Stochastic approximation with decision-dependent distributions: asymptotic normality and optimality

Joshua Cutler* Mateo Díaz† Dmitriy Drusvyatskiy‡

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 nicely 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 is locally minimax optimal.

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 from the same distribution. Recent literature on strategic classification [12] and performative prediction [23], 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 [22, 24, 33].

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 $\mathcal{X}$ be a convex set of feasible decisions. A variety of classical learning problems can be posed as solving the variational inequality

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

where $\mathcal{P}$ is some fixed distribution and $N_{\mathcal{X}}$ is the normal cone to $\mathcal{X}$. Two examples are worth keeping in mind: (i) standard problems of supervised learning amount to $G(x, z) = \nabla\ell(x, z)$ being the gradient of some loss function, while (ii) stochastic games correspond to $G(x, z)$ being a stacked gradient of the players' individual losses.

*Department of Mathematics, U. Washington, Seattle, WA 98195; jocutler@uw.edu.

†Department of Computing and Mathematical Sciences, Caltech, Pasadena, CA 91125, USA; http://users.cms.caltech.edu/~mateodd/.

‡Department of Mathematics, U. Washington, Seattle, WA 98195; www.math.washington.edu/~ddrusv. Research of Drusvyatskiy was supported by the NSF DMS 1651851 and CCF 1740551 awards.

Following the recent literature on performative prediction [12, 22, 23], 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 $\mathrm{VI}(\mathcal{D}(x^\star))$, or equivalently:

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

We will say that such points $x^\star$ are at *equilibrium*. In words, a learning system that deploys an equilibrium point $x^\star$ has no incentive to deviate from $x^\star$ based only on the solution of the variational inequality $\mathrm{VI}(\mathcal{D}(x^\star))$ induced by the response distribution $\mathcal{D}(x^\star)$. The setting of performative prediction [23] corresponds to the choice $G(x, z) = \nabla\ell(x, z)$ for some loss function ℓ.[1] More generally, decision-dependent games, proposed by [22, 24, 33], 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 algorithm for finding an equilibrium point $x^\star$ is the *stochastic forward-backward algorithm*:

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

where $\mathrm{proj}_{\mathcal{X}}$ is the nearest-point projection onto $\mathcal{X}$. Specializing to performative prediction [19] and its multiplayer extension [22], this algorithm reduces to a basic projected stochastic gradient iteration. The contribution of our paper can be informally summarized as follows.

We show that algorithm (1) 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

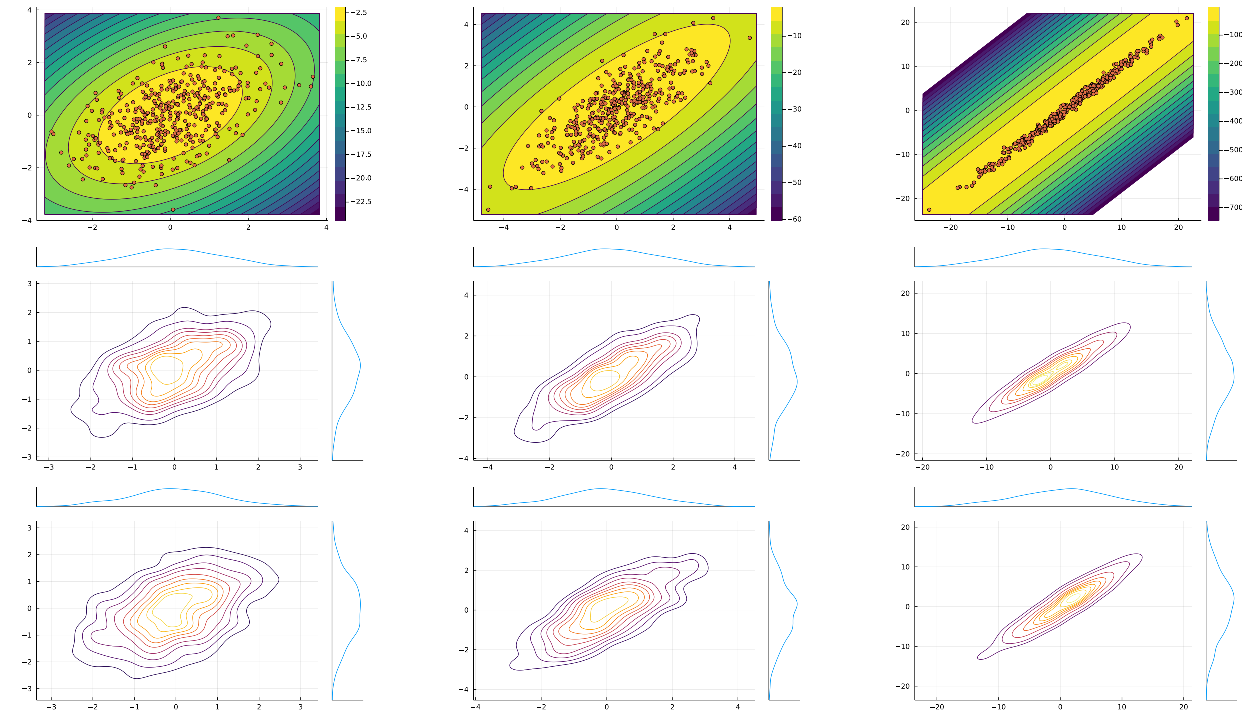
Arguing optimality of an algorithm is a two step process: (*i*) estimate the performance of the specific algorithm and (*ii*) derive a matching lower-bound that is valid among all possible estimation procedures. Beginning with the former, we build on the seminal work of Polyak and Juditsky [25]. Letting $\bar{x}_t = \frac{1}{t}\sum_{i=0}^{t-1} x_i$ denote the average iterate, 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 1.1 (Informal)**.** *Suppose that $G(\cdot, z)$ is α-strongly monotone and Lipschitz, $G(x, \cdot)$ is β-Lipschitz, and the distribution $\mathcal{D}(\cdot)$ is γ-Lipschitz in the Wasserstein-1 distance. Suppose moreover that $x^\star$ lies in the interior of $\mathcal{X}$ and we set $\eta_t = \eta_0 t^{-\nu}$ with $\eta_0 > 0$ and $\nu \in (\frac{1}{2}, 1)$. Then in the regime $\frac{\gamma\beta}{\alpha} < 1$, the average iterate $\bar{x}_t$ converges to $x^\star$ almost surely and satisfies*

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

[1] In the notation of [23], equilibria coincide with performatively stable points.

[2] The code can be found at https://github.com/mateodd25/Asymptotic-normality-in-performative-prediction.



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

Figure 1: Consider the problem corresponding to $G(x, z) = \nabla \ell(x, z)$ 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 (1) 400 times using $\eta_t = t^{-3/4}$ for 10^6 iterations. The first row depicts the resulting average iterates laid over the confidence regions (plotted in logarithmic scale) corresponding to the asymptotic normal distribution. The next two rows depict kernel density estimates from the asymptotic normal distribution (top) and the deviation $\sqrt{t}(\bar{x}_t - x^\star)$ (bottom).

where

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

A few comments are in order. First, the regime $\frac{\gamma\beta}{\alpha} < 1$ is, in essence, optimal because otherwise, equilibrium points may even fail to exist. Second, the effect of the distributional shift on the asymptotic covariance is entirely captured by the second "dynamic" term in W. Indeed, when this term is absent, the product $W^{-1} \cdot \Sigma \cdot W^{-\top}$ 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 1.1 follows by interpreting (1) as a stochastic approximation algorithm for finding a zero of the nonlinear map $R(x) = \mathbb{E}_{z \sim \mathcal{D}(x)} G(x, z)$, and then applying the results of Polyak and Juditsky [25]. The main technical work is in verifying the necessary assumptions needed in [25] for this particular nonlinear map and the corresponding noise sequence.

A reasonable question to ask is whether there exists an algorithm with better asymptotic guarantees than those of the stochastic forward-backward algorithm. We will show that in a strong

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

sense, the answer is no; that is, Algorithm (1) is asymptotically optimal. We will lower-bound the performance of *any* estimation procedure for finding an equilibrium point on an adversarially-chosen sequence of small perturbations of the target problem. That is, we will define a shrinking neighborhood of distribution maps $B_{1/k}$ around $\mathcal{D}$. Every map $\mathcal{D}' \in B_{1/k}$ defines a new problem with equilibrium point $x^\star_{\mathcal{D}'}$ near $x^\star$. Roughly speaking, we will show that for every estimator $\widehat{x}_k$ (i.e. a deterministic function of the observed samples), the local minimax bound holds:

$$\liminf_{k\to\infty} \sup_{\mathcal{D}'\in B_{1/k}} \mathbb{E}[\mathcal{L}(\sqrt{k}(\widehat{x}_k - x^\star_{\mathcal{D}'}))] \geq \mathbb{E}[\mathcal{L}(Z)], \tag{2}$$

where $Z \sim \mathsf{N}(0, W^{-1}\cdot\Sigma\cdot W^{-\top})$ and $\mathcal{L}$ is any nonnegative, symmetric, quasiconvex, lower semicontinuous function. The end result is the following theorem.

Theorem 1.2 (Informal)**.** *Suppose the same setting as in Theorem 1.1 and let $\mathcal{L}$ be any nonnegative, symmetric, quasiconvex, lower semicontinuous function. Fix any procedure for finding equilibrium points that outputs $\widehat{x}_k$ based on k samples. As $k \to \infty$, there is a family of perturbed distributions $\mathcal{D}_k(x)$ converging to $\mathcal{D}(x)$, along with corresponding equilibrium points $x^\star_k$ converging to $x^\star$, such that the expected error $\mathbb{E}[\mathcal{L}(\sqrt{k}(\widehat{x}_k - x^\star_k))]$ of the estimator $\widehat{x}_k$ on the perturbed problem is asymptotically lower bounded by $\mathbb{E}[\mathcal{L}(Z)]$ where $Z \sim \mathsf{N}(0, W^{-1}\cdot\Sigma\cdot W^{-\top})$.*

The formal statement of the theorem and its proof follow closely the classical work of Hájek and Le Cam [17, 29] on statistical lower bounds and the more recent work of Duchi and Ruan [8] on asymptotic optimality of the stochastic gradient method. In particular, the fundamental role of tilt-stability and the inverse function theorem highlighted in [8] is replaced by the implicit function theorem paradigm.

1.2 Related literature

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

Learning with decision-dependent distributions. The basic setup for decision-dependent problems that we use is inspired by the performative prediction framework introduced by [23] and its multiplayer extension developed independently in [22, 24, 33]. The stochastic gradient method for performative prediction was first introduced and analyzed in [18], while the stochastic forward-backward method for games was analyzed in [22]. A related work [7] 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 [23]. Seeking to develop algorithms for finding performatively optimal points, the paper [21] provides sufficient conditions for the prediction problem to be convex; extensions of such conditions to games appear in [22, 33]. Algorithms for finding performatively optimal points under a variety of different assumptions and oracle models appear in [14, 15, 21, 22, 33]. The performative prediction framework is largely motivated by the problem of strategic classification [12], which has been studied extensively from the perspective of causal inference [2, 20] and convex optimization [6]. Other lines of work [3, 4, 26, 34] in performative prediction have focused on the setting in which the environment evolves dynamically in time or experiences time drift.

Stochastic approximation. There is extensive literature on stochastic approximation. The most relevant results for us are those of [25] 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, such as [1, 10, 16, 28, 31]. We refer the reader to the recent paper [13], 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 [7, 18, 22, 23, 24, 33] 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 [32, 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 [17, 29], provide much finer problem-specific guarantees. Building on this framework, Duchi and Ruan [8] showed that the stochastic gradient method for standard single-stage stochastic optimization problems is, in an appropriate sense, locally 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 forward-backward algorithm is asymptotically normal (Theorem 1.1), and identify its asymptotic covariance. Finally, Section 5 proves the local minimax lower-bound (Theorem 1.2). We defer many of the technical proofs to the Appendix.

2 Basic notation

Throughout, we let $\mathbf{R}^d$ denote a d-dimensional Euclidean space, with inner product $\langle \cdot, \cdot \rangle$ and the induced Euclidean norm $\|x\| = \sqrt{\langle x, x \rangle}$. For any set $\mathcal{X} \subset \mathbf{R}^d$, the symbol $\mathrm{proj}_{\mathcal{X}}(x)$ will denote the set 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 $\{x \mid \mathcal{L}(x) \leq c\}$ 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. If A is a symmetric matrix, then the notation $A \succeq 0$ will mean that A is positive semidefinite (PSD). Given two symmetric matrices A and B, we write $A \succeq B$ as a shorthand for $A - B \succeq 0$.

A map $F \colon \mathcal{X} \to \mathbf{R}^d$ is called *α-strongly monotone* on $\mathcal{X} \subset \mathbf{R}^d$ if it satisfies

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

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

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

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

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

Here, $\mathrm{Lip}_1(\mathcal{Z})$ denotes the set of 1-Lipschitz continuous functions $h \colon \mathcal{Z} \to \mathbf{R}$. We will refer to a Borel measurable map between metric spaces simply as *measurable*. Given a map $H \colon \mathcal{X} \times \mathcal{Z} \to \mathbf{R}^k$ with $\mathcal{X} \subset \mathbf{R}^d$, the symbol $\nabla H(x, z)$ will always denote the Jacobian of $H(\cdot, z)$ evaluated at x.

Let P and Q be probability distributions over $\mathcal{Z}$ such that P is absolutely continuous with respect to Q (denoted $P \ll Q$). Then for any convex function $f\colon (0,\infty) \to \mathbf{R}$ with $f(1) = 0$, the f-divergence between P and Q is given by

$$\Delta_f(P \parallel Q) = \int_{\mathcal{Z}} f\left(\frac{dP}{dQ}\right) dQ, \tag{3}$$

where $\frac{dP}{dQ}$ denotes the Radon-Nikodym derivative of P with respect to Q and $f(0) := \lim_{t \downarrow 0} f(t)$. Abusing notation slightly, if P is not absolutely continuous with respect to Q, we set $\Delta_f(P \parallel Q) = \infty$.

