Journal of Machine Learning Research 25 (2024) 1-49 Submitted 7/22; Revised 12/23; Published 2/24

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

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

Mateo Díaz MATEODD@JHU.EDU
Department of Applied Mathematics and Statistics
Johns Hopkins University
Baltimore, MD 21218, USA

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

Editor: Francis Bach

Abstract

We analyze a stochastic approximation algorithm for decision-dependent 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 average iterate of the algorithm and the solution is asymptotically normal, with a covariance that clearly decouples the effects of the gradient noise and the distributional shift. Moreover, building on the work of Hájek and Le Cam, we show that the asymptotic performance of the algorithm with averaging is locally minimax optimal.

Keywords: stochastic approximation, decision-dependent distributions, performative prediction, asymptotic normality, local asymptotic minimax optimality

1. Introduction

The primary role of stochastic optimization in data science is to find a learning rule (e.g., a classifier) from a limited data sample which 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. Recent literature on strategic classification (Hardt et al., 2016) and performative prediction (Perdomo et al., 2020), however, has highlighted a variety of contemporary settings where this assumption is grossly violated. One common reason is that the data seen by a learning system may depend on or react to a deployed learning rule. 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—a phenomenon called gaming. 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 classifier that generalizes well under the response distribution. The situation may be further compounded by a population that reacts to multiple competing learners simultaneously (Narang et al., 2023; Wood and Dall'Anese, 2023; Piliouras and Yu, 2022).

In this work, 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

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

where $\mathcal{P}$ is some fixed distribution and $N_{\mathcal{X}}(x) = \{v \in \mathbf{R}^d \mid \langle v, y - x\rangle \leq 0 \text{ for all } y \in \mathcal{X}\}$ 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 both of these examples, VI($\mathcal{P}$) encodes the standard first-order optimality conditions. The benefit of variational inequalities is that they yield a single framework for analyzing a wide range of learning problems, notably in optimization and game theory. We refer the interested reader to Kinderlehrer and Stampacchia (2000) and Dontchev and Rockafellar (2009) for a historical perspective and further details on the use of variational inequalities in applications.

Following the recent literature on performative prediction (Hardt et al., 2016; Perdomo et al., 2020; Narang et al., 2023), 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 VI($\mathcal{D}(x^\star)$) induced by the response distribution $\mathcal{D}(x^\star)$. The setting of performative prediction (Perdomo et al., 2020) corresponds to the choice $G(x, z) = \nabla_x \ell(x, z)$ for some loss function ℓ.[1] More generally, decision-dependent games, proposed by Narang et al. (2023), Piliouras and Yu (2022), and Wood and Dall'Anese (2023), correspond to the choice $G(x, z) = \big(\nabla_1 \ell_1(x, z), \ldots, \nabla_k \ell_k(x, z)\big)$ 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

1. In the language of Perdomo et al. (2020), equilibria coincide with performatively stable points.

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 (Mendler-Dünner et al., 2020) and its multiplayer extension (Narang et al., 2023), this algorithm reduces to a basic projected stochastic gradient iteration. The contribution of our paper 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 for both single player and multiplayer performative prediction.

1.1 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 of Polyak and Juditsky (1992), 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 for an illustration.[2]

Theorem 1 (Asymptotic normality, informal; see Theorem 7) *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 continuous 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 $\nu \in \big(\frac{1}{2}, 1\big)$. 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*

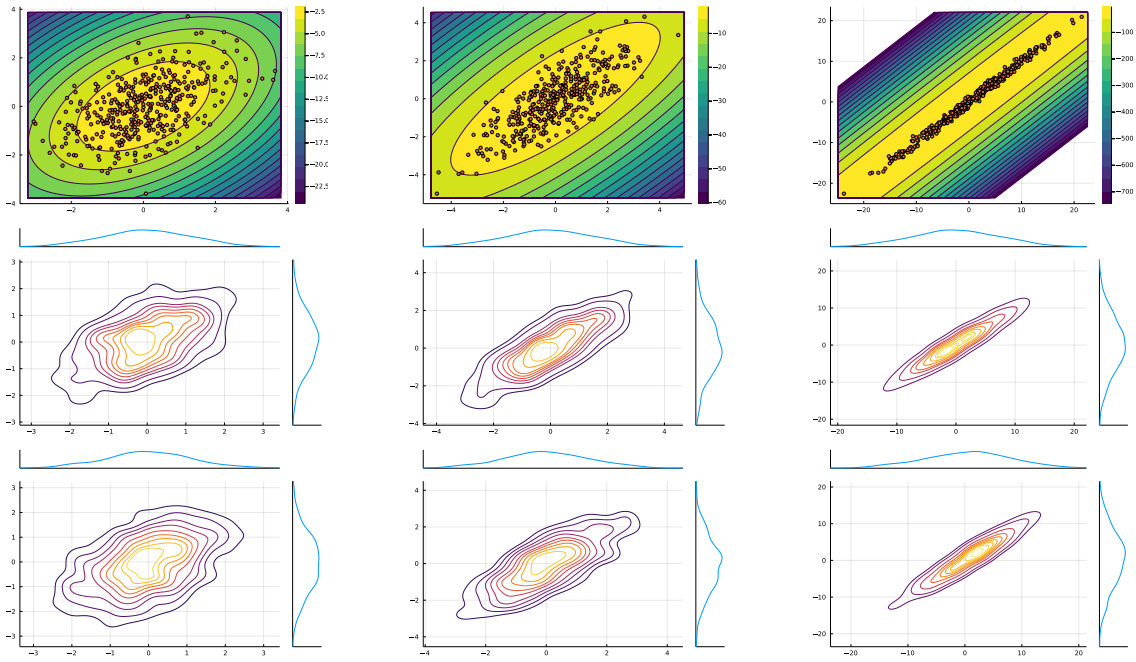
$$\sqrt{t}(\bar{x}_t - x^\star) \rightsquigarrow \mathsf{N}\big(0, W^{-1}\Sigma W^{-\top}\big),$$

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 conditions on the data of the problem reduce to the standard assumptions in performative prediction (Perdomo et al., 2020) when $G(x, z) = \nabla_x \ell(x, z)$. In particular, $G(\cdot, z)$ being α-strongly monotone and Lipschitz is then equivalent to the function $\ell(\cdot, z)$ being α-strongly convex and smooth with Lipschitz continuous gradient. The strong monotonicity requirement can be loosened to hold only in expectation; see Theorem 7 for the formal statement. Second, the regime $\frac{\gamma\beta}{\alpha} < 1$ is, in essence, optimal because otherwise equilibrium points may even fail to exist. Third, the effect of the distributional shift on the asymptotic covariance is entirely captured by the second "dynamic" term in W. Indeed,

2. Visit `https://github.com/mateodd25/Asymptotic-normality-in-performative-prediction` for code.



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

Figure 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 algorithm 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 $\sqrt{k}(\bar{x}_k - x^\star)$ (bottom).

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.[3] The proof of Theorem 1 follows by interpreting SFB as a stochastic 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 of Polyak and Juditsky (1992, 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; 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. The end result is summarized informally as follows.

Theorem 2 (Asymptotic optimality, informal; see Theorem 16) *Suppose the same setting as in Theorem 1 and let $\mathcal{L}\colon \mathbf{R}^d \to [0, \infty)$ be any symmetric, quasiconvex, lower*

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

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 $x_k^\star$ converging to $x^\star$, such that the following hold.

*(i) (**Lower bound**) The expected error $\mathbb{E}[\mathcal{L}(\sqrt{k}(\widehat{x}_k - x_k^\star))]$ of the estimator $\widehat{x}_k$ on the perturbed problem is asymptotically lower-bounded by $\mathbb{E}[\mathcal{L}(Z)]$, where $Z \sim \mathsf{N}(0, W^{-1}\Sigma W^{-\top})$.*

*(ii) (**Tightness of SFB**) Moreover, if $\mathcal{L}$ is bounded and continuous, then the lower bound in (i) 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 the theorem and its proof follow closely the classical work of Hájek and Le Cam (Le Cam and Yang, 2000; van der Vaart, 1998) on statistical lower bounds and the more recent work of Duchi and Ruan (2021) on asymptotic optimality of the stochastic gradient method. In particular, the fundamental role of tilt-stability and the inverse function theorem highlighted by Duchi and Ruan (2021) is replaced by the implicit function theorem paradigm.

Taken together, Theorems 1 and 2 provide a solid theoretical footing for the practical application of SFB, which generalizes stochastic gradient descent. These results precisely quantify the asymptotic uncertainty of SFB, with confidence regions that are optimally narrow (in an appropriate sense) among all methods for finding equilibrium points. In particular, algorithms that use momentum or try to learn and adapt to how the distributions vary cannot achieve better asymptotic performance. Thus, stronger modeling assumptions are necessary to develop algorithms with provably superior asymptotic sample efficiency. We also note that all results in the paper extend directly to a minibatch variant of SFB, where in each iteration the update direction $G(x_t, z_t)$ is replaced by the empirical average $\frac{1}{m}\sum_{i=1}^m G(x_t, z_{t,i})$ with $(z_{t,1}, \ldots, z_{t,m})$ sampled i.i.d. from $\mathcal{D}(x_t)$. The only effect of the batching is that the asymptotic covariance Σ is rescaled by $1/m$ in all results.

Before continuing, it is important to highlight a limitation of our results. In order to generate a sample $z_t \sim \mathcal{D}(x_t)$ in practice, one must first deploy the learning rule x_t and then wait for the population to adapt. Consequently, the sampling and deployment have different associated "costs." Our results can somewhat adapt to this imbalance by using minibatches, as explained above. Nonetheless, a more nuanced approach that balances sample complexity against the deployment cost is worth investigating in future work.

1.2 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 of Perdomo et al. (2020) and its multiplayer extension developed independently by Narang et al. (2023), Piliouras and Yu (2022), and Wood and Dall'Anese (2023). The stochastic gradient method for performative prediction was first introduced and analyzed by Mendler-Dünner et al. (2020), while the stochastic forward-backward method for games was analyzed by Narang et al. (2023). The related work of Drusvyatskiy and Xiao (2022) 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 Perdomo et al. (2020). Seeking to develop algorithms for finding performatively optimal points, the work of Miller et al. (2021) provides sufficient conditions for the prediction problem to be convex; extensions of such conditions to games appear in the papers of Narang et al. (2023) and Wood and Dall'Anese (2023). Algorithms for finding performatively optimal points under a variety of different assumptions and oracle models appear in the works of Izzo et al. (2021), Jagadeesan et al. (2022), Miller et al. (2021), Narang et al. (2023), and Wood and Dall'Anese (2023). The performative prediction framework is largely motivated by the problem of strategic classification (Hardt et al., 2016), which has been studied extensively from the perspective of causal inference (Bechavod et al., 2020; Miller et al., 2020) and convex optimization (Dong et al., 2018). Other lines of work (Brown et al., 2022; Cutler et al., 2023; Ray et al., 2022; Wood et al., 2021) 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 of Polyak and Juditsky (1992) 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; see, e.g., the works of Ahmed (2000), Dupačová (2006), Jonsbråten et al. (1998), Rubinstein and Shapiro (1993), and Varaiya and Wets (1988). We refer the reader to the recent paper of Hellemo et al. (2018), 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 (Perdomo et al., 2020; Narang et al., 2023; Wood and Dall'Anese, 2023; Piliouras and Yu, 2022; Drusvyatskiy and Xiao, 2022; Mendler-Dünner et al., 2020) and our current paper 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 Wainwright (2019, 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 (Le Cam and Yang, 2000; van der Vaart, 1998), provide much finer problem-specific guarantees. Building on this framework, Duchi and Ruan (2021) showed that the stochastic gradient method for standard single-stage stochastic optimization problems is, in an appropriate sense, locally asymptotically minimax optimal. Our paper builds heavily on this line of work.

