Sampling Bounds for Stochastic Optimization

Moses Charikar* Chandra Chekuri† Martin Pál‡

April 14, 2005

Abstract

A large class of stochastic optimization problems can be modeled as minimizing an objective function f that depends on a choice of a vector $x \in X$, as well as on a random external parameter $\omega \in \Omega$ given by a probability distribution π. The value of the objective function is a random variable and often the goal is to find an $x \in X$ to minimize the expected cost $E_\omega[f_\omega(x)]$. Each ω is referred to as a *scenario*. We consider the case when Ω is large or infinite and we are allowed to sample from π in a black-box fashion. A common method, known as the SAA method (sample average approximation), is to pick sufficiently many independent samples from π and use them to approximate π and correspondingly $E_\omega[f_\omega(x)]$. This is one of several scenario reduction methods used in practice.

There has been substantial recent interest in two-stage stochastic versions of combinatorial optimization problems which can be modeled by the framework described above. In particular, we are interested in the model where a parameter λ bounds the relative factor by which costs increase if decisions are delayed to the second stage. Although the SAA method has been widely analyzed, the known bounds on the number of samples required for a $(1+\varepsilon)$ approximation depend on the variance of π even when λ is assumed to be a fixed constant. Shmoys and Swamy [13, 14] proved that a polynomial number of samples suffice when f can be modeled as a linear or convex program. They used modifications to the ellipsoid method to prove this.

In this paper we give a different proof, based on earlier methods of Kleywegt, Shapiro, Homem-De-Mello [6] and others, that a polynomial number of samples suffice for the SAA method. Our proof is not based on computational properties of f and hence also applies to integer programs. We further show that small variations of the SAA method suffice to obtain a bound on the sample size even when we have only an approximation algorithm to solve the sampled problem. We are thus able to extend a number of algorithms designed for the case when π is given explicitly to the case when π is given as a black-box sampling oracle.

*Computer Science Dept. , Princeton University, Princeton, NJ 08544. Supported by NSF ITR grant CCR-0205594, DOE Early Career Principal Investigator award DE-FG02-02ER25540, NSF CAREER award CCR-0237113, an Alfred P. Sloan Fellowship and a Howard B. Wentz Jr. Junior Faculty Award. moses@cs.princeton.edu

†Lucent Bell Labs, 600 Mountain Avenue, Murray Hill, NJ 07974. chekuri@reserch.bell-labs.com

‡DIMACS Center, CoRE Bldg 4th Floor, Rutgers University, Piscataway, NJ 08854. mpal@dimacs.rutgers.edu

1 Introduction

Uncertainty in data is a common feature in a number of real world problems. Stochastic optimization models uncertain data using probability distributions. In this paper we consider problems that are modeled by the two-stage stochastic minimization program

$$\min_{x \in X} f(x) = c(x) + \mathbf{E}_\omega[q(x, \omega)]. \tag{1}$$

An important context in which the problem (1) arises is *two-stage stochastic optimization with recourse*. In this model, a *first-stage decision* $x \in X$ has to be made while having only probabilistic information about the future, represented by the probability distribution π on Ω. Then, after a particular future *scenario* $\omega \in \Omega$ is realized, a *recourse action* $r \in R$ may be taken to ensure that the requirements of the scenario ω are satisfied. In the two-stage model, $c(x)$ denotes the cost of taking the first-stage action x. The cost of the second stage in a particular scenario ω, given a first-stage action x, is usually given as the optimum of the second-stage minimization problem

$$q(x, \omega) = \min_{r \in R} \{\text{cost}_\omega(x, r) \mid (x, r) \text{ is a feasible solution for scenario } \omega\}.$$

We give an example to illustrate some of the concepts. Consider the following facility location problem. We are given an a finite metric space in the form of a graph that represents the distances in some underlying transportation network. A company wants to build service centers at a number of locations to best serve the demand for its service. The objective is to minimize the cost of building the service centers subject to the constraint that each demand point is within a distance B from its nearest center. However, at the time that the company plans to build the service centers, there could be uncertainty in the demand locations. One way to deal with this uncertainty is to make decisions in two or more stages. In the first stage certain service centers are built and in the second stage when there is a clearer picture of the demand, additional service centers might be built and so on. How does one make these decisions so that company minimizes the overall cost of building service centers? Clearly, if building centers in the second stage is no more expensive than building them in the first stage, then the company will wait. However, very often companies cannot wait to make decisions. Building centers takes time and there are other costs such as inflation which makes it advantageous to build some first stage centers. We can assume that building a center in the second stage costs at most some $\lambda \geq 1$ times more than building it in the first stage. This tradeoff is captured by (1) as follows. X is the set of all n-dimensional binary vectors where n is the number of potential locations for the service centers. A binary vector x indicates which centers are built in the first stage and $c(x)$ is the cost of building them. The uncertainty in demand locations is modeled by the probability distribution on Ω. A scenario $\omega \in \Omega$ is characterized by set of demand locations in ω. Thus, given ω and the first stage decision x, the recourse action is to build additional centers so that all demands in ω are within a distance B of some center. The cost of building these additional centers is given by $q(x, \omega)$. Note that $q(x, \omega)$ is itself an optimization problem very closely related to the original k-center problem.