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 metric space $(\mathcal{Z}, d_{\mathcal{Z}})$, and a map $\mathcal{D}\colon \mathcal{X} \to P_1(\mathcal{Z})$. Thus, $\{\mathcal{D}(x)\}$ is a family of probability distributions indexed by points $x \in \mathcal{X}$. The variational behavior of the map $\mathcal{D}(\cdot)$ will play a central role in our work. In particular, following [23], we will assume that $\mathcal{D}(\cdot)$ is Lipschitz continuous.

Assumption 1 (Lipschitz distribution map)**.** There is a constant $\gamma > 0$ such that

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

Next, we fix a measurable map $G\colon \mathcal{X} \times \mathcal{Z} \to \mathbf{R}^d$ such that each section $G(x, \cdot)$ is Lipschitz continuous, and we define the family of maps[4]

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

We impose standard regularity conditions on $G(x, z)$.

Assumption 2 (Loss regularity)**.** There are constants $\bar{L}, \beta, \alpha > 0$ satisfying the following.

(i) **(Lipschitz)** There is a measurable function $L\colon \mathcal{Z} \to [0, \infty)$ such that the Lipschitz bounds

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

hold for all $x, x' \in \mathcal{X}$ and $z, z' \in \mathcal{Z}$. Further, $\mathbb{E}_{z \sim \mathcal{D}(x)}[L(z)^2] \leq \bar{L}^2$ 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) simply asserts that the map $G(x, z)$ is Lipschitz continuous both in x and in z. An immediate consequence is that $G_x(\cdot)$ is $\bar{L}$-Lipschitz continuous. Condition (ii) is a standard convexity and monotonicity requirement. Condition (iii) ensures that the Lipschitz constant γ of $\mathcal{D}(\cdot)$ is sufficiently small in comparison with the monotonicty constant α, signifying that the dynamics are "mild". This condition is widely used in the existing literature; see, e.g., [22, 23, 24, 33].

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 [22, Lemma 2]; a short argument appears in Appendix A.

[4] Note that $G_x(y)$ is well defined because $\mathcal{D}(x)$ has a finite first moment and $G(y, \cdot)$ is Lipschitz continuous.

Lemma 3.1 (Deviation)**.** *Suppose that Assumptions 1 and 2 hold. Then the estimate*

$$\|G_x(y) - G_{x'}(y)\| \le \gamma\beta \cdot \|x - x'\| \qquad \textit{holds for all } \; x, x', y \in \mathcal{X}.$$

The following definition, originating in [23] for performative prediction and in [22] for its multiplayer extension, is the key solution concept that we will use.

Definition 3.2 (Equilibrium point)**.** We say that $x^\star$ is an *equilibrium point* if it satisfies:

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

In words, $x^\star$ is an equilibrium point if $x = x^\star$ solves the variational inequality VI($\mathcal{P}$) induced by the distribution $\mathcal{P} = \mathcal{D}(x^\star)$. Equivalently, these are exactly the fixed points of the map

$$\texttt{Sol}(x) := \{y \mid 0 \in G_x(y) + N_{\mathcal{X}}(y)\}, \tag{4}$$

which is single-valued on $\mathcal{X}$ by strong monotonicity. 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., [7, 18, 22, 23, 24, 33]. Most importantly, equilibrium points exist and are unique under Assumptions 1 and 2. The proof is identical to that of [22, Theorem 1]; we provide a short argument in Appendix A for completeness.

Theorem 3.3 (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 $\gamma\beta \ge \alpha$, equilibrium points may easily fail to exist; see, e.g., [23, 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 (4). 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. In the setting of performative prediction [18] and its multiplayer extension [22], the algorithm reduces to stochastic projected gradient methods.

Algorithm 1 Stochastic Forward-Backward Method (SFB)

Input: initial $x_0 \in \mathcal{X}$ and step sizes $(\eta_t)_{t\ge0} \subset (0, \infty)$
Step $t \ge 0$:

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

We will see that under very mild assumptions, the SFB iterates almost surely converge to the equilibrium point $x^\star$. To this end, let $(\Omega, \mathcal{F}, \mathbb{P})$ denote the probability space where this stochastic process takes place[5], and define the noise vector

$$\xi_x(z) := G(x,z) - G_x(x).$$

We impose the following standard second moment bound on $\xi_x(z)$.

Assumption 3 (Variance bound)**.** There is a constant $K \geq 0$ such that

$$\mathop{\mathbb{E}}_{z\sim\mathcal{D}(x)} \|\xi_x(z)\|^2 \leq K(1 + \|x - x^\star\|^2) \qquad \text{for all } x \in \mathcal{X}.$$

The following proposition shows that the SFB iterates almost surely converge to the equilibrium point. The proof, which follows from a simple one-step improvement bound for SFB and the Robbins-Siegmund theorem [27], appears in Appendix A.3.

Proposition 4.1 (Almost sure convergence)**.** *Suppose that Assumptions 1, 2, and 3 hold, and let the step size sequence satisfy $\eta_t > 0$, $\sum_{t=0}^{\infty} \eta_t = \infty$, and $\sum_{t=0}^{\infty} \eta_t^2 < \infty$. Then x_t converges to $x^\star$ almost surely. Moreover, if $\eta_t = \eta_0 t^{-\nu}$ with $\nu \in (\frac{1}{2}, 1)$, then there exists a constant $C \geq 0$ such that*

$$\mathbb{E}\|x_t - x^\star\|^2 \leq C\eta_t \qquad \textit{for all } t \geq 0.$$

The main result of this section is the asymptotic normality of the average iterates $\bar{x}_t := \frac{1}{t}\sum_{i=0}^{t-1} x_i$, for which we require the following additional assumption.

Assumption 4. The following three conditions hold.

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

2. **(Joint smoothness)** The map $(x,y) \mapsto G_x(y)$ is smooth and has a Lipschitz continuous Jacobian on some neighborhood of $(x^\star, x^\star)$.

3. **(Uniform integrability)** There is a neighborhood $\mathcal{V}$ around $x^\star$ in $\mathcal{X}$ such that for all $\delta > 0$, there exists a constant N_δ such that:

   (i) $\mathop{\mathbb{E}}_{z\sim\mathcal{D}(x)} \left[\|G(x^\star,z)\|^2 \mathbf{1}\{\|G(x^\star,z)\| \geq N_\delta\}\right] \leq \delta$ for all $x \in \mathcal{V}$;

   (ii) $\mathop{\mathbb{E}}_{z\sim\mathcal{D}(x)} \left[\|\xi_x(z)\|^2 \mathbf{1}\{\|\xi_x(z)\| \geq N_\delta\}\right] \leq \delta$ for all $x \in \mathcal{V}$.

A few comments are in order. First, the interiority condition 1 is a standard assumption for asymptotic normality results even in static settings [25]. The joint smoothness condition 2 is fairly mild. For example, it holds automatically provided the partial derivatives $\nabla_y G_x(y)$ and $\nabla_x G_x(y)$ exist and vary in a locally Lipschitz way with respect to (x,y); 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 base measure $\mu \gg \mathcal{D}(x)$ such that $G(\cdot,z)$ and $p(\cdot,z)$ are $C^{1,1}$-smooth[6] and sufficient integrability conditions hold to apply the dominated convergence theorem. Condition 3 is a standard uniform integrability condition. For instance, it holds when the random vectors $G(x^\star,z)$ and $\xi_x(z)$, with $z \sim \mathcal{D}(x)$, are sub-Gaussian with the same variance proxy σ^2 for all x near $x^\star$.

We are now ready to present our main result.

[5]Formally, in order to define a single probability space, we may proceed as follows. Suppose that there is probability space $(\mathcal{S},\mathcal{H},\mu)$ and a measurable map $\theta\colon \mathcal{S}\times\mathcal{X} \to \mathcal{Z}$ such that $\mathcal{D}(x)$ is the pushforward of μ by the section $\theta(\cdot,x)$. This means that for every Borel set $V \subset \mathcal{Z}$, the $\mathcal{D}(x)$-measure of V is equal to the μ-measure of the set $\{s \in \mathcal{S} : \theta(s,x) \in V\}$. Then, we may simply define $(\Omega,\mathcal{F},\mathbb{P})$ as the countable product $(\mathcal{S},\mathcal{H},\mu)^\infty$.

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

Theorem 4.2 (Asymptotic normality)**.** *Suppose that Assumptions 1, 2, 3, and 4 hold, and let the step size sequence be set to* $\eta_t = \eta_0 t^{-\nu}$ *with* $\eta_0 > 0$ *and* $\nu \in (\frac{1}{2}, 1)$*. Then the average iterates* $\bar{x}_t = \frac{1}{t}\sum_{i=0}^{t-1} x_i$ *converge to* $x^\star$ *almost surely, and*

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

where

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

Thus Theorem 4.2 asserts that under mild assumptions, the deviations $\sqrt{t}(\bar{x}_t - x^\star)$ converge in distribution to a Gaussian random vector with a covariance matrix $\nabla R(x^\star)^{-1} \cdot \Sigma \cdot \nabla R(x^\star)^{-\top}$. Explicitly, under mild regularity conditions we may write

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

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 variance 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:

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

Example 4.3 (Performative prediction with location-scale families)**.** As an explicit example of Theorem 4.2, let us look at the case when $G(x,z) = \nabla_x \ell(x,z)$ is a gradient of a loss function and $\mathcal{D}(x)$ are affine perturbations of a fixed base distribution $\mathcal{D}_0$. Such distributions are quite reasonable when modeling performative effects, as explained in [21]. In this case, we have

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

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

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

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

Example 4.4 (Multiplayer performative prediction with location-scale families)**.** More generally, let us look at the problem of multiplayer performative prediction [22]. In this case, the map G takes the form

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

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

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

As highlighted in [22], a natural assumption on the distributions is that there exist some 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 coordinates x_i. Thus the distributional assumption asserts that the distribution used by player i is a "linear perturbation" of some 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

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

under mild integrability conditions. Thus, the dynamic part of $\nabla R_i(x^\star)$ is governed by $A_i^\top$, $A_{-i}^\top$ and the matrix $\nabla^2_{zx} \ell_i(x^\star, z_i)$ of mixed partial derivatives. The former measure the performative effects of the distributional shift, while the latter measures the sensitivity of the gradient $\nabla_x \ell_i(x^\star, z_i)$ to changes in the data z_i.

4.1 Proof of Theorem 4.2

The proof of Theorem 4.2 is based on the stochastic approximation result of Polyak and Juditsky [25, Theorem 2], which we review in Appendix B. For the remainder of the section, we impose the assumptions of Theorem 4.2.

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 solution to the equation $R(x) = 0$. To simplify notation, let us define the direction of motion and noise vectors:

$$v_t := G(x_t, z_t) \qquad \text{and} \qquad \xi_t := \xi_{x_t}(z_t). \tag{5}$$

Observe that v_t and ξ_t satisfy the relation

$$v_t = R(x_t) + \xi_t,$$

and so we may write the iterates of Algorithm 1 as

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

where

$$\zeta_t = \eta_t^{-1} (x_t - \eta_t v_t - \operatorname{proj}_{\mathcal{X}}(x_t - \eta_t v_t)). \tag{7}$$

Our goal is to apply Theorem B.1 to the process (6). In what follows, we establish the necessary assumptions for the theorem, namely, Assumptions B.1, B.2, and B.3.

We note first that Proposition 4.1 establishes Assumption B.2. It also follows from Assumption 3 and Proposition 4.1 that ξ_t is square-integrable; further, letting $\mathbb{F} = (\mathcal{F}_t)_{t \geq 0}$ denote the filtration given by

$$\mathcal{F}_t := \sigma(z_0, \ldots, z_{t-1}) \tag{8}$$

(with $\mathcal{F}_0 = \{\emptyset, \Omega\}$), it follows that ξ_t is $\mathcal{F}_{t+1}$-measurable and

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

Thus, ξ_t constitutes a martingale difference process on $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$.

To verify condition (i) of Assumption B.1, we need the following lemma.

Lemma 4.5 (Lipschitz continuity)**.** *The map R is $(\bar{L} + \gamma\beta)$-Lipschitz continuous.*

Proof. Given $x, y \in \mathcal{X}$, we have

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

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

In light of Assumption 3 and Lemma 4.5, we have

$$\mathbb{E}[\|\xi_t\|^2 \mid \mathcal{F}_t] \le K(1 + \|x_t - x^\star\|^2) \qquad \text{and} \qquad \|R(x_t)\|^2 \le (\bar{L} + \gamma\beta)^2 \|x_t - x^\star\|^2;$$

thus, upon setting

$$\bar{K} := 2\big(K + (\bar{L} + \gamma\beta)^2\big)(1 + \|x^\star\|^2),$$

we have

$$\mathbb{E}[\|\xi_t\|^2 \mid \mathcal{F}_t] + \|R(x_t)\|^2 \le \bar{K}(1 + \|x_t\|^2).$$

Hence condition (i) of Assumption B.1 holds. Moreover, since $x_t \to x^\star$ almost surely, the second uniform integrability condition of Assumption 4 implies

$$\sup_{t\ge 0} \mathbb{E}[\|\xi_t\|^2 \mathbf{1}\{\|\xi_t\| > N\} \mid \mathcal{F}_t] \xrightarrow{\text{a.s.}} 0 \qquad \text{as } N \to \infty,$$

so condition (ii) of Assumption B.1 holds.

The next lemma shows that $\mathbb{E}[\xi_t\xi_t^\top \mid \mathcal{F}_t]$ 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, thereby establishing condition (iii) of Assumption B.1.

