Stochastic Approximation with Dynamic Distributions

Joshua Ross Cutler

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:
Dmitriy Drusvyatskiy, Chair
Zaid Harchaoui
Soumik Pal

Program Authorized to Offer Degree:
Mathematics

University of Washington

Abstract

Stochastic Approximation with Dynamic Distributions

Joshua Ross Cutler

Chair of the Supervisory Committee:
Dmitriy Drusvyatskiy
Department of Mathematics

We consider first 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. In this setting, we provide novel non-asymptotic convergence guarantees for the proximal stochastic gradient method 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 benefits significantly from a step decay schedule. Next, we analyze a stochastic forward-backward method (SFB) for decision-dependent stochastic approximation problems, wherein the data distribution used by the algorithm evolves along the iterate sequence. The primary examples of such problems appear in performative prediction and its multiplayer extensions. We show that under mild assumptions, the deviation between the averaged SFB iterate and the solution is asymptotically normal, with a covariance that clearly decouples the effects of the gradient noise and the distributional dynamics. Moreover, building on the work of Hájek and Le Cam, we show that the asymptotic performance of SFB with averaging is locally minimax optimal.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

The author wishes to express his sincere gratitude to his mentors and collaborators who have made this work possible, especially Dmitriy Drusvyatskiy, Zaid Harchaoui, and Mateo Díaz. The author would also like to thank his family, especially his wife Vanessa, for their continual support and encouragement.

Chapter 1
INTRODUCTION

Stochastic optimization underpins much of machine learning theory and practice for its ability to find a learning rule (e.g., a classifier) from a limited data sample that enables accurate prediction on unseen data. Classical theory crucially relies on the assumption that both the observed data and the unseen data are generated by the same distribution. There is no shortage of problems, however, where the assumption of a fixed data distribution throughout the run of a learning process is grossly violated. There are two main sources of such non-stationary distributions. 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 [27] and is the basis for adaptive algorithms for stochastic tracking [9]. 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. For example, members of the population may alter their features in response to a deployed classifier in order to increase their likelihood of being positively labeled. Even when the population is agnostic to the learning rule, 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 rule that generalizes well under the response distribution. This setting has been a subject of increased interest recently in the context of strategic classification [7, 15, 19, 35] and performative prediction [24, 54, 60]. The situation may be further compounded by a population that reacts to multiple competing learners simultaneously [58, 61, 82].

In the present work, we focus on two related but distinct objectives in the realm of stochastic approximation with dynamic distributions, described in the next two sections.

1.1 Stochastic Optimization under Distributional Drift

The first objective we consider is that of tracking the minimizers or minimum values corresponding to a sequence of stochastic optimization problems

$$\min_{x\in\mathbb{R}^d} \varphi_t(x) := f_t(x) + r_t(x) \tag{1.1}$$

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 [49,69], 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 [38, 83], wherein the true hypothesis may be changing over time.

The main goal of a learning algorithm for problem (1.1) 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 $f_t\colon \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex with L-Lipschitz continuous gradient (L-smooth), or equivalently,

$$\frac{\mu}{2}\|x-y\|^2 \le f_t(y) - f_t(x) - \langle \nabla f_t(x), y-x\rangle \le \frac{L}{2}\|x-y\|^2$$

for all $x, y \in \mathbb{R}^d$ and $t \in \mathbb{N}$. We also impose the standard assumption that each regularizer $r_t : \mathbb{R}^d \to \mathbb{R}\cup\{\infty\}$ is proper ($\operatorname{dom} r_t \neq \emptyset$ and $r_t > -\infty$), closed (lower semicontinuous), and convex. Under these assumptions, it follows that for each $t \in \mathbb{N}$, there exists a unique solution $x_t^\star \in \arg\min_{x\in\mathbb{R}^d} \varphi_t(x)$ to the problem (1.1); the solution $x_t^\star$ is characterized by the first-order optimality condition

$$0 \in \nabla f_t(x_t^\star) + \partial r_t(x_t^\star),$$

where

$$\partial r_t(x_t^\star) = \left\{v \in \mathbb{R}^d \mid r_t(x) \ge r_t(x_t^\star) + \langle v, x - x_t^\star\rangle \ \forall x \in \mathbb{R}^d\right\}$$

denotes the subdifferential of the convex function r_t at $x_t^\star$. We refer the reader to [8] for a review of basic concepts and notation in convex analysis and optimization.

The online proximal stochastic gradient method (PSG) naturally applies to the sequence of problems (1.1). At each iteration t, the method simply takes the step

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

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 Chapter 2, we analyze two types of tracking error for PSG: the squared distance $\|x_t - x_t^\star\|^2$ and the suboptimality gap $\varphi_t(\hat{x}_t) - \varphi_t(x_t^\star)$, where $\hat{x}_t$ denotes a weighted average of iterates up to time t. We next outline the main results of Chapter 2, the contents of which appear in the joint work:

Joshua Cutler, Dmitriy Drusvyatskiy, and Zaid Harchaoui. Stochastic Optimization under Distributional Drift. *Journal of Machine Learning Research*, 24(147):1–56, 2023.

1.1.1 Tracking the Minimizer

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

$$\mathbb{E}\|x_t - x_t^\star\|^2 \lesssim \underbrace{(1-\mu\eta)^t \|x_0 - x_0^\star\|^2}_{\text{optimization}} + \underbrace{\frac{\eta\sigma^2}{\mu}}_{\text{noise}} + \underbrace{\left(\frac{\Delta}{\mu\eta}\right)^2}_{\text{drift}}. \tag{1.2}$$

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 (1.2) 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 in [53, Theorem 3.2]. Though the estimate (1.2) is likely known, we were unable to find a precise reference in this generality.

Letting t tend to infinity in (1.2), 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 this section 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 (1.2) 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 in [33], and can be used here. The end result is an algorithm that produces a point x_t satisfying

$$\mathbb{E}\|x_t - x_t^\star\|^2 \lesssim \mathcal{E} \quad \text{in time} \quad t \lesssim \frac{L}{\mu} \log\left(\frac{\|x_0 - x_0^\star\|^2}{\mathcal{E}} \right) + \frac{\sigma^2}{\mu^2 \mathcal{E}}. \tag{1.3}$$

This efficiency estimate is remarkably similar to that in the static setting [33], with $\mathcal{E}$ playing the role of the target accuracy ε. An elementary computation shows that (1.3) 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 (1.3) 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 (1.3). The proof follows closely the probabilistic techniques developed in [36] for bounding moment generating functions.

1.1.2 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 [41], 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 $\mathbb{E}\|x_t^\star - x_{t+1}^\star\|^2 \leq \Delta^2$.

---

[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.

Analogous to (1.2), 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 (1.3), we show that PSG with step decay produces a point $\hat{x}_t$ satisfying

$$\mathbb{E}\big[\varphi_t(\hat{x}_t) - \varphi_t^\star\big] \lesssim \mathcal{G} \quad \text{in time} \quad t \lesssim \frac{L}{\mu}\log\left(\frac{\varphi_0(x_0) - \varphi_0^\star}{\mathcal{G}}\right) + \frac{\sigma^2}{\mu\mathcal{G}}. \tag{1.4}$$

Again, the similarity to the static setting [33], 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 (1.4) up to a factor of $\log\log(1/\delta)$. The proofs are based on the generalized Freedman inequality developed recently in [36]—a remarkably flexible tool for analyzing stochastic gradient-type algorithms.

1.1.3 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. Combining time-dependence and decision-dependence yields a class of problems (1.1) where the loss function $f_t(x)$ takes the special form $f_t(x) = \mathbb{E}_{\xi\sim\mathcal{D}(t,x)}\ell(x,\xi)$, where $\mathcal{D}(t,x)$ is a distribution that depends on both time t and the decision variable x. Thus, for any fixed time t, the problem (1.1) captures the performative risk problem considered in [54, 60]. 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

that minimize the objective they induce, that is,

$$\bar{x}_t \in \arg\min_x \mathop{\mathbb{E}}_{\xi\sim\mathcal{D}(t,\bar{x}_t)} \ell(x,\xi) + r_t(x).$$

Equilibrium points are sure to exist and are unique under mild Lipschitzness and strong convexity assumptions. We refer the reader to [60] for a compelling motivation for considering this notion of equilibrium. The problem of tracking equilibrium points is yet again an instance of (1.1), 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 $x_{t+1} = \mathrm{prox}_{\eta_t r}(x_t - \eta_t\nabla\ell(x_t,\xi_t))$. Notice that the sample gradient $\nabla\ell(x_t,\xi_t)$ is typically 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 $\mathcal{D}(t,x_t)$ instead of $\mathcal{D}(t,\bar{x}_t)$. Nonetheless, as pointed out in [24], under mild assumptions 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 of our 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.

1.1.4 Related Work

Significant progress has been made over the last two decades in the finite-time analysis of stochastic approximation algorithms [1,4,11–13,51,59,70,71]. Our current work fits within the 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 [49] for a survey. Early results can be traced back to [27] in sequential estimation and [31] in stochastic

optimization; see also [30, 68, 72–74]. Stochastic algorithms have also been extensively studied as adaptive algorithms for stochastic tracking [9, 49, 72], 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 [34].

Recent works have revisited these methods from a more modern viewpoint [10, 53, 80]. In particular, the paper [53] focuses on (accelerated) gradient methods for deterministic tracking problems, while [80] presents a framework for online stochastic gradient methods with parameter estimation. The work [10] 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 work share many aspects with those available in the literature, the results we obtain are distinct. In particular, the guarantees (1.3) and (1.4) 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 [20, 38] and dynamic regret [10, 41, 57, 83, 85, 86]. 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

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

In the paper [41], an adaptive step size strategy is applied to an optimistic mirror descent algorithm to obtain 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, it is shown in [57] that online projected gradient descent satisfies $\mathrm{Reg}_T^\star \leq O(1 + C_T^\star)$. For convex losses with bounded domain $\mathcal{X}$, an adaptive online gradient method is developed in [83] 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, $\mathrm{Reg}_T(u_1, \ldots, u_T) \leq O\big(\sqrt{T(1 + C_T)}\big)$. This last guarantee is enhanced in [85] through exploiting smoothness to replace the time horizon T by problem-dependent quantities that are at most $O(T)$ 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 works [5, 37, 50, 63]. There exists a variety of techniques for establishing high-probability guarantees based on Freedman's inequality and doubling tricks [5, 37]. A more recent line of work [36] establishes a generalized Freedman inequality that is tailored to analyzing stochastic gradient-type methods and results in the best known high-probability guarantees. Our arguments closely follow the paradigm in [36] based on the generalized Freedman inequality.

Performative prediction and decision-dependent learning. Recent works on strategic classification [7, 15, 19, 35] and performative prediction [54, 60] 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 [60] puts forth an elegant framework for thinking about such problems, while [54] develops stochastic algorithms for this setting. The subsequent work [24] 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 in [24], 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 the present work, the closely related and complementary paper [81] was posted on arXiv.[2] The paper [81] considers decision-dependent projected stochastic gradient descent under time drift in the distributional framework proposed in [60], 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 in [81] for obtaining high-probability guarantees is more general than the one we use. On the other hand, we analyze tracking of both the

[2]More precisely, a short version of our work [16] was submitted to NeurIPS in May '21, the paper [81] appeared on arXiv in July '21, and our full paper was posted on arXiv in August '21. Our work [16] was presented at NeurIPS in December '21.

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.

1.2 Stochastic Approximation with Decision-Dependent Distributions: Asymptotic Normality and Optimality

The second objective we consider focuses instead on the asymptotic performance of estimation procedures for finding the equilibrium point of a decision-dependent stochastic approximation problem. In Chapter 3, we model decision-dependent problems using variational inequalities. Namely, let $G(x, z)$ be a map that depends on the decision x and data z, and let the set $\mathcal{X}$ of feasible decisions be closed and convex. A variety of classical learning problems can be posed as solving the variational inequality

$$0 \in \mathop{\mathbb{E}}_{z\sim\mathcal{P}} G(x, z) + N_{\mathcal{X}}(x), \qquad \text{VI}(\mathcal{P})$$

where $\mathcal{P}$ is some fixed distribution and

$$N_{\mathcal{X}}(x) = \{v \in \mathbb{R}^d \mid \langle v, y - x\rangle \leq 0 \ \forall y \in \mathcal{X}\}$$

is the normal cone to $\mathcal{X}$ at $x \in \mathcal{X}$. Two examples are worth keeping in mind: (i) standard problems of supervised learning amount to $G(x, z) = \nabla_x \ell(x, z)$ being the gradient of some loss function to be minimized over $\mathcal{X}$, and (ii) stochastic games correspond to $G(x, z)$ being a stacked gradient of the players' individual losses. In these examples, VI($\mathcal{P}$) reduces to standard first-order optimality conditions.

Following the recent literature on performative prediction [35, 58, 60], we will be interested in settings where the distribution $\mathcal{P}$ is not fixed but rather varies with x. With this in mind, let $\mathcal{D}(x)$ be a family of distributions indexed by $x \in \mathcal{X}$. The interpretation is that $\mathcal{D}(x)$ is the response of the population to a newly deployed learning rule x. We posit that the goal of a learning system is to find a point $x^\star$ so that $x = x^\star$ solves the variational inequality VI($\mathcal{D}(x^\star)$), or equivalently:

$$0 \in \mathop{\mathbb{E}}_{z\sim\mathcal{D}(x^\star)} G(x^\star, z) + N_{\mathcal{X}}(x^\star).$$

We will say that such points $x^\star$ are at *equilibrium*. In words, a learning system that deploys an equilibrium point $x^\star$ has no incentive to deviate from $x^\star$ based only on the solution of the variational inequality $\mathrm{VI}(\mathcal{D}(x^\star))$ induced by the response distribution $\mathcal{D}(x^\star)$. The setting of performative prediction [60] corresponds to the choice $G(x,z) = \nabla_x \ell(x,z)$ for some loss function ℓ.[3] More generally, decision-dependent games, proposed in [58, 61, 82], correspond to the choice $G(x,z) = \big(\nabla_1 \ell_1(x,z), \dots, \nabla_k \ell_k(x,z)\big)$ where $\nabla_i \ell_i(x,z)$ is the gradient of the i'th player's loss with respect to their decision x_i and $\mathcal{D}(x) = \mathcal{D}_1(x) \times \cdots \times \mathcal{D}_k(x)$ is a product distribution. The specifics of these two examples will not affect our results, and therefore we work with general maps $G(x,z)$.

Following the prevalent viewpoint in machine learning, we suppose that the only access to the data distributions $\mathcal{D}(x)$ is by drawing samples $z \sim \mathcal{D}(x)$. With this in mind, a natural algorithm for finding an equilibrium point $x^\star$ is the *stochastic forward-backward algorithm*:

$$\begin{aligned} &\text{Sample } z_t \sim \mathcal{D}(x_t) \\ &\text{Set } x_{t+1} = \mathrm{proj}_{\mathcal{X}}\big(x_t - \eta_t G(x_t, z_t)\big), \end{aligned} \tag{SFB}$$

where $\mathrm{proj}_{\mathcal{X}}$ is the nearest-point projection onto $\mathcal{X}$. Specializing to performative prediction [54] and its multiplayer extension [58], this algorithm reduces to a basic projected stochastic gradient iteration. The primary contribution of Chapter 3 can be informally summarized as follows.

We show that averaged SFB is asymptotically optimal for finding equilibrium points.

In particular, our results imply asymptotic optimality of the basic stochastic gradient methods (with averaging) for both single player and multiplayer performative prediction. Let us now outline the main results of Chapter 3, the contents of which appear in the joint work:

> Joshua Cutler, Mateo Díaz, and Dmitriy Drusvyatskiy. Stochastic approximation with decision-dependent distributions: asymptotic normality and optimality. *arXiv:2207.04173*, 2022.

[3] In the language of [60], equilibria coincide with performatively stable points.

1.2.1 Summary of Main Results

Arguing optimality of an algorithm is a two-step process: (i) estimate the performance of the specific algorithm and (ii) derive a matching lower bound that is valid among all relevant estimation procedures. Beginning with the former, we build on the seminal work [62], wherein a central limit theorem is established for stochastic approximation algorithms for solving smooth equations. Letting $\bar{x}_t = \frac{1}{t}\sum_{i=1}^t x_i$ denote the running average of the SFB iterates, we show that the deviation $\sqrt{t}(\bar{x}_t - x^\star)$ is asymptotically normal with an appealingly simple covariance. See Figure 1.1 for an illustration.[4]

Theorem 1.1 (Asymptotic normality, informal; see Theorem 3.5)**.** *Suppose that $G(\cdot, z)$ is α-strongly monotone and Lipschitz continuous on $\mathcal{X}$, $G(x, \cdot)$ is β-Lipschitz continuous on $\mathcal{Z}$, and the distribution map $\mathcal{D}(\cdot)$ is γ-Lipschitz on $\mathcal{X}$ with respect to the Wasserstein-1 distance. Suppose moreover that $x^\star$ lies in the interior of $\mathcal{X}$ and $\eta_t \propto t^{-\nu}$ for some $\nu \in \left(\frac{1}{2}, 1\right)$. Then in the regime $\frac{\gamma\beta}{\alpha} < 1$, the SFB iterates x_t converge to the equilibrium point $x^\star$ almost surely, and the averaged SFB iterates $\bar{x}_t = \frac{1}{t}\sum_{i=1}^t x_i$ satisfy*

$$\sqrt{t}(\bar{x}_t - x^\star) \rightsquigarrow \mathsf{N}\big(0, W^{-1}\Sigma W^{-\top}\big),$$

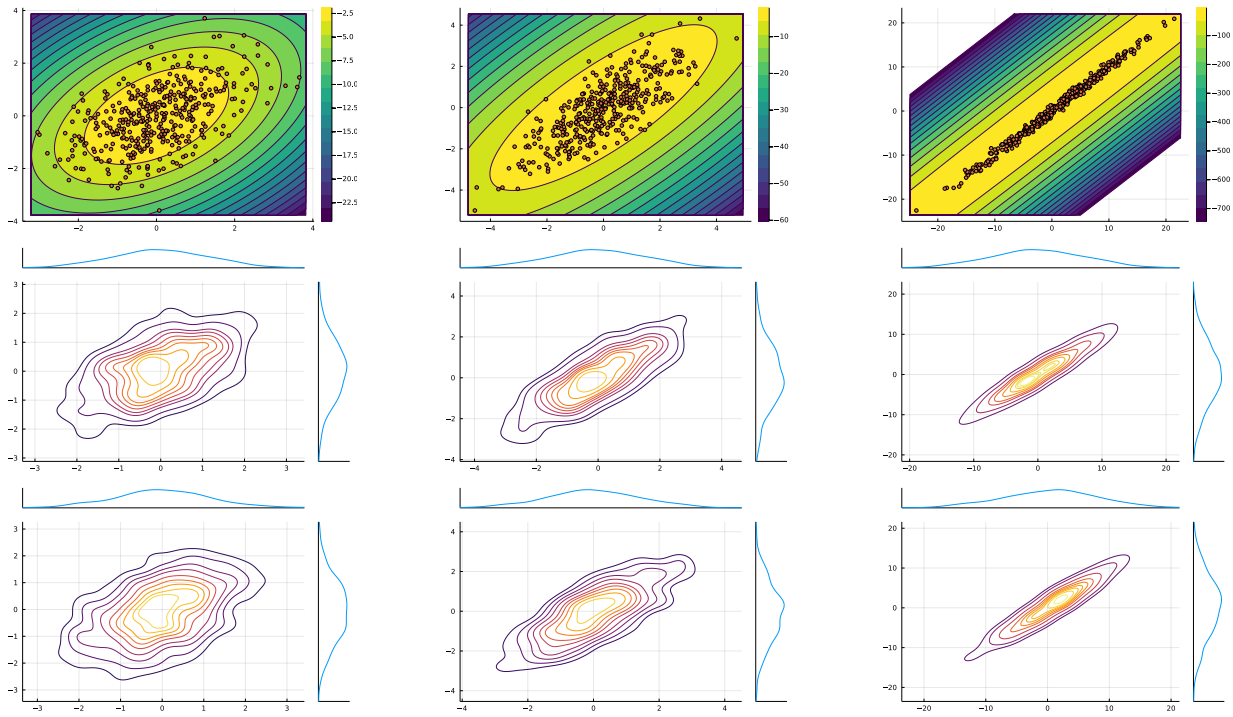
where

$$\Sigma = \mathop{\mathbb{E}}_{z\sim\mathcal{D}(x^\star)}\big[G(x^\star, z)G(x^\star, z)^\top\big] \quad \textit{and} \quad W = \underbrace{\mathop{\mathbb{E}}_{z\sim\mathcal{D}(x^\star)}[\nabla_x G(x^\star, z)]}_{\textit{static}} + \underbrace{\frac{d}{dy}\mathop{\mathbb{E}}_{z\sim\mathcal{D}(y)}[G(x^\star, z)]\Big|_{y=x^\star}}_{\textit{dynamic}}.$$

A few comments are in order. First, the regime $\frac{\gamma\beta}{\alpha} < 1$ is, in essence, optimal because otherwise equilibrium points may even fail to exist. Second, the effect of the distributional shift on the asymptotic covariance is entirely captured by the second "dynamic" term in W. Indeed, when this term is absent, the product $W^{-1}\Sigma W^{-\top}$ is precisely the asymptotic covariance of the stochastic forward-backward algorithm applied to the static problem $\mathrm{VI}(\mathcal{D}(x^\star))$ at equilibrium.[5] The proof of Theorem 1.1 follows by interpreting SFB as a stochastic

[4] Code is available online at https://github.com/mateodd25/Asymptotic-normality-in-performative-prediction.

[5] Of course, this analogy is entirely conceptual, since $\mathcal{D}(x^\star)$ is unknown a priori.



(a) $\rho = 0.25$ (b) $\rho = 0.5$ (c) $\rho = 0.9$

Figure 1.1: Consider the problem corresponding to $G(x, z) = \nabla_x \ell(x, z)$ with $\ell(x, z) = \frac{1}{2}\|x - z\|^2$ and $\mathcal{D}(x_1, x_2) = \mathsf{N}(\rho(x_2, x_1), I_2)$. A simple computation shows $\Sigma = I_2$ and $W = [1, -\rho; -\rho, 1]$. As ρ approaches one, W^{-1} becomes ill conditioned. We run SFB 400 times using $\eta_t = t^{-3/4}$ for 10^6 iterations. The first row depicts the resulting average iterates laid over the confidence regions (plotted in logarithmic scale) corresponding to the asymptotic normal distribution. The next two rows depict kernel density estimates from the asymptotic normal distribution (top) and the deviation $\sqrt{t}(\bar{x}_t - x^\star)$ (bottom).

approximation algorithm for finding the zero of the nonlinear map $R(x) = \mathbb{E}_{z\sim\mathcal{D}(x)} G(x, z)$ and then applying a variation of the classical asymptotic normality result [62, Theorem 2].

A reasonable question to ask is whether there exists an algorithm with better asymptotic guarantees than those of the stochastic forward-backward algorithm (with averaging). We

will show that in a strong sense, the answer is no; that is, averaged SFB is asymptotically optimal. In particular, we will obtain an optimal bound on the performance of *any* estimation procedure for finding the equilibrium point along an adversarially-chosen sequence of small perturbations of the target problem. More concretely, we will carefully construct for each $u \in \mathbb{R}^d$ a perturbation $\mathcal{D}^u$ of $\mathcal{D}$ such that, as $u \to 0$, the distribution map $\mathcal{D}^u$ induces a perturbed problem with equilibrium point $x^\star_u$ near $x^\star$. Using this family of perturbed problems parameterized by $u \in \mathbb{R}^d$, we will show that if $\widehat{x}_k \colon \mathcal{Z}^k \to \mathbb{R}^d$ is an arbitrary sequence of estimators (i.e., $\widehat{x}_k$ is a measurable function of k observed samples) and $\mathcal{L} \colon \mathbb{R}^d \to [0, \infty)$ is any loss functional that is "bowl-shaped" (symmetric and quasiconvex) and lower semicontinuous, then the following local asymptotic minimax bound holds:

$$\sup_{\mathcal{I} \subset \mathbb{R}^d,\, |\mathcal{I}| < \infty} \liminf_{k \to \infty} \max_{u \in \mathcal{I}} \mathbb{E}_{P_{k,u/\sqrt{k}}}\big[\mathcal{L}\big(\sqrt{k}\big(\widehat{x}_k - x^\star_{u/\sqrt{k}}\big)\big)\big] \geq \mathbb{E}[\mathcal{L}(Z)], \tag{1.5}$$

where $P_{k,v}$ denotes the distribution on $\mathcal{Z}^k$ induced by $\mathcal{D}^v$ along an arbitrary "dynamic estimation procedure" and $Z \sim \mathsf{N}(0, W^{-1}\Sigma W^{-\top})$ with Σ and W as in Theorem 1.1. Moreover, we will show that equality is achieved in (1.5) upon specializing to the dynamic estimation procedure corresponding to SFB with step sizes $\eta_k \propto k^{-\nu}$ (as in Theorem 1.1) and taking $\widehat{x}_k$ to be given by the averaged SFB iterate $\bar{x}_k = \frac{1}{k}\sum_{i=1}^k x_i$, provided $\mathcal{L}$ is bounded and continuous. The end result is summarized informally in the following theorem.

Theorem 1.2 (Asymptotic optimality, informal; see Theorem 3.16)**.** *Suppose the same setting as in Theorem 1.1 and let $\mathcal{L} \colon \mathbb{R}^d \to [0, \infty)$ be any bowl-shaped lower semicontinuous loss functional. Fix any procedure for finding equilibrium points that outputs an estimator $\widehat{x}_k$ based on k observed samples. As $k \to \infty$, there is a sequence of perturbed distribution maps $\mathcal{D}_k$ converging to $\mathcal{D}$, along with corresponding equilibrium points $x^\star_k$ converging to $x^\star$, such that the expected error $\mathbb{E}[\mathcal{L}(\sqrt{k}(\widehat{x}_k - x^\star_k))]$ of the estimator $\widehat{x}_k$ on the perturbed problem is asymptotically lower-bounded by $\mathbb{E}[\mathcal{L}(Z)]$, where $Z \sim \mathsf{N}(0, W^{-1}\Sigma W^{-\top})$. Moreover, if $\mathcal{L}$ is bounded and continuous, then this lower bound is achieved by the estimator given by the averaged SFB iterate $\bar{x}_k = \frac{1}{k}\sum_{i=1}^k x_i$.*

The formal statement of Theorem 1.2 and its proof follow closely the classical work of Hájek and Le Cam [52, 75] on statistical lower bounds and the more recent work [25] on asymptotic optimality of the stochastic gradient method. In particular, the fundamental role of tilt-stability and the inverse function theorem highlighted in [25] is replaced by the implicit function theorem paradigm.

1.2.2 Related Work

Our work builds on existing literature in machine learning and stochastic optimization.

Learning with decision-dependent distributions. The basic setup for decision-dependent problems that we use is inspired by the performative prediction framework introduced in [60] and its multiplayer extension developed independently in [58, 61, 82]. The stochastic gradient method for performative prediction was first introduced and analyzed in [54], while the stochastic forward-backward method for games was analyzed in [58]. The related work [24] showed that a variety of popular gradient-based algorithms for performative prediction can be understood as the analogous algorithms applied to a certain static problem corrupted by a vanishing bias. In general, performatively stable points (equilibria) are not "performatively optimal" in the sense of [60]. Seeking to develop algorithms for finding performatively optimal points, the paper [56] provides sufficient conditions for the prediction problem to be convex; extensions of such conditions to games appear in [58] and [82]. Algorithms for finding performatively optimal points under a variety of different assumptions and oracle models appear in [40, 42, 56, 58, 82]. The performative prediction framework is largely motivated by the problem of strategic classification [35], which has been studied extensively from the perspective of causal inference [7, 55] and convex optimization [23]. Other lines of work [14, 17, 64, 81] in performative prediction have focused on the setting in which the environment evolves dynamically in time.

Stochastic approximation. There is extensive literature on stochastic approximation. The most relevant results for us are those in [62] that quantify the limiting distribution of the average iterate of stochastic approximation algorithms. Stochastic optimization problems with decision-dependent uncertainties have appeared in the classical stochastic programming literature [2, 28, 44, 67, 77]. We refer the reader to the recent paper [39], which discusses taxonomy and various models of decision-dependent uncertainties. An important theme of these works is to utilize structural assumptions on how the decision variables impact the distributions. In contrast, much of the work on performative prediction [24, 54, 58, 60, 61, 82] and our current work are "model-free".

Local minimax lower bounds in estimation. There is a rich literature on minimax lower bounds in statistical estimation problems; we refer the reader to [79, Chapter 15] for a detailed treatment. Typical results of this type lower-bound the performance of any statistical procedure on a worst-case instance of that procedure. Minimax lower bounds can be quite loose as they do not consider the complexity of the particular problem that one is trying to solve but rather that of an entire problem class to which it belongs. More precise local minimax lower bounds, as developed by Hájek and Le Cam [52, 75], provide much finer problem-specific guarantees. Building on this framework, [25] showed that the stochastic gradient method for standard single-stage stochastic optimization problems is, in an appropriate sense, locally asymptotically minimax optimal. We build heavily on this line of work.

Chapter 2

STOCHASTIC OPTIMIZATION UNDER DISTRIBUTIONAL DRIFT

Joint work with D. Drusvyatskiy and Z. Harchaoui [17]

In this chapter, we present finite-time efficiency estimates in expectation and with high probability for the tracking error of the proximal stochastic gradient method under time drift. Our results concisely explain the interplay between the learning rate, the noise variance in the gradient oracle, and the strength of the time drift. 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".

The outline of the chapter is as follows. Section 2.1 formalizes the problem setting of time-dependent stochastic optimization and records the relevant assumptions. Sections 2.2–2.4 summarize the main results of the chapter. Specifically, Section 2.2 focuses on efficiency estimates for tracking the minimizer, Section 2.3 focuses on efficiency estimates for tracking the minimum value, and Section 2.4 develops an extension to the decision-dependent setting via tracking equilibria. Section 2.5 presents the proofs of the main results in a unified framework. Illustrative numerical results appear in Section 2.6. Appendix A describes the averaging technique used for tracking function values, and additional proofs appear in Appendix B.

2.1 Framework and Assumptions

Throughout Sections 2.1–2.3, we consider the sequence of stochastic optimization problems

$$\min_{x\in\mathbb{R}^d} \varphi_t(x) := f_t(x) + r_t(x) \tag{2.1}$$

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 $\mu, L > 0$.

(ii) Each regularizer $r_t \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is proper, closed, and convex.[1]

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 (2.1) that are worth keeping in mind and that guide our framework: stochastic tracking of a drifting target and online learning under distributional drift.

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

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

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

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

where 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 [34, 69]. ◊

[1] 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.

Example 2.2 (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. $\lozenge$

The main goal of a learning algorithm for problem (2.1) is to generate a sequence of points $\{x_t\}$ that minimize some natural performance metric. The most prevalent performance metrics in the literature are the *tracking error* and the *dynamic regret*. We will focus on two types of tracking error: 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

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

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.

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

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

Step $t = 0, \ldots, T-1$:

$$\text{Select } g_t = \widetilde{\nabla} f_t(x_t)$$

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

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 2.1), or when tracking the minimizer of an expected loss over a stochastically evolving data distribution (Example 2.2). 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 2.1–2.3.

Assumption 2.1 (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 2.1 formalizes the assertion that x_t and $x_t^\star$ are fully determined by information up to time t. The second item of Assumption 2.1 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 2.2 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 (2.1) 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.[2]

Lemma 2.3 (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 $\mu\|x_i^\star - x_t^\star\| \le \|0 - v\|$. □

2.2 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.

2.2.1 Bounds in Expectation

We begin with bounding the expected value $\mathbb{E}\|x_t - x_t^\star\|^2$. Proofs appear in Section 2.5.1. The starting point for our analysis is the following standard one-step improvement guarantee.

---

[2] Lemma 2.3 provides a bound similar in spirit to the bound $\mu\|x_i^\star - x_t^\star\|^2 \le 4\sup_{x\in\operatorname{dom} r}|f_i(x) - f_t(x)|$ in terms of variation in function value, which is also an elementary consequence of μ-strong convexity [84, Section 4.1].

Lemma 2.4 (One-step improvement)**.** *For all $x \in \mathbb{R}^d$, the iterates $\{x_t\}$ produced by Algorithm 1 with $\eta_t < 1/L$ satisfy the bound:*

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

For simplicity, we state the main results under the assumption that the second moments $\mathbb{E}\,\Delta_t^2$ 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 2.4 as well.

Assumption 2.2 (Bounded second moments)**.** There exist constants $\Delta, \sigma > 0$ such that the following two conditions hold for all $t \geq 0$:

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

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

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

Theorem 2.5 (Expected distance)**.** *Suppose that Assumption 2.2 holds. Then the iterates produced by Algorithm 1 with constant learning rate $\eta \leq 1/2L$ satisfy the bound:*

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

Interplay of optimization, noise, and drift. Theorem 2.5 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 [53, Theorem 3.2].

With Theorem 2.5 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 2.5 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 2.5 yields the estimate $\mathbb{E}\|x_t - x_t^\star\|^2 \lesssim \mathcal{E}$ in time $t \lesssim (\sigma^2/\mu^2\mathcal{E}) \log(\|x_0 - x_0^\star\|^2/\mathcal{E})$. This efficiency estimate can be significantly improved by gradually decaying the learning rate using a "step decay schedule", wherein the algorithm

is implemented in epochs with the new learning rate chosen to be the midpoint between the current learning rate and $\eta_\star$. Such schedules are well known to improve efficiency in the static setting, as was discovered in [33], and can be used here. The end result is the following theorem; see Theorem 2.32 for the formal statement.

Theorem 2.6 (Time to track in expectation, informal)**.** *Suppose that Assumption 2.2 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 2.6 is strikingly similar to the efficiency estimate in the static setting [33], with $\mathcal{E}$ playing the role of the target accuracy ε. An elementary computation shows that in the low drift-to-noise regime, Theorem 2.6 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 2.5 and 2.6 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.

2.2.2 High-Probability Guarantees

We next present high-probability guarantees for the tracking error $\|x_t - x_t^\star\|^2$. Proofs appear in Section 2.5.2. We make the following standard light-tail assumptions on the minimizer drift and gradient noise [36, 50, 59].

