
# ILL-DD NOTICE

To serve you better in the future:

On receipt of this article, please check for missing pages, truncated text, and blurry images. If any are found, ***please note the problem on our ILLiad request form*** and return it to us immediately!

Thanks and have a wonderful day.

# Decomposition and Parallel Processing for Large-Scale Electric Power System Planning Under Uncertainty

Mordecai Avriel, George B. Dantzig and Peter W. Glynn

**ABSTRACT**

Our research is concerned with developing software tools for finding optimal resource expansion plans for large scale power systems. An acceptable expansion plan must be capable of reliably supplying demand in the face of uncertainty about future demand and availability of generators and transmission lines due to equipment failures. The incorporation of uncertainty into the resource planning model of an electric power system is essential.

Unfortunately, the number of possible contingencies that must be hedged against can run in the millions. There are no off-the-shelf algorithms available to solve the problem in general. We believe, nevertheless, that these problems, though very very large in size, can be solved by applying improved methods such as decomposition techniques in combination with importance sampling and the exploitation of advanced computer technologies (parallel processors). The methodology described is being tested on a prototype stochastic Western Resource Planning Model for the west coast of the United States.

* Materials in this paper are drawn from two sources: Dantzig, G. B. and P. W. Glynn (1988) and Dantzig, Glynn, Avriel et al. (1989), see references at end of paper for complete title and where published.

## 1. SUMMARY

Our goal is to develop software tools for finding an optimal resource expansion plan for a large-scale power system. To be acceptable, a plan must be capable, when implemented, of reliably supplying demand in the face of uncertainty about future demand and uncertainty about how much generator and transmission line capacity will be lost due to equipment failure. For a plan to also be optimal, the discounted costs of building and maintaining the facilities plus the discounted expected operation costs must be less than any other plan of equal reliability.

To test our proposed algorithms, we have built a prototype computer model of the west coast power system. It allows for electric power to be exchanged between its six major utilities. Expansion, therefore, can take the form of more generators of various types, or more transmission lines linking the major utilities, or both.

The resource planning problem is well known in decision science as one of planning under uncertainty. In this paper we show that parallel processors, decomposition techniques, and importance sampling can be successfully combined to attack important classes of these problems, especially power system planning problems. This methodology represents the state-of-the-art techniques for solving large-scale linear programming models under uncertainty.

The *decomposition principle* is used to decompose the model into two parts: one part represents the deterministic facility expansion plan, while the second part is composed of a great many submodels. Each submodel is used to represent a different way the facilities would be operated when faced with one of the many different contingency conditions. The latter, called scenarios, have associated probabilities of occurrence and therefore are used to model the uncertainty.

Because the number of possible scenarios can be in the millions, we have developed a technique based on *importance sampling* which allows one to accurately estimate certain average information about the scenarios by examining only a very small fraction of all the scenarios. This information is used by the algorithm to improve the deterministic part of the expansion plan.

We believe that *parallel processors*, which are now becoming widely available, have the best computer architecture for efficiently carrying out the decomposition and importance sampling steps. Success of the entire approach, however, critically depends on the size of sample required being small. We believe that importance sampling can significantly reduce the size of sample required.

Details of the prototype west-coast model will be described first. This will be followed by the general theory of how mathematical programming and importance sampling can be used in conjunction with parallel processors to do resource planning.

## 2.DESCRIPTION OF THE PROTOTYPE MODEL

The six regions of North America represented by the prototype model (WRPM) are:

1. WSCC-CAN (Western Systems Coordinating Council—Canada);
2. NWPP (Northwest Power Pool Area);
3. RMPA (Rocky Mountain Power Area);
4. AZ-NM (Arizona-New Mexico Power Area);
5. NCAL (Northern California Power Area).
6. SCA-NV (Southern California-Nevada Power Area).

The objective is to determine optimal discounted least cost levels of generation and transmission facilities over time for each region in the system plus the cost of operating these facilities to meet demand. Electricity demand levels are given exogenously. Thus the objective function consists of the total discounted cost of supplying electricity (investments plus operating costs) to satisfy demand and certain reliability requirements. The supply of electricity is constrained by the initial availabilities and subsequent expansions/replacements of generation and transmission facilities over time in each region of the system network.

As with all models, the WRPM has a number of built-in limitations and simplifying assumptions in order to make it practical to solve. We believe that many improvements can be made to the model, particularly as to its size; these, however, have been postponed until the next stage of the research.

### Deterministic Version of WPRM:

In spite of its limitations, the *deterministic version* of the WRPM is a potentially useful model for long term planning purposes of electric power systems. It is basically a large-scale linear program. Source data is converted into a format required by the optimization algorithm by a powerful software package called GAMS (General Algebraic Modelling System). Although a large model, it can be solved easily on a moderate-size computer using a linear programming software package called MINOS. One of the nice features of the GAMS language is that it has an English-like structure, so that the programming model to be solved is described in terms that can be easily understood.

WRPM is a multiperiod model. The supply of electricity is optimally determined over $t = 1, 2, \cdots, T$ time periods, and consequently, most variables and parameters have several indices, one of which is the time period index $t$. There are options that allow the modeller to change the number of time periods and the length of each period. Because of the nature of electricity supply plans, the overall time span of the model has to be quite long, say 30–50 years. We recommend against having a model with 30 or 50 time periods of 1 year each since it could easily become intractable because of its large size . There is in fact, no real need to examine the behavior of the system on an annual basis because the type and size of electric networks that would be typically modelled by a WRPM type model would show little change from one year to the next. We have decided for our research purposes that it is more practical to model a 30-year time span consisting

of three to six time periods. Although the "time" index set for the description of the model is $t = 1, \cdots, T$, it is possible to assign a different time span to each period.

Since there are differences in demand patterns (and, to some extent also in supply availabilities) between summer and winter, each time period as discussed above is further indexed by seasons. There are two seasons in the model, corresponding to summer and winter respectively. In addition, the load duration curve for each demand region is assumed to consist of three load blocks: base load, intermediate load, and peak load. The introduction of seasonal and block indices into the model allows the modeller to more correctly represent peak demand in the summer for utility SCA-NV and in the winter for utility WSCC-CAN.

Various types of conventional and possible future electricity generation and transmission technologies are represented in the model. Generation facilities are classified according to the following technologies:

1. Liquid Fuel Fired;
2. Conventional Coal-Steam;
3. Pressurized Fluid-Bed Combustion;
4. Liquid/Gas Combustion Turbine;
5. Conventional Hydro;
6. Nuclear Light Water;
7. Geothermal;
8. Solar Central Receiver;
9. Wind;
10. Cogeneration;
11. Biomass-Municipal Refuse Steam;
12. Other.

The model can build and/or replace facilities of any type so as to achieve an optimal (cost effective) technology mix, subject to exogenously imposed regional and temporal constraints. Transmission lines are classified as either "short" (345 KV, 300 miles nominal length) or "long" (765 KV, 600 miles nominal length). Finally, generation and transmission facilities can be further classified by vintage as "old" and "new". The vintage classification is important for scheduling equipment retirement and replacement within the time horizon of the model.

Exogenous input data and assumptions constitute a significant portion of the model formulation. These data include among others demand for electricity (indexed by region, season, block) in the base (starting) year; similarly indexed demand growth rates; technological and cost-related data on generation and transmission facilities (indexed by type) and initial availabilities of these facilities (indexed by type and region, and in the case of hydro also by season); limitations on future developments of generation and transmission facilities (indexed by type, region, and time). The main sources for the data are NERC 1986 Electric Power Supply and Demand 1986-1995 [NERC (1986)], EPRI TAG-Technical Assessment Guide [EPRI (1986)], and the Wheeler Summer 1995 Snapshot model [Manne (1987)].

The variables in the model consist of the following main types:

- *Generation capacity levels*, indexed by region, generation technology type, vintage and time.
- *Transmission capacity levels*, indexed by source and target regions, transmission technology type, vintage and time.
- *Generation (operating) rate*, indexed by region, generation type, vintage, season, and block.
- *Transmission (operating) rate*, indexed by source and target regions, transmission type, time, season, and block.
- *Unserved demand*, indexed by region, time, season, and block.

In addition to these decision variables there are additional accounting variables that make the matrix structure have a more desirable pattern. In particular, we obtain a matrix with "staircase" structure, rather than merely a lower block triangular system.

