The Sample Average Approximation Method for 2-stage Stochastic Optimization

Chaitanya Swamy* David B. Shmoys†

March 23, 2008

1 Introduction

We consider the *Sample Average Approximation* (SAA) method for *2-stage stochastic optimization problems with recourse* and prove a polynomial time convergence theorem for the SAA method. In the 2-stage recourse model, where one makes decisions in two steps. First, given only distributional information about (some of) the data, one commits on initial (first-stage) actions, and then once the actual data is realized, according to the distribution, further *recourse actions* can be taken, so that one can augment the earlier solution to satisfy the revealed requirements, if necessary. Typically the recourse actions entail making decisions in rapid reaction to the observed scenario, that is, at the "last minute," and are therefore costlier than decisions made ahead of time. The goal is to choose the first stage elements so as to minimize the sum of the cost incurred in the first stage and the expected cost incurred in the second stage, where the expectation is taken over all problem instances and these instances are distributed according to the given probability distribution.

More formally, given a probability distribution on *scenarios* A and a vector x describing the first stage decisions, the cost incurred is given by $h(x) = c(x) + \mathrm{E}_A\big[f_A(x, r_A)\big]$ where vector r_A denotes the second stage decisions that are taken when scenario A materializes, $c(x)$ is the cost incurred in the first stage, and $f_A(x, r_A)$ is the cost of augmenting x to obtain the solution (x, r_A) for scenario A. We want to choose x that minimizes the total cost $h(x)$. Consider a discrete distribution and let p_A denote the probability of scenario A. Then the objective function is $h(x) = c(x) + \sum_{A\in\mathcal{A}} p_A f_A(x, r_A)$, where $\mathcal{A}$ denotes the set of all scenarios.

2 The Sample Average Approximation method

Assume that we have a black box that one can use to draw independent samples from the distribution on scenarios. A natural approach to computing near-optimal solutions for these problems is the sample average approximation approach: sample some $\mathcal{N}$ times from the distribution on scenarios, and estimate the probability p_A of scenario A by $\hat{p}_A = \mathcal{N}_A/\mathcal{N}$ where $\mathcal{N}_A$ denotes the number of times scenario A occurs. Now we consider the *sample average function* $\hat{h}(x) = c(x) + \sum_{A\in\mathcal{A}} \hat{p}_A f_A(x, r_A)$ and find $\hat{x}$ that minimizes $\hat{h}(.)$. Since the distribution $\{\hat{p}_A\}_{A\in\mathcal{A}}$ assigns a non-zero probability to at most the $\mathcal{N}$ sampled scenarios, if $\mathcal{N}$ is small, e.g., polynomial size then the sample average problem of minimizing $\hat{h}(.)$ is an easier task than minimizing $h(.)$. The issue here is to bound the sample size $\mathcal{N}$ required to guarantee that *every* (near-) optimal solution to the sample-average problem is a near-optimal solution to the true problem

*cswamy@ist.caltech.edu. Center for the Mathematics of Information, Caltech, Pasadena, CA 91125.

†shmoys@cs.cornell.edu. Dept. of Computer Science, Cornell University, Ithaca, NY 14853. Research supported partially by NSF grant CCF-0430682.

with high probability. Intuitively we need $\mathcal{N}$ large enough so that the function $\hat{h}(.)$, is in some sense, a close approximation to $h(.)$.

We show that for a large class of 2-stage stochastic linear programs, namely the class considered by Shmoys and Swamy [5], we can bound $\mathcal{N}$ by a polynomial in the input size, the inverse of the desired accuracy, and the maximum *ratio* λ between the second-stage and first-stage costs. This thus gives more efficient algorithms for solving this class of problems than the algorithm in [5], which is encumbered by the machinery of the ellipsoid method. In comparison, Kleywegt, Shapiro, and Homem-De-Mello [2] (see also [4]) gave a bound on the sample size that depends on the variance of a certain quantity that need not depend polynomially on the input size or λ. Very recently Nemirovskii and Shapiro (personal communication) independently showed that for the stochastic set-cover problem with second-stage costs that are non-scenario dependent, the bound of Kleywegt et al. is a polynomial bound, *provided that one preprocesses the input to eliminate certain first-stage decisions*, and then applies the SAA method to the reduced problem. In terms of lower bounds on the sample size, it was shown in [5] that the dependence on λ cannot be avoided.

