The Annals of Statistics
2021, Vol. 49, No. 1, 21–48
https://doi.org/10.1214/19-AOS1831

ASYMPTOTIC OPTIMALITY IN STOCHASTIC OPTIMIZATION

BY JOHN C. DUCHI[*] AND FENG RUAN[†]

Department of Statistics, Stanford University, [*]jduchi@stanford.edu; [†]fengruan@stanford.edu

We study local complexity measures for stochastic convex optimization problems, providing a local minimax theory analogous to that of Hájek and Le Cam for classical statistical problems. We give complementary optimality results, developing fully online methods that adaptively achieve optimal convergence guarantees. Our results provide function-specific lower bounds and convergence results that make precise a correspondence between statistical difficulty and the geometric notion of tilt-stability from optimization. As part of this development, we show how variants of Nesterov's dual averaging—a stochastic gradient-based procedure—guarantee finite time identification of constraints in optimization problems, while stochastic gradient procedures fail. Additionally, we highlight a gap between problems with linear and nonlinear constraints: standard stochastic-gradient-based procedures are suboptimal even for the simplest nonlinear constraints, necessitating the development of asymptotically optimal Riemannian stochastic gradient methods.

1. Introduction. In this paper, we consider smooth stochastic convex optimization problems of the form

$$\begin{aligned} &\underset{x}{\text{minimize}}\, f(x) := \mathbb{E}_P\big[f(x;S)\big] = \int_{\mathcal{S}} f(x;s)\,dP(s) \\ &\text{subject to } x \in \mathcal{X} := \big\{x \in \mathbb{R}^n : f_i(x) \le 0 \text{ for } i = 1,\ldots,m\big\}, \end{aligned} \tag{1}$$

where each $f_i : \mathbb{R}^n \to \mathbb{R}$ is convex and smooth ($\mathcal{C}^2$), $S \sim P$ is a random variable, and for $s \in \mathcal{S}$ the function $\mathbb{R}^n \ni x \mapsto f(x;s)$ is convex and continuously differentiable. We study algorithms that attempt to solve problem (1) using a sample $S_1, \ldots, S_k \overset{\text{iid}}{\sim} P$. In this setting, we investigate the optimality properties of stochastic optimization procedures, providing both problem-specific lower bounds on the performance of any method and giving optimal algorithms that adapt to problem structure.

Problems of the form (1) are of broad interest, as they encompass a variety of problems in statistics, machine learning and optimization [26]. Because of their wide applicability, it is important to carefully understand the difficulty of such problems. This includes understanding fundamental limits—how well the best algorithm can behave on problem (1)—as well as adaptivity, meaning the extent to which algorithms can adapt to the specific problem at hand. In this paper, we address these problems, showing function-specific difficulty measures and developing a variant of Nesterov's dual averaging algorithm [37] that is (often) optimal, though we demonstrate that alternative methods are necessary when the constraint functions f_i are nonlinear (and we provide one potential method). Unifying our results is an understanding of the stability of solutions to optimization problems under perturbations; we make precise connections between Poliquin and Rockafellar's "tilt stability" [39] and statistical and computational difficulty, giving an analogue of Fisher information for stochastic optimization problems (1).

Received August 2017; revised November 2018.
MSC2020 subject classifications. 62F10, 62F12, 68Q25.
Key words and phrases. Local asymptotic minimax theory, convex analysis, stochastic gradients, manifold identification.

A standard approach to providing optimality guarantees is the minimax risk [3, 36, 51]. Here, one defines a class $\mathcal{F}$ of functions of interest (such as Lipschitz convex functions) and measures algorithmic performance by the worst-case behavior over this function class. Minimax risk is an imprecise hammer: a function f may belong to a number of classes of functions, and the risk may differ substantially between these classes. The approach is also often too conservative: if f is decreasing quickly near the boundary of $\mathcal{X}$, it should be "easier" to solve problem (1). Hájek and Le Cam's local minimax theory [31, 51, 52] addresses these issues in classical statistical problems, giving *problem-specific* notions of difficulty and making rigorous the centrality of the Fisher information. In this paper, we build on these results to answer the following: how hard is it to solve the particular problem (1)?

The idea in this line of work (see also Zhu et al. [57]) is to define a shrinking neighborhood of problems, investigating worst-case complexity in this neighborhood. For stochastic optimization problems (1), the objective $(x, s) \mapsto f(x; s)$ is generally known, while the probability distribution P is not; with that in mind, we study neighborhoods $\mathcal{P}_k(P)$ whose elements are tilted variants $\widetilde{P}$ of the measure P satisfying $d\widetilde{P}(s) \in [1 \pm ck^{-\frac{1}{2}}]\,dP(s)$, so that $\mathcal{P}_k(P)$ shrinks to P as $k \to \infty$. Letting $\widetilde{x}$ denote the minimizer of the objective (1) when $\widetilde{P}$ replaces P and $L : \mathbb{R}^n \to \mathbb{R}$ be a loss, we consider local minimax complexity measures of the form

$$\inf_{\widehat{x}_k} \sup_{\widetilde{P} \in \mathcal{P}_k} \mathbb{E}_{\widetilde{P}}\bigl[L\bigl(\widehat{x}_k(S_1, \dots, S_k) - \widetilde{x}\bigr)\bigr], \tag{2}$$

where the expectation is taken over $S_i \overset{\text{iid}}{\sim} \widetilde{P}$. To describe our lower bound, we leverage the *tilt-stability* of an optimization problem [39], which describes the changes in solutions to problem (1) when the tilt $f_v(x) := f(x) - v^T x$ replaces $f(x)$. Letting x_v denote the minimizer of $f_v(x)$ over $\mathcal{X}$, let us assume the objective (1) is smoothly tilt stable, so $x_v = x^\star + Dv + o(\|v\|)$ for some matrix D; we show (Proposition 1) the precise dependence of D on the problem (1) via the objective f, distribution P, and constraints $\mathcal{X}$. Our first main result (Theorem 1) provides a lower bound on local complexity measures of the form (2). Here, the matrix $\Gamma := D\,\mathrm{Cov}(\nabla f(x^\star; S))D$ is analogous to the classical inverse Fisher information [51], and Theorem 1 shows that $\mathbb{E}[L(Z_k)]$, $Z_k \sim \mathsf{N}(0, k^{-1}\Gamma)$ is asymptotically a lower bound for the local complexity (2).

The next question we address is whether our problem-dependent lower bounds are accurate: are there procedures that achieve these guarantees, and can we adapt to specific problem geometry? The classical sample average approximation (or empirical risk minimization) approach [47], which sets $\widehat{x}_k = \mathrm{argmin} x \in \mathcal{X}\{\frac{1}{k}\sum_{i=1}^k f(x; S_i)\}$, is one approach. As we discuss in the sequel, it is optimal and adaptive. Given the scale of many modern problems, however, it is important to develop computationally efficient online procedures. To that end, our second contribution (Sections 4 and 5) is the development of stochastic-gradient-based procedures that are (asymptotically) optimal, achieving the infimum in the local complexity (2) for smooth enough functions f.

We develop a variant of Nesterov's dual averaging [37]; we iterate

$$x_{k+1} := \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \left(\sum_{i=1}^k \alpha_i \nabla f(x_i; S_i) \right)^T x + \frac{1}{2}\|x - x_0\|_2^2 \right\}, \tag{3}$$

where α_i denotes a stepsize sequence. In the case that $\mathcal{X} = \mathbb{R}^n$, this method reduces to the stochastic gradient method, and Polyak and Juditsky [40] show that the averages $\overline{x}_k = \frac{1}{k}\sum_{i=1}^k x_i$ are asymptotically normal with the optimal covariance we derive. In contrast, we show that (i) the iteration (3) converges a.s. and identifies the active constraints in problem (1) in finite time, and (ii) as long as the constraints f_i are linear, dual averaging is optimal and adaptive (Theorems 2–4). Stochastic projected gradient descent methods *do not* enjoy these guarantees. An intriguing gap arises when the constraints are nonlinear: our proposed algorithm and classical dual averaging [37] cannot be optimal with nonlinear constraints, even for $\mathcal{X} = \{x \in \mathbb{R}^n : \|x\|_2^2 \le 1\}$. To address this, we develop an asymptotically optimal manifold-based online algorithm (Theorem 5), showing that closing this gap is possible but nontrivial.

The unifying aspect of both threads—algorithms and lower bounds—throughout this work is the geometry of the problem (1). Letting $x^\star$ denote the minimizer of the problem (our coming assumptions make this unique), we give a perturbation analysis [8] of parameterized versions of problem (1) that shows how the active constraints $\{i : f_i(x^\star) = 0\}$ affect solutions to the perturbed problems (2). A similar perturbation analysis is also central to our results on optimal constraint identification and the asymptotic covariance structure of the iterates of dual averaging (3), providing a unifying geometric theme to our results and allowing us to provide computational and optimization-based analogues of the Fisher information.

1.1. *Related work.* That problem geometry strongly influences optimization algorithms is well known. In statistics, geometric conditions involving the continuity of the estimand with respect to the underlying probability measure are central to minimax analyses [6, 16, 17], and the Fisher information characterizes classical asymptotics [31, 51, 52]. Our approach to local asymptotic minimax lower bounds builds out of the literature on semi and nonparametric efficiency [5, 28, 48, 51], where one wishes to estimate a finite-dimensional parameter of an infinite-dimensional nuisance, thus studying hardest finite-dimensional subproblems; we connect these hardest subproblems to stability in optimization. In deterministic optimization, work by Burke and Moré [12] and Wright [55] shows how projected gradient and Newton methods identify active constraints and converge quickly once identified, and such identification underlies active set methods [38].

On the algorithmic side, there is a substantial literature on stochastic approximation and optimization procedures, with growing recent importance for large-sample problems [9, 20, 30, 35, 40, 41, 56, 58]. Early works, beginning with Robbins and Monro [41] and continuing through work by (among others) Ermoliev [22, 23], Venter [53], Fabian [24], Kushner [30] and Walk [54], develop probability one convergence with and without constraints, as well as asymptotic normality results in restricted situations [24, 53]. Polyak and Juditsky [40] show the importance of averaging stochastic gradient methods with “long stepping,” establishing a generic asymptotic normality result. Our results are a natural descendant of this work, but they require new development, and given the subtleties that nonlinear constraints introduce for asymptotics, we require extensions to and connections with Riemannian methods [2, 10, 49]. Recent progress on incremental gradient methods—which approximate the population expectation (1) by an empirical average—develops efficient estimators using limited computation [14, 29, 32, 34], though the methods do not apply in fully online stochastic scenarios.

1.2. *Notation and basic definitions.* We let $\mathbb{R}_+ = \{x \in \mathbb{R} : x \ge 0\}$ and $\mathbb{R}_{++} = \{x \in \mathbb{R} : x > 0\}$. For any $m \in \mathbb{N}$, we use $[m]$ to denote the set of integers $\{1, 2, \dots, m\}$. For a set C, we use $\operatorname{relint}(C)$ to denote its relative interior [43], Section 6, and $\mathbf{I}_C(x)$ to denote the extended real valued function

$$\mathbf{I}_C(x) = \begin{cases} 0 & x \in C, \\ +\infty & x \notin C. \end{cases}$$

For a vector v, $\|v\|$ denotes its Euclidean norm. For a matrix A, $A^\dagger$ is its Moore–Penrose inverse, and $|||A||| = \sup_{\|v\|=1} \|Av\|$ is its l_2 operator norm.

2. Background and assumptions.

Before moving to our main results, we collect important assumptions, definitions and recapitulate a few results on stochastic optimization. As we view our results through the lens of stability and perturbation, we also present a perturbation result on tilt-stability of optimization problems that underpins our development.

2.1. *Main assumptions.*

We begin by formalizing the problems we consider. This involves specifying smoothness and identifiability properties on f and $x^\star$, the unique minimizer of problem (1) (our assumptions ensure uniqueness).

ASSUMPTION A. There exists $L < \infty$ such that

$$\|\nabla f(x) - \nabla f(x^\star)\| \le L\|x - x^\star\| \quad \text{for all } x \in \mathcal{X}.$$

There exist $C, \epsilon \in (0, \infty)$ such that for $x \in \mathcal{X} \cap \{x : \|x - x^\star\| \le \epsilon\}$,

$$\|\nabla f(x) - \nabla f(x^\star) - \nabla^2 f(x^\star)(x - x^\star)\| \le C\|x - x^\star\|^2.$$

Because we study perturbation of solutions and rates of convergence, we require constraint qualifications to make precise guarantees. The normal cone to the set $\mathcal{X}$ at the point x is

$$\mathcal{N}_{\mathcal{X}}(x) := \{v \in \mathbb{R}^n : \langle v, y - x\rangle \le 0 \text{ for all } y \in \mathcal{X}\}.$$

The optimality conditions for convex programming [11, 27] for problem (1) are that $x^\star$ minimizes f over $\mathcal{X}$ if and only if $-\nabla f(x^\star) \in \mathcal{N}_{\mathcal{X}}(x^\star)$. The condition that $-\nabla f(x^\star) \in \mathcal{N}_{\mathcal{X}}(x^\star)$ is insufficient for our identification and perturbation results, so we make a standard constraint qualification [12, 55] and [25], Definition 2.4. Throughout, we let m_0 be the number of *active* constraints in problem (1), that is, the number of all indices i such that $f_i(x^\star) = 0$. Without loss of generality, we assume $f_1, \dots, f_{m_0}$ are the only active constraints.

ASSUMPTION B. The vector $\nabla f(x^\star)$ satisfies

$$-\nabla f(x^\star) \in \operatorname{relint} \mathcal{N}_{\mathcal{X}}(x^\star).$$

The constraint functions $\{f_1, \dots, f_m\}$ are $\mathcal{C}^2$ near $x^\star$. Additionally, the active constraints $\{f_1, \dots, f_{m_0}\}$ satisfy either:

i. The set $\{\nabla f_i(x^\star)\}_{i=1}^{m_0}$ is linearly independent
ii. The functions f_i are affine.

Assumption B implies there exists a strictly positive $\lambda^\star \in \mathbb{R}^{m_0}_{++}$ such that

$$\nabla f(x^\star) + \sum_{i=1}^{m_0} \lambda_i^\star \nabla f_i(x^\star) = 0, \tag{4}$$

and $\lambda^\star$ is unique under Assumption B.i. This follows by standard constraint qualifications [27], Chapter VII.2, for linear or independent constraints, which implies that $\mathcal{N}_{\mathcal{X}}(x^\star) = \{\sum_{i=1}^{m_0} \lambda_i \nabla f_i(x^\star), \lambda \in \mathbb{R}^{m_0}_{+}\}$, whose relative interior is the set with λ strictly positive. The set of $\lambda \in \mathbb{R}^{m_0}_{+}$ satisfying the KKT condition $\nabla f(x^\star) + \sum_i \lambda_i \nabla f_i(x^\star) = 0$ is a compact convex polyhedron.

We require two additional assumptions on the structure of the function f. We define the critical tangent set to $\mathcal{X}$ at x by

$$\mathcal{T}_{\mathcal{X}}(x) := \{w \in \mathbb{R}^n : \nabla f_i(x)^T w = 0 \text{ for } i \in [m] \text{ s.t. } f_i(x) = 0\}. \tag{5}$$

With this definition, we make the following standard second-order sufficiency, or restricted strong convexity, assumption [18, 46, 55].