1.3 Outline

The outline of the paper is as follows. Section 2 records some basic notation that we will use. Section 3 formally introduces/reviews the decision-dependent framework. In Section 4, we show that the running average of the stochastic forward-backward algorithm is asymptotically

normal (Theorem 1), and identify its asymptotic covariance. Finally, Section 5 presents the local minimax lower bound (Theorem 2). We defer many of the technical proofs to the appendices.

2. Notation and Definitions

Throughout, we let $\mathbf{R}^d$ 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 $\mathcal{X} \subset \mathbf{R}^d$, the symbol $\operatorname{proj}_{\mathcal{X}}(x)$ will denote the set $\arg\min_{y\in\mathcal{X}} \|y - x\|$ of nearest points of $\mathcal{X}$ to $x \in \mathbf{R}^d$. We say that a function $\mathcal{L}\colon \mathbf{R}^d \to \mathbf{R}$ is *symmetric* if it satisfies $\mathcal{L}(x) = \mathcal{L}(-x)$ for all $x \in \mathbf{R}^d$, and we say that $\mathcal{L}$ is *quasiconvex* if its sublevel set $\{x \mid \mathcal{L}(x) \le c\}$ is convex for any $c \in \mathbf{R}$. For any matrix $A \in \mathbf{R}^{m\times n}$, 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 $A, B \in \mathbf{R}^{n\times n}$, we write $A \succeq B$ if the matrix $A - B$ is positive semidefinite.

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

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

If $F = \nabla f$ for some C^1-smooth function f, then α-strong monotonicity of F is equivalent to α-strong convexity of f. We say that a map $F\colon \mathcal{X} \to \mathbf{R}^m$ is *smooth* on a set $\mathcal{X} \subset \mathbf{R}^d$ if F has a differentiable extension on an open neighborhood of each point of $\mathcal{X}$; further, we say that F is *β-smooth* on $\mathcal{X}$ if the Jacobian of F satisfies the Lipschitz condition

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

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:

$$W_1(\mu, \nu) = \sup_{\phi\in \mathrm{Lip}_1(\mathcal{Z})} \left\{ \mathop{\mathbb{E}}_{X\sim\mu}\big[\phi(X)\big] - \mathop{\mathbb{E}}_{Y\sim\nu}\big[\phi(Y)\big] \right\}. \tag{1}$$

Here, $\mathrm{Lip}_1(\mathcal{Z})$ denotes the set of 1-Lipschitz functions $\mathcal{Z} \to \mathbf{R}$. 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 $f\colon (0, \infty) \to \mathbf{R}$ with $f(1) = 0$, the f-divergence of μ from ν is given by

$$\Delta_f(\mu \,\|\, \nu) = \int_{\mathcal{Z}} f\left(\frac{d\mu}{d\nu}\right) d\nu, \tag{2}$$

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 $X_k \colon \Omega_k \to \mathbf{R}^m$ defined on probability spaces $(\Omega_k, \mathcal{S}_k, P_k)$ and a random vector $X \sim \mu$ in $\mathbf{R}^m$, 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 $\varphi \colon \mathbf{R}^m \to \mathbf{R}$). 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 $\|a_k\| \le C\|b_k\|$ for all but finitely many k; we write $a_k = o(b_k)$ if for every $\varepsilon > 0$, the inequality $\|a_k\| \le \varepsilon\|b_k\|$ holds for all but finitely many k; we write $a_k = \Theta(b_k)$ if there exist constants $c, C > 0$ such that $c\|b_k\| \le \|a_k\| \le C\|b_k\|$ 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. 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 $\mathcal{X} \subset \mathbf{R}^d$, 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 Perdomo et al. (2020), we will assume that $\mathcal{D} \colon \mathcal{X} \to P_1(\mathcal{Z})$ is Lipschitz continuous.

Assumption 1 (Lipschitz distribution map) There is a constant $\gamma > 0$ satisfying

$$W_1\big(\mathcal{D}(x), \mathcal{D}(x')\big) \le \gamma\|x - x'\| \qquad \text{for all } x, x' \in \mathcal{X}.^4$$

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

Assumption 2 (Loss regularity) There are constants $\beta, \bar{L} \ge 0$ 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

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

4. Assumption 1 implies in particular that $\{\mathcal{D}_x\}_{x\in\mathcal{X}}$ is a *Markov kernel* from $\mathcal{X}$ to $\mathcal{Z}$ (see Lemma 31); this is crucial to have a well-defined probability space on which to analyze decision-dependent stochastic approximation problems.

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 $G_x(\cdot)$ is $\bar{L}$-Lipschitz continuous. Condition (ii) is a standard monotonicity requirement; when $G(x, z) = \nabla_x \ell(x, z)$, this corresponds to α-strong convexity of the expected loss. 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; see, e.g., Perdomo et al. (2020), Piliouras and Yu (2022), Narang et al. (2023), and Wood and Dall'Anese (2023).

Assumptions 1 and 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 of Narang et al. (2023); a short argument appears in Section A.1.

Lemma 3 (Deviation) *Suppose that Assumptions 1 and 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 the work of Perdomo et al. (2020) for performative prediction and Narang et al. (2023) for its multiplayer extension, is the key solution concept that we will use.

Definition 4 (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

$$\texttt{Sol}(x) := \{y \mid 0 \in G_x(y) + N_{\mathcal{X}}(y)\}, \tag{3}$$

which is single-valued on $\mathcal{X}$ by the continuity and strong monotonicity of $G_x(\cdot)$ (e.g., see Rockafellar and Wets, 1998, 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; see, e.g., Drusvyatskiy and Xiao (2022), Mendler-Dünner et al. (2020), Narang et al. (2023), Perdomo et al. (2020), Piliouras and Yu (2022), and Wood and Dall'Anese

(2023). Most importantly, equilibrium points exist and are unique under Assumptions 1 and 2. The proof is identical to that of Theorem 7 of Narang et al. (2023); we provide a short argument in Section A.2 for completeness.

Theorem 5 (Existence) *Suppose that Assumptions 1 and 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., Perdomo et al. (2020, Proposition 3.6). Therefore, the regime $\gamma\beta < \alpha$ is the natural setting to consider when searching for equilibrium points.

4. 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). 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 1 below. In the setting of performative prediction (Mendler-Dünner et al., 2020) and its multiplayer extension (Narang et al., 2023), the algorithm reduces to projected stochastic gradient methods.

Algorithm 1 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 $t \geq 0$:

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

For the remainder of Section 4, we let $(x_t)_{t\geq 0}$ denote the stochastic process generated by Algorithm 1 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

$$\mathbb{P}(E_0 \times \cdots \times E_t \times \mathcal{Z}^{\mathbb{N}}) = \int_{E_0} \cdots \int_{E_t} d\mathcal{D}_{x_t}(z_t) \cdots d\mathcal{D}_{x_0}(z_0) \tag{4}$$

for all $E_0, \ldots, E_t \in \mathcal{B}(\mathcal{Z})$ and $t \geq 0$ (see Theorem 32). 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

$$\xi_x(z) := G(x,z) - G_x(x) \tag{5}$$

and impose the following standard bound on the conditional second moment of the noise. In words, this assumption stipulates that the variance of the noise $\xi_{x_t}(z)$ with respect to the distribution $\mathcal{D}_{x_t}$ induced by the iterate x_t grows at most quadratically with the distance of x_t to $x^\star$.

Assumption 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 1–3 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 (Narang et al., 2023, Theorem 24) and an application of the Robbins-Siegmund almost supermartingale convergence theorem (Robbins and Siegmund, 1971), appears in Section A.3.

Proposition 6 (Almost sure convergence) *Suppose that Assumptions 1–3 hold and the step size sequence in Algorithm 1 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 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 (Polyak and Juditsky, 1992). 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 locally $C^{1,1}$-smooth[5] and sufficient integrability conditions hold to invoke dominated convergence.

Asymptotic uniform integrability conditions such as (iii) are key for obtaining convergence of moments (see van der Vaart, 1998, 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 35). 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 34); 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.

We are now ready to present our main result.

Theorem 7 (Asymptotic normality) *Suppose that Assumptions 1–4 hold and the step size sequence in Algorithm 1 satisfies $\eta_t \propto t^{-\nu}$ for some $\nu \in \left(\frac{1}{2}, 1\right)$. Let $R\colon \mathcal{X} \to \mathbf{R}^d$ 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 (5). Then, as $t \to \infty$, the iterates x_t and their running averages $\bar{x}_t = \frac{1}{t}\sum_{i=1}^t x_i$ 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 7 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:

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

5. Recall that a map is said to be locally $C^{k,1}$-smooth if it is C^k-smooth and its k^{th}-order partial derivatives are locally Lipschitz continuous.

Example 1 (Performative prediction with location-scale families) As an explicit example of Theorem 7, 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 by Miller et al. (2021). In this case, we have

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

for some fixed matrix $A \in \mathbf{R}^{n \times d}$, where n is the dimension of the data z. 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 $\nabla^2_{zx} \ell(x^\star, z) \in \mathbf{R}^{d \times n}$ 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.

Example 2 (Multiplayer performative prediction with location-scale families) More generally, let us look at the problem of multiplayer performative prediction (Narang et al., 2023). In this case, the map G takes the form

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

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 by Narang et al. (2023), 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 $[A_i, A_{-i}]$.

4.1 Proof of Theorem 7

The proof of Theorem 7 is based on the stochastic approximation result of Polyak and Juditsky (Polyak and Juditsky, 1992, Theorem 2), which we review in Appendix B. For the remainder of this section, we impose the assumptions of Theorem 7.

Consider the map $R\colon \mathcal{X} \to \mathbf{R}^d$ given by $R(x) = G_x(x)$. In light of the interiority condition $x^\star \in \operatorname{int} \mathcal{X}$ of Assumption 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 1 as

$$x_{t+1} = x_t - \eta_t\big(R(x_t) + \xi_t + \zeta_t\big), \tag{6}$$

where

$$\zeta_t := \frac{x_t - \eta_t\big(R(x_t) + \xi_t\big) - \operatorname{proj}_{\mathcal{X}}\big(x_t - \eta_t\big(R(x_t) + \xi_t\big)\big)}{\eta_t}. \tag{7}$$

Our goal is to apply Theorem 26 to the process (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

$$\mathcal{F}_0 := \{\emptyset, \mathcal{Z}^{\mathbb{N}}\} \quad \text{and} \quad \mathcal{F}_t := \{A \times \mathcal{Z}^{\mathbb{N}} \mid A \in \mathcal{B}(\mathcal{Z}^t)\} \qquad \text{for all } t \geq 1 \tag{8}$$

and $\mathbb{P} = \bigotimes_{i=0}^{\infty} \mathcal{D}_{x_i}$ is given by (4). In what follows, we establish the necessary assumptions for Theorem 26.

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

Lemma 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

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

as a consequence of Lemma 3 and the $\bar{L}$-Lipschitz continuity of $G_y(\cdot)$. Similarly,

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

as a consequence of the α-strong monotonicity of $G_y(\cdot)$ and Lemma 3. ■

To establish Assumption 6, observe first that $\sup_{t\geq 0} \mathbb{E}\|\xi_t\|^2 < \infty$ by Assumption 3 and Proposition 6. 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

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

almost surely as $t \to \infty$.

