquad \frac{\Lambda_T}{T+2} \le \eta.$$

Insert these three inequalities into (65):

$$\begin{aligned}\mathbb{E}[F(\bar{x}_T^{\eta}) - F^*] &\le \frac{4R^2}{\eta(T+1)} + \frac{9}{2}\eta\sigma^2 + 3L\eta^2\sigma^2 \\ &\le \frac{4R^2}{\eta(T+1)} + \left(\frac{9}{2} + \frac{3}{32}\right)\eta\sigma^2 \le \frac{4R^2}{\eta(T+1)} + 5\eta\sigma^2.\end{aligned}$$

This gives the weighted-mean bound. Its right side is no larger than the common bound for the last iterate and uniform mean, so that bound applies to all three outputs.

Choice of the base stepsize. For the stepsize choice in Theorem 6,

$$\frac{1}{\eta} = \max\left\{32L, \frac{\sqrt{41}\sigma\sqrt{T+1}}{2R}\right\} \le 32L + \frac{\sqrt{41}\sigma\sqrt{T+1}}{2R}, \qquad \eta \le \frac{2R}{\sqrt{41}\sigma\sqrt{T+1}}.$$

Thus the two terms of the common bound satisfy

$$\begin{aligned}\frac{8\mathrm{e}R^2}{\eta(T+1)} &\le \frac{256\mathrm{e}LR^2}{T+1} + \frac{4\sqrt{41}\,\mathrm{e}\sigma R}{\sqrt{T+1}}, \\ 82\mathrm{e}\eta\sigma^2 &\le \frac{164\mathrm{e}\sigma R}{\sqrt{41}\sqrt{T+1}} = \frac{4\sqrt{41}\,\mathrm{e}\sigma R}{\sqrt{T+1}}.\end{aligned}$$

Adding proves the tuned bound. If $\sigma = 0$, choose $\eta = 1/(32L)$ and obtain the first term of the tuned bound. If $R = 0$, the separate zero-step choice returning $x_0 = x^*$ gives zero error. □

H Proof of Theorems 7 and 8

H.1 The linearly weighted mean

Proof of Theorem 7. Use the parameters in (32), and write $W_T = \sum_{t=0}^{T}(t+s) = (T+1)(T+2s)/2$. The choices satisfy $\beta \le 1/2$ and $\eta_t \le 1/(2L)$, so the companion lemma applies. For the remainder of this proof write

$$D_t = \mathbb{E}\left\|y_t - x^*\right\|^2, \qquad \delta_t = \mathbb{E}[F(y_t) - F^*].$$

The one-step inequality. Expand the squared distance in (55). The cross terms containing ν_t vanish after conditioning on $\mathcal{F}_t$, giving

$$\begin{aligned} D_{t+1} &\le D_t - 2\eta_t \mathbb{E}\left\langle \nabla F(y_t), y_t - x^*\right\rangle + \eta_t^2 \mathbb{E}\left\|\nabla F(y_t)\right\|^2 + \eta_t^2 S^2 \\ &\le (1-\lambda\eta_t) D_t - 2\eta_t(1 - L\eta_t)\delta_t + \eta_t^2 S^2 \\ &\le (1-\lambda\eta_t) D_t - \eta_t \delta_t + \eta_t^2 S^2. \end{aligned}$$

The second line uses $\langle \nabla F(y_t), y_t - x^*\rangle \ge F(y_t) - F^* + \lambda \left\|y_t - x^*\right\|^2/2$, together with (19). The third line uses $L\eta_t \le 1/2$. Rearranging and substituting the stepsize gives

$$\delta_t \le \frac{\lambda}{2}(t+s-1)D_t - \frac{\lambda}{2}(t+s+1)D_{t+1} + \frac{2S^2}{\lambda(t+s+1)}. \tag{76}$$

Summation with increasing weights. Multiply (76) by $t+s$:

$$(t+s)\delta_t \le \frac{\lambda}{2}\big((t+s)(t+s-1)D_t - (t+s)(t+s+1)D_{t+1}\big) + \frac{2S^2}{\lambda}\frac{t+s}{t+s+1}.$$