As mentioned earlier, the objective of the WRPM is to minimize the total discounted capital and operating costs of the system, subject to meeting demand and reliability constraints. The constraints of the model can be classified broadly into two general types: *intratemporal* and *intertemporal*.

The *intratemporal constraints* are of the following general type:

- *Energy-capacity.* These constraints express the requirement that the total energy potentially available by the installed capacities in each season and block of each time period exceeds the energy demand for the corresponding season, block and time period.
- *Instantaneous demand.* Demand for electric power in each region of the system in each season, block and time period must be supplied by the power generated in that region plus the power transmitted to that region, less the power transmitted from that region to other regions, in that season, block, and time period.
- *Capacity-generation rate.* The power supplied or transmitted by every facility is less than or equal to the available installed capacity of that facility, multiplied by an appropriate percent availability factor.
- *Unserved demand.* The level of unserved demand in each region, season, block, and time period must not exceed a small percentage of the total demand for power. It is algebraically treated in some of the constraints as if it were just another type of generated power, while in others the total unserved energy is explicitly limited to a small amount.
- *Nonnegativity constraints.* All variables in the model have to be nonnegative.

The *intertemporal constraints*, on the other hand, link together the time periods of the model. They generally express the relationships between the available installed generating and transmission capacity in each region at a given time period, to the amount of installed capacity in the previous period, plus the levels of new capacities built less the capacity of the retiring facilities.

The set of constraints described up to this point define a large-scale multiperiod linear program having a staircase matrix structure. This deterministic version of the model has been formulated for WRPM using GAMS software and solved using MINOS linear-programming solution software.

### Stochastic Version of WRPM:

As will be explained in the next section, the general problem of determining the optimal resource planning policies of a complex electric power system is actually a problem of planning under uncertainty. If one ignores all the uncertainties and assumes that all the parameters of the planning problem are known with certainty, then the resulting deterministic problem can be formulated and solved as an ordinary linear program. However, as explained earlier, ignoring uncertainty in the planning process can produce an unsatisfactory solution because it does not properly hedge against future contingencies. It is therefore desirable, if it is practical to do so, that planning policies of a system, such as WRPM, produce a resource plan that can handle a variety of different contingencies that might arise in the future. We now discuss how we can expand the deterministic WRPM to the case where values of certain parameters of the planning problem are not known with certainty at the time when the plan is prepared.

Contingencies that may arise are of two major types:

- *Demand level contingencies.*
- *Equipment availability contingencies.*

There may yet be another type of contingency, arising such as higher electricity generation costs, due to say, higher future fuel prices. The treatment of the latter type of contingency is much easier from the modelling point of view than the previous two.

It is clear that a realistic model formulation of WRPM should include uncertainties in one period that affect the kind of uncertainty that occurs in later periods. For example, high demand in one period is correlated with high demand in the next. However, in this phase of our research, we have been assuming uncertainties within periods are independent of the uncertainties in subsequent periods. This permits us to keep the model small. In formulating a stochastic linear programming model including intraperiod uncertainties we have for each period of the planning horizon a set of random problem parameters. The right-hand side constants of the deterministic version which represent the instantaneous demands for electricity are replaced by random parameters. Similarly, the constant coefficients of variables in the constraint equations and inequalities, representing availabilities of future generation and transmission levels, are replaced by coefficient parameters that can take on random values according to a specified distribution of percent of equipment available.

A systematic replacement of the deterministic problem parameters by the realizations of their random counter-parts (called "contingencies" or "scenarios") involves the following steps: First, we identify the design variables which are the ones that do not vary with each realization of the random parameters. We

also identify those variables which adjust directly to each contingency by varying their respective levels. These are the *operating* variables.

In the WRPM, the design variables are the planned generation and transmission capacities on hand in a period, or being built, or replaced. The operating variables are the power levels generated and/or transmitted in future periods. Second, in order to determine the optimal values of both the design and operating variables we have to rewrite (copy) with minor changes those constraints of the linear program in which random parameters appear in order that every possible combination of random outcomes is properly represented. For example, if in a certain constraint the coefficient of one variable and at the same time the right-hand side are random, then if in theory, each can take on two possible outcomes, we have to rewrite (copy) that constraint with appropriate changes *four* times by substituting the four possible combinations of the two random parameters. In addition, for each possible combination of outcomes we define appropriate operating variables. Hence, not only is the number of constraints increased, but also the number of variables. If the resulting linear program were small enough to be solved in a reasonably short time, the optimal solution would be a solution that minimizes the total expected capital and operating costs.

However, in a large program such as WRPM, the number of possible combinations can be so enormous that no existing hardware and software can generate the full model and solve it to provide an exact solution. As we will discuss we believe that parallel processors, decomposition techniques and efficient sampling schemes can be applied to produce solutions, that with high probability, are very close to optimal.

In order to test various procedures discussed in the next section, we have conducted a number of experiments on parts of the model. Recently Dr. Gerd Infinger has programmed in APL our entire proposed procedure and has been running tests on a "toy version" of the west-coast stochastic model.

## 3. GENERAL THEORY OF SOLUTION

In this section we discuss how decomposition, importance sampling techniques, and parallel processors can be combined to efficiently solve certain classes of stochastic programs.

Applications include any system which has a vector of facilities being considered for expansion. Electric power system planning under uncertainty is an example of the more general problem of multiperiod resource planning under uncertainty. Other examples are long-term fleet expansion of an airline, or long-term asset allocation by a financial institution. Such problems can usually be readily modelled as multiperiod stochastic linear (or nonlinear) programs. Their solution, however, up to now has been impractical because the size of the problem can grow exponentially with the number of periods. Fortunately, for the multiperiod electric power system planning problem under uncertainty, as will be explained, the size of the problem need not grow exponentially. Hence this type of problem may be amenable for solution.

## Formulation of Deterministic Dynamic Linear Programs

We begin with the deterministic case of multi-period planning in order to point out why the general problem grows exponentially in size with the number of periods when we extend the formulation to include stochastic effects. We then restrict the extension to certain important types of investment or resource planning problems. For this class, the size need not grow exponentially with the number of time periods $T$.

Lower block-triangular matrix structures are typical for planning problems modelled as linear programs because activities initiated in period $t$ have input and output coefficients in periods $t, t+1, \ldots$. For example, for $T = 3$ periods, the coefficient matrix has submatrix blocks of coefficients as displayed in (3-1) below:

$$\begin{bmatrix} A_{11} & & \\ A_{21} & A_{22} & \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \tag{3-1}$$

where the $A_{ij}$'s are submatrices, $X_t \geq 0$ the vector of activity levels in period $t$, and the $b_t$'s are the specified vectors of inputs and outputs of the system being modelled.

By the introduction of in-process inventories and other devices, linear programs of lower block-triangular type are mathematically equivalent to *staircase* or *multistage* problems of the form:
*Find* $\min Z$ *and vectors* $X_t \geq 0$, *such that*

$$\begin{aligned} b_1 &= A_1 X_1 \\ b_2 &= -B_1 X_1 + A_2 X_2 \\ &\vdots \qquad\qquad \ddots \\ b_t &= \qquad -B_{t-1} X_{t-1} + A_t X_t \\ &\vdots \qquad\qquad\qquad\quad \ddots \\ b_T &= \qquad\qquad\qquad -B_{T-1} X_{T-1} + A_T X_T \\ (\min)\ Z &= c_1 X_1 + \cdots + c_t X_t + \cdots + c_T X_T \end{aligned} \tag{3-2}$$

*where the matrices* $A_t, B_t$ *and vectors* $b_t, c_t$ *are given.*

A number of promising methods for solving dynamic deterministic linear programs are known. Some references are [Dantzig (1963)], [Dantzig and Wolfe (1960)], [Glassey (1973)], [Ho and Manne (1974)], [Dantzig and Perold (1979)], [Ho and Loute (1980)], [Fourer (1982, 1983, 1984)], [Nishiya (1983)], [Jackson and Lynch (1982)]; for a general reference, see [Dantzig, Dempster, and Kallio, eds. (1981)].

Robert Entriken, as part of his Ph.D. thesis, has developed and tested a multi-stage decomposition algorithm for a deterministic model on parallel computers at Oak Ridge National Laboratories and has more recently run tests on equipment available at Stanford and at IBM Almaden.

For a very general class of stochastic planning problems, the values of $b_t, B_{t-1}, A_t, c_t$ for $t > 1$ may not be known to the planner with certainty at the start of stage 1 but become known to him at some later time $\tau$. The value $\tau$ itself could be a random variable and there could be a different $\tau$ for every element of the matrices and left-hand side vectors. While the values of these matrices may not be known, their

probability distributions could be given and this we assume.