ASSUMPTION C. There exists $\mu > 0$ such that for any $w \in \mathcal{T}_{\mathcal{X}}(x^\star)$,

$$w^T \left[\nabla^2 f(x^\star) + \sum_{i=1}^{m_0} \lambda_i^\star \nabla^2 f_i(x^\star) \right] w \geq \mu \|w\|^2.$$

Assumption C guarantees the uniqueness of minimizers of the function f over $\mathcal{X}$; more, it implies f has the following growth properties.

LEMMA 2.1 (Wright [55], Theorem 3.2(i)). *Under Assumption* C, *there exists* $\epsilon > 0$ *such that*

$$\langle \nabla f(x), x - x^\star \rangle \geq f(x) - f(x^\star) \geq \epsilon \min\{\|x - x^\star\|^2, \|x - x^\star\|\} \quad \textit{for } x \in \mathcal{X}.$$

Finally, we make the standard assumption [35, 40, 42] that the noise in the functions f is not too substantial.

ASSUMPTION D. There exists $C < \infty$ such for all $x \in \mathcal{X}$,

$$\mathbb{E}[\|\nabla f(x; S) - \nabla f(x^\star; S)\|^2] \leq C \|x - x^\star\|^2.$$

The gradients $\nabla f(x^\star; S)$ have finite covariance $\Sigma := \mathrm{Cov}(\nabla f(x^\star; S))$.

We provide two remarks on Assumption D. First, Assumptions A and D, coupled with Jensen's inequality, imply that for any $x \in \mathcal{X}$ we have

$$\mathbb{E}[\|\nabla f(x; S) - \nabla f(x)\|^2] \leq \mathbb{E}[\|\nabla f(x; S)\|^2] \leq C(1 + \|x - x^\star\|^2), \tag{6}$$

where $C < \infty$ is some constant. Second, many statistical applications and stochastic programming problems, including linear and logistic regression, satisfy Assumption D. Verifying the assumptions for these is routine [40].

2.2. *Perturbation of optimal solutions and classical asymptotics.* The unifying thread throughout this work is the importance of perturbation results for optimal solutions of optimization problems, which form the building blocks of classical asymptotic results for problem (1) (cf. Shapiro [46]), for the local minimax lower bounds we develop, and for the identification and optimality results we provide for stochastic gradient-based algorithms.

With this in mind, we consider tilt-stability properties of solutions to problem (1). Tilt stability is the Lipschitz continuity of minimizers of *tilted* versions of an objective f, namely minimizers of $f_v(x) := f(x) - \langle v, x \rangle$ for v near 0; the notion has been influential in variational analysis and the development of optimization algorithms for some time [18, 19, 39]. In our case, we can provide an implicit function theorem for the KKT system associated with the optimality conditions for problem (1) under tilt-like perturbations of the objective. To make this concrete, let $v \in \mathbb{R}^n$ be a perturbation vector, and assuming that f_v is still convex, we consider approximate tilts of f satisfying

$$f_v(x) = f(x) - v^T x + c_v + o(\|v\|^2 + \|x - x_0\|^2) \tag{7}$$

for v near 0 and x near x_0, where x_0 minimizes $f_0(x)$ over $\mathcal{X}$ (i.e., $x_0 = x^\star$) and c_v depends only on v. We then consider the tilted problem

$$\text{minimize } f_v(x) \quad \text{subject to } f_i(x) \leq 0, \quad i = 1, \dots, m, \tag{8}$$

whose minimizer we denote by x_v. By assumption, the problem (8) is convex, so we equivalently assume that $\nabla_x f_v(x) = \nabla f_0(x) - v + o(\|v\| + \|x - x_0\|)$. Let $\mathcal{L}(x, \lambda) = f(x) +$

$\sum_{i=1}^{m} \lambda_i f_i(x)$ denote the Lagrangian for problem (1), and define the Hessian of the problem at optimality by

$$H^\star := \nabla_x^2 \mathcal{L}(x^\star, \lambda^\star) = \nabla^2 f(x^\star) + \sum_{i=1}^{m_0} \lambda_i^\star \nabla^2 f_i(x^\star).$$

Let $\mathsf{P}_{\mathcal{T}}$ denote the orthogonal projection onto the tangent set (5) at $x^\star$, which we recall is $\mathcal{T}_{\mathcal{X}}(x^\star) = \bigcap_{i=1}^{m_0} \{w : w^T \nabla f_i(x^\star) = 0\}$. That is, if $A \in \mathbb{R}^{m_0 \times n}$ denotes the matrix with rows $\nabla f_i(x^\star)^T$, then $\mathsf{P}_{\mathcal{T}} = I - A^T(AA^T)^\dagger A$. We then have the following perturbation result, an implicit function theorem for the KKT system generated by problem (8).

PROPOSITION 1. *Let Assumptions* A, B *and* C *hold. Assume that for any $v \in \mathbb{R}^n$, the function $f_v(x)$ is convex, and satisfies the Taylor expansion at equation* (7). *Then the minimizer x_v of equation* (8) *satisfies*

$$x_v = x_0 + \mathsf{P}_{\mathcal{T}} H^{\star\dagger} \mathsf{P}_{\mathcal{T}} v + o(\|v\|).$$

Though Proposition 1 is essentially known, because of its centrality in our development, we provide a proof based on [18], Theorem 2G.8, in Section 7.

2.3. *The classical M-estimator.* Proposition 1 underlies both achievability results for stochastic convex optimization [46] and, as we show in the sequel, local asymptotic minimax results. To illustrate, we give a heuristic sketch to show how Proposition 1 yields asymptotic normality of standard M-estimators for problem (1). Given a sample $S_1, \dots, S_k$, define

$$\widehat{x}_k \in \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \widehat{f}_k(x) := \frac{1}{k} \sum_{i=1}^{k} f(x; S_i) \right\}. \tag{9}$$

Taylor's theorem implies there are matrices $\widehat{E}_k(x)$ and $E(x)$, both $o(1)$ as $x \to x^\star$ (we assume heuristically this is uniform in k), such that

$$\nabla \widehat{f}_k(x) = \nabla \widehat{f}_k(x^\star) + (\nabla^2 \widehat{f}_k(x^\star) + \widehat{E}_k(x))(x - x^\star) \quad \text{and}$$
$$\nabla f(x) = \nabla f(x^\star) + (\nabla^2 f(x^\star) + E(x))(x - x^\star).$$

Then, defining $\widehat{v}_k = \nabla f(x^\star) - \nabla \widehat{f}_k(x^\star)$, we have that

$$\begin{aligned} \nabla \widehat{f}_k(x) &= \nabla f(x) - \widehat{v}_k + (\nabla^2 \widehat{f}_k(x^\star) - \nabla^2 f(x^\star) + \widehat{E}_k(x) - E(x))(x - x^\star) \\ &= \nabla f(x) - \widehat{v}_k + (o_p(1) + o(1)) \cdot (x - x^\star), \end{aligned}$$

where $o(1) \to 0$ as $x \to x^\star$, and the expansion (7) holds. Applying Proposition 1 yields that $\widehat{x}_k$ satisfies $\widehat{x}_k - x^\star = \mathsf{P}_{\mathcal{T}} H^{\star\dagger} \mathsf{P}_{\mathcal{T}} \widehat{v}_k + o_p(\|\widehat{v}_k\|)$, and finally noting that $\sqrt{k} \cdot \widehat{v}_k \overset{d}{\rightsquigarrow} \mathsf{N}(0, \Sigma)$ gives the following corollary.

COROLLARY 1 (Shapiro [46], Theorem 3.3). *Let Assumptions* A–D *hold and $\widetilde{x}_k \in \mathcal{X}$ satisfy $\widehat{f}_k(\widetilde{x}_k) - \inf_{x \in \mathcal{X}} \widehat{f}_k(x) = o_P(1/k)$. Then*

$$\sqrt{k}(\widetilde{x}_k - x^\star) \overset{d}{\rightsquigarrow} \mathsf{N}(0, \mathsf{P}_{\mathcal{T}} H^{\star\dagger} \mathsf{P}_{\mathcal{T}} \Sigma \mathsf{P}_{\mathcal{T}} H^{\star\dagger} \mathsf{P}_{\mathcal{T}}) \quad \textit{as } k \to \infty. \tag{10}$$

This result shows the M-estimator $\widehat{x}_k$ is asymptotically normal with the active constraints restricting (and improving) the covariance.

Corollary 1 leads to two questions. First, is the result improvable? In Section 3, we show that in a local minimax sense, the result is indeed optimal, so that it is essentially unimprovable. Second, the M-estimator (9) is not really a procedure, as it may require nontrivial computation. Because Corollary 1 allows estimators that are $o(1/k)$ accurate, recent efficient methods for minimization of finite sums using careful variance reduction and sampling techniques [14, 29, 32, 34] achieve the asymptotic normality (10), given a sample of size k, while computing $O(k \log k)$ gradients $\nabla f(x; S)$ in total (the methods require storing the entire dataset $S_1, \ldots, S_k$ and iterating through it multiple times). It is, however, not immediate that the rates (10) are achievable using online or purely stochastic gradient methods that compute a *single* stochastic gradient for each observation S_i. In Section 5, we show this is possible, developing asymptotically optimal online procedures.

3. Optimality guarantees.

With the asymptotic normality guarantee of Corollary 1, it is of interest to understand the best possible (statistical) behavior for optimization procedures. As we discuss in the Introduction, standard minimax complexity guarantees [3, 36] are too imprecise: they fail to provide guidance on specific to the problem at hand. With this in mind, we consider a local asymptotic minimax variant of problem (1). It is natural to assume that the loss $f(x; s)$ is specified—we have a way to measure performance of the decision vector x—but the distribution P may be unknown or is a nuisance parameter (we simply wish to find the minimizing x).

We thus consider the difficulty of solving problem (1) over small neighborhoods of P. To define these neighborhoods, for $d \in \mathbb{N}$, we parameterize P via a vector $u \in \mathbb{R}^d$ (where the original problem corresponds to $u = 0$ and P_0), denoting the objective of problem (1) by $f_0(x) = \mathbb{E}_{P_0}[f(x; S)]$ and its (unique) optimum by x_0. The perturbed distributions P_u dovetail with our results on stability of minimizers under tilt-perturbation (Proposition 1): in appropriate cases, we show that $f_u(x) = \mathbb{E}_{P_u}[f(x; S)] \approx f_0(x) - u^T \Sigma(x - x_0)$, where $\Sigma = \mathrm{Cov}(\nabla f(x_0; S))$. Our results elucidate the precise correspondence between tilt-stability and difficulty of stochastic optimization.

3.1. *Tilted distributions.*

To define the perturbed problems, let $h : \mathbb{R} \to [-1, 1]$ be any three-times continuously differentiable function, where

$$h(t) = t \quad \text{for } t \in [-1/2, 1/2], \tag{11}$$

the derivative $h' \geq 0$ is nonnegative, and the first three derivatives of h are bounded. (The choice $[-1/2, 1/2]$ is immaterial; any interval containing 0 on which $h(t) = t$ suffices.) Now, let

$$\mathcal{G}_d := \{g : \mathcal{S} \to \mathbb{R}^d \mid \mathbb{E}_{P_0}[g(S)] = 0, \mathbb{E}_{P_0}[\|g(S)\|^2] < \infty\}$$

(the maximal tangent set to the set of distributions on $\mathcal{S}$ at P_0, cf. [51], Chapter 25). Then for $g \in \mathcal{G}_d$ and $u \in \mathbb{R}^d$ we consider the tilted distribution

$$dP_u(s) = \frac{1 + h(u^T g(s))}{C_u} dP_0(s) \quad \text{where } C_u = 1 + \int h(u^T g(s)) \, dP_0(s). \tag{12}$$

This distribution approximates $dP_u(s) \propto e^{u^T g(s)} dP_0(s)$ as $u \to 0$, providing a slight reweighting in directions g specifies. Such tilted constructions are central to proving lower bounds for semiparametric inference problems (e.g., [28, 48] and [51], Example 25.16) where the goal is to infer a finite-dimensional parameter of a distribution P_0. The lower bound and essential geometric difficulty arise by embedding hardest one-dimensional subproblems into the broader problem. In this context, we identify the correct score (or influence) function [28, 51] for constrained stochastic optimization.

Thus, for $u \in \mathbb{R}^d$, we consider convex programs $\mathcal{P}_u$ defined by

$$\begin{aligned} &\underset{x}{\text{minimize}} \quad f_u(x) := \mathbb{E}_{P_u}[f(x; S)] = \int f(x; s) \, dP_u(s) \\ &\text{subject to} \quad f_i(x) \leq 0 \quad \text{for } i = 1, \ldots, m, \end{aligned} \tag{13}$$

letting x_u denote the minimizer of the tilted convex program (13). We develop a local asymptotic minimax theory as u varies in neighborhoods of zero of radius $\propto 1/\sqrt{k}$, where k denotes the sample size.

We require one additional assumption to show our lower bounds.

ASSUMPTION E. For P_0-almost all s, the function $f(\cdot; s)$ is $\mathcal{C}^2$ in a neighborhood of x_0. There are a remainder $\mathrm{Rem} : \mathcal{X} \times \mathcal{S} \to \mathbb{R}^{n \times n}$ and $M : \mathcal{S} \to \mathbb{R}_+$ satisfying

$$\nabla^2 f(x; s) = \nabla^2 f(x_0; s) + \mathrm{Rem}(x; s)$$

where for some $\delta > 0$,

$$\sup_{\|x - x_0\| \le \delta} |||\mathrm{Rem}(x; s)||| \le M(s) \quad \text{and} \quad \mathbb{E}_{P_0}[M(S)] < \infty.$$

Additionally, we have the following integrability conditions:

$$\mathbb{E}_{P_0}[M(S)\|\nabla f(x_0; S)\|] < \infty, \qquad \mathbb{E}_{P_0}[\|\nabla f(x_0; S)\| |||\nabla^2 f(x_0; S)|||] < \infty,$$

and for some $\delta > 0$

$$\sup_{\|x - x_0\| \le \delta} \mathbb{E}_{P_0}[|f(x; S)| \|\nabla f(x_0; S)\|^2] < \infty.$$

Note that $\mathrm{Rem}(x; s) \to 0$ as $x \to x_0$ by assumption that $f(\cdot; s)$ is $\mathcal{C}^2$.

3.2. *A local asymptotic minimax theorem.* With this assumption, we have the following theorem, which provides a local asymptotic minimax lower bound on optimization. In the theorem, we use the notation of Proposition 1, where $\mathsf{P}_{\mathcal{T}} \in \mathbb{R}^{n \times n}$ denotes the orthogonal projection onto the tangent space $\mathcal{T}$ and $H^\star = \nabla^2 f(x^\star) + \sum_{i=1}^{m_0} \lambda_i^\star \nabla^2 f_i(x^\star)$. We also recall that $L : \mathbb{R}^n \to \mathbb{R}$ is quasiconvex if for all $\alpha \in \mathbb{R}$ the sub-level sets $\{x \in \mathbb{R}^n : L(x) \le \alpha\}$ are convex, and let $\mathbb{E}_{P_u^k}$ denote expectation under k i.i.d. observations $S_i \sim P_u$.