Lemma 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 6), the uniform integrability condition (iii) of Assumption 4, and the Lipschitz condition (i) of Assumption 2, we may apply Lemma 35 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 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 6 are now established, and Lindeberg's condition (iii) of Assumption 6 holds by item (iv) of Assumption 4. Now consider the residual vector ζ_t given by (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 6 holds, and the verification of Assumption 6 is complete.

We turn now to Assumption 7. The first two conditions of Assumption 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 8), it follows that we have $\langle \nabla R(x^\star)v, v\rangle \geq \alpha - \gamma\beta$ for every unit vector $v \in \mathbf{S}^{d-1}$ 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 10 (Positivity of the Jacobian) *For any point $x \in \operatorname{int} \mathcal{X}$ at which R is differentiable, we have*

$$\langle \nabla R(x)v, v\rangle \geq \alpha - \gamma\beta \qquad \textit{for all } v \in \mathbf{S}^{d-1} \tag{9}$$

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 8, R is $(\alpha - \gamma\beta)$-strongly monotone on $\mathcal{X}$, so (9) follows immediately from the definitions of differentiability and strong monotonicity: for any unit vector $v \in \mathbf{S}^{d-1}$,

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

Next, observe that (9) implies $\lambda_{\min}(\nabla R(x) + \nabla R(x)^\top) \geq 2(\alpha - \gamma\beta)$. Now let $w \in \mathbf{C}^d$ be a normalized eigenvector of $\nabla R(x)$ with associated eigenvalue $\lambda \in \mathbf{C}$. 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 4. ■

Next, recall $\eta_t \propto t^{-\nu}$ for some $\nu \in \big(\frac{1}{2}, 1\big)$, 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 $\lim_{t\to\infty}\big((t+1)^r - t^r\big) = 0$ 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 7.

Finally, by Proposition 6, 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 (see Durrett, 2019, 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

$$\frac{1}{\sqrt{t}}\sum_{i=0}^{t-1}\|R(x_i) - \nabla R(x^\star)(x_i - x^\star)\| \xrightarrow{\text{a.s}} 0 \qquad \text{as } t \to \infty. \tag{10}$$

Since $\nabla R(x^\star)$ is positively stable (Lemma 10), this concludes the verification of Assumption 7. An application of Theorem 26 to the process (6) completes the proof of Theorem 7.

5. Asymptotic Optimality

In this section, we establish the local asymptotic optimality of Algorithm 1. Our result builds on classical ideas from Hájek and Le Cam (Le Cam and Yang, 2000; van der Vaart, 1998) on lower bounds for statistical estimation and the more recent work of Duchi and Ruan (2021) 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 $G\colon \mathcal{X}\times\mathcal{Z} \to \mathbf{R}^d$ satisfying Assumptions 1 and 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 11 (Admissible distribution map) A distribution map $\mathcal{D}'\colon \mathcal{X} \to P_1(\mathcal{Z})$ is *admissible* if Assumptions 1 and 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 is "optimal" among all algorithms for finding equilibrium points. To capture the notion of optimal covariance, a standard approach is to probe the normalized error with nonnegative "loss" functions $\mathcal{L}\colon \mathbf{R}^d \to [0,\infty)$ 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.[6]

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

6. 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 $u \in \mathbf{R}^d$, where $\mathcal{L}_u\colon \mathbf{R}^d \to [0,\infty)$ is given by $\mathcal{L}_u(x) = (u^\top x)^2 = u^\top(xx^\top)u$.

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 $u \in \mathbf{R}^d$ 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 $\widehat{x}_k \colon \mathcal{Z}^k \to \mathbf{R}^d$ is an arbitrary sequence of estimators (i.e., $\widehat{x}_k$ is a measurable function of k observed samples) and $\mathcal{L} \colon \mathbf{R}^d \to [0, \infty)$ is symmetric, quasiconvex, and lower semicontinuous, then the following lower bound holds:

$$\underbrace{\sup_{\mathcal{I} \subset \mathbf{R}^d,\, |\mathcal{I}| < \infty} \liminf_{k \to \infty} \max_{u \in \mathcal{I}} \mathbb{E}_{P_{k,u/\sqrt{k}}}\big[\mathcal{L}\big(\sqrt{k}\big(\widehat{x}_k - x^\star_{u/\sqrt{k}}\big)\big)\big]}_{\text{local asymptotic minimax risk}} \geq \mathbb{E}[\mathcal{L}(Z)], \tag{11}$$

where $P_{k,v} = \bigotimes_{i=0}^{k-1} \mathcal{D}^v_{\tilde{x}_i}$ 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.

The lower bound (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 (11) upon specializing to the dynamic estimation procedure corresponding to Algorithm 1 with step sizes $\eta_k \propto k^{-\nu}$ (as in Theorem 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 $P_{k,v} = \bigotimes_{i=0}^{k-1} \mathcal{D}^v_{\tilde{x}_i}$ appearing in (11).

Definition 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

$$\tilde{x}_k = \mathcal{A}_k(z_0, \ldots, z_{k-1}, \tilde{x}_0, \ldots, \tilde{x}_{k-1}) \tag{12}$$

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 $z_0, z_1, \ldots$ 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 (12) with $(z_0, \ldots, z_{k-1}) \sim \bigotimes_{i=0}^{k-1} \mathcal{D}_{\tilde{x}_i}$ 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 1, i.e.,

$$\mathcal{A}_{k+1}(z_0, \ldots, z_k, x_0, \ldots, x_k) = \operatorname{proj}_{\mathcal{X}}\big(x_k - \eta_k G(x_k, z_k)\big) \qquad \text{for all } k \geq 0, \tag{13}$$

is a dynamic estimation procedure under the assumptions of Proposition 6; 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.

5.1 Tilted Distributions

Following Duchi and Ruan (2021) and van der Vaart (1998, Section 25.3), for each distribution $\mathcal{D}_x := \mathcal{D}(x)$ we will construct "tilt perturbations" $\mathcal{D}_x^u$ parameterized by $u \in \mathbf{R}^d$. Henceforth, we fix an arbitrary nondecreasing C^3-smooth function $h\colon \mathbf{R} \to [-1,1]$ 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 $u \in \mathbf{R}^d$, the tilted distribution $\mathcal{D}_x^u \in P_1(\mathcal{Z})$ is defined by setting

$$\mathcal{D}_x^u(E) := \int_E \frac{1 + h(u^\top g_x(z))}{C_x^u}\, d\mathcal{D}_x(z) \qquad \text{for all } E \in \mathcal{B}(\mathcal{Z}), \tag{14}$$

where $g_x\colon \mathcal{Z} \to \mathbf{R}^d$ 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 $\{\mathcal{D}_x^u \mid u \in \mathbf{R}^d\}$ has score function g_x at zero, i.e.,

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

Thus, the collection of functions $\{u^\top g_x\colon \mathcal{Z} \to \mathbf{R} \,|\, u \in \mathbf{R}^d\}$ forms a "tangent space" of the model $\{\mathcal{D}_x^u \,|\, u \in \mathbf{R}^d\}$ at zero (see van der Vaart, 1998, Example 25.15). In the context of establishing the asymptotic optimality of Algorithm 1, 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 $g\colon \mathcal{X} \times \mathcal{Z} \to \mathbf{R}^d$ given by $g(x,z) = g_x(z)$. Despite being technical, these conditions (given in Assumption 5 and Definition 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 13 (Test functions) Given a compact metric space $\mathcal{K}$, we let $\mathcal{T}(\mathcal{K}, \mathcal{Z})$ consist of all bounded measurable functions $\phi\colon \mathcal{K} \times \mathcal{Z} \to \mathbf{R}$ 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 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 $\mathcal{X} \times \mathcal{K}$.[7]

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

(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 5 holds with $\vartheta(s) = \max\{\Lambda_p, L_p\} \cdot s$.

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

Definition 14 (Score functions) Let $\mathcal{G}$ consist of all measurable functions $g\colon \mathcal{X}\times\mathcal{Z} \to \mathbf{R}^d$ 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 on $\mathcal{Z}$.

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

(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 on $\mathcal{Z}$.

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

$$\xi(x,z) := G(x,z) - G_x(x), \tag{15}$$

which belongs to $\mathcal{G}$ as a consequence of Assumptions 2 and 5 and Lemma 30.

Henceforth, we fix $g \in \mathcal{G}$ and take $g_x(z) = g(x,z)$ in (14), thereby defining the tilted distribution map $\mathcal{D}^u\colon \mathcal{X} \to P_1(\mathcal{Z})$ given by $x \mapsto \mathcal{D}^u_x$. The following lemma guarantees that if Assumptions 1, 2, and 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 C.1, provides constants γ^u, $\bar{L}^u$, α^u that fulfill Assumptions 1 and 2 for $\mathcal{D}^u$ and deviate from γ, $\bar{L}$, α by $O(\|u\|)$ as $u \to 0$.

Lemma 15 (Tilted distributions are admissible) *Suppose that Assumptions 1, 2, and 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.*

Thus, the solutions $x^\star_{\mathcal{D}^u}$ are well defined for small enough u. For ease of notation, we set

$$x^\star_u := x^\star_{\mathcal{D}^u}$$

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

Theorem 16 (Asymptotic optimality) *Suppose that Assumptions 1, 2, and 5 hold with the equilibrium point $x^\star$ lying in the interior of $\mathcal{X}$, and suppose $g = \xi$.*[8] *Let $\mathcal{A}_k\colon \mathcal{Z}^k \times \mathcal{X}^k \to \mathcal{X}$*

8. Recall that g is the score function used to parameterize the perturbed distributions (14) and ξ is the noise (15).

be a dynamic estimation procedure, fix an initial point $\tilde{x}_0 \in \mathcal{X}$, and for each $u \in \mathbf{R}^d$, let $P_{k,u} = \bigotimes_{i=0}^{k-1} \mathcal{D}^u_{\tilde{x}_i}$ denote the distribution on $\mathcal{Z}^k$ induced by $\mathcal{D}^u$ along the sequence (12)*. Let $\widehat{x}_k \colon \mathcal{Z}^k \to \mathbf{R}^d$ be any sequence of estimators, and let $\mathcal{L} \colon \mathbf{R}^d \to [0,\infty)$ be symmetric, quasiconvex, and lower semicontinuous.*

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

$$\sup_{\mathcal{I} \subset \mathbf{R}^d,\, |\mathcal{I}| < \infty} \liminf_{k\to\infty} \max_{u \in \mathcal{I}} \mathbb{E}_{P_{k,u/\sqrt{k}}}\big[\mathcal{L}\big(\sqrt{k}\big(\widehat{x}_k - x^\star_{u/\sqrt{k}}\big)\big)\big] \geq \mathbb{E}[\mathcal{L}(Z)], \tag{16}$$

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

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

*(ii) (**Tightness of SFB**) If $\mathcal{A}_k$ is the dynamic estimation procedure* (13) *corresponding to Algorithm 1 with initial point $x_0 = \tilde{x}_0$ and step sizes $\eta_k \propto k^{-\nu}$ for some $\nu \in \big(\frac{1}{2}, 1\big)$, 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* (16) *whenever $\mathcal{L}$ is bounded and continuous.*

Most importantly, observe that the distribution of Z in Theorem 16 coincides with the asymptotic distribution of $\sqrt{t}(\bar{x}_t - x^\star)$ in Theorem 7, thereby justifying asymptotic optimality of the stochastic forward-backward method (Algorithm 1). The lower bound in Theorem 16 provides a decision-dependent analogue of the asymptotic optimality result of Duchi and Ruan (2021, Theorem 1) when the minimizer lies in the interior of $\mathcal{X}$. We believe that all the techniques developed here can be adapted to the more general setting where $x^\star$ may lie on the boundary of $\mathcal{X}$; since this generalization would require a significant technical overhead, we do not pursue it here. Taking the average of the SFB iterates to obtain an asymptotically optimal estimator as in item (ii) of Theorem 16 is important: even in the static case, the last iterate is known to be asymptotically suboptimal (Fabian, 1968).