How does one solve problems modeled by (1)? One key issue is how the probability distribution π is specified. In some cases the number of scenarios in Ω is small and π is explicitly known. In such cases the problem can be solved as a deterministic problem using whatever mathematical programming method (linear, integer, non-linear) is applicable for minimizing c and q. There are however situations in which Ω is too large or infinite and it is infeasible to solve the problem by explicitly listing all the scenarios. In such cases, a natural approach to minimize the function $f(\cdot)$ is to approximate it by taking some number N

of independent samples $\omega_1, \ldots, \omega_N$ from the distribution π and approximate the function f by the *sample average function*

$$\hat{f}(x) = c(x) + \frac{1}{N} \sum_{i=1}^{N} q(x, \omega_i). \tag{2}$$

If the number of samples N is not too large, finding an $\hat{x}$ that minimizes $\hat{f}(\hat{x})$ may be easier than the task of minimizing f. One might then hope that for a suitably chosen sample size N, a good solution $\hat{x}$ to the sample average problem would be a good solution to the problem (1). This approach is called the *sample average approximation* (SAA) method. The SAA method is an example of a *scenario reduction* technique, in that it replaces a complex distribution π over a large (or even infinite) number of scenarios by a simpler, empirical distribution π' over N observed scenarios. Since the function (2) is now deterministic, we can use tools and algorithms from deterministic optimization to attempt to find its exact or approximate optimum.

The SAA method is well known and falls under the broader area of Monte Carlo sampling. It is used in practice and has been extensively studied and analyzed in the stochastic programming literature. See [9, 10] for numerous pointers. In a number of settings, in particular for convex and integer programs, it is known that the SAA method converges to the true optimum as N tends to ∞. The number of samples required to obtain an *additive* ε approximation with a probability $(1 - \delta)$ has also been analyzed [6]; it is known to be polynomial in the dimension of X, $1/\varepsilon$, $\log 1/\delta$ and the quantity $V = \max_{x \in X} V(x)$ where $V(x)$ is the variance of the random variable $q(x, w)$. This factor V need not be polynomial in the input size even when π is given explicitly.

The two-stage model with recourse has gained recent interest in the theoretical computer science community following the work of Immorlica et al. [5] and Ravi and Sinha [11]. Several subsequent works have explored this topic [3, 4, 7, 1, 13, 14, 2]. The emphasis in these papers is primarily on *combinatorial optimization* problems such as shortest paths, spanning trees, Steiner trees, set cover, facility location, and so on. Most of these problems are NP-hard even when the underlying singe stage problem is polynomial time solvable, for example the spanning tree problem. Thus the focus has been on approximation algorithms and in particular on *relative* error guarantees. Further, for technical and pragmatic reasons, an additional parameter, the *inflation factor* has been introduced. Roughly speaking, the inflation factor, denoted by λ, upper bounds the relative factor by which the second stage decisions are more expensive when compared to the first stage. It is reasonable to expect that λ will be a small constant, say under 10 in many practical situations. The boosted sampling algorithm of [3] for example, takes number of samples proportional to λ.

In this new model, Swamy and Shmoys [13] showed for a large and interesting class of problems modeled by linear programs, a relative $(1 + \varepsilon)$ approximation can be obtained with a number of samples that is polynomial in the input size, λ, and $1/\varepsilon$. They also established that a polynomial dependence on λ is necessary. Thus the dependence on V is eliminated. Their first result [13] does not establish the guarantee for the SAA method but in subsequent work [14], they established similar bounds for the SAA method. Their proof is based on the ellipsoid method where the samples are used to compute approximate sub-gradients for the separation oracle. We note two important differences between these results when compared to earlier results of [6]. The first is that the new bounds obtained in [13, 14] do not guarantee that the *value* of f can be approximated, only that the optimum solution $\bar{x}$ to the sampled problem $\hat{f}$ satisfies the property that $f(\bar{x}) \leq (1 + \varepsilon) f(x^*)$ where x^* is an optimum solution to f. It can be shown that an approximation to the value requires the sample size to depend on V. Second, the new bounds, by relying on the ellipsoid method, limit the applicability to when X is a continuous space while the earlier methods applied even when X is a discrete set and hence could capture integer programming problems. In fact, in [6], the SAA method is analyzed for the discrete case, and the continuous case is analyzed by discretizing the space X using a fine grid. The discrete case is of particular interest in approximation algorithms for combinatorial optimization since they can be modeled as integer programs with X being the set of all possible integer solutions. Several recent results [4, 5, 7] have obtained algorithms that have provably good approximation ratios when π is given explicitly. An important and useful question is whether these results can be carried over to the case when π can only be sampled from. We answer this question in the positive. Details of our results follow.