THEOREM 1. *Let Assumptions A–E hold and let $L : \mathbb{R}^n \to \mathbb{R}$ be a symmetric quasiconvex loss. For any sequence of estimators $\widehat{x}_k : \mathcal{S}^k \to \mathbb{R}^n$,*

$$\sup_{d \in \mathbb{N}, g \in \mathcal{G}_d} \lim_{c \to \infty} \liminf_{k \to \infty} \sup_{\|u\|_2 \le c/\sqrt{k}} \mathbb{E}_{P_u^k}[L(\sqrt{k}(\widehat{x}_k - x_u))] \ge \mathbb{E}[L(Z)], \tag{14}$$

where

$$Z \sim \mathsf{N}(0, \mathsf{P}_{\mathcal{T}} H^{\star\dagger} \mathsf{P}_{\mathcal{T}} \mathrm{Cov}(\nabla f(x_0; S)) \mathsf{P}_{\mathcal{T}} H^{\star\dagger} \mathsf{P}_{\mathcal{T}}).$$

Moreover, $g(s) = \nabla f(x_0; s) - \mathbb{E}_{P_0}[\nabla f(x_0; S)]$ achieves the supremum (14).

Remarks. We provide the proof of Theorem 1 in Section 8, discussing it here. It is important that the limit in k is taken before that in c, as this provides the local nature of the result: the neighborhoods of problems, as given by the tilted distributions P_u in equation (12), have size decreasing as $O(1/\sqrt{k})$. The rescaling of the estimator error $\widehat{x}_k - x_u$ by $\sqrt{k}$ reflects our expectation that $\sqrt{k}(\widehat{x}_k - x_u)$ is $O(1)$ for good estimators $\widehat{x}_k$ by Corollary 1.

We may consider alternative choices of the neighborhood of P_0. One is to use ϕ-divergences [4, 13], where for ϕ convex with $\phi(1) = 0$, one defines

$$D_\phi(P \| Q) := \int \phi\left(\frac{dP}{dQ}\right) dQ \ge 0.$$

For example, KL-divergence has $\phi(t) = t \log t - t + 1$, the χ^2-divergence uses $\phi(t) = \frac{1}{2}(t - 1)^2$, and the squared Hellinger distance corresponds to $\phi(t) = \frac{1}{2}(\sqrt{t} - 1)^2$. It is no loss of generality to assume that $\phi'(1) = 0$ in the definition of D_ϕ, as $\phi_*(t) = \phi(t) - t\phi'(1) + \phi'(1)$ satisfies $D_\phi = D_{\phi_*}$. Consider now any ϕ-divergence with ϕ a $\mathcal{C}^2$ function in a neighborhood of 1 and $\phi''(1) > 0$. Then Lebesgue's dominated convergence theorem implies (see Section 8) that the normalization $C_u = 1 + o(\|u\|^2)$, and so

$$\begin{aligned} D_\phi(P_u \,\|\, P_0) &= \int \phi\biggl(\frac{1 + h(u^T g(s))}{C_u}\biggr) \, dP_0(s) \\ &= \frac{1}{2}\phi''(1) u^T \operatorname{Cov}(g(S)) u + o(\|u\|^2), \end{aligned}$$

where we use that $h(t) = t$ for t near 0. Replacing the supremum in the local minimax lower bound (14) by any ϕ-divergence ball, where we let x_P denote the minimizer of problem (1) with distribution P on the data S, yields the following.

COROLLARY 2. *Let the conditions of Theorem 1 hold and $\phi : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be convex. Assume that ϕ is $\mathcal{C}^2$ in a neighborhood of* 1. *Then*

$$\liminf_{c \to \infty} \liminf_{k \to \infty} \sup_{P : D_\phi(P \| P_0) \le c/k} \mathbb{E}_{P^k}\bigl[L\bigl(\sqrt{k}(\widehat{x}_k - x_P)\bigr)\bigr] \ge \mathbb{E}\bigl[L(Z)\bigr],$$

where $Z \sim \mathsf{N}(0, \mathsf{P}_{\mathcal{T}} H^{\star\dagger} \mathsf{P}_{\mathcal{T}} \Sigma \mathsf{P}_{\mathcal{T}} H^{\star\dagger} \mathsf{P}_{\mathcal{T}})$.

That is, our lower bounds imply lower bounds for natural nonparametric choices of the neighborhood of P_0.

It is possible to prove a somewhat stronger result than Theorem 1, which we do not do for simplicity, where instead of the inner supremum over all vectors u such that $\|u\|_2 \le c/\sqrt{k}$, we take an integral against the uniform measure π supported on the ball $\{u : \|u\|_2 \le c/\sqrt{k}\}$ (see the constructions in Le Cam and Yang [31], Chs. 6–7). We then have a superefficiency result [50]: if $\widehat{x}_k$ denotes an estimator based on the sample $S_1, \ldots, S_k$, the set of $u \in \mathbb{R}^d$ for problems (13) for which $\widehat{x}_k$ achieves $\limsup_k \mathbb{E}_{P_u^k}[L(\sqrt{k}(\widehat{x}_k - x_u))] < \mathbb{E}[L(Z)]$, for Z as in the theorem, has Lebesgue measure zero.

4. Convergence and manifold identification for dual averaging.

As we discuss following Corollary 1, the $\widehat{x}_k = \operatorname{argmin}_{x \in \mathcal{X}} \frac{1}{k} \sum_{i=1}^k f(x; S_i)$ achieves optimal asymptotic convergence. In this and the next section, we investigate the possibilities of efficient purely online stochastic gradient-based estimators. These have advantages—small storage space requirements, and they take a single pass through the data—that make them especially suitable for modern large-scale regimes [9, 35, 44, 58]. We study three aspects of these methods: identification of the active constraints (those i such that $f_i(x^\star) = 0$), almost sure convergence, and optimal asymptotic behavior. While stochastic gradient descent methods fail to even identify the active constraints, we develop a variant of Nesterov's dual averaging [37] that identifies active constraints in finite time and (as we show in the next section) is asymptotically optimal when the set $\mathcal{X}$ is a polytope; when the constraints are nonlinear, significant difficulties arise, which we also discuss.

We first consider the stochastic gradient method [35, 40, 41] for problem (1), to minimize $f(x)$ subject to $x \in \mathcal{X}$. This procedure requires a stochastic gradient oracle, which at each iteration provides a random vector g_k satisfying $\mathbb{E}[g_k \mid x_k] = \nabla f(x_k)$. In problem (1), drawing $S_k \sim P$ and computing $g_k = \nabla f(x_k; S_k)$ evidently satisfies this condition. Given stochastic gradients g_k, the stochastic gradient method iteratively updates

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \biggl\{ \langle g_k, x - x_k \rangle + \frac{1}{2\alpha_k} \|x - x_k\|_2^2 \biggr\}, \tag{15}$$

where $\alpha_k \propto k^{-\beta}$ for some $\beta \in [\frac{1}{2}, 1]$ is a stepsize. While the iterates (15) converge to the global optimum $x^\star$, they fail to identify optimal constraints [33]. As a simple example, we may consider a problem with $f(x) = x$ and $\mathcal{X} = [-1, 1] = \{x \mid x^2 - 1 \le 0\}$, which satisfies the assumptions of Theorem 1 and has $x^\star = -1$. Consider stochastic gradients $g_k = 1 + \xi_k$ for $\xi_k \overset{\text{iid}}{\sim} \mathsf{N}(0, 1)$; the iteration (15) satisfies $\mathbb{P}(x_k \ge -1 + \alpha_k) \ge 1 - \Phi(1)$, where Φ is the standard normal CDF. That is, $x_k \ge -1 + \alpha_k$ with constant probability at each iteration—it jumps off of the constraint infinitely often.

This instability is one of the motivations for Nesterov's dual averaging algorithm [37], which iterates

$$z_k = \sum_{i=1}^{k} g_i, \qquad x_{k+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \langle z_k, x \rangle + \frac{1}{2\alpha_k} \|x\|_2^2 \right\}. \tag{16}$$

Practically, this procedure has much better constraint identification properties [33, 56] because of the averaging effects in the definition of z_k. Xiao [56] notes its strong performance in application to ℓ_1-regularized problems, while Lee and Wright [33] give arguments showing that dual averaging spends most of its time on the "optimal manifold" for a variant of problem (1), which essentially corresponds to the set of zeros of the active constraints $\{x : f_i(x) = 0, i \in [m_0]\}$. The work [33] motivates this section, and we are able to show finite identification of the optimal constraints for a variant of the dual averaging method and its probability 1 convergence.

4.1. *Almost sure convergence.* We study a variant of dual averaging, which we view as a lazy-projected gradient algorithm, as it interpolates the stochastic gradient method and dual averaging. Given a sequence of positive stepsizes $\{\alpha_k\}_{k \in \mathbb{N}}$, initializing $z_0 = 0$, at each iteration k, we update

$$\begin{aligned} &\text{Update } x_k = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \langle z_{k-1}, x \rangle + \frac{1}{2} \|x\|_2^2 \right\} \\ &\text{Draw } S_k \overset{\text{iid}}{\sim} P, \text{ compute } g_k = \nabla f(x_k; S_k), \text{ set } z_k = z_{k-1} + \alpha_k g_k \end{aligned} \tag{17}$$

In contrast to the standard dual averaging update (16), procedure (17) constructs z_k as a weighted average and regularizes with $\frac{1}{2}\|x\|_2^2$. This has two consequences: first, in the unconstrained case, we recover the stochastic gradient method, which Polyak and Juditsky [40] show (when combined with averaging) is asymptotically normal with optimal covariance. The form (17) also allows us to prove the convergence $x_k \overset{a.s.}{\to} x^\star$ and finite time identification results. Without further comment, we assume the stepsizes α_k satisfy

$$\alpha_k = \alpha_0 k^{-\beta} \quad \text{where } \alpha_0 > 0 \text{ and } \frac{1}{2} < \beta < 1. \tag{18}$$

We may prove our results under slightly weaker conditions than the i.i.d. sampling assumed in the update (17), which we specify now for completeness. In particular, we assume that at each iteration k we observe a noisy gradient $g_k = \nabla f(x_k) + \xi_k(x_k)$, where $\xi_k : \mathcal{X} \to \mathbb{R}^n$ is a random function with the property that $\mathbb{E}[\xi_k(x)] = 0$ for all $x \in \mathcal{X}$. We make the following assumption.

ASSUMPTION D'. Define the filtration $\mathcal{F}_k := \sigma(\xi_1, \ldots, \xi_k)$. The noise ξ_k has the decomposable structure $\xi_k(x) = \xi_k^{(0)} + \xi_k^{(1)}(x)$, where $\xi_k^{(0)}$ and $\xi_k^{(1)}(x)$ are both martingale difference sequences adapted to the filtration $\mathcal{F}_k$. There exists a constant $C < \infty$ such that

$$\mathbb{E}[\|\xi_k^{(0)}\|^2 \mid \mathcal{F}_{k-1}] \le C \quad \text{and} \quad \mathbb{E}[\|\xi_k^{(1)}(x)\|^2 \mid \mathcal{F}_{k-1}] \le C\|x - x^\star\|^2.$$

Additionally, $\frac{1}{\sqrt{k}} \sum_{i=1}^{k} \xi_i^{(0)} \overset{d}{\rightsquigarrow} \mathsf{N}(0, \Sigma)$ for some $\Sigma \succeq 0$.

Assumptions A (smoothness of f) and D (variance bounds on $\nabla f(x; S)$) imply D' when $g_k = \nabla f(x_k; S_k) = \nabla f(x_k; S_k) - \nabla f(x^\star; S_k) + \nabla f(x^\star; S_k)$ as in the update (17). The additional generality causes no special difficulty in the proofs, so for the remainder of this paper we let Assumption D' hold.

We begin with the almost sure convergence of x_k. This a.s. convergence requires no constraint qualifications, just that there exists $\epsilon > 0$, such that $f(x) - f(x^\star) \ge \epsilon \|x - x^\star\|^2$ for $x \in \mathcal{X}$ near $x^\star$.

THEOREM 2. *Let x_k be generated by the dual averaging iterates* (17) *with stepsizes* (18), *let Assumptions* A *and* D' (*or* D) *hold, and let the growth condition on f in the conclusion of Lemma* 2.1 *hold. Then*

$$x_k \overset{a.s.}{\to} x^\star.$$

See Section 9.1 for a proof of the theorem.

4.2. *Constraint identification.* To segue into our results on identification of the optimal surface of the constraint set $\mathcal{X}$, note that Theorem 2 implies inactive constraints are inactive at some finite time: for some (random) $k < \infty$ we have $\sup_{l \ge k} f_i(x_l) < 0$ for $i > m_0$. Conversely, Theorem 2 says little about whether x_k identifies the constraints active at $x^\star$.

In brief, under the constraint qualifications of Assumption B, for the modified dual averaging iteration (17), there is a (random) iterate k_{ident} such that for $k \ge k_{\text{ident}}$, we have $f_i(x_k) = 0$ for $i \in [m_0]$. To provide this guarantee, we give our second set of results on perturbation of optimal solutions to convex programs, showing that solutions to linearized versions of problem (1) belong to $\{x : f_i(x) = 0, i \le m_0\}$. The linear approximation (as opposed to the quadratic approximations in Proposition 1) is a less immediate application of the results on parametrized optimization [8, 46, 55], but (nearly) linear minimization problems dovetail with the updates (17).

We give a few heuristics. Consider the problem

$$\underset{x}{\text{minimize}} \langle \nabla f(x^\star), x\rangle \quad \text{subject to } f_i(x) \le 0, \quad i = 1, \dots, m, \tag{19}$$

which has a linear objective. By Assumption B, the point $x^\star$ satisfies the KKT conditions for this problem and is optimal, but it may not be unique. The dual averaging iteration (17) eventually approximates a slightly perturbed version of the linear objective (19) because $x_k \overset{a.s.}{\to} x^\star$ and we expect $\sum_{i=1}^k \alpha_i g_i = \sum_{i=1}^k \alpha_i \nabla f(x_i) + o(\sum_{i=1}^k \alpha_i)$. This motivates the next two perturbation results, which we graphically describe in Figure 1. The intuition for each is that $-z_k$ is in $\mathcal{N}_{\mathcal{X}}(x)$ for some x near enough $x^\star$, in which case the constraint qualifications (Assumption B) imply that the projected point must lie on the set described by the active constraints at $x^\star$.