Assumption 2.3 (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 2.3 is equivalent to asserting that the minimizer drift Δ_t is sub-Gaussian conditioned on $\mathcal{F}_t$ [78, Lemma 2.7.6]. Clearly Assumption 2.3 implies Assumption 2.2 with the same constants Δ, σ. It is worthwhile to note some common settings in which Assumption 2.3 holds; the claims in Remark 2.7 below follow from standard results on sub-Gaussian random variables [43, 78].

Remark 2.7 (Common settings for Assumption 2.3)**.** 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 2.3 holds, then the expected bound on $\|x_t - x_t^\star\|^2$ derived in Theorem 2.5 holds with high probability.

Theorem 2.8 (High-probability distance tracking)**.** *Let $\{x_t\}$ be the iterates produced by Algorithm 1 with constant learning rate $\eta \le 1/2L$, and suppose that Assumption 2.3 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 $1-\delta$:*

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

The proof of Theorem 2.8 employs a technique used by [36]. 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 2.8, we can again implement a step decay schedule to obtain the following efficiency estimate with high probability; see Theorem 2.35 for the formal statement.

Theorem 2.9 (Time to track with high probability, informal)**.** *Suppose that Assumption 2.3 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*

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

2.3 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 2.5.3, where we apply a standard averaging technique (Appendix A) to a one-step improvement along x_t (Lemma 2.36) to obtain the desired progress along $\hat{x}_t$ (Proposition 2.37).

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

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

Step $t = 0, \dots, T-1$:

$$\text{Select } g_t = \widetilde{\nabla} f_t(x_t)$$

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

$$\text{Set } \hat{x}_{t+1} = \left(1 - \tfrac{\mu\eta_t}{2-\mu\eta_t}\right)\hat{x}_t + \tfrac{\mu\eta_t}{2-\mu\eta_t} x_{t+1}$$

Return $\hat{x}_T$

2.3.1 Bounds in Expectation

We begin with bounding the expected value $\mathbb{E}[\varphi_t(\hat{x}_t) - \varphi_t^\star]$. Proofs appear in Section 2.5.3. Analogous to Assumption 2.2, we make the following assumption regarding drift and noise.

Assumption 2.4 (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 2.3, it is clear that Assumption 2.4 implies the earlier Assumption 2.2 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 2.2, 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 [45] 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 2.2.

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

Theorem 2.10 (Expected function gap)**.** *Let $\{\hat{x}_t\}$ be the iterates produced by Algorithm 2 with constant learning rate $\eta \leq 1/2L$, and suppose that Assumption 2.4 holds. Then the following bound holds for all $t \geq 0$:*

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

The "noise + drift" error term in Theorem 2.10 coincides with μ times the error term in Theorem 2.5, as expected. With Theorem 2.10 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:

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

In the high drift-to-noise regime $\Delta/\sigma \geq \sqrt{\mu/16L^3}$, Theorem 2.10 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 2.10 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 2.39 for the formal statement.

Theorem 2.11 (Time to track in expectation, informal)**.** *Suppose that Assumption 2.4 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 2.11 improves the constant learning rate efficiency estimate when $\mathcal{G}$ is small, e.g., when $\mathcal{G} \le (\varphi_0(x_0) - \varphi_0^\star)/e^2$.

2.3.2 High-Probability Guarantees

Next, we obtain high-probability analogues of Theorems 2.10 and 2.11. Proofs appear in Section 2.5.4. 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 2.5 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 $0 \le i < t$; 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 2.1 and 2.2.

Assumption 2.5 (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 $0 \le i < t$:

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

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

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

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

Clearly the chain of implications holds:

$$\text{Assumption 2.5} \implies \text{Assumption 2.4} \implies \text{Assumption 2.2}.$$

The following theorem shows that if Assumption 2.5 holds, then the expected bound on $\varphi_t(\hat{x}_t) - \varphi_t^\star$ derived in Theorem 2.10 holds with high probability.

Theorem 2.12 (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 2.5 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 $1 - \delta$:*

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

The proof of Theorem 2.12 is based on combining the generalized Freedman inequality of [36] 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 $\sum_{i=0}^n \langle z_i, x_i - x_t^\star\rangle \zeta^{t-1-i}$, 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 2.12 in hand, we may implement a step decay schedule as before to obtain the following efficiency estimate; see Theorem 2.45 for the formal statement.

Theorem 2.13 (Time to track with high probability, informal)**.** *Suppose that Assumption 2.5 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*

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

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

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

2.4 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 in [60] 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 (2.1) 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

$$\min_{x \in \mathbb{R}^d} \; f_{t,x}(x) + r_t(x) \tag{2.2}$$

indexed by t. Tracking the solutions to (2.2) 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

$$\min_{u \in \mathbb{R}^d} \; f_{t,x}(u) + r_t(u) \tag{2.3}$$

indexed by both t and x. Instead of tracking the *optimal* decisions solving (2.2), our aim becomes to track the *stable* decisions

$$\bar{x}_t \in \operatorname*{arg\,min}_{u \in \mathbb{R}^d} \; f_{t,\bar{x}_t}(u) + r_t(u) \tag{2.4}$$

arising from (2.3). We call a point $\bar{x}_t$ satisfying (2.4) an *equilibrium point* of (2.3) at time t; observe that $\bar{x}_t$ is stable in the sense that it is a fixed point of the map $x \mapsto \operatorname{arg\,min}_u\{f_{t,x}(u) + r_t(u)\}$.

Reasonable regularity assumptions on the family $\{f_{t,x}(\cdot)\}$ ensure that (2.3) admits a

unique equilibrium point at each time t (see Assumption 2.6 and Lemma 2.15 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

$$\min_{u \in \mathbb{R}^d} f_{t,\bar{x}_t}(u) + r_t(u). \tag{2.5}$$

Formally, (2.5) is an example of (2.1), 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 in [60]. This will be a running example throughout the section.

Example 2.14 (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 those in [60] 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:

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

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 $\beta \ge 0$ 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 [24, Lemma 2.1]:

$$\sup_{u\in\mathbb{R}^d} \|\nabla f_{i,x}(u) - \nabla f_{t,y}(u)\| \le \theta\beta|i-t| + \varepsilon\beta\|x-y\| \quad \text{for all} \quad (i,x),(t,y) \in \mathbb{N}\times\mathbb{R}^d. \tag{2.6}$$

Suppose now that each expected loss $f_{t,x}(\cdot)$ is μ-strongly convex. In this case, [60] identifies the optimal parameter regime $\varepsilon\beta < \mu$ wherein the repeated minimization procedure $y_{k+1} = \arg\min_u\{f_{t,y_k}(u) + r_t(u)\}$ converges to the unique equilibrium point $\bar{x}_t$ at a linear rate as $k \to \infty$ [60, Theorem 3.5 and Proposition 3.6]. The gradient stability property (2.6) will lead us to the corresponding parameter regime in our generalized setting. ◊

2.4.1 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 2.4 the following stability property of the gradients with respect to the decision variable.

Assumption 2.6 (Gradient stability in the decision variable)**.** The following parameter regime holds: $\gamma < \mu$.

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

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

Returning to Example 2.14, it follows from (2.6) that Assumption 2.6 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 2.15 (Existence of equilibrium)**.** *For each $t \in \mathbb{N}$, the map $S_t \colon \mathbb{R}^d \to \mathbb{R}^d$ given by*

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

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

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

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

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

where the last inequality holds by the definition of γ. Hence $\|S_t(x) - S_t(y)\| \le (\gamma/\mu)\|x - y\|$, 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 2.16 (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

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

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

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

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

Next, we turn to tracking the equilibria $\bar{x}_t$ furnished by Lemma 2.15 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

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

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 `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 2.4.

Assumption 2.7 (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 2.7 formalizes the assertion that x_t and $\bar{x}_t$ are fully determined by information up to time t. The second item of Assumption 2.7 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 2.14 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 2.14, the estimate (2.6) implies $\bar{G}_{i,t} \le \theta\beta|i-t|$ and hence $\bar{\Delta}_t \le \theta\beta/\bar{\mu}$ by Lemma 2.3, provided the regularizers $r_t \equiv r$ are identical for all times t. We also set

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

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 2.1 with $x_t^\star = \bar{x}_t$.

2.4.2 Tracking the Equilibrium Point

In a nutshell, the results of Section 2.2 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 2.8 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 2.17 (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 2.4 and 2.17, we obtain the following equilibrium one-step improvement.

Lemma 2.18 (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 2.17, 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 2.4 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 2.18 as well.

Assumption 2.8 (Bounded second moments)**.** There exist constants $\bar{\Delta}, \sigma > 0$ such that the following two conditions hold for all $t \ge 0$:

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

(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 3, thereby extending Theorem 2.5; see Section 2.5.1 for the precise statement (Corollary 2.31) and proof.

Theorem 2.19 (Expected distance)**.** *Suppose that Assumption 2.8 holds. Then the iterates produced by Algorithm 3 with constant learning rate $\eta \le 1/2L$ 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 2.19 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 $\bar{\Delta}/\sigma$:

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

As before, the *high drift-to-noise regime* $\bar{\Delta}/\sigma \ge \sqrt{\bar{\mu}/16L^3}$ 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 2.6; see Theorem 2.32 for the formal statement and proof.

Theorem 2.20 (Time to track in expectation, informal)**.** *Suppose that Assumption 2.8 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 2.9 (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 2.9 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 2.14 with $\bar{\Delta} = \theta\beta/\bar{\mu}$ provided the regularizers $r_t \equiv r$ are identical for all times t. Clearly Assumption 2.9 implies Assumption 2.8 with the same constants $\bar{\Delta}, \sigma$. The following theorem shows that if Assumption 2.9 holds, then the expected bound on $\|x_t - \bar{x}_t\|^2$ derived in Theorem 2.19 holds with high probability.

Theorem 2.21 (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 2.9 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 $1 - \delta$:*

$$\|x_t - \bar{x}_t\|^2 \leq \left(1 - \tfrac{\bar{\mu}\eta}{2}\right)^t \|x_0 - \bar{x}_0\|^2 + c\left(\frac{\eta\sigma^2}{\bar{\mu}} + \left(\frac{\bar{\Delta}}{\bar{\mu}\eta}\right)^2\right)\log\left(\frac{e}{\delta}\right). \tag{2.7}$$

Theorem 2.21 is an extension of Theorem 2.8. As a consequence of Theorem 2.21, 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 2.9; see Section 2.5.2 for the precise statements (Theorems 2.34 and 2.35) and proofs.

Theorem 2.22 (Time to track with high probability, informal)**.** *Suppose that Assumption 2.9 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*

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

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

$$t \lesssim \frac{L}{\bar{\mu}} \log\left(\frac{\|x_0 - \bar{x}_0\|^2}{\bar{\mathcal{E}}}\right) + \frac{\sigma^2}{\bar{\mu}^2 \bar{\mathcal{E}}}.$$

2.4.3 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

$$\gamma < \mu/2. \tag{2.8}$$

The regime (2.8) matches the one used in Theorem 7.3 of [24] 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 (2.8), 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 2.3 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 2.10 below).

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 2.4.

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

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 $t = 0, \ldots, T-1$:

$$\text{Select } g_t = \widetilde{\nabla} f_{t,x_t}(x_t)$$

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

$$\text{Set } \hat{x}_{t+1} = \left(1 - \tfrac{\hat{\mu}\eta_t}{2-\mu\eta_t}\right)\hat{x}_t + \tfrac{\hat{\mu}\eta_t}{2-\mu\eta_t} x_{t+1}$$

Return $\hat{x}_T$

Assumption 2.10 (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 $0 \leq i < t$:

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

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

Taking into account Lemma 2.3, it is clear that Assumption 2.10 implies the earlier Assumption 2.8 with the same constants $\bar{\Delta}, \sigma$. Further, the condition on the drift holds trivially in the setting of Example 2.14 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 2.10; see Corollary 2.38 for the precise statement and proof.

Theorem 2.23 (Expected function gap)**.** *Let $\{\hat{x}_t\}$ be the iterates produced by Algorithm 4 with constant learning rate $\eta \leq 1/2L$, and suppose that Assumption 2.10 holds. Then the following bound holds for all $t \geq 0$:*

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

With Theorem 2.23 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 2.11; see Theorem 2.39 for the formal statement.

Theorem 2.24 (Time to track in expectation, informal)**.** *Suppose that Assumption 2.10 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 2.23 and 2.24. 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 2.11 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 $0 \le i < t$; 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 2.11 (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 $0 \le i < t$:

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

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

(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

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

Clearly the chain of implications

$$\text{Assumption 2.11} \implies \text{Assumption 2.10} \implies \text{Assumption 2.8}$$

holds, and the condition on the drift in Assumption 2.11 holds trivially in the setting of Example 2.14 with $\bar{\Delta} = \theta\beta/\hat{\mu}$ provided $\mu > 2\varepsilon\beta$. The following theorem shows that if Assumption 2.11 holds, then the expected bound on $\psi_t(\hat{x}_t) - \psi_t^\star$ derived in Theorem 2.23 holds with high probability, thereby extending Theorem 2.12.

Theorem 2.25 (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 2.11 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 $1 - \delta$:*

$$\psi_t(\hat{x}_t) - \psi_t^\star \le c\left(\left(1 - \tfrac{\hat{\mu}\eta}{2}\right)^t\left(\psi_0(x_0) - \psi_0^\star\right) + \eta\sigma^2 + \frac{\bar{\Delta}^2}{\hat{\mu}\eta^2}\right)\log\left(\frac{e}{\delta}\right). \tag{2.9}$$

With Theorem 2.25 in hand, we may implement a step decay schedule as before to obtain the following efficiency estimate, thereby extending Theorem 2.13; see Section 2.5.4 for the precise statements (Theorems 2.43 and 2.45) and proofs.

Theorem 2.26 (Time to track with high probability, informal)**.** *Suppose that Assumption 2.11 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*

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

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

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

2.5 Proofs of Main Results

Roadmap. In this section, we derive the results of the preceding sections under the unified framework presented in Section 2.4.1; we impose the assumptions and notation of Section 2.4.1

henceforth. Sections 2.5.1 and 2.5.2 handle distance tracking in expectation and with high probability, respectively; this corresponds to the results presented in Section 2.4.2 (entailing those of Sections 2.2.1 and 2.2.2). Then Sections 2.5.3 and 2.5.4 handle function gap tracking in expectation and with high probability, respectively; this corresponds to the results presented in Section 2.4.3 (entailing those of Sections 2.3.1 and 2.3.2).

2.5.1 Tracking the Equilibrium Point: Bounds in Expectation

The proof of Theorem 2.19 follows a familiar pattern in stochastic optimization. We begin by recalling Lemma 2.4, which gives a standard one-step improvement guarantee for the proximal stochastic gradient method on the fixed problem $\min \varphi_t$.

Lemma 2.27 (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

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

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

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

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

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

using the μ-strong convexity of f_t. Thus,

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

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

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

as claimed. □

It is critically important that the one-step improvement estimate in Lemma 2.27 holds with respect to any reference point x. In particular, as we already showed in Section 2.4.2, taking $x = \bar{x}_t$ and applying Lemma 2.17 yields Lemma 2.18:

Lemma 2.28 (Equilibrium one-step improvement)**.** *The iterates $\{x_t\}$ produced by Algorithm 3 with $\eta_t < 1/L$ satisfy the bound:*

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

With Lemma 2.28 in hand, we obtain the following recursion on $\|x_t - \bar{x}_t\|^2$.

Lemma 2.29 (Distance recursion)**.** *The iterates $\{x_t\}$ produced by Algorithm 3 with step size $\eta_t < 1/L$ satisfy the bound:*

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

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 2.28 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 2.29 recursively furnishes a bound on $\|x_t - \bar{x}_t\|^2$. When the step size is constant, the next proposition follows immediately.

Proposition 2.30 (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 2.30, we obtain the following precise version of Theorem 2.19.

Corollary 2.31 (Expected distance)**.** *Suppose that Assumption 2.8 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 2.31 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 2.20 (note that in the high drift-to-noise regime

$\bar{\Delta}/\sigma \geq \sqrt{\bar{\mu}/16L^3}$, Theorem 2.20 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 in [33].

Theorem 2.32 (Time to track in expectation)**.** *Suppose that Assumption 2.8 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[3]

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

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 2.31 directly implies

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

[3] 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.

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

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

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

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

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

Next, observe

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

so

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

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

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

This completes the proof. □

2.5.2 Tracking the Equilibrium Point: High-Probability Guarantees

The proof of Theorem 2.21 is based on recursively controlling the moment generating function of $\|x_t - \bar{x}_t\|^2$. Namely, Lemma 2.29 in the regime $\eta_t \le 1/2L$ directly yields

$$\|x_{t+1} - \bar{x}_{t+1}\|^2 \le (1 - \bar{\mu}\eta_t)\|x_t - \bar{x}_t\|^2 + 2\eta_t\langle z_t, v_t\rangle\|x_t - \bar{x}_t\| + 2\eta_t^2\|z_t\|^2 + \tfrac{2}{\bar{\mu}\eta_t}\bar{\Delta}_t^2, \qquad (2.10)$$

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 $\mathbb{E}[e^{\lambda\|x_t - \bar{x}_t\|^2}]$ through the recursive inequality (2.10). The basic probabilistic tool to achieve this in similar settings under bounded noise assumptions was developed in [36]; the following proposition is a slight generalization of Claim D.1 in [36] to the light-tail setting we require.

Proposition 2.33 (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

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

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

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

The proof is complete. □

We may now use Proposition 2.33 to derive the following precise version of Theorem 2.21.

Theorem 2.34 (High-probability distance tracking)**.** *Let $\{x_t\}$ be the iterates produced by Algorithm 3 with constant learning rate $\eta \le 1/2L$, and suppose that Assumption 2.9 holds. Then there exists an absolute constant*[4] *$c > 0$ such that for any specified $t \in \mathbb{N}$ and $\delta \in (0,1)$, the following estimate holds with probability at least $1-\delta$:*

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

Proof. Note first that under Assumption 2.9, there exists an absolute constant $c \ge 1$ 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 [43, see 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 2.9. Thus, in light of inequality (2.10), we may apply Proposition 2.33 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 $\nu_t = 2\eta_t^2 c\sigma^2 + 2\bar{\Delta}^2/\bar{\mu}\eta_t$,

[4] Explicitly, one can take any $c \ge 1$ 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.

yielding the estimate

$$\mathbb{E}\big[\exp\big(\lambda\|x_{t+1}-\bar{x}_{t+1}\|^2\big)\big] \le \exp\left(\lambda\left(2\eta_t^2 c\sigma^2 + \frac{2\bar{\Delta}^2}{\bar{\mu}\eta_t}\right)\right)\mathbb{E}\big[\exp\big(\lambda\big(1-\tfrac{\bar{\mu}\eta_t}{2}\big)\|x_t-\bar{x}_t\|^2\big)\big] \quad (2.11)$$

for all

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

Taking into account $\eta_t \equiv \eta$ and iterating the recursion (2.11), we deduce

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

for all

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

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 $c \ge 1$ and $\bar{\mu}\eta \le 1$, we have

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

and

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

Hence

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

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

With Theorem 2.34 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 2.22. The argument follows the same reasoning as in the proof of Theorem 2.32, with Theorem 2.34 playing the role of Corollary 2.31. The proof appears in Appendix B (see Section B.1).

Theorem 2.35 (Time to track with high probability)**.** *Suppose that Assumption 2.9 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*

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

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*.*

2.5.3 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 2.28 which holds in the static regularizer setting $r_t \equiv r$.

Lemma 2.36 (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 2.17, we have

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

Hence

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

Moreover, Young's inequality implies

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

Multiplying through by $2\eta_i$ and applying Lemma 2.27 completes the proof. □

Turning the estimate in Lemma 2.36 into an efficiency guarantee for the average iterate is essentially standard and follows for example from the averaging techniques used in [24, 32, 48]. 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 (2.8): $\gamma < \mu/2$.

Proposition 2.37 (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*

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

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

Proof. Setting $\alpha = \hat{\mu}/2$ in Lemma 2.36, 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 A.1 with $h = \psi_k - \psi_k^\star$ and then taking $k = t$. □

Taking expectations in Proposition 2.37 yields the following precise form of Theorem 2.23.

Corollary 2.38 (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 2.10 holds. Then*

$$\mathbb{E}\big[\psi_t(\hat{x}_t) - \psi_t^\star\big] \le (1-\hat{\rho})^t\,\mathbb{E}\big[\psi_t(x_0) - \psi_t^\star + \tfrac{\hat{\mu}}{4}\|x_0 - \bar{x}_t\|^2\big] + \eta\sigma^2 + \frac{8\bar{\Delta}^2}{\hat{\mu}\eta^2} \tag{2.12}$$

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 (2.12) follows by taking expectations in Proposition 2.37 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 2.10. Next, applying Lemma 2.17, Lemma 2.3, and Young's inequality together with the μ-strong convexity of ψ_0 yields

$$\psi_t(x_0) - \psi_t^\star + \tfrac{\hat{\mu}}{4}\|x_0 - \bar{x}_t\|^2 \le 3\big(\psi_0(x_0) - \psi_0^\star\big) + 5\bar{G}_{0,t}^2/\bar{\mu}, \tag{2.13}$$

and then taking expectations and invoking Assumption 2.10 gives

$$\mathbb{E}\big[\psi_t(x_0) - \psi_t^\star + \tfrac{\hat{\mu}}{4}\|x_0 - \bar{x}_t\|^2\big] \le 3\big(\psi_0(x_0) - \psi_0^\star\big) + 5\hat{\mu}\bar{\Delta}^2t^2. \tag{2.14}$$

Further, the inequality

$$e^{-\hat{\mu}\eta t/2}\hat{\mu}t^2 \le 16/\hat{\mu}\eta^2 \qquad \forall\hat{\mu},\eta,t>0 \tag{2.15}$$

combines with inequality (2.14) 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 2.38 to obtain a formal version of Theorem 2.24; the proof closely follows that of Theorem 2.32 and is included in Appendix B (see Section B.2).

Theorem 2.39 (Time to track in expectation)**.** *Suppose that Assumption 2.10 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* $D \geq \psi_0(x_0) - \psi_0^\star$. *Consider running Algorithm 4 in* $k = 0, \dots, K-1$ *epochs, namely, set* $X_0 = x_0$ *and iterate the process*

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

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

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

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

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

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

2.5.4 Tracking the Equilibrium Value: High-Probability Guarantees

In this section, we derive the high-probability analogues of the results in Section 2.5.3. In light of Proposition 2.37, 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 developed recently in [36] 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 $\sum_{i=0}^{n} d_i$ 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 2.36 and algebraic manipulation and is placed in Appendix B (see Section B.3).

Proposition 2.40 (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 in [36], or rather a direct consequence thereof [36, Lemma C.3].

Theorem 2.41 (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*

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

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 $\alpha_0, \ldots, \alpha_n \ge 0$,

$\delta \in [0,1]$, and $\beta(\delta) \geq 0$ satisfying

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

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

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

Combining Proposition 2.40 and Theorem 2.41 yields the following tail bound for $\sum_{i=0}^{t-1} d_i$.

Proposition 2.42 (Noise martingale tail bound)**.** *Let $\{x_t\}$ be the iterates produced by Algorithm 3 with constant step size $\eta \leq 1/2L$, set $\hat{\rho} := \hat{\mu}\eta/(2 - \mu\eta)$, and suppose that Assumption 2.11 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:*

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

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

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

Proof. By Assumption 2.11, 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 2.8, and Corollary 2.31. Now fix $t \geq 1$ and observe that Proposition 2.40 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

$$\mathbb{P}\left\{R_t \le \beta_t \log\left(\frac{3e}{\delta}\right)\right\} \ge 1-\delta \qquad \forall \delta \in (0,1). \tag{2.16}$$

To verify (2.16), observe first that for all $n \ge 0$, the sum $\sum_{i=0}^n \|z_i\|^2(1-\hat\rho)^{n-i}$ is sub-exponential with parameter $\sum_{i=0}^n c\sigma^2(1-\hat\rho)^{n-i} \le (c\sigma)^2/\hat\rho$, so Markov's inequality implies

$$\mathbb{P}\left\{\sum_{i=0}^n \|z_i\|^2(1-\hat\rho)^{n-i} \le \frac{(c\sigma)^2}{\hat\rho}\log\left(\frac{e}{\delta}\right)\right\} \ge 1-\delta \qquad \forall \delta \in (0,1). \tag{2.17}$$

Further, for all $0 \le n < t$, it follows from Assumption 2.11 and Lemma 2.3 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

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

so Markov's inequality implies

$$\mathbb{P}\left\{\|x_0-\bar x_t\|^2 \le 2(\|x_0-\bar x_0\|^2+\bar\Delta^2 t^2)\log\left(\frac{e}{\delta}\right)\right\} \ge 1-\delta \qquad \forall \delta\in(0,1) \tag{2.18}$$

and

$$\mathbb{P}\left\{\sum_{i=0}^n \bar G_{i,t}^2(1-\hat\rho)^{n-i} \le \frac{2(\hat\mu\bar\Delta)^2}{\hat\rho^3(1-\hat\rho)^{t-1-n}}\log\left(\frac{e}{\delta}\right)\right\} \ge 1-\delta \qquad \forall \delta\in(0,1). \tag{2.19}$$

Thus, (2.17)–(2.19) and a union bound yield (2.16). Consequently, Theorem 2.41 implies that the following bound holds for all $\delta \in (0,1)$ and $\tau > 0$:

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

as claimed. □

We may now deduce the following precise version of Theorem 2.25 using the tail bound furnished by Proposition 2.42.

Theorem 2.43 (Function gap with high probability)**.** *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 2.11 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 $1-\delta$:

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

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 2.42 to obtain

$$\mathbb{P}\left\{\sum_{i=0}^{t-1}\langle z_i, x_i - \bar{x}_t\rangle(1-\hat{\rho})^{t-1-i} < 5\sqrt{8\beta_t}\log\left(\frac{e}{\delta}\right)\right\} \geq 1 - 2\delta. \tag{2.20}$$

We may now combine (2.17), (2.19), and (2.20) together with Proposition 2.37 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 2.44. To see that Theorem 2.43 entails Theorem 2.25, observe first that in the setting of Theorem 2.43, 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 (2.15) implies

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

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 (2.13) and (2.15) together with Assumption 2.11 imply that the estimate

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

holds with probability at least $1-\delta$ for all $\delta\in(0,1)$. On the other hand, Theorem 2.43 shows that the estimate

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

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 2.25 to obtain a formal version of Theorem 2.26; the proof is analogous to that of Theorem 2.35 and appears in Appendix B (see Section B.4).

Theorem 2.45 (Time to track with high probability)**.** *Suppose that Assumption 2.11 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 $D\ge\psi_0(x_0)-\psi_0^\star$. Fix $\delta\in(0,1)$ and consider running Algorithm 4 in $k=0,\dots,K-1$ epochs, namely, set $X_0=x_0$ and iterate the process*

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

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

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

for all $k\ge1$, where $c>0$ is the absolute constant furnished by the bound (2.9). Then the time horizon $T=T_0+\cdots+T_{K-1}$ satisfies

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

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 $1 - K\delta$.

2.6 Numerical Illustrations

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

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 (2.1) 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 2.1 and 2.2, 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 2.1: $\Delta/\sigma < \sqrt{\mu/16L^3} = 1/4$. 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 $T = 100$.

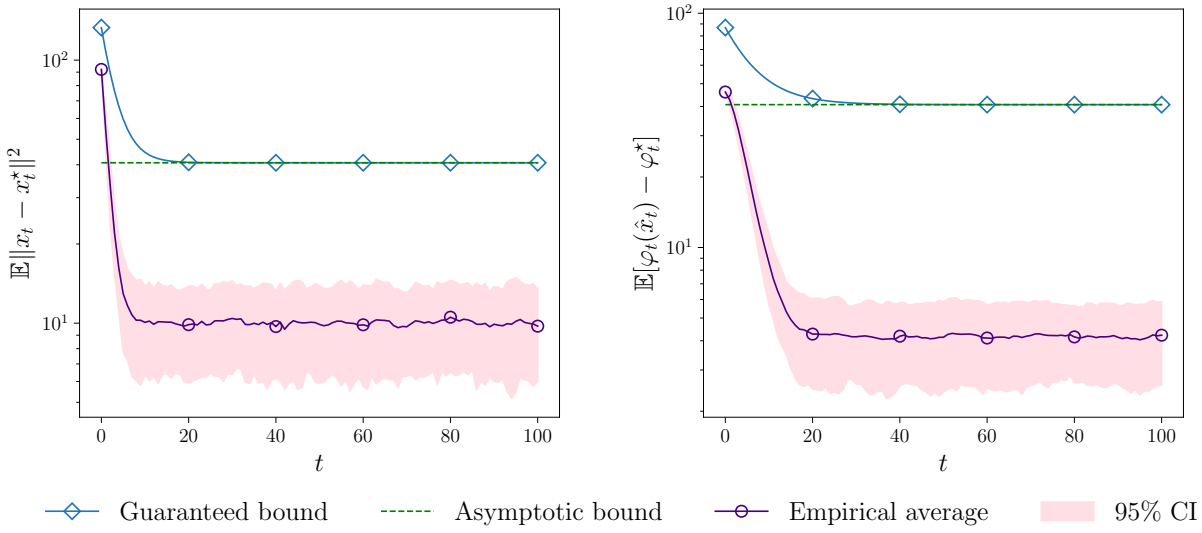


Figure 2.1: Semilog plots of guaranteed bounds and empirical tracking errors with respect to iteration t for least-squares recovery. Shaded regions indicate the 95% confidence intervals for $\|x_t - x_t^\star\|^2$ and $\varphi_t(\hat{x}_t) - \varphi_t^\star$; empirical averages and confidence intervals are computed over 100 trials. Default parameter values: $\mu = L = 1$, $\sigma = 10$, $\Delta = 1$, and $\eta = \eta_\star$.

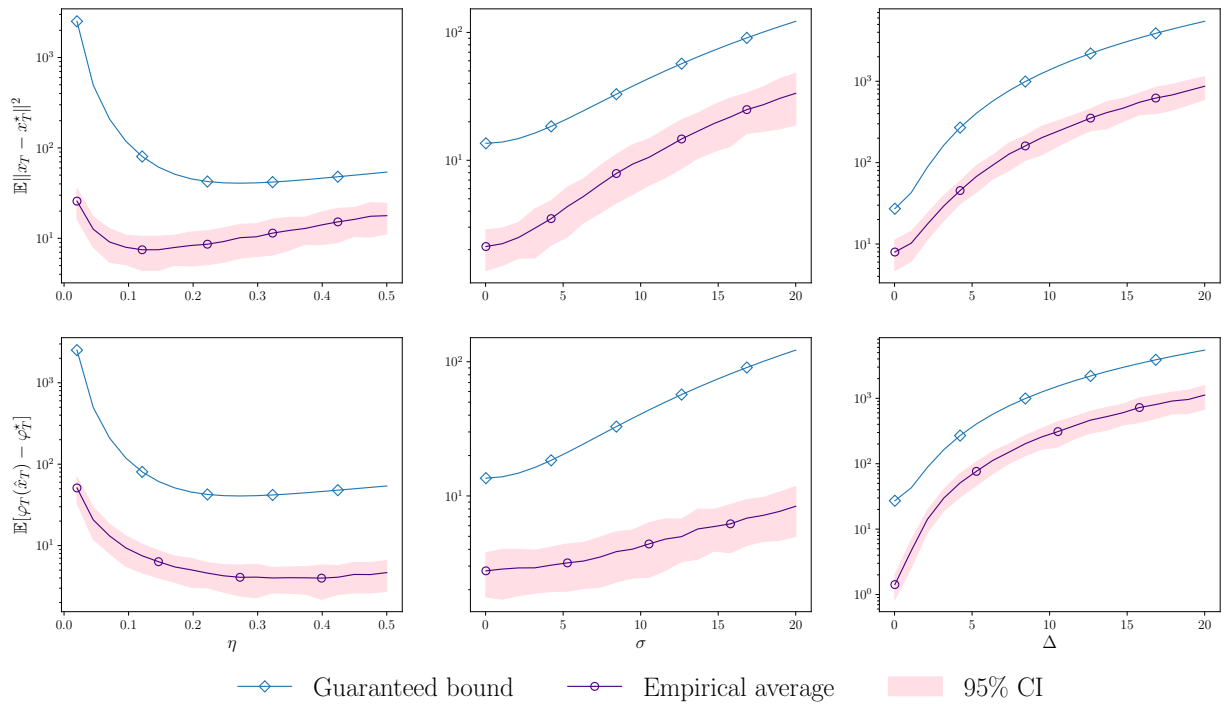


Figure 2.2: Semilog plots of guaranteed bounds and empirical tracking errors at horizon $T = 100$ with respect to η, σ, and Δ for least-squares recovery. Shaded regions indicate the 95% confidence intervals for $\|x_T - x_T^\star\|^2$ and $\varphi_T(\hat{x}_T) - \varphi_T^\star$; empirical averages and 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 $\mathbb{E}\|x_t^\star - x_{t+1}^\star\|^2 \le \Delta^2$. 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 $\operatorname{tr} \Sigma_t \le \sigma^2/L$. This amounts to the problem (2.1) 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

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

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 $\mathbb{E}\|\nabla f_t(x_t) - g_t\|^2 \le \sigma^2$.

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 2.3 and 2.4, 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 2.3: $\Delta/\sigma < \sqrt{\mu/16L^3} = 1/4$. 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 $T = 100$.