In such problems, the planner wants to make a decision $X_1$; let random events happen; make a decision in period $t = 2$; let random events happen; make a decision in period $t = 3$, etc. He may wish to make the choice $X_1$ so that the expected value of $Z$ is minimum. We now give reasons why this very general class of stochastic problems is likely to remain intractable in the foreseeable future with existing conventional (serial) computers or even with the availability of parallel processors. We will then discuss a less general, but very important class of stochastic programs which up to the present have not been practical to solve on serial mainframes but could become so using parallel processors.

## Two-Stage Stochastic Programming

We begin with the simplest *two-stage case* first studied in [Dantzig, 1955] and subsequently developed by [Wets (1966, 1984)]:

$$\begin{aligned} b_1 &= A_1 X_1, && (X_1, X_2) \geq 0, \\ b_2 &= -B_1 X_1 + A_2 X_2 \\ (min)\ Z &= c_1 X_1 \quad + c_2 X_2 \end{aligned} \tag{3-3}$$

where the first stage parameters $(b_1, A_1, c_1)$ are known with certainty while those of the second stage can take on possibly a continuum of values $(b_2(\omega), c_2(\omega), B_1(\omega), A_2(\omega))$ with probability (density) distribution $p(\omega)$ for $\omega$ in $\Omega$, or a discrete probability distribution $p(\omega)$ where $\omega = 1, 2, \ldots, K$. The range of $\omega$ in $\Omega$ may therefore be continuous or it may be discrete, finite or infinite.

For the purposes of the computational approach, we require that $\Omega$ be discrete with a finite number of elements. Practically speaking, this is no restriction since any distribution may be approximated by a probability mass function concentrated on a finite set of points. Then, assuming we label the sample points $\omega$ using the integers $\{1, 2, \ldots, k\}$, the random vectors and matrices $(b_2, c_2, B_1, A_2)$ take on the value $(b_2(\omega), c_2(\omega), B_1(\omega), A_2(\omega)), (1 \leq \omega \leq K)$ with known probability $p_2(\omega)$.

We now illustrate the approach for $K = 3$. The two-stage stochastic program of minimizing expected costs under uncertainty then has *as its certainty equivalent* the deterministic linear program of the following form:

**Find** $\min Z, X_1 \geq 0,\ X_2(\omega) \geq 0,\ \omega = 1, 2, 3$ such that:

$$b_1(1) = \quad A_1 X_1 \tag{3-4.1}$$

$$\begin{aligned} b_2(1) &= -B_1(1)X_1 + \quad A_2(1)X_2(1) \\ b_2(2) &= -B_1(2)X_1 \quad + \quad A_2(2)X_2(2) \\ b_2(3) &= -B_1(3)X_1 \quad + \quad A_2(3)X_2(3) \end{aligned} \tag{3-4.2}$$

$$Z = \quad c_1 X_1 \quad + p_2(1)\ c_2(1)X_2(1) \ + p_2(2)\ c_2(2)X_2(2) \ + p_2(3)\ c_2(3)X_2(3) \tag{3-4.3}$$

To simplify the discussion that follows, we assume a bounded optimal solution exists. It follows that we can always find $\pi_2(k)$ to multiply constraints corresponding to $b_2(k)$ above and subtract from the objective so that the adjusted $c_2(\omega) \geq 0$. Therefore we can assume without loss of generality $c_2(\omega) \geq 0$. Again to simplify our discussion here, we assume $B_1$ is independent of $\omega$, i.e., $B_1(\omega) = B_1$.

Typically this problem is solved using the dual decomposition algorithm (see [Benders, 1962]). The key idea is to replace the objective function contribution of the second period variables by a scalar $\theta_2$, and to replace the second period (stage) constraints — those shown in (3-4.2) between the dashed lines — by a set of inequalities expressed in terms of $X_1$ and $\theta_2$ only. These are called "cuts" and are *necessary conditions* which are satisfied by all feasible and optimal solutions to (3-4). These cuts are added sequentially ($\ell = 1, 2, \ldots$) to the first period problem, $A_1 X_1 = b_1(1), X_1 \geq 0$; and these together with a modified objective $c_1 X_1 + \theta_2$ constitute the *restricted master problem* whose $\min Z$ is a lower bound estimate for $\min Z$ of (3-4). Cuts are added to the restricted master problem until they become *sufficient* in the neighborhood of the optimum to solve (3-4). This happens when the current value of the objective $Z$ for a feasible solution to (3-4) equals the lower bound estimate of $\min Z$. In practice the iterative process is stopped when this difference is judged to be "small enough".

**The Restricted Master Problem** for Benders' Decomposition Method has the form:

$$\text{FIND} \quad \min Z,\ X_1 \geq 0,\ \theta_2 \geq 0\ , \tag{3-5.0}$$

$$b_1 = A_1 X_1, \tag{3-5.1}$$

$$\text{CUTS:} \quad g_1^\ell \leq -G_1^\ell X_1 + \delta_2^\ell \theta_2, \qquad \ell = 1, \ldots, L \tag{3-5.2}$$

$$Z = c_1 X_1 + \theta_2 \tag{3-5.3}$$

where $\delta_2^\ell = 0$ for "feasibility" cuts if the subproblem (3-6) below from which it was derived is infeasible, and $\delta_2^\ell = 1$ for "optimality" cuts if the subproblem is feasible.

**The Sub Problem** for Bender's Decomposition Method has the form: For each $\omega = 1, \ldots, K$, find $\min Z_2(\omega)$, $X_2(\omega) \geq 0$,

$$\begin{array}{lll} & & \text{Dual Prices} \\ A_2(\omega) X_2(\omega) & = b_2(\omega) + B_1 X_1^* & : \pi_2(\omega) \end{array} \tag{3-6}$$

$$p_2(\omega) c_2(\omega) X_2(\omega) = Z_2(\omega) \tag{3-7}$$

The optimal solution $X_1 = X_1^*$ to (3-5) is then "tested" to see if it is the first period component of an optimal solution $(X_1, X_2)$ for (3-4). This is accomplished by solving the set of subproblems (3-6) to see (i) if the contribution $B_1 X_1^*$ from the first period implies for the second period a feasible solution for every choice of $\omega$, and (ii) if it together with the set of optimal solutions to the second period for every $\omega$ provides

a global optimum to the original problem. Global optimality is easily tested by checking whether the lower bound estimate for $\min Z$ is equal to the value of $Z$ for the current feasible solution. If the answer to (i) or (ii) is negative, the optimal $\pi_2(\omega)$ to (3-6) is substituted in formula (3-8) or (3-9) below in order to generate cut $L+1$ which is then added to the $L$ already generated in (3-5.2).

If (3-6) is feasible, the optimal dual "prices" are computed, whereas, if (3-6) is infeasible, "infeasibility weights" are calculated. They are then used as follows: If any subproblem $\omega$ is infeasible, its infeasibility weights are used to generate a "feasibility" cut (3-8) with $\delta_2^{\ell+1}=0$:

$$g_1^{\ell+1}=\pi_2(\omega)b_2(\omega);\quad G_1^{\ell+1}=\pi_2(\omega)B_1(\omega)\,. \tag{3-8}$$

(Later when the concept of a "reliable" system is discussed, this strict notion of infeasibility will be relaxed somewhat.) If feasible for all $\omega\in\Omega$, then $X_1^*$ is tested for optimality by comparing the lower bound estimate of $\theta$ from the master problem with $\sum_\omega Z_2(\omega)$. If the test fails (i.e. $\theta_2\neq\sum_\omega Z_2(\omega)$), the *expected values*

$$g_1^{\ell+1}=\sum_\omega \pi_2(\omega)b_2(\omega);\quad G_1^{\ell+1}=\sum_\omega \pi_2(\omega)B_1(\omega)\,. \tag{3-9}$$

are used to generate new "optimality" cut conditions to augment those of (3-5.2) with $\delta_2^{\ell+1}=1$. Note that the quantities $g_1^{\ell+1}$ and $G_1^{\ell+1}$ are actually expected values because the $\pi_2(\omega)$'s as defined by (3-6), are weighted by the probabilities $p_2(\omega)$ appearing in the objective.

