Annals of Operations Research 39(1992)69–95

MONTE CARLO (IMPORTANCE) SAMPLING WITHIN A BENDERS DECOMPOSITION ALGORITHM FOR STOCHASTIC LINEAR PROGRAMS

Gerd INFANGER
Department of Operations Research, Stanford University, Stanford, CA 94305, USA

Abstract

This paper focuses on Benders decomposition techniques and Monte Carlo sampling (importance sampling) for solving two-stage stochastic linear programs with recourse, a method first introduced by Dantzig and Glynn [7]. The algorithm is discussed and further developed. The paper gives a complete presentation of the method as it is currently implemented. Numerical results from test problems of different areas are presented. Using small test problems, we compare the solutions obtained by the algorithm with universe solutions. We present the solutions of large-scale problems with numerous stochastic parameters, which in the deterministic formulation would have billions of constraints. The problems concern expansion planning of electric utilities with uncertainty in the availabilities of generators and transmission lines and portfolio management with uncertainty in the future returns.

1. Introduction

A stochastic linear program is a linear program whose parameters (coefficients, right-hand sides) are uncertain. The uncertain parameters are assumed to be known only by their distributions. This means that the values of some functions are numerical characteristics of random phenomena, e.g. mathematical expectations of functions dependent on decision variables and random parameters.

Suppose a function $z = EC(V)$ is an expectation of a function $C(v^\omega)$, $\omega \in \Omega$. V is a random parameter which has outcomes v^ω. Ω is the set of all possible random events. It can be finite, infinite, discrete or continuous. In the continuous case, the computation of the expected value requires the solution of the integral:

$$EC(V) = \int C(v^\omega) P(d\omega),$$

with P being the probability measure.

In a general case, V would consist of several components, e.g. $V = (V_1, \ldots, V_h)$ with outcomes v^ω, which we also will denote only by lower case letters, e.g. $v = (v_1, \ldots, v_h)$ and $p(v^\omega)$ alias $p(v)$ would denote the corresponding density

function. We assume the components of V to be independent. In addition, we will construct Ω by crossing the sets of outcomes Ω_i for vector entry v_i, $i = 1, \ldots, k$ as

$$\Omega = \Omega_1 \times \Omega_2 \times \ldots \Omega_k.$$

In this case, the above-mentioned integral takes the form of a multiple integral:

$$E\,C(V) = \int \cdots \int C(v)p(v)\mathrm{d}v_1 \ldots \mathrm{d}v_k.$$

In the case of Ω being discrete and finite, the expectation can be computed by a multiple sum:

$$E\,C(V) = \sum_{v_1} \cdots \sum_{v_k} C(v)p(v).$$

The main difficulties in stochastic linear programming deal with the evaluation of the multiple integral or the multiple sum. The numerical computation of the expectation requires a large number of function evaluations, and each function evaluation means a linear program to be solved. Different approaches attack this problem, e.g. Birge [3], Birge and Wets [5], Birge and Wallace [4], Frauendorfer [12], Frauendorfer and Kall [13], Ermoliev [10], Higle and Sen [18], Kall [20], Pereira et al. [26], Rockafellar and Wets [27], Ruszczynski [28], Wets [30], and others. See Ermoliev and Wets [11] for references. We follow the concept of Dantzig et al. [8] and Dantzig and Glynn [7].

2. Two-stage stochastic linear program

An important class of models concerns dynamic linear programs. Variables which describe activities initiated at time t have coefficients at time t and $t + 1$. Deterministic dynamic linear programs appear as staircase problems. The simplest staircase problem is that with two stages: X denotes the first, Y the second-stage decision variables, A, b represent the coefficients and right-hand sides of the first-stage constraints, and D, d concern the second period constraints together with B which couples the two periods. c, f are the objective function coefficients.

In the deterministic case, c, f, A, b, B, D, d are known with certainty to the planner. In the stochastic case, the parameters of the second stage are not known to the planner at time $t = 1$, but will be known at time $t = 2$. At time $t = 1$, only the distributions of these parameters are assumed to be known. The second-stage parameters can be seen as random variables which obtain certain outcomes with certain probabilities. We denote a certain outcome of these parameters with ω and the corresponding probability with p^ω, $\omega \in \Omega$, the set of possible outcomes.

$$
\begin{aligned}
\text{minimize} \quad & Z = cX + E^{\omega}(fY^{\omega}) \\
\text{subject to} \quad & AX = b, \\
& -B^{\omega}X + DY^{\omega} = d^{\omega}, \\
& X, \quad Y^{\omega} \geq 0, \quad \omega \in \Omega.
\end{aligned} \tag{1}
$$

In (1), a two-stage staircase problem is transformed into a two-stage stochastic linear program, with the parameters d and B being random variables. Given the two-stage stochastic linear program, one wants to make a decision X which is feasible for all scenarios and has the minimum expected costs.

We consider the case of Ω being discrete and finite, e.g. $\Omega = (1, \ldots, K)$, the parameter ω takes on K values. Then we can formulate an equivalent deterministic problem to the stochastic linear problem. This is tractable if K is small. For K scenarios, the deterministic equivalent problem is given in (2).

$$
\begin{aligned}
\text{Minimize} \quad & Z = cX + p^1 fY^1 + p^2 fY^2 + \ldots + p^K fY^K \\
\text{subject to} \quad & AX = b, \\
& -B^1X + DY^1 = d^1, \\
& -B^2X + DY^2 = d^2, \\
& \vdots \qquad\qquad \ddots \qquad\qquad \vdots \\
& -B^KX + DY^K = d^K, \\
& X, Y^1, Y^2, \ldots, Y^K \geq 0.
\end{aligned} \tag{2}
$$

Two-stage stochastic linear programs were first studied in Dantzig [6] and then developed by many authors. The method which we want to apply here uses Benders [2] decomposition. Van Slyke and Wets [29] suggested expressing the impact of the second period by a scalar θ and "cuts", which are necessary conditions to the problem and are expressed only in terms of the first period variables X and θ. Benders decomposition splits the original problem into a master problem and a subproblem which decomposes into a series of independent subproblems, one for each $\omega \in \Omega$. According to the L-shaped method the master problem, the subproblems and the cuts are represented in (3), (4) and (5).

The master problem:

$$
\begin{aligned}
\text{minimize} \quad & z_M = cX + \theta \\
\text{subject to} \quad & AX = b, \\
& -G^lX + \alpha^l\theta \geq g^l, \quad l = 1, \ldots, L, \\
& X, \quad \theta \geq 0.
\end{aligned} \tag{3}
$$

The subproblems:

$$\begin{aligned} &\text{minimize} \quad z^\omega = \quad p^\omega f Y^\omega \\ &\text{subject to} \quad p^\omega \pi^\omega : \quad DY^\omega = d^\omega + B^\omega X, \\ &\qquad\qquad\qquad\qquad Y^\omega \geq 0, \ \omega \in \Omega, \quad \text{e.g. } \Omega = \{1, 2, \ldots, K\}, \end{aligned} \tag{4}$$

where $p^\omega \pi^{\omega *}$ is the optimal dual solution of subproblem ω.

The cuts:

$$\begin{aligned} &g = \sum_\omega p^\omega \pi^{\omega *} d^\omega = E(\pi^{\omega *} d^\omega), \\ &G = \sum_\omega p^\omega \pi^{\omega *} B^\omega = E(\pi^{\omega *} B^\omega); \\ &\alpha^l = 0 \quad \text{feasibility cut}, \\ &\alpha^l = 1 \quad \text{optimality cut}. \end{aligned} \tag{5}$$

By solving the master problem, we obtain a solution X. Given X, we can solve K subproblems $\omega \in \Omega$ to compute a cut. The cut is a lower bound on the expected value of the second-stage costs represented as a function of X. Cuts are sequentially added to the master problem and new values of X are obtained until the optimality criterion is met. We distinguish between two types of cuts: feasibility cuts and optimality cuts. The first refers to infeasible subproblems for a given X and the latter to feasible and optimum subproblems, given X.

