The Annals of Statistics
2024, Vol. 52, No. 4, 1485–1508
https://doi.org/10.1214/24-AOS2401

ASYMPTOTIC NORMALITY AND OPTIMALITY IN NONSMOOTH STOCHASTIC APPROXIMATION

BY DAMEK DAVIS[1, a], DMITRIY DRUSVYATSKIY[2, c] AND LIWEI JIANG[1, b]

[1]*School of ORIE, Cornell University,* [a]*dsd95@cornell.edu,* [b]*lj282@cornell.edu*
[2]*Department of Mathematics, University of Washington,* [c]*ddrusv@uw.edu*

In their seminal work, Polyak and Juditsky showed that stochastic approximation algorithms for solving smooth equations enjoy a central limit theorem. Moreover, it has since been argued that the asymptotic covariance of the method is best possible among any estimation procedure in a local minimax sense of Hájek and Le Cam. A long-standing open question in this line of work is whether similar guarantees hold for important nonsmooth problems, such as stochastic nonlinear programming or stochastic variational inequalities. In this work, we show that this is indeed the case.

1. Introduction. Polyak and Juditsky [27] famously showed that the stochastic gradient method for minimizing smooth and strongly convex functions enjoys a central limit theorem: the error between the running average of the iterates and the minimizer, normalized by the square root of the iteration counter, converges to a normal random vector. Moreover, the asymptotic covariance matrix is in a precise sense "optimal" among any estimation procedure. A long standing open question is whether similar guarantees—asymptotic normality and optimality—exist for nonsmooth optimization and, more generally, for equilibrium problems. In this work, we obtain such guarantees under mild conditions that hold both in concrete circumstances (e.g., nonlinear programming) and under generic linear perturbations.

The types of problems we will consider are best modeled as stochastic variational inequalities. Setting the stage, consider the task of finding a solution x of the inclusion

$$0 \in \mathop{\mathbb{E}}_{z\sim P} A(x, z) + N_X(x). \tag{1.1}$$

Here, P is a probability distribution accessible only through sampling, $A(\cdot, z)$ is a smooth map for almost every $z \sim P$, and $N_X(x)$ denotes the normal cone to a closed set X. Stochastic variational inequalities (1.1) are ubiquitous in contemporary optimization. For example, optimality conditions for constrained optimization problems

$$\min_x \mathop{\mathbb{E}}_{z\sim P} f(x, z) \quad \text{subject to } x \in X,$$

fit into the framework (1.1) by setting $A(x, z) = \nabla f(x, z)$ in (1.1). More generally still, Nash equilibria $x = (x_1, \ldots, x_m)$ of stochastic games are solutions of the system

$$x_j \in \operatorname*{argmin}_{x_j \in X_j} \mathop{\mathbb{E}}_{z\sim P} f_j(x, z) \quad \text{for all } j = 1, \ldots, m,$$

where f_j and X_j, respectively, are the loss function and the strategy set of player j. First order optimality conditions for these k coupled inclusions can be modeled as (1.1) by setting $[A(x, z)]_j := \nabla_{x_j} f_j(x, z)$ and $X := X_1, \ldots, X_m$.

Received January 2023; revised February 2024.
MSC2020 subject classifications. Primary 90C15; secondary 65K05, 65K10.
Key words and phrases. Stochastic gradient, asymptotic normality, local asymptotic minimax theory, variational inequality, active manifold.

There are two standard strategies for solving (1.1): sample average approximation (SAA) and the stochastic forward–backward algorithm (SFB). The former proceeds by drawing a batch of samples $z_1, z_2, \ldots, z \overset{\text{iid}}{\sim} P$ and finding a solution x_k to the empirical approximation

$$0 \in \frac{1}{k} \sum_{i=1}^{k} A(x, z_i) + N_X(x). \tag{1.2}$$

In contrast, the stochastic forward–backward (SFB) algorithm proceeds in an online manner, drawing a single sample $z_k \sim P$ in each iteration k and declaring the next iterate x_{k+1} as

$$x_{k+1} \in P_X\left(x_k - \alpha_k \cdot A(x_k, z_k)\right). \tag{1.3}$$

Here, $P_X(\cdot)$ denotes the nearest-point projection onto X. In the case of constrained optimization, $A(x, z) = \nabla f(x, z)$ is the gradient of some loss function $f(x, z)$, and the process (1.3) reduces to the stochastic projected gradient algorithm. Online algorithms like SFB are usually preferable to SAA since each iteration is inexpensive and can be performed online, whereas SAA requires solving the auxiliary optimization problem (1.2). Although the asymptotic distribution of the SAA estimators is by now well understood [13, 14, 29], our understanding of the asymptotic performance of the SFB iterates is limited in nonsmooth and constrained settings. The goal of this paper is to fill this gap. The main result of our work is the following.

> Under reasonable assumptions, the running average of the SFB iterates exhibits the same asymptotic distribution as SAA. Moreover, both SAA and SFB are asymptotically optimal in a locally minimax sense of Hájek and Le Cam [15, 31].

We next describe our results, and their consequences, in some detail. Namely, it is classically known (e.g., [13, 14, 29]) that the asymptotic performance of SAA (1.2) is strongly influenced by the sensitivity of the solution x to perturbations of the left-hand side of (1.1). In order to isolate this effect, let $S(v)$ consist of solutions x to the perturbed system

$$v \in \mathop{\mathbb{E}}_{z \sim P} A(x, z) + N_X(x).$$

Throughout, we will assume that the solutions $S(v)$ vary smoothly near x. More precisely, we will assume that the graph of S locally around $(0, x)$ coincides with the graph of some smooth map $\sigma(\cdot)$. In the language of variational analysis [9], the map $\sigma(\cdot)$ is called a smooth localization of S around $(0, x)$. It is known that this assumption holds in a variety of concrete circumstances and under generic linear perturbations of semialgebraic problems [10].

Let us next provide the context and state our results. It is known from [14, 29] that under mild assumptions, the solutions x_k of SAA (1.2) are asymptotically normal:

$$\sqrt{k}\,(x_k - x) \xrightarrow{D} \mathrm{N}\left(0, \nabla\sigma(0) \cdot \mathrm{Cov}\left(A(x, z)\right) \cdot \nabla\sigma(0)^*\right). \tag{1.4}$$

Thus the Jacobian of the solution map $\nabla\sigma(0)$ appears in the asymptotic covariance of the SAA estimator. In fact, we will argue that this is unavoidable. The first contributions of our work is that we prove that the asymptotic performance of SAA is locally minimax optimal—in the sense of Hájek and Le Cam [15, 31]—among all estimation procedures. Roughly speaking, this means that for any estimation procedure that outputs $\hat{x}_k$ based on k samples, there exists a sequence of perturbations P_k with $\frac{dP_k}{dP} = 1 + O(k^{-1/2})$, such that the performance of $\hat{x}_k$ on the perturbed sequence of problems is asymptotically no better than the performance of SAA on the perturbed problems. We note that the analogous lower bound for stochastic nonlinear programming was obtained earlier in [12], and our arguments are motivated by the techniques therein. The fact that the SFB algorithm for smooth problems is asymptotically optimal was proved in [4], Theorem 5.6, by verifying that SFB is asymptotically equivariant

in law; we follow a similar argument here. Aside from the lower bound, the main result of our work is to show that under reasonable assumptions, the running average of the SFB iterates enjoys the same asymptotics as (1.4) and is thus asymptotically optimal.

The guarantees we develop are already interesting for stochastic nonlinear programming:

$$\min_x \; f(x) = \mathop{\mathbb{E}}_{z\sim P} f(x,z) \quad \text{subject to} \quad g_i(x) \le 0 \quad \forall i = 1,\ldots,m. \tag{1.5}$$

Here each g_i is a smooth function and the map $x \to f(x,z)$ is smooth for a.e. $z \sim P$. The optimality conditions for this problem can be modeled as the variational inequality (1.1) under the identification $A(x,z) = \nabla f(x,z)$ and $X = \{x : g_i(x) \le 0 \; \forall i = 1,\ldots,m\}$. The stochastic forward–backward algorithm then becomes the stochastic projected gradient method. Our results imply that under the three standard conditions—linear independence of active gradients, strict complementarity, and strong second-order sufficiency—the running average of the SFB iterates $\bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$ is asymptotically normal and optimal:

$$\sqrt{k}\,(\bar{x}_k - x^*) \xrightarrow{D} N\big(0, \nabla\sigma(0) \cdot \mathrm{Cov}\big(\nabla_x f(x^*,z)\big) \cdot \nabla\sigma(0)\big).$$

Moreover, as is classically known, the Jacobian $\nabla\sigma(0)$ admits an explicit description as

$$\nabla\sigma(0) = P_T \nabla^2_{xx} L(x^*,y^*) P_T, \tag{1.6}$$

where $\nabla^2_{xx} L(x^*,y^*)$ is the Hessian of the Lagrangian function, the symbol $\dagger$ denotes the Moore–Penrose pseudoinverse, and P_T is the projection onto the linear subspace $\{\nabla g_i(x^*)\}^{\perp}_{i\in I}$ and $I = \{i : g_i(x^*) = 0\}$ is the set of active indices. An illustrative example of the announced result is depicted in Figure 1, which plots the performance of the projected stochastic gradient method for minimizing a linear function over the intersection of two balls. A further illustration for a nonconvex problem of sparse recovery is depicted in Figure 2. This result may be surprising in light of the existing literature. Namely, Duchi and Ruan [12] uncover a striking gap between the estimation quality of SAA and at least one standard online method, called dual averaging [24, 33], for stochastic nonlinear optimization. Indeed, even for the problem of minimizing the expectation of a linear function over a ball, the dual averaging method exhibits a suboptimal asymptotic covariance ([12], Section 5.2). [1] In contrast, we see that the stochastic projected gradient method is asymptotically optimal.

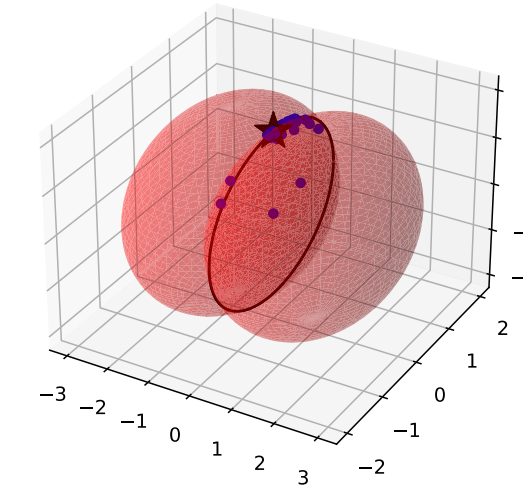
Let us now return to the general problem (1.1) and the stochastic forward–backward algorithm (1.3). In order to derive the claimed asymptotic guarantees for SFB, we will impose a few extra assumptions. First, in addition to assuming that $\sigma(\cdot)$ is smooth near the origin, we will assume that there exists a neighborhood U of the origin such that $\sigma(U)$ is a smooth manifold. This assumption is mild, since it holds automatically for example, if the matrix $\nabla\sigma(\cdot)$ has constant rank on a neighborhood of the origin. In the language of [11], the set $M = \sigma(U)$ is called an *active manifold* around $\bar{x}$. Returning to the case of stochastic nonlinear programming, the active manifold is simply the zero-set of the active inequalities

$$M = \{x : g_i(x) = 0 \; \forall i \in I\}.$$

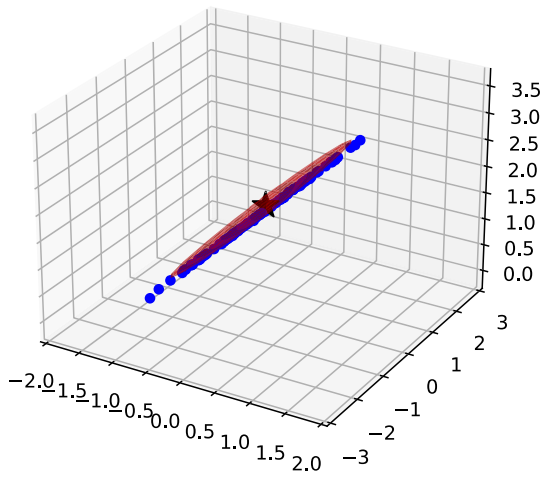
See Figure 1 for an illustration. Variants of active manifolds have been extensively studied in nonlinear programming, under the names of identifiable surfaces [32], partly smooth sets [19], UV-structures [18, 21], $g \circ F$ decomposable functions [30], and minimal identifiable sets [11].