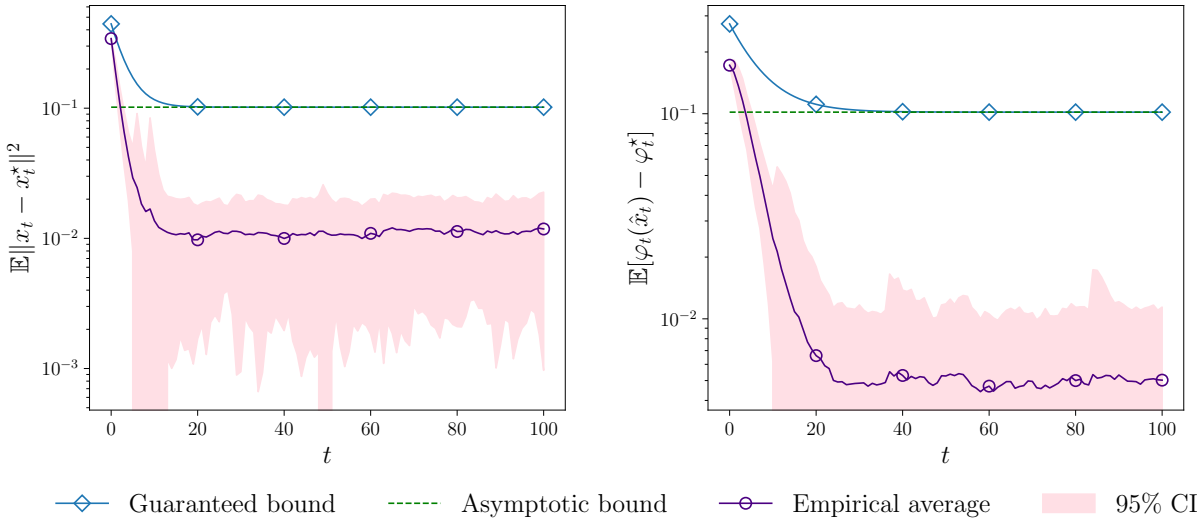


Figure 2.3: Semilog plots of guaranteed bounds and empirical tracking errors with respect to iteration t for sparse least-squares recovery. Shaded regions indicate the 95% confidence intervals for $\|x_t - x_t^\star\|^2$ and $\varphi_t(\hat{x}_t) - \varphi_t^\star$; empirical averages and confidence intervals are computed over 100 trials. Default parameter values: $\mu = L = 1$, $\sigma = 1/2$, $\Delta = 1/20$, and $\eta = \eta_\star$.

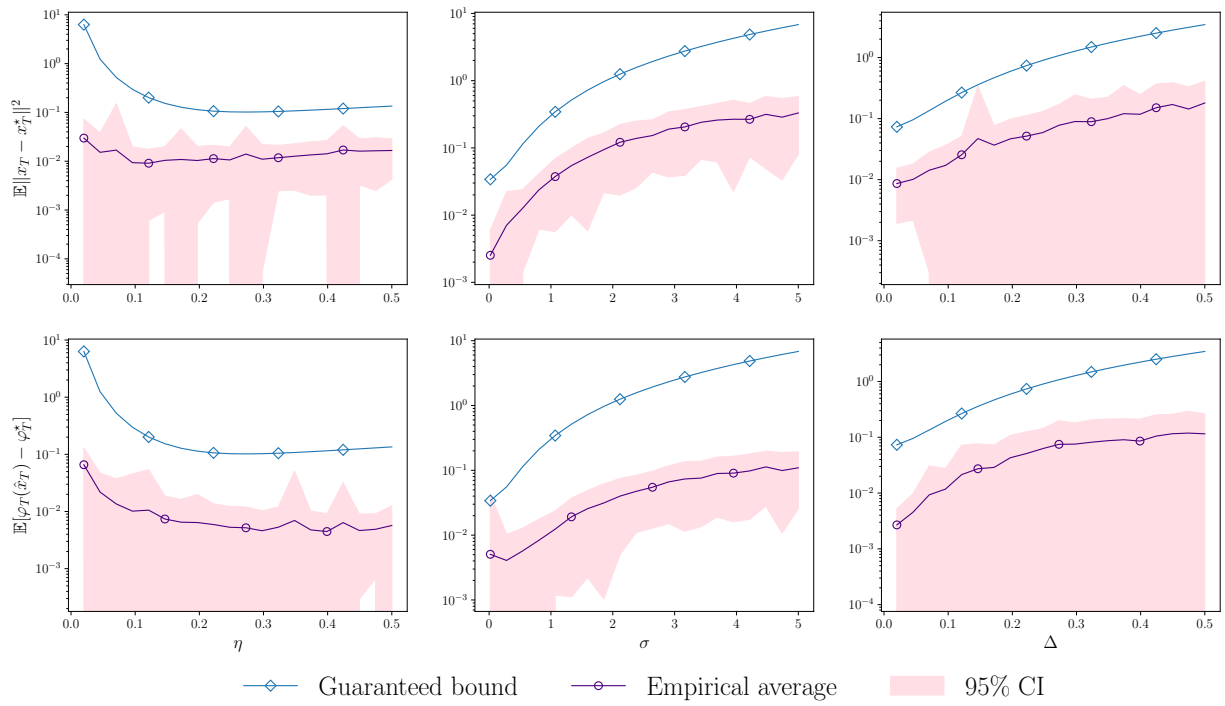


Figure 2.4: Semilog plots of guaranteed bounds and empirical tracking errors at horizon $T = 100$ with respect to η, σ, and Δ for sparse least-squares recovery. Shaded regions indicate the 95% confidence intervals for $\|x_T - x_T^\star\|^2$ and $\varphi_T(\hat{x}_T) - \varphi_T^\star$; empirical averages 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 $a_1, \dots, a_n \in \mathbb{R}^d$, $\{b_t\}$ 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 (2.1) 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 $\Delta = \frac{1}{\mu n}\max_{i=1,\dots,n}\|a_i\|$, 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 $k \sim \mathsf{Unif}\{1,\dots,n\}$ 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

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

In our simulations, we set $d = 20$ and $n = 200$, fix standard Gaussian vectors $x_0 \in \mathbb{R}^d$ and $a_1, \dots, a_n \in \mathbb{R}^d$, 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 2.5, we use default parameter values $\mu = 1$ and the corresponding asymptotically optimal step size $\eta = \eta_\star$. In Figure 2.6, 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 2.7, 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 2.7 illustrates that $\eta_\star$ is close to empirically optimal.

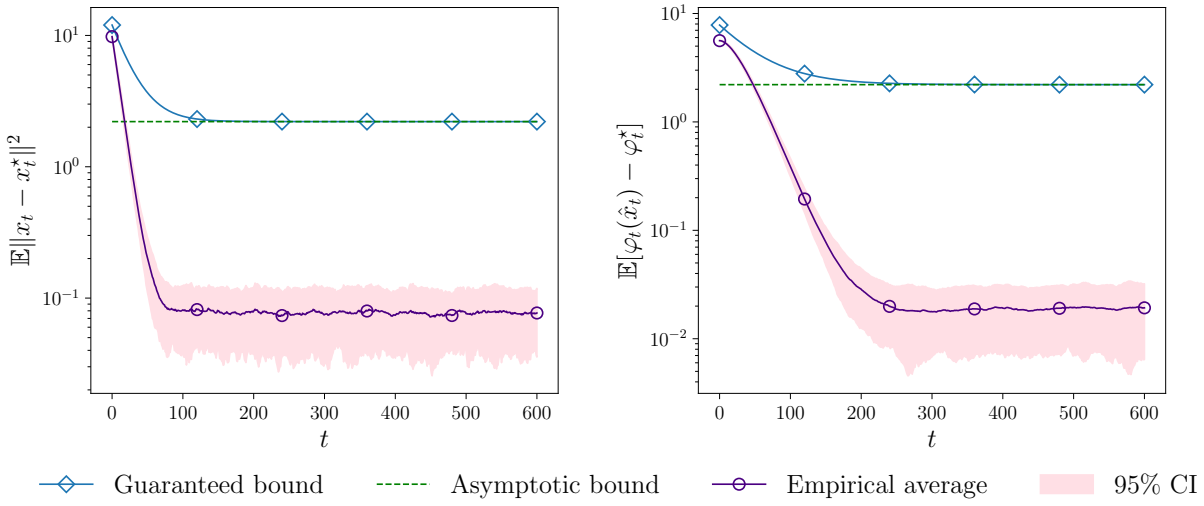


Figure 2.5: Semilog plots of guaranteed bounds and empirical tracking errors with respect to iteration t for ℓ_2^2-regularized logistic regression. Shaded regions indicate the 95% confidence intervals for $\|x_t - x_t^\star\|^2$ and $\varphi_t(\hat{x}_t) - \varphi_t^\star$; empirical averages and confidence intervals are computed over 100 trials. Default parameter values: $\mu = 1$ and $\eta = \eta_\star$.

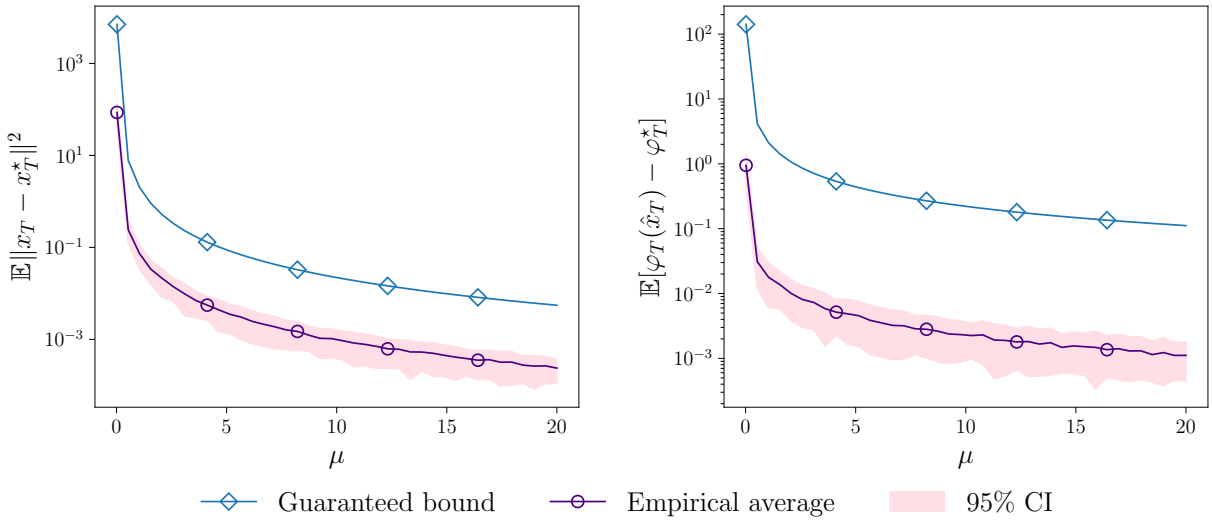


Figure 2.6: Semilog plots of guaranteed bounds and empirical tracking errors at horizon $T = 600$ with respect to the strong convexity parameter μ for ℓ_2^2-regularized logistic regression. Shaded regions indicate the 95% confidence intervals for $\|x_T - x_T^\star\|^2$ and $\varphi_T(\hat{x}_T) - \varphi_T^\star$; empirical averages and confidence intervals are computed over 100 trials, using the asymptotically optimal step size $\eta_\star$ (which itself depends on μ).

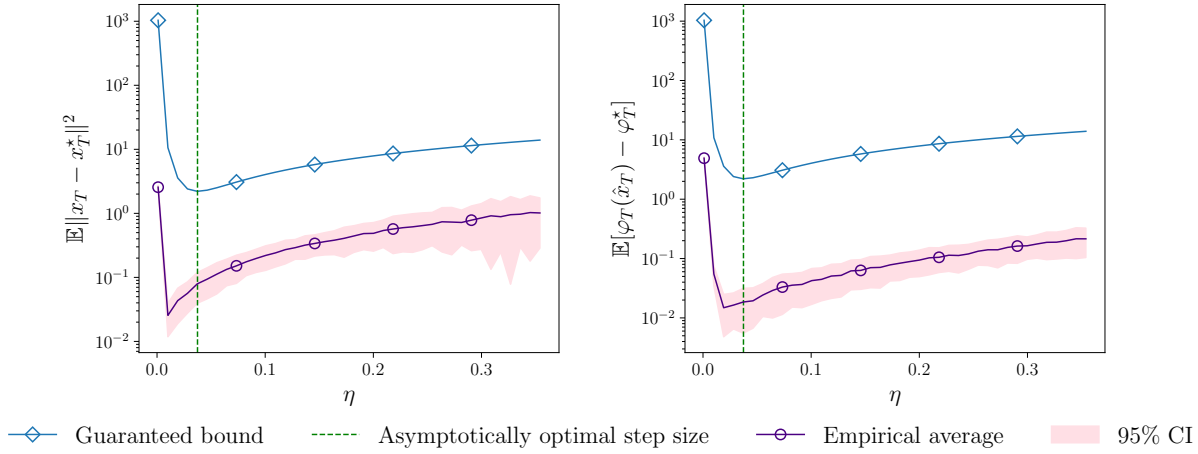


Figure 2.7: Semilog plots of guaranteed bounds and empirical tracking errors at horizon $T = 600$ with respect to the step size η for ℓ_2^2-regularized logistic regression. Shaded regions indicate the 95% confidence intervals for $\|x_T - x_T^\star\|^2$ and $\varphi_T(\hat{x}_T) - \varphi_T^\star$; empirical averages and confidence intervals are computed over 100 trials. Default parameter value: $\mu = 1$. Observe that $\eta_\star$ is close to empirically optimal.

Chapter 3

STOCHASTIC APPROXIMATION WITH DECISION-DEPENDENT DISTRIBUTIONS: ASYMPTOTIC NORMALITY AND OPTIMALITY

Joint work with M. Díaz and D. Drusvyatskiy [18]

The outline of the chapter is as follows. Section 3.1 records some basic notation and definitions that we will use. Section 3.2 formally introduces/reviews the decision-dependent framework. In Section 3.3, we show that the running average of the stochastic forward-backward algorithm (SFB) is asymptotically normal (Theorem 1.1), and identify its asymptotic covariance. Finally, Section 3.4 presents the local asymptotic minimax lower bound (Theorem 1.2). We defer many of the technical proofs to the appendices.

3.1 Basic Notation and Definitions

Throughout the chapter, we let $\mathbb{R}^d$ denote the standard d-dimensional Euclidean space equipped with the dot product $\langle x, y\rangle = x^\top y$ and the induced norm $\|x\| = \sqrt{\langle x, x\rangle}$. For any set $\mathcal{X} \subset \mathbb{R}^d$, the symbol $\mathrm{proj}_{\mathcal{X}}(x)$ will denote the set $\arg\min_{y\in\mathcal{X}} \|y - x\|$ of nearest points of $\mathcal{X}$ to $x \in \mathbb{R}^d$. We say that a function $\mathcal{L}\colon \mathbb{R}^d \to \mathbb{R}$ is *symmetric* if it satisfies $\mathcal{L}(x) = \mathcal{L}(-x)$ for all $x \in \mathbb{R}^d$, and we say that $\mathcal{L}$ is *quasiconvex* if its sublevel set $\{x \mid \mathcal{L}(x) \le c\}$ is convex for any $c \in \mathbb{R}$. For any matrix $A \in \mathbb{R}^{m\times n}$, the symbols $\|A\|_{\mathrm{op}}$ and $A^\dagger$ stand for the operator norm and Moore-Penrose pseudoinverse of A, respectively. For any two symmetric matrices $A, B \in \mathbb{R}^{n\times n}$, we write $A \succeq B$ if the matrix $A - B$ is positive semidefinite.

Strong monotonicity and smoothness. A map $F\colon \mathcal{X} \to \mathbb{R}^d$ is called *α-strongly monotone* on $\mathcal{X} \subset \mathbb{R}^d$ if $\alpha > 0$ and

$$\langle F(x) - F(x'), x - x' \rangle \geq \alpha \|x - x'\|^2 \qquad \text{for all } x, x' \in \mathcal{X}.$$

We say that a map $F\colon \mathcal{X} \to \mathbb{R}^m$ is *smooth* on a closed set $\mathcal{X} \subset \mathbb{R}^d$ if F extends to a differentiable map defined on an open neighborhood of $\mathcal{X}$; further, we say that F is *β-smooth* on $\mathcal{X}$ if the Jacobian of F satisfies the Lipschitz condition

$$\|\nabla F(x) - \nabla F(x')\|_{\text{op}} \leq \beta \|x - x'\| \qquad \text{for all } x, x' \in \mathcal{X}.$$

Probability measures. Given a nonempty Polish metric space $\mathcal{Z}$ (i.e., separable and complete), we equip $\mathcal{Z}$ with its Borel σ-algebra $\mathcal{B}(\mathcal{Z})$ and let $P_1(\mathcal{Z})$ denote the set of probability measures on $\mathcal{Z}$ with finite first moment. We will measure the deviation between two measures $\mu, \nu \in P_1(\mathcal{Z})$ using the Wasserstein-1 distance:

$$W_1(\mu, \nu) = \sup_{\phi \in \text{Lip}_1(\mathcal{Z})} \left\{ \mathop{\mathbb{E}}_{X \sim \mu}\big[\phi(X)\big] - \mathop{\mathbb{E}}_{Y \sim \nu}\big[\phi(Y)\big] \right\}. \tag{3.1}$$

Here, $\text{Lip}_1(\mathcal{Z})$ denotes the set of 1-Lipschitz functions $\mathcal{Z} \to \mathbb{R}$. Equipped with the metric W_1, the set $P_1(\mathcal{Z})$ becomes a Polish metric space.

For any two probability measures μ and ν on $\mathcal{Z}$ such that μ is absolutely continuous with respect to ν (denoted $\mu \ll \nu$) and any convex function $f\colon (0, \infty) \to \mathbb{R}$ with $f(1) = 0$, the f-divergence of μ from ν is given by

$$\Delta_f(\mu \,\|\, \nu) = \int_{\mathcal{Z}} f\left(\frac{d\mu}{d\nu}\right) d\nu, \tag{3.2}$$

where $\frac{d\mu}{d\nu}\colon \mathcal{Z} \to [0, \infty)$ denotes the Radon-Nikodym derivative of μ with respect to ν and we take $f(0) = \lim_{t \downarrow 0} f(t)$. Abusing notation slightly, if μ is not absolutely continuous with respect to ν, then we set $\Delta_f(\mu \,\|\, \nu) = \infty$.

We will refer to a Borel measurable map between metric spaces simply as *measurable*. Likewise, we will refer to Borel measurable sets simply as *measurable*.

Notions of convergence. Given a sequence of random vectors $X_k \colon \Omega_k \to \mathbb{R}^m$ defined on probability spaces $(\Omega_k, \mathcal{S}_k, P_k)$ and a random vector $X \sim \mu$ in $\mathbb{R}^m$, we write either $X_k \rightsquigarrow X$ or $X_k \rightsquigarrow \mu$ to indicate that X_k converges in distribution to X (i.e., $\lim_{k\to\infty} \mathbb{E}_{P_k}[\varphi(X_k)] = \mathbb{E}_{X\sim\mu}[\varphi(X)]$ for every bounded continuous function $\varphi \colon \mathbb{R}^m \to \mathbb{R}$). We write $X_k = o_{P_k}(1)$ if X_k tends to zero in P_k-probability (i.e., $\lim_{k\to\infty} P_k\{\|X_k\| < \varepsilon\} = 1$ for all $\varepsilon > 0$). If X and each X_k are defined on a common probability space $(\Omega, \mathcal{S}, P)$, then the notation $X_k \xrightarrow{p} X$ indicates that X_k converges to X in probability (i.e., $\lim_{k\to\infty} P\{\|X_k - X\| < \varepsilon\} = 1$ for all $\varepsilon > 0$), and the notation $X_k \xrightarrow{\text{a.s.}} X$ indicates that X_k converges to X almost surely (i.e., $P\{\omega \in \Omega \mid \lim_{k\to\infty} X_k(\omega) = X(\omega)\} = 1$).

For any pair of vector-valued sequences (a_k) and (b_k), we write $a_k = O(b_k)$ if there exists a constant $C > 0$ such that $\|a_k\| \leq C\|b_k\|$ for all but finitely many k; we write $a_k = o(b_k)$ if for every $\varepsilon > 0$, the inequality $\|a_k\| \leq \varepsilon\|b_k\|$ holds for all but finitely many k; we write $a_k = \Theta(b_k)$ if there exist constants $c, C > 0$ such that $c\|b_k\| \leq \|a_k\| \leq C\|b_k\|$ for all but finitely many k; and we write $a_k \propto b_k$ if there exists a constant c such that $a_k = cb_k$ for all but finitely many k.

3.2 Background on Learning with Decision-Dependent Distributions

In this section, we formally specify the class of problems that we consider along with relevant assumptions. In order to model decision-dependence, we fix a nonempty, closed, convex set $\mathcal{X} \subset \mathbb{R}^d$, a nonempty Polish metric space $(\mathcal{Z}, d_{\mathcal{Z}})$, and a map $\mathcal{D} \colon \mathcal{X} \to P_1(\mathcal{Z})$. For ease of notation, we set $\mathcal{D}_x := \mathcal{D}(x)$ for each $x \in \mathcal{X}$. Thus, $\{\mathcal{D}_x\}_{x\in\mathcal{X}}$ is a family of probability distributions on $\mathcal{Z}$ indexed by points $x \in \mathcal{X}$. The variational behavior of the map $\mathcal{D} \colon \mathcal{X} \to P_1(\mathcal{Z})$ will play a central role in our work. In particular, following [60], we will assume that $\mathcal{D} \colon \mathcal{X} \to P_1(\mathcal{Z})$ is Lipschitz continuous.

Assumption 3.1 (Lipschitz distribution map)**.** There is a constant $\gamma > 0$ satisfying

$$W_1\big(\mathcal{D}(x), \mathcal{D}(x')\big) \leq \gamma\|x - x'\| \qquad \text{for all } x, x' \in \mathcal{X}.$$

Importantly, Assumption 3.1 implies that the map $\mathcal{D} \colon \mathcal{X} \to P_1(\mathcal{Z})$ is measurable, which

in turn implies that $\{\mathcal{D}_x\}_{x\in\mathcal{X}}$ is a *Markov kernel* from $\mathcal{X}$ to $\mathcal{Z}$, i.e., for each $E \in \mathcal{B}(\mathcal{Z})$, the function $\mathcal{X} \to [0,1]$ given by $x \mapsto \mathcal{D}_x(E)$ is measurable (see Lemma F.1).

Next, we fix a measurable map $G\colon \mathcal{X} \times \mathcal{Z} \to \mathbb{R}^d$ such that each section $G(x,\cdot)\colon \mathcal{Z} \to \mathbb{R}^d$ is Lipschitz continuous, and we define the family of maps $G_x\colon \mathcal{X} \to \mathbb{R}^d$ by setting

$$G_x(y) = \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x} G(y,z)$$

for all $x, y \in \mathcal{X}$; since $\mathcal{D}_x$ has finite first moment, the Lipschitz continuity of $G(y,\cdot)$ guarantees that $G_x(y)$ is well defined. Additionally, we impose the following standard regularity conditions on $G\colon \mathcal{X} \times \mathcal{Z} \to \mathbb{R}^d$.

Assumption 3.2 (Loss regularity)**.** There are constants $\beta, \bar{L} \geq 0$ and $\alpha > 0$ and a measurable function $L\colon \mathcal{Z} \to [0,\infty)$ satisfying the following three conditions.

(i) **(Lipschitz continuity)** For all $x, x' \in \mathcal{X}$ and $z, z' \in \mathcal{Z}$, the Lipschitz bounds

$$\|G(x,z) - G(x',z)\| \leq L(z) \cdot \|x - x'\|,$$
$$\|G(x,z) - G(x,z')\| \leq \beta \cdot d_{\mathcal{Z}}(z,z')$$

hold. Further, the second moment bound $\mathbb{E}_{z\sim\mathcal{D}_x}[L(z)^2] \leq \bar{L}^2$ holds for all $x \in \mathcal{X}$.

(ii) **(Monotonicity)** For all $x \in \mathcal{X}$, the map $G_x(\cdot)$ is α-strongly monotone on $\mathcal{X}$.

(iii) **(Compatibility)** The inequality $\gamma\beta < \alpha$ holds.

A few comments are in order. Condition (i) asserts that the map $G(x,z)$ is separately Lipschitz continuous with respect to both x and z; an immediate consequence is that each map $G_x(\cdot)$ is $\bar{L}$-Lipschitz continuous on $\mathcal{X}$. Condition (ii) is a standard monotonicity requirement; when $G(x,z) = \nabla_x \ell(x,z)$ is given by the gradient of a loss function ℓ, this corresponds to α-strong convexity of the expected loss $y \mapsto \mathbb{E}_{z\sim\mathcal{D}_x}\ell(y,z)$. Condition (iii) ensures that the Lipschitz constant γ of $\mathcal{D}(\cdot)$ is sufficiently small in comparison with the monotonicity constant α, signifying that the dynamics are "mild". This condition is widely used in the existing literature [58, 60, 61, 82].

Assumptions 3.1 and 3.2 imply the following useful Lipschitz estimate on the deviation $G_x(y) - G_{x'}(y)$ arising from the shift in distribution from $\mathcal{D}_x$ to $\mathcal{D}_{x'}$. We will use this estimate often in what follows. The proof is identical to that of Lemma 5 in [58]; a short argument appears in Section C.1.

Lemma 3.1 (Deviation)**.** *Suppose that Assumptions 3.1 and 3.2 hold. Then the estimate*

$$\|G_x(y) - G_{x'}(y)\| \leq \gamma\beta \cdot \|x - x'\|$$

holds for all $x, x', y \in \mathcal{X}$.

Corresponding to each distribution $\mathcal{D}_x$ is the variational inequality

$$0 \in \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x} G(y, z) + N_{\mathcal{X}}(y). \qquad \mathrm{VI}(\mathcal{D}_x)$$

The following definition, originating in [60] for performative prediction and in [58] for its multiplayer extension, is the key solution concept that we will use.

Definition 3.2 (Equilibrium point)**.** We say that $x^\star$ is an *equilibrium point* of the family of variational inequalities $\{\mathrm{VI}(\mathcal{D}_x)\}_{x\in\mathcal{X}}$ if it satisfies:

$$0 \in G_{x^\star}(x^\star) + N_{\mathcal{X}}(x^\star).$$

Thus, $x^\star$ is an equilibrium point of $\{\mathrm{VI}(\mathcal{D}_x)\}_{x\in\mathcal{X}}$ if $y = x^\star$ is itself a solution to the variational inequality $\mathrm{VI}(\mathcal{D}_{x^\star})$ induced by the distribution $\mathcal{D}_{x^\star}$. Equivalently, these are exactly the fixed points of the map

$$\texttt{Sol}(x) := \{y \mid 0 \in G_x(y) + N_{\mathcal{X}}(y)\}, \tag{3.3}$$

which is single-valued on $\mathcal{X}$ by the continuity and strong monotonicity of $G_x(\cdot)$ [66, Example 12.7 and Proposition 12.54]. Equilibrium points have a clear intuitive meaning: a learning system that deploys a learning rule $x^\star$ that is at equilibrium has no incentive to deviate from $x^\star$ based only on the data drawn from $\mathcal{D}_{x^\star}$. The key role of equilibrium points in (multiplayer) performative prediction is by now well documented [24, 54, 58, 60, 61, 82]. Most importantly, equilibrium points exist and are unique under Assumptions 3.1 and 3.2. The proof is identical to that of Theorem 7 in [58]; we provide a short argument in Section C.2 for completeness.

Theorem 3.3 (Existence)**.** *Suppose that Assumptions 3.1 and 3.2 hold. Then the map* $\texttt{Sol}(\cdot)$ *is* $\frac{\gamma\beta}{\alpha}$*-contractive on* $\mathcal{X}$ *and therefore the problem admits a unique equilibrium point* $x^\star$*.*

We note in passing that when $\gamma\beta \geq \alpha$, equilibrium points may easily fail to exist; see, e.g., [60, Proposition 3.6]. Therefore, the regime $\gamma\beta < \alpha$ is the natural setting to consider when searching for equilibrium points.

3.3 Convergence and Asymptotic Normality

A central goal of performative prediction is the search for equilibrium points, which are simply the fixed points of the map $\texttt{Sol}(\cdot)$ defined in (3.3). Though the map $\texttt{Sol}(\cdot)$ is contractive, it cannot be evaluated directly since it involves evaluating the expectation $G_x(y) = \mathbb{E}_{z\sim\mathcal{D}_x} G(y,z)$. Employing the standard assumption that the only access to $\mathcal{D}_x$ is through sampling, one may instead in iteration t take a single stochastic forward-backward step on the problem corresponding to $\texttt{Sol}(x_t)$. The resulting procedure is recorded in Algorithm 5 below. In the setting of performative prediction [54] and its multiplayer extension [58], the algorithm reduces to projected stochastic gradient methods.

Algorithm 5 Stochastic Forward-Backward Method (SFB)

Input: initial $x_0 \in \mathcal{X}$ and step size sequence $(\eta_t)_{t\geq 0} \subset (0,\infty)$

Step $t \geq 0$:

$$\text{Sample } z_t \sim \mathcal{D}(x_t)$$

$$\text{Set } x_{t+1} = \operatorname{proj}_{\mathcal{X}}\big(x_t - \eta_t G(x_t, z_t)\big)$$

For the remainder of Section 3.3, we let $(x_t)_{t\geq 0}$ denote the stochastic process generated by Algorithm 5 on the probability space $(\mathcal{Z}^{\mathbb{N}}, \mathcal{B}(\mathcal{Z}^{\mathbb{N}}), \mathbb{P})$, where $\mathbb{P} = \bigotimes_{i=0}^{\infty} \mathcal{D}_{x_i}$ is the unique probability measure on the countable product space $\mathcal{Z}^{\mathbb{N}}$ satisfying

$$\mathbb{P}(E_0 \times \cdots \times E_t \times \mathcal{Z}^{\mathbb{N}}) = \int_{E_0} \cdots \int_{E_t} d\mathcal{D}_{x_t}(z_t) \cdots d\mathcal{D}_{x_0}(z_0) \tag{3.4}$$

for all $E_0, \ldots, E_t \in \mathcal{B}(\mathcal{Z})$ and $t \geq 0$ (see Theorem F.2). We will see that under very mild assumptions, the SFB iterates x_t almost surely converge to the equilibrium point $x^\star$. To this

end, we define for each $(x, z) \in \mathcal{X} \times \mathcal{Z}$ the noise vector

$$\xi_x(z) := G(x, z) - G_x(x) \tag{3.5}$$

and impose the following standard bound on the conditional second moment of the noise.

Assumption 3.3 (Variance bound)**.** There is a constant $K \geq 0$ such that for all $t \geq 0$, the following bound holds almost surely:

$$\mathop{\mathbb{E}}_{z_t \sim \mathcal{D}_{x_t}} \|\xi_{x_t}(z_t)\|^2 \leq K(1 + \|x_t - x^\star\|^2).$$

The subsequent proposition shows that the SFB iterates almost surely converge to the equilibrium point under Assumptions 3.1–3.3 and standard conditions restricting the rate of decrease of the step sizes η_t. The proof, which follows from a simple one-step improvement bound for the SFB method [58, Theorem 24] and an application of the Robbins-Siegmund almost supermartingale convergence theorem [65], appears in Section C.3.

Proposition 3.4 (Almost sure convergence)**.** *Suppose that Assumptions 3.1–3.3 hold and the step size sequence in Algorithm 5 satisfies* $\sum_{t=0}^{\infty} \eta_t = \infty$ *and* $\sum_{t=0}^{\infty} \eta_t^2 < \infty$*. Then* x_t *converges to* $x^\star$ *almost surely as* $t \to \infty$*, and* $\sum_{t=0}^{\infty} \eta_t \|x_t - x^\star\|^2 < \infty$ *almost surely. Moreover, if* $\eta_t = \Theta(t^{-\nu})$ *for some* $\nu \in \left(\frac{1}{2}, 1\right)$*, then* $\mathbb{E}\|x_t - x^\star\|^2 = O(t^{-\nu})$ *and hence* $\sum_{t=1}^{\infty} t^{-1/2} \|x_t - x^\star\|^2 < \infty$ *almost surely.*

The main result of this section is the asymptotic normality of the average iterates

$$\bar{x}_t := \frac{1}{t} \sum_{i=1}^{t} x_i,$$

for which we require the following additional assumption.

Assumption 3.4. The following four conditions hold.

(i) **(Interiority)** The equilibrium point $x^\star$ lies in the interior of $\mathcal{X}$.

(ii) **(Lipschitz Jacobian)** On a neighborhood of $x^\star$, the map $x \mapsto G_x(x)$ is differentiable with Lipschitz continuous Jacobian.

(iii) **(Asymptotic uniform integrability)** We have

$$\limsup_{t\to\infty} \mathop{\mathbb{E}}_{z_t\sim\mathcal{D}_{x_t}}\left[\|G(x^\star,z_t)\|^2\mathbf{1}_{\{\|G(x^\star,z_t)\|\geq N\}}\right] \xrightarrow{\text{a.s.}} 0 \qquad \text{as } N\to\infty$$

and

$$\mathop{\mathbb{E}}_{z\sim\mathcal{D}_{x^\star}}\left[\|G(x^\star,z)\|^2\mathbf{1}_{\{\|G(x^\star,z)\|\geq N\}}\right] \to 0 \qquad \text{as } N\to\infty.$$

(iv) **(Lindeberg's condition)** For all $\varepsilon > 0$,

$$\frac{1}{t}\sum_{i=0}^{t-1} \mathop{\mathbb{E}}_{z_i\sim\mathcal{D}_{x_i}}\left[\|\xi_{x_i}(z_i)\|^2\mathbf{1}_{\{\|\xi_{x_i}(z_i)\|\geq\varepsilon\sqrt{t}\}}\right] \xrightarrow{p} 0 \qquad \text{as } t\to\infty.$$

A few comments are in order. First, the interiority condition (i) is a standard assumption for asymptotic normality results even in static settings [62]. The smoothness condition (ii) is fairly mild. For example, it holds if the partial derivatives $\nabla_y G_x(y)$ and $\nabla_x G_x(y)$ exist and are Lipschitz continuous on a neighborhood of $(x^\star, x^\star)$; in turn, this holds if, on a neighborhood of $x^\star$, each distribution $\mathcal{D}(x)$ admits a density $p(x,z) = \frac{d\mathcal{D}(x)}{d\mu}(z)$ with respect to a common base measure $\mu \gg \mathcal{D}(x)$ such that $G(\cdot,z)$ and $p(\cdot,z)$ are $C^{1,1}$-smooth[1] and sufficient integrability conditions hold to invoke dominated convergence.