## Multi-Stage Stochastic Planning Problem

The two-staged stochastic program (3-4) just discussed is one amenable to solution using parallel computers. The corresponding "reduction" to the equivalent deterministic linear program, for the general multi-staged case, however, becomes intractable due to the exponential increase in the number of possible outcomes, see Figure 3-1 below for the case of $T=3$ where the various contingencies in the third stage $E_{11}^3,\ldots,E_{22}^3$, etc. depend on the various contingencies $E_1^2$ or $E_2^2$ that precede them in the second stage.

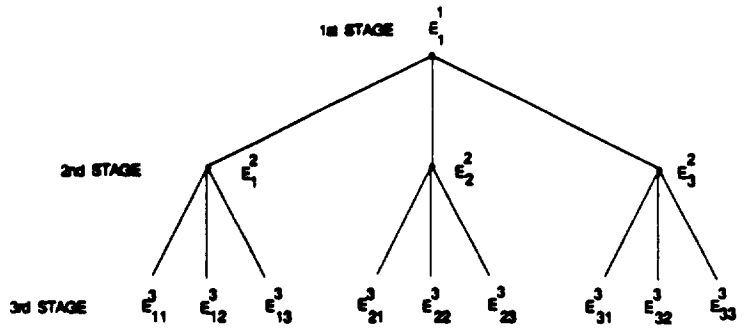


Figure 3-1. General Event Tree

It is now easy to see why, for large $K$, the proliferation of cases quickly becomes out of hand as the number of stages increases. Even if large numbers of inexpensive parallel processors were available, it does not seem to lead to a practical way to solve stochastic problems in general. In order to make headway with this fundamental problem, we propose for Electric Power systems to consider a certain restricted class of relevant problems which is tractable. In our approach, we focus on a subclass of multi-stage uncertainty problems whose dynamic deterministic part, once fixed, is assumed unaffected by stochastic events. Its "event tree" is displayed in Figure 3-2 below.

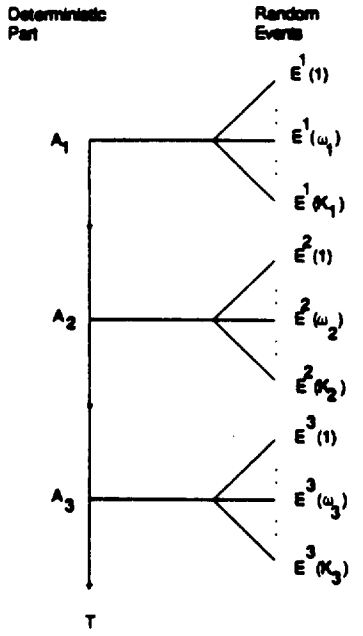


Figure 3-2. Event Tree for the "Here and Now" Type Decision

In the resource planning context, the planner is confronted with the problem of planning system resources for $T$ periods. The decision variables can be divided into two major groups: The first are the *design* variables that are determined by taking into account future uncertainties, but whose values, once determined, will not be affected by stochastic events. For example, the installed capacities of generation units and transmission lines are such design variables. The second group of variables are the *production* variables, or operating variables. In solving the multi-stage planning problems these variables may take on several values, corresponding to the outcomes of the stochastic events. For example, the actual power generated by the various generation units is a production variable, clearly affected by such stochastic events as demand variability and equipment failure.

Our formulation implies a "here and now" type decision, discussed in [Dantzig and Madansky 1961]. More specifically, Event Tree Figure 3-2 assumes that decision of how much resources to have in the future takes the form of a *here and now* decision of what our resource development path will be over time, and not a decision of what branches to take depending on what different contingencies may arise in the future. Under this assumption, high demand in year $t$ does not imply a higher expected demand in year $t+1$.

Whether or not a "here and now" decision is realistic depends on a number of factors: For many industries, such as electric power, resource planning is a major undertaking involving raising capital, and making long-term contractual arrangements. Thus there is a great deal of momentum to stick to a course of action for several years once a commitment is made. For such situations, our "here and now" assumption may not be a bad one.

In practice, of course, planners often act as if they are making a decision here and now of what resources they plan to have on hand in the future, even though they know that later on they are likely to change their mind. It probably would be better in such situations for planners to incorporate into their decision model the possibility of making some **"in course"** future corrections. From a computational point of view, there is no difficulty extending our approach to allow for some branching provided the number of branches is kept down to a few, say two or three, major ones. More branches than that could create computational difficulties. How to generalize the model to allow for stochastic events that are correlated from one period to the next will be described in greater detail later.

## Mathematical Formulation of the Multi-stage Resource Planning Problem

The multi-stage resource planning problem can be generically stated as follows: Given a finite number of periods, find the levels of the design and the production variables that will minimize the total expected costs of building and operating the resources up to the planning horizon.

Let us now formulate this problem as a multi-stage stochastic mathematical program. To simplify the presentation, we consider the two-period (3 stage) case first. For this case we assume the random events occurring in period 1 are independent of those occurring in period 2. Moreover, we assume that within a stage, there are random events concerning supply availability and random events concerning demand levels, and these two types of random events are independent of each other. Let $\tilde{\omega}_t, \bar{\omega}_t$ denote the random events in stage $t$ corresponding to supply availability and demand levels, respectively. The two period problem then has the form:

**Find** $\min Z$, *"design" vectors* $X_t \geq 0$, *and "production" vectors* $U_t(\omega_t) \geq 0$, *corresponding to each point* $\omega_t = (\tilde{\omega}_t, \bar{\omega}_t)$ *in* $\Omega_t$ *such that*

$$
\begin{array}{lllll}
b_1 & = A_1 X_1 & & & \\
0 & = -F_1(\tilde{\omega}_1) X_1 & & +D_1 U_1(\omega_1) & \\
d_1(\bar{\omega}_1) = & & & +E_1 U_1(\omega_1) & \\
b_2 & = -B_1 X_1 & +A_2 X_2 & & \\
0 & = & -F_2(\tilde{\omega}_2) X_2 & & +D_2 U_2(\omega_2) \\
d_2(\bar{\omega}_2) = & & & & +E_2 U_2(\omega_2) \\
Z & = c_1 X_1 & +c_2 X_2 & +\mathcal{E} f_1 U_1(\omega_1) & +\mathcal{E} f_2 U_2(\omega_2)
\end{array} \qquad (3\text{-}10)
$$

*where* $\mathcal{E}$ *denotes expectation over* $\omega_t = (\tilde{\omega}_t, \bar{\omega}_t) \in \Omega_t$.

Here $A_1$ and $b_1$ have to do with the initial conditions of the resources at the beginning of the planning span (stage 1) and are known with certainty. $F_t(\tilde{\omega}_t)$ and $D_t$ are used to relate supply availability with operating levels in period $t$ while $d_t(\bar{\omega}_t)$, $E_t$ are used to relate uncertain demands to operating levels in period $t$. Finally $B_1$, $A_2$ are used to relate the levels of the design variables in periods 1 and 2. Design variables are those having to do with the development, maintenance, and retirement of facilities.

The mathematical structure for the deterministic part of the resource planning problem is exactly the same as in (3-2), namely:

$$b_t = -B_{t-1}X_{t-1} + A_tX_t\ , \quad \text{for } t = 1,\ldots,T \text{ and } B_0X_0 \equiv 0\ , \tag{3-11}$$

where $X_t \geq 0$ are the planned design variables (to be determined) for period $t$. Their cost is $\bar{Z} = \Sigma c_tX_t$ where $b_t, A_t, B_t, c_t$ are all known with certainty. The subproblem for period $t$ is defined for some $X_t = X_t^*$ by first finding the level of design variables available for use in period $t$:

$$F_t(\tilde{\omega}_t)X_t^*, \qquad \tilde{\omega}_t \text{ in } \tilde{\Omega}_t\ , \tag{3-12}$$

which depends on $X_t^*$ and a random variable $\tilde{\omega}_t$ reflecting the uncertain proportion of unavailable $X_t^*$ requiring repair. The probability distribution of $\tilde{\omega}_t$ is assumed known. Letting $d_t(\bar{\omega}_t)$ be the uncertain demand we then solve for the dual multipliers and $Z_t$ that solves (3-13) below for each random choice of right-hand side, i.e., for $\tilde{\omega}_t$ in $\tilde{\Omega}_t$ and $\bar{\omega}_t$ in $\bar{\Omega}_t$. We do this in order to determine the expected values of the cuts generated by (3-14) below. If the number of discrete values of $\tilde{\omega}_t$, $\bar{\omega}_t$ are small, this can be done exactly. If the number is large, then their expected values have to be estimated by an importance sampling procedure described later. Let $\omega_t = (\tilde{\omega}_t, \bar{\omega}_t) \in \Omega_t$. The sub-subproblems are:

Find $U_t(\omega_t) \geq 0$ and $\min Z_t(\omega_t)$, $\omega_t = (\tilde{\omega}_t, \bar{\omega}_t) \in \Omega_t$:

$$\begin{array}{lll} & & \text{Dual Prices} \\ D_t\ U_t(\omega_t) = F_t(\tilde{\omega}_t)X_t^*\ , & \tilde{\omega}_t \text{ in } \tilde{\Omega}_t\ , & : \tilde{\pi}_t(\omega_t) \\ E_t\ U_t(\omega_t) = d_t(\bar{\omega}_t)\ , & \bar{\omega}_t \text{ in } \bar{\Omega}_t\ , & : \bar{\pi}_t(\omega_t) \\ f_tU_t(\omega_t) \ = Z_t(\omega_t)\ , & & \end{array} \tag{3-13}$$

where the corresponding dual prices are shown on the right. Sample scenarios $\tilde{\omega}$ and $\bar{\omega}$ are used to estimate expected cuts $\ell = 1, 2, \ldots$ of the form

$$g_t^\ell \leq -G_t^\ell X_t + \delta_t^\ell\theta_t\ ; \tag{3-14}$$

If the subproblems are all feasible, then we get

$$g_t^\ell = \sum_{\omega_t} p_t(\omega_t)\bar{\pi}_t(\omega_t)d_t(\bar{\omega}_t); \quad G_t^\ell = \sum_{\omega_t} p_t(\omega_t)\tilde{\pi}_t(\omega_t)F_t(\tilde{\omega}_t) \tag{3-15}$$

and obtain a cost form

$$Z = \sum_t (c_t X_t + \theta_t)\,. \tag{3-16}$$

which, together with (3-3), form the MASTER PROBLEM (3-17) below.

Several parallel processors could be assigned to each stage. The processors receive $X_t^*$ as inputs from the Master Program, generate random $\tilde{\omega}_t$, $\bar{\omega}_t$ and solve the subproblems, and finally give back to the MASTER approximate cuts (3-14) obtained by sampling. These cuts are then used to augment those obtained earlier.

### Master Problem.

**FIND** $\min Z,\ X_t \geq 0,\ \theta_t \geq 0$:

$$\begin{aligned}
b_1 &= A_1 X_1 \\
b_2 &= -B_1 X_1 + A_2 X_2 \\
g_1^{\ell_1} &\leq -G_1^{\ell_1} X_1 + \delta_1^{\ell_1}\theta_1 \\
g_2^{\ell_2} &\leq -G_2^{\ell_2} X_2 + \delta_2^{\ell_2}\theta_2 \\
Z &= c_1 X_1 + \theta_1 + c_2 X_2 + \theta_2
\end{aligned} \tag{3-17}$$

where $\ell_1 = 1, \ldots, L_1,\ \ell_2 = 1, \ldots, L_2$.

Note that the Master Problem has the form of a deterministic staircase system which can be solved directly; especially large staircase systems can be solved by using the dual nested decomposition approach, see for example [Abrahamson (1983)], [Wittrock (1983)], and [Entriken (1988)]. Entriken has developed a version for parallel processors. The advantage of the dual-decomposition approach is that the parallel processors at each stage can be used effectively to provide information in the form of $X_{t+1}^*$ to the stage below and to pass back cuts generated in stage $t$ to stage $t-1$.

### Interperiod Dependence of Random Events

As noted earlier, there may be situations where our previous assumptions on interperiod independence of the stochastic events are not appropriate. At the other extreme, when there is complete dependence as in Figure 3-1 the exponential growth of cases with number of periods, makes such stochastic planning problems in general too costly to find a numerical solution.

However, it would be practical to solve a model when there is a limited amount of stochastic dependence between periods. The contingent events occurring in the planning horizon (for $K = 3$) would have an event tree like that depicted in Figure 3-3 below where, note, *we force the consolidation of some of the stochastic events*, like $E_{12}^3$ and $E_{21}^3$, *into a single composite event.*

Such a consolidation is an approximation which, if acceptable causes the number of possible outcomes (in the example) to increase by two from one stage to the next. This avoids the exponential growth which can arise in the case of total dependence. Suppose, for example, that the demand for electricity can either grow, remain the same, or decline by a fixed amount from one period to another. If demand is, say, 100 at

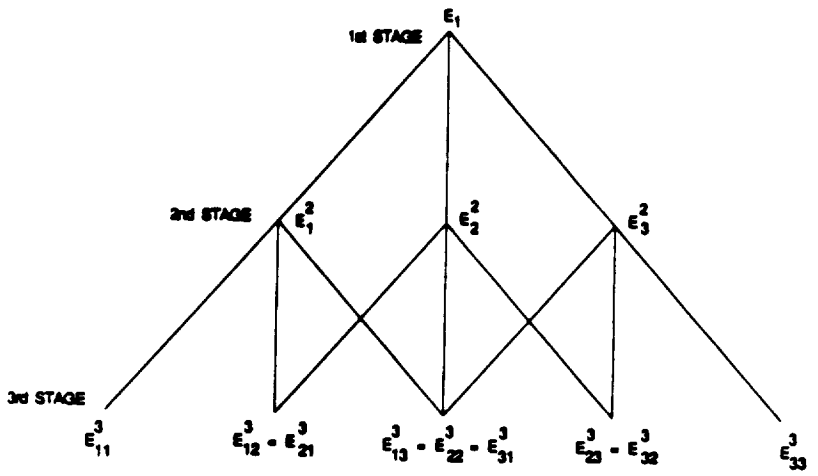


Figure 3-3. Event Tree with Consolidated Events

the beginning of the planning horizon, and can either grow by 10 units, remain the same, or decline by 10 units, we have:

$$E_1^1 = 110\ ;$$

$$E_1^2 = 90,\ E_2^2 = 100,\ E_3^2 = 110;$$

$$E_{11}^3 = 80,\ E_{12}^3 = E_{21}^3 = 90,\ E_{13}^3 = E_{22}^3 = E_{31}^3 = 100, E_{23}^3 = E_{32}^3 = 110, E_{33}^3 = 120.$$

Incorporation of this type of partial interperiod dependence into a multi-period resource planning framework requires new formulations and perhaps also new decomposition techniques, subjects that we propose to investigate at a future time.

## Reliability Considerations

An important aspect of electric power system planning models has to do with introducing constraints that guarantee the system is reliable. A 100% reliable system would be too costly to build. What is done in practice is to tolerate some failure. Specifically, one specifies an upper bound on the amount of expected unserved demand. We use a constraint of the form

$$\sum_{\omega \in \Omega_t} p_t(\omega) g_t\ \tilde{U}_t(\omega) \leq \alpha_t, \quad t = 1, 2, \ldots, T. \tag{3-18}$$

Here $g_t \cdot \tilde{U}_t(\omega)$ calculates the amount of unserved energy in period $t$ under contingency $\omega$.

Adding such a constraint means that the subproblem no longer decomposes into separate sub-subproblems for each $\omega \in \Omega_t$. In order not to lose this feature, we form a primal decomposition of the problem into relations (3-19) and (3-20) corresponding to a "high-level" Master Problem, in the sense of the primal decomposition approach:

$$\sum_t c_t(X_t) + \sum_t \sum_{\omega \in \Omega_t} p_t(\omega) f_t \tilde{U}_t(\omega) = \min Z \tag{3-19}$$

$$\sum_{\omega \in \Omega_t} p_t(\omega) g_t \tilde{U}_t(\omega) \leq \alpha_t, \qquad t = 1, \ldots, T\,, \tag{3-20}$$

For a basic reference, see [Dantzig-Wolfe, 1960] and [Dantzig, 1963]. All remaining constraints are treated as the constraints of the "high-level" subproblem. This essentially is a systematic way of assigning "penalty" prices $\rho_t$ to the reliability constraints. These are then used to replace the coefficients $f_t$ of (3-19) by $\bar{f}_t = f_t - \rho_t g_t$. The subproblem of this new problem is in exactly the same format as the problem we discussed earlier without reliability constraints. In this formulation infeasibility constraints no longer need to be generated because all the sub-subproblems (3-6) can now be formulated in such a way that they are always feasible.