Our proof technique is different from that of [2], and exploits the notion of subgradients and approximate subgradients that was used in [5]. The ellipsoid-based convex minimization algorithm given by Shmoys and Swamy [5] shows that, under an appropriately defined notion of approximate subgradient, one can minimize a convex function in polynomial time *using only approximate subgradient information* about the function. For a given class of convex functions, if one can compute these approximate subgradients efficiently by some uniform procedure, then one might be able to interpret these vectors as *exact* subgradients of another "nice" function, that is, in some sense, "fit" a nice function to these vectors, and thereby argue that minimizing this nice function yields a near-optimal solution to the original minimization problem. For the class of 2-stage problems considered in [5], one can compute approximate subgradients by simply sampling and averaging, and therefore it turns out that the "nice" function is the sample average function $\hat{h}(.)$.

We believe that our proof is simpler and may be of independent interest. The proof does not rely on anything specific to discrete probability distributions and therefore extends to the case of continuous distributions. In essence, our proof suggests that a performance guarantee statement about any algorithm that uses approximate subgradients computed by sampling in some uniform way, can be translated to a statement about the performance guarantee of the sample average method. We believe that our approach can be applied to prove convergence results for the sample average method for other stochastic models as well. In particular, Swamy and Shmoys [6] recently gave an algorithm for solving a class of multi-stage stochastic linear programs to near-optimality based on a (uniform) sampling-based procedure for computing approximate subgradients; this suggests that one might be able to use our approach to prove an analogous polynomial-time convergence theorem for the sample average method for this class of multi-stage programs.

3 Analysis of the SAA method

We consider the following generic 2-stage stochastic optimization problem considered in [5].

$$\min_{x\in\mathcal{P}} \quad h(x) = w^{\mathrm{I}}\cdot x + \sum_{A\in\mathcal{A}} p_A f_A(x), \tag{P}$$

$$\begin{aligned} \text{where}\quad f_A(x) \;=\; \min \quad & w^A\cdot r_A + q^A\cdot s_A \\ \text{s.t.}\quad & B^A s_A \ \geq h^A && (1)\\ & D^A s_A + T^A r_A \ \geq j^A - T^A x && (2)\\ & r_A, s_A \ \geq 0,\ r_A\in\mathbb{R}^m,\ s_A\in\mathbb{R}^n. \end{aligned}$$

Here (a) $T^A \geq \mathbf{0}$ for every scenario A, and (b) for every $x\in\mathcal{P}$, $\sum_{A\in\mathcal{A}} p_A f_A(x)\geq 0$ and that the primal and dual problems corresponding to $f_A(x)$ are feasible for every scenario A. A sufficient condition for (b) is to insist that $0 \leq f_A(x) < +\infty$ at every point $x \in \mathcal{P}$ and scenario $A \in \mathcal{A}$. $\mathcal{P} \subseteq \mathbb{R}^m_{\geq 0}$ denotes the bounded feasible region of first-stage decisions and $\mathbf{0} \in \mathcal{P}$. Define $\lambda = \max\bigl(1, \max_{A \in \mathcal{A}, S} \frac{w^A_S}{w^{\mathrm{I}}_S}\bigr)$; we assume that λ is known. Let OPT be the optimum value. Shmoys and Swamy adapted the ellipsoid method to approximately solve (P), given only a black box to draw independent samples from the scenario distribution.