If the expected values z, G, and g are computed exactly, that is, by evaluating all scenarios $\omega \in \Omega$, we refer to this as the universe case. As we will see later, the number of scenarios easily gets out of hand, and it is not always possible to solve the universe case. Therefore, methods are sought which guarantee a satisfying solution without solving the universe case.

3. Monte Carlo sampling

Each iteration of Benders decomposition requires the computation of expected values, e.g. the subproblem costs, the coefficients and the right-hand sides of the cuts. For each outcome $\omega \in \Omega$, a linear program has to be solved. The expected value of the subproblem costs is denoted by

$$z = E\,C(v^\omega) = E\,fY^{*\omega}, \quad \omega \in \Omega,$$

with $Y^{*\omega}$ being the optimum solution of subproblem ω. The number of elements of Ω is determined by the dimensionality of the stochastic vector $V = (V_1, \ldots, V_k)$.

Typically, the dimension h of V is quite large. For example, in expansion planning problems of electric power systems, one component of V denotes the availability of one type of generators or one demand of power in a node of a multi-area supply network, or the availability of one type of transmission line connecting two nodes. Consider several nodes and arcs and one demand and some options of generators at each node. The number of scenarios K in the universe case quickly gets out of hand, even if the distribution of each component of V is determined by just a small number K^i of discrete points. Suppose, for example, $h = 20$ and $K^i = 5$, $i = 1, \ldots, 20$. Then the total number number of terms in the expected value calculations is $K = 5^{20} = 10^{14}$, which is not practically solvable by direct summation. Monte Carlo methods appear promising to compute multiple integrals or multiple sums for h large [9]. See Hammersly and Handscomb [16] for a description of Monte Carlo sampling techniques.

3.1. CRUDE MONTE CARLO

Suppose v^ω, $\omega = 1, \ldots, n$ are scenarios, sampled independently from their joint probability mass function, then $C^\omega = C(v^\omega)$ are independent random variates with expectation z.

$$\bar{z} = (1/n) \sum_{\omega=1}^{n} C^\omega \tag{6}$$

is an unbiased estimator of z and its variance is

$$\sigma_{\bar{z}}^2 = \sigma^2/n,$$

$$\sigma^2 = \mathrm{var}(C(V)).$$

Thus, the standard error is decreasing with sample size n by $n^{-0.5}$. The convergence rate of $\bar{z}$ to z is independent of the dimension h of the random vector V.

3.2. IMPORTANCE SAMPLING

We rewrite

$$z = \sum_{\omega \in \Omega} C(v^\omega) p(v^\omega) = \sum_{\omega \in \Omega} \frac{C(v^\omega) p(v^\omega) q(v^\omega)}{q(v^\omega)}$$

by introducing a probability mass function $q(v^\omega)$. We can view q as a probability mass function of a random vector W; therefore, by a change of variables,

$$z = E \frac{C(W) p(W)}{q(W)}.$$

We obtain a new estimator of z,

$$\bar{z} = \frac{1}{n} \sum_{\omega=1}^{n} \frac{C(w^\omega)p(w^\omega)}{q(w^\omega)},$$

which has a variance of

$$\text{var}(\bar{z}) = \frac{1}{n} \sum_{\omega \in \Omega} \left(\frac{C(w^\omega)p(w^\omega)}{q(w^\omega)} - z \right)^2 q(w^\omega).$$

Choosing

$$q^*(w^\omega) = \frac{C(w^\omega)p(w^\omega)}{\sum_{\omega \in \Omega} C(w^\omega)p(w^\omega)}$$

would lead to $\text{var}(\bar{z}) = 0$; that means we could obtain a perfect estimate of the multiple sum by just one single observation. However, this is practically useless, since to sample $C \cdot p/q$ we have to know q, and to determine q we need to know $z = \sum_{\omega \in \Omega} C(w^\omega)p(w^\omega)$, which we eventually want to compute. Nevertheless, this result helps us to derive heuristics of how to choose q: it should be approximately proportional to the product $C(w^\omega)p(w^\omega)$ and have a form which can be integrated analytically. For instance, using the additive (separable in the components of the stochastic vector) approximation

$$C(V) \approx \sum_{i=1}^{h} C_i(V_i)$$

could be a possible way to compute a proper q:

$$q(w^\omega) = \frac{C(w^\omega)p(w^\omega)}{\sum_{i=1}^{h} \sum_{\omega \in \Omega} C_i(w^\omega)}.$$

In this case, one has to solve only h one-dimensional sums instead of one h-dimensional sum. Depending on how well the additive model approximates the original cost surface, the above-mentioned estimator will lead to smaller variances compared to crude Monte Carlo sampling. Of course, if the original cost surface has the property of additivity (separability), no sampling is required, since the multiple sum is computed exactly by h one-dimensional sums.

The advantage of this approach lies in the fact that even if the additive model is a poor approximation to the cost surface, the method works. The price that has to be paid is a high sample size. The variance reduction compared to crude Monte Carlo will be small. For the theory of importance sampling, we refer the reader to Glynn and Iglehart [15]. See also Dantzig and Glynn [7].

Entriken and Nakayama in [8] developed an importance sampling scheme using an additive model to approximate the cost function $EC(V)$. In fact, $C(v)$ is approximated by a marginal cost model, considering marginal costs in each dimension i of V and a base case, the point from which the approximation is developed. We will use this approach here. As we employ importance sampling within the Benders decomposition algorithm, the costs $C(v, \hat{X})$, the approximation of the costs $\Gamma(v, \hat{X})$, and thus the importance distribution $q(v, \hat{X})$, depend also on $\hat{X}$, the current solution of the master problem. Introducing the costs of the base case $C(\tau, \hat{X})$ makes the model more sensitive to the impact of the stochastic variables V.

$$C(V, \hat{X}) = \Gamma(V, \hat{X}) = C(\tau, \hat{X}) + \sum_{i=1}^{h} M_i(V_i, \hat{X}),$$
$$M_i(V_i, \hat{X}) = C(\tau_1, \ldots, \tau_{i-1}, V_i, \tau_{i+1}, \ldots, \tau_h, \hat{X}) - C(\tau, \hat{X}). \tag{7}$$

$\tau = (\tau_1, \ldots, \tau_h)$ can be any arbitrarily chosen point out of the set of values v_i, $i = 1, \ldots, h$. For example, we choose τ_i as that outcome of V_i which leads to the lowest costs, ceteris paribus. These values can be found easily. Note that the second-stage costs are computed by a linear program, where the uncertain parameters appear on the right-hand side. Therefore, the second-stage costs are convex in the stochastic parameters V. The sign of the dual variables associated with the stochastic parameter indicates the direction to lowest costs. In the context of expansion planning of power systems, this means selecting, respectively, lowest demands and highest availabilities of generators and transmission lines.

Defining

$$\overline{M}_i(\hat{X}) = E\, M_i(V_i, \hat{X}) = \sum_{\omega \in \Omega_i} M_i(v_i^{\omega}, \hat{X}) p(v_i^{\omega}) \tag{8}$$

and

$$F(v^{\omega}, \hat{X}) = \frac{C(v^{\omega}, \hat{X}) - C(\tau, \hat{X})}{\sum_{i=1}^{h} M_i(v_i^{\omega}, \hat{X})}, \tag{9}$$

where we assume that

$$\sum_{i=1}^{h} M_i(v_i^{\omega}, \hat{X}) > 0, \quad \text{that means at least one } M_i(v_i^{\omega}, \hat{X}) > 0,$$

we can express the expected value of the costs in the following form:

$$z(\hat{X}) = C(\tau, \hat{X}) + \sum_{i=1}^{h} \overline{M}_i(\hat{X}) \sum_{\omega \in \Omega} F(v^{\omega}, \hat{X}) \frac{M_i(v_i^{\omega}, \hat{X})}{\overline{M}_i(\hat{X})} \prod_{j=1}^{h} p_j(v_j^{\omega}). \tag{10}$$

Note that this formulation consists of a constant term and a sum of h expectations. Given a fixed sample size n, we partition n into h sub-samples, with sample sizes

n_i, $i = 1, \ldots, h$ such that $\sum n_i = n$ and $n_i \geq 1$, $i = 1, \ldots, n$ and n_i being approximately proportional to $\overline{M}_i$. The h expectations are separately approximated by sampling using marginal densities. The ith expectation corresponds of course to the ith component of V. Generating sample points in the ith expectation, we use the importance density $(p_i M_i / \overline{M}_i)$ for sampling the ith components of V and the original marginal densities for any other components. Denoting

$$\mu_i(\hat{X}) = \frac{1}{n_i} \sum_{j=1}^{n_i} F(v^j, \hat{X}) \tag{11}$$

the estimate of the ith sum, we obtain

$$\bar{z}(\hat{X}) = C(\tau, \hat{X}) + \sum_{i=1}^{h} \overline{M}_i(\hat{X}) \mu_i(\hat{X}), \tag{12}$$

the estimated expected value of the second-stage costs $\bar{z}(\hat{X})$.

