Stochastic optimization under time drift: iterate averaging, step decay, and high-probability guarantees

Joshua Cutler
University of Washington
jocutler@uw.edu

Dmitriy Drusvyatskiy
University of Washington
ddrusv@uw.edu

Zaid Harchaoui
University of Washington
zaid@uw.edu

Abstract

We consider the problem of minimizing a convex function that is evolving in time according to unknown and possibly stochastic dynamics. Such problems abound in the machine learning and signal processing literature, under the names of concept drift and stochastic tracking. We provide novel non-asymptotic convergence guarantees for stochastic algorithms with iterate averaging, focusing on bounds valid both in expectation and with high probability. Notably, we show that the tracking efficiency of the proximal stochastic gradient method depends only logarithmically on the initialization quality when equipped with a step-decay schedule.

1 Introduction

Stochastic optimization underpins much of machine learning theory and practice. Significant progress has been made over the last two decades in the finite-time analysis of stochastic approximation algorithms; see, e.g., [1, 2, 6, 7, 8, 23, 26, 31, 32]. The predominant assumption in this line of work is that the distribution generating the data is fixed throughout the run of the process. There is no shortage of problems, however, where this assumption is grossly violated for reasons beyond the learner's control. Indeed, data often shifts and evolves over time for reasons that may be independent of the learning process.

Two examples are worth highlighting. The first is a classical problem in signal processing related to stochastic tracking [21, 29], wherein the learning algorithm aims to track over time a moving target driven by an unknown stochastic process. The second example is the concept drift phenomenon in online learning [15, 30], wherein the true hypothesis may be changing over time, as in topic modeling or spam classification. An important goal in online problems, and the one we adopt here, is to track as closely as possible an unknown sequence of minimizers or minimal values. The tracking error efficiency of stochastic algorithms in online settings is much less developed than sample complexity guarantees for static problems.

We present finite-time efficiency estimates in expectation and with high probability for the tracking error of the proximal stochastic gradient method under time drift. Our results concisely explain the interplay between the learning rate, the noise variance in the gradient oracle, and the strength of the time drift. The high-probability results merely assume that the gradient noise and time drift have light tails. Moreover, none of the results require the objectives to have bounded domains. While conventional wisdom and previous work recommend the use of constant step sizes under time drift, we show in an important regime that a significantly better step size schedule is one that is geometrically decaying to a "critical step size".

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

1.1 Related work

Our current work fits within the broader literature on stochastic tracking, online optimization with controlled increments, and high-probability guarantees in stochastic optimization. We now survey the most relevant literature in these areas.

Stochastic tracking. Stochastic gradient-type algorithms for stochastic tracking and filtering have been the subject of extensive research in the past century. Most works have focused on the so-called least mean-squares (LMS) algorithm and its variants, which can be viewed as a stochastic gradient method on a least-squares loss-based objective. Other stochastic algorithms that have been studied in these settings with a larger cost per iteration include recursive least-squares and the Kalman filter [13]. Recent works have revisited these methods from a more modern viewpoint [5, 25, 34]. In particular, the paper [25] focuses on (accelerated) gradient methods for deterministic tracking problems, while [34] analyzes a stochastic gradient method for online problems that is adaptive to unknown parameters. The paper [5] analyzes the dynamic regret of stochastic algorithms for time-varying problems, focusing both on lower and upper complexity bounds. Though the proof techniques in our paper share many aspects with those available in the literature, the results we obtain are distinct.

Online optimization with controlled increments. Online learning under concept drift was first considered by [24] and further developed in several papers [3, 18]. In this literature, the data distribution is typically fixed over time and the rate of variation is stated in terms of the probability of disagreement of consecutive target functions, which is assumed to be upper bounded. Another line of work assumes a time partitioning with an expert in each time interval, and the goal is to compete with the expert in each segment. Closer to this work is [15, 17], where in the framework of online convex optimization the bounds are stated in terms of maximum regret over any contiguous time interval; see also [5, 9, 28]. In contrast to these works, in our framework we state our bounds in the same spirit as in classical stochastic approximation, that is, in terms of distance to optimum and objective function gap, and we present results holding both in expectation and with high probability.

High-probability guarantees in stochastic optimization. A large part of our work revolves around high-probability guarantees in stochastic optimization. Classical references on the subject in static settings and for minimizing regret in online optimization include [4, 16, 22, 27]. There exists a variety of techniques for establishing high-probability guarantees based on Freedman's inequality and doubling tricks; see, e.g., [4, 16]. A more recent line of work [14] establishes a generalized Freedman inequality that is custom-tailored for analyzing stochastic gradient-type methods and results in best known high-probability guarantees. Our arguments closely follow the paradigm of [14] based on the generalized Freedman inequality.

1.2 Outline

The outline of the paper is as follows. Section 2 formalizes the problem setting of time-dependent stochastic optimization and records the relevant assumptions. Sections 3 and 4 summarize the main results of the paper. Specifically, Section 3 focuses on efficiency estimates for tracking the minimizer, while Section 4 focuses on efficiency estimates for tracking the minimal value. Proofs of the main results appear in Section 5. Illustrative numerical results appear in Section 6, and additional proofs appear in Appendix A.

2 Framework and assumptions

2.1 Stochastic optimization under time drift

Throughout Sections 2-4, we consider the sequence of stochastic optimization problems

$$\min_{x} \ \varphi_t(x) := f_t(x) + r_t(x) \tag{1}$$

indexed by time $t \in \mathbb{N}$. We make the standard standing assumption that (i) each function $f_t \colon \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and C^1-smooth with L-Lipschitz continuous gradient for some common parameters $\mu, L > 0$, and (ii) each regularizer $r_t \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is proper, closed and convex. The minimizer and minimal value of (1) will be denoted by $x_t^\star$ and $\varphi_t^\star$, respectively. Throughout, $\|\cdot\|$ denotes the ℓ_2-norm on $\mathbb{R}^d$ induced by the dot product $\langle\cdot,\cdot\rangle$.

As motivation, we describe two classical examples of (1) that are worth keeping in mind and that guide our framework: stochastic tracking of a drifting target and online learning under a distributional drift.

Example 2.1 (Stochastic tracking of a drifting target)**.** The problem of stochastic tracking, related to the filtering problem in signal processing, is to track a moving target $x_t^\star$ from observations

$$b_t = c_t(x_t^\star) + \epsilon_t,$$

where $c_t(\cdot)$ is a known measurement map and ϵ_t is a mean-zero noise vector. A typical time-dependent problem formulation takes the form

$$\min_x \mathop{\mathbb{E}}_{\epsilon_t} \ell_t(b_t - c_t(x)) + r_t(x),$$

where $\ell_t(\cdot)$ derives from the distribution of ϵ_t and $r_t(\cdot)$ encodes available side information about the target $x_t^\star$. Common choices for r_t are the 1-norm and the squared 2-norm. The motion of the target $x_t^\star$ is typically driven by a random walk or a diffusion [13, 29].

Example 2.2 (Online learning under distributional drift)**.** The problem of online learning under a distributional drift is to learn while the data distribution may change over time. More formally, one problem formulation takes the form

$$\min_x \mathop{\mathbb{E}}_{w\sim\mathcal{D}(u_t)} \ell(x,w) + r(x).$$

where $\mathcal{D}(u_t)$ is a data distribution that depends on an unknown parameter sequence $\{u_t\}$, which itself may evolve stochastically. The evolution of u_t is often assumed to be piecewise constant in t in online learning [15, 30].

Algorithm 1 Online Proximal Stochastic Gradient $\texttt{PSG}(x_0, \{\eta_t\}, T)$

Input: initial x_0 and step size sequence $\{\eta_t\}_{t=0}^T \subset (0,\infty)$
Step $t = 0,\ldots,T-1$:

Set $g_t = \widetilde{\nabla} f_t(x_t)$
Set $x_{t+1} = \mathrm{prox}_{\eta_t r_t}(x_t - \eta_t g_t)$

Return x_T

Online proximal stochastic gradient method. The main goal of a learning algorithm for problem (1) is to generate a sequence of points $\{x_t\}$ that minimize some natural performance metric. The most prevalent performance metrics in the literature are the *tracking error* and the *dynamic regret*. We will focus on two types of tracking error, $\|x_t - x_t^\star\|^2$ and $\varphi_t(x_t) - \varphi_t(x_t^\star)$.

We make the standing assumption that at every time t, and at every query point x, the learner may obtain an *unbiased estimator* $\widetilde{\nabla} f_t(x)$ of the true gradient $\nabla f_t(x)$ in order to proceed with a stochastic gradient-like optimization algorithm. With this oracle access, the online proximal stochastic gradient method—recorded as Algorithm 1 above—in each iteration t simply takes a stochastic gradient step on f_t at x_t followed by a proximal operation on r_t:

$$x_{t+1} := \mathrm{prox}_{\eta_t r_t}(x_t - \eta_t g_t) = \mathop{\arg\min}_{u\in\mathbb{R}^d} \left\{ r_t(u) + \tfrac{1}{2\eta_t}\|u - (x_t - \eta_t g_t)\|^2 \right\}.$$

The goal of our work is to obtain efficiency estimates for this procedure that hold both in expectation and with high probability.

Minimizer drift. The guarantees we obtain allow both the iterates x_t *and the minimizers* $x_t^\star$ to evolve stochastically. This is convenient for example when tracking a moving target $x_t^\star$ whose motion may be governed by a stochastic process such as a random walk or a diffusion (see Example 2.1). Throughout, we define the *minimizer drift* at time t to be the random variable

$$\Delta_t := \|x_t^\star - x_{t+1}^\star\|.$$

Clearly, an efficiency estimate for Algorithm 1 must take into account the variation of the functions f_t in time t. Two of the most popular metrics for measuring such variations are the minimizer drift Δ_t and the gradient variation $\sup_x \|\nabla f_t(x) - \nabla f_{t+1}(x)\|$. Given identical regularizers, a bound on the gradient variation always implies a bound on the minimizer drift.

Lemma 2.3 (Gradient variation vs. minimizer drift)**.** *Suppose $i, t \geq 0$ are such that the regularizers r_i and r_t are identical. Then we have*

$$\mu\|x_i^\star - x_t^\star\| \leq \|\nabla f_i(x_t^\star) - \nabla f_t(x_t^\star)\|.$$

Proof. Let r denote the common regularizer: $r = r_i = r_t$. Then the first-order optimality condition

$$0 \in \partial\varphi_t(x_t^\star) = \nabla f_t(x_t^\star) + \partial r(x_t^\star)$$

implies $-\nabla f_t(x_t^\star) \in \partial r(x_t^\star)$, so the vector $v := \nabla f_i(x_t^\star) - \nabla f_t(x_t^\star)$ lies in $\partial\varphi_i(x_t^\star)$. Hence the μ-strong convexity of φ_i and the inclusion $0 \in \partial\varphi_i(x_i^\star)$ imply $\mu\|x_i^\star - x_t^\star\| \leq \|0 - v\|$. □

2.2 Running assumption on the stochastic process

Setting the stage, given $\{x_t\}$ and $\{g_t\}$ as in Algorithm 1 we let

$$z_t := \nabla f_t(x_t) - g_t$$

denote the *gradient noise* at time t and we impose the following assumption modeling stochasticity in the online problem throughout Sections 3 and 4.

Assumption 2.4 (Stochastic framework)**.** There exists a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ with filtration $\mathbb{F} = (\mathcal{F}_t)_{t\geq 0}$ such that $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and the following holds for all $t \geq 0$:

(i) $x_t, x_t^\star \colon \Omega \to \mathbb{R}^d$ are $\mathcal{F}_t$-measurable,

(ii) $z_t \colon \Omega \to \mathbb{R}^d$ is $\mathcal{F}_{t+1}$-measurable with $\mathbb{E}[z_t \mid \mathcal{F}_t] = 0$.

The first item of Assumption 2.4 simply says that x_t and $x_t^\star$ are fully determined by information up to time t. The second item of Assumption 2.4 asserts that the gradient noise z_t is fully determined by information up to time $t + 1$ and has zero mean conditioned on the information up to time t; for example, this holds naturally in Example 2.2 if we take $g_t = \nabla\ell(x_t, w_t)$ with $w_t \sim \mathcal{D}(u_t)$ provided the loss $\ell(\cdot, w_t)$ is C^1-smooth.

3 Tracking the minimizer with the last iterate

In this section, we present bounds on the tracking error $\|x_t - x_t^\star\|^2$ that are valid both in expectation and with high probability under light-tail assumptions. Further, we show that a geometrically decaying learning rate schedule may be superior to a constant learning rate in terms of efficiency.

3.1 Bounds in expectation

We begin with bounding the expected value $\mathbb{E}\|x_t - x_t^\star\|^2$. The starting point for our analysis is the following standard one-step improvement guarantee.

Lemma 3.1 (One-step improvement)**.** *For all $x \in \mathbb{R}^d$, the iterates $\{x_t\}$ produced by Algorithm 1 with $\eta_t < 1/L$ satisfy the bound:*

$$2\eta_t(\varphi_t(x_{t+1}) - \varphi_t(x)) \leq (1 - \mu\eta_t)\|x_t - x\|^2 - \|x_{t+1} - x\|^2 + 2\eta_t\langle z_t, x_t - x\rangle + \tfrac{\eta_t^2}{1-L\eta_t}\|z_t\|^2.$$

For simplicity, we state the main results under the assumption that the second moments $\mathbb{E}\big[\Delta_t^2\big]$ and $\mathbb{E}\|z_t\|^2$ are uniformly bounded; more general guarantees that take into account weighted averages of the moments and allow for time-dependent learning rates follow from Lemma 3.1 as well.

Assumption 3.2 (Bounded second moments)**.** There exist constants $\Delta, \sigma > 0$ such that the following holds for all $t \geq 0$.

(i) **(Drift)** The minimizer drift Δ_t satisfies $\mathbb{E}\left[\Delta_t^2\right] \leq \Delta^2$.

(ii) **(Noise)** The gradient noise z_t satisfies $\mathbb{E}\|z_t\|^2 \leq \sigma^2$.

The following theorem establishes an expected improvement guarantee for Algorithm 1, and serves as the basis for much of what follows; see Section 5.1 for the precise statements and proofs of the present section.

Theorem 3.3 (Expected distance)**.** *Suppose that Assumption 3.2 holds. Then the iterates produced by Algorithm 1 with constant learning rate $\eta \leq 1/2L$ satisfy the bound:*

$$\mathbb{E}\|x_t - x_t^\star\|^2 \lesssim \underbrace{(1-\mu\eta)^t\|x_0 - x_0^\star\|^2}_{\textit{optimization}} + \underbrace{\frac{\eta\sigma^2}{\mu}}_{\textit{noise}} + \underbrace{\left(\frac{\Delta}{\mu\eta}\right)^2}_{\textit{drift}}.$$

Interplay of optimization, noise, and drift. Theorem 3.3 states that when using a constant learning rate, the error $\mathbb{E}\|x_t - x_t^\star\|^2$ decays linearly in time t, until it reaches the "noise+drift" error $\eta\sigma^2/\mu + (\Delta/\mu\eta)^2$. Notice that the "noise+drift" error cannot be made arbitrarily small. This is perfectly in line with intuition: a learning rate that is too small prevents the algorithm from catching up with $x_t^\star$. We note that the individual error terms due to the optimization and noise are classically known to be tight for PSG; tightness of the drift term is proved in [25, Theorem 3.2].

With Theorem 3.3 in hand, we are led to define the following asymptotic tracking error of Algorithm 1 corresponding to $\mathbb{E}\|x_t - x_t^\star\|^2$, together with the corresponding optimal step size:

$$\mathcal{E} := \min_{\eta\in(0,1/2L]}\left\{\frac{\eta\sigma^2}{\mu} + \left(\frac{\Delta}{\mu\eta}\right)^2\right\} \quad \text{and} \quad \eta_\star := \min\left\{\frac{1}{2L}, \left(\frac{2\Delta^2}{\mu\sigma^2}\right)^{1/3}\right\}.$$

Plugging $\eta_\star$ into the definition of $\mathcal{E}$, we see that Algorithm 1 exhibits qualitatively different behaviors in settings corresponding to high or low drift-to-noise ratio Δ/σ, explicitly given by

$$\mathcal{E} \asymp \begin{cases} \frac{\sigma^2}{\mu L} + \left(\frac{L\Delta}{\mu}\right)^2 & \text{if } \frac{\Delta}{\sigma} \geq \sqrt{\frac{\mu}{16L^3}} \\ \left(\frac{\Delta\sigma^2}{\mu^2}\right)^{2/3} & \text{otherwise.} \end{cases}$$

Two regimes of variation are brought to light by the above computation: the *high drift-to-noise regime* $\Delta/\sigma \geq \sqrt{\mu/16L^3}$, and the *low drift-to-noise regime* $\Delta/\sigma < \sqrt{\mu/16L^3}$. The high drift-to-noise regime is uninteresting from the viewpoint of stochastic optimization because the optimal learning rate is as large as in the deterministic setting, $\eta_\star = 1/2L$. In contrast, the low drift-to-noise regime is interesting because the optimal learning rate $\eta_\star = (2\Delta^2/\mu\sigma^2)^{1/3}$ is smaller than $1/2L$ and exhibits a nontrivial scaling with the problem parameters.

Learning rate vs. rate of variation. A central question is to find a learning rate schedule that achieves a tracking error $\mathbb{E}\|x_t - x_t^\star\|^2$ that is within a constant factor of $\mathcal{E}$ in the shortest possible time. The answer is clear in the high drift-to-noise regime $\Delta/\sigma \geq \sqrt{\mu/16L^3}$. Indeed, in this case, Theorem 3.3 directly implies that Algorithm 1 with the constant learning rate $\eta_\star = 1/2L$ will find a point x_t satisfying $\mathbb{E}\|x_t - x_t^\star\|^2 \lesssim \mathcal{E}$ in time $t \lesssim (L/\mu)\log(\|x_0 - x_0^\star\|^2/\mathcal{E})$. Notice that the efficiency estimate is logarithmic in $1/\mathcal{E}$; intuitively, the reason for the absence of a sublinear component is that the error due to the drift Δ dominates the error due to the variance σ^2 in the stochastic gradient.

The low drift-to-noise regime $\Delta/\sigma < \sqrt{\mu/16L^3}$ is more subtle. Namely, the simplest strategy is to execute Algorithm 1 with the constant learning rate $\eta_\star = (2\Delta^2/\mu\sigma^2)^{1/3}$. Then a direct application of Theorem 3.3 yields the estimate $\mathbb{E}\|x_t - x_t^\star\|^2 \lesssim \mathcal{E}$ in time $t \lesssim (\sigma^2/\mu^2\mathcal{E})\log(\|x_0 - x_0^\star\|^2/\mathcal{E})$. This efficiency estimate can be significantly improved by gradually decaying the learning rate using a "step-decay schedule", wherein the algorithm is implemented in epochs with the new learning rate chosen to be the midpoint between the current learning rate and $\eta_\star$. Such schedules are well known to improve efficiency in the static setting, as was discovered in [11, 12], and can be used here. The end result is the following theorem (see Theorem 5.5 for the precise statement).