The main idea of our argument is to relate the nonsmooth dynamics of SFB to a smooth stochastic approximation algorithm on M. More precisely, we will show that under mild

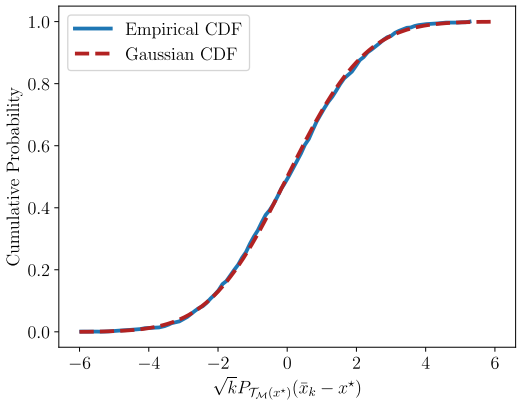
[1] In contrast, in the special case that X is polyhedral and convex, the dual averaging method is optimal [12].



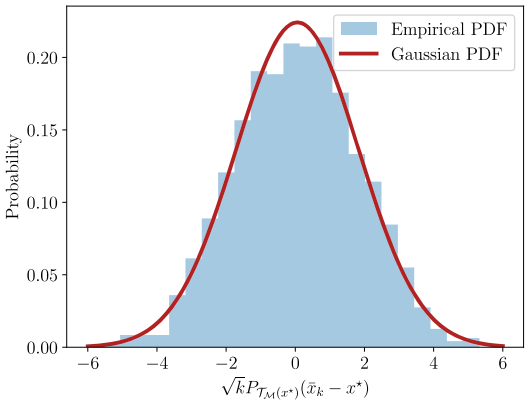
(a) Feasible region and the iterates x_k



(b) The deviations $\sqrt{k}(\bar{x}_k - x^\star)$ and the 95% confidence region.



(c) Empirical vs Gaussian CDF



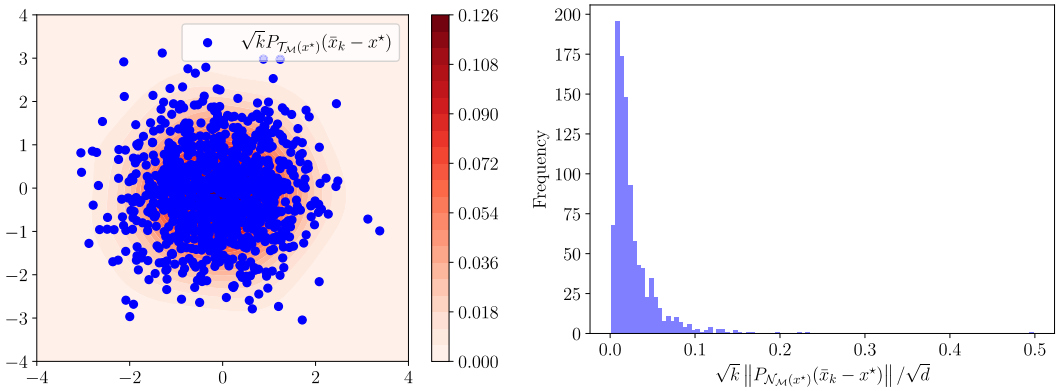
(d) Histogram vs Gaussian density.

FIG. 1. *The stochastic projected gradient method for minimizing* $\mathbb{E}g[-x_1 + g, x]$ *over the intersection of two balls centered around* $(-1, 0, 0)$ *and* $(1, 0, 0)$ *of radius two. The expectation is taken over a Gaussian* $g \sim N(0, I)$*. The optimal solution* $(0, 0, \sqrt{3})$ *(marked with a star) lies on the active manifold* M*, which is a circle depicted in black. The figure on the top left depicts the iterates generated by a single run of the process initialized at the origin with stepsize* $\eta_k = k^{-3/4}$ *and executed for* 1000 *iterations. The figure on the top right depicts the rescaled deviations* $\sqrt{k}(\bar{x}_K - x)$ *taken over* 100 *runs with* $K = 10^6$*. The two figures clearly show that the iterates rapidly approach the active manifold and asymptotically the deviations* $\sqrt{k}(\bar{x}_k - x)$ *are supported only along the tangent space to* M *at* x*. The two figures on the second row show the histogram and the empirical CDF, respectively, of the tangent components* $\sqrt{k}P_{T_M(x)}(\bar{x}_k - x)$*, overlaid with the analogous functions for a Gaussian.*

conditions, the shadow sequence $y_k := P_M(x_k)$ along the manifold M behaves smoothly up to a small error

$$y_{k+1} = y_k - \alpha_k P_{T_M(y_k)} A(y_k, z_k) + o(\alpha_k), \tag{1.7}$$

where $T_M(y_k)$ denotes the tangent space of M at y_k. Consequently, we may build on the techniques of Polyak and Juditsky [27] to obtain the asymptotics of the shadow sequence y_k, and then infer information about the original iterates x_k. We note that in the constrained optimization setting, the iteration (1.7) becomes an inexact Riemannian gradient method on the restriction of f to M.



(a) Kernel density estimation on tangent deviations. (b) Histogram of normal deviations.

FIG. 2. *The stochastic projected gradient method for minimizing* $\mathbb{E}_{(a,b)}[(\langle a, x\rangle - b)^2]$ *over the* ℓ_0 *ball* $X = \{x : \|x\|_0 \le 2\}$. *Here* $a \sim N(0, I)$ *and* $b = \langle a, x^\star\rangle + g$ *where* $g \sim N(0, 1)$ *and* $x^\star := e_1 + e_2$, *the sum of the first two standard basis vectors; in this example,* $d = 20$. *The optimal solution* $x^\star$ *lies on the active manifold* $M = \mathrm{span}\{e_1, e_2\}$. *The figure on the left depicts a kernel density estimation of the rescaled deviations* $\sqrt{K}\cdot P_{T_M(x^\star)}(\bar{x}_K - x^\star)$ *taken over* 1000 *runs of SGD* (*Gaussian kernel, bandwidth* 0·5); *here, the method is initialized at the origin with stepsize* $\eta_k = k^{-3/4}$ *and ran for* $K = 10^6$ *iterations. The figure on the right depicts the rescaled normal deviations* $\sqrt{K}\cdot P_{N_M(x^\star)}(\bar{x}_K - x^\star)/\sqrt{d}$. *Taken together, the figures again show that the iterates rapidly approach the active manifold and asymptotically the deviations* $\sqrt{k}(\bar{x}_k - x^\star)$ *are supported only along the tangent space to* M *at* $x^\star$.

The validity of (1.7) relies on two extra conditions, introduced in [6], which relate the "first-order" behavior of X to that of M. Namely, for $x \in X$ and $y \in M$ near $x^\star$, we assume:

- $N_X(y) \cap S^{d-1} \perp T_M(y)$ [(b)-regularity]
- $N_X(x) \cap S^{d-1} \subset N_M(y) + O(\|x - y\|)B$ [strong (a)-regularity]

Here S^{d-1} and B are the unit sphere and the closed unit ball in $\mathbb{R}^d$, respectively. The (b)-regularity condition simply asserts that the secant line joining $x \in X$ and $y \in M$ becomes tangent to M as x and y tend to the same point near $x^\star$. The strong-(a) regularity in contrast asserts that the normal cone $N_X(x)$ is contained in $N_M(y)$ up to a linear error $O(\|x - y\|)$ — a kind of Lipschitz condition. The two regularity conditions are introduced and thoroughly developed in [6], with numerous examples and calculus rules presented. In particular, both conditions hold automatically for stochastic nonlinear programming.

1.1. *Outline.* The outline of the rest of the paper is as follows. Section 2 presents the basic notation and constructions that will be used in the paper. Existence of smooth localizations $\sigma(\cdot)$ is a central assumption of our work. Section 3 develops asymptotic convergence guarantees for SAA, which motivate much of the subsequent sections. Section 4 presents the classes of algorithms that we consider. Section 5 states the main result on asymptotic normality of iterative methods. Section 6 present shows that SAA and SFB are both asymptotically local minimax optimal in the sense of Hájek and Le Cam.

2. Notation and basic constructions. This section records basic notation that we will use throughout the paper. To this end, the symbol $\mathbb{R}^d$ will denote a Euclidean space with inner product $\langle\cdot,\cdot\rangle$ and the induced norm $\|x\| = \sqrt{\langle x, x\rangle}$. The symbol B will stand for the closed unit ball in $\mathbb{R}^d$, while $B_r(x)$ will denote the closed ball of radius r around a point x. For any function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, the *domain*, *graph*, and *epigraph* are defined as

$$\mathrm{dom}\, f := \{x \in \mathbb{R}^d : f(x) < \infty\},$$

$$\operatorname{gph} f := \{(x, f(x)) \in \mathbb{R}^d \times \mathbb{R} : x \in \operatorname{dom} f\},$$
$$\operatorname{epi} f := \{(x, r) \in \mathbb{R}^d \times \mathbb{R} : r \geq f(x)\},$$

respectively. We say that f is *closed* if epi f is a closed set, or equivalently if f is lower-semicontinuous. The *proximal map of* f *with parameter* $\alpha > 0$ is given by

$$\operatorname{prox}_{\alpha f}(x) := \operatorname*{argmin}_y \left\{ f(y) + \frac{1}{2\alpha}\|y - x\|^2 \right\}.$$

The *distance* and the *projection* of a point $x \in \mathbb{R}^d$ onto a set $Q \subset \mathbb{R}^d$ are

$$d(x, Q) := \inf_{y \in Q} \|y - x\| \quad \text{and} \quad P_Q(x) := \operatorname*{argmin}_{y \in Q} \|y - x\|,$$

respectively. The indicator function of Q, denoted by $\delta_Q(\cdot)$, is defined to be zero on Q and $+\infty$ off it. The symbol $o(h)$ stands for any function $o(\cdot)$ satisfying $o(h)/h \to 0$ as $h \downarrow 0$.

2.1. *Smooth manifolds.* Next, we recall a few definitions from smooth manifolds; we refer the reader to [2, 16] for details. Throughout the paper, all smooth manifolds M are assumed to be embedded in $\mathbb{R}^d$ and we consider the tangent and normal spaces to M as subspaces of $\mathbb{R}^d$. Thus, a set $M \subset \mathbb{R}^d$ is a C^p *manifold* (with $p \geq 1$) if around any point $x \in M$ there exists an open neighborhood $U \subset \mathbb{R}^d$ and a C^p-smooth map F from U to some Euclidean space $\mathbb{R}^n$ such that the Jacobian $\nabla F(x)$ is surjective and equality $M \cap U = F^{-1}(0)$ holds. Then $F = 0$ are called the *local defining equations* for M, and the tangent and normal spaces to M at x are defined by $T_M(x) := \operatorname{Null}(\nabla F(x))$ and $N_M(x) := (T_M(x))^\perp$, respectively. Note that for C^p manifolds M with $p \geq 1$, the projection P_M is C^{p-1}-smooth on a neighborhood of each point x in M and is C^p-smooth on a neighborhood of the origin in the tangent space $T_M(x)$ [22]. Moreover, the inclusion range$(\nabla P_M(x)) \subseteq T_M(P_M(x))$ holds for all x near M and $\nabla P_M(x) = P_{T_M(x)}$ holds for all $x \in M$ ([20], Lemma 2.1).

Let $M \subset \mathbb{R}^d$ be a C^p-manifold for some $p \geq 1$. Then a function $f : M \to \mathbb{R}$ is called C^p-smooth around a point $x \in M$ if there exists a C^p function $\hat{f} : U \to \mathbb{R}$ defined on an open neighborhood $U \subset \mathbb{R}^d$ of x and that agrees with f on $U \cap M$. Then the *covariant gradient of* f *at* x is the vector $\nabla_M f(x) := P_{T_M(x)}(\nabla \hat{f}(x))$. When f and M are C^2-smooth, the *covariant Hessian of* f *at* x is defined to be the unique self-adjoint bilinear form $\nabla^2_M f(x) : T_M(x) \times T_M(x) \to \mathbb{R}$ satisfying

$$\frac{d^2}{dt^2} f\left(P_M(x + tw)\right)\Big|_{t=0} = \langle \nabla^2_M f(x)u, u\rangle \quad \text{for all } u \in T_M(x).$$

If M is C^3-smooth, then we can write $\nabla^2_M f(x)$ simply as

$$\nabla^2_M f(x) = P_{T_M(x)} \nabla^2 (f \circ P_M)(x) P_{T_M(x)},$$

while regarding the right-hand side as a linear operator on $T_M(x)$.

