Sampling-based Approximation Algorithms for Multi-stage Stochastic Optimization*

Chaitanya Swamy† cswamy@ist.caltech.edu

David B. Shmoys‡ shmoys@cs.cornell.edu

Abstract

Stochastic optimization problems provide a means to model uncertainty in the input data where the uncertainty is modeled by a probability distribution over the possible realizations of the actual data. We consider a broad class of these problems in which the realized input is revealed through a series of stages, and hence are called multi-stage stochastic programming problems. Our main result is to give the first fully polynomial approximation scheme for a broad class of multi-stage stochastic linear programming problems with any constant number of stages. The algorithm analyzed, known as the sample average approximation (SAA) method, is quite simple, and is the one most commonly used in practice. The algorithm accesses the input by means of a "black box" that can generate, given a series of outcomes for the initial stages, a sample of the input according to the conditional probability distribution (given those outcomes). We use this to obtain the first polynomial-time approximation algorithms for a variety of k-stage generalizations of basic combinatorial optimization problems.

1. Introduction

Stochastic optimization problems provide a means to model uncertainty in the input data where the uncertainty is modeled by a probability distribution over the possible realizations of the actual data. We shall consider a broad class of these problems in which the realized input is revealed through a series of stages, and hence are called *multi-stage stochastic programming problems*. Multi-stage stochastic linear programming is an area that has received a great deal of attention within the Operations Research community, both in terms of the asymptotic convergence results, as well as computational work in a wide variety of application domains. For example, a classic example of such a model seeks to minimize the expected cost of operating a water reservoir where one can decide, in each time period, the amount of irrigation water to be sold while maintaining the level of the reservoir within a specified range (where penalties are incurred for violating this constraint). The source of uncertainty is, of course, the variability in rainfall, and there is a simulation model that provides a means to sample from the distribution of inputs (of rainfall amounts per time period within the planning horizon) [2]. Observe that it is important to model this as a multi-stage process, rather than as a 2-stage one, since it allows us to capture essential conditional information, such as given a drought over the previous period, the next period is more likely to continue these conditions. Furthermore, within multi-stage stochastic linear programming, most work has focused on applications in which there are a small number of stages, including forest planning models electricity investment planning, bond investment planning, and currency options selection, as discussed in the recent survey of Ariyawansa and Felt [1].

Our main result is to give the first fully polynomial randomized approximation scheme (FPRAS) for a broad class of multi-stage stochastic linear programming problems with any constant number of stages. Although our results are much more general, we shall focus on a canonical example of the class of problems, a 3-stage stochastic variant of the fractional set covering problem. We are given a family of sets over a ground set and a probability distribution over the subsets that specifies a target set of ground elements that must be covered. We can view the three stages as specified by a scenario tree with 3 levels of nodes: the root, internal nodes, and leaves; the root corresponds to the initial state, each leaf is labeled with a target subset of elements that must be covered, and for each node in the tree there is a conditional distribution of the target sets at leaves within this subtree (where we condition on the fact that we have reached that node). One can buy (fractionally) sets at any node paying a cost that depends both on the set and the node at which it is bought. We want to be able to compute, given a node in the tree, the desired action, so as to minimize the expected total cost of fractionally covering the realized target set. This problem can be modeled as an exponentially

* A full version is available at ist.caltech.edu/~cswamy/papers/multistage.ps

† Center for Mathematics of Information, Caltech, Pasadena, CA 91125.

‡ Dept. of Computer Science, Cornell University, Ithaca, NY 14853.
Research supported partially by NSF grants CCF-0430682, DMI-0500263.

large linear program (LP) in which there is, for each set S and each node in the tree, a variable that indicates the fraction of S that is bought at that node. The constraints say that for each leaf, for each ground element e in its corresponding target set, the total fraction bought of sets S that contain e along this root-leaf path must be at least 1. If we view the probability of reaching a node as specified, it is straightforward to express the expected total cost as a linear function of these decision variables. As a corollary of our FPRAS, we also give the first approximation algorithms for the analogous class of multi-stage stochastic integer programs (IPs), such as the integer version of this set covering problem.

For a rich class of k-stage stochastic linear programming problems, where k is assumed to be constant and not part of the input, we show that, for any $\epsilon > 0$, we can compute, with high probability, a solution with expected cost guaranteed, for any probability distribution over inputs, to be within a $(1 + \epsilon)$ factor of the optimal expected cost, in time bounded by a polynomial in the input size, $\frac{1}{\epsilon}$, and a parameter λ that is an upper bound on the ratio between the cost of the same action (e.g., buying the set S) over successive stages. The algorithm accesses the input by means of a "black-box" (simulation) procedure that can generate, for any node in the scenario tree, a sample of the input according to the conditional distribution for this node. This is an extremely general model of the distribution, since it allows all types of correlated effects within different parts of the input. We improve upon our earlier work [13], which handles the very special case in which $k = 2$, not only by being able to handle *any fixed number of stages*, but whereas the earlier algorithm is based on the ellipsoid method, we can now show that the algorithm most commonly used in practice, the *sample average approximation* method (SAA), also yields the claimed approximation scheme.

The algorithm of Shmoys & Swamy[13] for 2-stage problems is based on computing an approximate subgradient with respect to a compact convex programming formulation, and this is done by estimating each component of the subgradient sufficiently accurately, and then applying the ellipsoid method using these approximate subgradients. In the sample average approximation method, we merely sample scenarios a given (polynomial) number of times N, and by computing the frequencies of occurrence in these samples, we derive a new LP that is a polynomial-sized approximation to the original exponential-sized LP, and the solve this compact LP explicitly. We first argue that using (approximate) subgradients one can establish a notion of closeness between two functions (e.g., the objective functions of the "true" LP and the SAA LP), so that if two functions are "close" in terms of their subgradients, then minimizing one function is equivalent to approximately minimizing the other. Next, we show that with a polynomially bounded sample size, the objective functions of the "true" problem and the sample-average problem satisfy this "closeness-in-subgradients" property with high probability, and therefore minimizing the sample-average problem yields a near-optimal solution to the true problem; thus we prove the polynomial-time convergence of the SAA method. Our proof does not rely on anything specific to discrete probability distributions, and therefore extends to the case of continuous distributions.

Compare now the 3-stage and 2-stage problems. In the 2-stage fractional set-covering problem, the compact convex program has variables corresponding only to the decisions made at the root to (fractionally) buy sets. Each component of the subgradient at the current point can be estimated by sampling a leaf from the scenario tree and using the optimal dual solution for the linear program that minimizes the cost to cover each element in this leaf's target set to the extent it is not already covered by the root variables. In the 3-stage version, a *2-stage stochastic LP* plays the analogous role of the linear program and we need to obtain a near-optimal dual solution for this exponentially large mathematical program to show the closeness property. Moreover, one difficulty that is not encountered in the 2-stage case, is that now this *2-stage recourse LP is different in the sample average and the "true" problems*, since the conditional distribution of scenarios given a second-stage outcome is only *approximated* in the sample average problem. Thus to show the closeness property one has to argue that solving the dual of the sample average 2-stage recourse LP yields a near-optimal solution to the "true" 2-stage recourse LP. We introduce a novel *compact non-linear formulation of this dual*, for which we can prove such a statement for the duals, and thereby obtain the "closeness-in-subgradients" property for the 3-stage problem. In fact, this formulation yields a new means to provide lower bounds on 2-stage stochastic LPs, which might be of interest in its own right. The analogous idea can be applied inductively to obtain the FPRAS for any fixed number of stages. We believe that our proof is of independent interest and that our approach of using subgradients will find applications in proving convergence results in other stochastic models as well.

Due to its simplicity and its use in practice, the SAA method has been studied extensively in the stochastic programming literature. Although it has been shown that the SAA method produces solutions that converge to the optimal solution as the number of samples N gets sufficiently large (see, e.g., [11] and its references), no results were known that bound the number of samples needed to obtain a $(1 + \epsilon)$-optimal solution by a polynomial in the input size, $\frac{1}{\epsilon}$ and λ. Prior to our work, for 2-stage stochastic optimization, bounds on the sample size required by the SAA method were proved in [9], but this bound depends on the variance of a certain quantity that need not depend polynomially on the input size or λ. Recently, Nemirovskii and

Shapiro (personal communication) showed that for 2-stage set-cover with non-scenario-dependent second-stage costs, the bound of [9] is a polynomial bound, provided that one applies the SAA method after some preprocessing to eliminate certain first-stage decisions.