Lemma 4.6 (Asymptotic covariance)**.** *The following limits hold:*

$$\mathbb{E}[v_t v_t^\top \mid \mathcal{F}_t] \xrightarrow{a.s.} \Sigma \qquad \textit{and} \qquad \mathbb{E}[\xi_t\xi_t^\top \mid \mathcal{F}_t] \xrightarrow{a.s.} \Sigma.$$

Proof. Taking into account almost sure convergence of x_t to $x^\star$, the first uniform integrability condition of Assumption 4, and the Lipschitz condition of Assumption 2, we may apply Lemma E.3 with $g = G$ along any sample path $x_t \to x^\star$ to obtain $\mathbb{E}[v_t v_t^\top \mid \mathcal{F}_t] \to \Sigma$ almost surely. Then we have

$$\mathbb{E}[\xi_t\xi_t^\top \mid \mathcal{F}_t] = \mathbb{E}[v_t v_t^\top \mid \mathcal{F}_t] - R(x_t)R(x_t)^\top \xrightarrow{\text{a.s.}} \Sigma,$$

as claimed. □

Conditions (i), (ii), and (iii) of Assumption B.1 are now established. Now consider the stochastic process $(\zeta_t)_{t\ge 0}$ on $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ given by (7). Since $x_t \to x^\star$ almost surely and $x^\star \in \operatorname{int} \mathcal{X}$, it follows that with probability 1, the vector ζ_t vanishes for all but finitely many t. Thus, condition (iv) of Assumption B.1 holds, and the verification of Assumption B.1 is complete.

Next, we establish Assumption B.3 by leveraging the joint smoothness condition of Assumption 4 to compute the Jacobian of R and show that $\min_j \operatorname{Re} \lambda_j(\nabla R(x^\star)) \ge \alpha - \gamma\beta > 0$. In the following lemma, the symbol $\nabla G_x(x)$ denotes the derivative of $G_x(\cdot)$ evaluated at x.

Lemma 4.7 (Positivity of the Jacobian)**.** *On a neighborhood of $x^\star$, the map R is smooth with Lipschitz Jacobian satisfying*

$$\nabla R(x) = \nabla G_x(x) + \frac{d}{dy} \mathop{\mathbb{E}}_{z\sim\mathcal{D}(y)}[G(x, z)]\Big|_{y=x} \tag{9}$$

and hence

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

In particular, the real part of every eigenvalue of $\nabla R(x)$ is no smaller than $\alpha - \gamma\beta$.

Proof. By the joint smoothness condition of Assumption 4, there exists a neighborhood U around $x^\star$ in $\mathcal{X}$ such that the map $(x, y) \mapsto G_x(y)$ is smooth on $U \times U$ with Lipschitz Jacobian. Thus, the chain rule reveals that R is smooth on U with Lipschitz Jacobian (9). Now let $x \in U$ and $v \in \mathbf{S}^{d-1}$. Strong monotonicity of $G_x(\cdot)$ implies $\langle \nabla G_x(x)v, v\rangle \ge \alpha$. Moreover, Lemma 3.1 shows that the map

$y \mapsto G_y(x)$ is $\gamma\beta$-Lipschitz on $\mathcal{X}$, and therefore

$$\left|\left\langle \left(\frac{d}{dy}G_y(x)\Big|_{y=x}\right)v, v\right\rangle\right| \le \gamma\beta.$$

Continuing with (9), we conclude

$$\langle \nabla R(x)v, v\rangle \ge \alpha - \gamma\beta \qquad \text{for all } v \in \mathbf{S}^{d-1}. \tag{10}$$

Finally, let $w \in \mathbf{C}^d$ be a normalized eigenvector of $\nabla R(x)$ with associated eigenvalue $\lambda \in \mathbf{C}$. The inequality (10) implies $\lambda_{\min}(\nabla R(x) + \nabla R(x)^\top) \ge 2(\alpha - \gamma\beta)$. Letting w^{H} denote conjugate transpose of w, we conclude

$$2(\alpha-\gamma\beta) \le w^{\mathrm{H}}(\nabla R(x) + \nabla R(x)^\top)w = w^{\mathrm{H}}\nabla R(x)w + (w^{\mathrm{H}}\nabla R(x)w)^{\mathrm{H}} = \lambda + \bar{\lambda} = 2(\mathrm{Re}\,\lambda),$$

where the first inequality follows from the Rayleigh-Ritz theorem. □

Note that since ∇R is Lipschitz on a neighborhood of $x^\star$ and $R(x^\star) = 0$, we have

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

Hence Assumption B.3 holds.

The only item left to prove in order to apply Theorem B.1 is that the step sizes η_t satisfy the required conditions. Recall that we have set $\eta_t = \eta_0 t^{-\nu}$ with $\nu \in (\frac{1}{2}, 1)$. Therefore, we compute

$$0 \le \frac{\eta_t - \eta_{t+1}}{\eta_t^2} = t^\nu \frac{(t+1)^\nu - t^\nu}{\eta_0 (t+1)^\nu} \le \frac{(t+1)^\nu - t^\nu}{\eta_0}.$$

Using the fact that $\lim_{t\to\infty}((t+1)^\nu - t^\nu) = 0$ for any $\nu \in (0,1)$, we obtain $\frac{\eta_t - \eta_{t+1}}{\eta_t^2} \to 0$. Moreover, because $\nu + 1/2 > 1$, we have

$$\sum_{t=1}^\infty \eta_t \cdot t^{-1/2} = \eta_0 \sum_{t=1}^\infty t^{-(\nu+1/2)} < \infty.$$

Thus, the step sizes satisfy the required conditions, and an application of Theorem B.1 to the process (6) completes the proof of Theorem 4.2.

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 [17, 29] on lower bounds for statistical estimation and the more recent work of Duchi and Ruan [8] on asymptotic optimality of the stochastic gradient method. Throughout the section, we fix a 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 primarily focus on lower-bounding the performance of estimation procedures when $\mathcal{D}$ is replaced by a small perturbation $\mathcal{D}'$. To this end, we make the following definition.

Definition 5.1 (Admissible distribution map)**.** A distribution map $\mathcal{D}'\colon \mathcal{X} \to P_1(\mathcal{Z})$ will be called *admissible* if Assumptions 1 and 2 hold for $\mathcal{D}'$ and G. The corresponding equilibrium point will then be denoted by $x^\star_{\mathcal{D}'}$.

We start with some intuition before delving into the details. Roughly speaking, we aim to show that the covariance matrix of the asymptotic distribution in Theorem 4.2 is smaller — in the positive semidefinite order — than the asymptotic covariance achieved by any other algorithm for finding equilibrium points. To capture this ordering, we will probe random vectors with functions $\mathcal{L}\colon \mathbf{R}^d \to [0,\infty)$ that are symmetric, quasiconvex, and lower semicontinuous, saying that $z_1 \sim \mathcal{P}_1$ is "better" than $z_2 \sim \mathcal{P}_2$ if $\mathbb{E}[\mathcal{L}(z_1)] \leq \mathbb{E}[\mathcal{L}(z_2)]$ for all such $\mathcal{L}$. This notion indeed implies the ordering of covariances $\mathbb{E}[z_1 z_1^\top] \preceq \mathbb{E}[z_2 z_2^\top]$.[7] We consider a local notion of complexity that evaluates an estimator's performance on problems close to the one we wish to solve. Since our problem is defined via the distribution map $\mathcal{D}$, we will define a set of close problems through a collection of admissible distribution maps $B_{1/k}$ near $\mathcal{D}$ that shrink as k increases. Every map $\mathcal{D}' \in B_{1/k}$ defines a new problem with equilibrium point $x^\star_{\mathcal{D}'}$. The rest of the section is devoted to showing roughly that for every "estimation procedure" $\widehat{x}_k$ we have

$$\liminf_{k\to\infty} \sup_{\mathcal{D}' \in B_{1/k}} \mathbb{E}[\mathcal{L}(\sqrt{k}(\widehat{x}_k - x^\star_{\mathcal{D}'}))] \geq \mathbb{E}[\mathcal{L}(Z)], \tag{11}$$

where $Z \sim \mathsf{N}(0, W^{-1} \cdot \Sigma \cdot W^{-\top})$ with Σ and W as in Theorem 1.1.

To proceed formally, we need to impose the following assumptions on the base distribution map $\mathcal{D}$ and the map G. Despite being technical, these conditions are very mild and essentially amount to quantifying the smoothness of $\mathcal{D}$ and G. In particular, we will make use of a certain set of test functions to be integrated against $\mathcal{D}(x)$.

Definition 5.2 (Test functions)**.** Given a compact metric space $\mathcal{K}$, we let $\mathcal{T}(\mathcal{K}, \mathcal{Z})$ consist of all bounded measurable functions $\phi\colon \mathcal{K} \times \mathcal{Z} \to \mathbf{R}$ which admit 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.

1. **(Compactness)** The set $\mathcal{X}$ is compact, and the set $\mathcal{Z}$ is bounded.
2. **(Smoothness of the distribution)** There exists a constant $\vartheta \geq 0$ such that for every 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 $\mathcal{X} \times \mathcal{K}$.[8]
3. **(Lipschitz Jacobian)** There exist a measurable function $\Lambda\colon \mathcal{Z} \to [0, \infty)$ and constants $\bar{\Lambda}, \beta' > 0$ 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 a chain of 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 the second condition of Assumption 5 holds with $\vartheta = \max\{\Lambda_p, L_p\}$.

Next, we describe the "estimation procedures" to which our desired lower bound will apply. We assume that the estimator $\widehat{x}_k$ in the lower bound depends on samples $z_0, \ldots, z_{k-1} \in \mathcal{Z}$. Unlike in classical settings, these samples are not necessarily drawn from a fixed distribution. Instead, they

[7] If Σ_1 and Σ_2 are the covariance matrices of z_1 and z_2, respectively, and $\mathcal{L}_u(z) = (u^\top z)^2$, then $\Sigma_1 \preceq \Sigma_2$ if and only if $\mathbb{E}[\mathcal{L}_u(z_1)] \leq \mathbb{E}[\mathcal{L}_u(z_2)]$ for all u.

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

are drawn from distributions $\mathcal{D}(\tilde{x}_0), \ldots, \mathcal{D}(\tilde{x}_{k-1})$ taken along a possibly random sequence of points $\tilde{x}_0, \ldots, \tilde{x}_{k-1}$. In order to model the dynamic nature of the estimation procedure, we introduce the following definition.

Definition 5.3 (Dynamic estimation procedure)**.** We say that a sequence of measurable maps $\mathcal{A}_k \colon \mathcal{Z}^k \times \mathcal{X}^k \to \mathcal{X}$ is a *dynamic estimation procedure* if for any admissible distribution map $\mathcal{D}'$ and initial point $\tilde{x}_0 \in \mathcal{X}$, the recursively defined sequences

$$z_k \sim \mathcal{D}'(\tilde{x}_k) \qquad \text{and} \qquad \tilde{x}_{k+1} = \mathcal{A}_{k+1}(z_0, \ldots, z_k, \tilde{x}_0, \ldots, \tilde{x}_k) \tag{12}$$

are such that $\tilde{x}_k$ converges to $x^\star_{\mathcal{D}'}$ almost surely.

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 the distribution map $\mathcal{D}'$. Importantly, $\mathcal{A}_k$ is assumed to be a deterministic function of its arguments. In particular, Proposition 4.1 shows that maps $\mathcal{A}_k$ corresponding to Algorithm 1, i.e.,

$$\mathcal{A}_{k+1}(z_0, \ldots, z_k, x_0, \ldots, x_k) = \mathrm{proj}_{\mathcal{X}}(x_k - \eta_k G(x_k, z_k)),$$

furnish a dynamic estimation procedure under standard conditions on η_k. Although this map $\mathcal{A}_{k+1}$ depends only on the last iterate x_k and the last sample z_k, general dynamic estimation procedures may depend arbitrarily on the previous samples and iterates.

Next, we define the neighborhoods $B_{1/k}$ of $\mathcal{D}$ in (11). These neighborhoods will be defined based on an f-divergence induced by any C^3-smooth convex function $f \colon (0, \infty) \to \mathbf{R}$ with $f(1) = 0$. For any distribution map $\mathcal{D}'$, we define the similarity measure

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

where $\Delta_f(\mathcal{D}'(x) \parallel \mathcal{D}(x))$ denotes the usual f-divergence between the probability measures $\mathcal{D}'(x)$ and $\mathcal{D}(x)$, defined in (3). We then define the admissible neighborhood

$$B_\varepsilon := \{\mathcal{D}' \colon \mathcal{X} \to P_1(\mathcal{Z}) \mid \mathcal{D}' \text{ is admissible and } \Delta_f(\mathcal{D}' \parallel \mathcal{D}) \le \varepsilon\}.$$

Equipped with these tools, we are ready to state the main result of this section.

Theorem 5.4 (Asymptotic optimality)**.** *Suppose that Assumptions 1, 2, and 5 hold and let* $\{\mathcal{A}_k\}$ *be a dynamic estimation procedure. Fix an initial point* $\tilde{x}_0 \in \mathcal{X}$*, and for any admissible distribution map* $\mathcal{D}'$*, let* $\mathbb{E}_{P'_k}[\cdot]$ *denote the expectation with respect to* $z_0, \ldots, z_{k-1}$ *produced by the recursion (12). Let* $\mathcal{L} \colon \mathbf{R}^d \to [0, \infty)$ *be symmetric, quasiconvex, and lower semicontinuous, and let* $\widehat{x}_k \colon \mathcal{Z}^k \to \mathbf{R}^d$ *be a sequence of estimators. Then for any* $c > 0$*, the inequality*

$$\liminf_{k \to \infty} \sup_{\mathcal{D}' \in B_{c/k}} \mathbb{E}_{P'_k}[\mathcal{L}(\sqrt{k}(\widehat{x}_k - x^\star_{\mathcal{D}'}))] \ge \mathbb{E}[\mathcal{L}(Z)]$$

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

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

Most importantly, observe that the distribution of Z in Theorem 5.4 coincides with the asymptotic distribution of $\sqrt{t}(\bar{x}_t - x^\star)$ in Theorem 4.2, thereby justifying the asymptotic optimality of the stochastic forward backward algorithm.