Asymptotic uniform integrability conditions such as (iii) are key for obtaining convergence of moments [75, Section 2.5]; it is used in our setting to establish

$$\mathop{\mathbb{E}}_{z_t\sim\mathcal{D}_{x_t}}\left[G(x_t,z_t)G(x_t,z_t)^\top\right] \xrightarrow{\text{a.s.}} \mathop{\mathbb{E}}_{z\sim\mathcal{D}_{x^\star}}\left[G(x^\star,z)G(x^\star,z)^\top\right] \qquad \text{as } t\to\infty$$

(see Theorem F.5). Condition (iii) holds, for instance, if there exists a neighborhood $\mathcal{V}$ of $x^\star$ satisfying $\sup_{x\in\mathcal{V}} \mathbb{E}_{z\sim\mathcal{D}_x}\left[\|G(x^\star,z)\|^2\mathbf{1}_{\{\|G(x^\star,z)\|\geq N\}}\right] \to 0$ as $N\to\infty$; in turn, this holds if $\sup_{x\in\mathcal{V}} \mathbb{E}_{z\sim\mathcal{D}_x}\left[\|G(x^\star,z)\|^q\right] < \infty$ for some $q\in(2,\infty)$, e.g., if each random vector $G(x^\star,z)$, with $z\sim\mathcal{D}_x$, is sub-Gaussian with the same variance proxy σ^2 for all $x\in\mathcal{V}$. Lindeberg's condition (iv) imposes a standard constraint on the sequence of noise vectors $\xi_{x_t}(z_t)$ for application of the martingale central limit theorem (see Theorem F.4); it holds, for example, if both $\sup_{t\geq 0}\mathbb{E}_{z_t\sim\mathcal{D}_{x_t}}\left[\|\xi_{x_t}(z_t)\|^2\right] < \infty$ almost surely and the asymptotic uniform integrability condition $\limsup_{t\to\infty}\mathbb{E}_{z_t\sim\mathcal{D}_{x_t}}\left[\|\xi_{x_t}(z_t)\|^2\mathbf{1}_{\{\|\xi_{x_t}(z_t)\|\geq N\}}\right] \xrightarrow{p} 0$ as $N\to\infty$ is fulfilled.

[1] That is, differentiable with locally Lipschitz continuous partial derivatives.

We are now ready to present the main result of this section.

Theorem 3.5 (Asymptotic normality)**.** *Suppose that Assumptions 3.1–3.4 hold and the step size sequence in Algorithm 5 satisfies $\eta_t \propto t^{-\nu}$ for some $\nu \in \left(\frac{1}{2}, 1\right)$. Let $R\colon \mathcal{X} \to \mathbb{R}^d$ and $\Sigma \succeq 0$ be given by*

$$R(x) = \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x}[G(x,z)] \qquad \text{and} \qquad \Sigma = \mathop{\mathbb{E}}_{z\sim\mathcal{D}_{x^\star}}\left[G(x^\star,z)G(x^\star,z)^\top\right],$$

and let $\xi_t = \xi_{x_t}(z_t)$ denote the noise vector at step t given by (3.5). Then, as $t \to \infty$, the average iterates $\bar{x}_t = \frac{1}{t}\sum_{i=1}^t x_i$ converge to $x^\star$ almost surely,

$$\sqrt{t}(\bar{x}_t - x^\star) = -\nabla R(x^\star)^{-1}\left(\frac{1}{\sqrt{t}}\sum_{i=0}^{t-1}\xi_i\right) + o_{\mathbb{P}}(1),$$

and hence

$$\sqrt{t}(\bar{x}_t - x^\star) \rightsquigarrow \mathsf{N}\big(0, \nabla R(x^\star)^{-1}\cdot\Sigma\cdot\nabla R(x^\star)^{-\top}\big).$$

Theorem 3.5 asserts that under mild assumptions, the deviations $\sqrt{t}(\bar{x}_t - x^\star)$ converge in distribution to a Gaussian random vector with covariance matrix $\nabla R(x^\star)^{-1}\cdot\Sigma\cdot\nabla R(x^\star)^{-\top}$. Moreover, under mild regularity conditions we may write

$$\nabla R(x^\star) = \underbrace{\mathop{\mathbb{E}}_{z\sim\mathcal{D}(x^\star)}[\nabla_x G(x^\star,z)]}_{\text{static}} + \underbrace{\frac{d}{dy}\mathop{\mathbb{E}}_{z\sim\mathcal{D}(y)}[G(x^\star,z)]\Big|_{y=x^\star}}_{\text{dynamic}}.$$

It is part of the theorem's conclusion that the matrix $\nabla R(x^\star)$ is invertible. It is worthwhile to note that the effect of the distributional shift on the asymptotic covariance is entirely captured by the second "dynamic" term in $\nabla R(x^\star)$. When the distributions $\mathcal{D}(x)$ admit a density $p(x,z) = \frac{d\mathcal{D}(x)}{d\mu}(z)$ as before, the Jacobian $\nabla R(x^\star)$ admits the simple description:

$$\nabla R(x^\star) = \mathop{\mathbb{E}}_{z\sim\mathcal{D}(x^\star)}[\nabla_x G(x^\star,z)] + \int G(x^\star,z)\nabla_x p(x^\star,z)^\top\, d\mu(z).$$

Example 3.6 (Performative prediction with location-scale families)**.** As an explicit example of Theorem 3.5, let us look at the case when $G(x,z) = \nabla_x \ell(x,z)$ is the gradient of a loss function and $\mathcal{D}(x)$ is a "linear perturbation" of a fixed base distribution $\mathcal{D}_0$. Such distributions are quite reasonable when modeling performative effects, as explained in [56]. In this case,

we have

$$z \sim \mathcal{D}(x) \quad \iff \quad z - Ax \sim \mathcal{D}_0$$

for some fixed matrix $A \in \mathbb{R}^{n \times d}$. Then a quick computation shows that we may write

$$\nabla R(x) = \mathop{\mathbb{E}}_{z \sim \mathcal{D}(x)} \left[\nabla^2_{xx} \ell(x, z) + \nabla^2_{zx} \ell(x, z) A \right]$$

under mild integrability conditions. Thus, the dynamic part of $\nabla R(x^\star)$ is governed by the product of the matrix of mixed partial derivatives $\nabla^2_{zx} \ell(x^\star, z) \in \mathbb{R}^{d \times n}$ with A. The former measures the sensitivity of the gradient $\nabla_x \ell(x^\star, z)$ at $x^\star$ to changes in the data z, while the latter measures the performative effects of the distributional shift. $\lozenge$

Example 3.7 (Multiplayer performative prediction with location-scale families)**.** More generally, let us look at the problem of multiplayer performative prediction [58]. In this case, the map G takes the form

$$G(x, z) = \big(\nabla_1 \ell_1(x, z_1), \ldots, \nabla_k \ell_k(x, z_k) \big)$$

where ℓ_i is a loss for each player i and $\nabla_i \ell_i$ denotes the gradient of ℓ_i with respect to the action x_i of player i. The distribution $\mathcal{D}(x)$ takes the product form

$$\mathcal{D}(x) = \mathcal{D}_1(x) \times \cdots \times \mathcal{D}_k(x).$$

As highlighted in [58], a natural parametric assumption is that there exist probability distributions $\mathcal{P}_i$ and matrices A_i, A_{-i} such that the following holds:

$$z_i \sim \mathcal{D}_i(x) \quad \iff \quad z_i - A_i x_i - A_{-i} x_{-i} \sim \mathcal{P}_i.$$

Here x_{-i} denotes the vector obtained from x by deleting the coordinate x_i; thus, the distribution used by player i is a "linear perturbation" of a fixed base distribution $\mathcal{P}_i$. We can interpret the matrices A_i and A_{-i} as quantifying the performative effects of player i's decisions and the rest of the players' decisions, respectively, on the distribution $\mathcal{D}_i$ governing player i's data. It is straightforward to check the expression

$$\nabla R_i(x) = \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_i(x)} \left[\nabla^2_{x x_i} \ell_i(x, z_i) + \nabla^2_{z_i x_i} \ell_i(x, z_i) [A_i, A_{-i}] \right]$$

under mild integrability conditions, where $[A_i, A_{-i}]x = A_i x_i + A_{-i} x_{-i}$. Thus, the dynamic part

of $\nabla R_i(x^\star)$ is governed by the product of the matrix of mixed partial derivatives $\nabla^2_{z_i x_i} \ell_i(x^\star, z_i)$ with $[A_i, A_{-i}]$. $\lozenge$

3.3.1 Proof of Theorem 3.5

The proof of Theorem 3.5 is based on the stochastic approximation result of Polyak and Juditsky [62, Theorem 2], which we review in Appendix D. For the remainder of this section, we impose the assumptions of Theorem 3.5.

Consider the map $R\colon \mathcal{X} \to \mathbb{R}^d$ given by $R(x) = G_x(x)$. In light of the interiority condition $x^\star \in \operatorname{int} \mathcal{X}$ of Assumption 3.4, the equilibrium point $x^\star$ is the unique solution to the equation $R(x) = 0$ on $\operatorname{int} \mathcal{X}$. Observe that the noise vector $\xi_t = \xi_{x_t}(z_t)$ satisfies the relation

$$G(x_t, z_t) = R(x_t) + \xi_t$$

and so we may write the iterates of Algorithm 5 as

$$x_{t+1} = x_t - \eta_t\big(R(x_t) + \xi_t + \zeta_t\big), \tag{3.6}$$

where

$$\zeta_t := \frac{x_t - \eta_t\big(R(x_t) + \xi_t\big) - \operatorname{proj}_{\mathcal{X}}\big(x_t - \eta_t\big(R(x_t) + \xi_t\big)\big)}{\eta_t}. \tag{3.7}$$

Our goal is to apply Theorem D.1 to the process (3.6) on the filtered probability space $(\mathcal{Z}^{\mathbb{N}}, \mathcal{B}(\mathcal{Z}^{\mathbb{N}}), \mathbb{F}, \mathbb{P})$, where $\mathbb{F} = (\mathcal{F}_t)_{t \geq 0}$ is the filtration given by

$$\mathcal{F}_0 := \{\emptyset, \mathcal{Z}^{\mathbb{N}}\} \quad \text{and} \quad \mathcal{F}_t := \{A \times \mathcal{Z}^{\mathbb{N}} \mid A \in \mathcal{B}(\mathcal{Z}^t)\} \qquad \text{for all } t \geq 1 \tag{3.8}$$

and $\mathbb{P} = \bigotimes_{i=0}^{\infty} \mathcal{D}_{x_i}$ is given by (3.4). In what follows, we establish the necessary assumptions for Theorem D.1.

To begin, we note that the map R is Lipschitz continuous and strongly monotone on $\mathcal{X}$; in particular, R is measurable.

Lemma 3.8 (Lipschitz continuity and strong monotonicity)**.** *The map R is $(\bar{L} + \gamma\beta)$-Lipschitz continuous and $(\alpha - \gamma\beta)$-strongly monotone on $\mathcal{X}$.*

Proof. Let $x, y \in \mathcal{X}$. Then

$$\|R(x) - R(y)\| \le \|G_x(x) - G_y(x)\| + \|G_y(x) - G_y(y)\| \le (\gamma\beta + \bar{L})\|x - y\|$$

as a consequence of Lemma 3.1 and the $\bar{L}$-Lipschitz continuity of $G_y(\cdot)$. Similarly,

$$\begin{aligned}\langle R(x) - R(y), x - y\rangle &= \langle G_x(x) - G_y(x), x - y\rangle + \langle G_y(x) - G_y(y), x - y\rangle \\ &\ge -\|G_x(x) - G_y(x)\|\|x - y\| + \alpha\|x - y\|^2 \\ &\ge (-\gamma\beta + \alpha)\|x - y\|^2\end{aligned}$$

as a consequence of the α-strong convexity of $G_y(\cdot)$ and Lemma 3.1. □

To establish Assumption D.1, observe first that $\sup_{t\ge 0} \mathbb{E}\|\xi_t\|^2 < \infty$ by Assumption 3.3 and Proposition 3.4. Clearly x_t is $\mathcal{F}_t$-measurable, ξ_t and ζ_t are $\mathcal{F}_{t+1}$-measurable, and ξ_t constitutes a martingale difference sequence satisfying

$$\mathbb{E}[\xi_t \,|\, \mathcal{F}_t] = \mathop{\mathbb{E}}_{z_t \sim \mathcal{D}_{x_t}}[G(x_t, z_t)] - G_{x_t}(x_t) = 0.$$

The following lemma shows that $\mathbb{E}[\xi_t \xi_t^\top \,|\, \mathcal{F}_t]$ converges to the positive semidefinite matrix

$$\Sigma = \mathop{\mathbb{E}}_{z \sim \mathcal{D}_x^\star}\left[G(x^\star, z) G(x^\star, z)^\top\right]$$

almost surely as $t \to \infty$.

Lemma 3.9 (Asymptotic covariance)**.** *As $t \to \infty$, we have*

$$\mathbb{E}\big[G(x_t, z_t)G(x_t, z_t)^\top \,|\, \mathcal{F}_t\big] \xrightarrow{\text{a.s.}} \Sigma \qquad \textit{and} \qquad \mathbb{E}\big[\xi_t \xi_t^\top \,|\, \mathcal{F}_t\big] \xrightarrow{\text{a.s.}} \Sigma.$$

Proof. Taking into account the almost sure convergence of x_t to $x^\star$ (Proposition 3.4), the uniform integrability condition (iii) of Assumption 3.4, and the Lipschitz condition (i) of Assumption 3.2, we may apply Lemma F.5 with $g = G$ along any sample path witnessing $x_t \to x^\star$ to obtain $\mathbb{E}[G(x_t, z_t)G(x_t, z_t)^\top \,|\, \mathcal{F}_t] \to \Sigma$ almost surely as $t \to \infty$. Therefore

$$\mathbb{E}\big[\xi_t \xi_t^\top \,|\, \mathcal{F}_t\big] = \mathbb{E}\big[G(x_t, z_t)G(x_t, z_t)^\top \,|\, \mathcal{F}_t\big] - R(x_t)R(x_t)^\top \xrightarrow{\text{a.s.}} \Sigma \qquad \text{as } t \to \infty$$

by virtue of the continuity of R and the relation $R(x^\star) = 0$. □

By Lemma 3.9, we have $\frac{1}{t}\sum_{i=0}^{t-1} \mathbb{E}\big[\xi_i \xi_i^\top \,|\, \mathcal{F}_i\big] \xrightarrow{\text{a.s.}} \Sigma$ as $t \to \infty$. Conditions (i) and (ii) of

Assumption D.1 are now established, and Lindeberg's condition (iii) of Assumption D.1 holds by item (iv) of Assumption 3.4. Now consider the residual vector ζ_t given by (3.7). Since $x^\star \in \operatorname{int} \mathcal{X}$ and $x_t \xrightarrow{\text{a.s.}} x^\star$ as $t \to \infty$, we have $\mathbb{P}\{\zeta_t = 0 \text{ for all but finitely many } t\} = 1$ and hence $\frac{1}{\sqrt{t}} \sum_{i=0}^{t-1} \|\zeta_i\| \xrightarrow{\text{a.s.}} 0$ as $t \to \infty$. Thus, condition (iv) of Assumption D.1 holds, and the verification of Assumption D.1 is complete.

We turn now to Assumption D.2. The first two conditions of Assumption 3.4 assert that the map R is differentiable on a neighborhood of $x^\star \in \operatorname{int} \mathcal{X}$. Since R is $(\alpha - \gamma\beta)$-strongly monotone on $\mathcal{X}$ (Lemma 3.8), it follows that we have $\langle \nabla R(x^\star)v, v\rangle \geq \alpha - \gamma\beta$ for every unit vector $v \in \mathbb{S}^{d-1}$ and hence every eigenvalue of $\nabla R(x^\star)$ has real part no smaller than $\alpha - \gamma\beta$. This is the content of the following lemma.

Lemma 3.10 (Positivity of the Jacobian)**.** *For any point $x \in \operatorname{int} \mathcal{X}$ at which R is differentiable, we have*

$$\langle \nabla R(x)v, v\rangle \geq \alpha - \gamma\beta \qquad \text{for all } v \in \mathbb{S}^{d-1} \tag{3.9}$$

and hence every eigenvalue of $\nabla R(x)$ has real part no smaller than $\alpha - \gamma\beta$. In particular, $\nabla R(x^\star)$ is positively stable.

Proof. Suppose R is differentiable at $x \in \operatorname{int} \mathcal{X}$. By Lemma 3.8, R is $(\alpha - \gamma\beta)$-strongly monotone on $\mathcal{X}$, so (3.9) follows immediately from the definitions of differentiability and strong monotonicity: for any unit vector $v \in \mathbb{S}^{d-1}$,

$$\langle \nabla R(x)v, v\rangle = t^{-2}\langle \nabla R(x+tv) - R(x), tv\rangle + o(1) \geq \alpha - \gamma\beta + o(1) \qquad \text{as } t \to 0.$$

Next, observe that (3.9) implies $\lambda_{\min}(\nabla R(x) + \nabla R(x)^\top) \geq 2(\alpha - \gamma\beta)$. Now let $w \in \mathbb{C}^d$ be a normalized eigenvector of $\nabla R(x)$ with associated eigenvalue $\lambda \in \mathbb{C}$. Letting w^* denote conjugate transpose of w, we conclude

$$2(\alpha - \gamma\beta) \leq w^*\big(\nabla R(x) + \nabla R(x)^\top\big)w = w^*\nabla R(x)w + (w^*\nabla R(x)w)^* = \lambda + \bar{\lambda} = 2(\operatorname{Re}\lambda),$$

where the first inequality follows from the Rayleigh-Ritz theorem. Thus, every eigenvalue of $\nabla R(x)$ has real part no smaller than $\alpha - \gamma\beta$. In particular, every eigenvalue of $\nabla R(x)$ has

positive real part, that is, $\nabla R(x)$ is positively stable. The last claim of the lemma follows since R is differentiable at $x^\star \in \operatorname{int} \mathcal{X}$ by Assumption 3.4. □

Next, recall $\eta_t \propto t^{-\nu}$ for some $\nu \in \left(\frac{1}{2}, 1\right)$, i.e., there exist a constant $c > 0$ and an index $T \geq 1$ such that $\eta_t = ct^{-\nu}$ for all $t \geq T$. Clearly $\eta_t = o(1)$. Moreover,

$$0 \leq \frac{\eta_t - \eta_{t+1}}{\eta_t^2} = \frac{t^\nu}{(t+1)^\nu} \cdot \frac{(t+1)^\nu - t^\nu}{c} \leq \frac{(t+1)^\nu - t^\nu}{c} \qquad \text{for all } t \geq T,$$

and since $\lim_{t\to\infty}\left((t+1)^r - t^r\right) = 0$ for any $r \in (0,1)$, we conclude

$$\frac{\eta_t - \eta_{t+1}}{\eta_t} = o(\eta_t).$$

This establishes condition (i) of Assumption D.2.

Finally, by Proposition 3.4, we have $x_t \xrightarrow{\text{a.s.}} x^\star$ and hence $\bar{x}_t \xrightarrow{\text{a.s.}} x^\star$ as $t \to \infty$, and $\sum_{t=1}^\infty t^{-1/2}\|x_t - x^\star\|^2 < \infty$ almost surely, which by Kronecker's lemma [29, Lemma 2.5.9] implies $\frac{1}{\sqrt{t}}\sum_{i=0}^{t-1}\|x_i - x^\star\|^2 \xrightarrow{\text{a.s.}} 0$ as $t \to \infty$; on the other hand,

$$R(x) - \nabla R(x^\star)(x - x^\star) = O(\|x - x^\star\|^2) \qquad \text{as } x \to x^\star$$

since ∇R is Lipschitz continuous on a neighborhood of $x^\star$ and $R(x^\star) = 0$. Therefore

$$\frac{1}{\sqrt{t}}\sum_{i=0}^{t-1}\|R(x_i) - \nabla R(x^\star)(x_i - x^\star)\| \xrightarrow{\text{a.s}} 0 \qquad \text{as } t \to \infty. \tag{3.10}$$

Since $\nabla R(x^\star)$ is positively stable (Lemma 3.10), this concludes the verification of Assumption D.2. An application of Theorem D.1 to the process (3.6) completes the proof of Theorem 3.5.

3.4 Asymptotic Optimality

In this section, we establish the local asymptotic optimality of Algorithm 5. Our result builds on classical ideas from Hájek and Le Cam [52, 75] on lower bounds for statistical estimation and the more recent work [25] on asymptotic optimality of the stochastic gradient method. Throughout, we fix a base distribution map $\mathcal{D}\colon \mathcal{X} \to P_1(\mathcal{Z})$ and a map $G\colon \mathcal{X} \times \mathcal{Z} \to \mathbb{R}^d$ satisfying Assumptions 3.1 and 3.2. We will be concerned with evaluating the performance of

estimation procedures for finding the equilibrium points induced by an adversarially-chosen sequence of small perturbations $\mathcal{D}'$ of $\mathcal{D}$, where each $\mathcal{D}'$ is "admissible" in the following sense.

Definition 3.11 (Admissible distribution map)**.** A distribution map $\mathcal{D}'\colon \mathcal{X} \to P_1(\mathcal{Z})$ is *admissible* if Assumptions 3.1 and 3.2 hold with $\mathcal{D}'$ in place of $\mathcal{D}$ (allowing for different constants γ', $\bar{L}'$, α' in place of γ, $\bar{L}$, α). For each admissible distribution map $\mathcal{D}'\colon \mathcal{X} \to P_1(\mathcal{Z})$, the corresponding equilibrium point is denoted by $x^\star_{\mathcal{D}'}$.

Let us start with some intuition before delving into the details. Roughly speaking, we aim to show that the asymptotic covariance of the normalized error $\sqrt{t}(\bar{x}_t - x^\star)$ in Theorem 1.1 is "optimal" among all algorithms for finding equilibrium points. To capture the notion of optimal covariance, a standard approach is to probe random vectors in $\mathbb{R}^d$ with nonnegative "loss" functions $\mathcal{L}\colon \mathbb{R}^d \to [0,\infty)$ that are symmetric, quasiconvex, and lower semicontinuous, interpreting the concentration of $X_1 \sim \mathcal{P}_1$ to be "better" than that of $X_2 \sim \mathcal{P}_2$ if the inequality $\mathbb{E}[\mathcal{L}(X_1)] \le \mathbb{E}[\mathcal{L}(X_2)]$ holds for all such $\mathcal{L}$; if X_1 and X_2 are square-integrable, this relation clearly entails the positive semidefinite ordering $\mathbb{E}[X_1X_1^\top] \preceq \mathbb{E}[X_2X_2^\top]$ of second-moment matrices.[2] Using this idea, we consider a local asymptotic notion of minimax risk that evaluates the performance of an arbitrary sequence of estimators on problems close to the one we wish to solve. Since our target problem models stochasticity using the base distribution map $\mathcal{D}$, we will parameterize close problems through perturbations of $\mathcal{D}$. More concretely, we will carefully construct for each $u \in \mathbb{R}^d$ a perturbation $\mathcal{D}^u$ of $\mathcal{D}$ such that, as $u \to 0$, the distribution map $\mathcal{D}^u$ is admissible with equilibrium point $x^\star_u := x^\star_{\mathcal{D}^u}$ near $x^\star$. The primary goal of this section is to show that if $\widehat{x}_k\colon \mathcal{Z}^k \to \mathbb{R}^d$ is an arbitrary sequence of estimators (i.e., $\widehat{x}_k$ is a measurable function of k observed samples) and $\mathcal{L}\colon \mathbb{R}^d \to [0,\infty)$ is symmetric, quasiconvex, and lower semicontinuous, then the following lower bound holds:

$$\underbrace{\sup_{\mathcal{I}\subset\mathbb{R}^d,\,|\mathcal{I}|<\infty} \liminf_{k\to\infty} \max_{u\in\mathcal{I}} \mathbb{E}_{P_{k,u/\sqrt{k}}}\big[\mathcal{L}\big(\sqrt{k}\big(\widehat{x}_k - x^\star_{u/\sqrt{k}}\big)\big)\big]}_{\text{local asymptotic minimax risk}} \ge \mathbb{E}[\mathcal{L}(Z)], \tag{3.11}$$

[2] Note $\mathbb{E}[X_1X_1^\top] \preceq \mathbb{E}[X_2X_2^\top]$ if and only if $\mathbb{E}[\mathcal{L}_u(X_1)] \le \mathbb{E}[\mathcal{L}_u(X_2)]$ for all $u \in \mathbb{R}^d$, where $\mathcal{L}_u\colon \mathbb{R}^d \to [0,\infty)$ is given by $\mathcal{L}_u(x) = (u^\top x)^2 = u^\top(xx^\top)u$.

where $P_{k,v} = \bigotimes_{i=0}^{k-1} \mathcal{D}_{\tilde{x}_i}^v$ denotes the distribution on $\mathcal{Z}^k$ induced by $\mathcal{D}^v$ along an arbitrary "dynamic estimation procedure" and $Z \sim \mathsf{N}(0, W^{-1}\Sigma W^{-\top})$ with Σ and W as in Theorem 1.1.

The lower bound (3.11) provides a precise expression of the optimality of the covariance of the limit distribution $\mathsf{N}(0, W^{-1}\Sigma W^{-\top})$. Moreover, we will show that equality is achieved in (3.11) upon specializing to the dynamic estimation procedure corresponding to Algorithm 5 with step sizes $\eta_k \propto k^{-\nu}$ (as in Theorem 1.1) and taking $\widehat{x}_k$ to be given by the average iterates $\bar{x}_k = \frac{1}{k}\sum_{i=1}^k x_i$, provided $\mathcal{L}$ is bounded and continuous.

To formalize the preceding discussion, we begin by defining the dynamic estimation procedure used to define the sequence of distributions $P_{k,v} = \bigotimes_{i=0}^{k-1} \mathcal{D}_{\tilde{x}_i}^v$ appearing in (3.11).

Definition 3.12 (Dynamic estimation procedure)**.** A *dynamic estimation procedure* is a sequence of measurable maps $\mathcal{A}_k \colon \mathcal{Z}^k \times \mathcal{X}^k \to \mathcal{X}$ such that for any initial point $\tilde{x}_0 \in \mathcal{X}$, the sequence of estimators $\tilde{x}_k \colon \mathcal{Z}^k \to \mathcal{X}$ defined recursively by

$$\tilde{x}_k = \mathcal{A}_k(z_0, \dots, z_{k-1}, \tilde{x}_0, \dots, \tilde{x}_{k-1}) \tag{3.12}$$

satisfies

$$\tilde{x}_k \xrightarrow{\text{a.s.}} x^\star \qquad \text{as } k \to \infty$$

with respect to the distribution $\bigotimes_{i=0}^{\infty} \mathcal{D}_{\tilde{x}_i}$ on $\mathcal{Z}^{\mathbb{N}}$.

Thus, the dynamic estimation procedure $\mathcal{A}_k$ plays the role of the decision-maker that selects the sequence of points at which to query a given distribution map; this generalizes the classical static setting wherein $z_0, z_1, \dots$ are i.i.d. samples drawn from a fixed distribution. In the dynamic setting, we are concerned with algorithms for estimating the equilibrium point $x^\star$, so it is sensible to require that the iterates $\tilde{x}_k$ produced by the recursion (3.12) with $(z_0, \dots, z_{k-1}) \sim \bigotimes_{i=0}^{k-1} \mathcal{D}_{\tilde{x}_i}$ converge almost surely to $x^\star$ as $k \to \infty$. Importantly, $\mathcal{A}_k$ is assumed to be a deterministic function of its arguments. For example, the sequence of maps $\mathcal{A}_k$ corresponding to Algorithm 5, i.e.,

$$\mathcal{A}_{k+1}(z_0, \dots, z_k, x_0, \dots, x_k) = \operatorname{proj}_{\mathcal{X}}\big(x_k - \eta_k G(x_k, z_k)\big) \qquad \text{for all } k \geq 0, \tag{3.13}$$

is a dynamic estimation procedure under the assumptions of Proposition 3.4; although this

particular map $\mathcal{A}_{k+1}$ depends directly only on the last iterate x_k and the last sample z_k, general dynamic estimation procedures may depend directly on any number of the previous samples and iterates.

We turn now to defining the perturbations $\mathcal{D}^u$ of $\mathcal{D}$ used to encode difficult instances near the target problem.

3.4.1 Tilted Distributions

Following [25] and [75, Section 25.3], for each distribution $\mathcal{D}_x := \mathcal{D}(x)$ we will construct "tilt perturbations" $\mathcal{D}_x^u$ parameterized by $u \in \mathbb{R}^d$. Henceforth, we fix an arbitrary nondecreasing C^3-smooth function $h\colon \mathbb{R} \to [-1, 1]$ such that the first three derivatives of h are bounded and $h(t) = t$ for all t in a neighborhood of zero. For each $x \in \mathcal{X}$ and $u \in \mathbb{R}^d$, the tilted distribution $\mathcal{D}_x^u \in P_1(\mathcal{Z})$ is defined by setting

$$\mathcal{D}_x^u(E) := \int_E \frac{1 + h(u^\top g_x(z))}{C_x^u}\, d\mathcal{D}_x(z) \qquad \text{for all } E \in \mathcal{B}(\mathcal{Z}), \tag{3.14}$$

where $g_x\colon \mathcal{Z} \to \mathbb{R}^d$ is $\mathcal{D}_x$-integrable with $\mathbb{E}_{z\sim\mathcal{D}_x}[g_x(z)] = 0$ and C_x^u is the normalizing constant $C_x^u = 1 + \mathbb{E}_{z\sim\mathcal{D}_x}[h(u^\top g_x(z))]$. The resulting parametric statistical model $\{\mathcal{D}_x^u \mid u \in \mathbb{R}^d\}$ has score function g_x at zero, i.e.,

$$\nabla_u \left(\log \frac{d\mathcal{D}_x^u}{d\mathcal{D}_x}(z) \right) \Bigg|_{u=0} = g_x(z).$$

Thus, the collection of functions $\{u^\top g_x\colon \mathcal{Z} \to \mathbb{R} \,|\, u \in \mathbb{R}^d\}$ forms a "tangent space" of the model $\{\mathcal{D}_x^u \,|\, u \in \mathbb{R}^d\}$ at zero [75, Example 25.15]. In the context of establishing the asymptotic optimality of Algorithm 5, we will see that the relevant score function is the noise $\xi_x(z) = G(x, z) - G_x(x)$.

To guarantee that the tilted distribution map given by $x \mapsto \mathcal{D}_x^u$ is admissible for small u, we require additional conditions on the base distribution map $\mathcal{D}$, the map G, and the function $g\colon \mathcal{X} \times \mathcal{Z} \to \mathbb{R}^d$ given by $g(x, z) = g_x(z)$. Despite being technical, these conditions (given in Assumption 3.5 and Definition 3.14 below) are mild and essentially amount to quantifying the smoothness of $\mathcal{D}$, G, and g. To quantify the smoothness of $\mathcal{D}$, we will make use of a certain set of test functions to be integrated against each distribution $\mathcal{D}_x$.

Definition 3.13 (Test functions)**.** Given a compact metric space $\mathcal{K}$, we let $\mathcal{T}(\mathcal{K}, \mathcal{Z})$ consist of all bounded measurable functions $\phi\colon \mathcal{K} \times \mathcal{Z} \to \mathbb{R}$ admitting a constant L_ϕ such that each section $\phi(\cdot, z)$ is L_ϕ-Lipschitz on $\mathcal{K}$. For any $\phi \in \mathcal{T}(\mathcal{K}, \mathcal{Z})$, we set $M_\phi := \sup|\phi|$.

Assumption 3.5. The following three conditions hold.

(i) **(Compactness)** The set $\mathcal{X}$ is compact, and the set $\mathcal{Z}$ is bounded.

(ii) **(Smooth distribution map)** There exists an increasing function $\vartheta\colon [0,\infty) \to [0,\infty)$ such that for every compact metric space $\mathcal{K}$ and test function $\phi \in \mathcal{T}(\mathcal{K}, \mathcal{Z})$, the function

$$x \mapsto \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x} \phi(y, z)$$

is C^1-smooth on $\mathcal{X}$ for each $y \in \mathcal{K}$ and the map

$$(x, y) \mapsto \nabla_x \Big(\mathop{\mathbb{E}}_{z\sim\mathcal{D}_x} \phi(y, z) \Big)$$

is $\vartheta(L_\phi + M_\phi)$-Lipschitz on $\mathcal{X} \times \mathcal{K}$.[3]

(iii) **(Lipschitz Jacobian)** There exist a measurable function $\Lambda\colon \mathcal{Z} \to [0,\infty)$ and constants $\bar{\Lambda}, \beta' \geq 0$ such that for every $z \in \mathcal{Z}$ and $x \in \mathcal{X}$, the section $G(\cdot, z)$ is $\Lambda(z)$-smooth on $\mathcal{X}$ with $\mathbb{E}_{z\sim\mathcal{D}_x}[\Lambda(z)] \leq \bar{\Lambda}$, and the section $\nabla_x G(x, \cdot)$ is β'-Lipschitz on $\mathcal{Z}$.

The first condition is imposed mainly for simplicity. The last two smoothness conditions are required in our arguments to apply dominated convergence and implicit function theorems. To illustrate with a concrete example, suppose that there exists a Borel probability measure μ on $\mathcal{Z}$ such that $\mathcal{D}(x) \ll \mu$ for all $x \in \mathcal{X}$, and consider the density $p(x, z) = \frac{d\mathcal{D}(x)}{d\mu}(z)$. If there exist constants $\Lambda_p, L_p \geq 0$ such that each section $p(\cdot, z)$ is Λ_p-smooth and $\sup_{x,z} \|\nabla_x p(x, z)\| \leq L_p$, then item (ii) of Assumption 3.5 holds with $\vartheta(s) = \max\{\Lambda_p, L_p\} \cdot s$.

Next, we specify the collection of functions $g\colon \mathcal{X} \times \mathcal{Z} \to \mathbb{R}^d$ satisfying the regularity conditions we require.

---

[3]The same conclusion then holds for all measurable maps $\phi\colon \mathcal{K} \times \mathcal{Z} \to \mathbb{R}^n$ with $n \in \mathbb{N}$, $L_\phi := \sup_z \mathrm{Lip}\big(\phi(\cdot, z)\big) < \infty$, and $M_\phi := \sup\|\phi\| < \infty$.

Definition 3.14 (Score functions)**.** Let $\mathcal{G}$ consist of all measurable functions $g\colon \mathcal{X} \times \mathcal{Z} \to \mathbb{R}^d$ satisfying the following three conditions.