Results: In this paper we show that the results of Shmoys and Swamy [13, 14] can also be derived using a modification to the basic analysis framework in the methods of [10, 6]. This results in a simple proof relying only on Chernoff bounds. Similar to earlier work [10, 6], the proof shows that the SAA method works because of *statistical* properties of the underlying space X and the function f, and not on *computational* properties of optimizing over X. This allows us to prove a result for the discrete case and hence we obtain bounds for integer programs. The sample size that we guarantee is strongly polynomial in the input size: in the discrete setting it depends on $\log |X|$ and in the continuous setting it depends on the dimension of the space containing X. We also extend the ideas to approximation algorithms. In this case, the plain SAA method achieves guarantee of $(1 + \varepsilon)$ times the guarantee of the underlying approximation algorithm only with $O(\epsilon)$ probability. To obtain a probability of $1 - \delta$ for any given δ, we analyze two minor variants of SAA. The first variant repeats SAA $O(\frac{1}{\varepsilon} \log \frac{1}{\delta})$ times independently and picks the solution with the smallest sample value. The second variant rejects a small fraction of high cost samples before running the approximation algorithm on the samples. This latter result was also obtained independently by Ravi and Singh [12].

2 Preliminaries

In the following, we consider the stochastic optimization problem in its general form (1). We consider stochastic two stage problems that satisfy the following properties.

(A1) **Non-negativity.** The functions $c(x)$ and $q(x, \omega)$ are non-negative for every first stage action x and every scenario ω.

(A2) **Empty First Stage.** We assume that there is an empty first-stage action, $0 \in X$. The empty action incurs no first-stage cost (i.e. $c(0) = 0$), but is least helpful in the second stage. That is, for every $x \in X$ and every scenario ω, $q(x, \omega) \leq q(0, \omega)$.

(A3) **Bounded Inflation Factor.** The inflation factor λ determines the relative cost of information. It compares the difference in cost of the "wait and see" solution $q(0, \omega)$ and the cost of the best solution with hindsight, $q(x, \omega)$, relative to the first stage cost of x. Formally, the inflation factor $\lambda \geq 1$ is the least number such that for every scenario $\omega \in \Omega$ and every $x \in X$, we have

$$q(0, \omega) - q(x, \omega) \leq \lambda c(x). \tag{3}$$

We note that the above assumptions capture both the discrete and continuous case problems considered in recent work [5, 11, 3, 13, 14, 1].

We work with exact and approximate minimizers of the sampled problem. We make the notion precise.

Definition 1 *An $x^* \in X$ is said to be an* exact minimizer *of the function $f(\cdot)$ if for all $x \in X$ it holds that $f(x^*) \leq f(x)$. An $\bar{x} \in X$ is an* α-approximate minimizer *of the function $f(\cdot)$, if for all $x \in X$ it holds that $f(x^*) \leq \alpha f(x)$.*

The main tool we will be using is the Chernoff bound. We will be using the following version of the bound (see e.g. [8]).

Fact 2 (Chernoff bound) *Let $X_1, \ldots, X_N$ be independent random variables with $X_i \in [0,1]$ and let $X = \sum_{i=1}^N X_i$. Then, for any $\varepsilon \geq 0$, we have $\Pr[|X - \mathbf{E}[X]| > \varepsilon N] \leq 2\exp(-\varepsilon^2 N)$.*

Throughout the rest of the paper when we refer to an event happening with probability β, it is with respect to the randomness in sampling from the distribution π over Ω.

3 Discrete Case

We start by discussing the case when the first stage decision x ranges over a finite set of choices X. In the following, let x^* denote an optimal solution to the true problem (1), and Z^* its value $f(x^*)$. In the following we assume that ε is small, say $\varepsilon < 0.1$.

Theorem 3 *Any exact minimizer $\bar{x}$ of the function $\hat{f}(\cdot)$ constructed with $\Theta(\lambda^2 \frac{1}{\epsilon^4} \log |X| \log \frac{1}{\delta})$ samples is, with probability $1 - 2\delta$, a $(1 + O(\varepsilon))$-approximate minimizer of the function $f(\cdot)$.*

In a natural attempt to prove Theorem 3, one might want to show that if N is large enough, the functions $\hat{f}$ will be close to f, in that with high probability $|f(x) - \hat{f}(x)| \leq \varepsilon f(x)$. Unfortunately this may not be the case, as for any particular x, the random variable $q(x, \omega)$ may have very high variance. However, intuitively, the high variance of $q(x, \omega)$ can only be caused by a few "disaster" scenarios of very high cost but low probability, whose cost is not very sensitive to the particular choice of the first stage action x. Hence these high cost scenarios would affect all x in the same way. We formalize this intuition below.

For the purposes of the analysis, we divide the scenarios into two classes. We call a scenario ω *high*, if its second stage "wait and see" cost $q(0, \omega)$ exceeds a threshold M, and *low* otherwise. We set the threshold M to be $\lambda Z^*/\varepsilon$.