5.1 Proof of Theorem 5.4

Our strategy for proving Theorem 5.4 follows a similar general pattern as in [8] and crucially relies on the Hájek-Le Cam minimax theorem. Throughout the section, we impose the assumptions and notation of Theorem 5.4. In order to simplify notation, we will use the shorthand $\mathcal{D}_x = \mathcal{D}(x)$.

5.1.1 Tilted distributions

The proof proceeds by constructing "tilt perturbations" of $\mathcal{D}$, in the language of [8], which are contained in $B_{1/k}$ and realize the claimed lower bound of Theorem 5.4. These distributions encode the difficult instances near the target problem. The tilted distributions will be parametrized by a function $h\colon \mathbf{R} \to [-1, 1]$, a map $g\colon \mathcal{X} \times \mathcal{Z} \to \mathbf{R}^d$, and a vector $u \in \mathbf{R}^d$. We start by defining the relevant class of maps $g \in \mathcal{G}$.

Definition 5.5. Let $\mathcal{G}$ consist of all measurable maps $g\colon \mathcal{X} \times \mathcal{Z} \to \mathbf{R}^d$ satisfying the following.

1. **(Unbiased)** $\mathbb{E}_{z\sim\mathcal{D}(x)}\big[g(x, z)\big] = 0$ for all $x \in \mathcal{X}$.
2. **(Lipschitz)** There exists a constant $\beta_g > 0$ such that for every $x \in \mathcal{X}$, the section $g(x, \cdot)$ is β_g-Lipschitz on $\mathcal{Z}$.
3. **(Smoothness)** There exist a measurable function $\Lambda_g\colon \mathcal{Z} \to [0, \infty)$ and constants $\bar{\Lambda}_g, \beta'_g > 0$ 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 $\mathbb{E}_{z\sim\mathcal{D}(x)}\big[\Lambda_g(z)\big] \le \bar{\Lambda}_g$, and the section $\nabla_x g(x, \cdot)$ is β'_g-Lipschitz on $\mathcal{Z}$.

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

$$\xi(x, z) := G(x, z) - G_x(x),$$

which belongs to $\mathcal{G}$ as a consequence of Assumptions 2 and 5 and Lemma C.4. For any $g \in \mathcal{G}$, we will employ the notation $g_x(z) := g(x, z)$.

Next, we fix $g \in \mathcal{G}$ and an arbitrary C^3-smooth function $h\colon \mathbf{R} \to [-1, 1]$ such that its first three derivatives are bounded and $h(t) = t$ for all $t \in [-1/2, 1/2]$. Given $u \in \mathbf{R}^d$, we define a new map $\mathcal{D}^u\colon \mathcal{X} \to P_1(\mathcal{Z})$ that maps x to the probability distribution $\mathcal{D}^u_x$ given by

$$d\mathcal{D}^u_x(z) := \frac{1 + h(u^\top g_x(z))}{C^u_x}\, d\mathcal{D}_x(z), \tag{13}$$

where C^u_x is the normalizing constant $C^u_x := 1 + \int h(u^\top g_x(z))\, d\mathcal{D}_x(z)$. Reassuringly, for all sufficiently small u, the perturbed distribution map $\mathcal{D}^u$ is admissible. This is the content of the following lemma, whose proof appears in Appendix C.1.

Lemma 5.6 (Perturbed distributions are admissible)**.** *There exists a convex neighborhood $\mathcal{U}$ of zero such that for all $u \in \mathcal{U}$, the map $\mathcal{D}^u$ is admissible.*[9]

With this lemma in mind, we set $x^\star_u := x^\star_{\mathcal{D}^u}$ for all $u \in \mathcal{U}$. For each $u \in \mathbf{R}^d$, we define the sequence of perturbed distributions $P_{k,u}$ on $\mathcal{Z}^k$ through the recursion

$$z_i \sim \mathcal{D}^u_{\tilde{x}_i} \qquad \text{and} \qquad \tilde{x}_{i+1} = \mathcal{A}(z_0, \ldots, z_i, \tilde{x}_0, \ldots, \tilde{x}_i) \qquad \text{for all } i \in \{0, \ldots, k-1\}.$$

Upon taking $g = \xi$, this collection of tilted distributions achieves the desired lower bound.

Theorem 5.7. *Let $\mathcal{L}\colon \mathbf{R}^d \to [0, \infty)$ be symmetric, quasiconvex, and lower semicontinuous, let $\widehat{x}_k\colon \mathcal{Z}^k \to \mathbf{R}^d$ be a sequence of estimators, and suppose $g = \xi$. Then for any $c > 0$, the inequality*

$$\liminf_{k\to\infty} \sup_{\|u\|\le\sqrt{c/k}} \mathbb{E}_{P_{k,u}}\big[\mathcal{L}(\sqrt{k}(\widehat{x}_k - x^\star_u))\big] \ge \mathbb{E}[\mathcal{L}(Z)] \tag{14}$$

holds, where $Z \sim \mathsf{N}(0, W^{-1} \cdot \Sigma \cdot W^{-\top})$.

[9] We actually show that constants γ^u, $\bar{L}^u$, and α^u for $\mathcal{D}^u$ from Assumptions 1 and 2 deviate from γ, $\bar{L}$, and α by at most $O(\|u\|)$ as $u \to 0$.

Theorem 5.4 follows directly from the above result. To see this, it suffices to show that for any given $c > 0$, there exists $\bar{c} > 0$ such that the condition $\|u\| \leq \sqrt{\bar{c}/k}$ implies $\mathcal{D}^u \in B_{c/k}$ for all sufficiently large k. To this end, we have to check two conditions: first, that $\mathcal{D}^u$ is admissible, and second, that $\Delta_f(\mathcal{D}^u \parallel \mathcal{D}) \leq c/k$. It follows from Lemma 5.6 that for any $\bar{c} > 0$, the condition $\|u\| \leq \sqrt{\bar{c}/k}$ implies that $\mathcal{D}^u$ is admissible for all sufficiently large k. Let us now consider the second condition.

Without loss of generality, we may assume that $f'(1) = 0$, since we can take $\tilde{f}(t) = f(t) - f'(1)(t-1)$ and $\Delta_f = \Delta_{\tilde{f}}$. Then for all $x \in \mathcal{X}$ and sufficiently small $u \in \mathbf{R}^d$, we have

$$
\begin{aligned}
\Delta_f(\mathcal{D}_x^u \parallel \mathcal{D}_x) &= \int f\Big(\frac{1 + h(u^\top g_x(z))}{C_x^u}\Big)\, d\mathcal{D}_x(z) \\
&= \int f\Big(\frac{1 + u^\top g_x(z)}{C_x^u}\Big)\, d\mathcal{D}_x(z) \qquad (15)\\
&= \frac{f''(1)}{2} u^\top \Big(\mathop{\mathbb{E}}_{z\sim\mathcal{D}_x} g_x(z) g_x(z)^\top\Big) u + r_x(u), \qquad (16)
\end{aligned}
$$

where $r_x \colon \mathbf{R}^d \to \mathbf{R}$ is such that $\sup_{x\in\mathcal{X}} |r_x(u)| = o(\|u\|^2)$ as $u \to 0$. The equality (15) follows since $g_x(z)$ is uniformly bounded over $\mathcal{X} \times \mathcal{Z}$ (see Lemma C.1) and hence for small u we have $h(u^\top g_x(z)) = u^\top g_x(z)$ for all $x \in \mathcal{X}$ and $z \in \mathcal{Z}$. The last equality (16) follows by the dominated convergence theorem; we defer the details to Lemma C.3. Since $g_x(z)$ is uniformly bounded, there is a constant $a > 0$ 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, by setting $\bar{c} = c/(\frac{a}{2}f''(1) + b)$, it follows that the condition $\|u\|^2 \leq \bar{c}/k$ implies $\Delta_f(\mathcal{D}^u \parallel \mathcal{D}) \leq c/k$ for all sufficiently large k. Thus, Theorem 5.4 is a corollary of Theorem 5.7. The remainder of this section focuses on establishing Theorem 5.7.

5.1.2 Proof of Theorem 5.7

The proof of this result is based on the classical Hájek-Le Cam minimax theorem. To state this result, we need two standard definitions from statistics. In what follows, we say that a sequence of random vectors X_k in $\mathbf{R}^n$ on probability spaces $(\Omega_k, \mathcal{F}_k, P_k)$ is $o_{P_k}(1)$ if X_k tends to zero in P_k probability, i.e., $\lim_{k\to\infty} P_k\{\|X_k\| \geq \varepsilon\} = 0$ for all $\varepsilon > 0$. We also employ the notation

$$X_k \xrightarrow[P_k]{d} \mathcal{L}_X$$

to indicate that $X_k \xrightarrow{d} X$ where $X \sim \mathcal{L}_X$ and the laws of X_k are taken with respect to P_k.

Definition 5.8 (Locally asymptotically normal)**.** Let $U \subset \mathbf{R}^d$ be a neighborhood of zero. For each $k \in \mathbf{N}$ and $u \in U$, let $Q_{k,u}$ be a probability measure on a measurable space $(\Omega_k, \mathcal{F}_k)$. The family $\{Q_{k,u}\}$ is *locally asymptotically normal (at zero)* with precision $V \succeq 0$ if there exist $Z_k \xrightarrow[Q_{k,0}]{d} \mathsf{N}(0, V)$ such that for every bounded sequence $\{u_k\}$ in U, we have

$$\log \frac{dQ_{k,u_k}}{dQ_{k,0}} = u_k^\top Z_k - \frac{1}{2} u_k^\top V u_k + o_{Q_{k,0}}(1). \qquad (17)$$

Definition 5.9 (Regular mapping sequence)**.** Let $U \subset \mathbf{R}^n$ be a neighborhood of zero. A sequence of mappings $\{\Gamma_k \colon U \to \mathbf{R}^d\}$ is *regular* with derivative $D \in \mathbf{R}^{d\times n}$ if

$$\lim_{k\to\infty} \sqrt{k}(\Gamma_k(u) - \Gamma_k(0)) = Du \qquad \text{for all } u \in U.$$

Equipped with these definitions, we are ready to state the minimax theorem; which appears for example in [17, Theorem 6.6.2], [30, Theorem 3.11.5], and [8, Lemma 8.2].

Theorem 5.10. *Let $\{\Omega_k, \mathcal{F}_k, Q_{k,u}\}_{u\in U}$ be a locally asymptotically normal family with precision $V \succeq 0$ and let $\{\Gamma_k : U \to \mathbf{R}^d\}$ be regular sequence with derivative D. Let $\mathcal{L} : \mathbf{R}^d \to [0, \infty)$ be symmetric, quasiconvex, and lower semicontinuous. Then, for any sequence of estimators $T_k : \Omega_k \to \mathbf{R}^d$, we have*

$$\sup_{U_0\subset U, |U_0|<\infty} \liminf_{k\to\infty} \max_{u\in U_0} \mathbb{E}_{Q_{k,u}}[\mathcal{L}(\sqrt{k}(T_k - \Gamma_k(u)))] \geq \mathbb{E}[\mathcal{L}(Z)], \tag{18}$$

where $Z \sim \mathsf{N}(0, DV^{-1}D^\top)$ when V is invertible; if V is singular, then (18) holds with $Z \sim \mathsf{N}(0, D(V + \lambda I)^{-1}D^\top)$ for all $\lambda > 0$.

We take $\Omega_k = \mathcal{Z}^k$ and set $\mathcal{F}_k = \mathcal{B}(\Omega_k)$ to be its Borel σ-algebra. To prove Theorem 5.7 we will apply Theorem 5.10 to the sequence of probability measures $Q_{k,u} := P_{k,u/\sqrt{k}}$ and regular mapping sequence $\Gamma_k : \mathcal{U} \to \mathbf{R}^d$ given by $u \mapsto x^\star_{u/\sqrt{k}}$, where $\mathcal{U}$ is the neighborhood of zero furnished by Lemma 5.6. We now state two key lemmas that will allow us to apply Theorem 5.10; we defer their proofs to Appendix C.2.

The following lemma shows that Γ_k constitutes a regular mapping sequence.

Lemma 5.11. *The sequence $\{\Gamma_k : \mathcal{U} \to \mathbf{R}^d\}$ is regular with derivative $-W^{-1}\Sigma_{g,G}^\top$, where*

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

Furthermore, the family of distributions $Q_{k,u}$ is locally asymptotically normal.

Lemma 5.12. *The family $\{\Omega_k, \mathcal{F}_k, Q_{k,u}\}_{u\in\mathbf{R}^d}$ is locally asymptotically normal with precision*

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

We now apply Theorem 5.10. Let $\mathcal{L} : \mathbf{R}^d \to [0, \infty)$ be symmetric, quasiconvex, and lower semicontinuous, $\widehat{x}_k : \mathcal{Z}^k \to \mathbf{R}^d$ be a sequence of estimators, and $c > 0$. To connect the inequality (18) to (14), take $U = \{u \in \mathcal{U} \mid \|u\| \leq \sqrt{c}\}$ and observe that for any finite subset $U_0 \subset U$, we have

$$\liminf_{k\to\infty} \sup_{\|u\|\leq\sqrt{c/k}} \mathbb{E}_{P_{k,u}}[\mathcal{L}(\sqrt{k}(\widehat{x}_k - x^\star_u))] \geq \liminf_{k\to\infty} \max_{u\in U_0} \mathbb{E}_{Q_{k,u}}[\mathcal{L}(\sqrt{k}(\widehat{x}_k - \Gamma_k(u)))].$$

Taking the supremum over all finite $U_0 \subset U$ and applying Theorem 5.10 with Lemmas 5.11 and 5.12 with $g = \xi$ (noting $\Sigma_{\xi,G} = \Sigma_\xi = \Sigma$) yields

$$\liminf_{k\to\infty} \sup_{\|u\|\leq\sqrt{c/k}} \mathbb{E}_{P_{k,u}}[\mathcal{L}(\sqrt{k}(\widehat{x}_k - x^\star_u))] \geq \mathbb{E}[\mathcal{L}(Z_\lambda)] \tag{19}$$

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 (19) establishes (14). Indeed, let $\Sigma = AA^\top$ be a Cholesky decomposition of Σ and observe that the pseudoinverse identities $A^\dagger = \lim_{\lambda\downarrow 0} A^\top(AA^\top + \lambda I)^{-1}$ and $AA^\dagger A = A$ imply

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

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 $0 < \lambda_1 \leq \lambda_2$, we have $\widetilde{\Sigma}_{\lambda_1} \succeq \widetilde{\Sigma}_{\lambda_2}$ and hence $\exp(-\frac{1}{2}u^\top\widetilde{\Sigma}^\dagger_{\lambda_1}u) \geq \exp(-\frac{1}{2}u^\top\widetilde{\Sigma}^\dagger_{\lambda_2}u)$ for all $u \in \mathbf{R}^d$. 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

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

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

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

Hence (19) entails (14) and the proof of Theorem 5.7 is complete.

