Sample Complexity of Smooth Stochastic Optimization*

Benjamin Armbruster†

December 28, 2007

Abstract

Let $N(\epsilon, \delta)$ be the number of samples needed when solving a stochastic program such that the objective function evaluated at the sample optimizer is within ϵ of the true optimum with probability $1 - \delta$. Previous results are of the form $N(\epsilon, \delta) = O(\epsilon^{-2} \log \delta^{-1})$. However, a smooth objective function is often locally quadratic at an interior optimum. For that case we use results on the convergence of the sample optimizers, to show that $N(\epsilon, \delta) = O(\epsilon^{-1} \log \delta^{-1})$. These results are both bounds and asymptotics. Hence we show for a common case (smooth objective functions with an interior optimum), that the number of samples needed is $O(\epsilon^{-1})$. This suggests that stochastic optimization is a practical approach for such problems.

1 Introduction

Let $F(x, \omega)$ be a function of a decision variable $x \in K \subseteq \mathbb{R}^d$ and a random variable ω. Consider a single-stage stochastic program:

$$\min_{x \in K \subseteq \mathbb{R}^d} f(x) \quad \text{where} \quad f(x) := \mathbb{E}_\omega F(x, \omega). \tag{1}$$

The goal is to choose the action x with the minimum expected cost. Since the expectations in (1) may be time consuming to compute, this problem is often approached (Shapiro and Nemirovski,

*AMS 2000 subject classifications 90C15

†Department of Management Science and Engineering, Stanford University, Stanford, California 94305-4026, USA.
B.A. gratefully acknowledges support of an NSF Graduate Research Fellowship. The author also thanks Anthony Man-Cho So, Hernan P. Awad, Yinyu Ye, Margaret L. Brandeau, Gerd Infanger, Jiawei Zhang, Sherrylyn Branchaw, and Peter W. Glynn for helpful discussions and constructive feedback.

2005) by taking n i.i.d. (independently identically distributed) samples $\omega^1, \ldots, \omega^n$ of the random variable and solving the associated deterministic *sample average problem*,

$$\min_{x \in K \subseteq \mathbb{R}^d} f_n(x; \omega^1, \ldots, \omega^n) \quad \text{where} \quad f_n(x; \omega^1, \ldots, \omega^n) := \frac{1}{n} \sum_{i=1}^{n} F(x, \omega^i). \tag{2}$$

Assume for simplicity that (1) and (2) have unique minimizers (in section 5 we discuss how this assumption may be lifted). Let x^* and x_n be the minimizers of (1) and (2) respectively, and let $z^* := f(x^*)$ be the minimum objective function value. Sometimes x^* and z^* are referred to as the *true minimizer* and the *true minimum* of the stochastic program while x_n is called the *sample minimizer* or *sample optimizer*.

We would like to find the minimum number of samples such that with high confidence (i.e., with probability $1 - \delta$), the minimizer of the sample average problem (2), x_n, is a "good solution" of the original stochastic program (1), one whose expected cost is within ϵ of the optimum:

$$N(\epsilon, \delta) := \min\{n : \Pr[f(x_n) - z^* \geq \epsilon] \leq \delta\}. \tag{3}$$

Motivating Example As a preliminary, we explicitly work out the sample complexity of a stochastic program with a quadratic objective function. This example provides intuition for our results (in subsequent sections) on the general case, because smooth objective functions are often locally quadratic at an interior optimum. Also, we will later use the explicit solution to this example to check the tightness of our general results.

Consider the stochastic program where we minimize the quadratic equation $F(x, \omega) = x^2 - 2x\omega$ on the reals, $K = \mathbb{R}$, with an uncertain parameter, $\omega \sim \mathcal{N}(0, \sigma^2)$. Since $f(x) = x^2$, the true minimizer and minimum are $x^* = z^* = 0$, and the sample minimizer is $x_n = \frac{1}{n} \sum_{i=1}^{n} \omega^i \stackrel{\mathcal{D}}{=} n^{-1/2}\omega$. Here $\stackrel{\mathcal{D}}{=}$ denotes equality in distribution. Hence the sample minimizer converges like $\mathbb{E}\,|x_n - x^*| = O(n^{-1/2})$ while (since f is quadratic) the objective function value $\mathbb{E}\, f(x_n) - z^* = \mathbb{E}\, x_n^2 = \sigma^2 n^{-1}$ converges at a faster rate. More precisely, $\Pr[f(x_n) - z^* \geq \epsilon] = \Pr[\mathcal{N}(0,1)^2 \leq n\epsilon/\sigma^2] = 2\Phi(-\sqrt{n\epsilon}/\sigma)$ where Φ is the cumulative distribution function of a standard normal random variable. Hence

$$N(\epsilon, \delta) = \left\lceil \frac{\sigma^2 (\Phi^{-1}(\delta/2))^2}{\epsilon} \right\rceil. \tag{4}$$

The main driver of the complexity is the ϵ^{-1} dependence. The complexity is less sensitive to changes in δ because $2\log\delta^{-1} \geq (\Phi^{-1}(\delta/2))^2$ and $(\Phi^{-1}(\delta/2))^2 \sim 2\log\delta^{-1}$ for small δ (see lemma 1). This inequality gives the bound

$$N(\epsilon,\delta) \leq \left\lceil \frac{2\sigma^2 \log\delta^{-1}}{\epsilon} \right\rceil. \tag{5}$$