(i) **(Lipschitz continuity)** There exists a constant $\beta_g \geq 0$ such that for every $x \in \mathcal{X}$, the section $g(x, \cdot)$ is β_g-Lipschitz continuous on $\mathcal{Z}$.

(ii) **(Unbiasedness)** For all $x \in \mathcal{X}$, we have $\mathbb{E}_{z\sim\mathcal{D}_x}[g(x, z)] = 0$.

(iii) **(Smoothness)** There exist a measurable function $\Lambda_g\colon \mathcal{Z} \to [0, \infty)$ and constants $\bar{\Lambda}_g, \beta_g' \geq 0$ such that for every $z \in \mathcal{Z}$ and $x \in \mathcal{X}$, the section $g(\cdot, z)$ is $\Lambda_g(z)$-smooth on $\mathcal{X}$ with $\mathbb{E}_{z\sim\mathcal{D}_x}[\Lambda_g(z)] \leq \bar{\Lambda}_g$, and the section $\nabla_x g(x, \cdot)$ is β_g'-Lipschitz continuous on $\mathcal{Z}$.

For our purposes, the most important map in $\mathcal{G}$ will be the noise

$$\xi(x, z) := G(x, z) - G_x(x), \tag{3.15}$$

which belongs to $\mathcal{G}$ as a consequence of Assumptions 3.2 and 3.5 and Lemma E.4.

Henceforth, we fix $g \in \mathcal{G}$ and take $g_x(z) = g(x, z)$ in (3.14), thereby defining the tilted distribution map $\mathcal{D}^u\colon \mathcal{X} \to P_1(\mathcal{Z})$ given by $x \mapsto \mathcal{D}_x^u$. The following lemma guarantees that if Assumptions 3.1, 3.2, and 3.5 hold, then $\mathcal{D}^u$ is admissible for all u in a neighborhood $\mathcal{U}$ of zero; the proof, which we defer to Section E.1, provides constants γ^u, $\bar{L}^u$, α^u that fulfill Assumptions 3.1 and 3.2 for $\mathcal{D}^u$ and deviate from γ, $\bar{L}$, α by $O(\|u\|)$ as $u \to 0$.

Lemma 3.15 (Tilted distributions are admissible)**.** *Suppose that Assumptions 3.1, 3.2, and 3.5 hold. Then there exists a neighborhood $\mathcal{U}$ of zero such that for all $u \in \mathcal{U}$, the map $\mathcal{D}^u$ is admissible.*

For ease of notation, we set

$$x_u^\star := x_{\mathcal{D}^u}^\star$$

for each u in the neighborhood $\mathcal{U}$ of zero furnished by Lemma 3.15. With the preceding definitions in place, we are now ready to state the main result of this section.

Theorem 3.16 (Asymptotic optimality)**.** *Suppose that Assumptions 3.1, 3.2, and 3.5 hold with the equilibrium point $x^\star$ lying in the interior of $\mathcal{X}$, and suppose $g = \xi$. Let $\mathcal{A}_k \colon \mathcal{Z}^k \times \mathcal{X}^k \to \mathcal{X}$ be a dynamic estimation procedure, fix an initial point $\tilde{x}_0 \in \mathcal{X}$, and for each $u \in \mathbb{R}^d$, let $P_{k,u} = \bigotimes_{i=0}^{k-1} \mathcal{D}_{\tilde{x}_i}^u$ denote the distribution on $\mathcal{Z}^k$ induced by $\mathcal{D}^u$ along the sequence (3.12). Let $\widehat{x}_k \colon \mathcal{Z}^k \to \mathbb{R}^d$ be any sequence of estimators, and let $\mathcal{L} \colon \mathbb{R}^d \to [0, \infty)$ be symmetric, quasiconvex, and lower semicontinuous.*

(i) ***(Lower bound)*** *The following lower bound on the local asymptotic minimax risk holds:*

$$\sup_{\mathcal{I} \subset \mathbb{R}^d,\, |\mathcal{I}| < \infty} \liminf_{k \to \infty} \max_{u \in \mathcal{I}} \mathbb{E}_{P_{k,u/\sqrt{k}}} \left[\mathcal{L}\left(\sqrt{k} \left(\widehat{x}_k - x^\star_{u/\sqrt{k}} \right) \right) \right] \geq \mathbb{E}[\mathcal{L}(Z)], \tag{3.16}$$

where $Z \sim \mathsf{N}\left(0, W^{-1} \Sigma W^{-\top}\right)$ with

$$\Sigma = \mathop{\mathbb{E}}_{z \sim \mathcal{D}_{x^\star}} \left[G(x^\star, z) G(x^\star, z)^\top \right] \quad \text{and} \quad W = \mathop{\mathbb{E}}_{z \sim \mathcal{D}_{x^\star}} \left[\nabla_x G(x^\star, z) \right] + \frac{d}{dy} \mathop{\mathbb{E}}_{z \sim \mathcal{D}_y} \left[G(x^\star, z) \right] \Big|_{y = x^\star}.$$

(ii) ***(Achieving the bound)*** *If $\mathcal{A}_k$ is the dynamic estimation procedure (3.13) corresponding to Algorithm 5 with initial point $x_0 = \tilde{x}_0$ and step sizes $\eta_k \propto k^{-\nu}$ for some $\nu \in \left(\frac{1}{2}, 1\right)$, and if the sequence of estimators $\widehat{x}_k$ is given by the average iterates $\bar{x}_k = \frac{1}{k} \sum_{i=1}^k x_i$, then equality holds in (3.16) whenever $\mathcal{L}$ is bounded and continuous.*

Most importantly, observe that the distribution of Z in Theorem 3.16 coincides with the asymptotic distribution of $\sqrt{t}(\bar{x}_t - x^\star)$ in Theorem 3.5, thereby justifying asymptotic optimality of the stochastic forward-backward method (Algorithm 5). The lower bound in Theorem 3.16 provides a decision-dependent analogue of the asymptotic optimality result established in [25, Theorem 1].

Remark 3.17 (Convergence of equilibria and tilted distributions)**.** In the setting of Theorem 3.16, it is easy to see that the following approximations hold:

$$\|x^\star_u - x^\star\| = O(\|u\|) \qquad \text{as } u \to 0 \tag{3.17}$$

and

$$\sup_{x \in \mathcal{X}} W_1(\mathcal{D}_x^u, \mathcal{D}_x) = O(\|u\|) \qquad \text{as } u \to 0. \tag{3.18}$$

Indeed, we will show in the forthcoming Lemma 3.24 that the map $u \mapsto x_u^\star$ is C^1-smooth on a neighborhood of zero, which implies (3.17) by the mean value theorem.

To verify the approximation (3.18), note first that for any 1-Lipschitz function $\phi \in \mathrm{Lip}_1(\mathcal{Z})$, the translate $\bar{\phi} = \phi - \inf \phi$ is bounded by $\mathrm{diam}(\mathcal{Z})$, and

$$\mathop{\mathbb{E}}_{z\sim\mathcal{D}_x^u}[\phi(z)] - \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x}[\phi(z)] = \frac{1}{C_x^u}\mathop{\mathbb{E}}_{z\sim\mathcal{D}_x}\big[\bar{\phi}(z)\big(1 + h\big(u^\top g_x(z)\big)\big)\big] - \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x}[\bar{\phi}(z)] \tag{3.19}$$

for any $x \in \mathcal{X}$ and $u \in \mathbb{R}^d$. Further, Lemma E.2 shows $\sup_{x\in\mathcal{X}} \mathbb{E}_{z\sim\mathcal{D}_x}|h(u^\top g_x(z))| = O(\|u\|)$ for all $u \in \mathbb{R}^d$ and $\sup_{x\in\mathcal{X}} \frac{1}{C_x^u} = 1 + O(\|u\|^3)$ as $u \to 0$, so (3.19) implies

$$\sup_{x\in\mathcal{X}} W_1(\mathcal{D}_x^u, \mathcal{D}_x) \le \mathrm{diam}(\mathcal{Z}) \cdot \sup_{x\in\mathcal{X}} \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x}|h(u^\top g_x(z))| + O(\|u\|^3) = O(\|u\|) \qquad \text{as } u \to 0.$$

This establishes (3.18), which in particular asserts that the collection of tilted distribution maps $\{\mathcal{D}^u\}_{u\in\mathbb{R}^d}$ converges uniformly to $\mathcal{D}$ as $u \to 0$. ◊

Remark 3.18 (f-divergence of tilted distributions)**.** We can also quantify the variation of the tilted distribution map $\mathcal{D}^u$ from the base distribution map $\mathcal{D}$ via f-divergence. Let $f\colon (0,\infty) \to \mathbb{R}$ be any C^3-smooth convex function with $f(1) = 0$. Then for any distribution map $\mathcal{D}'\colon \mathcal{X} \to P_1(\mathcal{Z})$, we may define the similarity measure

$$\Delta_f(\mathcal{D}' \,\|\, \mathcal{D}) := \sup_{x\in\mathcal{X}} \Delta_f(\mathcal{D}'_x \,\|\, \mathcal{D}_x),$$

where $\Delta_f(\mathcal{D}'_x \,\|\, \mathcal{D}_x)$ denotes the usual f-divergence of $\mathcal{D}'_x$ from $\mathcal{D}_x$ given by (3.2). The following approximation holds:

$$\Delta_f(\mathcal{D}^u \,\|\, \mathcal{D}) = O(\|u\|^2) \qquad \text{as } u \to 0. \tag{3.20}$$

To verify (3.20), observe that for all sufficiently small $u \in \mathbb{R}^d$ and all $x \in \mathcal{X}$, we have

$$\begin{aligned}
\Delta_f(\mathcal{D}_x^u \,\|\, \mathcal{D}_x) &= \int f\left(\frac{1 + h(u^\top g_x(z))}{C_x^u}\right) d\mathcal{D}_x(z) \\
&= \int f\left(\frac{1 + u^\top g_x(z)}{C_x^u}\right) d\mathcal{D}_x(z) &(3.21)\\
&= \frac{f''(1)}{2} u^\top \left(\mathop{\mathbb{E}}_{z\sim\mathcal{D}_x} g_x(z) g_x(z)^\top\right) u + r_x(u), &(3.22)
\end{aligned}$$

where $\sup_{x\in\mathcal{X}} |r_x(u)| = o(\|u\|^2)$ as $u \to 0$. The equality (3.21) holds for all sufficiently small $u \in \mathbb{R}^d$ and all $x \in \mathcal{X}$ because g is uniformly bounded over $\mathcal{X} \times \mathcal{Z}$ (see Lemma E.1)

and $h(t) = t$ for all t in a neighborhood of zero. The equality (3.22) follows from a second-order approximation and the dominated convergence theorem; we defer the details to Lemma E.3. Another appeal to the uniform boundedness of g yields a constant $a \geq 0$ for which $\sup_{x\in\mathcal{X}} \|\mathbb{E}_{z\sim\mathcal{D}_x}[g_x(z)g_x(z)^\top]\|_{\mathrm{op}} \leq a$. Further, given any $b > 0$, there is a neighborhood U of zero such that $\sup_{x\in\mathcal{X},\, u\in U} \|u\|^{-2}|r_x(u)| \leq b$. Therefore $\Delta_f(\mathcal{D}^u \,\|\, \mathcal{D}) \leq (\frac{a}{2}f''(1) + b)\|u\|^2$ for all sufficiently small $u \in \mathbb{R}^d$.

In light of (3.20), one may obtain from (3.16) a less refined local asymptotic minimax bound in terms of the "admissible neighborhoods" $B_f(\varepsilon)$ of $\mathcal{D}$ defined for each $\varepsilon > 0$ by

$$B_f(\varepsilon) := \big\{\mathcal{D}' \colon \mathcal{X} \to P_1(\mathcal{Z}) \mid \mathcal{D}' \text{ is admissible and } \Delta_f(\mathcal{D}' \,\|\, \mathcal{D}) \leq \varepsilon\big\},$$

namely,

$$\lim_{c\to\infty} \liminf_{k\to\infty} \sup_{\mathcal{D}'\in B_f(c/k)} \mathbb{E}_{P'_k}\big[\mathcal{L}\big(\sqrt{k}(\widehat{x}_k - x^\star_{\mathcal{D}'})\big)\big] \geq \mathbb{E}[\mathcal{L}(Z)], \tag{3.23}$$

where $P'_k = \bigotimes_{i=0}^{k-1} \mathcal{D}'_{\tilde{x}_i}$ denotes the distribution on $\mathcal{Z}^k$ induced by $\mathcal{D}'$ along the sequence (3.12). Indeed, (3.20) facilitates the elementary estimation

$$\begin{aligned}
&\lim_{c\to\infty} \liminf_{k\to\infty} \sup_{\mathcal{D}'\in B_f(c/k)} \mathbb{E}_{P'_k}\big[\mathcal{L}\big(\sqrt{k}(\widehat{x}_k - x^\star_{\mathcal{D}'})\big)\big] \\
&\qquad \geq \lim_{c\to\infty} \liminf_{k\to\infty} \sup_{\|u\|\leq c/\sqrt{k}} \mathbb{E}_{P_{k,u}}\big[\mathcal{L}\big(\sqrt{k}(\widehat{x}_k - x^\star_u)\big)\big] \\
&\qquad \geq \sup_{\mathcal{I}\subset\mathbb{R}^d,\, |\mathcal{I}|<\infty} \liminf_{k\to\infty} \max_{u\in\mathcal{I}} \mathbb{E}_{P_{k,u/\sqrt{k}}}\big[\mathcal{L}\big(\sqrt{k}\big(\widehat{x}_k - x^\star_{u/\sqrt{k}}\big)\big)\big]
\end{aligned}$$

and hence (3.16) implies (3.23). ◊

3.4.2 Proof of Theorem 3.16

The proof of Theorem 3.16 is based on the classical Hájek-Le Cam minimax theorem. To state this result, we require several standard definitions from statistics. In the sequel, we let $\{Q_{k,u} \,|\, u \in \mathbb{R}^d\}$ denote a sequence of parametric statistical models, where $Q_{k,u}$ is a probability measure on $(\Omega_k, \mathcal{S}_k)$ such that $Q_{k,u} \ll Q_{k,0}$ for each $k \in \mathbb{N}$ and $u \in \mathbb{R}^d$; following [76], we write either $X_k \overset{u}{\rightsquigarrow} X$ or $X_k \overset{u}{\rightsquigarrow} \mathsf{D}$ to indicate that a sequence of random vectors $X_k \colon \Omega_k \to \mathbb{R}^m$ converges in distribution to a random vector $X \sim \mathsf{D}$ with respect

to $Q_{k,u}$, i.e., $\lim_{k\to\infty} \mathbb{E}_{Q_{k,u}}[\varphi(X_k)] = \mathbb{E}_{X\sim \mathsf{D}}[\varphi(X)]$ for every bounded continuous function $\varphi\colon \mathbb{R}^m \to \mathbb{R}$.

Definition 3.19 (Locally asymptotically normal)**.** The sequence $\{Q_{k,u} \,|\, u \in \mathbb{R}^d\}$ is *locally asymptotically normal (LAN) with precision V at zero* if there exist a sequence of random vectors $Z_k\colon \Omega_k \to \mathbb{R}^d$ and a positive semidefinite matrix $V \in \mathbb{R}^{d\times d}$ such that $Z_k \overset{0}{\rightsquigarrow} \mathsf{N}(0, V)$ and, for each $u \in \mathbb{R}^d$,

$$\log \frac{dQ_{k,u}}{dQ_{k,0}} = u^\top Z_k - \frac{1}{2} u^\top V u + o_{Q_{k,0}}(1). \tag{3.24}$$

Definition 3.20 (Regular mapping sequence)**.** A sequence of mappings $\Gamma_k\colon \mathbb{R}^d \to \mathbb{R}^n$ is *regular with derivative $\dot\Gamma$ at zero* if there exists a matrix $\dot\Gamma \in \mathbb{R}^{n\times d}$ satisfying

$$\lim_{k\to\infty} \sqrt{k}\big(\Gamma_k(u) - \Gamma_k(0)\big) = \dot\Gamma u \quad \text{for all } u \in \mathbb{R}^d.$$

Example 3.21. Given any $\psi\colon \mathbb{R}^d \to \mathbb{R}^n$ such that ψ is differentiable at zero, the induced mapping sequence $\Gamma_k\colon \mathbb{R}^d \to \mathbb{R}^n$ given by $\Gamma_k(u) = \psi(u/\sqrt{k})$ is clearly regular with derivative $\dot\Gamma = \nabla\psi(0)$ at zero. We will see that this construction provides the relevant regular mapping sequence for establishing Theorem 3.16 by taking $\psi(u) = x^\star_u$ on a neighborhood of zero. $\lozenge$

Equipped with the preceding definitions, we are ready to state the following version of the Hájek-Le Cam minimax theorem, which appears for example in [25, Lemma 8.2] and [76, Theorem 3.11.5].

Theorem 3.22 (Local asymptotic minimax bound)**.** *Let $\{Q_{k,u} \,|\, u \in \mathbb{R}^d\}$ be locally asymptotically normal with precision V at zero, $\Gamma_k\colon \mathbb{R}^d \to \mathbb{R}^n$ be a regular mapping sequence with derivative $\dot\Gamma$ at zero, and $\mathcal{L}\colon \mathbb{R}^n \to [0,\infty)$ be symmetric, quasiconvex, and lower semicontinuous. Then, for any sequence of estimators $T_k\colon \Omega_k \to \mathbb{R}^n$, we have*

$$\sup_{\mathcal{I}\subset\mathbb{R}^d,\, |\mathcal{I}|<\infty} \liminf_{k\to\infty} \max_{u\in\mathcal{I}} \mathbb{E}_{Q_{k,u}}\big[\mathcal{L}\big(\sqrt{k}\big(T_k - \Gamma_k(u)\big)\big)\big] \geq \mathbb{E}[\mathcal{L}(Z)], \tag{3.25}$$

where $Z \sim \mathsf{N}(0, \dot\Gamma(V + \lambda I)^{-1}\dot\Gamma^\top)$ for any $\lambda > 0$; if V is invertible, then (3.25) also holds with $Z \sim \mathsf{N}(0, \dot\Gamma V^{-1}\dot\Gamma^\top)$.

To establish the lower bound (3.16) in Theorem 3.16, we will apply Theorem 3.22 as follows. Suppose henceforth that Assumptions 3.1, 3.2, and 3.5 hold with the equilibrium point $x^\star$ lying in the interior of $\mathcal{X}$. Let $\mathcal{A}_k \colon \mathcal{Z}^k \times \mathcal{X}^k \to \mathcal{X}$ be a dynamic estimation procedure and fix an initial point $\tilde{x}_0 \in \mathcal{X}$ and a score function $g \in \mathcal{G}$. For each $k \in \mathbb{N}$ and $u \in \mathbb{R}^d$, we let

$$P_{k,u} := \bigotimes_{i=0}^{k-1} \mathcal{D}^u_{\tilde{x}_i} \tag{3.26}$$

denote the distribution on $\mathcal{Z}^k$ induced by $\mathcal{D}^u$ along the sequence (3.12), and we set

$$Q_{k,u} := P_{k,u/\sqrt{k}}. \tag{3.27}$$

Further, we define $\psi \colon \mathbb{R}^d \to \mathbb{R}^d$ by

$$\psi(u) = \begin{cases} x^\star_u & \text{if } u \in \mathcal{U} \\ 0 & \text{otherwise} \end{cases}$$

and take $\Gamma_k \colon \mathbb{R}^d \to \mathbb{R}^d$ to be the induced mapping sequence given by

$$\Gamma_k(u) = \psi(u/\sqrt{k});$$

since $\mathcal{U}$ is a neighborhood of zero, it follows that for each $u \in \mathbb{R}^d$, we have $\Gamma_k(u) = x^\star_{u/\sqrt{k}}$ for all but finitely many $k \in \mathbb{N}$.

We now state two key lemmas that will allow us to apply Theorem 3.22; their proofs are deferred to Sections E.2 and E.3, respectively. The first lemma verifies that $\{Q_{k,u} \,|\, u \in \mathbb{R}^d\}$ is locally asymptotically normal at zero with precision

$$\Sigma_g := \mathop{\mathbb{E}}_{z \sim \mathcal{D}_{x^\star}} \left[g_{x^\star}(z) g_{x^\star}(z)^\top \right],$$

while the second lemma shows that ψ is C^1-smooth on a neighborhood of zero and computes $\nabla\psi(0) = -W^{-1}\Sigma_{g,G}^\top$, where

$$\Sigma_{g,G} := \mathop{\mathbb{E}}_{z \sim \mathcal{D}_{x^\star}} \left[g_{x^\star}(z) G(x^\star, z)^\top \right].$$

Lemma 3.23 (LAN). *Let $Z_k \colon \mathcal{Z}^k \to \mathbb{R}^d$ be the sequence of random vectors given by*

$$Z_k = \frac{1}{\sqrt{k}} \sum_{i=0}^{k-1} g_{\tilde{x}_i}(z_i).$$

Then $Z_k \overset{0}{\rightsquigarrow} \mathsf{N}(0, \Sigma_g)$, where $\overset{0}{\rightsquigarrow}$ denotes convergence in distribution with respect to $Q_{k,0}$.

Moreover, for each $u \in \mathbb{R}^d$,

$$\log \frac{dQ_{k,u}}{dQ_{k,0}} = u^\top Z_k - \frac{1}{2} u^\top \Sigma_g u + o_{Q_{k,0}}(1).$$

Hence $\{Q_{k,u} \,|\, u \in \mathbb{R}^d\}$ *is locally asymptotically normal with precision* Σ_g *at zero.*

Lemma 3.24 (Smooth equilibrium perturbation)**.** *The map* ψ *is* C^1*-smooth on a neighborhood of zero with* $\nabla\psi(0) = -W^{-1}\Sigma_{g,G}^\top$*. Hence* Γ_k *is regular with derivative* $\dot{\Gamma} = -W^{-1}\Sigma_{g,G}^\top$ *at zero.*

Importantly, taking g to be the noise map ξ given by (3.15) and noting $\xi(x^\star, z) = G(x^\star, z)$ yields

$$\Sigma_\xi = \Sigma_{\xi,G} = \mathop{\mathbb{E}}_{z \sim \mathcal{D}_{x^\star}} \big[G(x^\star, z) G(x^\star, z)^\top \big] = \Sigma.$$

We are now in position to apply Theorem 3.22. Let $\mathcal{L}\colon \mathbb{R}^d \to [0, \infty)$ be symmetric, quasiconvex, and lower semicontinuous, $\widehat{x}_k\colon \mathcal{Z}^k \to \mathbb{R}^d$ be any sequence of estimators, and suppose henceforth that $g = \xi$. Invoking Lemmas 3.23 and 3.24 and applying Theorem 3.22 yields

$$\begin{aligned} &\sup_{\mathcal{I} \subset \mathbb{R}^d,\, |\mathcal{I}| < \infty} \liminf_{k \to \infty} \max_{u \in \mathcal{I}} \mathbb{E}_{P_{k,u/\sqrt{k}}} \big[\mathcal{L}\big(\sqrt{k} \big(\widehat{x}_k - x^\star_{u/\sqrt{k}} \big) \big) \big] \\ &\qquad = \sup_{\mathcal{I} \subset \mathbb{R}^d,\, |\mathcal{I}| < \infty} \liminf_{k \to \infty} \max_{u \in \mathcal{I}} \mathbb{E}_{Q_{k,u}} \big[\mathcal{L}\big(\sqrt{k} \big(\widehat{x}_k - \Gamma_k(u) \big) \big) \big] \geq \mathbb{E}[\mathcal{L}(Z_\lambda)], \end{aligned} \tag{3.28}$$

where $Z_\lambda \sim \mathsf{N}(0, W^{-1}\Sigma(\Sigma + \lambda I)^{-1}\Sigma W^{-\top})$ for any $\lambda > 0$.

Letting $\lambda \downarrow 0$ in (3.28) establishes (3.16). Indeed, let $\Sigma = AA^\top$ be a Cholesky decomposition of Σ and observe that the pseudoinverse identities $A^\dagger = \lim_{\lambda \downarrow 0} A^\top \big(AA^\top + \lambda I\big)^{-1}$ and $AA^\dagger A = A$ imply

$$\lim_{\lambda \downarrow 0} \Sigma \big(\Sigma + \lambda I\big)^{-1} \Sigma = A \Big(\lim_{\lambda \downarrow 0} A^\top \big(AA^\top + \lambda I\big)^{-1} \Big) AA^\top = \big(AA^\dagger A\big) A^\top = AA^\top = \Sigma.$$

Thus, upon setting $\widetilde{\Sigma}_\lambda := W^{-1}\Sigma(\Sigma + \lambda I)^{-1}\Sigma W^{-\top}$ and $\widetilde{\Sigma} := W^{-1}\Sigma W^{-\top}$, we have $\widetilde{\Sigma}_\lambda \to \widetilde{\Sigma}$ as $\lambda \downarrow 0$. Further, for all $0 < \lambda_2 \leq \lambda_1$, we have $\exp\big(-\frac{1}{2} v^\top \widetilde{\Sigma}_{\lambda_2}^\dagger v\big) \geq \exp\big(-\frac{1}{2} v^\top \widetilde{\Sigma}_{\lambda_1}^\dagger v\big)$ for all $v \in \mathbb{R}^d$. Since the densities corresponding to $Z_\lambda \sim \mathsf{N}(0, \widetilde{\Sigma}_\lambda)$ and $Z \sim \mathsf{N}(0, \widetilde{\Sigma})$ with respect to the Lebesgue measure restricted to $S := \operatorname{range} \widetilde{\Sigma}$ are given by

$$p_\lambda(v) := \frac{\exp\big(-\frac{1}{2} v^\top \widetilde{\Sigma}_\lambda^\dagger v\big)}{\sqrt{(2\pi)^r \det{}^*\big(\widetilde{\Sigma}_\lambda\big)}} \qquad \text{and} \qquad p(v) := \frac{\exp\big(-\frac{1}{2} v^\top \widetilde{\Sigma}^\dagger v\big)}{\sqrt{(2\pi)^r \det{}^*\big(\widetilde{\Sigma}\big)}},$$

where r is the rank of Σ, we may therefore apply the monotone convergence theorem to obtain

$$\begin{aligned}\lim_{\lambda\downarrow 0}\mathbb{E}[\mathcal{L}(Z_\lambda)] &= \lim_{\lambda\downarrow 0}\frac{1}{\sqrt{(2\pi)^r\det{}^*\big(\widetilde{\Sigma}_\lambda\big)}}\int_S \mathcal{L}(v)\exp\big(-\tfrac{1}{2}v^\top\widetilde{\Sigma}_\lambda^\dagger v\big)\,dv\\ &= \frac{1}{\sqrt{(2\pi)^r\det{}^*\big(\widetilde{\Sigma}\big)}}\int_S \mathcal{L}(v)\exp\big(-\tfrac{1}{2}v^\top\widetilde{\Sigma}^\dagger v\big)\,dv\\ &= \mathbb{E}[\mathcal{L}(Z)].\end{aligned}$$

Hence (3.28) entails (3.16).

To prove the final claim of Theorem 3.16, we proceed by establishing a type of asymptotic equivariance (as in [75, Lemma 8.14]) of the average SFB iterates.

Lemma 3.25 (Asymptotic equivariance)**.** *Let $\mathcal{A}_k$ be the dynamic estimation procedure (3.13) corresponding to Algorithm 5 with initial point $x_0 = \tilde{x}_0$ and step sizes $\eta_k \propto k^{-\nu}$ for some $\nu \in \big(\frac{1}{2}, 1\big)$. Then the average iterates $\bar{x}_k = \frac{1}{k}\sum_{i=1}^k x_i$ are asymptotically equivariant-in-law with respect to $\{Q_{k,u}\,|\,u\in\mathbb{R}^d\}$ for estimating $x^\star$, that is, for each $u\in\mathbb{R}^d$,*

$$\sqrt{k}\big(\bar{x}_k - \Gamma_k(u)\big) \overset{u}{\rightsquigarrow} \mathsf{N}\big(0, W^{-1}\Sigma W^{-\top}\big). \tag{3.29}$$

Proof. Lemma 3.23 shows that the sequence of random vectors $Z_k\colon \mathcal{Z}^k \to \mathbb{R}^d$ given by

$$Z_k = \frac{1}{\sqrt{k}}\sum_{i=0}^{k-1}\xi_{x_i}(z_i)$$

satisfies

$$Z_k \overset{0}{\rightsquigarrow} \mathsf{N}(0,\Sigma), \tag{3.30}$$

and, for each $u\in\mathbb{R}^d$,

$$\log\frac{dQ_{k,u}}{dQ_{k,0}} = u^\top Z_k - \frac{1}{2}u^\top\Sigma u + o_{Q_{k,0}}(1). \tag{3.31}$$

Moreover, Theorem 3.5 reveals

$$\sqrt{k}(\bar{x}_k - x^\star) = -W^{-1}Z_k + o_{Q_{k,0}}(1). \tag{3.32}$$

Now let $\bar{Z}\sim\mathsf{N}(0,\Sigma)$, fix $u\in\mathbb{R}^d$, and consider the affine map $\varphi\colon\mathbb{R}^d\to\mathbb{R}^{d+1}$ given by

$$\varphi(z) = \begin{pmatrix}-W^{-1}\\ u^\top\end{pmatrix}z + \begin{pmatrix}0\\ -\frac{1}{2}u^\top\Sigma u\end{pmatrix}.$$

Then (3.30) implies $\varphi(Z_k) \overset{0}{\rightsquigarrow} \varphi(\bar{Z})$ and hence

$$\begin{pmatrix} \sqrt{k}(\bar{x}_k - x^\star) \\ \log \frac{dQ_{k,u}}{dQ_{k,0}} \end{pmatrix} \overset{0}{\rightsquigarrow} \begin{pmatrix} -W^{-1}\bar{Z} \\ u^\top \bar{Z} - \frac{1}{2} u^\top \Sigma u \end{pmatrix} \sim \mathsf{N}\left(\begin{pmatrix} 0 \\ -\frac{1}{2} u^\top \Sigma u \end{pmatrix}, \begin{pmatrix} W^{-1}\Sigma W^{-\top} & -W^{-1}\Sigma u \\ -u^\top \Sigma W^{-\top} & u^\top \Sigma u \end{pmatrix} \right) \tag{3.33}$$

by virtue of (3.31), (3.32), and the continuous mapping theorem [75, Theorems 2.3 and 2.7].

In light of (3.33), Le Cam's Third Lemma [75, Example 6.7] asserts

$$\sqrt{k}(\bar{x}_k - x^\star) \overset{u}{\rightsquigarrow} \mathsf{N}\big(-W^{-1}\Sigma u, W^{-1}\Sigma W^{-\top}\big). \tag{3.34}$$

On the other hand, Lemma 3.24 shows that Γ_k is a regular mapping sequence with derivative $\dot{\Gamma} = -W^{-1}\Sigma$ at zero, so

$$\sqrt{k}\big(x^\star - \Gamma_k(u)\big) = -\sqrt{k}\big(\Gamma_k(u) - \Gamma_k(0)\big) \to W^{-1}\Sigma u \qquad \text{as } k \to \infty. \tag{3.35}$$

Combining (3.34) and (3.35) yields (3.29). □

Finally, suppose that the assumptions of Lemma 3.25 hold. Let $\varphi\colon \mathbb{R}^d \to \mathbb{R}$ be any bounded continuous function and $Z \sim \mathsf{N}(0, W^{-1}\Sigma W^{-\top})$. Then (3.29) directly implies that for every finite subset $\mathcal{I} \subset \mathbb{R}^d$, we have

$$\lim_{k\to\infty} \max_{u\in\mathcal{I}} \mathbb{E}_{P_{k,u/\sqrt{k}}}\big[\varphi\big(\sqrt{k}\big(\bar{x}_k - x^\star_{u/\sqrt{k}}\big)\big)\big] = \max_{u\in\mathcal{I}} \lim_{k\to\infty} \mathbb{E}_{Q_{k,u}}\big[\varphi\big(\sqrt{k}\big(\bar{x}_k - \Gamma_k(u)\big)\big)\big] = \mathbb{E}[\varphi(Z)].$$

Hence

$$\sup_{\mathcal{I}\subset\mathbb{R}^d,\, |\mathcal{I}|<\infty} \liminf_{k\to\infty} \max_{u\in\mathcal{I}} \mathbb{E}_{P_{k,u/\sqrt{k}}}\big[\varphi\big(\sqrt{k}\big(\bar{x}_k - x^\star_{u/\sqrt{k}}\big)\big)\big] = \mathbb{E}[\varphi(Z)],$$

thereby demonstrating equality in (3.16) whenever $\mathcal{L}$ is bounded and continuous. The proof of Theorem 3.16 is complete.

Appendix A

AVERAGING LEMMA

We make use a variation of the averaging strategy used in [32]; our approach here follows [24, Appendix A] and [48, 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

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

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

$$1 + \sum_{i=1}^{t} \frac{\hat{\rho}_i}{\hat{\Gamma}_i} = \frac{1}{\hat{\Gamma}_t}. \tag{A.2}$$

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

$$\hat{x}_t = \hat{\Gamma}_t \left(x_0 + \sum_{i=1}^{t} \frac{\hat{\rho}_i}{\hat{\Gamma}_i} x_i \right). \tag{A.3}$$

The following lemma provides the key estimate we use.

Lemma A.1 (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 (A.3) expresses $\hat{x}_t$ as a convex combination of $x_0, \ldots, x_t$ by virtue of (A.2). Therefore, by the convexity of h, we may apply Jensen's inequality to obtain

$$h(\hat{x}_t) \leq \hat{\Gamma}_t \left(h(x_0) + \sum_{i=1}^{t} \frac{\hat{\rho}_i}{\hat{\Gamma}_i} h(x_i) \right). \tag{A.4}$$

On the other hand, for each $i \geq 1$, we may divide the recursion (A.1) 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 (A.4) 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

PROOFS DEFERRED FROM CHAPTER 2

B.1 Proof of Theorem 2.35

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 (2.7) in Theorem 2.21. Taking into account $\eta_k \geq \bar{\eta}_\star$ and our selection of c, Theorem 2.21 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

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

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

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

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

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

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 2.39

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 2.38 and inequality (2.13) directly imply

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

We will verify by induction that the estimate $\mathbb{E}\big[\psi_{t_k}(X_k) - \psi^\star_{t_k}\big] \leq 11\widehat{G}_{k-1}$ holds for all indices $k \geq 1$. To see the base case, observe that inequality (2.15) facilitates the estimation

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

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

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

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 2.40

Fix $t \geq 1$. Given $i \geq 1$ and $\alpha > 0$, the μ-strong convexity of ψ_t and Lemma 2.36 imply

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

hence

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

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

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

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

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

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

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

Hence the following estimation holds:

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

This completes the proof.

