Journal of Machine Learning Research 24 (2023) 1-56 Submitted 11/21; Revised 10/22; Published 2/23

Stochastic Optimization under Distributional Drift

Joshua Cutler JOCUTLER@UW.EDU
Department of Mathematics
University of Washington
Seattle, WA 98195-4322, USA

Dmitriy Drusvyatskiy DDRUSV@UW.EDU
Department of Mathematics
University of Washington
Seattle, WA 98195-4322, USA

Zaid Harchaoui ZAID@UW.EDU
Department of Statistics
University of Washington
Seattle, WA 98195-4322, USA

Editor: Alekh Agarwal

Abstract

We consider the problem of minimizing a convex function that is evolving according to unknown and possibly stochastic dynamics, which may depend jointly on time and on the decision variable itself. Such problems abound in the machine learning and signal processing literature, under the names of concept drift, stochastic tracking, and performative prediction. We provide novel non-asymptotic convergence guarantees for stochastic algorithms with iterate averaging, focusing on bounds valid both in expectation and with high probability. The efficiency estimates we obtain clearly decouple the contributions of optimization error, gradient noise, and time drift. Notably, we identify a low drift-to-noise regime in which the tracking efficiency of the proximal stochastic gradient method benefits significantly from a step decay schedule. Numerical experiments illustrate our results.

Keywords: stochastic gradient, stochastic tracking, concept drift, performative prediction, high-probability bounds

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 (Bottou, 2003; Bottou and Bousquet, 2007; Bottou, 2012; Srebro et al., 2011; Agarwal et al., 2014; Lang et al., 2019; Bach and Moulines, 2011; Sra et al., 2011; Nemirovski et al., 2009). The predominant assumption in much of the work on stochastic optimization for machine learning 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. There are two main sources of such distributional shifts. The first is temporal, wherein the distribution varies slowly in time due to reasons that are independent of the learning process. This setting is often called dynamic stochastic approximation (Dupač, 1965) and is the basis for adaptive algorithms for stochastic track-

©2023 Joshua Cutler, Dmitriy Drusvyatskiy, Zaid Harchaoui.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v24/21-1410.html.

ing (Benveniste et al., 1990). The second common source is due to a feedback mechanism, wherein the distribution generating the data may depend on, or react to, the decisions made by the learner. This setting has been a subject of increased interest recently in the context of strategic classification and performative prediction (Mendler-Dünner et al., 2020; Drusvyatskiy and Xiao, 2022).

In this work, 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. The results are presented in a single framework that encompasses both the purely temporal and the decision-dependent 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. While conventional wisdom and previous work recommend the use of a constant step size under time drift, we identify a low drift-to-noise regime in which tracking efficiency benefits significantly from a step size schedule that geometrically decays to a "critical step size".

Setting the stage, consider the sequence of stochastic optimization problems

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

indexed by time $t \in \mathbb{N}$. In typical machine learning and signal processing settings, the function f_t corresponds to an average loss that varies in time, while the regularizer r_t models constraints or promotes structure (e.g., sparsity) in the variable x. Two examples are worth highlighting. The first is a classical problem in signal processing related to stochastic tracking (Kushner and Yin, 1997; Sayed, 2003), 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 (Hazan and Seshadhri, 2009; Zhang et al., 2018), wherein the true hypothesis may be changing over time.

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. To make progress, we impose the standard assumption that each function f_t is μ-strongly convex with L-Lipschitz continuous gradient, while each regularizer r_t is proper, closed, and convex. The online proximal stochastic gradient method (PSG) naturally applies to the sequence of problems (1). At each iteration t, the method simply takes the step

$$x_{t+1} = \mathrm{prox}_{\eta_t r_t}(x_t - \eta_t g_t),$$

where the vector g_t is an unbiased estimator of the true gradient of f_t at x_t, the step size (learning rate) $\eta_t > 0$ is user-specified, and $\mathrm{prox}_{\eta_t r_t}(\cdot)$ is the proximal map of the scaled regularizer $\eta_t r_t$. In this work, we analyze two types of tracking error for PSG: the squared distance $\|x_t - x_t^\star\|^2$ and the suboptimality gap $\varphi_t(\hat{x}_t) - \varphi_t(x_t^\star)$. Here, $x_t^\star$ denotes the minimizer of the function φ_t which may evolve stochastically in time, and $\hat{x}_t$ denotes a weighted average of iterates up to time t. We next outline the main results of the paper; the results in Sections 1.1 and 1.2 below appeared in a preliminary version of this paper at NeurIPS (Cutler et al., 2021).

1.1 Tracking the Minimizer

We begin with a simple bound on distance tracking of the constant-step PSG:

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

Here $\eta \in (0, 1/2L]$ is the constant step size used by PSG, σ^2 upper-bounds the variance of the stochastic gradient, and Δ^2 upper-bounds the minimizer variations $\mathbb{E}\|x_t^\star - x_{t+1}^\star\|^2$; the symbol $\lesssim$ indicates an inequality that holds up to an absolute constant factor, i.e., up to multiplying the upper bound by a positive numerical constant independent of the problem parameters. Inequality (2) asserts that the tracking 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 by tuning η. This is perfectly in line with intuition: a step size η that is too small prevents the algorithm from catching up with the minimizers $x_t^\star$. 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 by Madden et al. (2021, Theorem 3.2). Though the estimate (2) is likely known, we were unable to find a precise reference in this generality.

Letting t tend to infinity in (2), the optimization error tends to zero, leaving only the "noise + drift" term. Optimizing this remaining term over η, it is natural to define the asymptotic distance tracking error of PSG and the corresponding optimal learning rate as

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

Two regimes of variation are brought to light: 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 in this case the optimal learning rate $\eta_\star \asymp 1/L$ is as large as in the deterministic setting (here, the symbol $\asymp$ indicates an equality that holds up to an absolute constant factor). In contrast, the low drift-to-noise regime is interesting because it necessitates using a smaller learning rate $\eta_\star \asymp (\Delta^2/\mu\sigma^2)^{1/3}$ that exhibits a nontrivial scaling with the problem parameters. Consequently, for the rest of the introduction we focus on the low drift-to-noise regime.

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 simplest strategy is to execute PSG with the constant learning rate $\eta_\star$. Then a direct application of (2) yields the efficiency 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 (stationary objective) setting, as was discovered by Ghadimi and Lan (2013), and can be used here. The end result is an algorithm that produces a point x_t satisfying

$$\mathbb{E}\|x_t - x_t^\star\|^2 \lesssim \mathcal{E} \quad \text{in time} \quad t \lesssim \frac{L}{\mu}\log\left(\frac{\|x_0 - x_0^\star\|^2}{\mathcal{E}}\right) + \frac{\sigma^2}{\mu^2\mathcal{E}}. \tag{3}$$

This efficiency estimate is remarkably similar to that in the static setting (Ghadimi and Lan, 2013), with $\mathcal{E}$ playing the role of the target accuracy ε. An elementary computation shows that (3) improves the constant learning rate efficiency estimate when $\mathcal{E}$ is small, e.g., when $\mathcal{E} \leq \|x_0 - x_0^\star\|^2/e^2$, where e denotes Euler's number.

The efficiency estimate (3) is a baseline guarantee for PSG with step decay. Since the result is stated in terms of the *expected* tracking error $\mathbb{E}\|x_t - x_t^\star\|^2$, it is only meaningful if the entire algorithm can be repeated from scratch multiple times on the same problem.[1] However, 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. These situations call for efficiency estimates that hold with high probability, rather than only in expectation. With this in mind, we show that under mild light-tail assumptions, PSG with step decay produces a point x_t satisfying $\|x_t - x_t^\star\|^2 \lesssim \mathcal{E}\log(1/\delta)$ with probability at least $1-\delta$ in the same order of iterations as in (3). The proof follows closely the probabilistic techniques developed by Harvey et al. (2019) for bounding moment generating functions.

1.2 Tracking the Minimum Value

The results outlined so far have focused on tracking the minimizer $x_t^\star$; stronger guarantees may be obtained for tracking the minimum value $\varphi_t^\star$. To this end, we require stronger assumptions on the variation of the functions f_t beyond control on the minimizer drift $\|x_t^\star - x_{t+1}^\star\|^2$. Similar in spirit to the measure of cumulative gradient variation in the dynamic online learning literature (e.g., see Jadbabaie et al., 2015), we will be concerned with the *gradient drift*

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

and assume the bound $\mathbb{E}[G_{i,t}^2/\mu^2] \leq \Delta^2|i-t|^2$ for all times i and t. Thus, the second moment of the gradient drift is assumed to grow at most quadratically in the time horizon. Assuming henceforth that the regularizers $r_t \equiv r$ are identical for all times t, this condition on the gradient drift implies the weaker assumption $\mathbb{E}\|x_t^\star - x_{t+1}^\star\|^2 \leq \Delta^2$ used in Section 1.1.

Analogous to (2), we show that PSG generates a point $\hat{x}_t$ (an average iterate) satisfying

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

This estimate again decouples nicely into three terms, signifying the error due to optimization, gradient noise, and time drift. Taking the limit as t tends to infinity, we obtain the asymptotic function gap tracking error $\mathcal{G} := \mu\mathcal{E}$. Similar to (3), we show that PSG with step decay produces a point $\hat{x}_t$ satisfying

$$\mathbb{E}\big[\varphi_t(\hat{x}_t) - \varphi_t^\star\big] \lesssim \mathcal{G} \quad \text{in time} \quad t \lesssim \frac{L}{\mu}\log\left(\frac{\varphi_0(x_0) - \varphi_0^\star}{\mathcal{G}}\right) + \frac{\sigma^2}{\mu\mathcal{G}}. \tag{4}$$

1. Specifically, error bounds holding in expectation yield concentration inequalities for the average of i.i.d. errors arising from executing a stochastic algorithm multiple times from scratch on the same problem (e.g., via Chebyshev's inequality); if the algorithm cannot be executed in this fashion to generate i.i.d. errors, then alternative high-probability error bounds are called for.

Again, the similarity to the static setting (Ghadimi and Lan, 2013), with $\mathcal{G}$ playing the role of a target accuracy, is striking. We then provide a high-probability extension of this estimate: under mild light-tail assumptions, PSG with step decay produces a point $\hat{x}_t$ satisfying $\varphi_t(\hat{x}_t) - \varphi_t^\star \lesssim \mathcal{G}\log(1/\delta)$ with probability at least $1-\delta$ in the same order of iterations as in (4) up to a factor of $\log\log(1/\delta)$. The proofs are based on the generalized Freedman inequality of Harvey et al. (2019)—a remarkably flexible tool for analyzing stochastic gradient-type algorithms.

1.3 Extension to Decision-Dependent Problems with Time Drift

We have so far focused on stochastic optimization problems that undergo a temporal shift. A primary reason for this phenomenon in machine learning, and data science more broadly, is that data distributions often evolve in time independently of the learning process. Recent literature, on the other hand, highlights a different source of distributional shift due to decision-dependent or performative effects. Namely, the distribution generating the data in iteration t may depend on, or react to, the current "decision" x_t. For example, deployment of a classifier by a learning system, when made public, often causes the population to adapt their attributes in order to increase the likelihood of being positively labeled—a process called "gaming". Even when the population is agnostic to the classifier, the decisions made by the learning system (e.g., loan approval) may inadvertently alter the profile of the population (e.g., credit score). The goal of the learning system therefore is to find a classifier that generalizes well under the response distribution. Recent research in strategic classification (Hardt et al., 2016; Brückner et al., 2012; Bechavod et al., 2021; Dalvi et al., 2004) and performative prediction (Perdomo et al., 2020; Mendler-Dünner et al., 2020) has highlighted the prevalence of this phenomenon.

Combining time-dependence and decision-dependence yields a class of problems (1) where the loss function $f_t(x)$ takes the special form $f_t(x) = \mathbb{E}_{\xi\sim\mathcal{D}(t,x)}\ell(x,\xi)$. Here $\mathcal{D}(t,x)$ is a distribution that depends on both time t and the decision variable x. Thus for any fixed time t, the problem (1) becomes the performative risk problem considered by Perdomo et al. (2020) and Mendler-Dünner et al. (2020). Following this line of work, instead of tracking the true minimizer of φ_t—typically a challenging task—we will settle for tracking the equilibrium points $\bar{x}_t$. These are the points satisfying

$$\bar{x}_t \in \operatorname*{argmin}_x \mathop{\mathbb{E}}_{\xi\sim\mathcal{D}(t,\bar{x}_t)} \ell(x,\xi) + r(x).$$

Equilibrium points are sure to exist and are unique under mild Lipschitzness and strong convexity assumptions. We refer the reader to Perdomo et al. (2020) for a compelling motivation for considering such equilibrium points. The problem of tracking equilibrium points is yet again an instance of (1), but now with the different function $f_t(x) = \mathbb{E}_{\xi\sim\mathcal{D}(t,\bar{x}_t)}\ell(x,\xi)$ induced by the equilibrium distributions. The PSG algorithm is not directly applicable here since the learner cannot typically sample from $\mathcal{D}(t,\bar{x}_t)$ directly. Instead, a natural algorithm for this problem class draws in each iteration t a sample ξ_t from the current distribution $\mathcal{D}(t,x_t)$ and declares $x_{t+1} = \operatorname{prox}_{\eta_t r}(x_t - \eta_t\nabla\ell(x_t,\xi_t))$. Notice that the sample gradient $\nabla\ell(x_t,\xi_t)$ is a biased estimator of the true gradient $\nabla f_t(x_t) = \mathbb{E}_{\xi\sim\mathcal{D}(t,\bar{x}_t)}\nabla\ell(x_t,\xi)$ because ξ_t is sampled from the wrong distribution. Nonetheless, as pointed out by Drusvyatskiy and Xiao (2022), the gradient bias is small for any fixed time, decaying linearly with the distance

to $\bar{x}_t$. Using this perspective, we show that all guarantees for PSG in the time-dependent setting naturally extend to this biased PSG algorithm for tracking equilibrium points, with essentially no loss in efficiency.

1.4 Related Work

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

Stochastic tracking. Stochastic optimization with time drift was considered soon after the Robbins-Monro approach for stochastic optimization was introduced; see Kushner and Yin (1997) for a survey. Early results can be traced back to Dupač (1965) in sequential estimation and Gaivoronskii (1978) in stochastic optimization; see also Fujita and Fukao (1972), Ruppert (1979), Tsypkin and Nikolic (1971), Tsypkin and Polyak (1992), and Uosaki (1974). Stochastic algorithms have also been extensively studied as adaptive algorithms for stochastic tracking (Tsypkin and Nikolic, 1971; Kushner and Yin, 1997; Benveniste et al., 1990), for their ability to indeed track parameters under time drift. 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 related Kalman filtering algorithms (Guo and Ljung, 1995).

Recent works have revisited these methods from a more modern viewpoint (Besbes et al., 2015; Wilson et al., 2019; Madden et al., 2021). In particular, the paper of Madden et al. (2021) focuses on (accelerated) gradient methods for deterministic tracking problems, while Wilson et al. (2019) present a framework for online stochastic gradient methods with parameter estimation. The work of Besbes et al. (2015) 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. In particular, the guarantees (3) and (4) for PSG with step decay, along with their high-probability variants, are new to the best of our knowledge.

Online optimization with dynamic regret. Online optimization for sequences of convex objectives with domain $\mathcal{X}$ has been studied through the lens of adaptive regret (Hazan and Seshadhri, 2009; Daniely et al., 2015) and dynamic regret (Besbes et al., 2015; Zinkevich, 2003; Zhang et al., 2018; Zhao et al., 2020; Jadbabaie et al., 2015; Mokhtari et al., 2016). The adaptive regret

$$\sup_{[r,s]\subset[T]} \left\{ \sum_{t=r}^{s} f_t(x_t) - \inf_{x\in\mathcal{X}} \sum_{t=r}^{s} f_t(x) \right\}$$

reads as the maximum static regret over any contiguous time interval; more relevant to our analysis is the dynamic regret

$$\mathrm{Reg}_T^\star = \sum_{t=1}^{T} \big(f_t(x_t) - f_t(x_t^\star) \big),$$

which reads as the cumulative difference between the instantaneous loss and the minimum loss. More generally, one can consider the dynamic regret against an arbitrary comparator sequence $\{u_t\}_{t=1}^T$ in $\mathcal{X}$, given by

$$\mathrm{Reg}_T(u_1, \ldots, u_T) = \sum_{t=1}^{T} \big(f_t(x_t) - f_t(u_t)\big).$$

Jadbabaie et al. (2015) apply an adaptive step size strategy to an optimistic mirror descent algorithm, thereby obtaining a comprehensive dynamic regret guarantee for $\mathrm{Reg}_T^\star$ in terms of the cumulative loss variation $V_T = \sum_{t=2}^T \sup_{x\in\mathcal{X}} |f_t(x) - f_{t-1}(x)|$, the cumulative gradient variation $D_T = \sum_{t=1}^T \|\nabla f_t(x_t) - M_t\|^2$ using a causally predictable sequence M_t available to the algorithm prior to time t (e.g., $M_t = \nabla f_{t-1}(x_{t-1})$), and the cumulative minimizer variation $C_T^\star = \sum_{t=2}^T \|x_t^\star - x_{t-1}^\star\|$. Under strong convexity, Mokhtari et al. (2016) show that online projected gradient descent satisfies $\mathrm{Reg}_T^\star \le \mathcal{O}(1 + C_T^\star)$. For convex losses with bounded domain $\mathcal{X}$, Zhang et al. (2018) present an adaptive online gradient method that achieves an optimal dynamic regret bound in terms of the cumulative comparator variation $C_T = \sum_{t=2}^T \|u_t - u_{t-1}\|$, namely, $\mathrm{Reg}_T(u_1, \ldots, u_T) \le \mathcal{O}\big(\sqrt{T(1 + C_T)}\big)$. This last guarantee is enhanced by Zhao et al. (2020) through exploiting smoothness to replace the time horizon T by problem-dependent quantities that are at most $\mathcal{O}(T)$ but often much smaller in easy problems.

As is standard in the dynamic online optimization literature, the preceding works assume that either the losses $f_t(x)$ or their gradients $\nabla f_t(x)$ are uniformly bounded in both t and x, and a priori knowledge of these uniform bounds is required for the aforementioned guarantees. In contrast, we take great care to make no such uniform boundedness assumptions and we work instead with bounded second moments or light tails of minimizer or gradient drift, which we allow to evolve stochastically. Furthermore, we only assume stochastic gradient access, and the presence of stochasticity in the drift and the gradient noise requires guarantees that hold both in expectation and with high probability. Our bounds depend on a characteristic quantity of the problem difficulty encapsulating the drift and the noise level and hence delineate two regimes depending on the drift-to-noise ratio. In general, regret bounds do not entail last-iterate bounds of the type presented in our work.

High-probability guarantees in stochastic optimization. A large part of our work revolves around high-probability guarantees for stochastic optimization. Classical references on the subject in static settings and for minimizing regret in online optimization include the work of Bartlett et al. (2008), Hazan and Kale (2014), Lan (2012), and Rakhlin et al. (2012). There exists a variety of techniques for establishing high-probability guarantees based on Freedman's inequality and doubling tricks (e.g., see Bartlett et al., 2008; Hazan and Kale, 2014). A more recent line of work by Harvey et al. (2019) establishes a generalized Freedman inequality that is custom-tailored for analyzing stochastic gradient-type methods and results in the best known high-probability guarantees. Our arguments closely follow the paradigm of Harvey et al. (2019) based on the generalized Freedman inequality.