Lemma 1. *$2\log\delta^{-1} \geq (\Phi^{-1}(\delta/2))^2$ for $\delta \in [0,1]$ and in addition $(\Phi^{-1}(\delta/2))^2 \sim 2\log\delta^{-1}$ as $\delta \to 0$.*

Proof. To prove the inequality we note that $\int_t^\infty e^{-s^2}\,\mathrm{d}s \leq \frac{e^{-t^2}}{t+\sqrt{t^2+4/\pi}}$ for $t \geq 0$ (Abramowitz and Stegun, 1972, 7.1.13). Since $\Phi(-t) = 1-\Phi(t) = \frac{1}{\sqrt{2\pi}}\int_t^\infty e^{-s^2/2}\,\mathrm{d}s$ and $\frac{1}{t+\sqrt{t^2+4/\pi}} \leq \frac{\sqrt{\pi}}{2}$ for $t \geq 0$, it follows that $\Phi(-t) \leq (1/2)e^{-t^2/2}$ for $t \geq 0$. Substituting $t = -\Phi^{-1}(\delta/2)$, we obtain $\delta \leq e^{-(\Phi^{-1}(\delta/2))^2/2}$ for $\delta \in [0,1]$. Hence $2\log\delta^{-1} \geq (\Phi^{-1}(\delta/2))^2$ for $\delta \in [0,1]$.

To prove the asymptotic claim we note that $\operatorname{erf}(t) \sim 1 - \frac{1}{t\sqrt{\pi}}e^{-t^2}$ as $t \to \infty$ (Abramowitz and Stegun, 1972, 7.1.23). Since $\Phi(-t) = 1 - \frac{1}{2}\operatorname{erf}(t/\sqrt{2})$, it follows that $\Phi(-t) \sim \frac{1}{2t\sqrt{\pi}}e^{-t^2/2}$ as $t \to \infty$. Hence $-2\log\Phi(-t) \sim t^2$ as $t \to \infty$. Substituting $t = -\Phi^{-1}(\delta/2)$, we obtain $(\Phi^{-1}(\delta/2))^2 \sim 2\log(2/\delta) \sim 2\log\delta^{-1}$ as $\delta \to 0$. □

Paper Overview The next section develops general asymptotics for interior minima along the lines of the above example. This is followed in section 3 by explicit complexity bounds. Direct attacks on the convergence of $f(x_n) \to z^*$ result in bounds of the form $N(\epsilon,\delta) = O(\epsilon^{-2}\log\delta^{-1})$. Though for interior optima, attacks on the convergence of f_n' give insight about the convergence of the sample minimizer and result in bounds of the form $N(\epsilon,\delta) = O(\epsilon^{-1}\log\delta^{-1})$. Section 4 provides a simple numerical asset allocation example. In section 5 we summarize our results, mention other cases with complexity bounds improving on $O(\epsilon^{-2})$, and suggest how to generalize our results. Proofs of the two main complexity bounds in section 3 are in the appendix.

2 Asymptotics

In this paper $'$ denotes the derivative with respect to x, ∂_i denotes the ith partial derivative with respect to x, and $\|\cdot\|$ denotes the vector or matrix 2-norm.

Beginning with Aitchison and Silvey (1958) and Huber (1967), statisticians have studied the asymptotics of optimizers in the context of maximum likelihood estimators. This work was later generalized to stochastic optimization problems with inequality constraints (Dupačová and Wets,

1988; Shapiro, 1989). (Further generalizations are in King and Rockafellar (1993) and Shapiro (1993).) These results show that $\sqrt{n}(x_n - x^*)$ converges in distribution to some random variable X, if certain regularity conditions are satisfied (see theorem 3.3 in Shapiro (1989) for a description of X). In particular, X is normally distributed for interior minima:

$$\sqrt{n}(x_n - x^*) \Rightarrow \mathcal{N}(0, \Sigma) \quad \text{where} \quad \Sigma := [f''(x^*)]^{-1}[\operatorname{Var} F'(x^*, \omega)][f''(x^*)]^{-1}. \tag{6}$$

Hence at an interior optimum, we exploit the locally quadratic convergence of $f(x_n) \to z^*$ as $x_n \to x^*$,

$$f(x_n) - z^* = \frac{1}{2} \left\| [f''(x^*)]^{1/2}(x_n - x^*) \right\|^2 + o(\|x_n - x^*\|^2) \tag{7}$$

$$n(f(x_n) - z^*) = \frac{1}{2} \left\| [f''(x^*)]^{1/2}\sqrt{n}(x_n - x^*) \right\|^2 + o(\left\|\sqrt{n}(x_n - x^*)\right\|^2), \tag{8}$$

and obtain the convergence in distribution of

$$n(f(x_n) - z^*) \Rightarrow \frac{1}{2} \left\| [f''(x^*)]^{1/2}\, \mathcal{N}(0, \Sigma) \right\|^2 = \left\| \mathcal{N}(0, \bar{\Sigma}) \right\|^2 \quad \text{by (6)} \tag{9}$$

where $\bar{\Sigma} := (1/2)[f''(x^*)]^{-1/2}[\operatorname{Var} F'(x^*, \omega)][f''(x^*)]^{-1/2}$.