For multi-stage problems with arbitrary distributions, to the best of our knowledge, there are no results known about the rate of convergence of the sample average approximation to the true optimal solution (with high probability). In fact, we are not aware of any work (even outside of the sample average approach) that proves worst-case bounds on the sample size required for solving multi-stage stochastic linear programs with arbitrary distributions in the black-box model. Very recently, Shapiro [12] proved bounds on the sample size required in the SAA method for multi-stage problems, under the strong assumption that *the distributions in the different stages are independent*. In particular, this implies that the distribution of the outcomes in any stage i, and hence of the scenarios in stage k, does not depend on the outcomes in the previous stages, which fails to capture the notion of learning new information about the uncertainty as one proceeds through the stages. Moreover, as in the 2-stage case, the bounds in [12] are not polynomial in the input size or λ, even when the number of stages is fixed. It is important to note that we prove that an optimal solution to the SAA LP is a near-optimal solution to true LP, not that the optimal value of the SAA LP is a good approximation to the true optimal value. Indeed, one interesting question is to show, for any class of stochastic IPs and LPs, if one could obtain an approximation algorithm to the case in which there are only a polynomial number of scenarios, then one can also obtain an approximation algorithm for the general case. Subsequent to the dissemination of early versions of our work [14], Charikar, Chekuri and Pál [3] have obtained such a result for 2-stage problems.

There has been a series of recent papers on approximation algorithms for 2-stage stochastic integer programming problems. Most of this work has focused on more restricted mechanisms for specifying the distribution of inputs [4, 10, 8]; Gupta, Pál, Ravi, and Sinha [5] were the first to consider the "black-box" model, and gave approximation algorithms for various 2-stage problems, but with the restriction that the second-stage costs be proportional to the first-stage costs. Shmoys and Swamy [13] showed that one could derive approximation algorithms for most of the stochastic integer programming problems considered in [4, 10, 8, 5] by adopting a natural LP rounding approach that, in effect, converted an LP-based approximation guarantee for the deterministic analogue to a guarantee for the stochastic generalization (where the performance guarantee degraded by a factor of 2 in the process).

An immediate consequence of our approximation scheme for multi-stage stochastic linear programs is that we obtain approximation algorithms for several natural multi-stage stochastic integer programming problems, by extending the rounding approach of [13]. The only other work on multi-stage problems in the black-box model is due to Hayrapetyan, Swamy, and Tardos [7], and Gupta et al. [6] (done concurrently with this work). Both present $O(k)$-approximation algorithms for a k-stage version of the Steiner tree problem under some restrictions on the costs; the latter also gives algorithms for the k-stage versions of the vertex cover and facility location problems under the same cost restrictions, but their approximation ratio is *exponential* in k. In contrast, in the black-box model without any cost restrictions, we obtain performance guarantees of $k \log n$ for k-stage set cover, $2k$ for k-stage vertex cover and k-stage multicut on trees, and $1.71(k-1)+1.52$ for the k-stage facility location problem. Finally, we obtain a FPRAS for a k-stage multicommodity flow problem as a direct consequence of our stochastic linear programming result.

2. Preliminaries

We state some definitions and basic facts that we will frequently use. Let $\|u\|$ denote the ℓ_2 norm of u. We say that function $g : \mathbb{R}^m \mapsto \mathbb{R}$, has *Lipschitz constant* K if $|g(v) - g(u)| \leq K\|v - u\|$ for all $u, v \in \mathbb{R}^m$.

Definition 2.1 *We say that d is a subgradient of a function $g : \mathbb{R}^m \mapsto \mathbb{R}$ at the point u if the inequality $g(v) - g(u) \geq d \cdot (v - u)$ holds for every $v \in \mathbb{R}^m$. We say that $\widehat{d}$ is an $(\omega, \Delta, \mathcal{D})$-subgradient of g at $u \in \mathcal{D}$ if for every $v \in \mathcal{D}$, we have $g(v) - g(u) \geq \widehat{d} \cdot (v - u) - \omega g(u) - \omega g(v) - \Delta$.*

The above definition is slightly weaker than the notion of an $(\omega, \mathcal{D})$-subgradient as defined in [13], but it is easy to see that one can also implement the algorithm in [13] using the notion of an approximate subgradient given by Definition 2.1. The following claim will be useful in bounding the Lipschitz constant of the functions encountered.

Claim 2.2 ([13]) *Let $d(x)$ denote a subgradient of a function $g : \mathbb{R}^m \mapsto \mathbb{R}$ at point x. Suppose $\|d(x)\| \leq K$ for every x. Then $g(.)$ has Lipschitz constant (at most) K.*

We will consider both convex minimization problems and concave maximization problems where we optimize over a polytope $\mathcal{P} \subseteq \mathbb{R}^m_{\geq 0}$. Analogous to Definition 2.1, we define a $\max$-*subgradient*, and an approximate version of it, that we use for concave maximization problems.

Definition 2.3 *We say that d is a $\max$-subgradient of a function $g : \mathbb{R}^m \mapsto \mathbb{R}$ at $u \in \mathbb{R}^m$ if for every point $v \in \mathbb{R}^m$, we have $g(v) - g(u) \leq d \cdot (v - u)$. We say that $\widehat{d}$ is an $(\omega, \Delta, \mathcal{D})$-$\max$-subgradient of $g(.)$ at $u \in \mathcal{D}$ if for every $v \in \mathcal{D}$ we have $g(v) - g(u) \leq \widehat{d} \cdot (v - u) + \omega g(u) + \Delta$.*

When $\mathcal{D}$ is clear from the context, we drop the $\mathcal{D}$ from $(\omega, \Delta, \mathcal{D})$-subgradient and $(\omega, \Delta, \mathcal{D})$-max-subgradient, and if $\Delta = 0$ we drop it from the notation. We will frequently use $(\omega, \Delta, \mathcal{P})$-subgradients which we abbreviate to (ω, Δ)-subgradients. We need the following sampling lemma which is proved using simple Chernoff bounds.

Lemma 2.4 *Let $X_i, i = 1, \ldots, \mathcal{N} = \frac{4(1+\alpha)^2}{c^2} \ln\left(\frac{2}{\delta}\right)$ be iid random variables where each $X_i \in [-a, b]$, $a, b > 0$, $\alpha = \max(1, a/b)$, and c is an arbitrary positive number. Let $X = \left(\sum_i X_i\right)/\mathcal{N}$ and $\mu = \mathrm{E}\left[X\right] = \mathrm{E}\left[X_i\right]$. Then $\Pr\left[X \in [\mu - cb, \mu + cb]\right] \geq 1 - \delta$.*

3. The Sample Average Approximation method

Suppose we have a black-box that can generate, for any sequence of outcomes for the initial stages, independent samples from the conditional distribution of scenarios given those initial outcomes. A natural approach to computing near-optimal solutions for these problems given such sampling access is the sample average approximation (SAA) approach: sample some $\mathcal{N}$ times from the distribution on scenarios, estimate the actual distribution by the distribution induced by the samples, and solve the multi-stage problem specified by this approximate distribution. For 2-stage programs, we just estimate the probability of a scenario A by its frequency in the sampled set; for k-stage programs we construct an approximate k-level distribution tree by sampling repeatedly at each level: we sample $\mathcal{T}_2$ times to obtain some stage 2 outcomes, for each sampled outcome we sample $\mathcal{T}_3$ times from the conditional distribution given that outcome and so on, and for each sampled outcome we estimate its conditional probability of occurrence given the previous-stage outcome by its frequency in the sampled set. The multi-stage problem specified by the approximate distribution is called the *sample average problem*, and its objective function is called the *sample average function*.

If the total number of samples $\mathcal{N}$ is polynomially bounded, then since the approximate distribution has support of size at most $\mathcal{N}$, the sample average problem can be solved efficiently by solving a polynomial size linear program. The issue here is the sample size $\mathcal{N}$ required to guarantee that *every optimal solution to the sample-average problem is a near-optimal solution to the original problem* with high probability. We show that for any given k (which is not part of the input), for a large class of k-stage stochastic LPs, one can bound $\mathcal{N}$ by a polynomial in the input size, the inverse of the desired accuracy, and the maximum *ratio* λ between the cost of an action in successive stages.

Intuitively, to prove such a theorem, we need to show that the sample-average function is a close approximation to the true function in some sense. One obvious approach would be to argue that, with high probability, the values of the sample average function and the true function are close to each other, at a sufficiently dense set of points. This however immediately runs into problems since the variance in the scenario costs could be quite (exponentially) large, so one cannot estimate the true function value, that is, the expected scenario cost, to within a reasonable accuracy using a small (polynomial) number of samples. The basic problem is that there could be very low-probability outcomes that contribute significantly towards the cost in the true problem, but will almost never be sampled with only a polynomial number of samples (so they contribute nothing to the sample average function). The key insight is that such *rare outcomes do not much influence the optimal first-stage decisions*, since one would defer decisions for such outcomes till later. The minimizer of a convex function is determined by its "slope" (i.e., gradient or subgradient), which suggests that perhaps we should compare the slopes of the sample-average and the true objective functions and show that they are close to each other, and argue that this is sufficient to prove the near-equivalence of the corresponding minimization problems. Our proof builds upon this intuition. A *subgradient* is the analogue of a gradient for a non-differentiable function, and is a measure of the "slope" of the function. We identify a notion of closeness between any two functions based on their (approximate) subgradients so that if two functions are close under this criterion, then minimizing one is approximately equivalent to minimizing the other. Next, we show that the objective functions of the original multi-stage problem, and the sample average problem with polynomially bounded sample size, satisfy this "closeness-in-subgradients" property, and thus we obtain the desired result.

