Models and Computational Strategies for Multistage Stochastic Programming under Endogenous and Exogenous Uncertainties

Robert M. Apap, Ignacio E. Grossmann[1]
Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, United States

Abstract
In this work, we address the modeling and solution of mixed-integer linear multistage stochastic programming problems involving both endogenous and exogenous uncertain parameters. We first propose a composite scenario tree that captures both types of uncertainty, and we exploit its unique structure to derive new theoretical properties that can drastically reduce the number of non-anticipativity constraints (NACs). Since the reduced model is often still intractable, we discuss two special solution approaches. The first is a sequential scenario decomposition heuristic in which we sequentially solve endogenous MILP subproblems to determine the binary investment decisions, fix these decisions to satisfy the first-period and exogenous NACs, and then solve the resulting model to obtain a feasible solution. The second is Lagrangean decomposition. We present numerical results for a process network and an oilfield development planning problem. The results clearly demonstrate the efficiency of the special solution methods over solving the reduced model directly.

Keywords: Multistage stochastic programming; endogenous uncertainty; exogenous uncertainty; scenario tree; non-anticipativity constraints; Lagrangean decomposition; oilfield planning.

1 Introduction
In the optimization of process systems, there is often some level of uncertainty in one or more of the input parameters. A major challenge for the decision-maker, then, is to determine how to best account for this uncertainty. Rather than optimizing for expected values, which can often lead to suboptimal or even infeasible solutions, these problems can usually be effectively approached with mathematical programming techniques such as stochastic programming (Birge and Louveaux, 2011), robust optimization (Ben-Tal et al., 2009), or chance-constrained optimization (Li et al., 2008).

In stochastic programming, the topic of this paper, a decision-maker must implement a set of decisions at the beginning of the planning horizon without knowing exactly what the true values of some of the input parameters will be. After the uncertainty in those parameters is resolved, the decision-maker can take corrective action based on this new information. Since this approach does not fix all of the decisions at the beginning of the planning horizon, it tends to be an appropriate choice for long-term planning projects that may span several decades (Grossmann et al., 2016).

In robust optimization, on the other hand, the general goal is to guarantee feasibility over a specified uncertainty set. This is typically more appropriate for short-term scheduling problems where feasibility is a major concern and where there is little scope for corrective action (Grossmann et al., 2016). Chance-constrained optimization also has a similar emphasis on constraint feasibility; specifically, some of the

[1] Author to whom correspondence should be addressed. Tel.: +1 412 268 3642; fax: +1 412 268 7139; email: grossmann@cmu.edu.

constraints must be satisfied with at least a given level of probability for all possible outcomes of the uncertain parameters present in those respective constraints (Calfa et al., 2015). As our intended applications are long-term planning problems in which corrective action is essential and probabilistic constraints are not required, we focus on a stochastic programming framework to effectively hedge against parameter uncertainties. Extensions of robust optimization and chance-constrained optimization that allow for some corrective action will not be considered here, but a discussion of these approaches can be found in Ben-Tal et al. (2004) (as well as Lappas and Gounaris (2016) and Zhang et al. (2016)) and Liu et al. (2016), respectively.

A second major concern for the decision-maker is the *type* of uncertainty. In general, there are two types: *exogenous*, where the true parameter values are revealed independently of decisions, and *endogenous*, where the parameter realizations are influenced by the decisions (Jonsbråten, 1998). In the context of process systems engineering, exogenous uncertainties often correspond to market uncertainties, such as crude-oil prices. The corresponding realizations occur automatically in each period of the planning horizon, independently of any decisions. For example, in an oilfield planning problem, we may rely on a forecast to predict the price of oil in the upcoming year. At the time the forecast is prepared, the true price is unknown. Once next year arrives, however, we will realize the true price of oil, regardless of the decisions that have been made.

For endogenous uncertainties, we may be dealing with at least two distinct types which we will refer to as Type 1 and Type 2 (Goel and Grossmann, 2006). In the case of Type-1 endogenous uncertainties, decisions influence the parameter realizations by altering the underlying probability distributions for the uncertain parameters. A simple example of this may be an oil company's decision to flood the market in order to force a competitor out of business. Here the uncertainty is no longer strictly exogenous, as the decision will make lower oil-price realizations more probable. This type has been considered in relatively few stochastic-programming publications; namely, as far as we are aware: Ahmed (2000), Viswanath et al. (2004), Flach (2010), Peeta et al. (2010), Escudero, Garín, Merino, and Pérez (2013) (which considers both exogenous and Type-1 endogenous uncertainties), Laumanns et al. (2014), and Hellemo (2016).

In the case of Type-2 endogenous uncertainties, decisions influence the parameter realizations by affecting the time at which we observe these realizations. This refers specifically to technical parameters, such as oilfield size, for which the true values cannot be determined until a particular investment decision is made (Goel and Grossmann, 2006). For instance, seismic studies may provide a good indication of the size of an oilfield, but we will not know the exact recoverable oil volume until we drill the field and begin producing from it (Goel and Grossmann, 2004). Note that Type-1 and Type-2 endogenous uncertainties are not mutually exclusive; for example, the choice of drilling technology may make higher oil recoveries more likely (Type 1), but the true recovery will only be revealed if we decide to develop that field (Type 2). This case is referred to as Type-3 endogenous uncertainty in Hellemo (2016).

It is worth noting that Powell (2011) classifies problems with either Type-1 or Type-2 endogenous uncertainty as "state-dependent information processes" and recommends the use of approximate dynamic programming (ADP) to solve them. In fact, dynamic programming methods have been successfully applied to optimization problems involving exogenous uncertainties (e.g., Powell (2011)), Type-1 endogenous uncertainties (e.g., Webster et al. (2012)), and Type-2 endogenous uncertainties (e.g., Choi et

al. (2004)). Such methods are outside the scope of this paper; however, we refer the reader to these selected references for further details.

For the endogenous uncertainties considered here, we will focus exclusively on Type 2, where decisions affect the timing of realizations. This is sometimes referred to as 'exogenous uncertainty with endogenous observation,' in view of the fact that the technical uncertainty itself is exogenous (as we cannot alter it), but the time at which this uncertainty is resolved is endogenous (since it depends on our investment decisions) (Colvin and Maravelias, 2011; Mercier and Van Hentenryck, 2011). For the remainder of this paper, we will drop the "Type 2" prefix and simply refer to these uncertainties as *endogenous*.

The literature on stochastic programming (SP) has focused primarily on problems with exogenous uncertainties. Reviews of this area are given in Birge (1997), Schultz (2003), and Sahinidis (2004). Endogenous uncertainty is a newer area and has received far less attention in the literature, with the first publication introduced by Jonsbråten et al. (1998) less than 20 years ago.[2]

In the area of process systems engineering, Goel and Grossmann (2004) and Goel et al. (2006) addressed a gas-field development problem in which the size and initial deliverability of reserves are uncertain, and these endogenous uncertainties are resolved immediately after the drilling decisions are made. Tarhan and Grossmann (2008) explored the synthesis of process networks with endogenous uncertainty in the process yields and relaxed the common assumption of immediate resolution of the uncertainty. Instead, the authors modeled the gradual resolution of uncertainty over time, which is more in line with reality in certain applications. Tarhan et al. (2009) applied this approach to the oil/gas-field development problem and considered nonlinearities in the reservoir model. Boland et al. (2008) studied the open pit mine production scheduling problem with endogenous uncertainty in the geological properties of the mined materials. The authors proposed a lazy-constraints approach for handling the large number of non-anticipativity constraints, whereby these constraints are only added to the problem as needed.

Colvin and Maravelias (2010) (an extension of Colvin and Maravelias (2008, 2009)) considered endogenous uncertainty in the scheduling of pharmaceutical clinical trials, and proposed a branch-and-cut method for this problem, as well as several theoretical reduction properties. Although many of these reduction properties are specific to the pharmaceutical scheduling problem, one applies to the general case considered here and will be discussed later in this paper. Solak et al. (2010) studied R&D project portfolio management under endogenous uncertainty, where the investment requirement for each project resolves gradually as a function of the progress of the respective project. The authors solved the resulting model with the sample average approximation method. The sample problems in this method were solved through the use of Lagrangean relaxation and a heuristic. In a related study, Colvin and Maravelias (2011) explored endogenous uncertainty in R&D activities in an R&D pipeline management problem, and also explored risk management strategies in this context.

Gupta and Grossmann (2011) discussed process networks with endogenous uncertainty in process yields, and proposed a general theoretical property that can considerably reduce the dimensionality of the model when there are uncertain parameters defined with three or more possible realizations. Gupta and

[2] Jonsbråten et al. (1998) is the first work to address the specific case considered here, where decisions must be made in order to gain more accurate process information. Pflug (1990) is the first work (of which we are aware) to consider the case of a decision-dependent stochastic process.

Grossmann (2014a) developed a scenario grouping Lagrangean decomposition algorithm for solving large-scale problems of this class (which is similar in concept to the scenario clustering approach of Escudero, Garín, Pérez, and Unzueta (2013) for two-stage exogenous problems). Gupta and Grossmann (2014b) also made advances in the modeling of the oilfield development planning problem under endogenous uncertainty. More recently, Christian and Cremaschi (2015) proposed two heuristic solution methods for the R&D pipeline management problem: a shrinking-horizon, multiple two-stage stochastic programming decomposition algorithm, and a knapsack decomposition algorithm. Other publications on stochastic programming under endogenous uncertainty which we will not discuss here, but may be of interest to the reader, include: multistage stochastic network interdiction (Held and Woodruff, 2005); the decision-rule approach to multistage stochastic programming (Vayanos et al., 2011); the optimal design of integrated chemical-production sites (Terrazas-Moreno et al., 2012); computational strategies for nonconvex, multistage mixed-integer nonlinear programs (Tarhan et al., 2013); and the dynamic single-vehicle routing problem with uncertain demands (Hooshmand Khaligh and MirHassani, 2016).

Although many problems contain both endogenous and exogenous uncertainties (e.g., uncertain field sizes *and* uncertain oil prices), optimization under both types has been largely unexplored in the literature.[3] To the best of our knowledge, Goel and Grossmann (2006) has been the only previous work to comprehensively explore multistage stochastic programming (MSSP) problems of this class. The authors introduced a hybrid mixed-integer linear disjunctive programming model for these problems and proposed two efficient theoretical properties for eliminating redundant constraints; however, their numerical studies considered only endogenous uncertainties in capacity expansion and sizing problems. Dupačová (2006) briefly discussed optimization under both types of uncertainty but did not provide a specific multistage formulation, new solution strategies, or numerical results. More recently, Bruni et al. (2015) proposed a stochastic programming approach for the operating theater scheduling problem, in which there is exogenous uncertainty in the arrival of emergency patients and endogenous uncertainty in the duration of surgery. The authors offered only brief details on the modeling of the endogenous uncertainty and employed a heuristic approach to solve the problem. As our focus is on a general framework for multistage stochastic programming, Goel and Grossmann (2006) will serve as the foundation for this paper.

The goals of this work are to: (1) efficiently model multistage stochastic programming problems that involve both endogenous and exogenous parameters; (2) develop effective solution methods for these problems; and (3) apply the proposed methods to challenging applications. Given the complexity of these problems and the fact that only little work has been reported on them, we begin in section 2 with a detailed review of the relevant background regarding multistage stochastic programming under exogenous uncertainty, as well as multistage stochastic programming under endogenous uncertainty. In section 3, we then introduce the definitions and notation necessary to model these types of uncertainties, and propose a composite scenario tree that captures all possible realizations of both endogenous and exogenous parameters. Next, in section 4, we present the multistage stochastic programming models for

[3] It is worth noting that there is also a significant lack of literature on stochastic programs with both exogenous and Type-1 endogenous uncertainties. Tong et al. (2012) addressed demand and yield uncertainties in a risk-averse oil supply chain planning problem; however, the authors represented the product yield fluctuations with a Markov chain and solved the problem using an iterative heuristic algorithm that integrates two-stage stochastic programming and simulation. For a discussion of modeling and solution considerations for multistage stochastic programs with both exogenous and Type-1 endogenous uncertainties, see Escudero, Garín, Merino, and Pérez (2013).

purely exogenous uncertainty, purely endogenous uncertainty, and both endogenous and exogenous uncertainties. After this point, we focus our attention on the latter case, and in section 5, we discuss reduction properties that can significantly reduce the dimensionality of these problems. Finally, in section 6, we introduce a sequential scenario decomposition heuristic and briefly review Lagrangean decomposition, and then apply these algorithms in section 7 to solve a process network example and an oilfield development planning problem.

2 Background

2.1 Stochastic Programming under Exogenous Uncertainty

A common approach for optimization under exogenous uncertainty is two-stage stochastic programming (Birge and Louveaux, 2011). In this approach, first-stage decisions are made 'here and now' at the beginning of the first time period, without knowing exactly how the uncertainty will unfold. The decision-maker then waits for the outcome. At some point following these decisions, the uncertainty is resolved and the true values of the exogenous-uncertain parameters become known. Second-stage, or recourse ('wait-and-see'), decisions are then taken by the decision-maker as corrective action. For example, in a problem spanning multiple time periods, the decision-maker's first-stage decisions may enforce an investment plan that is fixed for the entire horizon. Subsequent recourse decisions allow operating conditions to be specified in response to this plan, based on the realizations observed for the exogenous-uncertain parameters (see, for instance, Liu and Sahinidis (1996)).

In practice, however, it is often necessary for the decision-maker to have the additional freedom to make new here-and-now decisions at the beginning of each time period. This leads to a multistage stochastic programming formulation; decisions, realizations, and recourse actions occur sequentially, allowing for a more accurate description of the decision-making process for long-term planning projects. This is illustrated in Figure 1 for a three-stage problem with one exogenous-uncertain parameter, ξ_t. We use y_t^s and x_t^s to denote the vectors of here-and-now decisions and recourse decisions, respectively, in each time period t and scenario s. Note that $t = 0$ corresponds to the beginning of the first time period (stage 1), $t = 1$ corresponds to the end of the first time period/beginning of the second time period (stage 2), and $t = 2$ corresponds to the end of the second time period (stage 3). As will be discussed, multistage stochastic programming also provides a more suitable framework for endogenous uncertainties, as these realizations can occur at any point in the time horizon.

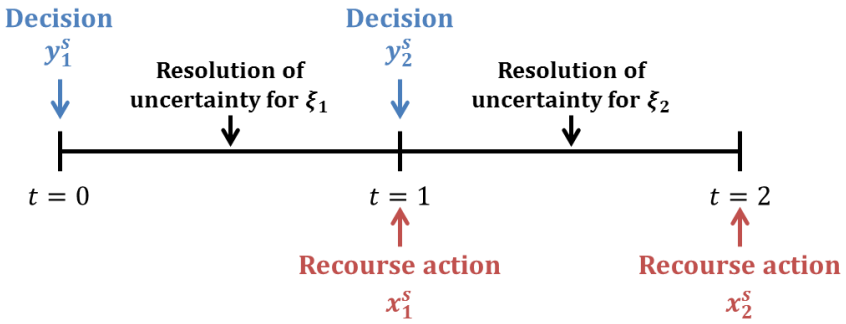
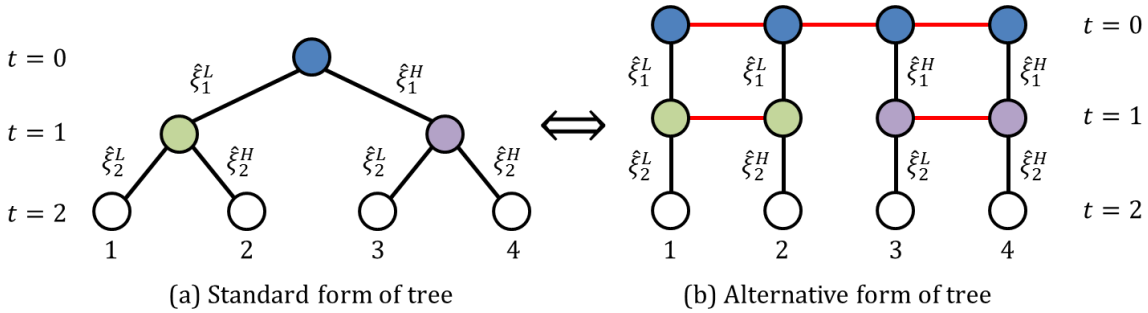


Figure 1. Sequence of events in multistage stochastic programming under exogenous uncertainty.

One fundamental assumption in stochastic programming, which we have already made here, is that the

time horizon is represented by a set of discrete time periods. A second very common assumption is that the possible realizations (possible values) for each uncertain parameter are available from a discretized probability distribution. With these two assumptions in place, the stochastic process can be represented by a scenario tree, like that shown in Figure 2a. Note that this is the scenario-tree representation of Figure 1, where the exogenous parameter ξ_t has two possible realizations (*low* (L) or *high* (H)) in each time period. Each node in the tree represents a different possible state of the system in time period t. Arcs indicate a possible transition from a state in time period t to a new state in time period $t+1$, with a given probability of this transition occurring. For example, the system shown in Figure 2a can transition from its initial state in time period 1 to either of two different states in time period 2 depending upon the realized value of ξ_1. A complete path from the root node to a leaf node represents a scenario, which corresponds to one possible combination of realizations for the uncertain parameters (e.g., $(\hat{\xi}_1^L, \hat{\xi}_2^L)$). Note that since the uncertainty is purely exogenous in this case, and exogenous realizations occur automatically in each time period, the structure of the scenario tree is known in advance.



(a) Standard form of tree (b) Alternative form of tree

Figure 2. An exogenous scenario tree and its alternative representation.

One complicating aspect of the standard form of the scenario tree (Figure 2a) is that the corresponding stochastic programming problem contains variables that are shared among two or more scenarios. For instance, in Figure 2a, all four scenarios share the variables of the root node (shown in blue), scenarios 1 and 2 share the variables of the green node at $t = 1$, and scenarios 3 and 4 share the variables of the purple node at $t = 1$. This prevents the direct application of scenario-decomposition approaches like Lagrangean decomposition which can be effective for solving large stochastic programs.

Ruszczyński (1997) proposed an alternative form of the scenario tree in which shared nodes are split such that each scenario is given its own unique set of nodes. This is shown in Figure 2b. The alternative form is more amenable to scenario decomposition, as variables are no longer shared and each scenario represents a different instance of the same deterministic problem with different realizations for the uncertain parameters. Notice, however, that in moving from the standard form of the tree to the alternative form, we have created several copies of the same states. For example, the root node in Figure 2a has been split into four separate nodes in Figure 2b. These four nodes all have identical information at that point in time. Accordingly, scenarios 1-4 are said to be *indistinguishable* at the beginning of the first time period. It follows that because these scenarios are indistinguishable at that time, we must treat them all in the same way, and we must make the same here-and-now decisions in all four scenarios at the beginning of the first time period. This equality between the states is enforced by the red horizontal lines connecting the nodes in Figure 2b. These red lines represent what are known as *non-anticipativity constraints* (Rockafellar and Wets, 1991; Ruszczyński, 1997). Without these constraints, it is clear that

the tree would decompose into independent scenarios in which we would be anticipating *one* particular outcome for each of the uncertain parameters. Since we do not have this level of information, these constraints are required. Using the notation from Figure 1, we express the non-anticipativity constraints (NACs) as $y_1^1 = y_1^2$, $y_1^2 = y_1^3$, and $y_1^3 = y_1^4$. Similarly, as can be seen from the green nodes at $t = 1$ in Figure 2b, we must make the same recourse decisions at the end of the first time period and the same here-and-now decisions at the beginning of the second time period in scenarios 1 and 2. Thus, the corresponding NACs are $x_1^1 = x_1^2$, and $y_2^1 = y_2^2$. A similar argument can be made regarding the purple nodes at $t = 1$ and the corresponding decisions in scenarios 3 and 4. Notice that by the end of the time horizon, all scenarios differ in the realizations of exogenous parameter ξ_t, and the leaf nodes refer to independent states. Accordingly, the scenarios are said to be *distinguishable* at that time, and non-anticipativity no longer applies (as noted by the absence of any red lines connecting the scenarios). In other words, at the end of the second time period, we are free to make independent recourse decisions in each of the four scenarios.

It is important to note that the alternative form of the scenario tree corresponds directly to the non-anticipativity formulation of the *deterministic equivalent* for stochastic programming problems (Birge and Louveaux, 2011). In this formulation, as the preceding discussion suggests, each scenario represents a different instance of the deterministic problem with different realizations for the uncertain parameters, and non-anticipativity constraints ensure that we make the same decisions in indistinguishable scenarios in each time period. This is the modeling approach that will be used in this paper. We will rely heavily on the concept that two scenarios are indistinguishable in time period t if they are identical in the realizations of all uncertain parameters that have been resolved up until that time;[4] and as soon as the scenarios differ in the realization of any uncertain parameter, they are distinguishable for the remainder of the time horizon. As we will describe in the next section, the alternative form of the scenario tree is also very useful in modeling endogenous uncertainties.

2.2 Stochastic Programming under Endogenous Uncertainty

In the case of stochastic programming under endogenous uncertainty, a multistage framework is generally the logical starting point. This can be seen when considering a problem such as the capacity expansion of process networks (Goel and Grossmann, 2006), where small installations are made in early time periods to determine the true yields of new process units. Capacity expansions can then be made at a later point in time to capitalize on that knowledge. This type of decision making is not possible with only two stages. Furthermore, in the two-stage case, if investments are not made at the beginning of the first time period (as this may not be optimal), the uncertainty in the endogenous parameters cannot be resolved during the time horizon.

The decision-making process in these types of multistage stochastic programming problems proceeds in a manner similar to that of the exogenous case (Figure 1). The primary difference here is that the timing of realizations depends on the decisions. Hence, uncertainty is not resolved automatically in each time period, and the uncertainty in some parameters may not be resolved at all. This is illustrated in Figure 3 for a three-stage problem with two endogenous-uncertain parameters, θ_1 and θ_2, where set $\overline{\mathcal{I}}_t^s$ indicates the parameters that are realized in each time period t of scenario s. It is important to note that rather than

[4] The phrase "indistinguishable in time period t" will be used as a shorthand way of stating: "indistinguishable at the end of time period t, after all realizations in that period have occurred."

being associated with a particular time period, endogenous parameters represent intrinsic properties of a given *source*, such as the size of an *oilfield* or the yield of a *process unit* (Goel and Grossmann, 2006). Accordingly, in the case of Figure 3, we state that θ_1 is an endogenous parameter associated with a given "Source 1," and θ_2 is an endogenous parameter associated with a given "Source 2."

Consider the case where we make an investment[5] in both Source 1 and Source 2 at the beginning of the first time period. Also, assume that the uncertainty is resolved immediately after we implement this decision. As indicated by the sequence of events in Figure 3, we will realize the values of θ_1 and θ_2 in the first time period, and no realizations will occur in the second time period (i.e., $\overline{\mathcal{I}}_1^s = \{1, 2\}$, and $\overline{\mathcal{I}}_2^s = \emptyset$). Notice that unlike the exogenous case, we do not know which parameters will be realized until we know which decisions we will make. This information is not known in advance and must be determined by solving the corresponding stochastic programming problem. We use dotted lines in Figure 3 to indicate that the timing of the realizations is conditional.