Nonlinear constraints. We begin with a perturbation result for the case in which the constraints are nonlinear, as the linear independence constraint qualification (Assumption B.i) makes the argument easier in this case. Let $x^\star$ be a point such that $f_i(x^\star) = 0$ for $1 \le i \le m_0$ and $f_i(x^\star) < 0$ for $m_0 + 1 \le i \le m$. Let $\lambda^\star \in \mathbb{R}^{m_0}$ with $\lambda^\star > 0$ be otherwise arbitrary, and define $g = -\sum_{i=1}^{m_0} \lambda_i^\star \nabla f_i(x^\star)$. Let $x_0 \in \mathbb{R}^n$, and $v \in \mathbb{R}^n$ and $\delta > 0$, and consider the tilted and quadratically perturbed version of problem (19)

$$\begin{aligned} &\underset{x}{\text{minimize}} \quad \langle g, x\rangle + \langle v, x\rangle + \frac{\delta}{2}\|x - x_0\|^2 \\ &\text{subject to} \quad f_i(x) \le 0, \quad i = 1, \dots, m. \end{aligned} \tag{20}$$

The problem (20) has a unique minimizer that we denote $x^\star_{v,\delta}$. Then we have the following lemma, whose proof we provide in the Supplementary Material [21], Section 11.1.

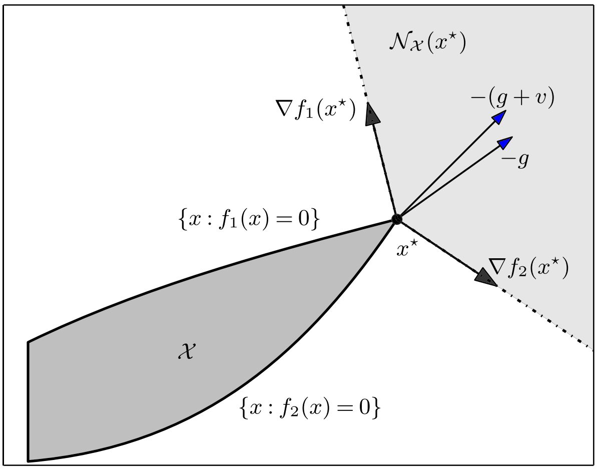

FIG. 1. *The set* $\mathcal{X} = \{x \in \mathbb{R}^2 : x_1 \ge 0, f_1(x) \le 0, f_2(x) \le 0\}$, *the top and bottom boundaries of* $\mathcal{X}$ *corresponding to* f_1 *and* f_2. *The normal cone* $\mathcal{N}_{\mathcal{X}}(x^\star)$ *is the convex hull of* $\nabla f_1(x^\star)$ *and* $\nabla f_2(x^\star)$. *The vectors* $-v$ *and its perturbation* $-(v+g)$ *both belong to* relint$\mathcal{N}_{\mathcal{X}}(x^\star)$.

LEMMA 4.1. *Let the sequence* $(v_k, \delta_k) \in \mathbb{R}^n \times \mathbb{R}_{++}$ *satisfy* $v_k \to 0$, $\delta_k \to 0$, *and that* $x_k := x^\star_{v_k,\delta_k} \to x^\star$ *as* $k \to \infty$. *Then there exists* $K < \infty$ *such that* $f_i(x_k) = 0$ *for* $i \in [m_0]$ *and* $k \ge K$.

Linear constraints. Considering linear constraints allows weaker assumptions than the case in which the constraints f_i are nonlinear. Assume that the matrix $A \in \mathbb{R}^{m_0 \times n}$ and vector $b \in \mathbb{R}^{m_0}$ represent the active constraints, while $C \in \mathbb{R}^{(m-m_0)\times n}$ and $d \in \mathbb{R}^{m-m_0}$ coincide with the inactive constraints, so that $Ax^\star = b$ and $Cx^\star < d$. Specializing the problem (19) and the tilted problem (20) to this setting, for $(v, \delta) \in \mathbb{R}^n \times \mathbb{R}_+$ we consider

$$\begin{aligned} &\underset{x}{\text{minimize}} \quad \langle g, x\rangle + \langle v, x\rangle + \frac{\delta}{2}\|x - x_0\|^2 \\ &\text{subject to} \quad Ax \le b, \qquad Cx \le d. \end{aligned} \tag{21}$$

As before, we assume that for some $\lambda^\star \in \mathbb{R}^{m_0}_{++}$ we have $g = A^T\lambda^\star$ so that $x^\star$ is a minimizer of problem (21) at $v = 0$, $\delta = 0$. The next lemma is the analogue of Lemma 4.1 for the linear case. As in Lemma 4.1, $x^\star_{v,\delta}$ denotes the unique optimum for the perturbed problem (21) with $\delta > 0$. We provide a proof of the lemma in the Supplementary Material, Section 11.2.

LEMMA 4.2. *Let the sequence* $(v_k, \delta_k) \in \mathbb{R}^n \times \mathbb{R}_{++}$ *satisfy* $v_k \to 0$, $\delta_k \to 0$, *and that* $x_k := x^\star_{v_k,\delta_k} \to x^\star$ *as* $k \to \infty$. *Then there exists* $K < \infty$ *such that* $Ax_k = b$ *for* $k \ge K$.

With the identification results provided by Lemmas 4.1 and 4.2, we can now show a result that demonstrates that our variant (17) of dual averaging identifies the optimal manifold in finite time with probability 1.

THEOREM 3. *Let Assumptions* A–D (*or* D') *hold. Then with probability one, there exists some* (*random*) $K < \infty$ *such that* $k \ge K$ *implies*

$$f_i(x_k) = 0 \quad \textit{for } i \le m_0 \quad \textit{and} \quad \sup_{k \ge K} f_i(x_k) < 0 \quad \textit{for } i > m_0.$$

We provide the proof of Theorem 3 in Section 9.2. The outline of the proof, though, is apparent from the above lemmas and Theorem 2. Letting $A_k = \sum_{i=1}^k \alpha_i$, the dual averaging iterates (17) perform the update

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \langle z_k, x \rangle + \frac{1}{2} \|x\|^2 \right\} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \langle \nabla f(x^\star) + v_k, x \rangle + \frac{1}{2A_k} \|x\|^2 \right\},$$

where $v_k = \frac{1}{A_k}(z_k - A_k \nabla f(x^\star)) = o(1)$, equivalent to problems (20) and (21).

5. Stochastic gradient procedures: Asymptotic normality.

Now that we have established that dual averaging converges almost surely and in finite time identifies constraints active at $x^\star$, we turn asymptotic normality results. We focus first on the case that the constraints are linear, where dual averaging is locally asymptotically minimax optimal. As we demonstrate, however, nonlinearity forces a departure from this optimality. Consequently, in Section 5.3 we develop a joint dual-averaging and Riemannian stochastic gradient procedure that is both online—it sequentially computes only a single gradient $\nabla f(x; S_i)$ from each observation—and asymptotically optimal.

5.1. *Dual averaging*: *Asymptotic normality.*

When the problem is unconstrained with $\mathcal{X} = \mathbb{R}^n$, Polyak and Juditsky [40] show that under our assumptions, the stochastic gradient method is asymptotically normal when combined with averaging. In the notation of Theorem 1, $\overline{x}_k = \frac{1}{k}\sum_{i=1}^k x_i$ satisfies $\sqrt{k}(\overline{x}_k - x^\star) \overset{d}{\rightsquigarrow} \mathsf{N}(0, \nabla^2 f(x^\star)^{-1} \operatorname{Cov}(\nabla f(x^\star; S)) \times \nabla^2 f(x^\star)^{-1})$, which is optimal. In the constrained case, identical results hold if we solve the problem over a subspace (i.e., $\{x : Ax = b\}$); there are no differences from the classical case [40]. We thus expect our dual averaging variant to behave as follows: eventually, we identify the active constraints, that is, we have $Ax_k = b$ and $Cx_k < d$ for all sufficiently large k. Once this occurs, the iterations of the dual averaging variant are *identical* to those of the stochastic gradient method in the subspace $\{x : Ax = b\}$. Thus, we expect asymptotic normality, with the asymptotic covariance reflecting variability only in the null space of A. While our development tracks this idea, the "sufficiently large k" for active set identification is random, and to have $Ax_k = b$ for all k depends on the entire future noise sequence $\{\xi_i\}_{i=k}^\infty$, making this intuitive argument fail. With a bit more delicacy, we can provide a similar argument that builds off of Polyak and Juditsky's treatment. Now, define the orthogonal projector onto the null space $\{w : Aw = 0\} = \mathcal{T}_{\mathcal{X}}(x^\star)$,

$$\mathsf{P}_A := I - A^T (AA^T)^\dagger A.$$

We then have the following theorem.

THEOREM 4. *Let Assumptions A–D' hold, and assume that $\alpha_k \propto k^{-\beta}$ for some $\beta \in (\frac{1}{2}, 1)$. Let $\Sigma = \operatorname{Cov}(\nabla f(x^\star; S))$. Then*

$$\frac{1}{\sqrt{k}} \sum_{i=1}^k (x_i - x^\star) \overset{d}{\rightsquigarrow} \mathsf{N}(0, \mathsf{P}_A (\nabla^2 f(x^\star))^\dagger \mathsf{P}_A \Sigma \mathsf{P}_A (\nabla^2 f(x^\star))^\dagger \mathsf{P}_A).$$

We defer the proof of Theorem 4 to the Supplementary Material, Section 14.

5.2. *Slow convergence for nonlinear constraint sets.*

Theorems 2 and 3 guarantee almost sure convergence and finite time constraint identification, but Theorem 4 provides an optimal convergence rate only when the constraints are linear, and this is fundamental. Indeed, we provide two results showing the suboptimality of dual averaging (both our variant and Nesterov's original version [37]) on a simple optimization problem.

To make this failure concrete, let e_1 be the first standard basis vector. Consider the problem (for $n \ge 2$) with $\mathcal{S} = \mathbb{R}^n$, $S \sim \mathsf{N}(0, I)$, $\mathcal{X} = \{x : \|x\|^2 - 1 \le 0\}$ and $f(x; s) = -(e_1 + s)^T x$. In this case, problem (1) becomes

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} -e_1^T x \quad \text{subject to } \|x\|_2^2 \le 1. \tag{22}$$

The optimum for program (22) is $x^\star = e_1$. The Lagrangian for the problem is $\mathcal{L}(x, \lambda) = -e_1^T x + \frac{\lambda}{2}(\|x\|_2^2 - 1)$ with optimal dual multiplier $\lambda^\star = 1$, whence Corollary 1 and the lower bound of Theorem 1 show that the optimal asymptotic covariance is $I - e_1 e_1^T$. As we show, however, dual averaging and our variant are suboptimal even with $g_k = e_1 + S_k$ for $S_k \overset{\text{iid}}{\sim} \mathsf{N}(0, I)$.

We first consider the variant (17) of dual averaging with $z_k = \sum_{i=1}^k \alpha_i g_i$.

OBSERVATION 5.1. Let the stepsizes $\alpha_i = i^{-\beta}$ for some $\beta \in (\frac{1}{2}, 1)$, and let the iterates x_k be generated by the dual averaging procedure (17). Then

$$\frac{1}{k^\beta} \sum_{i=1}^k (x_i - x^\star) \overset{d}{\rightsquigarrow} \mathsf{N}(0, \sigma_\beta^2 (I - e_1 e_1^T)) \quad \text{where } \sigma_\beta^2 := \frac{(1-\beta)^2}{\beta^2} \sum_{i=1}^\infty \alpha_i^2.$$

See the Supplementary Material, Section 12.1, for a proof. In this case, even the *rate* of convergence is lost: denoting $\overline{x}_k = \frac{1}{k}\sum_{i=1}^k x_i$, then asymptotic normality holds for $\overline{x}_k - x^\star$, but $\overline{x}_k - x^\star$ is order $k^{\beta - 1} \gg k^{-\frac{1}{2}}$,

Our second observation applies to dual averaging with $z_k = \sum_{i=1}^k g_i$.

OBSERVATION 5.2. Let the stepsize sequence $\alpha_k \propto k^{-\beta}$ for some $\beta \in [0, 1)$. Then the classical dual averaging (16) iterates satisfy

$$\frac{1}{\sqrt{k}} \sum_{i=1}^k (x_i - x^\star) \overset{d}{\rightsquigarrow} \mathsf{N}(0, 2(I - e_1 e_1^T)).$$

See the Supplementary Material, Section 12.2 for a proof.

We give a bit of intuition for the difficulty in Observations 5.1 and 5.2. We have that $\sum_{i=1}^k \alpha_i g_i = (\sum_{i=1}^k \alpha_i)\nabla f(x^\star) + \sum_{i=1}^k \alpha_i \xi_i$, where $\xi_i \overset{\text{iid}}{\sim} \mathsf{N}(0, I)$. But in projecting to the curved surface of the ball $\{x : \|x\|_2 \le 1\}$, there is still sufficient noise in the sum $\sum_{i=1}^k \alpha_i \xi_i$ to induce variance. In the case of linear constraints $Ax \le b$, the vector $z_k = \sum_{i=1}^k \alpha_i g_i$ eventually lies in the normal cone to the active face $\{x : Ax = b\}$, so that projections force all iterates into the subspace $\{x : Ax = b\}$, with no curvature for additional variance. Stochastic gradient descent—which fails to even identify the active constraints—similarly has sub-optimal rates for this problem.

5.3. *A Riemannian stochastic gradient procedure.* The challenges we outline in Section 5.2 for classical dual averaging and stochastic gradient methods necessitate alternative algorithms for asymptotically optimal online procedures. To that end, we develop an algorithm that alternates between dual averaging and a stochastic gradient method on the manifold of the active constraints. The intuition is that we use dual averaging (17) to identify the optimal manifold, then use a Riemannian stochastic gradient-like method [7, 49] on the active manifold. Letting $\mathcal{M} = \{x : f_i(x) = 0, i \in [m_0]\}$ denote the optimal manifold on which the solutions lie, two challenges arise in the analysis of any such method. First, projections onto $\mathcal{M}$ are not necessarily nonexpansive—a major component of most analyses of stochastic gradient-based methods—so that showing convergence of a pure Riemannian method is

Algorithm 1 Riemannian Stochastic Gradient with Dual Averaging

```
Initialize $k = 0$, $y_0 \in \mathcal{X}$, $\mathcal{M}_0 = \varnothing$. Input $q \in (0, 1)$, dual averaging times $T_{\mathrm{da}}$ with $1 \lesssim |\{i \in T_{\mathrm{da}} \mid i \le k\}|/k^{\rho} \lesssim 1$ for some $\rho \in (0, 1)$, and stepsizes $\alpha_k = \alpha_0 k^{-\beta}$ with $\beta \in (\frac{1}{2}, 1)$. Require that $q < \min\{\frac{1-2\beta}{(1-\beta)\rho}, \frac{1}{2(1-\rho)\beta}\}$.
for $k = 1, 2, \ldots, k \notin T_{\mathrm{da}}$ do
    Compute the manifold $\mathcal{M}_k$ that $x_k^{\mathrm{da}}$ identifies:
        $$\mathcal{M}_k = \bigcap_{i \in \mathcal{I}_k} \{x \in \mathbb{R}^n : f_i(x) = 0\} \quad \text{where } \mathcal{I}_k = \{i \in [m] : f_i(x_{t_k}^{\mathrm{da}}) = 0\}.$$
    Let $g_k = \nabla f(y_k) + \xi_k(y_k)$ and compute the iterate
        $$y_{k+1}^{\mathrm{man}} = \begin{cases} \Pi_{\mathcal{M}_k}(y_k - \alpha_k \mathsf{P}_{\mathcal{T}_{\mathcal{M}_k}(y_k)} g_k) & \text{if } \mathcal{M}_k = \mathcal{M}_{k-1} \\ x_k^{\mathrm{da}} & \text{otherwise.} \end{cases}$$
    Let $\overline{x}_k^{\mathrm{da}} := (\sum_{i \in T_{\mathrm{da}}}^{i \le k} \alpha_i^{\mathrm{da}})^{-1} \sum_{i \in T_{\mathrm{da}}}^{i \le k} \alpha_i^{\mathrm{da}} x_i^{\mathrm{da}}$.
    Let $\mathcal{B}_{k,1} = \mathcal{B}(\overline{x}_k^{\mathrm{da}}, \epsilon_k)$ and $\mathcal{B}_{k,3} = \mathcal{B}(\overline{x}_k^{\mathrm{da}}, 3\epsilon_k)$ for $\epsilon_k = (\sum_{i \in T_{\mathrm{da}}}^{i \le k} \alpha_i^{\mathrm{da}})^{-q}$. Compute
        $$y_{k+1} = \begin{cases} \Pi_{\mathcal{X}}(y_{k+1}^{\mathrm{man}}) & \text{if } \Pi_{\mathcal{X}}(y_{k+1}^{\mathrm{man}}) \in \mathcal{B}_{k,3} \\ \operatorname{argmin}\{\|x\| \mid x \in \mathcal{M}_k \cap \mathcal{X} \cap \mathcal{B}_{k,1}\} & \text{if } \Pi_{\mathcal{X}}(y_{k+1}^{\mathrm{man}}) \notin \mathcal{B}_{k,3}, \mathcal{M}_k \cap \mathcal{X} \cap \mathcal{B}_{k,1} \neq \varnothing \\ \operatorname{argmin}\{\|x\| \mid x \in \mathcal{M}_k \cap \mathcal{X}\} & \text{otherwise.} \end{cases}$$
end for
```

challenging.[1] Even in noiseless settings, gradient descent and other first-order methods do not enjoy global convergence results for minimization of convex $f : \mathbb{R}^n \to \mathbb{R}$ on Riemannian manifolds [1, 2, 10].