Let $\bar{\sigma}_i^2(\hat{X})$ be the estimated sample variance of the ith expectation, where $\sigma_i^2(\hat{X}) = 0$ if $n_i = 1$. The estimated variance of the mean $\sigma_{\bar{z}}^2(\hat{X})$ is then given by

$$\sigma_{\bar{z}}^2(\hat{X}) = \sum_{i=1}^{h} \frac{\overline{M}_i^2(\hat{X}) \bar{\sigma}_i^2(\hat{X})}{n_i}. \tag{13}$$

Using importance sampling, one can achieve significant variance reduction. The experiment of Nakayama in [8] claims a variance reduction of 1:20000 using importance sampling versus crude Monte Carlo sampling: for a given and optimal $\hat{X}$, the second-stage costs of a multi-area expansion planning model with 192 universe scenarios were sampled with a sample size of 10 using both methods and the results compared.

The above derivation concerned the estimation of the expected second-stage costs $z(\hat{X})$. To derive a cut, we use the same framework analogously. Note that a cut is defined as an outer linearization of the second-stage costs represented as a function of X, the first-stage variables. At $\hat{X}$, the value of the cut is exactly the expected second-stage costs $z(\hat{X})$. Note also that any choice of q is a valid choice. Since we do not want to derive different importance distributions for the coefficients and the right-hand side of a cut, we use the q already at hand from the cost estimation. Therefore, we employ directly the cost approximation scheme and the importance distribution to compute the gradient and the right-hand side of a cut. With $B(v^\omega) := B^\omega$ and $d(v^\omega) := d^\omega$ being the outcome of B and d in scenarios ω, $\omega \in \Omega$ and $\pi^*(v^\omega, \hat{X}) := \pi^{\omega *}(\hat{X})$, the optimum dual solution in scenario ω, we define

$$F^G(v^\omega, \hat{X}) = \frac{\pi^*(v^\omega, \hat{X}) B(v^\omega) - \pi^*(\tau, \hat{X}) B(\tau)}{\sum_{i=1}^{h} M_i(v_i^\omega, \hat{X})}, \tag{14}$$

$$F^g(v^\omega, \hat{X}) = \frac{\pi^*(v^\omega, \hat{X})d(v^\omega) - \pi^*(\tau, \hat{X})d(\tau)}{\sum_{i=1}^{h} M_i(v_i^\omega, \hat{X})}, \tag{15}$$

and compute

$$G(\hat{X}) = \pi^*(\tau, \hat{X})B(\tau) + \sum_{i=1}^{h} \overline{M}_i(\hat{X}) \sum_{\omega \in \Omega} F^G(v^\omega, \hat{X}) \frac{M_i(v_i^\omega, \hat{X})}{\overline{M}_i(\hat{X})} \prod_{j=1}^{h} p_j(v_j^\omega), \tag{16}$$

$$g(\hat{X}) = \pi^*(\tau, \hat{X})B(\tau) + \sum_{i=1}^{h} \overline{M}_i(\hat{X}) \sum_{\omega \in \Omega} F^g(v^\omega, \hat{X}) \frac{M_i(v_i^\omega)}{\overline{M}_i(\hat{X})} \prod_{j=1}^{h} p_j(v_j^\omega), \tag{17}$$

the coefficients and the right-hand side of a cut. We estimate the expected values again by sampling, using the same sample points as at hand from the computation of z.

Using Monte Carlo sampling, we obtain $\bar{z}(\hat{X})$, $\overline{G}(\hat{X})$, $\bar{g}(\hat{X})$, which are approximations of the expected values $z(\hat{X})$, $G(\hat{X})$, $g(\hat{X})$. We also obtain the estimated variance of the mean of the second-stage costs $\sigma_T(\hat{X})$. The impact of using approximations instead of the exact parameters on the Benders decomposition algorithm is analyzed in the following section.

4. Benders decomposition

In the following, we will derive the main steps of Benders decomposition algorithm for two-stage stochastic linear programs considering the "universe" case, which gives the exact solution of the equivalent deterministic problem ("certainty equivalent"). We will then analyze the impact of sampling of subproblems on Benders decomposition. See Geoffrion [14] for a derivation of Benders decomposition algorithm.

Given the equivalent deterministic problem in (2) and assuming K scenarios describe the universe case, we rewrite the problem applying projection onto the X variables and obtain (18). We assume for simplicity that (2) is feasible and has a finite optimum solution.

$$\begin{aligned}
\text{Minimize} \quad & Z = cX + \min[p^1 fY^1 + p^2 fY^2 + \ldots + p^K fY^K]; \\
& AX = b \qquad DY^1 \qquad\qquad\qquad\qquad = d^1 + B^1 X, \\
& X \geq 0 \qquad\qquad DY^2 \qquad\qquad\qquad = d^2 + B^2 X, \\
& \qquad\qquad\qquad\qquad \ddots \qquad\qquad \vdots \quad \vdots \\
& \qquad\qquad\qquad\qquad\qquad\qquad DY^K = d^K + B^K X, \\
& \qquad\qquad Y^1, \quad Y^2 \quad \ldots, \quad Y^K \geq 0.
\end{aligned} \tag{18}$$

The infimal value function in (18) corresponds to the following primal linear problem (19):

$$
\begin{array}{lllll}
\text{minimize} & z_P = p^1 f Y^1 + p^2 f Y^2 + \ldots + p^K f Y^K = E^{\omega}(f Y^{\omega}); & & & \\
p^1\pi^1: & DY^1 & & = d^1 + B^1 X, & \\
p^2\pi^2: & & DY^2 & = d^2 + B^2 X, & \\
\vdots & & & \vdots \quad \vdots & \\
p^K\pi^K: & & & DY^K = d^K + B^K X, & \\
& Y^1, & Y^2 \quad \ldots, & Y^K \geq 0; &
\end{array}
\tag{19}
$$

and to the dual linear problem (20):

$$
\begin{array}{lll}
\text{maximize} & z_D = p^1\pi^1(d^1 + B^1 X) + p^2\pi^2(d^2 + B^2 X) + \ldots + p^K\pi^K(d^K + B^K X); & \\
& \pi^1 D & \leq f, \\
& \pi^2 D & \leq f, \\
& \ddots & \vdots \\
& \pi^K D & \leq f.
\end{array}
$$

The primal problem is parameterized in the right-hand side by X. The assumption (2) being finite implies that (19) is finite for at least one value of X for which $X \geq 0$ and $AX = b$. Applying the Duality Theorem of Linear Programming, we state that (20) has to be feasible. The feasibility conditions

$$\pi^{\omega} D - f \leq 0$$

indicate that the feasible region $\{\pi^{\omega} \mid \pi^{\omega} D - f \leq 0\}$ is independent of X and ω, and just repeated for each scenario $\omega \in \Omega$.