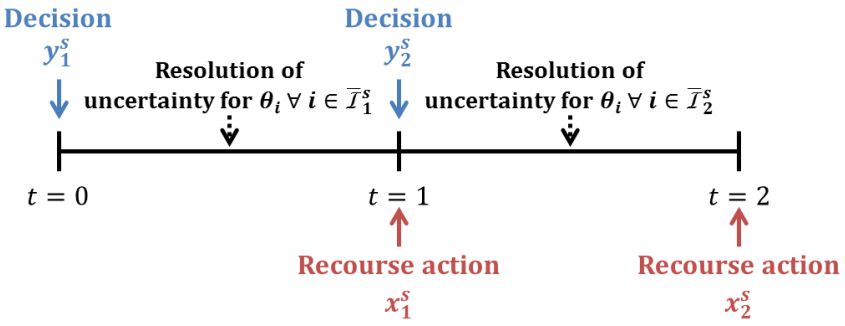
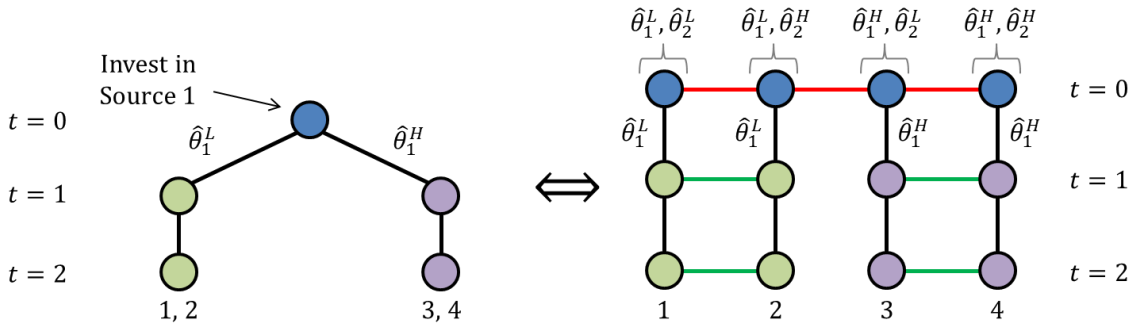


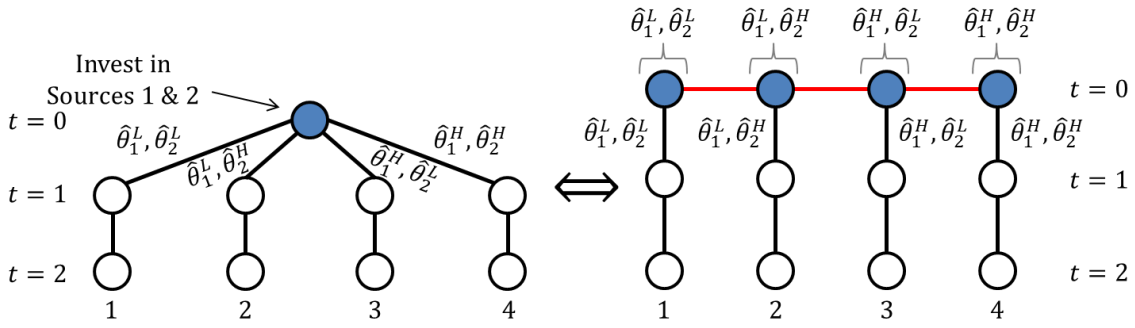
Figure 3. Sequence of events in multistage stochastic programming under endogenous uncertainty.

As this discussion suggests, the scenario-tree representation of these stochastic processes is also not as straightforward as the exogenous case. This is for the simple reason that there are many possible outcomes for the decisions, and accordingly, there will be many possible outcomes for the structure of the scenario tree. This is illustrated in Figure 4 with just a few of the many possible scenario-tree representations of Figure 3, where the endogenous parameters θ_1 and θ_2 each have two possible realizations (*low* (L) or *high* (H)). (Note that above each scenario in the alternative form of the tree, we indicate the possible realizations defined for that particular scenario.) We again assume that the uncertainty in a parameter is resolved immediately after an investment is made in its respective source. In the first case, Figure 4a, an investment is made in Source 1 at the beginning of the first time period. As a result, the value of θ_1 is immediately realized in all scenarios. Notice that non-anticipativity constraints still apply for the beginning of the first time period in the alternative form of the tree, just as they do in the exogenous case (i.e., we have the same red lines at $t = 0$ as we do in Figure 2b). This is because at the beginning of the time horizon (prior to the implementation of the decisions for the first time period), we have yet to make any decisions, and no realizations have occurred. Thus, all scenarios must be indistinguishable at that time, regardless of the type of uncertainty being considered.

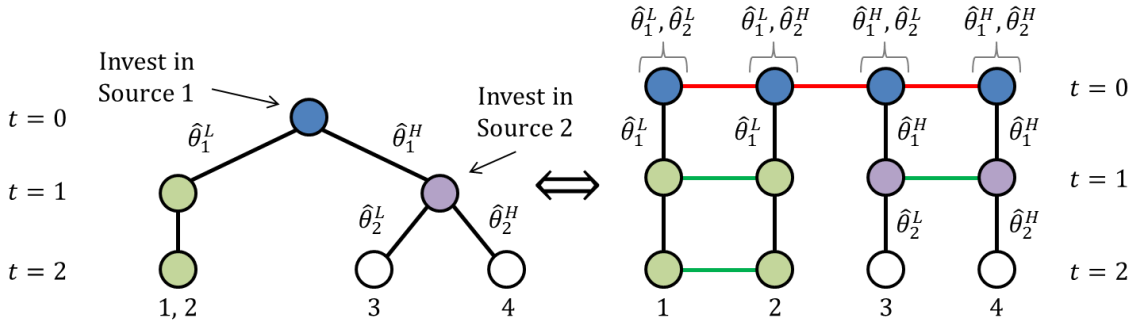
[5] An 'investment in a source' broadly refers to any here-and-now decision that allows us to realize the values of the endogenous parameters associated with that source.



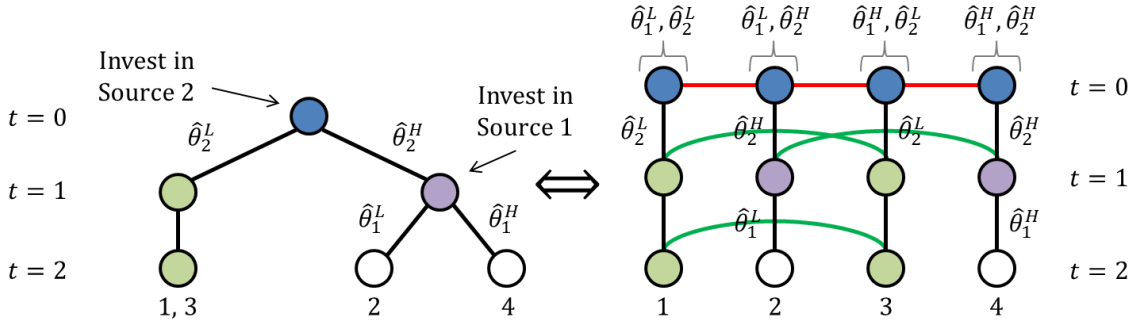
(a) θ_1 is realized during the first time period in all scenarios.



(b) θ_1 and θ_2 are both realized during the first time period in all scenarios.



(c) θ_1 is realized during the first time period in all scenarios. θ_2 is realized during the second time period in scenarios 3 and 4 only.



(d) θ_2 is realized during the first time period in all scenarios. θ_1 is realized during the second time period in scenarios 2 and 4 only.

Figure 4. Four possible structures for a scenario tree with two endogenous parameters.

Continuing with the discussion of Figure 4a, we note that no other investments are made after the first-stage decisions, so the value of θ_2 is never realized. Non-anticipativity constraints (shown in green) therefore restrict our decision-making such that, for the remainder of the time horizon, we must make all of the same decisions in scenarios 1 and 2, as well as all of the same decisions in scenarios 3 and 4. In the second case, Figure 4b, an investment is made in both Source 1 and Source 2 at the beginning of the first time period (this is the case that was previously described in relation to Figure 3). The values of θ_1 and θ_2 are immediately realized in all scenarios, and the four scenarios are distinguishable for the remainder of the time horizon. In other words, by the end of the first time period, we are free to make independent decisions in all scenarios. In Figure 4c and Figure 4d, an asymmetric scenario tree results from making an investment in only two of the four scenarios. By simply swapping the order of investments, the alternative tree in Figure 4d looks very different from that of Figure 4c, in the sense that non-anticipativity constraints no longer apply solely between adjacent scenarios. We again emphasize that many other outcomes for the tree are possible, even with only four scenarios.

Due to the conditional structure of the endogenous scenario tree, it is clearly impractical to model all possible outcomes with the standard form of the tree. To deal with this issue, we adopt the alternative form and create a superstructure in which non-anticipativity constraints are applied conditionally (as inspired by Gupta and Grossmann (2014a)). This is shown in Figure 5, where the dotted green lines represent these conditional NACs. Notice that the superstructure form of the tree accounts for all possible outcomes, and any of the alternative trees shown in Figure 4 can easily be recovered from Figure 5.

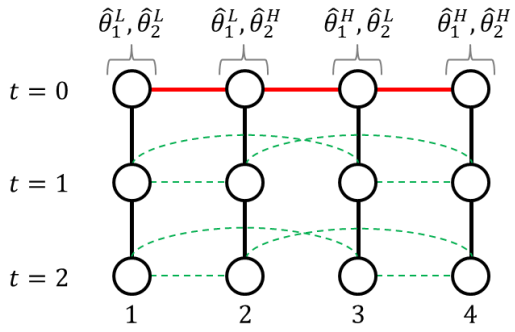


Figure 5. A superstructure representation for endogenous scenario trees.

Because we are now dealing with conditional NACs, the modeling approach is significantly different from the simple equality constraints for exogenous uncertainty. In the exogenous case, if two scenarios differ in the realization of uncertain parameter ξ_t in time period t_X^*, the scenarios will be distinguishable by the end of that time period (since realizations occur automatically). Therefore, we apply non-anticipativity constraints between these scenarios in all time periods up to, but not including, the end of t_X^*. In the endogenous case, however, it is not this simple. Scenarios that differ in the possible realization of an uncertain parameter θ_i will remain indistinguishable until the uncertainty in that parameter is resolved; up until that point, t_N^*, the scenarios are identical. Because we do not know the value of t_N^* for these scenarios, we must conditionally apply NACs for all decisions in all time periods (excluding the decisions made at the beginning of the first time period and in other initial time periods, as well). The indistinguishability is determined at each point in time as part of the stochastic programming problem, and the NACs are enforced if the scenarios are indistinguishable and ignored if they are not. As opposed

to a fixed scenario tree in the exogenous case, the optimal structure of the endogenous scenario tree is determined by solving this stochastic program. The modeling of NACs will be discussed in greater detail later in this paper.

3 Definitions and Notation

3.1 Mathematical Description of Exogenous Uncertainty

Let the time horizon be divided into a set of discrete time periods $\mathcal{T} \coloneqq \{t\colon\ t = 1, 2, \dots, T\}$, and let set $\mathcal{J} \coloneqq \{j\colon\ j = 1, 2, \dots, J\}$ define the index of each exogenous-uncertain parameter. We define $\xi_{j,t}$ as exogenous parameter $j \in \mathcal{J}$ in time period $t \in \mathcal{T}$. The exogenous parameter has a number of possible realizations given by the ordered set $\Xi_{j,t} \coloneqq \left\{\hat{\xi}_{j,t}^{r}\colon\ r = 1, 2, \dots, R_{j,t}\right\}$, where r refers to the index of one particular realization, and for convenience, we set $\hat{\xi}_{j,t}^{1} < \hat{\xi}_{j,t}^{2} < \cdots < \hat{\xi}_{j,t}^{R_{j,t}}$. As an example of how we use this notation, if $r = 2$ is the index of the actual realization for parameter j in time period t, we will have $\xi_{j,t} = \hat{\xi}_{j,t}^{2}$. The total number of possible realizations for this parameter is given by $\left|\Xi_{j,t}\right| = R_{j,t}$. Note that because the uncertainty in parameter j is exogenous, it is resolved automatically in each time period t, regardless of the decisions that have been made. In instances where there is only one exogenous parameter, we will frequently drop the *j* subscript to simplify the notation.

Each scenario in the model corresponds to *one* possible combination of realizations for the uncertain parameters. We assume that these parameters are independent (see the supplementary material) and that the complete set of scenarios corresponds to *all* possible combinations of their realizations. Accordingly, in the case where the uncertainty is purely exogenous, the complete set of scenarios $\mathcal{R}_X$ is represented by a Cartesian product over the sets of realizations for the exogenous parameters:

$$\mathcal{R}_X \coloneqq \times_{t\in\mathcal{T}}\left(\times_{j\in\mathcal{J}}\Xi_{j,t}\right) = \left\{\left(\hat{\xi}_{1,1}^{1}, \dots, \hat{\xi}_{J,T}^{1}\right), \dots, \left(\hat{\xi}_{1,1}^{R_{1,1}}, \dots, \hat{\xi}_{J,T}^{R_{J,T}}\right)\right\} \quad \textbf{(1)}$$

where we use the subscript X to indicate eXogenous. We enforce a lexicographical ordering on the Cartesian product (and all other Cartesian products in this paper) based on the index of each realization.[6] Set $\mathcal{R}_X$ corresponds to a scenario tree constructed from all possible combinations of realizations of the exogenous parameters; e.g., Figure 2. Note that in this figure there is only one exogenous parameter, so we have dropped the j subscript to simplify the notation, and we have $\mathcal{R}_X = \Xi_1 \times \Xi_2 = \left\{\hat{\xi}_1^L, \hat{\xi}_1^H\right\} \times \left\{\hat{\xi}_2^L, \hat{\xi}_2^H\right\} = \left\{\left(\hat{\xi}_1^L, \hat{\xi}_2^L\right),\ \left(\hat{\xi}_1^L, \hat{\xi}_2^H\right),\ \left(\hat{\xi}_1^H, \hat{\xi}_2^L\right),\ \left(\hat{\xi}_1^H, \hat{\xi}_2^H\right)\right\}$. The L and H superscripts here refer to the index of the *low* and *high* realizations for the uncertain parameter ($r = 1$ and $r = 2$), respectively. The cardinality of $\mathcal{R}_X$, or the number of exogenous scenarios, is simply equal to the product of the cardinality of all of the sets in the Cartesian product in Equation (1). In other words, the number of exogenous scenarios is equal to the product of the number of realizations for each exogenous parameter,

$$S_X \coloneqq \left|\mathcal{R}_X\right| = \prod_{t\in\mathcal{T}}\prod_{j\in\mathcal{J}} R_{j,t} \quad \textbf{(2)}$$

[6] For example, tuple $\left(\hat{\xi}_{1,1}^{1}, \hat{\xi}_{1,2}^{1}\right)$ would be placed before $\left(\hat{\xi}_{1,1}^{1}, \hat{\xi}_{1,2}^{2}\right)$ based on a comparison of the realization indices (i.e., superscripts) for each respective element: $1 = 1$ for the first element of the tuples, so we proceed to the second element and see that $1 \leq 2$. Tuple $\left(\hat{\xi}_{1,1}^{1}, \hat{\xi}_{1,2}^{2}\right)$ would be placed before $\left(\hat{\xi}_{1,1}^{2}, \hat{\xi}_{1,2}^{1}\right)$ since a comparison of the realization indices for the first element gives $1 < 2$ (and we do not consider the other elements in such a case).

This allows us to index the exogenous scenarios by defining the ordered set of indices $\mathcal{S}_X \coloneqq \{s\colon\ s = 1, 2, \dots, S_X\}$. Applying this analysis to Figure 2, we have $\mathcal{J} = \{1\}$, $\mathcal{T} = \{1, 2\}$, and $R_1 = R_2 = 2$, which gives $S_X = R_1 \cdot R_2 = 2 \cdot 2 = 4$. Accordingly, $\mathcal{S}_X = \{1, 2, 3, 4\}$. Note that if all exogenous parameters have the same number of realizations in all time periods (i.e., $\left|\Xi_{j,t}\right| = R\ \ \forall\, j \in \mathcal{J},\ \ t \in \mathcal{T}$), as they do in Figure 2, Equation (2) can be simplified to $S_X = R^{J \cdot T}$.

Since it will be necessary to know the realization of the exogenous parameter $\xi_{j,t}$ in each scenario $s \in \mathcal{S}_X$, we introduce the scenario index to this parameter to define $\xi_{j,t}^{s}$. Notice, however, that in order to assign a realization value to $\xi_{j,t}^{s}$ for each scenario, we must first establish a link between the scenario *index* (i.e., $s \in \mathcal{S}_X$) and the *actual scenario* that it represents (i.e., the corresponding tuple in $\mathcal{R}_X$). To do so, we first restate Equation (1) with the new notation: $\mathcal{R}_X = \left\{\left(\xi_{1,1}^{s}, \dots, \xi_{J,T}^{s}\right)\colon\ \forall\, s \in \mathcal{S}_X\right\}$. We then equate the right-hand side of this expression with the right-hand side of Equation (1) to give the value of $\xi_{j,t}^{s}$ for all $j \in \mathcal{J},\ \ t \in \mathcal{T}$, and $s \in \mathcal{S}_X$. For instance, for scenario $s = 1$, we are considering the first tuple in $\mathcal{R}_X$. Thus, we have $\left(\xi_{1,1}^{1}, \dots, \xi_{J,T}^{1}\right) = \left(\hat{\xi}_{1,1}^{1}, \dots, \hat{\xi}_{J,T}^{1}\right)$, which implies $\xi_{1,1}^{1} = \hat{\xi}_{1,1}^{1}, \dots, \xi_{J,T}^{1} = \hat{\xi}_{J,T}^{1}$. Similarly, for scenario $s = S_X$, we are considering the final tuple in $\mathcal{R}_X$. Now we have $\left(\xi_{1,1}^{S_X}, \dots, \xi_{J,T}^{S_X}\right) = \left(\hat{\xi}_{1,1}^{R_{1,1}}, \dots, \hat{\xi}_{J,T}^{R_{J,T}}\right)$, which implies $\xi_{1,1}^{S_X} = \hat{\xi}_{1,1}^{R_{1,1}}, \dots, \xi_{J,T}^{S_X} = \hat{\xi}_{J,T}^{R_{J,T}}$. The same reasoning applies for all other scenarios in $\mathcal{S}_X$.

3.2 Mathematical Description of Endogenous Uncertainty

Let set $\mathcal{I} \coloneqq \{i\colon\ i = 1, 2, \dots, I\}$ represent the sources of endogenous uncertainty, and let set $\mathcal{H}_i \coloneqq \{h\colon\ h = 1, 2, \dots, H_i\}$ define the index of each endogenous-uncertain parameter associated with source $i \in \mathcal{I}$. We define $\theta_{i,h}$ as endogenous parameter $h \in \mathcal{H}_i$ for source $i \in \mathcal{I}$. Recall that we must consider the *source* of uncertainty for each endogenous parameter because the realization for that parameter will only occur once a certain decision has been made for that source. For instance, if source $i = 1$ is an oilfield that has not yet been drilled, the values of the associated endogenous parameters (e.g., oilfield size and initial deliverability) will only be resolved once the oilfield has been drilled. Parameter $\theta_{i,h}$ has a number of possible realizations given by the ordered set $\Theta_{i,h} \coloneqq \left\{\hat{\theta}_{i,h}^{m}\colon\ m = 1, 2, \dots, M_{i,h}\right\}$, where m refers to the index of one particular realization, and for convenience, we set $\hat{\theta}_{i,h}^{1} < \hat{\theta}_{i,h}^{2} < \cdots < \hat{\theta}_{i,h}^{M_{i,h}}$. Thus, if $m = 2$ is the index of the actual realization for endogenous parameter h of source i, we will have $\theta_{i,h} = \hat{\theta}_{i,h}^{2}$. The total number of possible realizations for this parameter is given by $\left|\Theta_{i,h}\right| = M_{i,h}$. We emphasize that, unlike exogenous uncertainty, the resolution of uncertainty in $\theta_{i,h}$ depends on the timing of decisions related to source i and is not an automatic occurrence in each time period. When there is only one endogenous parameter associated with each source of uncertainty, we will often drop the h subscript to simplify the notation.

In the case where the uncertainty is purely endogenous and the uncertain parameters are independent, the complete set of scenarios $\mathcal{R}_N$ is represented by a Cartesian product over the sets of realizations for the endogenous parameters:

$$\mathcal{R}_N \coloneqq \times_{i \in \mathcal{I}} \left(\times_{h \in \mathcal{H}_i} \Theta_{i,h}\right) = \left\{\left(\hat{\theta}_{1,1}^{1}, \dots, \hat{\theta}_{I,H_I}^{1}\right), \dots, \left(\hat{\theta}_{1,1}^{M_{1,1}}, \dots, \hat{\theta}_{I,H_I}^{M_{I,H_I}}\right)\right\} \qquad \textbf{(3)}$$

where we use the subscript N to indicate eNdogenous. Set $\mathcal{R}_N$ corresponds to a scenario tree constructed from all possible combinations of realizations of the endogenous parameters; e.g., Figure 5. Note that in this figure there is only one endogenous parameter associated with each of the two sources, so we have dropped the h subscript for simplicity in the notation (as we did for the j subscript in Figure 2), and we have $\mathcal{R}_N = \Theta_1 \times \Theta_2 = \{\hat{\theta}_1^L, \hat{\theta}_1^H\} \times \{\hat{\theta}_2^L, \hat{\theta}_2^H\} = \{(\hat{\theta}_1^L, \hat{\theta}_2^L),\ (\hat{\theta}_1^L, \hat{\theta}_2^H),\ (\hat{\theta}_1^H, \hat{\theta}_2^L),\ (\hat{\theta}_1^H, \hat{\theta}_2^H)\}$. The cardinality of $\mathcal{R}_N$, or the number of endogenous scenarios, is simply equal to the product of the number of realizations for each endogenous parameter,

$$S_N \coloneqq |\mathcal{R}_N| = \prod_{i \in \mathcal{I}} \prod_{h \in \mathcal{H}_i} M_{i,h} \tag{4}$$

This allows us to index the endogenous scenarios by defining the ordered set of indices $\mathcal{S}_N \coloneqq \{s\colon\ s = 1, 2, \dots, S_N\}$. In the context of Figure 5, we have $\mathcal{I} = \{1, 2\}$, $\mathcal{H}_1 = \mathcal{H}_2 = \{1\}$, and $M_1 = M_2 = 2$. Thus, $S_N = M_1 \cdot M_2 = 2 \cdot 2 = 4$, and $\mathcal{S}_N = \{1, 2, 3, 4\}$. If all endogenous parameters have the same number of realizations (i.e., $|\Theta_{i,h}| = M\ \ \forall\, i \in \mathcal{I},\ \ h \in \mathcal{H}_i$), as is the case in Figure 5, Equation (4) can be simplified to $S_N = M^{\sum_{i \in \mathcal{I}} H_i}$.

As in the exogenous case, we also assign the index s to the endogenous parameter $\theta_{i,h}$ to indicate the parameter's realization in each scenario; i.e., $\theta_{i,h}^s$. Using this notation, we restate Equation (3) as $\mathcal{R}_N = \{(\theta_{1,1}^s, \dots, \theta_{I,H_I}^s)\colon\ \forall\, s \in \mathcal{S}_N\}$, and equate the right-hand side of this expression with the right-hand side of Equation (3) to give the value of $\theta_{i,h}^s$ for all $i \in \mathcal{I},\ \ h \in \mathcal{H}_i$, and $s \in \mathcal{S}_N$.