Remark 17 (Convergence of equilibria and tilted distributions) In the setting of Theorem 16, one can show that the following approximations hold:

$$\|x^\star_u - x^\star\| = O(\|u\|) \qquad \text{as } u \to 0 \tag{17}$$

and

$$\sup_{x\in\mathcal{X}} W_1(\mathcal{D}^u_x, \mathcal{D}_x) = O(\|u\|) \qquad \text{as } u \to 0. \tag{18}$$

Indeed, we will prove in the forthcoming Lemma 23 that the map $u \mapsto x^\star_u$ is C^1-smooth on a neighborhood of zero, which implies (17) by the mean value theorem.

To verify the approximation (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

$$\mathop{\mathbb{E}}_{z\sim\mathcal{D}^u_x}[\phi(z)] - \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x}[\phi(z)] = \frac{1}{C^u_x}\mathop{\mathbb{E}}_{z\sim\mathcal{D}_x}\big[\bar{\phi}(z)\big(1 + h\big(u^\top g_x(z)\big)\big)\big] - \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x}[\bar{\phi}(z)] \tag{19}$$

for any $x \in \mathcal{X}$ and $u \in \mathbf{R}^d$. Further, Lemma 28 shows $\sup_{x\in\mathcal{X}} \mathbb{E}_{z\sim\mathcal{D}_x}|h(u^\top g_x(z))| = O(\|u\|)$ for all $u \in \mathbf{R}^d$ and $\sup_{x\in\mathcal{X}} \frac{1}{C^u_x} = 1 + O(\|u\|^3)$ as $u \to 0$, so (19) implies

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

which follows from definition (1). This establishes (18), which in particular asserts that the collection of tilted distribution maps $\{\mathcal{D}^u\}_{u\in\mathbf{R}^d}$ converges uniformly to $\mathcal{D}$ as $u \to 0$.

Remark 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 an f-divergence. Let $f\colon (0,\infty) \to \mathbf{R}$ be any convex function that is $C^{2,1}$-smooth around $t = 1$ and satisfies $f(1) = 0$. 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 $\Delta_f(\mathcal{D}'_x \parallel \mathcal{D}_x)$ denotes the usual f-divergence of $\mathcal{D}'_x$ from $\mathcal{D}_x$ given by (2).[9] The following approximation holds:

$$\Delta_f(\mathcal{D}^u \,\|\, \mathcal{D}) = O(\|u\|^2) \qquad \text{as } u \to 0. \tag{20}$$

To verify (20), observe that for all sufficiently small $u \in \mathbf{R}^d$ and all $x \in \mathcal{X}$, we have

$$\begin{aligned}
\Delta_f(\mathcal{D}^u_x \,\|\, \mathcal{D}_x) &= \int f\left(\frac{1 + h(u^\top g_x(z))}{C^u_x}\right) d\mathcal{D}_x(z) \\
&= \int f\left(\frac{1 + u^\top g_x(z)}{C^u_x}\right) d\mathcal{D}_x(z) &(21)\\
&= \frac{f''(1)}{2} u^\top \left(\mathop{\mathbb{E}}_{z\sim\mathcal{D}_x} g_x(z) g_x(z)^\top \right) u + r_x(u), &(22)
\end{aligned}$$

where $\sup_{x\in\mathcal{X}} |r_x(u)| = o(\|u\|^2)$ as $u \to 0$. The equality (21) holds for all sufficiently small $u \in \mathbf{R}^d$ and all $x \in \mathcal{X}$ because g is uniformly bounded over $\mathcal{X} \times \mathcal{Z}$ (see Lemma 27) and $h(t) = t$ for all t in a neighborhood of zero. The equality (22) follows from a second-order approximation and the dominated convergence theorem; we defer the details to Lemma 29. 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 $u \in \mathbf{R}^d$.

In light of (20), one may obtain from (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,

$$\lim_{c\to\infty} \liminf_{k\to\infty} \sup_{\mathcal{D}'\in B_f(c/k)} \mathbb{E}_{P'_k}\big[\mathcal{L}\big(\sqrt{k}(\widehat{x}_k - x^\star_{\mathcal{D}'})\big)\big] \geq \mathbb{E}[\mathcal{L}(Z)], \tag{23}$$

9. Examples of f-divergences include the χ^2-divergence, KL-divergence, and squared Hellinger distance.

where $P'_k = \bigotimes_{i=0}^{k-1} \mathcal{D}'_{\bar{x}_i}$ denotes the distribution on $\mathcal{Z}^k$ induced by $\mathcal{D}'$ along the sequence (12). Indeed, (20) facilitates the elementary estimation

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

and hence (16) implies (23).

5.2 Proof of Theorem 16

The proof of Theorem 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 $\{Q_{k,u} \,|\, u \in \mathbf{R}^d\}$ 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 $u \in \mathbf{R}^d$; following van der Vaart and Wellner (1996), 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 $X_k \colon \Omega_k \to \mathbf{R}^m$ 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 $\varphi \colon \mathbf{R}^m \to \mathbf{R}$.

Definition 19 (Locally asymptotically normal) The sequence $\{Q_{k,u} \mid u \in \mathbf{R}^d\}$ is *locally asymptotically normal (LAN) with precision V at zero* if there exist a sequence of random vectors $Z_k \colon \Omega_k \to \mathbf{R}^d$ and a positive semidefinite matrix $V \in \mathbf{R}^{d\times d}$ such that $Z_k \overset{0}{\rightsquigarrow} \mathsf{N}(0, V)$ and, for each $u \in \mathbf{R}^d$,

$$\log \frac{dQ_{k,u}}{dQ_{k,0}} = u^\top Z_k - \frac{1}{2} u^\top V u + o_{Q_{k,0}}(1). \tag{24}$$

Definition 20 (Regular mapping sequence) A sequence of mappings $\Gamma_k \colon \mathbf{R}^d \to \mathbf{R}^n$ is *regular with derivative $\dot{\Gamma}$ at zero* if there exists a matrix $\dot{\Gamma} \in \mathbf{R}^{n\times d}$ satisfying

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

Example 3 Given any $\psi \colon \mathbf{R}^d \to \mathbf{R}^n$ such that ψ is differentiable at zero, the induced mapping sequence $\Gamma_k \colon \mathbf{R}^d \to \mathbf{R}^n$ 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 16 by taking $\psi(u) = x^\star_u$ on a neighborhood of zero.

Equipped with the preceding definitions, we are ready to state the following version of the Hájek-Le Cam minimax theorem, which appears for example Lemma 8.2 of Duchi and Ruan (2021) and Theorem 3.11.5 of van der Vaart and Wellner (1996).

Theorem 21 (Local asymptotic minimax bound) *Let $\{Q_{k,u} \mid u \in \mathbf{R}^d\}$ be locally asymptotically normal with precision V at zero, $\Gamma_k \colon \mathbf{R}^d \to \mathbf{R}^n$ be a regular mapping sequence with derivative $\dot{\Gamma}$ at zero, and $\mathcal{L} \colon \mathbf{R}^n \to [0,\infty)$ be symmetric, quasiconvex, and lower*

semicontinuous. Then, for any sequence of estimators $T_k\colon \Omega_k \to \mathbf{R}^n$*, we have*

$$\sup_{\mathcal{I}\subset\mathbf{R}^d,\,|\mathcal{I}|<\infty} \liminf_{k\to\infty} \max_{u\in\mathcal{I}} \mathbb{E}_{Q_{k,u}}\big[\mathcal{L}\big(\sqrt{k}\big(T_k - \Gamma_k(u)\big)\big)\big] \geq \mathbb{E}[\mathcal{L}(Z)], \tag{25}$$

where $Z \sim \mathsf{N}(0, \dot{\Gamma}(V+\lambda I)^{-1}\dot{\Gamma}^\top)$ *for any* $\lambda > 0$*; if* V *is invertible, then* (25) *also holds with* $Z \sim \mathsf{N}(0, \dot{\Gamma}V^{-1}\dot{\Gamma}^\top)$*.*

To establish the lower bound (16) in Theorem 16, we will apply Theorem 21 as follows. Suppose henceforth that Assumptions 1, 2, and 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 $u \in \mathbf{R}^d$, we let

$$P_{k,u} := \bigotimes_{i=0}^{k-1} \mathcal{D}^u_{\tilde{x}_i} \tag{26}$$

denote the distribution on $\mathcal{Z}^k$ induced by $\mathcal{D}^u$ along the sequence (12), and we set

$$Q_{k,u} := P_{k,u/\sqrt{k}}. \tag{27}$$

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

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

and take $\Gamma_k\colon \mathbf{R}^d \to \mathbf{R}^d$ 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 $u \in \mathbf{R}^d$, 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 21; their proofs are deferred to Sections C.2 and C.3, respectively. The first lemma verifies that $\{Q_{k,u} \,|\, u \in \mathbf{R}^d\}$ is locally asymptotically normal at zero with precision

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

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

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

Lemma 22 (LAN) *Let* $Z_k\colon \mathcal{Z}^k \to \mathbf{R}^d$ *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* $u \in \mathbf{R}^d$*,*

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

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

Lemma 23 (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 (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 21. Let $\mathcal{L}\colon \mathbf{R}^d \to [0,\infty)$ be symmetric, quasiconvex, and lower semicontinuous, $\widehat{x}_k\colon \mathcal{Z}^k \to \mathbf{R}^d$ be any sequence of estimators, and suppose henceforth that $g = \xi$. Invoking Lemmas 22 and 23 and applying Theorem 21 yields

$$\begin{aligned}\sup_{\mathcal{I}\subset\mathbf{R}^d,\,|\mathcal{I}|<\infty} \liminf_{k\to\infty} \max_{u\in\mathcal{I}} \mathbb{E}_{P_{k,u/\sqrt{k}}}\big[\mathcal{L}\big(\sqrt{k}\big(\widehat{x}_k - x^\star_{u/\sqrt{k}}\big)\big)\big]&\\ = \sup_{\mathcal{I}\subset\mathbf{R}^d,\,|\mathcal{I}|<\infty} \liminf_{k\to\infty} \max_{u\in\mathcal{I}} \mathbb{E}_{Q_{k,u}}\big[\mathcal{L}\big(\sqrt{k}\big(\widehat{x}_k - \Gamma_k(u)\big)\big)\big] &\geq \mathbb{E}[\mathcal{L}(Z_\lambda)], \qquad (28)\end{aligned}$$

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 (28) establishes (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 $v \in \mathbf{R}^d$. 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 (28) entails (16).

To prove the final claim of Theorem 16, we proceed by establishing a type of asymptotic equivariance of the average SFB iterates (e.g., see van der Vaart, 1998, Lemma 8.14).