Proof details. The proof is organized as follows. First, in Lemma 3.1 we show that given functions g and $\widehat{g}$ that agree in terms of their (approximate) subgradients at points in a polytope $\mathcal{P}$, *every* optimal solution to $\min_{x \in \mathcal{P}} \widehat{g}(x)$ is a near-optimal solution to $\min_{x \in \mathcal{P}} g(x)$. Some intuition about why this closeness-in-subgradients property is sufficient can be obtained by considering the ellipsoid-based algorithm for convex minimization given in [13]. This algorithm uses only (approximate) subgradient information about the convex function to be minimized, using a subgradient or an ω-subgradient of the function to derive a cut passing through the center of the current ellipsoid at a feasible point and make progress. Suppose at every feasible point $x \in \mathcal{P}$, there is a vector d_x that is both a subgradient of $\widehat{g}(.)$ and an ω-subgradient of $g(.)$ at x. One can then use d_x to generate the cut at x, so that the ellipsoid-based algorithm will run *identically* on both $\min_{x \in \mathcal{P}} g(x)$ and $\min_{x \in \mathcal{P}} \widehat{g}(x)$ and return a point that is *simultaneously* near-optimal for both objective functions. Lemma 3.1 makes this intuition precise while weakening the assumption and strengthening the conclusion: we only require that at every point x in a *sufficiently dense finite set* $G \subseteq \mathcal{P}$ there be a vector d_x that is both a subgradient of $\widehat{g}(.)$ and an ω-subgradient of $g(.)$, and we prove that *every* optimal solution to $\min_{x\in\mathcal{P}} \widehat{g}(x)$ is a near-optimal solution to $\min_{x\in\mathcal{P}} g(x)$. Lemma 3.2 proves an analogous result for maximization problems.

The second part of the proof is to show that the objective functions of the true problem and the sample average problem (with polynomial samples) satisfy this closeness-in-subgradients property. This is divided into three parts. For the class of 2-stage linear programs considered in [13], this is easy to show (Theorem 3.5) because the subgradient at any point is the expectation (according to the scenario distribution) of a quantity derived from the optimal solutions to the dual of the recourse LP for each scenario, and this recourse LP is the same in both the sample average and the true problems. Thus, since the subgradient components have bounded variance [13], the closeness property follows.

For the k-stage problem however, one needs to develop several substantial new ideas to show this closeness property, even when $k = 3$. We introduce these ideas in Section 4 by focusing on 3-stage problems, and in particular, on the LP relaxation of 3-stage set cover as an illustrative example. We then generalize these ideas to prove an SAA theorem for a large class of 3-stage linear programs, and in Section 5 inductively apply the arguments to a broad class of k-stage problems. The main difficulty, and the essential difference from the 2-stage case, is that now the recourse problem for each second-stage outcome is a 2-stage stochastic LP whose underlying distribution is only estimated in the sample average problem. So *the sample average problem and the true problem solve different recourse problems for each stage 2 outcome*. A (approximate) subgradient is obtained from the (approximately) optimal solutions to the dual of the 2-stage recourse LP for each scenario, therefore to show closeness in subgradients we need to argue that maximizing the sample average dual yields a near-optimal solution to the true dual, that is, prove an SAA theorem for the *dual* of a 2-stage stochastic primal program! Mimicking the approach for the primal problem, we prove this by showing that the two dual objective functions agree in terms of their $\max$-*subgradients*. However, simply considering the LP dual of the 2-stage primal recourse LP does not work; a $\max$-subgradient of the linear dual objective function is just the constant vector specifying the conditional probabilities of the stage 3 scenarios given the outcome in stage 2, and one cannot estimate the true conditional distribution using only a polynomial number of samples, in particular, because rare scenarios will almost never be sampled. To circumvent this problem, we introduce a novel *compact, non-linear* formulation of the dual, which turns the dual objective function into a concave function whose $\max$-subgradient can be computed by solving a 2-stage primal stochastic problem. We use the earlier SAA theorem for 2-stage programs to show that any optimal solution to this 2-stage LP in the sample average dual, is a near-optimal solution to the 2-stage LP in the true dual. This shows that the two dual objective functions (in this new representation) are close in terms of their $\max$-subgradients, thereby proving that an optimal solution the sample average dual is a near-optimal solution to the true dual. This in turn establishes the closeness in subgradients of the objective functions of the 3-stage sample average problem and the true 3-stage problem and yields the SAA theorem.

Sufficiency of closeness in subgradients. Let $g : \mathbb{R}^m \mapsto \mathbb{R}$ and $\widehat{g} : \mathbb{R}^m \mapsto \mathbb{R}$ be two functions with Lipschitz constant (at most) K. Let $\mathcal{P} \subseteq \mathbb{R}^m_{\geq 0}$ be the bounded feasible region and R be a radius such that $\mathcal{P}$ is contained in the ball $B(\mathbf{0}, R) = \{x : \|x\| \leq R\}$. Let $\epsilon, \gamma > 0$ be two parameters with $\gamma \leq 1$. Set $N = \log\left(\frac{2KR}{\epsilon}\right)$ and $\omega = \frac{\gamma}{8N}$. Let $G' = \{x \in \mathcal{P} : x_i = n_i \cdot \left(\frac{\epsilon}{KN\sqrt{m}}\right),\ n_i \in \mathbb{Z} \text{ for all } i = 1, \dots, m\}$ Set $G = G' \cup \{x + t(y - x), y + t(x - y) : x, y \in G',\ t = 2^{-i},\ i = 1, \dots, N\}$. We call G the *extended* $\frac{\epsilon}{KN\sqrt{m}}$*-grid* of the polytope $\mathcal{P}$. Note that for every $x \in \mathcal{P}$, there exists $x' \in G'$ such that $\|x - x'\| \leq \frac{\epsilon}{KN}$. Fix $\Delta > 0$. We first consider minimization problems. We say that functions g and $\widehat{g}$ satisfy property (A) if

$$\forall x \in G,\ \exists \widehat{d}_x \in \mathbb{R}^m :\ \widehat{d}_x \text{ is a subgradient of } \widehat{g}(.), \text{ and an } (\omega, \Delta)\text{-subgradient of } g(.) \text{ at } x. \quad \text{(A)}$$

Lemma 3.1 *Suppose functions g and $\widehat{g}$ satisfy property* (A). *Let $x^*, \widehat{x} \in \mathcal{P}$ be points that respectively minimize $g(.)$ and $\widehat{g}(.)$, with $g(x^*) \geq 0$, and let $x' \in \mathcal{P}$ be a point such that $\widehat{g}(x') \leq \widehat{g}(\widehat{x}) + \rho$. Then, (i) $g(\widehat{x}) \leq (1+\gamma)g(x^*) + 6\epsilon + 2N\Delta$; (ii) $g(x') \leq (1+\gamma)g(x^*) + 6\epsilon + 2N\Delta + 2N\rho$.*

Proof : We prove part (i); part (ii) is proved almost identically. Suppose first that $\widehat{x} \in G'$. Let $\tilde{x}$ be the point in G' closest to x^*, so $\|\tilde{x} - x^*\| \leq \frac{\epsilon}{KN}$ and therefore $g(\tilde{x}) \leq g(x^*) + \epsilon$. Let $y = \widehat{x}\left(1 - \frac{1}{2^N}\right) + \left(\frac{1}{2^N}\right)\tilde{x} \in G$ and consider the vector $\widehat{d}_y$ given by property (A). It must be that $\widehat{d}_y \cdot (\widehat{x} - y) = -\widehat{d}_y \cdot (\tilde{x} - y) \leq 0$, otherwise we would have $\widehat{g}(\widehat{x}) > \widehat{g}(y)$ contradicting the optimality of $\widehat{x}$. So, by the definition of an (ω, Δ)-subgradient, we have $g(y) \leq \frac{(1+\omega)g(\tilde{x})+\Delta}{1-\omega} \leq (1 + 4\omega)(g(\tilde{x}) + \Delta) \leq (1 + \gamma)g(x^*) + 2\epsilon + 2\Delta$ since $\omega = \frac{\gamma}{8N} \leq \frac{1}{4}$. Also $\|\widehat{x} - y\| = \frac{\|\widehat{x}-\tilde{x}\|}{2^N} \leq \frac{\epsilon}{K}$ since $\|\widehat{x} - \tilde{x}\| \leq 2R$. So, $g(\widehat{x}) \leq g(y) + \epsilon \leq (1 + \gamma)g(x^*) + 3\epsilon + 2\Delta$.