3.3 Mathematical Description of Endogenous and Exogenous Uncertainties

We now consider the case where we have both endogenous and exogenous uncertain parameters. Because these parameters are entirely independent of one another, we must ensure that we can observe *any* possible combination of realizations for the exogenous parameters, regardless of the outcome for the endogenous parameters (and vice versa). Accordingly, we generate the complete set of scenarios $\mathcal{R}$ by the Cartesian product of all possible combinations of realizations of the endogenous parameters and all possible combinations of realizations of the exogenous parameters, $\mathcal{R}_N \times \mathcal{R}_X$:

$$\mathcal{R} \coloneqq \mathcal{R}_N \times \mathcal{R}_X = \begin{Bmatrix} (\hat{\theta}_{1,1}^1, \dots, \hat{\theta}_{I,H_I}^1, \hat{\xi}_{1,1}^1, \dots, \hat{\xi}_{J,T}^1), \dots, (\hat{\theta}_{1,1}^1, \dots, \hat{\theta}_{I,H_I}^1, \hat{\xi}_{1,1}^{R_{1,1}}, \dots, \hat{\xi}_{J,T}^{R_{J,T}}), \dots, \\ (\hat{\theta}_{1,1}^{M_{1,1}}, \dots, \hat{\theta}_{I,H_I}^{M_{I,H_I}}, \hat{\xi}_{1,1}^1, \dots, \hat{\xi}_{J,T}^1), \dots, (\hat{\theta}_{1,1}^{M_{1,1}}, \dots, \hat{\theta}_{I,H_I}^{M_{I,H_I}}, \hat{\xi}_{1,1}^{R_{1,1}}, \dots, \hat{\xi}_{J,T}^{R_{J,T}}) \end{Bmatrix} \tag{5}$$

Set $\mathcal{R}$ corresponds to a "composite" scenario tree that includes all possible combinations of realizations of the endogenous and exogenous parameters. Although there are other ways to generate such a set (e.g., $\mathcal{R}_X \times \mathcal{R}_N$ (see the supplementary material)), we focus our attention on this approach since the resulting scenario tree has a structure that can be exploited to significantly reduce the dimensionality of the model (as will be discussed in section 5). For the remainder of this paper, we assume that the scenario tree has been generated in this manner. The total number of scenarios (i.e., the cardinality of $\mathcal{R}$) is equal to the product of the number of endogenous scenarios and the number of exogenous scenarios,

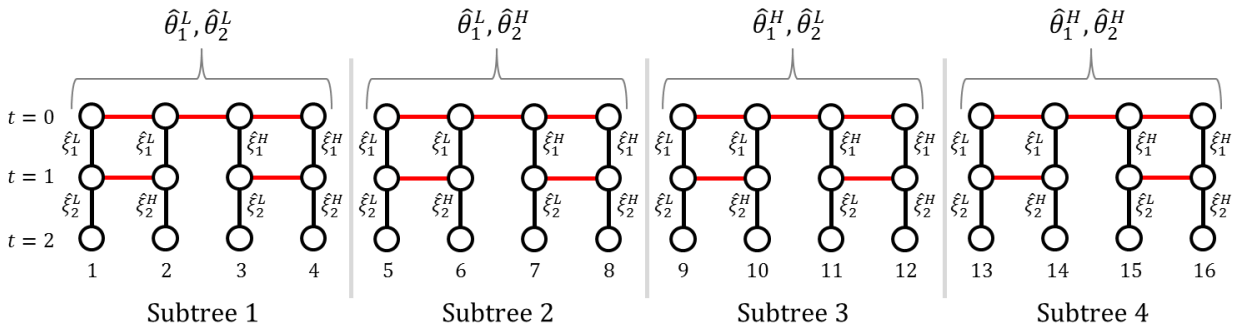
$$S \coloneqq |\mathcal{R}| = S_N \cdot S_X \tag{6}$$

We index the set of scenarios by defining the ordered set of indices $\mathcal{S} \coloneqq \{s\colon\ s = 1, 2, \dots, S\}$. We use this set to restate Equation (5) as $\mathcal{R} = \left\{\left(\theta^s_{1,1}, \dots, \theta^s_{I,H_I}, \xi^s_{1,1}, \dots, \xi^s_{J,T}\right)\colon\ \forall\, s \in \mathcal{S}\right\}$, and we equate the right-hand side of this expression with the right-hand side of Equation (5) to give the values of $\theta^s_{i,h}$ and $\xi^s_{j,t}$ for all $i \in \mathcal{I},\ h \in \mathcal{H}_i,\ j \in \mathcal{J},\ t \in \mathcal{T}$, and $s \in \mathcal{S}$.

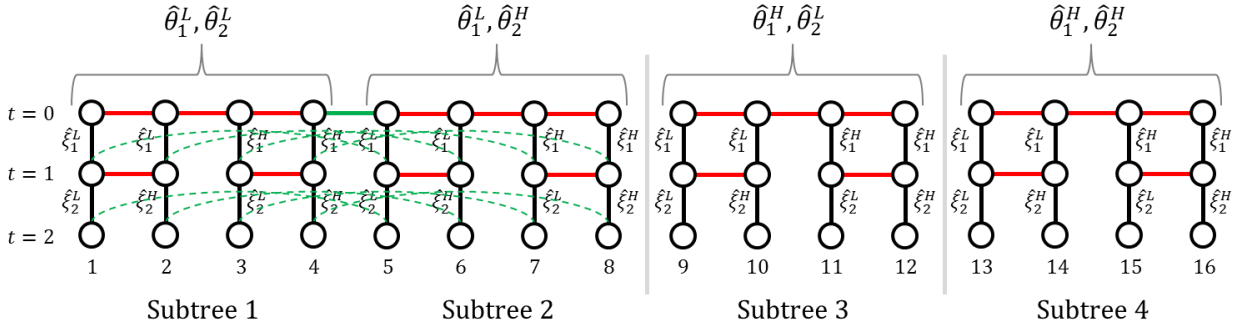
The generation of the composite scenario tree is shown in Figure 6. We consider the case where we have one exogenous parameter with two realizations (*low* or *high*) in each time period, two endogenous parameters each with two realizations (also *low* or *high*), and a time horizon consisting of two time periods (i.e., 3 stages). In generating the full set of scenarios, it follows that set $\mathcal{R}_N$ corresponds to the endogenous scenario tree in Figure 5, and set $\mathcal{R}_X$ corresponds to the exogenous scenario tree in Figure 2. By equation (5), the composite scenario tree resulting from all possible combinations of realizations of these parameters will consist of the scenarios given by $\mathcal{R} = \mathcal{R}_N \times \mathcal{R}_X = \left\{\left(\hat{\theta}^L_1, \hat{\theta}^L_2, \hat{\xi}^L_1, \hat{\xi}^L_2\right), \dots, \left(\hat{\theta}^H_1, \hat{\theta}^H_2, \hat{\xi}^H_1, \hat{\xi}^H_2\right)\right\}$. The number of scenarios in this composite tree is given by Equation (6), which yields $S = 4 \cdot 4 = 16$. Thus, $\mathcal{S} = \{1, 2, \dots, 16\}$. Figure 6a clarifies the mathematical procedure for generating the composite scenario tree by providing the graphical analogue: we simply copy the exogenous scenario tree (Figure 2) for each possible combination of realizations of the endogenous parameters. This gives rise to multiple "subtrees" (four in this case). Viewed another way, we have essentially replaced each scenario in the endogenous scenario tree (Figure 5) with an exogenous subtree. Here we use the alternative form of the exogenous tree, Figure 2b, so that we can easily apply scenario decomposition later in this paper. With the full set of scenarios in place, we then link these subtrees by adding first-period and endogenous non-anticipativity constraints (shown by solid and dotted green lines, respectively) which enforce equality between indistinguishable nodes. This process is partially illustrated in Figure 6b for the links between subtrees 1 and 2 only. By adding the remaining links between the subtrees, we end up with the complete composite scenario tree shown in Figure 6c. (We provide the reasoning behind our choice of these particular non-anticipativity constraints in section 5.) It is clear how quickly these problems can grow, as the composite tree is significantly more complex than either Figure 2 or Figure 5 alone. Note that in the figures, as before, we are only considering one exogenous parameter and one endogenous parameter for each of the two sources, so we have dropped the j and h subscripts, respectively.

Notice that within each subtree, all scenarios have the same possible endogenous realizations, and these endogenous realizations are the only distinguishing characteristic between each of the subtrees. Thus, if the uncertainty in the endogenous parameters is not resolved by the end of the time horizon, all of the subtrees will be exactly identical (since all of the conditional, dotted green lines will have become solid lines, enforcing non-anticipativity between the corresponding nodes).

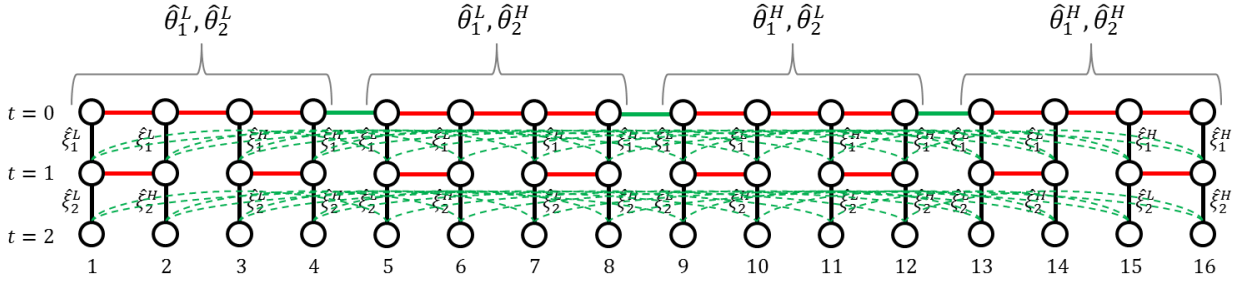
It is also interesting to note that if we assume expected values for each of the endogenous parameters, thereby neglecting the endogenous uncertainty, we recover the original exogenous tree (Figure 2). If we instead assume expected values for each of the exogenous parameters, thereby neglecting the exogenous uncertainty, we recover the original endogenous tree (Figure 5).



(a) **Step 1:** Copy the exogenous scenario tree for each possible combination of realizations of the endogenous parameters.



(b) **Step 2:** Link these 'subtrees' by adding first-period and endogenous non-anticipativity constraints (shown by solid and dotted green lines, respectively, between subtrees 1 and 2).



(c) Complete composite scenario tree.

Figure 6. Procedure for generating a 'composite' scenario tree. This tree captures all possible combinations of realizations for both the endogenous and exogenous uncertain parameters.

The concept of subtrees will be used extensively in the definitions of parameters and sets later in this paper, so we define parameter $Sub(s)$ to return the subtree number of each scenario in the composite tree. This number is calculated as the ceiling of the ratio of the scenario index s and the number of scenarios in each subtree, S_X:

$$Sub(s) \coloneqq \lceil s/S_X \rceil \qquad \forall\, s \in \mathcal{S} \tag{7}$$

Note that S_X, defined in Equation (2), is the number of scenarios in each subtree since each subtree is simply an exogenous tree. For Figure 6c, scenarios 1–4 are in subtree 1, scenarios 5–8 are in subtree 2, scenarios 9–12 are in subtree 3, and scenarios 13–16 are in subtree 4. Accordingly, Equation (7) returns $Sub(2) = \lceil 2/4 \rceil = 1$, $Sub(6) = \lceil 6/4 \rceil = 2$, $Sub(10) = \lceil 10/4 \rceil = 3$, and $Sub(14) = \lceil 14/4 \rceil = 4$.

3.4 Realization Probabilities

Following directly from the theory presented in the previous sections, we now briefly discuss realization probabilities. For each realization r of exogenous parameter $\xi_{j,t}$, there is a corresponding probability $\hat{v}_{j,t}^{r}$ of this value occurring. The realization values are defined in set $\Xi_{j,t} \coloneqq \{\hat{\xi}_{j,t}^{r}\colon\ r = 1, 2, \dots, R_{j,t}\}$, and we now define the set of probabilities $\Upsilon_{j,t} \coloneqq \{\hat{v}_{j,t}^{r}\colon\ r = 1, 2, \dots, R_{j,t}\}$. Note that this set is indexed in the same order as the realization values. For instance, if we consider the first realization, $r = 1$, the realization value is given by set $\Xi_{j,t}$ as $\hat{\xi}_{j,t}^{1}$, and the corresponding probability is given by set $\Upsilon_{j,t}$ as $\hat{v}_{j,t}^{1}$. Also, note that since these realizations represent all possible outcomes for parameter $\xi_{j,t}$ from a discretized probability distribution, the probabilities must sum to 1; i.e., $\sum_{r=1}^{R_{j,t}} \hat{v}_{j,t}^{r} = 1\ \ \forall\, j \in \mathcal{J},\ \ t \in \mathcal{T}$.

Each realization m of endogenous parameter $\theta_{i,h}$ also has a corresponding probability $\hat{\omega}_{i,h}^{m}$ that it will occur. The realization values are defined in set $\Theta_{i,h} \coloneqq \{\hat{\theta}_{i,h}^{m}\colon\ m = 1, 2, \dots, M_{i,h}\}$, and we define the set of probabilities $\Omega_{i,h} \coloneqq \{\hat{\omega}_{i,h}^{m}\colon\ m = 1, 2, \dots, M_{i,h}\}$. As is the case for the exogenous parameters, this set is indexed in the same order as the realization values. Thus, for $m = 1$, we have the realization value $\hat{\theta}_{i,h}^{1}$ from set $\Theta_{i,h}$, and the corresponding probability $\hat{\omega}_{i,h}^{1}$ from set $\Omega_{i,h}$. Again, these realizations represent all possible outcomes for parameter $\theta_{i,h}$ from a discretized probability distribution, so the probabilities must sum to 1; i.e., $\sum_{m=1}^{M_{i,h}} \hat{\omega}_{i,h}^{m} = 1\ \ \forall\, i \in \mathcal{I},\ \ h \in \mathcal{H}_i$.

Recall that set $\mathcal{R}_X$ (Equation (1)), set $\mathcal{R}_N$ (Equation (3)), and set $\mathcal{R}$ (Equation (5)) give the realization values for each scenario $s \in \mathcal{S}_X$, $s \in \mathcal{S}_N$, and $s \in \mathcal{S}$, respectively. Since each of these realizations has a corresponding probability, we can find the set of realization probabilities for each scenario by simply substituting the realization values in each expression with their corresponding probabilities. Specifically, for the case where the uncertainty is purely exogenous, the set of realization probabilities for each scenario is defined by,

$$\begin{aligned}\mathcal{P}_X \coloneqq& \times_{t\in\mathcal{T}} \left(\times_{j\in\mathcal{J}} \Upsilon_{j,t}\right)\\ =& \left\{\left(\hat{v}_{1,1}^{1}, \dots, \hat{v}_{J,T}^{1}\right), \dots, \left(\hat{v}_{1,1}^{R_{1,1}}, \dots, \hat{v}_{J,T}^{R_{J,T}}\right)\right\}\\ =& \left\{\left(v_{1,1}^{s}, \dots, v_{J,T}^{s}\right)\colon\ \forall\, s \in \mathcal{S}_X\right\}\end{aligned} \tag{8}$$

where $v_{j,t}^{s}$ refers to the realization probability of exogenous parameter $\xi_{j,t}$ in scenario s. In the case where the uncertainty is purely endogenous, the set of realization probabilities for each scenario is defined by,

$$\begin{aligned}\mathcal{P}_N \coloneqq& \times_{i\in\mathcal{I}} \left(\times_{h\in\mathcal{H}_i} \Omega_{i,h}\right)\\ =& \left\{\left(\hat{\omega}_{1,1}^{1}, \dots, \hat{\omega}_{I,H_I}^{1}\right), \dots, \left(\hat{\omega}_{1,1}^{M_{1,1}}, \dots, \hat{\omega}_{I,H_I}^{M_{I,H_I}}\right)\right\}\\ =& \left\{\left(\omega_{1,1}^{s}, \dots, \omega_{I,H_I}^{s}\right)\colon\ \forall\, s \in \mathcal{S}_N\right\}\end{aligned} \tag{9}$$

where $\omega_{i,h}^{s}$ refers to the realization probability of endogenous parameter $\theta_{i,h}$ in scenario s. And for the case of primary interest, where there are both endogenous and exogenous uncertainties, the set of realization probabilities for each scenario is defined by,

$$\begin{aligned}\mathcal{P} &:= \mathcal{P}_N \times \mathcal{P}_X \\ &= \left\{\left(\widehat{\omega}_{1,1}^{1}, \dots, \widehat{\omega}_{I,H_I}^{1}, \hat{v}_{1,1}^{1}, \dots, \hat{v}_{J,T}^{1}\right), \dots, \left(\widehat{\omega}_{1,1}^{M_{1,1}}, \dots, \widehat{\omega}_{I,H_I}^{M_{I,H_I}}, \hat{v}_{1,1}^{R_{1,1}}, \dots, \hat{v}_{J,T}^{R_{J,T}}\right)\right\} \\ &= \left\{\left(\omega_{1,1}^{s}, \dots, \omega_{I,H_I}^{s}, v_{1,1}^{s}, \dots, v_{J,T}^{s}\right):\ \forall\, s \in \mathcal{S}\right\}\end{aligned} \tag{10}$$

The probability of each scenario is given by p^s, and is equal to the product of all of the realization probabilities in scenario s:

$$p^s = \left(\prod_{i\in\mathcal{I}}\prod_{h\in\mathcal{H}_i} \omega_{i,h}^{s}\right)\cdot\left(\prod_{t\in\mathcal{T}}\prod_{j\in\mathcal{J}} v_{j,t}^{s}\right) \quad \forall\, s \in \mathcal{S} \tag{11}$$

Since the elements sum to 1 in each set of realization probabilities ($\Upsilon_{j,t}$ and $\Omega_{i,h}$), and we are simply taking the product of each possible combination of all of these elements, the sum over all of these products must also be 1 (see the supplementary material for the simple proof). In other words, the total probability over all scenarios must sum to 1: $\sum_{s\in\mathcal{S}} p^s = 1$.

4 Models

A simple MILP formulation for a deterministic multi-period planning problem is given in model (MPD). Variable vectors y_t represent investment and operation decisions that are made at the beginning of each time period t (e.g., whether or not to drill a particular oilfield, the processing capacity of a new offshore oil facility, etc.), and variable vectors x_t represent operation decisions that typically follow these investment decisions (e.g., the oil flow rate from a field to a newly-installed facility). Variable vectors w_t are commonly referred to as state variables and represent calculated quantities associated with each time period, such as intermediate flow rates and economic values like total operating cost. Vectors y_t, x_t, and w_t may each have integer and continuous components.

(MPD)

$$\min_{y,x} \phi_D = \sum_{t\in\mathcal{T}} \left({}^{y}c_t y_t + {}^{x}c_t x_t + {}^{w}c_t w_t\right) \tag{12}$$

$$\text{s.t.} \quad \sum_{\tau=1}^{t} \left({}^{y}A_{\tau,t} y_\tau + {}^{x}A_{\tau,t} x_\tau + {}^{w}A_{\tau,t} w_\tau\right) \le a_t \qquad \forall\, t \in \mathcal{T} \tag{13}$$

$$y_t \in \mathcal{Y}_t,\ x_t \in \mathcal{X}_t,\ w_t \in \mathcal{W}_t \qquad \forall\, t \in \mathcal{T} \tag{14}$$

The objective function, Equation (12), minimizes the total cost associated with decisions y_t and x_t, and state variables w_t. For convenience, we adopt the notation of Goel and Grossmann (2006) and specify the corresponding cost coefficients through row vectors ${}^{y}c_t$, ${}^{x}c_t$, and ${}^{w}c_t$, respectively. Equation (13) represents constraints that govern the decisions in each time period $t \in \mathcal{T}$, as well as constraints that link decisions across time periods. This equation also includes equality constraints such as those that assign

values to w_t. The constraint coefficients for variables y_t, x_t, and w_t are given by matrices ${}^{y}A_{\tau,t}$, ${}^{x}A_{\tau,t}$, and ${}^{w}A_{\tau,t}$, respectively, and the right-hand side is given by column vectors a_t. Bounds and integrality restrictions on the variables are specified by mixed-integer sets $\mathcal{Y}_t$, $\mathcal{X}_t$, and $\mathcal{W}_t$ in Equation (14).

In the following subsections, we will show how this model is transformed into a multistage stochastic programming problem in the case of: (1) exogenous uncertainty, (2) endogenous uncertainty, and (3) both endogenous and exogenous uncertainties. These stochastic programming models (largely inspired by the work of Goel and Grossmann (2006)) will be presented in deterministic-equivalent form using the non-anticipativity approach. For additional background, we refer the reader to Rockafellar and Wets (1991), Ruszczyński (1997), and Birge and Louveaux (2011).

4.1 MSSP Formulation for Exogenous Uncertainty

The multistage stochastic programming formulation of model (MPD) in the case of exogenous uncertainties is given in model (MSSP_X). Notice that variables y_t, x_t, and w_t have been indexed for each scenario $s \in \mathcal{S}_X$ to indicate the respective decisions and calculated quantities in each scenario. As we are now modeling under a multistage stochastic programming framework, the decision-making process is structured as shown in Figure 1. Specifically, variables y_t^s refer to here-and-now decisions, and variables x_t^s refer to recourse decisions. Recall that decisions y_t^s are implemented at the beginning of each time period t of scenario s. At some point after these decisions are made, but during t, the uncertainty in exogenous parameter $\xi_{j,t}$ is resolved. Recourse decisions x_t^s are then made as corrective action at the end of the period in response to this new information. Based on the values of y_t^s and x_t^s, state variables w_t^s are calculated.

(MSSP_X)

$$\min_{y,x} \phi_X = \sum_{s \in \mathcal{S}_X} p^s \sum_{t \in \mathcal{T}} ({}^{y}c_t^s y_t^s + {}^{x}c_t^s x_t^s + {}^{w}c_t^s w_t^s) \tag{15}$$

$$\text{s.t.} \quad \sum_{\tau=1}^{t} \left({}^{y}A_{\tau,t}^s y_\tau^s + {}^{x}A_{\tau,t}^s x_\tau^s + {}^{w}A_{\tau,t}^s w_\tau^s\right) \leq a_t^s \qquad \forall\, t \in \mathcal{T},\ s \in \mathcal{S}_X \tag{16}$$

$$y_1^s = y_1^{s'} \qquad \forall\, (s, s') \in \mathcal{SP}_F \tag{17}$$

$$x_t^s = x_t^{s'} \qquad \forall\, (t, s, s') \in \mathcal{SP}_X \tag{18}$$

$$y_{t+1}^s = y_{t+1}^{s'} \qquad \forall\, (t, s, s') \in \mathcal{SP}_X \tag{19}$$

$$y_t^s \in \mathcal{Y}_t^s,\ x_t^s \in \mathcal{X}_t^s,\ w_t^s \in \mathcal{W}_t^s \qquad \forall\, t \in \mathcal{T},\ s \in \mathcal{S}_X \tag{20}$$

Notice that only fairly simple changes are required to convert the deterministic model (MPD) to the multistage stochastic programming model (MSSP_X). In particular, the objective function, Equation (15), now minimizes the total *expected* cost by taking the weighted sum of the costs in each scenario based on the probability of each scenario, p^s. The cost coefficients have been indexed for all $s \in \mathcal{S}_X$ to allow for the possibility of different cost realizations in each scenario. Additionally, the constraints governing the decisions in each time period, represented by Equation (16), are simply applied for each $s \in \mathcal{S}_X$. Note that like the cost coefficients, the constraint coefficients and right-hand side have also been indexed for s to

allow for different realizations in each scenario. In other words, exogenous parameters $\xi_{j,t}$ may enter the model through the objective function and/or the constraints (via the constraint coefficients and/or the right-hand side).

The most significant difference between the models is the introduction of non-anticipativity constraints, given by Equations (17)-(19). Each scenario in model ($MSSP_X$) represents a different instance of the deterministic planning problem with different realizations for the uncertain parameters, and the non-anticipativity constraints link these scenarios together, as shown in Figure 2b.