We approximate the function f by taking N independent samples $\omega_1, \ldots, \omega_N$. We define the following two functions to account for the contributions of low and high scenarios respectively.

$$\hat{f}_l(x) = \frac{1}{N} \sum_{i:\omega_i \text{ low}} q(x, \omega_i) \quad \text{and} \quad \hat{f}_h(x) = \frac{1}{N} \sum_{i:\omega_i \text{ high}} q(x, \omega_i).$$

Note that $\hat{f}(x) = c(x) + \hat{f}_l(x) + \hat{f}_h(x)$. We make a similar definition for the function $f(\cdot)$. Let $p = \Pr_\omega[\omega \text{ is a high scenario}]$.

$$f_l(x) = \mathbf{E}[q(x, \omega) | \omega \text{ is low}] \cdot (1 - p) \quad \text{and} \quad f_h(x) = \mathbf{E}[q(x, \omega) | \omega \text{ is high}] \cdot p$$

so that $f(x) = c(x) + f_l(x) + f_h(x)$.

We need the following bound on p.

Fact 4 *The probability mass p of high scenarios is at most $\frac{\varepsilon}{(1-\varepsilon)\lambda}$.*

Proof: Recall that x^* is a minimizer of f, and hence $Z^* = f(x^*)$. We have

$$Z^* \geq f_h(x^*) = p \cdot \mathbf{E}[q(x^*, \omega) \mid \omega \text{ is high}] \geq p \cdot [M - \lambda c(x^*)],$$

where the inequality follows from the fact that for a high scenario ω, $q(0,\omega) \geq M$ and by Axiom (A3), $q(0,\omega) \geq q(x,\omega) - \lambda c(x)$. Substituting $M = \lambda Z^*/\varepsilon$, and using that $c(x^*) \leq Z^*$ we obtain

$$Z^* \geq Z^*(\frac{\lambda}{\varepsilon} - \lambda)p.$$

Solving for p proves the claim. ■

To prove Theorem 3, we show that each of the following properties hold with probability at least $1 - \delta$ (in fact, the last property holds with probability 1).

(P1) For every $x \in X$ it holds that $|f_l(x) - \hat{f}_l(x)| \leq \varepsilon Z^*$.

(P2) For every $x \in X$ it holds that $\hat{f}_h(0) - \hat{f}_h(x) \leq 2\varepsilon c(x)$.

(P3) For every $x \in X$ it holds that $f_h(0) - f_h(x) \leq 2\varepsilon c(x)$.

Proving that these properties hold is not difficult, and we will get to it shortly; but let us first show how they imply Theorem 3.

Proof of Theorem 3: With probability $1 - 2\delta$, we can assume that all three properties (P1–P3) hold. For any $x \in X$ we have

$$\begin{array}{rcl} f_l(x) & \leq & \hat{f}_l(x) + \varepsilon Z^* \quad \text{by (P1)} \\ f_h(x) & \leq & f_h(0) \quad \text{by (A2)} \\ 0 & \leq & \hat{f}_h(x) + 2\varepsilon c(x) - \hat{f}_h(0) \quad \text{by (P2)} \end{array}$$

Adding the above inequalities and using the definitions of functions f and $\hat{f}$ we obtain

$$f(x) - \hat{f}(x) \leq 2\varepsilon Z^* + 2\varepsilon c(\bar{x}) + f_h(0) - \hat{f}_h(0). \tag{4}$$

By a similar reasoning we get the opposite inequality

$$\hat{f}(x) - f(x) \leq 2\varepsilon Z^* + 2\varepsilon c(x) + \hat{f}_h(0) - f_h(0). \tag{5}$$

Now, let x^* and $\bar{x}$ be minimizers of the functions $f(\cdot)$ and $\hat{f}(\cdot)$ respectively. Hence we have $\hat{f}(\bar{x}) \leq \hat{f}(x^*)$. We now use (4) with $x = \bar{x}$ and (5) with $x = x^*$. Adding them up, together with the fact that $\hat{f}(\bar{x}) \leq \hat{f}(x^*)$ we get

$$f(\bar{x}) - 2\varepsilon c(\bar{x}) \leq f(x^*) + 2\varepsilon c(x^*) + 4\varepsilon Z^*.$$

Noting that $c(x) \leq f(x)$ holds for any x and that $Z^* = f(x^*)$, we get that $(1 - 2\varepsilon)f(\bar{x}) \leq (1 + 6\varepsilon)f(x^*)$. Hence, with probability $(1 - 2\delta)$, $f(\bar{x}) \leq (1 + O(\varepsilon))f(x^*)$. ■

Now we are ready to prove properties (P1–P3). We will make repeated use of the Chernoff bound stated in Fact 2. Properties (P2) and (P3) are an easy corrolary of Axiom A3 once we realize that the probability of drawing a high sample from the distribution π is small; and that the fraction of high samples we draw will be small as well with high probability. Let N_h denote the number of high samples in $\omega_1, \ldots, \omega_N$.

Fact 5 *With probability* $1 - \delta$, $N_h/N \leq 2\varepsilon/\lambda$.

Proof: Let X_i be an indicator variable that is equal to 1 if the sample ω_i is high and 0 otherwise. Then $N_h = \sum_{i=1}^{N} X_i$ is a sum of i.i.d. 0-1 variables, and $\mathbf{E}[N_h] = pN$. From Fact 4, $p \leq \frac{\varepsilon}{(1-\varepsilon)\lambda}$.