Theorem 3.4 (Time to track in expectation, informal)**.** *Suppose that Assumption 3.2 holds. Then there is a learning rate schedule $\{\eta_t\}$ such that Algorithm 1 produces a point x_t satisfying*

$$\mathbb{E}\|x_t - x_t^\star\|^2 \lesssim \mathcal{E} \quad \textit{after time} \quad t \lesssim \frac{L}{\mu}\log\left(\frac{\|x_0 - x_0^\star\|^2}{\mathcal{E}}\right) + \frac{\sigma^2}{\mu^2\mathcal{E}}.$$

Remarkably, the efficiency estimate in Theorem 3.4 looks identical to the efficiency estimate in the classical static setting [22], with $\mathcal{E}$ playing the role of the target accuracy ε. Theorems 3.3 and 3.4 provide useful baseline guarantees for the performance of Algorithm 1. Nonetheless, these guarantees are all stated in terms of the *expected* tracking error $\mathbb{E}\|x_t - x_t^\star\|^2$, and are therefore only meaningful if the entire algorithm can be repeated from scratch multiple times. There is no shortage of situations in which a learning algorithm is operating in real time and the time drift is irreversible; in such settings, the algorithm may only be executed once. Such settings call for efficiency estimates that hold with high probability, rather than only in expectation.

3.2 High-probability guarantees

We next present high-probability guarantees on the tracking error $\|x_t - x_t^\star\|^2$. To this end, we make the following standard light-tail assumptions on the minimizer drift and gradient noise [14, 22, 26].

Assumption 3.5 (Sub-Gaussian drift and noise)**.** There exist constants $\Delta, \sigma > 0$ such that the following holds for all $t \geq 0$.

(i) **(Drift)** The drift Δ_t^2 is sub-exponential conditioned on $\mathcal{F}_t$ with parameter Δ^2:

$$\mathbb{E}\big[\exp(\lambda\Delta_t^2) \,|\, \mathcal{F}_t\big] \leq \exp(\lambda\Delta^2) \quad \text{for all} \quad 0 \leq \lambda \leq \Delta^{-2}.$$

(ii) **(Noise)** The noise z_t is norm sub-Gaussian conditioned on $\mathcal{F}_t$ with parameter $\sigma/2$:

$$\mathbb{P}\big\{\|z_t\| \geq \tau \,|\, \mathcal{F}_t\big\} \leq 2\exp(-2\tau^2/\sigma^2) \quad \text{for all} \quad \tau > 0.$$

Note that the first item of Assumption 3.5 is equivalent to asserting that the minimizer drift Δ_t is sub-Gaussian conditioned on $\mathcal{F}_t$. Clearly Assumption 3.5 implies Assumption 3.2 with the same constants Δ, σ. It is worthwhile to note some common settings in which Assumption 3.5 holds; the claims in Remark 3.6 follow from standard results on sub-Gaussian random variables [19, 33].

Remark 3.6 (Common settings for Assumption 3.5)**.** Fix constants $\Delta, \sigma > 0$. If Δ_t is bounded by Δ, then clearly Δ_t^2 is sub-exponential (conditioned on $\mathcal{F}_t$) with parameter Δ^2. Similarly, if $\|z_t\|$ is bounded by σ, then z_t is norm sub-Gaussian (conditioned on $\mathcal{F}_t$) with parameter $\sigma/2$ (by Markov's inequality). Alternatively, if the increment $x_t^\star - x_{t+1}^\star$ is mean-zero sub-Gaussian conditioned on $\mathcal{F}_t$ with parameter $\Delta/\sqrt{d}$, then $x_t^\star - x_{t+1}^\star$ is mean-zero norm sub-Gaussian conditioned on $\mathcal{F}_t$ with parameter $2\sqrt{2}\cdot\Delta$ and hence Δ_t^2 is sub-exponential conditioned on $\mathcal{F}_t$ with parameter $c\cdot\Delta^2$ for some absolute constant $c > 0$. Similarly, if z_t is sub-Gaussian conditioned on $\mathcal{F}_t$ with parameter $\sigma/4\sqrt{2d}$, then z_t is norm sub-Gaussian conditioned on $\mathcal{F}_t$ with parameter $\sigma/2$.

The following theorem shows that if Assumption 3.5 holds, then the expected bound on $\|x_t - x_t^\star\|^2$ derived in Theorem 3.3 holds with high probability.

Theorem 3.7 (High-probability distance tracking)**.** *Suppose that Assumption 3.5 holds and let $\{x_t\}$ be the iterates produced by Algorithm 1 with constant learning rate $\eta \leq 1/2L$. Then there is an absolute constant $c > 0$ such that for any specified $t \in \mathbb{N}$ and $\delta \in (0,1)$, the estimate*

$$\|x_t - x_t^\star\|^2 \leq \left(1 - \tfrac{\mu\eta}{2}\right)^t\|x_0 - x_0^\star\|^2 + c\left(\frac{\eta\sigma^2}{\mu} + \left(\frac{\Delta}{\mu\eta}\right)^2\right)\log\left(\frac{e}{\delta}\right) \tag{2}$$

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

The proof of Theorem 3.7 employs a technique used in [14]. The main idea is to build a careful recursion for the moment generating function of $\|x_t - x_t^\star\|^2$, leading to a one-sided sub-exponential tail bound. As a consequence of Theorem 3.7, we can again implement a step-decay schedule in the low drift-to-noise regime to obtain the following efficiency estimate with high probability; see Section 5.2 for the formal statements and proofs.

Theorem 3.8 (Time to track with high probability, informal)**.** *Suppose that Assumption 3.5 holds and that we are in the low drift-to-noise regime $\Delta/\sigma < \sqrt{\mu/16L^3}$. Then there is a learning rate schedule $\{\eta_t\}$ such that for any specified $\delta \in (0,1)$, Algorithm 1 produces a point x_t satisfying*

$$\|x_t - x_t^\star\|^2 \lesssim \mathcal{E} \log\left(\frac{e}{\delta}\right)$$

with probability at least $1 - K\delta$ after time

$$t \lesssim \frac{L}{\mu} \log\left(\frac{\|x_0 - x_0^\star\|^2}{\mathcal{E}}\right) + \frac{\sigma^2}{\mu^2 \mathcal{E}}, \quad \textit{where } K \lesssim \log_2\left(\frac{1}{L} \cdot \left(\frac{\sigma^2 \mu}{\Delta^2}\right)^{1/3}\right).$$

4 Tracking the minimal value

The results outlined so far have focused on tracking the minimizer $x_t^\star$. In this section, we present results for tracking the minimal value $\varphi_t^\star$. These two goals are fundamentally different. Generally speaking, good bounds on the function gap along with strong convexity imply good bounds on the distance to the minimizer; the reverse implication is false. To this end, we require a stronger assumption on the variation of the functions f_t in time t: rather than merely controlling the minimizer drift Δ_t, we will assume control on the *gradient drift*

$$G_{i,t} := \sup_x \|\nabla f_i(x) - \nabla f_t(x)\|.$$

Our strategy is to track the minimal value along a running average $\hat{x}_t$ of the iterates x_t produced by Algorithm 1, as defined in Algorithm 2 below. The reason behind using this particular running average is brought to light in Section 5.3, where we apply a standard averaging technique (Lemma A.1) to a one-step improvement along x_t (Lemma 5.10) to obtain the desired progress along $\hat{x}_t$ (Proposition 5.11).

Algorithm 2 Averaged Online Proximal Stochastic Gradient $\overline{\texttt{PSG}}(x_0, \mu, \{\eta_t\}, T)$

Input: initial $x_0 =: \hat{x}_0$, strong convexity parameter μ, and step size sequence $\{\eta_t\}_{t=0}^T \subset (0, 2\mu^{-1})$
Step $t = 0, \ldots, T-1$:

Set $g_t = \widetilde{\nabla} f_t(x_t)$
Set $x_{t+1} = \operatorname{prox}_{\eta_t r_t}(x_t - \eta_t g_t)$
Set $\hat{x}_{t+1} = \left(1 - \frac{\mu\eta_t}{2-\mu\eta_t}\right)\hat{x}_t + \frac{\mu\eta_t}{2-\mu\eta_t} x_{t+1}$

Return $\hat{x}_T$

4.1 Bounds in expectation

We begin with bounding the expected value $\mathbb{E}[\varphi_t(\hat{x}_t) - \varphi_t^\star]$. Analogous to Assumption 3.2, we make the following assumption regarding drift and noise.

Assumption 4.1 (Bounded second moments)**.** The regularizers $r_t \equiv r$ are identical for all times t and there exist constants $\Delta, \sigma > 0$ such that the following properties hold for all $0 \le i < t$:

(i) **(Drift)** The gradient drift $G_{i,t}$ satisfies $\mathbb{E}\big[G_{i,t}^2\big] \le (\mu\Delta|i-t|)^2$.

(ii) **(Noise)** The gradient noise z_i satisfies $\mathbb{E}\|z_i\|^2 \le \sigma^2$ and $\mathbb{E}\langle z_i, x_t^\star\rangle = 0$.

These two assumptions are natural indeed. Taking into account Lemma 2.3, it is clear that Assumption 4.1 implies the earlier Assumption 3.2 with the same constants Δ, σ. The assumption on the drift intuitively asserts that gradient drift $G_{i,t}$ can grow only linearly in time $|i-t|$ (in expectation). In particular, returning to Example 2.2, suppose that the distribution map $\mathcal{D}(\cdot)$ is γ-Lipschitz continuous in the Wasserstein-1 distance, the loss $\ell(\cdot, w)$ is C^1 smooth for all w, and the gradient $\nabla\ell(x, \cdot)$ is β-Lipschitz continuous for all x. Then the Kantorovich-Rubinstein duality theorem directly implies $\mathbb{E}\big[G_{i,t}^2\big] \le (\gamma\beta)^2 \mathbb{E}\|u_i - u_t\|^2$. Therefore, as long as the second moment $\mathbb{E}\|u_i - u_t\|^2$ scales quadratically in $|i-t|$, the desired drift assumption holds. The assumption on the noise requires a uniform bound on the second moment $\mathbb{E}\|z_i\|^2$ and for the condition $\mathbb{E}\langle z_i, x_t^\star\rangle = 0$ to hold. The latter property confers a weak form of uncorrelatedness between the gradient noise z_i and the future minimizer $x_t^\star$, and holds automatically if the gradient noise and the minimizers evolve independently of each other, as would typically be the case for instance in Example 2.2.

The following theorem provides an expected improvement guarantee for Algorithm 2.

Theorem 4.2 (Expected function gap). *Suppose that Assumption 4.1 holds, and let $\{\hat{x}_t\}$ be the iterates produced by Algorithm 2 with constant learning rate $\eta \le 1/2L$. Then the following bound holds for all $t \ge 0$:*

$$\mathbb{E}\big[\varphi_t(\hat{x}_t) - \varphi_t^\star\big] \lesssim \underbrace{\left(1 - \tfrac{\mu\eta}{2}\right)^t\left(\varphi_0(x_0) - \varphi_0^\star\right)}_{\text{optimization}} + \underbrace{\eta\sigma^2}_{\text{noise}} + \underbrace{\frac{\Delta^2}{\mu\eta^2}}_{\text{drift}}.$$

The "noise+drift" error term in Theorem 4.2 coincides with μ times the corresponding error term in Theorem 3.3, as expected due to μ-strong convexity. With Theorem 4.2 in hand, we are led to define the following asymptotic tracking error of Algorithm 2 corresponding to $\mathbb{E}[\varphi_t(\hat{x}_t) - \varphi_t^\star]$:

$$\mathcal{G} := \mu\mathcal{E} = \min_{\eta\in(0,1/2L]} \left\{\eta\sigma^2 + \frac{\Delta^2}{\mu\eta^2}\right\}.$$

The corresponding asymptotically optimal choice of η is again given by $\eta_\star$, and the dichotomy governed by the drift-to-noise ratio Δ/σ remains:

$$\mathcal{G} \asymp \begin{cases} \frac{\sigma^2}{L} + \frac{(L\Delta)^2}{\mu} & \text{if } \frac{\Delta}{\sigma} \ge \sqrt{\frac{\mu}{16L^3}} \\ \mu\left(\frac{\Delta\sigma^2}{\mu^2}\right)^{2/3} & \text{otherwise.} \end{cases}$$

In the high drift-to-noise regime $\Delta/\sigma \ge \sqrt{\mu/16L^3}$, Theorem 4.2 directly implies that Algorithm 2 with the constant learning rate $\eta_\star = 1/2L$ finds a point $\hat{x}_t$ satisfying $\mathbb{E}[\varphi_t(\hat{x}_t) - \varphi_t^\star] \lesssim \mathcal{G}$ in time $t \lesssim (L/\mu)\log((\varphi_0(x_0) - \varphi_0^\star)/\mathcal{G})$. In the low drift-to-noise regime $\Delta/\sigma < \sqrt{\mu/16L^3}$, another direct application of Theorem 4.2 shows that Algorithm 2 with the constant learning rate $\eta_\star = (2\Delta^2/\mu\sigma^2)^{1/3}$ finds a point $\hat{x}_t$ satisfying $\mathbb{E}[\varphi_t(\hat{x}_t) - \varphi_t^\star] \lesssim \mathcal{G}$ in time $t \lesssim (\sigma^2/\mu\mathcal{G})\log((\varphi_0(x_0) - \varphi_0^\star)/\mathcal{G})$. As before, this efficiency estimate can be significantly improved by implementing a step-decay schedule. The end result is the following theorem; see Section 5.3 for the formal statements and proofs.

Theorem 4.3 (Time to track in expectation, informal). *Suppose that Assumption 4.1 holds. Then there is a learning rate schedule $\{\eta_t\}$ such that Algorithm 2 produces a point $\hat{x}_t$ satisfying*

$$\mathbb{E}[\varphi_t(\hat{x}_t) - \varphi_t^\star] \lesssim \mathcal{G} \quad \textit{after time} \quad t \lesssim \frac{L}{\mu}\log\left(\frac{\varphi_0(x_0) - \varphi_0^\star}{\mathcal{G}}\right) + \frac{\sigma^2}{\mu\mathcal{G}}.$$

4.2 High-probability guarantees

Our next result is an analogue of Theorem 4.2 that holds with high probability. Naturally, such a result should rely on light-tail assumptions on the gradient drift $G_{i,t}$ and the norm of the gradient noise $\|z_i\|$. We state the guarantee under the assumption that $G_{i,t}$ and $\|z_i\|$ are conditionally sub-Gaussian (Assumption 4.4). In particular, we require for the first time that the gradient noise z_i is mean-zero conditioned on the σ-algebra

$$\mathcal{F}_{i,t} := \sigma(\mathcal{F}_i, x_t^\star)$$

for all $0 \le i < t$; the property $\mathbb{E}[z_i \mid \mathcal{F}_{i,t}] = 0$ would follow from independence of the gradient noise z_i and the future minimizer $x_t^\star$ and is reasonable in light of Examples 2.1 and 2.2.

Assumption 4.4 (Sub-Gaussian drift and noise). The regularizers $r_t \equiv r$ are identical for all times t and there exist constants $\Delta, \sigma > 0$ such that the following properties hold for all $0 \le i < t$.

(i) **(Drift)** The square gradient drift $G_{i,t}^2$ is sub-exponential with parameter $(\mu\Delta|i-t|)^2$:

$$\mathbb{E}\big[\exp\left(\lambda G_{i,t}^2\right)\big] \le \exp\left(\lambda(\mu\Delta|i-t|)^2\right) \quad \text{for all} \quad 0 \le \lambda \le (\mu\Delta|i-t|)^{-2}.$$

(ii) **(Noise)** The gradient noise z_i is mean-zero norm sub-Gaussian conditioned on $\mathcal{F}_{i,t}$ with parameter $\sigma/2$, i.e., $\mathbb{E}[z_i \mid \mathcal{F}_{i,t}] = 0$ and

$$\mathbb{P}\big\{\|z_i\| \geq \tau \mid \mathcal{F}_{i,t}\big\} \leq 2\exp(-2\tau^2/\sigma^2) \quad \text{for all} \quad \tau \geq 0.$$

Clearly the chain of implications holds:

$$\text{Assumption 4.4} \implies \text{Assumption 4.1} \implies \text{Assumption 3.2.}$$

Example 4.5 (Sub-Gaussian feature model)**.** In the setting of logistic regression, sub-Gaussian gradient noise naturally arises from sampling a sub-Gaussian feature model. Indeed, in this case the objective takes the form $f(x) = \mathbb{E}_{A,b}\left[\sum_{i=1}^{n} \log(1 + \exp\langle a_i, x\rangle) - \langle Ax, b\rangle\right]$ and drawing (A, b) yields the sample gradient $\widetilde{\nabla} f(x) = A^T(S(Ax) - b)$, where $A \in \mathbb{R}^{n\times d}$ has rows $a_1, ..., a_n \in \mathbb{R}^d$ and S denotes the sigmoid function. Being that S and b are bounded, it therefore follows that if the rows of A are sub-Gaussian, then so is the gradient noise $\nabla f(x) - \widetilde{\nabla} f(x)$.

The following theorem shows that if Assumption 4.4 holds, then the expected bound on $\varphi_t(\hat{x}_t) - \varphi_t^\star$ derived in Theorem 4.2 holds with high probability.

Theorem 4.6 (Function gap with high probability)**.** *Suppose that Assumption 4.4 holds, and let $\{\hat{x}_t\}$ be the iterates produced by Algorithm 2 with constant learning rate $\eta \leq 1/2L$. Then there is an absolute constant $c > 0$ such that for any specified $t \in \mathbb{N}$ and $\delta \in (0, 1)$, the estimate*

$$\varphi_t(\hat{x}_t) - \varphi_t^\star \leq c\left(\left(1 - \tfrac{\mu\eta}{2}\right)^t(\varphi_0(x_0) - \varphi_0^\star) + \eta\sigma^2 + \frac{\Delta^2}{\mu\eta^2}\right)\log\left(\frac{e}{\delta}\right) \tag{3}$$

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

The proof of Theorem 4.6 is based on combining the generalized Freedman inequality of [14] with careful control on the drift and noise in improvement guarantees for the proximal stochastic gradient method. The key observation is that although we do not have simple recursive control on the moment generating function of $\varphi_t(\hat{x}_t) - \varphi_t^\star$ (as we do with $\|x_t - x_t^\star\|^2$), we *can* control the tracking error $\varphi_t(\hat{x}_t) - \varphi_t^\star$ by leveraging control on the martingale $\sum_{i=0}^{t-1}\langle z_i, x_i - x_t^\star\rangle\zeta^{t-1-i}$, where $\zeta = 1 - \mu\eta/(2 - \mu\eta)$. This martingale is self-regulating in the sense that its total conditional variance is bounded by the history of the process; the generalized Freedman inequality is precisely suited to bound such martingales with high probability.