Let $\epsilon(n; \delta)$ be the δ-probability confidence interval for n samples; that is $\delta = \Pr[f(x_n) - z^* \geq \epsilon(n; \delta)]$. Hence $\epsilon(n; \delta) = \bar{G}_n^{-1}(\delta)n^{-1}$ where $\bar{G}_n(r) := \Pr[n(f(x_n) - z^*) \geq r]$ is the distribution function of $n(f(x_n) - z^*)$. Above, we showed that this distribution converges to $\left\|\mathcal{N}(0, \bar{\Sigma})\right\|^2$: $\bar{G}_n \to \bar{G}$ pointwise, where $\bar{G}(r) := \Pr\left[\left\|\mathcal{N}(0, \bar{\Sigma})\right\|^2 \geq r\right]$. By the following lemma the inverses converge too, and thus for every δ,

$$\epsilon(n; \delta) \sim n^{-1}\bar{G}^{-1}(\delta) \quad (n \to \infty). \tag{10}$$

Lemma 2. *If the cumulative distribution functions $G_n \to G$ pointwise, then $G_n^{-1} \to G^{-1}$ pointwise.*

Proof. Consider $\delta \in [0, 1]$ and $\epsilon > 0$. Let $a := \min G^{-1}(\delta) - \epsilon$ and $b := \max G^{-1}(\delta) + \epsilon$ (the min and max are needed when $G^{-1}(\delta)$ is not a singleton set). Since $G(a) < \delta < G(b)$, $G_n(a) \to G(a)$, and $G_n(b) \to G(b)$, it follows that there exists n_0 such that $G_n(a) < \delta < G_n(b)$ for all $n \geq n_0$.

Hence $G_n^{-1}(\delta) \subset [a, b]$ for all $n \geq n_0$, proving our claim. □

Flipping (10) around,

$$N(\epsilon, \delta) \sim \epsilon^{-1} \bar{G}^{-1}(\delta) \quad (\epsilon \to 0). \tag{11}$$

This holds exactly (not just asymptotically) for the example given in the introduction (thus matching (4)), since in that problem $f''(x^*) = 2$, $\operatorname{Var} F'(x^*, \omega) = 4\sigma^2$, and hence $\bar{\Sigma} = \sigma^2$. Algorithms such as Genz (2003) can efficiently evaluate $\bar{G}$, allowing us to find $\bar{G}^{-1}$ using a bisection search. Let $H(n; a) := n^a e^{-n}$ and λ be the largest eigenvalue of $\bar{\Sigma}$. Lemma 4 then provides an asymptotic for $\bar{G}^{-1}(\delta)$, which when substituted into (11) gives the simple approximation,

$$N(\epsilon, \delta) \approx \frac{2\lambda \log \frac{1}{\delta}}{\epsilon}. \tag{12}$$

This matches (save for the $\lceil \cdot \rceil$) the bound for the example given in the introduction, (5), since in that problem $\lambda = \bar{\Sigma} = \sigma^2$.

Lemma 3. *$H^{-1}(\delta; a) \sim \log \delta^{-1}$ as $\delta \to 0$.*

Proof. $\log \frac{1}{H(n;a)} = n(1 + o(1))$ as $n \to \infty$. Letting $n := H^{-1}(\delta; a)$, $\log \delta^{-1} = H^{-1}(\delta; a)(1 + o(1))$ as $H^{-1}(\delta; a) \to \infty$ or equivalently as $\delta \to 0$. This proves the claim. □

Lemma 4. *$\bar{G}^{-1}(\delta) \sim 2\lambda \log \delta^{-1}$ as $\delta \to 0$.*

Proof. Note that $\lambda \chi_1^2 \leq \left\| \mathcal{N}(0, \bar{\Sigma}) \right\|^2 \leq \lambda \chi_d^2$. Hence $\bar{G}(r) = O(1) H(r/(2\lambda); O(1))$ as $r \to \infty$. Letting $\delta := \bar{G}(r)$, we have $r/(2\lambda) = H^{-1}(O(1)\delta; O(1))$. Applying lemma 3 then proves the claim. □

3 Bounds

In this section we develop bounds on $N(\epsilon, \delta)$ from bounds on $\Pr[f(x_n) - z^* \geq \epsilon]$.

Lemma 5. *Consider a function g_ϵ so that $\Pr[f(x_n) - z^* \geq \epsilon] \leq g_\epsilon(n)$. Then $N(\epsilon, \delta) \leq g_\epsilon^{-1}(\delta)$ for any δ in the range of g_ϵ.*

Proof. Follows immediately from the definition of $N(\epsilon, \delta)$ in (3). □

We now review $O(\epsilon^{-2})$ bounds before going on to derive $O(\epsilon^{-1})$ bounds for interior optima. In their survey, Shapiro and Nemirovski (2005) consider stochastic programs with an objective function $F(\cdot, \omega)$ which is Lipschitz with modulus L; defined on a feasible set $K \subset \mathbb{R}^d$ of diameter D; and of bounded magnitude $C_0 := \sup_{x,\omega} |F(x, \omega)|$. They show that

$$N(\epsilon, \delta) \leq \frac{O(1)C_0^2}{\epsilon^2} \left[\log \frac{1}{\delta} + d \log \frac{2DL}{\epsilon} + O(1) \right]. \tag{13}$$

This result is equation 2.22 of their paper where, in the notation of that paper, we assume $\delta = 0$ instead of $\delta = \epsilon/2$ and do not make the substitution $C = DL$. A similar result may possibly be obtained from theorem 3.1 of Dai et al. (2000) (after expanding β in ϵ). We prove a $O(\epsilon^{-2})$ bound by putting together pieces from Pflug (1999).