Now suppose $\widehat{x} \notin G'$. Let $\bar{x}$ be the point in G' closest to $\widehat{x}$, so $\|\bar{x} - \widehat{x}\| \leq \frac{\epsilon}{KN}$ and $\widehat{g}(\bar{x}) \leq \widehat{g}(\widehat{x}) + \frac{\epsilon}{N}$. For any $y \in G$, we have $\widehat{d}_y \cdot (\bar{x} - y) \leq \frac{\epsilon}{N}$, otherwise $\widehat{g}(\bar{x}) > \widehat{g}(y) + \frac{\epsilon}{N}$. Let $y_0 = \tilde{x}$, and $y_i = (\bar{x} + y_{i-1})/2$ for $i = 1, \dots, N$. For each y_i, $\widehat{d}_{y_i} \cdot (y_{i-1} - y_i) = -\widehat{d}_{y_i} \cdot (\bar{x} - y_i) \geq -\frac{\epsilon}{N}$, and since $\widehat{d}_{y_i}$ is an (ω, Δ)-subgradient of $g(.)$ at y_i, $g(y_i) \leq (1+4\omega)(g(y_{i-1}) + \frac{\epsilon}{N} + \Delta)$. This implies that $g(y_N) \leq (1+4\omega)^N(g(\tilde{x}) + \epsilon + N\Delta) \leq (1+\gamma)g(x^*) + 4\epsilon + 2N\Delta$. So $g(\widehat{x}) \leq g(y_N) + 2\epsilon \leq (1+\gamma)g(x^*) + 6\epsilon + 2N\Delta$. ■

Lemma 3.2 states an analogous result for maximization problems. We say that g and $\widehat{g}$ satisfy property (B) if

$$\forall x \in G,\ \exists \widehat{d}_x \in \mathbb{R}^m : \widehat{d}_x \text{ is a max-subgradient of } \widehat{g}(.), \text{ and an } (\omega, \Delta)\text{-max-subgradient of } g(.) \text{ at } x. \quad \text{(B)}$$

Lemma 3.2 *Suppose functions g and $\widehat{g}$ satisfy property* (B). *Let x^* and $\widehat{x}$ be points in $\mathcal{P}$ that respectively maximize functions $g(.)$ and $\widehat{g}(.)$, and suppose $g(x^*) \geq 0$. Then, $g(\widehat{x}) \geq (1-\gamma)g(x^*) - 4\epsilon - N\Delta$.*

Lemma 3.3 *Let G be the extended ϵ-grid of $\mathcal{P}$. Then $|G| \leq N\left(\frac{2R}{\epsilon}\right)^{2m}$.*

The SAA bound for 2-stage programs. We now prove a polynomial SAA bound for the class of 2-stage programs considered in [13] (this was stated with extra constraints $B^A s_A \geq h^A$ but these are handled below).

$$\min_{x \in \mathcal{P}} h(x) = w^{\mathrm{I}} \cdot x + \sum_{A \in \mathcal{A}} p_A f_A(x), \quad (\mathcal{P} \subseteq \mathbb{R}^m_{\geq 0}) \quad \text{(P)}$$

$$f_A(x) = \min\Big\{ w^A \cdot r_A + q^A \cdot s_A : r_A \in \mathbb{R}^m_{\geq 0}, s_A \in \mathbb{R}^n_{\geq 0}, D^A s_A + T^A r_A \geq j^A - T^A x \Big\}.$$

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 the primal and dual problems corresponding to $f_A(x)$ are feasible for every scenario A. It is assumed that $\mathcal{P} \subseteq B(\mathbf{0}, R)$ where $\ln R$ is polynomially bounded. Define $\lambda = \max\big(1, \max_{A \in \mathcal{A}, S} \frac{w_S^A}{w_S^{\mathrm{I}}}\big)$; we assume that λ is known. Let OPT be the optimum value and $\mathcal{I}$ denote the size of the input. The sample average problem is to minimize the sample average function $\widehat{h}(x) = w^{\mathrm{I}} \cdot x + \sum_{A \in \mathcal{A}} \widehat{p}_A f_A(x)$ over $x \in \mathcal{P}$, where $\widehat{p}_A = \mathcal{N}_A / \mathcal{N}$, $\mathcal{N}$ is the total number of samples and $\mathcal{N}_A$ is the number of times scenario A is sampled.

Lemma 3.4 *Let d be a subgradient of $h(.)$ at the point $x \in \mathcal{P}$, and suppose that $\widehat{d}$ is a vector such that $\widehat{d}_S \in [d_S - \omega w_S^{\mathrm{I}}, d_S + \omega w_S^{\mathrm{I}}]$ for all S. Then $\widehat{d}$ is an ω-subgradient (i.e., an $(\omega, 0)$-subgradient) of $h(.)$ at x.*

It is shown in [13] that at any point $x \in \mathcal{P}$, if (z_A^*) is an optimal solution to the dual of $f_A(x)$, then (i) $d_x = w^{\mathrm{I}} - \sum_A p_A (T^A)^{\mathrm{T}} z_A^*$ is a subgradient of $h(.)$; (ii) for any component S and any scenario A, component S of the vector $w^{\mathrm{I}} - (T^A)^{\mathrm{T}} z_A^*$ lies in $[-\lambda w_S^{\mathrm{I}}, w_S^{\mathrm{I}}]$; and therefore (iii) $\|d_x\| \leq \lambda \|w^{\mathrm{I}}\|$. The sample average function $\widehat{h}(.)$ has the same form as $h(.)$, but has a different distribution, so $\widehat{d}_x = w^{\mathrm{I}} - \sum_A \widehat{p}_A (T^A)^{\mathrm{T}} z_A^*$ is a subgradient of $\widehat{h}(.)$ at x, and $\|\widehat{d}_x\| \leq \lambda \|w^{\mathrm{I}}\|$. So (by Claim 2.2) the Lipschitz constant of $h, \widehat{h}$ is at most $K = \lambda \|w^{\mathrm{I}}\|$. Observe that $\widehat{d}_x$ is just $w^{\mathrm{I}} - (T^A)^{\mathrm{T}} z_A^*$ averaged over the random scenarios sampled to construct $\widehat{h}(.)$, and $\mathrm{E}\big[\widehat{d}_x\big] = d_x$ where the expectation is over these random samples.

Theorem 3.5 *For any $\epsilon, \gamma > 0$ ($\gamma < 1$) with probability at least $1 - \delta$, any optimal solution $\widehat{x}$ to the sample average problem constructed with $\mathsf{poly}\big(\mathcal{I}, \lambda, \frac{1}{\gamma}, \ln(\frac{1}{\epsilon}), \ln(\frac{1}{\delta})\big)$ samples, satisfies $h(\widehat{x}) \leq (1+\gamma) \cdot OPT + 6\epsilon$.*

Proof : We show that $h(.)$ and $\widehat{h}(.)$ satisfy property (A) with probability $1 - \delta$ with the stated sample size; the rest follows from Lemma 3.1. Define $N = \log\big(\frac{2KR}{\epsilon}\big)$, $\omega = \frac{\gamma}{8N}$ and let G be the extended $\frac{\epsilon}{KN\sqrt{m}}$-grid. Note that $\log(KR)$ is polynomially bounded in the input size. Let $n = |G|$. Using Lemma 2.4, if we sample $\mathcal{N} = \frac{4(1+\lambda)^2}{3\omega^2} \ln\big(\frac{2mn}{\delta}\big)$ times to construct $\widehat{h}(.)$ then at a given point x, subgradient $\widehat{d}_x$ of $\widehat{h}(.)$ is component-wise close to its expectation with probability at least $1 - \delta/n$, so by Lemma 3.4 $\widehat{d}_x$ is an ω-subgradient of $h(.)$ at x (with high probability). So with probability at least $1 - \delta$, $\widehat{d}_x$ is an ω-subgradient of $h(.)$ at *every* point $x \in G$. Using Lemma 3.3 to bound n, we get that $\mathcal{N} = O\big(m\lambda^2 \log^2(\frac{2KR}{\epsilon}) \ln(\frac{2KRm}{\epsilon\delta})\big)$. ■

Under the mild assumption that (a) the point $x = \mathbf{0}$ (i.e., deferring all decisions to stage 2) lies in $\mathcal{P}$, and (b) for every scenario A, either $f_A(x)$ is minimized with $x = \mathbf{0}$, or $w^{\mathrm{I}} \cdot x + f_A(x) \geq 1$ for every $x \in \mathcal{P}$, it was shown in [13] that by sampling $\lambda \ln\big(\frac{1}{\delta}\big)$ times initially one can detect with probability $1 - \delta$ (with $\delta \leq \frac{1}{2}$), that either $x = \mathbf{0}$ is an optimal solution to (P), or that $OPT \geq \frac{\delta}{\lambda \ln(1/\delta)}$. So if we detect that OPT is large, then we can set γ, ϵ appropriately to get a $(1+\kappa)$-optimal solution with probability $1 - 2\delta$, using the SAA method with $\mathsf{poly}\big(\mathcal{I}, \lambda, \frac{1}{\kappa}, \ln(\frac{1}{\delta})\big)$ samples.