Performative prediction and decision-dependent learning. Recent works on strategic classification (Hardt et al., 2016; Brückner et al., 2012; Bechavod et al., 2021; Dalvi et al., 2004) and performative prediction (Perdomo et al., 2020; Mendler-Dünner et al., 2020) have highlighted the importance of strategic behavior in machine learning. That is, common

learning systems exhibit a feedback mechanism, wherein the distribution generating the data in iteration t may depend on, or react to, the current "decision" of an algorithm x_t. The recent paper by Perdomo et al. (2020) put forth an elegant framework for thinking about such problems, while Mendler-Dünner et al. (2020) develop stochastic algorithms for this setting. The subsequent work of Drusvyatskiy and Xiao (2022) shows that a variety of stochastic algorithms for performative prediction can be understood as biased variants of the same algorithms on a certain static problem in equilibrium. Building on the techniques of Drusvyatskiy and Xiao (2022), we show how all our results for time-dependent problems extend to problems that simultaneously depend on time and on the decision variable. We note that during the final stage of completing this paper, the closely related and complementary work by Wood et al. (2022) was posted on arXiv.[2] The paper by Wood et al. (2022) considers decision-dependent projected stochastic gradient descent under time drift in the distributional framework proposed by Perdomo et al. (2020), establishing distance tracking bounds in expectation and with high probability under sub-Weibull gradient noise. In particular, the light-tail assumption on gradient noise used by Wood et al. (2022) for obtaining high-probability guarantees is more general than the one in our paper. On the other hand, we analyze tracking of both the minimizer and the minimum value of more general stochastically evolving objectives, allow presence of general convex regularizers, and propose a step decay schedule for improved efficiency.

1.5 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 3–5 summarize the main results of the paper. Specifically, Section 3 focuses on efficiency estimates for tracking the minimizer, Section 4 focuses on efficiency estimates for tracking the minimum value, and Section 5 develops an extension to the decision-dependent setting via tracking equilibria. Section 6 presents the proofs of the main results in a unified framework. Illustrative numerical results appear in Section 7. Appendix A describes the averaging technique used for tracking function values, and additional proofs appear in Appendix B.

2. Framework and Assumptions

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

$$\min_{x\in\mathbb{R}^d} \varphi_t(x) := f_t(x) + r_t(x) \tag{5}$$

indexed by time $t \in \mathbb{N}$, where $\mathbb{R}^d$ denotes a fixed d-dimensional Euclidean space with inner product $\langle\cdot,\cdot\rangle$ and Euclidean norm $\|x\| = \sqrt{\langle x, x\rangle}$, and the following standard regularity assumptions hold:

(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 $\mu, L > 0$.

2. More precisely, a short version of our paper (Cutler et al., 2021) was submitted to NeurIPS in May '21, the paper by Wood et al. (2022) appeared on arXiv in July '21, and our full paper was posted on arXiv in August '21. Our paper (Cutler et al., 2021) was presented at NeurIPS in December '21.

(ii) Each regularizer $r_t \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is proper, closed, and convex.[3]

The minimizer and minimum value of φ_t will be denoted by $x_t^\star$ and $\varphi_t^\star$, respectively. We will be concerned with settings in which φ_t evolves stochastically in time. As motivation, we describe two classical examples of (5) that are worth keeping in mind and that guide our framework: stochastic tracking of a drifting target and online learning under distributional drift.

Example 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 the loss $\ell_t(\cdot)$ derives from the distribution of ϵ_t and the regularizer $r_t(\cdot)$ 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 (Guo and Ljung, 1995; Sayed, 2003).

Example 2 (Online learning under distributional drift) The problem of online learning under distributional drift is to learn while the data distribution changes over time. More formally, a typical problem formulation takes the form

$$\min_x \mathop{\mathbb{E}}_{\xi \sim \mathcal{D}(v_t)} \ell(x, \xi) + r(x),$$

where $\mathcal{D}(v_t)$ is a data distribution that depends on an unknown parameter sequence $\{v_t\}$, which itself may evolve stochastically.

The main goal of a learning algorithm for problem (5) 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: the squared distance $\|x_t - x_t^\star\|^2$ and the suboptimality gap $\varphi_t(\hat{x}_t) - \varphi_t(x_t^\star)$, where $\hat{x}_t$ denotes a weighted average of iterates up to time t.

We make the standing assumption that at every time t, and at every query point x, the learner can select 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 below—selects in each iteration t the stochastic gradient $g_t = \widetilde{\nabla} f_t(x_t)$ and takes the step

$$x_{t+1} := \operatorname{prox}_{\eta_t r_t}(x_t - \eta_t g_t) = \operatorname*{argmin}_{u \in \mathbb{R}^d} \left\{ r_t(u) + \tfrac{1}{2\eta_t} \|u - (x_t - \eta_t g_t)\|^2 \right\}$$

using step size $\eta_t > 0$. The goal of our work is to obtain efficiency estimates for this procedure that hold both in expectation and with high probability.

3. We assume $\operatorname{dom} f_t = \mathbb{R}^d$ for simplicity, but this is not essential. For example, it suffices to assume that each function $f_t \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is closed and μ-strongly convex and that there exists an open convex set $U \subset \mathbb{R}^d$ such that for all $t \in \mathbb{N}$, $\operatorname{dom} r_t \subset U \subset \operatorname{dom} f_t$ and f_t is L-smooth on U.

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

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

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

Return x_T

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 (Example 1), or when tracking the minimizer of an expected loss over a stochastically evolving data distribution (Example 2). 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 on a fixed probability space $(\Omega, \mathcal{F}, \mathbb{P})$ throughout Sections 2–4.

Assumption 1 (Stochastic framework) There exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with filtration $(\mathcal{F}_t)_{t \geq 0}$ such that $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and the following two conditions hold 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 $\mathbb{E}[z_t \,|\, \mathcal{F}_t] = 0$.

The first item of Assumption 1 formalizes the assertion that x_t and $x_t^\star$ are fully determined by information up to time t. The second item of Assumption 1 formalizes the assertion 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, i.e., g_t is an unbiased estimator of $\nabla f_t(x_t)$; for example, this holds naturally in Example 2 under typical regularity assumptions if $g_t = \nabla \ell(x_t, \xi_t)$ with $\xi_t \sim \mathcal{D}(v_t)$, where $\nabla \ell(x_t, \xi_t)$ denotes the gradient of $\ell(\cdot, \xi_t)$ at x_t.

Efficiency estimates for Algorithm 1 must clearly take into account the variation of the problem (5) in time t. One of the standard metrics for measuring this variation is the *minimizer drift*

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

Another popular metric is the *gradient drift*

$$\sup_x \|\nabla f_t(x) - \nabla f_{t+1}(x)\|.$$

Our efficiency estimates for tracking the minimizer will depend on the minimizer drift, while our efficiency estimates for tracking the minimum value will depend on the gradient drift. As the following elementary lemma shows, the minimizer drift scaled by μ is dominated by the gradient drift whenever the regularizers do not vary in time.[4]

4. Lemma 1 provides a bound similar in spirit to the bound $\mu\|x_i^\star - x_t^\star\|^2 \leq 4 \sup_{x \in \operatorname{dom} r} |f_i(x) - f_t(x)|$ in terms of variation in function value, which is also an elementary consequence of μ-strong convexity (e.g., see Zhao and Zhang, 2021, Section 4.1).

Lemma 1 (Minimizer vs. gradient drift) *Suppose that i and t are indices for which the regularizers r_i and r_t are identical. Then*

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

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 $\mu\|x_i^\star - x_t^\star\| \le \|0 - v\|$. ■

3. Tracking the Minimizer

This section presents 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 $\mathbb{E}\|x_t - x_t^\star\|^2$. Proofs appear in Section 6.1. The starting point for our analysis is the following standard one-step improvement guarantee.

Lemma 2 (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:*

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

For simplicity, we state the main results under the assumption that the second moments $\mathbb{E}\,\Delta_t^2$ 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 2 as well.

Assumption 2 (Bounded second moments) There exist constants $\Delta, \sigma > 0$ such that the following two conditions hold for all $t \ge 0$:

(i) **(Drift)** The minimizer drift Δ_t satisfies $\mathbb{E}\,\Delta_t^2 \le \Delta^2$.

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

The following theorem establishes an expected improvement guarantee for Algorithm 1, and serves as the basis for much of what follows.

Theorem 3 (Expected distance) *Suppose that Assumption 2 holds. Then the iterates produced by Algorithm 1 with constant learning rate $\eta \le 1/2L$ 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 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 by Madden et al. (2021, Theorem 3.2).

With Theorem 3 in hand, we define the 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 with high or low drift-to-noise ratio Δ/σ. Explicitly,

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

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 in this case the optimal learning rate $\eta_\star \asymp 1/L$ is as large as in the deterministic setting. In contrast, the low drift-to-noise regime is interesting because it necessitates using a smaller learning rate $\eta_\star \asymp (\Delta^2/\mu\sigma^2)^{1/3}$ that 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 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 this 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 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 by Ghadimi and Lan (2013), and can be used here. The end result is the following theorem; see Theorem 28 for the formal statement.

Theorem 4 (Time to track in expectation, informal) *Suppose that Assumption 2 holds. Then there is a learning rate schedule $\{\eta_t\}$ such that Algorithm 1 produces a point x_t*

satisfying

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

The efficiency estimate in Theorem 4 is strikingly similar to the efficiency estimate in the static setting (Ghadimi and Lan, 2013), with $\mathcal{E}$ playing the role of the target accuracy ε. An elementary computation shows that in the low drift-to-noise regime, Theorem 4 improves the constant learning rate efficiency estimate when $\mathcal{E}$ is small, e.g., when $\mathcal{E} \le \|x_0 - x_0^\star\|^2/e^2$. Theorems 3 and 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. These situations 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 for the tracking error $\|x_t - x_t^\star\|^2$. Proofs appear in Section 6.2. We make the following standard light-tail assumptions on the minimizer drift and gradient noise (Harvey et al., 2019; Lan, 2012; Nemirovski et al., 2009).

Assumption 3 (Sub-Gaussian drift and noise) There exist constants $\Delta, \sigma > 0$ such that the following two conditions hold for all $t \ge 0$:

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

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

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

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

Note that the first item of Assumption 3 is equivalent to asserting that the minimizer drift Δ_t is sub-Gaussian conditioned on $\mathcal{F}_t$ (see Vershynin, 2018, Lemma 2.7.6). Clearly Assumption 3 implies Assumption 2 with the same constants Δ, σ. It is worthwhile to note some common settings in which Assumption 3 holds; the claims in Remark 5 below follow from standard results on sub-Gaussian random variables (Jin et al., 2019; Vershynin, 2018).

Remark 5 (Common settings for Assumption 3) 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 $\sigma/2$, 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 holds, then the expected bound on $\|x_t - x_t^\star\|^2$ derived in Theorem 3 holds with high probability.

Theorem 6 (High-probability distance tracking) *Let $\{x_t\}$ be the iterates produced by Algorithm 1 with constant learning rate $\eta \leq 1/2L$, and suppose that Assumption 3 holds. 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 $1 - \delta$:*

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

The proof of Theorem 6 employs a technique used by Harvey et al. (2019). The 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 6, we can again implement a step decay schedule to obtain the following efficiency estimate with high probability; see Theorem 31 for the formal statement.

Theorem 7 (Time to track with high probability, informal) *Suppose that Assumption 3 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 $1 - \delta$ in time

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

4. Tracking the Minimum Value

The results outlined so far have focused on tracking the minimizer $x_t^\star$. In this section, we present results for tracking the minimum 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 minimum value along the 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 6.3, where we apply a standard averaging technique (Appendix A) to a one-step improvement along x_t (Lemma 32) to obtain the desired progress along $\hat{x}_t$ (Proposition 33).

4.1 Bounds in Expectation

We begin with bounding the expected value $\mathbb{E}[\varphi_t(\hat{x}_t) - \varphi_t^\star]$. Proofs appear in Section 6.3. Analogous to Assumption 2, we make the following assumption regarding drift and noise.

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

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

$$\begin{aligned} &\text{Select } g_t = \widetilde{\nabla} f_t(x_t) \\ &\text{Set } x_{t+1} = \operatorname{prox}_{\eta_t r_t}(x_t - \eta_t g_t) \\ &\text{Set } \hat{x}_{t+1} = \left(1 - \tfrac{\mu\eta_t}{2-\mu\eta_t}\right)\hat{x}_t + \tfrac{\mu\eta_t}{2-\mu\eta_t} x_{t+1} \end{aligned}$$

Return $\hat{x}_T$

Assumption 4 (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 two conditions hold for all $0 \le i < t$:

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

(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 1, it is clear that Assumption 4 implies the earlier Assumption 2 with the same constants Δ, σ. The drift assumption intuitively asserts that second moment of $G_{i,t}$ grows at most quadratically in time $|i-t|$. In particular, returning to Example 2, suppose that the distribution map $\mathcal{D}(\cdot)$ is ε-Lipschitz continuous in the Wasserstein-1 distance, the loss $\ell(\cdot, \xi)$ is C^1-smooth for all ξ, and the gradient $\nabla\ell(x, \cdot)$ is β-Lipschitz continuous for all x. Then the Kantorovich-Rubinsteĭn duality theorem (Kantorovich and Rubinshteĭn, 1958) directly implies $\mathbb{E}\, G_{i,t}^2 \le (\varepsilon\beta)^2\, \mathbb{E}\|v_i - v_t\|^2$. Therefore, as long as the second moment $\mathbb{E}\|v_i - v_t\|^2$ scales quadratically in $|i-t|$, the desired drift assumption holds. The assumption on the gradient noise stipulates a uniform bound on the second moment $\mathbb{E}\|z_i\|^2$ and that the condition $\mathbb{E}\langle z_i, x_t^\star\rangle = 0$ holds. 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.

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

Theorem 8 (Expected function gap) *Let $\{\hat{x}_t\}$ be the iterates produced by Algorithm 2 with constant learning rate $\eta \le 1/2L$, and suppose that Assumption 4 holds. Then the following bound holds for all $t \ge 0$:*

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

The "noise + drift" error term in Theorem 8 coincides with μ times the error term in Theorem 3, as expected. With Theorem 8 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

$$\eta_\star = \min\left\{\frac{1}{2L}, \left(\frac{2\Delta^2}{\mu\sigma^2}\right)^{1/3}\right\},$$

and the dichotomy governed by the drift-to-noise ratio Δ/σ remains:

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

In the high drift-to-noise regime $\Delta/\sigma \geq \sqrt{\mu/16L^3}$, Theorem 8 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 8 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 Theorem 35 for the formal statement.

Theorem 9 (Time to track in expectation, informal) *Suppose that Assumption 4 holds. Then there is a learning rate schedule $\{\eta_t\}$ such that Algorithm 2 produces a point $\hat{x}_t$ satisfying*

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

In the low drift-to-noise regime, Theorem 9 improves the constant learning rate efficiency estimate when $\mathcal{G}$ is small, e.g., when $\mathcal{G} \leq (\varphi_0(x_0) - \varphi_0^\star)/e^2$.

4.2 High-Probability Guarantees

Next, we obtain high-probability analogues of Theorems 8 and 9. Proofs appear in Section 6.4. Naturally, such results 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 guarantees under an assumption of sub-Gaussian drift and noise (Assumption 5 below). In particular, we require 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 $0 \leq i < t$; the property $\mathbb{E}[z_i \,|\, \mathcal{F}_{i,t}] = 0$ would follow from independence of the gradient noise z_i and the future minimizer $x_t^\star$ and is very reasonable in light of Examples 1 and 2.

Assumption 5 (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 two conditions hold for all $0 \leq i < t$:

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

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

(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., $\mathbb{E}[z_i \,|\, \mathcal{F}_{i,t}] = 0$ and

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

Clearly the chain of implications holds:

$$\text{Assumption 5} \implies \text{Assumption 4} \implies \text{Assumption 2}.$$

The following theorem shows that if Assumption 5 holds, then the expected bound on $\varphi_t(\hat{x}_t) - \varphi_t^\star$ derived in Theorem 8 holds with high probability.

Theorem 10 (Function gap with high probability) *Let $\{\hat{x}_t\}$ be the iterates produced by Algorithm 2 with constant learning rate $\eta \leq 1/2L$, and suppose that Assumption 5 holds. 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 $1 - \delta$:*

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

The proof of Theorem 10 is based on combining the generalized Freedman inequality of Harvey et al. (2019) 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 instead 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 10 in hand, we may implement a step decay schedule as before to obtain the following efficiency estimate; see Theorem 41 for the formal statement.

Theorem 11 (Time to track with high probability, informal) *Suppose that Assumption 5 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*

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

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

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

5. Extension to the Decision-Dependent Setting

In this section, we extend the framework and results of the previous sections to a much wider class of tracking problems. In particular, the material in this section is a strict generalization of all the results in the previous sections and can model the performative prediction framework of Perdomo et al. (2020) in a time-dependent setting.

Setting the stage, suppose that we have a family of functions $\{f_{t,x}(\cdot)\}$ indexed by time $t \in \mathbb{N}$ and points $x \in \mathbb{R}^d$. Upon replacing the function f_t in the time-dependent problem (5) by the function $f_{t,x}$ depending not only on the time t but also on the decision variable x, we obtain the sequence of decision-dependent stochastic optimization problems

$$\min_{x \in \mathbb{R}^d} \; f_{t,x}(x) + r_t(x) \tag{6}$$

indexed by t. Tracking the solutions to (6) is typically a challenging task due to the dual dependency of $f_{t,x}(x)$ on the decision x. To obtain a more tractable tracking problem, we may decouple this dependency on x by introducing an auxiliary decision variable u and considering the family of stochastic optimization problems

$$\min_{u \in \mathbb{R}^d} \; f_{t,x}(u) + r_t(u) \tag{7}$$

indexed by both t and x. Instead of tracking the *optimal* decisions solving (6), our aim becomes to track the *stable* decisions

$$\bar{x}_t \in \operatorname*{argmin}_{u \in \mathbb{R}^d} \; f_{t,\bar{x}_t}(u) + r_t(u) \tag{8}$$

arising from (7). We call a point $\bar{x}_t$ satisfying (8) an *equilibrium point* of (7) at time t; observe that $\bar{x}_t$ is stable in the sense that it is a fixed point of the map $x \mapsto \operatorname{argmin}_u\{f_{t,x}(u) + r_t(u)\}$.

Reasonable regularity assumptions on the family $\{f_{t,x}(\cdot)\}$ ensure that (7) admits a unique equilibrium point at each time t (see Assumption 6 and Lemma 12 below). When the functions $f_{t,x}$ are independent of x, the equilibrium points are simply the minimizers of $f_t + r_t$—the content of the previous sections. Our goal in this section is to track the equilibrium points $\bar{x}_t$, or equivalently to track the minimizers of the time-dependent stochastic optimization problem

$$\min_{u \in \mathbb{R}^d} \; f_{t,\bar{x}_t}(u) + r_t(u). \tag{9}$$

Formally, (9) is an example of (5), but this viewpoint is not directly useful since $\bar{x}_t$ is unknown. This more general framework allows us to model more dynamic settings. The main example stems from the setting of performative prediction introduced by Perdomo et al. (2020). This will be a running example throughout the section.