Using Chernoff bounds,

$$\Pr\left[N_h - Np > \frac{\varepsilon}{\lambda} N \left(2 - \frac{1}{1-\varepsilon}\right)\right] \leq \exp\left(-\frac{\varepsilon^2}{\lambda^2}\frac{(1-2\varepsilon)^2}{(1-\varepsilon)^2} N\right).$$

With $\varepsilon < 1/3$ and N chosen as in Theorem 3, this probability is at most δ. ■

Corollary 6 *Property (P2) holds with probability* $1 - \delta$*, and property (P3) holds with probability 1.*

Proof: By fact 5, we can assume that the number of high samples $N_h < \frac{1}{1-\varepsilon} \leq 2\varepsilon/\lambda$ for $\varepsilon < 1/3$. Then,

$$\hat{f}_h(0) - \hat{f}_h(x) = \frac{1}{N} \sum_{i:\omega_i \text{ high}} q(0, \omega_i) - q(x, \omega_i) \leq \frac{N_h}{N} \lambda c(x).$$

The inequality comes from Axiom A3. Since $N_h \leq 2\varepsilon/\lambda$, the right hand side is at most $2\varepsilon c(x)$, and thus Property (P2) holds with probability $1 - \delta$.

Using Fact 5, by following the same reasoning (replacing sum by expectation) we obtain that Property (P3) holds with probability 1. ■

Now we are ready to prove that property (P1) holds with probability $1 - \delta$. This is done in Lemma 7. The proof of this lemma is the only place where we use the fact that X is finite. In Section 5 we show property (P1) to hold when $X \subseteq \mathbb{R}^n$, under an assumption that the function $c(\cdot)$ is linear and that $q(\cdot, \cdot)$ satisfies certain Lipschitz-type property.

Lemma 7 *With probability* $1 - \delta$ *it holds that for all* $x \in X$,

$$|f_l(x) - \hat{f}_l(x)| \leq \varepsilon Z^*.$$

Proof: First, consider a fixed first stage action $x \in X$. Note that we can view $\hat{f}_l(x)$ as the arithmetic mean of N independent copies $Q_1, \ldots, Q_N$ of the random variable Q distributed as

$$Q = \begin{cases} q(x, \omega) & \text{if } \omega \text{ is low} \\ 0 & \text{if } \omega \text{ is high} \end{cases}$$

Observe that $f_l(x) = \mathbf{E}[Q]$. Let Y_i be the variable Q_i/M and let $Y = \sum_{i=1}^{N} Y_i$. Note that $Y_i \in [0, 1]$ and $\mathbf{E}[Y] = \frac{N}{M} f_l(x)$. We apply the Chernoff bound from Fact 2 to obtain the following.

$$\Pr[|Y - \frac{N}{M} f_l(x)| > \frac{\varepsilon^2}{\lambda} N] \leq 2\exp\left(-\frac{\varepsilon^4}{\lambda^2} N\right).$$

With N as in Theorem 3, this probability is at most $\delta/|X|$. Now, taking the union bound over all $x \in X$, we obtain the desired claim. ■

4 Approximation algorithms and SAA

In many cases of our interest, finding an exact minimizer of the function $\hat{f}$ is computationally hard. However, we may have an algorithm that can find approximate minimizers of functions $\hat{f}$.

First, we explore the performance of the plain SAA method used with an α-approximation algorithm for minimizing the function $\hat{f}$. The following lemma is an adaptation of Theorem 3 to approximation algorithms.

Lemma 8 *Let $\bar{x}$ be a α-approximate minimizer for $\hat{f}$. Then, with probability $(1-2\delta)$,*

$$f(\bar{x})(1-2\varepsilon) \leq (1+6\varepsilon)\alpha f(x^*) + (\alpha-1)(\hat{f}_h(0) - f_h(0)).$$

Proof: Again, with probability $1-2\delta$, we can assume that Properties (P1–P3) hold. Following the proof of Theorem 3, the Lemma follows from Inequalities (4-5) and the fact that $\hat{f}(\bar{x}) \leq \alpha \hat{f}(x^*)$. ■

From the above lemma, we see that $\bar{x}$ is a good approximation to x^* if $\hat{f}_h(0) - f_h(0)$ is small. Since $\hat{f}_h(x^*)$ is an unbiased estimator of $f_h(x^*)$, by Markov's inequality we have that $\hat{f}_h(x^*) \leq (1+1/k)f_h(x^*)$ holds with probability $\frac{1}{k+1}$. Thus, if we want to achieve multiplicative error $(1+\varepsilon)$, we must be content with probability of success only proportional to $1/\varepsilon$. It is not difficult to construct distributions π where the Markov bound is tight.

There are various ways to improve the success probability of the SAA method used in conjunction with an approximation algorithm. We propose two of them in the following two sections. Our first option is to boost the success probability by repetition: in Section 4.1 we show that by repeating the SAA method $\varepsilon^{-1}\log\delta^{-1}$ times independently, we can achieve success probability $1-\delta$. An alternate and perhaps more elegant method is to reject the high cost samples; this does not significantly affect the quality of any solution, while significantly reducing the variance in evaluating the objective function. We discuss this in Section 4.2.