4. 3-stage stochastic programs

3-stage stochastic set cover. Our techniques yield a polynomial-sample bound for a broad class of 3-stage programs, but before considering a generic 3-stage program, we introduce and explain the main ideas involved by focusing on the 3-stage stochastic set cover problem.

In the stochastic set cover problem, we are given a universe U of n elements and a family $\mathcal{S}$ of m subsets of U, and the set of elements to cover is determined by a probability distribution. In the 3-stage problem this distribution is specified by a 3-level tree. We use A to denote an outcome in stage 2, and (A, B) to denote a stage 3 scenario where A was the stage 2 outcome. Let $\mathcal{A}$ be the set of all stage 2 outcomes, and for each $A \in \mathcal{A}$ let $\mathcal{B}_A = \{B : (A,B) \text{ is a scenario}\}$. Let p_A and $p_{A,B}$ be the probabilities of outcome A and scenario (A,B) respectively, and let $q_{A,B} = \frac{p_{A,B}}{p_A}$. Note that $\sum_{A\in\mathcal{A}} p_A = 1 = \sum_{B\in\mathcal{B}_A} q_{A,B}$ for every $A \in \mathcal{A}$. We have to cover the (random) set of elements $\mathcal{E}(A,B)$ in scenario (A,B), and we can buy a set S in stage 1, or in stage 2 outcome A, or in scenario (A,B) incurring a cost of w^{I}_S, w^A_S and $w^{A,B}_S$ respectively.

We use x, y_A and $z_{A,B}$ respectively to denote the decisions in stage 1, outcome A and scenario (A,B) respectively and consider the following fractional relaxation:

$$\min_{0\le x_S\le 1\ \forall S} h(x) = \sum_S w^{\mathrm{I}}_S x_S + \sum_{A\in\mathcal{A}} p_A f_A(x), \quad \text{(3SSC-P)}$$

$$f_A(x) = \min_{y_A\ge \mathbf{0}} \Big\{ \sum_S w^A_S y_{A,S} + \sum_{B\in\mathcal{B}_A} q_{A,B} f_{A,B}(x,y_A) \Big\}, \quad \text{(3Rec-P)}$$

$$f_{A,B}(x,y_A) = \min_{z_{A,B}\in\mathbb{R}^m_{\ge 0}} \Big\{ \sum_S w^{A,B}_S z_{A,B,S} : \sum_{S:e\in S} z_{A,B,S} \ge 1 - \sum_{S:e\in S} (x_S + y_{A,S})\ \forall e \in \mathcal{E}(A,B) \Big\}.$$

Let $\mathcal{P} = \{x \in \mathbb{R}^m : x_S \in [0,1]\ \ \forall S\}$ and $OPT = \min_{x\in\mathcal{P}} h(x)$. The total sample size in the sample average problem is $\mathcal{T}_2 \cdot \mathcal{T}_3$ where, (i) $\mathcal{T}_2$ is the sample size used to estimate probability p_A by the frequency $\widehat{p}_A = \mathcal{T}_{2;A}/\mathcal{T}_2$, and (ii) $\mathcal{T}_3$ is the number of samples generated from the conditional distribution of scenarios in $\mathcal{B}_A$ for each A with $\widehat{p}_A > 0$ to estimate $q_{A,B}$ by $\widehat{q}_{A,B} = \mathcal{T}_{3;A,B}/\mathcal{T}_3$. The sample average problem is similar to (3SSC-P) with $\widehat{p}_A$ replacing p_A, and $\widehat{q}_{A,B}$ replacing $q_{A,B}$ in the recourse problem $f_A(x)$. We use $\widehat{f}_A(x) = \min_{y_A\ge\mathbf{0}}\big(w^A \cdot y_A + \sum_{B\in\mathcal{B}_A} \widehat{q}_{A,B} f_A(x,y_A)\big)$ and $\widehat{h}(x) = w^{\mathrm{I}} \cdot x + \sum_{A\in\mathcal{A}} \widehat{p}_A \widehat{f}_A(x)$ to denote the sample average recourse problem for outcome A and the sample average function respectively.

As mentioned earlier, the main difficulty in showing that the sample average and the true functions satisfy the closeness-in-subgradients property, is that these two problems now solve different recourse problems, $\widehat{f}_A(x)$ and $f_A(x)$ respectively, for an outcome A. Since the subgradient is obtained from a dual solution, this entails first proving an SAA theorem for the dual which suggests that solving the dual of $\widehat{f}_A(x)$ yields a near-optimal solution to the dual of $f_A(x)$. To achieve this, we first formulate the dual as a compact concave maximization problem, then show that by slightly modifying the two dual programs, the dual objective functions become close in terms of their max-subgradients, and then use Lemma 3.2 to obtain the required SAA theorem (for the duals). A max-subgradient of the dual objective function is obtained from the optimal solution of a 2-stage primal problem and we use Theorem 3.5 to prove the closeness in max-subgradients of the sample average dual and the true dual. In Section 5 we show that this argument can be applied inductively to prove an SAA bound for a large class of k-stage stochastic LPs.

Let $f_A(\mathbf{0}; W)$ (respectively $\widehat{f}_A(\mathbf{0}; W)$) denote the recourse problem $f_A(x)$ (respectively $\widehat{f}_A(x)$) with $x = \mathbf{0}$ and costs $w^A = W$, that is, $f_A(\mathbf{0}; W) = \min_{y_A\ge\mathbf{0}}\big(W \cdot y_A + \sum_{B\in\mathcal{B}_A} q_{A,B} f_{A,B}(\mathbf{0}, y_A)\big)$. We formulate the following dual of the true and sample average recourse problems:

$$LD_A(x) = \max_{\mathbf{0}\le\alpha_A\le w^A} l_A(x;\alpha_A), \quad \widehat{LD}_A(x) = \max_{\mathbf{0}\le\alpha_A\le w^A} \widehat{l}_A(x;\alpha_A)$$

where $l_A(x;\alpha_A) = -\alpha_A \cdot x + f_A(\mathbf{0};\alpha_A)$ and $\widehat{l}_A(x;\alpha_A) = -\alpha_A \cdot x + \widehat{f}_A(\mathbf{0};\alpha_A)$.

Lemma 4.1 *At any point $x \in \mathcal{P}$ and outcome $A \in \mathcal{A}$, $f_A(x) = LD_A(x)$ and $\widehat{f}_A(x) = \widehat{LD}_A(x)$.*

Lemma 4.2 *Fix $x \in \mathcal{P}$. Let α_A be a solution to $LD_A(x)$ of value $l_A(x;\alpha_A) \ge (1-\varepsilon)LD_A(x) - \varepsilon w^{\mathrm{I}} \cdot x - \epsilon$ for every $A \in \mathcal{A}$. Then, (i) $d = w^{\mathrm{I}} - \sum_A p_A\alpha_A$ is an (ε,ϵ)-subgradient of $h(.)$ at x with $\|d\| \le \lambda\|w^{\mathrm{I}}\|$; (ii) if $\widehat{d}$ is a vector such that $d - \omega w^{\mathrm{I}} \le \widehat{d} \le d + \omega w^{\mathrm{I}}$, then $\widehat{d}$ is an $(\varepsilon+\omega,\epsilon)$-subgradient of $h(.)$ at x.*

Lemma 4.1 proves strong duality (in this new dual representation), which is used by Lemma 4.2. Lemma 4.2 also shows that any optimal solution $\widehat{\alpha}_A$ to $\widehat{LD}_A(x)$ yields $\widehat{d}_x = w^{\mathrm{I}} - \sum_A \widehat{p}_A\widehat{\alpha}_A$ as a subgradient of $\widehat{h}(.)$ at x, so to prove the closeness in subgradients of h and $\widehat{h}$ it suffices to argue that $\widehat{\alpha}_A$ is a near-optimal solution to $LD_A(x)$. (Note that both h and $\widehat{h}$ have Lipschitz constant at most $K = \lambda\|w^{\mathrm{I}}\|$.) We could try to argue this by showing that $l_A(x;.)$ and $\widehat{l}_A(x;.)$ are close in terms of their max-subgradients (i.e., satisfy property (B)), however some technical difficulties arise here. A max-subgradient of $l_A(x;.)$ at α_A is obtained from a solution to the 2-stage problem given by $f_A(\mathbf{0};\alpha_A)$, and to show closeness in max-subgradients at α_A we need to argue that an optimal solution $\widehat{y}_A$ to $\widehat{f}_A(\mathbf{0};\alpha_A)$ is a near-optimal solution to $f_A(\mathbf{0};\alpha_A)$. But this need not be true since the ratio $\max_S\big(\frac{w^{A,B}_S}{\alpha_{A,S}}\big)$ of the second- and first-stage costs in the 2-stage problem $f_A(\mathbf{0};\alpha_A)$, could be unbounded. To tackle this, we consider instead the modified dual problems $LD_{A;\rho}(x) = \max_{\rho w^{\mathrm{I}}\le\alpha_A\le w^A} l_A(x;\alpha_A)$ and $\widehat{LD}_{A;\rho}(x) = \max_{\rho w^{\mathrm{I}}\le\alpha_A\le w^A} \widehat{l}_A(x;\alpha_A)$ for a suitable $\rho > 0$. Define $h_\rho(x) = w^{\mathrm{I}} \cdot x + \sum_A p_A LD_{A;\rho}(x)$ and $\widehat{h}_\rho(x) = w^{\mathrm{I}} \cdot x + \sum_A \widehat{p}_A \widehat{LD}_{A;\rho}(x)$. As in Lemma 4.2, one can show that near-optimal solutions α_A to $LD_{A;\rho}(x)$ for every $A \in \mathcal{A}$ yield an approximate subgradient of $h_\rho(.)$ at x. But now the cost ratio in the 2-stage problem $f_A(\mathbf{0};\alpha_A)$ is at most $\frac{\lambda^2}{\rho}$ for any $A \in \mathcal{A}$, and this gives the SAA bound stated in Lemma 4.3 below; we present the proof after Theorem 4.6. Using Lemma 4.3, we show the closeness in subgradients of $h_\rho(.)$ and $\widehat{h}_\rho(.)$, and this suffices to show that if $\widehat{x}$ minimizes $\widehat{h}(.)$ then it is a near-optimal solution to $h(.)$.