Theorem 6. *Suppose $F(\cdot, \omega)$ is continuous for every ω, $K \subseteq [0, D]^d$ the closure of an open set, $C_0 := \sup_{x,\omega} |F(x, \omega)|$, $C_1 := \mathbb{E} \sup_x \|F'(x, \omega)\|^2$, $C(\cdot)$ is a universal function, and $C_2 := \left(C\left(\frac{D\sqrt{dC_1}}{2C_0}\right) / \sqrt{2d}) \right)^d$. Then for all $\epsilon \geq 0$,*

$$\Pr[f(x_n) \geq z^* + \epsilon] \leq C_2 H\left(\frac{n\epsilon^2}{8C_0^2}; \frac{d}{2}\right) \tag{14}$$

Then by lemma 5,

$$N(\epsilon, \delta) \leq \frac{8C_0^2 H^{-1}(\delta/C_2; d/2)}{\epsilon^2}. \tag{15}$$

Thus by lemma 3, asymptotically,

$$N(\epsilon, \delta) \preceq \frac{8C_0^2 \log \delta^{-1}}{\epsilon^2} \quad (\delta \to 0). \tag{16}$$

Turning now to $O(\epsilon^{-1})$ bounds, we start by applying Talagrand's inequality to the convergence of f_n' to f' (instead of $f_n \to f$ as done in theorem 6). In the proof we again combine various pieces from Pflug (1999).

Theorem 7. *Suppose $F'(\cdot, \omega)$ is continuous for every ω, $K \subseteq [0, D]^d$ the closure of an open, convex set, $\widehat{C}_0 := \sup_{x,i,\omega} |\partial_i F(x, \omega)|$, $\widehat{C}_1 := d^{2/d} \mathbb{E} \sup_x \|F''(x, \omega)\|^2$, $C(\cdot)$ is a universal function,*

and $\widehat{C}_2 := \left(C\left(\frac{D\sqrt{d\widehat{C}_1}}{2\widehat{C}_0}\right)/\sqrt{2d}\right)^d$. *Then for all* $\Delta \geq 0$,

$$\Pr\left[\sup_x \left\|f_n'(x) - f'(x)\right\| \geq \Delta\right] \leq \widehat{C}_2 H\left(\frac{n\Delta^2}{2d\widehat{C}_0^2}; \frac{d}{2}\right). \tag{17}$$

We then obtain a uniform bound on the convergence of the sample minimizer x_n to x^* (c.f., (6)) and consequently a bound on $N(\epsilon, \delta)$ by assuming global bounds on f. These bounds are analogous to the assumption, (7), that f is locally quadratic around x^* made in the previous section for the asymptotic result.

Corollary 8. *Under the assumptions of theorem 7 and assuming* $f'(x) \cdot (x - x^*) \geq 2c \left\|x - x^*\right\|^2$ *for all* x,

$$\Pr[\|x_n - x^*\| \geq \Delta] \leq \widehat{C}_2 H\left(\frac{2n\Delta^2 c^2}{d\widehat{C}_0^2}; \frac{d}{2}\right). \tag{18}$$

Further assuming that $f(x) \leq z^* + c_2 \left\|x - x^*\right\|^2$ *for all* x,

$$N(\epsilon, \delta) \leq \frac{c_2(\widehat{C}_0/c)^2 d H^{-1}(\delta/\widehat{C}_2; d/2)}{2\epsilon} \sim \frac{c_2(\widehat{C}_0/c)^2 d \log \delta^{-1}}{2\epsilon} \quad (\delta \to 0). \tag{19}$$

Proof of corollary 8. From the growth condition, $\|f'(x_n) - f_n'(x_n)\| = \|f'(x_n)\| \geq 2c \|x_n - x^*\|$. Applying theorem 7 gives the first equation. The second equation follows from lemmas 5 and 3. □

For the motivating example with ω a random variable taking on the value -1 and $+1$ with equal probability, the above bound and the asymptotic (11) have the same asymptotic properties for small δ. We do not get such a tight complexity bound if we replace the bound $f'(x) \cdot (x - x^*) \geq 2c \left\|x - x^*\right\|^2$ with an analogous bound on f:

Corollary 9. *Under the assumptions of theorem 7 and assuming* $f(x) \geq z^* + c \left\|x - x^*\right\|^2$ *for all* x,

$$\Pr[\|x_n - x^*\| \geq \Delta] \leq \widehat{C}_2 H\left(\frac{n\Delta^2 c^2}{2d\widehat{C}_0^2}; \frac{d}{2}\right). \tag{20}$$

Further assuming that $f(x) \leq z^* + c_2 \left\|x - x^*\right\|^2$ *for all* x,

$$N(\epsilon, \delta) \leq \frac{2c_2(\widehat{C}_0/c)^2 d H^{-1}(\delta/\widehat{C}_2; d/2)}{\epsilon} \sim \frac{2c_2(\widehat{C}_0/c)^2 d \log \delta^{-1}}{\epsilon} \quad (\delta \to 0). \tag{21}$$

Proof. Using the growth condition we show that

$$c\,\|x_n - x^*\|^2 \le f(x_n) - f(x^*) = (f(x_n) - f_n(x_n)) - (f(x^*) - f_n(x_n))$$
$$\le (f(x_n) - f_n(x_n)) - (f(x^*) - f_n(x^*)). \quad (22)$$

By the mean-value theorem and the convexity of K, there exists $x \in K$ so that $\|f'(x) - f_n'(x)\| \ge c\,\|x_n - x^*\|$. (Pflug (1999) states a similar inequality on page 22.) Thus,

$$\Pr\left[\|x_n - x^*\| \ge \Delta\right] \le \Pr\left[\sup_x \left\|f_n'(x) - f'(x)\right\| \ge \Delta c\right], \quad (23)$$

which together with (17) proves the first equation. Again the second equation follows from lemmas 5 and 3. □

The d dependence of these bounds, (19) and (21), is not likely to be necessary since it does not appear in the asymptotic result, (11).