A map $F : M \to \mathbb{R}^m$ is called C^p smooth near a point x if there exists a map $\hat{F} : U \to \mathbb{R}^d$ defined on some neighborhood $U \subset \mathbb{R}^d$ of x that agrees with F on M near x. In this case, we define the *covariant Jacobian* $\nabla F(x) : T_M(x) \to \mathbb{R}^m$ by the expression $\nabla_M F(x)u = \nabla \hat{F}(x)u$ for all $u \in T_M(x)$. An easy computation shows that in the particular case when $F(x) = \nabla_M f(x)$ for a C^3-smooth function $f : M \to \mathbb{R}$, the quadratic form defined by $\nabla_M F(x)$ on $T_M(x)$ coincides with $\nabla^2_M f(x)$. More precisely, equality holds:

$$P_{T_M(x)} \cdot \nabla_M F(x) \cdot P_{T_M(x)} = P_{T_M(x)} \cdot \nabla^2_M f(x) \cdot P_{T_M(x)}.$$

2.2. *Normal cones and subdifferentials.* Next, we require a few basic constructions of nonsmooth and variational analysis. Our discussion will follow mostly closely Rockafellar–Wets [28]. Other influential treatments of the subject include [1, 3, 23, 25]. The *Fréchet normal cone* to a set $Q \subset \mathbb{R}^d$ at a point $x \in \mathbb{R}^d$, denoted $\hat{N}_Q(x)$, consists of all vectors $v \in \mathbb{R}^d$ satisfying

$$\langle v, y - x \rangle \leq o(\|y - x\|) \quad \text{as } y \to x \text{ in } Q. \tag{2.1}$$

Thus v lies in $\hat{N}_Q(x)$ if up to first-order v makes an obtuse angle with all directions pointing from x into Q. Generally, Fréchet normals are highly discontinuous with respect to perturbations of the basepoint x. Consequently, we enlarge the Fréchet normal cone as follows. The *limiting normal cone* to Q at $x \in Q$, denoted by $N_Q(x)$, consists of all vectors $v \in \mathbb{R}^d$ for which there exist sequences $x_i \in Q$ and $v_i \in \hat{N}_Q(x_i)$ satisfying $(x_i, v_i) \to (x, v)$.

The analogous of normal cones for functions are subdifferentials, or sets of generalized derivatives. Namely, consider a function $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ and a point $x \in \operatorname{dom} f$. The *Fréchet subdifferential of* f *at* x, denoted $\hat{\partial} f(x)$, consists of all vectors $v \in \mathbb{R}^d$ satisfying the approximation property:

$$f(y) \geq f(x) + \langle v, y - x \rangle + o(\|y - x\|) \quad \text{as } y \to x.$$

The *limiting subdifferential of* f *at* x, denoted $\partial f(x)$, consists of all $v \in \mathbb{R}^d$ such that there exist sequences $x_i \in \mathbb{R}^d$ and Fréchet subgradients $v_i \in \hat{\partial} f(x_i)$ satisfying $(x_i, f(x_i), v_i) \to (x, f(x), v)$ as $i \to \infty$. A point x satisfying $0 \in \partial f(x)$ is called *critical* for f. The primary goal of algorithms for nonsmooth optimization is the search for critical points.

2.3. *Active manifolds.* Critical points of typical nonsmooth functions lie on a certain manifold that captures the activity of the problem in the sense that critical points of slight linear perturbation of the function do not leave the manifold. Such active manifolds have been modeled in a variety of ways, including identifiable surfaces [32], partial smoothness [19], UV-structures [18, 21], $g \circ F$ decomposable functions [30], and minimal identifiable sets [11]. In this work, we adopt the following formal model of activity, explicitly used in [11].

DEFINITION 2.1 (Active manifold). Consider a function $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ and fix a set $M \subseteq \operatorname{dom} f$ containing a point $\bar{x}$ satisfying $0 \in \partial f(\bar{x})$. Then M is called an *active* C^p*-manifold around* $\bar{x}$ if there exists a constant $\rho > 0$ satisfying the following:

- *(smoothness)* The set M is a C^p manifold near $\bar{x}$ and the restriction of f to M is C^p-smooth near $\bar{x}$.
- *(sharpness)* The lower bound holds:

$$\inf\{\|v\| : v \in \partial f(x), x \in U \setminus M\} > 0,$$

where we set $U = \{x \in B_\rho(\bar{x}) : |f(x) - f(\bar{x})| < \rho\}$.

More generally, we say that M *is an active manifold for* f *at* $\bar{x}$ *for* $\bar{v} \in \partial f(\bar{x})$ if M is an active manifold for the tilted function $f_{\bar{v}}(x) = f(x) - \langle \bar{v}, x \rangle$ at $\bar{x}$.

The sharpness condition simply means that the subgradients of f must be uniformly bounded away from zero at points off the manifold that are sufficiently close to $\bar{x}$ in distance and in function value. The localization in function value can be omitted for example if f is weakly convex or if f is continuous on its domain; see [11] for details.

Intuitively, the active manifold has the distinctive feature that the function varies smoothly along the manifold and grows linearly normal to it; see Figure 3 for an illustration. This is summarized by the following theorem from [5], Theorem D.2.

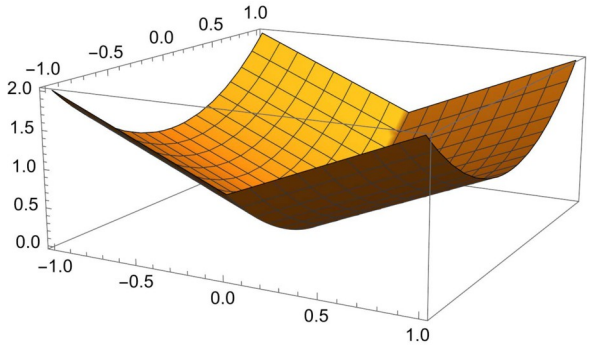


FIG. 3. $f(x_1, x_2) = |x_1| + x_2^2$.

PROPOSITION 2.2 (Identification implies sharpness). *Suppose that a closed function* $f: \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ *admits an active manifold* M *at a point* $\bar{x}$ *satisfying* $0 \in \hat{\partial} f(\bar{x})$. *Then there exist constants* $c > 0$ *such that*

$$f(x) - f(P_M(x)) \geq c \cdot \operatorname{dist}(x, M) \quad \forall x \in B(\bar{x}). \tag{2.2}$$

Notice that there is a nontrivial assumption $0 \in \hat{\partial} f(\bar{x})$ at play in Proposition 2.2. Indeed, under the weaker inclusion $0 \in \partial f(\bar{x})$ the growth condition (2.2) may easily fail, as the univariate example $f(x) = -|x|$ shows. It is worthwhile to note that under the assumption $0 \in \hat{\partial} f(\bar{x})$, the active manifold is locally unique around $\bar{x}$ ([11], Proposition 8.2).

In order to make progress, we will require two extra conditions to hold along the active manifold that tightly couple the subgradients of f on and off the manifold. Although these two conditions, introduced in [6], may look formidable, they are very mild indeed.

The motivation for the first regularity condition stems from a weakening of Taylor's theorem that is appropriate for nonsmooth functions. Namely, recall that any C^1-smooth function f satisfies the first-order approximation property

$$f(y) = f(x) + \langle \nabla f(x), y - x \rangle + o(\|y - x\|) \quad \text{as } x, y \to \bar{x}. \tag{2.3}$$

This estimate is fundamental to optimization theory and practice since it quantifies the approximation qualify of the linear model of f furnished by the gradient. When f is nonsmooth, the analogue of (2.3) with subgradients replacing the gradient can not possibly hold uniformly over all points x and y near $\bar{x}$, since it would imply differentiability. Instead, a reasonable requirement is to only require (2.3) to hold with points y lying on the active manifold M. In fact, for our purposes, it suffices to replace the equality with an inequality.

DEFINITION 2.3 ($(b_{\leq})$-regularity). Consider a function $f: \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ that is locally Lipschitz continuous on its domain. Fix a set $M \subset \operatorname{dom} f$ that is a C^1 manifold around $\bar{x}$ and such that the restriction of f to M is C^1-smooth near $\bar{x}$. We say that f is $(b_{\leq})$-*regular along* M *at* $\bar{x}$ if there exists $\rho > 0$ such that the estimates

$$f(y) \geq f(x) + \langle v, y - x \rangle + \frac{\rho}{1 + \|v\|} \cdot o(\|y - x\|), \tag{2.4}$$

hold for all $x \in \operatorname{dom} f \cap B_\rho(\bar{x})$, $y \in M \cap B_\rho(\bar{x})$, and $v \in \partial f(x)$.

Importantly, condition $(b_{\leq})$ along an active manifold M directly implies that all negative subgradients of f, taken at points x near $\bar{x}$, point towards the active manifold M. Indeed, this is a direct consequence of Proposition 2.2, and is summarized in the following corollary. Consequently, subgradient type methods always move in the direction of the active manifold—clearly a desirable property.

COROLLARY 2.4 (Proximal aiming). *Consider a closed function* $f : \mathbb{R}^d \to \mathbb{R}\cup\{\infty\}$ *that admits an active* C^1*-manifold* M *at a point* $\bar{x}$ *satisfying* $0 \in \hat{\partial} f(\bar{x})$. *Suppose that* f *is locally Lipschitz continuous on its domain and that* f *is* $(b_{\leq})$*-regular along* M *near* $\bar{x}$. *Then, there exists a constant* $\mu > 0$ *such that, the estimate*

$$\langle v, x - P_M(x)\rangle \geq \mu \cdot \operatorname{dist}(x, M) + o\left(1 + \|v\|\right) \cdot \operatorname{dist}(x, M), \tag{2.5}$$

holds for all $x \in \operatorname{dom} f$ *near* $\bar{x}$ *and for all* $v \in \partial f(x)$.

The second regularity condition has a different flavor, stemming from a weakening of Lipschitz continuity of the gradient. Namely, nonsmoothness by its nature implies that the deviation $\partial f(x) - \partial f(y)$ is not controlled well by the distance in the arguments $\|x - y\|$. On the other hand, it turns out that if we take $y \in M$ and only look at the subgradient deviations in tangent directions, the error $P_{T_M(y)}[\partial f(x) - \partial f(y)]$ is typically linearly bounded by $\|x - y\|$. Moreover, in typical circumstances $P_{T_M(y)}[\partial f(y)]$ consists only of a single vector, the covariant gradient $\nabla_M f(y)$.

DEFINITION 2.5 (Strong (a)-regularity). Consider a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ that is locally Lipschitz continuous on its domain. Fix a set $M \subset \operatorname{dom} f$ that is a C^1 manifold around $\bar{x}$ and such that the restriction of f to M is C^1-smooth near $\bar{x}$. We say that f is *strongly* (a)*-regular along* M *near* $\bar{x}$ if there exist constants $C, \varepsilon > 0$ satisfying

$$\left\|P_{T_M(y)}\left(v - \nabla_M f(y)\right)\right\| \leq C\left(1 + \|v\|\right)\|x - y\|, \tag{2.6}$$

for all $x \in \operatorname{dom} f \cap B_\varepsilon(\bar{x})$, $y \in M \cap B_\varepsilon(\bar{x})$, and $v \in \partial f(x)$.

The two regularity conditions easy extend to sets through their indicator functions. Namely, we say that a set $Q \subset \mathbb{R}^d$ is $(b_{\leq})$-regular (respectively strongly (a)-regular) along a C^1 manifold $M \subset Q$ at $\bar{x} \in M$ if the indicator function δ_Q is $(b_{\leq})$-regular (respectively, strongly (a)-regular) along M at $\bar{x}$.