Equation (17) enforces non-anticipativity between all scenarios at the beginning of the first time period. As previously stated, this is due to the fact that all scenarios are indistinguishable at this time, and we must make the same here-and-now decisions (first-stage decisions) in all scenarios. For the remainder of this paper, we will simply refer to these constraints as *first-period NACs*. Note that when discussing indistinguishability, we refer specifically to the indistinguishability *between two scenarios* s and s'. Accordingly, we will consider pairs of indistinguishable scenarios (s, s') in each time period t for which we must enforce non-anticipativity. We will also define sets of these tuples in order to simplify the notation in our models. For the first-period NACs, the corresponding *first-period scenario pairs* are elements of set $\mathcal{SP}_F$, given by:

$$\mathcal{SP}_F \coloneqq \mathcal{A} \tag{21}$$

where $\mathcal{A}$ is the set of scenario pairs for which s and s' are adjacent. This will be discussed in greater detail in section 5.1 (see Equation (47)).

Equations (18) and (19) represent non-anticipativity constraints for all remaining stages. In particular, if scenarios s and s' are indistinguishable in time period t in terms of the resolution of exogenous uncertainty, we must make the same recourse decisions at the end of this period (enforced by Equation (18)), as well as the same here-and-now decisions at the beginning of the next time period, $t+1$ (enforced by Equation (19)). We will refer to these constraints as *exogenous NACs*. The corresponding set of *exogenous scenario pairs* is given by set $\mathcal{SP}_X$ and is defined as:

$$\mathcal{SP}_X \coloneqq \left\{(t, s, s'):\ \ t \in \mathcal{T}\backslash\{T\},\ \ (s, s') \in \mathcal{A},\ \ Sub(s) = Sub(s'),\ \ Q_t^{s,s'} = True\right\} \tag{22}$$

where $Sub(s) = Sub(s')$ ensures that s and s' are in the same subtree, and $Q_t^{s,s'}$ is a Boolean parameter that indicates whether or not these scenarios are indistinguishable in time period t. This will be discussed in section 5.2 (see Equations (48) and (49)).

Notice that at the beginning of the final time period, the NACs for the here-and-now decisions correspond to Equation (19) with $t = T-1$. Also, specifically in the exogenous case, NACs never apply for the recourse decisions at the end of the final time period (final-stage decisions); this is because the leaf nodes must refer to independent states or else we would have duplicate scenarios in the tree (see Figure 2). It follows, then, that we can entirely exclude time period $t = T$ from the definition of set $\mathcal{SP}_X$, as indicated in Equation (22).

We also note that we never express NACs for state variables w_t^s, in *any* time period, since these variables are calculated based on the values of decision variables y_t^s and x_t^s. In other words, non-anticipativity for w_t^s is implicitly enforced by Equations (17)-(19).

Similar to the deterministic formulation, bounds and integrality restrictions on the variables are specified by the mixed-integer sets in Equation (20).

4.2 MSSP Formulation for Endogenous Uncertainty

The multistage stochastic programming formulation of model (MPD) in the case of endogenous uncertainties is given in model ($MSSP_N$). This model has been adapted from Goel and Grossmann (2006) and is presented in hybrid mixed-integer linear disjunctive form.

Previously, we used vector y_t^s to represent all here-and-now decisions in each time period t of scenario s. In the case of endogenous uncertainties, however, this approach does not provide us with particularly detailed information. As can be seen in Figure 3, it is not immediately obvious which decisions are associated with a source $i \in \mathcal{I}$. This is a very important modeling consideration, since such decisions uniquely determine whether or not the uncertainty in parameter $\theta_{i,h}$ can be resolved in scenario s. Accordingly, we define vector $b_{i,t}^s$ to identify those binary decisions that are strictly associated with a particular source i (e.g., to drill an oilfield of uncertain size and initial deliverability). To keep the notation simple, we will continue to use y_t^s to represent all other here-and-now decisions.

It is often the case that the uncertainty in some (or all) endogenous parameters cannot be resolved within the first few time periods of the planning horizon. For instance, in an oilfield development planning problem, we may assume that any oilfield must be in production for a certain number of years before the size of the field can be established. Before that amount of time has passed, the sizes of all fields are uncertain, and any scenarios that differ only in the possible realizations of field sizes must be indistinguishable. Thus, for these initial time periods, the corresponding conditional NACs can be expressed as equality constraints (Colvin and Maravelias, 2010; Gupta and Grossmann, 2014a). To model this, we denote the number of initial 'equality' periods as $T_E^{i'}$, and partition the set of time periods $\mathcal{T}$ into the set of these initial periods $\mathcal{T}_E^{i'} \coloneqq \{t:\ t = 1, \dots, T_E^{i'}\}$ and the set of remaining 'conditional' time periods $\mathcal{T}_C^{i'} \coloneqq \{t:\ t = T_E^{i'} + 1, \dots, T\}$, where $T_E^{i'} < T$, for all $i' \in \mathcal{I}$. We use index i' in these definitions so as not to conflict with index i of $b_{i,t}^s$. Note that if $T_E^{i'} = 0$, the corresponding sets reduce to $\mathcal{T}_E^{i'} \coloneqq \emptyset$ and $\mathcal{T}_C^{i'} \coloneqq \mathcal{T}$. Further note that these parameters and subsets are defined for each $i' \in \mathcal{I}$ since the number of initial periods may not be the same for all sources of endogenous uncertainty.

($MSSP_N$)

$$\min_{b,y,x} \phi_N = \sum_{s \in \mathcal{S}_N} p^s \sum_{t \in \mathcal{T}} \left({}^{y}c_t^s y_t^s + {}^{x}c_t^s x_t^s + {}^{w}c_t^s w_t^s + \sum_{i \in \mathcal{I}} {}^{b}c_{i,t}^s b_{i,t}^s \right) \tag{23}$$

$$\text{s.t.} \quad \sum_{\tau=1}^{t} \left({}^{y}A_{\tau,t}^s y_\tau^s + {}^{x}A_{\tau,t}^s x_\tau^s + {}^{w}A_{\tau,t}^s w_\tau^s + \sum_{i \in \mathcal{I}} {}^{b}A_{i,\tau,t}^s b_{i,\tau}^s \right) \le a_t^s \qquad \forall\, t \in \mathcal{T},\ s \in \mathcal{S}_N \tag{24}$$

$$b_{i,1}^s = b_{i,1}^{s'} \qquad \forall\, (s, s') \in \mathcal{SP}_F,\ i \in \mathcal{I} \tag{25}$$

$$y_1^s = y_1^{s'} \qquad \forall\, (s,s') \in \mathcal{SP}_F \tag{17}$$

$$x_t^s = x_t^{s'} \qquad \forall\, (t,s,s') \in \mathcal{SP}_N,\ t \in \mathcal{T}_E^{i'},\ \{i'\} = \widehat{\mathcal{D}}^{s,s'} \tag{26}$$

$$b_{i,t+1}^s = b_{i,t+1}^{s'} \qquad \forall\, (t,s,s') \in \mathcal{SP}_N,\ t \in \mathcal{T}_E^{i'},\ \{i'\} = \widehat{\mathcal{D}}^{s,s'},\ i \in \mathcal{I} \tag{27}$$

$$y_{t+1}^s = y_{t+1}^{s'} \qquad \forall\, (t,s,s') \in \mathcal{SP}_N,\ t \in \mathcal{T}_E^{i'},\ \{i'\} = \widehat{\mathcal{D}}^{s,s'} \tag{28}$$

$$\begin{bmatrix} Z_t^{s,s'} \\ x_t^s = x_t^{s'} \\ b_{i,t+1}^s = b_{i,t+1}^{s'}\ \forall\, i \in \mathcal{I},\ t < T \\ y_{t+1}^s = y_{t+1}^{s'} \qquad\qquad t < T \end{bmatrix} \vee \left[\neg Z_t^{s,s'}\right] \qquad \forall\, (t,s,s') \in \mathcal{SP}_N,\ t \in \mathcal{T}_C^{i'},\ \{i'\} = \widehat{\mathcal{D}}^{s,s'} \tag{29}$$

$$Z_t^{s,s'} \Leftrightarrow F\left(b_{i',1}^s, b_{i',2}^s, \dots, b_{i',t}^s\right) \qquad \forall\, (t,s,s') \in \mathcal{SP}_N,\ t \in \mathcal{T}_C^{i'},\ \{i'\} = \widehat{\mathcal{D}}^{s,s'} \tag{30}$$

$$b_{i,t}^s \in \{0,1\},\ y_t^s \in \mathcal{Y}_t^s,\ x_t^s \in \mathcal{X}_t^s,\ w_t^s \in \mathcal{W}_t^s \qquad \forall\, i \in \mathcal{I},\ t \in \mathcal{T},\ s \in \mathcal{S}_N \tag{31}$$

$$Z_t^{s,s'} \in \{True, False\} \qquad \forall\, (t,s,s') \in \mathcal{SP}_N,\ t \in \mathcal{T}_C^{i'},\ \{i'\} = \widehat{\mathcal{D}}^{s,s'} \tag{32}$$

The objective function, Equation (23), is very similar to that of model ($MSSP_X$). Notice that the only differences from Equation (15) are the following: we have now introduced decision variables $b_{i,t}^s$ and the corresponding row vector of cost coefficients, ${}^b c_{i,t}^s$ (which requires a summation over all sources $i \in \mathcal{I}$), and the set of scenarios is now given by $\mathcal{S}_N$. Likewise, Equation (24) only differs from Equation (16) by the same changes, except the corresponding coefficient matrix is ${}^b A_{i,\tau,t}^s$. Endogenous parameters $\theta_{i,h}$ may enter the model through the objective function and/or the constraints, as was the case for exogenous parameters $\xi_{j,t}$ in model ($MSSP_X$). First-period NACs still apply, and accordingly, we express them for our here-and-now decisions in Equation (25) and (from model ($MSSP_X$)) Equation (17).

Each scenario pair of time period $t \in \mathcal{T}_E^{i'}$ in set $\mathcal{SP}_N$ represents two scenarios s and s' that are indistinguishable at that time in terms of the resolution of endogenous uncertainty. This set of *endogenous scenario pairs*, $\mathcal{SP}_N$, is given by:

$$\begin{aligned} \mathcal{SP}_N^{i',h,l} \coloneqq \Big\{ (t,s,s') \colon\ & t \in \mathcal{T},\ s,s' \in \left({}^N\mathcal{G}_{i',h}^l \cap \mathcal{U}_t^{i',h}\right), \\ & s' = \min_{\hat{s}'} \left(\hat{s}' \in \left({}^N\mathcal{G}_{i',h}^l \cap \mathcal{U}_t^{i',h}\right),\ \hat{s}' > s\right), \\ & s < \max_{\hat{s}} \left(\hat{s} \in \left({}^N\mathcal{G}_{i',h}^l \cap \mathcal{U}_t^{i',h}\right)\right),\ \{(i',h)\} = \mathcal{D}^{s,s'} \Big\}\ \forall\, l \in \mathcal{L}_{i',h},\ i' \in \mathcal{I}, \\ & h \in \mathcal{H}_{i'} \end{aligned} \tag{33}$$

$$\mathcal{SP}_N \coloneqq \bigcup_{i' \in \mathcal{I}} \left(\bigcup_{h \in \mathcal{H}_{i'}} \left(\bigcup_{l \in \mathcal{L}_{i',h}} \mathcal{SP}_N^{i',h,l} \right) \right) \tag{34}$$

where, in Equation (33), we first determine the set of scenario pairs in each time period t corresponding to endogenous parameter $\theta_{i',h}$ for all $i' \in \mathcal{I}$ and $h \in \mathcal{H}_{i'}$. We then take the union of all of these sets in Equation (34).

Given the complexity of Equation (33), before continuing, we briefly describe the primary aspects of this expression. Sets ${}^{N}\mathcal{G}^{l}_{i',h}$, indexed by $l \in \mathcal{L}_{i',h}$, represent endogenous scenario groups corresponding to $\theta_{i',h}$. For each endogenous parameter $\theta_{i',h}$, set $\mathcal{U}_t^{i',h}$ provides a sufficient subset of scenarios that are available for pairing from that parameter's respective groups in time period t. These sets, $\mathcal{U}_t^{i',h}$, are defined in a sequential manner in which we successively eliminate scenarios based on the pairs that have already been formed. We then obtain a sufficient subset of each group specific to time period t via sets ${}^{N}\mathcal{G}^{l}_{i',h} \cap \mathcal{U}_t^{i',h}$, which we refer to as *reduced* endogenous scenario groups. We pair off consecutive scenarios in each of these reduced groups, where set $\mathcal{D}^{s,s'}$ indicates the specific parameter $\theta_{i',h}$ for which s and s' differ in possible realizations. This will be discussed in detail in section 5.3.

Accordingly, for each of the scenario pairs of time period $t \in \mathcal{T}_E^{i'}$ in set $\mathcal{SP}_N$, we enforce non-anticipativity between the respective scenarios s and s' as shown in Equations (26), (27), and (28), exactly as we would in the exogenous case (see Equations (18) and (19) for comparison). Notice that the only differences here are that we are considering different scenario pairs, and the set of time periods is source dependent. The particular source i' of set $\mathcal{T}_E^{i'}$ is determined by set $\widehat{\mathcal{D}}^{s,s'}$; specifically, this set indicates the source in which scenarios s and s' differ in the possible realization of some endogenous parameter (see Equation (66) in section 5.3). Note that we will refer to these equality constraints as *fixed endogenous NACs*.

Each scenario pair of time period $t \in \mathcal{T}_C^{i'}$ in set $\mathcal{SP}_N$ represents two scenarios s and s' that *may* be indistinguishable at that time (where the particular source i' of set $\mathcal{T}_C^{i'}$ is determined by set $\widehat{\mathcal{D}}^{s,s'}$). Recall that in the exogenous case, we know in advance whether two scenarios will differ in parameter realizations in time period t. For endogenous parameters, however, the timing of realizations depends on decisions $b_{i,t}^{s}$, so we can no longer use simple equality constraints to apply non-anticipativity. Instead, we *conditionally* enforce non-anticipativity between these scenarios (see Figure 5), as shown by the disjunctive constraints in Equation (29). Boolean variable $Z_t^{s,s'}$ indicates whether s and s' are indistinguishable by the end of time period t, and if so, the value is $True$ and the NACs are enforced. If they are distinguishable, the value is $False$ and the constraints are ignored. We will refer to these conditional constraints as *conditional endogenous NACs*. Note that since we make here-and-now decisions for the *next* time period $(t+1)$ based on indistinguishability information revealed up until the current time t, NACs for decisions $b_{i,t+1}^{s}$ and y_{t+1}^{s} must be restricted to $t < T$; this is, of course, because we cannot make new here-and-now decisions at the end of the time horizon. This restriction is implicit in the exogenous model because non-anticipativity does not apply at the end of the final time period.

Using a big-M reformulation (Trespalacios and Grossmann, 2014), we can rewrite the disjunctive constraints (29) as inequality constraints (35)-(37), where UB denotes the upper bound of the respective variable.[7]

[7] We substitute variable upper bounds for big-M parameters; however, for a specific problem instance, tighter bounds can often be established.

$$-x_t^{UB}\left(1-z_t^{s,s'}\right) \le x_t^s - x_t^{s'} \le x_t^{UB}\left(1-z_t^{s,s'}\right) \quad \forall\,(t,s,s') \in \mathcal{SP}_N,\ t \in \mathcal{T}_C^{i'},\ \{i'\} = \widehat{\mathcal{D}}^{s,s'} \tag{35}$$

$$-\left(1-z_t^{s,s'}\right) \le b_{i,t+1}^s - b_{i,t+1}^{s'} \le \left(1-z_t^{s,s'}\right) \quad \forall\,(t,s,s') \in \mathcal{SP}_N,\ t \in \mathcal{T}_C^{i'},\ t < T,\ \{i'\} = \widehat{\mathcal{D}}^{s,s'},\ i \in \mathcal{I} \tag{36}$$

$$-y_{t+1}^{UB}\left(1-z_t^{s,s'}\right) \le y_{t+1}^s - y_{t+1}^{s'} \le y_{t+1}^{UB}\left(1-z_t^{s,s'}\right) \quad \forall\,(t,s,s') \in \mathcal{SP}_N,\ t \in \mathcal{T}_C^{i'},\ t < T,\ \{i'\} = \widehat{\mathcal{D}}^{s,s'} \tag{37}$$

$$z_t^{s,s'} \in \{0,1\} \quad \forall\,(t,s,s') \in \mathcal{SP}_N,\ t \in \mathcal{T}_C^{i'},\ \{i'\} = \widehat{\mathcal{D}}^{s,s'} \tag{38}$$

Note that $z_t^{s,s'}$, defined in Equation (38), is the binary equivalent of Boolean variable $Z_t^{s,s'}$; i.e., $\left(z_t^{s,s'} = 1\right) \Leftrightarrow \left(Z_t^{s,s'} = True\right)$ and $\left(z_t^{s,s'} = 0\right) \Leftrightarrow \left(Z_t^{s,s'} = False\right)$. The fixed endogenous NACs can be viewed as a special case of these constraints with $z_t^{s,s'} = 1$. Specifically, in the initial time periods, each double-sided inequality collapses into a single equality constraint, thereby providing us with a smaller, tighter formulation (Colvin and Maravelias, 2010).

The value of $Z_t^{s,s'}$ is determined by an uncertainty-resolution rule, as stated in general form in Equation (30) (Gupta and Grossmann, 2014a). This rule uses the values of all decisions $b_{i',t}^s$ up to and including the current time period to determine whether uncertainty has been resolved in a given source $i' \in \mathcal{I}$.

The simplest uncertainty-resolution rule, given in Equation (39) and referred to as immediate resolution, assumes that the uncertainty in endogenous parameter $\theta_{i',h}$ is resolved immediately in scenario s after an investment is made in source $i' \in \mathcal{I}$ for the first time (i.e., $b_{i',t}^s = 1$ and $b_{i',\tau}^s = 0\ \forall\,\tau < t,\ t,\tau \in \mathcal{T}$) (Goel and Grossmann, 2006).[8]

$$Z_t^{s,s'} \Leftrightarrow \left[\bigwedge_{\tau=1}^{t}\left(\neg b_{i',\tau}^s\right)\right] \quad \forall\,(t,s,s') \in \mathcal{SP}_N,\ t \in \mathcal{T}_C^{i'},\ \{i'\} = \widehat{\mathcal{D}}^{s,s'} \tag{39}$$

Prior to this time t, $Z_\tau^{s,s'} = True$ and the scenarios s and s' that differ in the possible realization of $\theta_{i',h}$ are indistinguishable. After the investment at time t, $Z_t^{s,s'} = False$ and the scenarios are distinguishable; thus, non-anticipativity constraints no longer apply between s and s'. Note that we previously used this concept in the discussion of Figure 4. Logic constraints (39) can be rewritten as linear integer inequality constraints (40) and (41) by applying the reformulations described in Williams (2013) and Raman and Grossmann (1991).

$$1 - \sum_{\tau=1}^{t} b_{i',\tau}^s \le z_t^{s,s'} \quad \forall\,(t,s,s') \in \mathcal{SP}_N,\ t \in \mathcal{T}_C^{i'},\ \{i'\} = \widehat{\mathcal{D}}^{s,s'} \tag{40}$$

[8] We treat binary variable $b_{i',\tau}^s$ as Boolean to keep the notation simple.

$$z_t^{s,s'} \le 1 - b_{i',\tau}^{s} \quad \forall\, (t,s,s') \in \mathcal{SP}_N,\ t,\tau \in \mathcal{T}_C^{i'},\ \tau \le t,\ \{i'\} = \widehat{\mathcal{D}}^{s,s'} \tag{41}$$

Making this replacement and substituting disjunctive constraints (29) for constraints (35)-(38) transforms model ($MSSP_N$) into an MILP. Bounds and integrality restrictions on the variables are given in Equations (31), (32), and (38).

4.3 MSSP Formulation for Endogenous and Exogenous Uncertainties

In the case of both endogenous and exogenous uncertainties, the multistage stochastic programming formulation of model (MPD) is given by model (MSSP). This model is also adapted from the work of Goel and Grossmann (2006) and will be our primary focus for the remainder of this paper. Just as the scenario tree for this class of problems is represented by a composite scenario tree (Figure 6c), the corresponding model can also be seen as a composite of the exogenous model ($MSSP_X$) and the endogenous model ($MSSP_N$). In particular, all of their respective NACs and logic constraints appear together in (MSSP).

(MSSP)

$$\min_{b,y,x} \phi = \sum_{s\in\mathcal{S}} p^s \sum_{t\in\mathcal{T}} \left({}^{y}c_t^s y_t^s + {}^{x}c_t^s x_t^s + {}^{w}c_t^s w_t^s + \sum_{i\in\mathcal{I}} {}^{b}c_{i,t}^s b_{i,t}^s \right) \tag{42}$$

$$\text{s.t.} \quad \sum_{\tau=1}^{t} \left({}^{y}A_{\tau,t}^s y_\tau^s + {}^{x}A_{\tau,t}^s x_\tau^s + {}^{w}A_{\tau,t}^s w_\tau^s + \sum_{i\in\mathcal{I}} {}^{b}A_{i,\tau,t}^s b_{i,\tau}^s \right) \le a_t^s \qquad \forall\, t \in \mathcal{T},\ s \in \mathcal{S} \tag{43}$$

$$b_{i,1}^s = b_{i,1}^{s'} \qquad \forall\, (s,s') \in \mathcal{SP}_F,\ i \in \mathcal{I} \tag{25}$$

$$y_1^s = y_1^{s'} \qquad \forall\, (s,s') \in \mathcal{SP}_F \tag{17}$$

$$x_t^s = x_t^{s'} \qquad \forall\, (t,s,s') \in \mathcal{SP}_X \tag{18}$$

$$b_{i,t+1}^s = b_{i,t+1}^{s'} \qquad \forall\, (t,s,s') \in \mathcal{SP}_X,\ i \in \mathcal{I} \tag{44}$$

$$y_{t+1}^s = y_{t+1}^{s'} \qquad \forall\, (t,s,s') \in \mathcal{SP}_X \tag{19}$$

$$x_t^s = x_t^{s'} \qquad \forall\, (t,s,s') \in \mathcal{SP}_N,\ t \in \mathcal{T}_E^{i'},\ \{i'\} = \widehat{\mathcal{D}}^{s,s'} \tag{26}$$

$$b_{i,t+1}^s = b_{i,t+1}^{s'} \qquad \forall\, (t,s,s') \in \mathcal{SP}_N,\ t \in \mathcal{T}_E^{i'},\ \{i'\} = \widehat{\mathcal{D}}^{s,s'},\ i \in \mathcal{I} \tag{27}$$

$$y_{t+1}^s = y_{t+1}^{s'} \qquad \forall\, (t,s,s') \in \mathcal{SP}_N,\ t \in \mathcal{T}_E^{i'},\ \{i'\} = \widehat{\mathcal{D}}^{s,s'} \tag{28}$$

$$\begin{bmatrix} Z_t^{s,s'} \\ x_t^s = x_t^{s'} \\ b_{i,t+1}^s = b_{i,t+1}^{s'} \ \forall\, i \in \mathcal{I},\ t < T \\ y_{t+1}^s = y_{t+1}^{s'} \qquad\qquad t < T \end{bmatrix} \vee \left[\neg Z_t^{s,s'} \right] \qquad \forall\, (t,s,s') \in \mathcal{SP}_N,\ t \in \mathcal{T}_C^{i'},\ \{i'\} = \widehat{\mathcal{D}}^{s,s'} \tag{29}$$

$$Z_t^{s,s'} \Leftrightarrow F\left(b_{i',1}^s, b_{i',2}^s, \dots, b_{i',t}^s\right) \qquad \forall\, (t,s,s') \in \mathcal{SP}_N,\ t \in \mathcal{T}_C^{i'},\ \{i'\} = \widehat{\mathcal{D}}^{s,s'} \tag{30}$$

$$b_{i,t}^s \in \{0,1\},\ y_t^s \in \mathcal{Y}_t^s,\ x_t^s \in \mathcal{X}_t^s,\ w_t^s \in \mathcal{W}_t^s \qquad \forall\, i \in \mathcal{I},\ t \in \mathcal{T},\ s \in \mathcal{S} \tag{45}$$

$$Z_t^{s,s'} \in \{True, False\} \qquad \forall\, (t,s,s') \in \mathcal{SP}_N,\ t \in \mathcal{T}_C^{i'},\ \{i'\} = \widehat{\mathcal{D}}^{s,s'} \tag{32}$$

Like model ($MSSP_N$), this formulation represents a *hybrid* mixed-integer linear *disjunctive* programming problem due to the presence of the conditional endogenous constraints (29) and logic constraints (30). The disjunctive constraints can be replaced by constraints (35)-(38), and if immediate resolution of uncertainty is assumed, the logic constraints can be replaced by inequalities (40) and (41). These steps transform model (MSSP) into standard mixed-integer linear form.