Example 3 (Performative prediction) Within the framework of performative prediction, the functions take the form $f_{t,x}(u) = \mathbb{E}_{\xi \sim \mathcal{D}(t,x)} \ell(u, \xi)$ for some family of distributions $\mathcal{D}(t, x)$ indexed by both the time t and the decision variable x. The motivation for the dependence of the distribution on x is that often deployment of a learning rule parametrized by x causes the population to change their profile to increase the likelihood of a better personal outcome—a process called "gaming". In other words, the population data is a function of the decision taken by the learner. Moreover, the dependence of the population data on time appears naturally when the population evolves due to exogenous temporal effects (e.g., seasonal, economic). The equilibrium points $\bar{x}_t$ have a clear meaning in this context. Namely, $\bar{x}_t$ is an equilibrium point if the learner has no reason deviate from the learning rule $\bar{x}_t$ based on the response distribution $\mathcal{D}(t, \bar{x}_t)$ alone.

Whenever we refer back to this example, we will impose the following assumptions that are direct extensions of Perdomo et al. (2020) to the time-dependent setting. Namely, fix a nonempty metric space M equipped with its Borel σ-algebra and let $P_1(M)$ denote the

space of Radon probability measures on M with finite first moment, equipped with the Wasserstein-1 distance W_1. We make the natural assumption that there exist constants $\theta, \varepsilon \geq 0$ such that the distribution map $\mathcal{D}(\cdot, \cdot)$ satisfies the following Lipschitz condition:

$$W_1\big(\mathcal{D}(i,x), \mathcal{D}(t,y)\big) \leq \theta|i-t| + \varepsilon\|x-y\| \quad \text{for all} \quad (i,x),(t,y) \in \mathbb{N} \times \mathbb{R}^d.$$

Moreover, we suppose that the loss function $\ell\colon \mathbb{R}^d \times M \to \mathbb{R}$ has the following three properties: $\ell(u,\cdot) \in L^1(\pi)$ for all $u \in \mathbb{R}^d$ and $\pi \in P_1(M)$; $\ell(\cdot,\xi)$ is C^1-smooth for all $\xi \in M$; and there is a constant $\beta \geq 0$ such that the map $\xi \mapsto \nabla\ell(u,\xi)$ is β-Lipschitz continuous for all $u \in \mathbb{R}^d$, where $\nabla\ell(u,\xi)$ denotes the gradient of $\ell(\cdot,\xi)$ evaluated at u. These assumptions directly imply the following stability property of the gradients with respect to distributional perturbations (see Drusvyatskiy and Xiao, 2022, Lemma 2.1):

$$\sup_{u\in\mathbb{R}^d} \|\nabla f_{i,x}(u) - \nabla f_{t,y}(u)\| \leq \theta\beta|i-t| + \varepsilon\beta\|x-y\| \quad \text{for all} \quad (i,x),(t,y) \in \mathbb{N} \times \mathbb{R}^d. \quad (10)$$

Suppose now that each expected loss $f_{t,x}(\cdot)$ is μ-strongly convex. In this case, Perdomo et al. (2020) identified the optimal parameter regime $\varepsilon\beta < \mu$ wherein the repeated minimization procedure $y_{k+1} = \operatorname{argmin}_u\{f_{t,y_k}(u) + r_t(u)\}$ converges to the unique equilibrium point $\bar{x}_t$ at a linear rate as $k \to \infty$ (see Perdomo et al., 2020, Theorem 3.5 and Proposition 3.6). The gradient stability property (10) will lead us to the corresponding parameter regime in our generalized setting.

5.1 Decision-Dependent Framework

We begin by recording the assumptions of our framework. Similar to the previous sections, we assume that the following standard regularity conditions hold:

(i) Each function $f_{t,x}\colon \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and C^1-smooth with L-Lipschitz continuous gradient for some common parameters $\mu, L > 0$.

(ii) Each regularizer $r_t\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is proper, closed, and convex.

For each $t \in \mathbb{N}$ and $x, u \in \mathbb{R}^d$, we let $\nabla f_{t,x}(u)$ denote the gradient of the function $f_{t,x}(\cdot)$ evaluated at u. In order to control the variation of the family $\{f_{t,x}(\cdot)\}$ in the decision variable x, we introduce the parameter

$$\gamma := \sup_{\substack{t\in\mathbb{N},\, u,x,y\in\mathbb{R}^d \\ x\neq y}} \frac{\|\nabla f_{t,x}(u) - \nabla f_{t,y}(u)\|}{\|x-y\|}$$

and impose throughout Section 5 the following stability property of the gradients with respect to the decision variable.

Assumption 6 (Gradient stability in the decision variable) The following parameter regime holds: $\gamma < \mu$.

In particular, γ is finite and the following Lipschitz bound holds:

$$\sup_{t\in\mathbb{N},\, u\in\mathbb{R}^d} \|\nabla f_{t,x}(u) - \nabla f_{t,y}(u)\| \leq \gamma\|x-y\| \quad \text{for all} \quad x, y \in \mathbb{R}^d.$$

Returning to Example 3, it follows from (10) that Assumption 6 holds whenever $\varepsilon\beta < \mu$. As the following lemma shows, the requirement $\gamma < \mu$ guarantees that for each $t \in \mathbb{N}$, the equilibrium point $\bar{x}_t$ is well defined and unique.

Lemma 12 (Existence of equilibrium) *For each $t \in \mathbb{N}$, the map $S_t \colon \mathbb{R}^d \to \mathbb{R}^d$ given by*

$$S_t(x) = \operatorname*{argmin}_{u \in \mathbb{R}^d} f_{t,x}(u) + r_t(u)$$

is (γ/μ)-contractive and therefore has a unique fixed point $\bar{x}_t$ to which the repeated minimization procedure $y_{k+1} = S_t(y_k)$ converges at a linear rate as $k \to \infty$.

Proof Note first that S_t is well defined by the strong convexity of each function $\varphi_{t,x} := f_{t,x} + r_t$. Next, given $x, y \in \mathbb{R}^d$, observe that we have the first-order optimality conditions $0 \in \partial\varphi_{t,x}(S_t(x))$ and $0 \in \partial\varphi_{t,y}(S_t(y))$; this last inclusion implies $-\nabla f_{t,y}(S_t(y)) \in \partial r_t(S_t(y))$ and hence $\nabla f_{t,x}(S_t(y)) - \nabla f_{t,y}(S_t(y)) \in \partial\varphi_{t,x}(S_t(y))$. On the other hand, the μ-strong convexity of $\varphi_{t,x}$ implies that for all $u, u' \in \operatorname{dom}\varphi_{t,x}$, $w \in \partial\varphi_{t,x}(u)$, and $w' \in \partial\varphi_{t,x}(u')$, we have

$$\mu\|u - u'\| \leq \|w - w'\|.$$

Thus, taking $u = S_t(x)$, $w = 0$, $u' = S_t(y)$, and $w' = \nabla f_{t,x}(S_t(y)) - \nabla f_{t,y}(S_t(y))$ yields

$$\mu\|S_t(x) - S_t(y)\| \leq \|\nabla f_{t,x}(S_t(y)) - \nabla f_{t,y}(S_t(y))\| \leq \gamma\|x - y\|,$$

where the last inequality holds by the definition of γ. Hence $\|S_t(x) - S_t(y)\| \leq (\gamma/\mu)\|x - y\|$, so S_t is (γ/μ)-contractive since $\gamma < \mu$. An application of the Banach fixed-point theorem completes the proof. ∎

It is easy to see that the parameter regime $\gamma < \mu$ is optimal in the sense that equilibrium points can fail to exist whenever $\gamma \geq \mu$, as illustrated in the following example.

Example 4 (Optimality of the regime $\gamma < \mu$) Consider the time-independent family of functions given by $f_x(u) = \frac{1}{2}\|u - ax - b\|^2$ for any fixed constant $a \geq 1$ and vector $b \in \mathbb{R}^d$. Then f_x is 1-strongly convex and smooth with 1-Lipschitz continuous gradient, and

$$\gamma = \sup_{\substack{u,x,y \in \mathbb{R}^d \\ x \neq y}} \frac{\|\nabla f_x(u) - \nabla f_y(u)\|}{\|x - y\|} = a \geq 1 = \mu.$$

Now let $\mathcal{X} \subset \mathbb{R}^d$ be a nonempty closed convex set and take $r = \delta_{\mathcal{X}}$ to be the convex indicator of $\mathcal{X}$. Then $\bar{x}$ is an equilibrium point of the decoupled family of problems $\min_u\{f_x(u) + r(u)\}$ if and only if

$$\bar{x} \in \operatorname*{argmin}_{u \in \mathbb{R}^d}\{f_{\bar{x}}(u) + r(u)\} = \{\operatorname{proj}_{\mathcal{X}}(a\bar{x} + b)\}.$$

Taking $a = 1$, $b \neq 0$, and $\mathcal{X} = \mathbb{R}^d$, we see that $\gamma = \mu$ and no equilibrium point exists. On the other hand, if we take $\mathcal{X} = \mathbb{R}^d_+$ to be the nonnegative orthant and $b > 0$, then for any $a \geq 1$ and $x \in \mathcal{X}$ we have

$$x < ax + b = \operatorname{proj}_{\mathcal{X}}(ax + b)$$

and hence no equilibrium point exists for this problem with any value $\gamma \in [\mu, \infty)$.

Next, we turn to tracking the equilibria $\bar{x}_t$ furnished by Lemma 12 using a decision-dependent proximal stochastic gradient method. Specifically, we make the standing assumption that at every time t, and at every query point x, the learner may obtain an *unbiased estimator* $\widetilde{\nabla} f_{t,x}(x)$ of the true gradient $\nabla f_{t,x}(x)$. With this oracle access, the decision-dependent proximal stochastic gradient method—recorded as Algorithm 3 below—selects in

each iteration t the stochastic gradient $g_t = \widetilde{\nabla} f_{t,x_t}(x_t)$ and takes the step

$$x_{t+1} := \operatorname{prox}_{\eta_t r_t}(x_t - \eta_t g_t) = \operatorname*{argmin}_{u \in \mathbb{R}^d} \left\{ r_t(u) + \tfrac{1}{2\eta_t}\|u - (x_t - \eta_t g_t)\|^2 \right\}$$

using step size $\eta_t > 0$. As before, our goal is to obtain efficiency estimates for this procedure that hold both in expectation and with high probability.

Algorithm 3 Decision-Dependent PSG $\qquad$ D-PSG$(x_0, \{\eta_t\}, T)$

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

$$\text{Select } g_t = \widetilde{\nabla} f_{t,x_t}(x_t)$$
$$\text{Set } x_{t+1} = \operatorname{prox}_{\eta_t r_t}(x_t - \eta_t g_t)$$

Return x_T

The guarantees we obtain allow both the iterates x_t *and the equilibria* $\bar{x}_t$ to evolve stochastically. Given $\{x_t\}$ and $\{g_t\}$ as in Algorithm 3, we let

$$z_t := \nabla f_{t,x_t}(x_t) - g_t$$

denote the *gradient noise* at time t and we impose the following assumption modeling stochasticity on a fixed probability space $(\Omega, \mathcal{F}, \mathbb{P})$ throughout Section 5.

Assumption 7 (Stochastic framework) There exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with filtration $(\mathcal{F}_t)_{t \geq 0}$ such that $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and the following two conditions hold for all $t \geq 0$:

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

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

The first item of Assumption 7 formalizes the assertion that x_t and $\bar{x}_t$ are fully determined by information up to time t. The second item of Assumption 7 formalizes the assertion 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, i.e., g_t is an unbiased estimator of $\nabla f_{t,x_t}(x_t)$; for example, this holds naturally in Example 3 if we take $g_t = \nabla \ell(x_t, \xi_t)$ with $\xi_t \sim \mathcal{D}(t, x_t)$.

Finally, we fix some notation to be used henceforth. We define the positive parameter

$$\bar{\mu} := \mu - \gamma,$$

and we define the *equilibrium drift* $\bar{\Delta}_t$ and the *temporal gradient drift* $\bar{G}_{i,t}$ to be the random variables

$$\bar{\Delta}_t := \|\bar{x}_t - \bar{x}_{t+1}\| \quad \text{and} \quad \bar{G}_{i,t} := \sup_{u,x \in \mathbb{R}^d} \|\nabla f_{i,x}(u) - \nabla f_{t,x}(u)\|.$$

Note that in the setting of Example 3, the estimate (10) implies $\bar{G}_{i,t} \leq \theta\beta|i-t|$ and hence $\bar{\Delta}_t \leq \theta\beta/\bar{\mu}$ by Lemma 1, provided the regularizers $r_t \equiv r$ are identical for all times t. We also set

$$\varphi_t := f_{t,x_t} + r_t, \quad x_t^\star := \operatorname{argmin} \varphi_t, \quad \varphi_t^\star := \min \varphi_t$$

and

$$\psi_t := f_{t,\bar{x}_t} + r_t \quad \text{and} \quad \psi_t^\star := \min \psi_t.$$

In particular, the equilibrium point $\bar{x}_t$ is the minimizer of the *equilibrium function* ψ_t, and $\psi_t^\star$ denotes its minimum value. Observe that when $\gamma = 0$, we have $\varphi_t = \psi_t + c_t$ for some constant of integration c_t and hence we recover the setting of Section 2 with $x_t^\star = \bar{x}_t$.

5.2 Tracking the Equilibrium Point

In a nutshell, the results of Section 3 extend directly to tracking the equilibrium points $\bar{x}_t$, with μ replaced by $\bar{\mu}$ and Δ replaced by $\bar{\Delta}$ (defined in Assumption 8 below). We begin with bounding the expected value $\mathbb{E}\|x_t - \bar{x}_t\|^2$. Due to the fact that Algorithm 3 takes steps on the current functions φ_t but the minimizers we aim to track are those of the equilibrium functions ψ_t, we will rely at the outset on controlling the function gaps

$$\big[f_{i,x}(u) - f_{i,x}(v)\big] - \big[f_{t,y}(u) - f_{t,y}(v)\big]$$

(at first for $i = t$, and later for general i and t). We achieve this control in terms of the temporal gradient drift.

Lemma 13 (Function gap variation) *For all $(i,x),(t,y) \in \mathbb{N} \times \mathbb{R}^d$ and $u, v \in \mathbb{R}^d$, we have*

$$\big|\big[f_{i,x}(u) - f_{i,x}(v)\big] - \big[f_{t,y}(u) - f_{t,y}(v)\big]\big| \le \big(\bar{G}_{i,t} + \gamma\|x - y\|\big)\|u - v\|.$$

Proof Fix $(i,x),(t,y) \in \mathbb{N} \times \mathbb{R}^d$ and $u, v \in \mathbb{R}^d$, and set $u_\tau := v + \tau(u - v)$ for all $\tau \in [0,1]$. By the fundamental theorem of calculus and Cauchy-Schwarz, we have

$$\begin{aligned}\big[f_{i,x}(u) - f_{i,x}(v)\big] - \big[f_{t,y}(u) - f_{t,y}(v)\big] &= \int_0^1 \big\langle \nabla f_{i,x}(u_\tau) - \nabla f_{t,y}(u_\tau), u - v\big\rangle \, d\tau \\ &\le \big(\bar{G}_{i,t} + \gamma\|x - y\|\big)\|u - v\|.\end{aligned}$$

Switching (i,x) and (t,y) completes the proof. ∎

Using Lemmas 2 and 13, we obtain the following equilibrium one-step improvement.

Lemma 14 (Equilibrium one-step improvement) *The iterates $\{x_t\}$ produced by Algorithm 3 with $\eta_t < 1/L$ satisfy the bound:*

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

Proof By Lemma 13, we have

$$\begin{aligned}\big[\psi_t(x_{t+1}) - \psi_t(\bar{x}_t)\big] - \big[\varphi_t(x_{t+1}) - \varphi_t(\bar{x}_t)\big] \\ &= \big[f_{t,\bar{x}_t}(x_{t+1}) - f_{t,\bar{x}_t}(\bar{x}_t)\big] - \big[f_{t,x_t}(x_{t+1}) - f_{t,x_t}(\bar{x}_t)\big] \\ &\le \gamma\|x_t - \bar{x}_t\|\|x_{t+1} - \bar{x}_t\|.\end{aligned}$$

Hence

$$\psi_t(x_{t+1}) - \psi_t^\star \le \varphi_t(x_{t+1}) - \varphi_t(\bar{x}_t) + \gamma\|x_t - \bar{x}_t\|\|x_{t+1} - \bar{x}_t\|.$$

Moreover, Young's inequality implies

$$\gamma\|x_t - \bar{x}_t\|\|x_{t+1} - \bar{x}_t\| \le \tfrac{\gamma}{2}\|x_t - \bar{x}_t\|^2 + \tfrac{\gamma}{2}\|x_{t+1} - \bar{x}_t\|^2.$$

Multiplying through by $2\eta_t$ and applying Lemma 2 completes the proof. ∎

For simplicity, we state the main results under the assumption that the second moments $\mathbb{E}\,\bar{\Delta}_t^2$ 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 14 as well.

Assumption 8 (Bounded second moments) There exist constants $\bar{\Delta}, \sigma > 0$ such that the following two conditions hold for all $t \geq 0$:

(i) **(Drift)** The equilibrium drift $\bar{\Delta}_t$ satisfies $\mathbb{E}\,\bar{\Delta}_t^2 \leq \bar{\Delta}^2$.

(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 3, thereby extending Theorem 3; see Section 6.1 for the precise statement (Corollary 27) and proof.

Theorem 15 (Expected distance) *Suppose that Assumption 8 holds. Then the iterates produced by Algorithm 3 with constant learning rate $\eta \leq 1/2L$ satisfy the bound:*

$$\mathbb{E}\|x_t - \bar{x}_t\|^2 \lesssim \underbrace{(1-\bar{\mu}\eta)^t \|x_0 - \bar{x}_0\|^2}_{\textit{optimization}} + \underbrace{\frac{\eta\sigma^2}{\bar{\mu}}}_{\textit{noise}} + \underbrace{\left(\frac{\bar{\Delta}}{\bar{\mu}\eta}\right)^2}_{\textit{drift}}.$$

With Theorem 15 in hand, we are led to define the following asymptotic tracking error of Algorithm 3 corresponding to $\mathbb{E}\|x_t - \bar{x}_t\|^2$, together with the corresponding optimal step size:

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

Plugging $\bar{\eta}_\star$ into the definition of $\bar{\mathcal{E}}$, we see that Algorithm 3 exhibits qualitatively different behaviors in settings corresponding to high or low drift-to-noise ratio $\bar{\Delta}/\sigma$:

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

As before, the *high drift-to-noise regime* $\bar{\Delta}/\sigma \geq \sqrt{\bar{\mu}/16L^3}$ is uninteresting from the viewpoint of stochastic optimization and we focus on the *low drift-to-noise regime* $\bar{\Delta}/\sigma < \sqrt{\bar{\mu}/16L^3}$. The following theorem extends Theorem 4; see Theorem 28 for the formal statement and proof.

Theorem 16 (Time to track in expectation, informal) *Suppose that Assumption 8 holds. Then there is a learning rate schedule $\{\eta_t\}$ such that Algorithm 3 produces a point x_t satisfying*

$$\mathbb{E}\|x_t - \bar{x}_t\|^2 \lesssim \bar{\mathcal{E}} \quad \textit{in time} \quad t \lesssim \frac{L}{\bar{\mu}} \log\left(\frac{\|x_0 - \bar{x}_0\|^2}{\bar{\mathcal{E}}}\right) + \frac{\sigma^2}{\bar{\mu}^2\bar{\mathcal{E}}}.$$

Next, we present high-probability guarantees for the tracking error $\|x_t - \bar{x}_t\|^2$ under the following standard light-tail assumption on the equilibrium drift and gradient noise.

