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

Chaitanya Swamy† David B. Shmoys‡

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*. Multi-stage stochastic programming and in particular, multi-stage stochastic linear programs with full recourse, is a domain that has received a great deal of attention within the Operations Research community, but mostly from the perspective of computational results in application settings.

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 approximation algorithms for a variety of k-stage generalizations of basic combinatorial optimization problems including the set cover, vertex cover, multicut on trees, facility location, and multicommodity flow 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) [3]. 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

*A preliminary version [15] will appear in the Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science, 2005.

†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 grants CCF-0430682, DMI-0500263.

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 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. It is easy to imagine the constraints: 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 this result, we also give the first approximation algorithms for the analogous class of multi-stage stochastic integer programming problems, 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 [14], 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[14] 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., [12] 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, convergence rate results that bound the sample size required by the SAA method were proved in [10]. But the bound proved in [10] 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 [10] 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 *any* 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 [13] 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 [13] 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 integer and linear programming problems, 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 an early version of our work [16], Charikar, Chekuri and Pál [4] 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 [5, 11, 9]; Gupta, Pál, Ravi, and Sinha [6] 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 [14] showed that one could derive approximation algorithms for most of the stochastic integer programming problems considered in [5, 11, 9, 6] 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 [14]. The only other work on multi-stage problems in the black-box model is due to Hayrapetyan, Swamy, and Tardos [8], and Gupta et al. [7] (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 version of the facility location problem. (It is interesting to note that the textbook [3] gives an example of an application that is formulated as a 3-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 first state some basic definitions and facts that we will frequently use. Let $\|u\|$ denote the ℓ_2 norm of u. We say that a function $g : \mathbb{R}^m \mapsto \mathbb{R}$, has *Lipschitz constant* (at most) K if $|g(v) - g(u)| \leq K\|v-u\|$ for all $u, v \in \mathbb{R}^m$.

Definition 2.1 *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 $v \in \mathbb{R}^m$. We say that $\widehat{d}$ is an $(\omega, \Delta, \mathcal{D})$-subgradient of g at the point $u \in \mathcal{D}$ 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 of an $(\omega, \Delta, \mathcal{D})$-subgradient is slightly weaker than the notion of an $(\omega, \mathcal{D})$-subgradient as defined in [14] where one requires $g(v) - g(u) \geq \widehat{d} \cdot (v-u) - \omega g(u)$. This distinction is however superficial; one could also implement the algorithm in [14] using the notion of an approximate subgradient given by Definition 2.1.

We will consider convex minimization problems $\min_{x \in \mathcal{P}} g(x)$ where $\mathcal{P} \subseteq \mathbb{R}^m_{\geq 0}$ is a polytope and $g(.)$ is convex. It is well known (see [2]) that a convex function has a subgradient at every point. The following claim will be useful in bounding the Lipschitz constant of the functions encountered.

Claim 2.2 *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.*

Proof : Consider any two points $u, v \in \mathbb{R}^m$ and let d, d' denote the subgradients at u, v respectively, with $\|d\|, \|d'\| \leq K$, then we have $g(v) - g(u) \geq d \cdot (v-u) \geq -\|d\| \, \|v-u\| \geq -K\|v-u\|$, and similarly $g(u) - g(v) \geq -\|d'\| \, \|u-v\| \geq -K\|u-v\|$. ∎

We will also encounter concave maximization problems $\max_{x \in \mathcal{P}} g(x)$, where $g(.)$ is concave. Analogous to the definition of a subgradient, we define a $\max$-subgradient and an approximate version of a $\max$-subgradient.

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 abbreviate $(\omega, \Delta, \mathcal{D})$-subgradient and $(\omega, \Delta, \mathcal{D})$-$\max$-subgradient to (ω, Δ)-subgradient and (ω, Δ)-$\max$-subgradient respectively. If $\Delta = 0$, we will use $(\omega, \mathcal{D})$-subgradient and $(\omega, \mathcal{D})$-$\max$-subgradient, instead of $(\omega, \Delta, \mathcal{D})$-subgradient and $(\omega, \Delta, \mathcal{D})$-$\max$-subgradient respectively. We will frequently use $(\omega, \Delta, \mathcal{P})$-subgradients which we abbreviate and denote as (ω, Δ)-subgradients from now on. We will 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 $\Pr\left[X \in [\mu - cb, \mu + cb]\right] \geq 1 - \delta$.*

Proof : Let $Y_i = X_i + a \in [0, a + b]$ and $Y = \sum_i Y_i$. Let $\mu' = \mathrm{E}\left[Y_i\right] = \mu + a$. We have $\Pr[X > \mu + cb] = \Pr[Y > \mathrm{E}\left[Y\right](1 + cb/\mu')]$, and $\Pr[X < \mu - cb] = \Pr[Y < \mathrm{E}\left[Y\right](1 - cb/\mu')]$. 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, $\Pr[Y < \mathrm{E}\left[Y\right](1 - \nu)]$, is at most $e^{-\frac{\nu^2 s\mu'}{2(a+b)}} = e^{-\frac{(cb)^2 s}{2\mu'(a+b)}} \leq \frac{\delta}{2}$. To bound $\Pr[Y > \mathrm{E}\left[Y\right](1 + \nu)]$ we consider two cases. If $\nu > 2e - 1$, then this quantity is at most $2^{-\frac{(1+\nu)s\mu'}{a+b}}$ which is bounded by $2^{-\frac{\nu s\mu'}{a+b}} \leq \frac{\delta}{2}$. If $\nu \leq 2e - 1$, then the probability is at most $e^{-\frac{\nu^2 s\mu'}{4(a+b)}} = e^{-\frac{(cb)^2 s}{4\mu'(a+b)}} \leq \frac{\delta}{2}$. So using the union bound, $\Pr\left[X \notin [\mu - cb, \mu + cb]\right] \leq \delta$. ∎

3 The Sample Average Approximation method

Suppose that 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 the approximate distribution. For 2-stage programs, we just estimate the probability of scenario A by its frequency in the sampled set; for k-stage programs we construct an approximate k-level distribution tree by sampling repeatedly for each level: we sample $\mathcal{T}_2$ times to obtain some stage 2 outcomes, for each such outcome we sample $\mathcal{T}_3$ times from the conditional distribution given that outcome to generate some stage 3 outcomes 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 linear programs 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 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 that one cannot hope to estimate the true function value, which gives the expected scenario cost, to within a reasonable accuracy with a small (polynomial) number of samples. Essentially, the problem is that there could be extremely low-probability outcomes which contribute significantly towards the cost in the true problem, but will almost never be sampled with only a polynomial number of samples, and so they contribute nothing in the sample average function. Hence one cannot hope to estimate the true expected cost within a reasonable accuracy using polynomially many samples. 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., its 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. For a non-differentiable function, a *subgradient* provides the analogue of a gradient, and is a measure of the "slope" of the function. We identify a notion of closeness between any two functions based on their 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 Section 4 we show that closeness of subgradients is sufficient to prove the near-equivalence of the corresponding minimization (or maximization) problems. In Lemma 4.1 we show that given two functions $g, \widehat{g} : \mathbb{R}^m \mapsto \mathbb{R}$ that agree in terms of their (approximate) subgradients at points in a polytope $\mathcal{P}$ (we make this precise later), *every* optimal solution to $\min_{x \in \mathcal{P}} \widehat{g}(x)$ is a near-optimal solution $\min_{x \in \mathcal{P}} g(x)$. Some intuition about why this closeness-in-subgradient property is sufficient can be obtained by considering the ellipsoid-based algorithm for convex minimization given in [14]. This algorithm makes use of only (approximate) subgradient information about the convex function to be minimized, using at each feasible point, a subgradient or an ω-subgradient of the function to derive a cut passing through the center of the current ellipsoid and make progress. Suppose at every point $x \in \mathcal{P}$, there is a vector $\widehat{d}_x$ that is both a subgradient of $\widehat{g}(.)$ and an ω-subgradient of $g(.)$. One can then use $\widehat{d}_x$ to generate the cut at x, and thus cause the ellipsoid-based algorithm to 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 4.1 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 $\widehat{d}_x$ that is both 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 4.3 proves an analogous result for concave maximization problems using the concept of $\max$-subgradients.

The second part of the proof, where we show that the objective functions of the true multi-stage problem and the sample average problem (with polynomially many samples) satisfy this closeness-in-subgradient property, is divided into three parts. For the class of 2-stage linear programs considered in [14], this is easy to show because in both the sample average problem and the true problem, a subgradient at any point is computed by taking the expectation, according to the respective scenario distribution, of a quantity derived from the optimal solutions to the dual of the recourse LP (i.e., the LP that determines the recourse cost for a scenario), and this recourse LP is the same in both the sample average and the true problems. Thus, since the components of the subgradient vector have bounded variance [14], and the samples in the sample average problem are drawn from the original distribution, it is easy to show the closeness-in-subgradients property.

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 6 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 7 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 approximated in the sample average problem. So *the sample average problem and the true problem solve different recourse problems for each stage 2 outcome*. Like in the 2-stage case, a (approximate) subgradient is obtained form 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 could try to prove this by showing that the two dual objective functions are close in terms of their $\max$-*subgradients*. However, simply considering the (standard) 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 as we argued earlier one cannot hope to estimate the true conditional distribution using only a polynomial number of samples (because of rare scenarios that will almost never be sampled). To circumvent this problem, we introduce a novel *compact, non-linear* formulation of the dual, which turns the dual problem into a concave maximization problem with a 2-stage primal LP embedded inside it. A $\max$-subgradient of this new dual objective function can be computed by solving this 2-stage primal stochastic LP. We now use the earlier SAA theorem for 2-stage programs to show that, any optimal solution to the 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 to the sample average dual optimal solution 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.

It is useful to view the entire argument from a broader perspective. The ellipsoid-based algorithm of Shmoys and Swamy shows that one can minimize convex functions by only using only approximate subgradient information about the function. For a given class of convex functions, if one can compute these approximate subgradients 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 is sufficient to yield a near-optimal solution to the original problem. For our class of multi-stage problems, we essentially argue that ω-subgradients can be computed efficiently by sampling and averaging, and therefore it turns out that this "nice" function is precisely the sample average objective function.

4 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 $G' = \{x \in \mathcal{P} : x_i = n_i \cdot \left(\frac{\epsilon}{KN\sqrt{m}}\right),\ n_i \in \mathbb{Z} \text{ for all } i = 1, \ldots, m\}$ Set $G = G' \cup \left\{x + t(y - x), y + t(x - y) : x, y \in G',\ t = 2^{-i},\ i = 1, \ldots, N\right\}$. We call G' and G respectively, the $\frac{\epsilon}{KN\sqrt{m}}$-*grid*, and 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

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

Lemma 4.1 *Suppose g and $\widehat{g}$ are functions satisfying property* (A)*. Let $x^*, \widehat{x} \in \mathcal{P}$ be points that respectively minimize $g(.)$ and $\widehat{g}(.)$ over $\mathcal{P}$, and suppose $g(x^*) \geq 0$. Then, $g(\widehat{x}) \leq (1+\gamma)g(x^*) + 6\epsilon + 2N\Delta$.*

Proof : For ease of understanding, consider first the case when $\widehat{x} \in G'$. We will argue that there is a point x near $\widehat{x}$ such that $g(x)$ is close to $g(x^*)$, and from this it will follow that $g(\widehat{x})$ is close to $g(x^*)$. Let $\tilde{x}$ be the point in G' closest to x^*, so $\|\tilde{x} - x^*\| \leq \frac{\epsilon}{KN}$ and therefore $g(\tilde{x}) \leq g(x^*) + \epsilon$. Let $y = \widehat{x}\big(1 - \frac{1}{2^N}\big) + \big(\frac{1}{2^N}\big)\tilde{x} \in G$ 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 $y \in G$, if we consider $\widehat{d}_y$ given by property (A), it need not be that $\widehat{d}_y \cdot (\bar{x} - y) \leq 0$, so we have to argue a little differently. Note that however $\widehat{d}_y \cdot (\bar{x} - y) \leq \frac{\epsilon}{N}$, otherwise we would have $\widehat{g}(\widehat{x}) \geq \widehat{g}(\bar{x}) - \frac{\epsilon}{N} > \widehat{g}(y)$. 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 $\widehat{d}_{y_i} \cdot (y_{i-1} - y_i) = -\widehat{d}_{y_i} \cdot (\bar{x} - y_i) \geq -\frac{\epsilon}{N}$, and because $\widehat{d}_{y_i}$ 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 $g(\widehat{x}) \leq g(y_N) + 2\epsilon \leq (1+\gamma)g(x^*) + 6\epsilon + 2N\Delta$. ■

Corollary 4.2 *Let functions $g, \widehat{g}$ and points $x^*, \widehat{x} \in \mathcal{P}$ be as in Lemma 4.1. Let $x' \in \mathcal{P}$ be such that $\widehat{g}(x') \leq \widehat{g}(\widehat{x}) + \rho$. Then, $g(\widehat{x}) \leq (1+\gamma)g(x^*) + 6\epsilon + 2N\Delta + 2N\rho$.*