References

[1] S. Ahmed. *Strategic planning under uncertainty: Stochastic integer programming approaches.* University of Illinois at Urbana-Champaign, 2000.

[2] Y. Bechavod, K. Ligett, Z. S. Wu, and J. Ziani. Causal feature discovery through strategic modification. *arXiv:2002.07024*, 2020.

[3] G. Brown, S. Hod, and I. Kalemaj. Performative prediction in a stateful world. *arXiv preprint arXiv:2011.03885*, 2020.

[4] J. Cutler, D. Drusvyatskiy, and Z. Harchaoui. Stochastic optimization under time drift: iterate averaging, step-decay schedules, and high probability guarantees. *Advances in Neural Information Processing Systems*, 34, 2021.

[5] D. Davis, D. Drusvyatskiy, and L. Jiang. Subgradient methods near active manifolds: saddle point avoidance, local convergence, and asymptotic normality. *arXiv preprint arXiv:2108.11832*, 2021.

[6] J. Dong, A. Roth, Z. Schutzman, B. Waggoner, and Z. S. Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 55–70, 2018.

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

[8] J. C. Duchi and F. Ruan. Asymptotic optimality in stochastic optimization. *The Annals of Statistics*, 49(1):21–48, 2021.

[9] M. Duflo. *Random iterative models*, volume 34. Berlin: Springer, 1997. ISBN 3-540-57100-0.

[10] J. Dupacová. Optimization under exogenous and endogenous uncertainty. *University of West Bohemia in Pilsen*, 2006.

[11] W. Feller. *An introduction to probability theory and its applications. Vol II. 2nd ed.* Wiley Ser. Probab. Math. Stat. John Wiley & Sons, Hoboken, NJ, 1971.

[12] M. Hardt, N. Megiddo, C. H. Papadimitriou, and M. Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pages 111–122. ACM, 2016.

[13] L. Hellemo, P. I. Barton, and A. Tomasgard. Decision-dependent probabilities in stochastic programs with recourse. *Computational Management Science*, 15(3):369–395, 2018.

[14] Z. Izzo, L. Ying, and J. Zou. How to learn when data reacts to your model: performative gradient descent. In *International Conference on Machine Learning*, pages 4641–4650. PMLR, 2021.

[15] M. Jagadeesan, T. Zrnic, and C. Mendler-Dünner. Regret minimization with performative feedback. *arXiv preprint arXiv:2202.00628*, 2022.

[16] T. W. Jonsbråten, R. J. Wets, and D. L. Woodruff. A class of stochastic programs withdecision dependent random elements. *Annals of Operations Research*, 82:83–106, 1998.

[17] L. Le Cam, L. M. LeCam, and G. L. Yang. *Asymptotics in statistics: some basic concepts*. Springer Science & Business Media, 2000.

[18] C. Mendler-Dünner, J. Perdomo, T. Zrnic, and M. Hardt. Stochastic optimization for performative prediction. In *Advances in Neural Information Processing Systems*, volume 33, pages 4929–4939. Curran Associates, Inc., 2020.

[19] C. Mendler-Dünner, J. C. Perdomo, T. Zrnic, and M. Hardt. Stochastic optimization for performative prediction. *arXiv preprint arXiv:2006.06887*, 2020.

[20] J. Miller, S. Milli, and M. Hardt. Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning*, pages 6917–6926. PMLR, 2020.

[21] J. P. Miller, J. C. Perdomo, and T. Zrnic. Outside the echo chamber: Optimizing the performative risk. In *International Conference on Machine Learning*, pages 7710–7720. PMLR, 2021.

[22] A. Narang, E. Faulkner, D. Drusvyatskiy, M. Fazel, and L. J. Ratliff. Multiplayer performative prediction: Learning in decision-dependent games. *arXiv preprint arXiv:2201.03398*, 2022.

[23] J. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. 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.

[24] G. Piliouras and F.-Y. Yu. Multi-agent performative prediction: From global stability and optimality to chaos. *arXiv preprint arXiv:2201.10483*, 2022.

[25] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992. ISSN 0363-0129. doi: 10.1137/0330046.

[26] M. Ray, L. J. Ratliff, D. Drusvyatskiy, and M. Fazel. Decision-dependent risk minimization in geometrically decaying dynamic environments. *Proceedings of the AAAI International Conference on Artificial Intelligence (AAAI)*, 2022.

[27] H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. 1971.

[28] R. Y. Rubinstein and A. Shapiro. Sensitivity analysis and stochastic optimization by the score function method, 1993.

[29] A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[30] A. W. van der Vaart and J. A. Wellner. *Convolution and Minimax Theorems*, pages 412–422. Springer New York, New York, NY, 1996. ISBN 978-1-4757-2545-2. doi: 10.1007/978-1-4757-2545-2_37. URL https://doi.org/10.1007/978-1-4757-2545-2_37.

[31] P. Varaiya and R.-B. Wets. Stochastic dynamic optimization approaches and computation. 1988.

[32] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[33] K. Wood and E. Dall'Anese. Stochastic saddle point problems with decision-dependent distributions. *arXiv preprint arXiv:2201.02313*, 2022.

[34] K. Wood, G. Bianchin, and E. Dall'Anese. Online projected gradient descent for stochastic optimization with decision-dependent distributions. *IEEE Control Systems Letters*, 6:1646–1651, 2021.

A Proofs from Sections 3 and 4

A.1 Proof of Lemma 3.1

We successively estimate

$$\begin{aligned}\|G_x(y) - G_{x'}(y)\| &= \left\| \mathop{\mathbb{E}}_{z\sim\mathcal{D}(x)} G(y,z) - \mathop{\mathbb{E}}_{z\sim\mathcal{D}(x')} G(y,z) \right\| \\ &= \sup_{\|v\|\le 1} \left\{ \mathop{\mathbb{E}}_{z\sim\mathcal{D}(x)} \langle G(y,z), v\rangle - \mathop{\mathbb{E}}_{z\sim\mathcal{D}(x')} \langle G(y,z), v\rangle \right\} \\ &\le \beta \cdot W_1(\mathcal{D}(x), \mathcal{D}(x')) \qquad (20) \\ &\le \beta\gamma \cdot \|x - x'\|,\end{aligned}$$

where inequality (20) uses that the map $z \mapsto \langle G(y,z), v\rangle$ is β-Lipschitz continuous and the definition of the W_1 distance. The proof is complete.

A.2 Proof of Theorem 3.3

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

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

Strong monotonicity therefore ensures

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

where the last inequality follows from Lemma 3.1. Dividing through by $\alpha\|y - y'\|$ 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 4.1

We will use the following classical convergence result [9, Theorem 1.3.12].

Lemma A.1 (Robbins-Siegmund)**.** *Let $(A_t), (B_t), (C_t), (D_t)$ be sequences of nonnegative finite 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*

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

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

Set $\Delta_t = x_t - x^\star$ and $\bar{\alpha} = \alpha - \gamma\beta$; note that $\bar{\alpha} > 0$ by Assumption 2. Let v_t and ξ_t be the direction of motion and noise vectors given by (5), and let $(\mathcal{F}_t)$ be the filtration given by (8). Observe that Lemma 3.1 and Assumption 3 immediately yield the bias/variance bounds

$$\|\mathbb{E}_t[v_t] - G_{x^\star}(x_t)\| \leq \gamma\beta\|x_t - x^\star\| \qquad \text{and} \qquad \mathbb{E}_t\|\xi_t\|^2 \leq K(1 + \|x_t - x^\star\|^2), \tag{21}$$

where $\mathbb{E}_t[\cdot]$ denotes the conditional expectation with respect to $\mathcal{F}_t$. In light of (21) and the fact that $\eta_t \to 0$ (since $\sum_t \eta_t^2 < \infty$), we see that for all sufficiently large t, the result [22, Theorem 8 (Benign bias)] directly implies the one-step improvement bound

$$\mathbb{E}_t\|\Delta_{t+1}\|^2 \leq \|\Delta_t\|^2 + 4K\eta_t^2 - \tfrac{1}{2}\bar{\alpha}\eta_t\|\Delta_t\|^2. \tag{22}$$

We now apply Lemma A.1 with $A_t = \|\Delta_t\|^2$, $B_t = 0$, $C_t = 4K\eta_t^2$, and $D_t = \frac{1}{2}\bar{\mu}\eta_t\|\Delta_t\|^2$. Notice that the event $\{\sum_t B_t < \infty, \sum_t C_t < \infty\}$ holds since $\sum_t \eta_t^2 < \infty$. Thus, there exists a nonnegative finite random variable A_∞ such that $A_t \to A_\infty$ and $\sum_t D_t < \infty$ almost surely. Hence $\|\Delta_t\|^2 \to A_\infty$ and $\sum_t \eta_t\|\Delta_t\|^2 < \infty$ almost surely. Since $\sum_t \eta_t$ diverges, we conclude that $\lim_t \|\Delta_t\|^2 = A_\infty = 0$ almost surely.

Next, to establish the in-expectation rate, we note that by (22) and the tower rule, we have

$$\mathbb{E}\|\Delta_{t+1}\|^2 \leq (1 - \tfrac{1}{2}\bar{\alpha}\eta_t)\mathbb{E}\|\Delta_t\|^2 + 4K\eta_t^2$$

for all sufficiently large t. Thus, upon supposing $\eta_t = \eta_0 t^{-\nu}$ with $\nu \in (\frac{1}{2}, 1)$, a standard inductive argument (see [5, Lemma 3.11.8]) yields a constant $C \geq 0$ such that $\mathbb{E}\|\Delta_t\|^2 \leq C\eta_t$ for all $t \geq 0$. This completes the proof.

B Review of asymptotic normality

In this section, we briefly review the asymptotic normality results of Polyak and Juditsky [25]. Consider a set $\mathcal{X} \subset \mathbf{R}^d$ and a map $R\colon \mathcal{X} \to \mathbf{R}^d$. Suppose that there exists a solution $x^\star \in \operatorname{int}\mathcal{X}$ 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

$$x_{t+1} = x_t - \eta_t(R(x_t) + \xi_t + \zeta_t), \tag{23}$$

where η_t is a deterministic positive step size and ξ_t and ζ_t are random vectors in $\mathbf{R}^d$ such that $x_t \in \mathcal{X}$ for all t. We impose the following three assumptions.

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

(i) There is a constant $\bar{K} \geq 0$ such that
$$\mathbb{E}\big[\|\xi_t\|^2 \mid \mathcal{F}_t\big] + \|R(x_t)\|^2 \leq \bar{K}(1 + \|x_t\|^2) \qquad \text{for all } t \geq 0.$$

(ii) The variables $\sup_{t\geq 0} \mathbb{E}\big[\|\xi_t\|^2 \mathbf{1}\{\|\xi_t\| > N\} \mid \mathcal{F}_t\big]$ converge to zero in probability as $N \to \infty$.

(iii) The sequence $\mathbb{E}\big[\xi_t\xi_t^\top \mid \mathcal{F}_t\big]$ converges to a matrix $\Sigma \succeq 0$ in probability as $t \to \infty$.

(iv) With probability 1, the vector ζ_t vanishes for all but finitely many t.

Assumption B.2 (Convergence)**.** The iterates x_t converge to $x^\star$ almost surely, and there exists a constant $C \geq 0$ such that

$$\mathbb{E}\|x_t - x^\star\|^2 \leq C\eta_t \qquad \text{for all } t \geq 0.$$

Assumption B.3 (Smoothness)**.** The map R is differentiable at $x^\star$, and the Jacobian $\nabla R(x^\star)$ satisfies

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

Further, the real parts of the eigenvalues of $\nabla R(x^\star)$ are positive, i.e., $\min_j \operatorname{Re} \lambda_j(\nabla R(x^\star)) > 0$.

Theorem B.1 ([25, Theorem 2])**.** *Suppose that Assumptions B.1, B.2, and B.3 hold. Further, suppose that the step sizes satisfy* $\frac{\eta_t - \eta_{t+1}}{\eta_t^2} \to 0$ *and* $\sum_{t=1}^{\infty} \eta_t \cdot t^{-1/2} < \infty$*. Then the average iterates* $\bar{x}_t = \frac{1}{t}\sum_{i=0}^{t-1} x_i$ *converge to* $x^\star$ *almost surely, and*

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

We remark that the assumptions of Theorem B.1 differ slightly from those in [25]. The proof of the theorem follows by slightly modifying the arguments in [25]. We provide a quick sketch of the small modifications that are needed. There are two differences between the assumptions of Theorem B.1 and the one in [25]: (a) Assumption B.2 replaces a Lyapunov function assumption (Assumption 3.1 in [25]) and (b) our process (23) includes the additional term ζ_t. The proof of [25, Theorem 2] is divided into four parts:

Parts 1 and 2: use the existence of a Lyapunov function to establish almost sure convergence of x_t to $x^\star$ and to derive an in-expectation rate of convergence. Both of these conclusions follow directly from Assumption B.2.