Notice that the objective function (42) and constraints (43) have only been updated from their respective counterparts in model ($MSSP_N$) to reflect the fact that the set of scenarios is now given by $\mathcal{S}$. This is also true for the bounds and integrality restrictions specified in Equation (45). The only new addition to the model is Equation (44), which gives the exogenous non-anticipativity constraints for $b_{i,t}^s$, as these variables were not originally defined in the exogenous model ($MSSP_X$). Exogenous parameters $\xi_{j,t}$ and endogenous parameters $\theta_{i,h}$ may enter the model through the objective function and/or the constraints.

It is interesting to note that if we assume expected values for the endogenous parameters, then we have $\mathcal{SP}_N = \emptyset,\ \mathcal{S} = \mathcal{S}_X$, and model (MSSP) reduces to the exogenous model ($MSSP_X$). Similarly, if we assume expected values for the exogenous parameters, then we have $\mathcal{SP}_X = \emptyset,\ \mathcal{S} = \mathcal{S}_N$, and model (MSSP) reduces to the endogenous model ($MSSP_N$).

The multistage stochastic programming problem (MSSP) may appear to be simply a larger version of the purely-exogenous and purely-endogenous formulations previously discussed; however, there is a great deal of complexity contained in scenario-pair sets $\mathcal{SP}_F$, $\mathcal{SP}_X$, and $\mathcal{SP}_N$. Specifically, we must carefully account for the presence of both types of uncertainty when defining these sets. Notice that in the exogenous formulation ($MSSP_X$), NACs are applied in time period t for all pairs of scenarios that are indistinguishable in terms of the resolution of exogenous uncertainty. In the endogenous formulation ($MSSP_N$), NACs are applied in time period t for all pairs of scenarios that must be indistinguishable, and conditionally applied for those that *may* be indistinguishable, in terms of the resolution of endogenous uncertainty. As can be seen in the composite scenario tree (Figure 6c), this is not the case when endogenous and exogenous uncertainties are both present. First-period NACs link all scenarios at the beginning of the first time period, as always, but exogenous NACs now link scenarios in time period t that are indistinguishable in terms of the resolution of exogenous uncertainty *and* are identical in all possible realizations of the endogenous parameters. In other words, exogenous NACs are applied between scenarios *within each subtree*. Endogenous NACs now link scenarios in time period t that differ in the possible realization of one endogenous parameter *and* are identical in all realizations of the exogenous parameters. Thus, endogenous NACs are applied between scenarios *in different subtrees*. This is an interesting modeling challenge and will be discussed in detail in the next section.

5 Scenario Pairs and Reduction Properties

In defining each of the scenario-pair sets $\mathcal{SP}_F$, $\mathcal{SP}_X$, and $\mathcal{SP}_N$, we begin with a naïve approach in which we specify only $s, s' \in \mathcal{S}$ and $s \neq s'$, along with the additional indistinguishability conditions specific to either first-period NACs, exogenous NACs, or endogenous NACs; i.e.,

$$\{(s,s'):\ s,s' \in \mathcal{S},\ s \neq s',\ \text{conditions for indistinguishability}\} \quad \textbf{(46)}$$

As stated in the following property, however, the condition $s \neq s'$ is not particularly restrictive and leaves us with many redundant scenario pairs.

Property 1. Scenario pairs (s,s') and (s',s) refer to the same pair. Thus, it is sufficient to enforce non-anticipativity constraints for only pairs (s,s') where $s < s'$ (Goel and Grossmann, 2006).

Proof. See Appendix section A.1. A brief, qualitative proof can also be found in Goel and Grossmann (2006).

This simple symmetry argument eliminates half of the scenario pairs generated by Equation (46). We place special emphasis on *reduction properties* such as Property 1 since NACs are expressed for each pair of scenarios, and the number of pairs can be extremely large in instances with a large number of scenarios. In the following sections, we will define additional reduction properties to exclude all redundant pairs from each of our set definitions. We begin with scenario-pair set $\mathcal{SP}_F$ for first-period NACs.

5.1 First-period Scenario Pairs

As was the case for purely-exogenous and purely-endogenous uncertainties, at the beginning of the first time period, no decisions have been implemented and no uncertainties have been resolved. Hence, all scenarios are indistinguishable at that time and we must make the same here-and-now decisions in all scenarios. To define the set of scenario pairs required for these non-anticipativity constraints, we rely on the following property.

Property 2a. For first-period NACs, it is sufficient to consider only scenario pairs (s,s') for which s and s' are adjacent.

Proof. See Appendix section A.2.

Accordingly, we define the set of all pairs of adjacent scenarios, $\mathcal{A}$:

$$\mathcal{A} \coloneqq \{(s,s'):\ s,s' \in \mathcal{S},\ s' = s+1,\ s < S\} \quad \textbf{(47)}$$

Note that the condition $s < s'$ is implicit in this definition since we are only considering consecutive scenarios in the 'forward' direction. The set of all scenario pairs for first-period NACs is then simply equal to set $\mathcal{A}$, as we define in Equation (21).

$$\mathcal{SP}_F \coloneqq \mathcal{A} \quad \textbf{(21)}$$

This is the minimum number of scenario pairs, as stated in the following proposition.

Proposition 1. First-period scenario-pair set $\mathcal{SP}_F$ contains the minimum number of scenario pairs.

Proof. See Appendix section A.3.

Note that the respective scenario pairs in set $\mathcal{SP}_F$ are *non-unique*. In other words, different formulations with the same cardinality are possible; e.g., we may instead choose to link the first scenario to every other scenario. Such alternative pairing approaches have been shown to perform better in Lagrangean decomposition (Oliveira et al., 2013); however, for convenience, we limit our current discussion to the consecutive-pairing approach.

5.2 Exogenous Scenario Pairs

Excluding the beginning of the first time period, scenarios s and s' are indistinguishable in time period t if they are identical in the realizations of all exogenous parameters up to this point *and* they have all of the same possible realizations for the endogenous parameters. These scenario pairs are required for exogenous non-anticipativity constraints.

Rather than explicitly checking that each pair of scenarios has the same possible endogenous realizations, it is clear from Figure 6 that due to the manner in which we generate the scenario set, this condition is implicitly satisfied for any s and s' in the same subtree. Recall that this is because each subtree represents an exogenous scenario tree, and by definition, all scenarios in this tree must have the same endogenous realizations (see section 3.3). Furthermore, different subtrees have different possible endogenous realizations, so s and s' can *only* be in the same subtree. This argument also allows us to invoke the following reduction property.

Property 2b. For exogenous NACs, it is sufficient to consider only scenario pairs (s, s') for which s and s' are adjacent.

Proof. See Appendix section A.4.

Hence, we state that adjacent scenarios s and s' will be indistinguishable in the first time period if they have the same realizations for all exogenous parameters in this period and they are in the same subtree. Let Boolean parameter $Q_t^{s,s'}$ represent the indistinguishability of adjacent scenarios s and s' in time period t, where $Q_t^{s,s'} = True$ if the scenarios are indistinguishable, and $Q_t^{s,s'} = False$ otherwise. Then,

$$Q_1^{s,s'} = \begin{cases} True, & \text{if } \xi_{j,1}^{s} = \xi_{j,1}^{s'} \ \forall j \in \mathcal{J} \\ False, & \text{otherwise} \end{cases} \qquad \forall (s, s') \in \mathcal{A},\ Sub(s) = Sub(s') \tag{48}$$

where the subtree condition $Sub(s) = Sub(s')$ relies on the definition provided by Equation (7).

For all subsequent time periods, the scenarios are indistinguishable if they were indistinguishable in the previous time period, they have the same exogenous realizations in the current time period, and they are in the same subtree:

$$Q_t^{s,s'} = \begin{cases} True, & \text{if } Q_{t-1}^{s,s'} = True \text{ and } \xi_{j,t}^{s} = \xi_{j,t}^{s'} \ \forall j \in \mathcal{J} \\ False, & \text{otherwise} \end{cases} \qquad \begin{array}{l} t = 2, 3, \dots, T,\ \forall (s, s') \in \mathcal{A}, \\ Sub(s) = Sub(s') \end{array} \tag{49}$$

As an example, scenarios 1 and 2 in Figure 6c have the same realizations for the exogenous parameter in the first time period; i.e., $\xi_1^1 = \xi_1^2$. Thus, these scenarios are indistinguishable at the end of this period and $Q_1^{1,2} = True$. They have different realizations in the second time period (i.e., $\xi_2^1 \neq \xi_2^2$), so the scenarios

are distinguishable at that time and $Q_2^{1,2} = False$. Since the leaf nodes in each subtree refer to independent states, it is in fact the case that all adjacent scenarios in the same subtree will be distinguishable by the end of the final time period; i.e., $Q_T^{s,s'} = False$. Thus, it is unnecessary to evaluate Equation (49) for $t = T$. We also note that because $Q_t^{s,s'}$ is the same for all subtrees, it is most efficient to calculate $Q_t^{s,s'}$ only for the first subtree and then to duplicate the results for all others.

The set of all scenario pairs (s, s') in each time period t, such that s and s' are indistinguishable in terms of the resolution of exogenous uncertainty and are identical in all possible realizations of the endogenous parameters, can then be defined as:

$$\mathcal{SP}_X \coloneqq \left\{(t, s, s'):\ t \in \mathcal{T}\backslash\{T\},\ (s, s') \in \mathcal{A},\ Sub(s) = Sub(s'),\ Q_t^{s,s'} = True\right\} \quad \textbf{(22)}$$

Equation (22) is also applicable in purely-exogenous problems since, in that case, $Sub(s) = 1$ for all $s \in \mathcal{S}$. This is the reasoning behind the use of set $\mathcal{SP}_X$ in model (MSSP$_X$).

We now define exogenous scenario 'groups' in each time period $t \in \mathcal{T}\backslash\{T\}$, where each group is a set of indistinguishable scenarios that refer to the same state.[9] Specifically, each group is the direct result of splitting a single node into indistinguishable copies for each scenario, as discussed in the proof of Property 2b. For example, at the end of the first time period in Figure 6c, scenarios 1 and 2 refer to the same unique state and can be grouped together. Scenarios 3 and 4 refer to another unique state and can be placed into a second group. Continuing this process, we end up with 8 different groups of two scenarios each, as shown in Figure 7. Blue groups consist of scenarios with a *low* realization for exogenous parameter ξ_1, and green groups consist of scenarios with a *high* realization for that parameter. We typically do not define scenario groups for the final time period, since (as previously mentioned) adjacent leaf nodes in the same subtree are unique; in other words, there would be S groups of one scenario each in time period T (e.g., 16 groups of one scenario each in Figure 7).

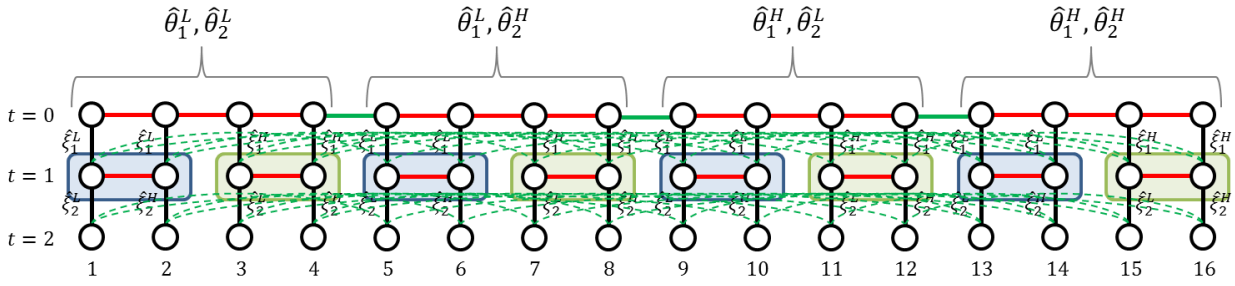


Figure 7. Exogenous scenario groups.

To generalize this grouping process, we first define parameter $G_X(t, s)$ to return the group number of each scenario $s \in \mathcal{S}$ in time period $t \in \mathcal{T}\backslash\{T\}$. Next, we assign the first scenario in each of these time periods to group 1 by specifying $G_X(t, 1) = 1\ \ \forall\, t \in \mathcal{T}\backslash\{T\}$. We then use Equation (50) to assign group numbers to all other scenarios:

[9] We will frequently refer to exogenous scenario groups *in* time period t. This will be understood to mean the *end* of time period t, after all realizations in that period have occurred.

$$G_X(t,s) = G_X(t,s-1) + \sum_{(t,s-1,s)\notin \mathcal{SP}_X} [1] \qquad \forall\, t \in \mathcal{T}\backslash\{T\},\ s = 2,3,\dots,S \tag{50}$$

The general idea behind this equation is that the group number of scenario s will be equal to the group number of the previous scenario $s-1$, given by $G_X(t,s-1)$, as long as these two adjacent scenarios are indistinguishable based on the definition of set $\mathcal{SP}_X$. If they are not indistinguishable in this sense (i.e., $(t,s-1,s) \notin \mathcal{SP}_X$), then the scenarios have different realizations for some of the uncertain parameters and scenario s belongs in a new group; thus, the group number is incremented by 1. For instance, at $t=1$ in Figure 7, scenario 1 is first assigned a group number of 1. Scenario 2 is indistinguishable from scenario 1, so must also be assigned to group 1. Scenario 3, however, is distinguishable from scenario 2 since $\xi_1^2 \neq \xi_1^3$, and $(1,2,3) \notin \mathcal{SP}_X$. Thus, we increment the group number and assign scenario 3 to group 2. We repeat this process for all remaining scenarios in this time period.

We index these groups by defining the set of indices $\mathcal{K}_t$,

$$\mathcal{K}_t \coloneqq \{k:\ \ k = 1,2,\dots,G_X(t,S)\}\ \ \forall\, t \in \mathcal{T}\backslash\{T\} \tag{51}$$

where $G_X(t,S)$ gives the total number of groups in time period t (since it is the group number for the final scenario in time period t). In Figure 7, this corresponds to $G_X(1,16) = 8$; therefore, $\mathcal{K}_1 \coloneqq \{1,2,\dots,8\}$.

We then use the group numbers to define the set of scenarios for each group:

$${}^{X}\mathcal{G}_t^k \coloneqq \{s:\ \ s \in \mathcal{S},\ \ G_X(t,s) = k\}\ \ \forall\, k \in \mathcal{K}_t,\ \ t \in \mathcal{T}\backslash\{T\} \tag{52}$$

For example, at $t=1$ in Figure 7, scenario 1 has a group number of 1 (i.e., $G_X(1,1)=1$) and scenario 2 has a group number of 1 (i.e., $G_X(1,2)=1$). Accordingly, exogenous scenario group 1 in the first time period is given by ${}^{X}\mathcal{G}_1^1 = \{1,2\}$. We similarly define ${}^{X}\mathcal{G}_1^2 = \{3,4\}$, ${}^{X}\mathcal{G}_1^3 = \{5,6\}, \dots,$ ${}^{X}\mathcal{G}_1^8 = \{15,16\}$.

The exogenous scenario-group definitions allow us to state the following proposition.

Proposition 2. Exogenous scenario-pair set $\mathcal{SP}_X$ contains the minimum number of scenario pairs.

Proof. See Appendix section A.5.

Like set $\mathcal{SP}_F$, the respective scenario pairs in set $\mathcal{SP}_X$ are non-unique. The concept of exogenous scenario groups will be used again in the next section to derive endogenous scenario-pair set $\mathcal{SP}_N$. As will be shown, the definition of this set is quite complex.

5.3 Endogenous Scenario Pairs

Excluding the beginning of the first time period, scenarios s and s' are indistinguishable in the initial time periods $t \in \mathcal{T}_E^{i'}$ if they differ in the possible realizations of one or more endogenous parameters *and* they are identical in the realizations of all exogenous parameters that have been realized up until that time.

Recall that these scenarios must be indistinguishable here because the endogenous uncertainty cannot yet be resolved. These scenario pairs are used to generate fixed endogenous NACs.

For the remaining time periods $t \in \mathcal{T}_C^{i'}$, the uncertainty *can* be resolved at some point, but we do not know when this will occur (or if it will at all). Scenarios s and s' will be indistinguishable until this unknown point in time. Thus, we state that under the same conditions given for $t \in \mathcal{T}_E^{i'}$, scenarios s and s' in $t \in \mathcal{T}_C^{i'}$ *may* be indistinguishable. These scenario pairs are used to generate conditional endogenous NACs. Notice that due to the conditional nature of these constraints, set $\mathcal{SP}_N$ may contain several scenario pairs that we do not need. This is in sharp contrast to the exogenous scenario-pair set $\mathcal{SP}_X$, where every scenario pair is required because all of the NACs are fixed.

Before we derive the endogenous scenario-pair set $\mathcal{SP}_N$, it is possible to significantly strengthen the indistinguishability requirements. We begin with the following reduction property.

Property 3. For endogenous NACs, it is sufficient to consider only scenario pairs (s, s') for which s and s' differ in the possible realization of a single endogenous parameter and are identical in the realizations of all exogenous parameters in all time periods.

Proof. See Goel and Grossmann (2006).

Example. Due to the complexity of Property 3, and its importance to this work, we provide an illustrative example of the proof. Consider Figure 8, where we have isolated scenarios 1, 5, and 13 from Figure 6c for an arbitrary time period $t = \tau$. Scenario 1 differs from scenario 5 in the possible realization of endogenous parameter θ_2. Scenario 5 differs from scenario 13 in the possible realization of endogenous parameter θ_1. Scenario 1 differs from scenario 13, however, in the possible realizations of *both* θ_1 *and* θ_2. The three scenarios have identical realizations for exogenous parameter ξ_t in all time periods.

Disregarding the possibility of initial 'equality' periods, we will have three conditional links between the scenarios, as shown at the top of Figure 8: $(1, 5)$ and $(5, 13)$, as shown in green, and $(1, 13)$, as shown in orange. There are four possible outcomes depending upon the way the uncertainty is resolved. In Case 1, both endogenous parameters have been realized by the end of this time period. Accordingly, the scenarios are distinguishable and non-anticipativity does not apply. In Case 2, only the value of θ_1 has been realized and NACs are enforced between scenarios 1 and 5. If we consider only variables y_τ^s, the corresponding NAC is $y_\tau^1 = y_\tau^5$. Similarly, in Case 3, only the value of θ_2 has been realized and NACs are enforced between scenarios 5 and 13; e.g., $y_\tau^5 = y_\tau^{13}$. When neither of the parameters has been realized in Case 4, all three conditional links are enforced: $y_\tau^1 = y_\tau^5$, $y_\tau^5 = y_\tau^{13}$, and $y_\tau^1 = y_\tau^{13}$.

Notice that Case 4 is the *only* case in which we apply non-anticipativity for scenario pair $(1, 13)$, and it applies only at the same time as the non-anticipativity for pairs $(1, 5)$ and $(5, 13)$. Thus, by a simple transitivity argument, it is clear that constraint $y_\tau^1 = y_\tau^{13}$ is implied by constraints $y_\tau^1 = y_\tau^5$ and $y_\tau^5 = y_\tau^{13}$. Accordingly, scenario pair $(1, 13)$ can be excluded entirely. This leaves us with two pairs that differ only in the possible realization of a single endogenous parameter.

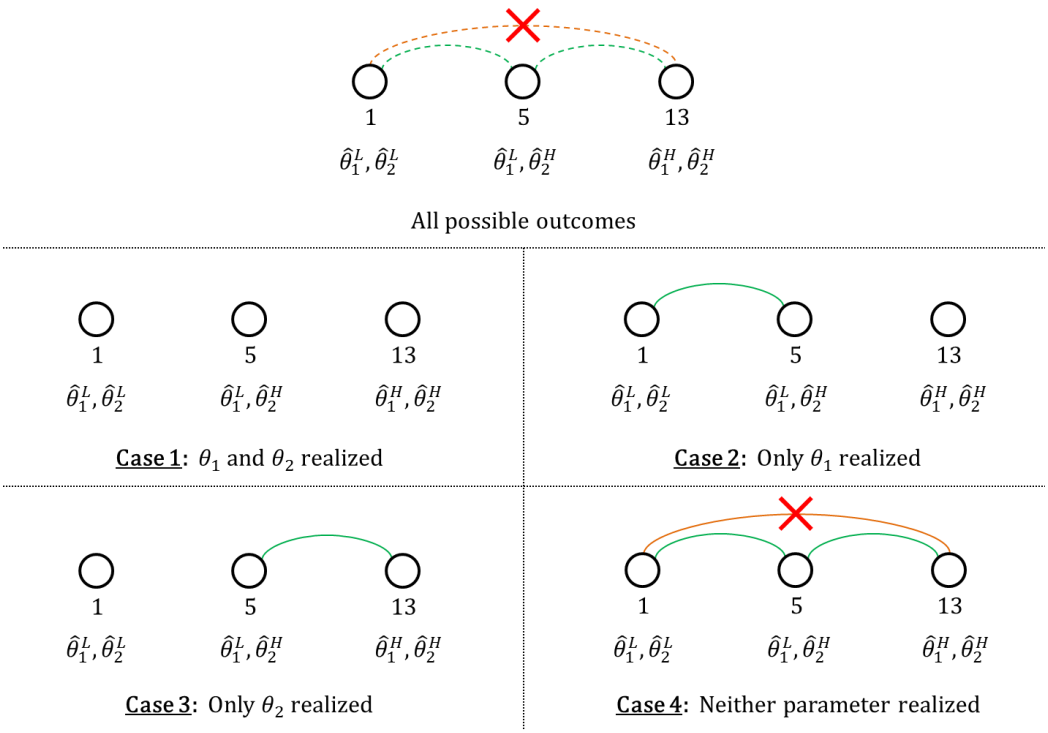


Figure 8. Illustration of Property 3.

Recall that scenarios 1, 5, and 13 have identical realizations for the exogenous parameter in all time periods. We now extend this example to include scenario 2, which has a different exogenous realization in the second period. Here, we illustrate the second part of Property 3; specifically, that the corresponding scenario pairs (s, s') consist of scenarios s and s' that are identical in the realizations of all exogenous parameters *in all time periods*, rather than just identical in the exogenous realizations that have been *revealed up until the current time period*. This is shown in Figure 9.

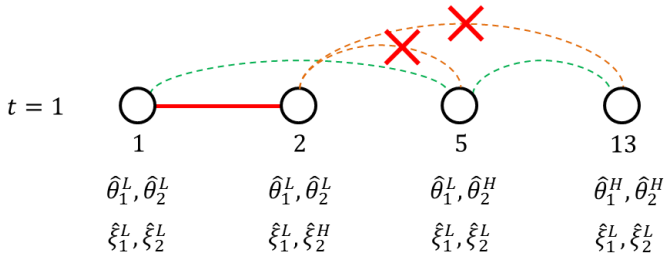


Figure 9. Property 3 as applied to endogenous and exogenous uncertainties.