At $t=0$, the distance terms are $s(s-1)D_0 - s(s+1)D_1$. At $t=1$, they are $s(s+1)D_1 - (s+1)(s+2)D_2$. Thus the D_1 terms cancel, and the same cancellation occurs at every interior index. Summing gives

$$\begin{aligned} \sum_{t=0}^{T}(t+s)\delta_t &\le \frac{\lambda}{2}\big(s(s-1)D_0 - (T+s)(T+s+1)D_{T+1}\big) + \frac{2S^2}{\lambda}\sum_{t=0}^{T}\frac{t+s}{t+s+1} \\ &\le \frac{\lambda s(s-1)R^2}{2} + \frac{2(T+1)S^2}{\lambda}. \end{aligned}$$

Here $D_{T+1} \ge 0$, every fraction is at most one, and $D_0 = R^2$.

Transfer to the STORM iterates. Convexity and (58) imply

$$\begin{aligned} \mathbb{E}[F(\bar{x}_T) - F^*] &\le \frac{1}{W_T}\sum_{t=0}^{T}(t+s)\mathbb{E}[F(x_t) - F^*] \\ &\le \frac{2}{W_T}\sum_{t=0}^{T}(t+s)\delta_t + \frac{3L\beta^2\sigma^2}{\lambda^2} \\ &\le \frac{\lambda s(s-1)R^2}{W_T} + \frac{4(T+1)S^2}{\lambda W_T} + \frac{3L\beta^2\sigma^2}{\lambda^2}. \end{aligned}$$

Substitute $W_T = (T+1)(T+2s)/2$, $S^2 \le 9\sigma^2/4$, and $\beta = 1/(T+s+1)$ to obtain the sharper three-term bound

$$\mathbb{E}[F(\bar{x}_T) - F^*] \le \frac{2\lambda s(s-1)R^2}{(T+1)(T+2s)} + \frac{18\sigma^2}{\lambda(T+2s)} + \frac{3\kappa\sigma^2}{\lambda(T+s+1)^2}. \tag{77}$$

For the simpler bound, use $s \ge 4\kappa \ge 4$:

$$\begin{aligned} \frac{18\sigma^2}{\lambda(T+2s)} + \frac{3\kappa\sigma^2}{\lambda(T+s+1)^2} &\le \left(18 + \frac{3\kappa}{T+s+1}\right)\frac{\sigma^2}{\lambda(T+s+1)} \\ &\le \left(18 + \frac{3}{4}\right)\frac{\sigma^2}{\lambda(T+s+1)} \le \frac{19\sigma^2}{\lambda(T+s+1)}. \end{aligned}$$

If J is independent of the run with $\Pr(J=t) = (t+s)/W_T$, the weighted average of expected STORM function gaps equals $\mathbb{E}[F(x_J) - F^*]$. The same calculation therefore proves the random-output assertion. □

H.2 A BLOCK COMPARISON FOR AN SGD COMPANION

Lemma 16 (From an earlier block to the last point)**.** *Let F be differentiable and convex, and consider*

$$y_{t+1} = y_t - \eta_t h_t, \qquad \mathbb{E}[h_t \mid \mathcal{F}_t] = \nabla F(y_t),$$

where $(\mathcal{F}_t)$ is a filtration and y_t is $\mathcal{F}_t$-measurable. Let $T \geq 1$ be an integer and set

$$M = \lceil (T+1)/4 \rceil, \qquad K = \lceil (T+1)/2 \rceil, \qquad \gamma_t = \frac{c}{t+s+1}, \qquad c > 0,\ s \geq 0.$$

Use the boundaries in (25) *and the stepsizes $\eta_t = 2^{-i}\gamma_t$ for $U_i \leq t < U_{i+1}$. Assume the displayed function-value expectations are finite. If $\mathbb{E}\|h_t\|^2 \leq G^2$ for $M-1 \leq t \leq T$, then*

$$\mathbb{E}F(y_T) \leq \min_{M-1 \leq t \leq K-1} \mathbb{E}F(y_t) + 64G^2\gamma_T. \tag{78}$$

It is enough to introduce y_{T+1} virtually; the output is y_T.

Proof. Since $T \geq 1$, we have $m \geq 1$. We use the geometry proved in Lemma 12: $M \geq (T+1)/4$, $K \leq 2M$, and $n = K - M + 1 \geq (T+1)/4$.