Proof : Let $\bar{x}$ and $\tilde{x}$ be the points in G' that are closest to $\widehat{x}$ and x^* respectively. So $\|\bar{x} - \widehat{x}\| \leq \frac{\epsilon}{KN}$ which implies that $\widehat{g}(\bar{x}) \leq \widehat{g}(\widehat{x}) + \frac{\epsilon}{N}$ and similarly $g(\tilde{x}) \leq g(x^*) + \epsilon$. For any $y \in G$, if we consider the vector $\widehat{d}_y$ given by property (A) then $\widehat{d}_y \cdot (\bar{x} - y) \leq \frac{\epsilon}{N} + \rho$, otherwise we get a contradiction. The rest of the proof is as in Lemma 4.1. ■

We prove an analogous statement for maximization problems. Recall the definition of an exact and approximate $\max$-subgradient (Definition 2.3). We say that g and $\widehat{g}$ satisfy property (B) if

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

Lemma 4.3 *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, $g(\widehat{x}) \geq (1-\gamma)g(x^*) - 4\epsilon - N\Delta$.*

Proof : The proof closely follows the proof of Lemma 4.1. Again first suppose that $\widehat{x} \in G'$. Let $\tilde{x}$ be the point in G' closest to x^*, so $g(\tilde{x}) \geq g(x^*) - \epsilon$. Let $y = \widehat{x}\big(1 - \frac{1}{2^N}\big) + \big(\frac{1}{2^N}\big)\tilde{x} \in G$ and consider the vector $\widehat{d}_y$ given by property (B). It must be that $\widehat{d}_y \cdot (\widehat{x} - y) = -\widehat{d}_y \cdot (\tilde{x} - y) \geq 0$, otherwise we would have $\widehat{g}(\widehat{x}) < \widehat{g}(y)$. Since $\widehat{d}_y$ is an (ω, Δ)-$\max$-subgradient of $g(.)$ at y, we have $g(y) \geq \frac{g(\tilde{x}) - \Delta}{1+\omega} \geq (1-\gamma)g(x^*) - \epsilon - \Delta$ and since $\|\widehat{x} - y\| \leq \frac{\epsilon}{K}$, we get that $g(\widehat{x}) \geq (1-\gamma)g(x^*) - 2\epsilon - \Delta$.

Suppose $\widehat{x} \notin G'$. Let $\bar{x}$ be the point in G' closest to $\widehat{x}$, so $\widehat{g}(\bar{x}) \geq \widehat{g}(\widehat{x}) - \frac{\epsilon}{N}$. At any $y \in G$, the vector $\widehat{d}_y$ given by property (B), must satisfy $\widehat{d}_y \cdot (\bar{x} - y) \geq -\frac{\epsilon}{N}$, otherwise we contradict the optimality of $\widehat{x}$. 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 $\widehat{d}_{y_i} \cdot (y_{i-1} - y_i) = -\widehat{d}_{y_i} \cdot (\bar{x} - y_i) \leq \frac{\epsilon}{N}$, and because $\widehat{d}_{y_i}$ is an (ω, Δ)-$\max$-subgradient of $g(.)$ at y_i, $g(y_i) \geq g(y_{i-1})/(1+\omega) - \big(\frac{\epsilon}{N} + \Delta\big)/(1+\omega)$. This implies that $g(y_N) \geq g(\tilde{x})/(1+\omega)^N - (\epsilon + N\Delta) \geq (1-\gamma)g(x^*) - 2\epsilon - N\Delta$ So $g(\widehat{x}) \geq g(y_N) - 2\epsilon \geq (1-\gamma)g(x^*) - 4\epsilon - N\Delta$. ■

As in Corollary 4.2, we can show that an approximate maximizer of $\widehat{g}$ is also an approximate maximizer of g, but we will not need this in the sequel.

Lemma 4.4 *Let G' be the ϵ-grid of $\mathcal{P}$ and G be the corresponding extended grid. Then $|G'| \leq \left(\frac{2R}{\epsilon}\right)^m$ and $|G| \leq N|G'|^2$.*

Proof : It is clear that $|G| \leq |G'| + 2N\binom{|G'|}{2} \leq N|G'|^2$. Each grid cell of G' contains a ball of radius $r = \frac{\epsilon}{2}$ and therefore has volume at least $r^m V_m$ where V_m is the volume of the unit ball in m dimensions. The grid cells are pairwise disjoint (volume-wise), and have total volume at most $\mathsf{vol}\big(B(\mathbf{0}, R)\big) \leq R^m V_m$ since $\mathcal{P} \subseteq B(\mathbf{0}, R)$. So $|G'| \leq \left(\frac{2R}{\epsilon}\right)^m$. ■

5 The SAA bound for 2-stage stochastic programs

We now prove a polynomial bound on the number of samples required by the SAA method to solve to near-optimality the class of 2-stage stochastic programs considered in [14][1].

$$\min \quad h(x) \;=\; w^{\mathrm{I}} \cdot x + \sum_{A \in \mathcal{A}} p_A f_A(x) \quad \text{subject to} \quad x \in \mathcal{P} \subseteq \mathbb{R}^m_{\geq 0}, \tag{2Gen-P}$$

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

Here we assume that (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. To prevent an exponential blowup in the input, we consider an oracle model where an oracle supplied with scenario A reveals the scenario-dependent data $(w^A, q^A, j^A, D^A, T^A)$. Define $\lambda = \max\big(1, \max_{A \in \mathcal{A}, S} \frac{w^A_S}{w^{\mathrm{I}}_S}\big)$; we assume that λ is known. Let OPT be the optimum value, $\mathcal{I}$ denote the input size.

The sample average function is $\widehat{h}(x) = w^{\mathrm{I}} \cdot x + \sum_{A \in \mathcal{A}} \widehat{p}_A f_A(x)$ where $\widehat{p}_A = \mathcal{N}_A / \mathcal{N}$, with $\mathcal{N}$ being the total number of samples and $\mathcal{N}_A$ being the number of times scenario A is sampled. The sample average problem is $\min_{x \in \mathcal{P}} \widehat{h}(x)$. We show that with a polynomially bounded $\mathcal{N}$, $h(.)$ and $\widehat{h}(.)$ satisfy property (A) (closeness in subgradients) with high probability.

Lemma 5.1 ([14]) *Let d be a subgradient of $h(.)$ at the point $x \in \mathcal{P}$, and suppose that $\widehat{d}$ is a vector such that $\widehat{d}_S \in [d_S - \omega w^{\mathrm{I}}_S, d_S + \omega w^{\mathrm{I}}_S]$ for all S. Then $\widehat{d}$ is an ω-subgradient (i.e., an $(\omega, 0)$-subgradient) of $h(.)$ at x.*

It is shown in [14] 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 $[-\lambda w^{\mathrm{I}}_S, w^{\mathrm{I}}_S]$; and therefore (iii) $\|d_x\| \leq \lambda \|w^{\mathrm{I}}\|$. The sample average function $\widehat{h}(.)$ is of the same form as $h(.)$, only with 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 5.2 *For any $\epsilon, \gamma > 0$ ($\gamma \leq 1$), 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 only need show that property (A) holds with probability $1 - \delta$ with the stated sample size; the rest follows from Lemma 4.1. Define $N = \log\big(\frac{2KR}{\epsilon}\big)$, $\omega = \frac{\gamma}{8N}$ and the extended $\frac{\epsilon}{KN\sqrt{m}}$-grid G of

[1] This was stated in [14] with extra constraints $B^A s_A \geq h^A$, but this is equivalent to $\binom{B^A}{D^A} s_A + \binom{\mathbf{0}}{T^A} r_A \geq \binom{h^A}{j^A} - \binom{\mathbf{0}}{T^A} x$.

$\mathcal{P}$. 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 the sample average function $\widehat{h}(.)$ then at any given point x, subgradient $\widehat{d}_x$ of $\widehat{h}(.)$ is component-wise close to its expectation with probability at least $1 - dt/n$, so by Lemma 5.1, $\widehat{d}_x$ 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 4.4 to bound n, we get that $\mathcal{N} = O\big(m\lambda^2 \log^2(\frac{2KR}{\epsilon})\ln(\frac{2KRm}{\epsilon\delta})\big)$. ■

One can convert the above guarantee into a purely multiplicative $(1+\kappa)$-approximation guarantee by setting γ and ϵ appropriately, provided that we have a lower bound on OPT (that is at least inverse exponential in the input size). It was shown in [14] that under some mild assumptions, one can perform an initial sampling step to obtain such a lower bound (with high probability). We detail this lower-bounding step, which is common to 2-stage, 3-stage, and k-stage problems (differing only in the number of samples required), in Section 7.1. Using this we obtain that (under some mild assumptions) the SAA method returns a $(1+\kappa)$-optimal solution to (2Gen-P) with high probability.

6 3-stage stochastic programs

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 a stochastic version of the set cover problem, namely the 3-stage stochastic set cover problem.

6.1 An illustrative example: 3-stage stochastic set cover

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 $q_{A,B} = p_{A,B}/p_A$. 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:

$$\min \quad h(x) \;=\; \sum_S w^{\mathrm{I}}_S x_S + \sum_{A\in\mathcal{A}} p_A f_A(x) \quad \text{subject to} \quad 0 \le x_S \le 1 \quad \text{for all } S, \tag{3SSC-P}$$

$$\text{where} \quad f_A(x) \;=\; \min\Big\{\sum_S w^A_S y_{A,S} \;+\; \sum_{B\in\mathcal{B}_A} q_{A,B} f_{A,B}(x, y_A) :\; y_{A,S} \ge 0 \quad \text{for all } S\Big\}, \tag{3SSCR-P}$$

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

Let $\mathcal{P} = \{x \in \mathbb{R}^m : 0 \le x_S \le 1 \text{ for all } S\}$ and $OPT = \min_{x\in\mathcal{P}} h(x)$. The sample average problem is parametrized by (i) the sample size $\mathcal{T}_2$ used to estimate probability p_A by the frequency $\widehat{p}_A = \mathcal{T}_{2;A}/\mathcal{T}_2$, and (ii) the number of samples $\mathcal{T}_3$ 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 $\widehat{q}_{A,B} = \mathcal{T}_{3;A,B}/\mathcal{T}_3$. So the total sample size is $\mathcal{T}_2 \cdot \mathcal{T}_3$. 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 $\widehat{f}_A(x) = \min_{y_A\ge\mathbf{0}}\big(w^A\cdot y_A + \sum_{B\in\mathcal{B}_A}\widehat{q}_{A,B} f_A(x,y_A)\big)$ to denote the sample average recourse problem for outcome A, 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 function.

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 4.3 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 5.2 to prove the closeness in $\max$-subgradients of the sample average dual and the true dual. In Section 7 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, $f_A(\mathbf{0}; W) = \min_{y_A \geq \mathbf{0}} \big(W \cdot y_A + \sum_{B \in \mathcal{B}_A} q_{A,B} f_{A,B}(\mathbf{0}, y_A) \big)$. We formulate the following dual of the true and sample average recourse problems:

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

where $l_A(x; \alpha_A) = -\alpha_A \cdot x + f_A(\mathbf{0}; \alpha_A)$ and $\widehat{l}_A(x; \alpha_A) = -\alpha_A \cdot x + \widehat{f}_A(\mathbf{0}; \alpha_A)$. [2]

Lemma 6.1 *At any point $x \in \mathcal{P}$ and outcome $A \in \mathcal{A}$, $f_A(x) = LD_A(x)$ and $\widehat{f}_A(x) = \widehat{LD}_A(x)$.*

Proof : We prove that $f_A(x) = LD_A(x)$; an identical argument shows that $\widehat{f}_A(x) = \widehat{LD}_A(x)$. $f_A(x)$ can be written as the following linear program:

$$\begin{aligned}
\min \quad & \sum_S w^A_S y_{A,S} + \sum_{B \in \mathcal{B}_A} q_{A,B} w^{A,B}_S z_{A,B,S} && && \text{(SR-P)} \\
\text{s.t.} \quad & \sum_{S: e \in S} y_{A,S} + \sum_{S: e \in S} z_{A,B,S} \;\geq\; 1 - \sum_{S: e \in S} x_S && \text{for all } B \in B_A, e \in \mathcal{E}(A,B). && (1) \\
& y_{A,S}, z_{A,B,S} \geq 0 && \forall B \in \mathcal{B}_A, S.
\end{aligned}$$