Figure 9 includes additional scenario pairs $(2, 5)$ and $(2, 13)$, as shown in orange. Notice, however, that scenario 2 is identical to scenario 1 aside from the different exogenous realization in the second time period (i.e., $\xi_2^1 \neq \xi_2^2$). This means that in the first period, non-anticipativity for pairs $(2, 5)$ and $(1, 5)$ will

apply at the same time, non-anticipativity for pairs $(2, 13)$ and $(1, 13)$ will apply at the same time,[10] and we have the exogenous non-anticipativity constraint $y_\tau^1 = y_\tau^2$ between scenarios 1 and 2 (shown in red).

Thus, by transitivity, pair $(2, 5)$ can be eliminated since constraints $y_\tau^1 = y_\tau^2$ and $y_\tau^1 = y_\tau^5$ imply $y_\tau^2 = y_\tau^5$. Likewise, pair $(2, 13)$ can also be eliminated since constraints $y_\tau^1 = y_\tau^2$, $y_\tau^1 = y_\tau^5$, and $y_\tau^5 = y_\tau^{13}$ imply $y_\tau^2 = y_\tau^{13}$. Notice that we have eliminated any endogenous scenario pairs (s, s') for which s and s' are not identical in the realizations of all exogenous parameters in all time periods.

As proved rigorously in Goel and Grossmann (2006), Property 3 always holds, provided that the set of scenarios consists of all possible combinations of realizations of the endogenous parameters. The authors also showed that this property extends to the general case where there are multiple parameters associated with each source of endogenous uncertainty (as we have considered throughout this paper with the use of parameter $\theta_{i,h}$).

By Property 3, we may now state that scenarios s and s' are indistinguishable in time period t if they differ in the possible realization of *exactly one* endogenous parameter and they are identical in the realizations of all exogenous parameters *in all time periods*. We first address the latter part of this statement.

Recall from the previous section that scenarios in the same subtree must have the same endogenous realizations. Thus, for s and s' to differ in *any* endogenous realizations, they must belong to different subtrees. Furthermore, for these scenarios to have exactly the same exogenous realizations, they must have the same *position* in both subtrees; for example, the first scenario in both, as in scenarios 1 and 5. This is because we generate the composite tree by starting with a single exogenous tree that has no duplicate scenarios. It follows that when we duplicate the exogenous tree for each possible combination of realizations of the endogenous parameters, scenarios in the same position in different subtrees have originated from the same scenario. Therefore, they must have all of the same exogenous realizations. Because there were no duplicates in the original exogenous tree, these are the only scenarios for which this holds.

We define parameter $Pos(s)$ to return the position of scenario s from the viewpoint of its respective subtree; in other words, the index that scenario s would have if it were in subtree 1:

$$Pos(s) = s - S_X(Sub(s) - 1) \quad \forall\, s \in \mathcal{S} \tag{53}$$

Equation (53) calculates this normalized scenario index for s by subtracting off the appropriate number of scenarios according to the subtree that s belongs to. Recall that S_X is just the number of scenarios in each subtree. As a simple example, consider scenarios 1, 5, 9, and 13 in Figure 6c. Since these scenarios refer to the first scenario in each subtree, respectively, Equation (53) gives $Pos(1) = 1 - 4(1 - 1) = 1$, $Pos(5) = 5 - 4(2 - 1) = 1$, $Pos(9) = 9 - 4(3 - 1) = 1$, and $Pos(13) = 13 - 4(4 - 1) = 1$.

[10] Recall from the discussion surrounding Figure 8 that scenario pair $(1, 13)$ is implied by pairs $(1, 5)$ and $(5, 13)$.

Thus, to indicate that s and s' are identical in all exogenous realizations, but differ in at least one possible endogenous realization, it is sufficient to state $Pos(s) = Pos(s')$, with $s < s'$. Note that this implies that the two scenarios are in different subtrees, so it is unnecessary to specify $Sub(s) \neq Sub(s')$.

We now address the first part of Property 3; namely, that scenarios s and s' differ in the possible realization of *exactly one* endogenous parameter. To do so, we define sets $\mathcal{D}^{s,s'}$, composed of pairs of indices (i', h), to indicate the endogenous parameters $\theta_{i',h}$ for which scenarios s and s' differ in possible realizations:[11]

$$\mathcal{D}^{s,s'} := \left\{(i', h):\ i' \in \mathcal{I},\ h \in \mathcal{H}_{i'},\ \theta^{s}_{i',h} \neq \theta^{s'}_{i',h}\right\}\ \forall\, s, s' \in \mathcal{S},\ s < s',\ Pos(s) = Pos(s') \qquad \textbf{(54)}$$

Property 3 then requires that $\left|\mathcal{D}^{s,s'}\right| = 1$ for all endogenous scenario pairs. In other words, the corresponding set of pairs for all time periods is given by:

$$\mathcal{SP}_{N^3} := \left\{(t, s, s'):\ t \in \mathcal{T},\ s, s' \in \mathcal{S},\ s < s',\ Pos(s) = Pos(s'),\ \left|\mathcal{D}^{s,s'}\right| = 1\right\} \qquad \textbf{(55)}$$

Note that the same pairs are present in each period.

As pointed out by Gupta and Grossmann (2011), however, when we consider 3 or more possible realizations for any of the endogenous parameters, there are additional redundant scenario pairs that are not removed by this property. This is illustrated in Figure 10. Here we consider a group of three scenarios ($\hat{s}$, $\hat{s}'$, and $\hat{s}''$), in an arbitrary time period $t = \tau$, that all differ in the possible realization of a single endogenous parameter $\theta_{\hat{\imath},\hat{h}}$. These scenarios will be distinguishable in time period τ if parameter $\theta_{\hat{\imath},\hat{h}}$ has been realized (Case 1), or indistinguishable if the parameter has not yet been realized (Case 2).

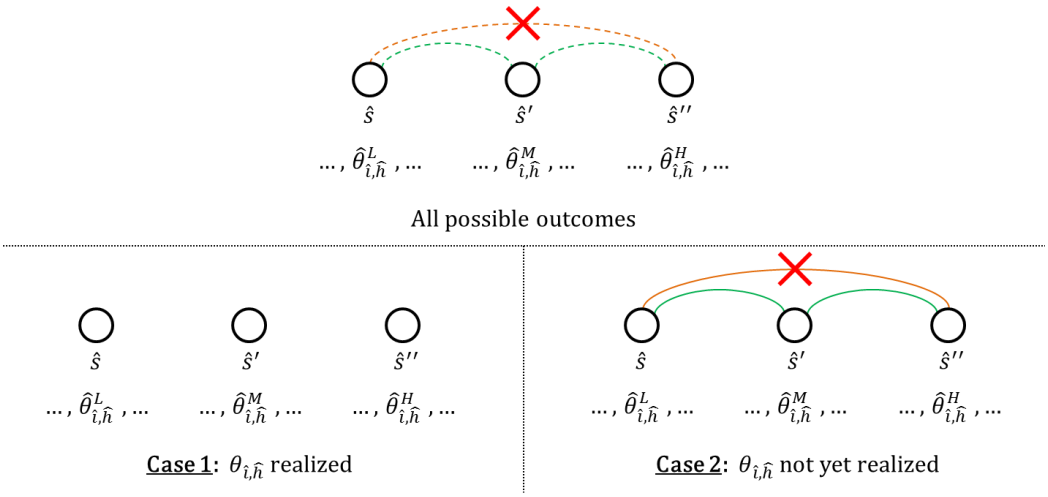


Figure 10. Property 3 fails to eliminate all redundant scenario pairs when there are 3 or more possible realizations for any of the endogenous parameters.

[11] Recall that we index the sources with i' so as not to conflict with index i of $b^{s}_{i,t}$ in models (MSSP$_N$) and (MSSP).

Using Property 3, we generate three scenario pairs: $(\hat{s}, \hat{s}')$ and $(\hat{s}', \hat{s}'')$, as shown in green, and $(\hat{s}, \hat{s}'')$, as shown in orange. Since the corresponding NACs all apply at the same time or are all ignored at the same time, it is clear that scenario pair $(\hat{s}, \hat{s}'')$ is redundant and can be eliminated. This follows directly from the simple transitivity arguments previously used in the example of Property 3. Because $\left|\mathcal{D}^{\hat{s},\hat{s}''}\right| = 1$ and yet $(\hat{s}, \hat{s}'')$ is redundant, it is also clear that we must rely on an alternative approach to exclude such scenario pairs.

A simple remedy for this, as proposed by Gupta and Grossmann (2011), is to first generate all 'groups' of scenarios like that shown in Figure 10, and then link consecutive scenarios in each of these groups.[12] Each group is the set of all scenarios that differ only in the possible realization of a single endogenous parameter h of source i' (i.e., $\theta_{i',h}$). As previously noted, these scenarios will be indistinguishable as long as this parameter is unrealized.

Thus, for each $i' \in \mathcal{I}$ and $h \in \mathcal{H}_{i'}$, we define parameter $G_N(i', h, s)$ to identify the index of the group that scenario $s \in \mathcal{S}$ belongs to. We refer to this as the group number, represented here by index l, and propose the following algorithm to assign group numbers to all scenarios. Notice that unlike the exogenous case, an algorithm is required here because the groups consist of nonconsecutively-indexed scenarios.

Endogenous Scenario-Group Algorithm

Step 1: Initialize the group numbers to zero for all scenarios; i.e., $G_N(i', h, s) \coloneqq 0 \quad \forall\, i' \in \mathcal{I},\ h \in \mathcal{H}_{i'},\ s \in \mathcal{S}$. Also, define a group counter, GroupCount, to keep track of the current group number in each iteration.

Step 2: For each endogenous parameter, define all groups of scenarios that differ in the possible realization of *only* this parameter. This is done as follows.

For each $i' \in \mathcal{I}$ and $h \in \mathcal{H}_{i'}$:

Step 2a: Reset the group counter (i.e., $\text{GroupCount} \coloneqq 0$).

Step 2b: Fix s to the next available scenario in $\mathcal{S}$ (i.e., s has not already been assigned to a group, so $G_N(i', h, s) = 0$), and then search for all other scenarios from which s differs in the possible realization of only $\theta_{i',h}$. Such scenarios must be in the same group as s.

Specifically, for $s = 1, 2, \ldots, S$, where $G_N(i', h, s) = 0$:

i) Increment the group counter (i.e., $\text{GroupCount} \coloneqq \text{GroupCount} + 1$).

ii) Set the group number of scenario s to the current group number:

$$G_N(i', h, s) \coloneqq \text{GroupCount} \tag{56}$$

[12] The scope of Gupta and Grossmann (2011) is limited to purely-endogenous MSSP problems with no initial 'equality' time periods and only one parameter associated with each source of uncertainty.

iii) Search for scenarios $s' \in \mathcal{S}$ that differ from s in the possible realization of the same endogenous parameter; i.e., $\mathcal{D}^{s,s'} = \{(i', h)\}$. For each s' that satisfies this condition, set the group number of that scenario to the same group number as scenario s; i.e.,

$$G_N(i', h, s') \coloneqq G_N(i', h, s) \quad \forall\, s' \in \mathcal{S},\ s' > s,\ Pos(s') = Pos(s),\ \mathcal{D}^{s,s'} = \{(i', h)\} \tag{57}$$

For instance, in Figure 10, assume that scenario $\hat{s}$ is in group $\hat{l}$ corresponding to endogenous parameter $\theta_{\hat{\imath},\hat{h}}$. Also, assume that $s = \hat{s}$. In this step of the algorithm, we would first identify $\hat{s}'$ as belonging to the same group as $\hat{s}$, and then the same for $\hat{s}''$, since $\mathcal{D}^{\hat{s},\hat{s}'} = \mathcal{D}^{\hat{s},\hat{s}''} = \{(\hat{\imath}, \hat{h})\}$. Thus, we would have $G_N(\hat{\imath}, \hat{h}, \hat{s}) = G_N(\hat{\imath}, \hat{h}, \hat{s}') = G_N(\hat{\imath}, \hat{h}, \hat{s}'') = \hat{l}$.

Notice that, aside from the fact that the s index is fixed, the restrictions on the scenarios in Equation (57) are the same as those for Property 3 (see the definition of set $\mathcal{SP}_{N^3}$ in Equation (55)), with the condition $\mathcal{D}^{s,s'} = \{(i', h)\}$ in place of $\left|\mathcal{D}^{s,s'}\right| = 1$. This condition is inspired by Gupta and Grossmann (2011) and implies that $\left|\mathcal{D}^{s,s'}\right| = 1$.

Step 2c: Use the final group number to define the set of indices for all groups corresponding to $\theta_{i',h}$:

$$\mathcal{L}_{i',h,} \coloneqq \{l:\ l = 1, 2, \dots, \text{GroupCount}\} \tag{58}$$

We index the endogenous scenario groups as $l \in \mathcal{L}_{i',h}$.

For each endogenous parameter $\theta_{i',h}$, the group-number parameter gives the particular index $\hat{l}$ for each $s \in \mathcal{S}$ (i.e., $G_N(i', h, s) = \hat{l}$). We can use this information to define the set of scenarios for each group:

$${}^{N}\mathcal{G}^{l}_{i',h} \coloneqq \{s:\ s \in \mathcal{S},\ G_N(i', h, s) = l\} \quad \forall\, l \in \mathcal{L}_{i',h},\ i' \in \mathcal{I},\ h \in \mathcal{H}_{i'} \tag{59}$$

Note that it is unnecessary to define these groups for every time period, since endogenous realizations are not explicitly associated with any particular time t.

We now define the corresponding set of endogenous scenario pairs, which is at least as restrictive as $\mathcal{SP}_{N^3}$ (we will prove this momentarily), by first linking consecutive scenarios in each group. This is handled separately for each group, as shown in Equation (60).

$$\begin{aligned}\mathcal{SP}^{i',h,l}_{N^4} \coloneqq \Big\{(t, s, s'):\ t \in \mathcal{T},\ s, s' \in {}^{N}\mathcal{G}^{l}_{i',h},\ s' = \min_{\hat{s}'}\left(\hat{s}' \in {}^{N}\mathcal{G}^{l}_{i',h},\ \hat{s}' > s\right),\\ s < \max_{\hat{s}}\left(\hat{s} \in {}^{N}\mathcal{G}^{l}_{i',h}\right),\ \{(i', h)\} = \mathcal{D}^{s,s'}\Big\} \quad \forall\, l \in \mathcal{L}_{i',h},\ i' \in \mathcal{I},\ h \in \mathcal{H}_{i'}\end{aligned} \tag{60}$$

Although seemingly complex, the expression $s' = \min_{\hat{s}'}\left(\hat{s}' \in {}^{N}\mathcal{G}^{l}_{i',h},\ \hat{s}' > s\right)$ simply ensures that scenario s' is the next-highest-indexed scenario immediately following scenario s. The expression $s < \max_{\hat{s}}\left(\hat{s} \in {}^{N}\mathcal{G}^{l}_{i',h}\right)$ simply excludes the highest-indexed scenario from the group, since there is no scenario following it with which to form a pair. This is *the same concept* used to define the set of adjacent scenarios, $\mathcal{A}$, previously defined in Equation (47) and used in our consecutive pairing approach for first-

period and exogenous scenario pairs. The endogenous case is merely a more general formulation that allows us to pair off consecutive scenarios that are nonconsecutively indexed. To prove that this is the case, consider the following: if we replace ${}^{N}\mathcal{G}^{l}_{i',h}$ with $\mathcal{S}$ in the two expressions under discussion, we arrive at $s' = s + 1$ from the first and $s < S$ from the second. These are the same two conditions that appear in the definition of set $\mathcal{A}$.

Returning to Equation (60), for a given scenario pair (s, s'), the indices i' and h are given by $\{(i', h)\} = \mathcal{D}^{s,s'}$ and correspond to the specific endogenous parameter $\theta_{i',h}$ for which scenarios s and s' differ in possible realizations. Also, notice that the pairs for each group are explicitly generated for every time period, even though they are the same in each period (the reasoning here will become apparent later in this section). Finally, to offer a brief insight into the use of this equation, consider an arbitrary group $\hat{l}$ in the context of Figure 10: ${}^{N}\mathcal{G}^{\hat{l}}_{\hat{i},\hat{h}} = \{\hat{s}, \hat{s}', \hat{s}''\}$.[13] By Equation (60), we generate scenario pairs $(\hat{s}, \hat{s}')$ and $(\hat{s}', \hat{s}'')$ for each time period; the third, redundant pair $(\hat{s}, \hat{s}'')$ is implicitly eliminated. (More specifically, for an arbitrary time period $t = \tau$, we will have tuples $(\tau, \hat{s}, \hat{s}'), (\tau, \hat{s}', \hat{s}'') \in \mathcal{SP}^{\hat{i},\hat{h},\hat{l}}_{N^4}$, and $(\tau, \hat{s}, \hat{s}'') \notin \mathcal{SP}^{\hat{i},\hat{h},\hat{l}}_{N^4}$.)

After evaluating Equation (60), there will be one set of pairs for each endogenous scenario group. The union of all of these sets gives the complete set of endogenous scenario pairs, as shown in Equation (61).

$$\mathcal{SP}_{N^4} \coloneqq \bigcup_{i' \in \mathcal{I}} \left(\bigcup_{h \in \mathcal{H}_{i'}} \left(\bigcup_{l \in \mathcal{L}_{i',h}} \mathcal{SP}^{i',h,l}_{N^4} \right) \right) \tag{61}$$

Since this set is at least as restrictive as $\mathcal{SP}_{N^3}$, as previously noted, we claim that $\mathcal{SP}_{N^4} \subseteq \mathcal{SP}_{N^3}$. We now formally state Property 4, by which we prove this claim.

Property 4. For endogenous NACs, it is sufficient to consider only scenario pairs (s, s') for which s and s' are consecutive scenarios in an endogenous scenario group.

Proof. See Appendix section A.6.

The following proposition states that, under special circumstances, the proposed approach leads to the minimum number of endogenous scenario pairs.

Proposition 3. In the case of purely endogenous uncertainty, with no initial 'equality' periods and only one parameter associated with each source, the approach described in Property 4 gives the minimum number of endogenous scenario pairs.

Proof. See Appendix section A.7.

[13] Note that we cannot provide an example in the context of Figure 6c, since there are only 2 possible realizations for each endogenous parameter in that case.

In the general case considered here, however, it is clear that Proposition 3 does not apply. For instance, with both endogenous and exogenous parameters present in the model, some of the endogenous NACs can be implied through the use of exogenous NACs. A simple example of this can be seen with scenario pairs $(1, 5)$ and $(2, 6)$ at the end of the first time period/beginning of the second time period in Figure 6c. We isolate the corresponding scenarios (1, 2, 5, and 6) in Figure 11 to clearly illustrate the issue. Notice that if we consider only variables y_τ^s, we have the exogenous NACs $y_2^1 = y_2^2$ and $y_2^5 = y_2^6$ for the beginning of the second period (shown in red). We also have the conditional endogenous NACs $y_2^1 = y_2^5$ (shown in green) and $y_2^2 = y_2^6$ (shown in orange), which are enforced together as long as endogenous parameter θ_2 is unrealized. Recall that the exogenous NACs *always* hold. We can thus use the two exogenous constraints to rewrite the first endogenous constraint as $y_2^2 = y_2^6$. This, of course, is the second endogenous constraint. Accordingly, we can eliminate the endogenous scenario pair $(2, 6)$ since it is already implied by existing pairs.

Because all scenarios in an exogenous scenario group refer to the same state at that point in time, it is only necessary to consider a single endogenous scenario pair between any two exogenous scenario groups. This transitivity argument is formally stated in the following reduction property.

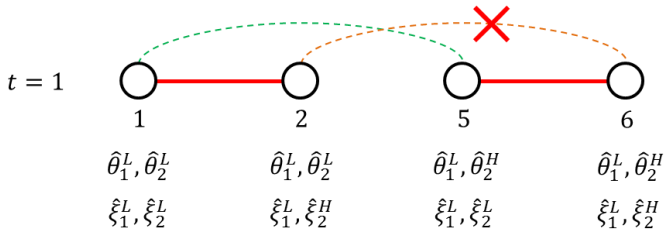


Figure 11. Illustration of Property 5.

Property 5. For any two exogenous scenario groups in time period $t = \tau$ (say, ${}^X\mathcal{G}_\tau^{\hat{k}}$ and ${}^X\mathcal{G}_\tau^{\tilde{k}}$), it is sufficient to consider only *one* endogenous scenario pair (s, s') such that s is in one group and s' is in the other (i.e., $s \in {}^X\mathcal{G}_\tau^{\hat{k}}$ and $s' \in {}^X\mathcal{G}_\tau^{\tilde{k}}$, or vice versa).

Proof. See Appendix section A.8.

Since we require only one endogenous scenario pair between each exogenous scenario group, it is sufficient to consider only a *subset* of scenarios when generating these endogenous pairs. Specifically, for each time period $t \in \mathcal{T}\backslash\{T\}$, rather than considering all scenarios in $\mathcal{S}$, we select a single 'representative' scenario from each exogenous scenario group. This gives us a set of *unique* scenarios, $\widetilde{\mathcal{U}}_t$, in each period. We use the term 'unique' because all of the scenarios in set $\widetilde{\mathcal{U}}_t$ have different realizations for the exogenous parameters up until that point in time and/or different possible realizations for the endogenous parameters.

In selecting these 'representative' scenarios, we must ensure that the corresponding scenario pairs can satisfy Property 3; i.e., $Pos(s) = Pos(s')$, where $s < s'$, and $\left|\mathcal{D}^{s,s'}\right| = 1$. We do this by selecting one

scenario from each exogenous scenario group *in the first subtree*, and then selecting only scenarios *with the same position* in every other subtree. This procedure is repeated for all $t \in \mathcal{T}\backslash\{T\}$.

For example, consider $t = 1$ in Figure 7. If we select scenario 1 from the first group in subtree 1, we must also select scenario 5 from subtree 2, scenario 9 from subtree 3, and scenario 13 from subtree 4. The resulting pairs can satisfy Property 3 since $Pos(1) = Pos(5) = Pos(9) = Pos(13) = 1$. Similarly, if we select scenario 4 from the second group in subtree 1, we must also select scenario 8 from subtree 2, scenario 12 from subtree 3, and scenario 16 from subtree 4. The corresponding pairs can satisfy Property 3 since $Pos(4) = Pos(8) = Pos(12) = Pos(16) = 4$. The set of unique scenarios in this case is then given by $\widetilde{\mathcal{U}}_1 = \{1, 4, 5, 8, 9, 12, 13, 16\}$.

For convenience, we simply select the lowest-indexed scenario from each exogenous scenario group (i.e., the first scenario in each group), as shown in Equation (62). Specifically, $\widetilde{\mathcal{U}}_t$ is expressed as the union of all of these single-scenario sets:

$$\widetilde{\mathcal{U}}_t \coloneqq \bigcup_{k \in \mathcal{K}_t} \left\{ s\colon \ s = \min_{\hat{s}} \left(\hat{s} \in {}^{X}\mathcal{G}_t^k \right) \right\} \ \forall\, t \in \mathcal{T}\backslash\{T\} \tag{62}$$

We let $\widetilde{\mathcal{U}}_T \coloneqq \mathcal{S}$, since there are no exogenous scenario groups defined for $t = T$. Notice that the time index, which was not strictly required in Equation (60), will now play a significant role in the definition of the set of endogenous scenario pairs.