B.4 Proof of Theorem 2.45

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 (2.9), it follows that for any specified index k, the estimate

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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.

Appendix C

PROOFS DEFERRED FROM SECTIONS 3.2 AND 3.3

C.1 Proof of Lemma 3.1

For any $x, x', y \in \mathcal{X}$, we successively estimate

$$\begin{aligned}
\|G_x(y) - G_{x'}(y)\| &= \left\| \mathop{\mathbb{E}}_{z\sim\mathcal{D}(x)} G(y,z) - \mathop{\mathbb{E}}_{z\sim\mathcal{D}(x')} G(y,z) \right\| \\
&= \sup_{\|v\|\le 1} \left\{ \mathop{\mathbb{E}}_{z\sim\mathcal{D}(x)} \langle G(y,z), v\rangle - \mathop{\mathbb{E}}_{z\sim\mathcal{D}(x')} \langle G(y,z), v\rangle \right\} \\
&\le \beta \cdot W_1\big(\mathcal{D}(x), \mathcal{D}(x')\big) \qquad \text{(C.1)} \\
&\le \beta\gamma \cdot \|x - x'\|,
\end{aligned}$$

where inequality (C.1) follows from the β-Lipschitz continuity of the function $z \mapsto \langle G(y,z), v\rangle$ and the characterization (3.1) of W_1.

C.2 Proof of Theorem 3.3

Fix any two points $x, x' \in \mathcal{X}$ and set $y := \texttt{Sol}(x)$ and $y' := \texttt{Sol}(x')$. Note that the definition of the normal cone implies

$$\langle G_x(y), y - y'\rangle \le 0 \qquad \text{and} \qquad \langle G_{x'}(y'), y' - y\rangle \le 0.$$

Strong monotonicity therefore ensures

$$\begin{aligned}
\alpha\|y - y'\|^2 &\le \langle G_x(y) - G_x(y'), y - y'\rangle \\
&\le \langle G_{x'}(y') - G_x(y'), y - y'\rangle \\
&\le \|G_{x'}(y') - G_x(y')\| \cdot \|y - y'\| \\
&\le \gamma\beta\|x - x'\| \cdot \|y - y'\|,
\end{aligned}$$

where the last inequality follows from Lemma 3.1. Dividing through by $\alpha\|y - y'\|$ reveals $\texttt{Sol}(\cdot)$ is contraction on $\mathcal{X}$ with parameter $\frac{\gamma\beta}{\alpha}$. An application of the Banach fixed-point theorem completes the proof.

C.3 Proof of Proposition 3.4

We will use the following classical result known as the Robbins-Siegmund almost supermartingale convergence theorem (see [26, Theorem 1.3.12] for a proof).

Lemma C.1 (Robbins-Siegmund)**.** *Let $(A_t), (B_t), (C_t), (D_t)$ be sequences of finite nonnegative random variables on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ adapted to the filtration $\mathbb{F} = (\mathcal{F}_t)$ and satisfying*

$$\mathbb{E}[A_{t+1} \,|\, \mathcal{F}_t] \leq (1 + B_t)A_t + C_t - D_t$$

for all t. Then on the event $\{\sum_t B_t < \infty, \sum_t C_t < \infty\}$, there is a finite random variable A_∞ such that $A_t \to A_\infty$ and $\sum_t D_t < \infty$ almost surely.

Toward applying Lemma C.1 with $A_t = \|x_t - x^\star\|^2$, let $(\mathcal{F}_t)$ be the filtration given by (3.8) and observe that the SFB iterate sequence (x_t) is given by

$$x_{t+1} = \mathrm{proj}_{\mathcal{X}}\big(x_t - \eta_t(R(x_t) + \xi_t)\big),$$

where the map $R\colon \mathcal{X} \to \mathbb{R}^d$ given by $R(x) = G_x(x)$ is Lipschitz continuous and strongly monotone on $\mathcal{X}$ with constants $\bar{L} + \gamma\beta$ and $\bar{\alpha} = \alpha - \gamma\beta$, respectively (see Lemma 3.8), and the noise vector $\xi_t = G(x_t, z_t) - R(x_t)$ satisfies $\mathbb{E}[\xi_t \,|\, \mathcal{F}_t] = 0$ (zero bias) with variance bound $\mathbb{E}[\|\xi_t\|^2 \,|\, \mathcal{F}_t] \leq K(1 + \|x_t - x^\star\|^2)$ for all $t \geq 0$ (Assumption 3.3). Thus, since $\eta_t \to 0$ (recall $\sum_t \eta_t^2 < \infty$), we see that for all sufficiently large t, we may apply the one-step improvement bound [58, Theorem 24] with zero bias to obtain

$$\mathbb{E}\big[\|x_{t+1} - x^\star\|^2 \mid \mathcal{F}_t\big] \leq \frac{1 + 2K\eta_t^2}{1 + \bar{\alpha}\eta_t}\|x_t - x^\star\|^2 + \frac{2K\eta_t^2}{1 + \bar{\alpha}\eta_t} \tag{C.2}$$

$$\leq (1 - \tfrac{1}{2}\bar{\alpha}\eta_t)\|x_t - x^\star\|^2 + \frac{2K\eta_t^2}{1 + \bar{\alpha}\eta_t} \tag{C.3}$$

$$\leq \|x_t - x^\star\|^2 + 2K\eta_t^2 - \tfrac{1}{2}\bar{\alpha}\eta_t\|x_t - x^\star\|^2. \tag{C.4}$$

(For (C.2), it suffices to require $\eta_t \le \frac{\bar{\alpha}}{2(\bar{L}+\gamma\beta)^2}$; for (C.3), it suffices to require $\eta_t \le \frac{\bar{\alpha}}{4K+\bar{\alpha}^2}$.) Using (C.4), we may now apply Lemma C.1 with $A_t = \|x_t - x^\star\|^2$, $B_t = 0$, $C_t = 2K\eta_t^2$, and $D_t = \frac{1}{2}\bar{\alpha}\eta_t\|x_t - x^\star\|^2$. By assumption, we have $\sum_t \eta_t^2 < \infty$, so Lemma C.1 yields a finite random variable A_∞ such that $A_t \to A_\infty$ and $\sum_t D_t < \infty$ almost surely. Hence $\|x_t - x^\star\|^2 \to A_\infty$ and $\sum_t \eta_t\|x_t - x^\star\|^2 < \infty$ almost surely. Since $\sum_t \eta_t = \infty$, we conclude $A_\infty = \lim_t \|x_t - x^\star\|^2 = 0$ almost surely, i.e., $x_t \to x^\star$ almost surely.

Next, to establish the in-expectation rate, note that (C.4) and the tower rule imply

$$\mathbb{E}\|x_{t+1} - x^\star\|^2 \le \big(1 - \tfrac{1}{2}\bar{\alpha}\eta_t\big)\mathbb{E}\|x_t - x^\star\|^2 + 2K\eta_t^2$$

for all sufficiently large t. Thus, upon supposing $\eta_t = \Theta(t^{-\nu})$ for some $\nu \in \big(\frac{1}{2}, 1\big)$, a standard inductive argument (see, e.g., [21, Lemma 3.11.8]) yields a constant $C > 0$ such that $\mathbb{E}\|x_t - x^\star\|^2 \le Ct^{-\nu}$ for all $t \ge 1$. Therefore

$$\mathbb{E}\left[\sum_{t=1}^{\infty} t^{-1/2}\|x_t - x^\star\|^2\right] \le C\sum_{t=1}^{\infty} t^{-(\nu+1/2)} < \infty$$

and hence $\sum_{t=1}^{\infty} t^{-1/2}\|x_t - x^\star\|^2 < \infty$ almost surely. This completes the proof.

Appendix D

REVIEW OF ASYMPTOTIC NORMALITY

In this appendix, we present a variation of the asymptotic normality result in [62]. Consider a measurable set $\mathcal{X} \subset \mathbb{R}^d$ and a measurable map $R\colon \mathcal{X} \to \mathbb{R}^d$. Suppose that there exists a solution $x^\star \in \mathcal{X}$ to the equation $R(x) = 0$. The goal is to approximate $x^\star$ while only having access to noisy evaluations of R. Given $x_0 \in \mathcal{X}$, consider the iterative process

$$x_{t+1} = x_t - \eta_t\big(R(x_t) + \xi_t + \zeta_t\big), \tag{D.1}$$

where η_t is a deterministic positive step size, ξ_t is a random vector in $\mathbb{R}^d$ representing noise with zero mean conditioned on prior information, and ζ_t is a random vector in $\mathbb{R}^d$ representing a residual element that both ensures $x_{t+1} \in \mathcal{X}$ and quantifies the difference between x_{t+1} and the basic step $x_t - \eta_t(R(x_t) + \xi_t)$ in the unbiased direction $-(R(x_t) + \xi_t)$; for example, taking

$$\zeta_t = \frac{x_t - \eta_t\big(R(x_t) + \xi_t\big) - \operatorname{proj}_{\mathcal{X}}\big(x_t - \eta_t\big(R(x_t) + \xi_t\big)\big)}{\eta_t}$$

in (D.1) yields the stochastic forward-backward method $x_{t+1} = \operatorname{proj}_{\mathcal{X}}(x_t - \eta_t(R(x_t) + \xi_t))$.

The following assumption formalizes the stochastic framework for our analysis.

Assumption D.1 (Stochastic framework)**.** The sequences $(x_t)_{t\geq 0}$, $(\xi_t)_{t\geq 0}$, and $(\zeta_t)_{t\geq 0}$ in (D.1) are stochastic processes defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ equipped with a filtration $(\mathcal{F}_t)_{t\geq 0}$ such that x_t is $\mathcal{F}_t$-measurable, ξ_t and ζ_t are $\mathcal{F}_{t+1}$-measurable, and ξ_t constitutes a martingale difference sequence satisfying $\mathbb{E}[\xi_t \,|\, \mathcal{F}_t] = 0$. Additionally, the following four conditions hold.

(i) **(L^2-bounded noise)** $\sup_{t\geq 0} \mathbb{E}\|\xi_t\|^2 < \infty$.

(ii) **(Asymptotic covariance)** There is a deterministic positive semidefinite matrix Σ

satisfying

$$\frac{1}{t}\sum_{i=0}^{t-1}\mathbb{E}\big[\xi_i\xi_i^\top \,|\, \mathcal{F}_i\big] \xrightarrow{p} \Sigma \qquad \text{as } t\to\infty.$$

(iii) **(Lindeberg's condition)** For all $\varepsilon > 0$,

$$\frac{1}{t}\sum_{i=0}^{t-1}\mathbb{E}\big[\|\xi_i\|^2\mathbf{1}_{\{\|\xi_i\|\geq\varepsilon\sqrt{t}\}} \,|\, \mathcal{F}_i\big] \xrightarrow{p} 0 \qquad \text{as } t\to\infty.$$

(iv) **(Negligible residual)** $\frac{1}{\sqrt{t}}\sum_{i=0}^{t-1}\|\zeta_i\| \xrightarrow{p} 0$ as $t\to\infty$.

Next, we stipulate the stability conditions regulating the dynamics of (D.1) that we require to establish asymptotic normality of the average iterates. Recall that a matrix $A\in\mathbb{R}^{d\times d}$ is said to be *positively stable* if every eigenvalue of A has a positive real part.

Assumption D.2 (Stable dynamics)**.** There is a positively stable matrix $A\in\mathbb{R}^{d\times d}$ for which the following two conditions hold.

(i) **(Step size)** The step size sequence $(\eta_t)_{t\geq 0}$ satisfies either

$$\eta_t \equiv \eta \qquad \text{and} \qquad 0<\eta<2\Big(\min_j \operatorname{Re}\lambda_j(A)\Big)^{-1} \tag{D.2}$$

or

$$\eta_t = o(1) \qquad \text{and} \qquad \frac{\eta_t-\eta_{t+1}}{\eta_t} = o(\eta_t) \qquad \text{as } t\to\infty. \tag{D.3}$$

(ii) **(Linear approximation)** The iterate sequence $(x_t)_{t\geq 0}$ satisfies

$$\frac{1}{\sqrt{t}}\sum_{i=0}^{t-1}\|R(x_i)-A(x_i-x^\star)\| \xrightarrow{p} 0 \qquad \text{as } t\to\infty. \tag{D.4}$$

Theorem D.1 (Theorem 2 in [62])**.** *Suppose that Assumptions D.1 and D.2 hold. Then, as $t\to\infty$, the average iterates $\bar{x}_t = \frac{1}{t}\sum_{i=1}^t x_i$ satisfy*

$$\sqrt{t}(\bar{x}_t - x^\star) = -A^{-1}\left(\frac{1}{\sqrt{t}}\sum_{i=0}^{t-1}\xi_i\right) + o_{\mathbb{P}}(1),$$

and hence

$$\sqrt{t}(\bar{x}_t - x^\star) \rightsquigarrow \mathsf{N}\big(0, A^{-1}\Sigma A^{-\top}\big).$$

We remark that the assumptions of Theorem D.1 are somewhat more general than those of Theorem 2 in [62], but the proof technique is the same. The primary differences are as follows:

(a) The residual term ζ_t in (D.1) need not satisfy $\mathbb{E}[\zeta_t \,|\, \mathcal{F}_t] = 0$, but this causes no difficulty as we assume ζ_t is negligible in the sense of condition (iv) of Assumption D.1. The rest of our stochastic setting stipulates conditions on ξ_t tailored to an application of the martingale central limit theorem (Theorem F.4); we note that Lindeberg's condition (iii) of Assumption D.1 is holds if the asymptotic uniform integrability condition $\limsup_{t\to\infty} \mathbb{E}\big[\|\xi_t\|^2 \mathbf{1}_{\{\|\xi_t\| \geq N\}} \,|\, \mathcal{F}_t\big] \xrightarrow{p} 0$ as $N \to \infty$ is fulfilled and $\sup_{t\geq 0} \mathbb{E}\big[\|\xi_t\|^2 \,|\, \mathcal{F}_t\big] < \infty$ almost surely.

(b) Theorem 2 in [62] requires $A = \nabla R(x^\star)$ with

$$R(x) - \nabla R(x^\star)(x - x^\star) = O(\|x - x^\star\|^q) \qquad \text{as } x \to x^\star \tag{D.5}$$

for some $q \in (1, 2]$, and assumes that the step size sequence $(\eta_t)_{t\geq 0}$ satisfies

$$\sum_{t=1}^{\infty} \eta_t^{q/2} t^{-1/2} < \infty$$

in addition to (D.3); together with a further Lyapunov function assumption, this suffices to demonstrate that the iterate sequence $(x_t)_{t\geq 0}$ satisfies both $x_t \xrightarrow{\text{a.s.}} x^\star$ and $\frac{1}{\sqrt{t}} \sum_{i=0}^{t-1} \|x_i - x^\star\|^q \xrightarrow{\text{a.s.}} 0$ as $t \to \infty$, which by (D.5) implies (D.4).

Proof. For each $t \geq 0$, let $\Delta_t = x_t - x^\star$ denote the error of the process (D.1) at time t, with corresponding average errors given by

$$\bar{\Delta}_t = \frac{1}{t} \sum_{j=1}^{t} \Delta_j = \bar{x}_t - x^\star \qquad \text{for all } t \geq 1.$$

Let A denote the matrix furnished by Assumption D.2 and observe that (D.1) yields the following recursion for all $t \geq 0$:

$$\begin{aligned} \Delta_{t+1} &= \Delta_t - \eta_t\big(R(x_t) + \xi_t + \zeta_t\big) \\ &= (I - \eta_t A)\Delta_t - \eta_t\big(R(x_t) - A\Delta_t + \xi_t + \zeta_t\big). \end{aligned} \tag{D.6}$$

Unrolling the recursion (D.6) gives

$$\Delta_j = \left(\prod_{k=0}^{j-1}(I-\eta_k A)\right)\Delta_0 - \sum_{i=0}^{j-1}\left(\prod_{k=i+1}^{j-1}(I-\eta_k)A\right)\eta_i\big(R(x_i) - A\Delta_i + \xi_i + \zeta_i\big)$$

for all $j \geq 0$ and hence

$$\begin{aligned} t\bar{\Delta}_t &= \sum_{j=1}^{t}\left(\prod_{k=0}^{j-1}(I-\eta_k A)\right)\Delta_0 - \sum_{j=1}^{t}\sum_{i=0}^{j-1}\left(\prod_{k=i+1}^{j-1}(I-\eta_k A)\right)\eta_i\big(R(x_i) - A\Delta_i + \xi_i + \zeta_i\big) \\ &= \sum_{j=1}^{t}\left(\prod_{k=0}^{j-1}(I-\eta_k A)\right)\Delta_0 - \sum_{i=0}^{t-1}\sum_{j=i+1}^{t}\left(\prod_{k=i+1}^{j-1}(I-\eta_k A)\right)\eta_i\big(R(x_i) - A\Delta_i + \xi_i + \zeta_i\big) \end{aligned}$$

for all $t \geq 1$ (interpreting empty products as the identity matrix and empty sums as zero). Thus, upon defining for each $t \geq 1$ and $i \geq 0$ the matrices

$$B_t = \sum_{j=1}^{t}\left(\prod_{k=0}^{j-1}(I-\eta_k A)\right), \qquad B_i^t = \eta_i \sum_{j=i+1}^{t}\left(\prod_{k=i+1}^{j-1}(I-\eta_k A)\right), \qquad A_i^t = B_i^t - A^{-1},$$

we have

$$\begin{aligned} t\bar{\Delta}_t &= B_t\Delta_0 - \sum_{i=0}^{t-1} B_i^t\big(R(x_i) - A\Delta_i + \xi_i + \zeta_i\big) \\ &= B_t\Delta_0 - \sum_{i=0}^{t-1} B_i^t\xi_i - \sum_{i=0}^{t-1} B_i^t\big(R(x_i) - A\Delta_i\big) - \sum_{i=0}^{t-1} B_i^t\zeta_i \\ &= B_t\Delta_0 - A^{-1}\sum_{i=0}^{t-1}\xi_i - \sum_{i=0}^{t-1} A_i^t\xi_i - \sum_{j=0}^{t-1} B_i^t\big(R(x_i) - A\Delta_i\big) - \sum_{i=0}^{t-1} B_i^t\zeta_i \end{aligned}$$

and hence

$$\begin{aligned} \sqrt{t}(\bar{x}_t - x^\star) + A^{-1}\left(\frac{1}{\sqrt{t}}\sum_{i=0}^{t-1}\xi_i\right) = \frac{1}{\sqrt{t}}B_t(x_0 - x^\star) - \frac{1}{\sqrt{t}}\sum_{i=0}^{t-1} A_i^t\xi_i \\ -\frac{1}{\sqrt{t}}\sum_{i=0}^{t-1} B_i^t\big(R(x_i) - A(x_i - x^\star)\big) - \frac{1}{\sqrt{t}}\sum_{i=0}^{t-1} B_i^t\zeta_i. \end{aligned} \tag{D.7}$$

We claim the the right-hand side of (D.7) is $o_{\mathbb{P}}(1)$ as $t \to \infty$. Indeed, since A is positively stable and the step size condition (i) of Assumption D.2 holds, it follows from [62, Lemma 1] that the collection of matrices $\{A_i^t, B_i^t, B_t \mid t \geq 1, i \geq 0\}$ is bounded with respect to the operator norm and

$$\lim_{t\to\infty}\frac{1}{t}\sum_{i=0}^{t-1}\|A_i^t\|_{\mathrm{op}} = 0. \tag{D.8}$$

Let $C = \sup\{\|A_i^t\|_{\mathrm{op}}, \|B_i^t\|_{\mathrm{op}}, \|B_t\|_{\mathrm{op}}, \mathbb{E}\|\xi_i\|^2 \mid t \geq 1, i \geq 0\}$; by the L^2-boundedness condition (i) of Assumption D.1, we have $C < \infty$. Therefore

$$\left\| \frac{1}{\sqrt{t}} B_t (x_0 - x^\star) \right\| \leq \frac{C\|x_0 - x^\star\|}{\sqrt{t}} \xrightarrow{\text{a.s.}} 0 \qquad \text{as } t \to \infty, \tag{D.9}$$

and since $(\xi_i)_{i \geq 0}$ is a martingale difference sequence, we deduce from (D.8) the following convergence in mean square:

$$\mathbb{E}\left\| \frac{1}{\sqrt{t}} \sum_{i=0}^{t-1} A_i^t \xi_i \right\|^2 = \frac{1}{t} \sum_{i=0}^{t-1} \mathbb{E}\|A_i^t \xi_i\|^2 \leq \frac{C}{t} \sum_{i=0}^{t-1} \|A_i^t\|_{\mathrm{op}}^2 \leq \frac{C^2}{t} \sum_{i=0}^{t-1} \|A_i^t\|_{\mathrm{op}} \to 0 \qquad \text{as } t \to \infty,$$

which by Markov's inequality implies

$$\frac{1}{\sqrt{t}} \sum_{i=0}^{t-1} A_i^t \xi_i \xrightarrow{p} 0 \qquad \text{as } t \to \infty. \tag{D.10}$$

Moreover, the linear approximation condition (ii) of Assumption D.2 implies

$$\left\| \frac{1}{\sqrt{t}} \sum_{i=0}^{t-1} B_i^t \big(R(x_i) - A(x_i - x^\star)\big) \right\| \leq \frac{C}{\sqrt{t}} \sum_{i=0}^{t-1} \|R(x_i) - A(x_i - x^\star)\| \xrightarrow{p} 0 \qquad \text{as } t \to \infty, \tag{D.11}$$

while the negligible residual condition (iv) of Assumption D.1 implies

$$\left\| \frac{1}{\sqrt{t}} \sum_{i=0}^{t-1} B_i^t \zeta_i \right\| \leq \frac{C}{\sqrt{t}} \sum_{i=0}^{t-1} \|\zeta_i\| \xrightarrow{p} 0 \qquad \text{as } t \to \infty. \tag{D.12}$$

By (D.9)–(D.12), we conclude that the right-hand side of (D.7) is $o_{\mathbb{P}}(1)$ as $t \to \infty$, so

$$\sqrt{t}(\bar{x}_t - x^\star) = -A^{-1} \left(\frac{1}{\sqrt{t}} \sum_{i=0}^{t-1} \xi_i \right) + o_{\mathbb{P}}(1) \qquad \text{as } t \to \infty.$$

Finally, by virtue of Assumption D.1, we may apply the martingale central limit theorem (Theorem F.4) to the square-integrable martingale $M_t = \sum_{i=0}^{t-1} \xi_i$ to obtain $t^{-1/2} M_t \rightsquigarrow \mathsf{N}(0, \Sigma)$ and hence, by the continuous mapping theorem [75, Theorem 2.3],

$$-A^{-1} \left(\frac{1}{\sqrt{t}} \sum_{i=0}^{t-1} \xi_i \right) \rightsquigarrow \mathsf{N}\big(0, A^{-1} \Sigma A^{-\top}\big) \qquad \text{as } t \to \infty.$$

This completes the proof. □

Appendix E

PROOFS DEFERRED FROM SECTION 3.4

This appendix contains the proofs deferred from Section 3.4. We assume throughout that the assumptions used in Section 3.4 are valid; in particular, $\mathcal{X}$ is compact, $\mathcal{Z}$ is bounded, and $g \in \mathcal{G}$ (see Definition 3.14). To begin, we present three preliminary lemmas.

Lemma E.1. *We have*

$$\sup_{x\in\mathcal{X},\, z\in\mathcal{Z}} \|g_x(z)\| < \infty \qquad \textit{and} \qquad \sup_{x\in\mathcal{X},\, z\in\mathcal{Z}} \|\nabla_x g_x(z)\|_{\mathrm{op}} < \infty.$$

Proof. Fix $x^\circ \in \mathcal{X}$ and $z^\circ \in \mathcal{Z}$. Since $\mathcal{X}$ and $\mathcal{Z}$ are bounded, we compute

$$\begin{aligned} M_g' := \sup_{x\in\mathcal{X},\, z\in\mathcal{Z}} \|\nabla_x g_x(z)\|_{\mathrm{op}} &\le \|\nabla_x g_{x^\circ}(z^\circ)\|_{\mathrm{op}} + \sup_{x\in\mathcal{X}} \|\nabla_x g_x(z^\circ) - \nabla_x g_{x^\circ}(z^\circ)\|_{\mathrm{op}} \\ &\quad + \sup_{x\in\mathcal{X},\, z\in\mathcal{Z}} \|\nabla_x g_x(z) - \nabla_x g_x(z^\circ)\|_{\mathrm{op}} \\ &\le \|\nabla_x g_{x^\circ}(z^\circ)\|_{\mathrm{op}} + \Lambda_g(z^\circ)\operatorname{diam}(\mathcal{X}) + \beta_g' \operatorname{diam}(\mathcal{Z}) < \infty. \end{aligned}$$

Hence every section $g(\cdot, z)$ is M_g'-Lipschitz on $\mathcal{X}$, and the estimate

$$\begin{aligned} M_g := \sup_{x\in\mathcal{X},\, z\in\mathcal{Z}} \|g_x(z)\| &\le \|g_{x^\circ}(z^\circ)\| + \sup_{x\in\mathcal{X}} \|g_x(z^\circ) - g_{x^\circ}(z^\circ)\| + \sup_{x\in\mathcal{X},\, z\in\mathcal{Z}} \|g_x(z) - g_x(z^\circ)\| \\ &\le \|g_{x^\circ}(z^\circ)\| + M_g' \operatorname{diam}(\mathcal{X}) + \beta_g \operatorname{diam}(\mathcal{Z}) \end{aligned}$$

completes the proof. □

Lemma E.2. *Let* $L_h = \sup|h'|$, $L_{h''} = \sup|h'''|$, $A_g = \sup_{x\in\mathcal{X}} \mathbb{E}_{z\sim\mathcal{D}_x}\|g_x(z)\|$, *and* $B_g = \sup_{x\in\mathcal{X}} \mathbb{E}_{z\sim\mathcal{D}_x}\|g_x(z)\|^3$.

(i) Let $u \in \mathbb{R}^d$. *Then*

$$|h(u^\top g_x(z))| \le L_h \|g_x(z)\| \|u\| \tag{E.1}$$

for all $x \in \mathcal{X}$ *and* $z \in \mathcal{Z}$ *and hence*

$$\sup_{x\in\mathcal{X}} \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x} |h(u^\top g_x(z))| \le L_h A_g \|u\| = O(\|u\|). \tag{E.2}$$

(ii) For each $x \in \mathcal{X}$*, the function* $u \mapsto C_x^u = 1 + \mathbb{E}_{z\sim\mathcal{D}_x} h(u^\top g_x(z))$ *is* C^2*-smooth on* $\mathbb{R}^d$ *with* $L_{h''}B_g$*-Lipschitz continuous Hessian, and we have* $C_x^0 = 1$*,* $\nabla_u C_x^u|_{u=0} = 0$*, and* $\nabla^2_{uu} C_x^u|_{u=0} = 0$*. Therefore*

$$\sup_{x\in\mathcal{X}} |\mathbb{E}_{z\sim\mathcal{D}_x} h(u^\top g_x(z))| \le \frac{L_{h''}B_g}{6}\|u\|^3 \tag{E.3}$$

for all $u \in \mathbb{R}^d$ *and hence*

$$\sup_{x\in\mathcal{X}} \frac{1}{C_x^u} = 1 + O(\|u\|^3) \qquad \text{as } u \to 0. \tag{E.4}$$

Proof. Note first that $h(t) = t$ for all t in a neighborhood of zero and the first three derivatives of h are bounded by assumption, while $A_g, B_g < \infty$ by Lemma E.1. Since $h(0) = 0$ and h is L_h-Lipschitz continuous, the inequalities (E.1) and (E.2) follow immediately. Next, let $x \in \mathcal{X}$ and observe that the dominated convergence theorem yields

$$\nabla_u \Big(\mathop{\mathbb{E}}_{z\sim\mathcal{D}_x} h\big(u^\top g_x(z)\big)\Big) = \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x} h'\big(u^\top g_x(z)\big) g_x(z)$$

and

$$\nabla^2_{uu} \Big(\mathop{\mathbb{E}}_{z\sim\mathcal{D}_x} h\big(u^\top g_x(z)\big)\Big) = \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x} h''\big(u^\top g_x(z)\big) g_x(z) g_x(z)^\top$$

for all $u \in \mathbb{R}^d$. Thus, $u \mapsto C_x^u$ is C^2-smooth on $\mathbb{R}^d$, and since h'' is $L_{h''}$-Lipschitz continuous, it follows at once that $u \mapsto \nabla^2_{uu} C_x^u$ is $L_{h''}B_g$-Lipschitz continuous on $\mathbb{R}^d$.

Clearly $C_x^0 = 1$ since $h(0) = 0$. Further,

$$\nabla_u C_x^u|_{u=0} = \nabla_u \Big(\mathop{\mathbb{E}}_{z\sim\mathcal{D}_x} h\big(u^\top g_x(z)\big)\Big)\Big|_{u=0} = \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x} g_x(z) = 0$$

since $h'(0) = 1$ and $g \in \mathcal{G}$, while

$$\nabla^2_{uu} C_x^u|_{u=0} = \nabla^2_{uu} \Big(\mathop{\mathbb{E}}_{z\sim\mathcal{D}_x} h\big(u^\top g_x(z)\big)\Big)\Big|_{u=0} = 0$$

since $h''(0) = 0$. The second-order Taylor polynomial of the function $u \mapsto \mathbb{E}_{z\sim\mathcal{D}_x} h(u^\top g_x(z))$ about $u = 0$ is therefore identically zero, so $L_{h''}B_g$-Lipschitzness of the Hessian implies (E.3).

Finally, the estimate

$$\frac{1}{1+t} = 1 - \frac{t}{1+t} \le 1 + 2|t| \qquad \text{for all } t \ge -\frac{1}{2}$$

together with (E.3) yields (E.4). □

Lemma E.3. *Let $f\colon (0,\infty) \to \mathbb{R}$ be a function that is C^3-smooth around $t = 1$ and satisfies $f(1) = 0$. Then for all sufficiently small $u \in \mathbb{R}^d$ and all $x \in \mathcal{X}$, we have*

$$\int f\left(\frac{1 + u^\top g_x(z)}{C_x^u}\right) d\mathcal{D}_x(z) = \frac{f''(1)}{2} u^\top \left(\mathop{\mathbb{E}}_{z\sim\mathcal{D}_x} g_x(z) g_x(z)^\top \right) u + r_x(u), \tag{E.5}$$

where $\sup_{x\in\mathcal{X}} |r_x(u)| = O(\|u\|^3)$ as $u \to 0$.

Proof. Fix $x \in \mathcal{X}$ and define $\varphi_x(u) := \mathbb{E}_{z\sim\mathcal{D}_x} f\left(\frac{1+u^\top g_x(z)}{C_x^u}\right)$. By the dominated convergence theorem, φ_x is C^2-smooth on a neighborhood of zero with

$$\nabla_u \varphi_x(u) = \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x}\left[f'\left(\frac{1 + u^\top g_x(z)}{C_x^u}\right) \left(\frac{g_x(z) C_x^u - \big(1 + u^\top g_x(z)\big) \nabla_u C_x^u}{(C_x^u)^2} \right) \right]$$

and $(C_x^u)^4 \cdot \nabla_{uu}^2 \varphi_x(u)$ equal to

$$\begin{aligned}
\mathop{\mathbb{E}}_{z\sim\mathcal{D}_x}\Bigg[& f''\left(\frac{1 + u^\top g_x(z)}{C_x^u}\right) \Big(g_x(z) C_x^u - \big(1 + u^\top g_x(z)\big) \nabla_u C_x^u \Big) \Big(g_x(z) C_x^u - \big(1 + u^\top g_x(z)\big) \nabla_u C_x^u \Big)^\top \\
& + f'\left(\frac{1 + u^\top g_x(z)}{C_x^u}\right) \Big((C_x^u)^2 \Big(g_x(z) (\nabla_u C_x^u)^\top - (\nabla_u C_x^u) g_x(z)^\top - \big(1 + u^\top g_x(z)\big) \nabla_{uu}^2 C_x^u \Big) \\
& \qquad - 2 C_x^u \Big(g_x(z) C_x^u - \big(1 + u^\top g_x(z)\big) \nabla_u C_x^u \Big) (\nabla_u C_x^u)^\top \Big) \Bigg].
\end{aligned}$$

Thus, taking a second-order Taylor expansion of φ_x at $u = 0$ with remainder r_x and applying the equalities $C_x^0 = 1$, $\nabla_u C_x^u|_{u=0} = 0$, $\nabla_{uu}^2 C_x^u|_{u=0} = 0$, and $f(1) = 0$ yields (E.5). It remains to verify $\sup_{x\in\mathcal{X}} |r_x(u)| = O(\|u\|^3)$ as $u \to 0$.

Lemmas E.1 and E.2 ensure that $C_x^u, \nabla_u C_x^u$, and $\nabla_{uu}^2 C_x^u$ are Lipschitz continuous and bounded on a compact neighborhood of $u = 0$, with Lipschitz constants and bounds independent of x. Further, since f is C^3-smooth around $t = 1$, we have that f' and f'' are Lipschitz continuous and bounded on a compact neighborhood of $t = 1$. It follows that $\nabla_{uu}^2 \varphi_x$ is $\tilde{L}$-Lipschitz on a neighborhood U of $u = 0$, with constant $\tilde{L}$ independent of x. Thus we deduce $|r_x(u)| \le \frac{\tilde{L}}{6}\|u\|^3$ for all $(x, u) \in \mathcal{X} \times U$, and the result follows. □

E.1 Proof of Lemma 3.15

The proof of Lemma 3.15 is divided into four steps: the first step verifies Assumption 3.1 and the next three steps establish Assumption 3.2. The strategy in all steps is to prove that various quantities of interest change continuously with u near zero. One of the main tools we will use to this end is the following elementary lemma (which we will also use crucially later in the proof of Lemma 3.24). Its proof consists of several applications of the dominated convergence theorem and is deferred to Section E.4.