To that end, we present Algorithm 1, which is complex and perhaps of more intellectual than practical interest, but fulfills our desiderata of being (i) fully online, (ii) convergent with probability 1 and (iii) asymptotically optimal. To describe the algorithm and its convergence, we require somewhat more notation. For a closed set $\mathcal{M}$, let $\Pi_{\mathcal{M}}(x) = \operatorname{argmin}_{y \in \mathcal{M}}\{\|x - y\|\}$ denote the Euclidean projection of x onto $\mathcal{M}$, with an arbitrary rule for choosing the projecting if it is nonunique. When the set $\mathcal{M} = \{x \in \mathbb{R}^n : G(x) = 0\}$ for a continuously differentiable $G : \mathbb{R}^n \to \mathbb{R}^l$, we let $\nabla G(x) = [\nabla g_1(x) \cdots \nabla g_l(x)] \in \mathbb{R}^{n \times l}$ and denote the tangent space to $\mathcal{M}$ at x by

$$\mathcal{T}_{\mathcal{M}}(x) := \{v \in \mathbb{R}^n : \nabla G(x)^T v = 0\},$$

and we define the orthogonal projector

$$\mathsf{P}_{\mathcal{T}_{\mathcal{M}}(x)} = I - \nabla G(x)\big(\nabla G(x)^T \nabla G(x)\big)^{\dagger} \nabla G(x)^T \in \mathbb{R}^{n \times n}.$$

With this notation established, we can describe Algorithm 1. The algorithm alternates between asymptotically infrequent iterates of dual averaging at iterates $k \in T_{\mathrm{da}}$, constructing a sequence x_k^{da}, and frequent iterates of Riemannian stochastic gradient-like method that projects onto the active constraints, the smooth manifold $\mathcal{M}_k = \{x : f_i(x) = 0 \text{ for } i \in \mathcal{I}_k\}$ where $\mathcal{I}_k = \{i \in [m] \mid f_i(x_k^{\mathrm{da}}) = 0\}$ denotes the constraints dual averaging identifies. The method takes a stepsize sequence $\{\alpha_k\}$ for the Riemannian stochastic gradient method where

[1] Many papers on Riemannian stochastic gradient methods assume convergence, or that iterates remain in a small neighborhood of $x^{\star}$, as a condition; cf. [7] and [49], Assumption 2.

$\alpha_k = \alpha_0 k^{-\beta}$. For the dual averaging iteration times $k \in T_{\mathrm{da}}$, we set the dual averaging stepsizes via $\alpha_k^{\mathrm{da}} = \alpha_{t_k}$ where $t_k = |\{i \in T_{\mathrm{da}} \mid i \le k\}|$, the same stepsize scaling as the Riemannian method. At each step $k \in T_{\mathrm{da}}$, the method updates the dual averaging iterate via the update (17) (with $z_k = \sum_{i \in T_{\mathrm{da}}}^{i \le k} \alpha_i^{\mathrm{da}} g_i$). Then for $k \notin T_{\mathrm{da}}$, the method performs a stochastic gradient step (line 4) but projects the stochastic gradient g_k onto the tangent space of the active manifold $\mathcal{M}_k$. The final step of the algorithm (line 6) guarantees that the iterates y_k of the method stay near enough the dual averaging iterates, which allows us to circumvent the difficulties of global convergence for Riemannian methods. As we demonstrate in the proof, this asymptotically iterates a stochastic gradient method in the tangent space $\mathcal{T} = \{v : \langle \nabla f_i(x^\star), v\rangle = 0, i \le m_0\}$, and only updates line 4 occur, as $\mathcal{M}_k = \mathcal{M}_{k-1}$ and $y_k^{\mathrm{man}} \in \mathcal{X}$.

We prove the following theorem in the Supplementary Material, Section 15, using the notation of Proposition 1, where $H^\star = \nabla^2 f(x^\star) + \sum_{i=1}^{m_0} \lambda_i^\star \nabla^2 f_i(x^\star)$ and $\mathsf{P}_{\mathcal{T}}$ is the projection onto the tangent space $\mathcal{T} = \{v \in \mathbb{R}^n : v^T \nabla f_i(x^\star) = 0, i \in [m_0]\}$.

THEOREM 5. *Let Assumptions A, B, C and D' hold. Then the iterates y_k of Algorithm 1 satisfy*

$$\frac{1}{\sqrt{k}} \sum_{i=1}^{k} (y_i - x^\star) \overset{d}{\rightsquigarrow} \mathsf{N}(0, \mathsf{P}_{\mathcal{T}} H^{\star\dagger} \mathsf{P}_{\mathcal{T}} \Sigma \mathsf{P}_{\mathcal{T}} H^{\star\dagger} \mathsf{P}_{\mathcal{T}}).$$

The extended Riemannian stochastic gradient method, coupled with identification results that dual averaging supplies, is asymptotically optimal.

6. Numerical experiments.

In this section, we perform a small simulation study to compare dual averaging (17) with stochastic (and Riemannian) gradient methods on nonnegative least squares and ridge regression. We take our observations $(a_i, b_i) \in \mathbb{R}^n \times \mathbb{R}$ and use the squared loss $f(x; (a, b)) = \frac{1}{2}(\langle a, x\rangle - b)^2$. Both problems are of the form (1), where for nonnegative least squares, we use the constraints $\mathcal{X} = \mathbb{R}_+^n$, and for the ridge regression problem, we set $\mathcal{X} = \{x \in \mathbb{R}^n : \|x\|^2 \le \lambda\}$, where $\lambda > 0$.

Now we describe our experimental setting. To allow easier visualization, we use dimension $n = 2$ and generate $b_i = \langle a_i, x^{\mathrm{true}}\rangle + \xi_i$ for $a_i \overset{\mathrm{iid}}{\sim} \mathsf{N}(0, I_2)$ and $\xi_i \overset{\mathrm{iid}}{\sim} \mathsf{N}(0, 1)$. For the nonnegative least squares problem, we set $x^{\mathrm{true}} = (1, -1)$, while for the ridge regression problem, we set $x^{\mathrm{true}} = (1, 1)$ and $\lambda = 1$, giving solutions $x^\star = (1, 0)$ and $x^\star = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$, respectively. For both problems, the unique solution $x^\star$ lies on the boundary of the feasible set $\mathcal{X}$. To fairly compare the performance of the algorithms, we use the same parameters, initializing at $x = 0$ and using stepsizes $\alpha_k = k^{-\beta}$ for $\beta = 3/4$. In each experiment, we run each method for K iterations, and we perform T independent replications.

6.1. *Constraint identification.*

Our first set of numerical results shows that the stochastic gradient method fails to identify active constraints, while dual averaging identifies them. We present the results graphically in Figure 2. For each of the two plots, the horizontal axis indexes the iteration k (over $K = 100$ iterations) and the vertical axis represents the proportion of the $T = 1000$ tests in which the iterate x_k lies on the active constraints. Both plots show that the dual averaging iterates (the solid red curve) identify the constraints (with 100% accuracy by iteration 40), while the stochastic gradient method (the dotted blue curve) does not.

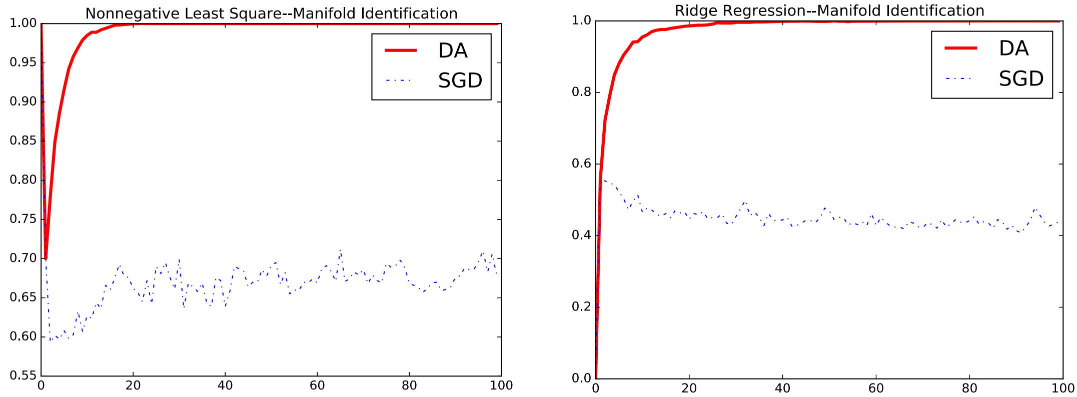

FIG. 2. *Success rate of manifold identification.*

6.2. *Accuracy.* Our second set of numerical results shows the improved performance of dual averaging relative to projected stochastic gradient descent, and that the manifold aware algorithm 1 exhibits better behavior on nonlinear constraints, which we illustrate in Figure 3. Each of the red triangles (resp., blue circles or green diamonds) represents an averaged dual averaging (resp., stochastic gradient or Riemannian method 1) iterate $\bar{x}_K = \frac{1}{K}\sum_{i=1}^{K} x_i$ (we set $K = 100$) out of $T = 20$ experiments. The dual averaging results are typically closer to $x^\star$ (the black cross) and to the active constraints (the grey dotted curve) than the stochastic gradient averages, while the right plot in Figure 3 shows the improved performance of the Riemannian method we outline in Algorithm 1. The distance of the dual averaging iterates to $x^\star$ is typically shorter along the normals to the active constraints, leading to better accuracy estimating $x^\star$.

6.3. *Linear versus nonlinear constraints.* For our third set of numerical results, we investigate the asymptotic variance of the variant dual averaging method (17) versus that of the Riemannian method (Algorithm 1) and the optimal asymptotic variance that Theorem 1 provides. For the nonnegative least squares problem, the linear constraints have tangent set $\mathcal{T} = \{tv^\star\}_{t\in\mathbb{R}}$, where $v^\star = (1,0)$, while the ridge problem has $\mathcal{T} = \{tv^\star\}_{t\in\mathbb{R}}$ where $v^\star = (1,-1)$. In each case, we compute the variance of $\sqrt{k}\langle v^\star, \bar{x}_k - x^\star\rangle$ for $\bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$ for $k \le K = 10^4$ over $T = 1000$ independent trials. We present the results in Figure 4. In each of the two plots, the red dashed curve shows the variance the dual averaging iterates and the gray dotted line shows the optimal asymptotic variance (Theorem 1). In the left plot,

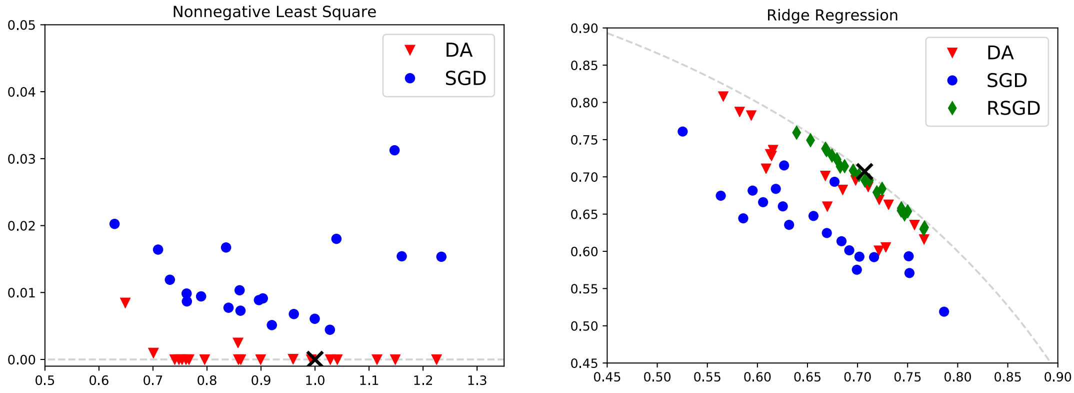

FIG. 3. *The averaged iterates* $\bar{x}_K = \frac{1}{K}\sum_{i=1}^{K} x_i$.

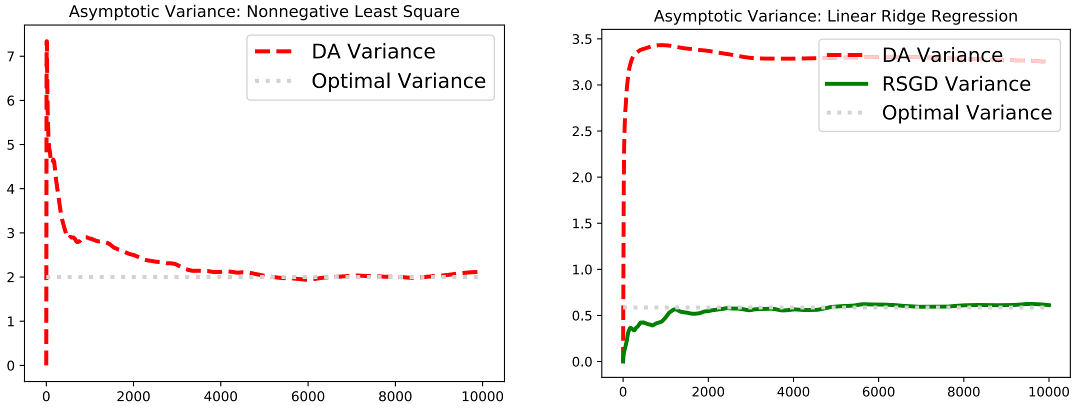

FIG. 4. *Asymptotic variance under linear and nonlinear constraints.*

we see that dual averaging converges with the asymptotically optimal rate. In the right, the Riemannian method (the solid green line) has asymptotically optimal variance, while the dual averaging procedure has variance between 3 and 3.5, which is suboptimal; these results suggest the accuracy of our theoretical predictions.