The assumption (2) being feasible requires feasibility of the primal problem (19) for at least one X for which $X \geq 0$ and $AX = b$. We define $\pi := (\pi^1, \pi^2, \ldots, \pi^K)$ to be the vector of dual variables of problem (20). By the Duality Theorem again (20) has to be finite. Let π^j, $j = 1, \ldots, p$ be the extreme points and π^j, $j = p+1, \ldots, p+q$ be representatives of the extreme rays of the feasible region of (20), where $\pi^j := (\pi^{1j}, \pi^{2j}, \ldots, \pi^{Kj})$. Problem (20) is finite if and only if

$$\pi^{\omega j}(d^{\omega} + B^{\omega} X) \leq 0, \quad j = p+1, \ldots, p+q, \quad \omega \in \Omega. \tag{21}$$

Constraints (21) may be appended to problem (18) to ensure that the problem is bounded.

Next, we outer linearize the infimal value function in (18), whose value is exactly

$$\underset{j=1,\ldots,p}{\text{maximum}} \sum_{\omega \in \Omega} p^{\omega} \pi^{\omega j} (d^{\omega} + B^{\omega} X). \tag{22}$$

By expressing the infimal value function by the outer linearized dual problem and using θ as the smallest upper bound, the problem can be represented in the following form:

$$\begin{aligned} &\text{minimize} \quad Z = cX + \theta; \\ &\qquad\qquad AX \quad = b, \\ &\qquad\qquad X \quad \geq 0, \\ &\theta \geq \sum_{\omega \in \Omega} p^{\omega} \pi^{\omega j} (d^{\omega} + B^{\omega} X), \quad j = 1, \ldots, p, \\ &\pi^{j} (d^{\omega} + B^{\omega} X) \leq 0, \qquad\qquad j = p+1, \ldots, p+q, \quad \omega \in \Omega. \end{aligned} \tag{23}$$

Relaxation is applied to solve problem (23) since we do not want to know all $\pi^j, j = 1, \ldots, p+q$ in advance: given a solution $(\hat{X}, \hat{\theta})$ from the master problem, one solves problem (19) or problem (20) in fact by solving the individual problems (4) or the dual problems (24) of these:

$$\begin{aligned} &z^{\omega^*}(\hat{X}) = \max z_D^{\omega} = \pi^{\omega}(d^{\omega} + B^{\omega}\hat{X}); \\ &\qquad\qquad \pi^{\omega} D \qquad\qquad \leq f, \quad \omega \in \Omega. \end{aligned} \tag{24}$$

We call $\pi^{\omega^*}(\hat{X})$ the optimum dual solution vector. If primal infeasibility or dual unboundedness is detected, with $\pi^{\omega^*}(\hat{X})$ denoting the corresponding extreme ray, a feasibility cut

$$\pi^{\omega^*}(\hat{X}) \cdot (d^{\omega} + B^{\omega} X) \leq 0 \tag{25}$$

is added to the master problem. If all primal problems are feasible or all dual problems are bounded, an optimality cut

$$\theta \geq \sum_{\omega \in \Omega} p^{\omega} \pi^{\omega^*}(\hat{X}) \cdot (d^{\omega} + b^{\omega} X) \tag{26}$$

is added to the master problem. We call

$$L(X) := \sum_{\omega \in \Omega} p^{\omega} \pi^{\omega^*}(\hat{X}) \cdot (d^{\omega} + B^{\omega} X) \tag{27}$$

an outer linearization of the second-stage costs, which are defined by

$$z(\hat{X}) := \sum_{\omega \in \Omega} z^{\omega}(\hat{X}). \tag{28}$$

The relation

$$L(X) \leq z(X) \tag{29}$$

formulates the main property of the outer linearization. Any cut independent of $\hat{X}$ from which it was originally derived is a valid cut as long as it does not violate the main property of outer linearization.

Benders decomposition algorithm provides upper and lower bounds to the solution in each iteration.

In the lth iteration

$$LB^l := c\hat{X}^l + \hat{\theta}^l, \tag{30}$$

with $\hat{X}^l$, $\hat{\theta}^l$ being the optimum solution of the master problem, is defined to be a lower bound and

$$UB^l := \min\{UB^{l-1}, c\hat{X}^l + z(\hat{X}^l)\}, \quad UB^0 = \infty, \tag{31}$$

with $z(\hat{X}^l)$ the second-stage costs, to be an upper bound to the solution of the problem. If

$$UB^l - LB^l)/LB^l \leq TOL, \tag{32}$$

where TOL is a given tolerance, the problem is said to be solved with a sufficient accuracy.

4.1. PROBABILISTIC CUTS

Employing Monte Carlo sampling techniques means not to solve all problems $\omega \in \Omega$, but solving problems $\omega \in S$, S being a subset of Ω. Instead of the exact expected values $z(\hat{X})$, $G(\hat{X})$, $g(\hat{X})$, we compute the estimates $\bar{z}(\hat{X})$, $\bar{G}(\hat{X})$, $\bar{g}(\hat{X})$ by importance sampling. We also estimate the error of the estimation of $z(\hat{X})$ by the variance $\text{var}(\bar{z}(\hat{X})) = \sigma_{\bar{z}}^2(\hat{X})$. Thus, e.g. in the case of the second-stage costs, the estimation results in an estimated mean with some error distribution. There is good reason to assume the error being normally distributed [9]. We define $\tilde{z}(\hat{X})$ to be random, normally distributed with mean $\bar{z}(\hat{X})$ and variance $\sigma_{\bar{z}}^2(\hat{X})$:

$$\tilde{z}(\hat{X}) := N(\bar{z}(\hat{X}), \sigma_{\bar{z}}^2(\hat{X})). \tag{33}$$

A cut obtained by sampling differs in general from a cut computed by solving the universe scenarios. The outer linearizations $L(X) = G(\hat{X})X + g(\hat{X})$ with respect to

the universe case, and $\bar{L}(X) = \bar{G}(\hat{X})X + \bar{g}(\hat{X})$ with respect to the estimation, differ in the gradient and the right-hand side. At $X = \hat{X}$, where $L(\hat{X}) = z(\hat{X})$ and $\bar{L}(\hat{X}) = \bar{z}(\hat{X})$, we substitute the variable θ for $\tilde{z}(\hat{X})$ when defining a cut. By this substitution, θ takes on the distribution of $\tilde{z}$; therefore, $\theta := N(\bar{\theta}, \sigma_{\tilde{z}}^2)$. This is only true at $X = \hat{X}$. However, we assume this error distribution to be constant with respect to X. This means we see the error mainly concentrated on the right-hand side of the cut and we assign the variance $\sigma_{\tilde{z}}(\hat{X})$ also to the right-hand side and define

$$\tilde{g}(\hat{X}) := N(\bar{g}(\hat{X}), \sigma_{\tilde{z}}(\hat{X})) \tag{34}$$

to be the random right-hand side of the cut, normally distributed with mean $\bar{g}(\hat{X})$ and $\sigma_{\tilde{z}}^2(\hat{X})$. We can expect that in the final solution, cuts will be binding at an X very close to $\hat{X}$, where they were originally derived. The assumption of a constant error distribution of θ is therefore intuitively plausible. See also Dantzig and Glynn [7] in this respect. In general, S is a sufficiently large subset of Ω so that the variance $\sigma_{\tilde{z}}^2$ is small.

Cuts computed by sampling do not necessarily meet the condition of outer linearization. Violating this condition, a cut intersects and separates parts of the feasible region of the second-stage problem. A sampled cut is therefore not a valid cut.

4.2. UPPER AND LOWER BOUNDS

For random second-stage costs $\tilde{z}(\hat{X}^l)$ and random right-hand sides $\tilde{g}^l$, $l = 1, \ldots, L$, the upper and lower bounds of the problem as provided by Benders decomposition algorithm are probabilistic.

The upper bounds

$$\tilde{UB}^l := c\hat{X}^l + \tilde{z}(\hat{X}^l), \qquad l = 1, \ldots, L \tag{35}$$

are random parameters, normally distributed with means $\bar{UB}^l$ and variances $\sigma_{\tilde{z}}(\hat{X}^l)$:

$$\tilde{UB}^l := N(\bar{UB}^l, \sigma_{\tilde{z}}^2(\hat{X}^l)), \qquad l = 1, \ldots, L. \tag{36}$$

We define the lowest upper bound to be the upper bound with the lowest mean

$$\tilde{UB}_{\min}^L : \bar{UB}_{\min}^L := \min_{l=1,\ldots,L} \{\bar{UB}^l\} \tag{37}$$

with corresponding variance $\sigma_{\tilde{UB}_{\min}^L}^2$.