Let $\big(y^*_A, \{z^*_{A,B}\}\big)$ be an optimal solution to (SR-P) and $\big(\{\beta^*_{A,B}\}\big)$ be an optimal solution to the (standard) LP dual of (SR-P) where $\beta^*_{A,B,e}$ is the dual multiplier corresponding to the inequality (1) for element $e \in \mathcal{E}(A,B)$ where $B \in \mathcal{B}_A$. Let α^*_A be an optimal solution to $LD_A(x)$. Setting $y_A = x + y^*_A$ yields a *feasible solution* to the minimization problem $f_A(\mathbf{0}; \alpha^*_A)$. So $LD_A(x)$ is at most $(y_A - x) \cdot \alpha^*_A + \sum_{B \in \mathcal{B}_A} q_{A,B} f_{A,B}(\mathbf{0}, y_A) = \alpha^*_A \cdot y^*_A + \sum_{B \in \mathcal{B}_A} q_{A,B} f_{A,B}(x, y^*_A)$ which is at most $f_A(x)$ since $\alpha^*_A \leq w_A$. For the other direction, consider the vector α_A with $\alpha_{A,S} = \sum_{B \in \mathcal{B}_A} \sum_{e \in S \cap \mathcal{E}(A,B)} \beta^*_{A,B,e}$. α_A is a feasible solution to $LD_A(x)$ since the dual of (SR-P) has $\sum_{B \in \mathcal{B}_A} \sum_{e \in S \cap \mathcal{E}(A,B)} \beta_{A,B,e} \leq w^A_S$ as a constraint for each set S. If we consider the LP dual of $f_A(\mathbf{0}; \alpha_A)$, then observe that $\big(\{\beta^*_{A,B}\}\big)$ yields a feasible solution to the dual and has value $\sum_{B \in \mathcal{B}_A} \sum_{e \in \mathcal{E}(A,B)} \beta^*_{A,B,e}$, which is therefore a lower bound on $f_A(\mathbf{0}; \alpha_A)$. Therefore we can lower bound $LD_A(x)$ by $l_A(x; \alpha_A) = -\sum_S \alpha_{A,S} x_S + \sum_{B \in \mathcal{B}_A} \sum_{e \in \mathcal{E}(A,B)} \beta^*_{A,B,e}$ which is equal to $\sum_{B \in \mathcal{B}_A} \sum_{e \in \mathcal{E}(A,B)} (1 - \sum_{S: e \in S} x_S) \beta^*_{A,B,e} = f_A(x)$ by LP duality. ∎

Lemma 6.1 proves strong duality (in this new dual representation). Using this strong duality, we show that a (approximate) subgradient to $h(.)$ at x can be computed from the (near-) optimal solutions to the dual problems $LD_A(x)$ for each outcome A.

[2] This dual representation can be obtained by adding the (redundant) constraints $x_S + y_{A,S} \geq r_S$ to $f_A(x)$, writing the objective function of $f_A(x)$ as $\sum_S w_{A,S} y_{A,S} + \sum_{B \in \mathcal{B}_A} q_{A,B} f_{A,B}(\mathbf{0}, r)$, and then taking the Lagrangian dual of the resulting program by dualizing only the $x_S + y_{A,S} \geq r_{A,S}$ constraint using $\alpha_{A,S}$ as the Lagrangian multiplier.

Lemma 6.2 *Fix $x \in \mathcal{P}$. Let α_A be a solution to $LD_A(x)$ of value $l_A(x;\alpha_A) \geq (1-\varepsilon)LD_A(x) - \varepsilon w^{\mathrm{I}} \cdot x - \epsilon$ 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 $\|d\| \leq \lambda \|w^{\mathrm{I}}\|$; (ii) if $\widehat{d}$ is a vector such that $d - \omega w^{\mathrm{I}} \leq \widehat{d} \leq d + \omega w^{\mathrm{I}}$, then $\widehat{d}$ is an $(\varepsilon + \omega, \epsilon)$-subgradient of $h(.)$ at x.*

Proof : Consider any $x' \in \mathcal{P}$. Since $l_A(x;\alpha_A) \geq (1-\varepsilon)LD_A(x) - \varepsilon w^{\mathrm{I}} \cdot x - \epsilon$ for every $A \in \mathcal{A}$, we have

$$h(x) = w^{\mathrm{I}} \cdot x + \sum_A p_A LD_A(x) \leq (1+\varepsilon) w^{\mathrm{I}} \cdot x + \sum_A p_A \Big(-\alpha_A \cdot x + f_A(\mathbf{0}; \alpha_A) + \varepsilon LD_A(x) \Big) + \epsilon.$$

At x', α_A is a feasible solution to $LD_A(x')$ for every outcome A. So $h(x') \geq w^{\mathrm{I}} \cdot x + \sum_A p_A(-\alpha_A \cdot x' + f_A(\mathbf{0}; \alpha_A))$. Subtracting we get that $h(x') - h(x)$ is at least $d \cdot (x' - x) - \varepsilon(w^{\mathrm{I}} \cdot x + \sum_A p_A LD_A(x)) - \epsilon = d \cdot (x' - x) - \varepsilon h(x) - \epsilon$. Since $\alpha_A \leq w^A \leq \lambda w^{\mathrm{I}}$, $\|d\| \leq \lambda \|w^{\mathrm{I}}\|$.

We know that $h(x') - h(x) \geq d \cdot (x' - x) - \varepsilon h(x) - \epsilon = (d - \widehat{d}) \cdot (x' - x) + \widehat{d} \cdot (x' - x) - \varepsilon h(x) - \epsilon$. Since $x_S, x'_S \geq 0$ for all S, we have $(d - \widehat{d}) \cdot x' \geq -\omega w^{\mathrm{I}} \cdot x' \geq -\omega h(x')$ and $(\widehat{d} - d) \cdot x \geq -\omega w^{\mathrm{I}} \cdot x \geq \omega h(x)$. This proves (ii). ■

Since $\widehat{h}(.)$ is of the same form as $h(.)$, Lemma 6.2 also shows that $\widehat{d}_x = w^{\mathrm{I}} - \sum_A \widehat{p}_A \widehat{\alpha}_A$ is a subgradient of $\widehat{h}(.)$ at x where $\widehat{\alpha}_A$ is an optimal solution to $\widehat{LD}_A(x)$. Thus, to prove the closeness in subgradients of h and $\widehat{h}$ it suffices to argue that any optimal solution to $\widehat{LD}_A(x)$ 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 (that is, 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 $f_A(\mathbf{0}; \alpha_A)$ (see Lemma 6.7), 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 $f_A(\mathbf{0}; \alpha_A)$. We would like to use Theorem 5.2 here, but this statement need not be true (with a polynomial sample size) since the ratio $\max_S \big(\frac{w_S^{A,B}}{\alpha_{A,S}} \big)$ 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

$$LD_{A;\rho}(x) = \max_{\rho w^{\mathrm{I}} \leq \alpha_A \leq w^A} l_A(x; \alpha_A) \quad \text{and} \quad \widehat{LD}_{A;\rho}(x) = \max_{\rho w^{\mathrm{I}} \leq \alpha_A \leq w^A} \widehat{l}_A(x; \alpha_A)$$

for a suitable $\rho > 0$. Observe that the cost ratio in the 2-stage problem $f_A(\mathbf{0}; \alpha_A)$ is bounded by $\frac{\lambda^2}{\rho}$ for any $A \in \mathcal{A}$. In Section 6.1.1, we prove the following SAA bound for the duals of the true and sample average recourse problems.

Lemma 6.3 *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 $1 - \delta$.*

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 6.2, 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. So using Lemma 6.3 we can show the closeness in subgradients of $h_\rho(.)$ and $\widehat{h}_\rho(.)$, and this will suffice to show that if $\widehat{x}$ minimizes $\widehat{h}(.)$ then it is a near-optimal solution to $h(.)$. Thus we get an SAA bound for our class of 3-stage programs.

First, in Lemma 6.4, we bound the number of samples required to ensure that at a single point $x \in \mathcal{P}'$, a subgradient of $\widehat{h}_\rho(.)$ is an (ω, ϵ)-subgradient of $h_\rho(.)$. The proof is somewhat involved because if we consider the random variable taking the value $w_S^{\mathrm{I}} - \widehat{\alpha}_{A,S}$ when outcome A is sampled, where $\widehat{\alpha}_A$ is an optimal solution to $\widehat{LD}_{A;\rho}(x)$, then the random variables corresponding to the different samples from stage 2 are not independent since we always use the same solution $\widehat{\alpha}_A$. We defer the proof till after Theorem 6.6.

Lemma 6.4 *Consider the sample average function generated using $\mathcal{N}_2 = \mathcal{T}_2(\omega,\delta) = \frac{16(1+\lambda)^2}{\omega^2}\ln\left(\frac{4m}{\delta}\right)$ samples from stage 2, and $\mathcal{T}\left(\epsilon,\rho,\frac{\omega}{2},\frac{\delta}{2\mathcal{N}_2}\right)$ 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 6.5 *For any $x \in \mathcal{P}$, $h_\rho(x) \le h(x) \le h_\rho(x) + \rho w^{\mathrm{I}} \cdot x$. Similarly $\widehat{h}_\rho(x) \le \widehat{h}(x) \le \widehat{h}_\rho(x) + \rho w^{\mathrm{I}} \cdot x$.*

Proof : We prove this for $h(.)$ and $h_\rho(.)$; the second statement is proved identically. The first inequality holds since we are maximizing over a larger feasible region in $LD_A(x)$. The second inequality follows because if α_A^* is such that $LD_A(x) = l_A(x;\alpha_A^*)$, then taking $\alpha'_A = \min(\alpha_A^* + \rho w^{\mathrm{I}}, w^A)$ gives $LD_{A;\rho}(x) \ge l_A(x;\alpha'_A) \ge l_A(x;\alpha_A^*) - \rho w^{\mathrm{I}} \cdot x$ since $f_A(\mathbf{0};\alpha_A)$ is increasing in α_A. So $h_\rho(x) \ge h(x) - \rho w^{\mathrm{I}} \cdot x$. ■

Theorem 6.6 *For any $\epsilon,\gamma > 0$ ($\gamma \le 1$), one can construct $\widehat{h}$ with $\mathrm{poly}\left(\mathcal{I},\lambda,\frac{1}{\gamma},\ln(\frac{1}{\epsilon}),\ln(\frac{1}{\delta})\right)$ samples, and with probability at least $1-\delta$, any optimal solution $\widehat{x}$ to $\min_{x\in\mathcal{P}}\widehat{h}(x)$ satisfies $h(\widehat{x}) \le (1+7\gamma)\cdot OPT + 18\epsilon$.*

Proof : Let $N = \log\left(\frac{2KR}{\epsilon}\right)$ 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 6.5, $\widehat{h}_\rho(\widehat{x}) \le \widehat{h}(\widehat{x}) \le \widehat{h}(\tilde{x}) \le \widehat{h}_\rho(\tilde{x}) + \rho w^{\mathrm{I}} \cdot \tilde{x}$, and $0 \le OPT_\rho = \min_{x\in\mathcal{P}} h_\rho(x) \le \min_{x\in\mathcal{P}} h(x) = OPT$.

Let G be the extended $\frac{\epsilon}{KN\sqrt{m}}$-grid of $\mathcal{P}$ and $n = |G|$. Let $\mathcal{N}' = \frac{16(1+\lambda)^2}{\omega^2}\ln\left(\frac{4mn}{\delta}\right)$ which is a polynomial in $\mathcal{I},\frac{\lambda}{\gamma},\ln\left(\frac{1}{\epsilon}\right)$ and $\ln\left(\frac{1}{\delta}\right)$, where we use Lemma 4.4 to bound n. We construct $\widehat{h}(.)$ using $\mathcal{N} = \mathcal{N}' \cdot \mathcal{T}\left(\epsilon',\rho,\frac{\omega}{2},\frac{\delta}{2n\mathcal{N}'}\right)$ samples. Since $\mathcal{N}'$ is polynomially bounded, Lemma 6.3 shows that so is $\mathcal{N}$. Using Lemma 6.4 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 Lemma 4.1, we have that $h_\rho(\tilde{x}) \le (1+\gamma)OPT_\rho + 6\epsilon + 2N\epsilon'$ with high probability. Since $\widehat{h}_\rho(\widehat{x}) \le \widehat{h}_\rho(\tilde{x}) + \rho w^{\mathrm{I}} \cdot \tilde{x}$, we also obtain by Corollary 4.2 that

$$h_\rho(\widehat{x}) \le (1+\gamma)OPT_\rho + 6\epsilon + 2N(\rho w^{\mathrm{I}} \cdot \tilde{x} + \epsilon'). \tag{2}$$

The bound $h(\tilde{x}) \le h_\rho(\tilde{x}) + \rho w^{\mathrm{I}} \cdot x$ (Claim 6.5) implies that $(1-\rho)w^{\mathrm{I}} \cdot \tilde{x} \le h_\rho(\tilde{x})$. Similarly $(1-\rho)h(\widehat{x}) \le h_\rho(\widehat{x})$. Combining these with the bound $OPT_\rho \le OPT$, and plugging in ϵ' and ρ in (2), we get that $h(\widehat{x}) \le (1+7\gamma)OPT + 18\epsilon$. ■

Under the very mild assumption that for every scenario (A,B) with $\mathcal{E}(A,B) \ne \emptyset$ (a "non-null" scenario), for every $x \in \mathcal{P}$ and $y_A \ge \mathbf{0}$ the total cost $w^{\mathrm{I}} \cdot x + w^A \cdot y_A + f_{A,B}(x,y_A)$ is at least 1, the sampling procedure in Section 7.1 gives a lower bound on OPT (Lemma 7.6). Thus we obtain a $(1+\kappa)$-optimal solution to (3SSC-P) with the SAA method (with high probability) using polynomially many samples.