Comparison with an earlier point. Fix deterministic indices $M-1 \leq p \leq q \leq T$. For every $p \leq t \leq q$, expand the update around y_p:

$$\|y_{t+1} - y_p\|^2 = \|y_t - y_p\|^2 - 2\eta_t \langle h_t, y_t - y_p \rangle + \eta_t^2 \|h_t\|^2 .$$

Although y_p is random, it is $\mathcal{F}_t$-measurable because $p \leq t$. The same is true of y_t. Consequently,

$$\begin{aligned} \mathbb{E}[\langle h_t, y_t - y_p \rangle \mid \mathcal{F}_t] &= \langle \mathbb{E}[h_t \mid \mathcal{F}_t], y_t - y_p \rangle \\ &= \langle \nabla F(y_t), y_t - y_p \rangle \\ &\geq F(y_t) - F(y_p). \end{aligned}$$

The last inequality is convexity applied with expansion point y_t:

$$F(y_p) \geq F(y_t) + \langle \nabla F(y_t), y_p - y_t \rangle .$$

Taking expectations in the distance identity therefore gives

$$2\eta_t\big(\mathbb{E}F(y_t) - \mathbb{E}F(y_p)\big) \leq \mathbb{E}\|y_t - y_p\|^2 - \mathbb{E}\|y_{t+1} - y_p\|^2 + \eta_t^2 G^2.$$

Summing from $t = p$ to q cancels intermediate distance. The first distance is $\|y_p - y_p\|^2 = 0$, so

$$2\sum_{t=p}^{q} \eta_t\big(\mathbb{E}F(y_t) - \mathbb{E}F(y_p)\big) \leq -\mathbb{E}\|y_{q+1} - y_p\|^2 + G^2 \sum_{t=p}^{q} \eta_t^2 \leq G^2 \sum_{t=p}^{q} \eta_t^2.$$

We will use the resulting inequality in the form

$$2\sum_{t=p}^{q} \eta_t\big(\mathbb{E}F(y_t) - \mathbb{E}F(y_p)\big) \leq G^2 \sum_{t=p}^{q} \eta_t^2. \tag{79}$$

All distances in this calculation have finite expectations: indeed, $y_t - y_p = -\sum_{j=p}^{t-1} \eta_j h_j$ is a finite sum of random vectors with finite second moments.

Comparison of adjacent blocks. Use the comparison blocks (62). They are consecutive from $M-1$ through T, and $B_m = \{T\}$. Choose index $\tau_i \in B_i$ minimizing $\mathbb{E}F(y_t)$ on B_i, and write

$$H_i = \mathbb{E}F(y_{\tau_{i+1}}) - \mathbb{E}F(y_{\tau_i}) \qquad (0 \leq i < m).$$

The index τ_i is deterministic because the quantities being minimized are expectations, not observed sample-path values.

Apply (79) with $p = \tau_i$ and $q = \max B_{i+1}$. The left sum consists of the remaining part of the source block B_i and all of the target block B_{i+1}. On the source part, the choice of τ_i gives

$$\mathbb{E}F(y_t) - \mathbb{E}F(y_{\tau_i}) \geq 0.$$

On the target block, the choice of τ_{i+1} gives

$$\mathbb{E}F(y_t) - \mathbb{E}F(y_{\tau_i}) \geq \mathbb{E}F(y_{\tau_{i+1}}) - \mathbb{E}F(y_{\tau_i}) = H_i.$$

All stepsizes are positive. Thus the whole left side of (79) is at least $2H_i \sum_{t\in B_{i+1}} \eta_t$. Dividing by this positive weight sum gives

$$H_i \leq \frac{G^2 \sum_{t=\tau_i}^{\max B_{i+1}} \eta_t^2}{2\sum_{t\in B_{i+1}} \eta_t}. \tag{80}$$

This reasoning is valid whether H_i is positive, zero, or negative.