The lower bounds are obtained from the solution of the master problem. To determine the distribution of a lower bound, consider the master problem at iteration L:

$$
\begin{aligned}
\bar{LB}^L = \bar{z}_M^{*L} = \min \bar{z}_M^L = \quad & cX + \theta, \\
\text{subject to } \rho_1^0: \quad & AX = b, \\
\rho^1: \quad & -G^1X + \theta \geq \bar{g}^1, \\
& \vdots \qquad\qquad \vdots \\
\rho^L: \quad & -G^LX + \theta \geq \bar{g}^L, \\
& X, \quad \theta \geq 0,
\end{aligned}
$$

where L optimality cuts have been added to the originally relaxed master problem. We do not consider feasibility cuts for the following argument, since they are exact. The vector ρ^0 and the scalars ρ^l, $l = 1, \ldots, L$, denote the dual prices. The right-hand sides $\bar{g}^l$, $l = 1, \ldots, L$, are independent stochastic parameters, normally distributed. We assume independence since the cuts are generated from independent samples, neglecting the dependency that $\hat{X}^l$, $l = 1, \ldots, L$, are weakly connected by Benders decomposition algorithm.

With the random parameters $\bar{g}^l$, $l = 1, \ldots, L$, on the right-hand side, also the optimum solution $\bar{z}_M^*$ will be random. We define the optimum solution of the master problem

$$
\bar{z}_M^{*L} := N\left(\bar{z}_M^L, \operatorname{var}\left(\bar{z}_M^{*L}\right)\right) \tag{39}
$$

to be a random parameter, normally distributed, with mean $\bar{z}_M^L$ and variance $\operatorname{var}(\bar{z}_M^{*L})$. Hence, one could experimentally obtain the distribution of $\bar{z}_M^{*L}$ by randomly varying the right-hand sides according to N samples, $j = 1, \ldots, N$, drawn from the normal distributions of $\bar{g}^l$, $l = 1, \ldots, L$, and by solving the master problem for all N samples. One could estimate the mean and the variance of the distribution from the samples $j = 1, \ldots, N$. Since this is a very expensive way to obtain an estimate of the lower bound distribution, we proceed instead in the following way. We have already stated that we choose a sample size $|S|$, such that the variances σ_T^l, $l = 1, \ldots, L$, are small. If the variances are small, we can assume that for all outcomes of the random right-hand sides $\bar{g}^l$, $l = 1, \ldots, L$, the optimum solution of the master problem has the same basis. Then we can compute the mean of the lower bound estimate

$$
\begin{aligned}
\bar{z}_M^{*L} = \min \; z_M = \quad & cX + \theta, \\
\text{subject to } \rho^0: \quad & AX = b, \\
\rho^1: \quad & -G^1X + \theta \geq \bar{g}^1, \\
& \vdots \qquad\qquad \vdots \\
\rho^L: \quad & -G^LX + \theta \geq \bar{g}^L, \\
& X, \quad \theta \geq 0,
\end{aligned}
$$

by substituting the means $\bar{g}^l$, $l = 1, \ldots, L$, for the random parameters $\tilde{g}^l$, $l = 1, \ldots, L$, and the variance $\text{var}(\tilde{z}_M^*)$ by using the dual solution

$$\text{var}(\tilde{z}_M^*) = \sum_{l=1}^{L} \rho^{l^2} \text{var}(\tilde{g}^l) = \sum_{l=1}^{L} \rho^{l^2} \sigma_T^2(\hat{X}^l). \tag{41}$$

As the lower bound means increase monotonically with the number of iterations, we obtain the largest lower bound by $\bar{LB}^L = \bar{z}_M^{*L}$ and $\tilde{LB}^L := N(\bar{LB}^L, \text{var}(\tilde{LB}^L))$.

4.3. STOPPING RULE

In analogy to the deterministic Benders decomposition algorithm, we stop if the upper and lower bound are sufficiently close. In the case of probabilistic bounds, the algorithm has to be stopped if the upper and lower bound are indistinguishable in distribution. We check this condition by using Student's t-test to determine if $s^l > 0$ with 95% probability, where

$$s^l = \bar{UB}^l - \bar{LB}^l + TOL \tag{42}$$

and TOL being a given tolerance.

The employment of Student's t-test requires independency of the upper and lower bound distributions. Since independency is not ensured in the first place as an upper bound and a binding cut in the master problem could be obtained from the same set of samples, we obtain independency by resampling the lowest upper bound before employing Student's t-test. The $\hat{X}$ corresponding to the lowest upper bound and the corresponding importance distribution have to be stored. If upper and lower bounds are close to each other, which is checked by using Student's t-test without fulfilling the independence requirement, we use new samples to compute an independent upper bound. Now we check if $s^l > 0$ by Student's t-test.

4.4. CONFIDENCE INTERVAL

After passing the Student's t-test in the last iteration, which means that the upper and lower bound means are indistinguishable, we obtain the optimum solution $\hat{X}^L, \hat{\theta}$ from the master problem. We derive from the distributions $\tilde{LB}^L$ and $\tilde{UB}^L$ a 95% confidence interval: on the left-hand side by using the lower bound distribution and on the right-hand side by using the upper bound distribution. We define

$$C_{\text{left}} = 1.96\sqrt{\text{var}(\tilde{LB}^L)}, \quad C_{\text{right}} = 1.96\sqrt{\text{var}(\tilde{UB}^L)}, \tag{43}$$

and obtain the confidence interval

$$\bar{LB} - C_{\text{left}} \leq Z^* \leq \bar{UB} + C_{\text{right}} \tag{44}$$

for the final solution Z^*.

If $(C_{left} + C_{right})/\underline{L}B^l \leq C_{tol}$, where C_{tol} is a predefined quality criteria for the confidence interval, the obtained solution is satisfactory. Otherwise, the sample size has to be increased and the problem has to be solved again with the increased sample size.

4.5. IMPROVEMENT OF THE SOLUTION

Suppose the solution with a certain sample size was not satisfactory. Instead of starting from the beginning with an increased sample size, we want to use the information that we have already collected. To do this, we look for the binding cuts in the final solution, increase the sample size and recompute the binding cuts at the same $\hat{X}^l$ they were originally computed. This of course means that one has to store the values of $\hat{X}^l$ and the associated importance distributions, or recompute the latter. The enlarged sample size leads to smaller variances of the binding cuts and eventually to a smaller confidence interval of the final solution. Berry–Esséen, e.g. Hall [17], give upper bounds on the rates of convergence in the central limit theorem. Solving the master problem again with the improved binding cuts will not in general result in an indistinguishable lower and upper bound. Therefore, some more iterations may be necessary to obtain the optimal solution according to the increased sample size. This improvement procedure could be employed iteratively until a satisfactory solution is obtained. It is a possible way to improve a non-satisfactory solution. It may not be very efficient and there may be better ways to do so. In general, we choose a sample size such that the obtained confidence interval is satisfactory. We can now state the algorithm as follows.

4.6. THE ALGORITHM

Step 0 Initialize:

$$l = 0, \quad \overline{U}B^0 = \infty.$$

Step 1 Solve the relaxed master problem and obtain a lower bound:

$$\underline{L}B^l = c\hat{X} + \hat{\theta}^l.$$

Step 3 $l = l + 1$.

Solve subproblems and obtain an upper bound:

$\overline{U}B^l = \min\{\overline{U}B^{l-1}, c\hat{X}^l + \overline{z}(\hat{X}^l)\}$, compute and add a cut to the master problem using Monte Carlo (importance) sampling.

Step 3 Solve the master problem and obtain a lower bound:

$$\underline{L}B^l = c\hat{X}^l + \overline{\theta}^l.$$

Step 4 $s = \overline{U}B^l - \overline{L}B^l + TOL$

If $s > 0$ (Student's t-test), go to step 2.