Lemma 24 (Asymptotic equivariance) *Let $\mathcal{A}_k$ be the dynamic estimation procedure* (13) *corresponding to Algorithm 1 with initial point $x_0 = \tilde{x}_0$ and step sizes $\eta_k \propto k^{-\nu}$ for*

some $\nu \in \left(\frac{1}{2}, 1\right)$. *Then the average iterates* $\bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$ *are asymptotically equivariant-in-law with respect to* $\{Q_{k,u} \,|\, u \in \mathbf{R}^d\}$ *for estimating* $x^\star$, *that is, for each* $u \in \mathbf{R}^d$,

$$\sqrt{k}\big(\bar{x}_k - \Gamma_k(u)\big) \overset{u}{\rightsquigarrow} \mathsf{N}\big(0, W^{-1}\Sigma W^{-\top}\big). \tag{29}$$

Proof Lemma 22 shows that the sequence of random vectors $Z_k \colon \mathcal{Z}^k \to \mathbf{R}^d$ given by

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

satisfies

$$Z_k \overset{0}{\rightsquigarrow} \mathsf{N}(0, \Sigma), \tag{30}$$

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

$$\log \frac{dQ_{k,u}}{dQ_{k,0}} = u^\top Z_k - \frac{1}{2} u^\top \Sigma u + o_{Q_{k,0}}(1). \tag{31}$$

Moreover, Theorem 7 reveals

$$\sqrt{k}(\bar{x}_k - x^\star) = -W^{-1} Z_k + o_{Q_{k,0}}(1). \tag{32}$$

Now let $\bar{Z} \sim \mathsf{N}(0, \Sigma)$, fix $u \in \mathbf{R}^d$, and consider the affine map $\varphi \colon \mathbf{R}^d \to \mathbf{R}^{d+1}$ 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 (30) implies $\varphi(Z_k) \overset{0}{\rightsquigarrow} \varphi(\bar{Z})$ and hence

$$\begin{pmatrix} \sqrt{k}(\bar{x}_k - x^\star) \\ \log \frac{dQ_{k,u}}{dQ_{k,0}} \end{pmatrix} \overset{0}{\rightsquigarrow} \begin{pmatrix} -W^{-1}\bar{Z} \\ u^\top \bar{Z} - \frac{1}{2} u^\top \Sigma u \end{pmatrix} \sim \mathsf{N}\left(\begin{pmatrix} 0 \\ -\frac{1}{2} u^\top \Sigma u \end{pmatrix}, \begin{pmatrix} W^{-1}\Sigma W^{-\top} & -W^{-1}\Sigma u \\ -u^\top \Sigma W^{-\top} & u^\top \Sigma u \end{pmatrix} \right) \tag{33}$$

by virtue of (31), (32), and the continuous mapping theorem (see van der Vaart, 1998, Theorems 2.3 and 2.7).

In light of (33), Le Cam's Third Lemma asserts

$$\sqrt{k}(\bar{x}_k - x^\star) \overset{u}{\rightsquigarrow} \mathsf{N}\big(-W^{-1}\Sigma u, W^{-1}\Sigma W^{-\top}\big) \tag{34}$$

(see van der Vaart, 1998, Example 6.7). On the other hand, Lemma 23 shows that Γ_k is a regular mapping sequence with derivative $\dot{\Gamma} = -W^{-1}\Sigma$ at zero, so

$$\sqrt{k}\big(x^\star - \Gamma_k(u)\big) = -\sqrt{k}\big(\Gamma_k(u) - \Gamma_k(0)\big) \to W^{-1}\Sigma u \qquad \text{as } k \to \infty. \tag{35}$$

Combining (34) and (35) yields (29). ∎

Finally, suppose that the assumptions of Lemma 24 hold. Let $\varphi \colon \mathbf{R}^d \to \mathbf{R}$ be any bounded continuous function and $Z \sim \mathsf{N}(0, W^{-1}\Sigma W^{-\top})$. Then (29) directly implies that for every finite subset $\mathcal{I} \subset \mathbf{R}^d$, 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

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

thereby demonstrating equality in (16) whenever $\mathcal{L}$ is bounded and continuous. The proof of Theorem 16 is complete.

Acknowledgments

Research of Drusvyatskiy was supported by the NSF DMS 1651851 and CCF 1740551 awards. We thank the anonymous reviewers for their insightful comments.

Appendix A. Proofs Deferred from Sections 3 and 4

A.1 Proof of Lemma 3

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

$$\begin{aligned}
\|G_x(y) - G_{x'}(y)\| &= \left\| \mathop{\mathbb{E}}_{z\sim\mathcal{D}(x)} G(y,z) - \mathop{\mathbb{E}}_{z\sim\mathcal{D}(x')} G(y,z) \right\| \\
&= \sup_{\|v\|\leq 1} \left\{ \mathop{\mathbb{E}}_{z\sim\mathcal{D}(x)} \langle G(y,z), v\rangle - \mathop{\mathbb{E}}_{z\sim\mathcal{D}(x')} \langle G(y,z), v\rangle \right\} \\
&\leq \beta \cdot W_1\big(\mathcal{D}(x), \mathcal{D}(x')\big) \qquad (36) \\
&\leq \beta\gamma \cdot \|x - x'\|,
\end{aligned}$$

where inequality (36) follows from the β-Lipschitz continuity of the function $z \mapsto \langle G(y,z), v\rangle$ and the characterization (1) of W_1.

A.2 Proof of Theorem 5

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

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

Strong monotonicity therefore ensures

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

where the last inequality follows from Lemma 3. Dividing through by $\alpha\|y - y'\|$ guarantees that $\texttt{Sol}(\cdot)$ is indeed a contraction on $\mathcal{X}$ with parameter $\frac{\gamma\beta}{\alpha}$. The result follows immediately from the Banach fixed point theorem.

A.3 Proof of Proposition 6

We will use the following classical result known as the Robbins-Siegmund almost supermartingale convergence theorem (for a proof, see Duflo, 1997, Theorem 1.3.12).

Lemma 25 (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 25 with $A_t = \|x_t - x^\star\|^2$, let $(\mathcal{F}_t)$ be the filtration given by (8) and observe that the SFB iterate sequence (x_t) is given by

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

where the map $R\colon \mathcal{X} \to \mathbf{R}^d$ 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 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). 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 of Narang et al. (2023, Theorem 24) with zero bias to obtain

$$\mathbb{E}\big[\|x_{t+1} - x^\star\|^2 \mid \mathcal{F}_t\big] \leq \frac{1 + 2K\eta_t^2}{1 + \bar{\alpha}\eta_t}\|x_t - x^\star\|^2 + \frac{2K\eta_t^2}{1 + \bar{\alpha}\eta_t} \tag{37}$$

$$\leq (1 - \tfrac{1}{2}\bar{\alpha}\eta_t)\|x_t - x^\star\|^2 + \frac{2K\eta_t^2}{1 + \bar{\alpha}\eta_t} \tag{38}$$

$$\leq \|x_t - x^\star\|^2 + 2K\eta_t^2 - \tfrac{1}{2}\bar{\alpha}\eta_t\|x_t - x^\star\|^2. \tag{39}$$

(For (37), it suffices to require $\eta_t \leq \frac{\bar{\alpha}}{2(\bar{L}+\gamma\beta)^2}$; for (38), it suffices to require $\eta_t \leq \frac{\bar{\alpha}}{4K+\bar{\alpha}^2}$.)

Using (39), we may now apply Lemma 25 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 25 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 (39) and the tower rule imply

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

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., Davis et al., 2021, Lemma 3.11.8) yields a constant $C > 0$ such that $\mathbb{E}\|x_t - x^\star\|^2 \leq Ct^{-\nu}$ for all $t \geq 1$. Therefore

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

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

Appendix B. Review of Asymptotic Normality

In this appendix, we present a variation of the asymptotic normality result of Polyak and Juditsky (1992, Theorem 2). Consider a measurable set $\mathcal{X} \subset \mathbf{R}^d$ and a measurable map $R\colon \mathcal{X} \to \mathbf{R}^d$. 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

$$x_{t+1} = x_t - \eta_t\big(R(x_t) + \xi_t + \zeta_t\big), \tag{40}$$

where η_t is a deterministic positive step size, ξ_t is a random vector in $\mathbf{R}^d$ representing noise with zero mean conditioned on prior information, and ζ_t is a random vector in $\mathbf{R}^d$ 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

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

in (40) yields the stochastic forward-backward method $x_{t+1} = \mathrm{proj}_{\mathcal{X}}(x_t - \eta_t(R(x_t) + \xi_t))$.

The following assumption formalizes the stochastic framework for our analysis.

Assumption 6 (Stochastic framework) The sequences $(x_t)_{t\geq 0}$, $(\xi_t)_{t\geq 0}$, and $(\zeta_t)_{t\geq 0}$ in (40) 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 (40) that we require to establish asymptotic normality of the average iterates. Recall that a matrix $A \in \mathbf{R}^{d\times d}$ is said to be *positively stable* if every eigenvalue of A has a positive real part.

Assumption 7 (Stable dynamics) There is a positively stable matrix $A \in \mathbf{R}^{d\times d}$ for which the following two conditions hold.

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

$$\eta_t \equiv \eta \qquad \text{and} \qquad 0 < \eta < 2\Big(\min_j \mathrm{Re}\,\lambda_j(A)\Big)^{-1} \tag{41}$$

or

$$\eta_t = o(1) \qquad \text{and} \qquad \frac{\eta_t - \eta_{t+1}}{\eta_t} = o(\eta_t) \qquad \text{as } t \to \infty. \tag{42}$$

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

$$\frac{1}{\sqrt{t}}\sum_{i=0}^{t-1}\|R(x_i)-A(x_i-x^\star)\| \xrightarrow{p} 0 \qquad \text{as } t\to\infty. \tag{43}$$

Theorem 26 (Polyak and Juditsky (1992, Theorem 2)) *Suppose that Assumptions 6 and 7 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 26 are somewhat more general than those of Theorem 2 of Polyak and Juditsky (1992), but the proof technique is the same. The primary differences are as follows:

(a) The residual term ζ_t in (40) 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 6. The rest of our stochastic setting stipulates conditions on ξ_t tailored to an application of the martingale central limit theorem (Theorem 34); we note that Lindeberg's condition (iii) of Assumption 6 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 of Polyak and Juditsky (1992) requires $A = \nabla R(x^\star)$ with

$$R(x) - \nabla R(x^\star)(x-x^\star) = O(\|x-x^\star\|^q) \qquad \text{as } x\to x^\star \tag{44}$$

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 (42); 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 (44) implies (43).