The paper [6] presents a wide array of functions that admit active manifolds along which both conditions $(b_{\leq})$ and strong (a) hold. Here, we discuss in detail a single example of nonlinear programming, and refer the reader to [6] for many more examples.

EXAMPLE 1 (Nonlinear programming). Consider the problem of nonlinear programming

$$\begin{aligned} &\min_x f(x) \\ &\text{s.t.} \quad g_i(x) \leq 0 \quad \text{for } i = 1, \ldots, m, \\ &\qquad g_i(x) = 0 \quad \text{for } i = m+1, \ldots, n, \end{aligned} \tag{2.7}$$

where f and g_i are C^p-smooth functions on $\mathbb{R}^d$. Let X denotes the set of all feasible points to the problem. Consider now a point $\bar{x} \in X$ that is critical for the function $f + \delta_X$ and define the active index set

$$I = \{i : g_i(\bar{x}) = 0\}.$$

Suppose the following is true:

- *(LICQ)* the gradients $\{\nabla g_i(\bar{x})\}_{i \in I}$ are linearly independent.

Then the set

$$M = \{x : g_i(x) = 0 \ \forall i \in I\}$$

is a C^p smooth manifold locally around $\bar{x}$. Moreover, all three functions f, δ_X, and $f + \delta_X$ are $(b_{\leq})$-regular and strongly (a)-regular along M near $\bar{x}$. In order to ensure that M is an active manifold of $f + \delta_X$, an extra condition is required. Define the Lagrangian function

$$L(x, y) := f(x) + \sum_{i=1}^{n+m} y_i g_i(x).$$

Criticality of $\bar{x}$ and LICQ ensures that there exists a (unique) Lagrange multiplier vector $\bar{y} \in \mathbb{R}^m_+ \times \mathbb{R}^n$ satisfying $\nabla_x L(\bar{x}, \bar{y}) = 0$ and $\bar{y}_i = 0$ for all $i \notin I$. Suppose the following standard assumption is true:

- *(Strict complementarity)* $\bar{y}_i > 0$ for all $i \in I \cap \{1, \ldots, m\}$

Then M is indeed an active C^p manifold for $f + \delta_X$ at $\bar{x}$.

EXAMPLE 2 (ℓ_1-regularization). Consider the stochastic optimization problem with ℓ_1 regularization

$$\min_x \ g(x) = f(x) + \lambda \|x\|_1, \tag{2.8}$$

where $f(x) = \mathbb{E}_{z \in P}[f(x, z)]$ is a C^p-smooth function in $\mathbb{R}^d$. Consider now a point $\bar{x} \in \mathbb{R}^d$ that is critical for the function g and define the index set $I = \{i : \bar{x}_i = 0\}$. Then the set

$$M = \{x : x_i = \bar{x}_i, \ \forall i \in I\}$$

is an affine space, hence a smooth manifold. Note that the definition of criticality ensures that $0 \in \partial g(\bar{x})$, so we always have

$$-\nabla f(\bar{x})_i \in [-\lambda, \lambda] \quad \forall i \in I.$$

Suppose the following condition is true:

- *(Strict complementarity)* $-(\nabla f(\bar{x}))_i \in (-\lambda, \lambda)$ for all $i \in I$.

Then M is indeed an active C^p manifold for g at $\bar{x}$. Moreover, $(b_{\leq})$-regularity and strong (a)-regularity hold trivially for g along M at $\bar{x}$. Note that there is usually a bias between the center of the asymptotic distribution $\bar{x}$ and the ground truth due to the regularization term.

2.4. *Smoothly invertible maps and active manifolds.* Performance of statistical estimation procedures strongly depends on the sensitivity of the problem to perturbation. A variety of estimation problems can in turn be modeled as the task of solving an inclusion $0 \in F(x)$ for some set-valued map F, whose values we can only approximate by sampling. We next review basic perturbation theory based on the inverse/implicit function theorem paradigm, while closely following the monograph [9].

A *set-valued map* $F : \mathbb{R}^d \rightrightarrows \mathbb{R}^m$ is an assignment that maps a point $x \in \mathbb{R}^d$ to a set $F(x) \subset \mathbb{R}^m$. Set-valued maps always admit a set-valued inverse:

$$F^{-1}(y) = \{x : y \in F(x)\}.$$

The domain and graph of F are defined, respectively, as

$$\operatorname{dom} F := \{x : F(x) = \varnothing\} \quad \text{and} \quad \operatorname{gph} F := \{(x, y) : y \in F(x)\}.$$

We will be interested in the sensitivity of the solutions to the system $v \in F(x)$ with respect to perturbations of the left-hand side v, or equivalently, the variational behavior of the map $v \to F^{-1}(v)$. In particular, we will be interested in settings when the graph of F^{-1} coincides locally around a base point (v, x) with a graph of a smooth map. This is the content of the following definition.

DEFINITION 2.6 (Smooth invertibility). Consider a set-valued map $F : \mathbf{R}^d \rightrightarrows \mathbf{R}^m$ and a pair $(\bar{x}, \bar{v}) \in \operatorname{gph} F$. We say that F is C^p *invertible around* $(\bar{x}, \bar{v})$ *with inverse* $\sigma(\cdot)$ if there exists a single-valued C^p-smooth map $\sigma(\cdot)$ and a neighborhood U of $(\bar{v}, \bar{x})$ satisfying

$$U \cap \operatorname{gph} F^{-1} = U \cap \operatorname{gph} \sigma.$$

The definition might seem odd at first: there is nothing "smooth" about F, and yet we require the graph of F^{-1} to coincide with a graph of a smooth function near $(\bar{v}, \bar{x})$. On the contrary, we will see that in a variety of settings this assumption is indeed valid. In particular, smooth invertibility is typical in nonlinear programming.

EXAMPLE 3 (Nonlinear programming). Returning to Example 1 with $p \geq 2$, define the set-valued map

$$F(x) = \nabla f(x) + N_X(x).$$

Then it is classically known that F is C^{p-1} invertible at $(\bar{x}, 0)$ if and only if the matrix

$$:= P_{T_M(\bar{x})} \nabla^2_{xx} L(\bar{x}, \bar{y}) P_{T_M(\bar{x})}$$

is nonsingular on $T_M(\bar{x})$. In this case, the Jacobian of the inverse map is $\nabla\sigma(0) = \quad^{\dagger}$, where $\dagger$ denotes the Moore–Penrose pseudoinverse. It is worthwhile to note that can be equivalently written as $P_{T_M(\bar{x})} \nabla^2_M f(\bar{x}) P_{T_M(\bar{x})}$.

EXAMPLE 4 ($_1$-regularization). Returning to Example 2 with $p \geq 2$, define the set-valued map $F(x) = \nabla f(x) + \lambda \partial(\cdot \quad _1)(x)$. Then F is C^{p-1} invertible at $(\bar{x}, 0)$ if and only if the matrix $:= P_{T_M(\bar{x})} \nabla^2 f(\bar{x}) P_{T_M(\bar{x})}$ is nonsingular on $T_M(\bar{x})$.

Smooth invertibility is closely tied to active manifolds, and Example 3 and Example 4 are simple consequences. Indeed the following much more general statement is true. This result follows from a standard argument combining active manifolds and the implicit function theorem. The proof appears in Section 7.1 of the Supplementary Material [7]. We will require a mild assumption on the considered functions. Following [26], Definition 2.1, a function f is called *subdifferentially continuous* at a point $\bar{x}$ if for any sequences $(x_i, v_i) \in \operatorname{gph} \partial f$ converging to some pair $(\bar{x}, \bar{v}) \in \operatorname{gph} \partial f$, the function values $f(x_i)$ converge to $f(\bar{x})$. In particular, functions that are continuous on their domains and closed convex functions are subdifferentially continuous.

THEOREM 2.7 (Smooth invertibility and active manifolds). *Consider the map*

$$F(x) := A(x) + \partial f(x),$$

where $A : \mathbf{R}^d \to \mathbf{R}^d$ *is* C^p*-smooth and* $f : \mathbf{R}^d \to \mathbf{R} \cup \{\infty\}$ *is subdifferentially continuous near a point* $\bar{x}$*. Suppose that* f *admits a* C^{p+1} *active manifold* M *at some point* $\bar{x}$ *for* $-A(\bar{x}) \in \hat{\partial} f(\bar{x})$*. Let* $G(x) = 0$ *be any* C^{p+1}*-smooth local defining equations for* M *near* $\bar{x}$ *and let* $\hat{f}$ *be a* C^{p+1}*-smooth function that agrees with* f *on a neighborhood of* $\bar{x}$ *in* M*. Define the map*

$$H(x, y) := A(x) + \nabla \hat{f}(x) + \nabla G(x) \quad y.$$

Then there exists a unique multiplier vector $\bar{y}$ *satisfying the condition* $0 = H(\bar{x}, \bar{y})$*. Moreover,* F *is* C^p*-invertible around* $(0, \bar{x})$ *with inverse* $\sigma(\cdot)$ *if and only if the matrix*

$$:= P_{T_M(\bar{x})} \nabla_x H(\bar{x}, \bar{y}) P_{T_M(\bar{x})}$$

is nonsingular on $T_M(\bar{x})$*, and in this case equality* $\nabla\sigma(0) = \quad^{\dagger}$ *holds.*

3. Asymptotic normality of SAA. Before analyzing the asymptotic performance of stochastic approximation algorithms, it is instructive to first recall guarantees for sample average approximation (SAA), where the assumptions, conclusions, and arguments are much simpler to state. This is the content of this section: we derive the asymptotic distribution of the SAA estimator for nonsmooth problems.Throughout the section, we focus on the problem of finding a point x satisfying the variational inclusion

$$0 \in A(x) + H(x) \quad \text{where } A(x) = \mathop{\mathbb{E}}_{z\sim P} A(x, z). \tag{3.1}$$

Here $H : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is a set-valued map with closed graph, P is a fixed probability distribution on some measure space $(Z, \mathcal{F})$, and $A: \mathbb{R}^d \times Z \to \mathbb{R}^d$ is a measurable map. We will impose the following assumption throughout the rest of the section.

ASSUMPTION A. *The map $F := A + H$ is C^1-smoothly invertible near $(0, \bar{x})$ with inverse $\sigma(\cdot)$.*

The SAA approach to solving (3.1) proceeds as follows. Let $S = \{z_1, \ldots, z_k\}$ be i.i.d. samples drawn from P and let x_k be a solution of the problem

$$0 \in A_S(x) + H(x) \quad \text{where } A_S(x) := \frac{1}{k}\sum_{i=1}^{k} A(x, z_i), \tag{3.2}$$

assuming one exists. We will now show that the solutions of sample average approximations are asymptotically normal with covariance $\nabla\sigma(0) \cdot \mathrm{Cov}(A(\bar{x}, z)) \cdot \nabla\sigma(0)$. Though variants of this result are known [13, 14, 29], we provide a short proof in Section 8.1 of the Supplementary Material [7] highlighting the use of the solution map $\sigma(\cdot)$. To this end, we impose the following standard assumption.

ASSUMPTION B (Integrability and smoothness). *Suppose that there exists a neighborhood U around $\bar{x}$ satisfying the following.*

1. *For almost every z, the map $A(\cdot, z)$ is differentiable at every $x \in U$.*
2. *The second moment bounds hold*:

$$\sup_{x\in U} \mathop{\mathbb{E}}_{z\sim P} \|A(x, z)\|^2 < \infty \quad \text{and} \quad \sup_{x\in U} \mathop{\mathbb{E}}_{z\sim P} \|\nabla A(x, z)\|_{\mathrm{op}}^2 < \infty.$$

The following theorem shows that as long as x_k eventually stay in a sufficiently small neighborhood of $\bar{x}$, the error $\sqrt{k}(x_k - \bar{x})$ is asymptotically normal with covariance $\nabla\sigma(0) \cdot \mathrm{Cov}(M(\bar{x}, z)) \cdot \nabla\sigma(0)$. Verifying that the problem (3.2) admits solutions x_k that are sufficiently close to $\bar{x}$ is a separate and well-studied topic and we do not discuss it here.