With Theorem 4.6 in hand, we may implement a step-decay schedule as before to obtain the following efficiency estimate; see Section 5.4 for the formal statements and proofs.

Theorem 4.7 (Time to track with high probability, informal)**.** *Suppose that Assumption 4.4 holds and that we are in the low drift-to-noise regime $\Delta/\sigma < \sqrt{\mu/16L^3}$. Fix $\delta \in (0, 1)$. Then there is a learning rate schedule $\{\eta_t\}$ such that Algorithm 2 produces a point $\hat{x}_t$ satisfying*

$$\varphi_t(\hat{x}_t) - \varphi_t^\star \lesssim \mathcal{G}\log\left(\frac{e}{\delta}\right)$$

with probability at least $1 - K\delta$ after time

$$t \lesssim \frac{L}{\mu}\log\left(\frac{\varphi_0(x_0) - \varphi_0^\star}{\mathcal{G}}\right) + \frac{\sigma^2}{\mu\mathcal{G}}\log\left(\log\left(\frac{e}{\delta}\right)\right), \quad \textit{where } K \lesssim \log_2\left(\frac{1}{L}\cdot\left(\frac{\sigma^2\mu}{\Delta^2}\right)^{1/3}\right).$$

5 Proofs of results in Sections 3 and 4

Roadmap. Throughout this section, we enforce the assumptions and notation of Section 2 and let $\{x_t\}$ denote the iterates generated by Algorithm 1 with $\eta_t < 1/L$. Sections 5.1 and 5.2 handle distance tracking under time drift: Section 5.1 derives the results of Section 3.1, while Section 5.2 derives the results of Section 3.2. Then Sections 5.3 and 5.4 handle function gap tracking under time drift: Section 5.3 derives the results of Section 4.1, while Section 5.4 derives the results of Section 4.2.

5.1 Tracking the minimizer: bounds in expectation

The proof of Theorem 3.3 follows a familiar pattern in stochastic optimization. We begin by recalling Lemma 3.1, which gives a standard one-step improvement guarantee [22] for the proximal stochastic gradient method on the fixed problem $\min \varphi_t$.

Lemma 5.1 (One-step improvement)**.** *The estimate*

$$2\eta_t(\varphi_t(x_{t+1}) - \varphi_t(x)) \le (1-\mu\eta_t)\|x_t - x\|^2 - \|x_{t+1} - x\|^2 + 2\eta_t\langle z_t, x_t - x\rangle + \tfrac{\eta_t^2}{1-L\eta_t}\|z_t\|^2$$

holds for all points $x \in \mathbb{R}^d$ and for all indices $t \ge 0$.

Proof. Since f_t is L-smooth, we have

$$\begin{aligned}\varphi_t(x_{t+1}) &= f_t(x_{t+1}) + r_t(x_{t+1})\\ &\le f_t(x_t) + \langle \nabla f_t(x_t), x_{t+1} - x_t\rangle + \tfrac{L}{2}\|x_{t+1} - x_t\|^2 + r_t(x_{t+1})\\ &= f_t(x_t) + r_t(x_{t+1}) + \langle g_t, x_{t+1} - x_t\rangle + \tfrac{L}{2}\|x_{t+1} - x_t\|^2 + \langle z_t, x_{t+1} - x_t\rangle.\end{aligned}$$

Next, given any $\delta_t > 0$, Young's inequality yields

$$\langle z_t, x_{t+1} - x_t\rangle \le \tfrac{\delta_t}{2}\|z_t\|^2 + \tfrac{1}{2\delta_t}\|x_{t+1} - x_t\|^2.$$

Therefore, given any $x \in \mathbb{R}^d$, we have

$$\begin{aligned}\varphi_t(x_{t+1}) &\le f_t(x_t) + r_t(x_{t+1}) + \langle g_t, x_{t+1} - x_t\rangle + \tfrac{\delta_t^{-1}+L}{2}\|x_{t+1} - x_t\|^2 + \tfrac{\delta_t}{2}\|z_t\|^2\\ &= f_t(x_t) + r_t(x_{t+1}) + \langle g_t, x_{t+1} - x_t\rangle + \tfrac{1}{2\eta_t}\|x_{t+1} - x_t\|^2\\ &\quad + \tfrac{\delta_t^{-1}+L-\eta_t^{-1}}{2}\|x_{t+1} - x_t\|^2 + \tfrac{\delta_t}{2}\|z_t\|^2\\ &\le f_t(x_t) + r_t(x) + \langle g_t, x - x_t\rangle + \tfrac{1}{2\eta_t}\|x - x_t\|^2 - \tfrac{1}{2\eta_t}\|x - x_{t+1}\|^2\\ &\quad + \tfrac{\delta_t^{-1}+L-\eta_t^{-1}}{2}\|x_{t+1} - x_t\|^2 + \tfrac{\delta_t}{2}\|z_t\|^2,\end{aligned}$$

where the last inequality holds because $x_{t+1} = \mathrm{prox}_{\eta_t r_t}(x_t - \eta_t g_t)$ is the minimizer of the η_t^{-1}-strongly convex function $r_t + \langle g_t, \cdot - x_t\rangle + \frac{1}{2\eta_t}\|\cdot - x_t\|^2$. Now we estimate

$$\begin{aligned}f_t(x_t) + r_t(x) + \langle g_t, x - x_t\rangle &= f_t(x_t) + \langle \nabla f_t(x_t), x - x_t\rangle + r_t(x) + \langle z_t, x_t - x\rangle\\ &\le f_t(x) - \tfrac{\mu}{2}\|x - x_t\|^2 + r_t(x) + \langle z_t, x_t - x\rangle\\ &= \varphi_t(x) - \tfrac{\mu}{2}\|x - x_t\|^2 + \langle z_t, x_t - x\rangle\end{aligned}$$

using the μ-strong convexity of f_t. Thus,

$$\begin{aligned}\varphi_t(x_{t+1}) \le \varphi_t(x) - \tfrac{\mu}{2}\|x - x_t\|^2 + \langle z_t, x_t - x\rangle + \tfrac{1}{2\eta_t}\|x - x_t\|^2 - \tfrac{1}{2\eta_t}\|x - x_{t+1}\|^2\\ + \tfrac{\delta_t^{-1}+L-\eta_t^{-1}}{2}\|x_{t+1} - x_t\|^2 + \tfrac{\delta_t}{2}\|z_t\|^2.\end{aligned}$$

Finally, taking $\delta_t = \eta_t/(1 - L\eta_t)$ and rearranging (note that $\varphi_t(x_{t+1})$ is finite) yields

$$2\eta_t(\varphi_t(x_{t+1}) - \varphi_t(x)) \le (1-\mu\eta_t)\|x_t - x\|^2 - \|x_{t+1} - x\|^2 + 2\eta_t\langle z_t, x_t - x\rangle + \tfrac{\eta_t^2}{1-L\eta_t}\|z_t\|^2,$$

as claimed. □

It is critically important that the one-step improvement estimate in Lemma 5.1 holds with respect to any reference point x. In particular, setting $x = x_t^\star$ yields the following lemma.

Lemma 5.2 (Distance recursion)**.** *The estimate*

$$\|x_{t+1}-x_{t+1}^\star\|^2 \le (1-\mu\eta_t)\|x_t-x_t^\star\|^2 + 2\eta_t\langle z_t, x_t-x_t^\star\rangle + \tfrac{\eta_t^2}{1-L\eta_t}\|z_t\|^2 + \left(1+\tfrac{1}{\mu\eta_t}\right)\Delta_t^2$$

holds for all indices $t \ge 0$.

Proof. Note that the μ-strong convexity of φ_t implies $\frac{\mu}{2}\|x_{t+1}-x_t^\star\|^2 \le \varphi_t(x_{t+1})-\varphi_t^\star$. Combining this estimate with Lemma 5.1 under the identification $x = x_t^\star$ yields

$$(1+\mu\eta_t)\|x_{t+1}-x_t^\star\|^2 \le (1-\mu\eta_t)\|x_t-x_t^\star\|^2 + 2\eta_t\langle z_t, x_t-x_t^\star\rangle + \tfrac{\eta_t^2}{1-L\eta_t}\|z_t\|^2.$$

Next, an application of Young's inequality reveals

$$\|x_{t+1}-x_{t+1}^\star\|^2 \le (1+\mu\eta_t)\|x_{t+1}-x_t^\star\|^2 + (1+(\mu\eta_t)^{-1})\|x_t^\star-x_{t+1}^\star\|^2,$$

thereby completing the proof. □

Applying Lemma 5.2 recursively furnishes a bound on $\|x_t-x_t^\star\|^2$. When the step size is constant, the next proposition follows immediately.

Proposition 5.3 (Last-iterate progress)**.** *Suppose $\eta_t \equiv \eta$. Then the following bound holds for all $t \ge 0$:*

$$\|x_t-x_t^\star\|^2 \le (1-\mu\eta)^t\|x_0-x_0^\star\|^2 + 2\eta\sum_{i=0}^{t-1}\langle z_i, x_i-x_i^\star\rangle(1-\mu\eta)^{t-1-i}$$
$$+ \tfrac{\eta^2}{1-L\eta}\sum_{i=0}^{t-1}\|z_i\|^2(1-\mu\eta)^{t-1-i} + \left(1+\tfrac{1}{\mu\eta}\right)\sum_{i=0}^{t-1}\Delta_i^2(1-\mu\eta)^{t-1-i}.$$

By taking expectations in Proposition 5.3, we obtain the following precise version of Theorem 3.3.

Corollary 5.4 (Expected distance)**.** *Suppose that Assumption 3.2 holds. Then the iterates $\{x_t\}$ generated by Algorithm 1 with constant learning rate $\eta \le 1/2L$ satisfy the bound:*

$$\mathbb{E}\|x_t-x_t^\star\|^2 \le (1-\mu\eta)^t\|x_0-x_0^\star\|^2 + 2\left(\frac{\eta\sigma^2}{\mu} + \left(\frac{\Delta}{\mu\eta}\right)^2\right).$$

With Corollary 5.4 in hand, we can now prove an expected efficiency estimate for the online proximal stochastic gradient method using a step-decay schedule, wherein the algorithm is implemented in epochs with the new learning rate chosen to be the midpoint between the current learning rate and $\eta_\star$. The following is the formal statement of Theorem 3.4 (as previously noted, in the high drift-to-noise regime $\Delta/\sigma \ge \sqrt{\mu/16L^3}$, Theorem 3.4 holds trivially with the constant learning rate $\eta_\star = 1/2L$). The argument is close in spirit to the justifications of the restart schemes in [11, 12].

Theorem 5.5 (Time to track in expectation)**.** *Suppose that Assumption 3.2 holds and that we are in the low drift-to-noise regime $\Delta/\sigma < \sqrt{\mu/16L^3}$. Set $\eta_\star = (2\Delta^2/\mu\sigma^2)^{1/3}$ and $\mathcal{E} = (\Delta\sigma^2/\mu^2)^{2/3}$. Suppose moreover that we have available a positive upper bound on the initial square distance $D \ge \|x_0-x_0^\star\|^2$. Consider running Algorithm 1 in $k = 0,\ldots,K-1$ epochs, namely, set $X_0 = x_0$ and iterate the process*

$$X_{k+1} = \texttt{PSG}(X_k,\eta_k,T_k) \quad \textit{for} \quad k = 0,\ldots,K-1,$$

where the number of epochs is

$$K = 1 + \left\lceil \log_2\left(\frac{1}{L}\cdot\left(\frac{\sigma^2\mu}{\Delta^2}\right)^{1/3}\right)\right\rceil$$

and we set

$$\eta_0 = \frac{1}{2L}, \quad T_0 = \left\lceil \frac{2L}{\mu}\log\left(\frac{\mu L D}{\sigma^2}\right)^+\right\rceil \quad \textit{and} \quad \eta_k = \frac{\eta_{k-1}+\eta_\star}{2}, \quad T_k = \left\lceil\frac{\log(4)}{\mu\eta_k}\right\rceil \quad \forall k \ge 1.$$

Then the time horizon $T = T_0 + \cdots + T_{K-1}$ satisfies

$$T \lesssim \frac{L}{\mu}\log\left(\frac{\mu L D}{\sigma^2}\right)^+ + \frac{\sigma^2}{\mu^2\mathcal{E}} \le \frac{L}{\mu}\log\left(\frac{D}{\mathcal{E}}\right)^+ + \frac{\sigma^2}{\mu^2\mathcal{E}},$$

while the corresponding tracking error satisfies $\mathbb{E}\|X_K - X_K^\star\|^2 \lesssim \mathcal{E}$, where $X_K^\star$ denotes the minimizer of φ_T.

Proof. For each index k, let $t_k := T_0 + \cdots + T_{k-1}$ (with $t_0 := 0$), $X_k^\star$ be the minimizer of the corresponding function φ_{t_k}, and

$$E_k := \frac{2}{\mu}\left(\eta_k\sigma^2 + \frac{\Delta^2}{\mu\eta_\star^2}\right).$$

Then taking into account $\eta_k \ge \eta_\star$, Corollary 5.4 directly implies

$$\begin{aligned}\mathbb{E}\|X_{k+1} - X_{k+1}^\star\|^2 &\le (1-\mu\eta_k)^{T_k}\mathbb{E}\|X_k - X_k^\star\|^2 + \frac{2}{\mu}\left(\eta_k\sigma^2 + \frac{\Delta^2}{\mu\eta_k^2}\right)\\ &\le e^{-\mu\eta_k T_k}\mathbb{E}\|X_k - X_k^\star\|^2 + E_k.\end{aligned}$$

We will verify by induction that the estimate $\mathbb{E}\|X_{k+1} - X_{k+1}^\star\|^2 \le 2E_k$ holds for all indices k. To see the base case, observe

$$\mathbb{E}\|X_1 - X_1^\star\|^2 \le e^{-\mu\eta_0 T_0}\|X_0 - X_0^\star\|^2 + E_0 \le 2E_0.$$

Assume next that the claim holds for index $k-1$. We then conclude

$$\begin{aligned}\mathbb{E}\|X_{k+1} - X_{k+1}^\star\|^2 &\le e^{-\mu\eta_k T_k}\mathbb{E}\|X_k - X_k^\star\|^2 + E_k\\ &\le \frac{1}{4}\mathbb{E}\|X_k - X_k^\star\|^2 + E_k \le \frac{E_k}{2E_{k-1}}\mathbb{E}\|X_k - X_k^\star\|^2 + E_k \le 2E_k,\end{aligned}$$

thereby completing the induction. Hence $\mathbb{E}\|X_K - X_K^\star\|^2 \le 2E_{K-1}$. Next, observe

$$E_{K-1} - \sqrt[3]{54}\left(\frac{\Delta\sigma^2}{\mu^2}\right)^{2/3} = \frac{2\sigma^2}{\mu}(\eta_{K-1} - \eta_\star) = \frac{2\sigma^2}{\mu}\cdot\frac{\eta_0 - \eta_\star}{2^{K-1}} \le \left(\frac{\Delta\sigma^2}{\mu^2}\right)^{2/3},$$

so

$$\mathbb{E}\|X_K - X_K^\star\|^2 \le 2(1+\sqrt[3]{54})\left(\frac{\Delta\sigma^2}{\mu^2}\right)^{2/3} \asymp \mathcal{E}.$$

Finally, note

$$T \lesssim \frac{L}{\mu}\log\left(\frac{\mu L D}{\sigma^2}\right)^+ + \frac{1}{\mu}\sum_{k=1}^{K-1}\frac{1}{\eta_k}$$

and

$$\sum_{k=1}^{K-1}\frac{1}{\eta_k} \le 2L\sum_{k=1}^{K-1}2^k \le 2L\cdot 2^K = 8L\cdot 2^{K-2} \le 8\left(\frac{\sigma^2\mu}{\Delta^2}\right)^{1/3} = \frac{8\sigma^2}{\mu}\cdot\left(\frac{\Delta\sigma^2}{\mu^2}\right)^{-2/3} \asymp \frac{\sigma^2}{\mu\mathcal{E}}.$$

This completes the proof. □

5.2 Tracking the minimizer: high-probability guarantees

The proof strategy of Theorem 3.7 follows a similar argument as in [14, Claim D.1], which recursively controls the moment generating function of $\|x_t - x_t^\star\|^2$. Namely, Lemma 5.2 in the regime $\eta_t \le 1/2L$ directly yields

$$\|x_{t+1} - x_{t+1}^\star\|^2 \le (1-\mu\eta_t)\|x_t - x_t^\star\|^2 + 2\eta_t\langle z_t, u_t\rangle\|x_t - x_t^\star\| + 2\eta_t^2\|z_t\|^2 + \tfrac{2}{\mu\eta_t}\Delta_t^2, \quad (4)$$

where we set $u_t := \frac{x_t - x_t^\star}{\|x_t - x_t^\star\|}$ if x_t is distinct from $x_t^\star$ and set it to zero otherwise. The right-hand side has the form of a contraction factor, gradient noise, and drift. The goal is now to control the moment generating function $\mathbb{E}\big[e^{\lambda\|x_t - x_t^\star\|^2}\big]$ through this recursion. The basic probabilistic tool for similar settings under bounded noise assumptions was developed in [14]. The following proposition is a slight generalization of [14, Claim D.1] to a light-tail setting.

Proposition 5.6 (Recursive control on MGF)**.** *Consider scalar stochastic processes (V_t), (D_t), and (X_t) on a probability space with filtration $(\mathcal{H}_t)$, which are linked by the inequality*

$$V_{t+1} \leq \alpha_t V_t + D_t \sqrt{V_t} + X_t + \kappa_t$$

for some deterministic constants $\alpha_t \in (-\infty, 1]$ and $\kappa_t \in \mathbb{R}$. Suppose the following properties hold.

- *V_t is nonnegative and $\mathcal{H}_t$-measurable.*
- *D_t is mean-zero sub-Gaussian conditioned on $\mathcal{H}_t$ with deterministic parameter σ_t:*
$$\mathbb{E}[\exp(\lambda D_t) \,|\, \mathcal{H}_t] \leq \exp(\lambda^2 \sigma_t^2/2) \quad \textit{for all} \quad \lambda \in \mathbb{R}.$$
- *X_t is nonnegative and sub-exponential conditioned on $\mathcal{H}_t$ with deterministic parameter ν_t:*
$$\mathbb{E}[\exp(\lambda X_t) \,|\, \mathcal{H}_t] \leq \exp(\lambda \nu_t) \quad \textit{for all} \quad 0 \leq \lambda \leq 1/\nu_t.$$

Then the estimate

$$\mathbb{E}[\exp(\lambda V_{t+1})] \leq \exp\big(\lambda(\nu_t + \kappa_t)\big)\mathbb{E}\left[\exp\left(\lambda\left(\frac{1+\alpha_t}{2}\right)V_t\right)\right]$$

holds for any λ satisfying $0 \leq \lambda \leq \min\left\{\frac{1-\alpha_t}{2\sigma_t^2}, \frac{1}{2\nu_t}\right\}$.