Proof For each $t\geq 0$, let $\Delta_t = x_t - x^\star$ denote the error of the process (40) 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 7 and observe that (40) yields the following recursion for all $t\geq 0$:

$$\begin{aligned}\Delta_{t+1} &= \Delta_t - \eta_t\big(R(x_t) + \xi_t + \zeta_t\big)\\ &= (I-\eta_t A)\Delta_t - \eta_t\big(R(x_t) - A\Delta_t + \xi_t + \zeta_t\big).\end{aligned} \tag{45}$$

Unrolling the recursion (45) 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

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

and hence

$$\begin{aligned} \sqrt{t}(\bar{x}_t - x^\star) + A^{-1}\left(\frac{1}{\sqrt{t}}\sum_{i=0}^{t-1}\xi_i\right) = \frac{1}{\sqrt{t}}B_t(x_0 - x^\star) - \frac{1}{\sqrt{t}}\sum_{i=0}^{t-1} A_i^t\xi_i \\ - \frac{1}{\sqrt{t}}\sum_{i=0}^{t-1} B_i^t\big(R(x_i) - A(x_i - x^\star)\big) - \frac{1}{\sqrt{t}}\sum_{i=0}^{t-1} B_i^t\zeta_i. \end{aligned} \tag{46}$$

We claim that the right-hand side of (46) is $o_{\mathbb{P}}(1)$ as $t \to \infty$. Indeed, since A is positively stable and the step size condition (i) of Assumption 7 holds, it follows from Lemma 1 of Polyak and Juditsky (1992) 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

$$\lim_{t\to\infty}\frac{1}{t}\sum_{i=0}^{t-1}\|A_i^t\|_{\mathrm{op}} = 0. \tag{47}$$

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 6, we have $C < \infty$. Therefore

$$\left\|\frac{1}{\sqrt{t}}B_t(x_0 - x^\star)\right\| \leq \frac{C\|x_0 - x^\star\|}{\sqrt{t}} \xrightarrow{\text{a.s.}} 0 \qquad \text{as } t \to \infty, \tag{48}$$

and since $(\xi_i)_{i\geq 0}$ is a martingale difference sequence, we deduce from (47) 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

$$\frac{1}{\sqrt{t}}\sum_{i=0}^{t-1} A_i^t \xi_i \xrightarrow{p} 0 \qquad \text{as } t\to\infty. \tag{49}$$

Moreover, the linear approximation condition (ii) of Assumption 7 implies

$$\left\|\frac{1}{\sqrt{t}}\sum_{i=0}^{t-1} B_i^t\big(R(x_i) - A(x_i - x^\star)\big)\right\| \leq \frac{C}{\sqrt{t}}\sum_{i=0}^{t-1}\|R(x_i) - A(x_i - x^\star)\| \xrightarrow{p} 0 \qquad \text{as } t\to\infty, \tag{50}$$

while the negligible residual condition (iv) of Assumption 6 implies

$$\left\|\frac{1}{\sqrt{t}}\sum_{i=0}^{t-1} B_i^t \zeta_i\right\| \leq \frac{C}{\sqrt{t}}\sum_{i=0}^{t-1}\|\zeta_i\| \xrightarrow{p} 0 \qquad \text{as } t\to\infty. \tag{51}$$

By (48)–(51), we conclude that the right-hand side of (46) 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 6, we may apply the martingale central limit theorem (Theorem 34) 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 (see van der Vaart, 1998, 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 C. Proofs Deferred from Section 5

This appendix presents contains all of the proofs deferred from Section 5. We assume throughout that the assumptions used in Section 5 are valid; in particular, $\mathcal{X}$ is compact, $\mathcal{Z}$ is bounded, and $g \in \mathcal{G}$ (see Definition 14). To begin, we present three preliminary lemmas.

Lemma 27 *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 28 *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 $u \in \mathbf{R}^d$. Then

$$|h(u^\top g_x(z))| \le L_h \|g_x(z)\|\|u\| \tag{52}$$

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

$$\sup_{x\in\mathcal{X}} \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x} |h(u^\top g_x(z))| \le L_h A_g \|u\| = O(\|u\|). \tag{53}$$

(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 $\mathbf{R}^d$ 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

$$\sup_{x\in\mathcal{X}} \left|\mathbb{E}_{z\sim\mathcal{D}_x} h(u^\top g_x(z))\right| \le \frac{L_{h''}B_g}{6}\|u\|^3 \tag{54}$$

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

$$\sup_{x\in\mathcal{X}} \frac{1}{C_x^u} = 1 + O(\|u\|^3) \qquad \text{as } u \to 0. \tag{55}$$

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 27. Since $h(0) = 0$ and h is L_h-Lipschitz continuous, the inequalites (52) and (53) 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 $u \in \mathbf{R}^d$. Thus, $u \mapsto C_x^u$ is C^2-smooth on $\mathbf{R}^d$, 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 $\mathbf{R}^d$.

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

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

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 (54). 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 (54) yields (55). ∎

Lemma 29 *Let $f\colon (0,\infty) \to \mathbf{R}$ be a function that is $C^{2,1}$-smooth around $t = 1$ and satisfies $f(1) = 0$. Then for all sufficiently small $u \in \mathbf{R}^d$ and all $x \in \mathcal{X}$, we have*

$$\int f\left(\frac{1 + u^\top g_x(z)}{C^u_x}\right) d\mathcal{D}_x(z) = \frac{f''(1)}{2} u^\top \left(\mathop{\mathbb{E}}_{z\sim\mathcal{D}_x} g_x(z) g_x(z)^\top \right) u + r_x(u), \tag{56}$$

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

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

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

and $(C^u_x)^4 \cdot \nabla^2_{uu}\varphi_x(u)$ equal to

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

Thus, taking a second-order Taylor expansion of φ_x at $u = 0$ with remainder r_x and applying the equalities $C^0_x = 1$, $\nabla_u C^u_x|_{u=0} = 0$, $\nabla^2_{uu} C^u_x|_{u=0} = 0$, and $f(1) = 0$ yields (56). It remains to verify $\sup_{x\in\mathcal{X}} |r_x(u)| = O(\|u\|^3)$ as $u \to 0$.

Lemmas 27 and 28 ensure that $C^u_x, \nabla_u C^u_x$, and $\nabla^2_{uu} C^u_x$ are Lipschitz continuous and bounded on a compact neighborhood of $u = 0$, with Lipschitz constants and bounds independent of x. Further, since f is $C^{2,1}$-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 $\nabla^2_{uu}\varphi_x$ 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. ∎

C.1 Proof of Lemma 15

The proof of this lemma is divided into four steps: the first step verifies Assumption 1 and the next three steps establish Assumption 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 23). Its proof consists of several applications of the dominated convergence theorem and is deferred to Section C.4.

Lemma 30 (Inferring smoothness) *Suppose that $T\colon \mathcal{X} \times \mathcal{Z} \to \mathbf{R}^n$ 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 $\mathcal{W} \subset \mathbf{R}^d$ of zero, the maps $\bar{H}\colon \mathcal{X}\times\mathcal{X}\times\mathcal{W} \to \mathbf{R}^n$ and $H\colon \mathcal{X}\times\mathcal{W}\to\mathbf{R}^n$ 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 1) First, we show that the perturbed distribution map $\mathcal{D}^u$ satisfies Assumption 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 $\mathbf{R}^d$ and apply Lemma 30 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 30 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 30 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 $\sup_{x\in\mathcal{X}} \|\nabla_x \rho_\phi(x,0)\| \le \gamma$ by Assumption 1. Thus, the triangle inequality yields

$$\|\nabla_x \rho_\phi(x,u)\| \le \|\nabla_x \rho_\phi(x,0)\| + \|\nabla_x \rho_\phi(x,u) - \nabla_x \rho_\phi(x,0)\| \le \gamma + L_1\|u\| = \gamma^u$$

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 1 with Lipschitz constant $\gamma^u = \gamma + O(\|u\|)$ as $u \to 0$.

Step 2 (Lipschitz continuity) Next, we establish Assumption 2(i) for the problem with the perturbed distribution map $\mathcal{D}^u$. Observe that the Lipschitz bounds in Assumption 2(i) remain unchanged, and that we only need to identify for all sufficiently small u a constant $\bar{L}^u \ge 0$ such that $\sup_{x\in\mathcal{X}} \mathbb{E}_{z\sim\mathcal{D}_x^u}[L(z)^2] \le (\bar{L}^u)^2$. 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 $\sup_{x\in\mathcal{X}} \mathbb{E}_{z\sim\mathcal{D}_x}[L(z)^2] \le \bar{L}^2$ furnished by Assumption 2(i). Indeed, for all $x \in \mathcal{X}$ and $u \in \mathbf{R}^d$, 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 28 yields

$$\sup_{x\in\mathcal{X}} \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x^u}\big[L(z)^2\big] \le \big(1 + O(\|u\|^3)\big)\big(1 + L_h M_g \|u\|\big)\bar{L}^2 = \bar{L}^2 + O(\|u\|) \qquad \text{as } u \to 0,$$

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 $u \in \mathbf{R}^d$, we have

$$\begin{aligned}\big\langle G_x^u(y) - G_x^u(y'), y - y'\big\rangle &= \big\langle G_x(y) - G_x(y'), y - y'\big\rangle \\ &\quad + \big\langle \big(G_x^u(y) - G_x(y)\big) - \big(G_x^u(y') - G_x(y')\big), y - y'\big\rangle \\ &\ge \alpha\|y - y'\|^2 - \big\|\big(G_x^u(y) - G_x(y)\big) - \big(G_x^u(y') - G_x(y')\big)\big\| \cdot \|y - y'\|\end{aligned}$$

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 30) and applying the dominated convergence theorem together with Lemma 28, we obtain

$$\begin{aligned}\ell^u &:= \sup_{x,y\in\mathcal{X}} \big\|\nabla_y\big(G_x^u(y) - G_x(y)\big)\big\|_{\mathrm{op}} \\ &= \sup_{x,y\in\mathcal{X}} \left\| \frac{1}{C_x^u} \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x}\big[\nabla_y G(y,z)\big(1 + h\big(u^\top g_x(z)\big)\big)\big] - \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x}\big[\nabla_y G(y,z)\big] \right\|_{\mathrm{op}} \\ &\le \underbrace{\sup_{x,y\in\mathcal{X}} \left\| \left(\frac{1}{C_x^u} - 1\right) \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x}\big[\nabla_y G(y,z)\big] \right\|_{\mathrm{op}}}_{O(\|u\|^3)} + \underbrace{\sup_{x,y\in\mathcal{X}} \left\| \frac{1}{C_x^u} \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x}\big[\nabla_y G(y,z) h\big(u^\top g_x(z)\big)\big] \right\|_{\mathrm{op}}}_{(1+O(\|u\|^3))\cdot O(\|u\|)} \\ &= O(\|u\|) \qquad \text{as } u \to 0.\end{aligned}$$

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 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.

C.2 Proof of Lemma 22

Fix $u \in \mathbf{R}^d$. For each $k \in \mathbb{N}$, it follows immediately from the definitions (14), (26), and (27) that for all $E_0, \ldots, E_{k-1} \in \mathcal{B}(\mathcal{Z})$, 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

$$\log \frac{dQ_{k,u}}{dQ_{k,0}} = \sum_{i=0}^{k-1} \log\left(1 + h\left(\frac{u^\top g_{\tilde{x}_i}(z_i)}{\sqrt{k}}\right)\right) - \sum_{i=0}^{k-1} \log C^{u/\sqrt{k}}_{\tilde{x}_i}. \tag{57}$$

By Lemma 28, 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 (57) 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 27, 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 (57) satisfies

$$\begin{aligned}\sum_{i=0}^{k-1} \log\biggl(1 + h\biggl(\frac{u^\top g_{\tilde{x}_i}(z_i)}{\sqrt{k}}\biggr)\biggr) \\ &= u^\top \biggl(\frac{1}{\sqrt{k}} \sum_{i=0}^{k-1} g_{\tilde{x}_i}(z_i)\biggr) - \frac{1}{2} u^\top \biggl(\frac{1}{k} \sum_{i=0}^{k-1} g_{\tilde{x}_i}(z_i) g_{\tilde{x}_i}(z_i)^\top\biggr) u + k \cdot o(k^{-1}) \\ &= u^\top Z_k - \frac{1}{2} u^\top V_k u + o(1) \qquad \text{as } k \to \infty,\end{aligned}$$

where $Z_k \colon \mathcal{Z}^k \to \mathbf{R}^d$ and $V_k \colon \mathcal{Z}^k \to \mathbf{R}^{d \times d}$ 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 $\{Q_{k,u} \,|\, u \in \mathbf{R}^d\}$ 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:

$$V_k \xrightarrow{\text{a.s.}} \Sigma_g \qquad \text{as } k \to \infty. \tag{58}$$