THEOREM 3.1 (Sample average approximation). *Suppose that Assumptions A and B hold. In particular, there exist $\delta_1, \delta_2 > 0$ and a C^1-smooth map $\sigma : B_{\delta_1}(0) \to B_{\delta_2}(\bar{x})$ with*

$$\mathrm{gph}\,\sigma = \left(B_{\delta_1}(0) \times B_{\delta_2}(\bar{x})\right) \cap \mathrm{gph}\, F^{-1}.$$

Suppose moreover that there exists a square integrable function $L(z)$ satisfying

$$\|\nabla A(x_1, z) - \nabla A(x_2, z)\| \le L(z)\|x_1 - x_2\| \quad \forall x_1, x_2 \in B_{\delta_2}(\bar{x}). \tag{3.3}$$

Shrinking δ_2, if necessary, let us ensure that $\delta_2 \le \min\{\frac{\mathrm{lip}(\sigma)^{-1}}{2\mathbb{E}L}, \frac{1}{2\mathbb{E}L}\}$. Let $S = \{z_1, \ldots, z_k\}$ be i.i.d. samples drawn from P and let x_k be a measurable selection of the solutions (3.2) such that $\Pr[x_k \in B_{\delta_2}(\bar{x})] \to 1$ as k tends to infinity. Then the expansion holds:

$$\sqrt{k}(x_k - \bar{x}) = -\nabla\sigma(0) \cdot \sqrt{k}\, A(\bar{x}) - A_S(\bar{x}) + o_P(1),$$

and therefore

$$\sqrt{k}(x_k - \bar{x}) \xrightarrow{D} \mathrm{N}\left(0, \nabla\sigma(0) \cdot \mathrm{Cov}\, A(\bar{x}, z) \cdot \nabla\sigma(0)\right). \tag{3.4}$$

Note that Theorem 3.1 assumes existence of a measurable selection of the solutions (3.2) such that $\Pr[x_k \in B_{\varepsilon_2}(\bar{x})] \to 1$ as k tends to infinity. This is a very mild assumption and follows for example, from uniform convergence and smooth invertibility.

THEOREM 3.2 (Existence of measurable selections). *Suppose that F is C^1-smoothly invertible around near $(0, \bar{x})$ and that $A(x, z)$ and $\nabla A(x, z)$ converge uniformly on some ball around $\bar{x}$, that is there exists $\varepsilon > 0$ such that*

$$\sup_{x \in B_\varepsilon(\bar{x})} \|\nabla A_S(x) - \nabla A(x)\| = o_p(1) \quad \text{and} \quad \sup_{x \in B_\varepsilon(\bar{x})} \|A_S(x) - A(x)\| = o_p(1).$$

Then there exists $\delta > 0$ and a measurable selection of the solutions (3.2) such that $\Pr[x_k \in B_\delta(\bar{x})] \to 1$ as k tends to infinity.

PROOF. Standard results on the implicit function theorem (see proof of [9], Theorem 3G.3) imply that there exist sufficiently small $\varepsilon_2, \varepsilon_3 > 0$ such that whenever $\sup_{x \in B_\varepsilon(\bar{x})} \|A_S(x) - A(x)\| < \varepsilon_2$ and $\sup_{x \in B_\varepsilon(\bar{x})} \|\nabla A_S(x) - \nabla A(x)\| < \varepsilon_2$, the map $A_S + H$ is guaranteed to be smoothly invertible on $B_{\varepsilon_3}(\bar{x}) \times B_{\varepsilon_3}(0)$. In particular, the solution $x_k \in B_{\varepsilon_3}(\bar{x})$ of (3.2) exists and is unique. To see measurability of x_k, observe that the maps A_S and H are both measurable ([28], Exercise 14.9), and therefore so is the map $D(S) = \mathrm{gph}(A_S + H)$. Notice now that $x_k(S)$ uniquely satisfies the inclusion $(x_k(S), 0) \in D(S)$. Therefore, [28], Theorem 14.16, implies that x_k is measurable.

Our goal in the rest of the paper is to show that a simple online algorithm, namely the stochastic forward backward (SFB) method, under reasonable assumptions enjoys the same guarantees as (3.4) for SAA. Moreover, in the final section of the paper (Section 6), we will show that this performance is best possible among any estimation procedure in a local minimax sense, and therefore both SAA and SFB are asymptotically local minimax optimal.

4. Stochastic approximation: Assumptions & examples. We now move to stochastic approximation algorithms, and in this section set forth the algorithms we will consider and the relevant assumptions. The concrete examples we will present will all be geared to solving variational inclusions, but the specifics of this problem class is somewhat distracting. Therefore, we will instead only isolate the essential ingredients that are needed for our results to take hold. Setting the stage, our goal is to find a point x satisfying the inclusion

$$0 \in F(x), \tag{4.1}$$

where $F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is a set-valued map. Throughout, we fix one such solution $\bar{x}$ of (4.1). We will assume that in a certain sense, the problem (4.1) is "variationally smooth". That is, there exists a distinguished manifold M —the active manifold in concrete examples—containing $\bar{x}$ and such that the map $x \to P_{T_M(x)} F(x)$ is single-valued and C^1-smooth on M near $\bar{x}$. The following assumption makes this precise.

ASSUMPTION C (Smooth reduction). *Suppose that there exists a C^p manifold $M \subset \mathbb{R}^d$ such that the following properties are true.*

(C1) *The map $F_M : M \to \mathbb{R}^d$ defined by*

$$F_M(x) := P_{T_M(x)} F(x)$$

is single-valued on some neighborhood of $\bar{x}$ in M.

(C2) *There exists a neighborhood* U *of* $(\bar{x}, 0)$ *such that*

$$U \cap \operatorname{gph} F = U \cap \operatorname{gph}(F_M + N_M).$$

We note that smooth invertibility of F can be easily characterized in terms of the covariant Jacobian $\nabla_M F_M(\bar{x})$. This is the content of the following lemma.

LEMMA 4.1 (Jacobian of the solution map). *The map* F *is* C^p*-smoothly invertible around* $(\bar{x}, 0)$ *with localization* $\sigma(\cdot)$ *if and only if the linear map* $P_{T_M(\bar{x})} \nabla F_M(\bar{x}) P_{T_M(\bar{x})}$ *is nonsingular on* $T_M(\bar{x})$. *In this case, the Jacobian of the localization is given by*

$$\nabla \sigma(0) = \left[P_{T_M(\bar{x})} \nabla_M F_M(\bar{x}) P_{T_M(\bar{x})}\right]^{\dagger}.$$

PROOF. Let be a smooth extension of F to a neighborhood $V \subset \mathbb{R}^d$ of $\bar{x}$. In light of Assumption (C2), the graphs of F and $+N_M$ coincide near $(\bar{x}, 0)$, and therefore we can focus on existence of smooth localizations of $(+N_M)^{-1}$. Applying Lemma 7.2 in the Supplementary Material [7] with $\bar{y} = 0$, we see that $+N_M$ is C^p-smoothly invertible around $(\bar{x}, 0)$ if and only if the linear map $P_{T_M(\bar{x})} \nabla(\bar{x}) P_{T_M(\bar{x})}$ is nonsingular on $T_M(\bar{x})$. In this case, the Jacobian of the localization is given by $\nabla \sigma(0) = (P_{T_M(\bar{x})} \nabla(\bar{x}) P_{T_M(\bar{x})})^{\dagger}$. Noting the equality $\nabla F_M(\bar{x}) P_{T_M(\bar{x})} = \nabla_M(\bar{x}) P_{T_M(\bar{x})}$ completes the proof.

The stochastic approximation algorithms we consider assume access to a *generalized gradient mapping*:

$$G: \mathbb{R}_{++} \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d.$$

Given $x_0 \in \mathbb{R}^d$, the algorithm iterates the update

$$x_{k+1} = x_k - \alpha_k G_{\alpha_k}(x_k, v_k), \tag{4.2}$$

where $\alpha_k > 0$ is a control sequence and v_k is stochastic noise. We will place relevant assumptions on the noise v_k later in Section 5.

We make two assumptions on G. The first (Assumption D) is similar to classical Lipschitz assumptions and ensures the steplength can only scale linearly in v.

ASSUMPTION D (Steplength). *We suppose that there exists a constant* $C > 0$ *and a neighborhood* U *of* $\bar{x}$ *such that the estimate*

$$\sup_{x \in U_F} \|G_\alpha(x, v)\| \le C(1 + \|v\|),$$

holds for all $v \in \mathbb{R}^d$ *and* $\alpha > 0$, *where we set* $U_F := U \cap \operatorname{dom} F$.

The second assumption makes precise the relationship between the mapping G and F_M.

ASSUMPTION E (Strong (a) and aiming). *We suppose that there exist constants* $C, \mu > 0$ *and a neighborhood* U *of* $\bar{x}$ *such that the following hold for all* $v \in \mathbb{R}^d$ *and* $\alpha > 0$, *where we set* $U_F := U \cap \operatorname{dom} F$.

(E1) *(Tangent comparison) For all* $x \in U_F$, *we have*

$$\left\| P_{T_M(P_M(x))} G_\alpha(x, v) - F(P_M(x)) - v \right\| \le C(1 + \|v\|)\left[\operatorname{dist}^2(x, M) + \alpha\right].$$

(E2) *(Proximal Aiming) For* $x \in U_F$, *we have*

$$\langle G_\alpha(x, v) - v, x - P_M(x) \rangle \ge \mu \cdot \operatorname{dist}(x, M) - (1 + \|v\|)\left[o(\operatorname{dist}^2(x, M)) + C\alpha\right].$$

Some comments are in order. Assumption (E1) ensures that the direction of motion $G_{\alpha_k}(x_k, v_k)$ approximates well $F_M(P_M(x))$ in tangent directions to the manifold M. Assumption (E2) ensures that after subtracting the noise from $G_{\alpha_k}(x_k, v_k)$, the update direction $x_k - x_{k+1}$ locally points towards the manifold M. Note that the little-o term in (E2) depends only on $\operatorname{dist}(x, M)$ and not on α. We will later show that this ensures the iterates x_k approach the manifold M at a controlled rate.

4.1. *Examples of stochastic approximation for variation inclusions.* The rest of the section is devoted to examples of algorithms satisfying Assumptions D and E. In all cases, we will consider the task of solving the variational inclusion

$$0 \in A(x) + \partial g(x) + \partial f(x). \tag{4.3}$$

Here $A\colon \mathbf{R}^d \to \mathbf{R}^d$ is any single-valued continuous map, $f : \mathbf{R}^d \to \overline{\mathbf{R}}$ is a closed function, and $g : \mathbf{R}^d \to \overline{\mathbf{R}}$ is a closed function that is bounded from below. [2] As explained in the Introduction, variational inclusions encompass a variety of problems, most-notably first-order optimality conditions for nonlinear programming and Nash equilibria of games. In order to identify (4.3) with (4.1), we define the set-valued map F to be

$$F(x) := A(x) + \partial g(x) + \partial f(x).$$

Throughout, we fix a point x satisfying the inclusion (4.3).

A classical algorithm for problem (4.3) is the stochastic forward–backward iteration, which proceeds by taking "forward-steps" on $A + \partial g$ and proximal steps on f. Specifically, given a current iterate x_t, the algorithm performs the update

$$\begin{aligned} &\text{Choose } w_t \in \partial g(x_t) \\ &\text{Choose } x_{t+1} \in \operatorname{prox}_{\alpha_t f}\big(x_t - \alpha_t\big(A(x_t) + w_t + v_t\big)\big) \end{aligned}, \tag{4.4}$$

where v_t is a noise sequence. The operator $G_\alpha(x, v)$ corresponding to this algorithm is simply

$$G_\alpha(x, v) := \frac{x - s_f(x - \alpha(A(x) + s_g(x) + v))}{\alpha},$$

where $s_g(x)$ is any selection of the subdifferential $\partial g(x)$ and $s_f(x)$ is any selection of the proximal map $\operatorname{prox}_{\alpha f}(x)$. The goal of this section is to verify Assumption E for this operator under a number of reasonable assumptions on A, g, and f.

In particular, the local boundedness condition D for G is widely used in the literature, with a variety of sufficient conditions known. The following lemma describes a number of such conditions, which we will use in what follows. The proof appear in Section 9.1 of the Supplementary Material [7].