Assumption 9 (Sub-Gaussian drift and noise) There exist constants $\bar{\Delta}, \sigma > 0$ such that the following two conditions hold for all $t \geq 0$:

(i) **(Drift)** The drift $\bar{\Delta}_t^2$ is sub-exponential conditioned on $\mathcal{F}_t$ with parameter $\bar{\Delta}^2$:
$$\mathbb{E}\big[\exp(\lambda \bar{\Delta}_t^2) \,|\, \mathcal{F}_t\big] \leq \exp(\lambda \bar{\Delta}^2) \quad \text{for all} \quad 0 \leq \lambda \leq \bar{\Delta}^{-2}.$$

(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 9 is equivalent to asserting that the equilibrium drift $\bar{\Delta}_t$ is sub-Gaussian conditioned on $\mathcal{F}_t$, and that this condition holds trivially in the setting of Example 3 with $\bar{\Delta} = \theta\beta/\bar{\mu}$ provided the regularizers $r_t \equiv r$ are identical for all times t. Clearly Assumption 9 implies Assumption 8 with the same constants $\bar{\Delta}, \sigma$. The following theorem shows that if Assumption 9 holds, then the expected bound on $\|x_t - \bar{x}_t\|^2$ derived in Theorem 15 holds with high probability.

Theorem 17 (High-probability distance tracking) *Let $\{x_t\}$ be the iterates produced by Algorithm 3 with constant learning rate $\eta \leq 1/2L$, and suppose that Assumption 9 holds. 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 $1 - \delta$:*
$$\|x_t - \bar{x}_t\|^2 \leq \big(1 - \tfrac{\bar{\mu}\eta}{2}\big)^t \|x_0 - \bar{x}_0\|^2 + c\left(\frac{\eta\sigma^2}{\bar{\mu}} + \left(\frac{\bar{\Delta}}{\bar{\mu}\eta}\right)^2\right)\log\Big(\frac{e}{\delta}\Big). \tag{11}$$

Theorem 17 is an extension of Theorem 6. As a consequence of Theorem 17, we can again implement a step decay schedule in the low drift-to-noise regime to obtain the following efficiency estimate with high probability, thereby extending Theorem 7; see Section 6.2 for the precise statements (Theorems 30 and 31) and proofs.

Theorem 18 (Time to track with high probability, informal) *Suppose that Assumption 9 holds and that we are in the low drift-to-noise regime $\bar{\Delta}/\sigma < \sqrt{\bar{\mu}/16L^3}$. Then there is a learning rate schedule $\{\eta_t\}$ such that for any specified $\delta \in (0,1)$, Algorithm 3 produces a point x_t satisfying*
$$\|x_t - \bar{x}_t\|^2 \lesssim \bar{\mathcal{E}}\log\Big(\frac{e}{\delta}\Big)$$
with probability at least $1 - \delta$ in time
$$t \lesssim \frac{L}{\bar{\mu}}\log\left(\frac{\|x_0 - \bar{x}_0\|^2}{\bar{\mathcal{E}}}\right) + \frac{\sigma^2}{\bar{\mu}^2\bar{\mathcal{E}}}.$$

5.3 Tracking the Equilibrium Value

The results outlined so far have focused on tracking the equilibrium point $\bar{x}_t$, i.e., the minimizer of ψ_t. In this section, we present results for tracking the equilibrium value $\psi_t^\star$ in the parameter regime
$$\gamma < \mu/2. \tag{12}$$

The regime (12) matches the one used in Theorem 7.3 of Drusvyatskiy and Xiao (2022) to obtain function gap bounds for biased PSG along an average iterate, and we employ a similar averaging technique to obtain our bounds.

Imposing the regime (12), we define the positive parameter

$$\hat{\mu} := \mu - 2\gamma.$$

Our strategy is to track the equilibrium value $\psi_t^\star$ along the running average $\hat{x}_t$ of the iterates x_t produced by Algorithm 3, as defined in Algorithm 4 below. In a nutshell, the results of Section 4 extend directly to tracking the equilibrium value $\psi_t^\star$, with μ replaced by $\hat{\mu}$ and Δ replaced by $\bar{\Delta}$ (defined in Assumption 10 below).

Algorithm 4 Averaged Decision-Dependent PSG $\qquad$ $\texttt{D-}\overline{\texttt{PSG}}(x_0, \mu, \gamma, \{\eta_t\}, T)$

Input: initial $x_0 = \hat{x}_0$, strong convexity parameter μ, gradient drift parameter $\gamma \in [0, \mu/2)$, and step sizes $\{\eta_t\}_{t=0}^{T-1} \subset (0, 1/\bar{\mu})$, where $\bar{\mu} = \mu - \gamma$; set $\hat{\mu} = \mu - 2\gamma$
Step $t = 0, \dots, T-1$:

$$\begin{aligned}
&\text{Select } g_t = \widetilde{\nabla} f_{t,x_t}(x_t) \\
&\text{Set } x_{t+1} = \operatorname{prox}_{\eta_t r_t}(x_t - \eta_t g_t) \\
&\text{Set } \hat{x}_{t+1} = \Big(1 - \tfrac{\hat{\mu}\eta_t}{2-\mu\eta_t}\Big)\hat{x}_t + \tfrac{\hat{\mu}\eta_t}{2-\mu\eta_t} x_{t+1}
\end{aligned}$$

Return $\hat{x}_T$

We begin with bounding the expected value $\mathbb{E}[\psi_t(\hat{x}_t) - \psi_t^\star]$. This requires a weak form of uncorrelatedness between the gradient noise z_i and the future equilibrium point $\bar{x}_t$, which we stipulate in the following analogue of Assumption 4.

Assumption 10 (Bounded second moments) The regularizers $r_t \equiv r$ are identical for all times t and there exist constants $\bar{\Delta}, \sigma > 0$ such that the following two conditions hold for all $0 \le i < t$:

(i) **(Drift)** The temporal gradient drift $\bar{G}_{i,t}$ satisfies $\mathbb{E}\,\bar{G}_{i,t}^2 \le (\hat{\mu}\bar{\Delta}|i-t|)^2$.

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

Taking into account Lemma 1, it is clear that Assumption 10 implies the earlier Assumption 8 with the same constants $\bar{\Delta}, \sigma$. Further, the condition on the drift holds trivially in the setting of Example 3 with $\bar{\Delta} = \theta\beta/\hat{\mu}$ provided $\mu > 2\varepsilon\beta$. The following theorem presents an expected improvement guarantee for Algorithm 4, thereby extending Theorem 8; see Corollary 34 for the precise statement and proof.

Theorem 19 (Expected function gap) *Let $\{\hat{x}_t\}$ be the iterates produced by Algorithm 4 with constant learning rate $\eta \le 1/2L$, and suppose that Assumption 10 holds. Then the following bound holds for all $t \ge 0$:*

$$\mathbb{E}\big[\psi_t(\hat{x}_t) - \psi_t^\star\big] \lesssim \underbrace{\big(1 - \tfrac{\hat{\mu}\eta}{2}\big)^t \big(\psi_0(x_0) - \psi_0^\star\big)}_{\textit{optimization}} + \underbrace{\eta\sigma^2}_{\textit{noise}} + \underbrace{\frac{\bar{\Delta}^2}{\hat{\mu}\eta^2}}_{\textit{drift}}.$$

With Theorem 19 in hand, we are led to define the following asymptotic tracking error of Algorithm 4 corresponding to $\mathbb{E}[\psi_t(\hat{x}_t) - \psi_t^\star]$, together with the corresponding optimal step size:

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

A familiar dichotomy governed by the drift-to-noise ratio $\bar{\Delta}/\sigma$ arises. We again focus on the *low drift-to-noise regime* $\bar{\Delta}/\sigma < \sqrt{\hat{\mu}/16L^3}$. The following theorem extends Theorem 9; see Theorem 35 for the formal statement.

Theorem 20 (Time to track in expectation, informal) *Suppose that Assumption 10 holds. Then there is a learning rate schedule $\{\eta_t\}$ such that Algorithm 4 produces a point $\hat{x}_t$ satisfying*

$$\mathbb{E}[\psi_t(\hat{x}_t) - \psi_t^\star] \lesssim \widehat{\mathcal{G}} \quad \textit{in time} \quad t \lesssim \frac{L}{\hat{\mu}} \log\left(\frac{\psi_0(x_0) - \psi_0^\star}{\widehat{\mathcal{G}}} \right) + \frac{\sigma^2}{\hat{\mu}\widehat{\mathcal{G}}}.$$

Next, we obtain high-probability analogues of Theorems 19 and 20. Naturally, such results should rely on light-tail assumptions on the temporal gradient drift $\bar{G}_{i,t}$ and the norm of the gradient noise $\|z_i\|$. We state the guarantees under an assumption of sub-Gaussian drift and noise (Assumption 11 below). In particular, we require that the gradient noise z_i is mean-zero conditioned on the σ-algebra

$$\mathcal{F}_{i,t} := \sigma(\mathcal{F}_i, \bar{x}_t)$$

for all $0 \leq i < t$; the property $\mathbb{E}[z_i \,|\, \mathcal{F}_{i,t}] = 0$ would follow from independence of the gradient noise z_i and the future equilibrium point $\bar{x}_t$.

Assumption 11 (Sub-Gaussian drift and noise) The regularizers $r_t \equiv r$ are identical for all times t and there exist constants $\bar{\Delta}, \sigma > 0$ such that the following two conditions hold for all $0 \leq i < t$:

(i) **(Drift)** The drift $\bar{G}_{i,t}^2$ is sub-exponential with parameter $(\hat{\mu}\bar{\Delta}|i-t|)^2$:

$$\mathbb{E}\big[\exp\big(\lambda \bar{G}_{i,t}^2\big)\big] \leq \exp\big(\lambda(\hat{\mu}\bar{\Delta}|i-t|)^2\big) \quad \text{for all} \quad 0 \leq \lambda \leq (\hat{\mu}\bar{\Delta}|i-t|)^{-2}.$$

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

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

Clearly the chain of implications

$$\text{Assumption 11} \implies \text{Assumption 10} \implies \text{Assumption 8}$$

holds, and the condition on the drift in Assumption 11 holds trivially in the setting of Example 3 with $\bar{\Delta} = \theta\beta/\hat{\mu}$ provided $\mu > 2\varepsilon\beta$. The following theorem shows that if Assumption 11 holds, then the expected bound on $\psi_t(\hat{x}_t) - \psi_t^\star$ derived in Theorem 19 holds with high probability, thereby extending Theorem 10.

Theorem 21 (Function gap with high probability) *Let $\{\hat{x}_t\}$ be the iterates produced by Algorithm 4 with constant learning rate $\eta \le 1/2L$, and suppose that Assumption 11 holds. 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 $1-\delta$:*

$$\psi_t(\hat{x}_t) - \psi_t^\star \le c\left(\left(1 - \tfrac{\hat{\mu}\eta}{2}\right)^t\left(\psi_0(x_0) - \psi_0^\star\right) + \eta\sigma^2 + \frac{\bar{\Delta}^2}{\hat{\mu}\eta^2}\right)\log\Big(\frac{e}{\delta}\Big). \tag{13}$$

With Theorem 21 in hand, we may implement a step decay schedule as before to obtain the following efficiency estimate, thereby extending Theorem 11; see Section 6.4 for the precise statements (Theorems 39 and 41) and proofs.

Theorem 22 (Time to track with high probability, informal) *Suppose that Assumption 11 holds and that we are in the low drift-to-noise regime $\bar{\Delta}/\sigma < \sqrt{\hat{\mu}/16L^3}$. Fix $\delta \in (0,1)$. Then there is a learning rate schedule $\{\eta_t\}$ such that Algorithm 4 produces a point $\hat{x}_t$ satisfying*

$$\psi_t(\hat{x}_t) - \psi_t^\star \lesssim \widehat{\mathcal{G}}\log\Big(\frac{e}{\delta}\Big)$$

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

$$t \lesssim \frac{L}{\hat{\mu}}\log\left(\frac{\psi_0(x_0) - \psi_0^\star}{\widehat{\mathcal{G}}}\right) + \frac{\sigma^2}{\hat{\mu}\widehat{\mathcal{G}}}\log\Big(\log\Big(\frac{e}{\delta}\Big)\Big), \quad \textit{where} \quad K \lesssim \log_2\left(\frac{1}{L}\cdot\left(\frac{\sigma^2\hat{\mu}}{\bar{\Delta}^2}\right)^{1/3}\right).$$

6. Proofs of Main Results

Roadmap. In this section, we derive the results of the preceding sections under the unified framework presented in Section 5.1; we impose the assumptions and notation of Section 5.1 henceforth. Sections 6.1 and 6.2 handle distance tracking in expectation and with high probability, respectively; this corresponds to the results presented in Section 5.2 (entailing those of Sections 3.1 and 3.2). Then Sections 6.3 and 6.4 handle function gap tracking in expectation and with high probability, respectively; this corresponds to the results presented in Section 5.3 (entailing those of Sections 4.1 and 4.2).

6.1 Tracking the Equilibrium Point: Bounds in Expectation

The proof of Theorem 15 follows a familiar pattern in stochastic optimization. We begin by recalling Lemma 2, which gives a standard one-step improvement guarantee for the proximal stochastic gradient method on the fixed problem $\min \varphi_t$.

Lemma 23 (One-step improvement) *For all $x \in \mathbb{R}^d$, the iterates $\{x_t\}$ produced by Algorithm 3 with $\eta_t < 1/L$ satisfy the bound:*

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

Proof Since $f_t := f_{t,x_t}$ 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 $\delta_t > 0$, Cauchy-Schwarz and Young's inequality yield

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

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

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

where the last inequality holds because $x_{t+1} = \operatorname{prox}_{\eta_t r_t}(x_t - \eta_t g_t)$ 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

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

using the μ-strong convexity of f_t. Thus,

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

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

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

as claimed. ■

It is critically important that the one-step improvement estimate in Lemma 23 holds with respect to any reference point x. In particular, as we already showed in Section 5.2, taking $x = \bar{x}_t$ and applying Lemma 13 yields Lemma 14:

Lemma 24 (Equilibrium one-step improvement) *The iterates $\{x_t\}$ produced by Algorithm 3 with $\eta_t < 1/L$ satisfy the bound:*

$$\begin{aligned}
2\eta_t(\psi_t(x_{t+1}) - \psi_t^\star) \leq (1 - \bar{\mu}\eta_t)\|x_t - \bar{x}_t\|^2 - (1 - \gamma\eta_t)\|x_{t+1} - \bar{x}_t\|^2 \\
+ 2\eta_t\langle z_t, x_t - \bar{x}_t \rangle + \tfrac{\eta_t^2}{1-L\eta_t}\|z_t\|^2.
\end{aligned}$$

With Lemma 24 in hand, we obtain the following recursion on $\|x_t - \bar{x}_t\|^2$.

Lemma 25 (Distance recursion) *The iterates $\{x_t\}$ produced by Algorithm 3 with step size $\eta_t < 1/L$ satisfy the bound:*

$$\|x_{t+1} - \bar{x}_{t+1}\|^2 \leq (1 - \bar{\mu}\eta_t)\|x_t - \bar{x}_t\|^2 + 2\eta_t\langle z_t, x_t - \bar{x}_t \rangle + \tfrac{\eta_t^2}{1-L\eta_t}\|z_t\|^2 + \left(1 + \tfrac{1}{\bar{\mu}\eta_t}\right)\bar{\Delta}_t^2.$$

Proof First, note

$$\begin{aligned}\|x_{t+1} - \bar{x}_{t+1}\|^2 &= \|x_{t+1} - \bar{x}_t\|^2 + \|\bar{x}_t - \bar{x}_{t+1}\|^2 + 2\langle x_{t+1} - \bar{x}_t, \bar{x}_t - \bar{x}_{t+1}\rangle \\ &\leq (1 + \bar{\mu}\eta_t)\|x_{t+1} - \bar{x}_t\|^2 + \left(1 + \tfrac{1}{\bar{\mu}\eta_t}\right)\|\bar{x}_t - \bar{x}_{t+1}\|^2\end{aligned}$$

by Cauchy-Schwarz and Young's inequality. Further, the μ-strong convexity of ψ_t implies $\frac{\mu}{2}\|x_{t+1} - \bar{x}_t\|^2 \leq \psi_t(x_{t+1}) - \psi_t^\star$, which together with Lemma 24 implies

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

The result follows. ■

Applying Lemma 25 recursively furnishes a bound on $\|x_t - \bar{x}_t\|^2$. When the step size is constant, the next proposition follows immediately.

Proposition 26 (Last-iterate progress) *The iterates $\{x_t\}$ produced by Algorithm 3 with constant step size $\eta < 1/L$ satisfy the bound:*

$$\begin{aligned}\|x_t - \bar{x}_t\|^2 \leq (1 - \bar{\mu}\eta)^t\|x_0 - \bar{x}_0\|^2 + 2\eta\sum_{i=0}^{t-1}\langle z_i, x_i - \bar{x}_i\rangle(1 - \bar{\mu}\eta)^{t-1-i} \\ + \tfrac{\eta^2}{1 - L\eta}\sum_{i=0}^{t-1}\|z_i\|^2(1 - \bar{\mu}\eta)^{t-1-i} + \left(1 + \tfrac{1}{\bar{\mu}\eta}\right)\sum_{i=0}^{t-1}\bar{\Delta}_i^2(1 - \bar{\mu}\eta)^{t-1-i}.\end{aligned}$$

By taking expectations in Proposition 26, we obtain the following precise version of Theorem 15.

Corollary 27 (Expected distance) *Suppose that Assumption 8 holds. Then the iterates $\{x_t\}$ generated by Algorithm 3 with constant learning rate $\eta \leq 1/2L$ satisfy the bound:*

$$\mathbb{E}\|x_t - \bar{x}_t\|^2 \leq (1 - \bar{\mu}\eta)^t\|x_0 - \bar{x}_0\|^2 + 2\left(\frac{\eta\sigma^2}{\bar{\mu}} + \left(\frac{\bar{\Delta}}{\bar{\mu}\eta}\right)^2\right).$$

With Corollary 27 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 the asymptotically optimal learning rate $\bar{\eta}_\star$. The following theorem provides a formal version of Theorem 16 (note that in the high drift-to-noise regime $\bar{\Delta}/\sigma \geq \sqrt{\bar{\mu}/16L^3}$, Theorem 16 holds trivially with the constant learning rate $\bar{\eta}_\star = 1/2L$). The argument is close in spirit to the justifications of the restart schemes used by Ghadimi and Lan (2013).

Theorem 28 (Time to track in expectation) *Suppose that Assumption 8 holds and that we are in the low drift-to-noise regime $\bar{\Delta}/\sigma < \sqrt{\bar{\mu}/16L^3}$. Set $\bar{\eta}_\star = (2\bar{\Delta}^2/\bar{\mu}\sigma^2)^{1/3}$ and $\bar{\mathcal{E}} = (\bar{\Delta}\sigma^2/\bar{\mu}^2)^{2/3}$. Suppose moreover that we have available a positive upper bound on the initial squared distance $D \geq \|x_0 - \bar{x}_0\|^2$. Consider running Algorithm 3 in $k = 0, \ldots, K-1$ epochs, namely, set $X_0 = x_0$ and iterate the process*

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

where the number of epochs is

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

and we set[5]

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

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

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

and the corresponding tracking error satisfies $\mathbb{E}\|X_K - \bar{X}_K\|^2 \lesssim \bar{\mathcal{E}}$, where $\bar{X}_K$ denotes the minimizer of ψ_T.