7. Proof of Proposition 1.

This result is a consequence of Shapiro [45], Theorem 5.1, or Dontchev and Rockafellar [18], Theorem 2G.8. First, consider the Lagrangian for the tilted problem (8),

$$\mathcal{L}_v(x, \lambda) = f_v(x) + \sum_{i=1}^{m} \lambda_i f_i(x).$$

We perform a second-order Taylor approximation to $\mathcal{L}_v(x, \lambda)$ around x_0, linearizing the active constraints $f_i(x)$ for $i \in [m_0]$, and minimizers of this quadratic over linear constraints are $o(\|v\|)$-close to x_v. We make this precise.

Let $\Lambda_0 \subset \mathbb{R}_+^m$ denote the set of optimal Lagrange multipliers for problem (1) (the tilted problem (8) at $v = 0$), recalling that by Assumption B, this set is a compact polyhedron (and is a singleton under B.i). In either case of Assumption B, the set $\{\nabla_x^2 \mathcal{L}_0(x_0, \lambda) : \lambda \in \Lambda_0\}$ is a singleton, so $H^\star = \nabla_x^2 \mathcal{L}_0(x^\star, \lambda^\star) = \nabla^2 \mathcal{L}_0(x_0, \lambda)$ for any $\lambda \in \Lambda_0$. At $v = 0$, our assumptions imply $\nabla_v^2 \mathcal{L}_v(x_0, \lambda) = 0$ and $\nabla_{vx}^2 \mathcal{L}_v(x_0, \lambda) = 0$. Define the quadratic

$$\zeta_v(w) := \frac{1}{2} w^T H^\star w - v^T w = \frac{1}{2} w^T \nabla_x^2 \mathcal{L}_0(x_0, \lambda) w - v^T w, \tag{23}$$

which approximates $\mathcal{L}_v(x_0 + w, \lambda) \approx f_0(x_0) + \zeta_v(w)$ for w, v small, because $\nabla_x \mathcal{L}_0(x_0, \lambda) = 0$ for $\lambda \in \Lambda_0$. For $\lambda \in \Lambda_0$, define the sets $I_0(\lambda) := \{i \in [m_0] \mid \lambda_i = 0\}$ and $I_+(\lambda) := \{i \in [m_0] \mid \lambda_i > 0\}$, and consider the tangent cone

$$\overline{\mathcal{T}} := \bigcup_{\lambda \in \Lambda_0} \{w : w^T \nabla f_i(x_0) = 0 \text{ for } i \in I_+(\lambda), w^T \nabla f_i(x_0) \le 0 \text{ for } i \in I_0(\lambda)\}.$$

The minimizers of the quadratic function (23) over $\overline{\mathcal{T}}$ approximate those of the tilted problem (8) as follows [18], Theorem 2G.8: if for v near 0 the function $\zeta_v(w)$ has a unique minimizer w_v over $\overline{\mathcal{T}}$, then

$$\lim_{t \downarrow 0} \frac{x_{tv} - x_0}{t} = w_v. \tag{24}$$

Moreover [18], Theorem 2G.8 and Definition 2.4 (semiderivative), if w_v is linear in v, then $v \mapsto x_v$ is differentiable at $v = 0$ with $x_v = x_0 + w_v + o(\|v\|)$. We consider the two cases of Assumption B to give the result.

Case I: *Linearly independent constraints.* As noted following Assumption B, the set $\Lambda_0 = \{\lambda^\star\}$, a singleton. Thus $\overline{\mathcal{T}} = \mathcal{T}$ and following the quadratic expansion (23), we solve:

$$\operatorname*{minimize}_{w} \frac{1}{2} w^T H^\star w - v^T w \quad \text{subject to } w^T \nabla f_i(x_0) = 0, \quad i \in [m_0].$$

This quadratic problem has solution $\mathsf{P}_{\mathcal{T}} H^{\star\dagger} \mathsf{P}_{\mathcal{T}} v$, which is unique by Assumption C. Expression (24) gives the proposition in this case.

Case II: *Affine constraints.* In Assumption B.ii, the active f_i are affine. We claim that, though Λ_0 may not be a singleton,

$$\overline{\mathcal{T}} = \{w \mid Aw = 0\}.$$

To see this, let $u = -\nabla f_0(x_0)$, whence we know that $u = A^T \lambda^\star = 0$ for some $\lambda^\star > 0$ by Assumption B.ii. Writing $A = [a_1 \cdots a_{m_0}]^T$, we see that for any $\lambda \in \Lambda_0$, if we have

$$w \in \{w \mid a_i^T w = 0 \text{ for } i \text{ s.t. } \lambda_i > 0, a_i^T w \leq 0 \text{ otherwise}\},$$

then $u = A^T \lambda$ and $u^T w = \lambda^T A w = \sum_{i=1}^{m_0} \lambda_i a_i^T w = 0$, because $a_i^T w = 0$ whenever $\lambda_i \neq 0$. But of course, we know that $A^T \lambda^\star = u$, so that

$$0 = w^T u = w^T A^T \lambda^\star = \sum_{i=1}^{m_0} \lambda_i^\star a_i^T w$$

so each index i satisfies $a_i^T w = 0$ as $a_i^T w \leq 0$ and $\lambda_i^\star > 0$. The simplification $\overline{\mathcal{T}} = \mathcal{T}$ as in Case I applies; the remainder of the proof is identical.

8. Proof of Theorem 1: Local minimax lower bounds.

We briefly outline the approach. We divide the proof into two parts: an analytic part studying properties of the perturbed solutions x_u (Section 8.1), and a stochastic part applying Le Cam's local asymptotic normality theory (Section 8.2). In the first part, we investigate the perturbation properties of the solutions x_u as $u \to 0$ via the implicit function result of Proposition 1. We show that our choice (12) of P_u gives $f_u(x) \approx f_0(x) + u^T \Sigma_g (x - x_0)$ for an appropriate Σ_g, so that $x_u = x_0 + Du + o(\|u\|)$ as $u \to 0$ for a matrix D by Proposition 1. This allows application of Le Cam's local asymptotic normality theory [31, 51, 52]; heuristically, we may place a Gaussian prior on u concentrated at rate $1/k$, so that minimization in the problem (13) indexed by u is asymptotically equivalent to estimating the Gaussian shift Du. By our construction of the tilting (12), the vector u is asymptotically normally distributed (we make this precise in Section 8.2), which allows us to apply standard normality optimality guarantees. We unify our arguments in Section 8.3.

8.1. *Perturbation of optimal solutions.*

We first consider optimal solutions to the problem $\mathcal{P}_u$ defined in equation (13). We begin with a lemma that describes the perturbation of f_u from f_0.

LEMMA 8.1. *Let the conditions of Theorem 1 hold. Then $(x, u) \mapsto f_u(x)$ is $\mathcal{C}^2$ near $u = 0$ and $x = x_0$, and*

$$f_u(x) = f_0(x) + u^T \Sigma_{g,f}(x - x_0) + c_u + o(\|x - x_0\|^2 + \|u\|^2),$$

where $\Sigma_{g,f} := \mathbb{E}[g(S)(\nabla f(x_0; S) - \nabla f(x_0))^T]$ and c_u depends only on u.

The lemma consists of a number of applications of Lebesgue's dominated convergence theorem; we defer proof to the Supplementary Material, Section 10.1.

Evidently, Proposition 1 applies to the minimizers x_u, as the problem $\mathcal{P}_u$ is asymptotically equivalent to a linear tilt, exactly as in equation (8). Thus, it is immediate that the minimizers x_u of $f_u(x) = \int f(x; s)\, dP_u(s)$ over $\mathcal{X}$ satisfy

$$\sqrt{k}(x_{u/\sqrt{k}} - x_0) \underset{k\uparrow\infty}{\to} -\mathsf{P}_{\mathcal{T}}\left(\nabla^2 f_0(x_0) + \sum_{i=1}^{m_0} \lambda_i^\star \nabla^2 f_i(x_0)\right)^\dagger \mathsf{P}_{\mathcal{T}} \Sigma_{g,f}^T u, \tag{25}$$

where we recall that $\mathsf{P}_{\mathcal{T}}$ denotes projection onto the tangent set (5) and $\lambda^\star$ are optimal Lagrange multipliers for problem (1).

8.2. *Local asymptotic normality.* The tilts P_u are a locally asymptotically normal [31, 52] family of distributions indexed $u \in \mathbb{R}^d$, which, when coupled with the differentiability result (25), allows us to apply the Hájek–Le Cam local minimax theory. We first recall definitions due to Le Cam [31] that we use to develop our problems with asymptotically Gaussian structure.

DEFINITION 8.1. Let $U \subset \mathbb{R}^d$ be an open set containing 0. For each $k \in \mathbb{N}$ and $u \in U$, let $P_{k,u}$ be a probability measure on a measurable space $(\mathcal{S}_k, \mathcal{F}_k)$, and let S^k be a sample from $P_{k,u}$. The sequence $\{\mathcal{S}_k, \mathcal{F}_k, P_{k,u}\}_{u\in U}$ is *locally asymptotically normal with precision* $K \succeq 0$ *(LAN)* if

$$\log \frac{dP_{k,u}(S^k)}{dP_{k,0}(S^k)} = \langle u, Z_k\rangle - \frac{1}{2} u^T K u + o_{P_0}(1),$$

where $Z_k \overset{d}{\rightsquigarrow} \mathsf{N}(0, K)$ under the distribution P_0.

A second important definition is the regular estimand [51, 52].

DEFINITION 8.2. Let $U \subset \mathbb{R}^d$ be a neighborhood of 0 and $\kappa_k : U \to \mathbb{R}^n$. The sequence $\{\kappa_k\}_{k\in\mathbb{N}}$ is *regular with derivative* $D \in \mathbb{R}^{n\times d}$ if

$$\sqrt{k}(\kappa_k(u) - \kappa_k(0)) \to Du \quad \text{for all } u \in U.$$

With these definitions, the following local asymptotic minimax result, a variant of the Hájek–Le Cam minimax theorem, holds.

LEMMA 8.2 (Local minimax theorem, Theorem 3.11.5 [52] or Lemma 6.6.1 and Theorem 6.6.2 [31]). *Let the sequence* $\{\mathcal{S}_k, \mathcal{F}_k, P_{k,u}\}_{u\in U}$ *be locally asymptotically normal with precision* K *(Definition 8.1) and let* $\kappa_k : U \to \mathbb{R}^{n'}$ *be regular with derivative* D *(Definition 8.2). Let* $L : \mathbb{R}^n \to \mathbb{R}_+$ *be symmetric and quasi-convex. Then for any sequence* $T_k : \mathcal{S}_k \to \mathbb{R}^n$ *of estimators,*

$$\sup_{U_0 \subset U, |U_0|<\infty} \liminf_{k\to\infty} \max_{u\in U_0} \mathbb{E}_{P_{u,k}}[L(\sqrt{k}(T_k(S^k) - \kappa_k(u)))] \geq \mathbb{E}[L(Z)],$$

where $Z \sim \mathsf{N}(0, DK^{-1}D^T)$ *when* $K \succ 0$. *If* K *is singular and* $\mathrm{range}(D^T) \cap \mathrm{null}(K) \neq \varnothing$, *the result holds for* $Z \sim \mathsf{N}(0, D(K + \lambda I)^{-1} D^T)$ *for any* $\lambda > 0$.

Eq. (25) shows that $\kappa_k(u) := \mathrm{argmin}_{x\in\mathcal{X}} f_{u/\sqrt{k}}(x)$ is regular (Def. 8.2): recalling the definition of the Hessian $H^\star = \nabla_x^2 \mathcal{L}(x^\star, \lambda^\star)$ in the statement of the theorem, the sequence is regular with derivative $\mathsf{P}_{\mathcal{T}} H^{\star\dagger} \mathsf{P}_{\mathcal{T}} \Sigma_{g,f}^T$. It remains to establish the local asymptotic normality properties of P_u.

LEMMA 8.3. *Let P_u be as in expression* (12). *Let $u \in \mathbb{R}^d$ and define $P_k = P^k_{u/\sqrt{k}}$, the k-fold product of $P_{u/\sqrt{k}}$. Let $\Sigma_g = \mathbb{E}_{P_0}[g(S)g(S)^T]$. Then*

$$\log \frac{dP_k(S_1, \ldots, S_k)}{dP_0(S_1, \ldots, S_k)} = -\frac{1}{\sqrt{k}} u^T \sum_{i=1}^k g(S_i) - \frac{1}{2} u^T \Sigma_g u + o_{P_0}(1).$$

See the Supplementary Material, Section 10.2 for a proof. In particular, we see that if $\mathcal{F}_k$ denotes the σ-algebra on the product $\mathcal{S}^k$, then the sequence

$$\{\mathcal{S}^k, \mathcal{F}_k, P^k_{u/\sqrt{k}}\}_{u \in \mathbb{R}^n}$$

is LAN with precision Σ_g for g with $\mathbb{E}_{P_0}[g] = 0$ and $\mathbb{E}_{P_0}[\|g\|^2] < \infty$.