The corresponding sample average problem is

$$\min_{x \in \mathcal{P}} \quad \hat{h}(x) = w^{\mathrm{I}} \cdot x + \sum_{A \in \mathcal{A}} \hat{p}_A f_A(x). \tag{SAA-P}$$

where $\hat{p}_A = \mathcal{N}_A / \mathcal{N}$ is the estimated probability of scenario A, $\mathcal{N}$ is the total number of samples, and $\mathcal{N}_A$ denotes the number of times that scenario A occurs in those samples. We assume that the polytope $\mathcal{P}$ is contained in the ball $B(\mathbf{0}, R) = \{x : \|x\| \leq R\}$ and contains a ball of radius V, such that $\ln R$ and $\ln\bigl(\frac{1}{V}\bigr)$ are polynomially bounded; Lemmas 6.2.4–6.2.6 in [1] show that one can always obtain such an R and V. We may assume that $V \leq 1$ without loss of generality. We show that for any $\epsilon, \gamma > 0$, we can bound $\mathcal{N}$ by $\text{poly}\bigl(\text{input size}, \lambda, \frac{1}{\gamma}, \ln\bigl(\frac{1}{\epsilon}\bigr)\bigr)$, and have that with high probability, $h(\hat{x}) \leq (1+\gamma) \cdot OPT + 8\epsilon$ where $\hat{x}$ is any optimal solution to (SAA-P). Our proof uses the notion of a subgradient and an approximate subgradient as defined below.

Definition 3.1 *Let $g : \mathbb{R}^m \mapsto \mathbb{R}$ be a function. We say that d is a subgradient of g at the point u if the inequality $g(v) - g(u) \geq d \cdot (v - u)$ holds for every $v \in \mathbb{R}^m$.*

Definition 3.2 *We say that $\hat{d}$ is a $(\omega, \mathcal{D})$-subgradient of a function $g : \mathbb{R}^m \mapsto \mathbb{R}$ at the point $u \in \mathcal{D}$ if for every $v \in \mathcal{D}$, we have $g(v) - g(u) \geq \hat{d} \cdot (v - u) - \omega g(v) - \omega g(u)$.*

The above definition of an $(\omega, \mathcal{D})$-subgradient is slightly different and weaker than the notion of an $(\omega, \mathcal{D})$-subgradient as defined in [5] where one requires $g(v) - g(u) \geq \hat{d} \cdot (v - u) - \omega g(u)$, since any vector that is an $(\omega, \mathcal{D})$ subgradient according to the definition in [5] is clearly also an $(\omega, \mathcal{D})$-subgradient according to Definition 3.2. This distinction is however inconsequential; note that one could implement the algorithm in [5] using the notion of an approximate subgradient given by Definition 3.2. In the sequel, we will only use $(\omega, \mathcal{P})$-subgradients, which we abbreviate and denote as ω-subgradients from now on. It is straightforward to show that both $h(.)$ and $\hat{h}(.)$ are convex functions, and hence for both $h(.)$ and $\hat{h}(.)$, a well-defined subgradient exists at any given point.

The proof is based on two main ideas. As mentioned earlier, in the convex minimization algorithm of [5], the only information needed about the convex function to be minimized, is its subgradient or ω-subgradient at any given point. The algorithm generates a sequence of ellipsoids of successively smaller volume starting with a ball that encloses the feasible region, using at each step a cut passing through the center of the current ellipsoid to chop off a half-ellipsoid and make progress; this cut is derived either through an infeasible inequality (which one can assume is determined uniquely using an arbitrary tie-breaking rule), or if the current center is feasible, then by a subgradient or an ω-subgradient cut. Thus, if we have two functions $g, \hat{g} : \mathbb{R}^m \mapsto \mathbb{R}$ that agree in terms of their (approximate) subgradients on $\mathcal{P}$, specifically suppose at every $x \in \mathcal{P}$ there is a vector d_x that is both a subgradient of $\hat{g}(.)$ and an ω-subgradient of $g(.)$, then using d_x to generate the cut at x would make the algorithm run *identically* on both the problems $\min_{x \in \mathcal{P}} g(x)$ and $\min_{x \in \mathcal{P}} \hat{g}(x)$. So we would obtain a point that is simultaneously near-optimal for both functions g and $\hat{g}$. This only argues that there is one specific point that is near-optimal for both g and $\hat{g}$. But in fact, we will show that if g and $\hat{g}$ agree in terms of their subgradients even on a sufficiently dense finite set $G \subseteq \mathcal{P}$ (property (A) makes this precise), then *every* optimal solution to $\min_{x \in \mathcal{P}} \hat{g}(x)$ is a near-optimal solution to $\min_{x \in \mathcal{P}} g(x)$.