Proof For each index k, let $t_k := T_0 + \cdots + T_{k-1}$ (with $t_0 := 0$), $\bar{X}_k$ be the minimizer of the corresponding equilibrium function ψ_{t_k}, and

$$\bar{E}_k := \frac{2}{\bar{\mu}} \left(\eta_k \sigma^2 + \frac{\bar{\Delta}^2}{\bar{\mu} \bar{\eta}_\star^2}\right).$$

Then taking into account $\eta_k \geq \bar{\eta}_\star$, Corollary 27 directly implies

$$\begin{aligned}
\mathbb{E}\|X_{k+1} - \bar{X}_{k+1}\|^2 &\leq (1 - \bar{\mu}\eta_k)^{T_k} \mathbb{E}\|X_k - \bar{X}_k\|^2 + \frac{2}{\bar{\mu}}\left(\eta_k \sigma^2 + \frac{\bar{\Delta}^2}{\bar{\mu}\eta_k^2}\right) \\
&\leq e^{-\bar{\mu}\eta_k T_k} \mathbb{E}\|X_k - \bar{X}_k\|^2 + \bar{E}_k.
\end{aligned}$$

We will verify by induction that the estimate $\mathbb{E}\|X_k - \bar{X}_k\|^2 \leq 2\bar{E}_{k-1}$ holds for all indices $k \geq 1$. To see the base case, observe

$$\mathbb{E}\|X_1 - \bar{X}_1\|^2 \leq e^{-\bar{\mu}\eta_0 T_0} \|X_0 - \bar{X}_0\|^2 + \bar{E}_0 \leq 2\bar{E}_0.$$

Now assume that the claim holds for some index $k \geq 1$. We then conclude

$$\begin{aligned}
\mathbb{E}\|X_{k+1} - \bar{X}_{k+1}\|^2 &\leq e^{-\bar{\mu}\eta_k T_k} \mathbb{E}\|X_k - \bar{X}_k\|^2 + \bar{E}_k \\
&\leq \frac{1}{4} \mathbb{E}\|X_k - \bar{X}_k\|^2 + \bar{E}_k \\
&\leq \frac{\bar{E}_k}{2\bar{E}_{k-1}} \mathbb{E}\|X_k - \bar{X}_k\|^2 + \bar{E}_k \leq 2\bar{E}_k,
\end{aligned}$$

thereby completing the induction. Hence $\mathbb{E}\|X_K - \bar{X}_K\|^2 \leq 2\bar{E}_{K-1}$.

Next, observe

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

so

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

5. We use here the notation $a^+ = a \vee 0 = \max\{a, 0\}$ to denote the positive part of a real number a; note that for small D, the logarithms $\log(\bar{\mu} L D/\sigma^2)$ and $\log(D/\bar{\mathcal{E}})$ may be negative.

Finally, note

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

and

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

This completes the proof. ■

6.2 Tracking the Equilibrium Point: High-Probability Guarantees

The proof of Theorem 17 is based on recursively controlling the moment generating function of $\|x_t - \bar{x}_t\|^2$. Namely, Lemma 25 in the regime $\eta_t \leq 1/2L$ directly yields

$$\|x_{t+1} - \bar{x}_{t+1}\|^2 \leq (1 - \bar{\mu}\eta_t)\|x_t - \bar{x}_t\|^2 + 2\eta_t \langle z_t, v_t \rangle \|x_t - \bar{x}_t\| + 2\eta_t^2 \|z_t\|^2 + \tfrac{2}{\bar{\mu}\eta_t} \bar{\Delta}_t^2, \qquad (14)$$

where we set

$$v_t := \begin{cases} \frac{x_t - \bar{x}_t}{\|x_t - \bar{x}_t\|} & \text{if } x_t \neq \bar{x}_t \\ 0 & \text{otherwise.} \end{cases}$$

The goal is now to control the moment generating function $\mathbb{E}\big[e^{\lambda \|x_t - \bar{x}_t\|^2}\big]$ through the recursive inequality (14). The basic probabilistic tool to achieve this in similar settings under bounded noise assumptions was developed by Harvey et al. (2019); the following proposition is a slight generalization of Claim D.1 of Harvey et al. (2019) to the light-tail setting we require.

Proposition 29 (Recursive control on MGF) *Consider scalar stochastic processes (V_t), (D_t), and (X_t) on a probability space with filtration $(\mathcal{H}_t)$ such that V_t is nonnegative and $\mathcal{H}_t$-measurable and the inequality*

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

holds for some deterministic constants $\alpha_t \in (-\infty, 1]$ and $\kappa_t \in \mathbb{R}$. Suppose that the moment generating functions of D_t and X_t conditioned on $\mathcal{H}_t$ satisfy the following inequalities for some deterministic constants $\sigma_t, \nu_t > 0$:

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

Then the inequality

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

holds for all $0 \leq \lambda \leq \min\big\{\frac{1-\alpha_t}{2\sigma_t^2}, \frac{1}{2\nu_t}\big\}$.

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

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

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

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

The proof is complete. ■

We may now use Proposition 29 to derive the following precise version of Theorem 17.

Theorem 30 (High-probability distance tracking) *Let $\{x_t\}$ be the iterates produced by Algorithm 3 with constant learning rate $\eta \leq 1/2L$, and suppose that Assumption 9 holds. Then there exists an absolute constant*[6] *$c > 0$ such that for any specified $t \in \mathbb{N}$ and $\delta \in (0,1)$, the following estimate holds with probability at least $1-\delta$:*

$$\|x_t - \bar{x}_t\|^2 \leq \big(1 - \tfrac{\bar{\mu}\eta}{2}\big)^t\|x_0 - \bar{x}_0\|^2 + \left(\frac{8\eta(c\sigma)^2}{\bar{\mu}} + 4\left(\frac{\bar{\Delta}}{\bar{\mu}\eta}\right)^2\right)\log\Big(\frac{e}{\delta}\Big).$$

Proof Note first that under Assumption 9, 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 (see Jin et al., 2019, Lemma 3). Therefore $\langle z_t, v_t\rangle$ is mean-zero sub-Gaussian conditioned on $\mathcal{F}_t$ with parameter $c\sigma$, while $\bar{\Delta}_t^2$ is sub-exponential conditioned on $\mathcal{F}_t$ with parameter $\bar{\Delta}^2$ by Assumption 9. Thus, in light of inequality (14), we may apply Proposition 29 with $\mathcal{H}_t = \mathcal{F}_t$, $V_t = \|x_t - \bar{x}_t\|^2$, $D_t = 2\eta_t\langle z_t, v_t\rangle$, $X_t = 2\eta_t^2\|z_t\|^2 + 2\bar{\Delta}_t^2/\bar{\mu}\eta_t$, $\alpha_t = 1 - \bar{\mu}\eta_t$, $\kappa_t = 0$, $\sigma_t = 2\eta_t c\sigma$, and $\nu_t = 2\eta_t^2 c\sigma^2 + 2\bar{\Delta}^2/\bar{\mu}\eta_t$, yielding the estimate

$$\mathbb{E}\big[\exp\big(\lambda\|x_{t+1} - \bar{x}_{t+1}\|^2\big)\big] \leq \exp\left(\lambda\left(2\eta_t^2 c\sigma^2 + \frac{2\bar{\Delta}^2}{\bar{\mu}\eta_t}\right)\right)\mathbb{E}\big[\exp\big(\lambda\big(1 - \tfrac{\bar{\mu}\eta_t}{2}\big)\|x_t - \bar{x}_t\|^2\big)\big] \quad (15)$$

for all

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

6. 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.

Taking into account $\eta_t \equiv \eta$ and iterating the recursion (15), we deduce

$$\begin{aligned}\mathbb{E}\big[\exp\big(\lambda\|x_t - \bar{x}_t\|^2\big)\big] &\leq \exp\left(\lambda\big(1 - \tfrac{\bar{\mu}\eta}{2}\big)^t\|x_0 - \bar{x}_0\|^2 + \lambda\left(2\eta^2 c\sigma^2 + \frac{2\bar{\Delta}^2}{\bar{\mu}\eta}\right)\sum_{i=0}^{t-1}\big(1 - \tfrac{\bar{\mu}\eta}{2}\big)^i\right)\\ &\leq \exp\left(\lambda\left(\big(1 - \tfrac{\bar{\mu}\eta}{2}\big)^t\|x_0 - \bar{x}_0\|^2 + \frac{4\eta c\sigma^2}{\bar{\mu}} + 4\left(\frac{\bar{\Delta}}{\bar{\mu}\eta}\right)^2\right)\right)\end{aligned}$$

for all

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

Moreover, setting

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

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

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

and

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

Hence

$$\mathbb{E}\left[\exp\left(\lambda\left(\|x_t - \bar{x}_t\|^2 - \big(1 - \tfrac{\bar{\mu}\eta}{2}\big)^t\|x_0 - \bar{x}_0\|^2\right)\right)\right] \leq \exp(\lambda\nu) \quad \text{for all} \quad 0 \leq \lambda \leq 1/\nu.$$

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

With Theorem 30 in hand, we can now prove a high-probability efficiency estimate for Algorithm 3 using a step decay schedule. The following theorem provides a formal version of Theorem 18. The argument follows the same reasoning as in the proof of Theorem 28, with Theorem 30 playing the role of Corollary 27. The proof appears in Appendix B (see Section B.1).

Theorem 31 (Time to track with high probability) *Suppose that Assumption 9 holds and that we are in the low drift-to-noise regime $\bar{\Delta}/\sigma < \sqrt{\bar{\mu}/16L^3}$. Set $\bar{\eta}_\star = (2\bar{\Delta}^2/\bar{\mu}\sigma^2)^{1/3}$ and $\bar{\mathcal{E}} = (\bar{\Delta}\sigma^2/\bar{\mu}^2)^{2/3}$. Suppose moreover that we have available an upper bound on the initial squared distance $D \geq \|x_0 - \bar{x}_0\|^2$. Consider running Algorithm 3 in $k = 0, \ldots, K-1$ epochs, namely, set $X_0 = x_0$ and iterate the process*

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

where the number of epochs is

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

and we set

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

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

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

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

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

with probability at least $1-\delta$, where $\bar{X}_K$ denotes the minimizer of ψ_T.

6.3 Tracking the Equilibrium Value: Bounds in Expectation

We turn now to tracking the equilibrium value. To begin, we require a more flexible version of Lemma 24 which holds in the static regularizer setting $r_t \equiv r$.

Lemma 32 (Equilibrium one-step improvement) *The iterates $\{x_t\}$ produced by Algorithm 3 with $r_t \equiv r$ and $\eta_t < 1/L$ satisfy the following bound for all indices $i, t \in \mathbb{N}$ and arbitrary $\alpha > 0$:*

$$2\eta_i\big(\psi_t(x_{i+1}) - \psi_t^\star\big) \leq (1-\bar{\mu}\eta_i)\|x_i - \bar{x}_t\|^2 - \big(1-(\gamma+\alpha)\eta_i\big)\|x_{i+1} - \bar{x}_t\|^2 \\ + 2\eta_i\langle z_i, x_i - \bar{x}_t\rangle + \tfrac{\eta_i^2}{1-L\eta_i}\|z_i\|^2 + \tfrac{\eta_i}{\alpha}\bar{G}_{i,t}^2.$$

Proof Taking into account $r_t \equiv r$ and applying Lemma 13, we have

$$\begin{aligned}\big[\psi_t(x_{i+1}) - \psi_t(\bar{x}_t)\big] - \big[\varphi_i(x_{i+1}) - \varphi_i(\bar{x}_t)\big] \\ &= \big[f_{t,\bar{x}_t}(x_{i+1}) - f_{t,\bar{x}_t}(\bar{x}_t)\big] - \big[f_{i,x_i}(x_{t+1}) - f_{i,x_i}(\bar{x}_t)\big] \\ &\leq \big(\bar{G}_{i,t} + \gamma\|x_i - \bar{x}_t\|\big)\|x_{i+1} - \bar{x}_t\|.\end{aligned}$$

Hence

$$\psi_t(x_{i+1}) - \psi_t^\star \leq \varphi_i(x_{i+1}) - \varphi_i(\bar{x}_t) + \big(\bar{G}_{i,t} + \gamma\|x_i - \bar{x}_t\|\big)\|x_{i+1} - \bar{x}_t\|.$$

Moreover, Young's inequality implies

$$\big(\bar{G}_{i,t} + \gamma\|x_i - \bar{x}_t\|\big)\|x_{i+1} - \bar{x}_t\| \leq \tfrac{\gamma}{2}\|x_i - \bar{x}_t\|^2 + \tfrac{\gamma+\alpha}{2}\|x_{i+1} - \bar{x}_t\|^2 + \tfrac{1}{2\alpha}\bar{G}_{i,t}^2.$$

Multiplying through by $2\eta_i$ and applying Lemma 23 completes the proof. ■

Turning the estimate in Lemma 32 into an efficiency guarantee for the average iterate is essentially standard and follows for example from the averaging techniques used by Drusvyatskiy and Xiao (2022), Ghadimi and Lan (2012), and Kulunchakov and Mairal (2020). The resulting progress along the average iterate is summarized in the following proposition, while the description of the key averaging lemma is placed in Appendix A. Henceforth, we impose the regime (12): $\gamma < \mu/2$.

Proposition 33 (Progress along the average iterate) *The iterates $\{\hat{x}_t\}$ produced by Algorithm 4 with $r_t \equiv r$ and constant step size $\eta \le 1/2L$ satisfy the bound*

$$\psi_t(\hat{x}_t) - \psi_t^\star \le (1-\hat{\rho})^t\big(\psi_t(x_0) - \psi_t^\star + \tfrac{\hat{\mu}}{4}\|x_0 - \bar{x}_t\|^2\big) + \hat{\rho}\sum_{i=0}^{t-1}\langle z_i, x_i - \bar{x}_t\rangle(1-\hat{\rho})^{t-1-i}$$
$$+ \hat{\rho}\eta\sum_{i=0}^{t-1}\|z_i\|^2(1-\hat{\rho})^{t-1-i} + \tfrac{\hat{\rho}}{\hat{\mu}}\sum_{i=0}^{t-1}\bar{G}_{i,t}^2(1-\hat{\rho})^{t-1-i},$$

where $\hat{\rho} := \hat{\mu}\eta/(2-\mu\eta)$.

Proof Setting $\alpha = \hat{\mu}/2$ in Lemma 32, we obtain the following recursion for all indices $k \ge 0$ and $t \ge 1$:

$$\rho\big(\psi_k(x_t) - \psi_k^\star\big) \le (1 - c_1\rho)V_{t-1} - (1 + c_2\rho)V_t + \omega_t,$$

where $\rho = 2\eta$, $c_1 = \bar{\mu}/2$, $c_2 = -\mu/4$, $V_i = \|x_i - \bar{x}_k\|^2$, and $\omega_t = 2\eta\langle z_{t-1}, x_{t-1} - \bar{x}_k\rangle + 2\eta^2\|z_{t-1}\|^2 + (2\eta/\hat{\mu})\bar{G}_{t-1,k}^2$. The result follows by applying Lemma 42 with $h = \psi_k - \psi_k^\star$ and then taking $k = t$. ∎

Taking expectations in Proposition 33 yields the following precise version of Theorem 19.

Corollary 34 (Expected function gap) *Let $\{\hat{x}_t\}$ be the iterates produced by Algorithm 4 with constant step size $\eta \le 1/2L$, set $\hat{\rho} := \hat{\mu}\eta/(2-\mu\eta)$, and suppose that Assumption 10 holds. Then*

$$\mathbb{E}\big[\psi_t(\hat{x}_t) - \psi_t^\star\big] \le (1-\hat{\rho})^t\,\mathbb{E}\big[\psi_t(x_0) - \psi_t^\star + \tfrac{\hat{\mu}}{4}\|x_0 - \bar{x}_t\|^2\big] + \eta\sigma^2 + \frac{8\bar{\Delta}^2}{\hat{\mu}\eta^2} \tag{16}$$

for all $t \ge 0$. Consequently, we have

$$\mathbb{E}\big[\psi_t(\hat{x}_t) - \psi_t^\star\big] \lesssim (1-\hat{\rho})^t\big(\psi_0(x_0) - \psi_0^\star\big) + \eta\sigma^2 + \frac{\bar{\Delta}^2}{\hat{\mu}\eta^2}$$

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

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

Proof The bound (16) follows by taking expectations in Proposition 33 and noting

$$\sum_{i=0}^{t-1}\mathbb{E}\|z_i\|^2(1-\hat{\rho})^{t-1-i} \le \frac{\sigma^2}{\hat{\rho}} \qquad \text{and} \qquad \sum_{i=0}^{t-1}\mathbb{E}\,\bar{G}_{i,t}^2(1-\hat{\rho})^{t-1-i} \le \frac{(\hat{\mu}\bar{\Delta})^2(2-\hat{\rho})}{\hat{\rho}^3}$$

by Assumption 10. Next, applying Lemma 13, Lemma 1, and Young's inequality together with the μ-strong convexity of ψ_0 yields

$$\psi_t(x_0) - \psi_t^\star + \tfrac{\hat{\mu}}{4}\|x_0 - \bar{x}_t\|^2 \le 3\big(\psi_0(x_0) - \psi_0^\star\big) + 5\bar{G}_{0,t}^2/\bar{\mu}, \tag{17}$$

and then taking expectations and invoking Assumption 10 gives

$$\mathbb{E}\big[\psi_t(x_0) - \psi_t^\star + \tfrac{\hat{\mu}}{4}\|x_0 - \bar{x}_t\|^2\big] \le 3\big(\psi_0(x_0) - \psi_0^\star\big) + 5\hat{\mu}\bar{\Delta}^2t^2. \tag{18}$$

Further, the inequality

$$e^{-\hat{\mu}\eta t/2}\hat{\mu}t^2 \le 16/\hat{\mu}\eta^2 \qquad \forall\hat{\mu},\eta,t > 0 \tag{19}$$

combines with inequality (18) to yield

$$(1-\hat{\rho})^t\,\mathbb{E}\big[\psi_t(x_0)-\psi_t^\star+\tfrac{\hat{\mu}}{4}\|x_0-\bar{x}_t\|^2\big]\le 3(1-\hat{\rho})^t\big(\psi_0(x_0)-\psi_0^\star\big)+\frac{80\bar{\Delta}^2}{\hat{\mu}\eta^2},$$

and the remaining assertions of the corollary follow. ∎

We may now apply Corollary 34 to obtain a formal version of Theorem 20; the proof closely follows that of Theorem 28 and is included in Appendix B (see Section B.2).