8.3. *Finalizing the argument.* Now that we have the regularity of the sequence $x_{u/\sqrt{k}}$ as $k \to \infty$ (the convergence guarantee (25)) and the asymptotic normality of Lemma 8.3, we may apply Lemma 8.2. Indeed, let $P_{u,k} = P^k_{u/\sqrt{k}}$ be the distribution of an i.i.d. sample $S_i \overset{\text{iid}}{\sim} P_{u/\sqrt{k}}$ for $i = 1, \ldots, k$, and let $\widehat{x}_k$ be an arbitrary estimator based on $S_{1:k}$. Lemma 8.2 implies

$$\sup_{U_0 \subset \mathbb{R}^d, |U_0| < \infty} \liminf_{k \to \infty} \max_{u \in U_0} \mathbb{E}_{P^k_{u/\sqrt{k}}} \left[L\left(\sqrt{k}(\widehat{x}_k - x_{u/\sqrt{k}})\right)\right] \geq \mathbb{E}\left[L(Z_\lambda)\right]$$

for any $\lambda > 0$, where

$$Z_\lambda \sim \mathsf{N}\left(0, \mathsf{P}_\mathcal{T} H^{\star\dagger} \mathsf{P}_\mathcal{T} \Sigma^T_{g,f} (\Sigma_g + \lambda I)^{-1} \Sigma_{g,f} \mathsf{P}_\mathcal{T} H^{\star\dagger} \mathsf{P}_\mathcal{T}\right).$$

The theorem follows by taking $\lambda \downarrow 0$, noting that for any two mean-zero random vectors Z and Y, we have

$$\mathbb{E}[YZ^T]\mathbb{E}[ZZ^T]^\dagger \mathbb{E}[ZY^T] \preceq \mathbb{E}[YY^T], \tag{26}$$

and that (by Anderson's lemma [51], Lemma 8.5) if $\Sigma_1 \preceq \Sigma_2$ and $Z_i \sim \mathsf{N}(0, \Sigma_i)$, then $\mathbb{E}[L(Z_1)] \leq \mathbb{E}[L(Z_2)]$. To see inequality (26), we may without loss of generality assume that $\mathbb{E}[ZZ^T] \preceq I$, as by letting $\Sigma = \mathbb{E}[ZZ^T]^\dagger$, we have $\mathbb{E}[\Sigma^{1/2} ZZ^T \Sigma^{1/2}] = \Sigma^{1/2} \Sigma^\dagger \Sigma^{1/2} \preceq I$; to show inequality (26), it is thus equivalent to show that $\mathbb{E}[YZ^T]\mathbb{E}[ZY^T] \preceq I$ for all Z such that $\mathbb{E}[ZZ^T] \preceq I$. To see this, let v be arbitrary, and note that by Cauchy-Schwarz we have

$$\|\mathbb{E}[v^T YZ]\|_2^2 = \sup_{\|u\| \leq 1} \mathbb{E}[v^T YZ^T u]^2 \leq \mathbb{E}[(v^T Y)^2] \sup_{\|u\| \leq 1} \mathbb{E}[(u^T Z)^2] \leq v^T \mathbb{E}[YY^T] v.$$

9. Proofs of convergence for dual averaging.

Here we collect the major arguments for our proofs of the almost sure convergence and finite time constraint identification for our variant (17) of dual averaging. We highlight new results and techniques, deferring technical details.

9.1. *Proof of Theorem 2: Almost sure convergence.* First, we establish a few technical properties of the stepsize sequence. We begin with the following lemma, whose proof is immediate when $\alpha_k \propto k^{-\beta}$ for $\beta \in (\frac{1}{2}, 1)$.

LEMMA 9.1. *For α_k satisfying condition* (18), $\sum_{k=1}^\infty \frac{\alpha_k}{\sum_{i=1}^k \alpha_i} = \infty$.

Now we state a classical result that is useful for showing the almost convergence of stochastic approximation algorithms.

LEMMA 9.2 (Robbins and Siegmund [42]). *Let V_k, A_k, B_k, C_k be nonnegative random variables adapted to a filtration $\mathcal{F}_k$. Assume that*

$$\mathbb{E}[V_{k+1} \mid \mathcal{F}_k] \le (1 + A_k)V_k + B_k - C_k.$$

Then on the event $\{\sum_k A_k < \infty, \sum_k B_k < \infty\}$, there is a random variable $V_\infty < \infty$ such that $V_k \overset{a.s.}{\to} V_\infty$ and $\sum_k C_k < \infty$ a.s.

We use Lemma 9.2 to show that the quantity

$$R_k := \langle z_k + x_{k+1}, x^\star - x_{k+1}\rangle + \frac{1}{2}\|x_{k+1} - x^\star\|_2^2 \tag{27}$$

converges a.s. to some random variable $R_\infty < \infty$, where $z_k := \sum_{i=1}^k \alpha_i g_i$. We can decompose R_k as the sum of two nonnegative random variables,

$$R_k = G_k + V_k, \qquad G_k = \langle z_k + x_{k+1}, x^\star - x_{k+1}\rangle \ge 0 \quad \text{and} \quad V_k = \frac{1}{2}\|x_{k+1} - x^\star\|_2^2.$$

Here we have $G_k \ge 0$ because x_{k+1} minimizes $\langle z_k, x\rangle + \frac{1}{2}\|x\|_2^2$ over $x \in \mathcal{X}$, so that $\langle z_k + x_{k+1}, y - x_{k+1}\rangle \ge 0$ for all $y \in \mathcal{X}$ (and $x^\star \in \mathcal{X}$ by definition), while $V_k \ge 0$ clearly. Recall the definition (Assumption D') of the filtration

$$\mathcal{F}_k := \sigma(\xi_1, \dots, \xi_k)$$

as the σ-field generated by the noise sequence through time k. Then we have the measurability $R_k, G_k, V_k \in \mathcal{F}_k$ and the following convergence.

LEMMA 9.3. *Let R_k be as in (27) and assume that $\sum_k \alpha_k^2 < \infty$. Then for some finite random variable R_∞, we have $R_k \overset{a.s.}{\to} R_\infty$. Moreover,*

$$\sum_{i=1}^{\infty} \alpha_i [f(x_i) - f(x^*)] < \infty \quad \textit{with probability } 1.$$

PROOF. Let $h(x) = \frac{1}{2}\|x\|_2^2 + \mathbf{I}_{\mathcal{X}}(x)$ and define its conjugate $h^*(z) = \sup_{x\in\mathcal{X}}\{\langle z, x\rangle - \frac{1}{2}\|x\|_2^2\}$. Then h^* has 1-Lipschitz continuous gradient with $\nabla h^*(z) = \operatorname{argmax}_{x\in\mathcal{X}}\{\langle z, x\rangle - \frac{1}{2}\|x\|_2^2\}$ [27], Chapter X, and

$$R_k = \langle z_k, x^\star - x_{k+1}\rangle + \frac{1}{2}\|x^\star\|_2^2 - \frac{1}{2}\|x_{k+1}\|_2^2 = \langle z_k, x^\star\rangle + \frac{1}{2}\|x^\star\|_2^2 + h^*(-z_k).$$

Using $\nabla h^*(-z_{k-1}) = x_k$ and the Lipschitz continuity of ∇h^*, we have

$$\begin{aligned} h^*(-z_k) &\le h^*(-z_{k-1}) + \langle \nabla h^*(-z_{k-1}), z_{k-1} - z_k\rangle + \frac{1}{2}\|z_k - z_{k-1}\|_2^2 \\ &= h^*(-z_{k-1}) - \alpha_k\langle g_k, x_k\rangle + \frac{\alpha_k^2}{2}\|g_k\|_2^2. \end{aligned}$$

That is, we have for any k that

$$\begin{aligned} R_k &\le \langle z_k, x^\star\rangle + \frac{1}{2}\|x^\star\|_2^2 + h^*(-z_{k-1}) - \alpha_k\langle g_k, x_k\rangle + \frac{\alpha_k^2}{2}\|g_k\|_2^2 \\ &= \underbrace{\langle z_{k-1} + x_k, x^\star - x_k\rangle + \frac{1}{2}\|x_k - x^\star\|_2^2}_{=R_{k-1}} - \alpha_k\langle g_k, x_k - x^\star\rangle + \frac{\alpha_k^2}{2}\|g_k\|_2^2. \end{aligned}$$

Taking conditional expectations and using that $\mathbb{E}[g_k \mid \mathcal{F}_{k-1}] = \nabla f(x_k)$ yields

$$\begin{aligned}\mathbb{E}[R_k \mid \mathcal{F}_{k-1}] &\le R_{k-1} - \alpha_k \langle \nabla f(x_k), x_k - x^\star \rangle + \frac{\alpha_k^2}{2}\mathbb{E}[\|g_k\|_2^2 \mid \mathcal{F}_{k-1}] \\ &\overset{(i)}{\le} G_{k-1} + V_{k-1} - \alpha_k \langle \nabla f(x_k), x_k - x^\star \rangle + \frac{\alpha_k^2}{2}(C\|x_k - x^\star\|_2^2 + C) \\ &\overset{(ii)}{\le} (1 + C\alpha_k^2)[G_{k-1} + V_{k-1}] - \alpha_k \langle \nabla f(x_k), x_k - x^\star \rangle + C\alpha_k^2,\end{aligned}$$

where inequality (i) follows by Assumption D' and the discussion (Eq. (6)) immediately following Assumption D, and inequality (ii) because $G_{k-1} \ge 0$ and $V_{k-1} = \frac{1}{2}\|x_k - x^\star\|_2^2$. In particular, we have

$$\mathbb{E}[R_k \mid \mathcal{F}_{k-1}] \le (1 + C\alpha_k^2)R_{k-1} - \alpha_k \langle \nabla f(x_k), x_k - x^\star \rangle + C\alpha_k^2.$$

Because $f(x^\star) \ge f(x_k) + \langle \nabla f(x_k), x^\star - x_k \rangle$, or $\langle \nabla f(x_k), x_k - x^\star \rangle \ge f(x_k) - f(x^\star) \ge 0$, Lemma 9.2 applies. Thus we must have $R_k \overset{a.s.}{\to} R_\infty$ for some finite random variable R_∞, and moreover

$$\sum_{i=1}^\infty \alpha_i [f(x_i) - f(x^*)] \le \sum_{i=1}^\infty \alpha_i \langle \nabla f(x_i), x_i - x^* \rangle < \infty,$$

where we have used the standard first-order convexity inequality. □

With these lemmas as background, we finally provide the proof of Theorem 2, by showing that with R_k defined as in expression (27),

$$R_k \overset{a.s.}{\to} 0 \quad \text{so that } x_k \overset{a.s.}{\to} x^\star. \tag{28}$$

We introduce a bit of notation. Let $A_k = \sum_{i=1}^k \alpha_i$, and recall that $z_k = \sum_{i=1}^k \alpha_i g_i$. Define $\bar{z}_k = \sum_{i=1}^k \alpha_i \nabla f(x_i)$ to be the weighted partial sum of the (nonnoisy) gradients $\nabla f(x_i)$, and we let $z_k^\star = A_k \nabla f(x^\star)$.

We first claim that the error sequence is asymptotically negligible:

$$\frac{1}{\sqrt{A_k}} \sum_{i=1}^k \alpha_i \xi_i \overset{a.s.}{\to} 0. \tag{29}$$

To see the claim (29), we use the following lemma.

LEMMA 9.4 (Dembo [15], Exercise 5.3.35). *Let $Z_k \in \mathbb{R}^n$ be a martingale adapted to $\mathcal{F}_k$ and let $b_k > 0$ be a nonrandom sequence increasing to ∞. If $\sum_{k=1}^\infty b_k^{-2}\mathbb{E}[\|Z_k - Z_{k-1}\|^2 \mid \mathcal{F}_{k-1}] < \infty$, we have $b_k^{-1} Z_k \overset{a.s.}{\to} 0$.*

Since $\{\sum_{i=1}^k \alpha_i \xi_i\}_{k=1}^\infty$ is a martingale difference sequence, Lemma 9.4 shows that to obtain the claim (29) it is sufficient to show that

$$\sum_{k=1}^\infty \frac{1}{A_k}\mathbb{E}[\|\alpha_k \xi_k\|^2 \mid \mathcal{F}_{k-1}] < \infty.$$

By Assumption D', the left side of the preceding display has upper bound $\frac{C}{A_1}\sum_{i=1}^\infty \alpha_i^2(1 + \|x_i - x^\star\|^2)$, so that showing $\sum_{i=1}^\infty \alpha_i^2(1 + \|x_i - x^\star\|^2) < \infty$ proves the claim (29). With that in mind, recall Lemma 2.1, which guarantees an $\epsilon > 0$ such that $f(x) - f(x^\star) \ge$

$\epsilon(\|x - x^\star\|^2 \wedge \|x - x^\star\|)$. Using Lemma 9.3, we know that $M := \sup_i \|x_i - x^\star\| \vee 1 < \infty$. Thus we have

$$f(x_i) - f(x^\star) \ge \epsilon \min\{\|x_i - x^\star\| M/M, \|x_i - x^\star\|^2\} \ge c\|x_i - x^\star\|^2$$

where $c > 0$ is a random positive constant that depends on the bound M. Combining this result with Lemma 9.3, we have

LEMMA 9.5. *Under the conditions of Theorem 2, we have*

$$\sum_{i=1}^{\infty} \alpha_i \|x_i - x^\star\|^2 \le \frac{1}{c}\sum_{i=1}^{\infty} c\alpha_i \|x_i - x^\star\|^2 \le \frac{1}{c}\sum_{i=1}^{\infty} \alpha_i [f(x_i) - f(x^\star)] < \infty.$$

Here the final inequality follows from Lemma 9.3. Noting that $\sum_{i=1}^{\infty} \alpha_i^2 < \infty$ by assumption (18), we obtain the claim (29). Moreover, this implies that

$$\frac{z_k - \bar{z}_k}{\sqrt{A_k}} = \frac{1}{\sqrt{A_k}} \sum_{i=1}^{k} \alpha_i \xi_i \overset{a.s.}{\to} 0. \tag{30}$$

Now that we have the convergence guarantee (30), that $R_k \overset{a.s.}{\to} R_\infty < \infty$ (Lemma 9.3), and that $\sum_{i=1}^{\infty} \alpha_i \|x_i - x^\star\|^2 < \infty$ with probability 1, we define

$$\Omega_0 := \left\{\sum_{i=1}^{\infty} \alpha_i \|x_i - x^\star\|^2 < \infty, R_k \to R_\infty < \infty, \frac{z_k - \bar{z}_k}{\sqrt{A_k}} \to 0\right\}, \quad \mathbb{P}(\Omega_0) = 1.$$

On the set Ω_0, using the Lipschitz continuity Assumption A, we may define

$$\sigma_\infty^2 := \sum_{i=1}^{\infty} \alpha_i \|\nabla f(x_i) - \nabla f(x^*)\|^2 \le L^2 \sum_{i=1}^{\infty} \alpha_i \|x_i - x^*\|^2 < \infty.$$

Then using Jenson's inequality and recalling the definition $z_k^\star = A_k \nabla f(x^\star)$,

$$\|\bar{z}_k - z_k^\star\|^2 = \left\|\sum_{i=1}^{k} \alpha_i (\nabla f(x_i) - \nabla f(x^\star))\right\|^2 \le A_k \sigma_\infty^2.$$

Hence, we have $\|\bar{z}_k - z_k^\star\| \le \sqrt{A_k}\sigma_\infty$. Now, we see that on Ω_0,

$$\infty > \sum_{i=1}^{\infty} \alpha_i \|x_i - x^\star\|^2 = \sum_{i=1}^{\infty} \frac{\alpha_i}{A_i} A_i \|x_i - x^\star\|^2.$$

Lemma 9.1 (that $\sum_i \frac{\alpha_i}{A_i} = \infty$) implies there exists a subsequence $\{k_i\}$ with

$$\lim_{i \to \infty} A_{k_i} \|x_{k_i} - x^\star\|^2 = 0,$$

and moreover,

$$\|\bar{z}_{k_i} - z_{k_i}^\star\| \|x_{k_i} - x^\star\| \le \sigma_\infty \sqrt{A_{k_i}} \|x_{k_i} - x^\star\| \to 0.$$

Keep the subsequence $\{k_i\}$ fixed, and note that on Ω_0, we have that $R_{k_i - 1} \to R_\infty$. Let us expand the terms in the definition of R_k to see that we must have $R_\infty = 0$. Indeed, we have

$$\begin{aligned}
R_{k-1} &= \langle z_{k-1} + x^\star, x^\star - x_k\rangle - \frac{1}{2}\|x_k - x^\star\|^2 \\
&\le \langle z_{k-1} + x^\star, x^\star - x_k\rangle \\
&= \langle z_{k-1} - z_{k-1}^\star, x^\star - x_k\rangle + \langle z_{k-1}^\star, x^\star - x_k\rangle + \langle x^\star, x^\star - x_k\rangle \\
&\le \|z_{k-1} - z_{k-1}^\star\| \|x^\star - x_k\| + A_k\langle \nabla f(x^\star), x^\star - x_k\rangle + \|x^\star\| \|x^\star - x_k\|.
\end{aligned}$$

The optimality conditions for $x^\star$ imply $\langle \nabla f(x^\star), x^\star - x_k\rangle \le 0$. On the subequence k_i, we have

$$\limsup_{i\to\infty} \|z_{k_i-1} - z^\star_{k_i-1}\| \|x^\star - x_{k_i}\| \le \limsup_{i\to\infty} \sigma_\infty \sqrt{A_{k_i-1}} \|x^\star - x_{k_i}\| = 0$$

and $\limsup_{i\to\infty} \|x^\star\| \|x^\star - x_{k_i}\| = 0$. In particular, that $R_k \ge 0$ implies

$$0 \le \liminf_{i\to\infty} R_{k_i-1} \le \limsup_{i\to\infty} R_{k_i-1} = 0.$$

Because $R_k \to R_\infty$ on Ω_0, it must thus be the case that $R_\infty = 0$.