Lemma E.4 (Inferring smoothness)**.** *Suppose that $T\colon \mathcal{X} \times \mathcal{Z} \to \mathbb{R}^n$ is a map satisfying the following two conditions.*

*(i) (**Lipschitz continuity**) There exists a constant $\beta_T \geq 0$ such that for every $x \in \mathcal{X}$, the section $T(x, \cdot)$ is β_T-Lipschitz on $\mathcal{Z}$.*

*(ii) (**Smoothness**) There exist a measurable function $\Lambda_T\colon \mathcal{Z} \to [0, \infty)$ and constants $\bar{\Lambda}_T, \beta_T' \geq 0$ such that for every $z \in \mathcal{Z}$ and $x \in \mathcal{X}$, the section $T(\cdot, z)$ is $\Lambda_T(z)$-smooth on $\mathcal{X}$ with $\mathbb{E}_{z\sim\mathcal{D}_x}[\Lambda_T(z)] \leq \bar{\Lambda}_T$, and the section $\nabla_x T(x, \cdot)$ is β_T'-Lipschitz on $\mathcal{Z}$.*

Set

$$M_T := \sup_{x\in\mathcal{X},\, z\in\mathcal{Z}} \|T(x,z)\| \qquad \text{and} \qquad M_T' := \sup_{x\in\mathcal{X},\, z\in\mathcal{Z}} \|\nabla_x T(x,z)\|_{\mathrm{op}}.$$

Then M_T and M_T' are finite. Moreover, given any fixed compact neighborhood $\mathcal{W} \subset \mathbb{R}^d$ of zero, the maps $\bar{H}\colon \mathcal{X} \times \mathcal{X} \times \mathcal{W} \to \mathbb{R}^n$ and $H\colon \mathcal{X} \times \mathcal{W} \to \mathbb{R}^n$ given by

$$\bar{H}(x,y,u) = \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x}\big[T(y,z)\big(1 + h\big(u^\top g_y(z)\big)\big)\big] \qquad \text{and} \qquad H(x,u) = \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x^u} T(x,z)$$

are smooth with Lipschitz continuous Jacobians with constants depending on T only through $\beta_T, \bar{\Lambda}_T, \beta_T'$, M_T, and M_T'; further, we have

$$\nabla_x H(x,0) = \nabla_x\Big(\mathop{\mathbb{E}}_{z\sim\mathcal{D}_x} T(x,z)\Big) \qquad \text{and} \qquad \nabla_u H(x,0) = \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x}\big[T(x,z) g_x(z)^\top\big]$$

for all $x \in \mathcal{X}$.

Step 1 (Assumption 3.1) First, we show that the perturbed distribution map $\mathcal{D}^u$ satisfies Assumption 3.1 with Lipschitz constant $\gamma^u = \gamma + O(\|u\|)$ as $u \to 0$, where γ is the Lipschitz constant for $\mathcal{D}$. To this end, we take $\mathcal{W}$ to be the unit ball in $\mathbb{R}^d$ and apply Lemma E.4 to identify a constant $L_1 \geq 0$ such that for every 1-Lipschitz function $\phi \in \mathrm{Lip}_1(\mathcal{Z})$ and every $u \in \mathcal{W}$, the function

$$\rho_\phi(x, u) := \mathop{\mathbb{E}}_{z \sim \mathcal{D}_x^u} \phi(z)$$

is Lipschitz in the x-component with constant $\gamma^u := \gamma + L_1\|u\|$. Indeed, for every $\phi \in \mathrm{Lip}_1(\mathcal{Z})$, the translate $\bar{\phi} = \phi - \inf \phi$ is 1-Lipschitz and bounded by $\mathrm{diam}(\mathcal{Z})$, and $\rho_{\bar{\phi}} = \rho_\phi - \inf \phi$. Thus, Lemma E.4 yields a constant L_1 such that for every $\phi \in \mathrm{Lip}_1(\mathcal{Z})$, the function $\rho_{\bar{\phi}}$ is L_1-smooth on $\mathcal{X} \times \mathcal{W}$ and hence so is ρ_ϕ. Moreover, Lemma E.4 shows

$$\nabla_x \rho_\phi(x, 0) = \nabla_x \Big(\mathop{\mathbb{E}}_{z \sim \mathcal{D}_x} \phi(z) \Big)$$

for all $x \in \mathcal{X}$, so $\sup_{x \in \mathcal{X}} \|\nabla_x \rho_\phi(x, 0)\| \leq \gamma$ by Assumption 3.1. Thus, by the triangle inequality,

$$\|\nabla_x \rho_\phi(x, u)\| \leq \|\nabla_x \rho_\phi(x, 0)\| + \|\nabla_x \rho_\phi(x, u) - \nabla_x \rho_\phi(x, 0)\| \leq \gamma + L_1\|u\| = \gamma^u$$

for all $(x, u) \in \mathcal{X} \times \mathcal{W}$. Therefore $\rho_\phi(\cdot, u)$ is γ^u-Lipschitz on $\mathcal{X}$ for all $\phi \in \mathrm{Lip}_1(\mathcal{Z})$ and $u \in \mathcal{W}$, so $\mathcal{D}^u$ satisfies Assumption 3.1 with Lipschitz constant $\gamma^u = \gamma + O(\|u\|)$ as $u \to 0$.

Step 2 (Lipschitz continuity) Next, we establish Assumption 3.2(i) for the problem with the perturbed distribution map $\mathcal{D}^u$. Observe that the Lipschitz bounds in Assumption 3.2(i) remain unchanged, and that we only need to identify for all sufficiently small u a constant $\bar{L}^u \geq 0$ such that $\sup_{x \in \mathcal{X}} \mathbb{E}_{z \sim \mathcal{D}_x^u}[L(z)^2] \leq (\bar{L}^u)^2$. We will show more, namely, that we can select $(\bar{L}^u)^2 = \bar{L}^2 + O(\|u\|)$ as $u \to 0$, where $\bar{L}$ is the constant satisfying $\sup_{x \in \mathcal{X}} \mathbb{E}_{z \sim \mathcal{D}_x}[L(z)^2] \leq \bar{L}^2$ furnished by Assumption 3.2(i). Indeed, for all $x \in \mathcal{X}$ and $u \in \mathbb{R}^d$, we have

$$\mathop{\mathbb{E}}_{z \sim \mathcal{D}_x^u}\big[L(z)^2\big] = \frac{1}{C_x^u} \mathop{\mathbb{E}}_{z \sim \mathcal{D}_x}\big[L(z)^2\big(1 + h\big(u^\top g_x(z)\big)\big)\big].$$

Thus, an application of Lemma E.2 yields

$$\sup_{x \in \mathcal{X}} \mathop{\mathbb{E}}_{z \sim \mathcal{D}_x^u}\big[L(z)^2\big] \leq \big(1 + O(\|u\|^3)\big)\big(1 + L_h M_g \|u\|\big)\bar{L}^2 = \bar{L}^2 + O(\|u\|) \qquad \text{as } u \to 0,$$

where $L_h = \sup |h'|$ and $M_g = \sup \|g\|$.

Step 3 (Monotonicity) We prove that for all $x \in \mathcal{X}$, the map $G_x^u(\cdot)$ given by

$$G_x^u(y) := \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x^u} G(y,z)$$

is strongly monotone on $\mathcal{X}$ with constant $\alpha^u = \alpha + O(\|u\|)$ as $u \to 0$, where α is the strong monotonicity constant of $G_x(\cdot)$. Given $x \in \mathcal{X}$ and $u \in \mathbb{R}^d$, we have

$$\begin{aligned}\big\langle G_x^u(y) - G_x^u(y'), y - y'\big\rangle &= \big\langle G_x(y) - G_x(y'), y - y'\big\rangle \\ &\quad + \big\langle \big(G_x^u(y) - G_x(y)\big) - \big(G_x^u(y') - G_x(y')\big), y - y'\big\rangle \\ &\geq \alpha\|y - y'\|^2 - \big\|\big(G_x^u(y) - G_x(y)\big) - \big(G_x^u(y') - G_x(y')\big)\big\| \cdot \|y - y'\|\end{aligned}$$

for all $y, y' \in \mathcal{X}$ by the α-strong monotonicity of $G_x(\cdot)$. We claim that for all sufficiently small u, there exists $\ell^u = O(\|u\|)$ independent of x such that the map $y \mapsto G_x^u(y) - G_x(y)$ is ℓ^u-Lipschitz on $\mathcal{X}$ for all $x \in \mathcal{X}$. Indeed, upon noting $\sup_{x\in\mathcal{X},\, z\in\mathcal{Z}} \|\nabla_x G(x,z)\|_{\mathrm{op}} < \infty$ (see Lemma E.4) and applying the dominated convergence theorem and Lemma E.2, we obtain

$$\begin{aligned}\ell^u &:= \sup_{x,y\in\mathcal{X}} \big\|\nabla_y\big(G_x^u(y) - G_x(y)\big)\big\|_{\mathrm{op}} \\ &= \sup_{x,y\in\mathcal{X}} \left\| \frac{1}{C_x^u} \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x}\big[\nabla_y G(y,z)\big(1 + h\big(u^\top g_x(z)\big)\big)\big] - \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x}\big[\nabla_y G(y,z)\big] \right\|_{\mathrm{op}} \\ &\leq \underbrace{\sup_{x,y\in\mathcal{X}} \left\| \left(\frac{1}{C_x^u} - 1\right) \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x}\big[\nabla_y G(y,z)\big] \right\|_{\mathrm{op}}}_{O(\|u\|^3)} + \underbrace{\sup_{x,y\in\mathcal{X}} \left\| \frac{1}{C_x^u} \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x}\big[\nabla_y G(y,z) h\big(u^\top g_x(z)\big)\big] \right\|_{\mathrm{op}}}_{(1+O(\|u\|^3))\cdot O(\|u\|)} \\ &= O(\|u\|) \qquad \text{as } u \to 0.\end{aligned}$$

Setting $\alpha^u := \alpha - \ell^u$ for all u in a neighborhood of zero, we conclude that for all $x, y, y' \in \mathcal{X}$,

$$\big\langle G_x^u(y) - G_x^u(y'), y - y'\big\rangle \geq \alpha^u\|y - y'\|^2$$

and hence $G_x^u(\cdot)$ is strongly monotone on $\mathcal{X}$ with constant $\alpha^u = \alpha + O(\|u\|)$ as $u \to 0$.

Step 4 (Compatibility) Finally, we verify that Assumption 3.2(iii) holds for the perturbed problem corresponding to $\mathcal{D}^u$. Indeed, as a consequence of the previous steps, we have $\gamma^u \to \gamma$ and $\alpha^u \to \alpha$ as $u \to 0$, so the compatibility inequality $\gamma\beta < \alpha$ corresponding to $\mathcal{D}$ implies $\gamma^u\beta < \alpha^u$ for all sufficiently small u.

E.2 Proof of Lemma 3.23

Fix $u \in \mathbb{R}^d$. For each $k \in \mathbb{N}$, it follows immediately from the definitions (3.14), (3.26), and (3.27) that for all $E_0, \dots, E_{k-1} \in \mathcal{B}(\mathcal{Z})$, the $Q_{k,u}$-measure of the rectangle $E = E_0 \times \cdots \times E_{k-1}$ is given by

$$\begin{aligned} Q_{k,u}(E) &= \int_{E_0} \cdots \int_{E_{k-1}} d\mathcal{D}^{u/\sqrt{k}}_{\tilde{x}_{k-1}}(z_{k-1}) \cdots d\mathcal{D}^{u/\sqrt{k}}_{\tilde{x}_0}(z_0) \\ &= \int_{E_0} \cdots \int_{E_{k-1}} \prod_{i=0}^{k-1} \frac{1+h(u^\top g_{\tilde{x}_i}(z_i)/\sqrt{k})}{C^{u/\sqrt{k}}_{\tilde{x}_i}} \, d\mathcal{D}_{\tilde{x}_{k-1}}(z_{k-1}) \cdots d\mathcal{D}_{\tilde{x}_0}(z_0) \\ &= \int_E \prod_{i=0}^{k-1} \frac{1+h(u^\top g_{\tilde{x}_i}(z_i)/\sqrt{k})}{C^{u/\sqrt{k}}_{\tilde{x}_i}} \, dQ_{k,0}. \end{aligned}$$

Therefore

$$\frac{dQ_{k,u}}{dQ_{k,0}} = \prod_{i=0}^{k-1} \frac{1+h(u^\top g_{\tilde{x}_i}(z_i)/\sqrt{k})}{C^{u/\sqrt{k}}_{\tilde{x}_i}}$$

and hence

$$\log \frac{dQ_{k,u}}{dQ_{k,0}} = \sum_{i=0}^{k-1} \log\left(1 + h\left(\frac{u^\top g_{\tilde{x}_i}(z_i)}{\sqrt{k}}\right)\right) - \sum_{i=0}^{k-1} \log C^{u/\sqrt{k}}_{\tilde{x}_i}. \tag{E.6}$$

By Lemma E.2, we have $C^u_x = 1 + r_x(u)$ with $\sup_{x \in \mathcal{X}} |r_x(u)| = o(\|u\|^2)$ as $u \to 0$, so the first-order approximation $\log(1+t) = t + o(t)$ as $t \to 0$ reveals that the last sum in (E.6) satisfies

$$\sum_{i=0}^{k-1} \log C^{u/\sqrt{k}}_{\tilde{x}_i} = \sum_{i=0}^{k-1} \Big(r_{\tilde{x}_i}\big(u/\sqrt{k}\big) + o\big(r_{\tilde{x}_i}\big(u/\sqrt{k}\big)\big)\Big) = k \cdot o(k^{-1}) = o(1) \qquad \text{as } k \to \infty.$$

Further, since $h(t) = t$ for all t in a neighborhood of zero and $c := \sup_{x \in \mathcal{X},\, z \in \mathcal{Z}} |u^\top g_x(z)|$ is finite by Lemma E.1, it follows that for all sufficiently large $k \in \mathbb{N}$, we have

$$h\left(\frac{u^\top g_{\tilde{x}_i}(z_i)}{\sqrt{k}}\right) = \frac{u^\top g_{\tilde{x}_i}(z_i)}{\sqrt{k}} \in \left[-\frac{c}{\sqrt{k}}, \frac{c}{\sqrt{k}}\right] \qquad \text{for all } i \geq 0.$$

Thus, the second-order approximation $\log(1+t) = t - \frac{1}{2}t^2 + o(t^2)$ as $t \to 0$ reveals that the

first sum in (E.6) satisfies

$$\begin{aligned}\sum_{i=0}^{k-1} \log\left(1 + h\left(\frac{u^\top g_{\tilde{x}_i}(z_i)}{\sqrt{k}}\right)\right) \\ &= u^\top \left(\frac{1}{\sqrt{k}} \sum_{i=0}^{k-1} g_{\tilde{x}_i}(z_i)\right) - \frac{1}{2} u^\top \left(\frac{1}{k} \sum_{i=0}^{k-1} g_{\tilde{x}_i}(z_i) g_{\tilde{x}_i}(z_i)^\top\right) u + k \cdot o(k^{-1}) \\ &= u^\top Z_k - \frac{1}{2} u^\top V_k u + o(1) \qquad \text{as } k \to \infty,\end{aligned}$$

where $Z_k \colon \mathcal{Z}^k \to \mathbb{R}^d$ and $V_k \colon \mathcal{Z}^k \to \mathbb{R}^{d \times d}$ are given by

$$Z_k = \frac{1}{\sqrt{k}} \sum_{i=0}^{k-1} g_{\tilde{x}_i}(z_i) \qquad \text{and} \qquad V_k = \frac{1}{k} \sum_{i=0}^{k-1} g_{\tilde{x}_i}(z_i) g_{\tilde{x}_i}(z_i)^\top.$$

Therefore

$$\log \frac{dQ_{k,u}}{dQ_{k,0}} = u^\top Z_k - \frac{1}{2} u^\top V_k u + o(1) \qquad \text{as } k \to \infty.$$

Hence, to complete the verification that $\{Q_{k,u} \,|\, u \in \mathbb{R}^d\}$ is locally asymptotically normal at zero with precision Σ_g, it only remains to demonstrate $Z_k \overset{0}{\rightsquigarrow} \mathsf{N}(0, \Sigma_g)$ and $V_k = \Sigma_g + o_{Q_{k,0}}(1)$.

The assertion $V_k = \Sigma_g + o_{Q_{k,0}}(1)$ is equivalent to $V_k \xrightarrow{p} \Sigma_g$ as $k \to \infty$ on the filtered probability space $(\mathcal{Z}^\mathbb{N}, \mathcal{B}(\mathcal{Z}^\mathbb{N}), \mathbb{F}, \mathbb{P})$, where $\mathbb{F} = (\mathcal{F}_k)_{k \geq 0}$ is the filtration given by

$$\mathcal{F}_0 := \{\emptyset, \mathcal{Z}^\mathbb{N}\} \quad \text{and} \quad \mathcal{F}_k := \{E \times \mathcal{Z}^\mathbb{N} \mid E \in \mathcal{B}(\mathcal{Z}^k)\} \qquad \text{for all } k \geq 1$$

and $\mathbb{P} := \bigotimes_{i=0}^\infty \mathcal{D}_{\tilde{x}_i}$. We will show more, namely, that almost sure convergence holds:

$$V_k \xrightarrow{\text{a.s.}} \Sigma_g \qquad \text{as } k \to \infty. \tag{E.7}$$

This is a consequence of the martingale strong law of large numbers (Theorem F.3). Indeed, for each $i \geq 0$, set

$$\begin{aligned}X_{i+1} &= g_{\tilde{x}_i}(z_i) g_{\tilde{x}_i}(z_i)^\top - \mathbb{E}\big[g_{\tilde{x}_i}(z_i) g_{\tilde{x}_i}(z_i)^\top \,|\, \mathcal{F}_i\big] \\ &= g_{\tilde{x}_i}(z_i) g_{\tilde{x}_i}(z_i)^\top - \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_{\tilde{x}_i}} \big[g_{\tilde{x}_i}(z_i) g_{\tilde{x}_i}(z_i)^\top\big],\end{aligned}$$

thereby defining a martingale difference sequence X in $\mathbb{R}^{d \times d}$ adapted to $\mathbb{F}$; note that we have $\sup_i \mathbb{E}\|X_i\|_\mathrm{F}^2 < \infty$ by Lemma E.1, so $\sum_{i=1}^\infty i^{-2} \mathbb{E}\|X_i\|_\mathrm{F}^2 < \infty$ and hence Theorem F.3 implies

$$V_k - \frac{1}{k} \sum_{i=0}^{k-1} \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_{\tilde{x}_i}} \big[g_{\tilde{x}_i}(z_i) g_{\tilde{x}_i}(z_i)^\top\big] = \frac{1}{k} \sum_{i=1}^{k} X_i \xrightarrow{\text{a.s.}} 0 \qquad \text{as } k \to \infty. \tag{E.8}$$

On the other hand, we have $\tilde{x}_i \xrightarrow{\text{a.s.}} x^\star$ as $i \to \infty$ by Definition 3.12, so Lemma F.5 implies

$$\mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_{\tilde{x}_i}} \left[g_{\tilde{x}_i}(z_i) g_{\tilde{x}_i}(z_i)^\top \right] \xrightarrow{\text{a.s.}} \mathop{\mathbb{E}}_{z \sim \mathcal{D}_{x^\star}} \left[g_{x^\star}(z) g_{x^\star}(z)^\top \right] = \Sigma_g \qquad \text{as } i \to \infty$$

and hence the arithmetic mean satisfies

$$\frac{1}{k} \sum_{i=0}^{k-1} \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_{\tilde{x}_i}} \left[g_{\tilde{x}_i}(z_i) g_{\tilde{x}_i}(z_i)^\top \right] \xrightarrow{\text{a.s.}} \Sigma_g \qquad \text{as } k \to \infty. \tag{E.9}$$

Combining (E.8) and (E.9) gives (E.7).

Finally, we establish $Z_k \overset{0}{\rightsquigarrow} \mathsf{N}(0, \Sigma_g)$ by applying the martingale central limit theorem (Theorem F.4). Set $M_0 = 0$ and $M_k = \sum_{i=0}^{k-1} g_{\tilde{x}_i}(z_i)$ for each $k \geq 1$; then M is a square-integrable martingale in $\mathbb{R}^d$ adapted to the filtration $\mathbb{F}$. Indeed, the increments of M are clearly uniformly bounded (Lemma E.1), M_k is $\mathcal{F}_k$-measurable, and

$$\mathbb{E}\left[M_{k+1} \,|\, \mathcal{F}_k \right] = M_k + \mathop{\mathbb{E}}_{z_k \sim \mathcal{D}_{\tilde{x}_k}} \left[g_{\tilde{x}_k}(z_k) \right] = M_k$$

by the unbiasedness condition of Definition 3.14. The predictable quadratic variation of M is given by

$$\langle M \rangle_k = \sum_{i=1}^{k} \mathbb{E}\left[(M_i - M_{i-1})(M_i - M_{i-1})^\top \,|\, \mathcal{F}_{i-1} \right] = \sum_{i=0}^{k-1} \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_{\tilde{x}_i}} \left[g_{\tilde{x}_i}(z_i) g_{\tilde{x}_i}(z_i)^\top \right].$$

Thus, by (E.9), we have

$$k^{-1} \langle M \rangle_k = \frac{1}{k} \sum_{i=0}^{k-1} \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_{\tilde{x}_i}} \left[g_{\tilde{x}_i}(z_i) g_{\tilde{x}_i}(z_i)^\top \right] \xrightarrow{\text{a.s.}} \Sigma_g \qquad \text{as } k \to \infty.$$

The assumptions of Theorem F.4 are therefore fulfilled with $a_k = k$ (note that Lindeberg's condition holds trivially by the uniform boundedness of the increments of M). Hence

$$Z_k = k^{-1/2} M_k \overset{0}{\rightsquigarrow} \mathsf{N}(0, \Sigma_g).$$

This completes the proof.

E.3 Proof of Lemma 3.24

Let $F \colon \mathcal{X} \times \mathbb{R}^d \to \mathbb{R}^d$ be the map given by

$$F(x, u) = \mathop{\mathbb{E}}_{z \sim \mathcal{D}_x^u} \left[G(x, z) \right] = \frac{1}{C_x^u} \mathop{\mathbb{E}}_{z \sim \mathcal{D}_x} \left[\big(1 + h\big(u^\top g_x(z) \big) \big) G(x, z) \right],$$

where we recall $C_x^u = 1 + \mathbb{E}_{z\sim\mathcal{D}_x} h(u^\top g_x(z))$. Lemma E.4 directly implies that F is C^1-smooth. Consider now the family of smooth nonlinear equations

$$F(x,u) = 0 \tag{E.10}$$

parametrized by $u \in \mathbb{R}^d$. Note $F(x^\star, 0) = G_{x^\star}(x^\star) = 0$ since $x^\star \in \operatorname{int}\mathcal{X}$. More generally, the equality (E.10) with $(x,u) \in (\operatorname{int}\mathcal{X}) \times \mathcal{U}$ holds precisely when x is equal to $x_u^\star$. We will apply the implicit function theorem to show that (E.10) determines $x_u^\star$ as a smooth function of u on a neighborhood of zero. To this end, observe that Lemma E.4 reveals

$$\nabla_x F(x^\star, 0) = \nabla_x \Big(\mathop{\mathbb{E}}_{z\sim\mathcal{D}_x} G(x,z) \Big)\Big|_{x=x^\star} = W,$$

which is invertible by Lemma 3.10. Consequently, the implicit function theorem yields open neighborhoods $U \subset \mathcal{U}$ of 0 and $V \subset \operatorname{int}\mathcal{X}$ of $x^\star$ and a C^1-smooth map $U \to V$ given by $u \mapsto x_u^\star$ with Jacobian $-W^{-1}\nabla_u F(x^\star, 0)$ at $u = 0$. This yields the first-order approximation

$$x_u^\star = x^\star - W^{-1}\nabla_u F(x^\star, 0)u + o(\|u\|) \qquad \text{as } u \to 0. \tag{E.11}$$

By Lemma E.4, we have

$$\nabla_u F(x,0) = \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x}\big[G(x,z)g_x(z)^\top\big]$$

for all $x \in \mathcal{X}$. In particular, $\nabla_u F(x^\star, 0) = \Sigma_{g,G}^\top$. Thus, (E.11) asserts

$$x_u^\star = x^\star - W^{-1}\Sigma_{g,G}^\top u + o(\|u\|) \qquad \text{as } u \to 0.$$

Consequently, for any fixed $u \in \mathbb{R}^d$, we have

$$\sqrt{k}\big(x^\star_{u/\sqrt{k}} - x^\star\big) = -W^{-1}\Sigma_{g,G}^\top u + \sqrt{k}\cdot o\left(\frac{1}{\sqrt{k}}\right) \to -W^{-1}\Sigma_{g,G}^\top u \qquad \text{as } k \to \infty.$$

The proof is complete.

E.4 Proof of Lemma E.4

Recall first that the quantities $M_g' := \sup_{x\in\mathcal{X},\, z\in\mathcal{Z}} \|\nabla_x g_x(z)\|_{\mathrm{op}}$ and $M_g := \sup_{x\in\mathcal{X},\, z\in\mathcal{Z}} \|g_x(z)\|$ are finite by Lemma E.1. The same argument shows that M_T' and M_T are finite.

Next, we turn to establishing that the map $H\colon \mathcal{X}\times\mathcal{W} \to \mathbb{R}^n$ given by

$$H(x,u) = \frac{1}{C_x^u} \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x}\big[T(x,z)\big(1 + h\big(u^\top g_x(z)\big)\big)\big]$$

is smooth with Lipschitz Jacobian on the compact set $\mathcal{K} := \mathcal{X} \times \mathcal{W}$. By Lemma F.7, it is enough to show that $(x, u) \mapsto C_x^u$ and

$$\hat{H}(x, u) := \mathop{\mathbb{E}}_{z \sim \mathcal{D}_x} \big[T(x, z) \big(1 + h\big(u^\top g_x(z) \big) \big) \big]$$

are smooth with Lipschitz Jacobians on $\mathcal{K}$; in turn, it suffices to establish this fact for $\hat{H}$ since we can then take $T \equiv 1$ to derive the result for C_x^u.

We reason this via the chain rule. Namely, consider the map $\bar{H} \colon \mathcal{X} \times \mathcal{X} \times \mathcal{W} \to \mathbb{R}^n$ given by

$$\bar{H}(x, y, u) = \mathop{\mathbb{E}}_{z \sim \mathcal{D}_x} \big[T(y, z) \big(1 + h\big(u^\top g_y(z) \big) \big) \big].$$

Clearly $\hat{H} = \bar{H} \circ J$ with $J(x, u) := (x, x, u)$ and therefore the chain rule implies $\nabla \hat{H}(x, u) = \nabla \bar{H}(x, x, u) \nabla J(x, u)$ provided $\bar{H}$ is smooth. Thus, it suffices to show that $\bar{H}$ is smooth with Lipschitz Jacobian. To this end, we demonstrate that the three partial derivatives of $\bar{H}$ are all Lipschitz with constants depending on T only through $\beta_T, \bar{\Lambda}_T, \beta_T'$, M_T, and M_T'.

We begin with the partial derivative of $\bar{H}$ with respect to x. Consider the function $\phi \colon \mathcal{K} \times \mathcal{Z} \to \mathbb{R}^n$ given by

$$\phi(y, u, z) = T(y, z) \big(1 + h\big(u^\top g_y(z) \big) \big).$$

Let us verify that ϕ is a test function to which item (ii) of Assumption 3.5 applies. Clearly ϕ is measurable and bounded with $\sup \|\phi\| \leq 2M_T$. Further, for each $z \in \mathcal{Z}$, it follows readily that the section $\phi(\cdot, z)$ is Lipschitz on $\mathcal{K}$ with constant

$$L_\phi := 2M_T' + M_T L_h \big(\operatorname{diam}(\mathcal{W}) M_g' + M_g \big),$$

where $L_h := \sup |h'|$. Thus, item (ii) of Assumption 3.5 implies that the map

$$x \mapsto \mathop{\mathbb{E}}_{z \sim \mathcal{D}_x} \phi(y, u, z) = \bar{H}(x, y, u)$$

is smooth on $\mathcal{X}$ for each $(y, u) \in \mathcal{K}$, and that the map

$$(x, y, u) \mapsto \nabla_x \bar{H}(x, y, u)$$

is Lipschitz on $\mathcal{X} \times \mathcal{X} \times \mathcal{W}$ with constant $\vartheta(L_\phi + 2M_T)$, which depends on T only through M_T and M_T'.

Next, we consider the partial derivative of $\bar{H}$ with respect to y. Given $(x, u) \in \mathcal{X} \times \mathcal{W}$, the dominated convergence theorem ensures that $\bar{H}(x, y, u)$ is smooth in y with

$$\nabla_y \bar{H}(x, y, u) = \mathop{\mathbb{E}}_{z \sim \mathcal{D}_x} \nabla_y \phi(y, u, z) \tag{E.12}$$

provided $\|\nabla_y \phi(y, u, z)\|_{\mathrm{op}}$ is dominated by a $\mathcal{D}_x$-integrable random variable independent of y. Using the product rule, we have

$$\nabla_y \phi(y, u, z) = \big(\nabla_y T(y, z)\big)\big(1 + h\big(u^\top g_y(z)\big)\big) + h'\big(u^\top g_y(z)\big)\big(T(y, z)u^\top\big)\nabla_y g_y(z) \tag{E.13}$$

and hence

$$\begin{aligned}\|\nabla_y \phi(y, u, z)\|_{\mathrm{op}} &\le 2\|\nabla_y T(y, z)\|_{\mathrm{op}} + (\sup |h'|)\|T(y, z)\|\|u\|\|\nabla_y g_y(z)\|_{\mathrm{op}} \\ &\le 2M_T' + \operatorname{diam}(\mathcal{W}) L_h M_T M_g',\end{aligned}$$

so $\nabla_y \phi$ is in fact uniformly bounded. Therefore $\bar{H}(x, y, u)$ is smooth in y and (E.12) holds. Moreover, it follows from (E.13) that the map

$$(x, y, u) \mapsto \nabla_y \bar{H}(x, y, u)$$

is Lipschitz on $\mathcal{X} \times \mathcal{X} \times \mathcal{W}$; we will verify this by computing Lipschitz constants separately in x, y, and u. To begin, note that it follows from (E.13) that $z \mapsto \nabla_y \phi(y, u, z)$ is Lipschitz on $\mathcal{Z}$ with constant

$$a := 2\beta_T' + \operatorname{diam}(\mathcal{W}) M_T' L_h \beta_g + \operatorname{diam}(\mathcal{W}) L_h (\beta_T M_g' + M_T \beta_g') + \operatorname{diam}(\mathcal{W})^2 M_T M_g' L_{h'} \beta_g,$$

where $L_{h'} := \sup |h''|$. Hence (E.12) and Assumption 3.1 imply that $x \mapsto \nabla_y \bar{H}(x, y, u)$ is Lipschitz on $\mathcal{X}$ with constant γa, which depends on T only through β_T, β_T', M_T, and M_T'. Likewise, it follows from (E.13) that $y \mapsto \nabla_y \phi(y, u, z)$ is Lipschitz on $\mathcal{X}$ with constant

$$2\Lambda_T(z) + \operatorname{diam}(\mathcal{W}) M_T' L_h M_g' + \operatorname{diam}(\mathcal{W}) L_h \big(M_T \Lambda_g(z) + M_T' M_g'\big) + \operatorname{diam}(\mathcal{W})^2 M_T L_{h'} (M_T')^2.$$

Hence (E.12) implies that $y \mapsto \nabla_y \bar{H}(x, y, u)$ is Lipschitz on $\mathcal{X}$ with constant

$$2\bar{\Lambda}_T + \operatorname{diam}(\mathcal{W}) L_h \big(M_T \bar{\Lambda}_g + 2M_T' M_g'\big) + \operatorname{diam}(\mathcal{W})^2 M_T L_{h'} (M_T')^2.$$

Similarly, it follows from (E.13) that $u \mapsto \nabla_y \phi(y, u, z)$ is Lipschitz on $\mathcal{W}$ with constant

$$M_T' M_g L_h + M_T M_g' \big(L_h + \operatorname{diam}(\mathcal{W}) M_g L_{h'}\big),$$

so (E.12) implies that $u \mapsto \nabla_y \bar{H}(x, y, u)$ is Lipschitz on $\mathcal{W}$ with the same constant. We conclude therefore that the map $(x, y, u) \mapsto \nabla_y \bar{H}(x, y, u)$ is Lipschitz on $\mathcal{X} \times \mathcal{X} \times \mathcal{W}$ with constant depending on T only through $\beta_T, \bar{\Lambda}_T, \beta'_T$, M_T, and M'_T.

Finally, we consider the partial derivative of $\bar{H}$ with respect to u. Given $(x, y) \in \mathcal{X} \times \mathcal{X}$, the dominated convergence theorem ensures that $\bar{H}(x, y, u)$ is smooth in u with

$$\nabla_u \bar{H}(x, y, u) = \mathop{\mathbb{E}}_{z \sim \mathcal{D}_x} \nabla_u \phi(y, u, z) \tag{E.14}$$

provided $\|\nabla_u \phi(y, u, z)\|_{\text{op}}$ is dominated by a $\mathcal{D}_x$-integrable random variable independent of u. In this case, we have

$$\nabla_u \phi(y, u, z) = h'\big(u^\top g_y(z)\big) T(y, z) g_y(z)^\top \tag{E.15}$$

and hence

$$\|\nabla_u \phi(y, u, z)\|_{\text{op}} \le (\sup |h'|) \|T(y, z)\| \|g_y(z)\| \le L_h M_T M_g.$$

Therefore $\bar{H}(x, y, u)$ is smooth in u and (E.14) holds. Moreover, it follows from (E.15) that the map

$$(x, y, u) \mapsto \nabla_u \bar{H}(x, y, u)$$

is Lipschitz on $\mathcal{X} \times \mathcal{X} \times \mathcal{W}$; as before, we will verify this by computing Lipschitz constants separately in x, y, and u. First, note that it follows from (E.15) that $z \mapsto \nabla_u \phi(y, u, z)$ is Lipschitz on $\mathcal{Z}$ with constant

$$b := L_h(\beta_T M_g + M_T \beta_g) + \operatorname{diam}(\mathcal{W}) M_T M_g L_{h'} \beta_g.$$

Hence (E.14) and Assumption 3.1 imply that $x \mapsto \nabla_u \bar{H}(x, y, u)$ is Lipschitz on $\mathcal{X}$ with constant γb, which depends on T only through β_T and M_T. Likewise, it follows from (E.15) that $y \mapsto \nabla_u \phi(y, u, z)$ is Lipschitz on $\mathcal{X}$ with constant

$$L_h(M'_T M_g + M_T M'_g) + \operatorname{diam}(\mathcal{W}) M_T M_g L_{h'} M'_g,$$

hence so is $y \mapsto \nabla_u \bar{H}(x, y, u)$ by (E.14). Similarly, it follows from (E.15) that $u \mapsto \nabla_u \phi(y, u, z)$ is $L_{h'} M_T M_g^2$-Lipschitz on $\mathcal{W}$, hence so is $u \mapsto \nabla_u \bar{H}(x, y, u)$ by (E.14). We conclude therefore

that the map $(x, y, u) \mapsto \nabla_u \bar{H}(x, y, u)$ is Lipschitz on $\mathcal{X} \times \mathcal{X} \times \mathcal{W}$ with constant depending on T only through β_T, M_T, and M_T'.