The effect of rounded block lengths. First consider $i = 0$. By (64), the numerator in (80) contains at most $3|B_1|$ terms. Furthermore, $T + 1 \leq 4M$ and $s \geq 0$ give

$$\eta_t \leq \gamma_{M-1} = \frac{c}{M+s} \leq 4\gamma_T \quad \text{throughout the first comparison span.}$$

On B_1, the multiplier is $1/2$, and $t \leq T$ implies $\gamma_t \geq \gamma_T$. Each target stepsize is therefore at least $\gamma_T/2$. Substituting these upper and lower bounds into (80) yields

$$H_0 \leq \frac{3|B_1|G^2(4\gamma_T)^2}{2|B_1|(\gamma_T/2)} = 48G^2\gamma_T.$$

Next take $1 \leq i < m$. By (64), the comparison span contains at most $4|B_{i+1}|$ indices. All its indices satisfy $t \geq U_1 = K \geq (T+1)/2$, whence

$$\gamma_t = \frac{c}{t+s+1} \leq 2\gamma_T.$$

The largest multiplier in this span is 2^{-i}, so its largest stepsize is at most $2^{1-i}\gamma_T$. On the target block the multiplier is 2^{-i-1}, and every stepsize is at least $2^{-i-1}\gamma_T$. Consequently,

$$H_i \leq \frac{4|B_{i+1}|G^2(2^{1-i}\gamma_T)^2}{2|B_{i+1}|2^{-i-1}\gamma_T} = 16\,2^{-i}G^2\gamma_T.$$

When $m = 1$, there are no comparisons of this second type.

Summation over the blocks. Since $\tau_m = T$, consecutive differences telescope:

$$\begin{aligned}\mathbb{E}F(y_T) - \mathbb{E}F(y_{\tau_0}) &= \sum_{i=0}^{m-1} H_i \\ &\leq \left(48 + 16\sum_{i=1}^{m-1} 2^{-i}\right) G^2\gamma_T \\ &\leq 64G^2\gamma_T.\end{aligned}$$

The defining property of τ_0 now proves (78). The last comparison uses the update at $t = T$ only to produce the nonnegative terminal distance $\mathbb{E}\left\|y_{T+1} - y_{\tau_{m-1}}\right\|^2$, which is discarded. The point y_{T+1} need not be returned or computed by the algorithm. □

H.3 Proof of the last-iterate theorem

Proof of Theorem 8. Apply Lemma 5. In particular, the companion starts at $y_0 = x_0$ and has centered innovations satisfying

$$\mathbb{E}[\nu_t \mid \mathcal{F}_t] = 0, \qquad \mathbb{E}[\|\nu_t\|^2 \mid \mathcal{F}_t] \leq S^2, \qquad S^2 = (1+\beta)^2\sigma^2 \leq \frac{9}{4}\sigma^2.$$

Write

$$D_t = \mathbb{E}\left\|y_t - x^*\right\|^2, \qquad \delta_t = \mathbb{E}[F(y_t) - F^*], \qquad \gamma_t = \frac{2}{\lambda(t+s+1)},$$

and let $M = \lceil (T+1)/4 \rceil$, $K = \lceil (T+1)/2 \rceil$. The schedule (34) satisfies $\eta_t = \gamma_t$ for $0 \leq t < K$, and $\eta_t \leq \gamma_t$ everywhere. Since $s \geq 4\kappa^2 \geq 4\kappa$, every actual and virtual stepsize is at most $1/(2L)$.

Distance decrease in the first half. Expand one companion update around x^*:

$$y_{t+1} - x^* = y_t - x^* - \eta_t \nabla F(y_t) - \eta_t \nu_t.$$

The first three terms are $\mathcal{F}_t$-measurable, so the cross term with ν_t vanishes after taking expectation. Gradient-map contraction and $\nabla F(x^*) = 0$ then imply

$$\begin{aligned} D_{t+1} &\le (1 - \lambda\eta_t)^2 D_t + \eta_t^2 S^2 \\ &\le (1 - \lambda\eta_t) D_t + \eta_t^2 S^2. \end{aligned} \tag{81}$$

The second line follows from $0 \le \lambda\eta_t \le 1$. During the first block, where $\eta_t = \gamma_t$, we claim that

$$D_t \le \frac{(s+1)^2 R^2}{(t+s+1)^2} + \frac{4S^2}{\lambda^2(t+s+1)}, \qquad 0 \le t \le K. \tag{82}$$