Proof of Lemma 6.4 : Let $\delta' = \frac{\delta}{2\mathcal{N}_2}$ and $\omega' = \frac{\omega}{2}$. Observe that the sampling of outcomes from $\mathcal{A}$ only determines whether or not we sample from $\mathcal{B}_A$ but does not influence the probability of any event determined by the samples from $\mathcal{B}_A$. So, we may view the sampling process as follows: (1) for each outcome A, we independently sample from the conditional distribution on $\mathcal{B}_A$ to construct ($\widehat{f}_A(x)$ and) $\widehat{LD}_{A;\rho}(x)$; (2) we sample stage 2 outcomes from $\mathcal{A}$ to determine the probabilities $\widehat{p}_A$, which are the weights used to combine the functions $\widehat{LD}_{A;\rho}(x)$ and construct $\widehat{h}_\rho(x)$. Let Ω_2 be the probability space of all random choices involved in sampling the $\mathcal{N}_2$ stage 2 outcomes from $\mathcal{A}$, and let Ω_A be the space of all random choices involved in sampling from $\mathcal{B}_A$. So the entire probability space is $\Omega = \Omega_2 \times \prod_{A\in\mathcal{A}}\Omega_A$.

Let $Z_{A,i}$ be 1 if the i^{th} sample results in outcome A and 0 otherwise. Let O_A^* be the set of all solutions α_A to $LD_{A;\rho}(x)$ satisfying $l_A(x;\alpha_A) \geq (1-\omega')LD_A(x) - \omega' w^{\mathrm{I}} \cdot x - \epsilon$. Define the random vector Ψ_A to be an optimal solution (breaking ties arbitrarily) to $\widehat{LD}_{A;\rho}(x)$. Let $G_A \subseteq \Omega_A$ be the event that $\Psi_A \in O_A^*$. By Lemma 6.3, we know that $\Pr_{\Omega_A}[G_A] \geq 1-\delta'$, where for clarity we use the subscript to indicate that the probability is wrt. the space Ω_A. We may assume without loss of generality that this probability is exactly $1-\delta'$ since we can simply choose $G_A \subseteq \Omega_A$ so that this holds. Let $G_i \subseteq \Omega = \bigcup_{A\in\mathcal{A}}(\{Z_{A,i} = 1\} \times G_A \times \prod_{A'\in\mathcal{A},A'\neq A} \Omega_{A'})$. So G_i is the event representing "if A is the stage 2 outcome generated by the i^{th} sample then event G_A occurs". We have $\Pr[G_i] = 1-\delta'$. We will condition on the event $\mathcal{G} = \bigcap_i G_i$. Note that $\Pr[\mathcal{G}] \geq 1-\delta/2$. For each component S of x, define $X_{i,S} = \sum_{A\in\mathcal{A}} Z_{A,i}(w_S^{\mathrm{I}} - \Psi_{A,S})$ and $X_S = \left(\sum_{i=1}^{\mathcal{N}_2} X_{i,S}\right)/\mathcal{N}_2$. The subgradient $\widehat{d}_x$ is the random vector X. We argue that conditioned on $\mathcal{G}$, with probability at least $1-\delta/2$, there exist solutions $\alpha_A \in O_A^*$ for every A, such that for every component S, $|X_S - \sum_A p_A(w_S^{\mathrm{I}} - \alpha_{A,S})| \leq \omega' w_S^{\mathrm{I}}$. Therefore conditioned on $\mathcal{G}$, by Lemma 6.2, X is an $(2\omega',\epsilon) = (\omega,\epsilon)$-subgradient of $h_\rho(.)$ at x with probability $1-\delta/2$; since $\Pr[\mathcal{G}] \geq 1-\delta/2$, with probability at least $1-\delta$, $\widehat{d}_x = X$ is an (ω,ϵ)-subgradient of $h_\rho(.)$ at x.

Select some solution $\overline{\alpha}_A \in O_A^*$ for each outcome A. We have to show that $\Pr[E \mid \mathcal{G}] \leq \delta/2$ where E is the bad event $\left\{\neg\left[\exists \alpha = (\alpha_A)_{A\in\mathcal{A}} \in \prod_{A\in\mathcal{A}} O_A^* \text{ such that } \forall S,\ |X_S - \sum_A p_A(w_S^{\mathrm{I}} - \alpha_{A,S})| \leq \omega' w_S^{\mathrm{I}}\right]\right\}$. Note that although the variables $Z_{A,i}, i = 1,\ldots,\mathcal{N}_2$ are independent, the $X_{i,S}$ variables for $i = 1,\ldots,\mathcal{N}_2$ are *not independent* because they are coupled by the $\Psi_{A,S}$ variables. But if we condition on the Ψ_A variables, the $X_{i,S}$ variables do become independent. Let $\Psi'_A = \Psi_A$ if $\Psi_A \in O_A^*$ and $\overline{\alpha}_A$ otherwise. Conditioning on $\overline{\Psi} = (\Psi_A)_{A\in\mathcal{A}}$, we have

$$\begin{aligned}
\Pr\left[E \;\middle|\; \mathcal{G}, \overline{\Psi}\right] &\leq \Pr\left[\exists S \text{ s.t. } \left|X_S - \textstyle\sum_A p_A(w_S^{\mathrm{I}} - \Psi'_{A,S})\right| > \omega' w_S^{\mathrm{I}} \;\middle|\; \mathcal{G}, \overline{\Psi}\right] \\
&\leq \sum_S \Pr\left[\left|X_S - \textstyle\sum_A p_A(w_S^{\mathrm{I}} - \Psi'_{A,S})\right| > \omega' w_S^{\mathrm{I}} \;\middle|\; \mathcal{G}, \overline{\Psi}\right] \qquad (3)
\end{aligned}$$

where the first inequality follows since event E implies that given the solutions $\left\{\Psi'_A\right\}_{A\in\mathcal{A}}$, there exists some component S such that $|X_S - \sum_A p_A \Psi'_{A,S}| > \omega' w_S^{\mathrm{I}}$. Since we have conditioned on $\mathcal{G}$, if $\Psi_A \notin O_A^*$ it follows that $\sum_i Z_{A,i} = 0$. Therefore we can write $X_S = \left(\sum_{i=1}^{\mathcal{N}_2} Y_{i,S}\right)/\mathcal{N}_2$ where $Y_{i,S} = \sum_{A\in\mathcal{A}} Z_{A,i}(w_S^{\mathrm{I}} - \Psi'_{A,S})$. The variables $Y_{i,S}$ are iid, so by Lemma 2.4, $\Pr\left[|X_S - \sum_A p_A(w_S^{\mathrm{I}} - \Psi'_{A,S})| > \omega' w_S^{\mathrm{I}} \;\middle|\; \mathcal{G}, \overline{\Psi}\right] \leq \delta/2m$, and using (3), we have $\Pr\left[E \;\middle|\; \mathcal{G}, \overline{\Psi}\right] \leq \delta/2$. Since this holds for every $\overline{\Psi}$, this also holds if we remove the conditioning on $\overline{\Psi}$. Therefore $\Pr[E \mid \mathcal{G}] \leq \delta/2$ which completes the proof. ■

6.1.1 An SAA bound for $\widehat{LD}_{A;\rho}(x)$

We now prove Lemma 6.3. Throughout this section ϵ, ε and ρ are fixed parameters given by the statement of Lemma 6.3. Let $\mathcal{D}_A = \{\alpha_A \in \mathbb{R}^m : \rho w^{\mathrm{I}} \leq \alpha_A \leq \omega^A\}$. Recall that the (true) dual problem $LD_{A;\rho}(x)$ is to maximize $l_A(x;\alpha_A)$ over the region $\mathcal{D}_A$ where $l_A(x;\alpha_A) = -\alpha_A \cdot x + f_A(\mathbf{0};\alpha_A)$. In the sample average dual problem $\widehat{LD}_{A;\rho}(x)$, we have $\widehat{l}_A(x;\alpha_A) = -\alpha_A \cdot x + \widehat{f}_A(\mathbf{0};\alpha_A)$ instead of $l_A(x;\alpha_A)$. 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 4.3. 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 the $\mathcal{D}_A$. A max-subgradient to $l_A(x;.)$ (respectively $\widehat{l}_A(x;.)$) at α_A is obtained from the solution to the 2-stage problem $f_A(\mathbf{0};\alpha_A)$ (respectively $\widehat{f}_A(\mathbf{0};\alpha_A)$).

Lemma 6.7 *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.*

Proof : Let $C = \sum_{B \in \mathcal{B}_A} q_{A,B} f_{A,B}(\mathbf{0}, y_A)$. So $\alpha_A \cdot y_A + C \leq (1+\omega')f_A(\mathbf{0}; \alpha_A) + \epsilon'$ and $l_A(x; \alpha_A) = -\alpha_A \cdot x + f_A(\mathbf{0}; \alpha_A) \geq (y_A - x) \cdot \alpha_A + C - \omega' \cdot f_A(\mathbf{0}; \alpha_A) - \epsilon'$. At any other point α'_A, y_A gives a feasible solution to the 2-stage problem $f_A(\mathbf{0}; \alpha'_A)$. So $l_A(x; \alpha'_A) \leq (y_A - x) \cdot \alpha'_A + C$. Subtracting we get that

$$l_A(x; \alpha'_A) - l_A(x; \alpha_A) \leq d \cdot (\alpha'_A - \alpha_A) + \omega' \cdot f_A(\mathbf{0}; \alpha_A) + \epsilon' \leq d \cdot (\alpha'_A - \alpha_A) + \omega' \cdot l_A(x; \alpha_A) + \epsilon' + \omega' \alpha_A \cdot x.$$

The last term is at most $\omega w^{\mathrm{I}} \cdot x$ since $\alpha_A \leq w^A \leq \lambda w^{\mathrm{I}}$ and $x \geq \mathbf{0}$. Thus d is an $(\omega, \omega w^{\mathrm{I}} \cdot x + \epsilon')$- max-subgradient. ■

We can bound the Lipschitz constant of $l_A(x; .)$ and $\widehat{l}_A(x; .)$ by $K' = \sqrt{m}$, since $x_S, y_{A,S} \leq 1$. The feasible region of the 2-stage problem $f_A(\mathbf{0}; \alpha_A)$ is contained in the ball $B(\mathbf{0}, \sqrt{m})$, and since $\alpha_A \in \mathcal{D}_A$, the ratio of costs in the two stages is at most $\frac{\lambda^2}{\rho}$. Thus, we can use Theorem 5.2 to argue that any optimal solution $\widehat{y}_A$ to $\widehat{f}_A(\mathbf{0}; \alpha_A)$ is a near-optimal solution to $f_A(\mathbf{0}; \alpha_A)$, and this will prove the closeness in max-subgradients of $\widehat{l}_A(x; .)$ and $l_A(x; .)$.

Proof of Lemma 6.3 : 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 5.2, 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 6.7 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 4.3, 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 $\mathcal{T}(\epsilon, \rho, \varepsilon, \delta) = \mathsf{poly}\big(\mathcal{I}, \frac{\lambda}{\rho\varepsilon}, \ln(\frac{1}{\epsilon}), \ln(\frac{1}{\delta})\big)$. ■

6.2 A class of solvable 3-stage programs

The above arguments can be adapted to prove an SAA bound for a broad class of 3-stage stochastic programs, which includes the 3-stage stochastic set cover problem considered above. As before, 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, and x, y_A and $z_{A,B}$ respectively to denote the decisions in stage 1, outcome A and scenario (A, B) respectively. $\mathcal{A}$ denotes 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 $q_{A,B} = p_{A,B}/p_A$. We consider the following class of 3-stage problems.

$$\min \quad h(x) \;=\; w^{\mathrm{I}} \cdot x + \sum_{A \in \mathcal{A}} p_A f_A(x) \quad \text{subject to} \quad x \in \mathcal{P} \subseteq \mathbb{R}^m_{\geq 0}, \tag{3Gen-P}$$

$$\text{where} \quad f_A(x) \;=\; \min_{y_A \in \mathbb{R}^m_{\geq 0}} \Big\{ w^A \cdot y_A + \sum_{B \in \mathcal{B}_A} q_{A,B} f_{A,B}(x, y_A) : T^A y_A \geq j^A - T^A x \Big\}, \quad \text{and} \tag{3Rec-P}$$

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

where for every outcome $A \in \mathcal{A}$ and scenario (A, B), (a) $T^A, T^{A,B} \geq \mathbf{0}$; (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 $\lambda = \max_{S, A \in \mathcal{A}, B \in \mathcal{B}_A} \max\big(1, \frac{w^A_S}{w_S}, \frac{w^{A,B}_S}{w^A_S}\big)$; we assume that λ is known.

As before, we assume that $\mathcal{P} \subseteq B(\mathbf{0}, R)$ where $\ln R$ is polynomially bounded. Further we assume that for any $x \in \mathcal{P}$ and any $A \in \mathcal{A}$, the feasible region of $f_A(x)$ can be restricted to $B(\mathbf{0}, R)$ without affecting the solution quality, that is, 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. Let OPT be the optimum value and $\mathcal{I}$ be the input size.