Part 3: considers a linearized version of the stochastic process (23) — which we dub x_t^1 — and uses a martingale central limit theorem to show that it has the desired asymptotic distribution. To do so, the error $\sqrt{t}(\bar{x}_t^1 - x^\star)$ is decomposed into three sums, two of which tend to zero in mean square and the third realizing the stated asymptotic normal distribution. The addition of ζ_t is straightforward: it results in extra sums involving ζ_t in the error decomposition, and these additional sums tend to zero almost surely by virtue of condition (iv) of Assumption B.1.

Part 4: shows that the linearized stochastic process x_t^1 and (23) are asymptotically equivalent. This step follows without changes since it is solely based on the in-expectation rate obtained in Parts 1 and 2 and an application of Kronecker's lemma.

C Proofs from Section 5

This section presents all the missing proofs 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 $g \in \mathcal{G}$.

Lemma C.1. *We have*

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

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

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

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

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

completes the proof. □

Lemma C.2. *For any $x \in \mathcal{X}$, the function $C_x(u) := 1 + \mathbb{E}_{z\sim\mathcal{D}_x} h(u^\top g_x(z))$ is C^2-smooth on $\mathbf{R}^d$ with $A_h M_g^3$-Lipschitz continuous Hessian, where $A_h := \sup |h'''|$ and $M_g := \sup_{x\in\mathcal{X},\, z\in\mathcal{Z}} \|g_x(z)\|$. Moreover, we have $\sup_{x\in\mathcal{X}} |\mathbb{E}_{z\sim\mathcal{D}_x} h(u^\top g_x(z))| = o(\|u\|^2)$ as $u \to 0$.*

Proof. Given $x \in \mathcal{X}$, the dominated convergence theorem yields

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

and

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

Thus, $C_x(\cdot)$ is C^2-smooth on $\mathbf{R}^d$, and the claimed Lipschitz constant of the Hessian follows trivially.

To prove the final claim of the lemma, recall that $h'(0) = 1$ and $h''(0) = 0$. Therefore

$$\nabla_u \Big(\mathop{\mathbb{E}}_{z\sim\mathcal{D}_x} h(u^\top g_x(z)) \Big)\Big|_{u=0} = \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x} g_x(z) = 0 \qquad \text{and} \qquad \nabla_u^2 \Big(\mathop{\mathbb{E}}_{z\sim\mathcal{D}_x} h(u^\top g_x(z)) \Big)\Big|_{u=0} = 0.$$

Consequently, the second-order Taylor polynomial of $\mathbb{E}_{z\sim\mathcal{D}_x} h(u^\top g_x(z))$ in u based at 0 is identically zero. Lipschitzness of the Hessian therefore implies

$$\Big| \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x} h(u^\top g_x(z)) \Big| \le \frac{A_h M_g^3}{6} \|u\|^3$$

for all $x \in \mathcal{X}$ and $u \in \mathbf{R}^d$. Hence $\sup_{x\in\mathcal{X}} |\mathbb{E}_{z\sim\mathcal{D}_x} h(u^\top g_x(z))| = o(\|u\|^2)$ as $u \to 0$. □

Lemma C.3. *Let $f\colon (0, \infty) \to \mathbf{R}$ be a function that is C^3-smooth around $t = 1$ and satisfies $f(1) = f'(1) = 0$. Then for all $x \in \mathcal{X}$ and sufficiently small $u \in \mathbf{R}^d$, we have*

$$\int f\left(\frac{1 + u^\top g_x(z)}{C_x^u} \right) d\mathcal{D}_x(z) = \frac{f''(1)}{2} u^\top \Big(\mathop{\mathbb{E}}_{z\sim\mathcal{D}_x} g_x(z) g_x(z)^\top \Big) u + r_x(u) \tag{24}$$

with $r_x\colon \mathbf{R}^d \to \mathbf{R}$ satisfying $\sup_{x\in\mathcal{X}} |r_x(u)| = o(\|u\|^2)$ as $u \to 0$.

Proof. Fix $x \in \mathcal{X}$ and define $\varphi_x(u) := \mathbb{E}_{z\sim\mathcal{D}_x} f\big(\frac{1+u^\top g_x(z)}{C_x^u}\big)$. By the dominated convergence theorem,

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

Similarly, the Hessian $\nabla_{uu}^2 \varphi_x(u)$ of φ_x at u is equal to

$$\frac{1}{(C_x^u)^4} \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x} \Bigg[f''\left(\frac{1 + u^\top g_x(z)}{C_x^u} \right) \Big(g_x(z) C_x^u - (1 + u^\top g_x(z)) \nabla_u C_x^u \Big) \Big(g_x(z) C_x^u - (1 + u^\top g_x(z)) \nabla_u C_x^u \Big)^\top$$
$$+ f'\left(\frac{1 + u^\top g_x(z)}{C_x^u} \right) \bigg((C_x^u)^2 \Big(g_x(z) (\nabla_u C_x^u)^\top - (\nabla_u C_x^u) g_x(z)^\top - (1 + u^\top g_x(z)) \nabla_{uu}^2 C_x^u \Big)$$
$$- 2 C_x^u \Big(g_x(z) C_x^u - (1 + u^\top g_x(z)) \nabla_u C_x^u \Big) (\nabla_u C_x^u)^\top \bigg) \Bigg].$$

Thus, φ_x is C^2-smooth. Taking a Taylor expansion at $u = 0$ and using the equalities $C_x^0 = 1$, $\nabla_u C_x^u|_{u=0} = 0$, and $f(1) = f'(1) = 0$ yields (24) with second-order Taylor remainder $r_x\colon \mathbf{R}^d \to \mathbf{R}$.

It remains to verify that $\sup_{x\in\mathcal{X}} |r_x(u)| = o(\|u\|^2)$ as $u \to 0$.

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

C.1 Proof of Lemma 5.6

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. The main tool we will use to infer continuity is the following elementary lemma. Its proof consists of several applications of the dominated convergence theorem and is deferred to Appendix D.

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

1. *There exists $\beta_T > 0$ such that for every $x \in \mathcal{X}$, the section $T(x, \cdot)$ is β_T-Lipschitz on $\mathcal{Z}$.*
2. *There exist a measurable function $\Lambda_T\colon \mathcal{Z} \to [0, \infty)$ and constants $\bar{\Lambda}_T, \beta'_T > 0$ 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 $\mathbb{E}_{z\sim\mathcal{D}_x}[\Lambda_T(z)] \le \bar{\Lambda}_T$, 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. Further, given any compact subset $\mathcal{W} \subset \mathbf{R}^d$, the map $H\colon \mathcal{X} \times \mathcal{W} \to \mathbf{R}^n$ given by

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

is Λ_H-smooth, where Λ_H is a constant depending on T only through $\beta_T, \bar{\Lambda}_T, \beta'_T$, M_T, and M'_T.

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 C.4 to identify a constant L_1 such that for every 1-Lipschitz function $\phi\colon \mathcal{Z} \to \mathbf{R}$ 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 $\gamma^u := \gamma + L_1\|u\|$. Without loss of generality, we may assume that ϕ is bounded by $\mathrm{diam}(\mathcal{Z})$, since we can always shift the function $\phi \leftarrow \phi - \inf \phi$ and this will not change the Lipschitz constant of $\rho_\phi(\cdot, u)$. Then Lemma C.4 yields a constant L_1 independent of ϕ such that ρ_ϕ is L_1-smooth on $\mathcal{X} \times \mathcal{W}$.

To establish the Lipschitz constant of $\rho_\phi(\cdot, u)$, we derive a bound on the gradient of ρ_ϕ. By the triangle inequality, the following holds for all $(x, u) \in \mathcal{X} \times \mathcal{W}$:

$$\begin{aligned}\|\nabla_x \rho_\phi(x, u)\| &\le \|\nabla_x \rho_\phi(x, 0)\| + \|\nabla_x \rho_\phi(x, u) - \nabla_x \rho_\phi(x, 0)\| \\ &\le \gamma + L_1\|u\|.\end{aligned}$$

Therefore $\mathcal{D}^u$ satisfies Assumption 1 with Lipschitz constant $\gamma^u = \gamma + O(\|u\|)$ as $u \to 0$.

Step 2 (Lipschitz) 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 $\bar{L}^u > 0$ 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 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). To this end, observe first that for all $x \in \mathcal{X}$ and $u \in \mathbf{R}^d$, we have

$$\mathop{\mathbb{E}}_{z\sim\mathcal{D}_x^u}[L(z)^2] = \frac{1}{C_x^u} \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x}[L(z)^2(1 + h(u^\top g_x(z)))].$$

Using the Taylor series of $(1+t)^{-1}$ for $|t| < 1$, it follows from Lemma C.2 that

$$\sup_{x\in\mathcal{X}} \frac{1}{C_x^u} = 1 + o(\|u\|^2) \qquad \text{as } u \to 0. \tag{25}$$

Moreover, upon noting that $h(0) = 0$ and the function $u \mapsto h(u^\top g_x(z))$ is Lipschitz on $\mathbf{R}^d$ with constant $(\sup|h'|)M_g$ for each $x \in \mathcal{X}$ and $z \in \mathcal{Z}$, we obtain the following for all $u \in \mathbf{R}^d$:

$$\sup_{x\in\mathcal{X},\, z\in\mathcal{Z}} |h(u^\top g_x(z))| \le (\sup|h'|)M_g\|u\|$$

and hence

$$\sup_{x\in\mathcal{X}} \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x}\Big[L(z)^2(1 + h(u^\top g_x(z)))\Big] \le \bar{L}^2 \cdot (1 + (\sup|h'|)M_g\|u\|).$$

We deduce

$$\sup_{x\in\mathcal{X}} \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x^u}[L(z)^2] \le (1 + o(\|u\|^2)) \cdot \bar{L}^2 \cdot (1 + (\sup|h'|)M_g\|u\|) = \bar{L}^2 + O(\|u\|) \qquad \text{as } u \to 0.$$

Step 3 (Monotonicity) We prove that for all $x \in \mathcal{X}$, the map $G_x^u(\cdot)$ is strongly monotone on $\mathcal{X}$ with constant $\alpha^u = \alpha + o(\|u\|^2)$ 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

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

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 $\ell^u = o(\|u\|^2)$ 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 $\sup_{x\in\mathcal{X},\, z\in\mathcal{Z}} \|\nabla_x G(x,z)\|_{\mathrm{op}} < \infty$ (as in Lemma C.4) and applying the dominated convergence theorem together with (25) and Lemma C.2, we obtain

$$\begin{aligned}\ell^u &:= \sup_{x,y\in\mathcal{X}} \|\nabla_y(G_x^u(y) - G_x(y))\|_{\mathrm{op}} \\ &= \sup_{x,y\in\mathcal{X}} \left\| \frac{1}{C_x^u} \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x}[\nabla_y G(y,z)(1 + h(u^\top g_x(z)))] - \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x}[\nabla_y G(y,z)] \right\|_{\mathrm{op}} \\ &\le \underbrace{\sup_{x,y\in\mathcal{X}} \left\| \frac{1}{C_x^u} \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x}[\nabla_y G(y,z)] - \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x}[\nabla_y G(y,z)] \right\|_{\mathrm{op}}}_{o(\|u\|^2)} + \underbrace{\sup_{x,y\in\mathcal{X}} \left\| \frac{1}{C_x^u} \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x}[\nabla_y G(y,z) h(u^\top g_x(z))] \right\|_{\mathrm{op}}}_{(1+o(\|u\|^2))\cdot o(\|u\|^2)} \\ &= o(\|u\|^2) \qquad \text{as } u \to 0.\end{aligned}$$

Setting $\alpha^u := \alpha - \ell^u$ for all u in a neighborhood of zero, we conclude that for all $x, y, y' \in \mathcal{X}$,

$$\langle G_x^u(y) - G_x^u(y'), y - y'\rangle \ge \alpha^u\|y - y'\|^2$$

and hence $G_x^u(\cdot)$ is strongly monotone on $\mathcal{X}$ with constant $\alpha^u = \alpha + o(\|u\|^2)$ 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, the previous steps show $\gamma^u = \gamma + O(\|u\|)$ and $\alpha^u = \alpha + o(\|u\|^2)$ 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 Lemmas used in the proof of Theorem 5.7

In this section, we collect the proofs of the two key lemmas used in the proof of Theorem 5.7.

C.2.1 Proof of Lemma 5.11

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

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

where we recall $C_x^u = 1 + \mathbb{E}_{z\sim\mathcal{D}_x} h(u^\top g_x(z))$. Lemma C.4 directly implies that F is C^1-smooth. Consider now the family of smooth nonlinear equations

$$F(x,u) = 0 \tag{26}$$

parametrized by $u \in \mathbf{R}^d$. Observe that $F(x^\star, 0) = 0$ since $x^\star \in \operatorname{int}\mathcal{X}$. More generally, the equality (26) with $x \in \operatorname{int}\mathcal{X}$ holds precisely when x is equal to $x_u^\star$. We will apply the implicit function theorem to $x_u^\star$ as a smooth implicit function of u. To this end, we compute the the partial Jacobian

$$\nabla_x F(x^\star, 0) = \nabla_x \Big(\mathop{\mathbb{E}}_{z\sim\mathcal{D}_x} G(x,z)\Big)\Big|_{x=x^\star} = W.$$

Lemma 4.7 therefore implies that $\nabla_x F(x^\star, 0)$ is invertible. Consequently, the implicit function theorem yields open sets $V \subset \operatorname{int}\mathcal{X}$ and $U \subset \mathbf{R}^d$ such that $x^\star \in V$, $0 \in U$, and for each $u \in U$ there exists a unique point $x_u^\star \in V$ satisfying $F(x_u^\star, u) = 0$. Further, the map $u \mapsto x_u^\star$ is C^1-smooth at $u = 0$ with Jacobian $-W^{-1}\nabla_u F(x^\star, 0)$. This yields the first-order Taylor expansion

$$x_u^\star = x^\star - W^{-1}\nabla_u F(x^\star, 0)u + o(\|u\|) \qquad \text{as } u \to 0. \tag{27}$$

To compute $\nabla_u F(x^\star, 0)$, we apply the quotient rule and the dominated convergence theorem to obtain

$$\nabla_u F(x,u) = \frac{1}{(C_x^u)^2} \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x} \big[C_x^u h'(u^\top g_x(z)) G(x,z) g_x(z)^\top - (1 + h(u^\top g_x(z))) G(x,z)(\nabla_u C_x^u)^\top\big]$$

for all $x \in \mathcal{X}$ and $u \in \mathbf{R}^d$. Then the equalities $C_x^0 = 1$, $h'(0) = 1$, and

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

imply

$$\nabla_u F(x, 0) = \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x} \big[G(x,z) g_x(z)^\top\big]$$

for all $x \in \mathcal{X}$. In particular, $\nabla_u F(x^\star, 0) = \Sigma_{g,G}^\top$. Thus, the expansion (27) becomes

$$x_u^\star = x^\star - W^{-1}\Sigma_{g,G}^\top u + o(\|u\|) \qquad \text{as } u \to 0.$$

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

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

The proof is complete.