4 Asset Allocation Example

Consider an asset allocation problem where one decides to invest one dollar (or one million dollars) among $d+1$ assets having returns $\omega \in \mathbb{R}^{d+1}$ in order to maximize expected utility after one year. Letting $\tilde{x} \in \mathbb{R}^{d+1}$ be the allocation and $U(\cdot)$ the utility function, we have the following problem

$$\begin{aligned} &\max_{\tilde{x}} \ \mathbb{E}\,U(\tilde{x}^\top \omega) \\ &s.t.\ \tilde{x}^\top \mathbf{1} = 1. \end{aligned} \quad (24)$$

Let $\mathbf{1} \in \mathbb{R}^d$ be a vector of ones, $x := \tilde{x}_{1:d}$ be the allocation among the first d assets, $\omega_{1:d}$ be the returns of the first d assets, and ω_{d+1} be the return of the last asset. Hence $\tilde{x} = (x, 1 - x^\top \mathbf{1})$ and $\omega = (\omega_{1:d}, \omega_{d+1})$. Since our results are for unconstrained minimization problems, we rewrite (24) as

$$\min_x \ -\mathbb{E}\,U(x^\top \omega_{1:d} + (1 - x^\top \mathbf{1})\omega_{d+1}). \quad (25)$$

We choose a quadratic utility function $U(y) := -y^2 + 2yy_0$ valid for $y \le y_0$, so that the stochastic

program, (24), will have an analytic solution. Our numerical calculations set $y_0 = 2$ and use four assets, $d = 3$: long-term Treasury bonds, long-term corporate bonds, large stocks, and small stocks. Their annual returns, ω, are modeled by a log-normal distribution parameterized using the mean, standard deviation, and correlation of the annual, total, real returns from 1926–2000 (Ibbotson, 2001, Tables 6-5 and 6-7):

$$\mathbb{E}\,\omega = 1 + \begin{bmatrix} 5.7 \\ 6.0 \\ 13.0 \\ 17.3 \end{bmatrix} \cdot 10^{-2}, \quad \sigma_\omega = \begin{bmatrix} 9.4 \\ 8.7 \\ 20.2 \\ 33.4 \end{bmatrix} \cdot 10^{-2}, \quad \rho_\omega = \begin{bmatrix} 1.00 & 0.95 & 0.24 & 0.04 \\ 0.95 & 1.00 & 0.30 & 0.12 \\ 0.24 & 0.30 & 1.00 & 0.79 \\ 0.04 & 0.12 & 0.79 & 1.00 \end{bmatrix}. \tag{26}$$

Figure 1 compares the actual convergence of the 90% percentile (i.e., $\delta = 0.1$) with the asymptotic result (11), and the $O(\epsilon^{-2})$ bound from the previous section (specifically its small-δ asymptotic, (16)). The small-δ approximation (12) of the asymptotic result is not shown because it is visually indistinguishable from the asymptotic result, (11), ($2\lambda \log \delta^{-1} \approx 3.6$ and $\bar{G}^{-1}(\delta) \approx 3.8$). Since the complexity bounds from the previous section require a bounded set of allocations and returns, they are calculated for a variant of the problem. In this variant we assume that K is a small neighborhood of x^* and that the distribution of returns is truncated to $[0.5, 1.8]^4$ (this includes 95% of the original probability mass). Larger sets of feasible allocations, K, and larger supports for the distribution of returns will lead to higher complexity bounds. The $O(\epsilon^{-1})$ complexity bounds from the previous section, (19) and (21), are not shown because they would be off the scale, $c_2(\widehat{C}_0/c)^2 \log \delta^{-1} \approx 10^8$. The figure shows that the asymptotic result is quite accurate even for a small number of samples. On the other hand, the bounds from the previous section are not of much practical use, despite their theoretical importance.

5 Conclusion

This paper defines a complexity measure $N(\epsilon, \delta)$ for the sample size needed to adequately approximate a stochastic program with its sample average problem. It then gives an asymptotic result on $N(\epsilon, \delta)$, (11), (and a handy approximation, (12)) and proves several complexity bounds (and their small-δ asymptotics), (15), (16), (19), and (21). These results (except for (15) and (16)) are the