Step 5 Compute the confidence interval and obtain a solution: Z^*, $\hat{X}$, $\hat{\theta}$. Stop.

Improvement of the solution:

Step 6 If $(C_{left} + C_{right})/\overline{L}B \leq C_{tol}$, stop;
otherwise, go to step 7.

Step 7 Increase sample size and initialize $\overline{U}B^0 = \infty$.

Step 8 Recompute binding cuts.
Upper bound: $\overline{U}B^l = \min\{\overline{U}B^{l-1}, C\hat{X} + \bar{\tau}(\hat{X}^l)\}$.

Step 9 Go to step 3.

5. Numerical results

The method has been implemented. The FORTRAN code for solving general large-scale two-stage stochastic linear problems with recourse using Benders decomposition and importance sampling uses MINOS [24], which has been adapted for this purpose, as a subroutine for solving the linear programs of the master problem and the subproblems. Alternatively, the code can also use a modified version of Tomlin's [31] LPM1 code of the revised simplex method as a subroutine. Versions of the code are installed on several computers, such as IBM-3090, a Microvax workstation, and on personal computers. All the following test results were computed on a Toshiba laptop personal computer T5200. First, we present an illustrative example, a toy problem of expansion planning of power systems, which we discuss in detail. Then we derive numerical results from other small test problems. Eventually, we demonstrate the solution of large-scale test problems with numerous stochastic parameters.

The illustrative example, test problem APL1P, is a model of a simple power network with one demand region. There are two generators with different investment and operating costs, and the demand is given by a load duration curve with three load levels: base, medium, and peak. We index the generators with $j = 1, 2$, and the demands with $i = 1, 2, 3$. The variables x_j, $j = 1, 2$, denote the capacities which can be built and operated to meet demands d_i, $i = 1, 2, 3$. The variable y_{ij} denotes the operating level for generator j in load level i with operating cost f_{ij}. The variable y_{is} defines the unserved demand in load level i which can be purchased with penalty cost $f_{is} > f_{ij}$. The subscript s is not an index, but denotes only an unserved demand variable. The per-unit cost to build generator j is c_j. Finally, the model is formulated with complete recourse, which means that at any given choice of x, demand is satisfied for all outcomes. In this model, building new generators competes with

Table 1

APL1P test problem data.

Generator capacity costs (10^5\$/MW, a)

$c_1 = 4.0,\ c_2 = 2.5$

Generator operating costs (10^5\$/MW, a)

$f_{11} = 4.3 \quad f_{21} = 8.7$
$f_{12} = 2.0 \quad f_{22} = 4.0$
$f_{13} = 0.5 \quad f_{23} = 1.0$

Unserved demand penalities (10^5\$/MW, a)

$f_{1s} = f_{2s} = f_{3s} = 10.0$

Minimum generator capacities (MW)

$b_1 = b_2 = 1000$

Demands (MW)

No.	1	2	3	4
Outcome	900	1000	1100	1200
Probability	0.15	0.45	0.25	0.15

Availabilities of generators

Generator 1 (β_1)

No.	1	2	3	4
Outcome	1.0	0.9	0.5	0.1
Probability	0.2	0.3	0.4	0.1

Generator 2 (β_2)

No.	1	2	3	4	5
Outcome	1.0	0.9	0.7	0.1	0.0
Probability	0.1	0.2	0.5	0.1	0.1

purchasing unserved demand through the cost function, yet there is a minimum capacity b_j which has to be built for each load level. The availabilities of the two generators β_j, $j = 1, 2$, and the demands in each load level d_i, $i = 1, 2, 3$, are uncertain. Generator one has four possibilities, while generator two has five, and each demand has four. All of the data values are given in table 1 and the problem can be formulated as follows:

$$\text{minimize} \quad \sum_{j=1}^{2} c_j x_j + E\left\{ \sum_{j=1}^{2} \sum_{i=1}^{3} f_{ij} y_{ij}^{\omega} + \sum_{i=1}^{2} f_{is} y_{is}^{\omega} \right\}$$

$$\text{subject to} \quad x_j \geq b_j, \qquad j = 1, 2,$$

$$-\alpha_j^{\omega} x_j + \sum_{i=1}^{3} y_{ij}^{\omega} \leq 0, \qquad j = 1, 2,$$

$$\sum_{j=1}^{2} y_{ij}^{\omega} + y_{is}^{\omega} \geq d_i^{\omega}, \qquad i = 1, 2, 3,$$

$$x_j,\ y_{ij}^{\omega},\ y_{is}^{\omega} \qquad j = 1, 2,\ i = 1, 2, 3.$$

We will take $\omega \in \Omega$ when solving the universe problem and $\omega \in S$ when solving a problem with sampling.

The number of possible demands and availabilities results in $4*5*4^3 = 1280$ possible outcomes in Ω, and thus 1280 subproblems have to be solved in each iteration of Benders decomposition for the universe case. We compare the universe solution with solutions gained by the importance sampling algorithm. Table 2 shows the results in the case of 20 samples out of the possible 1280 combinations and

Table 2

Model APL1P, 20 samples (100 replications of the experiment).

	Correct	Mean	95% conf [%]	Bias [%]
No. univ	1280			
No. iter		7.6		
G1	1800.0	1666.5	57.0	−7.4
G2	1571.4	1732.5	52.5	10.2
θ	13513.7	13729.4	21.3	1.6
Obj	24642.3	24726.7	2.1	0.3
Est. conf [%]	left	1.5		
Est. conf [%]	right	1.9		
Coverage		0.90		

without an improvement phase. One hundred replications of the same experiment with different seeds were run to obtain statistical information about the accuracy of the solution and the estimated confidence interval. The mean over the 100 replications of the objective function value (total costs) differs from the universe solution by 0.3%. From the distribution of the optimum objective function value derived from the 100 replications of the experiment, a 95% confidence interval is computed: ±2.1%. In each replication, a 95% confidence interval of the solution is estimated. The mean over all replications of the estimated confidence interval is

1.5% on the left-hand side and 1.9% on the right-hand side. In the worst case, an objective function value of 26233.9 was computed. This is about 6.4% off the correct answer. The estimated 95% confidence interval in this case did not cover the correct answer. The coverage rate of 90% expresses that in 90% of the 100 replications, the correct answer of the universe solution is covered by the estimated confidence interval. This shows that if we use a sample size of 20, we are slightly underestimating the confidence interval; if the computation of the 95% confidence interval was exact, we would expect a coverage rate of 95%. The reason for the underestimation of the 95% confidence interval in the case of sample size 20 lies in the underlying assumptions of the estimation method, e.g. constant error distribution along a cut, same basis for all outcomes of the random right-hand sides of the cuts. Especially the latter assumption is only true if the variances are small. A larger sample size reduces the variances and we expect a better coverage rate of the 95% confidence interval. The bias and the confidence interval of the optimum strategies (the loads x to be installed) are larger than those of the optimum objective function value. The objective function near the optimal solution appears to be flat: several different strategies lead close to the optimum costs. Confidence intervals of about 57% and 52% are computed. In the above example, a sample size of 20 was chosen. Note that additional computational effort is also needed to obtain the importance distribution, e.g. 17 subproblems have to be solved in each iteration to obtain the marginal costs M_i. Compared to the universe solution, the method e.g. achieves, with about 2.9% of computational effort, a solution which has 95% confidence within an interval of ±2.1% of the correct answer. Importance sampling seems to be a promising approach to solving stochastic linear programs. Table 3 represents the results when using 200 samples: one can see decreasing bias, decreasing confidence intervals, and improving estimations of the confidence intervals with increased

Table 3

Model APL1P, 200 samples (100 replications of the experiment).

	Correct	Mean	95% conf [%]	Bias [%]
No. univ	1280			
No. iter		7.9		
G1	1800.0	1728.7	31.5	-4.0
G2	1571.4	1681.7	29.2	7.0
θ	13513.7	13554.7	12.2	0.3
Obj	24642.3	24673.8	0.4	0.1
Est. conf [%]	left	0.4		
Est. conf [%]	right	0.7		
Coverage		0.95		

sample size. The coverage of the 95% confidence interval, computed by 100 replications of the experiment with different seeds, is now 95%.