Proof. For any index t and any scalar $\lambda \geq 0$, the tower rule implies

$$\begin{aligned}\mathbb{E}[\exp(\lambda V_{t+1})] &\leq \mathbb{E}\big[\exp\big(\lambda\big(\alpha_t V_t + D_t\sqrt{V_t} + X_t + \kappa_t\big)\big)\big] \\ &= \exp(\lambda\kappa_t)\mathbb{E}\Big[\exp(\lambda\alpha_t V_t)\mathbb{E}\big[\exp\big(\lambda D_t\sqrt{V_t}\big)\exp(\lambda X_t) \,|\, \mathcal{H}_t\big]\Big].\end{aligned}$$

Hölder's inequality in turn yields

$$\begin{aligned}\mathbb{E}\big[\exp\big(\lambda D_t\sqrt{V_t}\big)\exp(\lambda X_t) \,|\, \mathcal{H}_t\big] &\leq \sqrt{\mathbb{E}\big[\exp\big(2\lambda\sqrt{V_t}D_t\big) \,|\, \mathcal{H}_t\big] \cdot \mathbb{E}[\exp(2\lambda X_t) \,|\, \mathcal{H}_t]} \\ &\leq \sqrt{\exp(2\lambda^2 V_t\sigma_t^2)\exp(2\lambda\nu_t)} \\ &= \exp(\lambda^2\sigma_t^2 V_t)\exp(\lambda\nu_t)\end{aligned}$$

provided $0 \leq \lambda \leq \frac{1}{2\nu_t}$. Thus, if $0 \leq \lambda \leq \min\left\{\frac{1-\alpha_t}{2\sigma_t^2}, \frac{1}{2\nu_t}\right\}$, then the following estimate holds:

$$\begin{aligned}\mathbb{E}[\exp(\lambda V_{t+1})] &\leq \exp(\lambda\kappa_t)\mathbb{E}\big[\exp(\lambda\alpha_t V_t)\exp(\lambda^2\sigma_t^2 V_t)\exp(\lambda\nu_t)\big] \\ &= \exp\big(\lambda(\nu_t + \kappa_t)\big)\mathbb{E}\big[\exp(\lambda(\alpha_t + \lambda\sigma_t^2)V_t)\big] \\ &\leq \exp\big(\lambda(\nu_t + \kappa_t)\big)\mathbb{E}\left[\exp\left(\lambda\left(\frac{1+\alpha_t}{2}\right)V_t\right)\right].\end{aligned}$$

The proof is complete. □

We may now use Proposition 5.6 to derive the following precise version of Theorem 3.7.

Theorem 5.7 (High-probability distance tracking)**.** *Suppose that Assumption 3.5 holds and let $\{x_t\}$ be the iterates produced by Algorithm 1 with constant learning rate $\eta \leq 1/2L$. Then there exists an absolute constant*[1] *$c > 0$ such that for any specified $t \in \mathbb{N}$ and $\delta \in (0, 1)$, the estimate*

$$\|x_t - x_t^\star\|^2 \leq \left(1 - \tfrac{\mu\eta}{2}\right)^t \|x_0 - x_0^\star\|^2 + \left(\frac{8\eta(c\sigma)^2}{\mu} + 4\left(\frac{\Delta}{\mu\eta}\right)^2\right)\log\left(\frac{e}{\delta}\right)$$

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

[1] Explicitly, one can take any $c \geq 1$ such that $\|z_t\|^2$ is sub-exponential conditioned on $\mathcal{F}_t$ with parameter $c\sigma^2$ and z_t is mean-zero sub-Gaussian conditioned on $\mathcal{F}_t$ with parameter $c\sigma$ for all t.

Proof. Note first that under Assumption 3.5, there exists an absolute constant $c \geq 1$ such that $\|z_t\|^2$ is sub-exponential conditioned on $\mathcal{F}_t$ with parameter $c\sigma^2$ and z_t is mean-zero sub-Gaussian conditioned on $\mathcal{F}_t$ with parameter $c\sigma$ for all t. Therefore $\langle z_t, u_t\rangle$ is mean-zero sub-Gaussian conditioned on $\mathcal{F}_t$ with parameter $c\sigma$, while Δ_t^2 is sub-exponential conditioned on $\mathcal{F}_t$ with parameter Δ^2 by assumption. Thus, in light of inequality (4), we may apply Proposition 5.6 with $\mathcal{H}_t = \mathcal{F}_t$, $V_t = \|x_t - x_t^\star\|^2$, $D_t = 2\eta_t\langle z_t, u_t\rangle$, $X_t = 2\eta_t^2\|z_t\|^2 + 2\Delta_t^2/\mu\eta_t$, $\alpha_t = 1-\mu\eta_t$, $\kappa_t = 0$, $\sigma_t = 2\eta_t c\sigma$, and $\nu_t = 2\eta_t^2 c\sigma^2 + 2\Delta^2/\mu\eta_t$, yielding the estimate

$$\mathbb{E}\big[\exp\big(\lambda\|x_{t+1} - x_{t+1}^\star\|^2\big)\big] \leq \exp\left(\lambda\left(2\eta_t^2 c\sigma^2 + \frac{2\Delta^2}{\mu\eta_t}\right)\right)\mathbb{E}\big[\exp\big(\lambda\big(1-\tfrac{\mu\eta_t}{2}\big)\|x_t - x_t^\star\|^2\big)\big] \tag{5}$$

for all

$$0 \leq \lambda \leq \min\left\{\frac{\mu}{8\eta_t(c\sigma)^2}, \frac{1}{4\eta_t^2 c\sigma^2 + 4\Delta^2/\mu\eta_t}\right\}.$$

Taking into account $\eta_t \equiv \eta$ and iterating the recursion (5), we deduce

$$\begin{aligned}\mathbb{E}\big[\exp\big(\lambda\|x_t - x_t^\star\|^2\big)\big] &\leq \exp\left(\lambda\big(1-\tfrac{\mu\eta}{2}\big)^t\|x_0 - x_0^\star\|^2 + \lambda\left(2\eta^2 c\sigma^2 + \frac{2\Delta^2}{\mu\eta}\right)\sum_{i=0}^{t-1}\big(1-\tfrac{\mu\eta}{2}\big)^i\right)\\ &\leq \exp\left(\lambda\left(\big(1-\tfrac{\mu\eta}{2}\big)^t\|x_0 - x_0^\star\|^2 + \frac{4\eta c\sigma^2}{\mu} + 4\Big(\frac{\Delta}{\mu\eta}\Big)^2\right)\right)\end{aligned}$$

for all

$$0 \leq \lambda \leq \min\left\{\frac{\mu}{8\eta(c\sigma)^2}, \frac{1}{4\eta^2 c\sigma^2 + 4\Delta^2/\mu\eta}\right\}.$$

Moreover, setting

$$\nu := \frac{8\eta(c\sigma)^2}{\mu} + 4\Big(\frac{\Delta}{\mu\eta}\Big)^2$$

and taking into account $c \geq 1$ and $\mu\eta \leq 1$, we have

$$\frac{4\eta c\sigma^2}{\mu} + 4\Big(\frac{\Delta}{\mu\eta}\Big)^2 \leq \nu$$

and

$$\frac{1}{\nu} = \frac{\mu}{8\eta(c\sigma)^2 + 4\Delta^2/\mu\eta^2} \leq \min\left\{\frac{\mu}{8\eta(c\sigma)^2}, \frac{1}{4\eta^2 c\sigma^2 + 4\Delta^2/\mu\eta}\right\}.$$

Hence

$$\mathbb{E}\Big[\exp\Big(\lambda\Big(\|x_t - x_t^\star\|^2 - \big(1-\tfrac{\mu\eta}{2}\big)^t\|x_0 - x_0^\star\|^2\Big)\Big)\Big] \leq \exp(\lambda\nu) \quad \text{for all} \quad 0 \leq \lambda \leq 1/\nu.$$

Taking $\lambda = 1/\nu$ and applying Markov's inequality completes the proof. □

With Theorem 3.7 in hand, we can now prove a high-probability efficiency estimate for the online proximal stochastic gradient method using a step-decay schedule. The following theorem is the precise form of Theorem 3.8. The argument follows the same reasoning as in the proof of Theorem 5.5, with Theorem 3.7 playing the role of Corollary 5.4 while using a union bound over the epochs. The proof appears in the appendix (see Section A.1).

Theorem 5.8 (Time to track with high probability)**.** *Suppose that Assumption 3.5 holds and that we are in the low drift-to-noise regime $\Delta/\sigma < \sqrt{\mu/16L^3}$. Set $\eta_\star = (2\Delta^2/\mu\sigma^2)^{1/3}$ and $\mathcal{E} = (\Delta\sigma^2/\mu^2)^{2/3}$. Suppose moreover that we have available an upper bound on the initial square distance $D \geq \|x_0 - x_0^\star\|^2$. Consider running Algorithm 1 in $k = 0, \ldots, K-1$ epochs, namely, set $X_0 = x_0$ and iterate the process*

$$X_{k+1} = \texttt{PSG}(X_k, \eta_k, T_k) \quad \textit{for} \quad k = 0, \ldots, K-1,$$

where the number of epochs is

$$K = 1 + \left\lceil \log_2\left(\frac{1}{L}\cdot\left(\frac{\sigma^2\mu}{\Delta^2}\right)^{1/3}\right)\right\rceil$$

and we set

$$\eta_0 = \frac{1}{2L}, \quad T_0 = \left\lceil \frac{4L}{\mu} \log\left(\frac{\mu LD}{\sigma^2}\right)^+ \right\rceil \quad \textit{and} \quad \eta_k = \frac{\eta_{k-1} + \eta_\star}{2}, \quad T_k = \left\lceil \frac{2\log(4)}{\mu \eta_k} \right\rceil \qquad \forall k \geq 1.$$

Then the time horizon $T = T_0 + \cdots + T_{K-1}$ *satisfies*

$$T \lesssim \frac{L}{\mu} \log\left(\frac{\mu LD}{\sigma^2}\right)^+ + \frac{\sigma^2}{\mu^2 \mathcal{E}} \leq \frac{L}{\mu} \log\left(\frac{D}{\mathcal{E}}\right)^+ + \frac{\sigma^2}{\mu^2 \mathcal{E}},$$

and for any specified $\delta \in (0,1)$*, the corresponding tracking error satisfies*

$$\|X_K - X_K^\star\|^2 \lesssim \mathcal{E} \log\left(\frac{e}{\delta}\right)$$

with probability at least $1 - K\delta$*, where* $X_K^\star$ *denotes the minimizer of* φ_T*.*

5.3 Tracking the minimal value: bounds in expectation

We turn now to tracking the minimal value. Henceforth, we suppose $\eta_t \leq 1/2L$ and that the regularizers $r_t \equiv r$ are identical for all times t. Setting the stage, fix a time horizon t. Then Lemma 5.1 directly yields the following one-step improvement guarantee for all indices i:

$$2\eta_i(\varphi_i(x_{i+1}) - \varphi_i(x_t^\star)) \leq (1 - \mu\eta_i)\|x_i - x_t^\star\|^2 - \|x_{i+1} - x_t^\star\|^2 + 2\eta_i\langle z_i, x_i - x_t^\star\rangle + 2\eta_i^2\|z_i\|^2.$$

Notice that this provides an estimate on the "wrong quantity" $\varphi_i(x_{i+1}) - \varphi_i(x_t^\star)$, whereas we would like to obtain an estimate on the suboptimality gap $\varphi_t(x_{i+1}) - \varphi_t(x_t^\star)$. In words, we would like to replace φ_i with φ_t, while the controlling the incurred error. Lemma 5.9 shows that the incurred error can by controlled by the gradient drift $G_{i,t}$, while Lemma 5.10 deduces the desired one-step improvement guarantee on φ_t.

Lemma 5.9 (Gradient drift vs. gap variation)**.** *For all indices* $i, t \in \mathbb{N}$ *and points* $x, y \in \operatorname{dom} r$*, the estimate holds:*

$$\big|[\varphi_i(y) - \varphi_i(x)] - [\varphi_t(y) - \varphi_t(x)]\big| \leq G_{i,t}\|y - x\|.$$

Proof. Taking into account $r_t \equiv r$ and using the fundamental theorem of calculus, we may write

$$\begin{aligned}[\varphi_i(y) - \varphi_i(x)] - [\varphi_t(y) - \varphi_t(x)] &= \int_0^1 \langle \nabla f_i(x + s(y-x)) - \nabla f_t(x + s(y-x)), y - x\rangle \, ds \\ &\leq G_{i,t}\|y - x\|,\end{aligned}$$

where the last inequality follows from Cauchy-Schwarz. Switching x and y completes the proof. □

Lemma 5.10 (One-step improvement)**.** *For all indices* $i, t \in \mathbb{N}$*, points* $x \in \operatorname{dom} r$*, and arbitrary* $\alpha > 0$*, we have*

$$\begin{aligned}2\eta_i(\varphi_t(x_{i+1}) - \varphi_t(x)) \leq (1 - \mu\eta_i)\|x_i - x\|^2 &- (1 - \alpha\eta_i)\|x_{i+1} - x\|^2 \\ &+ 2\eta_i\langle z_i, x_i - x\rangle + 2\eta_i^2\|z_i\|^2 + \tfrac{\eta_i}{\alpha} G_{i,t}^2.\end{aligned}$$

Proof. This follows immediately from combining Lemmas 5.1 and 5.9 and Young's inequality, $2G_{i,t}\|x_{i+1} - x\| \leq \alpha^{-1}G_{i,t}^2 + \alpha\|x_{i+1} - x\|^2$. □

Turning the estimate in Lemma 5.10 into an efficiency guarantee on the average iterate is essentially standard and follows for example from the averaging techniques in [10, 11, 12, 20]. The resulting progress along the average iterate is summarized in the following proposition, while the description of the key averaging lemma is placed in the appendix (see Section A.2).

Proposition 5.11 (Progress along the average iterate)**.** *Let* $\{\hat{x}_t\}$ *be the iterates produced by Algorithm 2 with constant step size* $\eta \leq 1/2L$*; thus, setting* $\hat{\rho} := \mu\eta/(2 - \mu\eta)$*, we have* $\hat{x}_0 = x_0$ *and*

$\hat{x}_t = (1-\hat{\rho})\,\hat{x}_{t-1} + \hat{\rho} x_t$ *for all* $t \geq 1$. *Then the following bound holds for all* $t \geq 0$ *and* $x \in \operatorname{dom} r$*:*

$$\varphi_t(\hat{x}_t) - \varphi_t(x) \leq (1-\hat{\rho})^t \left(\varphi_t(x_0) - \varphi_t(x) + \tfrac{\mu}{4}\|x_0 - x\|^2\right) + \hat{\rho}\sum_{i=0}^{t-1}\langle z_i, x_i - x\rangle (1-\hat{\rho})^{t-1-i}$$
$$+ \hat{\rho}\eta \sum_{i=0}^{t-1}\|z_i\|^2(1-\hat{\rho})^{t-1-i} + \tfrac{\hat{\rho}}{\mu}\sum_{i=0}^{t-1} G_{i,t}^2 (1-\hat{\rho})^{t-1-i}.$$

Proof. Setting $\alpha = \mu/2$ in Lemma 5.10, we obtain the following recursion for all indices $k \geq 0$ and $t \geq 1$ and points $x \in \operatorname{dom} r$:

$$\rho(\varphi_k(x_t) - \varphi_k(x)) \leq (1 - c_1\rho)V_{t-1} - (1 + c_2\rho)V_t + \omega_t,$$

where $\rho = 2\eta$, $c_1 = \mu/2$, $c_2 = -\mu/4$, $V_i = \|x_i - x\|^2$, and $\omega_t = 2\eta\langle z_{t-1}, x_{t-1} - x\rangle + 2\eta^2\|z_{t-1}\|^2 + (2\eta/\mu)G_{t-1,k}^2$. The result follows by applying the averaging Lemma A.1 with $h = \varphi_t - \varphi_t(x)$. ☐

Taking expectations in Proposition 5.11, we obtain the following precise version of Theorem 4.2.

Corollary 5.12 (Expected function gap)**.** *Suppose that Assumption 4.1 holds, let* $\{\hat{x}_t\}$ *be the iterates produced by Algorithm 2 with constant step size* $\eta \leq 1/2L$*, and set* $\hat{\rho} := \mu\eta/(2-\mu\eta)$*. Then the following bound holds for all* $t \geq 0$*:*

$$\mathbb{E}\big[\varphi_t(\hat{x}_t) - \varphi_t^\star\big] \leq (1-\hat{\rho})^t \cdot \mathbb{E}\big[\varphi_t(x_0) - \varphi_t^\star + \tfrac{\mu}{4}\|x_0 - x_t^\star\|^2\big] + \eta\sigma^2 + \frac{8\Delta^2}{\mu\eta^2}. \tag{6}$$

Consequently, we have

$$\mathbb{E}\big[\varphi_t(\hat{x}_t) - \varphi_t^\star\big] \lesssim (1-\hat{\rho})^t\big(\varphi_0(x_0) - \varphi_0^\star\big) + \eta\sigma^2 + \frac{\Delta^2}{\mu\eta^2}$$

for all $t \geq 0$*, and the following asymptotic error bound holds:*

$$\limsup_{t\to\infty} \mathbb{E}\big[\varphi_t(\hat{x}_t) - \varphi_t^\star\big] \leq \eta\sigma^2 + \frac{8\Delta^2}{\mu\eta^2}.$$

Proof. The bound (6) follows by setting $x = x_t^\star$ in Proposition 5.11, taking expectations, and noting

$$\sum_{i=0}^{t-1}\mathbb{E}\|z_i\|^2(1-\hat{\rho})^{t-1-i} \leq \frac{\sigma^2}{\hat{\rho}} \qquad \text{and} \qquad \sum_{i=0}^{t-1}\mathbb{E}\big[G_{i,t}^2\big](1-\hat{\rho})^{t-1-i} \leq \frac{(\mu\Delta)^2(2-\hat{\rho})}{\hat{\rho}^3}$$

by Assumption 4.1. Next, applying Lemma 5.9 and Young's inequality together with the μ-strong convexity of φ_0 and Lemma 2.3 yields

$$\varphi_t(x_0) - \varphi_t^\star + \tfrac{\mu}{4}\|x_0 - x_t^\star\|^2 \leq 3\big(\varphi_0(x_0) - \varphi_0^\star\big) + 2\mu^{-1}G_{0,t}^2, \tag{7}$$

and then taking expectations and invoking Assumption 4.1 gives

$$\mathbb{E}\big[\varphi_t(x_0) - \varphi_t^\star + \tfrac{\mu}{4}\|x_0 - x_t^\star\|^2\big] \leq 3\big(\varphi_0(x_0) - \varphi_0^\star\big) + 2\mu\Delta^2 t^2. \tag{8}$$

Further, the inequality

$$e^{-\mu\eta t/2}\mu t^2 \leq 16/\mu\eta^2 \qquad \forall \mu,\eta,t > 0 \tag{9}$$

combines with inequality (8) to yield

$$(1-\hat{\rho})^t \cdot \mathbb{E}\big[\varphi_t(x_0) - \varphi_t^\star + \tfrac{\mu}{4}\|x_0 - x_t^\star\|^2\big] \leq 3(1-\hat{\rho})^t\big(\varphi_0(x_0) - \varphi_0^\star\big) + \frac{32\Delta^2}{\mu\eta^2}$$

and the remaining assertions of the corollary follow. ☐

We may now apply Corollary 5.12 to obtain the formal version of Theorem 4.3; the proof closely follows that of Theorem 5.5 and is included in the appendix (see Section A.3).