C.2.2 Proof of Lemma 5.12

We start by deriving the expression

$$dQ_{k,u}(z_0, \ldots, z_{k-1}) = d\mathcal{D}_{\tilde{x}_{k-1}}^{u/\sqrt{k}}(z_{k-1}) \cdots d\mathcal{D}_{\tilde{x}_0}^{u/\sqrt{k}}(z_0). \tag{28}$$

To see this, let $B \in \mathcal{B}(\mathcal{Z}^k)$ be a Borel measurable set. By repeatedly applying the tower rule, we get

$$\begin{aligned}
\int \mathbf{1}_B \, dQ_{k,u} &= \mathbb{E}_{P_{k,u/\sqrt{k}}}[\mathbf{1}_B] \\
&= \mathop{\mathbb{E}}_{z_0 \sim \mathcal{D}_{\tilde{x}_0}^{u/\sqrt{k}}} \big[\mathbb{E}[\mathbf{1}_B \mid z_0]\big] \\
&= \int \mathbb{E}[\mathbf{1}_B \mid z_0] \, d\mathcal{D}_{\tilde{x}_0}^{u/\sqrt{k}}(z_0) \\
&= \int \mathop{\mathbb{E}}_{z_1 \sim \mathcal{D}_{\tilde{x}_1}^{u/\sqrt{k}}} \big[\mathbb{E}[\mathbf{1}_B \mid z_0, z_1]\big] \, d\mathcal{D}_{\tilde{x}_0}^{u/\sqrt{k}}(z_0) \\
&= \int \mathbb{E}[\mathbf{1}_B \mid z_0, z_1] \, d\mathcal{D}_{\tilde{x}_1}^{u/\sqrt{k}}(z_1) \, d\mathcal{D}_{\tilde{x}_0}^{u/\sqrt{k}}(z_0) \\
&\;\;\vdots \\
&= \int \mathbf{1}_B \, d\mathcal{D}_{\tilde{x}_{k-1}}^{u/\sqrt{k}}(z_{k-1}) \cdots d\mathcal{D}_{\tilde{x}_1}^{u/\sqrt{k}}(z_1) \, d\mathcal{D}_{\tilde{x}_0}^{u/\sqrt{k}}(z_0)
\end{aligned}$$

using that $\tilde{x}_j$ is a deterministic function of the samples $(z_0, \ldots, z_{j-1})$.

Now let $\{u_k\}$ be any bounded sequence in $\mathbf{R}^d$ and observe that (28) together with (13) yield

$$\log \frac{dQ_{k,u_k}}{dQ_{k,0}}(z_0, \ldots, z_{k-1}) = -\sum_{j=0}^{k-1} \log C_{\tilde{x}_j}^{u_k/\sqrt{k}} + \sum_{j=0}^{k-1} \log\Big(1 + h(u_k^\top g_{\tilde{x}_j}(z_j)/\sqrt{k})\Big). \tag{29}$$

Setting $e_x(u) := \mathbb{E}_{z \sim \mathcal{D}_x} h(u^\top g_x(z))$ gives

$$C_x^u = 1 + e_x(u),$$

and Lemma C.2 shows $\sup_{x \in \mathcal{X}} |e_x(u)| = o(\|u\|^2)$ as $u \to 0$. Using the fact that $\log(1+t) = t + o(t)$ as $t \to 0$, we derive

$$\sum_{j=0}^{k-1} \log C_{\tilde{x}_j}^{u_k/\sqrt{k}} = \sum_{j=0}^{k-1} \Big(e_{\tilde{x}_j}(u_k/\sqrt{k}) + o(e_{\tilde{x}_j}(u_k/\sqrt{k}))\Big) = k \cdot o\Big(\frac{\|u_k\|^2}{k}\Big) \to 0 \qquad \text{as } k \to \infty. \tag{30}$$

We can therefore ignore the first term in (29).

Next, note that the boundedness of $\{u_k\}$ and g (Lemma C.1) implies that $\sup_{j,k} |u_k^\top g_{\tilde{x}_j}(z_j)| < \infty$. Therefore $h(u_k^\top g_{\tilde{x}_j}(z_j)/\sqrt{k}) = u_k^\top g_{\tilde{x}_j}(z_j)/\sqrt{k} \in [-1/2, 1/2]$ for all sufficiently large k. Then, using the second-order mean value theorem for $\log(1+t)$ around zero, we have

$$\sum_{j=0}^{k-1} \log\Big(1 + \frac{u_k^\top g_{\tilde{x}_j}(z_j)}{\sqrt{k}}\Big) = u_k^\top \Big(\frac{1}{\sqrt{k}} \sum_{j=0}^{k-1} g_{\tilde{x}_j}(z_j)\Big) - \frac{1}{2} u_k^\top \Big(\frac{1}{k} \sum_{j=0}^{k-1} g_{\tilde{x}_j}(z_j) g_{\tilde{x}_j}(z_j)^\top\Big) u_k + \sum_{j=0}^{k-1} \tau_j \frac{(u_k^\top g_{\tilde{x}_j}(z_j))^3}{3k^{3/2}} \tag{31}$$

where the coefficients τ_j satisfy $|\tau_j| \le 8$ for all j. The first term on the right-hand side matches the corresponding term in (17) upon setting $Z_k := k^{-1/2} \sum_{j=0}^{k-1} g_{\tilde{x}_j}(z_j)$. Before establishing the distributional convergence of Z_k, let us study the last two terms of (31).

We start with the middle term on the right-hand side of (31); we claim

$$\frac{1}{k}\sum_{j=0}^{k-1} g_{\tilde{x}_j}(z_j)g_{\tilde{x}_j}(z_j)^\top = \Sigma_g + o_{Q_{k,0}}(1) \tag{32}$$

as a consequence of Theorem E.1. Indeed, let $\mathbb{F} = (\mathcal{F}_k)$ be the filtration given by $\mathcal{F}_k = \sigma(z_0, \ldots, z_{k-1})$ where $(z_0, \ldots, z_{k-1}) \sim Q_{k,0}$, and for each $j \geq 0$, set

$$\begin{aligned} X_{j+1} &= g_{\tilde{x}_j}(z_j)g_{\tilde{x}_j}(z_j)^\top - \mathbb{E}[g_{\tilde{x}_j}(z_j)g_{\tilde{x}_j}(z_j)^\top \mid \mathcal{F}_j] \\ &= g_{\tilde{x}_j}(z_j)g_{\tilde{x}_j}(z_j)^\top - \mathop{\mathbb{E}}_{z_j\sim\mathcal{D}_{\tilde{x}_j}}[g_{\tilde{x}_j}(z_j)g_{\tilde{x}_j}(z_j)^\top], \end{aligned}$$

thereby defining a martingale difference sequence X adapted to $\mathbb{F}$; note that Lemma C.1 implies $\sup_j \|X_j\|_{\mathrm{F}}^2 < \infty$, so X is square-integrable and

$$\sum_{j=1}^{\infty} j^{-2}\mathbb{E}\|X_j\|_{\mathrm{F}}^2 \leq \tfrac{\pi^2}{6}\sup_j \|X_j\|_{\mathrm{F}}^2 < \infty.$$

Therefore Theorem E.1 implies

$$\frac{1}{k}\sum_{j=0}^{k-1}\Big(g_{\tilde{x}_j}(z_j)g_{\tilde{x}_j}(z_j)^\top - \mathop{\mathbb{E}}_{z_j\sim\mathcal{D}_{\tilde{x}_j}}[g_{\tilde{x}_j}(z_j)g_{\tilde{x}_j}(z_j)^\top]\Big) = \frac{1}{k}\sum_{j=1}^{k} X_j \xrightarrow{\text{a.s.}} 0. \tag{33}$$

On the other hand, we have $\tilde{x}_j \to x^\star$ almost surely by Assumption 5.3, so

$$\mathop{\mathbb{E}}_{z_j\sim\mathcal{D}_{\tilde{x}_j}}[g_{\tilde{x}_j}(z_j)g_{\tilde{x}_j}(z_j)^\top] \xrightarrow{\text{a.s.}} \mathop{\mathbb{E}}_{z\sim\mathcal{D}_{x^\star}}[g_{x^\star}(z)g_{x^\star}(z)^\top] = \Sigma_g$$

by Lemma E.3. Thus, Σ_g is a Cesàro mean of covariances:

$$\frac{1}{k}\sum_{j=0}^{k-1}\mathop{\mathbb{E}}_{z_j\sim\mathcal{D}_{\tilde{x}_j}}[g_{\tilde{x}_j}(z_j)g_{\tilde{x}_j}(z_j)^\top] \xrightarrow{\text{a.s.}} \Sigma_g. \tag{34}$$

Combining (33) and (34), we deduce

$$\frac{1}{k}\sum_{j=0}^{k-1} g_{\tilde{x}_j}(z_j)g_{\tilde{x}_j}(z_j)^\top \xrightarrow{\text{a.s.}} \Sigma_g,$$

thereby establishing (32). As for the last term in (31), we have

$$\left|\sum_{j=0}^{k-1}\tau_j\frac{(u_k^\top g_{\tilde{x}_j}(z_j))^3}{3k^{3/2}}\right| \leq \frac{8(\sup_{i,j}|u_i^\top g_{\tilde{x}_j}(z_j)|^3)}{3\sqrt{k}} \to 0 \qquad \text{as } k\to\infty. \tag{35}$$

Hence, it only remains to demonstrate that $Z_k \xrightarrow[Q_{k,0}]{d} \mathsf{N}(0,\Sigma_g)$. To see this, we apply the martingale central limit theorem (Theorem E.2). Set $M_0 = 0$ and $M_k = \sum_{j=0}^{k-1} g_{\tilde{x}_j}(z_j)$ for all $k \geq 1$; then M is a square-integrable martingale in $\mathbf{R}^d$ adapted to the filtration $\mathbb{F}$. Indeed, M_k is clearly bounded (Lemma C.1), $\mathcal{F}_k$-measurable, and

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

by the unbiasedness assumption of Definition 5.5, which also facilitates the inductive calculation of the predictable square variation $\langle M\rangle$ of M:

$$\langle M\rangle_k = \sum_{j=0}^{k-1}\mathop{\mathbb{E}}_{z_j\sim\mathcal{D}_{\tilde{x}_j}}[g_{\tilde{x}_j}(z_j)g_{\tilde{x}_j}(z_j)^\top].$$

Thus, by (34), we have

$$k^{-1}\langle M\rangle_k = \frac{1}{k}\sum_{j=0}^{k-1} \mathop{\mathbb{E}}_{z_j\sim\mathcal{D}_{\tilde{x}_j}}\left[g_{\tilde{x}_j}(z_j)g_{\tilde{x}_j}(z_j)^\top\right] \xrightarrow{\text{a.s.}} \Sigma_g.$$

The assumptions of Theorem E.2 are therefore fulfilled with $a_k = k$ (note that Lindeberg's condition holds trivially since $\sup_k \|M_k - M_{k-1}\| < \infty$ by Lemma C.1). Hence

$$Z_k = k^{-1/2}M_k \xrightarrow[Q_{k,0}]{d} \mathsf{N}(0,\Sigma_g).$$

This completes the proof.