Next, we show that with a polynomially bounded sample size $\mathcal{N}$, functions $h(.)$ and $\hat{h}(.)$ satisfy this "closeness-in-subgradient" property with high probability, so every optimal solution to the sample average problem is a near-optimal solution to (P). Our method of using subgradients on a dense set G to identify closeness between $h(.)$ and $\hat{h}(.)$ and prove the polynomial time convergence of the SAA method, is different from that of Kleywegt et al. [2] who show that if x^* is an optimal solution to (P), then the quantities $h(x) - h(x^*)$ and $\hat{h}(x) - \hat{h}(x^*)$ should be close to each other at every grid point x, and use this to prove the convergence result.

Let the functions h and $\hat{h}$ have Lipschitz constant (at most) K, $\epsilon, \gamma > 0$ be two input parameters with $\gamma \leq 1$ without loss of generality. Let $N = \log\left(\frac{2KR}{\epsilon}\right)$ and $\omega = \frac{\gamma}{8N}$. We first construct a suitably dense "gridding" of the polytope $\mathcal{P}$. Let $G' = \left\{x \in \mathcal{P} : x_i = n_i \cdot \left(\frac{\epsilon V}{8KNR\sqrt{m}}\right),\ n_i \in \mathbb{Z} \text{ for all } i = 1, \ldots, m\right\}$ and $G = G' \cup \left\{x + t(y - x), y + t(x - y) : x, y \in G',\ t = 2^{-i},\ i = 1, \ldots, N\right\}$. The following bound on the volume of a cell of G' will be useful. Let vol_m denote the volume of the unit ball in m dimensions. Let $s = \frac{\epsilon V}{8KNR\sqrt{m}}$ be the size of a cell of G'.

Claim 3.3 *The volume of a grid cell of G' is at least $\left(\frac{s}{2}\right)^m \mathsf{vol}_m$.*

Proof : Each grid cell of G' contains a ball of radius $\frac{s}{2}$ centered at the center of the cell, so its volume is at least $\left(\frac{s}{2}\right)^m \mathsf{vol}_m$. ■

Lemma 3.4 *For every point $x \in \mathcal{P}$ there exists $x' \in G'$ such that $\|x - x'\| \leq \frac{\epsilon}{KN}$.*

Proof : Let $r = \frac{\epsilon}{4KNR}$. Consider the affine transformation $T(y) = rI_m(y - x) + x = ry + (1 - r)x$ where I_m is the $m \times m$ identity matrix, and let $W = T(\mathcal{P})$, so W is a shrunken version of $\mathcal{P}$ "centered" around x. Observe that $W \subseteq \mathcal{P}$ because $\mathcal{P}$ is convex, and any point $T(y) \in W$ is a convex combination of $y \in \mathcal{P}$ and $x \in \mathcal{P}$. Let $\hat{y} \in \mathcal{P}$ be the center of the ball of radius V contained in $\mathcal{P}$ and let $\hat{x} = T(\hat{y}) \in W$. Note that W contains the ball $B(\hat{x}, rV)$, and that $\|\hat{x} - x\| = r\|\hat{y} - x\| \leq \frac{\epsilon}{2KN}$ since $x, \hat{y} \in B(\mathbf{0}, R)$. We will show that there exists $x' \in G'$ such that $\|\hat{x} - x'\| \leq \frac{\epsilon}{2KN}$, which will complete the proof. Define n_i such that $\hat{x}_i \in [z_i + n_i s, z_i + (n_i + 1)s]$ for all $i = 1, \ldots, m$. Consider any point x' where $x'_i \in \{z_i + n_i s, z_i + (n_i + 1)s\}$ for all i. Then $\|\hat{x} - x'\| \leq s\sqrt{m} < rV < \frac{\epsilon}{2KN}$, so $x' \in B(\hat{x}, rV) \subseteq \mathcal{P}$, which implies that $x' \in G'$. ■