LEMMA 4.2 (Local boundedness). *Suppose that $A(\cdot)$ and $s_g(\cdot)$ are locally bounded around $\bar{x}$. Then Assumption D is guaranteed to hold in any of the following settings.*

1. *f is the indicator function of a closed set X.*
2. *f is convex and the function $x \to \operatorname{dist}(0, \partial f(x))$ is bounded on $\operatorname{dom} f$ near $\bar{x}$.*
3. *f is Lipschitz continuous on $\operatorname{dom} g \cap \operatorname{dom} f$.*

We next verify Assumption E in a number of reasonable settings; all proofs appear in the Supplementary Material [7]. In particular, it will be useful to note the following expression for F_M. We will use this lemma throughout the section, without explicit reference.

[2] In particular, $\operatorname{prox}_{\alpha f}(x)$ is nonempty for all $x \in \mathbf{R}^d$ and all $\alpha > 0$.

LEMMA 4.3 (Local tangent reduction). *Suppose that f and g are Lipschitz continuous on their domains, A is C^p-smooth, $f + g$ admits an active C^{p+1} manifold at x for $-A(x)$, and f and g are both C^{p+1}-smooth and strongly (a) regular along M near x. Then Assumption C holds and F_M admits the simple form*

$$F_M(x) = P_{T_M(x)}\, A(x) + \nabla_M g(x) + \nabla_M f(x), \tag{4.5}$$

for all $x \in M$ near x.

4.1.1. *Stochastic forward algorithm ($f = 0$).* We begin with the simplest case of (4.3) where $f = 0$. In this case, the iteration (4.2) reduces to a pure stochastic forward algorithm and the map G takes the simple form

$$G_\alpha(x, v) := A(x) + s_g(x) + v,$$

which is independent of α. Let us introduce the following assumption on the problem data.

ASSUMPTION F (Assumptions for the forward algorithm). *Suppose that $f = 0$ and that both $g(\cdot)$ and $A(\cdot)$ are Lipschitz continuous around $\bar{x}$. Suppose that $M \subseteq X$ is a C^p-smooth manifold for g at $\bar{x}$.*

(F1) *(Strong (a)) The function g is strongly (a)-regular along M at $\bar{x}$.*

(F2) *(Proximal aiming) There exists $\mu > 0$ such that the inequality holds:*

$$\langle A(\bar{x}) + v, x - P_M(x) \rangle \geq \mu \cdot \operatorname{dist}(x, M) \quad \text{for all } x \text{ near } \bar{x} \text{ and } v \in \partial g(x). \tag{4.6}$$

Note that Corollary 2.4 shows that the aiming condition (F2) holds as long as the inclusion $-A(\bar{x}) \in \hat{\partial} g(\bar{x})$ holds, M is an active manifold for g at $\bar{x}$ for $v = -A(\bar{x})$, and g is $(b_\leq)$-regular along M at $\bar{x}$. The following proposition shows that Assumption F suffices to ensure Assumption E. The proof appears in the Supplementary Material.

PROPOSITION 4.4 (Forward method). *Assumption F implies Assumption E.*

The following is now immediate.

COROLLARY 4.5 (Active manifolds). *Suppose $f = 0$ and that both $g(\cdot)$ and $A(\cdot)$ are Lipschitz continuous around $\bar{x}$. Suppose moreover that the inclusion $-A(\bar{x}) \in \hat{\partial} g(\bar{x})$ holds, that g admits a C^2 active manifold around $\bar{x}$ for $\bar{v} = -A(\bar{x})$, and that g is both $(b_\leq)$-regular and strongly (a)-regular along M at $\bar{x}$. Then Assumption E holds.*

4.1.2. *Stochastic projected forward algorithm ($f = \delta_X$).* Next, we focus on the particular instance of (4.3) where f is an indicator function of a closed set X. In this case, the iteration (4.2) reduces to a stochastic projected forward algorithm and the map G takes the form

$$G_\alpha(x, v) := \frac{x - s_X(x - \alpha(A(x) + s_g(x) + v))}{\alpha},$$

where $s_X(x)$ is a selection of the projection map $P_X(x)$. In order to ensure Assumption E for the stochastic projected forward method, we introduce the following assumption.

ASSUMPTION G (Assumptions for the projected gradient mapping). *Suppose that f is the indicator function of a closed set X and both $g(\cdot)$ and $A(\cdot)$ are Lipschitz continuous around $\bar{x}$. Let $M \subseteq X$ be a C^2 manifold containing $\bar{x}$ and suppose that f is C^2 on M near $\bar{x}$.*

(G1) *(Strong (a)) The function* g *and set* X *are strongly* (a)*-regular along* M *at* $\bar{x}$.

(G2) *(Proximal aiming) There exists* $\mu > 0$ *such that the inequality holds*

$$\langle A(\bar{x}) + v, x - P_M(x) \rangle \geq \mu \cdot \operatorname{dist}(x, M) \quad \forall x \in X \text{ near } \bar{x} \text{ and } v \in \partial g(x). \tag{4.7}$$

(G3) *(Condition (b)) The set* X *is* $(b_{\leq})$*-regular along* M *at* $\bar{x}$.

Note that a similar argument as Corollary 2.4 shows that the aiming condition (G2) holds as long as the inclusion $-A(\bar{x}) \in \hat{\partial}(g + f)(\bar{x})$ holds, M is an active manifold of $g + f$ at $\bar{x}$ for $\bar{v} = -A(\bar{x})$, and g is $(b_{\leq})$-regular along M at $\bar{x}$.

The following proposition shows that Assumption G is sufficient to ensure Assumption E.

PROPOSITION 4.6 (Projected forward method). *If Assumptions* D *and* G *hold, then so does Assumption* E.

The following is now immediate.

COROLLARY 4.7 (Active manifolds). *Suppose that* f *is the indicator function of a closed set* X *and both* $g(\cdot)$ *and* $A(\cdot)$ *are Lipschitz continuous around* $\bar{x}$. *Suppose moreover the inclusion* $-A(\bar{x}) \in \hat{\partial}(g + f)(\bar{x})$ *holds,* $g + f$ *admits a* C^2 *active manifold around* $\bar{x}$ *for the vector* $\bar{v} = -A(\bar{x})$, *and both* g *and* f *are* $(b_{\leq})$*-regular and strongly* (a)*-regular along* M *at* $\bar{x}$. *Then Assumption* E *holds.*

4.1.3. *Stochastic forward–backward method* $(g = 0)$. Finally, we focus on the particular instance of (4.3) where $g = 0$. In this case, the iteration (4.2) reduces to a stochastic forward–backward algorithm and the map G becomes

$$G_\alpha(x, v) := \frac{x - s_f(x - \alpha(A(x) + v))}{\alpha}.$$

In order to ensure Assumption E for the stochastic proximal gradient method, we introduce the following assumptions.

ASSUMPTION H (Assumptions for the forward–backward method). *Suppose* $g = 0$ *and* $f(\cdot)$ *and* $A(\cdot)$ *are Lipschitz continuous on* $\operatorname{dom} f$ *near* $\bar{x}$. *Suppose moreover that there exists a* C^2 *manifold* $M \subset X$ *containing* $\bar{x}$ *and such that* f *is* C^2*-smooth on* M *near* $\bar{x}$.

(H1) *(Strong (a)) The function* f *is strongly* (a)*-regular along* M *at* $\bar{x}$.

(H2) *(Proximal Aiming) There exists* $\mu > 0$ *such that the inequality*

$$\langle A(\bar{x}) + v, x - P_M(x) \rangle \geq \mu \cdot \operatorname{dist}(x, M) - \|1 + v\| \, o(\operatorname{dist}(x, M)) \tag{4.8}$$

holds for all $x \in \operatorname{dom} f$ *near* $\bar{x}$ *and* $v \in \partial f(x)$.

Note that Corollary 2.4 shows that the aiming condition (H2) holds as long as the inclusion $-A(\bar{x}) \in \hat{\partial} f(\bar{x})$ holds, M is an active manifold for f at $\bar{x}$ for $\bar{v} = -A(\bar{x})$, and f is $(b_{\leq})$-regular along M at $\bar{x}$.

The following proposition shows that Assumption H is sufficient to ensure Assumption E.

PROPOSITION 4.8 (Forward–backward method). *If Assumptions* D *and* H *hold, then so does Assumption* E.

The following is now immediate.

COROLLARY 4.9 (Active manifolds). *Suppose* $g = 0$ *and both* f *and* $A(\cdot)$ *are Lipschitz continuous on* $\operatorname{dom} f$ *near* $\bar{x}$. *Suppose moreover the inclusion* $-A(\bar{x}) \in \hat{\partial} f(\bar{x})$ *holds. Suppose that* f *admits a* C^2 *active manifold around* $\bar{x}$ *for* $\bar{v} = -A(\bar{x})$ *and* f *is both* $(b)_{\leq}$*-regular and strongly* (a)*-regular along* M *at* $\bar{x}$. *Then Assumption* E *holds.*

5. Asymptotic normality. Next, we impose two assumptions on the step-size α_k and the noise sequence ν_k. The first is standard, and is summarized next.

ASSUMPTION I (Standing assumptions). *Assume the following.*

(I1) *The map* G *is measurable.*

(I2) *There exist constants* $c_1, c_2 > 0$ *and* $\gamma \in (1/2, 1]$ *such that*

$$\frac{c_1}{k^\gamma} \le \alpha_k \le \frac{c_2}{k^\gamma}.$$

(I3) $\{\nu_k\}$ *is a martingale difference sequence w.r.t. to the increasing sequence of* σ*-fields*

$$\mathcal{F}_k = \sigma(x_j : j \le k \text{ and } \nu_j : j < k),$$

and there exists a function $q : \mathbb{R}^d \to \mathbb{R}_+$ *that is bounded on bounded sets with*

$$\mathbb{E}[\nu_k | \mathcal{F}_k] = 0 \quad \text{and} \quad \mathbb{E}\|\nu_k\|^4 | \mathcal{F}_k < q(x_k).$$

We let $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot | \mathcal{F}_k]$ *denote the conditional expectation.*

(I4) *The inclusion* $x_k \in \operatorname{dom} F$ *holds for all* $k \ge 1$.

All items in Assumption I are standard in the literature on stochastic approximation methods and mirror for example those found in [8], Assumption C. The only exception is the fourth moment bound on $\|\nu_k\|$, which stipulates that ν_k has slightly lighter tails. This bound appears to be necessary for the setting we consider.

To prove our asymptotic normality results, we impose a further assumption on the noise sequence ν_k, which also appears in [12], Assumption D. Before stating it, as motivation, consider the stochastic variational inequality (4.3) given by:

$$0 \in A(x) + \partial f(x) + \partial g(x) \qquad \text{where } A(x) = \mathop{\mathbb{E}}_{z \sim P} A(x, z).$$

Then the noise ν_k in the algorithm (4.4) takes the form

$$\nu_k = A(x_k; z_k) - A(x_k).$$

Equivalently, we may decompose the right-hand side as

$$\nu_k = \underbrace{A(\bar{x}; z_k) - A(\bar{x})}_{=:\nu_k^{(1)}} + \underbrace{A(x_k; z_k) - A(\bar{x}; z_k) + A(\bar{x}) - A(x_k)}_{=:\nu_k^{(2)}(x_k)}.$$

The two components $\nu_k^{(1)}$ and $\nu_k^{(2)}(x_k)$ are qualitatively different in the following sense. On one hand, the sum $\frac{1}{\sqrt{k}} \sum_{i=1}^{k} \nu_i^{(1)}$ clearly converges to a zero-mean normal vector as long as the covariance $\operatorname{Cov}(A(\bar{x}, z))$ exists. On the other hand, $\nu_k^{(2)}(x_k)$ is small in the sense that $\mathbb{E}_k\|\nu_k^{(2)}(x_k)\|^2 \le 2 \cdot \mathbb{E}_z[L(z)^2] \cdot \|x_k - \bar{x}\|^2$, where $L(z)$ is a Lipschitz constant of $A(\cdot, z)$. With this example in mind, we introduce the following assumption on the noise sequence.