The sample average problem is of the same form as (3Gen-P), where p_A and $q_{A,B}$ are replaced by their estimates $\widehat{p}_A$ and $\widehat{q}_{A,B}$ respectively, the frequencies of occurrence of outcome A and scenario (A, B) in the appropriate sampled sets. Let $\widehat{h}(x) = w^{\mathrm{I}} \cdot x + \sum_{A \in \mathcal{A}} \widehat{p}_A \widehat{f}_A(x)$ denote the sample average function where

$$\widehat{f}_A(x) \;=\; \min_{y_A \geq \mathbf{0}} \Big\{ w^A \cdot y_A + \sum_{B \in \mathcal{B}_A} \widehat{q}_{A,B} f_{A,B}(x, y_A) : T^A y_A \geq j^A - T^A x \Big\} \qquad \text{(3SARec-P)}$$

is the sample average recourse problem.

Let $f_A(\mathbf{0}; W)$ (respectively $\widehat{f}_A(\mathbf{0}; W)$) denote the recourse problem (3Rec-P) (respectively (3SARec-P)) with $x = \mathbf{0}$ and costs $w^A = W$. The dual of the recourse problem is formulated as before, $LD_A(x) = \max_{\mathbf{0} \leq \alpha_A \leq w^A} l_A(x; \alpha_A)$ where $l_A(x; \alpha_A) = -\alpha_A \cdot x + f_A(\mathbf{0}; \alpha_A)$. We use $\widehat{LD}_A(x)$ and $\widehat{l}_A(x; \alpha_A)$ to denote the corresponding quantities for the sample average problem.

The only portion of the argument in Section 6.1 that needs to be modified is the proof of Lemma 6.1 which proves strong duality in the new dual representation. The proof is along the same lines.

Lemma 6.8 *At any point $x \in \mathcal{P}$ and outcome $A \in \mathcal{A}$, $f_A(x) = LD_A(x)$ and $\widehat{f}_A(x) = \widehat{LD}_A(x)$. Moreover, we can restrict y_A so that $\|y_A\| \leq 2R$ in the problems $f_A(0; \alpha_A)$ and $\widehat{f}_A(\mathbf{0}; \alpha_A)$, without affecting the values of $LD_A(x)$ and $\widehat{LD}_A(x)$.*

Proof : We prove that $f_A(x) = LD_A(x)$; an identical argument shows that $\widehat{f}_A(x) = \widehat{LD}_A(x)$. $f_A(x)$ can be written as the following linear program:

$$\begin{aligned}
\min \quad & w^A \cdot y_A \;+\; \sum_{B \in \mathcal{B}_A} q_{A,B}(w^{A,B} \cdot z_{A,B} \;+\; c^{A,B} \cdot s_{A,B}) && && \text{(R-P)} \\
\text{s.t.} \quad & T^A y_A \;\geq j^A - T^A x && && \\
& D^{A,B} s_{A,B} \;+\; T^{A,B} y_A \;+\; T^{A,B} z_{A,B} \;\geq j^{A,B} - T^{A,B} x && \forall B \in \mathcal{B}_A, && (4) \\
& s_{A,B} \in \mathbb{R}^n, \quad s_{A,B}, y_A, z_{A,B} \;\geq \mathbf{0} && \forall B \in \mathcal{B}_A. &&
\end{aligned}$$

Let $\big(y_A^*, \{s_{A,B}^*, z_{A,B}^*\}\big)$ be an optimal solution to (R-P) and $\big(\theta_A^*, \{\beta_{A,B}^*\}\big)$ be an optimal solution to the (standard, LP) dual of (R-P) where $\beta_{A,B}$ is the dual multiplier corresponding to inequalities (4) for each $B \in \mathcal{B}_A$. Let α_A^* be an optimal solution to $LD_A(x)$. Setting $y_A = x + y_A^*$ yields a *feasible solution* to the minimization problem $f_A(\mathbf{0}; \alpha_A^*)$. So $LD_A(x)$ is at most $(y_A - x) \cdot \alpha_A^* + \sum_{B \in \mathcal{B}_A} q_{A,B} f_{A,B}(\mathbf{0}, y_A) = \alpha_A^* \cdot y_A^* + \sum_{B \in \mathcal{B}_A} q_{A,B} f_{A,B}(x, y_A^*)$ which is at most $w^A \cdot y_A^* + \sum_{B \in \mathcal{B}_A} q_{A,B} f_{A,B}(x, y_A^*) = f_A(x)$. For the other direction, consider the solution $\alpha_A = (T^A)^{\mathrm{T}} \theta_A^* + \sum_{B \in \mathcal{B}_A} (T^{A,B})^{\mathrm{T}} \beta_{A,B}^*$. This is a feasible solution to $LD_A(x)$ since the dual of (R-P) has $(T^A)^{\mathrm{T}} \theta_A + \sum_{B \in \mathcal{B}_A} (T^{A,B})^{\mathrm{T}} \beta_{A,B} \leq w^A$ as a constraint. If we consider the LP dual of $f_A(\mathbf{0}; \alpha_A)$, then observe that $\big(\theta_A^*, \beta_{A,B}^*\big)$ yields a feasible solution to the dual that has value $j^A \cdot \theta_A^* + \sum_{B \in \mathcal{B}_A} j^{A,B} \cdot \beta_{A,B}^*$. Therefore we can lower bound $LD_A(x)$ by $-\alpha_A \cdot x + j^A \cdot \theta_A^* + \sum_{B \in \mathcal{B}_A} j^{A,B} \cdot \beta_{A,B}^*$ which is equal to $(j^A - T^A x) \cdot \theta_A^* + \sum_{B \in \mathcal{B}_A} (j^{A,B} - T^{A,B} x) \cdot \beta_{A,B}^* = f_A(x)$ by LP duality.

Notice that the upper-bound argument also holds if we restrict y_A to lie in the ball $B(\mathbf{0}, 2R)$ in the problem $f_A(\mathbf{0}; \alpha_A)$ embedded in the dual problem $LD_A(x)$, that is, $\max_{\mathbf{0} \leq \alpha_A \leq w_A} \big(-\alpha_A \cdot x + f'_A(\mathbf{0}; \alpha_A)\big) \leq f_A(x)$ where $f'_A(\mathbf{0}; \alpha_A)$ is the same as $f_A(\mathbf{0}; \alpha_A)$ except that we restrict y_A to lie in $B(\mathbf{0}, 2R)$. Since this restriction can only increase the value of the minimization problem, $f'_A(\mathbf{0}; \alpha_A) \geq f(\mathbf{0}; \alpha_A)$, and so

$\max_{\mathbf{0} \le \alpha_A \le w_A} \big(-\alpha_A \cdot x + f'_A(\mathbf{0}; \alpha_A)\big) \ge LD_A(x) = f_A(x)$. This shows that we may assume $\|y_A\| \le 2R$ in the problem $f_A(\mathbf{0}; \alpha_A)$ (respectively $\widehat{f}_A(\mathbf{0}; \alpha_A)$) without changing the value of $LD_A(x)$ (respectively $\widehat{LD}_A(x)$). ■

It need not be true that for an arbitrary cost vector $\alpha_A, \mathbf{0} \le \alpha_A \le w_A$, there exists an optimal solution to $f_A(\mathbf{0}; \alpha_A)$ which lies in $B(\mathbf{0}, 2R)$. However, since $f_A(\mathbf{0}; \alpha_A)$ (respectively $\widehat{f}_A(\mathbf{0}; \alpha_A)$) is only "used" while embedded in the maximization problem $LD_A(x)$ (respectively $\widehat{LD}_A(x)$), and by Lemma 6.8 its value is not affected by imposing the constraint $\|y_A\| \le 2R$, we will assume that this constraint is implicitly included in $f_A(\mathbf{0}; \alpha_A)$, and this will not affect the validity of our arguments. That is, when we say $f_A(\mathbf{0}; \alpha_A)$ we actual mean the minimization problem $\min_{y_A \ge \mathbf{0}: \|y_A\| \le 2R} \big\{\alpha_A \cdot y_A + \sum_{B \in \mathcal{B}_A} q_{A,B} f_{A,B}(\mathbf{0}, y_A) : T^A y_A \ge j^A\big\}$; this saves us from having to introduce extra cumbersome notation.

Lemma 6.3 and its proof in Section 6.1.1 remain almost unchanged. The only place where we used problem-specific information was in bounding $y_{A,S} \le 1$ in the 2-stage problems $f_A(\mathbf{0}; \alpha_A)$ and $\widehat{f}_A(\mathbf{0}; \alpha_A)$ which allowed us to, a) bound the Lipschitz constant of $l_A(x; .)$ and $\widehat{l}_A(x; .)$, and b) to show that the feasible region of $f_A(\mathbf{0}; \alpha_A)$ is bounded (so that Theorem 5.2 could be applied). As argued above, y_A can be restricted to the ball $B(\mathbf{0}, 2R)$ in the problems $f_A(\mathbf{0}; \alpha_A)$ and $\widehat{f}_A(\mathbf{0}; \alpha_A)$. So using Lemma 6.7 (which remains unchanged), we can bound the Lipschitz constant of $l_A(x; .)$ and $\widehat{l}_A(x; .)$ by $K' = 3R$ (note that $\ln R = \mathsf{poly}(\mathcal{I})$, so $\ln K'$ is polynomially bounded), and since $f_A(\mathbf{0}; \alpha_A)$ is a 2-stage program of the form (2Gen-P) with a bounded feasible region, we can still apply Theorem 5.2 to $f_A(\mathbf{0}; \alpha_A)$ (when $\alpha_A \ge \rho w^A$). So the proof in Section 6.1.1 is essentially unchanged, and thus using essentially the same arguments that we used for the 3-stage set cover problem, we obtain the following theorem.

Theorem 6.9 *For any parameters $\epsilon, \gamma > 0$ ($\gamma \le 1$), one can construct the sample average problem $\widehat{h}$ using $\mathsf{poly}\big(\mathcal{I}, \lambda, \frac{1}{\gamma}, \ln(\frac{1}{\epsilon}), \ln(\frac{1}{\delta})\big)$ samples so that, with probability at least $1 - \delta$, any optimal solution $\widehat{x}$ to $\widehat{h}$ has value $h(\widehat{x}) \le (1 + 7\gamma) \cdot OPT + 18\epsilon$.*

The sampling step described in Section 7.1 yields a lower bound on OPT for a subclass of (3Gen-P) where the recourse problem $f_A(x)$ does not have any constraints (for instance, as in the relaxation of the 3-stage set cover problem (3SSCR-P)). This allows us to obtain a purely multiplicative $(1 + \kappa)$-guarantee for this subclass of 3-stage programs.

7 The SAA bound for k-stage programs

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

In the k-stage problem, the scenario distribution is specified by a k-level tree, called the *distribution tree*. We start at the root r of this tree at level 1, which represents the first-stage. Let $\mathsf{level}(i)$ denote the set of nodes at level i, so $\mathsf{level}(1) = \{r\}$. Each such node u represents an outcome in stage i and its ancestors correspond to the outcomes in the previous stages; so node u represents a particular evolution of the uncertainty through stages $1, \ldots, i$. At a leaf node, the uncertainty has completely resolved itself and we know the input precisely. As before, for clarity, a *scenario* will always refer to a stage k outcome, that is, 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 $\mathsf{path}(u)$ be the set of all nodes (including u) on u's path to the root. Let $\mathsf{child}(u)$ be the set of all children of u; this is the set of possible outcomes in the next stage given that u is the current outcome. Let p_u be the probability that outcome u occurs, and q_u be the *conditional probability* that u occurs given the outcome in the previous stage. *We do not assume anything about the distribution*, and it can incorporate

various correlation effects from previous stages. Note that $p_u = \prod_{v \in \mathsf{path}(u)} q_v$. Clearly we have $p_r = q_r = 1$, for any i $\sum_{v \in \mathsf{level}(i)} p_v = 1$, and for any node u, $\sum_{v \in \mathsf{child}(u)} q_v = 1$.