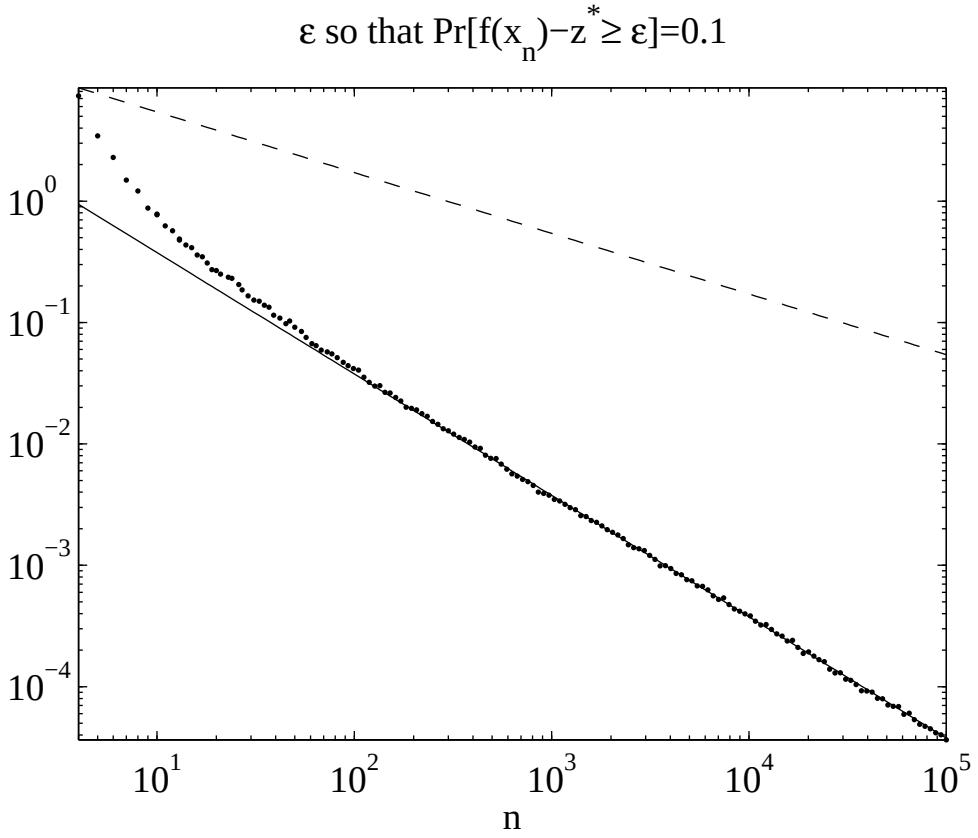

Figure 1: This figure shows the 90% percentile of $f(x_n) - z^*$ from the asset allocation example: with 90% probability, $f(x_n) - z^*$ is below the indicated value. The dots show the actual results calculated via simulation, the solid line shows the asymptotic result, (11), and the dashed line shows the small-δ asymptotic of the $O(\epsilon^{-2})$ complexity bound, (16).

first we know of using only $O(\epsilon^{-1})$ samples of the random variable. Since (11) is an asymptotic result, the exponent cannot be improved. These results achieve $O(\epsilon^{-1})$ complexity by assuming that the objective function, f, is in some sense quadratic (either globally or locally at x^* depending on the type of result) and then study the convergence of $x_n \to x^*$. These results improve upon the $O(\epsilon^{-2})$ bounds obtained using direct approaches in the literature and this paper, (15).

Our results are more general than they may first seem. Even if F is nonsmooth, it is still quite likely that f is smooth. For example, if $F(x, \omega) = |x - \omega|$ and $\omega \sim \mathrm{U}[-1, 1]$, then f is quadratic around x^* and our asymptotic result from section 2 implies $O(\epsilon^{-1})$ sample complexity. Other stochastic programs with better than $O(\epsilon^{-2})$ complexity include those with discrete distributions (Shapiro and Homem-de-Mello, 2000) and those where the set of possible sample minimizers, $\{x_n : n, \omega^1, \ldots, \omega^n\}$, is finite. In both cases one can argue that $N(\epsilon, \delta) = O(\log \delta^{-1})$ and does not depend

on ϵ. When the set of possible sample minimizers, $X := \{x_n : n, \omega^1, \ldots, \omega^n\}$, is finite, there is no ϵ dependence when ϵ is small, only the question of whether $x_n = x^*$. In this case, we merely need to compare $f(x)$ for all $x \in X$, and thus, N only has a δ dependence: $N(\epsilon, \delta) = O(\log \delta^{-1})$. Thus there are many classes of stochastic programs with sample complexity better than $O(\epsilon^{-2})$. Our future work will study further types of stochastic programs.

A number of our assumptions are technical in nature and are probably not necessary for our results. For example, our assumption that the minimizers of (1) and (2) are unique may not be necessary. In addition, the restriction on K in theorems 6 and 7 (that K is the closure of an open, convex set) can probably be weakened to allow for the case where K is a subset of a manifold (e.g., arising from some equality constraints); then d would be the dimension of the manifold.

The assumptions that F, F', and K are bounded are more deeply embedded in our bounds and may be necessary for the results. Nevertheless, it is of great practical interest to determine what results hold without these assumptions. (To apply our complexity bounds to the asset allocation example we had to bound the possible allocations and returns, which we did not need to for our asymptotic result.) In addition, our $O(\epsilon^{-1})$ bounds ((19) and (21)) would be more useful if they more closely matched the asymptotic result.

The apparently large number of samples needed to achieve accurate results is a primary reason for the infrequent use of stochastic optimization. In this paper we showed that for a common case (smooth objective functions with an interior optimum), that the number of samples needed is $O(\epsilon^{-1})$. This suggests that stochastic optimization is a practical approach for such problems.

Appendix: Proofs of Theorems 6 and 7

Define the set of functions $\mathcal{F} := \{F(x, \cdot) : x \in K\}$ and $\mathcal{F}_i := \{\partial_i F(x, \cdot) : x \in K\}$ for $i = 1, \ldots, d$.

Definition 10 (Definition 3 of Pflug (1999)). *A set $T \subset \mathbb{R}^d$ is said to be of covering type (ν, V) if $\left\lfloor \left(\frac{V+o(1)}{\epsilon} \right)^{\nu} \right\rfloor$ balls of diameter of ϵ suffice to cover T as $\epsilon \to 0$.*

By stacking the cubes inscribed in balls of diameter ϵ, we show that $[0, D]^d$ is of covering type $(d, D\sqrt{d})$. So, if $K \subseteq [0, D]^d$, then it also is of covering type $(d, D\sqrt{d})$.