This is a consequence of the martingale strong law of large numbers (Theorem 33). Indeed, for each $i \geq 0$, set

$$\begin{aligned}X_{i+1} &= g_{\tilde{x}_i}(z_i) g_{\tilde{x}_i}(z_i)^\top - \mathbb{E}\bigl[g_{\tilde{x}_i}(z_i) g_{\tilde{x}_i}(z_i)^\top \,|\, \mathcal{F}_i\bigr] \\ &= g_{\tilde{x}_i}(z_i) g_{\tilde{x}_i}(z_i)^\top - \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_{\tilde{x}_i}} \bigl[g_{\tilde{x}_i}(z_i) g_{\tilde{x}_i}(z_i)^\top\bigr],\end{aligned}$$

thereby defining a martingale difference sequence X in $\mathbf{R}^{d \times d}$ adapted to $\mathbb{F}$; note that we have $\sup_i \mathbb{E}\|X_i\|_{\mathrm{F}}^2 < \infty$ by Lemma 27, so $\sum_{i=1}^{\infty} i^{-2} \mathbb{E}\|X_i\|_{\mathrm{F}}^2 < \infty$ and hence Theorem 33 implies

$$V_k - \frac{1}{k} \sum_{i=0}^{k-1} \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_{\tilde{x}_i}} \bigl[g_{\tilde{x}_i}(z_i) g_{\tilde{x}_i}(z_i)^\top\bigr] = \frac{1}{k} \sum_{i=1}^{k} X_i \xrightarrow{\text{a.s.}} 0 \qquad \text{as } k \to \infty. \tag{59}$$

On the other hand, we have $\tilde{x}_i \xrightarrow{\text{a.s.}} x^\star$ as $i \to \infty$ by Definition 12, so Lemma 35 implies

$$\mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_{\tilde{x}_i}} \bigl[g_{\tilde{x}_i}(z_i) g_{\tilde{x}_i}(z_i)^\top\bigr] \xrightarrow{\text{a.s.}} \mathop{\mathbb{E}}_{z \sim \mathcal{D}_{x^\star}} \bigl[g_{x^\star}(z) g_{x^\star}(z)^\top\bigr] = \Sigma_g \qquad \text{as } i \to \infty$$

and hence the arithmetic mean satisfies

$$\frac{1}{k} \sum_{i=0}^{k-1} \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_{\tilde{x}_i}} \bigl[g_{\tilde{x}_i}(z_i) g_{\tilde{x}_i}(z_i)^\top\bigr] \xrightarrow{\text{a.s.}} \Sigma_g \qquad \text{as } k \to \infty. \tag{60}$$

Combining (59) and (60) gives (58).

Finally, we establish $Z_k \overset{0}{\rightsquigarrow} \mathsf{N}(0, \Sigma_g)$ by applying the martingale central limit theorem (Theorem 34). 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 $\mathbf{R}^d$ adapted to the filtration $\mathbb{F}$. Indeed, the increments of M are clearly uniformly bounded (Lemma 27), M_k is $\mathcal{F}_k$-measurable, and

$$\mathbb{E}\big[M_{k+1} \,|\, \mathcal{F}_k\big] = M_k + \mathop{\mathbb{E}}_{z_k \sim \mathcal{D}_{\tilde{x}_k}}\big[g_{\tilde{x}_k}(z_k)\big] = M_k$$

by the unbiasedness condition of Definition 14. The predictable quadratic variation of M is given by

$$\langle M \rangle_k = \sum_{i=1}^{k} \mathbb{E}\big[(M_i - M_{i-1})(M_i - M_{i-1})^\top \,|\, \mathcal{F}_{i-1}\big] = \sum_{i=0}^{k-1} \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_{\tilde{x}_i}}\big[g_{\tilde{x}_i}(z_i) g_{\tilde{x}_i}(z_i)^\top\big].$$

Thus, by (60), we have

$$k^{-1}\langle M \rangle_k = \frac{1}{k} \sum_{i=0}^{k-1} \mathop{\mathbb{E}}_{z_i \sim \mathcal{D}_{\tilde{x}_i}}\big[g_{\tilde{x}_i}(z_i) g_{\tilde{x}_i}(z_i)^\top\big] \xrightarrow{\text{a.s.}} \Sigma_g \qquad \text{as } k \to \infty.$$

The assumptions of Theorem 34 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.

C.3 Proof of Lemma 23

Let $F \colon \mathcal{X} \times \mathbf{R}^d \to \mathbf{R}^d$ be the map given by

$$F(x, u) = \mathop{\mathbb{E}}_{z \sim \mathcal{D}_x^u}\big[G(x, z)\big] = \frac{1}{C_x^u} \mathop{\mathbb{E}}_{z \sim \mathcal{D}_x}\big[\big(1 + h\big(u^\top g_x(z)\big)\big) G(x, z)\big],$$

where we recall $C_x^u = 1 + \mathbb{E}_{z \sim \mathcal{D}_x}[h(u^\top g_x(z))]$. Lemma 30 directly implies that F is C^1-smooth. Consider now the family of smooth nonlinear equations

$$F(x, u) = 0 \tag{61}$$

parameterized by $u \in \mathbf{R}^d$. Note $F(x^\star, 0) = G_{x^\star}(x^\star) = 0$ since $x^\star \in \operatorname{int} \mathcal{X}$. More generally, the equality (61) 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 (61) determines $x_u^\star$ as a smooth function of u on a neighborhood of zero. To this end, observe that Lemma 30 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 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

$$x_u^\star = x^\star - W^{-1} \nabla_u F(x^\star, 0) u + o(\|u\|) \qquad \text{as } u \to 0. \tag{62}$$

By Lemma 30, 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, (62) 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 $u \in \mathbf{R}^d$, we have

$$\sqrt{k}\big(x^\star_{u/\sqrt{k}} - x^\star\big) = -W^{-1}\Sigma_{g,G}^\top u + \sqrt{k}\cdot o\bigg(\frac{1}{\sqrt{k}}\bigg) \to -W^{-1}\Sigma_{g,G}^\top u \qquad \text{as } k \to \infty.$$

The proof is complete.

C.4 Proof of Lemma 30

Recall first that the quantities

$$M_g' := \sup_{x\in\mathcal{X},\, z\in\mathcal{Z}} \|\nabla_x g_x(z)\|_{\mathrm{op}} \qquad \text{and} \qquad M_g := \sup_{x\in\mathcal{X},\, z\in\mathcal{Z}} \|g_x(z)\|$$

are finite by Lemma 27. The same argument shows that M_T' and M_T are finite.

Next, we turn to establishing that the map $H\colon \mathcal{X}\times\mathcal{W} \to \mathbf{R}^n$ 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 37, 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 $\bar{H}\colon \mathcal{X}\times\mathcal{X}\times\mathcal{W} \to \mathbf{R}^n$ 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 $\phi\colon \mathcal{K}\times\mathcal{Z} \to \mathbf{R}^n$ 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 5 applies. Clearly ϕ is measurable and bounded with $\sup\|\phi\| \le 2M_T$. 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 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

$$\nabla_y \bar{H}(x,y,u) = \mathop{\mathbb{E}}_{z \sim \mathcal{D}_x} \nabla_y \phi(y,u,z) \tag{63}$$

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

$$\nabla_y \phi(y,u,z) = \big(\nabla_y T(y,z)\big)\big(1 + h\big(u^\top g_y(z)\big)\big) + h'\big(u^\top g_y(z)\big)\big(T(y,z)u^\top\big)\nabla_y g_y(z) \tag{64}$$

and hence

$$\begin{aligned}\|\nabla_y \phi(y,u,z)\|_{\mathrm{op}} &\le 2\|\nabla_y T(y,z)\|_{\mathrm{op}} + (\sup|h'|)\|T(y,z)\|\|u\|\|\nabla_y g_y(z)\|_{\mathrm{op}} \\ &\le 2M_T' + \mathrm{diam}(\mathcal{W})L_h M_T M_g',\end{aligned}$$

so $\nabla_y \phi$ is in fact uniformly bounded. Therefore $\bar{H}(x,y,u)$ is smooth in y and (63) holds. Moreover, it follows from (64) 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 (64) that $z \mapsto \nabla_y \phi(y,u,z)$ is Lipschitz on $\mathcal{Z}$ with constant

$$a := 2\beta_T' + \mathrm{diam}(\mathcal{W})M_T' L_h \beta_g + \mathrm{diam}(\mathcal{W})L_h(\beta_T M_g' + M_T \beta_g') + \mathrm{diam}(\mathcal{W})^2 M_T M_g' L_{h'} \beta_g,$$

where $L_{h'} := \sup|h''|$. Hence (63) and Assumption 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 (64) that $y \mapsto \nabla_y \phi(y,u,z)$ is Lipschitz on $\mathcal{X}$ with constant

$$2\Lambda_T(z) + \mathrm{diam}(\mathcal{W})M_T' L_h M_g' + \mathrm{diam}(\mathcal{W})L_h\big(M_T \Lambda_g(z) + M_T' M_g'\big) + \mathrm{diam}(\mathcal{W})^2 M_T L_{h'} (M_T')^2.$$

Hence (63) implies that $y \mapsto \nabla_y \bar{H}(x,y,u)$ is Lipschitz on $\mathcal{X}$ with constant

$$2\bar{\Lambda}_T + \mathrm{diam}(\mathcal{W})L_h\big(M_T \bar{\Lambda}_g + 2M_T' M_g'\big) + \mathrm{diam}(\mathcal{W})^2 M_T L_{h'} (M_T')^2.$$

Similarly, it follows from (64) that $u \mapsto \nabla_y \phi(y,u,z)$ is Lipschitz on $\mathcal{W}$ with constant

$$M_T' M_g L_h + M_T M_g'\big(L_h + \mathrm{diam}(\mathcal{W})M_g L_{h'}\big),$$

so (63) 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

$$\nabla_u \bar{H}(x,y,u) = \mathop{\mathbb{E}}_{z \sim \mathcal{D}_x} \nabla_u \phi(y,u,z) \tag{65}$$

provided $\|\nabla_u \phi(y,u,z)\|_{\mathrm{op}}$ is dominated by a $\mathcal{D}_x$-integrable random variable independent of u. In this case, we have

$$\nabla_u \phi(y,u,z) = h'\big(u^\top g_y(z)\big) T(y,z) g_y(z)^\top \tag{66}$$

and hence

$$\|\nabla_u \phi(y,u,z)\|_{\mathrm{op}} \leq (\sup |h'|) \|T(y,z)\| \|g_y(z)\| \leq L_h M_T M_g.$$

Therefore $\bar{H}(x,y,u)$ is smooth in u and (65) holds. Moreover, it follows from (66) 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 (66) 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 (65) and Assumption 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 (66) 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 (65). Similarly, it follows from (66) 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 (65). 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 37 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 D. Underlying Probability Space

In this appendix, we formally construct the probability space where decision-dependent dynamics take place. The following lemma shows that Assumption 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 31 (Markov kernel) *Let $\mathcal{Z}$ be a nonempty Polish metric space. Then, for any bounded measurable function $\varphi\colon \mathcal{Z} \to \mathbf{R}$, the function $P_1(\mathcal{Z}) \to \mathbf{R}$ given by $\mu \mapsto \int \varphi \, d\mu$ is*

measurable. In particular, 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 $\mathcal{Z} \to \mathbf{R}$. For each $\varphi \in \mathcal{M}_b$, let $I_\varphi \colon P_1(\mathcal{Z}) \to \mathbf{R}$ 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 (e.g., see Kechris, 1995, Exercise 11.7) to show that $\mathcal{C}$ possesses the following two properties:

(i) Every bounded continuous function $\mathcal{Z} \to \mathbf{R}$ is contained in $\mathcal{C}$.