At $t = 0$, the first term equals $D_0 = R^2$ and the second is nonnegative. For the induction step, put $u = t + s + 1$. Here $u \ge 5$, so $1 - 2/u \ge 0$ and substitution of the induction upper bound into (81) preserves the inequality. Since $\eta_t = 2/(\lambda u)$, this gives

$$\begin{aligned} D_{t+1} &\le \left(1 - \frac{2}{u}\right)\left(\frac{(s+1)^2 R^2}{u^2} + \frac{4S^2}{\lambda^2 u}\right) + \frac{4S^2}{\lambda^2 u^2} \\ &= \frac{(s+1)^2 R^2 (u-2)}{u^3} + \frac{4S^2(u-1)}{\lambda^2 u^2} \\ &\le \frac{(s+1)^2 R^2}{(u+1)^2} + \frac{4S^2}{\lambda^2(u+1)}. \end{aligned}$$

The last line follows term by term from

$$(u-2)(u+1)^2 = u^3 - 3u - 2 \le u^3, \qquad (u-1)(u+1) = u^2 - 1 \le u^2.$$

This proves (82).

Second moments of the search directions. For any $t \ge M - 1$, discard contraction in (81) and sum from $M - 1$ to $t - 1$:

$$D_t \le D_{M-1} + S^2 \sum_{j=M-1}^{t-1} \eta_j^2 \le D_{M-1} + S^2 \sum_{j=M-1}^{\infty} \gamma_j^2.$$

The infinite sum is only an upper bound on a finite sum; it does not introduce additional algorithmic iterations. By monotonicity of the summand and $M + s \ge 1$,

$$\begin{aligned} \sum_{j=M-1}^{\infty} \gamma_j^2 &= \frac{4}{\lambda^2} \sum_{k=M}^{\infty} \frac{1}{(k+s)^2} \\ &\le \frac{4}{\lambda^2}\left(\frac{1}{(M+s)^2} + \int_M^{\infty} \frac{du}{(u+s)^2}\right) \\ &= \frac{4}{\lambda^2}\left(\frac{1}{(M+s)^2} + \frac{1}{M+s}\right) \le \frac{8}{\lambda^2(M+s)}. \end{aligned}$$

Use (82) at $t = M - 1$ to conclude that

$$D_t \le \frac{(s+1)^2 R^2}{(M+s)^2} + \frac{12S^2}{\lambda^2(M+s)}, \qquad M - 1 \le t \le T. \tag{83}$$

Now write $h_t = \nabla F(y_t) + \nu_t$ for the companion direction. Since y_t is $\mathcal{F}_t$-measurable, conditional centering gives $\mathbb{E}\,\langle \nabla F(y_t), \nu_t \rangle = 0$. Standard smoothness and $\nabla F(x^*) = 0$ imply $\|\nabla F(y_t)\| \le L \|y_t - x^*\|$. Consequently,

$$\begin{aligned} \mathbb{E}\,\|h_t\|^2 &= \mathbb{E}\,\|\nabla F(y_t)\|^2 + \mathbb{E}\,\|\nu_t\|^2 \\ &\le L^2 D_t + S^2 \\ &\le \frac{L^2(s+1)^2 R^2}{(M+s)^2} + \left(1 + \frac{12\kappa^2}{M+s}\right) S^2 \\ &\le \frac{L^2(s+1)^2 R^2}{(M+s)^2} + 4S^2 =: G^2. \end{aligned}$$

For the last line, $M + s \geq s \geq 4\kappa^2$ gives $12\kappa^2/(M+s) \leq 3$. Thus the moment assumption of Lemma 16 holds on the required trajectory segment. This argument uses smoothness of F only; it imposes neither sample smoothness nor a globally bounded stochastic gradient.

A reference function value. For $M-1 \leq t \leq K-1$, the stepsize is γ_t. Lemma 11 applies because $\gamma_t \leq 1/(2L)$, and gives

$$D_{t+1} \leq D_t - \gamma_t \delta_t + \gamma_t^2 S^2.$$

Rearranging and summing makes the interior distances cancel:

$$\sum_{t=M-1}^{K-1} \gamma_t \delta_t \leq D_{M-1} - D_K + S^2 \sum_{t=M-1}^{K-1} \gamma_t^2 \leq D_{M-1} + S^2 \sum_{t=M-1}^{K-1} \gamma_t^2.$$