Definition 11 (Definition 5 of Pflug (1999)). *A family $\mathcal{G}$ of random variables is said to*

be of covering type (ν, V) *if as* $\epsilon \to 0$, $\left\lfloor \left(\frac{V+o(1)}{\epsilon} \right)^{\nu} \right\rfloor$ *pairs of random variables* $(\underline{X}_i, \overline{X}_i)$ *with* $\mathbb{E}(\overline{X}_i - \underline{X}_i)^2 \leq \epsilon^2$ *suffice to cover* $\mathcal{G}$*: so that for any* $X \in \mathcal{G}$*, there exists a pair* i *such that* $\underline{X}_i \leq X \leq \overline{X}_i$ *a.s.*

Here are lemma 10 (taking $\beta = 1$) and part (i) of lemma 11 from Pflug (1999) (where in the latter, we replace a Lipschitz condition by a stronger differentiability condition).

Lemma 12. *If* K *is of covering type* (ν, V)*, then* $\mathcal{F}$ *is of covering type* (ν, VC) *if for all balls* B *in* $\mathbb{R}^d$*,* $diam\{F(x,\cdot) : x \in B\} \leq C diam(B)$*.*

Lemma 13. *For all balls* B *in* $\mathbb{R}^d$*,* $diam\{F(x,\cdot) : x \in B\} \leq diam(B)\sqrt{\mathbb{E}\sup_x \|F'(x,\omega)\|^2}$*.*

The following corollary follows from these lemmas.

Corollary 14. *Let* $C_1 := \mathbb{E}\sup_x \|F'(x,\omega)\|^2$ *and* $\widehat{C}_1 := d^{2/d}\,\mathbb{E}\sup_x \|F''(x,\omega)\|^2$*. If* $K \subseteq [0,D]^d$*, then* $\mathcal{F}$ *is of covering type* $(d, D\sqrt{dC_1})$ *and* $\cup_i \mathcal{F}_i$ *is of type* $(d, D\sqrt{d\widehat{C}_1})$*.*

Proof. Since K of covering type $(d, D\sqrt{d})$, the covering type of $\mathcal{F}$ follows directly from applying the above two lemmas. Applying them to the partial derivatives, we find for $i = 1, \ldots, d$ that $\mathcal{F}_i$ is of covering type $(d, D\sqrt{dC_{1i}})$ where $C_{1i} := \mathbb{E}\sup_x \|(\partial_i F(x,\omega))'\|^2$. Since $\|(\partial_i F(x,\omega))'\| \leq \|F''(x,\omega)\|$ for all x and ω, it follows that $C_{1i} \leq d^{-2/d}\widehat{C}_1$. So as $\epsilon \to 0$,

$$\sum_i \left\lfloor \left(\frac{D\sqrt{dC_{1i}} + o(1)}{\epsilon} \right)^d \right\rfloor \leq d \left\lfloor \left(\frac{D\sqrt{dd^{-2/d}\widehat{C}_1} + o(1)}{\epsilon} \right)^d \right\rfloor$$
$$\leq \left\lfloor d \left(\frac{D\sqrt{dd^{-2/d}\widehat{C}_1} + o(1)}{\epsilon} \right)^d \right\rfloor = \left\lfloor \left(\frac{D\sqrt{d\widehat{C}_1} + o(1)}{\epsilon} \right)^d \right\rfloor \quad (27)$$

random variables suffice to cover $\cup_i \mathcal{F}_i$. Hence $\cup_i \mathcal{F}_i$ is of covering type $(d, D\sqrt{d\widehat{C}_1})$. □

The following theorem (a variant of theorem 1.3 of Talagrand (1994)) is at the heart of our analysis. It is theorem 6 of Pflug (1999) with $M\sqrt{n}$ substituted for M. Remark 7 of Pflug (1999) notes that the theorem holds even though we have a $o(1)$ term in the previous two definitions of covering type. It is likely that slightly tighter results could be achieved using theorem 3 or 4 of Massert (2000). In particular he gives a result for vectors.

Theorem 15. *Let $C(\cdot)$ be a universal function. If $\mathcal{F}$ is of covering type (ν, V) and $|F(x,\omega)| \leq C_0$ for all x, ω, then for all $M \geq 0$,*

$$\Pr\left[\sup_{x\in K} |f_n(x) - f(x)| \geq M\right] \leq \left(\frac{C(V/2C_0)}{\sqrt{2\nu}}\right)^{\nu} H\left(\frac{nM^2}{2C_0^2}; \frac{\nu}{2}\right), \tag{28}$$

if the supremum is measurable.

Lemma 8 of Pflug (1999) gives conditions for the supremum to be measurable:

Lemma 16. *The random variable $\sup_{x\in K} |f_n(x) - f(x)|$ is measurable if (i) K is the closure of $K \cap \mathbb{Q}^d$, (ii) $F(\cdot, \omega)$ is lower semicontinuous for every ω, and (iii) $f(\cdot)$ is continuous.*

The proof of the following lemma follows from the density of the rationals, $\mathbb{Q}$.

Lemma 17. *If K is the closure of an open set, then K is the closure of $K \cap \mathbb{Q}^d$.*

We now prove our main results by combining theorem 15 with the above lemma and corollary 14.