We use y_u to refer to the decisions taken in outcome u and w^u to denote the costs in outcome u; thus the costs may depend on the history of outcomes in the previous stages. Note that y_u may only depend on the decisions in the previous outcomes, that is, on the y_v's where $v \in \mathsf{path}(u)$. For convenience we use $x \equiv y_r$ to denote the first-stage decisions, and w^{I} to denote the first-stage costs. We consider the following generic k-stage linear program.

$$f_{k,r} \;=\; \min \quad h(x) = w^{\mathrm{I}} \cdot x + \sum_{u \in \mathsf{child}(r)} q_u f_{k-1,u}(x) \quad \text{subject to} \quad x \in \mathcal{P} \subseteq \mathbb{R}^m_{\geq 0} \qquad (k\text{Gen-P})$$

where $f_{k-1,u}(x)$ gives the expected cost of stages $2, \ldots, k$ given the first-stage decision x and when u is the stage 2 outcome. Thus $f_{k-1,u}(x)$ is the cost of the $(k-1)$-*stage problem* that is obtained when u is the second-stage outcome, and x is the first-stage decision. In general, consider an outcome $u \in \mathsf{level}(i)$ and let $v \in \mathsf{level}(i-1)$ be it's parent. Let $\mathbf{y}_v = (y_r, \ldots, y_v)$, where $\{y_r, \ldots, y_v\} = \mathsf{path}(v)$, denote the collective tuple of decisions taken in the previous stages; for the root r, $\mathbf{y}_r \equiv y_r \equiv x$. The function $f_{k-i+1,u}(\mathbf{y}_v)$ is a $(k-i+1)$-stage stochastic program that determines the expected cost of stages $i, \ldots, k$ given the decisions in the previous stages $\mathbf{y}_v$, and when u is the outcome in stage i. It is defined recursively as

$$f_{k-i+1,u}(\mathbf{y}_v) \;=\; \min \quad \Big\{ w^u \cdot y_u + \sum_{u' \in \mathsf{child}(u)} q_{u'} f_{k-i,u'}(\mathbf{y}_v, y_u) : y_u \in \mathbb{R}^m_{\geq 0}, \quad T^u y_u \geq j^u - \sum_{t \in \mathsf{path}(v)} T^u y_t \Big\},$$

for a non-leaf node $u \in \mathsf{level}(i)$, $2 \leq i < k$. For a leaf u at level k,

$$f_{1,u}(\mathbf{y}_v) \;=\; \min \quad \Big\{ w^u \cdot y_u + c^u \cdot s_u : y_u \in \mathbb{R}^m_{\geq 0},\ s_u \in \mathbb{R}^n_{\geq 0}, \quad D^u s_u + T^u y_u \geq j^u - \sum_{t \in \mathsf{path}(v)} T^u y_t \Big\}.$$

The variables s_u appearing in $f_{1,u}(.)$, capture the fact that at a scenario u when we know the input precisely, one might need to make some additional decisions. We require that (a) $T^u \geq \mathbf{0}$ for every node u; (b) $0 \leq f_{k-i+1,u}(\mathbf{y}_v) < \infty$ for every node $u \in \mathsf{level}(i)$ with parent v, and feasible decisions $\mathbf{y}_v$ — this ensures that the primal problem $f_{k-i+1,u}(\mathbf{y}_v)$ and its dual are feasible for every feasible $\mathbf{y}_v$; and (c) there is some R with $\ln R$ polynomially bounded such that for every internal node u, the feasible region of $f_{k-i+1,u}(\mathbf{y}_v)$ can be restricted to $B(\mathbf{0}, R)$ without affecting the solution quality (so $\mathcal{P} \subseteq B(\mathbf{0}, R)$), that is, for each $f_{k-i+1,u}(\mathbf{y}_v)$ there is some optimal solution y^*_u such that $\|y^*_u\| \leq R$. Let $\mathcal{I}$ denote the input size, λ be the ratio $\max\big(1, \max_{v,u \in \mathsf{child}(v)} \frac{w^u}{w^v}\big)$, and K be the Lipschitz constant of $h(.)$. Define $OPT = f_{k,r}$.

The sample average problem is of the same form as (kGen-P), where the probability q_u is replaced by its estimate $\widehat{q}_u$, which is the frequency of occurrence of outcome u in the appropriate sampled set. It is constructed as follows: we sample $\mathcal{T}_2$ times from the entire distribution and estimate the probability q_u of a node $u \in \mathsf{level}(2)$ by its frequency of occurrence $\widehat{q}_u = \mathcal{T}_{2;u}/\mathcal{T}_2$; for each u such that $\widehat{q}_u > 0$, we sample $\mathcal{T}_3$ times from the conditional distribution of scenarios in the tree rooted at u and estimate the probability $q_{u'}$ for each $u' \in \mathsf{child}(u)$ by the frequency $\widehat{q}_{u'} = \mathcal{T}_{3;u'}/\mathcal{T}_3$. We continue this way, sampling for each node u such that $\widehat{q}_u > 0$, the leaves of the tree rooted at u to estimate the probabilities of the children of u, till we reach the leaves of the distribution tree. Let $\widehat{p}_u = \prod_{v \in \mathsf{path}(u)} \widehat{q}_v$ denote the probability of occurrence of outcome u in the sample average problem. We use $\widehat{f}_{k,r}$ to denote the k-stage sample average problem; correspondingly for node $u \in \mathsf{level}(i)$ (where $\widehat{p}_u > 0$) with parent v, $\widehat{f}_{k-i+1,u}(\mathbf{y}_v)$ is the $(k-i+1)$-stage program in the sample average problem that determines the expected cost of stages $i, \ldots, k$ when outcome u occurs and given the decisions $\mathbf{y}_v$ in the previous stages. Note that for a leaf u, $f_{1,u}(\mathbf{y}_v)$ is simply a (1-stage) deterministic linear program, so $\widehat{f}_{1,u}(\mathbf{y}_v) = f_{1,u}(\mathbf{y}_v)$. Let $\widehat{h}(x)$ be the objective function of the k-stage sample average program, so $\widehat{f}_{k,r} = \min_{x \in \mathcal{P}} \widehat{h}(x)$.

In Sections 5 and 6 we proved a polynomial SAA bound for the generic 2-stage problem $f_{2,r}$ and 3-stage problem $f_{3,r}$ respectively. We now extend this argument inductively to prove an SAA bound for the k-stage problem $f_{k,r}$. We will show that assuming inductively a polynomial SAA bound $\mathcal{N}_{k-1}$ for the $(k-1)$-stage problem $f_{k-1,r}$, one can construct the sample average problem $\widehat{f}_{k,r}$ with a sufficiently large polynomial sample size, so that, with high probability, any optimal solution to $\widehat{f}_{k,r}$ is a near-optimal solution to $f_{k,r}$. Combined with the results in Sections 5 and 6 which provide the base case in this argument, this establishes a polynomial SAA bound for k-stage programs of the form (kGen-P).

We dovetail the approach used for 3-stage programs in Section 6. For a node $u \in \mathsf{level}(2)$, we use $f_{k-1,u}(\mathbf{0}; W)$ to denote the $(k-1)$-stage problem $f_{k-1,u}(x)$ with $x = \mathbf{0}$ and costs $w^u = W$; $\widehat{f}_{k-1,u}(\mathbf{0}; W)$ denotes the corresponding quantity in the sample average problem. Like in Section 6, we formulate a concave maximization problem $LD_{k-1,u}(x)$ that is dual to $f_{k-1,u}(x)$, which has a $(k-1)$-stage primal problem of the type $f_{k-1,r}$ embedded inside it. This dual is defined as $LD_{k-1,u}(x) = \max_{\mathbf{0} \le \alpha_u \le w^u} l_{k-1,u}(x; \alpha_u)$ where $l_{k-1,u}(x; \alpha_u) = -\alpha_u \cdot x + f_{k-1,u}(\mathbf{0}; \alpha_u)$. We use $\widehat{l}_{k-1,u}(x; \alpha_u)$ and $\widehat{LD}_{k-1,u}(x)$ to denote the analogues in the sample average problem.

We want to show that the true function $h(.)$, and the sample average function $\widehat{h}(.)$ are close in terms of their subgradients. As in Section 6, to avoid some technical difficulties, we consider slightly modified versions of these functions, h_ρ and $\widehat{h}_\rho$ respectively, and show that they are close in terms of their subgradients and this will suffice to prove an SAA bound. Define $h_\rho(x) = w^{\mathrm{I}} \cdot x + \sum_{u \in \mathsf{child}(r)} q_u LD_{k-1,u;\rho}(x)$ and $\widehat{h}_\rho(x) = w^{\mathrm{I}} \cdot x + \sum_{u \in \mathsf{child}(r)} \widehat{q}_u \widehat{LD}_{k-1,u;\rho}(x)$. where $LD_{k-1,u;\rho}(x)$ and $\widehat{LD}_{k-1,u;\rho}(x)$ are respectively the maximum of $l_{k-1,u}(x; \alpha_u)$ and $\widehat{l}_{k-1,u}(x; \alpha_u)$ over the region $\mathcal{D}_u = \{\alpha_u \in \mathbb{R}^m : \rho w^{\mathrm{I}} \le \alpha_u \le w^u\}$. A subgradient to $h_\rho(.)$ and $\widehat{h}_\rho(.)$ at point x is obtained from the solutions to the dual recourse problems $LD_{k-1,u;\rho}(x)$ and $\widehat{LD}_{k-1;\rho,u}(x)$ respectively; so we first argue that an optimal solution to $\widehat{LD}_{k-1;\rho,u}(x)$ is a near-optimal solution to $LD_{k-1,u;\rho}(x)$. To do this we show that the dual objective functions are close in terms of their $\max$-subgradients. A (approximate) $\max$-subgradient of $l_{k-1,u}(x; .)$ at the point α_u is obtained from an (near-) optimal solution to $f_{k-1;\rho,u}(\mathbf{0}; \alpha_u)$, which is a $(k-1)$-stage program belonging to our class with bounded cost ratio (this is the reason why we consider functions h_ρ and $\widehat{h}_\rho$ instead of h and $\widehat{h}$). We use the inductive hypothesis to argue that an optimal solution to the $(k-1)$-stage program $\widehat{f}_{k-1,u}(\mathbf{0}; \alpha_u)$ in the sample average dual yields a near-optimal solution to the $(k-1)$-stage program $f_{k-1,u}(\mathbf{0}; \alpha_u)$ in the true dual, and therefore the $\max$-subgradients of the objective functions of the sample-average dual and the true dual are close to each other. Unfolding the chain of arguments, this shows that an optimal solution to $\widehat{LD}_{k-1,u}(x)$ is a near-optimal solution to $LD_{k-1,u}(x)$, which shows the closeness in subgradients of the objective functions h and $\widehat{h}$. This in turn leads to an SAA bound for the k-stage program $f_{k,r}$.

To reduce clutter we adopt the following terminology: for a minimization problem, we call a solution a (γ, ϵ)-optimal solution if it has cost at most $(1+\gamma) \cdot (\text{minimum}) + \epsilon$; for a maximization problem, a (γ, ϵ)-optimal solution is a solution that has value at least $(1-\gamma) \cdot (\text{maximum}) - \epsilon$. We first state the induction hypothesis precisely.

Induction Hypothesis *For a $(k-1)$-stage problem of the type $f_{k-1,r}$ with input size $\mathcal{I}$, cost ratio λ, and satisfying requirements (a), (b), and (c), one can construct the sample average problem $\widehat{f}_{k-1,r}$ using $\mathcal{N}_{k-1}(\mathcal{I}, \lambda, \gamma, \epsilon, \delta) = \mathsf{poly}\big(\mathcal{I}, \frac{\lambda}{\gamma}, \ln(\frac{1}{\epsilon\delta})\big)$ samples, with probability at least $1-\delta$, any optimal solution to $\widehat{f}_{k-1,r}$ is a (γ, ϵ)-optimal solution to $f_{k-1,r}$.*

Like in Section 6, we show that strong duality holds (with the new dual representation), and state a structural lemma about the subgradients of the objective function which paves the way for showing the closeness in subgradients. The proofs of these two lemmas are very similar to those of Lemmas 6.8 and 6.2. We use Γ_u to denote the subtree of the distribution tree rooted at node u.

Lemma 7.1 *At any $x \in \mathcal{P}$ and node $u \in \mathsf{level}(2)$, $f_{k-1,u}(x) = LD_{k-1,u}(x)$ and $\widehat{f}_{k-1,u}(x) = \widehat{LD}_{k-1,u}(x)$. Moreover in the $(k-1)$-stage problems $f_{k-1,u}(\mathbf{0};\alpha_u)$ and $\widehat{f}_{k-1,u}(\mathbf{0};\alpha_u)$, we can restrict y_u to $B(\mathbf{0},2R)$ and y_t to $B(\mathbf{0},R)$ for any internal node $t \in \Gamma_u$, without affecting the values of $LD_{k-1,u}(x)$ and $\widehat{LD}_{k-1,u}(x)$.*

Proof : The proof proceeds as in Lemma 6.8 and we only briefly sketch the details. One can expand $f_{k-1,u}(x)$ into a minimization LP with objective function $w^u \cdot y_u + \sum_{t\in\Gamma_u} \frac{p_t}{p_u} w^t \cdot y_t + \sum_{t\in\Gamma_u \cap \mathsf{level}(k)} \frac{p_t}{p_u} c^t \cdot s_t$. The constraints are $T^t(\sum_{t'\in\mathsf{path}(t)\setminus\{r\}} y_{t'}) \geq j^t - T^t x$ for every non-leaf node $t \in \Gamma_u$, and $D^t s_t + T^t(\sum_{t'\in\mathsf{path}(t)\setminus\{r\}} y_{t'}) \geq j^t - T^t x$ for every leaf $t \in \Gamma_u$. Let $\big(\{y_t^*\},\{s_t^*\}\big)$ be an optimal solution to this LP, and $\big(\{\theta_t^*\}\big)$ be a solution to the dual maximization LP. Let α_u^* be an optimal solution to $LD_{k-1,u}(x)$.