ASSUMPTION J. *Fix a point* $\bar{x} \in \operatorname{dom} F$ *at which Assumption C holds and let* U *be a matrix whose column vectors form an orthogonal basis of* $T_M(\bar{x})$. *Recall that* $\mathbb{E}_k[\cdot]$ *denote the conditional expectation with respect to* $\mathcal{F}_k$. *We suppose the noise sequence has decomposable structure* $\nu_k = \nu_k^{(1)} + \nu_k^{(2)}(x_k)$, *where* $\nu_k^{(2)} : \operatorname{dom} F \to \mathbb{R}^d$ *is a random function satisfying*

$$\mathbb{E}_k\|U^* \nu_k^{(2)}(x)\|^2 \le C\|x - \bar{x}\|^2 \quad \text{for all } x \in \operatorname{dom} F \text{ near } \bar{x},$$

and some $C > 0$. *In addition, we suppose that for all* $x \in \operatorname{dom} F$, *we have* $\mathbb{E}_k[v_k^{(1)}] = \mathbb{E}_k[v_k^{(2)}(x)] = 0$ *and the following limit holds*:

$$\frac{1}{\sqrt{k}} \sum_{i=1}^{k} U^\top v_i^{(1)} \xrightarrow{D} N(0, U^\top \Sigma U)$$

for some symmetric positive semidefinite matrix Σ.

Note that Assumption J only requires that $v_k^{(1)}$ and $v_k^{(2)}$ have zero conditional mean, which is weaker than being independent of the previous iterates. We are now ready to state the main result of this work—asymptotic normality for stochastic approximation algorithms.

THEOREM 5.1 (Asymptotic normality). *Suppose that Assumption C, D, E, I, and J hold. Suppose that* $\gamma \in (\frac{1}{2}, 1)$ *and that the sequence* x_k *generated by the process* (4.2) *converges to* $\bar{x}$ *with probability one. Suppose that there exists a constant* $\mu > 0$ *satisfying*

$$\nabla_M F_M(\bar{x})v, v \geq \mu \|v\|^2 \quad \text{for all } v \in T_M(\bar{x}). \tag{5.1}$$

Then F *is* C^p*-smoothly invertible around* $(\bar{x}, 0)$ *with inverse* $\sigma(\cdot)$ *and the average iterate* $\bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$ *admits the expansion*

$$\sqrt{k}(\bar{x}_k - \bar{x}) = -\frac{1}{\sqrt{k}} \sum_{i=1}^{k} U \left[U^\top \nabla_M F_M(\bar{x}) U\right]^{-1} U^\top v_i^{(1)} + o_P(1),$$

and hence

$$\sqrt{k}(\bar{x}_k - \bar{x}) \xrightarrow{D} N\left(0, \nabla\sigma(0) \cdot \Sigma \cdot \nabla\sigma(0)^\top\right).$$

Moreover, $\nabla\sigma(0)$ *can be equivalently written as* $\nabla\sigma(0) = (P_{T_M(\bar{x})} \nabla_M F_M(\bar{x}) P_{T_M(\bar{x})})^\dagger$.

The conclusion of this theorem is surprising: although the sequence x_k never reaches the manifold, the limiting distribution of $\sqrt{k}(\bar{x}_k - \bar{x})$ is supported on the tangent space $T_M(\bar{x})$. Thus asymptotically, the "directions of nonsmoothness," which are normal to M, are quickly "averaged out." When $G_{\alpha_k}(x_k, v_k)$ is bounded away from 0 for all k, this means that x_k must oscillate across the manifold, instead of approaching it from one direction.

5.1. *Asymptotic normality in nonlinear programming.* As a simple illustration of Theorem 5.1, we now spell out the consequence for the stochastic projected gradient method for stochastic nonlinear programming, already discussed in Example 1. Namely, consider the problem (2.7) and let $\bar{x}$ be a local minimizer. Suppose that g_i are C^3-smooth near $\bar{x}$ and f takes the form $f(x) = \mathbb{E}_{z\sim P} f(x, z)$ for some probability distribution P and each function $f(\cdot, z)$ is C^3-smooth near $\bar{x}$. Consider the following stochastic projected gradient method:

$$\begin{aligned} &\text{Sample: } z_k \sim P, \\ &\text{Update: } x_{k+1} \in P_X\left(x_k - \alpha_k \nabla f(x_k; z_k)\right). \end{aligned} \tag{5.2}$$

In order to understand the asymptotics of the algorithm, as in Example 1, let $\bar{y}$ be the Lagrange multiplier vector and suppose that LICQ and strict complementarity holds. Suppose moreover the second-order sufficient conditions: there exists $\mu > 0$ such that

$$w^\top \nabla^2_{xx} L(\bar{x}, \bar{y})\, w \geq \mu \|w\|^2 \quad \text{for all } w \in T_M(\bar{x}). \tag{5.3}$$

Note that, as explained in Example 3, this condition is simply the requirement that the covariant Hessian of $f := f_0 + \delta_X$

$$\nabla^2_M f(\bar{x}) = P_{T_M(\bar{x})} \nabla^2_{xx} L(x, y) P_{T_M(\bar{x})}$$

is positive definite on $T_M(\bar{x})$. Finally, to ensure our noise sequence

$$\begin{aligned} v_k &= \nabla f(x_k; z_k) - \nabla f(x_k) \\ &= \underbrace{\nabla f(\bar{x}; z_k)}_{=:v_k^{(1)}} - \underbrace{\nabla f_0(\bar{x})}_{} + \underbrace{\nabla f(x_k; z_k) - \nabla f(\bar{x}; z_k)}_{=:v_k^{(2)}(x_k)} + \underbrace{\nabla f(\bar{x}) - \nabla f(x_k)}_{}, \end{aligned}$$

satisfies Assumptions I and J, we assume the stochasticity is sufficiently well behaved:

ASSUMPTION K (Stochastic gradients). *As a function of x, the fourth moment*

$$x \in X \to \mathsf{E}_{z\sim P} \|\nabla f(x; z) - \nabla f(x)\|^4$$

is bounded on bounded sets. Moreover, there exists $C > 0$ such that

$$\mathsf{E}_{z\sim P} \|\nabla f(x; z) - \nabla f(\bar{x}; z)\|^2 \le C\|x - \bar{x}\|^2 \quad \text{for all } x \in X.$$

Finally, the gradients $P_{T_M(\bar{x})}\nabla f(\bar{x}; z)$ have finite covariance $= \mathrm{Cov}(P_{T_M(\bar{x})}\nabla f(\bar{x}; z))$.

With these assumptions in hand, we have the following asymptotic normality result for nonlinear programming—a direct corollary of Theorem 5.1.

COROLLARY 5.2 (Asymptotic normality in nonlinear programming). *Suppose that LICQ, strict complementary, second-order sufficient conditions, and Assumption K hold. Suppose that $\gamma \in (\frac{1}{2}, 1)$ and consider the iterates x_k generated by the stochastic projected gradient method* (5.2). *Then if x_k converges to $\bar{x}$ with probability* 1, *the average iterate $\bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$ admits the expansion*

$$\sqrt{k}(\bar{x}_k - \bar{x}) = -\frac{1}{\sqrt{k}} \sum_{i=1}^{k} U\,U^{\top}\nabla^2_{xx}L(\bar{x}, \bar{y})U^{-1}U^{\top}v_i^{(1)} + o_P(1),$$

where the columns of U form an orthonormal basis of $T_M(\bar{x})$. Consequently, asymptotic normality holds:

$$\sqrt{k}(\bar{x}_k - \bar{x}) \xrightarrow{d} N\big(0, \nabla\sigma(0) \cdot \mathrm{Cov}(\nabla f(\bar{x}; z)) \cdot \nabla\sigma(0)\big),$$

where $\nabla\sigma(0) = (P_{T_M(\bar{x})}\nabla^2_{xx}L(\bar{x}, \bar{y})P_{T_M(\bar{x})})^{\dagger}$.

As stated in the Introduction, this appears to be the first asymptotic normality guarantee for the standard stochastic projected gradient method in general nonlinear programming problems with C^3 data, even in the convex setting. Finally, we note that even for simple optimization problems, dual averaging procedures can achieve suboptimal convergence [12]. This is surprising since such methods identify the active manifold [17] (also see [12], Section 4.1), while projected stochastic gradient methods do not.

EXAMPLE 5. It is instructive to look at three problem formulations for sparse recovery:

(regularized) $$\min_x \mathsf{E}\, f(x, z) + \lambda\|x\|_1,$$

(l_1 constraint) $$\min_{\|x\|_1 \le A} \mathsf{E}\, f(x, z),$$

(l_0 constraint) $$\min_{|\mathrm{supp}(x)| \le s} \mathsf{E}\, f(x, z).$$

Problem (regularized) is typically solved by the stochastic proximal algorithm, while (l_1 constraint) and (l_0 constraint) are solved by the stochastic projected gradient method.

Both methods are trivially examples of the algorithm (4.4) that we have studied in the section. Let us now look at the asymptotic covariance of these methods corresponding to the three problems. To this end, let x denote the optimal solution for the three problems and suppose that $x_1 = A$ and $|\mathrm{supp}(x)| = s$. Without loss of generality suppose moreover $\mathrm{supp}(x) = \{1, \ldots, s\}$. In all cases, under the regularity conditions discussed in the section, the asymptotic covariance of the average iterate is

$$\nabla\sigma(0) \cdot \mathrm{Cov}\,\nabla f(x, z) \cdot \nabla\sigma(0).$$

Thus the only distinction is in Jacobian of the solution map $\nabla\sigma(0)$. It is straightforward to see that the active manifold (under strict complementarity) for (regularized) and (l_0 constraint) is

$$M_{1,3} = \mathbb{R}^s \times \{0\}^{d-s},$$

while the active manifold for (l_1 constraint) is

$$M_2 = M_{1,3} \cap \Big\{x : \sum_{i=1}^{s} |x_i| = A\Big\}.$$

Because the active manifold M in all cases is piecewise linear, an application of Theorem 2.7 yields the expression:

$$\nabla\sigma(0) = P_{T_M(x)}\mathrm{E}\,\nabla^2 f(x, z)\, P_{T_M(x)}.$$

For the problem (regularized) and (l_0 constraint), the tangent space is simply $T_{M_{1,3}}(x) = \mathbb{R}^s \times \{0\}^{d-s}$, while for ($l_1$ constraint), the tangent space is smaller

$$T_{M_2}(x) = \Big\{v \in T_{M_{1,3}}(x) : \sum_{i=1}^{s} \mathrm{sign}(x_i)\, v_i = 0\Big\}.$$

In particular, the asymptotic covariance corresponding to (l_1 constraint) is no larger in the Loewner order than that of (regularized) and (l_0 constraint). Consequently, the formulation (l_1 constraint) may be preferable when $A = x_1$ is known. The caveat, however, is that LASSO solution is biased due to the regularization term and therefore a comparison of the three formulation purely based on the asymptotic covariance is not entirely justified.

6. Asymptotic optimality of SAA and SFB. In this section, we show that the asymptotic covariance in (3.4) is the best possible among all estimators of $\bar{x}$, and therefore both SAA and SFB are asymptotically optimal. Namely, we will lower-bound the performance of *any* estimation procedure for finding a solution of an adversarially-chosen sequence of small perturbations of the target problem. In order to specify this sequence, define the set

$$G := \Big\{g : Z \to \mathbb{R}^d : \mathop{\mathrm{E}}_{z\sim P} g(z) = 0,\ \mathop{\mathrm{E}}_{z\sim P} \|g(z)\|^2 < \infty\Big\}.$$

Fix now a function $g \in G$ and an arbitrary C^3-smooth function $h\colon \mathbb{R} \to [-1, 1]$ such that its first three derivatives are bounded and $h(t) = t$ for all $t \in [-1/2, 1/2]$. Now for each $u \in \mathbb{R}^d$, define a new probability distribution D^u whose density is given by

$$dP_u(z) := \frac{1 + h(u^T g(z))}{C(u)}\, dP(z), \tag{6.1}$$

where $C(u)$ is the normalizing constant $C(u) := 1 + \int h(u^T g(z))\, dP(z)$. Thus each vector $u \in \mathbb{R}^d$ induces the perturbed problem

$$0 \in L(x, u) + H(x) \quad \text{where } L(x, u) = \mathop{\mathrm{E}}_{z\sim P_u} A(x, z). \tag{6.2}$$

Reassuringly, the following lemma shows that map $(x, u) \to L(x, u)$ is C^1 near $(\bar{x}, 0)$. All proofs of results in this section appear in Section 11 of the Supplementary Material [7].