Lemma 4.3 *For any parameters $\epsilon, \rho, \varepsilon > 0$, any $x \in \mathcal{P}$, and any outcome $A \in \mathcal{A}$, if we use $\mathcal{T}(\epsilon, \rho, \varepsilon, \delta) = \mathsf{poly}\big(\mathcal{I}, \frac{\lambda}{\rho\varepsilon}, \ln(\frac{1}{\epsilon}), \ln(\frac{1}{\delta})\big)$ samples to construct the recourse problem $\widehat{f}_A(x)$, then any optimal solution $\widehat{\alpha}_A$ to $\widehat{LD}_{A;\rho}(x)$ satisfies $l_A(x; \widehat{\alpha}_A) \geq (1-\varepsilon)LD_{A;\rho}(x) - \varepsilon w^{\mathrm{I}} \cdot x - \epsilon$ with probability at least $1-\delta$.*

Lemma 4.4 *Consider the sample average function $\widehat{h}$ generated using $\mathcal{N}' = \mathcal{T}_2(\omega, \delta) = \frac{16(1+\lambda)^2}{\omega^2}\ln\big(\frac{4m}{\delta}\big)$ samples from stage 2, and $\mathcal{T}\big(\epsilon, \rho, \frac{\omega}{2}, \frac{\delta}{2\mathcal{N}'}\big)$ samples from stage 3 for each outcome A with $\widehat{p}_A > 0$. At any point $x \in \mathcal{P}$, subgradient $\widehat{d}_x$ of $\widehat{h}_\rho(.)$ is an (ω, ϵ)-subgradient of $h_\rho(.)$ with probability at least $1-\delta$.*

Claim 4.5 *For any $x \in \mathcal{P}$, $h_\rho(x) \leq h(x) \leq h_\rho(x) + \rho w^{\mathrm{I}} \cdot x$. Similarly $\widehat{h}_\rho(x) \leq \widehat{h}(x) \leq \widehat{h}_\rho(x) + \rho w^{\mathrm{I}} \cdot x$.*

Theorem 4.6 *For any $\epsilon, \gamma > 0$ ($\gamma \leq 1$), one can construct $\widehat{h}$ with $\mathsf{poly}\big(\mathcal{I}, \lambda, \frac{1}{\gamma}, \ln(\frac{1}{\epsilon}), \ln(\frac{1}{\delta})\big)$ samples, and with probability at least $1-\delta$, any optimal solution $\widehat{x}$ to $\min_{x\in\mathcal{P}} \widehat{h}(x)$ satisfies $h(\widehat{x}) \leq (1+7\gamma) \cdot OPT + 18\epsilon$.*

Proof : Assume that $\gamma \leq 1$ without loss of generality. Let $N = \log\big(\frac{2KR}{\epsilon}\big)$ and $\omega = \frac{\gamma}{8N}$. Note that $\log(KR)$ is polynomially bounded in the input size. Set $\epsilon' = \frac{\epsilon}{N}$ and $\rho = \frac{\gamma}{4N}$. We show that (i) a near-optimal solution to $\min_{x\in\mathcal{P}} \widehat{h}_\rho(x)$ yields a near-optimal solution to $\min_{x\in\mathcal{P}} h_\rho(x)$, and (ii) minimizing $h(.)$ and $\widehat{h}(.)$ over $\mathcal{P}$ is roughly the same as approximately minimizing $h_\rho(.)$ and $\widehat{h}_\rho(.)$ respectively over $\mathcal{P}$.

Let $\tilde{x}$ be an optimal solution to $\min_{x\in\mathcal{P}} \widehat{h}_\rho(x)$. By Claim 4.5, $\widehat{h}_\rho(\widehat{x}) \leq \widehat{h}_\rho(\tilde{x}) + \rho w^{\mathrm{I}} \cdot \tilde{x}$, and $0 \leq OPT_\rho = \min_{x\in\mathcal{P}} h_\rho(x) \leq OPT$. Let G be the extended $\frac{\epsilon}{KN\sqrt{m}}$-grid of $\mathcal{P}$ and $n = |G|$. Let $\mathcal{N}' = \frac{16(1+\lambda)^2}{\omega^2}\ln\big(\frac{4mn}{\delta}\big)$ which is a polynomial in $\mathcal{I}, \frac{\lambda}{\gamma}, \ln\big(\frac{1}{\epsilon}\big)$ and $\ln\big(\frac{1}{\delta}\big)$, where we use Lemma 3.3 to bound n. We construct $\widehat{h}(.)$ using $\mathcal{N} = \mathcal{N}' \cdot \mathcal{T}\big(\epsilon', \rho, \frac{\omega}{2}, \frac{\delta}{2n\mathcal{N}'}\big)$ samples. Since $\mathcal{N}'$ is polynomially bounded, Lemma 4.3 shows that so is $\mathcal{N}$. Using Lemma 4.4 and the union bound over all points in G, probability at least $1-\delta$, at every point $x \in G$, subgradient $\widehat{d}_x$ of $\widehat{h}_\rho(.)$ is an (ω, ϵ')-subgradient of $h_\rho(.)$. So by parts (i) and (ii) of Lemma 3.1, we have that $h_\rho(\tilde{x}) \leq (1+\gamma)OPT_\rho + 6\epsilon + 2N\epsilon'$ and $h_\rho(\widehat{x}) \leq (1+\gamma)OPT_\rho + 6\epsilon + 2N(\rho w^{\mathrm{I}} \cdot \tilde{x} + \epsilon')$ with high probability (since $\widehat{h}_\rho(\widehat{x}) \leq \widehat{h}_\rho(\tilde{x}) + \rho w^{\mathrm{I}} \cdot \tilde{x}$). Combining this with the bound on OPT_ρ, the bounds $(1-\rho)w^{\mathrm{I}} \cdot \tilde{x} \leq h_\rho(\tilde{x})$, $(1-\rho)h(\widehat{x}) \leq h_\rho(\widehat{x})$ (Claim 4.5), and plugging in ϵ' and ρ, we get that $h(\widehat{x}) \leq (1+7\gamma)OPT + 18\epsilon$. ■

Under the very mild assumption that for every scenario (A, B) with $\mathcal{E}(A, B) \neq \emptyset$, for every $x \in \mathcal{P}$ and $y_A \geq \mathbf{0}$, $w^{\mathrm{I}} \cdot x + w^A \cdot y_A + f_{A,B}(x, y_A) \geq 1$, we have the following.

Lemma 4.7 *By sampling $M = \lambda^2 \ln\big(\frac{1}{\delta}\big)$ times, one can detect with probability $1-\delta$ ($\delta < \frac{1}{2}$) that either $x = \mathbf{0}$ is an optimal solution to* (3SSC-P)*, or that $OPT \geq \frac{\delta}{M}$.*

Thus, as in the 2-stage case, we can obtain a $(1+\kappa)$-optimal solution with the SAA method (with high probability) using polynomially many samples.