Theorem 5.13 (Time to track in expectation)**.** *Suppose that Assumption 4.1 holds and that we are in the low drift-to-noise regime* $\Delta/\sigma < \sqrt{\mu/16L^3}$*. Set* $\eta_\star = (2\Delta^2/\mu\sigma^2)^{1/3}$ *and* $\mathcal{G} = \mu(\Delta\sigma^2/\mu^2)^{2/3}$*. Suppose moreover that we have available a positive upper bound on the initial gap* $D \geq \varphi_0(x_0) - \varphi_0^\star$*.*

Consider running Algorithm 2 in $k = 0, \ldots, K-1$ epochs, namely, set $X_0 = x_0$ and iterate the process

$$X_{k+1} = \overline{\texttt{PSG}}(X_k, \mu, \eta_k, T_k) \quad \textit{for} \quad k = 0, \ldots, K-1,$$

where the number of epochs is

$$K = 1 + \left\lceil \log_2 \left(\frac{1}{L} \cdot \left(\frac{\sigma^2 \mu}{\Delta^2} \right)^{1/3} \right) \right\rceil$$

and we set

$$\eta_0 = \frac{1}{2L}, \quad T_0 = \left\lceil \frac{4L}{\mu} \log \left(\frac{LD}{\sigma^2} \right)^+ \right\rceil \quad \textit{and} \quad \eta_k = \frac{\eta_{k-1} + \eta_\star}{2}, \quad T_k = \left\lceil \frac{2\log(12)}{\mu \eta_k} \right\rceil \quad \forall k \geq 1.$$

Then the time horizon $T = T_0 + \cdots + T_{K-1}$ satisfies

$$T \lesssim \frac{L}{\mu} \log \left(\frac{LD}{\sigma^2} \right)^+ + \frac{\sigma^2}{\mu \mathcal{G}} \leq \frac{L}{\mu} \log \left(\frac{D}{\mathcal{G}} \right)^+ + \frac{\sigma^2}{\mu \mathcal{G}}$$

and the corresponding tracking error satisfies $\mathbb{E}[\varphi_T(X_K) - \varphi_T^\star] \lesssim \mathcal{G}$.

5.4 Tracking the minimal value: high-probability guarantees

In this section, we derive the high-probability analogues of the results in Section 5.3. In light of Proposition 5.11, we seek upper bounds on the sums

$$\sum_{i=0}^{t-1} \langle z_i, x_i - x_t^\star \rangle (1-\hat{\rho})^{t-1-i}, \quad \sum_{i=0}^{t-1} \|z_i\|^2 (1-\hat{\rho})^{t-1-i}, \quad \sum_{i=0}^{t-1} G_{i,t}^2 (1-\hat{\rho})^{t-1-i}$$

that hold with high probability. The last two sums can easily be estimated under boundedness or light-tail assumptions on $\|z_i\|$ and $G_{i,t}$. Controlling the first sum is more challenging because the error $\|x_i - x_t^\star\|$ may in principle grow large. In order to control this term, we use a remarkable generalization of Freedman's inequality, recently proved in [14] for the purpose of analyzing the stochastic gradient method on static nonsmooth problems (without a regularizer).

The main idea is as follows. Fix a horizon t, assume $\mathbb{E}[z_i \mid \mathcal{F}_{i,t}] = 0$ for all $0 \leq i < t$ (recall that $\mathcal{F}_{i,t} := \sigma(\mathcal{F}_i, x_t^\star)$), and define the martingale difference sequence

$$d_i := \langle z_i, x_i - x_t^\star \rangle (1-\hat{\rho})^{t-1-i}$$

adapted to the filtration $(\mathcal{F}_{i+1,t})_{i=0}^{t-1}$. Roughly speaking, under mild light-tail assumptions, the total conditional variance of the corresponding martingale $\sum_{i=0}^{t-1} d_i$ can be bounded above with high probability by an affine transformation of itself, i.e., by an affine combination of the sequence $\{d_i\}_{i=0}^{t-1}$. In this way, the martingale is self-regulating. This is the content of the following proposition. The proof follows from Lemma 5.10 and algebraic manipulation and is placed in the appendix (see Section A.4).

Proposition 5.14 (Self-regulation)**.** *The iterates $\{x_t\}$ produced by Algorithm 1 with $r_t \equiv r$ and constant step size $\eta \leq 1/2L$ satisfy the following bound for all $\lambda \in (0, \mu\eta]$:*

$$\begin{aligned}\sum_{i=0}^{t-1} \|x_i - x_t^\star\|^2 (1-\lambda)^{2(t-1-i)} \leq \sum_{j=0}^{t-2} &\left(2\eta \sum_{i=j+1}^{t-1} (1-\lambda)^{t-2-i} \right) \langle z_j, x_j - x_t^\star \rangle (1-\lambda)^{t-1-j} \\ &+ \tfrac{1}{\lambda}(1-\lambda)^{t-1} \|x_0 - x_t^\star\|^2 + \tfrac{2\eta^2}{\lambda} \sum_{j=0}^{t-2} \|z_j\|^2 (1-\lambda)^{t-2-j} \\ &+ \tfrac{\eta}{\mu\lambda} \sum_{j=0}^{t-2} G_{j,t}^2 (1-\lambda)^{t-2-j}.\end{aligned}$$

In order to bound the self-regulating martingale $\sum_{i=0}^{t-1} d_i$, we use the following direct consequence of the generalized Freedman inequality developed in [14].

Theorem 5.15 (Consequence of generalized Freedman)**.** *Let $(D_i)_{i=0}^n$ and $(V_i)_{i=0}^n$ be scalar stochastic processes on a probability space with filtration $(\mathcal{H}_i)_{i=0}^{n+1}$ satisfying*

$$\mathbb{E}[\exp(\lambda D_i) \mid \mathcal{H}_i] \leq \exp(\lambda^2 V_i/2) \quad \textit{for all} \quad \lambda \geq 0.$$

Suppose that D_i is $\mathcal{H}_{i+1}$-measurable with $\mathbb{E}|D_i| < \infty$ and $\mathbb{E}[D_i \mid \mathcal{H}_i] = 0$, and that V_i is nonnegative and $\mathcal{H}_i$-measurable. Suppose moreover that there are constants $\alpha_0, \ldots, \alpha_n \geq 0$, $\delta \in [0,1]$, and $\beta(\delta) \geq 0$ satisfying

$$\mathbb{P}\left\{\sum_{i=0}^n V_i \leq \sum_{i=0}^n \alpha_i D_i + \beta(\delta)\right\} \geq 1 - \delta.$$

Set $\alpha := \max\{\alpha_0, \ldots, \alpha_n\}$. Then for all $\tau > 0$, the following bound holds:

$$\mathbb{P}\left\{\sum_{i=0}^n D_i \geq \tau\right\} \leq \delta + \exp\left(-\frac{\tau}{4\alpha + 8\beta(\delta)/\tau}\right).$$

Combining Proposition 5.14 and Theorem 5.15 yields the following tail bound for $\sum_{i=0}^{t-1} d_i$.

Proposition 5.16 (Noise martingale tail bound)**.** *Suppose that Assumption 4.4 holds, let $\{x_t\}$ be the iterates produced by Algorithm 1 with constant step size $\eta \leq 1/2L$, and set $\hat{\rho} := \mu\eta/(2-\mu\eta)$. Then there is an absolute constant $c > 0$ such that for any specified $t \in \mathbb{N}$, $\delta \in (0,1)$, and $\tau > 0$, the following bound holds:*

$$\mathbb{P}\left\{\sum_{i=0}^{t-1} \langle z_i, x_i - x_t^\star\rangle (1-\hat{\rho})^{t-1-i} \geq \tau\right\} \leq \delta + \exp\left(-\frac{\tau}{4\alpha + 8\beta_t \log(3e/\delta)/\tau}\right),$$

where $\alpha := 3\eta(c\sigma)^2/\hat{\rho}$ and

$$\beta_t := (1-\hat{\rho})^{t-1}\left(\|x_0 - x_0^\star\|^2 + \Delta^2 t^2\right)\frac{2(c\sigma)^2}{\hat{\rho}} + \frac{2\eta^2(c\sigma)^4}{\hat{\rho}^2} + \frac{3\mu\Delta^2\eta(c\sigma)^2}{\hat{\rho}^4}.$$

Proof. By Assumption 4.4, there exists an absolute constant $c \geq 1$ such that $\|z_i\|^2$ is sub-exponential conditioned on $\mathcal{F}_{i,t}$ with parameter $c\sigma^2$ and z_i is mean-zero sub-Gaussian conditioned on $\mathcal{F}_{i,t}$ with parameter $c\sigma$ for all indices $0 \leq i < t$. Then for each $0 \leq i < t$, the $\mathcal{F}_{i+1,t}$-measurable random variable $\langle z_i, x_i - x_t^\star\rangle$ is mean-zero sub-Gaussian conditioned on $\mathcal{F}_{i,t}$ with parameter $c\sigma\|x_i - x_t^\star\|$, so

$$\mathbb{E}\big[\exp\left(\lambda\langle z_i, x_i - x_t^\star\rangle(1-\hat{\rho})^{t-1-i}\right) \mid \mathcal{F}_{i,t}\big] \leq \exp\left(\lambda^2(c\sigma)^2\|x_i - x_t^\star\|^2(1-\hat{\rho})^{2(t-1-i)}/2\right) \quad \forall \lambda \in \mathbb{R}.$$

Now fix $t \geq 1$ and observe that Proposition 5.14 yields the total conditional variance bound

$$\sum_{i=0}^{t-1}(c\sigma)^2\|x_i - x_t^\star\|^2(1-\hat{\rho})^{2(t-1-i)} \leq \sum_{j=0}^{t-2}\alpha_j\langle z_j, x_j - x_t^\star\rangle(1-\hat{\rho})^{t-1-j} + R_t,$$

where $0 \leq \alpha_j \leq \alpha$ for all $0 \leq j \leq t-2$ and

$$R_t := \tfrac{(c\sigma)^2}{\hat{\rho}}(1-\hat{\rho})^{t-1}\|x_0 - x_t^\star\|^2 + \tfrac{2\eta^2(c\sigma)^2}{\hat{\rho}}\sum_{j=0}^{t-2}\|z_j\|^2(1-\hat{\rho})^{t-2-j} + \tfrac{\eta(c\sigma)^2}{\mu\hat{\rho}}\sum_{j=0}^{t-2}G_{j,t}^2(1-\hat{\rho})^{t-2-j}.$$

We claim that

$$\mathbb{P}\left\{R_t \leq \beta_t \log\left(\frac{3e}{\delta}\right)\right\} \geq 1 - \delta \quad \forall \delta \in (0,1). \tag{10}$$

To verify (10), observe first that for all $n \geq 0$, the sum $\sum_{i=0}^n \|z_i\|^2(1-\hat{\rho})^{n-i}$ is sub-exponential with parameter $\sum_{i=0}^n c\sigma^2(1-\hat{\rho})^{n-i} \leq (c\sigma)^2/\hat{\rho}$, so Markov's inequality implies

$$\mathbb{P}\left\{\sum_{i=0}^n \|z_i\|^2(1-\hat{\rho})^{n-i} \leq \frac{(c\sigma)^2}{\hat{\rho}}\log\left(\frac{e}{\delta}\right)\right\} \geq 1 - \delta \quad \forall \delta \in (0,1). \tag{11}$$

Further, for all $0 \leq n < t$, it follows from Assumption 4.4 and Lemma 2.3 that $\|x_0 - x_t^\star\|^2$ is sub-exponential with parameter $2\left(\|x_0 - x_0^\star\|^2 + \Delta^2 t^2\right)$ and $\sum_{i=0}^n G_{i,t}^2(1-\hat{\rho})^{n-i}$ is sub-exponential with parameter

$$\sum_{i=0}^{n}(\mu\Delta)^2(t-i)^2(1-\hat{\rho})^{n-i} = (\mu\Delta)^2(1-\hat{\rho})^{n+1-t}\sum_{i=0}^{n}(t-i)^2(1-\hat{\rho})^{t-i-1} \leq \frac{2(\mu\Delta)^2}{\hat{\rho}^3(1-\hat{\rho})^{t-1-n}},$$

so Markov's inequality implies

$$\mathbb{P}\left\{\|x_0 - x_t^\star\|^2 \leq 2\left(\|x_0 - x_0^\star\|^2 + \Delta^2 t^2\right)\log\left(\frac{e}{\delta}\right)\right\} \geq 1-\delta \quad \forall \delta \in (0,1) \tag{12}$$

and

$$\mathbb{P}\left\{\sum_{i=0}^{n} G_{i,t}^2(1-\hat{\rho})^{n-i} \leq \frac{2(\mu\Delta)^2}{\hat{\rho}^3(1-\hat{\rho})^{t-1-n}}\log\left(\frac{e}{\delta}\right)\right\} \geq 1-\delta \quad \forall \delta \in (0,1). \tag{13}$$

Thus, (11)–(13) and a union bound yield (10). Consequently, Theorem 5.15 implies that the following bound holds for all $\delta \in (0,1)$ and $\tau > 0$:

$$\mathbb{P}\left\{\sum_{i=0}^{t-1}\langle z_i, x_i - x_t^\star\rangle(1-\hat{\rho})^{t-1-i} \geq \tau\right\} \leq \delta + \exp\left(-\frac{\tau}{4\alpha + 8\beta_t\log(3e/\delta)/\tau}\right),$$

as claimed. □

We may now deduce the following precise version of Theorem 4.6 using the tail bound furnished by Proposition 5.16.

Theorem 5.17 (Function gap with high probability)**.** *Suppose that Assumption 4.4 holds, let $\{\hat{x}_t\}$ be the iterates produced by Algorithm 2 with constant step size $\eta \leq 1/2L$, and set $\hat{\rho} := \mu\eta/(2-\mu\eta)$. Then there is an absolute constant $c > 0$ such that for any specified $t \in \mathbb{N}$ and $\delta \in (0,1)$, the following estimate holds with probability at least $1-\delta$:*

$$\varphi_t(\hat{x}_t) - \varphi_t^\star \leq (1-\hat{\rho})^t\left(\varphi_t(x_0) - \varphi_t^\star + \tfrac{\mu}{4}\|x_0 - x_t^\star\|^2\right) + \left(\eta(c\sigma)^2 + \frac{8\Delta^2}{\mu\eta^2} + 9\hat{\rho}\sqrt{8\beta_t}\right)\log\left(\frac{4e}{\delta}\right),$$

where

$$\beta_t := (1-\hat{\rho})^{t-1}\left(\|x_0 - x_0^\star\|^2 + \Delta^2 t^2\right)\frac{2(c\sigma)^2}{\hat{\rho}} + \frac{2\eta^2(c\sigma)^4}{\hat{\rho}^2} + \frac{3\mu\Delta^2\eta(c\sigma)^2}{\hat{\rho}^4}.$$

Proof. A quick computation shows that given any $\delta \in (0,1)$, we may take

$$\tau = (2+\sqrt{5})\sqrt{8\beta_t}\log\left(\frac{3e}{\delta}\right)$$

in Proposition 5.16 to obtain

$$\mathbb{P}\left\{\sum_{i=0}^{t-1}\langle z_i, x_i - x_t^\star\rangle(1-\hat{\rho})^{t-1-i} < (2+\sqrt{5})\sqrt{8\beta_t}\log\left(\frac{3e}{\delta}\right)\right\} \geq 1-2\delta. \tag{14}$$

We may now combine (11), (13), and (14) together with Proposition 5.11 and a union bound to conclude that for all $\delta \in (0,1)$, the estimate

$$\varphi_t(\hat{x}_t) - \varphi_t^\star \leq (1-\hat{\rho})^t\left(\varphi_t(x_0) - \varphi_t^\star + \tfrac{\mu}{4}\|x_0 - x_t^\star\|^2\right) + \left(\eta(c\sigma)^2 + \frac{2\mu\Delta^2}{\hat{\rho}^2}\right)\log\left(\frac{e}{\delta}\right)$$
$$+ (2+\sqrt{5})\hat{\rho}\sqrt{8\beta_t}\log\left(\frac{3e}{\delta}\right)$$

holds with probability at least $1-4\delta$; hence

$$\varphi_t(\hat{x}_t) - \varphi_t^\star \leq (1-\hat{\rho})^t\left(\varphi_t(x_0) - \varphi_t^\star + \tfrac{\mu}{4}\|x_0 - x_t^\star\|^2\right) + \left(\eta(c\sigma)^2 + \frac{8\Delta^2}{\mu\eta^2} + 9\hat{\rho}\sqrt{8\beta_t}\right)\log\left(\frac{e}{\delta}\right).$$

with probability at least $1-4\delta$. □

Remark 5.18. To see that Theorem 5.17 entails Theorem 4.6, fix $t \in \mathbb{N}$ and observe that upon setting $C := \max\{c, 1\}$, we have

$$\hat{\rho}\sqrt{8\beta_t} \le 4C^2\left(\sqrt{(1-\hat{\rho})^t\left(\|x_0 - x_0^\star\|^2 + \Delta^2 t^2\right)\mu\eta\sigma^2} + \eta\sigma^2 + \sqrt{6}\frac{\Delta\sigma}{\sqrt{\mu\eta}}\right),$$

while the AM-GM inequality implies

$$2\sqrt{(1-\hat{\rho})^t\left(\|x_0 - x_0^\star\|^2 + \Delta^2 t^2\right)\mu\eta\sigma^2} \le (1-\hat{\rho})^t\left(\mu\|x_0 - x_0^\star\|^2 + \mu\Delta^2 t^2\right) + \eta\sigma^2,$$

inequality (9) implies

$$(1-\hat{\rho})^t\left(\mu\|x_0 - x_0^\star\|^2 + \mu\Delta^2 t^2\right) \le 2(1-\hat{\rho})^t\big(\varphi_0(x_0) - \varphi_0^\star\big) + \frac{16\Delta^2}{\mu\eta^2},$$

and Young's inequality implies

$$\frac{2\Delta\sigma}{\sqrt{\mu\eta}} \le \eta\sigma^2 + \frac{\Delta^2}{\mu\eta^2}.$$

Hence

$$\hat{\rho}\sqrt{8\beta_t} \lesssim (1-\hat{\rho})^t\left(\varphi_0(x_0) - \varphi_0^\star\right) + \eta\sigma^2 + \frac{\Delta^2}{\mu\eta^2}.$$

Further, inequalities (7) and (9) together with Assumption 4.4 imply that for all $\delta \in (0,1)$, the estimate

$$(1-\hat{\rho})^t\left(\varphi_t(x_0) - \varphi_t^\star + \tfrac{\mu}{4}\|x_0 - x_t^\star\|^2\right) \le 3(1-\hat{\rho})^t\big(\varphi_0(x_0) - \varphi_0^\star\big) + \frac{32\Delta^2}{\mu\eta^2}\log\left(\frac{e}{\delta}\right)$$

holds with probability at least $1-\delta$. Thus, under the assumptions of Theorem 5.17, a union bound reveals that for all $t \in \mathbb{N}$ and $\delta \in (0,1)$, the estimate

$$\begin{aligned}
\varphi_t(\hat{x}_t) - \varphi_t^\star &\le (1-\hat{\rho})^t\left(\varphi_t(x_0) - \varphi_t^\star + \tfrac{\mu}{4}\|x_0 - x_t^\star\|^2\right) + \left(\eta(c\sigma)^2 + \frac{8\Delta^2}{\mu\eta^2} + 9\hat{\rho}\sqrt{8\beta_t}\right)\log\left(\frac{8e}{\delta}\right)\\
&\lesssim \left((1-\hat{\rho})^t\left(\varphi_0(x_0) - \varphi_0^\star\right) + \eta\sigma^2 + \frac{\Delta^2}{\mu\eta^2}\right)\log\left(\frac{e}{\delta}\right)
\end{aligned}$$

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

We may now apply Theorem 4.6 to obtain the formal version of Theorem 4.7; the proof is analogous to that of Theorem 5.8 and appears in the appendix (see Section A.5).