9.2. *Manifold identification*: *Theorem* 3. Recall that $z_k = \sum_{i=1}^k \alpha_i g_i$ is the weighted partial sum of the noisy gradients, and let $A_k = \sum_{i=1}^k \alpha_i$. The following lemma is a nearly immediate consequence of our previous results and, given the perturbation results in Lemmas 4.1 and 4.2, is the key to our finite identification result.

LEMMA 9.6. *Under the conditions of the Theorem* 3, $\frac{1}{A_k} z_k \overset{a.s.}{\to} \nabla f(x^\star)$.

PROOF. We first remove the randomness of ξ_i. By Jenson's inequality,

$$\left\| \frac{z_k}{A_k} - \nabla f(x^\star) \right\|^2 \le 2 \left\| A_k^{-1} \sum_{i=1}^k \alpha_i (\nabla f(x_i) - \nabla f(x^\star)) \right\|^2 + 2 \left\| A_k^{-1} \sum_{i=1}^k \alpha_i \xi_i \right\|^2.$$

The second term converges almost surely to zero by the almost sure convergence (29) in the proof of Theorem 2. We thus focus on the first term.

By Lemma 9.5 in the proof of Theorem 2 and the Lipschitz Assumption A, we know that $\sum_{i=1}^\infty \alpha_i \|\nabla f(x_i) - \nabla f(x^\star)\|^2 \le C \sum_{i=1}^\infty \alpha_i \|x_i - x^\star\|^2 < \infty$ with probability 1. Thus, by Jenson's inequality,

$$\frac{1}{A_k^2} \left\| \sum_{i=1}^k \alpha_i (\nabla f(x_i) - \nabla f(x^\star)) \right\|^2 \le \frac{1}{A_k} \sum_{i=1}^\infty \alpha_i \|\nabla f(x_i) - \nabla f(x^\star)\|^2.$$

Taking $A_k \to \infty$ gives the result. □

Applying Assumption B, there exist $\lambda_i > 0$ and $\nu_i = 0$ such that $\nabla f(x^\star) + \sum_{i=1}^{m_0} \lambda_i \times \nabla f_i(x^\star) + \sum_{i=m_0+1}^m \nu_i \nabla f_i(x^\star) = 0$. Applying the standard KKT conditions, we immediately see that $x^\star$ is an optimum of the convex problem

$$\underset{x}{\text{minimize}} \quad \langle \nabla f(x^\star), x\rangle \quad \text{subject to } f_i(x) \le 0 \quad \text{for } i \in [m].$$

The dual averaging update (17) chooses x_{k+1} via

$$x_{k+1} = \underset{x}{\operatorname{argmin}} \left\{ \langle \nabla f(x^\star), x\rangle + \langle v_k, x\rangle + \frac{1}{2A_k} \|x\|^2 \;\middle|\; f_i(x) \le 0, i \in [m] \right\},$$

where $v_k = \frac{z_k}{A_k} - \nabla f(x^\star)$. Theorem 2 guarantees that $x_k \to x^\star$, while Lemma 9.6 shows that $A_k^{-1} z_k - \nabla f(x^\star) \to 0$ with probability 1. The perturbation results (Lemmas 4.1 and 4.2) immediately yield the theorem.

10. Discussion. We have developed asymptotic theory for stochastic optimization problems, showing a local asymptotic minimax lower bound and making precise connections between tilt stability in optimization and the (statistical) difficulty of solving risk minimization problems. These optimal rates of convergence are achievable by the classical M-estimator $\widehat{x}_k = \mathrm{argmin}_{x \in \mathcal{X}} \frac{1}{k} \sum_{i=1}^{k} F(x; S_i)$ (Corollary 1) and approximate versions thereof, for example, from modern incremental gradient methods [14, 29, 32, 34]. Our dual averaging (lazy projected gradient) and Riemannian stochastic gradient methods are also asymptotically optimal, though subtleties arise for nonlinear constraint sets. There are open questions about whether simpler methods—for example, methods that do not explicitly track the active manifold—can achieve these rates, and developing finite sample analogues of this theory remains an open question.

Acknowledgments. JCD and FR partially supported by NSF Award 1553086 and the Toyota Research Institute. FR additionally supported by the E.K. Potter Stanford Graduate Fellowship.

SUPPLEMENTARY MATERIAL

Supplement to “Asymptotic optimality in stochastic optimization” (DOI: 10.1214/19-AOS1831SUPP; .pdf). Supplementary information.

REFERENCES

[1] ABSIL, P.-A., BAKER, C. G. and GALLIVAN, K. A. (2007). Trust-region methods on Riemannian manifolds. *Found. Comput. Math.* **7** 303–330. MR2335248 https://doi.org/10.1007/s10208-005-0179-9

[2] ABSIL, P.-A., MAHONY, R. and SEPULCHRE, R. (2008). *Optimization Algorithms on Matrix Manifolds*. Princeton Univ. Press, Princeton, NJ. MR2364186 https://doi.org/10.1515/9781400830244

[3] AGARWAL, A., BARTLETT, P. L., RAVIKUMAR, P. and WAINWRIGHT, M. J. (2012). Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Trans. Inf. Theory* **58** 3235–3249. MR2952543 https://doi.org/10.1109/TIT.2011.2182178

[4] ALI, S. M. and SILVEY, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *J. Roy. Statist. Soc. Ser. B* **28** 131–142. MR0196777

[5] BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York. Reprint of the 1993 original. MR1623559

[6] BIRGÉ, L. (1983). Approximation dans les espaces métriques et théorie de l’estimation. *Z. Wahrsch. Verw. Gebiete* **65** 181–237. MR0722129 https://doi.org/10.1007/BF00532480

[7] BONNABEL, S. (2013). Stochastic gradient descent on Riemannian manifolds. *IEEE Trans. Automat. Control* **58** 2217–2229. MR3101606 https://doi.org/10.1109/TAC.2013.2254619

[8] BONNANS, J. F. and SHAPIRO, A. (1998). Optimization problems with perturbations: A guided tour. *SIAM Rev.* **40** 228–264. MR1624098 https://doi.org/10.1137/S0036144596302644

[9] BOTTOU, L. and BOUSQUET, O. (2007). The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems* 20.

[10] BOUMAL, N. (2014). Optimization and estimation on manifolds. Ph.D. thesis, Univ. Catholique de Louvain.

[11] BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge. MR2061575 https://doi.org/10.1017/CBO9780511804441

[12] BURKE, J. V. and MORÉ, J. J. (1994). Exposing constraints. *SIAM J. Optim.* **4** 573–595. MR1287817 https://doi.org/10.1137/0804032

[13] CSISZÁR, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.* **2** 299–318. MR0219345

[14] DEFAZIO, A., BACH, F. and SAGA, S. L.-J. (2014). A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems* 27.

[15] DEMBO, A. (2016). Lecture notes on probability theory: Stanford statistics 310. Accessed October 1, 2016. http://statweb.stanford.edu/~adembo/stat-310b/lnotes.pdf.

[16] DONOHO, D. L. and LIU, R. C. (1987). Geometrizing rates of convergence I. Technical Report 137, Dept. Statistics, Univ. California, Berkeley.

[17] DONOHO, D. L. and LIU, R. C. (1991). Geometrizing rates of convergence II. *Ann. Statist.* **19** 633–667. MR1105839 https://doi.org/10.1214/aos/1176348114

[18] DONTCHEV, A. L. and ROCKAFELLAR, R. T. (2014). *Implicit Functions and Solution Mappings*: *A View from Variational Analysis*, 2nd ed. *Springer Series in Operations Research and Financial Engineering*. Springer, New York. MR3288139 https://doi.org/10.1007/978-1-4939-1037-3

[19] DRUSVYATSKIY, D. and LEWIS, A. S. (2013). Tilt stability, uniform quadratic growth, and strong metric regularity of the subdifferential. *SIAM J. Optim.* **23** 256–267. MR3033107 https://doi.org/10.1137/120876551

[20] DUCHI, J., HAZAN, E. and SINGER, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* **12** 2121–2159. MR2825422

[21] DUCHI, J. C. and RUAN, F. (2021). Supplement to "Asymptotic optimality in stochastic optimization." https://doi.org/10.1214/19-AOS1831SUPP

[22] ERMOLIEV, Y. (1969). On the stochastic quasi-gradient method and stochastic quasi-Feyer sequences. *Kibernetika* **2** 72–83.

[23] ERMOLIEV, Y. (1983). Stochastic quasigradient methods and their application to system optimization. *Stochastics* **9** 1–36. MR0703846 https://doi.org/10.1080/17442508308833246

[24] FABIAN, V. (1973). Asymptotically efficient stochastic approximation; the RM case. *Ann. Statist.* **1** 486–495. MR0381189

[25] HARE, W. L. and LEWIS, A. S. (2004). Identifying active constraints via partial smoothness and prox-regularity. *J. Convex Anal.* **11** 251–266. MR2158904

[26] HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2722294 https://doi.org/10.1007/978-0-387-84858-7

[27] HIRIART-URRUTY, J.-B. and LEMARÉCHAL, C. (1993). *Convex Analysis and Minimization Algorithms I & II*. Springer, Berlin. MR1261420

[28] IBRAGIMOV, I. A. and HAS'MINSKII, R. Z. (1981). *Statistical Estimation*: *Asymptotic Theory*. Springer.

[29] JOHNSON, R. and ZHANG, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems* 26.

[30] KUSHNER, H. J. (1974). Stochastic approximation algorithms for constrained optimization problems. *Ann. Statist.* **2** 713–723. MR0365955

[31] 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

[32] LE ROUX, N., SCHMIDT, M. and BACH, F. (2012). A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems* 25.

[33] 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

[34] LIN, H., MAIRAL, J. and HARCHAOUI, Z. (2015). A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems* 3384–3392.

[35] NEMIROVSKI, A., JUDITSKY, A., LAN, G. and SHAPIRO, A. (2008). Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.* **19** 1574–1609. MR2486041 https://doi.org/10.1137/070704277

[36] NEMIROVSKY, A. S. and YUDIN, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization*. *Wiley-Interscience Series in Discrete Mathematics*. Wiley, New York. Translated from the Russian and with a preface by E. R. Dawson. MR0702836

[37] 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

[38] NOCEDAL, J. and WRIGHT, S. J. (2006). *Numerical Optimization*, 2nd ed. *Springer Series in Operations Research and Financial Engineering*. Springer, New York. MR2244940

[39] POLIQUIN, R. A. and ROCKAFELLAR, R. T. (1998). Tilt stability of a local minimum. *SIAM J. Optim.* **8** 287–299. MR1618790 https://doi.org/10.1137/S1052623496309296

[40] 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

[41] ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* **22** 400–407. MR0042668 https://doi.org/10.1214/aoms/1177729586

[42] ROBBINS, H. and SIEGMUND, D. (1971). A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics* (*Proc. Sympos.*, *Ohio State Univ.*, *Columbus*, *Ohio*, 1971) 233–257. MR0343355

[43] ROCKAFELLAR, R. T. (1970). *Convex Analysis*. *Princeton Mathematical Series* **28**. Princeton Univ. Press, Princeton, NJ. MR0274683

[44] SHALEV-SWHARTZ, S. and SREBRO, N. (2008). SVM optimization: Inverse dependence on training set size. In *Proceedings of the* 25*th International Conference on Machine Learning*.

[45] SHAPIRO, A. (1988). Sensitivity analysis of nonlinear programs and differentiability properties of metric projections. *SIAM J. Control Optim.* **26** 628–645. MR0937676 https://doi.org/10.1137/0326037

[46] SHAPIRO, A. (1989). Asymptotic properties of statistical estimators in stochastic programming. *Ann. Statist.* **17** 841–858. MR0994271 https://doi.org/10.1214/aos/1176347146

[47] SHAPIRO, A., DENTCHEVA, D. and RUSZCZYŃSKI, A. (2009). *Lectures on Stochastic Programming*: *Modeling and Theory*. *MPS/SIAM Series on Optimization* **9**. SIAM, Philadelphia, PA; Mathematical Programming Society (MPS), Philadelphia, PA. MR2562798 https://doi.org/10.1137/1.9780898718751

[48] STEIN, C. (1956). Efficient nonparametric testing and estimation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. I* 187–195. Univ. California Press, Berkeley and Los Angeles. MR0084921

[49] TRIPURANENI, N., STERN, M., JIN, C., REGIER, J. and JORDAN, M. I. (2017). Stochastic cubic regularization for fast nonconvex optimization. arXiv:1711.02838 [cs.LG].

[50] VAN DER VAART, A. W. (1997). Superefficiency. In *Festschrift for Lucien Le Cam* 397–410. Springer, New York. MR1462961

[51] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

[52] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*: *With Applications to Statistics*. *Springer Series in Statistics*. Springer, New York. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2

[53] VENTER, J. H. (1967). An extension of the Robbins–Monro procedure. *Ann. Math. Stat.* **38** 181–190. MR0205396 https://doi.org/10.1214/aoms/1177699069

[54] WALK, H. (1983/84). Stochastic iteration for a constrained optimization problem. *Seq. Anal.* **2** 369–385. MR0752415 https://doi.org/10.1080/07474948408836045

[55] WRIGHT, S. J. (1993). Identifiable surfaces in constrained optimization. *SIAM J. Control Optim.* **31** 1063–1079. MR1227547 https://doi.org/10.1137/0331048

[56] XIAO, L. (2010). Dual averaging methods for regularized stochastic learning and online optimization. *J. Mach. Learn. Res.* **11** 2543–2596. MR2738777

[57] ZHU, Y., CHATTERJEE, S., DUCHI, J. and LAFFERTY, J. (2016). Local minimax complexity of stochastic convex optimization. In *Advances in Neural Information Processing Systems* 29.

[58] ZINKEVICH, M. (2003). Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning*.