(ii) If (φ_n) is a uniformly bounded sequence in $\mathcal{C}$ with pointwise limit $\varphi \colon \mathcal{Z} \to \mathbf{R}$ (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 (e.g., see Ambrosio et al., 2008, Proposition 7.1.5); indeed, if $\varphi \colon \mathcal{Z} \to \mathbf{R}$ 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 $\varphi \colon \mathcal{Z} \to \mathbf{R}$, 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 $\varphi \colon \mathcal{Z} \to \mathbf{R}$; 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 $x_t \colon \mathcal{Z}^t \to \mathcal{X}$ (corresponding to iterates of a decision-dependent algorithm). This is provided by the following theorem, which may be viewed as a special case of either the Kolmogorov extension theorem (see Bass, 2011, Appendix D) or the Ionescu–Tulcea extension theorem (see Klenke, 2020, Theroem 14.35).

Theorem 32 (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 $\varphi\colon \mathcal{Z}^{t+1} \to \overline{\mathbf{R}}$ 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

$$\mathbb{E}[\varphi \circ \pi_{t+1} \,|\, \mathcal{F}_t] = \int_{\mathcal{Z}} \varphi(z_0, \ldots, z_t)\, d\mathcal{D}_{x_t}(z_t) = \mathop{\mathbb{E}}_{z_t \sim \mathcal{D}_{x_t}} [\varphi(z_0, \ldots, z_t)],$$

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 $\mathcal{F}_0 = \{\emptyset, \mathcal{Z}^{\mathbb{N}}\}$).

Appendix E. Supplementary Results

In this appendix, we record some supplementary results fundamental to our analysis. First, we record suitably general versions of the Strong Law of Large Numbers (see Dembo, 2021, Exercise 5.3.35) and the Central Limit Theorem (see Duflo, 1997, Corollary 2.1.10) for square-integrable martingales.

Theorem 33 (Martingale Strong Law of Large Numbers) *Let X be a square-integrable martingale difference sequence in $\mathbf{R}^n$ 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 34 (Martingale Central Limit Theorem) *Let M be a square-integrable martingale in $\mathbf{R}^n$ 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 35 (Asymptotic covariance) *Let $x_t \in \mathcal{X}$ be a sequence in some set $\mathcal{X} \subset \mathbf{R}^d$ 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 $g\colon \mathcal{X} \times \mathcal{Z} \to \mathbf{R}^n$ 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

$$\limsup_{t\to\infty} \mathop{\mathbb{E}}_{z\sim\mu_t}\left[\|g_{x^\star}(z)\|^2\right] \leq N_\delta^2 + \delta. \tag{67}$$

On the other hand, for all t, we also have the decomposition

$$\begin{aligned}\mathop{\mathbb{E}}_{z\sim\mu_t}\left[g_{x_t}(z)g_{x_t}(z)^\top\right] = &\mathop{\mathbb{E}}_{z\sim\mu_t}\left[g_{x^\star}(z)g_{x^\star}(z)^\top\right] + \mathop{\mathbb{E}}_{z\sim\mu_t}\left[g_{x^\star}(z)\big(g_{x_t}(z) - g_{x^\star}(z)\big)^\top\right] \\ &+ \mathop{\mathbb{E}}_{z\sim\mu_t}\left[\big(g_{x_t}(z) - g_{x^\star}(z)\big)g_{x_t}(z)^\top\right].\end{aligned} \tag{68}$$

The last two summands in (68) 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 (67) to conclude

$$\begin{aligned}\left\|\mathop{\mathbb{E}}_{z\sim\mu_t}\left[g_{x^\star}(z)\big(g_{x_t}(z) - g_{x^\star}(z)\big)^\top\right]\right\|_{\mathrm{op}} &\leq \mathop{\mathbb{E}}_{z\sim\mu_t}\left[\|g_{x^\star}(z)\| \cdot \|g_{x_t}(z) - g_{x^\star}(z)\|\right] \\ &\leq \|x_t - x^\star\| \sqrt{\mathop{\mathbb{E}}_{z\sim\mu_t}\left[\|g_{x^\star}(z)\|^2\right] \cdot \mathop{\mathbb{E}}_{z\sim\mu_t}\left[L(z)^2\right]} \\ &\to 0 \qquad \text{as } t \to \infty\end{aligned}$$

and

$$\begin{aligned}
&\Big\| \mathop{\mathbb{E}}_{z\sim\mu_t}\big[\big(g_{x_t}(z) - g_{x^\star}(z)\big)g_{x_t}(z)^\top\big]\Big\|_{\mathrm{op}} \\
&\qquad \le \mathop{\mathbb{E}}_{z\sim\mu_t}\big[\|g_{x_t}(z) - g_{x^\star}(z)\| \cdot \|g_{x_t}(z)\|\big] \\
&\qquad \le \|x_t - x^\star\| \sqrt{\mathop{\mathbb{E}}_{z\sim\mu_t}\big[L(z)^2\big] \cdot \mathop{\mathbb{E}}_{z\sim\mu_t}\big[\|g_{x_t}(z)\|^2\big]} \\
&\qquad \le \|x_t - x^\star\| \sqrt{2\mathop{\mathbb{E}}_{z\sim\mu_t}\big[L(z)^2\big]\Big(\mathop{\mathbb{E}}_{z\sim\mu_t}\big[\|g_{x^\star}(z)\|^2\big] + \mathop{\mathbb{E}}_{z\sim\mu_t}\big[\|g_{x_t}(z) - g_{x^\star}(z)\|^2\big]\Big)} \\
&\qquad \le \|x_t - x^\star\| \sqrt{2\mathop{\mathbb{E}}_{z\sim\mu_t}\big[L(z)^2\big]\Big(\mathop{\mathbb{E}}_{z\sim\mu_t}\big[\|g_{x^\star}(z)\|^2\big] + \|x_t - x^\star\|^2 \cdot \mathop{\mathbb{E}}_{z\sim\mu_t}\big[L(z)^2\big]\Big)} \\
&\qquad \to 0 \qquad \text{as } t \to \infty.
\end{aligned}$$

To complete the proof, it now suffices by (68) 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 $q \in \mathbf{R}$ the step-like function $\varphi_q\colon \mathbf{R} \to \mathbf{R}$ 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 $\Phi\colon \mathbf{R}^n \to \mathbf{R}^{n\times n}$ 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,

$$\begin{aligned}\|B_t\|_{\mathrm{op}} &= \Big\| \mathop{\mathbb{E}}_{z\sim\mu_t}\big[\phi(z)\big] - \mathop{\mathbb{E}}_{z\sim\mu^\star}\big[\phi(z)\big]\Big\|_{\mathrm{op}} \\ &= \sup_{\|u\|,\|v\|\le 1}\left\{ \mathop{\mathbb{E}}_{z\sim\mu_t}\big[\langle\phi(z)u,v\rangle\big] - \mathop{\mathbb{E}}_{z\sim\mu^\star}\big[\langle\phi(z)u,v\rangle\big]\right\} \\ &\le C\cdot W_1(\mu_t,\mu^\star)\to 0 \qquad \text{as } t\to\infty,\end{aligned}$$

where the inequality follows from the C-Lipschitz continuity of the function $z\mapsto\langle\phi(z)u,v\rangle$. Hence

$$\limsup_{t\to\infty}\Big\| \mathop{\mathbb{E}}_{z\sim\mu_t}\big[g_{x^\star}(z)g_{x^\star}(z)^\top\big]-\Sigma\Big\|_{\mathrm{op}} \le \limsup_{t\to\infty}\big(\|A_t\|_{\mathrm{op}}+\|B_t\|_{\mathrm{op}}\big)\le 2\delta.$$

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 36 *Let $\mathcal{K}$ be a metric space and suppose that $f\colon\mathcal{K}\to\mathbf{R}^{n\times q}$ and $g\colon\mathcal{K}\to\mathbf{R}^{q\times m}$ are bounded and Lipschitz. Then the product $fg\colon\mathcal{K}\to\mathbf{R}^{n\times m}$ 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 37 *Let $\mathcal{K}\subset\mathbf{R}^m$ be a compact set and suppose that $f\colon\mathcal{K}\to\mathbf{R}^n$ and $g\colon\mathcal{K}\to\mathbf{R}\setminus\{0\}$ 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

$$\nabla(f/g)=(1/g)(\nabla f)-(f/g^2)(\nabla g)^\top. \tag{69}$$

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, (69) and Lemma 36 show that $\nabla(f/g)$ is the difference of two Lipschitz maps. Therefore $\nabla(f/g)$ is Lipschitz. ■

References

Shabbir Ahmed. *Strategic Planning Under Uncertainty: Stochastic Integer Programming Approaches*. University of Illinois at Urbana-Champaign, 2000.

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.

Richard F. Bass. *Stochastic Processes*. Cambridge University Press, 2011.

Yahav Bechavod, Katrina Ligett, Zhiwei Steven Wu, and Juba Ziani. Causal feature discovery through strategic modification. *arXiv:2002.07024*, 2020.

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.

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

Damek Davis, Dmitriy Drusvyatskiy, and Liwei Jiang. Subgradient methods near active manifolds: saddle point avoidance, local convergence, and asymptotic normality. *arXiv:2108.11832*, 2021.

Amir Dembo. *Probability Theory: STAT310/MATH230*. Stanford University, 2021.

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.

Asen L. Dontchev and R. Tyrrell Rockafellar. *Implicit Functions and Solution Mappings: A View from Variational Analysis*. Springer, 2009.

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

John C. Duchi and Feng Ruan. Asymptotic optimality in stochastic optimization. *The Annals of Statistics*, 49(1):21–48, 2021.

Marie Duflo. *Random Iterative Models*. Stochastic Modeling and Applied Probability. Springer Berlin, Heidelberg, 1997.

Jitka Dupačová. Optimization under exogenous and endogenous uncertainty. *University of West Bohemia in Pilsen*, 2006.

Rick Durrett. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 5th edition, 2019.

Vaclav Fabian. On Asymptotic Normality in Stochastic Approximation. *The Annals of Mathematical Statistics*, 39(4):1327–1332, 1968.

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.

Lars Hellemo, Paul I. Barton, and Asgeir Tomasgard. Decision-dependent probabilities in stochastic programs with recourse. *Computational Management Science*, 15(3):369–395, 2018.

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.

Meena Jagadeesan, Tijana Zrnic, and Celestine Mendler-Dünner. Regret minimization with performative feedback. In *International Conference on Machine Learning*, pages 9760–9785. PMLR, 2022.

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.

Alexander S. Kechris. *Classical Descriptive Set Theory*. Springer, 1995.

David Kinderlehrer and Guido Stampacchia. *An Introduction to Variational Inequalities and Their Applications*. SIAM, 2000.

Achim Klenke. *Probability Theory: A Comprehensive Course*. Springer, 3rd edition, 2020.

Lucien Le Cam and Grace Lo Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer, 2000.

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.

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.

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.

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.

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.

Georgios Piliouras and Fang-Yi Yu. Multi-agent performative prediction: From global stability and optimality to chaos. *arXiv:2201.10483*, 2022.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

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.

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.

R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*. Springer, 1998.

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.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

Pravin Varaiya and Roger J.-B. Wets. Stochastic dynamic optimization approaches and computation. IIASA WP-88-87, Laxenburg, Austria, 1988.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

Killian Wood and Emiliano Dall'Anese. Stochastic saddle point problems with decision-dependent distributions. *SIAM Journal on Optimization*, 33(3):1943–1967, 2023.

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, 2021.