Theorem 5.19 (Time to track with high probability)**.** *Suppose that Assumption 4.4 holds and that we are in the low drift-to-noise regime $\Delta/\sigma < \sqrt{\mu/16L^3}$. Set $\eta_\star = (2\Delta^2/\mu\sigma^2)^{1/3}$ and $\mathcal{G} = \mu(\Delta\sigma^2/\mu^2)^{2/3}$. Suppose moreover that we have available a positive upper bound on the initial gap $D \ge \varphi_0(x_0) - \varphi_0^\star$. Fix $\delta \in (0,1)$ and consider running Algorithm 2 in $k = 0,\dots,K-1$ epochs, namely, set $X_0 = x_0$ and iterate the process*

$$X_{k+1} = \overline{\texttt{PSG}}(X_k, \mu, \eta_k, T_k) \quad \text{for} \quad k = 0,\dots,K-1,$$

where the number of epochs is

$$K = 1 + \left\lceil \log_2\left(\frac{1}{L}\cdot\left(\frac{\sigma^2\mu}{\Delta^2}\right)^{1/3}\right)\right\rceil$$

and we set

$$\eta_0 = \frac{1}{2L}, \quad T_0 = \left\lceil \frac{4L}{\mu}\log\left(\frac{LD}{\sigma^2}\right)^+\right\rceil \quad \text{and} \quad \eta_k = \frac{\eta_{k-1} + \eta_\star}{2}, \quad T_k = \left\lceil \frac{2\log\left(4c\log(e/\delta)\right)^+}{\mu\eta_k}\right\rceil$$

for all $k \ge 1$, where $c > 0$ is the absolute constant furnished by the bound (3). Then the time horizon $T = T_0 + \cdots + T_{K-1}$ satisfies

$$T \lesssim \frac{L}{\mu}\log\left(\frac{LD}{\sigma^2}\right)^+ + \frac{\sigma^2}{\mu\mathcal{G}}\left(1 \vee \log\log\frac{e}{\delta}\right) \le \frac{L}{\mu}\log\left(\frac{D}{\mathcal{G}}\right)^+ + \frac{\sigma^2}{\mu\mathcal{G}}\left(1 \vee \log\log\frac{e}{\delta}\right)$$

and the corresponding tracking error satisfies

$$\varphi_T(X_K) - \varphi_T^\star \lesssim \mathcal{G} \log\left(\frac{e}{\delta}\right)$$

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

6 Numerical illustrations

We investigate the empirical behavior of our finite-time bounds on numerical examples with synthetic data. We consider examples of a) least-squares recovery; b) sparse least-squares recovery; c) ℓ_2^2-regularized logistic regression; and investigate the behavior of $\|x_t - x_t^\star\|^2$ and $\varphi_t(\hat{x}_t) - \varphi_t^\star$ in each case. The main findings are that our bounds exhibit: 1) the correct dependence on η, σ, and Δ; 2) excellent coverage in Monte-Carlo simulations. Code is available online at `https://github.com/joshuacutler/TimeDriftExperiments`.

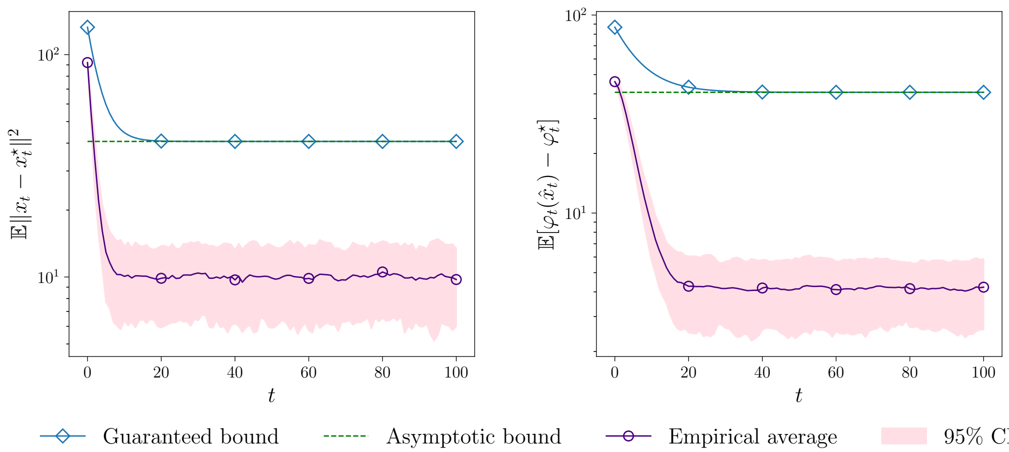

Figure 1: Semilog plots of guaranteed bounds and empirical tracking errors with respect to iteration t for least-squares recovery. Shaded regions indicate the 95% confidence intervals for $\|x_t - x_t^\star\|^2$ and $\varphi_t(\hat{x}_t) - \varphi_t^\star$; empirical averages and confidence intervals are computed over 100 trials. Default parameter values: $\mu = 1$, $L = 1$, $\sigma = 10$, $\Delta = 1$, and $\eta = \eta_\star$.

Least-squares recovery. Fix $x_0, x_0^\star \in \mathbb{R}^d$ with standard Gaussian entries, and consider a Gaussian random walk $(x_t^\star)$ given by $x_{t+1}^\star = x_t^\star + v_t$, where v_t is drawn uniformly from the sphere of radius Δ in $\mathbb{R}^d$. Given a fixed matrix $A \in \mathbb{R}^{n\times d}$, we aim to recover the vectors $(x_t^\star)$ via least-squares:

$$\min_{x\in\mathbb{R}^d} \mathop{\mathbb{E}}_{w\sim\mathcal{P}_t} \frac{1}{2}\|Ax - w\|^2,$$

where $\mathcal{P}_t = \mathcal{N}(Ax_t^\star, C)$ with $C = (\sigma^2/n\|A\|_{\text{op}}^2)I_n$. This amounts to the target problem (1) under the identifications $f_t(x) = \mathbb{E}_{w\sim\mathcal{P}_t}\frac{1}{2}\|Ax - w\|^2$ and $r_t = 0$; clearly $\|x_t^\star - x_{t+1}^\star\| = \Delta$ and $\sup_x \|\nabla f_t(x) - \nabla f_{t+1}(x)\| \le \|A\|_{\text{op}}^2\Delta$. We implement Algorithms 1 and 2 initialized at x_0 using the sample gradient $g_t = A^T(Ax_t - w)$ at step t, where $w \sim \mathcal{P}_t$; hence $\mathbb{E}\|\nabla f_t(x_t) - g_t\|^2 \le \sigma^2$.

In our simulations, we take $d = 50$, $n = 100$, and randomly generate A via its singular value decomposition (using Haar-distributed orthogonal matrices) so that its minimal singular value is $\sqrt{\mu}$ and its maximal singular value is $\sqrt{L}$. In Figures 1 and 2, we use default parameter values $\mu = 1$, $L = 1$, $\sigma = 10$, $\Delta = 1$, and the corresponding asymptotically optimal step size $\eta = \eta_\star$. Since f_t is μ-strongly convex and L-smooth, this puts us in the low drift/noise regime in Figure 1: $\Delta/\sigma < \sqrt{\mu/16L^3} = 1/4$. To estimate the empirical averages and confidence intervals of $\|x_t - x_t^\star\|^2$ and $\varphi_t(\hat{x}_t) - \varphi_t^\star$, we run 100 trials with horizon $T = 100$. The results confirm our bounds and show that they capture the correct dependence on η, σ, and Δ.

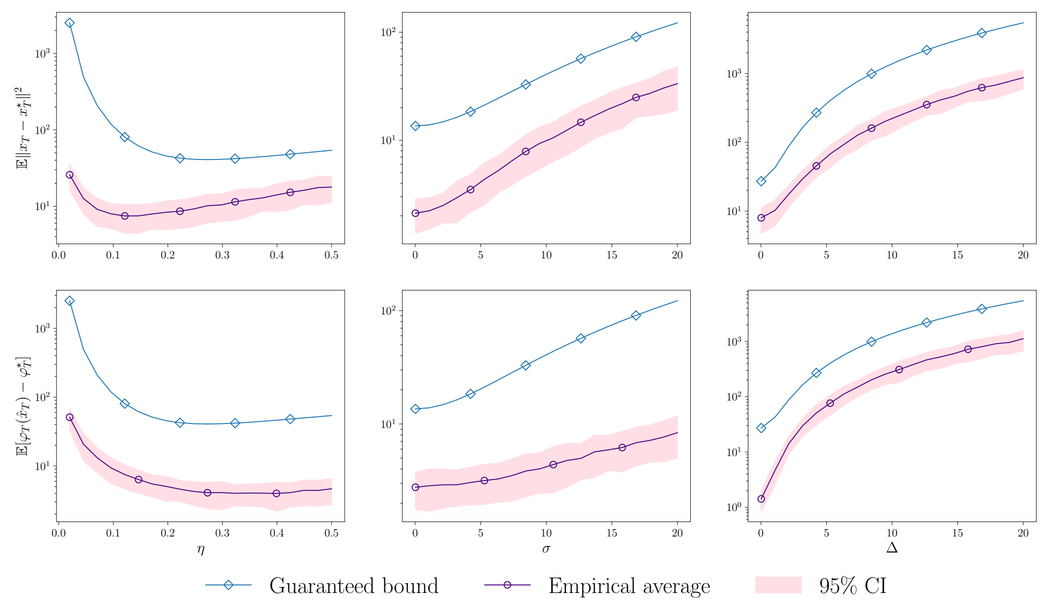

Figure 2: Semilog plots of guaranteed bounds and empirical tracking errors at horizon $T = 100$ with respect to η, σ, and Δ for least-squares recovery. Shaded regions indicate the 95% confidence intervals for $\|x_T - x_T^\star\|^2$ and $\varphi_T(\hat{x}_T) - \varphi_T^\star$; empirical averages and confidence intervals are computed over 100 trials. Default parameter values: $\mu = 1$, $L = 1$, $\sigma = 10$, $\Delta = 1$, and $\eta = \eta_\star$.

Sparse least-squares recovery. Next, we consider least-squares recovery constrained to the closed ℓ_1-ball in $\mathbb{R}^d$, which we denote by B_1. We aim to recover a sparse sequence of vectors in B_1 defined as follows. Set $s = \lfloor \log d \rfloor$, draw a vector u uniformly from the ℓ_1-ball in $\mathbb{R}^s$, fix $x_0^\star = (u, 0) \in \mathbb{R}^d$, and select $\Delta \in (0, \sqrt{2}]$. At step t, with probability $p = (4-2\Delta^2)/(4-\Delta^2)$, we set $x_{t+1}^\star = x_t^\star + v$, where v is selected to have the same support as $x_t^\star$ and satisfy $\|v\| = \Delta/\sqrt{2}$ and $x_t^\star + v \in B_1$; otherwise, with probability $1 - p$, we obtain $x_{t+1}^\star$ from $x_t^\star$ by swapping precisely one nonzero coordinate with a zero coordinate. Then the resulting sparse sequence $(x_t^\star)$ in B_1 satisfies $\mathbb{E}\|x_t^\star - x_{t+1}^\star\|^2 \leq \Delta^2$. Given a fixed matrix $A \in \mathbb{R}^{n \times d}$, we aim to recover $(x_t^\star)$ via constrained least-squares:

$$\min_{x \in B_1} \mathop{\mathbb{E}}_{w \sim \mathcal{P}_t} \frac{1}{2}\|Ax - w\|^2,$$

where $\mathcal{P}_t = \mathcal{N}(Ax_t^\star, C)$ with $C = (\sigma^2/n\|A\|_{\text{op}}^2)I_n$. This amounts to the target problem (1) under the identifications $f_t(x) = \mathbb{E}_{w \sim \mathcal{P}_t} \frac{1}{2}\|Ax - w\|^2$ and $r_t = \delta_{B_1}$ (the convex indicator of B_1); clearly $\mathbb{E}[\sup_x \|\nabla f_t(x) - \nabla f_{t+1}(x)\|^2] \leq (\|A\|_{\text{op}}^2 \Delta)^2$. Fixing x_0 drawn uniformly from B_1, we implement Algorithms 1 and 2 initialized at x_0 using the sample gradient $g_t = A^T(Ax_t - w)$ at step t, where $w \sim \mathcal{P}_t$; hence $\mathbb{E}\|\nabla f_t(x_t) - g_t\|^2 \leq \sigma^2$.

In our simulations, we take $d = 50$, $n = 100$, and randomly generate A via its singular value decomposition (using Haar-distributed orthogonal matrices) so that its minimal singular value is $\sqrt{\mu}$ and its maximal singular value is $\sqrt{L}$. In Figures 3 and 4, we use default parameter values $\mu = 1$, $L = 1$, $\sigma = 1/2$, $\Delta = 1/20$, and the corresponding asymptotically optimal step size $\eta = \eta_\star$. Since f_t is μ-strongly convex and L-smooth, this puts us in the low drift/noise regime in Figure 3: $\Delta/\sigma < \sqrt{\mu/16L^3} = 1/4$. To estimate the empirical averages and confidence intervals of $\|x_t - x_t^\star\|^2$ and $\varphi_t(\hat{x}_t) - \varphi_t^\star$, we run 100 trials with horizon $T = 100$. The results confirm our bounds and show that they capture the correct dependence on η, σ, and Δ.

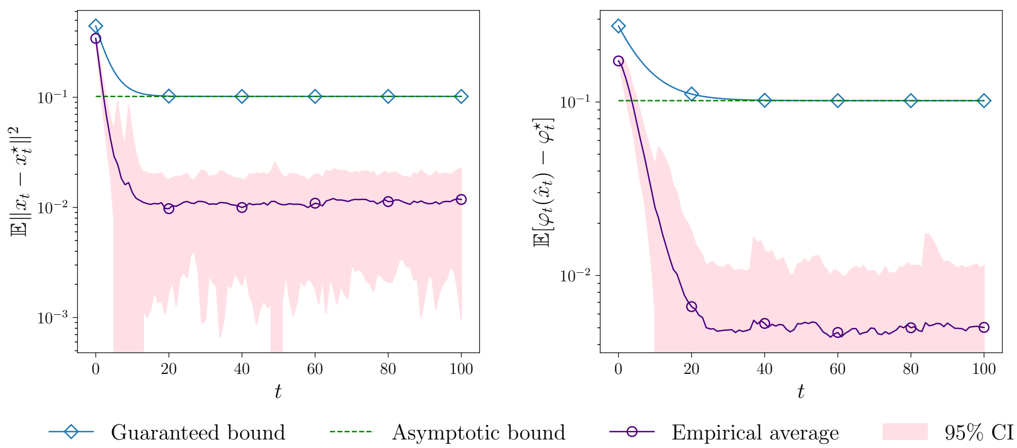

Figure 3: Semilog plots of guaranteed bounds and empirical tracking errors with respect to iteration t for sparse least-squares recovery. Shaded regions indicate the 95% confidence intervals for $\|x_t - x_t^\star\|^2$ and $\varphi_t(\hat{x}_t) - \varphi_t^\star$; empirical averages and confidence intervals are computed over 100 trials. Default parameter values: $\mu = 1$, $L = 1$, $\sigma = 1/2$, $\Delta = 1/20$, and $\eta = \eta_\star$.

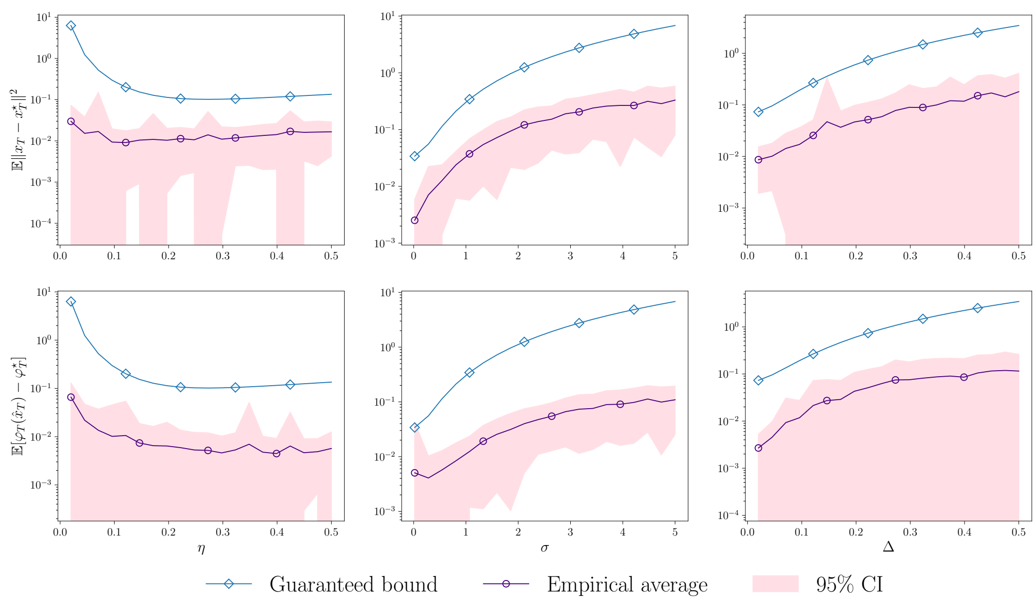

Figure 4: Semilog plots of guaranteed bounds and empirical tracking errors at horizon $T = 100$ with respect to η, σ, and Δ for sparse least-squares recovery. Shaded regions indicate the 95% confidence intervals for $\|x_T - x_T^\star\|^2$ and $\varphi_T(\hat{x}_T) - \varphi_T^\star$; empirical averages and confidence intervals are computed over 100 trials. Default parameter values: $\mu = 1$, $L = 1$, $\sigma = 1/2$, $\Delta = 1/20$, and $\eta = \eta_\star$.