Part of our research concerns the best way to order the flow of information to the dual masters to form the cuts of various periods and to the primal master to adjust $\bar{f}_t = f_t - \rho_t g_t$.

## 4. SOLUTION METHODS BASED ON SAMPLING

The decomposition algorithm just described is clearly only practical when $K$ is small. When $K$ is large, we propose to use parallel processors as high-speed quadrature devices based on *Monte Carlo sampling* to effectively solve such problems. One idea we are currently considering is to have a processor at the Master level serve as an *integrator* which sequentially receives as input estimates of the cuts (3-5.2). The Master Problem is then solved to optimality with the estimates it has received so far, and is used to generate as output revised $X_1 = X_1^*$ that are sent to other parallel processors which are busy solving (3-6) for various choices of $\omega$. This process also provides a *lower bound* estimate for $\min Z$ which most of the time strictly increases with each solution of the master problem.

The amount of space needed to store the generated cuts in the computer memory need not be large. Assuming $B_1$ is independent of $\omega$, no more than $L \leq m_2$ of the cuts will be tight on any major iteration, where $m_2$ is the number of rows in $B_1$. This is so because $G_1^\ell$, generated by linear combinations of the rows of $B_1$, has rank $\leq r$ where $r \leq m_2$ is the rank of $B_1$. The remainder may be dropped (possibly to be regenerated on some later iteration).

Several parallel processors could be employed at the Sub level, each having as input the latest value of $X_1^*$ and solving (3-6) in dual form for many random or stratified choices of $\omega$. When $c_2, A_2$ are the same for all $\omega$, the dual of (3-6) is a linear program with only the dual objective $b_2(\omega)$ changing. By judiciously stratifying the random sampling of $\Omega$ we hope to use the optimal basic dual feasible solution for one choice of $\omega$ to find quickly the optimal one for the next $\omega$. To provide cuts for the Master, the parallel processors

are to be used to determine the expected values $g_1^{\ell+1}$ and $G_1^{\ell+1}$ defined by (3-7.2) or to approximate them by means of a large enough *importance sample.*

If it is practical to solve (3-6) for all $\omega$, the set of solutions to (3-6) would generate a valid cut and a correct *upper bound* estimate for $\min Z$. In that case, the difference between the lower bound and upper bound estimates can be used to test optimality of $X_1^*$ for the original problem. When $Z$, according to some specified tolerance, is close enough to the lower bound estimate for $\min Z$ the iterative process is stopped and $X_1^*$ is declared "optimal".

However, when $K$ is large, it is no longer possible to solve (3-6) for all $\omega$. Instead our approach is to use random sampling to choose a subset of $\omega$'s for which (3-6) will be solved. A major methodological problem to be resolved is the choice of sampling strategy.

Critical to success of our whole approach is the size of sample required. We have found in tests described in [Dantzig et al. (1989)] that the *importance sampling* technique can reduce the size of sample dramatically. This is why we believe our approach will be successful in solving large-scale resource planning of power systems under uncertainty.

Therefore, a major thrust of our research effort has been to incorporate Monte Carlo sampling ideas into a linear programming decomposition framework. Specifically, each iteration of the Bender's decomposition requires the computation of an expected value of costs and expected value of dual prices. Since both are done in an analogous way, only the cost case will be illustrated. Expected costs, denoted by $\alpha$, is of the form

$$\alpha = Ec(\vec{X}); \tag{4-1}$$

Here, $c(\vec{X})$, calculated by solving a linear program, is the operating plus penalty cost associated with (random) scenario $\vec{X}$. The scenario $\vec{X}$, in turn, describes the equipment status of the system. The various components of $\vec{X}$ list the number of transmission lines and generators that are down at various locations in the system. Typically, we assume that these components are independent random variables. This basically means that we are asserting that generation/transmission failures in one part of the system are created independently of what occurs elsewhere. Although this independence assumption simplifies much of the following discussion, the basic ideas and methods that we will be describing are equally pertinent to power systems in which generation/transmission failure is modelled in a dependent fashion.

Because of the multi-regional nature of the power systems that we are considering, $\vec{X} = (X_1, \cdots, X_d)$ will typically incorporate many components. (Recall that $\vec{X}$ includes a description of generating capacity within each of the model regions, so that the number of components in $\vec{X}$ increases at least as fast as the number of regions.) Hence, $\alpha$ takes the form

$$\alpha = \sum_{\vec{x}} c(\vec{x}) p_1(x_1) \cdots p_d(x_d), \tag{4-2}$$

where the dimension $d$ is large. $(p_i(x_i) = p\{X_i = x_i\})$. The large size of $d$ means that the number of summands comprising $\alpha$ will (typically) be enormous. For example, if $d = 20$ and the number of possible

operating levels for the $i$'th component $X_i$ is set to 5 $(1 \leq i \leq 20)$ , we find that the total number of summands is $5^{20}$. Clearly, this renders the computation of $\alpha$ by direct summation impractical in our setting.

Standard references on numerical integration [e.g., Davis and Rabinowitz (1984)] typically recommend Monte Carlo sampling as the method of choice for performing computations of the above type when $d$ is large. The reason for this choice is that the convergence rate of Monte Carlo sampling is relatively insensitive to the dimensionality effects typical of most conventional integration (or summation) schemes. Specifically, if $\vec{X}_1 \vec{X}_2, \cdots$ are sampled independently from the joint probability mass function $p_1(x_1) \cdots p_d(x_d)$, the central limit theorem states that

$$n^{1/2} \left( \frac{1}{n} \sum_{i=1}^{n} c(\vec{X}_i) - \alpha \right) \Rightarrow \sigma N(0,1) \tag{4-3}$$

as $n \rightarrow \infty$ ($\Rightarrow$ denotes convergence in distribution), where $\sigma^2 = \text{var}(c(\vec{X}_i))$. A more descriptive way of re-writing the above limit theorem is as follows:

$$\frac{1}{n} \sum_{i=1}^{n} c(\vec{X}_i) \overset{D}{\approx} \alpha + \frac{\sigma}{\sqrt{n}} N(0,1) \tag{4-4}$$

($\overset{D}{\approx}$ denotes approximate equality in distribution). Hence, our estimator for $\alpha$ (namely, $n^{-1} \sum_{i=1}^{n} c(\vec{X}_i)$) has an error term which is approximately normally distributed and goes down with sample size $n$ as $1/\sqrt{n}$. The important feature of this result is that the convergence rate $1/\sqrt{n}$ is independent of the dimension $d$. This is in contrast to conventional summation schemes, in which the convergence rate is typically sensitive to $d$ and degrades quickly as $d$ increases. In fact, the typical "worst case" convergence rate for conventional numerical schemes takes the form $n^{-1/d}$ ($n = \#$ number of function evaluations of $c(\vec{x})$). Thus, if $d = 20$, the convergence rate for such schemes would be impossibly slow.

The above convergence rate analysis suggests that Monte Carlo sampling is the appropriate method for performing the types of expected value computations that arise in our power system decomposition context. A further advantage of these sampling methods is that they are easily adapted to a parallel computation setting. Assuming that processors can compute independently of one another, we permit each processor to sample $\vec{X}$ independently; the corresponding linear program is then solved on the appropriate processor, and $c(\vec{X})$ returned. This replication-based approach has the potential to decrease computing time by a rough factor of $p$ (the number of processors). This, of course, is the best possible decrease that one can expect with a parallel machine. Hence, sampling methods are particularly suitable for purposes of parallel computation.

## Variance Reduction: Importance Sampling

As the central limit theorem described above suggests, the convergence rate of Monte Carlo sampling (although dimension independent) is quite slow, namely of the order $n^{-1/2}$ in the number of function evaluations $n$. This puts a premium on the development and application of *variance reduction techniques*

for computation of $\alpha$. The idea here is to try to reduce the so-called variance constant $\sigma^2 = \text{var}(c(\vec{X}))$. This then reduces the error of the estimates in proportion to the reduction in standard deviation achieved. A particularly promising approach to obtaining significant variance reduction in the electric power capacity planning problem is to apply importance sampling. In our problem context, we expect the solution to hedge to some degree against contingencies in which significant capacity is disabled because of (random) failure. Loss of capacity due to failure is measured in the linear programming formulation primarily by the terms in the constraint on expected unsupplied demand.