4.1 Approximation Algorithms and Repeating SAA

As we saw in Lemma 8, there is a reasonable probability of success with SAA even with an approximation algorithm for the sampled problem. To boost the probability of success we independently repeat the SAA some number of times and we pick the solution of lowest cost. The precise parameters are formalized in the theorem below.

Theorem 9 *Consider a collection of k functions $\hat{f}^1, \hat{f}^2, \ldots, \hat{f}^k$, such that $k = \Theta(\varepsilon^{-1}\log\delta^{-1})$ and the $\hat{f}^i$ are independent sample average approximations of the function f, using $N = \Theta(\lambda^2\epsilon^{-4}\cdot k \cdot \log|X|\log\delta^{-1})$ samples each. For $i = 1,\ldots,k$, let $\bar{x}^i$ be an α-approximate minimizer of the function $\hat{f}^i$. Let $i =$ $\mathrm{argmin}_j \hat{f}^j(\bar{x}^j)$. Then, with probability $1-3\delta$, $\bar{x}^i$ is an $(1+O(\varepsilon))\alpha$-approximate minimizer of the function $f(\cdot)$.*

Proof: We call the i-th function $\hat{f}^i$ good if it satisfies Properties (P1) and (P2). The number of samples N has been picked so that each $\hat{f}^i$ is good with probability at least $1-2\delta/k$ and hence all samples are good with probability $1-2\delta$.

By Markov inequality, the probability that $\hat{f}^i_h(0) < (1+\varepsilon)f_h(0)$ is at least $1/\varepsilon$. Hence the probability that none of these events happens is at most $(1-\varepsilon)^k < \delta$. Thus, with probability $1-\delta$ we can assume that there is an index j for which $\hat{f}_h^j(0) \le (1+\varepsilon) f_h(0)$. As $f_h(0) \le (1+2\varepsilon)Z^*$ easily follows from Property (P2), with probability $1-\delta$ we have

$$\hat{f}_h^j(0) \le (1+4\varepsilon)Z^*. \tag{6}$$

Let $i = \text{argmin}_\ell \hat{f}^\ell(\bar{x}^\ell)$. Using the fact that $\hat{f}^i(\bar{x}^i) \le \hat{f}^j(\bar{x}^j)$ and that x^i and x^j are both α-approximate minimizers of their respective functions, we get

$$\hat{f}^i(\bar{x}^i) \;\le\; \frac{1}{\alpha}\hat{f}^i(\bar{x}^i) + \frac{\alpha-1}{\alpha}\hat{f}^j(\bar{x}^j) \;\le\; \hat{f}^i(x^*) + (\alpha-1)\hat{f}^j(x^*). \tag{7}$$

Substituting $\bar{x}^i$ and x^* for x in Inequalities (4) and (5) respectively, we obtain

$$f(\bar{x}^i) \le \hat{f}^i(\bar{x}^i) + 2\varepsilon Z^* + 2\varepsilon c(\bar{x}^i) + f_h(0) - \hat{f}_h(0) \tag{8}$$

$$\hat{f}(x^*) \le f(x^*) + 2\varepsilon Z^* + 2\varepsilon c(x^*) + \hat{f}_h(0) - f_h(0). \tag{9}$$

Adding inequalities (7), (8), and (9), we get

$$f(\bar{x}^i) - 2\varepsilon c(\bar{x}^i) \le (\alpha-1)\hat{f}^j(x^*) + f(x^*) + 4\varepsilon Z^* + 2\varepsilon c(x^*) \tag{10}$$

Using Equation 6 with Lemma 8 to bound $\hat{f}^j(x^*)$ finishes the proof. ■

4.2 Approximation Algorithms and Sampling with Rejection

Instead of repeating the SAA method multiple times to get a good approximation algorithm, we can use it only once, but ignore the high cost samples. The following lemma makes this statement precise.

Lemma 10 *Let $g : X \mapsto \mathbb{R}$ be a function satisfying $|f_l(x) + c(x) - g(x)| = O(\varepsilon)Z^*$ for every $x \in X$. Then any α-approximate minimizer $\bar{x}$ of the function $g(\cdot)$ is also an $\alpha(1+O(\varepsilon))$-approximate minimizer of the function $f(\cdot)$.*

Proof: Let $\bar{x}$ be an α-approximate minimizer of g. We have

$$\begin{aligned} f(\bar{x}) &\le g(\bar{x}) + O(\varepsilon Z^*) + f_h(\bar{x}) \\ g(\bar{x}) &\le \alpha(c(x^*) + f_l(x^*) + O(\varepsilon Z^*)) \end{aligned}$$