In order to define the set of pairs for each endogenous scenario group, $\mathcal{SP}_{N^5}^{i',h,l}$, corresponding to the addition of Property 5, we first restate our earlier definition corresponding to Property 4 (see Equation (60)). Our only change is to replace set ${}^{N}\mathcal{G}_{i',h}^{l}$ with set ${}^{N}\mathcal{G}_{i',h}^{l} \cap \mathcal{S}$, as follows:

$$\begin{aligned} \mathcal{SP}_{N^4}^{i',h,l} \coloneqq \Big\{ (t, s, s')\colon\ & t \in \mathcal{T},\ s, s' \in \left({}^{N}\mathcal{G}_{i',h}^{l} \cap \mathcal{S} \right), \\ & s' = \min_{\hat{s}'} \left(\hat{s}' \in \left({}^{N}\mathcal{G}_{i',h}^{l} \cap \mathcal{S} \right),\ \hat{s}' > s \right), \\ & s < \max_{\hat{s}} \left(\hat{s} \in \left({}^{N}\mathcal{G}_{i',h}^{l} \cap \mathcal{S} \right) \right),\ \{(i',h)\} = \mathcal{D}^{s,s'} \Big\} \ \forall\, l \in \mathcal{L}_{i',h},\ i' \in \mathcal{I}, \\ & h \in \mathcal{H}_{i'} \end{aligned} \tag{63}$$

Because $\mathcal{S}$ refers to the complete set of scenarios, the intersection of ${}^{N}\mathcal{G}_{i',h}^{l}$ and $\mathcal{S}$ is redundant; there are no scenarios removed from each group, and accordingly, Equation (63) is equivalent to Equation (60). For Property 5, however, we simply replace set $\mathcal{S}$ in this intersection with a subset of unique scenarios, $\widetilde{\mathcal{U}}_t$. The resulting set, ${}^{N}\mathcal{G}_{i',h}^{l} \cap \widetilde{\mathcal{U}}_t$, further restricts $\mathcal{SP}_{N^4}^{i',h,l}$ such that the endogenous scenario pairs can only be formed among *unique* scenarios in each of the endogenous scenario groups in each time period. This further-restricted set is defined as $\mathcal{SP}_{N^5}^{i',h,l}$ in Equation (64).

$$\mathcal{SP}_{N^5}^{i',h,l} \coloneqq \Big\{(t,s,s'):\ t \in \mathcal{T},\ s,s' \in \left({}^{N}\mathcal{G}_{i',h}^{l} \cap \widetilde{\mathcal{U}}_t\right),\ s' = \min_{\hat{s}'}\left(\hat{s}' \in \left({}^{N}\mathcal{G}_{i',h}^{l} \cap \widetilde{\mathcal{U}}_t\right),\ \hat{s}' > s\right),\ s < \max_{\hat{s}}\left(\hat{s} \in \left({}^{N}\mathcal{G}_{i',h}^{l} \cap \widetilde{\mathcal{U}}_t\right)\right),\ \{(i',h)\} = \mathcal{D}^{s,s'}\Big\}\ \forall\, l \in \mathcal{L}_{i',h},\ i' \in \mathcal{I},\ h \in \mathcal{H}_{i'} \tag{64}$$

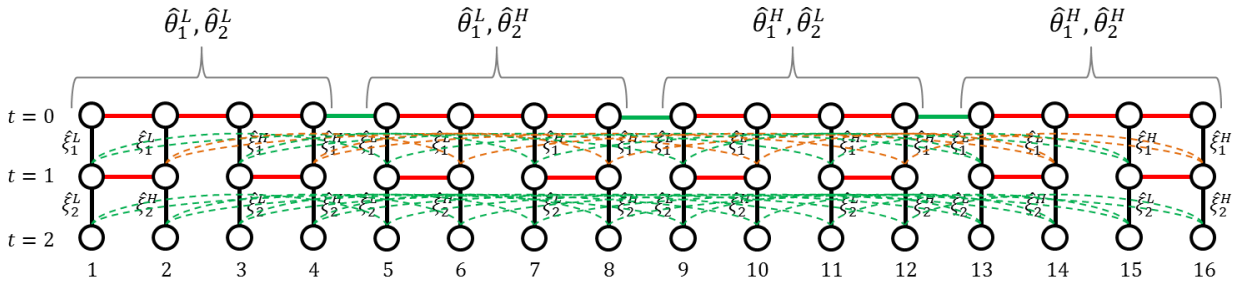
Note that we will refer to sets ${}^{N}\mathcal{G}_{i',h}^{l} \cap \widetilde{\mathcal{U}}_t$ as *reduced* endogenous scenario groups, since for each $l \in \mathcal{L}_{i',h}$, $i' \in \mathcal{I}$, and $h \in \mathcal{H}_{i'}$, this intersection produces a subset of group ${}^{N}\mathcal{G}_{i',h}^{l}$ specific to time period $t \in \mathcal{T}$. (For the case of $t = T$, it is also worth noting that ${}^{N}\mathcal{G}_{i',h}^{l} \cap \widetilde{\mathcal{U}}_T = {}^{N}\mathcal{G}_{i',h}^{l}$ since $\widetilde{\mathcal{U}}_T \coloneqq \mathcal{S}$.)

We then take the union of all of the sets of pairs from Equation (64) in order to produce the complete set of endogenous scenario pairs, $\mathcal{SP}_{N^5}$, as shown in Equation (65). Note that this is the same approach previously used in Equation (61) in the context of Property 4.

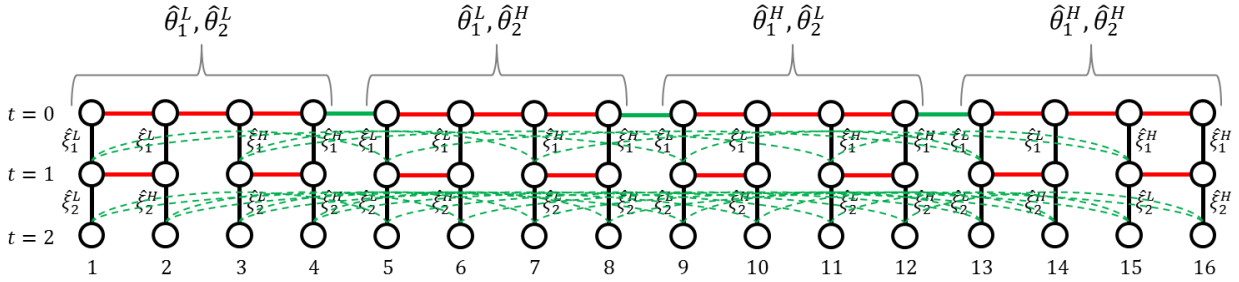
$$\mathcal{SP}_{N^5} \coloneqq \bigcup_{i' \in \mathcal{I}}\left(\bigcup_{h \in \mathcal{H}_{i'}}\left(\bigcup_{l \in \mathcal{L}_{i',h}} \mathcal{SP}_{N^5}^{i',h,l}\right)\right) \tag{65}$$

We state that $\mathcal{SP}_{N^5} \subseteq \mathcal{SP}_{N^4}$ based on the proof of Property 5 (see Appendix section A.8); in other words, Property 5 may eliminate additional pairs that cannot be removed by Property 4. This conclusion can also be reached by comparing Equation (64) to Equation (63).

For illustrative purposes, we apply Property 5 to Figure 6c in order to remove endogenous scenario pair $(2,6)$ and all other similar pairs at the end of the first time period/beginning of the second time period. Here, we have endogenous scenario groups ${}^{N}\mathcal{G}_1^1 = \{1,9\}$, ${}^{N}\mathcal{G}_1^2 = \{2,10\}$, ${}^{N}\mathcal{G}_1^3 = \{3,11\}, \dots,$ ${}^{N}\mathcal{G}_1^8 = \{8,16\}$ corresponding to θ_1, and ${}^{N}\mathcal{G}_2^1 = \{1,5\}$, ${}^{N}\mathcal{G}_2^2 = \{2,6\}$, ${}^{N}\mathcal{G}_2^3 = \{3,7\}, \dots,$ ${}^{N}\mathcal{G}_2^8 = \{12,16\}$ corresponding to θ_2. The set of unique scenarios from Property 5 is given by $\widetilde{\mathcal{U}}_1 = \{1,3,5,7,9,11,13,15\}$. The intersection ${}^{N}\mathcal{G}_{i',h}^{l} \cap \widetilde{\mathcal{U}}_1$ then yields the following: ${}^{N}\mathcal{G}_1^1 \cap \widetilde{\mathcal{U}}_1 = \{1,9\}$, ${}^{N}\mathcal{G}_1^2 \cap \widetilde{\mathcal{U}}_1 = \emptyset$, ${}^{N}\mathcal{G}_1^3 \cap \widetilde{\mathcal{U}}_1 = \{3,11\}, \dots,$ ${}^{N}\mathcal{G}_1^8 \cap \widetilde{\mathcal{U}}_1 = \emptyset$, and ${}^{N}\mathcal{G}_2^1 \cap \widetilde{\mathcal{U}}_1 = \{1,5\}$, ${}^{N}\mathcal{G}_2^2 \cap \widetilde{\mathcal{U}}_1 = \emptyset$, ${}^{N}\mathcal{G}_2^3 \cap \widetilde{\mathcal{U}}_1 = \{3,7\}, \dots,$ ${}^{N}\mathcal{G}_2^8 \cap \widetilde{\mathcal{U}}_1 = \emptyset$, respectively. For the groups listed, this corresponds to scenario pairs $(1,5)$, $(1,9)$, $(3,7)$, and $(3,11)$ (or, more specifically, tuples $(1,1,5),(1,1,9),(1,3,7),(1,3,11) \in \mathcal{SP}_{N^5}$). The respective pairs are illustrated in Figure 12, along with all remaining (non-listed) pairs for the end of the first time period and the end of the second time period. Notice that at $t = 1$, all conditional endogenous NACs involving non-unique scenarios have been removed. Also, note that this reduction does not apply in the final time period; at that time, there are no exogenous scenarios groups that we can exploit (see Figure 7), and all pairs in $\mathcal{SP}_{N^4}$ are present in $\mathcal{SP}_{N^5}$ for $t = T$. This can easily be seen by comparing Equation (63) to Equation (64) with $\widetilde{\mathcal{U}}_T \coloneqq \mathcal{S}$.



(a) Before applying Property 5, there are several redundant scenario pairs at the end of the first time period/beginning of the second time period (shown in orange).



(b) After applying Property 5, the redundant scenario pairs have been eliminated.

Figure 12. Property 5 as applied to Figure 6c.

In certain cases (such as Figure 6c), the addition of Property 5 leads to the minimum number of endogenous scenario pairs. This is formally stated in the following proposition.

Proposition 4. In the case of both endogenous and exogenous uncertainties, with no initial 'equality' periods and only one parameter associated with each source, the approach described in Property 4 and supplemented by Property 5 gives the minimum number of endogenous scenario pairs.

Proof. See Appendix section A.9.

As was the case with Proposition 3, this proposition does not apply in the general case considered here. This is because we may have: (1) endogenous parameters that cannot be realized in some of the initial time periods; and/or (2) multiple endogenous parameters associated with some of the sources of uncertainty. Both of these possibilities have a similar effect on the model.

For the first case, we have *fixed endogenous NACs*, as previously introduced in section 4.2 (see Equations (26)-(28)). An example of this is shown in Figure 13. Here we consider scenarios 1, 5, 9, and 13 from Figure 6c and assume that endogenous parameter θ_2 cannot be realized in the first time period. The four scenarios have identical realizations for the exogenous parameter, and we have four endogenous scenario pairs: $(1, 5)$ and $(9, 13)$, as indicated by solid green lines; $(1, 9)$, as indicated by a dotted green line; and $(5, 13)$, as indicated by a dotted orange line. Notice that scenarios 1 and 5 differ in the possible realization of θ_2 but must be indistinguishable in the first time period because θ_2 cannot be realized at that time. The same is true of scenarios 9 and 13.

If we consider only variables y_τ^s, we have the fixed endogenous NACs $y_2^1 = y_2^5$ and $y_2^9 = y_2^{13}$ for the beginning of the second period. We also have the conditional endogenous NACs $y_2^1 = y_2^9$ and $y_2^5 = y_2^{13}$, which must be enforced together as long as endogenous parameter θ_1 is unrealized. It follows that we can use the two fixed endogenous constraints to rewrite the first conditional endogenous constraint as $y_2^5 = y_2^{13}$. Notice that this is the second conditional endogenous constraint. Accordingly, we can eliminate the endogenous scenario pair $(5, 13)$ since it is already implied by existing pairs. Recall that this result is very similar to what we previously observed in Figure 11 with Property 5.

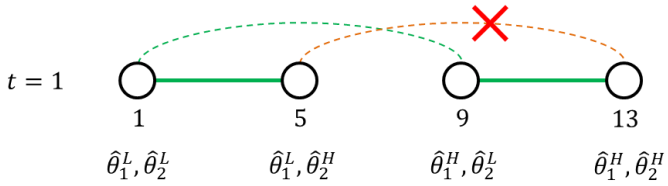


Figure 13. Illustration of Property 6 for endogenous parameters that cannot be realized in some of the initial time periods.

For the second case, we have multiple endogenous parameters associated with some of the sources of uncertainty. We use Figure 14 to illustrate this and consider 4 scenarios ($\hat{s}$, $\hat{s}'$, $\hat{s}''$, and $\hat{s}'''$) in an arbitrary time period $t = \tau$. There are 2 endogenous parameters ($h = 1$ and $h = 2$) associated with a single source $\hat{\imath}$. It is assumed that the scenarios have identical realizations for all exogenous parameters.

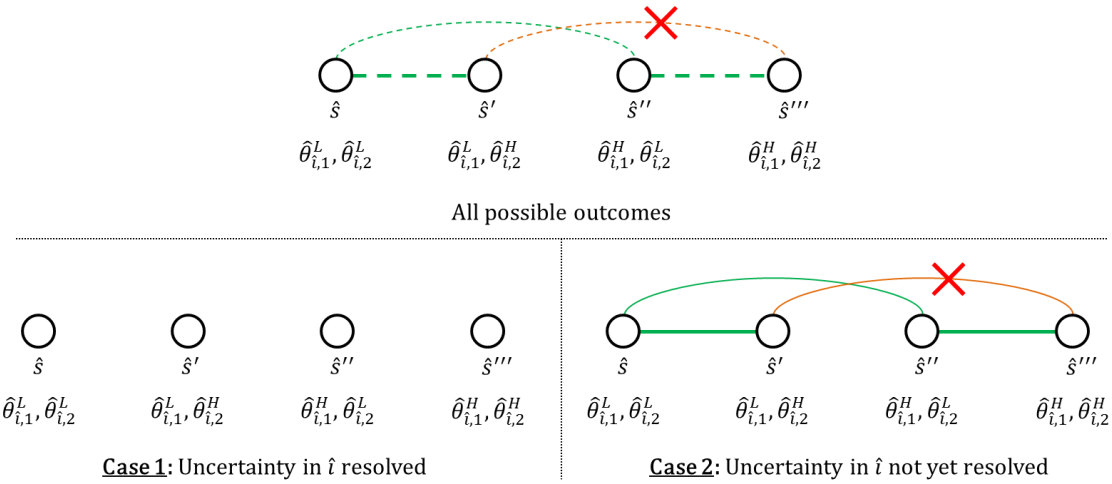


Figure 14. Illustration of Property 6 for multiple parameters associated with a single source of endogenous uncertainty.

By our existing reduction properties, we generate four scenario pairs: $(\hat{s}, \hat{s}')$, $(\hat{s}, \hat{s}'')$, and $(\hat{s}'', \hat{s}''')$, as shown in green, and $(\hat{s}', \hat{s}''')$, as shown in orange. Each of these pairs consists of scenarios s and s' that differ in the possible realization of an endogenous parameter of *the same source* $\hat{\imath}$. This means that if the uncertainty in source $\hat{\imath}$ has been resolved by the end of time period τ, then all of the scenarios will be distinguishable, and the corresponding NACs will be jointly ignored (Case 1). If the uncertainty has not

yet been resolved, then all of the scenarios will be indistinguishable, and the NACs will be jointly enforced (Case 2). Notice that from a modeling perspective, it is only necessary for us to consider Case 2, where the NACs can be viewed as equality constraints. From this viewpoint, it is clear that scenario pair $(\hat{s}', \hat{s}''')$ is redundant and can be eliminated. Recall that this discussion is very similar to what we previously observed in Figure 10 in relation to Property 4.

What we see in these two cases is that some of the endogenous NACs can in fact be implied by other endogenous NACs. To eliminate the corresponding redundant scenario pairs, we extend our definition of unique scenarios.

First, recall that set $\widetilde{\mathcal{U}}_t$ from Property 5 provides a sufficient subset of scenarios (in place of the complete set of scenarios, $\mathcal{S}$) that can be considered when generating endogenous scenario pairs. This is based on the presence of exogenous scenario pairs. In a similar manner, provided that the endogenous scenario pairs are generated in sequential order, we may use the existing endogenous pairs to eliminate additional scenarios from set $\widetilde{\mathcal{U}}_t$ at each step. For example, in the context of Figure 13, after generating scenario pairs $(1, 5)$ and $(9, 13)$ from the endogenous scenario groups corresponding to θ_2, it is clear that there is no further need to consider scenarios 5 and 13; thus, we may remove these scenarios from the groups for θ_1.

We use this concept to define set $\mathcal{U}_t^{i',h}$ for each endogenous parameter $\theta_{i',h}$ and time period $t \in \mathcal{T}$. Each set indicates the unique scenarios available for forming pairs from the groups corresponding to $\theta_{i',h}$, taking into account all endogenous pairs formed *before this point*. The specific definitions for these sets, as well as the order in which to define them, are given by the unique scenarios algorithm, which we present in Appendix section A.10. Note that there is no need to index sets $\mathcal{U}_t^{i',h}$ for $l \in \mathcal{L}_{i',h}$, since the groups corresponding to $\theta_{i',h}$ each contain different scenarios, and any reductions would thus have no effect until we begin forming pairs from the groups of the *next* endogenous parameter.

Next, we formally state the final reduction property, by which we justify the use of the unique scenarios algorithm.

Property 6. For endogenous NACs, it is sufficient to consider only scenario pairs (s, s') for which s and s' are unique, as defined by the unique scenarios algorithm.

Proof. See Appendix section A.11.

This leads to the following proposition.

Proposition 5. In the general case considered throughout this paper, the approach described in Property 4 and supplemented by Property 5 and Property 6 gives the minimum number of endogenous scenario pairs.

Proof. See Appendix section A.12.

We now define the set of all scenario pairs (s, s') in each time period t, such that s and s' differ in the possible realization of one endogenous parameter and are identical in all exogenous realizations, with

additional redundant pairs eliminated by Properties 4–6. We apply the same general approach as described in Equations (60) and (64). Specifically, for each endogenous parameter $\theta_{i',h}$, we first link consecutive scenarios in each of the associated endogenous scenario groups $l \in \mathcal{L}_{i',h}$. We define a separate set for each of these groups in Equation (33). Note that in keeping with the notation of the previous sets in this section (e.g., $\mathcal{SP}_{N^4}^{i',h,l}$ and $\mathcal{SP}_{N^5}^{i',h,l}$), this set should be named $\mathcal{SP}_{N^6}^{i',h,l}$; however, since we will make no further modifications to the following definition, we will simply refer to this set as $\mathcal{SP}_N^{i',h,l}$.

$$\begin{aligned}\mathcal{SP}_N^{i',h,l} \coloneqq \Big\{(t,s,s')\colon\ & t \in \mathcal{T},\ s,s' \in \left({}^N\mathcal{G}_{i',h}^{l} \cap \mathcal{U}_t^{i',h}\right), \\ & s' = \min_{\hat{s}'}\left(\hat{s}' \in \left({}^N\mathcal{G}_{i',h}^{l} \cap \mathcal{U}_t^{i',h}\right),\ \hat{s}' > s\right), \\ & s < \max_{\hat{s}}\left(\hat{s} \in \left({}^N\mathcal{G}_{i',h}^{l} \cap \mathcal{U}_t^{i',h}\right)\right),\ \{(i',h)\} = \mathcal{D}^{s,s'}\Big\}\ \ \forall\, l \in \mathcal{L}_{i',h},\ \ i' \in \mathcal{I}, \\ & h \in \mathcal{H}_{i'}\end{aligned} \tag{33}$$

Notice that the only change from Equation (64) is that we have replaced the reduced endogenous scenario groups ${}^N\mathcal{G}_{i',h}^{l} \cap \widetilde{\mathcal{U}}_t$ with a further reduced set, ${}^N\mathcal{G}_{i',h}^{l} \cap \mathcal{U}_t^{i',h}$, based on Property 6. A brief example of the use of Equation (33) is provided at the end of Appendix section A.10.

Finally, we take the union of all of the individual scenario-pair sets in Equation (34) to obtain set $\mathcal{SP}_N$, the complete set of endogenous scenario pairs. (Again, note that since we will make no further modifications to the following definition, we will refer to this set as $\mathcal{SP}_N$ rather than $\mathcal{SP}_{N^6}$.) This is simply an updated form of Equation (65) in which we have replaced set $\mathcal{SP}_{N^5}^{i',h,l}$ with $\mathcal{SP}_N^{i',h,l}$.

$$\mathcal{SP}_N \coloneqq \bigcup_{i' \in \mathcal{I}}\left(\bigcup_{h \in \mathcal{H}_{i'}}\left(\bigcup_{l \in \mathcal{L}_{i',h}} \mathcal{SP}_N^{i',h,l}\right)\right) \tag{34}$$

Since Property 6 may eliminate additional scenario pairs that cannot be removed by Property 5, we state that $\mathcal{SP}_N \subseteq \mathcal{SP}_{N^5}$ by the proof of Property 6 (see Appendix section A.11). This can also be seen by comparing Equation (33) to Equation (64). It follows that $\mathcal{SP}_N \subseteq \mathcal{SP}_{N^5} \subseteq \mathcal{SP}_{N^4} \subseteq \mathcal{SP}_{N^3}$. Like sets $\mathcal{SP}_F$ and $\mathcal{SP}_X$, the respective scenario pairs in set $\mathcal{SP}_N$ are also non-unique.

In the case of purely endogenous uncertainty, it is worth noting that $\widetilde{\mathcal{U}}_t = \mathcal{S}$ for all $t \in \mathcal{T}$ (since, in each time period, scenario 1 will be assigned to exogenous scenario group 1, and every other scenario will be assigned to a separate group by Equation (50) (due to $\mathcal{SP}_X = \emptyset$)). Thus, Equations (33) and (34) are also applicable for purely-endogenous problems, as previously suggested with the use of $\mathcal{SP}_N$ in model (MSSP_N).

We now formulate one final set in this section. Recall that in model (MSSP) (and accordingly, model (MSSP_N)), there are many cases where we require the *source* i' of the endogenous parameter for which scenarios s and s' differ in possible realizations. The reason, of course, is that there is some information that is specific to the source itself.

For example, we may make an investment in a source to reveal uncertain parameter values, and there may be a certain number of initial time periods (i.e., a lead time) before we can observe these values. The investment decision $b^{S}_{i',t}$ and hence the indistinguishability of scenarios are both specific to the source i'. We require this index to evaluate our uncertainty-resolution rule (see Equation (30)). The set of initial 'equality' time periods $\mathcal{T}_E^{i'}$ and thus the remaining 'conditional' periods $\mathcal{T}_C^{i'}$ are also both specific to the source. Accordingly, the index i' is required in all endogenous non-anticipativity constraints. To further emphasize our point, notice that $b^{S}_{i',t}$, $\mathcal{T}_E^{i'}$, and $\mathcal{T}_C^{i'}$ are *not* indexed for any particular parameter h.