The left side is at least the smallest δ_t in the range times the positive sum of weights. Hence

$$\min_{M-1 \leq t \leq K-1} \delta_t \leq \frac{D_{M-1} + S^2 \sum_{t=M-1}^{K-1} \gamma_t^2}{\sum_{t=M-1}^{K-1} \gamma_t}.$$

There are $n = K - M + 1 \geq (T+1)/4$ indices. Their largest denominator in γ_t is $K+s$ and their smallest is $M+s$, so

$$\sum_{t=M-1}^{K-1} \gamma_t \geq \frac{2n}{\lambda(K+s)}, \qquad \sum_{t=M-1}^{K-1} \gamma_t^2 \leq \frac{4n}{\lambda^2 (M+s)^2}.$$

Also $K \leq 2M$, and therefore $K + s \leq 2(M+s)$. It follows that

$$\begin{aligned}
\min_{M-1 \leq t \leq K-1} \delta_t &\leq \frac{\lambda(K+s)}{2n} D_{M-1} + \frac{2S^2(K+s)}{\lambda(M+s)^2} \\
&\leq \frac{4\lambda(M+s)}{T+1} D_{M-1} + \frac{4S^2}{\lambda(M+s)} \\
&\leq \frac{4\lambda(s+1)^2 R^2}{(T+1)(M+s)} + \frac{16S^2}{\lambda(T+1)} + \frac{4S^2}{\lambda(M+s)} \\
&\leq \frac{4\lambda(s+1)^2 R^2}{(T+1)(M+s)} + \frac{32S^2}{\lambda(T+1)}.
\end{aligned}$$

In the third line, we use the sharper bound (82) at $M-1$, whose noise coefficient is 4, rather than the tail bound with coefficient 12. For the last line, use $M + s \geq (T+1)/4$.

The last companion iterate. Lemma 16, with $c = 2/\lambda$, adds at most

$$\begin{aligned}
64G^2\gamma_T &= \frac{128L^2(s+1)^2R^2}{\lambda(T+s+1)(M+s)^2} + \frac{512S^2}{\lambda(T+s+1)} \\
&\leq \frac{32\lambda(s+1)^2R^2}{(T+1)(M+s)} + \frac{512S^2}{\lambda(T+1)}.
\end{aligned}$$

For the initialization term, divide its coefficient by the coefficient of $\lambda(s+1)^2R^2/[(T+1)(M+s)]$:

$$\frac{128\kappa^2(T+1)}{(T+s+1)(M+s)} = 128\frac{\kappa^2}{M+s}\frac{T+1}{T+s+1} \leq 32.$$

Combining this with the preceding bound on the minimum over $M-1 \leq t \leq K-1$, and using $4 + 32 = 36$ and $32 + 512 = 544$, gives

$$\mathbb{E}[F(y_T) - F^*] \leq \frac{36\lambda(s+1)^2R^2}{(T+1)(M+s)} + \frac{544S^2}{\lambda(T+1)}. \tag{84}$$

Transfer to STORM. The comparison bound (58) gives

$$\mathbb{E}[F(x_T) - F^*] \leq 2\mathbb{E}[F(y_T) - F^*] + \frac{3L\beta^2\sigma^2}{\lambda^2}.$$

Substitute (84), then use $S^2 \le 9\sigma^2/4$ and $\beta = 1/(T+s+1)$:

$$\begin{aligned}\mathbb{E}[F(x_T) - F^*] &\le \frac{72\lambda(s+1)^2R^2}{(T+1)(M+s)} + \frac{1088S^2}{\lambda(T+1)} + \frac{3\kappa\sigma^2}{\lambda(T+s+1)^2} \\ &\le \frac{72\lambda(s+1)^2R^2}{(T+1)(M+s)} + \frac{2448\sigma^2}{\lambda(T+1)} + \frac{3\kappa\sigma^2}{\lambda(T+s+1)^2}.\end{aligned}$$

Finally, $s \ge 4\kappa^2 \ge \kappa$ and $(T+s+1)^2 \ge 4(T+1)s \ge \kappa(T+1)$ imply

$$\frac{3\kappa\sigma^2}{\lambda(T+s+1)^2} \le \frac{3\sigma^2}{\lambda(T+1)}.$$

Finally, $M + s \ge (T+1)/4$, so $72/[(T+1)(M+s)] \le 288/(T+1)^2$, which proves the bound in Theorem 8. □