ℓ_2^2-regularized logistic regression. Finally, we consider the time-varying ℓ_2^2-regularized logistic regression problem

$$\min_{x \in \mathbb{R}^d} \frac{1}{n} \left(\sum_{i=1}^{n} \log(1 + \exp\langle a_i, x \rangle) - \langle Ax, b_t \rangle \right) + \frac{\mu}{2}\|x\|^2,$$

where the fixed matrix $A \in \mathbb{R}^{n \times d}$ has standard Gaussian rows $a_1, \ldots, a_n \in \mathbb{R}^d$, (b_t) is a random sequence of label vectors in $\{0,1\}^n$ such that b_t and b_{t+1} differ in precisely one coordinate for each t, and $\mu > 0$. This amounts to the target problem (1) under the identifications $f_t(x) = \frac{1}{n}\left(\sum_{i=1}^{n} \log(1 + \exp\langle a_i, x\rangle) - \langle Ax, b_t\rangle\right) + \frac{\mu}{2}\|x\|^2$ and $r_t = 0$; setting $L = \frac{1}{4n}\|A\|_{\mathrm{op}}^2 + \mu$, it follows

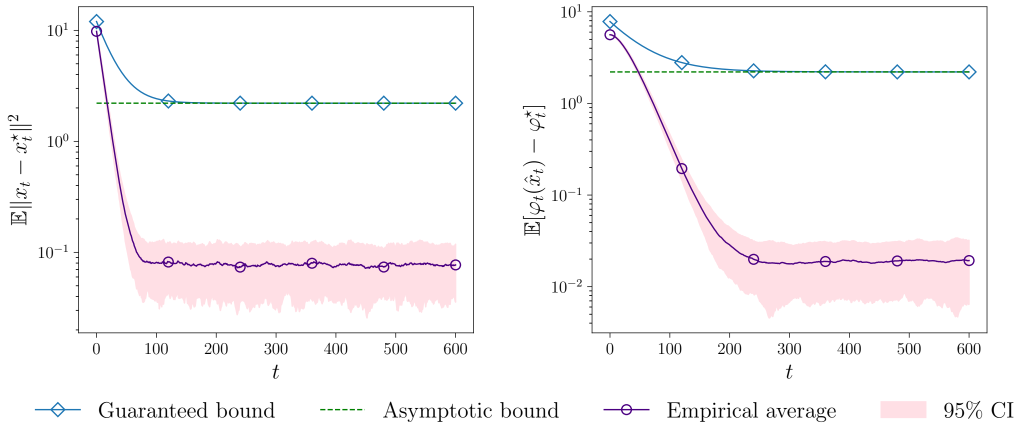

Figure 5: Semilog plots of guaranteed bounds and empirical tracking errors with respect to iteration t for ℓ_2^2-regularized logistic regression. Shaded regions indicate the 95% confidence intervals for $\|x_t - x_t^\star\|^2$ and $\varphi_t(\hat{x}_t) - \varphi_t^\star$; empirical averages and confidence intervals are computed over 100 trials. Default parameter values: $\mu = 1$ and $\eta = \eta_\star$.

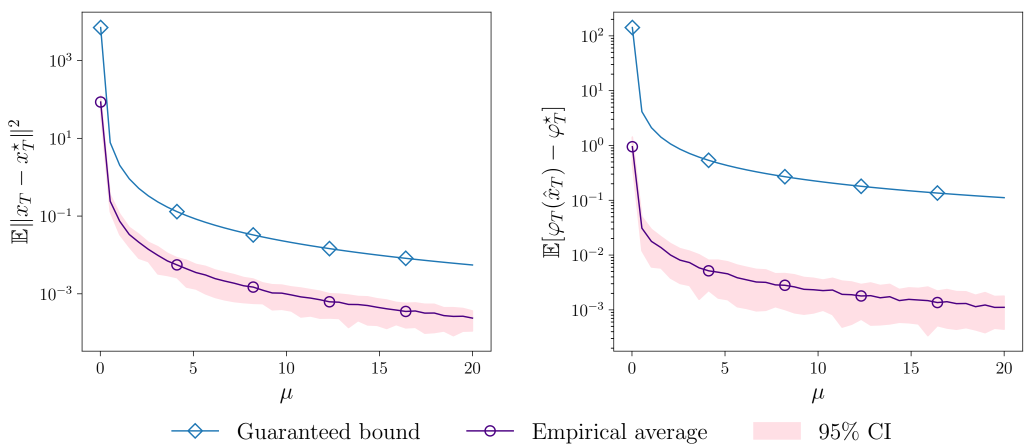

Figure 6: Semilog plots of guaranteed bounds and empirical tracking errors at horizon $T = 600$ with respect to the strong convexity parameter μ for ℓ_2^2-regularized logistic regression. Shaded regions indicate the 95% confidence intervals for $\|x_T - x_T^\star\|^2$ and $\varphi_T(\hat{x}_T) - \varphi_T^\star$; empirical averages and confidence intervals are computed over 100 trials, using the asymptotically optimal step size $\eta_\star$ (which itself depends on μ).

that f_t is μ-strongly convex and L-smooth. Letting $(x_t^\star)$ denote the corresponding sequence of minimizers and setting $\Delta = \frac{1}{\mu n}\max_{i=1,\ldots,n}\|a_i\|$, it follows that $\sup_x \|\nabla f_t(x) - \nabla f_{t+1}(x)\| \le \mu\Delta$ and hence $\|x_t^\star - x_{t+1}^\star\| \le \Delta$. Fixing the initial label b_0 (drawn uniformly from $\{0,1\}^n$) and a standard Gaussian vector $x_0 \in \mathbb{R}^d$, we implement Algorithms 1 and 2 initialized at x_0 using the following random summand sample gradient at each step t:

$$g_t = \left(\frac{\exp\langle a_k, x_t\rangle}{1+\exp\langle a_k, x_t\rangle} - b_t^k\right) a_k + \mu x_t,$$

where $k \sim \mathcal{U}\{1,\ldots,n\}$ and b_t^k denotes the k^{th} coordinate of b_t. Then $\mathbb{E}\|\nabla f_t(x_t) - g_t\|^2 \le \sigma^2$, where

$$\sigma^2 = \frac{1}{n^2}\left((n-2)\sum_{i=1}^n \|a_i\|^2 + \sum_{i,j=1}^n \|a_i\|\|a_j\|\right) \le \max_{i=1,\ldots,n} 2\|a_i\|^2.$$

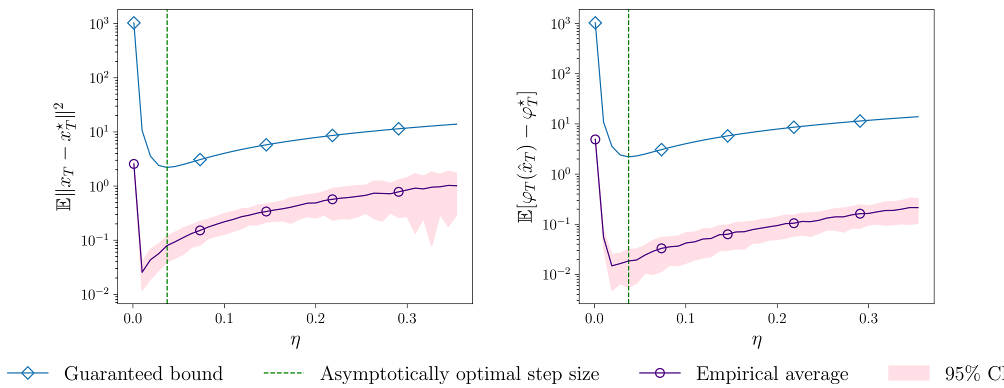

Figure 7: Semilog plots of guaranteed bounds and empirical tracking errors at horizon $T = 600$ with respect to the step size η for ℓ_2^2-regularized logistic regression. Shaded regions indicate the 95% confidence intervals for $\|x_T - x_T^\star\|^2$ and $\varphi_T(\hat{x}_T) - \varphi_T^\star$; empirical averages and confidence intervals are computed over 100 trials. Default parameter value: $\mu = 1$. Observe that $\eta_\star$ is close to empirically optimal.

In our simulations, we take $d = 20$ and $n = 200$, and we generate b_{t+1} from b_t by flipping a single coordinate selected uniformly at random. In Figure 5, we use default parameter values $\mu = 1$ and the corresponding asymptotically optimal step size $\eta = \eta_\star$. In Figure 6, we illustrate the dependence of tracking error on the regularization parameter μ; here, the asymptotically optimal step size $\eta_\star$ is used (which itself depends on μ). In Figure 7, we use the default parameter value $\mu = 1$. To estimate the empirical averages and confidence intervals of $\|x_t - x_t^\star\|^2$ and $\varphi_t(\hat{x}_t) - \varphi_t^\star$, we run 100 trials with horizon $T = 600$. The results confirm our bounds and show that they capture the correct dependence on μ and η. In particular, Figure 7 illustrates that $\eta_\star$ is close to empirically optimal.

Acknowledgments

This work was supported by NSF CCF-2019844, NSF DMS-2023166, NSF DMS-1651851, CIFAR-LMB, and faculty research awards. Part of this work was done while Z. Harchaoui was visiting the Simons Institute for the Theory of Computing. Added Research of Drusvyatskiy was supported by the NSF DMS 1651851 and CCF 2023166 awards.

References

[1] A. Agarwal, O. Chapelle, M. Dudík, and J. Langford. A reliable effective terascale linear learning system. *Journal of Machine Learning Research*, 15(1):1111–1133, 2014.

[2] F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, volume 24, pages 451–459. Curran Associates, Inc., 2011.

[3] P. L. Bartlett, S. Ben-David, and S. R. Kulkarni. Learning changing concepts by exploiting the structure of change. *Machine Learning*, 41(2):153–174, 2000.

[4] P. L. Bartlett, V. Dani, T. P. Hayes, S. M. Kakade, A. Rakhlin, and A. Tewari. High-probability regret bounds for bandit online linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory*. Omnipress, 2008.

[5] O. Besbes, Y. Gur, and A. Zeevi. Non-stationary stochastic optimization. *Operations Research*, 63(5):1227–1244, 2015.

[6] L. Bottou. Stochastic learning. In *Advanced Lectures on Machine Learning, ML Summer Schools 2003*, volume 3176 of *Lecture Notes in Computer Science*, pages 146–168. Springer, 2003.

[7] L. Bottou. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade - Second Edition*, volume 7700 of *Lecture Notes in Computer Science*, pages 421–436. Springer, 2012.

[8] L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems*, volume 20, pages 161–168. Curran Associates, Inc., 2007.

[9] C.-K. Chiang, T. Yang, C.-J. Lee, M. Mahdavi, C.-J. Lu, R. Jin, and S. Zhu. Online optimization with gradual variations. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 6.1–6.20. PMLR, 2012.

[10] D. Drusvyatskiy and L. Xiao. Stochastic optimization with decision-dependent distributions. *arXiv:2011.11173*, 2020.

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

[12] S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, ii: shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23(4):2061–2089, 2013.

[13] L. Guo and L. Ljung. Exponential stability of general tracking algorithms. *IEEE Transactions on Automatic Control*, 40(8):1376–1387, 1995.

[14] N. J. A. Harvey, C. Liaw, Y. Plan, and S. Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Proceedings of the 32nd Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1579–1613. PMLR, 2019.

[15] E. Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 2016.

[16] E. Hazan and S. Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15(1):2489–2512, 2014.

[17] E. Hazan and C. Seshadhri. Efficient learning algorithms for changing environments. In *Proceedings of the 26th International Conference on Machine Learning*. Omnipress, 2009.

[18] M. Herbster and M. K. Warmuth. Tracking the best expert. *Machine Learning*, 32(2):151–178, 1998.

[19] C. Jin, P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan. A short note on concentration inequalities for random vectors with subgaussian norm. *arXiv:1902.03736*, 2019.

[20] A. Kulunchakov and J. Mairal. Estimate sequences for stochastic composite optimization: Variance reduction, acceleration, and robustness to noise. *Journal of Machine Learning Research*, 21(155):1–52, 2020.

[21] H. J. Kushner and G. G. Yin. *Stochastic Approximation Algorithms and Applications*, volume 35 of *Applications of Mathematics*. Springer, 1997.

[22] G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.

[23] H. Lang, L. Xiao, and P. Zhang. Using statistics to automate stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 32, pages 9536–9546. Curran Associates, Inc., 2019.

[24] P. M. Long. The complexity of learning according to two models of a drifting environment. *Machine Learning*, 37(3):337–354, 1999.

[25] L. Madden, S. Becker, and E. Dall'Anese. Bounds for the tracking error of first-order online optimization methods. *Journal of Optimization Theory and Applications*, 189(2):437–457, 2021.

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

[27] A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning*. Omnipress, 2012.

[28] A. Rakhlin and K. Sridharan. Online learning with predictable sequences. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 993–1019. PMLR, 2013.

[29] A. H. Sayed. *Fundamentals of Adaptive Filtering*. John Wiley & Sons, 2003.

[30] R. E. Schapire and Y. Freund. *Boosting: Foundations and Algorithms*. The MIT Press, 2012.

[31] S. Sra, S. Nowozin, and S. J. Wright. *Optimization for Machine Learning*. The MIT Press, 2011.

[32] N. Srebro, K. Sridharan, and A. Tewari. On the universality of online mirror descent. In *Advances in Neural Information Processing Systems*, volume 24, pages 2645–2653. Curran Associates, Inc., 2011.

[33] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2018.

[34] C. Wilson, V. V. Veeravalli, and A. Nedić. Adaptive sequential stochastic optimization. *IEEE Transactions on Automatic Control*, 64(2):496–509, 2018.

A Additional proofs

A.1 Proof of Theorem 5.8

For each index k, let $t_k := T_0 + \cdots + T_{k-1}$ (with $t_0 := 0$), $X_k^\star$ be the minimizer of the corresponding function φ_{t_k}, and

$$E_k := c\left(\frac{\eta_k \sigma^2}{\mu} + \left(\frac{\Delta}{\mu \eta_\star}\right)^2\right),$$

where $c \geq 1$ is an absolute constant satisfying the bound (2) in Theorem 3.7. Taking into account $\eta_k \geq \eta_\star$ and our selection of c, Theorem 3.7 implies that for any specified $k \geq 0$ and $\delta \in (0,1)$ the following estimate holds with probability at least $1-\delta$:

$$\begin{aligned}\|X_{k+1} - X_{k+1}^\star\|^2 &\leq \left(1 - \frac{\mu\eta_k}{2}\right)^{T_k}\|X_k - X_k^\star\|^2 + c\left(\frac{\eta_k\sigma^2}{\mu} + \left(\frac{\Delta}{\mu\eta_k}\right)^2\right)\log\left(\frac{e}{\delta}\right)\\ &\leq e^{-\mu\eta_k T_k/2}\|X_k - X_k^\star\|^2 + E_k \log\left(\frac{e}{\delta}\right).\end{aligned}$$

We will verify by induction that for all indices $k \geq 1$, the estimate $\|X_k - X_k^\star\|^2 \leq 3E_{k-1}\log(e/\delta)$ holds with probability at least $1 - k\delta$ for all $\delta \in (0,1)$. To see the base case, observe that the estimate

$$\|X_1 - X_1^\star\|^2 \leq e^{-\mu\eta_0 T_0/2}\|X_0 - X_0^\star\|^2 + E_0\log\left(\frac{e}{\delta}\right) \leq 3E_0\log\left(\frac{e}{\delta}\right)$$

holds with probability at least $1-\delta$ for all $\delta \in (0,1)$. Now assume the claim holds for some index $k \geq 1$, and let $\delta \in (0,1)$; then $\|X_k - X_k^\star\|^2 \leq 3E_{k-1}\log(e/\delta)$ with probability at least $1 - k\delta$. Thus, since we also have

$$\begin{aligned}\|X_{k+1} - X_{k+1}^\star\|^2 &\leq e^{-\mu\eta_k T_k/2}\|X_k - X_k^\star\|^2 + E_k\log\left(\frac{e}{\delta}\right)\\ &\leq \frac{1}{4}\|X_k - X_k^\star\|^2 + E_k\log\left(\frac{e}{\delta}\right)\\ &\leq \frac{E_k}{2E_{k-1}}\|X_k - X_k^\star\|^2 + E_k\log\left(\frac{e}{\delta}\right)\end{aligned}$$

with probability at least $1-\delta$, a union bound reveals $\|X_{k+1} - X_{k+1}^\star\|^2 \leq 3E_k\log(e/\delta)$ with probability at least $1-(k+1)\delta$, thereby completing the induction. Hence, upon fixing $\delta \in (0,1)$, we have $\|X_K - X_K^\star\|^2 \leq 3E_{K-1}\log(e/\delta)$ with probability at least $1 - K\delta$.

Next, observe

$$\frac{2}{c}E_{K-1} - \sqrt[3]{54}\left(\frac{\Delta\sigma^2}{\mu^2}\right)^{2/3} = \frac{2\sigma^2}{\mu}(\eta_{K-1} - \eta_\star) = \frac{2\sigma^2}{\mu}\cdot\frac{\eta_0 - \eta_\star}{2^{K-1}} \leq \left(\frac{\Delta\sigma^2}{\mu^2}\right)^{2/3} = \mathcal{E},$$

so

$$\|X_K - X_K^\star\|^2 \leq \frac{3c}{2}\big(1 + \sqrt[3]{54}\big)\mathcal{E}\log\left(\frac{e}{\delta}\right) \asymp \mathcal{E}\log\left(\frac{e}{\delta}\right)$$

with probability at least $1 - K\delta$. Finally, note

$$T \lesssim \frac{L}{\mu}\log\left(\frac{\mu L D}{\sigma^2}\right)^+ + \frac{1}{\mu}\sum_{k=1}^{K-1}\frac{1}{\eta_k}$$

and

$$\sum_{k=1}^{K-1}\frac{1}{\eta_k} \leq 2L\sum_{k=1}^{K-1}2^k \leq 2L\cdot 2^K = 8L\cdot 2^{K-2} \leq 8\left(\frac{\sigma^2\mu}{\Delta^2}\right)^{1/3} = \frac{8\sigma^2}{\mu}\cdot\left(\frac{\Delta\sigma^2}{\mu^2}\right)^{-2/3} \asymp \frac{\sigma^2}{\mu\mathcal{E}}.$$

This completes the proof.