D Proof of Lemma C.4

First, we recall that the quantities $M_g' := \sup_{x\in\mathcal{X},\, z\in\mathcal{Z}} \|\nabla_x g_x(z)\|_{\text{op}}$ and $M_g := \sup_{x\in\mathcal{X},\, z\in\mathcal{Z}} \|g_x(z)\|$ are finite by Lemma C.1. Exactly 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

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

is smooth with Lipschitz Jacobian on the compact set $\mathcal{K} := \mathcal{X}\times\mathcal{W}$. By Lemma E.5, it is enough to show that $(x,u)\mapsto C_x^u$ and

$$\hat{H}(x,u) := \mathop{\mathbb{E}}_{z\sim\mathcal{D}_x}\left[T(x,z)\big(1+h(u^\top g_x(z))\big)\right]$$

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

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

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

$$\phi(y,u,z) = T(y,z)\big(1+h(u^\top g_y(z))\big).$$

Let us verify that ϕ is a test function to which the second item 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, the second item 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

$$\nabla_y \bar{H}(x,y,u) = \mathop{\mathbb{E}}_{z \sim \mathcal{D}_x} \nabla_y \phi(y,u,z) \tag{36}$$

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

$$\nabla_y \phi(y,u,z) = \big(\nabla_y T(y,z)\big)\big(1 + h(u^\top g_y(z))\big) + h'(u^\top g_y(z))\big(T(y,z)u^\top\big)\nabla_y g_y(z) \tag{37}$$

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 (36) holds. Moreover, it follows from (37) 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 (37) that $z \mapsto \nabla_y \phi(y,u,z)$ is Lipschitz on $\mathcal{Z}$ with constant

$$a := 2\beta'_T + \mathrm{diam}(\mathcal{W}) M'_T L_h \beta_g + \mathrm{diam}(\mathcal{W}) L_h(\beta_T M'_g + M_T \beta'_g) + \mathrm{diam}(\mathcal{W})^2 M_T M'_g L_{h'} \beta_g.$$

Hence (36) 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 (37) that $y \mapsto \nabla_y \phi(y,u,z)$ is Lipschitz on $\mathcal{X}$ with constant

$$2\Lambda_T(z) + \mathrm{diam}(\mathcal{W}) M'_T L_h M'_g + \mathrm{diam}(\mathcal{W}) L_h(M_T \Lambda_g(z) + M'_T M'_g) + \mathrm{diam}(\mathcal{W})^2 M_T L_{h'} (M'_T)^2,$$

where $L_{h'} := \sup|h''|$. Hence (36) implies that $y \mapsto \nabla_y \bar{H}(x,y,u)$ is Lipschitz on $\mathcal{X}$ with constant

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

Similarly, it follows from (37) that $u \mapsto \nabla_y \phi(y,u,z)$ is Lipschitz on $\mathcal{W}$ with constant

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

so (36) 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

$$\nabla_u \bar{H}(x,y,u) = \mathop{\mathbb{E}}_{z \sim \mathcal{D}_x} \nabla_u \phi(y,u,z) \tag{38}$$

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

$$\nabla_u \phi(y,u,z) = h'(u^\top g_y(z)) T(y,z) g_y(z)^\top \tag{39}$$

and hence

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

Therefore $\bar{H}(x,y,u)$ is smooth in u and (38) holds. Moreover, it follows from (39) 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 (39) 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 (38) 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 (39) 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 (38). Similarly, it follows from (39) 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 (38). 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 on β_T, M_T, and M'_T, it follows from Lemma E.5 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. This completes the proof.

E Supplementary results

In this section, we provide some supplementary results required in our proofs. First, we record fairly general versions of the Strong Law of Large Numbers and the Central Limit Theorem.

Theorem E.1 (Martingale Strong Law of Large Numbers [11, Theorem VII.9.3])**.** *Let X be a square-integrable martingale difference sequence in $\mathbf{R}^n$, and let (a_k) be a deterministic sequence of nonzero real numbers increasing to ∞. If*

$$\sum_{j=1}^{\infty} a_j^{-2} \mathbb{E}\|X_j\|^2 < \infty,$$

then

$$a_k^{-1} \sum_{j=1}^{k} X_j \xrightarrow{a.s.} 0.$$

Theorem E.2 (Martingale Central Limit Theorem [9, Corollary 2.1.10])**.** *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 square variation of M, given by*

$$\langle M \rangle_k - \langle M \rangle_{k-1} = \mathbb{E}\big[M_k M_k^\top \mid \mathcal{F}_{k-1}\big] - M_{k-1} M_{k-1}^\top$$

with $\langle M \rangle_0 = 0$. Further, let (a_k) be a deterministic sequence of nonzero real numbers increasing to ∞, and suppose that the following two assumptions hold.

1. *(**Square variation**) There exists a deterministic positive semidefinite matrix Σ such that*

$$a_k^{-1} \langle M \rangle_k \xrightarrow{p} \Sigma.$$

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

$$a_k^{-1} \sum_{j=1}^{k} \mathbb{E}[\|M_j - M_{j-1}\|^2 \mathbf{1}\{\|M_j - M_{j-1}\| \geq \varepsilon a_k^{1/2}\} \mid \mathcal{F}_{j-1}] \xrightarrow{p} 0.$$

Then

$$a_k^{-1} M_k \xrightarrow{a.s.} 0 \quad \textit{and} \quad a_k^{-1/2} M_k \xrightarrow{d} \mathsf{N}(0, \Sigma).$$

The following lemma is used multiple times in our arguments to control covariance matrices.

Lemma E.3 (Asymptotic covariance)**.** *Let* $x_t \in \mathcal{X}$ *be a sequence in some set* $\mathcal{X} \subset \mathbf{R}^d$ *that converges to some point* $x^\star \in \mathcal{X}$*. Let* $\mu_t \in P_1(\mathcal{Z})$ *be a sequence of probability measures on a metric space* $\mathcal{Z}$ *converging to some measure* $\mu^\star \in P_1(\mathcal{Z})$ *in the Wasserstein-1 distance. Let* $g\colon \mathcal{X} \times \mathcal{Z} \to \mathbf{R}^n$ *be a map satisfying the following two conditions.*

1. *For every* $\delta > 0$*, there exists a constant* N_δ *such that*

$$\mathop{\mathbb{E}}_{z \sim \mu^\star} [\|g(x^\star, z)\|^2 \mathbf{1}\{\|g(x^\star, z)\|^2 \geq N_\delta\}] \leq \delta,$$
$$\limsup_{t \to \infty} \mathop{\mathbb{E}}_{z \sim \mu_t} [\|g(x^\star, z)\|^2 \mathbf{1}\{\|g(x^\star, z)\|^2 \geq N_\delta\}] \leq \delta.$$

2. *There exist a neighborhood* $\mathcal{V}$ *of* $x^\star$*, a measurable function* $L\colon \mathcal{Z} \to [0, \infty)$*, and constants* $\bar{L}, \beta > 0$ *such that for every* $z \in \mathcal{Z}$ *and* $x \in \mathcal{V}$*, 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

$$\mathop{\mathbb{E}}_{z \sim \mu_t} [g(x_t, z) g(x_t, z)^\top] \xrightarrow{t \to \infty} \mathop{\mathbb{E}}_{z \sim \mu^\star} [g(x^\star, z) g(x^\star, z)^\top].$$

Proof. For notational convenience, we set $g_x(z) = g(x, z)$ and

$$\Sigma = \mathop{\mathbb{E}}_{z \sim \mu^\star} [g_{x^\star}(z) g_{x^\star}(z)^\top].$$

Since $x_t \to x^\star$, it follows that all but finitely many x_t lie in $\mathcal{V}$. Suppose $x_t \in \mathcal{V}$. Condition 1 of the lemma implies a uniform upper bound on the second moment of $g_{x^\star}$:

$$\mathop{\mathbb{E}}_{z \sim \mu_t} [\|g_{x^\star}(z)\|^2] = \mathop{\mathbb{E}}_{z \sim \mu_t} [\|g_{x^\star}(z)\|^2 \mathbf{1}\{\|g_{x^\star}(z)\| < N_1\}] + \mathop{\mathbb{E}}_{z \sim \mu_t} [\|g_{x^\star}(z)\|^2 \mathbf{1}\{\|g_{x^\star}(z)\| \geq N_1\}],$$

and therefore

$$\limsup_{t \to \infty} \mathop{\mathbb{E}}_{z \sim \mu_t} [\|g_{x^\star}(z)\|^2] \leq N_1^2 + 1. \tag{40}$$

Now observe that we have the decomposition

$$\mathop{\mathbb{E}}_{z \sim \mu_t} [g_{x_t}(z) g_{x_t}(z)^\top] = \mathop{\mathbb{E}}_{z \sim \mu_t} [g_{x^\star}(z) g_{x^\star}(z)^\top] + \mathop{\mathbb{E}}_{z \sim \mu_t} [g_{x^\star}(z) (g_{x_t}(z) - g_{x^\star}(z))^\top]$$
$$+ \mathop{\mathbb{E}}_{z \sim \mu_t} [(g_{x_t}(z) - g_{x^\star}(z)) g_{x_t}(z)^\top].$$

The last two summands may be bounded as follows:

$$\left\| \mathop{\mathbb{E}}_{z \sim \mu_t} [g_{x^\star}(z) (g_{x_t}(z) - g_{x^\star}(z))^\top] \right\|_{\mathrm{op}} \leq \mathop{\mathbb{E}}_{z \sim \mu_t} [\|g_{x^\star}(z)\| \cdot \|g_{x_t}(z) - g_{x^\star}(z)\|]$$
$$\leq \|x_t - x^\star\| \sqrt{\mathop{\mathbb{E}}_{z \sim \mu_t} [\|g_{x^\star}(z)\|^2] \mathop{\mathbb{E}}_{z \sim \mu_t} [L(z)^2]} \to 0$$

and

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

where we used Hölder's inequality, condition 2 of the lemma, and (40).

To complete the proof, it now suffices to show $\mathbb{E}_{z\sim\mu_t}\left[g_{x^\star}(z)g_{x^\star}(z)^\top\right]\to\Sigma$. To this end, we define the step-like function $\varphi_q\colon\mathbf{R}\to\mathbf{R}$ for any $q\in\mathbf{R}$ by setting

$$\varphi_q(x)=\begin{cases}1 & \text{if } x\leq q,\\ -x+q+1 & \text{if } q\leq x\leq q+1,\\ 0 & \text{if } q+1\leq x.\end{cases}$$

Fix a constant $\delta>0$. Then we have the decomposition

$$\begin{aligned}&\mathop{\mathbb{E}}_{z\sim\mu_t}\left[g_{x^\star}(z)g_{x^\star}(z)^\top\right]-\Sigma\\ &\quad=\mathop{\mathbb{E}}_{z\sim\mu_t}\left[g_{x^\star}(z)g_{x^\star}(z)^\top\right]-\mathop{\mathbb{E}}_{z\sim\mu^\star}\left[g_{x^\star}(z)g_{x^\star}(z)^\top\right]\\ &\quad=\underbrace{\mathop{\mathbb{E}}_{z\sim\mathcal{D}_{x_t}}\left[(1-\varphi_{N_\delta}(\|g_{x^\star}(z)\|))g_{x^\star}(z)g_{x^\star}(z)^\top\right]-\mathop{\mathbb{E}}_{z\sim\mu^\star}\left[(1-\varphi_{N_\delta}(\|g_{x^\star}(z)\|))g_{x^\star}(z)g_{x^\star}(z)^\top\right]}_{A_t}\\ &\qquad+\underbrace{\mathop{\mathbb{E}}_{z\sim\mu_t}\left[\varphi_{N_\delta}(\|g_{x^\star}(z)\|)g_{x^\star}(z)g_{x^\star}(z)^\top\right]-\mathop{\mathbb{E}}_{z\sim\mu^\star}\left[\varphi_{N_\delta}(\|g_{x^\star}(z)\|)g_{x^\star}(z)g_{x^\star}(z)^\top\right]}_{B_t},\end{aligned}$$

where N_δ comes from condition 1 of the lemma. We bound A_t using the triangle inequality:

$$\begin{aligned}\|A_t\|_{\mathrm{op}}&\leq\left\|\mathop{\mathbb{E}}_{z\sim\mu_t}\left[(1-\varphi_{N_\delta}(\|g_{x^\star}(z)\|))g_{x^\star}(z)g_{x^\star}(z)^\top\right]\right\|_{\mathrm{op}}+\left\|\mathop{\mathbb{E}}_{z\sim\mu^\star}\left[(1-\varphi_{N_\delta}(\|g_{x^\star}(z)\|))g_{x^\star}(z)g_{x^\star}(z)^\top\right]\right\|_{\mathrm{op}}\\ &\leq\mathop{\mathbb{E}}_{z\sim\mu_t}\left[\|g_{x^\star}(z)\|^2\mathbf{1}\{\|g_{x^\star}(z)\|\geq N_\delta\}\right]+\mathop{\mathbb{E}}_{z\sim\mu^\star}\left[\|g_{x^\star}(z)\|^2\mathbf{1}\{\|g_{x^\star}(z)\|\geq N_\delta\}\right]\end{aligned}$$

and therefore

$$\limsup_{t\to\infty}\|A_t\|_{\mathrm{op}}\leq 2\delta.$$

In order to bound B_t, we first define the map $h(z)=\varphi_{N_\delta}(\|g_{x^\star}(z)\|)g_{x^\star}(z)g_{x^\star}(z)^\top$ and note

$$B_t=\mathop{\mathbb{E}}_{z\sim\mu_t}\left[h(z)\right]-\mathop{\mathbb{E}}_{z\sim\mu^\star}\left[h(z)\right].$$

Observe that we can rewrite $h=\Phi\circ g_{x^\star}$, where $\Phi(w)=\varphi_{N_\delta}(\|w\|)ww^\top$. Clearly Φ is Lipschitz continuous on any compact set and zero outside of a ball of radius $N_\delta+1$ around the origin. Therefore Φ is globally Lipschitz. Since $g_{x^\star}$ is β-Lipschitz by the second assumption of the lemma, we conclude that h is globally Lipschitz with a constant $C<\infty$ that depends only on N_δ and β.

Therefore, we deduce

$$\begin{aligned}\|B_t\|_{\mathrm{op}} &= \left\| \mathop{\mathbb{E}}_{z\sim\mu_t}[h(z)] - \mathop{\mathbb{E}}_{z\sim\mu^\star}[h(z)] \right\|_{\mathrm{op}} \\ &= \sup_{\|u\|,\|v\|\le 1} \left\{ \mathop{\mathbb{E}}_{z\sim\mu_t}[u^\top h(z) v] - \mathop{\mathbb{E}}_{z\sim\mu^\star}[u^\top h(z) v] \right\} \\ &\le C \cdot W_1(\mu_t, \mu^\star) \to 0,\end{aligned}$$

where the inequality follows from the fact that the map $z \mapsto u^\top h(z) v$ is C-Lipschitz. Hence

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

Since $\delta > 0$ is arbitrary, we have $\mathbb{E}_{z\sim\mu_t}[g_{x^\star}(z) g_{x^\star}(z)^\top] \to \Sigma$, as claimed. □

Finally, we record two lemmas about products and quotients of Lipschitz functions.

Lemma E.4. *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 E.5. *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

$$\nabla(f/g) = (1/g)(\nabla f) - (f/g^2)(\nabla g)^\top. \tag{41}$$

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, (41) and Lemma E.4 show that $\nabla(f/g)$ is the difference of two Lipschitz maps. Therefore $\nabla(f/g)$ is Lipschitz. □