LEMMA 6.1. *The map* $(x, u) \to L(x, u)$ *is* C^1 *near* $(\bar{x}, 0)$ *with partial derivatives*

$$\nabla_x L(\bar{x}, 0) = \nabla A(\bar{x}) \quad \text{and} \quad \nabla_u L(x, 0) = \mathop{\mathbb{E}}_{z\sim P} A(\bar{x}, z)g(z) \ .$$

The family of problems (6.2) would not be particularly useful if their solution would vary wildly in u. On the contrary, the following lemma shows that for all small u, each problem (6.2) admits a unique solution in U, which moreover varies smoothly in u. We will use the following standard notation. A map $\sigma(\cdot)$ is called a *localization* of a set-valued map $F(\cdot)$ around a pair $(\bar{u}, \bar{v}) \in \operatorname{gph} F$ if the two sets, $\operatorname{gph}\sigma$ and $\operatorname{gph} F$, coincide locally around $(\bar{u}, \bar{v})$.

LEMMA 6.2 (Derivative of the solution map). *The solution map*

$$S(u) = \{x : 0 \in L(x, u) + H(x)\}$$

admits a single-valued localization $s(\cdot)$ *around around* $(0, \bar{x})$ *that is differentiable at* 0 *with Jacobian*

$$\nabla s(0) = -\nabla\sigma(0) \cdot \mathop{\mathbb{E}}_{z\sim P} A(\bar{x}, z)g(z) \ .$$

In light of Lemma 6.2, for all small u, we define the solution $\bar{x}_u := s(u)$. The following theorem provides an asymptotic lower bound on the performance of any estimator when applied to the problems within our parametric family. We let $\mathbb{E}_{P_u^k}$ denote the expectation with respect to k i.i.d. observations $z_i \sim P_u$.

THEOREM 6.3 (Local minimax). *Let* $L : \mathbb{R}^d \to [0, \infty)$ *be symmetric, quasiconvex, and lower semicontinuous, let* $x_k : Z^k \to U$ *be a sequence of estimators, and set* $g(z) := A(\bar{x}, z) - A(\bar{x})$. *Then the inequality*

$$\sup_{I \subset \mathbb{R}^d,\ |I| < \infty} \liminf_{k\to\infty} \max_{u\in I} \mathbb{E}_{P_{u/\sqrt{k}}^k} L\left(\sqrt{k}(x_k - \bar{x}_{u/\sqrt{k}})\right) \ge \mathbb{E}\, L(Z) \tag{6.3}$$

holds, where $Z \sim N(0, \nabla\sigma(0) \cdot \operatorname{Cov}(A(\bar{x}, z)) \cdot \nabla\sigma(0))$.

In particular, applying Theorem 6.3 with quadratics L yields a lower bound on the achievable covariance among any estimator. We will now show that both SAA and SFB fulfill (6.3) with equality, and therefore in a precise sense *asymptotically minimax optimal*. Note that we already know that the asymptotic covariance $\nabla\sigma(0) \cdot \operatorname{Cov}(A(\bar{x}, z)) \cdot \nabla\sigma(0)$ is achieved by both SAA (Theorem 3.1) and SFB (Theorem 5.1) when applied to the *fixed problem* $u = 0$. It remains therefore to argue that $\sqrt{k}(x_k - \bar{x}_u)$ along the perturbed sequence of problems is asymptotically independent of u. This is the content of the following theorem.

THEOREM 6.4 (Tightness of SAA). *Under the same assumptions as Theorem 3.1, the sample average approximation estimator* $x_k := x_k$ *satisfies* (6.3) *with equality for any bounded continuous function* $L : \mathbb{R}^d \to [0, \infty)$.

SFB enjoys completely analogous results, which we summarize next.

THEOREM 6.5 (Tightness of SFB). *Suppose the same setting as Theorem 5.1 and that* $v_i^{(1)} = A(\bar{x}, z_i) - A(\bar{x})$ *with* $z_i \sim^{\text{iid}} P$ *and such that* $\mathbb{E}\, v_i^{(1)\,2} < \infty$. *Then the average iterate* $x_k = \frac{1}{k}\sum_{i=1}^{k} x_i$ *satisfies* (6.3) *with equality for any bounded continuous function* $L : \mathbb{R}^d \to [0, \infty)$.

THEOREM 6.6 (Tightness of SFB for nonlinear programming). *Under the same assumptions as Corollary 5.2, the average iterate* $\bar{x}_k := \frac{1}{k}\sum_{i=1}^{k} x_i$ *satisfies* (6.3) *with equality for any bounded continuous function* $L : \mathbb{R}^d \to [0, \infty)$.

We note that asymptotic optimality of SFB for smooth problems was proved in [4], Theorem 5.6, and the proof we present of the three theorems above is an adaptation of the argument therein.

Acknowledgments. The authors thank John Duchi for insightful comments about the paper and encouraging us to prove that SFB attains the asymptotic lower bound in (6.3).

Funding. Research of D. Davis supported by an Alfred P. Sloan research fellowship and NSF Grant DMS-2047637.

Research of D. Drusvyatskiy was supported by NSF Grants DMS-1651851, NSF DMS-2306322, and CCF-2023166.

SUPPLEMENTARY MATERIAL

Supplement to "Asymptotic normality and optimality in nonsmooth stochastic approximation" (DOI: 10.1214/24-AOS2401SUPP; .pdf). Supplementary information.

REFERENCES

[1] BORWEIN, J. and LEWIS, A. S. (2017). *Convex Analysis and Nonlinear Optimization*: *Theory and Examples*. Springer, Berlin.

[2] BOUMAL, N. (2023). *An Introduction to Optimization on Smooth Manifolds*. Cambridge Univ. Press, Cambridge. MR4533407

[3] CLARKE, F. H., LEDYAEV, Y. S., STERN, R. J. and WOLENSKI, P. R. (1998). *Nonsmooth Analysis and Control Theory*. *Graduate Texts in Mathematics* **178**. Springer, New York. MR1488695

[4] CUTLER, J., DÍAZ, M. and DRUSVYATSKIY, D. (2024). Stochastic approximation with decision-dependent distributions: Asymptotic normality and optimality. *J. Mach. Learn. Res.* **25** Paper No. 90, 49. MR4749126

[5] DAVIS, D., DRUSVYATSKIY, D. and CHARISOPOULOS, V. (2019). Stochastic algorithms with geometric step decay converge linearly on sharp functions. arXiv preprint. Available at arXiv:1907.09547.

[6] DAVIS, D., DRUSVYATSKIY, D. and JIANG, L. (2021). Active manifolds, stratifications, and convergence to local minima in nonsmooth optimization. arXiv preprint. Available at arXiv:2108.11832v2.

[7] DAVIS, D., DRUSVYATSKIY, D. and JIANG, L. (2024). Supplement to "Asymptotic normality and optimality in nonsmooth stochastic approximation." https://doi.org/10.1214/24-AOS2401SUPP

[8] DAVIS, D., DRUSVYATSKIY, D., KAKADE, S. and LEE, J. D. (2020). Stochastic subgradient method converges on tame functions. *Found. Comput. Math.* **20** 119–154. MR4056927 https://doi.org/10.1007/s10208-018-09409-5

[9] DONTCHEV, A. L. and ROCKAFELLAR, R. T. (2009). *Implicit Functions and Solution Mappings*: *A view from variational analysis*. *Springer Monographs in Mathematics*. Springer, Dordrecht. MR2515104 https://doi.org/10.1007/978-0-387-87821-8

[10] DRUSVYATSKIY, D., IOFFE, A. D. and LEWIS, A. S. (2016). Generic minimizing behavior in semialgebraic optimization. *SIAM J. Optim.* **26** 513–534. MR3461323 https://doi.org/10.1137/15M1020770

[11] DRUSVYATSKIY, D. and LEWIS, A. S. (2014). Optimality, identifiability, and sensitivity. *Math. Program.* **147** 467–498. MR3258532 https://doi.org/10.1007/s10107-013-0730-4

[12] DUCHI, J. C. and RUAN, F. (2021). Asymptotic optimality in stochastic optimization. *Ann. Statist.* **49** 21–48. MR4206668 https://doi.org/10.1214/19-AOS1831

[13] DUPAČOVÁ, J. and WETS, R. (1988). Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems. *Ann. Statist.* **16** 1517–1549. MR0964937 https://doi.org/10.1214/aos/1176351052

[14] KING, A. J. and ROCKAFELLAR, R. T. (1993). Asymptotic theory for solutions in statistical estimation and stochastic programming. *Math. Oper. Res.* **18** 148–162. MR1250111 https://doi.org/10.1287/moor.18.1.148

[15] LE CAM, L. and YANG, G. L. (2000). *Asymptotics in Statistics*: *Some Basic Concepts*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR1784901 https://doi.org/10.1007/978-1-4612-1166-2
[16] LEE, J. M. (2013). Smooth manifolds. In *Introduction to Smooth Manifolds*, 2nd ed. *Graduate Texts in Mathematics* **218** 1–31. Springer, New York. MR2954043
[17] LEE, S. and WRIGHT, S. J. (2012). Manifold identification in dual averaging for regularized stochastic online learning. *J. Mach. Learn. Res.* **13** 1705–1744. MR2956341
[18] LEMARÉCHAL, C., OUSTRY, F. and SAGASTIZÁBAL, C. (2000). The $\mathcal{U}$-Lagrangian of a convex function. *Trans. Amer. Math. Soc.* **352** 711–729. MR1487623 https://doi.org/10.1090/S0002-9947-99-02243-6
[19] LEWIS, A. S. (2002). Active sets, nonsmoothness, and sensitivity. *SIAM J. Optim.* **13** 702–725. MR1972212 https://doi.org/10.1137/S1052623401387623
[20] LEWIS, A. S. and MALICK, J. (2008). Alternating projections on manifolds. *Math. Oper. Res.* **33** 216–234. MR2393548 https://doi.org/10.1287/moor.1070.0291
[21] MIFFLIN, R. and SAGASTIZÁBAL, C. (2005). A *VU*-algorithm for convex minimization. *Math. Program.* **104** 583–608. MR2179252 https://doi.org/10.1007/s10107-005-0630-3
[22] MILLER, S. A. and MALICK, J. (2005). Newton methods for nonsmooth convex minimization: Connections among $\mathcal{U}$-Lagrangian, Riemannian Newton and SQP methods. *Math. Program.* **104** 609–633. MR2179253 https://doi.org/10.1007/s10107-005-0631-2
[23] MORDUKHOVICH, B. S. (2006). *Variational Analysis and Generalized Differentiation. I*: *Basic Theory*. *Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **330**. Springer, Berlin. MR2191744
[24] NESTEROV, Y. (2009). Primal-dual subgradient methods for convex problems. *Math. Program.* **120** 221–259. MR2496434 https://doi.org/10.1007/s10107-007-0149-x
[25] PENOT, J.-P. (2012). *Calculus Without Derivatives*. *Graduate Texts in Mathematics* **266**. Springer, New York. MR2986672 https://doi.org/10.1007/978-1-4614-4538-8
[26] POLIQUIN, R. A. and ROCKAFELLAR, R. T. (1996). Prox-regular functions in variational analysis. *Trans. Amer. Math. Soc.* **348** 1805–1838. MR1333397 https://doi.org/10.1090/S0002-9947-96-01544-9
[27] POLYAK, B. T. and JUDITSKY, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.* **30** 838–855. MR1167814 https://doi.org/10.1137/0330046
[28] ROCKAFELLAR, R. T. and WETS, R. J.-B. (2009). *Variational Analysis* **317**. Springer, Berlin.
[29] SHAPIRO, A. (1989). Asymptotic properties of statistical estimators in stochastic programming. *Ann. Statist.* **17** 841–858. MR0994271 https://doi.org/10.1214/aos/1176347146
[30] SHAPIRO, A. (2003). On a class of nonsmooth composite functions. *Math. Oper. Res.* **28** 677–692. MR2015908 https://doi.org/10.1287/moor.28.4.677.20512
[31] VAN DER VAART, A. W. (2000). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256
[32] WRIGHT, S. J. (1993). Identifiable surfaces in constrained optimization. *SIAM J. Control Optim.* **31** 1063–1079. MR1227547 https://doi.org/10.1137/0331048
[33] XIAO, L. (2009). Dual averaging method for regularized stochastic learning and online optimization. In *Advances in Neural Information Processing Systems* **22**.