We investigated the performance of the algorithm on two other examples which are small enough to compute the universe solution. PGP2, derived from Louvaux [22], is a power generation planning model used to determine the capacities of various types of equipment required to ensure that consumer demand is met. The demands in three demand regions are stochastic and represented by discrete random variables with nine, nine and eight outcomes. CEP1 is a capacity planning model for a manufacturing plant in which several parts are produced on several machines. If the demand for the parts exceeds the production capability, the residual parts are purchased from external sources at a price much higher than the production costs to meet the demand. There are three stochastic parameters (demands for parts) with discrete and uniform distributions with ten outcomes each. The formulations and data for CEP1 and PGP2 may be found in Higle et al. [19].

In the case of PGP2, we obtained very accurate results using a sample size of 50. By computing 100 replications of the experiment, the mean of the objective function values differs by 0.1% from the correct answer. The 95% confidence interval of the objective function value, computed by the 100 replications of the experiment, is ±0.76%, the mean of the confidence intervals estimated in each replication is 0.62% on the left-hand side and 0.9% on the right-hand side. In 98% of instances, the correct solution is covered by the 95% confidence interval. In the worst case, the solution differed by 0.77% from the correct answer and was not covered by the 95% confidence interval.

In the case of CEP1, a high sample size is needed to obtain accurate results. The estimation of the second-stage costs appears to be more difficult. The reason lies in the fact that the (penalty) costs of buying parts from external sources are much higher than the costs of production. For this problem, the additive approximation function is not a very good approximation of the true cost function since it does not cover the very high costs in scenarios where all three demands are high. The estimated confidence interval seems to be large; we computed 4.65% on the left-hand side and 4.62% on the right-hand side (mean over 100 replications of the experiment). The estimations of the confidence interval are accurate, as indicated by the coverage rate of 95% of the correct answer by the 95% confidence interval. In the worst case, a different of 8.07% of the objective function value to the correct answer was computed. The worst case solution is not covered by the estimated confidence interval. In these examples, it is easier to compute the value of the first-stage variables than to estimate the second-stage costs. In most cases, the correct answer of the first-stage variables was obtained. We have developed methods which adaptively improve the approximation function if sample information shows that the variance of the estimation is too large. A discussion of the adaptive approach is not the subject of this paper. Tables 4 and 5 represent the computational results of PGP2 and CEP1 and show the sizes of the test problems.

Table 4

Model PGP2, 50 samples (100 replications of the experiment).

	Correct	Mean	95% conf [%]	Bias [%]
No. univ	648			
No. iter		9.1		
Obj	392.2	392.5	0.76	0.1
Est. conf [%]	left	0.62		
Est. conf [%]	right	0.9		
Coverage		0.98		
Comp. time [min]		0.28		
Problem size				
Master:	rows	3		
	columns	7		
	nonzeros	16		
Sub:	rows	8		
	columns	16		
	nonzeros	52		

Table 5

Model CEP1, 200 samples (100 replications of the experiment).

	Correct	Mean	95% conf [%]	Bias [%]
No. univ	1000			
No. iter		6.4		
Obj	57790.7	58832.7	4.63	1.8
Est. conf [%]	left	4.65		
Est. conf [%]	right	4.62		
Coverage		0.95		
Comp. time [min]		0.28		
Problem size				
Master:	rows	12		
	columns	10		
	nonzeros	36		
Sub:	rows	9		
	columns	16		
	nonzeros	53		

In the following, we report on the solution of some large test problems with several stochastic parameters which are too large to be solved by computing the universe solution.

WRPM is a prototype multi-area capacity expansion planning problem for the western USA and Canada. The model is detailed, covering six regions, three demand blocks, two seasons, and several kinds of generation and transmission technologies. The objective is to determine optimum discounted least-cost levels of generation and transmission facilities for each region of the system over time. The model minimizes the total discounted costs of supplying electricity (investment and operating costs) to meet the exogenously given demand subject to expansion and operating constraints. A description of the model can be found in Dantzig et al. [8] and Avriel et al. [1]. In the stochastic version of the model, the availabilities of generators and transmission lines and demands are subject to uncertainty. There are thirteen stochastic parameters per time period (eight stochastic availabilities of generators and transmission lines and five uncertain demands) with discrete distributions with three or four outcomes. The operating subproblems in each period are stochastically independent. The test problem WRPM1 covers a time horizon of one future period and WRPM2 covers two future periods. There are differences in the parameters between WRPM1 and WRPM2. Note that in the deterministic equivalent formulation, the problem would have more than 1.5 billion (WRPM1) and more than 3 billion (WRPM2) equations.

FI2 is a portfolio management test problem, formulated as a network problem. It is a modified version of test problems found in Mulvey and Vladimirou [23]. The problem is to select a portfolio which maximizes expected returns in future periods, taking into account the possibility of revising the portfolio in each period. There are also transaction costs and bounds on the holdings and turnovers. The test problem FI2 covers a planning horizon of two future periods. The returns of the stocks in the two future periods are stochastic parameters. The problem is formulated as a two-stage problem. Rather than solving the problem by looking at a certain number of preselected scenarios (18 to 72 in the case of Mulvey and Vladimirou), we instead assumed the returns of the stocks in the future periods to be independent random parameters, discretely distributed with three outcomes each. Since there are thirteen stocks with uncertain returns, the problem has twenty-six stochastic parameters. The universe number of scenarios (2.5×10^{12}) is very large, so that the deterministic equivalent formulation of the problem has more than 10^{14} rows. The stochastic parameters appear in the B-matrix as well as in the D-matrix.

Computational results of the large-scale test problems are represented in table 6. In addition to the solution of the stochastic problems, table 6 also shows the results of solving the expected value problem. In this case, the stochastic parameters are substituted by their expectations to obtain a deterministic problem. The expected value solution is then used as a starting point for the stochastic solution. We also report on the estimated expected costs of the expected value solution. These are the total expected cost which would occur if the expected value solution is implemented

Table 6

Large test problems: computational results.

		WRPM1	WRPM2	FI2
No. iter stoch. (exp. val.)		139 (82)	131 (83)	4 (2)
Sample size		100	100	200
Exp. val. solution obj		286323.2	140041.0	1.0766
Exp. val. solution, exp. cost		295473.7	147227.3	1.172
Stochastic solution		289644.2	143109.2	1.169
Est.	conf. left [%]	0.0913	0.0962	0.454
	conf. right [%]	0.063	0.1212	0.371
Solution time [min]		75	187	2
Problem size				
Master:	rows	44	86	48
	columns	76	151	33
	nonzeros	153	334	130
Sub:	rows	302	302	61
	columns	289	289	45
	nonzeros	866	866	194
No. univ. scenarios		5038848	10077696	2.5×10^{12}

in a stochastic environment. The objective function value of the true stochastic solution has to lie between the objective function value of the expected value solution and the expected costs of the expected value solution.

In the case of WRPM1 and WRPM2, we chose a sample size of 100. The estimate of the objective function value of the stochastic solution (289644.2 in the case of WRPM1 and 143109.2 in the case of WRPM2) turns out to be amazingly accurate. The 95% confidence interval was computed at 0.0913% on the left-hand side and 0.063% on the right-hand side (WRPM1) and 0.0962% on the left-hand side and 0.1212% on the right-hand side (WRPM2). Thus, the objective function value of the stochastic solution lies with 95% probability within $289379.7 \leq z^* \leq 289826.0$ (WRPM1) and $142971.5 \leq z^* \leq 143282.6$ (WRPM2). In both cases, the expected costs of the expected value solution and the expected costs of the stochastic solution differ significantly. The solution time on a Toshiba T5200 laptop PC with a 80387 mathematic coprocessor was 75 minutes (WRPM1) and 187 minutes (WRPM2). During this time, about 7500 (WRPM1) and 15700 (WRPM2) subproblems (linear programs of the size of 302 rows and 289 columns) are solved.