We say that functions $g : \mathbb{R}^m \mapsto \mathbb{R}$ and $\hat{g} : \mathbb{R}^m \mapsto \mathbb{R}$ satisfy property (A) if

$$\begin{aligned} &\text{at every point } x \in G, \text{ there exists } d_x \in \mathbb{R}^m \text{ such that} \\ &\quad 1)\quad d_x \text{ is a subgradient of } \hat{g}(.) \text{ at } x, \text{ and} \\ &\quad 2)\quad d_x \text{ is an } \omega\text{-subgradient of } g(.) \text{ at } x. \end{aligned} \tag{A}$$

Lemma 3.5 *Let $g : \mathbb{R}^m \mapsto \mathbb{R}$ and $\hat{g} : \mathbb{R}^m \mapsto \mathbb{R}$ be any two convex functions with Lipschitz constant (at most) K that satisfy property* (A). *Then, $g(\hat{x}) \leq (1 + \gamma)g(x^*) + 8\epsilon$ where x^* and $\hat{x}$ are points in $\mathcal{P}$ that respectively minimize functions $g(.)$ and $\hat{g}(.)$.*

Proof : For ease of understanding, consider first the case when $\hat{x} \in G'$. We will argue that there is a point x near $\hat{x}$ such that $g(x)$ is close to $g(x^*)$, and from this it will follow that $g(\hat{x})$ is close to $g(x^*)$. Let $\tilde{x}$ be the point in G' closest to x^*, so $\|\tilde{x} - x^*\| \leq \frac{\epsilon}{KN}$ (Lemma 3.4) and $g(\tilde{x}) \leq g(x^*) + \epsilon$. Let $y = \hat{x}\left(1 - \frac{1}{2^N}\right) + \left(\frac{1}{2^N}\right)\tilde{x} \in G$ and consider the vector d_y given by property (A). It must be that $d_y \cdot (\hat{x} - y) \leq 0$, otherwise we would have $\hat{g}(\hat{x}) > \hat{g}(y)$ contradicting the optimality of $\hat{x}$. So, by the definition of an ω-subgradient, we have $g(y) \leq \frac{1+\omega}{1-\omega} \cdot g(\tilde{x}) \leq (1 + 4\omega)g(\tilde{x}) \leq (1 + \gamma)g(x^*) + 2\epsilon$ since $\omega = \frac{\gamma}{8N} \leq \frac{1}{4}$. Also $\|\hat{x} - y\| = \frac{\|\hat{x} - \tilde{x}\|}{2^N} \leq \frac{\epsilon}{K}$ since $\|\hat{x} - \tilde{x}\| \leq 2R$. So, $g(\hat{x}) \leq g(y) + \epsilon \leq (1 + \gamma)g(x^*) + 3\epsilon$.

Now consider the case when $\hat{x} \notin G'$. Let $\bar{x}$ be the point in G' closest to $\hat{x}$, so $\|\bar{x} - \hat{x}\| \leq \frac{\epsilon}{KN}$ and $\hat{g}(\bar{x}) \leq \hat{g}(\hat{x}) + \frac{\epsilon}{N}$. For any $y \in G$, if we consider d_y given by property (A), then whereas $d_y \cdot (\hat{x} - y) \leq 0$ it need not be that $d_y \cdot (\bar{x} - y) \leq 0$, so we have to argue a little differently. Note that however $d_y \cdot (\bar{x} - y) \leq \frac{\epsilon}{N}$, otherwise we would have $\hat{g}(\hat{x}) > \hat{g}(y)$. Let $y_0 = \tilde{x}$, and $y_i = (\bar{x} + y_{i-1})/2$ for $i = 1, \ldots, N$. Since each $y_i \in G$, we have $d_{y_i} \cdot (y_{i-1} - y_i) = -d_{y_i} \cdot (\bar{x} - y_i) \geq -\frac{\epsilon}{N}$, and because d_{y_i} is an ω-subgradient of $g(.)$ at y_i, $g(y_i) \leq (1+4\omega)g(y_{i-1}) + \frac{\epsilon}{N(1-\omega)}$. This implies that $g(y_N) \leq (1+4\omega)^N g(\tilde{x}) + \frac{\epsilon(1+4\omega)^N}{(1-\omega)} \leq (1+\gamma)g(x^*) + 2\epsilon + 4\epsilon$. So $g(\hat{x}) \leq g(y_N) + 2\epsilon \leq (1+\gamma)g(x^*) + 8\epsilon$. ■