By Axiom (A2) and Property (P3) we can replace $f_h(\bar{x})$ in the first inequality by $f_h(x^*) + \varepsilon c(x^*)$. Adding up, we obtain

$$f(\bar{x}) \;\le\; \alpha(c(x^*) + f_l(x^*)) + \alpha O(\varepsilon Z^*) + \varepsilon c(x^*) + f_h(x^*) \;\le\; (1+O(\varepsilon))\alpha Z^*.$$

■

According to Lemma 10, a good candidate for the function g would be $g = c(x) + \hat{f}_l(x)$, since by Lemma 7 we know that $f_l(x) - \hat{f}_l(x)$ holds with probability $1-\delta$. However, in order to evaluate $f_l(x)$, we need to know the value Z^* to be able to classify samples as high or low. If Z^* is not known, we can approximate f_l by the function

$$\bar{f}_l(x) = \frac{1}{N} \sum_{i=1}^{N-2\varepsilon N/\lambda} q(x, \omega_i)$$

where we assume that the samples were reordered so that $q(0, \omega_1) \leq q(0, \omega_2) \leq \cdots \leq q(0, \omega_N)$. In other words, we throw out $2\varepsilon N/\lambda$ samples with highest recourse cost.

Fact 11 *With probability* $1 - \delta$*, for all* $x \in X$ *it holds that* $|\bar{f}_l(x) - f_l(x)| \leq 3\varepsilon Z^*$.

Proof: By Fact 5, we can assume that $\bar{f}_l$ does not contain any high samples, and hence $\bar{f}_l(x) \leq \hat{f}_l(x)$. Since there can be at most $2\varepsilon N/\lambda$ samples that contribute to $\hat{f}_l$ but not to $\bar{f}_l$, and all of them are low, we get $\hat{f}_l(x) - \bar{f}_l(x) \leq M \cdot 2\varepsilon/\lambda = 2\varepsilon Z^*$, and hence $|\bar{f}_l(x) - \hat{f}_l(x)| \leq 2\varepsilon Z^*$. Finally, by Lemma 7 we have that $|\hat{f}_l(x) - f_l(x)| \leq \varepsilon Z^*$. ■

Theorem 12 *Let* $\omega_1, \omega_2, \ldots, \omega_N$ *be independent samples from the distribution* π *on* Ω *with* $N = \Theta(\lambda^2 \epsilon^{-4} \cdot \log |X| \log \delta^{-1})$. *Let* $\omega'_1, \omega'_2, \ldots, \omega'_N$ *be a reordering of the samples such that* $q(0, \omega'_1) \leq q(0, \omega'_2) \ldots \leq q(0, \omega'_N)$. *Then any* α*-approximate minimizer* $\bar{x}$ *of the function* $\bar{f}(x) = c(x) + \frac{1}{N} \sum_{i=1}^{N'} q(x, \omega'_i)$ *with* $N' = (1 - 2\varepsilon/\lambda)N$ *is a* $(1 + O(\varepsilon))\alpha$*-approximate minimizer of* $f(\cdot)$.

In many situations, computing the value of $q(x, \omega)$ (or even $q(0, \omega)$) requires us to solve an NP-hard problem. Hence we cannot order the samples as we require in the above theorem. However, if we have an approximation algorithm with ratio β for computing $q(\cdot, \cdot)$, we can use it to order the samples instead. For the above theorem to be applicable with such an approximation algorithm, the number of samples, N, needs to increase by a factor of β^2 and N' needs to be $(1 - 2\varepsilon/(\beta\lambda))N$.

5 From the Discrete to the Continuous

So far we have assumed that X, the set of first-stage decisions, is a finite set. In this section we demonstrate that this assumption is not crucial, and extend Theorems 3, 9 and 12 to the case when $X \subseteq \mathbb{R}^n$, under reasonable Lipschitz type assumptions on the functions $c(\cdot)$ and $q(\cdot, \cdot)$.

Since all our theorems depend only on the validity of the three properties (P1–P3), they continue to hold in all settings where (P1–P3) can be shown to hold. Properties (P2) and (P3) are a simple consequence of Axiom (A3) and hold irrespective of the underlying action space X. Hence, to extend our results to a continuous setting, we only need to check the validity of Property (P1). In the rest of this section, we show (P1) to hold for $X \subseteq \mathbb{R}^n_+$, assuming that the first stage costs is linear, i.e. $c(x) = c^t \cdot x$ for some real vector $c \geq 0$, and that the recourse function satisfies the following property.

Definition 13 We say that the recourse function $q(\cdot, \cdot)$ is (λ, c)-*Lipschitz*, if the following inequality holds for every scenario ω:

$$|q(x, \omega) - q(x', \omega)| \leq \lambda \sum_{i=1}^{n} c_i |x_i - x'_i|.$$

Note that the Lipschitz property implies Axiom (A3) in that any (λ, c)-Lipschitz recourse function q satisfies $q(0,\omega) - q(x,\omega) \le \lambda c(x)$. [1]

We use a standard meshing argument: if two functions $\hat{f}$ and f do not differ by much on a dense enough finite mesh, because of bounded gradient, they must approximately agree in the whole region covered by the mesh. This idea is by no means original; it has been used in the context of stochastic optimization by various authors (among others, [6, 14]). We give the argument for the sake of completeness.