A sample size of 200 was chosen for solving test problem FI2. The problem is solved in only four iterations. The objective function value of the stochastic solution is computed as 1.1695 with a 95% confidence interval of 0.454% on the left-hand side and 0.371% on the right-hand side. Thus, with 95% probability the

optimal solution lies between $1.164 \leq z^* \leq 1.174$. The estimated expected costs of the expected value solution (1.172) lie within the 95% confidence interval of the costs of the stochastic solution; however, also in this case, expected costs of the expected value solution and expected costs of the stochastic solution differ significantly.

6. Conclusion

We have discussed a promising approach to solving two-stage stochastic linear programs with recourse and obtained first numerical results; employing importance sampling within the Benders decomposition algorithm, we obtained very accurate solutions to the test problems with only a small sample size. The technique enables us to solve large-scale problems with a large number of stochastic parameters on a laptop computer. The test problems solved so far include up to twenty-six stochastic parameters and the subproblems have a size of several hundred rows and columns. In the deterministic equivalent formulation, the problems would have more than several billions of equations. The small confidence intervals of the solutions indicate that an extension to even more stochastic parameters is possible. The analysis in this paper concentrated on discrete distributions. The method, however, can be easily extended to continuous distributions. Current research concentrates on testing the technique on large-scale problems of different areas with large numbers of stochastic parameters. We investigated possibilities to adaptively improve the approximation function should it prove that the error of the estimate exceeds a predefined level. If some problems require a much higher sample size, the use of parallel processors will enable us to quickly solve large numbers of samples to obtain low variances of the estimations. A parallel implementation of the method is in preparation.

Acknowledgements

Planning under uncertainty and solving stochastic problems by combining decomposition techniques, sampling techniques, and parallel processors is a theme composed by Professor George B. Dantzig. The research leading to this paper was conducted while the author was a visiting scholar at the Department of Operations Research at Stanford University. The authors wishes to thank Professor George B. Dantzig for his outstanding personal and professional support during this visit. The author also wants to thank Peter Glynn and John Stone for valuable discussions on this topic. The author is grateful to James K. Ho, R.P. Sundarraj and Kingsley Gnanendran for their help in adapting the LPM1 optimizer for our purpose, to Manuel Nunez for his assistance in adapting the MINOS software, and to Alamuru Krishna who assisted in preparing some of the test problems. The author is grateful to Mordecai Avriel, Julie Higle, Suvrajeet Sen and anonymous referees for helpful suggestions concerning previous versions of this paper.

References

[1] M. Avriel, G.B. Dantzig and P.W. Glynn, Decomposition and parallel processing techniques for large scale electric power system planning under uncertainty, *Proc. Workshop on Resource Planning Under Uncertainty*, Stanford University (1989).

[2] J.F. Benders, Partitioning procedures for solving mixed-variable programming problems, Numer. Math. 4(1962)238–252.

[3] J.R. Birge, Decomposition and partitioning methods for multi-stage stochastic linear programming, Oper. Res. 33(1985)989–1007.

[4] J.R. Birge and S.W. Wallace, A separable piecewise linear upper bound for stochastic linear programs, SIAM J. Control Optim. 26(1988)3.

[5] J.R. Birge and R.J. Wets, Designing approximation schemes for stochastic optimization problems, in particular for stochastic programs with recourse, Math. Progr. Study 27(1986)54–102.

[6] G.B. Dantzig, Linear programming under uncertainty, Manag. Sci. 1(1955)197–206.

[7] G.B. Dantzig and P.W. Glynn, Parallel processors for planning under uncertainty, Ann. Oper. Res. 22(1990)1–21.

[8] G.B. Dantzig, P.W. Glynn, M. Avriel, J. Stone, R. Entriken and M. Nakayama, Decomposition techniques for multi-area generation and transmission planning under uncertainty, EPRI Report 2940-1 (1989).

[9] P.J. Davis and P. Rabinowitz, *Methods of Numerical Integration* (Academic Press, London, 1984).

[10] Y. Ermoliev, Stochastic quasi-gradient methods and their applications to systems optimization, Stochastics 9(1983)1–36.

[11] Y. Ermoliev and R.J. Wets (eds.), *Numerical Techniques for Stochastic Optimization* (Springer, 1988).

[12] K. Frauendorfer, Solving SLP recourse problems with arbitrary multivariate distributions – the dependent case, Math. Oper. Res. 13(1988)377–394.

[13] K. Frauendorfer and P. Kall, Solving SLP recourse problems with arbitrary multivariate distributions – the independent case, Probl. Control Inf. Theory 17(1988)177–205.

[14] A.M. Geoffrion, Elements of large-scale mathematical programming, Manag. Sci. 16, no. 11 (1974).

[15] P.W. Glynn and D.L. Iglehart, Importance sampling for stochastic simulation, Manag. Sci. 35(1989) 1367–1392.

[16] J.M. Hammersly and D.C. Handscomb, *Monte Carlo Methods* (Methuen, London, 1964).

[17] P. Hall, Rates of convergence in the central limit theorem, Res. Notes Math. 62 (1982).

[18] J.L. Higle and S. Sen, Stochastic decomposition: An algorithm for two stage linear programs with recourse, Math. Oper. Res. (1989), to appear.

[19] J.L. Higle, S. Sen and D. Yakowitz, Finite master programs in stochastic decomposition, Technical Report, Department of Systems and Industrial Engineering, University of Arozina, Tucson, AZ 85721.

[20] P. Kall, Computational methods for two stage stochastic linear programming problems, Z. angew. Math. Phys. 30(1979)261–271.

[21] P. Kall, A. Ruszczynski and K. Frauendorfer, Approximation techniques in stochastic programming, in: *Numerical Techniques for Stochastic Optimization*, ed. Y. Ermoliev and R.J. Wets (1988).

[22] F.V. Louveaux and Y. Smeers, Optimal investment for electricity generation: A stochastic model and a test problem, in: *Numerical Techniques for Stochastic Optimization*, ed. Y. Ermoliev and R.J. Wets (1988).

[23] J.M. Mulvey and H. Vladimirou, Specifying the progressive hedging algorithm to stochastic generalized networks, Sstatistics and Operations Research Series Report SOR-89-9, Princeton University, Princeton, NJ 08544 (1989).

[24] B.A. Murtagh and M.A. Saunders, MINOS user's guide, SOL 82-20, Department of Operations Research, Stanford University, Stanford, CA 94305 (1983).

[25] L. Nazareth and R.J. Wets, Algorithms for stochastic programs: The case of nonstochastic tenders, Math. Progr. Study 28(1987)48–62.
[26] M.V. Pereira, L.M.V.G. Pinto, G.C. Oliveira and S.H.F. Cunha, A technique for solving LP-problems with stochastic right-hand sides, CEPEL, Centro del Pesquisas de Energia Electria, Rio de Janeiro, Brazil (1989).
[27] R.T. Rockafellar and R.J. Wets, Scenario and policy aggregation in optimization under uncertainty, Math. Oper. Res. (1989), to appear.
[28] A. Ruszczynski, A regularized decomposition method for minimizing a sum of polyhedral functions, Math. Progr. 35(1986)309–333.
[29] R.M. Van Slyke and R.J. Wets, L-shaped linear programs with applications to optimal control and stochastic programming, SIAM J. Appl. Math. 17(1969)638–663.
[30] R.J. Wets, Programming under uncertainty: The equivalent convex program, SIAM J. Appl. Math. 14(1984)89–105.
[31] J. Tomlin, LPM1 user's guide, Manuscript, Systems Optimization Laboratory, Stanford University (1973), unpublished.

Research and reproduction of this report were partially supported by the Office of Naval Research Contract N00014-89-J-1659; the National Science Foundation Grants ECS-8906260, DMS-8913089; the Electric Power Research Institute Contract RP 8010-09 at Stanford University. Additional research was supported by the Electric Power Research Institute Contract CSA-4005335, and the Austrian Science Foundation, "Fonds zur Förderung der wissenschaftlichen Forschung," Grant J0323-PHY.