Now we show that $h(.)$ and $\hat{h}(.)$ satisfy property (A). First, as in [5], we show that to get an ω-subgradient, it suffices to approximate each component of a subgradient to within a certain additive error. Then, we argue that with a large enough sample size, at any point $x \in \mathcal{P}$, there is a subgradient of $\hat{h}(.)$ that is component-wise close to a subgradient of $h(.)$ with high probability. Finally, we bound the size of G, and show that we can set $\mathcal{N}$ large enough, keeping it polynomially bounded, so that property (A) holds with high probability.

Lemma 3.6 *Let d be a subgradient of $h(.)$ at the point $x \in \mathcal{P}$, and suppose that $\hat{d}$ is a vector such that $\hat{d}_S \in [d_S - \omega w_S^{\mathrm{I}}, d_S + \omega w_S^{\mathrm{I}}]$ for all S. Then $\hat{d}$ is an ω-subgradient of $h(.)$ at x.*

Proof : The proof is almost exactly as in [5]. Let $y \in \mathcal{P}$. Then, $h(y) - h(x) \geq d \cdot (y - x) = \hat{d} \cdot (y - x) + (d - \hat{d}) \cdot (y - x)$. Since $x_S, y_S \geq 0$ for all S, the latter term is at least

$$\sum_{S: d_S \leq \hat{d}_S} (d_S - \hat{d}_S) y_S + \sum_{S: \hat{d}_S \leq d_S} (\hat{d}_S - d_S) x_S \geq \sum_S (-\omega w_S^{\mathrm{I}} y_S - \omega w_S^{\mathrm{I}} x_S) \geq -\omega h(y) - \omega h(x).$$

■

Recall that $\lambda = \max\bigl(1, \max_{A \in \mathcal{A}, S} \frac{w_S^A}{w_S^{\mathrm{I}}}\bigr)$. Consider any point $x \in \mathcal{P}$, and let (u_A^*, z_A^*) be an optimal solution to the dual of $f_A(x)$. where z_A^* is the dual multiplier corresponding to inequalities (2). It is shown in [5] that the vector $d = w^{\mathrm{I}} - \sum_A p_A (T^A)^{\mathrm{T}} z_A^*$ is a subgradient of $h(.)$ at x and $\|d\| \leq \lambda \|w^{\mathrm{I}}\|$. So the Lipschitz constant of $h(.)$ is at most $K = \lambda \|w^{\mathrm{I}}\|$. The sample average function $\hat{h}(.)$ is of the same form as $h(.)$, only with a different probability distribution, so

$$\hat{d} = w^{\mathrm{I}} - \sum_A \hat{p}_A (T^A)^{\mathrm{T}} z_A^* \tag{3}$$

is a subgradient of $\hat{h}(.)$ at x and the Lipschitz constant of $\hat{h}(.)$ is also at most K. Observe that $\hat{d}$ is just $w^{\mathrm{I}} - (T^A)^{\mathrm{T}} z_A^*$ averaged over the scenarios sampled to construct $\hat{h}(.)$ since $\hat{p}_A = \mathcal{N}_A/\mathcal{N}$, and and $\mathrm{E}\bigl[\hat{d}\bigr] = d$ where the expectation is over these samples. Also for any scenario A, component S of $w^{\mathrm{I}} - (T^A)^{\mathrm{T}} z_A^*$ lies in $[-\lambda w_S^{\mathrm{I}}, w_S^{\mathrm{I}}]$ since the dual has the constraint $(T^A)^{\mathrm{T}} z_A \leq w_S^{\mathrm{II}}$. The following lemma shows that $\hat{d}$ will be component-wise close to d with high probability, and is therefore an approximate subgradient of $g(.)$ at x by Lemma 3.6.