Setting $y_u = x + y_u^*$ in $f_{k-1,u}(\mathbf{0};\alpha_u^*)$ shows that $LD_{k-1,u}(x) \leq f_{k-1,u}(x)$. Note that this upper bound also holds when we require that, $y_u \in B(\mathbf{0},2R)$ and $y_t \in B(\mathbf{0},R)$ for all other internal nodes $t \in \Gamma_u$, in the problem $f_{k-1,u}(\mathbf{0};\alpha_u)$ embedded in the dual maximization problem $LD_{k-1,u}(x)$. We can lower bound $LD_{k-1,u}(x)$ by $f_{k-1,u}(x)$, by computing the value of the feasible solution where $\alpha_u = \sum_{t\in\Gamma_u}(T^t)^{\mathrm{T}}\theta_t^*$ and the solution to the LP dual of $f_{k-1,u}(\mathbf{0};\alpha_u)$ is given by $\big(\{\theta_t^*\}\big)$. Hence, $f_{k-1,u}(x) = LD_{k-1,u}(x)$, and constraining $\|y_u\| \leq 2R$ and $\|y_t\| \leq R$ for every internal node $t \in \Gamma_u \setminus \{u\}$ in the problem $f_{k-1,u}(\mathbf{0};\alpha_u)$ does not affect the value of $LD_{k-1,u}(x)$. The arguments for $\widehat{f}_{k-1,u}(x)$ and $\widehat{LD}_{k-1,u}(x)$ are identical. ■

Lemma 7.2 *Let $x \in \mathcal{P}$ and α_u be an $(\varepsilon, \varepsilon w^{\mathrm{I}}\cdot x + \epsilon)$-optimal solution to $LD_{k-1,u}(x)$ for every node $u \in \mathsf{level}(2)$. (i) $d = w^{\mathrm{I}} - \sum_{u\in\mathsf{level}(2)} p_u\alpha_u$ is an (ε,ϵ)-subgradient of $h(.)$ at x with $\|d\| \leq \lambda\|w^{\mathrm{I}}\|$; (ii) if $\widehat{d}$ is a vector such that $d - \omega w^{\mathrm{I}} \leq \widehat{d} \leq d + \omega w^{\mathrm{I}}$, then $\widehat{d}$ is an $(\varepsilon+\omega,\epsilon)$-subgradient of $h(.)$ at x.*

As in Section 6.2, given Lemma 7.1, we abuse notation and use $f_{k-1,u}(\mathbf{0};\alpha_u)$ to actually refer to the problem where we have imposed the constraints that y_u lie in $B(\mathbf{0},2R)$ and y_t lie in $B(\mathbf{0},R)$ for every internal node $t \in \Gamma_u$. Observe that for any $u \in \mathsf{level}(2)$, when $\alpha_u \geq \rho w^{\mathrm{I}}$, in the $(k-1)$-stage problem $f_{k-1,u}(\mathbf{0};\alpha_u)$ the ratio of costs $\frac{w^{t'}}{w^t}$ for any t lying in the tree rooted at u and any $t' \in \mathsf{child}(t)$, is bounded by $\frac{\lambda^2}{\rho}$. Let $P_k(\mathcal{I},\lambda,\gamma,\epsilon,\delta) = \mathsf{poly}\big(\mathcal{I},\lambda,\frac{1}{\gamma},\ln(\frac{1}{\epsilon}),\ln(\frac{1}{\delta})\big)$ be a sufficiently large polynomial. To avoid clutter we suppress the dependence on $(\mathcal{I},\ldots,\delta)$.

Lemma 7.3 *For any $\epsilon,\rho,\varepsilon > 0$, any $x \in \mathcal{P}$, and any node $u \in \mathsf{level}(2)$, if we construct the recourse problem $\widehat{f}_{k-1,u}(x)$ with $\mathcal{T}(\epsilon,\rho,\varepsilon,\delta) = \mathcal{N}_{k-1}\big(\mathcal{I},\frac{\lambda^2}{\rho},\frac{\varepsilon}{8N'\lambda},\frac{\epsilon}{16N'},\frac{\delta}{n'}\big)$ samples, for a suitable $N', \ln n' = \mathsf{poly}\big(\mathcal{I},\ln(\frac{1}{\epsilon})\big)$, then any optimal solution to $\widehat{LD}_{k-1,u;\rho}(x)$ is an $(\varepsilon,\varepsilon w^{\mathrm{I}}\cdot x+\epsilon)$-optimal solution to $LD_{k-1,u;\rho}(x)$ with probability at least $1-\delta$.*

Proof : We show that $l_{k-1,u}(x;.)$ and $\widehat{l}_{k-1,u}(x;.)$ are close in terms of their $\max$-subgradients and then use Lemma 4.3. Recall that $\mathcal{D}_u = \{\alpha_u \in \mathbb{R}^m : \rho w^{\mathrm{I}} \leq \alpha_u \leq w^u\}$. Let $R' = \|w^u\| \leq \lambda\|w^{\mathrm{I}}\|$, so $\mathcal{D}_u \subseteq B(\mathbf{0},R')$. In the sequel we will only consider $(\omega,\Delta,\mathcal{D}_u)$-$\max$-subgradients, so we will omit the $\mathcal{D}_u$.

As in Lemma 6.7, for any $\omega > 0$, one can show that if y_u is an (ω',ϵ')-optimal solution to $f_{k-1,u}(\mathbf{0};\alpha_u)$, where $\omega' = \frac{\omega}{\lambda}$, then $y_u - x$ is an $(\omega,\omega w^{\mathrm{I}}\cdot x + \epsilon')$-$\max$-subgradient of $l_{k-1,u}(x;.)$ at α_u. This follows because $l_{k-1,u}(x;\alpha_u) \geq (y_u - x)\cdot\alpha_u + C - \omega' f_{k-1,u}(\mathbf{0};\alpha_u) - \epsilon'$ where $C = \sum_{u'\in\mathsf{child}(u)} q_{u'} f_{k-2,u'}(y_u)$, and at any other point α_u', we have $l_{k-1,u}(x;\alpha_u') \leq (y_u - x)\cdot\alpha_u' + C$. This also shows that if $\widehat{y}_u$ is an optimal solution to $\widehat{f}_{k-1,u}(\mathbf{0};\alpha_u)$ then $\widehat{y}_u - x$ is a $\max$-subgradient of $\widehat{l}_{k-1,u}(x;.)$ at α_u. We may assume that $\|y_u\| \leq 2R$ by Lemma 7.1, so the Lipschitz constant of $l_{k-1,u}(x;.)$ and $\widehat{l}_{k-1,u}(x;.)$ can be bounded by $K' = 3R$. $f_{k-1,u}(\mathbf{0};\alpha_u)$ is a $(k-1)$-stage problem of the form $f_{k-1,r}$ such that for every internal node t in the tree Γ_u we have $\|y_t\| \leq 2R$, so we can apply the induction hypothesis to it.

Set $\epsilon' = \frac{\epsilon}{8}$. Let $N' = \log\big(\frac{2K'R'}{\epsilon'}\big)$ and $\omega = \frac{\varepsilon}{8N'}$. Observe that $\log(K'R') = \mathsf{poly}(\mathcal{I})$. Let G be the extended $\frac{\epsilon'}{K'N'\sqrt{m}}$-grid of $\mathcal{D}_u$ and $n' = |G|$. Suppose that we use $\mathcal{N}_{k-1}\big(\mathcal{I},\frac{\lambda^2}{\rho},\frac{\omega}{\lambda},\frac{\epsilon'}{2N'},\frac{\delta}{n'}\big)$ samples to

construct the recourse problem $\widehat{f}_{k-1,u}(x)$, and hence the dual $\widehat{LD}_{k-1,u;\rho}(x)$. At any given $\alpha_u \in G$, applying the induction hypothesis to $f_{k-1,u}(\mathbf{0};\alpha_u)$, an optimal solution $\widehat{y}_u$ to $\widehat{f}_{k-1,u}(\mathbf{0};\alpha_u)$ is an $\big(\frac{\omega}{\lambda}, \frac{\epsilon'}{2N'}\big)$-optimal solution to $f_{k-1,u}(\mathbf{0};\alpha_u)$ with probability at least $1-\frac{\delta}{n'}$. Thus, with probability $1-\delta$, at every $\alpha_u \in G$, $\widehat{y}_u - x$ is both a max-subgradient of $\widehat{l}_{k-1,u}(x;.)$ and an $(\omega, \omega w^{\mathrm{I}}\cdot x + \frac{\epsilon'}{2N'})$- max-subgradient of $l_{k-1,u}(x;.)$ at α_u. So by Lemma 4.3 we get that if $\widehat{\alpha}_u \in \mathcal{D}_u$ maximizes $\widehat{l}_{k-1,u}(x;\alpha_u)$ then it is an $(\varepsilon, \varepsilon w^{\mathrm{I}}\cdot x + \epsilon)$-optimal solution to $LD_{k-1,u;\rho}(x)$. Thus we obtain $\mathcal{T}(\epsilon,\rho,\varepsilon,\delta) = \mathcal{N}_{k-1}\big(\mathcal{I}, \frac{\lambda^2}{\rho}, \frac{\varepsilon}{8N'\lambda}, \frac{\epsilon}{16N'}, \frac{\delta}{n'}\big)$. ■

Now we can prove our main theorem. First we state the analogue of Lemma 6.4.

Lemma 7.4 *Consider the sample average function $\widehat{h}$ constructed using $\mathcal{N}_2 = \mathcal{T}_2(\omega,\delta) = \frac{16(1+\lambda)^2}{\omega^2}\ln\big(\frac{4m}{\delta}\big)$ samples from stage 2, and using $\mathcal{T}\big(\epsilon,\rho,\frac{\omega}{2},\frac{\delta}{2\mathcal{N}_2}\big)$ samples from the tree rooted at u (to generate $\widehat{f}_{k-1,u}(x)$) for each $u \in \mathsf{level}(2)$ with $\widehat{q}_u > 0$. 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 $1-\delta$.*

Theorem 7.5 *For any $\epsilon,\gamma > 0$ ($\gamma < 1$), with probability at least $1-\delta$, any optimal solution $\widehat{x}$ to the k-stage sample average problem constructed using $\mathsf{poly}(\mathcal{I},\lambda,\gamma,\epsilon,\delta)$ samples satisfies $h(\widehat{x}) \le (1+\gamma)\cdot f_{k,r} + \epsilon$.*

Proof : Set $\gamma' = \frac{\gamma}{7}$ and $\epsilon' = \frac{\epsilon}{18}$. Let $N = \log\big(\frac{2KR}{\epsilon'}\big)$ and $\omega = \frac{\gamma'}{8N}$. Note that $\log(KR) = \mathsf{poly}(\mathcal{I})$. Set $\epsilon'' = \frac{\epsilon'}{N}$ and $\rho = \frac{\gamma'}{4N}$. Let G be the extended $\frac{\epsilon'}{KN\sqrt{m}}$-grid of $\mathcal{P}$ and $n = |G|$. Let $\mathcal{N}' = \frac{16(1+\lambda)^2}{\omega^2}\ln\big(\frac{4mn}{\delta}\big)$. Using Lemma 7.4 and the union bound over all points in G, by constructing $\widehat{h}(.)$ (and hence $\widehat{h}_\rho(.)$) using $\mathcal{N} = \mathcal{N}'\cdot\mathcal{T}\big(\epsilon'',\rho,\frac{\omega}{2},\frac{\delta}{2n\mathcal{N}'}\big)$ samples, with 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 $h_\rho(.)$. Mimicking the proof of Theorem 6.9 we obtain that $h(\widehat{x}) \le (1+7\gamma')OPT + 18\epsilon'$.