Theorem 35 (Time to track in expectation) *Suppose that Assumption 10 holds and that we are in the low drift-to-noise regime $\bar{\Delta}/\sigma<\sqrt{\hat{\mu}/16L^3}$. Set $\hat{\eta}_\star=(2\bar{\Delta}^2/\hat{\mu}\sigma^2)^{1/3}$ and $\widehat{\mathcal{G}}=\hat{\mu}(\bar{\Delta}\sigma^2/\hat{\mu}^2)^{2/3}$. Suppose moreover that we have available a positive upper bound on the initial gap $D\ge\psi_0(x_0)-\psi_0^\star$. Consider running Algorithm 4 in $k=0,\ldots,K-1$ epochs, namely, set $X_0=x_0$ and iterate the process*

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

where the number of epochs is

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

and we set

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

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

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

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

6.4 Tracking the Equilibrium Value: High-Probability Guarantees

In this section, we derive the high-probability analogues of the results in Section 6.3. In light of Proposition 33, we seek upper bounds on the sums

$$\sum_{i=0}^{t-1}\langle z_i,x_i-\bar{x}_t\rangle(1-\hat{\rho})^{t-1-i},\quad\sum_{i=0}^{t-1}\|z_i\|^2(1-\hat{\rho})^{t-1-i},\quad\sum_{i=0}^{t-1}\bar{G}_{i,t}^2(1-\hat{\rho})^{t-1-i}$$

that hold with high probability. The last two sums can easily be estimated under boundedness or light-tail assumptions on $\|z_i\|$ and $\bar{G}_{i,t}$. Controlling the first sum is more challenging because the error $\|x_i-\bar{x}_t\|$ may in principle grow large. In order to control this term, we will use a remarkable generalization of Freedman's inequality, recently proved by Harvey et al. (2019) 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 $\mathbb{E}[z_i\,|\,\mathcal{F}_{i,t}]=0$ for all $0\le i<t$ (recall that $\mathcal{F}_{i,t}:=\sigma(\mathcal{F}_i,\bar{x}_t)$), and define the martingale difference sequence

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

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 32 and algebraic manipulation and is placed in Appendix B (see Section B.3).

Proposition 36 (Self-regulation) *The iterates $\{x_t\}$ produced by Algorithm 3 with $r_t \equiv r$ and constant step size $\eta \le 1/2L$ satisfy the following bound for all $\lambda \in (0, \bar{\mu}\eta]$:*

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

In order to bound the self-regulating martingale $\sum_{i=0}^{t-1} d_i$, we will use the generalized Freedman inequality developed by Harvey et al. (2019), or rather a direct consequence thereof (see Harvey et al., 2019, Lemma C.3).

Theorem 37 (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*

$$\mathbb{E}[\exp(\lambda D_i) \,|\, \mathcal{H}_i] \le \exp(\lambda^2 V_i/2) \quad \text{for all} \quad \lambda \ge 0.$$

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

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

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

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

Combining Proposition 36 and Theorem 37 yields the following tail bound for $\sum_{i=0}^{t-1} d_i$.

Proposition 38 (Noise martingale tail bound) *Let $\{x_t\}$ be the iterates produced by Algorithm 3 with constant step size $\eta \le 1/2L$, set $\hat{\rho} := \hat{\mu}\eta/(2 - \mu\eta)$, and suppose that Assumption 11 holds. 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:*

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

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

$$\beta_t := (1-\hat{\rho})^{t-1}\big(\|x_0 - \bar{x}_0\|^2 + \bar{\Delta}^2 t^2\big)\frac{2(c\sigma)^2}{\hat{\rho}} + \frac{2\eta^2(c\sigma)^4}{\hat{\rho}^2} + \frac{3\hat{\mu}\bar{\Delta}^2\eta(c\sigma)^2}{\hat{\rho}^4}.$$

Proof By Assumption 11, 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 $\langle z_i, x_i - \bar{x}_t\rangle$ is mean-zero sub-Gaussian conditioned on $\mathcal{F}_{i,t}$ with parameter $c\sigma\|x_i - \bar{x}_t\|$, so

$$\mathbb{E}\big[\exp\big(\lambda\langle z_i, x_i - \bar{x}_t\rangle(1-\hat{\rho})^{t-1-i}\big) \,|\, \mathcal{F}_{i,t}\big] \leq \exp\big(\lambda^2(c\sigma)^2\|x_i - \bar{x}_t\|^2(1-\hat{\rho})^{2(t-1-i)}/2\big) \quad \forall \lambda \in \mathbb{R};$$

note also that $\mathbb{E}|\langle z_i, x_i - \bar{x}_t\rangle| < \infty$ by Hölder's inequality, Assumption 8, and Corollary 27. Now fix $t \geq 1$ and observe that Proposition 36 yields the total conditional variance bound

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

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

$$R_t := \tfrac{(c\sigma)^2}{\hat{\rho}}(1-\hat{\rho})^{t-1}\|x_0 - \bar{x}_t\|^2 + \tfrac{2\eta^2(c\sigma)^2}{\hat{\rho}}\sum_{j=0}^{t-2}\|z_j\|^2(1-\hat{\rho})^{t-2-j} + \tfrac{\eta(c\sigma)^2}{\bar{\mu}\hat{\rho}}\sum_{j=0}^{t-2}\bar{G}_{j,t}^2(1-\hat{\rho})^{t-2-j}.$$

We claim

$$\mathbb{P}\left\{R_t \leq \beta_t \log\left(\frac{3e}{\delta}\right)\right\} \geq 1-\delta \qquad \forall \delta \in (0,1). \tag{20}$$

To verify (20), 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

$$\mathbb{P}\left\{\sum_{i=0}^n \|z_i\|^2(1-\hat{\rho})^{n-i} \leq \frac{(c\sigma)^2}{\hat{\rho}}\log\left(\frac{e}{\delta}\right)\right\} \geq 1-\delta \qquad \forall \delta \in (0,1). \tag{21}$$

Further, for all $0 \leq n < t$, it follows from Assumption 11 and Lemma 1 that $\|x_0 - \bar{x}_t\|^2$ is sub-exponential with parameter $2(\|x_0 - \bar{x}_0\|^2 + \bar{\Delta}^2 t^2)$ and $\sum_{i=0}^n \bar{G}_{i,t}^2(1-\hat{\rho})^{n-i}$ is sub-exponential with parameter

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

so Markov's inequality implies

$$\mathbb{P}\left\{\|x_0 - \bar{x}_t\|^2 \leq 2(\|x_0 - \bar{x}_0\|^2 + \bar{\Delta}^2 t^2)\log\left(\frac{e}{\delta}\right)\right\} \geq 1-\delta \qquad \forall \delta \in (0,1) \tag{22}$$

and

$$\mathbb{P}\left\{\sum_{i=0}^n \bar{G}_{i,t}^2(1-\hat{\rho})^{n-i} \leq \frac{2(\hat{\mu}\bar{\Delta})^2}{\hat{\rho}^3(1-\hat{\rho})^{t-1-n}}\log\left(\frac{e}{\delta}\right)\right\} \geq 1-\delta \qquad \forall \delta \in (0,1). \tag{23}$$

Thus, (21)–(23) and a union bound yield (20). Consequently, Theorem 37 implies that the following bound holds for all $\delta \in (0,1)$ and $\tau > 0$:

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

as claimed. ∎

We may now deduce the following precise version of Theorem 21 using the tail bound furnished by Proposition 38.

Theorem 39 (Function gap with high probability) *Let $\{\hat{x}_t\}$ be the iterates produced by Algorithm 4 with constant step size $\eta \leq 1/2L$, set $\hat{\rho} := \hat{\mu}\eta/(2-\mu\eta)$, and suppose that Assumption 11 holds. 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 $1-\delta$:*

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

where

$$\beta_t := (1-\hat{\rho})^{t-1}\big(\|x_0 - \bar{x}_0\|^2 + \bar{\Delta}^2 t^2\big)\frac{2(c\sigma)^2}{\hat{\rho}} + \frac{2\eta^2(c\sigma)^4}{\hat{\rho}^2} + \frac{3\hat{\mu}\bar{\Delta}^2\eta(c\sigma)^2}{\hat{\rho}^4}.$$

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

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

in Proposition 38 to obtain

$$\mathbb{P}\left\{\sum_{i=0}^{t-1}\langle z_i, x_i - \bar{x}_t\rangle(1-\hat{\rho})^{t-1-i} < 5\sqrt{8\beta_t}\log\left(\frac{e}{\delta}\right)\right\} \geq 1 - 2\delta. \tag{24}$$

We may now combine (21), (23), and (24) together with Proposition 33 and a union bound to conclude that for all $\delta \in (0,1)$, the estimate

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

holds with probability at least $1-4\delta$; noting $\hat{\rho} \geq \hat{\mu}\eta/2$ completes the proof. ∎

Remark 40 To see that Theorem 39 entails Theorem 21, observe first that in the setting of Theorem 39, upon setting $C := \max\{c, 1\}$ and selecting any $t \in \mathbb{N}$, we have

$$\hat{\rho}\sqrt{8\beta_t} \leq 4C^2\left(\sqrt{(1-\hat{\rho})^t\big(\|x_0 - \bar{x}_0\|^2 + \bar{\Delta}^2 t^2\big)\hat{\mu}\eta\sigma^2} + \eta\sigma^2 + \sqrt{6}\frac{\bar{\Delta}\sigma}{\sqrt{\hat{\mu}\eta}}\right),$$

while the AM-GM inequality implies

$$2\sqrt{(1-\hat{\rho})^t\big(\|x_0 - \bar{x}_0\|^2 + \bar{\Delta}^2 t^2\big)\hat{\mu}\eta\sigma^2} \leq (1-\hat{\rho})^t\big(\hat{\mu}\|x_0 - \bar{x}_0\|^2 + \hat{\mu}\bar{\Delta}^2 t^2\big) + \eta\sigma^2,$$

inequality (19) implies

$$(1-\hat{\rho})^t\big(\hat{\mu}\|x_0-\bar{x}_0\|^2+\hat{\mu}\bar{\Delta}^2t^2\big)\leq 2(1-\hat{\rho})^t\big(\psi_0(x_0)-\psi_0^\star\big)+\frac{16\bar{\Delta}^2}{\hat{\mu}\eta^2},$$

and Young's inequality implies

$$\frac{2\bar{\Delta}\sigma}{\sqrt{\hat{\mu}\eta}}\leq \eta\sigma^2+\frac{\bar{\Delta}^2}{\hat{\mu}\eta^2}.$$

Hence

$$\hat{\rho}\sqrt{8\beta_t}\lesssim (1-\hat{\rho})^t\big(\psi_0(x_0)-\psi_0^\star\big)+\eta\sigma^2+\frac{\bar{\Delta}^2}{\hat{\mu}\eta^2}.$$

Further, inequalities (17) and (19) together with Assumption 11 imply that the estimate

$$(1-\hat{\rho})^t\big(\psi_t(x_0)-\psi_t^\star+\tfrac{\hat{\mu}}{4}\|x_0-\bar{x}_t\|^2\big)\leq 3(1-\hat{\rho})^t\big(\psi_0(x_0)-\psi_0^\star\big)+\frac{80\bar{\Delta}^2}{\hat{\mu}\eta^2}\log\Big(\frac{e}{\delta}\Big)$$

holds with probability at least $1-\delta$ for all $\delta\in(0,1)$. On the other hand, Theorem 39 shows that the estimate

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

holds with probability at least $1-\delta$ for all $\delta\in(0,1)$. Thus, a union bound reveals that the estimate

$$\psi_t(\hat{x}_t)-\psi_t^\star\lesssim\left((1-\hat{\rho})^t\big(\psi_0(x_0)-\psi_0^\star\big)+\eta\sigma^2+\frac{\bar{\Delta}^2}{\hat{\mu}\eta^2}\right)\log\Big(\frac{e}{\delta}\Big)$$

holds with probability at least $1-\delta$ for all $\delta\in(0,1)$.

We may now apply Theorem 21 to obtain a formal version of Theorem 22; the proof is analogous to that of Theorem 31 and appears in Appendix B (see Section B.4).

Theorem 41 (Time to track with high probability) *Suppose that Assumption 11 holds and that we are in the low drift-to-noise regime $\bar{\Delta}/\sigma < \sqrt{\hat{\mu}/16L^3}$. Set $\hat{\eta}_\star = (2\bar{\Delta}^2/\hat{\mu}\sigma^2)^{1/3}$ and $\widehat{\mathcal{G}} = \hat{\mu}(\bar{\Delta}\sigma^2/\hat{\mu}^2)^{2/3}$. Suppose moreover that we have available a positive upper bound on the initial gap $D\geq\psi_0(x_0)-\psi_0^\star$. Fix $\delta\in(0,1)$ and consider running Algorithm 4 in $k=0,\ldots,K-1$ epochs, namely, set $X_0=x_0$ and iterate the process*

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

where the number of epochs is

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

and we set

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

for all $k\geq 1$, where $c>0$ is the absolute constant furnished by the bound (13). Then the time horizon $T=T_0+\cdots+T_{K-1}$ satisfies

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

and the corresponding tracking error satisfies

$$\psi_T(X_K) - \psi_T^\star \lesssim \widehat{\mathcal{G}} \log\Big(\frac{e}{\delta}\Big)$$

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

7. 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`.

Least-squares recovery. Fix $x_0, x_0^\star \in \mathbb{R}^d$ 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 rank-d matrix $A \in \mathbb{R}^{n\times d}$ with minimum singular value $\sqrt{\mu}$ and maximum singular value $\sqrt{L}$, we aim to recover $\{x_t^\star\}$ via the online least-squares problem

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

where $\mathcal{P}_t = \mathsf{N}(Ax_t^\star, \Sigma_t)$ with covariance matrix Σ_t satisfying $\operatorname{tr}\Sigma_t \le \sigma^2/L$. This amounts to the problem (5) with $f_t(x) = \mathbb{E}_{y\sim\mathcal{P}_t}\frac{1}{2}\|Ax - y\|^2$ and $r_t = 0$, and the minimizer and gradient drift satisfy

$$\|\nabla f_t(x) - \nabla f_{t+1}(x)\| = \|A^\top A(x_t^\star - x_{t+1}^\star)\| \le L\|x_t^\star - x_{t+1}^\star\| = L\Delta$$

for all $x \in \mathbb{R}^d$. We implement Algorithms 1 and 2 using the sample gradient $g_t = A^T(Ax_t - y_t)$ at step t with $y_t \sim \mathcal{P}_t$; the gradient noise $z_t = A^\top(y_t - Ax_t^\star) \sim \mathsf{N}(0, A^\top\Sigma_t A)$ satisfies $\mathbb{E}\|z_t\|^2 \le L\operatorname{tr}\Sigma_t \le \sigma^2$.

In our simulations, we set $d = 50$, $n = 100$, and $\Sigma_t = (\sigma^2/nL)I_n$ for all t, where I_n denotes the $n \times n$ identity matrix. We initialize x_0 and $x_0^\star$ using standard Gaussian entries and generate A via singular value decomposition with Haar-distributed orthogonal matrices. In Figures 1 and 2, we use default parameter values $\mu = L = 1$, $\sigma = 10$, $\Delta = 1$, 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 expected values 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 $T = 100$.

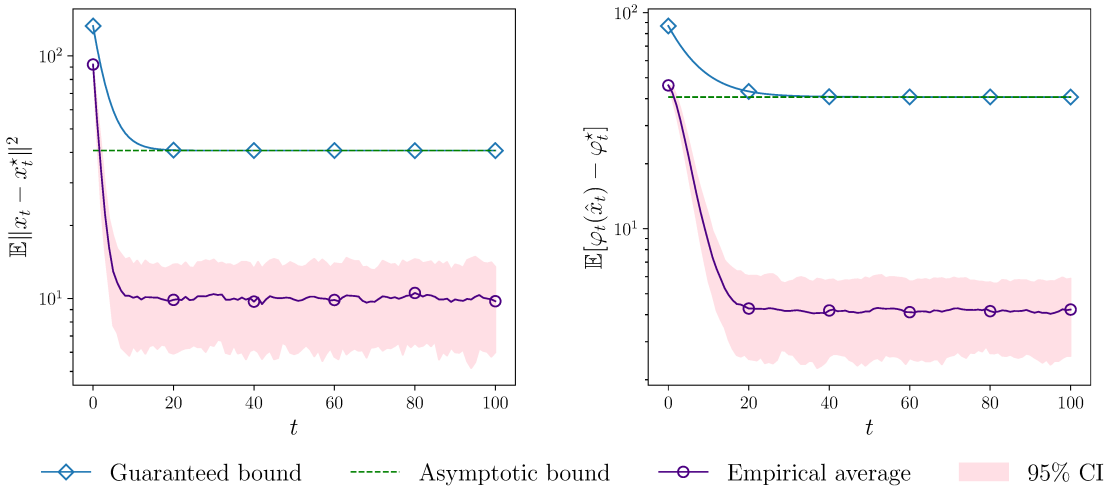

Figure 1: Semilog plots of guaranteed bounds and empirical tracking errors with respect to iteration t for least-squares recovery. Shaded regions indicate the 95% confidence intervals for $\|x_t - x_t^\star\|^2$ and $\varphi_t(\hat{x}_t) - \varphi_t^\star$; empirical averages and confidence intervals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

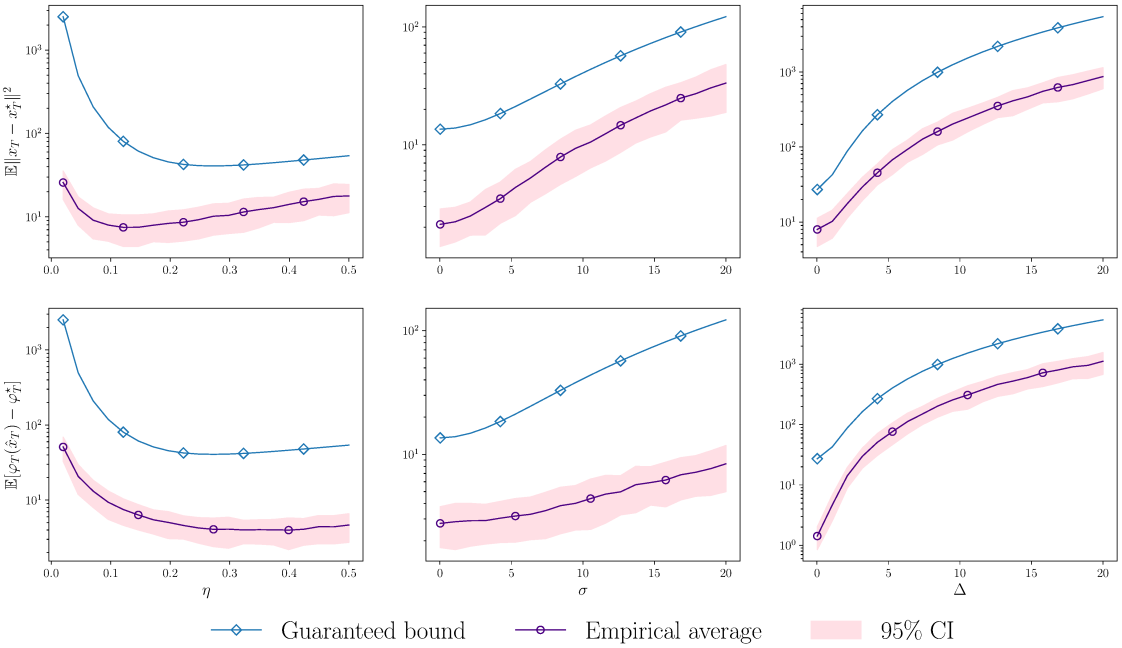

Figure 2: Semilog plots of guaranteed bounds and empirical tracking errors at horizon $T = 100$ with respect to η, σ, and Δ for least-squares recovery. Shaded regions indicate the 95% confidence intervals for $\|x_T - x_T^\star\|^2$ and $\varphi_T(\hat{x}_T) - \varphi_T^\star$; empirical averages and confidence intervals are computed over 100 trials. Default parameter values: $\mu = L = 1$, $\sigma = 10$, $\Delta = 1$, 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 sequence of sparse vectors in B_1 generated 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 $x_{t+1}^\star = x_t^\star + v_t$, where v_t is selected to have the same support as $x_t^\star$ and satisfy $\|v_t\| = \Delta/\sqrt{2}$ and $x_t^\star + v_t \in B_1$; 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. The resulting sequence $\{x_t^\star\}$ in B_1 satisfies $\mathbb{E}\|x_t^\star - x_{t+1}^\star\|^2 \leq \Delta^2$. Given a fixed rank-$d$ matrix $A \in \mathbb{R}^{n \times d}$ with minimum singular value $\sqrt{\mu}$ and maximum singular value $\sqrt{L}$, we aim to recover $\{x_t^\star\}$ via the online constrained least-squares problem

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

where $\mathcal{P}_t = \mathsf{N}(Ax_t^\star, \Sigma_t)$ with covariance matrix Σ_t satisfying $\operatorname{tr} \Sigma_t \leq \sigma^2/L$. This amounts to the problem (5) with $f_t(x) = \mathbb{E}_{y \sim \mathcal{P}_t} \frac{1}{2}\|Ax - y\|^2$ and $r_t = \delta_{B_1}$ (the convex indicator of B_1), and the minimizer and gradient drift satisfy

$$\mathbb{E}\big[\sup_x \|\nabla f_t(x) - \nabla f_{t+1}(x)\|^2\big] \leq L^2\, \mathbb{E}\|x_t^\star - x_{t+1}^\star\|^2 \leq (L\Delta)^2.$$

Fixing x_0 drawn uniformly from B_1, we implement Algorithms 1 and 2 initialized at x_0 using the sample gradient $g_t = A^T(Ax_t - y_t)$ at step t with $y_t \sim \mathcal{P}_t$; hence $\mathbb{E}\|\nabla f_t(x_t) - g_t\|^2 \leq \sigma^2$.