Because of the safety margin built into electric power systems, significant capacity must be disabled for the demand constraint to come into play. Clearly, this event typically has low probability, if the power system has been adequately designed from a reliability standpoint. Thus, the events that have a major impact on the solution's ability to hedge are events that will only rarely be picked up by a naive sampling procedure based on the original sampling frequencies.

This, in turn, leads to a high variance in the estimator associated with naive sampling. The idea behind importance sampling is to change the sampling procedure so that those events that matter more are sampled more. In the power systems context, this typically will mean that we wish to sample contingencies in which multiple failures occur at a higher frequency than they occur naturally. Multiple failures lead to loss of significant system capacity, which in turn forces the solution to hedge properly. Thus, importance sampling appears intuitively to be a useful tool in the power systems setting.

We now briefly describe the theory that underlies importance sampling. Note that if $q(\vec{x})$ is a (positive) probability mass function, we can re-express $\alpha$ as

$$\begin{aligned} \alpha &= \sum_{\vec{x}} c(\vec{x}) \frac{p(\vec{x})}{q(\vec{x})} q(\vec{x}) \\ &= E c(\vec{Y}) \frac{p(\vec{Y})}{q(\vec{Y})}, \end{aligned} \tag{4-5}$$

where $\vec{Y}$ is a random vector having probability mass function $q(\cdot)$ (i.e., $P\{\vec{Y}=\vec{y}\} = q(\vec{y})$)). Here, $p(\vec{y}) = p_1(y_1) \cdots p_d(y_d)$, where the $p_i(\cdot)$'s are the marginal probability mass functions associated with generator/transmission capacity on an area-by-area basis. The variance of the new estimator based on the above transformation is just

$$\text{var}\left( \frac{1}{n} \sum_{i=1}^{n} c(\vec{Y}_i) \frac{p(\vec{Y}_i)}{q(\vec{Y}_i)} \right) = \frac{1}{n} \left( E c(\vec{Y})^2 \frac{p(\vec{Y})^2}{q(\vec{Y})^2} - \alpha^2 \right). \tag{4-6}$$

It turns out that there exists an optimal choice $q*$ of the mass function $q$, namely

$$q^*(\vec{x}) = \frac{|c(\vec{x})| \cdot p(\vec{x})}{\sum_y |c(\vec{y})| \cdot p(\vec{y})}. \tag{4-7}$$

In fact, the variance of $c(\vec{Y})p(\vec{Y})/q^*(Y)$ is easily seen to be zero when $c$ is non-negative. In other words, we obtain a perfect estimate for $\alpha$ in just one observation.

Of course, the situation is clearly more complicated than this. The basic difficulty is that generating variates from $q^*(\cdot)$ is impossible; in particular, when $c$ is non-negative, $q$ involves the actual sum that we are trying to calculate, in its denominator. As a consequence, the above result can only be used as a heuristic guide on how to choose a good sampling population $q$. The heuristic that one follows is: Choose a sampling population $q$ having the property that $q(\vec{x})$ is approximately proportional to $|c(\vec{x})| \cdot p(\vec{x})$, subject to certain computational constraints on $q$. The constraints are that $q$ should be relatively easy to calculate and that the generation of variates from $q$ should be computationally inexpensive.

Assume, for the remainder of this section, that the cost function is non-negative. (This is true of virtually all applications.) Suppose, for argument's sake, that

$$c(\vec{x}) \approx c_1(x_1) \cdots c_d(x_d) \tag{4-8}$$

i.e., the function $c(\vec{x})$ is multiplicative in its arguments. Then,

$$q(\vec{x}) = c_1(x)p_1(x_1) \cdots c_d(x_d)p_d(x_d). \tag{4-9}$$

Here, the random vectors $\vec{Y}$ can easily be generated from marginal distributions: $\vec{Y} = (Y_1, \cdots, Y_d)$ where the $Y_i$'s are independent and $P\{Y_i = y_i\} = q_i(y_i)$, with

$$q_i(y_i) = \frac{c_i(y_i)p_i(y_i)}{\sum_{x_i} c_i(x_i)p_i(x_i)}. \tag{4-10}$$

Thus, assuming the $q_i$'s have nice variate generation structure, generating the random vectors $\vec{Y}_1, \vec{Y}_2, \cdots$, is easy. To calculate a good approximation to $q_i(\cdot)$, we note that under the multiplicative model,

$$\frac{c_i(x_i)}{c_i(\tau_i)} \approx \frac{c(\tau_1, \cdots, \tau_{i-1}, x_i, \tau_i, \cdots, \tau_d)}{c(\tau_1, \cdots, \tau_i, \cdots, \tau_d)}, \tag{4-11}$$

where $\tau_1, \cdots, \tau_d$ are suitably chosen (fixed) constants. (If the summation is over the unit hypercube, we could, for example, choose $\tau$ as the origin of the $d$-dimensional hypercubes or as the vector of marginal means.) The above relationship determines the marginal cost functions $c_1, \cdots, c_d$ up to proportionality constants. Since the $q_i(\cdot)$'s are mass functions, this is sufficient to uniquely determine the joint mass function $q$. Hence, the multiplicative model assumption permits us to approximate the surface $c(x)p(x)$ with a relatively small number of cost function evaluations. (Recall that a cost function evaluation amounts to solving a linear program, so that it is desirable to minimize the work devoted to this task.)

The problem with this approach is that a multiplicative representation for our cost function seems unreasonable, even as an approximation. A better heuristic is to assume that

$$c(\vec{x}) \approx \sum_{i=1}^{d} c_i(x_i).$$

i.e., the cost function is well approximated by an additive model. In this case, we can again reduce the computation of the joint mass function $q(x)$ to something involving the marginal distributions. The $d$-dimensional problem is then basically reduced to $d$ 1-dimensional problems. More specifically, if $c$ takes on the additive form, then

$$q(\vec{x}) \approx \frac{\sum_{i=1}^{d} c_i(x_i)p_1(x_1)\cdots p_d(x_d)}{\sum_{i=1}^{d}\sum_{x_i} c_i(x_i)p_i(x_i)} \tag{4-12}$$

$$= \sum_{i=1}^{d} \alpha_i q_i(x_i) \prod_{j\neq i} p_j(x_j)$$

where

$$\alpha_i = \frac{\sum_{x_i} c_i(x_i)p_i(x_i)}{\sum_{i=1}^{d}\sum_{x_i} c_i(x_i)p_i(x_i)} \tag{4-13}$$

$$q_i(x_i) = \frac{c_i(x_i)p_i(x_i)}{\sum_{x_i} c_i(x_i)p_i(x_i)} \tag{4-14}$$

Here, a variate $\vec{Y}$ can again be easily generated from the marginal distributions $q_1, \cdots, q_d$ of $Y_1, \cdots, Y_d$: With probability $\alpha_i$, independently generate $Y_i$ from $q_i$ and $Y_j (j \neq i)$ from the original marginal distributions $p_j$ . This results in a random vector $Y$ having the above joint probability mass function $q$.

Thus, as in the multiplicative case, we are led to the question of how to calculate the $q_i$'s . This can again be done by evaluating the cost function on a relatively small lattice of points:

$$c_i(x_i) \approx c(\tau_1, \cdots, \tau_{i-1}, x_i, \tau_{i+1}, \cdots, \tau_d) - c(\tau_1, \cdots, \tau_i, \cdots, \tau_d). \tag{4-15}$$

This determines the $c_i$'s up to the additive constant $c(\tau_1, \cdots, \tau_d)$. This additive constant is disposed of by writing the cost function $c(\vec{X})$ as $c(\vec{X}) = c(\tau_1, \cdots, \tau_d) + \Delta C(\vec{X})$. The function $\Delta C(\vec{X})$ is again of additive form, but has the advantage that we know a priori that we may take $\Delta c_i(\tau_i) = 0$. This eliminates the additive constant from the picture.

The degree of variance reduction obtained by applying this "additive importance sampling method" depends on the extent to which the true cost surface is fit by an additive representation. Experimental results based on the additive approach are described by M. Nakayama in [Dantzig, Glynn, Avriel, Stone, Entriken, and Nakayama (1989)]. He applies this method to the electric power system capacity planning model described in Section 1 of this paper. It turns out to be a highly effective means of variance reduction.

## REFERENCES

Abrahamson, P.G. (1983), A nested decomposition approach for solving staircase linear programs, Report SOL 83-4, Department of Operations Research, Stanford University, Stanford, Ca.

Beale, E.M.L., G. B. Dantzig and R.D. Watson (1986), A First Order Approach to a Class of Multi-Time-Period Stochastic Programming Problems, *Mathematical Programming Study*, **27**, 103-117.