Our mesh is an n-dimensional grid of points with $\varepsilon/(n\alpha\lambda c_i)$ spacing in each dimension $1 \le i \le n$.

Since the i-th coordinate $\bar{x}_i$ of any α-approximate minimizer $\bar{x}$ cannot be larger than $\alpha Z^*/c_i$ (as otherwise the first stage cost would exceed αZ^*), we can assume that the feasible region lies within the bounding box $0 \le x_i \le \alpha Z^*/c_i$ for $1 \le i \le n$. Thus, the set X' of mesh points can be written as

$$X' = \left\{ \left(i_1 \frac{\varepsilon Z^*}{n\lambda c_1}, i_2 \frac{\varepsilon Z^*}{n\lambda c_2}, \ldots, i_n \frac{\varepsilon Z^*}{n\lambda c_n} \right) \;\middle|\; (i_1, i_2, \ldots, i_n) \in \{0, 1, \ldots, \lceil n\alpha\lambda/\varepsilon \rceil\}^n \right\}.$$

We claim the following analog of Lemma 7.

Lemma 14 *If $N \ge \theta(\lambda^2 \frac{1}{\varepsilon^4} n \log(n\lambda/\varepsilon) \log \delta)$, then with probability $1-\delta$, $|\hat{f}_l(x) - f_l(x)| \le 3\varepsilon Z^*$ holds for every $x \in X$.*

Proof: The size of X' is $(1+n\alpha\lambda/\varepsilon)^n$, and hence $\log|X'| = n\log(n\lambda/\varepsilon)$. Hence Lemma 7 guarantees that with probability $1-\delta$, $|\hat{f}_l(x') - f_l(x')| \le \varepsilon Z^*$ holds for every $x' \in X'$.

For a general point $x \in X$, there must be a nearby mesh point $x' \in X'$ such that $\sum_{i=1}^n c_i |x_i - x'_i| \le \varepsilon Z^*/\lambda$. By Lipschitz continuity of q we have that $|f_l(x) - f_l(x')| \le \varepsilon Z^*$ and $|\hat{f}_l(x) - \hat{f}_l(x')| \le \varepsilon Z^*$. By triangle inequality,

$$|\hat{f}_l(x) - f_l(x)| \le |\hat{f}_l(x) - \hat{f}_l(x')| + |\hat{f}_l(x') - f_l(x')| + |f_l(x') - f_l(x)| \le 3\varepsilon Z^*.$$

■

Acknowledgments: We thank Retsef Levi, R. Ravi, David Shmoys, Mohit Singh and Chaitanya Swamy for useful discussions.

References

[1] K. Dhamhere, R. Ravi and M. Singh. On two-stage Stochastic Minimum Spanning Trees. *Proc. of IPCO*, 2005.

[2] A. Flaxman, A. Frieze, M. Krivelevich. On the random 2-stage minimum spanning tree. *Proc. of SODA*, 2005.

[3] A. Gupta, M. Pál, R. Ravi, and A. Sinha. Boosted sampling: Approximation algorithms for stochastic optimization. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, 2004.

[4] A. Gupta, R. Ravi, and A. Sinha. An edge in time saves nine: Lp rounding approximation algorithms. In *Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science*, 2004.

[1] This not necessarily true for non-linear $c(\cdot)$.

[5] N. Immorlica, D. Karger, M. Minkoff, and V. Mirrokni. On the costs and benefits of procrastination: Approximation algorithms for stochastic combinatorial optimization problems. In *Proceedings of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2004.

[6] A. J. Kleywegt, A. Shapiro, and T. Homem-De-Mello. The sample average approximation method for stochastic discrete optimization. *SIAM J. on Optimization*, 12:479-502, 2001.

[7] M. Mahdian. Facility Location and the Analysis of Algorithms through Factor-Revealing Programs. *Ph.D. Thesis*, MIT, June 2004.

[8] R. Motwani and P. Raghavan. **Randomized Algorithms**. Cambridge University Press, 1995.

[9] *Stochastic Programming*. A. Ruszczynski, and A. Shapiro editors. Vol 10 of *Handbook in Operations Research and Management Science*, Elsevier 2003.

[10] A. Shapiro. Montecarlo sampling methods. In A. Ruszczynski, and A. Shapiro editors, *Stochastic Programming*, Vol 10 of *Handbook in Operations Research and Management Science*, Elsevier 2003.

[11] R. Ravi and A. Sinha. Hedging uncertainty: Approximation algorithms for stochastic optimization problems. In *Proceedings of the 10th International Conference on Integer Programming and Combinatorial Optimization (IPCO)*, 2004. *GSIA Working Paper 2003-E68.*

[12] R. Ravi and M. Singh. Personal communication, February 2005.

[13] D. Shmoys and C. Swamy. Stochastic optimization is (almost) as easy as deterministic optimization. *Proc. of FOCS*, 2004.

[14] D. Shmoys and C. Swamy. The Sample Average Approximation Method for 2-stage Stochastic Optimization. Manuscript, November 2004.