A.2 The averaging lemma

We will use a small variation of the averaging lemma in [11]. To this end, consider a convex function $h\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ and let $\{x_t\}_{t\geq 0}$ be a sequence of vectors in $\operatorname{dom} h$. Suppose that there are constants $c_1, c_2 \in \mathbb{R}$, a nonnegative sequence of weights $\{\rho_t\}_{t\geq 1}$, and scalar sequences $\{V_t\}_{t\geq 0}$ and $\{\omega_t\}_{t\geq 1}$ satisfying the recursion

$$\rho_t h(x_t) \leq (1 - c_1\rho_t)V_{t-1} - (1 + c_2\rho_t)V_t + \omega_t \tag{15}$$

for all $t \geq 1$. The goal is to bound the function value $h(\hat{x}_t)$ evaluated along an "average iterate" $\hat{x}_t$.

Suppose that the relations $c_1 + c_2 > 0$, $1 - c_1\rho_t > 0$, and $1 + c_2\rho_t > 0$ hold for all $t \geq 1$. Define the augmented weights and products

$$\hat{\rho}_t = \frac{\rho_t(c_1 + c_2)}{1 + c_2\rho_t} \qquad \text{and} \qquad \hat{\Gamma}_t = \prod_{i=1}^{t}(1 - \hat{\rho}_i)$$

for each $t \geq 1$, while setting $\hat{\Gamma}_0 = 1$. A straightforward induction yields the relation

$$1 + \sum_{i=1}^{t} \frac{\hat{\rho}_i}{\hat{\Gamma}_i} = \frac{1}{\hat{\Gamma}_t}.$$

Now set $\hat{x}_0 = x_0$ and recursively define the average iterates

$$\hat{x}_t = (1 - \hat{\rho}_t)\hat{x}_{t-1} + \hat{\rho}_t x_t$$

for all $t \geq 1$. Unrolling this recursion, we may equivalently write

$$\hat{x}_t = \hat{\Gamma}_t \left(x_0 + \sum_{i=1}^{t} \frac{\hat{\rho}_i}{\hat{\Gamma}_i} x_i \right). \tag{16}$$

The following is the key estimate we will need.

Lemma A.1 (Averaging)**.** *The following estimate holds for all $t \geq 0$:*

$$\frac{h(\hat{x}_t)}{c_1 + c_2} + V_t \leq \hat{\Gamma}_t \left(\frac{h(x_0)}{c_1 + c_2} + V_0 + \sum_{i=1}^{t} \frac{\omega_i}{\hat{\Gamma}_i(1 + c_2\rho_i)} \right).$$

Proof. Observe that (16) expresses $\hat{x}_t$ as a convex combination of $x_0, \ldots, x_t$. Therefore, by the convexity of h we may apply Jensen's inequality to obtain

$$h(\hat{x}_t) \leq \hat{\Gamma}_t h(x_0) + \sum_{i=1}^{t} \frac{\hat{\Gamma}_t \hat{\rho}_i}{\hat{\Gamma}_i} h(x_i).$$

On the other hand, for each $i \geq 1$, we may divide the recursion (15) by $\hat{\Gamma}_i(1 + c_2\rho_i)$ to obtain

$$\frac{\hat{\rho}_i}{\hat{\Gamma}_i(c_1 + c_2)} h(x_i) \leq \frac{V_{i-1}}{\hat{\Gamma}_{i-1}} - \frac{V_i}{\hat{\Gamma}_i} + \frac{\omega_i}{\hat{\Gamma}_i(1 + c_2\rho_i)},$$

which telescopes to yield

$$\frac{1}{c_1 + c_2} \sum_{i=1}^{t} \frac{\hat{\rho}_i}{\hat{\Gamma}_i} h(x_i) \leq V_0 - \frac{V_t}{\hat{\Gamma}_t} + \sum_{i=1}^{t} \frac{\omega_i}{\hat{\Gamma}_i(1 + c_2\rho_i)}.$$

Hence

$$\frac{h(\hat{x}_t)}{c_1 + c_2} \leq \hat{\Gamma}_t \left(\frac{h(x_0)}{c_1 + c_2} + V_0 - \frac{V_t}{\hat{\Gamma}_t} + \sum_{i=1}^{t} \frac{\omega_i}{\hat{\Gamma}_i(1 + c_2\rho_i)} \right),$$

as claimed. □

A.3 Proof of Theorem 5.13

For each index k, let $t_k := T_0 + \cdots + T_{k-1}$ (with $t_0 := 0$) and $G_k := \eta_k\sigma^2 + 8\Delta^2/\mu\eta_\star^2$. Then taking into account $\eta_k \geq \eta_\star$, Corollary 5.12 and inequality (7) directly imply

$$\begin{aligned}\mathbb{E}\big[\varphi_{t_{k+1}}(X_{k+1}) - \varphi^\star_{t_{k+1}}\big] &\leq \Big(1 - \frac{\mu\eta_k}{2}\Big)^{T_k}\mathbb{E}\big[3\big(\varphi_{t_k}(X_k) - \varphi^\star_{t_k}\big) + 2\mu\Delta^2 T_k^2\big] + \eta_k\sigma^2 + \frac{8\Delta^2}{\mu\eta_k^2}\\ &\leq 3e^{-\mu\eta_k T_k/2}\mathbb{E}\big[\varphi_{t_k}(X_k) - \varphi^\star_{t_k}\big] + 2e^{-\mu\eta_k T_k/2}\mu\Delta^2 T_k^2 + G_k.\end{aligned}$$

We will verify by induction that the estimate $\mathbb{E}\big[\varphi_{t_{k+1}}(X_{k+1}) - \varphi^\star_{t_{k+1}}\big] \leq 7G_k$ holds for all indices k. To see the base case, observe that inequality (9) facilitates the estimation

$$\mathbb{E}\big[\varphi_{t_1}(X_1) - \varphi^\star_{t_1}\big] \leq 3e^{-\mu\eta_0 T_0/2}(\varphi_0(x_0) - \varphi_0^\star) + 2e^{-\mu\eta_0 T_0/2}\mu\Delta^2 T_0^2 + G_0 \leq 7G_0.$$

Assume next that the claim holds for index $k-1$. We then conclude

$$\begin{aligned}\mathbb{E}\big[\varphi_{t_{k+1}}(X_{k+1}) - \varphi^\star_{t_{k+1}}\big] &\leq 3e^{-\mu\eta_k T_k/2}\mathbb{E}\big[\varphi_{t_k}(X_k) - \varphi^\star_{t_k}\big] + 2e^{-\mu\eta_k T_k/2}\mu\Delta^2 T_k^2 + G_k\\ &\leq \frac{1}{4}\mathbb{E}\big[\varphi_{t_k}(X_k) - \varphi^\star_{t_k}\big] + \frac{16\Delta^2}{\mu\eta_k^2} + G_k\\ &\leq \frac{G_k}{2G_{k-1}}\mathbb{E}\big[\varphi_{t_k}(X_k) - \varphi^\star_{t_k}\big] + \frac{16\Delta^2}{\mu\eta_k^2} + G_k \leq 7G_k,\end{aligned}$$

completing the induction. Hence $\mathbb{E}\big[\varphi_T(X_K) - \varphi^\star_T\big] \leq 7G_{K-1}$.

Next, observe

$$G_{K-1} - \sqrt[3]{250}\cdot\mu\bigg(\frac{\Delta\sigma^2}{\mu^2}\bigg)^{2/3} = \sigma^2(\eta_{K-1} - \eta_\star) = \sigma^2\cdot\frac{\eta_0 - \eta_\star}{2^{K-1}} \leq \frac{\mu}{2}\bigg(\frac{\Delta\sigma^2}{\mu^2}\bigg)^{2/3} = \frac{1}{2}\mathcal{G},$$

so

$$\mathbb{E}\big[\varphi_T(X_K) - \varphi^\star_T\big] \leq 7\big(\tfrac{1}{2} + \sqrt[3]{250}\big)\cdot\mu\bigg(\frac{\Delta\sigma^2}{\mu^2}\bigg)^{2/3} \asymp \mathcal{G}.$$

Finally, note

$$T \lesssim \frac{L}{\mu}\log\bigg(\frac{LD}{\sigma^2}\bigg)^+ + \frac{1}{\mu}\sum_{k=1}^{K-1}\frac{1}{\eta_k}$$

and

$$\sum_{k=1}^{K-1}\frac{1}{\eta_k} \leq 2L\sum_{k=1}^{K-1}2^k \leq 2L\cdot 2^K = 8L\cdot 2^{K-2} \leq 8\bigg(\frac{\sigma^2\mu}{\Delta^2}\bigg)^{1/3} = 8\sigma^2\cdot\mu^{-1}\bigg(\frac{\Delta\sigma^2}{\mu^2}\bigg)^{-2/3} \asymp \frac{\sigma^2}{\mathcal{G}}.$$

This completes the proof.

A.4 Proof of Proposition 5.14

Fix $t \geq 1$. Given $i \geq 1$ and $\alpha > 0$, the μ-strong convexity of φ_t and Lemma 5.10 imply

$$\begin{aligned}\mu\eta\|x_i - x_t^\star\|^2 \leq 2\eta(\varphi_t(x_i) - \varphi_t^\star) &\leq (1-\mu\eta)\|x_{i-1} - x_t^\star\|^2 - (1-\alpha\eta)\|x_i - x_t^\star\|^2\\ &\quad + 2\eta\langle z_{i-1}, x_{i-1} - x_t^\star\rangle + 2\eta^2\|z_{i-1}\|^2 + \tfrac{\eta}{\alpha}G_{i-1,t}^2,\end{aligned}$$

hence

$$\begin{aligned}\big(1 + (\mu-\alpha)\eta\big)\|x_i - x_t^\star\|^2 &\leq (1-\mu\eta)\|x_{i-1} - x_t^\star\|^2 + 2\eta\langle z_{i-1}, x_{i-1} - x_t^\star\rangle\\ &\quad + 2\eta^2\|z_{i-1}\|^2 + \tfrac{\eta}{\alpha}G_{i-1,t}^2.\end{aligned}$$

Taking $\alpha = \mu$, we obtain

$$\|x_i - x_t^\star\|^2 \leq (1-\mu\eta)\|x_{i-1} - x_t^\star\|^2 + 2\eta\langle z_{i-1}, x_{i-1} - x_t^\star\rangle + 2\eta^2\|z_{i-1}\|^2 + \tfrac{\eta}{\mu}G_{i-1,t}^2.$$

Thus, given any $\lambda \in (0, \mu\eta]$ and proceeding by induction, we conclude

$$\|x_i - x_t^\star\|^2 \le (1-\lambda)^i \|x_0 - x_t^\star\|^2 + 2\eta \sum_{j=0}^{i-1} \langle z_j, x_j - x_t^\star \rangle (1-\lambda)^{i-1-j} + 2\eta^2 \sum_{j=0}^{i-1} \|z_j\|^2 (1-\lambda)^{i-1-j} + \tfrac{\eta}{\mu} \sum_{j=0}^{i-1} G_{j,t}^2 (1-\lambda)^{i-1-j}$$

for all $i \ge 1$. Therefore

$$\begin{aligned}\sum_{i=0}^{t-1} \|x_i - x_t^\star\|^2 (1-\lambda)^{2(t-1-i)} &\le \|x_0 - x_t^\star\|^2 \sum_{i=0}^{t-1} (1-\lambda)^{2(t-1)-i} + 2\eta \sum_{i=1}^{t-1}\sum_{j=0}^{i-1} \langle z_j, x_j - x_t^\star \rangle (1-\lambda)^{2t-3-j-i} \\ &\quad + 2\eta^2 \sum_{i=1}^{t-1}\sum_{j=0}^{i-1} \|z_j\|^2 (1-\lambda)^{2t-3-j-i} + \tfrac{\eta}{\mu} \sum_{i=1}^{t-1}\sum_{j=0}^{i-1} G_{j,t}^2 (1-\lambda)^{2t-3-j-i}.\end{aligned}$$

Now we compute

$$\sum_{i=0}^{t-1} (1-\lambda)^{2(t-1)-i} = (1-\lambda)^{t-1} \sum_{i=0}^{t-1} (1-\lambda)^{t-1-i} < \tfrac{1}{\lambda}(1-\lambda)^{t-1}$$

and observe that for any scalar sequence $(X_j)_{j=0}^{t-2}$, we have

$$\sum_{i=1}^{t-1}\sum_{j=0}^{i-1} X_j (1-\lambda)^{2t-3-j-i} = \sum_{j=0}^{t-2} \left(\sum_{i=j+1}^{t-1} (1-\lambda)^{t-2-i} \right) X_j (1-\lambda)^{t-1-j}.$$

Further, if $X_j \ge 0$ for all $j = 0, \ldots, t-2$, then we have

$$\begin{aligned}\sum_{i=1}^{t-1}\sum_{j=0}^{i-1} X_j (1-\lambda)^{2t-3-j-i} &= \sum_{j=0}^{t-2} \left(\sum_{i=j+1}^{t-1} (1-\lambda)^{t-1-i} \right) X_j (1-\lambda)^{t-2-j} \\ &\le \tfrac{1}{\lambda} \sum_{j=0}^{t-2} X_j (1-\lambda)^{t-2-j}.\end{aligned}$$

Hence the following estimation holds:

$$\begin{aligned}\sum_{i=0}^{t-1} \|x_i - x_t^\star\|^2 (1-\lambda)^{2(t-1-i)} &\le \sum_{j=0}^{t-2} \left(2\eta \sum_{i=j+1}^{t-1} (1-\lambda)^{t-2-i} \right) \langle z_j, x_j - x_t^\star \rangle (1-\lambda)^{t-1-j} \\ &\quad + \tfrac{1}{\lambda}(1-\lambda)^{t-1} \|x_0 - x_t^\star\|^2 + \tfrac{2\eta^2}{\lambda} \sum_{j=0}^{t-2} \|z_j\|^2 (1-\lambda)^{t-2-j} \\ &\quad + \tfrac{\eta}{\mu\lambda} \sum_{j=0}^{t-2} G_{j,t}^2 (1-\lambda)^{t-2-j}.\end{aligned}$$

This completes the proof.

A.5 Proof of Theorem 5.19

For each index k, let $t_k := T_0 + \cdots + T_{k-1}$ (with $t_0 := 0$) and $G_k := \eta_k\sigma^2 + \Delta^2/\mu\eta_\star^2$. Then taking into account $\eta_k \geq \eta_\star$ and our selection of the absolute constant c via (3), it follows that for all indices k the estimate

$$\begin{aligned}\varphi_{t_{k+1}}(X_{k+1}) - \varphi^\star_{t_{k+1}} &\leq c\bigg(\Big(1 - \frac{\mu\eta_k}{2}\Big)^{T_k}(\varphi_{t_k}(X_k) - \varphi^\star_{t_k}) + \eta_k\sigma^2 + \frac{\Delta^2}{\mu\eta_k^2}\bigg)\log\Big(\frac{e}{\delta}\Big) \\ &\leq c\Big(e^{-\mu\eta_k T_k/2}(\varphi_{t_k}(X_k) - \varphi^\star_{t_k}) + G_k\Big)\log\Big(\frac{e}{\delta}\Big)\end{aligned}$$

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

We will verify by induction that for all indices $k \geq 1$, the estimate

$$\varphi_{t_k}(X_k) - \varphi^\star_{t_k} \leq 3cG_{k-1}\log\Big(\frac{e}{\delta}\Big)$$

holds with probability at least $1-k\delta$. To see the base case, observe that the estimate

$$\varphi_{t_1}(X_1) - \varphi^\star_{t_1} \leq c\Big(e^{-\mu\eta_0 T_0/2}(\varphi_0(x_0) - \varphi^\star_0) + G_0\Big)\log\Big(\frac{e}{\delta}\Big) \leq 3cG_0\log\Big(\frac{e}{\delta}\Big)$$

holds with probability at least $1-\delta$. Now assume the claim holds for some index $k \geq 1$. Then because we also have

$$\begin{aligned}\varphi_{t_{k+1}}(X_{k+1}) - \varphi^\star_{t_{k+1}} &\leq c\Big(e^{-\mu\eta_k T_k/2}(\varphi_{t_k}(X_k) - \varphi^\star_{t_k}) + G_k\Big)\log\Big(\frac{e}{\delta}\Big) \\ &\leq c\Big(\frac{1}{4c\log(e/\delta)}(\varphi_{t_k}(X_k) - \varphi^\star_{t_k}) + G_k\Big)\log\Big(\frac{e}{\delta}\Big) \\ &\leq c\Big(\frac{G_k}{2cG_{k-1}\log(e/\delta)}(\varphi_{t_k}(X_k) - \varphi^\star_{t_k}) + G_k\Big)\log\Big(\frac{e}{\delta}\Big)\end{aligned}$$

with probability at least $1-\delta$, a union bound reveals that the estimate

$$\varphi_{t_{k+1}}(X_{k+1}) - \varphi^\star_{t_{k+1}} \leq 3cG_k\log\Big(\frac{e}{\delta}\Big)$$

holds with probability at least $1-(k+1)\delta$, thereby completing the induction. In particular, $\varphi_T(X_K) - \varphi^\star_T \leq 3cG_{K-1}\log(e/\delta)$ with probability at least $1-K\delta$.

Next, observe

$$G_{K-1} - \sqrt[3]{\tfrac{27}{4}}\cdot\mu\bigg(\frac{\Delta\sigma^2}{\mu^2}\bigg)^{2/3} = \sigma^2(\eta_{K-1} - \eta_\star) = \sigma^2\cdot\frac{\eta_0 - \eta_\star}{2^{K-1}} \leq \frac{\mu}{2}\bigg(\frac{\Delta\sigma^2}{\mu^2}\bigg)^{2/3},$$

so

$$\varphi_T(X_K) - \varphi^\star_T \leq 3c\Big(\tfrac{1}{2} + \sqrt[3]{\tfrac{27}{4}}\Big)\cdot\mu\bigg(\frac{\Delta\sigma^2}{\mu^2}\bigg)^{2/3}\log\Big(\frac{e}{\delta}\Big) \asymp \mathcal{G}\log\Big(\frac{e}{\delta}\Big)$$

with probability at least $1-K\delta$. Finally, note

$$T \lesssim \frac{L}{\mu}\log\Big(\frac{LD}{\sigma^2}\Big)^+ + \Big(1 \vee \log\log\frac{e}{\delta}\Big)\frac{1}{\mu}\sum_{k=1}^{K-1}\frac{1}{\eta_k}$$

and

$$\sum_{k=1}^{K-1}\frac{1}{\eta_k} \leq 2L\sum_{k=1}^{K-1}2^k \leq 2L\cdot 2^K = 8L\cdot 2^{K-2} \leq 8\bigg(\frac{\sigma^2\mu}{\Delta^2}\bigg)^{1/3} = 8\sigma^2\cdot\mu^{-1}\bigg(\frac{\Delta\sigma^2}{\mu^2}\bigg)^{-2/3} \asymp \frac{\sigma^2}{\mathcal{G}}.$$

This completes the proof.