The preceding reveals that $\bar{H}$ and hence $\hat{H} = \bar{H} \circ J$ are smooth, with Lipschitz Jacobians with constants depending on T only through $\beta_T, \bar{\Lambda}_T, \beta_T'$, M_T, and M_T'. Taking $T \equiv 1$, we conclude that $(x, u) \mapsto C_x^u$ is smooth, with Lipschitz Jacobian with constant independent of T. Upon observing in the same way as above that $\bar{H}$ and hence $\hat{H}$ are Lipschitz with constants depending on T only through β_T, M_T, and M_T', it follows from Lemma F.7 and its proof that H is smooth, with Lipschitz Jacobian with constant depending on T only through $\beta_T, \bar{\Lambda}_T, \beta_T'$, M_T, and M_T'.

Finally, given any $x \in \mathcal{X}$, the equalities

$$\nabla_x H(x, 0) = \nabla_x \Big(\mathop{\mathbb{E}}_{z \sim \mathcal{D}_x} T(x, z) \Big) \qquad \text{and} \qquad \nabla_u H(x, 0) = \mathop{\mathbb{E}}_{z \sim \mathcal{D}_x} \big[T(x, z) g_x(z)^\top \big]$$

follow from straightforward computations (using the quotient rule, dominated convergence theorem, and chain and product rules). This completes the proof.

Appendix F

SUPPLEMENTARY RESULTS

In this appendix, we record some supplementary results fundamental to our analysis in Chapter 3. The following lemma shows that Assumption 3.1 implies $\{\mathcal{D}_x\}_{x\in\mathcal{X}}$ is a Markov kernel from $\mathcal{X}$ to $\mathcal{Z}$, i.e., for each $E \in \mathcal{B}(\mathcal{Z})$, the function $\mathcal{X} \to [0,1]$ given by $x \mapsto \mathcal{D}_x(E)$ is measurable.

Lemma F.1 (Markov kernel)**.** *Let $\mathcal{Z}$ be a nonempty Polish metric space. For any bounded measurable function $\varphi\colon \mathcal{Z} \to \mathbb{R}$, the function $P_1(\mathcal{Z}) \to \mathbb{R}$ given by $\mu \mapsto \int \varphi\, d\mu$ is measurable. Thus, for any measurable space $\mathcal{X}$ and any measurable map $x \mapsto \mathcal{D}_x$ from $\mathcal{X}$ to $P_1(\mathcal{Z})$, it follows that $\{\mathcal{D}_x\}_{x\in\mathcal{X}}$ is a Markov kernel from $\mathcal{X}$ to $\mathcal{Z}$.*

Proof. Let $\mathcal{M}_b$ denote the set of all bounded measurable functions $\mathcal{Z} \to \mathbb{R}$. For each $\varphi \in \mathcal{M}_b$, let $I_\varphi\colon P_1(\mathcal{Z}) \to \mathbb{R}$ be the function given by $I_\varphi(\mu) = \int \varphi\, d\mu$. Now consider the set

$$\mathcal{C} = \big\{\varphi \in \mathcal{M}_b \mid I_\varphi \text{ is measurable}\big\}.$$

To demonstrate $\mathcal{C} = \mathcal{M}_b$, it suffices by the functional monotone class theorem [46, Exercise 11.7] to show that $\mathcal{C}$ possesses the following two properties:

(i) Every bounded continuous function $\mathcal{Z} \to \mathbb{R}$ is contained in $\mathcal{C}$.

(ii) If (φ_n) is a uniformly bounded sequence in $\mathcal{C}$ with pointwise limit $\varphi\colon \mathcal{Z} \to \mathbb{R}$ (i.e., $\sup_{n,z} |\varphi_n(z)| < \infty$ and $\lim_{n\to\infty} \varphi_n(z) = \varphi(z)$ for all $z \in \mathcal{Z}$), then $\varphi \in \mathcal{C}$.

To this end, note first that (i) holds because W_1-convergence in $P_1(\mathcal{Z})$ implies weak convergence [3, Proposition 7.1.5]; indeed, if $\varphi\colon \mathcal{Z} \to \mathbb{R}$ is bounded and continuous, then for any sequence (μ_n) in $P_1(\mathcal{Z})$ such that $W_1(\mu_n, \mu) \to 0$ for some $\mu \in P_1(\mathcal{Z})$, we have $I_\varphi(\mu_n) \to I_\varphi(\mu)$, so I_φ

is continuous and hence measurable. On the other hand, (ii) follows from the dominated convergence theorem: if (φ_n) is a uniformly bounded sequence in $\mathcal{C}$ with pointwise limit $\varphi\colon \mathcal{Z} \to \mathbb{R}$, then $\varphi \in \mathcal{M}_b$ and

$$I_\varphi(\mu) = \int \lim_{n\to\infty} \varphi_n(z)\, d\mu(z) = \lim_{n\to\infty} \int \varphi_n(z)\, d\mu(z) = \lim_{n\to\infty} I_{\varphi_n}(\mu)$$

for all $\mu \in P_1(\mathcal{Z})$, so I_φ is measurable as the pointwise limit of the sequence of measurable functions (I_{φ_n}). Hence $\mathcal{C} = \mathcal{M}_b$, i.e., I_φ is measurable for every bounded measurable function $\varphi\colon \mathcal{Z} \to \mathbb{R}$; in particular, the last claim of the lemma follows by taking φ to be the indicator function $\mathbf{1}_E$ of any measurable set $E \in \mathcal{B}(\mathcal{X})$. □

We will require the existence of the probability measure $\bigotimes_{i=0}^{\infty} \mathcal{D}_{x_i}$ on the countable product space $\mathcal{Z}^{\mathbb{N}}$ with marginals given by recursive application of the Markov kernel $\{\mathcal{D}_x\}_{x\in\mathcal{X}}$ from $\mathcal{X}$ to $\mathcal{Z}$ along a sequence of measurable maps $x_t\colon \mathcal{Z}^t \to \mathcal{X}$. The following theorem may be viewed as a special case of either the Kolmogorov extension theorem [6, Appendix D] or the Ionescu–Tulcea extension theorem [47, Theroem 14.35].

Theorem F.2 (Ionescu-Tulcea)**.** *Let $\mathcal{X}$ be a measurable space, $\mathcal{Z}$ be a nonempty Polish metric space, $\{\mathcal{D}_x\}_{x\in\mathcal{X}}$ be a Markov kernel from $\mathcal{X}$ to $\mathcal{Z}$, and $x_t\colon \mathcal{Z}^t \to \mathcal{X}$ be a sequence of measurable maps (with $x_0 \in \mathcal{X}$). For each $t \geq 1$, let $\mathbb{P}_t = \bigotimes_{i=0}^{t-1} \mathcal{D}_{x_i}$ be the probability measure on $\mathcal{Z}^t$ defined recursively by setting $\mathbb{P}_1 = \mathcal{D}_{x_0}$ and*

$$\mathbb{P}_{t+1}(A \times E) = \int_A \mathcal{D}_{x_t}(E)\, d\,\mathbb{P}_t \qquad \text{for all } A \in \mathcal{B}(\mathcal{Z}^t) \text{ and } E \in \mathcal{B}(\mathcal{Z}),$$

and let $\pi_t\colon \mathcal{Z}^{\mathbb{N}} \to \mathcal{Z}^t$ denote the projection from the countable product space $\mathcal{Z}^{\mathbb{N}}$ onto the first t coordinates. Then there exists a unique probability measure $\mathbb{P} = \bigotimes_{i=0}^{\infty} \mathcal{D}_{x_i}$ on $\mathcal{Z}^{\mathbb{N}}$ satisfying $(\pi_t)_{\#}\,\mathbb{P} = \mathbb{P}_t$ for all $t \geq 1$, that is,

$$\mathbb{P}(A \times \mathcal{Z}^{\mathbb{N}}) = \mathbb{P}_t(A) \qquad \text{for all } A \in \mathcal{B}(\mathcal{Z}^t) \text{ and } t \geq 1.$$

Thus, for every $t \geq 0$ and every measurable function $\varphi\colon \mathcal{Z}^{t+1} \to \overline{\mathbb{R}}$ that is nonnegative or $\mathbb{P}_{t+1}$-integrable, we have

$$\mathbb{E}[\varphi \circ \pi_{t+1}] = \int_{\mathcal{Z}^{t+1}} \varphi\, d\,\mathbb{P}_{t+1} = \int_{\mathcal{Z}} \cdots \int_{\mathcal{Z}} \varphi(z_0, \ldots, z_t)\, d\mathcal{D}_{x_t}(z_t) \cdots d\mathcal{D}_{x_0}(z_0)$$

and

$$\mathbb{E}[\varphi \circ \pi_{t+1} \,|\, \mathcal{F}_t] = \int_{\mathcal{Z}} \varphi(z_0, \dots, z_t)\, d\mathcal{D}_{x_t}(z_t) = \mathop{\mathbb{E}}_{z_t \sim \mathcal{D}_{x_t}} [\varphi(z_0, \dots, z_t)],$$

where $\mathcal{F}_t = \{A \times \mathcal{Z}^{\mathbb{N}} \mid A \in \mathcal{B}(\mathcal{Z}^t)\}$ *denotes the* σ*-algebra generated by* π_t *(with* $\mathcal{F}_0 = \{\emptyset, \mathcal{Z}^{\mathbb{N}}\}$*).*

Next, we record suitably general versions of the Strong Law of Large Numbers and the Central Limit Theorem for square-integrable martingales.

Theorem F.3 (Martingale Strong Law of Large Numbers [22, Exercise 5.3.35])**.** *Let* X *be a square-integrable martingale difference sequence in* $\mathbb{R}^n$ *adapted to a filtration* $(\mathcal{F}_k)$ *and* (a_k) *be a sequence of positive constants such that* $a_k \uparrow \infty$ *as* $k \to \infty$*. Then on the event* $\big\{ \sum_{i=1}^{\infty} a_i^{-2}\, \mathbb{E}[\|X_i\|^2 \,|\, \mathcal{F}_{i-1}] < \infty \big\}$*, we have* $a_k^{-1} \sum_{i=1}^{k} X_i \to 0$ *almost surely as* $k \to \infty$*. In particular, if* $\sum_{i=1}^{\infty} a_i^{-2}\, \mathbb{E}\|X_i\|^2 < \infty$*, then* $a_k^{-1} \sum_{i=1}^{k} X_i \to 0$ *almost surely as* $k \to \infty$*.*

Theorem F.4 (Martingale Central Limit Theorem [26, Corollary 2.1.10])**.** *Let* M *be a square-integrable martingale in* $\mathbb{R}^n$ *adapted to a filtration* $(\mathcal{F}_k)$*, and let* $\langle M \rangle$ *denote the predictable quadratic variation of* M*:*

$$\langle M \rangle_k = \sum_{i=1}^{k} \mathbb{E}\big[(M_i - M_{i-1})(M_i - M_{i-1})^\top \,|\, \mathcal{F}_{i-1}\big] \qquad \text{for all } k \geq 1.$$

Let (a_k) *be a sequence of positive constants such that* $a_k \uparrow \infty$ *as* $k \to \infty$*. Suppose that the following two assumptions hold.*

*(i) (**Asymptotic covariance**) There is a deterministic positive semidefinite matrix* Σ *satisfying*

$$a_k^{-1} \langle M \rangle_k \xrightarrow{p} \Sigma \qquad \text{as } k \to \infty.$$

*(ii) (**Lindeberg's condition**) For all* $\varepsilon > 0$*,*

$$a_k^{-1} \sum_{i=1}^{k} \mathbb{E}\big[\|M_i - M_{i-1}\|^2 \mathbf{1}_{\{\|M_i - M_{i-1}\| \geq \varepsilon a_k^{1/2}\}} \,|\, \mathcal{F}_{i-1}\big] \xrightarrow{p} 0 \qquad \text{as } k \to \infty.$$

Then

$$a_k^{-1} M_k \xrightarrow{\text{a.s.}} 0 \quad \text{and} \quad a_k^{-1/2} M_k \rightsquigarrow \mathsf{N}(0, \Sigma) \qquad \text{as } k \to \infty.$$

The following lemma is used multiple times in our arguments to compute limits of covariance matrices.

Lemma F.5 (Asymptotic covariance)**.** *Let $x_t \in \mathcal{X}$ be a sequence in some set $\mathcal{X} \subset \mathbb{R}^d$ converging to some point $x^\star \in \mathcal{X}$, and let $\mu_t \in P_1(\mathcal{Z})$ be a sequence of probability measures on a nonempty Polish space $\mathcal{Z}$ converging to some measure $\mu^\star \in P_1(\mathcal{Z})$ in the Wasserstein-1 metric. Suppose that $g\colon \mathcal{X} \times \mathcal{Z} \to \mathbb{R}^n$ is a measurable map satisfying the following two conditions.*

*(i) (**Asymptotic uniform integrability**) For every $\delta > 0$, there exists a constant $N_\delta \geq 0$ such that*

$$\limsup_{t\to\infty} \mathop{\mathbb{E}}_{z\sim\mu_t}\left[\|g(x^\star,z)\|^2 \mathbf{1}_{\{\|g(x^\star,z)\|\geq N_\delta\}}\right] \leq \delta,$$
$$\mathop{\mathbb{E}}_{z\sim\mu^\star}\left[\|g(x^\star,z)\|^2 \mathbf{1}_{\{\|g(x^\star,z)\|\geq N_\delta\}}\right] \leq \delta.$$

*(ii) (**Lipschitz continuity**) There exist a neighborhood $\mathcal{V}$ of $x^\star$, a measurable function $L\colon \mathcal{Z} \to [0,\infty)$, and constants $\beta, \bar{L} \geq 0$ such that for every $z \in \mathcal{Z}$, the section $g(\cdot, z)$ is $L(z)$-Lipschitz on $\mathcal{V}$ with $\limsup_{t\to\infty} \mathbb{E}_{z\sim\mu_t}[L(z)^2] \leq \bar{L}^2$, and the section $g(x^\star, \cdot)$ is β-Lipschitz on $\mathcal{Z}$.*

Then

$$\lim_{t\to\infty} \mathop{\mathbb{E}}_{z\sim\mu_t}\left[g(x_t,z)g(x_t,z)^\top\right] = \mathop{\mathbb{E}}_{z\sim\mu^\star}\left[g(x^\star,z)g(x^\star,z)^\top\right].$$

Proof. For notational convenience, set $g_x(z) = g(x,z)$ and

$$\Sigma = \mathop{\mathbb{E}}_{z\sim\mu^\star}\left[g_{x^\star}(z)g_{x^\star}(z)^\top\right].$$

For any $\delta > 0$, the decomposition

$$\mathop{\mathbb{E}}_{z\sim\mu_t}\left[\|g_{x^\star}(z)\|^2\right] = \mathop{\mathbb{E}}_{z\sim\mu_t}\left[\|g_{x^\star}(z)\|^2 \mathbf{1}_{\{\|g_{x^\star}(z)\|<N_\delta\}}\right] + \mathop{\mathbb{E}}_{z\sim\mu_t}\left[\|g_{x^\star}(z)\|^2 \mathbf{1}_{\{\|g_{x^\star}(z)\|\geq N_\delta\}}\right]$$

holds for all t, so condition (i) implies

$$\limsup_{t\to\infty} \mathop{\mathbb{E}}_{z\sim\mu_t}\left[\|g_{x^\star}(z)\|^2\right] \leq N_\delta^2 + \delta. \tag{F.1}$$

On the other hand, for all t, we also have the decomposition

$$\begin{aligned}\mathop{\mathbb{E}}_{z\sim\mu_t}\left[g_{x_t}(z)g_{x_t}(z)^\top\right] = \mathop{\mathbb{E}}_{z\sim\mu_t}\left[g_{x^\star}(z)g_{x^\star}(z)^\top\right] + \mathop{\mathbb{E}}_{z\sim\mu_t}\left[g_{x^\star}(z)\big(g_{x_t}(z)-g_{x^\star}(z)\big)^\top\right] \\ + \mathop{\mathbb{E}}_{z\sim\mu_t}\left[\big(g_{x_t}(z)-g_{x^\star}(z)\big)g_{x_t}(z)^\top\right].\end{aligned} \tag{F.2}$$

The last two summands in (F.2) tend to zero as $t\to\infty$. Indeed, since $x_t\to x^\star$ as $t\to\infty$, we have $x_t\in\mathcal{V}$ for all but finitely many t and so we may apply condition (ii) together with Hölder's inequality and (F.1) to conclude

$$\begin{aligned}\left\|\mathop{\mathbb{E}}_{z\sim\mu_t}\left[g_{x^\star}(z)\big(g_{x_t}(z)-g_{x^\star}(z)\big)^\top\right]\right\|_{\mathrm{op}} &\le \mathop{\mathbb{E}}_{z\sim\mu_t}\left[\|g_{x^\star}(z)\|\cdot\|g_{x_t}(z)-g_{x^\star}(z)\|\right] \\ &\le \|x_t-x^\star\|\sqrt{\mathop{\mathbb{E}}_{z\sim\mu_t}\left[\|g_{x^\star}(z)\|^2\right]\mathop{\mathbb{E}}_{z\sim\mu_t}\left[L(z)^2\right]} \\ &\to 0 \qquad \text{as } t\to\infty\end{aligned}$$

and

$$\begin{aligned}&\left\|\mathop{\mathbb{E}}_{z\sim\mu_t}\left[\big(g_{x_t}(z)-g_{x^\star}(z)\big)g_{x_t}(z)^\top\right]\right\|_{\mathrm{op}} \\ &\qquad\le \mathop{\mathbb{E}}_{z\sim\mu_t}\left[\|g_{x_t}(z)-g_{x^\star}(z)\|\cdot\|g_{x_t}(z)\|\right] \\ &\qquad\le \|x_t-x^\star\|\sqrt{\mathop{\mathbb{E}}_{z\sim\mu_t}\left[L(z)^2\right]\mathop{\mathbb{E}}_{z\sim\mu_t}\left[\|g_{x_t}(z)\|^2\right]} \\ &\qquad\le \|x_t-x^\star\|\sqrt{2\mathop{\mathbb{E}}_{z\sim\mu_t}\left[L(z)^2\right]\Big(\mathop{\mathbb{E}}_{z\sim\mu_t}\left[\|g_{x^\star}(z)\|^2\right]+\mathop{\mathbb{E}}_{z\sim\mu_t}\left[\|g_{x_t}(z)-g_{x^\star}(z)\|^2\right]\Big)} \\ &\qquad\le \|x_t-x^\star\|\sqrt{2\mathop{\mathbb{E}}_{z\sim\mu_t}\left[L(z)^2\right]\Big(\mathop{\mathbb{E}}_{z\sim\mu_t}\left[\|g_{x^\star}(z)\|^2\right]+\|x_t-x^\star\|^2\cdot\mathop{\mathbb{E}}_{z\sim\mu_t}\left[L(z)^2\right]\Big)} \\ &\qquad\to 0 \qquad \text{as } t\to\infty.\end{aligned}$$

To complete the proof, it now suffices by (F.2) to show $\mathbb{E}_{z\sim\mu_t}[g_{x^\star}(z)g_{x^\star}(z)^\top]\to\Sigma$ as $t\to\infty$. To this end, define for each $q\in\mathbb{R}$ the step-like function $\varphi_q\colon\mathbb{R}\to\mathbb{R}$ by setting

$$\varphi_q(x)=\begin{cases}1 & \text{if } x\le q,\\ -x+q+1 & \text{if } q\le x\le q+1,\\ 0 & \text{if } q+1\le x.\end{cases}$$

Let $\delta > 0$ be arbitrary. Then for any given t, we have the decomposition

$$\begin{aligned}
&\mathop{\mathbb{E}}_{z\sim\mu_t}\big[g_{x^\star}(z)g_{x^\star}(z)^\top\big] - \Sigma \\
&\quad= \mathop{\mathbb{E}}_{z\sim\mu_t}\big[g_{x^\star}(z)g_{x^\star}(z)^\top\big] - \mathop{\mathbb{E}}_{z\sim\mu^\star}\big[g_{x^\star}(z)g_{x^\star}(z)^\top\big] \\
&\quad= \underbrace{\mathop{\mathbb{E}}_{z\sim\mu_t}\big[\big(1-\varphi_{N_\delta}(\|g_{x^\star}(z)\|)\big)g_{x^\star}(z)g_{x^\star}(z)^\top\big] - \mathop{\mathbb{E}}_{z\sim\mu^\star}\big[\big(1-\varphi_{N_\delta}(\|g_{x^\star}(z)\|)\big)g_{x^\star}(z)g_{x^\star}(z)^\top\big]}_{A_t} \\
&\qquad+ \underbrace{\mathop{\mathbb{E}}_{z\sim\mu_t}\big[\varphi_{N_\delta}(\|g_{x^\star}(z)\|)g_{x^\star}(z)g_{x^\star}(z)^\top\big] - \mathop{\mathbb{E}}_{z\sim\mu^\star}\big[\varphi_{N_\delta}(\|g_{x^\star}(z)\|)g_{x^\star}(z)g_{x^\star}(z)^\top\big]}_{B_t}.
\end{aligned}$$

By the triangle inequality, $\|A_t\|_{\mathrm{op}}$ is bounded above by

$$\begin{aligned}
&\Big\|\mathop{\mathbb{E}}_{z\sim\mu_t}\big[\big(1-\varphi_{N_\delta}(\|g_{x^\star}(z)\|)\big)g_{x^\star}(z)g_{x^\star}(z)^\top\big]\Big\|_{\mathrm{op}} + \Big\|\mathop{\mathbb{E}}_{z\sim\mu^\star}\big[\big(1-\varphi_{N_\delta}(\|g_{x^\star}(z)\|)\big)g_{x^\star}(z)g_{x^\star}(z)^\top\big]\Big\|_{\mathrm{op}} \\
&\qquad\le \mathop{\mathbb{E}}_{z\sim\mu_t}\big[\|g_{x^\star}(z)\|^2\mathbf{1}_{\{\|g_{x^\star}(z)\|\ge N_\delta\}}\big] + \mathop{\mathbb{E}}_{z\sim\mu^\star}\big[\|g_{x^\star}(z)\|^2\mathbf{1}_{\{\|g_{x^\star}(z)\|\ge N_\delta\}}\big],
\end{aligned}$$

so

$$\limsup_{t\to\infty}\|A_t\|_{\mathrm{op}} \le 2\delta.$$

In order to bound B_t, consider the map $\Phi\colon \mathbb{R}^n \to \mathbb{R}^{n\times n}$ given by $\Phi(w) = \varphi_{N_\delta}(\|w\|)ww^\top$, set $\phi = \Phi \circ g_{x^\star}$, and note

$$B_t = \mathop{\mathbb{E}}_{z\sim\mu_t}\big[\phi(z)\big] - \mathop{\mathbb{E}}_{z\sim\mu^\star}\big[\phi(z)\big].$$

Clearly Φ is Lipschitz continuous on any compact set and zero outside of the ball of radius $N_\delta + 1$ centered at the origin. Therefore Φ is globally Lipschitz. Since $g_{x^\star}$ is β-Lipschitz on $\mathcal{Z}$ by condition (ii), we conclude that ϕ is Lipschitz on $\mathcal{Z}$ with a constant C that depends only on N_δ and β. Consequently,

$$\begin{aligned}
\|B_t\|_{\mathrm{op}} &= \Big\|\mathop{\mathbb{E}}_{z\sim\mu_t}\big[\phi(z)\big] - \mathop{\mathbb{E}}_{z\sim\mu^\star}\big[\phi(z)\big]\Big\|_{\mathrm{op}} \\
&= \sup_{\|u\|,\|v\|\le 1}\Big\{\mathop{\mathbb{E}}_{z\sim\mu_t}\big[\langle\phi(z)u, v\rangle\big] - \mathop{\mathbb{E}}_{z\sim\mu^\star}\big[\langle\phi(z)u, v\rangle\big]\Big\} \\
&\le C\cdot W_1(\mu_t, \mu^\star) \to 0 \qquad \text{as } t\to\infty,
\end{aligned}$$

where the inequality follows from the C-Lipschitz continuity of the function $z \mapsto \langle\phi(z)u, v\rangle$.

Hence

$$\limsup_{t\to\infty} \left\| \mathop{\mathbb{E}}_{z\sim\mu_t} \left[g_{x^\star}(z) g_{x^\star}(z)^\top \right] - \Sigma \right\|_{\mathrm{op}} \le \limsup_{t\to\infty} \left(\|A_t\|_{\mathrm{op}} + \|B_t\|_{\mathrm{op}} \right) \le 2\delta.$$

Since $\delta > 0$ is arbitrary, we deduce $\mathbb{E}_{z\sim\mu_t}[g_{x^\star}(z) g_{x^\star}(z)^\top] \to \Sigma$ as $t \to \infty$. □

Finally, we record two basic lemmas about products and quotients of Lipschitz functions.

Lemma F.6. *Let $\mathcal{K}$ be a metric space and suppose that $f\colon \mathcal{K} \to \mathbb{R}^{n\times q}$ and $g\colon \mathcal{K} \to \mathbb{R}^{q\times m}$ are bounded and Lipschitz. Then the product $fg\colon \mathcal{K} \to \mathbb{R}^{n\times m}$ is Lipschitz.*

Proof. Let L_f and L_g be the Lipschitz constants of f and g with respect to the operator norm $\|\cdot\|$. Then for all $x, y \in \mathcal{K}$, we have

$$\begin{aligned}
\|f(x)g(x) - f(y)g(y)\| &\le \|f(x)(g(x) - g(y))\| + \|(f(x) - f(y))g(y)\| \\
&\le \sup_{z\in\mathcal{K}} \|f(z)\| \cdot \|g(x) - g(y)\| + \|f(x) - f(y)\| \cdot \sup_{z\in\mathcal{K}} \|g(z)\| \\
&\le \left(L_g \cdot \sup_{z\in\mathcal{K}} \|f(z)\| + L_f \cdot \sup_{z\in\mathcal{K}} \|g(z)\| \right) \cdot d_{\mathcal{K}}(x, y).
\end{aligned}$$

Since f and g are bounded, this demonstrates that fg is Lipschitz. □

Lemma F.7. *Let $\mathcal{K} \subset \mathbb{R}^m$ be a compact set and suppose that $f\colon \mathcal{K} \to \mathbb{R}^n$ and $g\colon \mathcal{K} \to \mathbb{R} \setminus \{0\}$ are C^1-smooth with Lipschitz Jacobians. Then f/g is C^1-smooth with Lipschitz Jacobian.*

Proof. Since f and g are C^1-smooth, it follows immediately from the quotient rule that f/g is C^1-smooth with Jacobian given by

$$\nabla(f/g) = (1/g)(\nabla f) - (f/g^2)(\nabla g)^\top. \tag{F.3}$$

By assumption, ∇f and ∇g are Lipschitz, and they are bounded by the compactness of $\mathcal{K}$. Further, the functions $1/g$ and f/g^2 are C^1-smooth, so they are locally Lipschitz by the mean value theorem; hence $1/g$ and f/g^2 are Lipschitz and bounded by the compactness of $\mathcal{K}$. Thus, (F.3) and Lemma F.6 show that $\nabla(f/g)$ is the difference of two Lipschitz maps. Therefore $\nabla(f/g)$ is Lipschitz. □

BIBLIOGRAPHY

[1] 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.

[2] Shabbir Ahmed. *Strategic Planning Under Uncertainty: Stochastic Integer Programming Approaches*. University of Illinois at Urbana-Champaign, 2000.

[3] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 2nd edition, 2008.

[4] 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.

[5] 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*, pages 335–342. Omnipress, 2008.

[6] Richard F. Bass. *Stochastic Processes*. Cambridge University Press, 2011.

[7] Yahav Bechavod, Katrina Ligett, Zhiwei Steven Wu, and Juba Ziani. Causal feature discovery through strategic modification. *arXiv:2002.07024*, 2020.

[8] Amir Beck. *First-Order Methods in Optimization*. MOS-SIAM Series on Optimization. Society for Industrial & Applied Mathematics, 2017.

[9] Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive algorithms and stochastic approximations*, volume 22. Springer Science & Business Media, 2012.

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

[11] 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.

[12] 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.

[13] 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.

[14] Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 6045–6061. PMLR, 2022.

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

[16] Joshua Cutler, Dmitriy Drusvyatskiy, and Zaid Harchaoui. Stochastic optimization under time drift: iterate averaging, step decay, and high-probability guarantees. *Advances in Neural Information Processing Systems*, 34:11859–11869, 2021.

[17] Joshua Cutler, Dmitriy Drusvyatskiy, and Zaid Harchaoui. Stochastic Optimization under Distributional Drift. *Journal of Machine Learning Research*, 24(147):1–56, 2023.

[18] Joshua Cutler, Mateo Díaz, and Dmitriy Drusvyatskiy. Stochastic approximation with decision-dependent distributions: asymptotic normality and optimality. *arXiv:2207.04173*, 2022.

[19] 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.

[20] 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.

[21] Damek Davis, Dmitriy Drusvyatskiy, and Liwei Jiang. Subgradient methods near active manifolds: saddle point avoidance, local convergence, and asymptotic normality. *arXiv:2108.11832*, 2021.

[22] Amir Dembo. *Probability Theory: STAT310/MATH230*. Stanford University, 2021.

[23] Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 55–70, 2018.

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

[25] John C. Duchi and Feng Ruan. Asymptotic optimality in stochastic optimization. *The Annals of Statistics*, 49(1):21–48, 2021.

[26] Marie Duflo. *Random Iterative Models*. Stochastic Modeling and Applied Probability. Springer Berlin, Heidelberg, 1997.

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

[28] Jitka Dupačová. Optimization under exogenous and endogenous uncertainty. *University of West Bohemia in Pilsen*, 2006.

[29] Rick Durrett. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 5th edition, 2019.

[30] 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.

[31] AA Gaivoronskii. Nonstationary stochastic programming problems. *Cybernetics*, 14(4):575–579, 1978.

[32] 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.

[33] 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.

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

[35] 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.

[36] 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.

[37] 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.

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

[39] Lars Hellemo, Paul I. Barton, and Asgeir Tomasgard. Decision-dependent probabilities in stochastic programs with recourse. *Computational Management Science*, 15(3):369–395, 2018.

[40] Zachary Izzo, Lexing Ying, and James Zou. How to learn when data reacts to your model: performative gradient descent. In *International Conference on Machine Learning*, pages 4641–4650. PMLR, 2021.

[41] 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.

[42] Meena Jagadeesan, Tijana Zrnic, and Celestine Mendler-Dünner. Regret minimization with performative feedback. In *International Conference on Machine Learning*, pages 9760–9785. PMLR, 2022.

[43] 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.

[44] Tore W. Jonsbråten, Roger J.-B. Wets, and David L. Woodruff. A class of stochastic programs with decision dependent random elements. *Annals of Operations Research*, 82:83–106, 1998.

[45] 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.

[46] Alexander S. Kechris. *Classical Descriptive Set Theory*. Springer, 1995.

[47] Achim Klenke. *Probability Theory: A Comprehensive Course*. Springer International Publishing, 3rd edition, 2020.

[48] 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.

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

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

[51] 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.

[52] Lucien Le Cam and Grace Lo Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer, 2000.

[53] 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.

[54] 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.

[55] John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning*, pages 6917–6926. PMLR, 2020.

[56] John Miller, Juan Perdomo, and Tijana Zrnic. Outside the echo chamber: Optimizing the performative risk. In *International Conference on Machine Learning*, pages 7710–7720. PMLR, 2021.

[57] 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.

[58] Adhyyan Narang, Evan Faulkner, Dmitriy Drusvyatskiy, Maryam Fazel, and Lillian J. Ratliff. Multiplayer performative prediction: Learning in decision-dependent games. *Journal of Machine Learning Research*, 24(202):1–56, 2023.

[59] 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.

[60] 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.

[61] Georgios Piliouras and Fang-Yi Yu. Multi-agent performative prediction: From global stability and optimality to chaos. *arXiv:2201.10483*, 2022.

[62] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[63] 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.

[64] Mitas Ray, Lillian J. Ratliff, Dmitriy Drusvyatskiy, and Maryam Fazel. Decision-dependent risk minimization in geometrically decaying dynamic environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8081–8088, 2022.

[65] H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics*, pages 233–257. Academic Press, 1971.

[66] R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*. Springer, 1998.

[67] R. Y. Rubinstein and A. Shapiro. *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method*. Wiley Series in Probability and Mathematical Statistics. Wiley, 1993.

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

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

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

[71] 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.

[72] Y. Z. Tsypkin and Z. J. Nikolic. *Adaptation and Learning in Automatic Systems*. Elsevier Science, 1971.

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

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

[75] Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[76] Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

[77] Pravin Varaiya and Roger J.-B. Wets. Stochastic dynamic optimization approaches and computation. IIASA WP-88-87, International lnst tute for Applied Systems Analysis, Laxenburg, Austria, 1988.

[78] 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.

[79] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

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

[81] 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.

[82] Killian Wood and Emiliano Dall'Anese. Stochastic saddle point problems with decision-dependent distributions. *arXiv:2201.02313*, 2022.

[83] 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.

[84] 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.

[85] 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.

[86] 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.