Proof of theorem 6. From corollary 14 it follows that $\mathcal{F}$ is of covering type $(d, D\sqrt{dC_1})$. When we now apply theorem 15 the supremum is measurable because condition (i) of lemma 16 holds due to lemma 17 and conditions (ii) and (iii) hold by the continuity of $F(\cdot, \omega)$. Thus for all $M \geq 0$,

$$\Pr\left[\sup_{x} |f_n(x) - f(x)| \geq M\right] \leq C_2 H\left(\frac{nM^2}{2C_0^2}; \frac{d}{2}\right). \tag{29}$$

Substituting $M = \epsilon/2$,

$$\Pr\left[\sup_{x} |f_n(x) - f(x)| \geq \epsilon/2\right] \leq C_2 H\left(\frac{n\epsilon^2}{8C_0^2}; \frac{d}{2}\right). \tag{30}$$

Note that,

$$\Pr[|f(x_n) - f_n(x_n)| + |f_n(x^*) - f(x^*)| \geq \epsilon] \leq \Pr\left[2\sup_{x} |f_n(x) - f(x)| \geq \epsilon\right]. \tag{31}$$

Applying the triangle inequality,

$$\Pr[f(x_n) - f_n(x_n) + f_n(x^*) - f(x^*) \geq \epsilon] \leq \Pr\left[\sup_{x} |f_n(x) - f(x)| \geq \epsilon/2\right], \tag{32}$$

and the fact that $f_n(x^*) \geq f_n(x_n)$, we obtain

$$\Pr[f(x_n) - z^* \geq \epsilon] \leq \Pr\left[\sup_x |f_n(x) - f(x)| \geq \epsilon/2\right]. \tag{33}$$

Substituting the above inequality into (30) proves (14). □

Proof of theorem 7. From corollary 14 it follows that $\cup_i \mathcal{F}_i$ is of covering type $(d, D\sqrt{d\widehat{C}_1})$. We now apply theorem 15 to $\cup_i \mathcal{F}_i$. Thus for all $M \geq 0$,

$$\Pr\left[\sup_{x,i} |\partial_i f_n(x) - \partial_i f(x)| \geq M\right] \leq \widehat{C}_2 H\left(\frac{nM^2}{2\widehat{C}_0^2}; \frac{d}{2}\right) \tag{34}$$

when the supremum is measurable. For each i, the random variable $\sup_x |\partial_i f_n(x) - \partial_i f(x)|$ is measurable because condition (i) of lemma 16 holds due to lemma 17 and conditions (ii) and (iii) hold by the continuity of $F'(\cdot, \omega)$. Taking the maximum of these random variables over $i = 1, \ldots, d$ implies that $\sup_{x,i} |\partial_i f_n(x) - \partial_i f(x)|$ is also measurable. Having established the measurability of (34) we apply a norm inequality,

$$\Pr\left[\sup_x \left\|f_n'(x) - f'(x)\right\| \geq M\sqrt{d}\right] \leq \widehat{C}_2 H\left(\frac{nM^2}{2\widehat{C}_0^2}; \frac{d}{2}\right), \tag{35}$$

and substitute $M = \Delta/\sqrt{d}$,

$$\Pr\left[\sup_x \left\|f_n'(x) - f'(x)\right\| \geq \Delta\right] \leq \widehat{C}_2 H\left(\frac{n\Delta^2}{2d\widehat{C}_0^2}; \frac{d}{2}\right). \tag{36}$$

□

References

M. Abramowitz and I. A. Stegun, editors. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. National Bureau of Standards, 1972. URL http://mintaka.sdsu.edu/faculty/wfw/ABRAMOWITZ-STEGUN/frameindex.htm.

J. Aitchison and S. D. Silvey. Maximum-likelihood estimation of parameters subject to restraints. *Ann. Math. Statist.*, 29:813–828, 1958.

L. Dai, C. H. Chen, and J. R. Birge. Convergence properties of two-stage stochastic programming. *J. Optim. Theory Appl.*, 106(3):489–509, 2000.

J. Dupačová and R. Wets. Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems. *Ann. Statist.*, 16(4):1517–1549, 1988.

A. Genz. MVNLPS: a Matlab function for the numerical computation of multivariate normal distribution values for ellipsoidal integration regions. 2003. URL http://www.math.wsu.edu/faculty/genz.

P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 221–233. 1967.

Ibbotson. *Stocks, Bonds, Bills, and Inflation, 2001 Yearbook*. Ibbotson Associates, 2001.

A. J. King and R. T. Rockafellar. Asymptotic theory for solutions in statistical estimation and stochastic programming. *Math. Oper. Res.*, 18(1):148–162, 1993.

P. Massert. About the constants in Talagrand's concentration inequalities for empirical processes. *Annals of Probability*, 28(2):863–884, 2000.

G. C. Pflug. Stochastic programs and statistical data. *Ann. Oper. Res.*, 85:59–78, 1999.

A. Shapiro. Asymptotic properties of statistical estimators in stochastic programming. *Ann. Statist.*, 17(2):841–858, 1989.

A. Shapiro. Asymptotic behavior of optimal solutions in stochastic programming. *Math. Oper. Res.*, 18(4):829–845, 1993.

A. Shapiro and T. Homem-de-Mello. On the rate of convergence of optimal solutions of Monte Carlo approximations of stochastic programs. *SIAM J Optim*, 11(1):70–86, 2000.

A. Shapiro and A. Nemirovski. On complexity of stochastic programming problems. In V. Jeyakumar and A. M. Rubinov, editors, *Continuous Optimization: Current Trends and Applications*, pages 111–144. Springer, 2005. URL http://www.optimization-online.org/DB_HTML/2004/10/978.html.

M. Talagrand. Sharper bounds for Gaussian and empirical processes. *Ann. Probab.*, 22(1):28–76, 1994.