In our simulations, we set $d = 50$, $n = 100$, and $\Sigma_t = (\sigma^2/nL) I_n$ for all t. We generate A via singular value decomposition with Haar-distributed orthogonal matrices. In Figures 3 and 4, we use default parameter values $\mu = L = 1$, $\sigma = 1/2$, $\Delta = 1/20$, 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 expected values 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 $T = 100$.

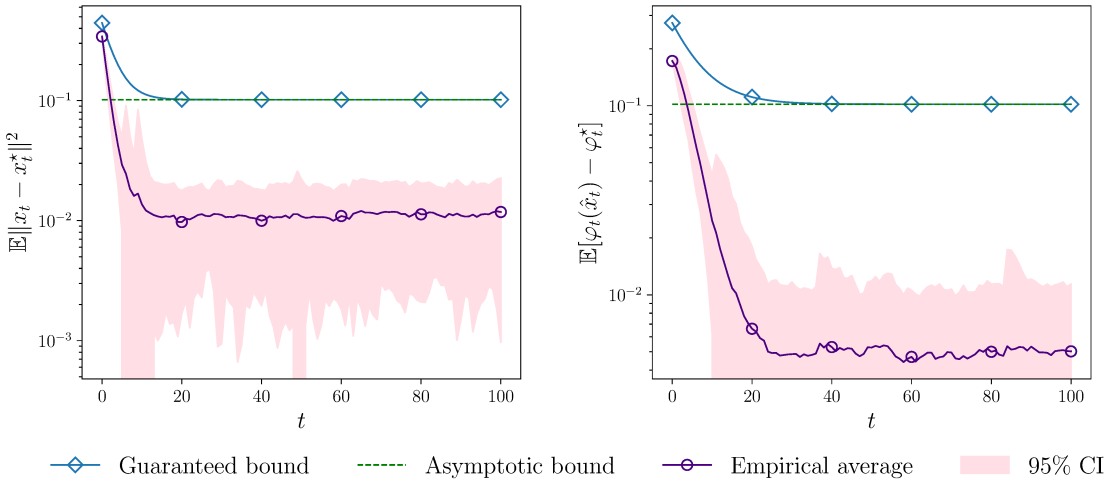

Figure 3: Semilog plots of guaranteed bounds and empirical tracking errors with respect to iteration t for sparse least-squares recovery. Shaded regions indicate the 95% confidence intervals for $\|x_t - x_t^\star\|^2$ and $\varphi_t(\hat{x}_t) - \varphi_t^\star$; empirical averages and confidence intervals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

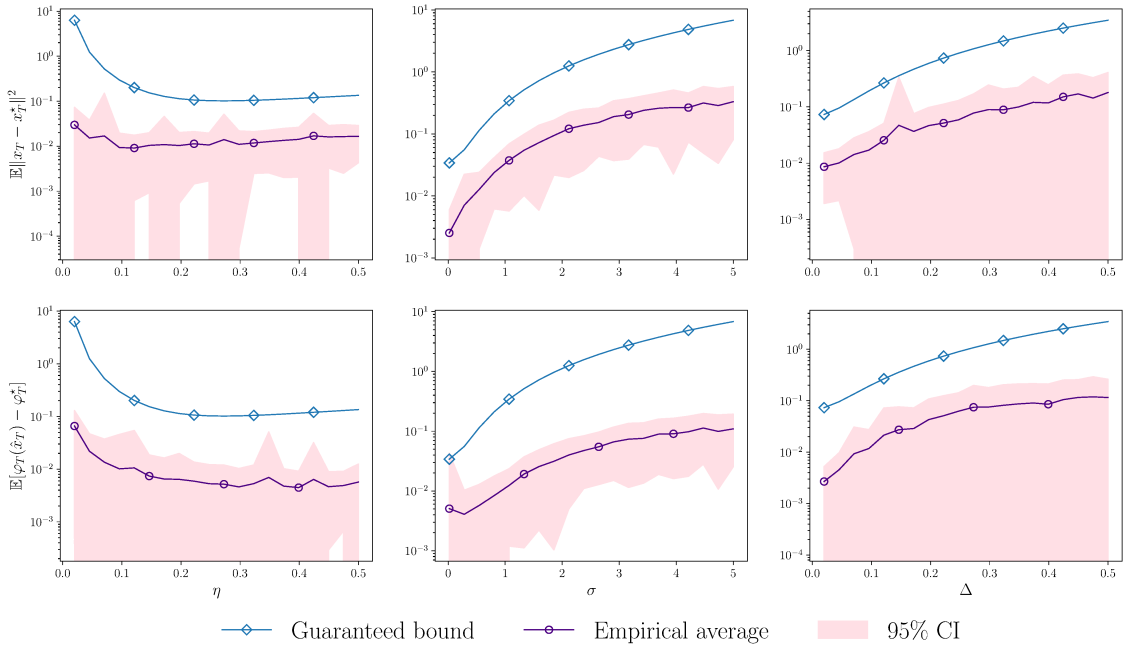

Figure 4: Semilog plots of guaranteed bounds and empirical tracking errors at horizon $T = 100$ with respect to η, σ, and Δ for sparse least-squares recovery. Shaded regions indicate the 95% confidence intervals for $\|x_T - x_T^\star\|^2$ and $\varphi_T(\hat{x}_T) - \varphi_T^\star$; empirical averages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

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

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

where the matrix $A \in \mathbb{R}^{n \times d}$ has fixed 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 problem (5) with $f_t(x) = \frac{1}{n}(\sum_{i=1}^n \log(1 + \exp\langle a_i, x\rangle) - \langle Ax, b_t\rangle) + \frac{\mu}{2}\|x\|^2$ and $r_t = 0$; setting $L = \frac{1}{4n}\|A\|_{\mathrm{op}}^2 + \mu$, it follows 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 the minimizer and gradient drift satisfy

$$\mu\|x_t^\star - x_{t+1}^\star\| \le \sup_x \|\nabla f_t(x) - \nabla f_{t+1}(x)\| \le \mu\Delta.$$

We implement Algorithms 1 and 2 using the random summand sample gradient

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

at step t, where $k \sim \mathsf{Unif}\{1, \ldots, n\}$ and b_t^k denotes the k^{th} coordinate of b_t; the gradient noise satisfies $\mathbb{E}\|\nabla f_t(x_t) - g_t\|^2 \le \sigma^2$, where

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

In our simulations, we set $d = 20$ and $n = 200$, fix standard Gaussian vectors $x_0 \in \mathbb{R}^d$ and $a_1, \ldots, a_n \in \mathbb{R}^d$, fix b_0 drawn uniformly from $\{0,1\}^n$, and 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 expected values 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.

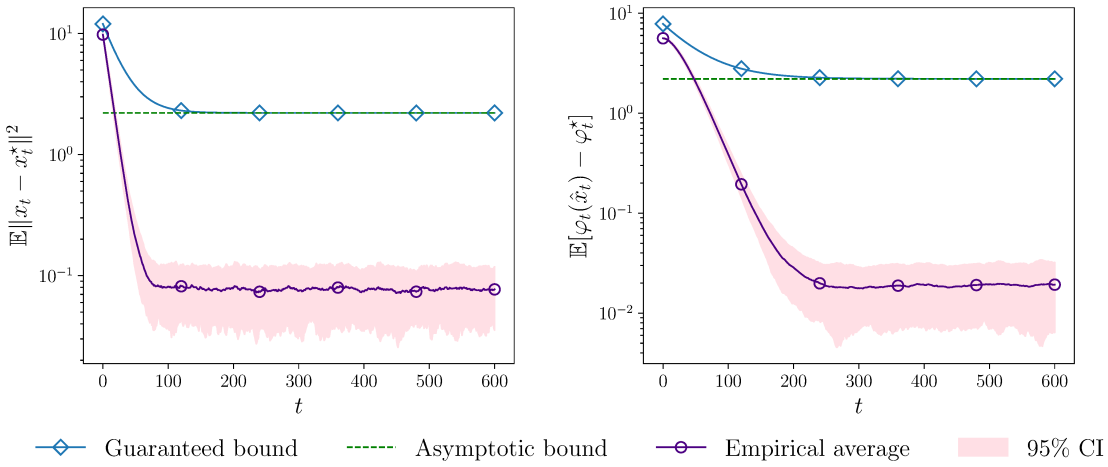

Figure 5: Semilog plots of guaranteed bounds and empirical tracking errors with respect to iteration t for ℓ_2^2-regularized logistic regression. Shaded regions indicate the 95% confidence intervals for $\|x_t - x_t^\star\|^2$ and $\varphi_t(\hat{x}_t) - \varphi_t^\star$; empirical averages and confidence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

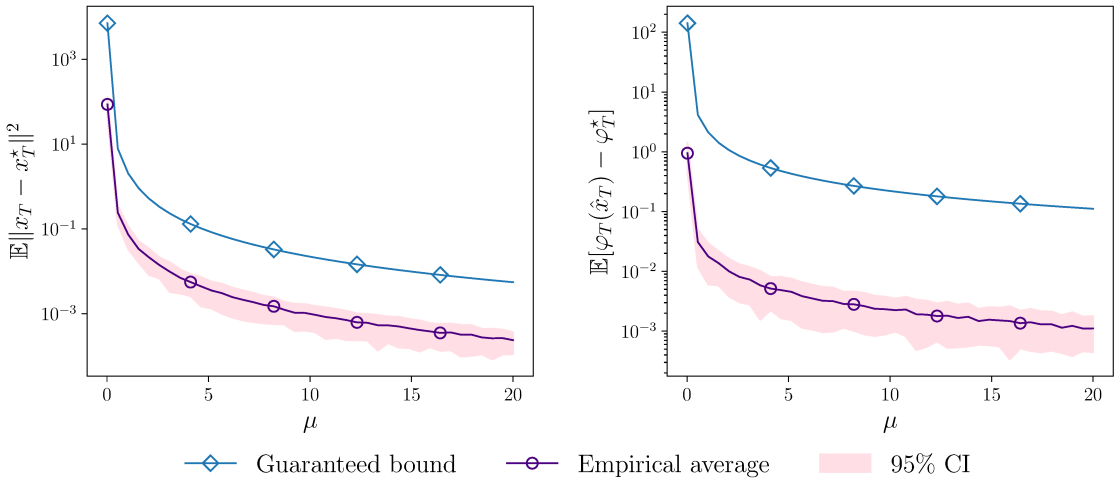

Figure 6: Semilog plots of guaranteed bounds and empirical tracking errors at horizon $T = 600$ with respect to the strong convexity parameter μ for ℓ_2^2-regularized logistic regression. Shaded regions indicate the 95% confidence intervals for $\|x_T - x_T^\star\|^2$ and $\varphi_T(\hat{x}_T) - \varphi_T^\star$;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

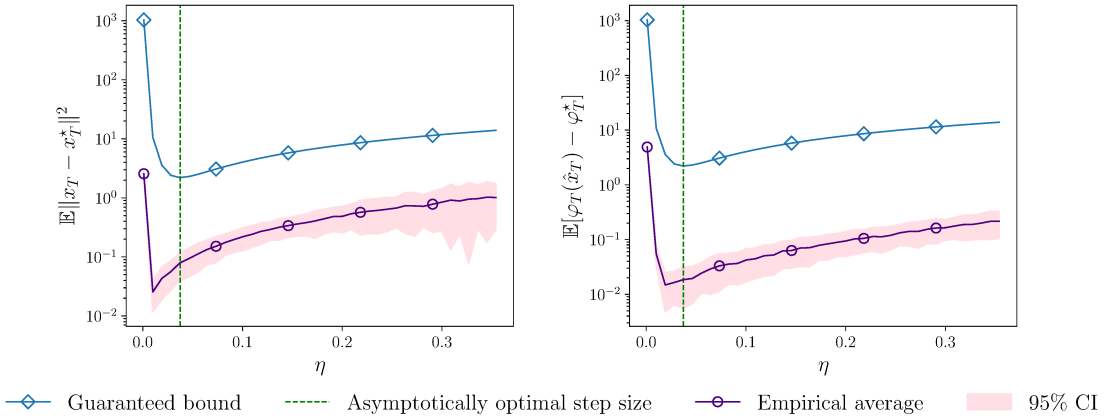

Figure 7: Semilog plots of guaranteed bounds and empirical tracking errors at horizon $T = 600$ with respect to the step size η for ℓ_2^2-regularized logistic regression. Shaded regions indicate the 95% confidence intervals for $\|x_T - x_T^\star\|^2$ and $\varphi_T(\hat{x}_T) - \varphi_T^\star$; empirical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

Acknowledgments

This work was supported by NSF DMS-2023166, DMS-2134012, DMS-1651851, DMS-2133244, CCF-2019844, and CIFAR LMB. Part of this work was done while Z. Harchaoui was visiting the Simons Institute for the Theory of Computing.

Appendix A. Averaging Lemma

We will use a variation of the averaging strategy used by Ghadimi and Lan (2012); our approach here follows Drusvyatskiy and Xiao (2022, Section A) and Kulunchakov and Mairal (2020, Sections A.2 and A.3). To begin, 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 $\mathbb{R}^d$. Suppose that there are constants $c_1, c_2 \in \mathbb{R}$, a sequence of nonnegative weights $\{\rho_t\}_{t\geq 1}$, and scalar sequences $\{V_t\}_{t\geq 0}$ and $\{\omega_t\}_{t\geq 1}$ satisfying the recursion

$$\rho_t h(x_t) \leq (1 - c_1\rho_t)V_{t-1} - (1 + c_2\rho_t)V_t + \omega_t \tag{25}$$

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

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

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

$$1 + \sum_{i=1}^{t} \frac{\hat{\rho}_i}{\hat{\Gamma}_i} = \frac{1}{\hat{\Gamma}_t}. \tag{26}$$

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

$$\hat{x}_t = \hat{\Gamma}_t \left(x_0 + \sum_{i=1}^{t} \frac{\hat{\rho}_i}{\hat{\Gamma}_i} x_i \right). \tag{27}$$

The following lemma provides the key estimate we will need.

Lemma 42 (Averaging) *The 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 (27) expresses $\hat{x}_t$ as a convex combination of $x_0, \ldots, x_t$ by virtue of (26). Therefore, by the convexity of h, we may apply Jensen's inequality to obtain

$$h(\hat{x}_t) \leq \hat{\Gamma}_t \left(h(x_0) + \sum_{i=1}^{t} \frac{\hat{\rho}_i}{\hat{\Gamma}_i} h(x_i) \right). \tag{28}$$

On the other hand, for each $i \geq 1$, we may divide the recursion (25) by $\hat{\Gamma}_i(1+c_2\rho_i)$ to obtain

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

which telescopes to yield

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

Multiplying this inequality by $\hat{\Gamma}_t$ and applying (28) yields

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

as claimed. ∎

Appendix B. Additional Proofs

B.1 Proof of Theorem 31

For each index k, let $t_k := T_0 + \cdots + T_{k-1}$ (with $t_0 := 0$), $\bar{X}_k$ be the minimizer of the corresponding function ψ_{t_k}, and

$$\bar{E}_k := c\left(\frac{\eta_k \sigma^2}{\bar{\mu}} + \left(\frac{\bar{\Delta}}{\bar{\mu}\bar{\eta}_\star} \right)^2 \right),$$

where $c \geq 1$ is an absolute constant satisfying the bound (11) in Theorem 17. Taking into account $\eta_k \geq \bar{\eta}_\star$ and our selection of c, Theorem 17 implies that for any specified index k and $\delta \in (0,1)$, the following estimate holds with probability at least $1-\delta$:

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

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

$$\|X_1 - \bar{X}_1\|^2 \leq e^{-\bar{\mu}\eta_0 T_0/2} \|X_0 - \bar{X}_0\|^2 + \bar{E}_0 \log\left(\frac{e}{\delta}\right) \leq 3\bar{E}_0 \log\left(\frac{e}{\delta}\right)$$

holds with probability at least $1-\delta$ for all $\delta \in (0,1)$. Now assume that the claim holds for some index $k \geq 1$, and let $\delta \in (0,1)$; then $\|X_k - \bar{X}_k\|^2 \leq 3\bar{E}_{k-1}\log(2e/\delta)$ with probability

at least $1-\delta/2$. Thus, since we also have

$$\begin{aligned}\|X_{k+1}-\bar{X}_{k+1}\|^2 &\leq e^{-\bar{\mu}\eta_k T_k/2}\|X_k-\bar{X}_k\|^2+\bar{E}_k\log\biggl(\frac{2e}{\delta}\biggr)\\ &\leq \frac{1}{12}\|X_k-\bar{X}_k\|^2+\bar{E}_k\log\biggl(\frac{2e}{\delta}\biggr)\\ &\leq \frac{\bar{E}_k}{6\bar{E}_{k-1}}\|X_k-\bar{X}_k\|^2+\bar{E}_k\log\biggl(\frac{2e}{\delta}\biggr)\end{aligned}$$

with probability at least $1-\delta/2$, a union bound reveals

$$\|X_{k+1}-\bar{X}_{k+1}\|^2 \leq \frac{3}{2}\bar{E}_k\log\biggl(\frac{2e}{\delta}\biggr) \leq 3\bar{E}_k\log\Bigl(\frac{e}{\delta}\Bigr)$$

with probability at least $1-\delta$, thereby completing the induction. Hence, upon fixing $\delta\in(0,1)$, we have $\|X_K-\bar{X}_K\|^2\leq 3\bar{E}_{K-1}\log(e/\delta)$ with probability at least $1-\delta$.

Next, observe

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

so

$$\|X_K-\bar{X}_K\|^2\leq\frac{3c}{2}\bigl(1+\sqrt[3]{54}\bigr)\bar{\mathcal{E}}\log\Bigl(\frac{e}{\delta}\Bigr)\asymp\bar{\mathcal{E}}\log\Bigl(\frac{e}{\delta}\Bigr)$$

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

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

and

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

This completes the proof.