Let N', n' be as given by Lemma 7.3. We can choose $P_k(\mathcal{I},\lambda,\gamma,\epsilon,\delta)$ to be a large enough polynomial so that the following hold: $\mathcal{N}' \le P_k$, $\frac{1}{\rho} \le P_k$, $\frac{\omega}{16N'\lambda} = \frac{\gamma}{O(1)NN'\lambda} \le \frac{\gamma}{P_k}$, $\frac{\epsilon''}{16N'} = \frac{\epsilon}{O(1)NN'} \le \frac{\epsilon}{P_k}$, $\frac{\delta}{2n\mathcal{N}'n'} \le \frac{\delta}{2^{P_k}}$. So using Lemma 7.3 we can bound $\mathcal{T}\big(\epsilon'',\rho,\frac{\omega}{2},\frac{\delta}{2n\mathcal{N}'}\big)$ by $\mathcal{N}_{k-1}\big(\mathcal{I},P_k\lambda^2,\frac{\gamma}{P_k},\frac{\epsilon}{P_k},\frac{\delta}{2^{P_k}}\big)$. Unfolding the recurrence (note that k is a constant), and using Theorem 5.2 for the base case, we get that $\mathcal{N}_k(\mathcal{I},\lambda,\gamma,\epsilon,\delta)$ is a polynomial in $\mathcal{I}$, $\frac{\lambda}{\gamma}$, $\ln\big(\frac{1}{\epsilon\delta}\big)$. ■

7.1 Obtaining a lower bound on OPT

The bounds obtained thus far on the quality of an optimal solution to the sample average problem in Theorem 5.2, Theorem 6.9, and Theorem 7.5 are all of the form $h(\widehat{x}) \le \big(1+O(\gamma)\big)\cdot OPT + O(\epsilon)$ (where $\gamma,\epsilon > 0$ are parameters) containing both multiplicative and additive approximation factors. This can be converted into a purely multiplicative $(1+\kappa)$-guarantee by setting γ and ϵ appropriately provided that we have a lower bound on OPT (that is at least inverse exponential in the input size). We now show that, under some mild assumptions, one can obtain such a lower bound for a *subclass* of (kGen-P), where for every node u in $\mathsf{level}(i)$, $2 \le i < k$, the recourse problem $f_{k-i+1,u}(x,\mathbf{y}_v)$ does not have any constraints. That is, we consider the following subclass of (kGen-P):

$$g_{k,r} \;=\; \min\; h(x) = w^{\mathrm{I}}\cdot x \;+ \sum_{u\in\mathsf{child}(r)} q_u g_{k-1,u}(x) \quad \text{subject to} \quad x \in \mathcal{P} \subseteq \mathbb{R}^m_{\ge 0}, \quad \text{where}$$

$$g_{k-i+1,u}(\mathbf{y}_v) \;=\; \min\Big\{w^u\cdot y_u + \sum_{u'\in\mathsf{child}(u)} q_{u'} g_{k-i,u'}(\mathbf{y}_v, y_u) : \quad y_u \in \mathbb{R}^m_{\ge 0}\Big\}, \quad \text{for } u \in \mathsf{level}(i),\ 2 \le i < k,$$

$$g_{1,u}(\mathbf{y}_v) \;=\; \min \quad \Big\{w^u\cdot y_u + c^u\cdot s_u : y_u \in \mathbb{R}^m_{\ge 0},\ s_u \in \mathbb{R}^n_{\ge 0}, \quad D^u s_u + T^u y_u \ge j^u - \sum_{t\in\mathsf{path}(v)} T^u y_t\Big\}.$$

Note that for 2-stage programs, the above class is *the same as* (2Gen-P).

We make the mild assumption that (a) $x = \mathbf{0}$ lies in $\mathcal{P}$, and (b) for every scenario u with parent v, either $f_{1,u}(\mathbf{y}_v)$ is minimized by setting $y_t = \mathbf{0}$ for all $t \in \mathsf{path}(v)$, or the total cost $\sum_{t \in \mathsf{path}(u)} w^t y^t + c^u s^u \geq 1$ for any feasible decisions $(\mathbf{y}_u, s_u)$. For example, for the 3-stage set cover problem considered in Section 6.1, (a) just requires that we are allowed to not pick any set in the first-stage, (b) is satisfied if the total cost incurred in every scenario (A, B) with $\mathcal{E}(A, B) \neq \emptyset$ is at least 1 (if $\mathcal{E}(A, B) = \emptyset$ then the cost incurred is 0 for every x, y_A). Under these assumptions, we show that we can sample initially to detect if OPT is large.

Let $\mathsf{Null} = \{u \in \mathsf{level}(k) : f_{1,u}(\mathbf{y}_v)$ is minimized at $\mathbf{y}_v = \mathbf{0}\}$; we call a scenario $u \in \mathsf{Null}$ a "null-scenario". The basic idea is that if the non-null scenarios account for a probability mass of at least $\frac{1}{\lambda^k}$, then $OPT \geq \frac{1}{\lambda^k}$ since the cost incurred in each such scenario is at least 1. Otherwise we show that $x = \mathbf{0}$ is an optimal solution, by arguing that for any solution $x \neq \mathbf{0}$ we can substitute the x-decisions with recourse actions y_u in each scenario u, and the overall cost decreases since the low probability of occurrence of a non-null scenario outweighs the increase in the cost of such a scenario (at most a factor of λ^k).

Lemma 7.6 *By sampling $M = \lambda^k \ln\left(\frac{1}{\delta}\right)$ times, one can detect with probability at least $1 - \delta$ ($\delta < \frac{1}{2}$), that either $x = \mathbf{0}$ is an optimal solution to* (3SSC-P)*, or that $OPT \geq \frac{\delta}{M}$.*

Proof : Let X be the number of times we sample a non-null scenario, i.e., a scenario not in Null. Note that given a scenario u, one can decide in polynomial time if u with parent v is a null-scenario by solving the polynomial-size LP $\min_{\mathbf{y}_v \geq \mathbf{0}} f_{1,u}(\mathbf{y}_v)$. If $X = 0$, we return $x = \mathbf{0}$ as an optimal solution, otherwise we assert that $OPT \geq \frac{\delta}{M}$. In every non-null scenario we incur a cost of at least 1, so $OPT \geq q$ where $q = \sum_{u \in \mathsf{level}(k) \setminus \mathsf{Null}} p_u$ is the probability of occurrence of a non-null scenario. Let $r = \Pr[X = 0] = (1 - q)^M$. So $r \leq e^{-qM}$ and $r \geq 1 - qM$. If $q \geq \frac{1}{\lambda^k}$, then $\Pr[X = 0] \leq \delta$. So with probability at least $1 - \delta$ we will say that $OPT \geq \frac{\delta}{M}$ which is true since $OPT \geq q$. We show that if $q < \frac{1}{\lambda^k}$, then $x = \mathbf{0}$ is an optimal solution. So if $q \leq \delta/M$, then $\Pr[X = 0] \geq 1 - \delta$, and we return the correct answer with probability at least $1 - \delta$. If $\delta/M < q < \ln\left(\frac{1}{\delta}\right)/M$, then we always return a correct answer since it is both true that $x = \mathbf{0}$ is an optimal solution, and that $OPT \geq q \geq \frac{\delta}{M}$.

We now show that if $q < \frac{1}{\lambda^k}$, then $x = \mathbf{0}$ is an optimal solution. Consider any solution $\left(x', \{y'_u\}\right)$. The cost of this solution is

$$h(x') \geq w^{\mathrm{I}} \cdot x' + \sum_{u \in \mathsf{level}(i), 1 < i < k} p_u w^u \cdot y'_u + \sum_{u \in \mathsf{Null} \text{ with parent } v} p_u f_{1,u}(\mathbf{y}'_v) + \sum_{u \notin \mathsf{Null}} p_u (w^u \cdot y'_u + c^u \cdot s_u).$$

For any scenario u with parent v, since $T^u \geq \mathbf{0}$, $f_{1,u}(\mathbf{y}_v)$ is a decreasing function of y_t for every $t \in \mathsf{path}(v)$. So for a null-scenario u, since $f_{1,u}(\mathbf{y}_v)$ is minimized at $\mathbf{y}_v = \mathbf{0}$, we have that $f_{1,u}(\mathbf{y}_v) = f_{1,u}(\mathbf{0})$ for any feasible decisions $\mathbf{y}_v$. The solution with $x = \mathbf{0}$ and $y_u = y'_u + x'$ for every scenario u, and $y_u = y'_u$ for every other node u is also feasible, and has cost

$$\sum_{u \in \mathsf{level}(i), 1 < i < k} p_u w^u \cdot y'_u + \sum_{u \in \mathsf{Null} \text{ with parent } v} p_u f_{1,u}(\mathbf{0}) + \sum_{u \notin \mathsf{Null}} p_u \left(w^u \cdot (y'_u + x) + c^u \cdot s_u\right).$$

This is at most $h(x') - w^{\mathrm{I}} \cdot x' + q\lambda^k w^{\mathrm{I}} \cdot x' < h(x')$ since $w^u \leq \lambda^k w^{\mathrm{I}}$ for any scenario u and $q < \frac{1}{\lambda^k}$. ■

We can use the above lemma to convert a guarantee of the form $h(\widehat{x}) \leq (1 + c_1\gamma) \cdot OPT + c_2\epsilon$ into a purely multiplicative $(1 + \kappa)$-guarantee. We perform the above sampling step, and after this if we detect that $OPT \geq \varrho/\lambda^k$ where $\varrho = \frac{\delta}{\ln(1/\delta)}$, then we can set $\gamma = \kappa/(2c_1)$ and $\epsilon = \kappa\varrho/(2c_2\lambda^k)$ to obtain a $(1 + \kappa)$-guarantee.

8 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 with restrictions on the distribution or input [11, 9, 6], and later without any restrictions [14]. For a k-stage integer optimization problem, we obtain a near-optimal solution to its linear relaxation by solving the sample average problem as argued in Section 7, and round this solution using an extension of the rounding scheme in [14].

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. This problem can be formulated as a k-stage stochastic LP: $f_{k-i+1,u}(\mathbf{y}_v) = \min_{\mathbf{0} \le y_u \le \Gamma} \big(c^u \cdot y_u + \sum_{u' \in \mathsf{child}(u)} q_{u'} f_{k-i,u'}(\mathbf{y}_v, y_u) \big)$ for a non-leaf node u at level i; for a leaf u, $f_{1,u}(\mathbf{y}_v) = \min_{\mathbf{0} \le y_u \le \Gamma} c^u \cdot y_u$ subject to the constraints that the total flow routed for (s_i, t_i) is at least d_i^u, and the flow on edge e is at most $\min(\Gamma_e, \sum_{u' \in \mathsf{path}(u)} y_{u',e})$. We can apply our algorithm to get $(1+\epsilon)$-optimal solution to this program.

Covering problems We consider the k-stage versions of set cover, vertex cover and the multicut problem on tree. 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. The LP relaxation for the k-stage problem has a variable $y_{u,S}$ indicating if set S is bought in outcome u, and constraints stating that for every leaf, and every element e in its corresponding target set, we must buy some set S that contains e along this root-leaf path.

We can generalize the rounding theorem of Shmoys and Swamy [14] to show that one can use a ρ-approximation algorithm for the deterministic analogue, where the guarantee is with respect to its natural LP relaxation, to round any fractional solution to the k-stage problem to an integer solution losing a factor of $k\rho$; combined with the algorithm in Section 7, this 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 a performance guarantee of $(k \log n + \epsilon)$ for the k-stage set cover problem, and $(2k + \epsilon)$ for the k-stage vertex cover problem and the k-stage multicut problem on trees.

Facility location problems In the k-stage uncapacitated facility location (UFL) problem, we are given a set of candidate facility locations $\mathcal{F}$, 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. This is captured by the k-stage program where $g_{k-i+1,u}(\mathbf{y}_v) = \min_{y_u \ge \mathbf{0}} \big(\sum_i f_i^u y_{u,i} + g_{k-i,u'}(\mathbf{y}_v, y_u) \big)$ for a stage i outcome u, and for a stage k scenario u, $g_{1,u}(\mathbf{y}_v)$ is the minimum of $\sum_i f_i^u y_{u,i} + \sum_j d_j^u c_{ij} x_{u,ij}$ subject to the constraint that for every client j, $\sum_i x_{u,ij}$ is at least 1 if $d_u^j > 0$ and 0 otherwise, and for every i, j, $x_{u,ij} \le \sum_{u' \in \mathsf{path}(u)} y_{u',i}$. We can obtain a $(1+\epsilon)$-optimal solution to this program. Adapting the rounding procedure in [14], we obtain a $1.71(k-1) + 1.52 + \epsilon = O(k)$-approximation algorithm for k-stage UFL. This rounding procedure extends to give $O(k)$-approximation algorithms for 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. Borwein and A. S.Lewis. *Convex Analysis and Nonlinear Optimization* Springer-Verlag, NY, 2000.

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

[4] M. Charikar, C. Chekuri, and M. Pál. Sampling bounds for stochastic optimization. In *Proceedings of 9th International Workshop on Randomization and Computation*, pages 257–269, 2005.

[5] 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.

[6] A. Gupta, M. Pál, R. Ravi, and A. Sinha. Boosted sampling: approximation algorithms for stochastic optimization. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, pages 417–426, 2004.

[7] A. Gupta, M. Pál, R. Ravi, & A. Sinha. What about Wednesday? Approximation algorithms for multistage stochastic optimization. In *Proceedings of the 8th International Workshop on Approximation Algorithms for Combinatorial Optimization*, pages 86–98, 2005.

[8] A. Hayrapetyan, C. Swamy, and É. Tardos Network design for information networks. In *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms*, 933–942, 2005.

[9] N. Immorlica, D. Karger, M. Minkoff, and V. Mirrokni. On the costs and benefits of procrastination: approximation algorithms for stochastic combinatorial optimization problems. In *Proceedings of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 684–693, 2004.

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

[11] R. Ravi and A. Sinha. Hedging uncertainty: approximation algorithms for stochastic optimization problems. In *Proceedings of the 10th International Conference on Integer Programming and Combinatorial Optimization*, pages 101–115, 2004.

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

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

[14] 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.

[15] C. Swamy and D. B. Shmoys. Sampling-based approximation algorithms for multi-stage stochastic optimization. To appear in *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*, 2005.

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