Lemma 3.7 *Let $X_i, i = 1, \ldots, \mathcal{N} = \frac{4(1+\alpha)^2}{c^2} \ln\bigl(\frac{2}{\delta}\bigr)$ be iid random variables where each $X_i \in [-a, b]$, $a, b > 0$, $\alpha = \max(1, a/b)$, and $c \in [0, 1]$. Let $X = \bigl(\sum_i X_i\bigr)/\mathcal{N}$ and $\mu = \mathrm{E}\bigl[X\bigr] = \mathrm{E}\bigl[X_i\bigr]$. Then $\Pr\bigl[X \in [\mu - cb, \mu + cb]\bigr] \geq 1 - \delta$.*

Proof : Let $Y_i = X_i + a \in [0, a+b]$ and $Y = \sum_i Y_i$. Let $\mu' = \mathrm{E}\bigl[Y_i\bigr] = \mu + a$. We have $\Pr[X > \mu + cb] = \Pr[Y > \mathrm{E}\bigl[Y\bigr](1 + cb/\mu')]$, and $\Pr[X < \mu - cb] = \Pr[Y < \mathrm{E}\bigl[Y\bigr](1 - cb/\mu')]$. Let $\nu = cb/\mu'$. Note that $\mu' \leq a + b$. Since the variables Y_i are independent we can use Chernoff bounds here. The latter probability, $\Pr[Y < \mathrm{E}\bigl[Y\bigr](1-\nu)]$, is at most $e^{-\frac{\nu^2 \mathcal{N} \mu'}{2(a+b)}} = e^{-\frac{(cb)^2 \mathcal{N}}{2\mu'(a+b)}} \leq \frac{\delta}{2}$. To bound $\Pr[Y > \mathrm{E}\bigl[Y\bigr](1+\nu)]$ we consider two cases. If $\nu > 2e - 1$, then this quantity is at most $2^{-\frac{(1+\nu)\mathcal{N}\mu'}{a+b}}$ (see, e.g., [3], Chapter 4), which is bounded by $2^{-\frac{\nu\mathcal{N}\mu'}{a+b}} \leq \frac{\delta}{2}$. If $\nu \leq 2e-1$, then the probability is at most $e^{-\frac{\nu^2\mathcal{N}\mu'}{4(a+b)}} = e^{-\frac{(cb)^2\mathcal{N}}{4\mu'(a+b)}} \leq \frac{\delta}{2}$. So using the union bound, $\Pr\big[X \notin [\mu - cb, \mu + cb]\big] \leq \delta$. ■

Theorem 3.8 *With probability at least* $1-\delta$*, any optimal solution* $\hat{x}$ *to the sample average problem constructed with at most* $\mathsf{poly}\big($*input size*$, \frac{1}{\gamma}, \ln(\frac{1}{\epsilon}), \ln(\frac{1}{\delta})\big)$ *samples, satisfies,* $h(\hat{x}) \leq (1+\gamma)\cdot OPT + 8\epsilon$.

Proof : We will satisfy property (A) with probability $1-\delta$. Let $n = |G|$. Recall that $N = \log\big(\frac{2KR}{\epsilon}\big)$ and $\omega = \frac{\gamma}{8N}$. Note that $\log\big(\frac{KR}{V}\big)$ is polynomially bounded in the input size. Using Lemmas 3.7 and 3.6 by taking $\mathcal{N} = \frac{4(1+\lambda)^2}{3\omega^2}\ln\big(\frac{2mn}{\delta}\big)$ samples to construct $\hat{h}(.)$, at any point x, the subgradient $\hat{d}_x$ of $\hat{h}(.)$ given by (3) is an ω-subgradient of $h(.)$ with probability at least $1-\delta/n$. So with probability at least $1-\delta$, $\hat{d}_x$ is an ω-subgradient of $h(.)$ at *every* point $x \in G$.