Benders, J.F. (1962), Partitioning procedures for solving mixed-variable programming problems, *Numerische Mathematik*, **4**, 238-252.

Birge, J.R. (1984), Aggregation in Stochastic Linear Programming, *Mathematical Programming*, **31**, 25-41.

Birge, J.R. and S.W. Wallace (1988), A Separable Piecewise Linear Upper Bound for Stochastic Linear Programs, *SIAM J. Control and Optimization*, **26**, No. 3.

Birge, J.R. and J.B. Wets (1986), Designing Approximation Schemes for Stochastic Optimization Problems, in Particular for Stochastic Programs with Recourse, *Mathematical Programming Study*, **27**, 54-102.

Bratley, P., B. Fox and L. Schrage (1983), *A Guide to Simulation*, Springer-Verlag, New York.

Brodt, A.I. (1984), International Bank Asset and Liability Management, *J. of Bank Research*, **15**, No. 2, 82–94.

Chow, Y.S. and H. Robbins (1965), On the Asymptotic Theory of Fixed Width Sequential Confidence Intervals for the Mean, *Ann. Math. Stat.*, **36**, 457-462.

Cochran, W.G. (1977), *Sampling Techniques*, John Wiley, New York.

Cohen, K.J. and S. Thore (1970), Programming Bank Portfolios Under Uncertainty, *J. of Bank Research*, Vol. 1, No. 1, 42–61.

Dantzig, G.B. (1955), Linear programming under uncertainty, *Management Science*, **1**, 197-206.

Dantzig, G.B. (1963), *Linear Programming and Extensions*, Princeton University Press, Princeton.

Dantzig, G.B. (1982), Time-staged methods in linear programs, In *Studies in Management Science and Systems, Vol. 7 Large- Scale Systems* (Y.Y. Haims, ed.), North Holland Publishing Company, Amsterdam, 19-30.

Dantzig, G.B. (1988), Planning under uncertainty using parallel computing, *Annals of Operations Research*, in press.

Dantzig, G.B., M.A.H. Dempster and M.J. Kallio (Eds.) (1981), *Large-Scale Linear Programming (Volumes 1, 2)*, IIASA Collaborative Proceedings Series, CP-81-51, IIASA, Laxenburg, Austria.

Dantzig, G.B. and P.W. Glynn (1988), Parallel Processors for Planning Under Uncertainty, Technical Report SOL 88-8, Department of Operations Research, Stanford University. To appear in *Annals of Operations Research.*

Dantzig, G.B. and M. Madansky (1961), On the Solution of Two-Staged Linear Programs under Uncertainty, In *Proceedings Fourth Berkeley Symposium on Mathematical Statistics and Probability I* (J. Neyman, ed.), 165-176.

Dantzig, G.B. and M.V.F. Pereira, et al. (1988), Mathematical Decomposition Techniques for Power System Expansion Planning, EPRI EL-5299, Volumes 1 - 5, Electric Power Research Institute, Palo Alto, Ca.

Dantzig G.B. and A.F. Perold (1979), A basic factorization method for block triangular linear programs, Report SOL 78-7, Department of Operations Research, Stanford University, Stanford, Ca.; *Sparse Matrix Proceedings*, 1978, Iain S. Duff and G. W. Steward, Eds. SIAM, 283-312.

Dantzig, G.B. and P. Wolfe (1960), The decomposition principle for linear programs, *Operations Research*, **8**, 110-111.

Dantzig, G. B., P. W. Glynn, M. Avriel, J. C. Stone, R. Entriken, and M. Nakayama (1989), *Decomposition Techniques for Multi-Area Transmission Planning Under Uncertainty* (1989). Final report of EPRI Project RP 2940–1, prepared by Systems Optimization Laboratory, Operations Research Department, Stanford University, Ca. 94305.

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

Emoliev, Y. (1983), Stochastic Quasigradient Methods and Their Applications to Systems Optimization, *Stochastics*, **9**, 1-36.

Entriken, R. (1988), A Parallel Decomposition Algorithm for Staircase Linear Programs, Oak Ridge National Laboratory Report ORNL/TM 11011, November 1988.

EPRI (1986), TAG, Technical Assessment Guide, Vol. 1, Electricity Supply—1985, Electric Power Research Institute, 1986.

Fourer, R.H. (1982), Solving staircase linear programs by the simplex method, 1: Inversion, *Mathematical Programming*, **23**, 274-313.

Fourer, R.H. (1983), Solving staircase linear programs by the simplex method, 2: Pricing, *Mathematical Programming*, **25** 251-292.

Fourer, R.H. (1984), Staircase matrices and systems, *SIAM Review*, **26**, 1-70.

Glassey, R. (1973), Nested decomposition and multi-stage linear programs, *Management Science*, **20**, 282-292.

Glynn, P.W. and W. Whitt (1988), Efficiency of Simulation Estimates. Submitted for publication.

Glynn, P.W. and D.L. Iglehart (1988), Importance Sampling for Stochastic Simulation. Submitted for publication.

Hammersley, J.M. and D.C. Handscomb (1964), *Monte Carlo Methods*, Mathuen, London.

Ho, J.K., C. Tak and R.P. Sundarraj (1988), Decomposition of linear programs using parallel computation (revised), Invited paper at the Symposium on Parallel Optimization, Madison, Wis.

Ho, J.K. and E. Loute (1980), An advanced implementation of the Dantzig-Wolfe decomposition algorithm for linear programming, Discussion Paper 8014, Center for Operations Research and Econometrics (CORE), Belgium.

Ho, J.K. and A.S. Manne (1974), Nested decomposition for dynamic models, *Mathematical Programming*, **6**, 121-140.

Jackson, P.L. and D.F. Lynch (1982), Revised Dantzig-Wolfe Decomposition for Staircase-Structured Linear Programs, Technical Report 558, School of Operations Research and Industrial Engineering, Cornell University (revised 1985), To appear in *Mathematical Programming.*

Kall, P. (1979), Computational Methods for Solving Two-Stage Stochastic Linear Programming Problems, *Z. Angew. Math. Phys.*, P **30**, 261-271.

Lavenberg, S.S. and P.D. Welch (1981), A perspective on the use of control variables to increase the efficiency of Monte Carlo simulation. *Management Science*, **27**, 322-335.

Louveaux, F.V. (1986), Multistage stochastic programs with block-separable recourse, *Mathematical Programming Study*, **28**, 48-62.

Manne, A.S. (1987), Interregional Electricity Trade—A Case Study of the Western Region, EA - Research Project, Final Report, August 1987.

Nazareth, L. and R.J-B. Wets (1986), Algorithms for Stochastic Programs: The Case of Nonstochastic Tenders, *Mathematical Programming Study*, **28**, 1-28.

NERC (1986), Electric Power Supply and Demand 1986-1995, Annual Data Summary Report for the Regional Reliability Councils of NERC, North American Electric Reliability Council, 1986.

Niederreiter, H. (1986), Multidimensional numerical integration using pseudo random numbers, *Mathematical Programming Study*, **27**, 17-38.

Nishiya, T. (1983), A basis factorization method for multi-stage linear programming with an application to optimal operation of an energy plant, (draft report),

Prekopa, A., (1978), Dynamic Type Stochastic Programming Models, In *Studies in Applied Stochastic Programming* (A. Prekopa, ed.), Hungarian Academy of Science, Budapest, 179-209.

Strazicky, B. (1980), Computational Experience with an Algorithm for Discrete Recourse Problems, In *Stochastic Programming* (M. Dempster, ed.), Academic Press, London, 263-274.

Wets, R.J. (1984), Programming under uncertainty: the equivalent convex program, *J. SIAM Applied Math.*, **14**, (1), 89-105.

Wets, R.J. (1984), Large-scale linear programming techniques in stochastic programming, In *Numerical Methods for Stochastic Optimization*, Y. Ermoliev and R. Wets, Eds. Springer-Verlag. IIASA WP-84-90.

Wets, R.J. (1985), On Parallel Processor Design for Solving Stochastic Programs, In *Proceedings of the 6th Mathematical Programming Symposium*, Japanese Mathematical Programming Society, Tokyo, 13-36.

Whittrock, R.J. (1983), Advances in a nested decomposition algorithm for solving staircase linear programs, Report SOL 83-2, Department of Operations Research, Stanford University, Stanford, Ca.