Proof of Lemma 4.3 : Let $\mathcal{D}_A = \{\alpha_A \in \mathbb{R}^m : \rho w^{\mathrm{I}} \leq \alpha_A \leq \omega^A\}$. Clearly we may assume that $y_{A,S} \leq 1$ in the problems $f_A(\mathbf{0}; \alpha_A)$ and $\widehat{f}_A(\mathbf{0}; \alpha_A)$. Let $R' = \|w^A\| \leq \lambda\|w^{\mathrm{I}}\|$, so $\mathcal{D}_A \subseteq B(\mathbf{0}, R')$. We want to show that if $\widehat{\alpha}_A$ solves $\widehat{LD}_{A;\rho}(x)$, then $l_A(x; \widehat{\alpha}_A) \geq (1-\varepsilon)LD_{A;\rho}(x) - \varepsilon w^{\mathrm{I}} \cdot x - \epsilon$ with high probability. By a now familiar approach, we will show that $\widehat{l}_A(x; .)$ and $l_A(x; .)$ are close in terms of their max-subgradients and then use Lemma 3.2. Let $g(\alpha_A; y_A) = \alpha_A \cdot y_A + \sum_{B\in\mathcal{B}_A} q_{A,B} f_{A,B}(\mathbf{0}, y_A)$. We only consider $(\omega, \Delta, \mathcal{D}_A)$-$\max$-subgradients, so we drop $\mathcal{D}_A$ from now on. A $\max$-subgradient to $l_A(x; .)$ (resp. $\widehat{l}_A(x; .)$) at α_A is obtained from the solution to the 2-stage problem $f_A(\mathbf{0}; \alpha_A)$ (resp. $\widehat{f}_A(\mathbf{0}; \alpha_A)$):

Lemma 4.8 *Fix $x \in \mathcal{P}$ and $\alpha_A \in \mathcal{D}_A$. Let $\omega' = \frac{\omega}{\lambda}$. If y_A is a solution to $f_A(\mathbf{0}; \alpha_A)$ of value $g(\alpha_A; y_A) \leq (1+\omega')f_A(\mathbf{0}; \alpha_A) + \epsilon'$, then $d = y_A - x$ is an $(\omega, \omega w^{\mathrm{I}} \cdot x + \epsilon')$-$\max$-subgradient of $l_A(x; .)$ at α_A.*

We can bound the Lipschitz constant of $l_A(x; .)$ and $\widehat{l}_A(x; .)$ by $K' = \sqrt{m}$, since $x_S, y_{A,S} \leq 1$. Since $\alpha_A \in \mathcal{D}_A$, the ratio of costs in the two stages in the 2-stage problem $f_A(\mathbf{0}; \alpha_A)$ is at most $\frac{\lambda^2}{\rho}$.

Set $\gamma = \varepsilon$ and $\epsilon' = \frac{\epsilon}{8}$. Set $N = \log\big(\frac{2K'R'}{\epsilon'}\big)$ and $\omega = \frac{\gamma}{8N}$. Observe that $\log(K'R')$ is polynomially bounded. Recall that $\widehat{\alpha}_A$ is an optimal solution to $\widehat{LD}_{A;\rho}(x)$. Let G be the extended $\frac{\epsilon'}{KN\sqrt{m}}$-grid of $\mathcal{D}_A$ and $n = |G|$. By Theorem 3.5, if we use $\mathcal{T}(\epsilon, \rho, \varepsilon, \delta) = \mathsf{poly}\big(\mathcal{I}, \frac{\lambda^2}{\rho}, \frac{\lambda}{\omega}, \ln(\frac{2N}{\epsilon}), \ln(\frac{n}{\delta})\big)$ samples from $\mathcal{B}_A$ to construct $\widehat{LD}_{A;\rho}(x)$, then with probability at least $1 - \frac{\delta}{n}$, at a given point $\alpha_A \in \mathcal{D}_A$, any optimal solution $\widehat{y}_A$ to $\widehat{f}(\mathbf{0}; \alpha_A)$ satisfies $g(\alpha_A; \widehat{y}_A) \leq \big(1 + \frac{\omega}{\lambda}\big)f_A(\mathbf{0}; \alpha_A) + \frac{\epsilon}{2N}$. So by applying Lemma 4.8 and the union bound over all points in G, with probability at least $1-\delta$, at each point $\alpha_A \in G$, the $\max$-subgradient $\widehat{y}_A - x$ of $\widehat{l}_A(x; .)$ at α_A is an $(\omega, \omega w^{\mathrm{I}} \cdot x + \frac{\epsilon}{2N})$-$\max$-subgradient of $l_A(x; .)$ at α_A. By Lemma 3.2, we have $l_A(x; \widehat{\alpha}_A) \geq (1-\gamma)LD_{A;\rho}(x) - 4\epsilon' - N\omega w^{\mathrm{I}} \cdot x - \frac{\epsilon}{2}$ which is at least $(1-\varepsilon)LD_{A;\rho}(x) - \varepsilon w^{\mathrm{I}} \cdot x - \epsilon$.

Since $\ln n$ and N are $\mathsf{poly}\big(\mathcal{I}, \ln(\frac{1}{\epsilon})\big)$, we get that $\mathcal{T}(\epsilon, \rho, \varepsilon, \delta) = \mathsf{poly}\big(\mathcal{I}, \frac{\lambda}{\rho\varepsilon}, \ln(\frac{1}{\epsilon}), \ln(\frac{1}{\delta})\big)$. ■

A class of solvable 3-stage programs. The above arguments can be adapted to prove an SAA bound for the following broad class of 3-stage stochastic programs.

$$\min_{x \in \mathcal{P} \subseteq \mathbb{R}^m_{\geq 0}} h(x) \;=\; w^{\mathrm{I}} \cdot x + \sum_{A \in \mathcal{A}} p_A f_A(x), \qquad \text{(3G-P)}$$

$$f_A(x) = \min_{y_A \geq \mathbf{0}} \Big\{ w^A \cdot y_A + \sum_{B \in \mathcal{B}_A} q_{A,B} f_{A,B}(x, y_A) \Big\},$$

$$f_{A,B}(x, y_A) = \min_{\substack{z_{A,B} \in \mathbb{R}^m_{\geq 0} \\ s_{A,B} \in \mathbb{R}^n_{\geq 0}}} \Big\{ w^{A,B} \cdot z_{A,B} + c^{A,B} \cdot s_{A,B} :$$

$$D^{A,B} s_{A,B} + T^{A,B} z_{A,B} \;\geq\; j^{A,B} - T^{A,B}(x + y_A) \Big\},$$

where for every scenario (A, B), (a) $T^{A,B} \geq \mathbf{0}$, and (b) for every $x \in \mathcal{P}$ and $y_A \geq \mathbf{0}$, $0 \leq f_A(x), f_{A,B}(x, y_A) < +\infty$. Let $\lambda = \max_{S, A \in \mathcal{A}, B \in \mathcal{B}_A} \max\big(1, \frac{w_S^A}{w_S}, \frac{w_S^{A,B}}{w_S^A}\big)$; we assume that λ is known. Let OPT be the optimum value and $\mathcal{I}$ denote the input size. We assume that there is some R with $\ln R = \mathsf{poly}(\mathcal{I})$ such that $\mathcal{P} \subseteq B(\mathbf{0}, R)$, and further we assume that for any $x \in \mathcal{P}$ and any $A \in \mathcal{A}$, there is an optimal solution to $f_A(x)$ lying in $B(\mathbf{0}, R)$. These assumptions are fairly mild and unrestrictive; in particular, they hold trivially for the fractional relaxations of 0-1 integer programs and many combinatorial optimization problems. We obtain the guarantee stated in Theorem 4.9 below; as before, we can use Lemma 4.7 to convert this to a $(1+\kappa)$-guarantee under the mild assumption that $x = \mathbf{0} \in \mathcal{P}$, and for every scenario (A, B) either $f_{A,B}(x, y_A)$ is minimized at $x = y_A = \mathbf{0}$ or for every $x \in \mathcal{P}$ and $y_A \geq \mathbf{0}$, $w^{\mathrm{I}} \cdot x + w^A \cdot y_A + f_{A,B}(x, y_A) \geq 1$.

Theorem 4.9 *For any $\epsilon, \gamma > 0$, one can construct the sample average approximation $\widehat{h}$ to* (3G-P) *with* $\mathsf{poly}\big(\mathcal{I}, \lambda, \frac{1}{\gamma}, \ln(\frac{1}{\epsilon}), \ln(\frac{1}{\delta})\big)$ *samples, and with probability at least $1 - \delta$, any optimal solution $\widehat{x}$ to $\min_{x \in \mathcal{P}} \widehat{h}(x)$ satisfies $h(\widehat{x}) \leq (1 + 7\gamma) \cdot OPT + 18\epsilon$.*

Remark: Theorem 4.9 also handles programs, with constraints $T^A y_A \geq j^A - T^A x$, with $T_A \geq \mathbf{0}$, in $f_A(x)$.

5. The bound for k-stage stochastic programs

We now extend our techniques to solve k-stage stochastic linear programs. Here k is a constant that is not part of the input; the running time will be exponential in k.