To bound n, note that $n \leq 2N\binom{|G'|}{2} \leq N|G'|^2$. By Claim 3.3, each grid cell of G' has volume at least $(s/2)^m\mathsf{vol}_m$, where $s = \frac{\epsilon V}{8KNR\sqrt{m}}$. The grid cells are pairwise disjoint (volume-wise), and have total volume at most $\mathsf{vol}\big(B(\mathbf{0},R)\big) \leq R^m\mathsf{vol}_m$ since $\mathcal{P} \subseteq B(\mathbf{0},R)$. So $|G'| \leq \big(\frac{16KNR^2\sqrt{m}}{\epsilon V}\big)^m$. Plugging this above, we get that $\mathcal{N} = O\big(\lambda^2N^2\ln(\frac{2mN|G'|}{\delta})/\gamma^2\big) = O\big(m\lambda^2\log^2(\frac{2KR}{\epsilon})\ln(\frac{KRm}{\epsilon V\delta})/\gamma^2\big)$ which is $\mathsf{poly}\big(\text{input size}, \frac{1}{\gamma}, \ln(\frac{1}{\epsilon}), \ln(\frac{1}{\delta})\big)$. ■

As shown in [5], under the slight assumption that for every $x \in \mathcal{P}$ and scenario A, either $f_A(x)$ is minimized at $x = \mathbf{0}$, or the total cost $w^{\mathrm{I}}\cdot x + f_A(x) \geq 1$, by sampling $\lambda\ln\big(\frac{1}{\delta}\big)$ times initially, one can detect with probability at least $1-\delta$ ($\delta \leq \frac{1}{2}$), that either $x = \mathbf{0}$ is an optimal solution to (P), or that $OPT \geq \varrho/\lambda$ where $\varrho = \frac{\delta}{\ln(1/\delta)}$. So to get a multiplicative $(1+\kappa)$-guarantee, if we detect that OPT is large after this initial sampling, then setting $\gamma = \kappa/2$ and $\epsilon = \kappa\varrho/(16\lambda)$ above, we get that $\mathsf{poly}\big(\text{input size}, \lambda, \frac{1}{\kappa}, \ln(\frac{1}{\delta})\big)$ samples suffice to ensure that any optimal solution to (SAA-P) is a $(1+\kappa)$-optimal solution to (P) with probability at least $1-2\delta$.

Extension to continuous distributions. Notice that nothing in the preceding analysis relied on the fact that we have a discrete probability distribution. In particular Lemma 3.7 and Corollary 3.6 also hold for continuous distributions. This shows that the SAA method with a polynomial number of samples, returns a near-optimal solution to the class of programs (P) where the second stage scenario is specified by a parameter ξ that is continuously distributed with probability density function $p(\xi)$, the objective function is $h(x) = w^{\mathrm{I}}\cdot x + \mathrm{E}_\xi\big[f(x,\xi)\big]$, where $\mathrm{E}_\xi\big[f(x,\xi)\big] = \int p(\xi)f(x,\xi)\,d\xi$ and $f(x,\xi)$ is the cost of scenario ξ determined by the minimization problem in (P) with parameters $w(\xi)$, $q(\xi)$, $h(\xi)$, $j(\xi)$, $B(\xi)$, $D(\xi)$ and $T(\xi)$. Here $\lambda = \max\big(1, \sup_{\xi,S}\frac{w(\xi)_S}{w^{\mathrm{I}}_S}\big)$. As before we have that at every feasible point x and scenario ξ, (a) $T(\xi) \geq \mathbf{0}$, (b) $\int p(\xi)f(x,\xi)\,d\xi \geq 0$, and that the primal and dual problems corresponding to $f(x,\xi)$ are feasible.

References

[1] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer-Verlag, New York, 1988.

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

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

[4] A. Shapiro. Monte Carlo sampling methods. In A. Ruszczynski and A. Shapiro, editors, *Stochastic Programming*, volume 10 of *Handbooks in Operations Research and Management Science*, North-Holland, Amsterdam, 2003.

[5] D. B. Shmoys and C. Swamy. Stochastic optimization is (almost) as easy as deterministic optimization. In *Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science*, pages 228–237, 2004.

[6] C. Swamy and D. B. Shmoys. Approximation algorithms for multi-stage stochastic optimization. *Submitted*.