B.2 Proof of Theorem 35

For each index k, let $t_k:=T_0+\cdots+T_{k-1}$ (with $t_0:=0$) and $\widehat{G}_k:=\eta_k\sigma^2+8\bar{\Delta}^2/\hat{\mu}\hat{\eta}_\star^2$. Then taking into account $\eta_k\geq\hat{\eta}_\star$, Corollary 34 and inequality (17) directly imply

$$\begin{aligned}\mathbb{E}\bigl[\psi_{t_{k+1}}(X_{k+1})-\psi_{t_{k+1}}^\star\bigr] &\leq \bigl(1-\tfrac{\hat{\mu}\eta_k}{2}\bigr)^{T_k}\mathbb{E}\bigl[3\bigl(\psi_{t_k}(X_k)-\psi_{t_k}^\star\bigr)+5\hat{\mu}\bar{\Delta}^2T_k^2\bigr]+\eta_k\sigma^2+\frac{8\bar{\Delta}^2}{\hat{\mu}\eta_k^2}\\ &\leq 3e^{-\hat{\mu}\eta_kT_k/2}\mathbb{E}\bigl[\psi_{t_k}(X_k)-\psi_{t_k}^\star\bigr]+5e^{-\hat{\mu}\eta_kT_k/2}\hat{\mu}\bar{\Delta}^2T_k^2+\widehat{G}_k.\end{aligned}$$

We will verify by induction that the estimate $\mathbb{E}\bigl[\psi_{t_k}(X_k)-\psi_{t_k}^\star\bigr]\leq 11\widehat{G}_{k-1}$ holds for all indices $k\geq 1$. To see the base case, observe that inequality (19) facilitates the estimation

$$\mathbb{E}\bigl[\psi_{t_1}(X_1)-\psi_{t_1}^\star\bigr]\leq 3e^{-\hat{\mu}\eta_0T_0/2}\bigl(\psi_0(x_0)-\psi_0^\star\bigr)+5e^{-\hat{\mu}\eta_0T_0/2}\hat{\mu}\bar{\Delta}^2T_0^2+\widehat{G}_0\leq 11\widehat{G}_0.$$

Now assume that the claim holds for some index $k \geq 1$. We then conclude

$$\begin{aligned}\mathbb{E}\big[\psi_{t_{k+1}}(X_{k+1}) - \psi^\star_{t_{k+1}}\big] &\leq 3e^{-\hat{\mu}\eta_k T_k/2}\mathbb{E}\big[\psi_{t_k}(X_k) - \psi^\star_{t_k}\big] + 5e^{-\hat{\mu}\eta_k T_k/2}\hat{\mu}\bar{\Delta}^2 T_k^2 + \widehat{G}_k \\ &\leq \frac{1}{4}\mathbb{E}\big[\psi_{t_k}(X_k) - \psi^\star_{t_k}\big] + \frac{13\bar{\Delta}^2}{\hat{\mu}\eta_k^2} + \widehat{G}_k \\ &\leq \frac{\widehat{G}_k}{2\widehat{G}_{k-1}}\mathbb{E}\big[\psi_{t_k}(X_k) - \psi^\star_{t_k}\big] + \frac{13\bar{\Delta}^2}{\hat{\mu}\eta_k^2} + \widehat{G}_k \leq 11\widehat{G}_k,\end{aligned}$$

thereby completing the induction. Hence $\mathbb{E}\big[\psi_T(X_K) - \psi^\star_T\big] \leq 11\widehat{G}_{K-1}$.

Next, observe

$$\widehat{G}_{K-1} - \sqrt[3]{250}\cdot\hat{\mu}\left(\frac{\bar{\Delta}\sigma^2}{\hat{\mu}^2}\right)^{2/3} = \sigma^2(\eta_{K-1} - \hat{\eta}_\star) = \sigma^2\cdot\frac{\eta_0 - \hat{\eta}_\star}{2^{K-1}} \leq \frac{\hat{\mu}}{2}\left(\frac{\bar{\Delta}\sigma^2}{\hat{\mu}^2}\right)^{2/3} = \frac{1}{2}\widehat{\mathcal{G}},$$

so

$$\mathbb{E}\big[\psi_T(X_K) - \psi^\star_T\big] \leq 11\big(\tfrac{1}{2} + \sqrt[3]{250}\big)\cdot\hat{\mu}\left(\frac{\bar{\Delta}\sigma^2}{\hat{\mu}^2}\right)^{2/3} \asymp \widehat{\mathcal{G}}.$$

Finally, note

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

and

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

This completes the proof.

B.3 Proof of Proposition 36

Fix $t \geq 1$. Given $i \geq 1$ and $\alpha > 0$, the μ-strong convexity of ψ_t and Lemma 32 imply

$$\begin{aligned}\mu\eta\|x_i - \bar{x}_t\|^2 \leq 2\eta\big(\psi_t(x_i) - \psi^\star_t\big) &\leq (1-\bar{\mu}\eta)\|x_{i-1} - \bar{x}_t\|^2 - \big(1 - (\gamma+\alpha)\eta\big)\|x_i - \bar{x}_t\|^2 \\ &\quad + 2\eta\langle z_{i-1}, x_{i-1} - \bar{x}_t\rangle + 2\eta^2\|z_{i-1}\|^2 + \tfrac{\eta}{\alpha}\bar{G}^2_{i-1,t},\end{aligned}$$

hence

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

Taking $\alpha = \bar{\mu}$, we obtain

$$\|x_i - \bar{x}_t\|^2 \leq (1-\bar{\mu}\eta)\|x_{i-1} - \bar{x}_t\|^2 + 2\eta\langle z_{i-1}, x_{i-1} - \bar{x}_t\rangle + 2\eta^2\|z_{i-1}\|^2 + \tfrac{\eta}{\bar{\mu}}\bar{G}^2_{i-1,t}.$$

Thus, given any $\lambda \in (0, \bar{\mu}\eta]$ and proceeding by induction, we conclude that the following estimate holds for all $i \geq 1$:

$$\begin{aligned}\|x_i - \bar{x}_t\|^2 \leq (1-\lambda)^i\|x_0 - \bar{x}_t\|^2 &+ 2\eta\sum_{j=0}^{i-1}\langle z_j, x_j - \bar{x}_t\rangle(1-\lambda)^{i-1-j} \\ &+ 2\eta^2\sum_{j=0}^{i-1}\|z_j\|^2(1-\lambda)^{i-1-j} + \tfrac{\eta}{\bar{\mu}}\sum_{j=0}^{i-1}\bar{G}^2_{j,t}(1-\lambda)^{i-1-j}.\end{aligned}$$

Therefore

$$\sum_{i=0}^{t-1}\|x_i-\bar{x}_t\|^2(1-\lambda)^{2(t-1-i)}$$
$$\leq \|x_0-\bar{x}_t\|^2\sum_{i=0}^{t-1}(1-\lambda)^{2(t-1)-i}+2\eta\sum_{i=1}^{t-1}\sum_{j=0}^{i-1}\langle z_j,x_j-\bar{x}_t\rangle(1-\lambda)^{2t-3-j-i}$$
$$+2\eta^2\sum_{i=1}^{t-1}\sum_{j=0}^{i-1}\|z_j\|^2(1-\lambda)^{2t-3-j-i}+\tfrac{\eta}{\bar{\mu}}\sum_{i=1}^{t-1}\sum_{j=0}^{i-1}\bar{G}_{j,t}^2(1-\lambda)^{2t-3-j-i}.$$

Next, 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 $\{a_j\}_{j=0}^{t-2}$, we have

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

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

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

Hence the following estimation holds:

$$\sum_{i=0}^{t-1}\|x_i-\bar{x}_t\|^2(1-\lambda)^{2(t-1-i)}\leq\sum_{j=0}^{t-2}\left(2\eta\sum_{i=j+1}^{t-1}(1-\lambda)^{t-2-i}\right)\langle z_j,x_j-\bar{x}_t\rangle(1-\lambda)^{t-1-j}$$
$$+\tfrac{1}{\lambda}(1-\lambda)^{t-1}\|x_0-\bar{x}_t\|^2+\tfrac{2\eta^2}{\lambda}\sum_{j=0}^{t-2}\|z_j\|^2(1-\lambda)^{t-2-j}$$
$$+\tfrac{\eta}{\bar{\mu}\lambda}\sum_{j=0}^{t-2}\bar{G}_{j,t}^2(1-\lambda)^{t-2-j}.$$

This completes the proof.

B.4 Proof of Theorem 41

For each index k, let $t_k:=T_0+\cdots+T_{k-1}$ (with $t_0:=0$) and $\widehat{G}_k:=\eta_k\sigma^2+\bar{\Delta}^2/\hat{\mu}\hat{\eta}_\star^2$. Then taking into account $\eta_k\geq\hat{\eta}_\star$ and our selection of the absolute constant $c>0$ via (13), it

follows that for any specified index k, the estimate

$$\begin{aligned}\psi_{t_{k+1}}(X_{k+1}) - \psi^\star_{t_{k+1}} &\le c\left(\left(1 - \tfrac{\hat{\mu}\eta_k}{2}\right)^{T_k}\left(\psi_{t_k}(X_k) - \psi^\star_{t_k}\right) + \eta_k\sigma^2 + \frac{\bar{\Delta}^2}{\hat{\mu}\eta_k^2}\right)\log\left(\frac{e}{\delta}\right)\\ &\le c\left(e^{-\hat{\mu}\eta_k T_k/2}\left(\psi_{t_k}(X_k) - \psi^\star_{t_k}\right) + \widehat{G}_k\right)\log\left(\frac{e}{\delta}\right)\end{aligned}$$

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

We will verify by induction that for each index $k \ge 1$, the estimate

$$\psi_{t_k}(X_k) - \psi^\star_{t_k} \le 3c\cdot\widehat{G}_{k-1}\log\left(\frac{e}{\delta}\right)$$

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

$$\psi_{t_1}(X_1) - \psi^\star_{t_1} \le c\left(e^{-\hat{\mu}\eta_0 T_0/2}\left(\psi_0(x_0) - \psi^\star_0\right) + \widehat{G}_0\right)\log\left(\frac{e}{\delta}\right) \le 3c\cdot\widehat{G}_0\log\left(\frac{e}{\delta}\right)$$

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

$$\begin{aligned}\psi_{t_{k+1}}(X_{k+1}) - \psi^\star_{t_{k+1}} &\le c\left(e^{-\hat{\mu}\eta_k T_k/2}\left(\psi_{t_k}(X_k) - \psi^\star_{t_k}\right) + \widehat{G}_k\right)\log\left(\frac{e}{\delta}\right)\\ &\le c\left(\frac{1}{4c\log(e/\delta)}\left(\psi_{t_k}(X_k) - \psi^\star_{t_k}\right) + \widehat{G}_k\right)\log\left(\frac{e}{\delta}\right)\\ &\le c\left(\frac{\widehat{G}_k}{2c\cdot\widehat{G}_{k-1}\log(e/\delta)}\left(\psi_{t_k}(X_k) - \psi^\star_{t_k}\right) + \widehat{G}_k\right)\log\left(\frac{e}{\delta}\right)\end{aligned}$$

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

$$\psi_{t_{k+1}}(X_{k+1}) - \psi^\star_{t_{k+1}} \le 3c\cdot\widehat{G}_k\log\left(\frac{e}{\delta}\right)$$

holds with probability at least $1-(k+1)\delta$, thereby completing the induction. In particular, $\psi_T(X_K) - \psi^\star_T \le 3c\cdot\widehat{G}_{K-1}\log(e/\delta)$ with probability at least $1-K\delta$.

Next, observe

$$\widehat{G}_{K-1} - \sqrt[3]{\tfrac{27}{4}}\cdot\hat{\mu}\left(\frac{\bar{\Delta}\sigma^2}{\hat{\mu}^2}\right)^{2/3} = \sigma^2(\eta_{K-1} - \hat{\eta}_\star) = \sigma^2\cdot\frac{\eta_0 - \hat{\eta}_\star}{2^{K-1}} \le \frac{\hat{\mu}}{2}\left(\frac{\bar{\Delta}\sigma^2}{\hat{\mu}^2}\right)^{2/3} = \frac{1}{2}\widehat{\mathcal{G}},$$

so

$$\psi_T(X_K) - \psi^\star_T \le 3c\left(\tfrac{1}{2} + \sqrt[3]{\tfrac{27}{4}}\right)\cdot\hat{\mu}\left(\frac{\bar{\Delta}\sigma^2}{\hat{\mu}^2}\right)^{2/3}\log\left(\frac{e}{\delta}\right) \asymp \widehat{\mathcal{G}}\log\left(\frac{e}{\delta}\right)$$

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

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

and

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

This completes the proof.

References

Alekh Agarwal, Olivier Chapelle, Miroslav Dudík, and John Langford. A reliable effective terascale linear learning system. *Journal of Machine Learning Research*, 15(1):1111–1133, 2014.

Francis Bach and Eric 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.

Peter L. Bartlett, Varsha Dani, Thomas P. Hayes, Sham M. Kakade, Alexander Rakhlin, and Ambuj Tewari. High-probability regret bounds for bandit online linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory*. Omnipress, 2008.

Yahav Bechavod, Katrina Ligett, Zhiwei Steven Wu, and Juba Ziani. Gaming helps! Learning from strategic interactions in natural dynamics. In *The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1234–1242. PMLR, 2021.

Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive Algorithms and Stochastic Approximations*. Springer, 1990.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Operations Research*, 63(5):1227–1244, 2015.

Léon Bottou. Stochastic learning. In *Advanced Lectures on Machine Learning: ML Summer Schools 2003*, volume 3176 of *Lecture Notes in Artificial Intelligence*, pages 146–168. Springer, 2003.

Léon Bottou. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade*, volume 7700 of *Lecture Notes in Computer Science*, pages 421–436. Springer, 2012.

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

Michael Brückner, Christian Kanzow, and Tobias Scheffer. Static prediction games for adversarial learning problems. *Journal of Machine Learning Research*, 13(1):2617–2654, 2012.

Joshua Cutler, Dmitriy Drusvyatskiy, and Zaid Harchaoui. Stochastic optimization under time drift: iterate averaging, step decay, and high-probability guarantees. In *Advances in Neural Information Processing Systems*, volume 34. Curran Associates, Inc., 2021.

Nilesh N. Dalvi, Pedro M. Domingos, Mausam, Sumit K. Sanghai, and Deepak Verma. Adversarial classification. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 99–108. ACM, 2004.

Amit Daniely, Alon Gonen, and Shai Shalev-Shwartz. Strongly adaptive online learning. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1405–1411. JMLR, 2015.

Dmitriy Drusvyatskiy and Lin Xiao. Stochastic optimization with decision-dependent distributions. *Mathematics of Operations Research*, 2022.

Václav Dupač. A dynamic stochastic approximation method. *The Annals of Mathematical Statistics*, 36(6):1695–1702, 1965.

S. Fujita and T. Fukao. Convergence conditions of dynamic stochastic approximation method for nonlinear stochastic discrete-time dynamic systems. *IEEE Transactions on Automatic Control*, 17(5):715–717, 1972.

A. A. Gaivoronskii. Nonstationary stochastic programming problems. *Cybernetics*, 14(4): 575–579, 1978.

Saeed Ghadimi and Guanghui 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.

Saeed Ghadimi and Guanghui 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.

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

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 111–122. ACM, 2016.

Nicholas J. A. Harvey, Christopher Liaw, Yaniv Plan, and Sikander 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.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15(1): 2489–2512, 2014.

Elad Hazan and C. Seshadhri. Efficient learning algorithms for changing environments. In *Proceedings of the 26th International Conference on Machine Learning*, pages 393–400. ACM, 2009.

Ali Jadbabaie, Alexander Rakhlin, Shahin Shahrampour, and Karthik Sridharan. Online optimization: Competing with dynamic comparators. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, volume 38, pages 398–406. PMLR, 2015.

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M. Kakade, and Michael I. Jordan. A short note on concentration inequalities for random vectors with subgaussian norm. *arXiv:1902.03736*, 2019.

L. V. Kantorovich and G. Sh. Rubinshteĭn. On a space of completely additive functions. *Vestnik Leningradskogo Universiteta. Matematika, Mekhanika, Astronomiya*, 13(2):52–59, 1958.

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

Harold J. Kushner and Gang George Yin. *Stochastic Approximation Algorithms and Applications*, volume 35 of *Applications of Mathematics*. Springer, 1997.

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

Hunter Lang, Lin Xiao, and Pengchuan Zhang. Using statistics to automate stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 32, pages 9536–9546. Curran Associates, Inc., 2019.

Liam Madden, Stephen Becker, and Emiliano Dall'Anese. Bounds for the tracking error of first-order online optimization methods. *Journal of Optimization Theory and Applications*, 189(2):437–457, 2021.

Celestine Mendler-Dünner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. In *Advances in Neural Information Processing Systems*, volume 33, pages 4929–4939. Curran Associates, Inc., 2020.

Aryan Mokhtari, Shahin Shahrampour, Ali Jadbabaie, and Alejandro Ribeiro. Online optimization in dynamic environments: Improved regret rates for strongly convex problems. In *55th IEEE Conference on Decision and Control*, pages 7195–7201. IEEE, 2016.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7599–7609. PMLR, 2020.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning*. Omnipress, 2012.

David Ruppert. A new dynamic stochastic approximation procedure. *The Annals of Statistics*, 7(6):1179–1195, 1979.

Ali H. Sayed. *Fundamentals of Adaptive Filtering*. John Wiley & Sons, 2003.

Suvrit Sra, Sebastian Nowozin, and Stephen J. Wright. *Optimization for Machine Learning*. The MIT Press, 2011.

Nati Srebro, Karthik Sridharan, and Ambuj Tewari. On the universality of online mirror descent. In *Advances in Neural Information Processing Systems*, volume 24, pages 2645–2653. Curran Associates, Inc., 2011.

Ya. Z. Tsypkin and Z. J. Nikolic. *Adaptation and Learning in Automatic Systems*. Elsevier Science, 1971.

Ya. Z. Tsypkin and B. T. Polyak. Optimal recurrent algorithms for identification of nonstationary plants. *Computers and Electrical Engineering*, 18(5):365–371, 1992.

Katsuji Uosaki. Some generalizations of dynamic stochastic approximation processes. *The Annals of Statistics*, 2(5):1042–1048, 1974.

Roman 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.

Craig Wilson, Venugopal V. Veeravalli, and Angelia Nedić. Adaptive sequential stochastic optimization. *IEEE Transactions on Automatic Control*, 64(2):496–509, 2019.

Killian Wood, Gianluca Bianchin, and Emiliano Dall'Anese. Online projected gradient descent for stochastic optimization with decision-dependent distributions. *IEEE Control Systems Letters*, 6:1646–1651, 2022.

Lijun Zhang, Shiyin Lu, and Zhi-Hua Zhou. Adaptive online learning in dynamic environments. In *Advances in Neural Information Processing Systems*, volume 31, pages 1330–1340. Curran Associates, Inc., 2018.

Peng Zhao and Lijun Zhang. Improved analysis for dynamic regret of strongly convex and smooth functions. In *Proceedings of the 3rd Annual Conference on Learning for Dynamics and Control*, volume 144 of *Proceedings of Machine Learning Research*, pages 48–59. PMLR, 2021.

Peng Zhao, Yu-Jie Zhang, Lijun Zhang, and Zhi-Hua Zhou. Dynamic regret of convex and smooth functions. In *Advances in Neural Information Processing Systems*, volume 33, pages 12510–12520. Curran Associates, Inc., 2020.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, pages 928–936. AAAI Press, 2003.