In the k-stage problem, the scenario distribution is specified by a k-level tree, called the *scenario tree*. We start at the root r of this tree at level 1, which represents the first-stage. Each node u at level i represents an outcome in stage i. At a leaf node, the uncertainty has completely resolved itself and we know the input precisely. A *scenario* always refers to a stage k outcome, i.e., a leaf of the tree. The goal is to choose the first stage elements so as to minimize the total expected cost, i.e., $\sum_{i=1}^{k} \mathrm{E}\big[\text{stage } i \text{ cost}\big]$ where the expectation is taken over all scenarios. Let $\mathsf{child}(u)$ be the set of all children of u, $\mathsf{path}(u)$ be the set of all ancestors (including u) of u. Let p_u be the probability that outcome u occurs, and q_u be the *conditional probability* that u occurs given the previous-stage outcome. *The distribution can be arbitrary*, and can incorporate correlation effects from previous stages. Let w^u denote the costs in outcome u, y_u be the decisions taken in outcome u and $\mathbf{y}_v = \big(y_r, \ldots, y_v\big)$, where $\{r, \ldots, v\} = \mathsf{path}(v)$ and v is u's parent, be all the decisions taken in the previous stages. Let $x \equiv y_r \equiv \mathbf{y}_r$ for the root r. We consider the following generic k-stage LP.

$$f_{k,r} = \min_{x \in \mathcal{P}} \Big(h(x) = w^{\mathrm{I}} \cdot x + \sum_{u \in \mathsf{child}(r)} q_u f_{k-1,u}(x) \Big), \quad (k\text{G-P})$$

$$f_{k-i+1,u}(\mathbf{y}_v) = \min_{y_u \geq \mathbf{0}} \Big\{ w^u \cdot y_u + \sum_{u' \in \mathsf{child}(u)} q_{u'} f_{k-i,u'}(\mathbf{y}_v, y_u) \Big\},$$

$$\text{for } u \text{ at level } i,\; 2 \leq i < k,$$

$$f_{1,u}(\mathbf{y}_v) = \min \Big\{ w^u \cdot y_u + c^u \cdot s_u : y_u \in \mathbb{R}^m_{\geq 0},\; s_u \in \mathbb{R}^n_{\geq 0},$$

$$D^u s_u + T^u y_u \geq j^u - \sum_{t \in \mathsf{path}(v)} T^u y_t \Big\}.$$

We need that for every u with parent v and feasible decisions $\mathbf{y}_v$, $T^u \geq \mathbf{0}$ and $0 \leq f_{k-i+1,u}(\mathbf{y}_v) < \infty$, and that there is some R with $\ln R$ polynomially bounded such that each $f_{k-i+1,u}(\mathbf{y}_v)$, $i < k$, has an optimal solution in $B(\mathbf{0}, R)$. Let $\mathcal{I}$ be the input size, λ be the ratio $\max\big(1, \max_{v, u \in \mathsf{child}(v)} \frac{w^u}{w^v}\big)$. The sample average problem is of the same form as (kG-P), where the probability q_u is estimated by the frequency $\widehat{q}_u$ of outcome u in the appropriate sampled set.

The SAA bound for programs of the form (kG-P) follows by induction. Theorem 3.5 supplies the base case; the induction step, where we show that an SAA bound for $(k-1)$-stage programs of the form $f_{k-1,r}$ yields an SAA bound for $f_{k,r}$, follows by dovetailing the arguments in Section 4. As in the 3-stage case, we formulate a concave-maximization problem $LD_{k-1,u}(x)$ that is dual to $f_{k-1,u}(x)$, which is a $\max$-$\min$ problem with a $(k-1)$-stage primal problem embedded inside it. Using the induction hypothesis, we argue that the objective functions of the (slightly modified) true and sample-average dual programs are close in terms of their $\max$-subgradients, so that any optimal solution to the (modified) sample-average dual is a near-optimal solution to the (modified) true dual. This in turn allows us to show (essentially) the closeness in subgradients of the sample average and true functions. Lemma 3.1 completes the induction step. We obtain the following result, which also yields a $(1+\kappa)$-guarantee under some mild assumptions.

Theorem 5.1 *For any $\epsilon, \gamma > 0$, with probability $1-\delta$, any optimal solution $\widehat{x}$ to the sample average problem constructed using* $\mathsf{poly}\big(\mathcal{I}, \frac{\lambda}{\gamma}, \ln(\frac{1}{\delta\epsilon})\big)$ *samples satisfies $h(\widehat{x}) \leq (1+\gamma)\cdot f_{k,r} + \epsilon$.*

6. Applications

We consider a number of k-stage stochastic optimization problems, where k is a constant, for which we prove the first known performance guarantees. Our algorithms do not assume anything about the distribution or the cost structure of the input. Previously, algorithms for these problems were known only in the 2-stage setting initially in restricted settings [10, 8, 5], and later without any restrictions [13]. For each of these k-stage problems, we can write a k-stage LP for the linear relaxation of the problem for which Theorem 5.1 applies, and round the near-optimal fractional solution obtained by solving the sample average problem using an extension of the rounding scheme in [13].

Multicommodity flow We consider a stochastic version of the concurrent multicommodity flow problem where we have to buy capacity to install on the edges so that one can concurrently ship demand of each commodity i from its source s_i to its sink t_i. The demand is uncertain and is revealed in k-stages. We can buy capacity on edge e in any stage i outcome u at a cost of c_e^u; and the total amount of capacity that we can install on an edge is limited by its capacity Γ_e. The goal is to minimize the expected capacity installation cost. We obtain a $(1+\epsilon)$-optimal solution.

Covering problems We consider the k-stage versions of set cover, vertex cover and the multicut problem on trees. In each of these problems, there are elements in some universe that need to covered by sets. In the k-stage stochastic problem, the target set of elements to cover is determined by a probability distribution, and becomes known after a sequence of k stages. In each outcome u, we can purchase a set S at a price of c_S^u. We have to determine which sets to buy in stage I so as to minimize the (expected) total cost of buying sets. We can generalize the rounding theorem in [13] to show that a ρ-approximation algorithm for the deterministic analogue, that uses the natural LP relaxation as a lower bound, yields a $(k\rho+\epsilon)$-approximation algorithm for the k-stage problem. In general, to compute the decisions in a stage i outcome, we solve a $(k-i+1)$-stage problem, and round the solution. We get performance guarantees of $(k\log n+\epsilon)$ for k-stage set cover, and $(2k+\epsilon)$ for k-stage vertex cover and k-stage multicut on trees.

Facility location problems In the k-stage uncapacitated facility location (UFL) problem, we are given some candidate facility locations, a set of clients, and a probability distribution on the client demands that evolves over k-stages. In each stage, one can buy facilities paying a certain facility opening cost; in stage k, we know the exact demands and we have to assign each client's demand to an open facility incurring a client assignment cost. The goal is to minimize the expected total cost. Adapting the procedure in [13], we obtain a $O(k)$-approximation algorithms for k-stage UFL, and k-stage UFL with penalties, or with soft capacities.

References

[1] K. A. Ariyawansa and A. J. Felt On a new collection of stochastic linear programming test problems. *INFORMS Journal on Computing*, 16(3):291–299, 2004.

[2] J. R. Birge and F. V. Louveaux. *Introduction to Stochastic Programming*. Springer-Verlag, NY, 1997.

[3] M. Charikar, C. Chekuri, and M. Pál. Sampling bounds for stochastic optimization. *Proc. RANDOM*, 2005. To appear.

[4] S. Dye, L. Stougie, and A. Tomasgard. The stochastic single resource service-provision problem. *Naval Research Logistics*, 50(8):869–887, 2003. Also appeared as COSOR-Memorandum 99-13, Dept. of Mathematics and Computer Sc., Eindhoven, Tech. Univ., Eindhoven, 1999.

[5] A. Gupta, M. Pál, R. Ravi, and A. Sinha. Boosted sampling: approximation algorithms for stochastic optimization. *Proc. 36th STOC*, pages 417–426, 2004.

[6] A. Gupta, M. Pál, R. Ravi, & A. Sinha. What about Wednesday? Approximation algorithms for multistage stochastic optimization. *Proc. 8th APPROX*, 2005. To appear.

[7] A. Hayrapetyan, C. Swamy, and É. Tardos Network design for information networks. *Proc. 16th SODA*, 933–942, 2005.

[8] N. Immorlica, D. Karger, M. Minkoff, and V. Mirrokni. On the costs and benefits of procrastination: approximation algorithms for stochastic combinatorial optimization problems. *Proc. 15th SODA*, pages 684–693, 2004.

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

[10] R. Ravi and A. Sinha. Hedging uncertainty: approximation algorithms for stochastic optimization problems. *Proc. 10th IPCO*, pages 101–115, 2004.

[11] A. Shapiro. Monte Carlo sampling methods. In A. Ruszczynski and A. Shapiro, editors, *Stochastic Programming*, volume 10 of *Handbooks in Oper. Res. and Management Sc.*, North-Holland, Amsterdam, 2003.

[12] A. Shapiro. On complexity of multistage stochastic programs. *Optimization Online*, 2005. http://www.optimization-online.org/DB_FILE/2005/01/1041.pdf.

[13] D. B. Shmoys and C. Swamy. Stochastic optimization is (almost) as easy as deterministic optimization. *Proc. 45th FOCS*, pages 228–237, 2004.

[14] C. Swamy and D. Shmoys. The sample average approximation method for 2-stage stochastic optimization. November 2004. http://ist.caltech.edu/~cswamy/papers/SAAproof.pdf.