The previously-defined set $\mathcal{D}^{s,s'}$ indicates the specific *parameter* $\theta_{i',h}$ for which s and s' differ in possible realizations. We now define set $\widehat{\mathcal{D}}^{s,s'}$ to indicate only the associated *source*, i':

$$\widehat{\mathcal{D}}^{s,s'} := \left\{ i' :\ i' \in \mathcal{I},\ \exists\, h \in \mathcal{H}_{i'} \text{ s.t. } (i',h) \in \mathcal{D}^{s,s'} \right\}\ \forall\, s,s' \in \mathcal{S},\ s < s',\ Pos(s) = Pos(s') \qquad \textbf{(66)}$$

where we specify that there exists *at least one* endogenous parameter h associated with source i' for which scenarios s and s' differ in possible realizations. (Due to Property 3, there will be *exactly one* endogenous parameter h in each case.)

5.4 Summary of Scenario Pairs and Reduction Properties

In the previous sections, we have presented 6 theoretical reduction properties that eliminate all redundant scenario pairs. This, in turn, eliminates all redundant non-anticipativity constraints, which can significantly reduce the dimensionality of our multistage stochastic programming model, (MSSP), as compared to the case where no reduction properties are applied.

Note that in the reduced form of the model, first-period scenario-pair set $\mathcal{SP}_F$ is defined in Equations (21) and (47), exogenous scenario-pair set $\mathcal{SP}_X$ is defined in Equation (22), and endogenous scenario-pair set $\mathcal{SP}_N$ is defined in Equations (33) and (34). The NACs in model (MSSP) are expressed in terms of these sets. In other words, with the stated definitions, this model is in reduced form, and no further reduction is possible.[14]

6 Solution Methods

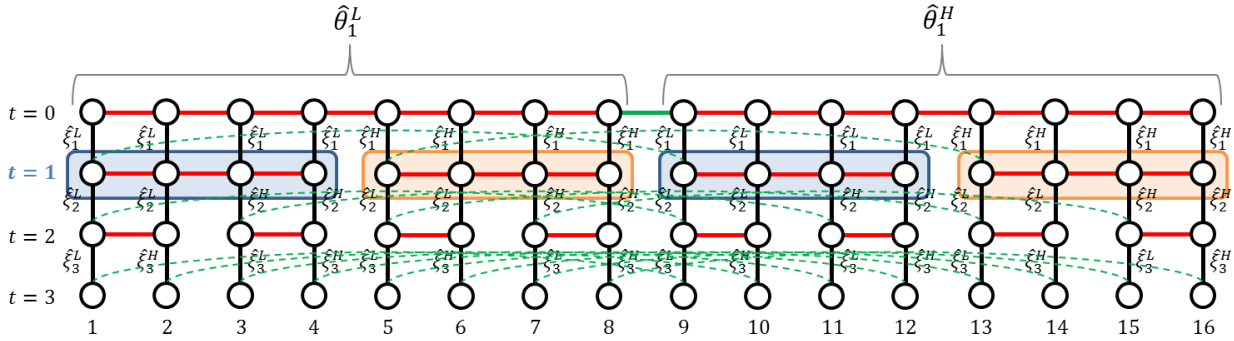
Even after eliminating redundant scenario pairs with properties 1–6, model (MSSP) is often still too large to solve directly with commercial MILP solvers. We thus rely on alternative solution methods. Specifically, we consider a novel sequential scenario decomposition heuristic and Lagrangean decomposition.

6.1 Sequential Scenario Decomposition Heuristic

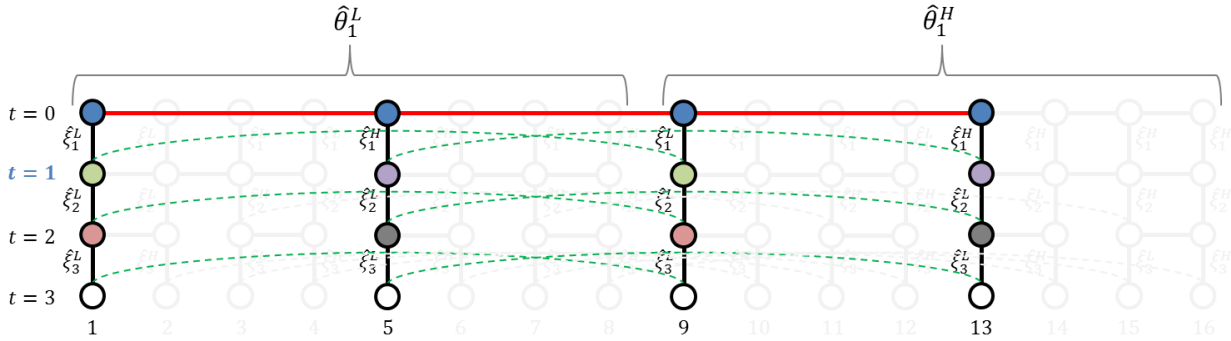
The first alternative solution method that we will discuss is a heuristic that we refer to as sequential scenario decomposition (SSD). The basic idea behind this algorithm is that we sequentially solve endogenous MILP subproblems to determine the binary investment decisions, fix these decisions to

[14] The same can be said of models ($MSSP_X$) and ($MSSP_N$), which are simply special cases of model (MSSP). Also, note that this statement applies to the general formulations considered in this paper; further reduction may be possible in specific problem instances.

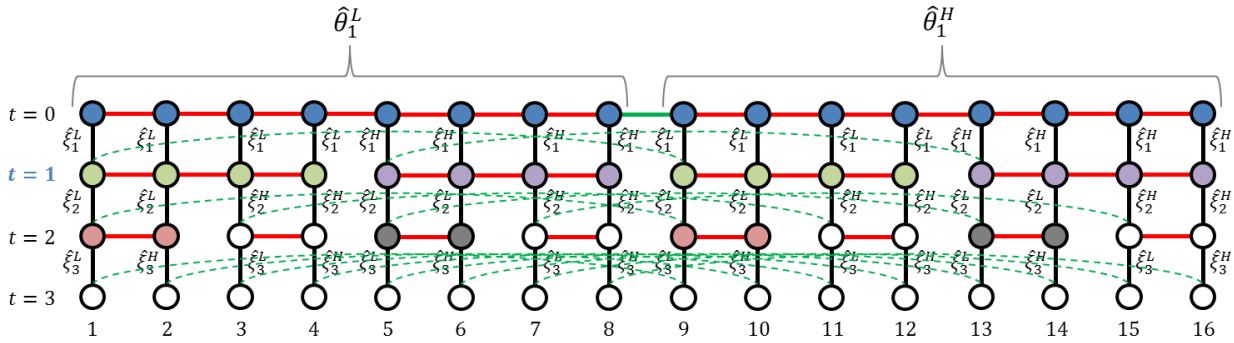
satisfy the corresponding first-period and exogenous NACs, and then solve the resulting model to obtain a feasible solution to the original problem.



(a) Start at $t = 1$ and select one scenario from each exogenous scenario group.



(b) Solve an MILP subproblem that consists of only the selected scenarios. This subproblem includes first-period NACs and endogenous NACs, but no exogenous NACs.



(c) Extract the binary investment decisions from the subproblem solution, and fix these decisions in the scenarios of the original problem in order to satisfy first-period and exogenous NACs.

Figure 15. Sequential scenario decomposition heuristic (first subproblem).

More specifically, we start at $t = 1$ in model (MSSP) and select one scenario from each exogenous scenario group. This subset of scenarios will be connected by only first-period and endogenous NACs, since we have effectively removed all of the exogenous constraints by disregarding many of the scenarios. We then solve this endogenous MILP subproblem (a modified form of model ($MSSP_N$)) and extract the

binary investment decisions from the solution. Returning to the original problem, we fix the respective binary first-stage decisions in all scenarios, and for all other time periods, we fix the binary here-and-now decisions in all scenarios that belong to the same exogenous scenario groups as the subproblem scenarios. We then proceed to the next time period and repeat this process (excluding the consideration of binary first-stage decisions, as these have already been fixed), selecting only scenarios that have not been considered in any previous subproblem. We continue until we reach $t = T - 1$; this is the last subproblem, as we are solving for binary here-and-now decisions for the next time period, and there are no such decisions for $t = T$. After this process is complete, all binary investment decisions will be fixed in model (MSSP). This means that the scenario tree is fixed and we no longer have conditional constraints. The solution of this model gives a feasible solution to the original problem. In Figure 15, we demonstrate the first iteration of the algorithm.

The primary motivation for this procedure is that the subproblems should be considerably easier to solve than the original model. Furthermore, as shown in Figure 15, the first "easy" subproblem includes all of the unique scenarios in the first time period; thus, at the beginning of the planning horizon, we have the same level of information as the original model. The quality of information gradually deteriorates as we proceed forward in time since (by design) some required scenarios are not considered until later subproblems. For instance, in Figure 15, scenarios 3, 7, 11, and 15 are excluded from the first subproblem and thus the model is unaware of the possibility of a *high* demand in the second time period. This demand is accounted for in the next subproblem, after investment decisions have already been fixed in all scenarios at the beginning of the first and second time periods and in half of the scenarios at the beginning of the third time period, based on partial information (see Figure 15c). To our benefit, however, this is typically not a significant concern. In problems with endogenous uncertainty, investment decisions are often made early in the planning horizon, at which point we still have "mostly complete" information. Hence, the subproblem data may not be extensive enough to determine optimal values for the continuous variables, but should be sufficient to approximate the optimal "yes" or "no" investment decisions.

We assume that fixing binary decisions for time period t does not render any later-period subproblems infeasible. Note that when we refer to "binary decisions," we are referring to *all* binary here-and-now decisions $b_{i,t}^s$, as well as any binary components of variable vector y_t^s. For convenience of notation, however, we will represent all such binary decisions as $b_{i,t}^s$ in this section. We next present the algorithm.

Sequential Scenario Decomposition Algorithm

Step 1: Generate all parameters and sets required for model (MSSP).

Step 2: Determine the set of scenarios $\mathcal{S}_{SSD}^{\hat{t}}$ for each subproblem $\hat{t} \in \mathcal{T}\backslash\{T\}$. This is done as follows: for each subproblem $\hat{t}$, select *one* scenario from each exogenous scenario group in this time period (i.e., $s \in \widetilde{\mathcal{U}}_{\hat{t}}$), excluding all scenarios in previous subproblems (i.e., $s \notin \cup_{\hat{\tau}\in\mathcal{T},\ \hat{\tau}<\hat{t}} \mathcal{S}_{SSD}^{\hat{\tau}}$). We exclude the final time period because there are no exogenous scenario groups defined for $t = T$, and we cannot make new here-and-now decisions at the end of the time horizon. Set $\mathcal{S}_{SSD}^{\hat{t}}$ is then given by:

$$\mathcal{S}_{SSD}^{\hat{t}} \coloneqq \left\{ s:\ s \in \widetilde{\mathcal{U}}_{\hat{t}} \backslash \bigcup_{\hat{\tau}\in\mathcal{T},\ \hat{\tau}<\hat{t}} \mathcal{S}_{SSD}^{\hat{\tau}} \right\} \quad \forall\, \hat{t} \in \mathcal{T},\ \hat{t} < T \tag{67}$$

In Figure 15, the set of scenarios for the first subproblem is given by $\mathcal{S}^1_{SSD} \coloneqq \{s\colon\ s \in \{1,5,9,13\}\backslash \emptyset\} = \{1,5,9,13\}$. For the second subproblem (not shown), $\mathcal{S}^2_{SSD} \coloneqq \{s\colon\ s \in \{1,3,5,7,9,11,13,15\}\backslash \{1,5,9,13\}\} = \{3,7,11,15\}$.

Step 3: For $\hat{t} = 1, 2, \ldots, T-1$:

Step 3a: Redefine set $\mathcal{A}$ (Equation (47)), and thus set $\mathcal{SP}_F$ (Equation (21)), using the set of scenarios for subproblem $\hat{t}$ (i.e., $\mathcal{S} \coloneqq \mathcal{S}^{\hat{t}}_{SSD}$ and $S \coloneqq \left|\mathcal{S}^{\hat{t}}_{SSD}\right|$).

Step 3b: Generate subproblem $\hat{t}$, as shown in Figure 15b. This is a modified form of model ($MSSP_N$), where $\mathcal{S}_N \coloneqq \mathcal{S}^{\hat{t}}_{SSD}$ and the non-anticipativity constraints for $b^s_{i,t}$ depend on the subproblem number, $\hat{t}$. Accordingly, replace first-period NACs (25), fixed endogenous NACs (27), and conditional endogenous NACs (36) (assuming a big-M reformulation is used) with constraints (68)–(70), respectively. The idea behind these modifications is that first-period NACs for $b^s_{i,t}$ are no longer needed after the corresponding decisions are fixed in the first subproblem. Thus, Equation (68) enforces them for only the first subproblem, $\hat{t} = 1$. Similarly for the endogenous constraints, at time t, decisions $b^s_{i,t}$ will have been fixed in all earlier time periods $t < \hat{t}$ by previous subproblems; hence, we consider these constraints for only $\hat{t} \le t < T$ in Equations (69) and (70).

$$b^s_{i,1} = b^{s'}_{i,1} \qquad \hat{t} = 1,\ \forall\,(s,s') \in \mathcal{SP}_F,\ i \in \mathcal{I} \tag{68}$$

$$b^s_{i,t+1} = b^{s'}_{i,t+1} \quad \forall\,(t,s,s') \in \mathcal{SP}_N,\ t \in \mathcal{T}^{i'}_E,\ t \ge \hat{t},\ \{i'\} = \widehat{\mathcal{D}}^{s,s'},\ i \in \mathcal{I} \tag{69}$$

$$-\left(1 - z^{s,s'}_t\right) \le b^s_{i,t+1} - b^{s'}_{i,t+1} \le \left(1 - z^{s,s'}_t\right) \quad \forall\,(t,s,s') \in \mathcal{SP}_N,\ t \in \mathcal{T}^{i'}_C,\ \hat{t} \le t < T,\ \{i'\} = \widehat{\mathcal{D}}^{s,s'},\ i \in \mathcal{I} \tag{70}$$

Next, solve subproblem $\hat{t}$. Note that we preserve all endogenous NACs in each time period, so there is no need to update set $\mathcal{SP}_N$.

In some cases, the heuristic subproblem may be too difficult to solve directly. One viable option here is Lagrangean decomposition; specifically, we may apply the endogenous scenario grouping approach described in Gupta and Grossmann (2014a), since these are purely-endogenous problems.

Step 3c: If $\hat{t} = 1$, this is the first subproblem. Use the binary first-stage decisions from this subproblem (i.e., $b^{\hat{s}}_{i,1}\ \forall\, i \in \mathcal{I},\ \hat{s} \in \mathcal{S}^{\hat{t}}_{SSD}$) to fix the binary first-stage decisions in *all* scenarios (i.e., $b^s_{i,1}\ \forall\, i \in \mathcal{I},\ s \in \mathcal{S}$). This is shown in Figure 15c, where the nodes at the beginning of the first time period (originally white in Figure 15a) have now been shaded in blue. Because the first-stage decisions must be identical in all scenarios, we arbitrarily use only the decisions from the first scenario, $\hat{s} = 1$, instead of considering all $\hat{s} \in \mathcal{S}^{\hat{t}}_{SSD}$. Decisions are fixed as shown in Equation (71). Note that this step allows us to satisfy the first-period NACs in the original problem, (MSSP).

$$b_{i,1}^{s} \coloneqq b_{i,1}^{\hat{s}} \qquad \hat{s} = 1,\ \forall\, i \in \mathcal{I},\ s \in \mathcal{S} \tag{71}$$

Step 3d: Fix binary here-and-now decisions in all other time periods. This is done as follows: for each subproblem scenario $\hat{s} \in \mathcal{S}_{SSD}^{\hat{t}}$, start at $t = \hat{t}$ and fix decisions $b_{i,\hat{t}+1}^{\hat{s}}$ in all scenarios in the same exogenous scenario group as $\hat{s}$. We use the condition $G_X(t,s) = G_X(t,\hat{s})$ to check that scenario $s \in \mathcal{S}$ is in the same group as $\hat{s}$.[15] For each remaining time period $t < T$, we repeat this process of fixing decisions $b_{i,t+1}^{\hat{s}}$ in the respective scenario groups. In Figure 15c, for instance, scenario $\hat{s} = 1$ is considered in the first subproblem. Since scenarios $s = 1$, 2, 3, and 4 are all in the same group as $\hat{s}$ in time period 1, we fix their binary decisions in that period based on those of $\hat{s}$. In time period 2, scenarios $s = 1$ and 2 are in the same group as $\hat{s}$, and we fix their binary decisions in an identical manner. This is represented by a change in color of the nodes as compared to Figure 15a. Note that we solve each subproblem for the full time horizon $t \in \mathcal{T}$, but we only fix decisions for $\hat{t} \le t < T$ since the decisions for all previous time periods have already been fixed in the previous subproblems. This step allows us to satisfy the exogenous NACs in the original problem, (MSSP).

$$b_{i,t+1}^{s} \coloneqq b_{i,t+1}^{\hat{s}} \quad \forall\, i \in \mathcal{I},\ t \in \mathcal{T},\ \hat{t} \le t < T,\ s \in \mathcal{S},\ \hat{s} \in \mathcal{S}_{SSD}^{\hat{t}},\ G_X(t,s) = G_X(t,\hat{s}) \tag{72}$$

Step 4: At this point, the binary here-and-now decisions $b_{i,t}^{s}$ have been fixed for all $i \in \mathcal{I}$, $t \in \mathcal{T}$, and $s \in \mathcal{S}$. (In Figure 15, this would occur after one more iteration.) Thus, in model (MSSP), drop all NACs related to these decisions. This includes the first-period NACs given by Equation (25), the exogenous NACs given by Equation (44), the fixed endogenous NACs given by Equation (27), and the conditional endogenous NACs given by either Equation (36) or the third line of Equation (29). Note that the scenario tree is fixed at this point, since indistinguishability can be determined by directly calculating $Z_t^{s,s'}$ (and thus $z_t^{s,s'}$) from the known values of $b_{i,t}^{s}$.

Next, redefine set $\mathcal{A}$ and set $\mathcal{SP}_F$ using the complete set of scenarios (i.e., $S \coloneqq |\mathcal{R}|$ by Equation (6) and $\mathcal{S} \coloneqq \{s:\ s = 1, 2, \dots, S\}$). Then, solve the resulting form of model (MSSP). This provides a feasible, but not necessarily optimal, solution to (MSSP).

Note that since the scenario tree is fixed in the final form of model (MSSP), the timing of all realizations is known in advance; thus, all uncertainties can be viewed as exogenous. This model, however, is not in the form of a purely-exogenous stochastic program (i.e., (MSSP_X)). For large instances where a direct-solution approach is impractical, we have two basic options: (1) preserve the structure and apply Lagrangean decomposition, as discussed in the next section; or (2) reformulate the problem into the form of model (MSSP_X), as shown graphically in section 7.1. In the latter case, we can take advantage of effective solution methods for purely-exogenous MSSP problems, such as the branch-and-fix coordination scheme by Escudero et al. (2009).

This heuristic can be used to obtain an initial upper bound in a Lagrangean decomposition algorithm, as discussed in the next section.

[15] Rather than the conditions $s \in \mathcal{S},\ G_X(t,s) = G_X(t,\hat{s})$ in Equation (72), we could state $s \in {}^{X}\mathcal{G}_t^{\hat{k}}$, where $\hat{k}$ is the group number corresponding to $\hat{s}$, given by $\hat{k} = G_X(t,\hat{s})$.

6.2 Lagrangean Decomposition

From Figure 6c, it is clear that if we remove all non-anticipativity constraints, then the scenario tree decomposes into independent scenarios. This is shown in Figure 16. The appealing aspect of this structure is that independent scenario subproblems should be considerably easier to solve than the full model. Such reasoning is the primary motivation behind Lagrangean decomposition, in which 'complicating' (i.e., 'linking') constraints are dualized in order to achieve a similar relaxation of the original model (Carøe and Schultz, 1999; Goel and Grossmann, 2006; Gupta and Grossmann, 2011; Escudero et al., 2016). In this context, the complicating constraints are the NACs.

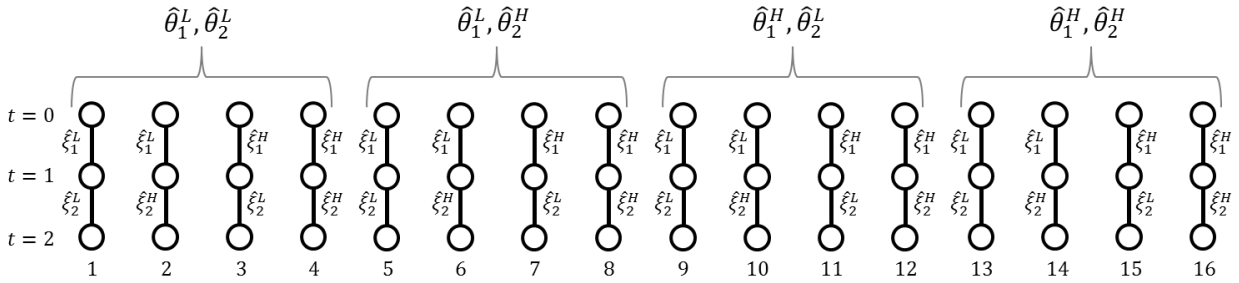


Figure 16. A scenario tree decomposes into independent scenarios when all NACs are removed.

As described in Gupta and Grossmann (2014a), in the case of standard Lagrangean decomposition for MSSP problems with endogenous uncertainties, the first step is to relax all of the conditional endogenous NACs. We then form the Lagrangean relaxation (Guignard, 2003) by dualizing the first-period and fixed endogenous NACs. This entails moving these constraints to the objective function as penalty terms multiplied by Lagrange multipliers. In our case, we must also dualize the exogenous NACs (Goel and Grossmann, 2006). We use a simplified form of model (MSSP) to illustrate this, (MSSPS), where for simplicity we keep only decision variables y_t^s. We also assume that the set of initial 'equality' periods is identical for all sources of endogenous uncertainty; i.e., $\mathcal{T}_E^{i'} = \mathcal{T}_E$, and thus $\mathcal{T}_C^{i'} = \mathcal{T}_C$, for all $i' \in \mathcal{I}$.

(MSSPS)

$$\min_{y} \phi = \sum_{s\in\mathcal{S}} p^s \sum_{t\in\mathcal{T}} {}^{y}c_t^s y_t^s \tag{73}$$

$$\text{s.t.} \quad \sum_{\tau=1}^{t} {}^{y}A_{\tau,t}^s y_\tau^s \le a_t^s \qquad \forall\, t \in \mathcal{T},\ s \in \mathcal{S} \tag{74}$$

$$y_1^s = y_1^{s'} \qquad \forall\, (s,s') \in \mathcal{SP}_F \tag{17}$$

$$y_{t+1}^s = y_{t+1}^{s'} \qquad \forall\, (t,s,s') \in \mathcal{SP}_X \tag{19}$$

$$y_{t+1}^s = y_{t+1}^{s'} \qquad \forall\, (t,s,s') \in \mathcal{SP}_N,\ t \in \mathcal{T}_E \tag{28}$$

$$-y_{t+1}^{UB}\left(1 - z_t^{s,s'}\right) \le y_{t+1}^s - y_{t+1}^{s'} \le y_{t+1}^{UB}\left(1 - z_t^{s,s'}\right) \qquad \forall\, (t,s,s') \in \mathcal{SP}_N,\ t \in \mathcal{T}_C,\ t < T \tag{37}$$

$$Z_t^{s,s'} \Leftrightarrow F(y_1^s, y_2^s, \dots, y_t^s) \qquad \forall\, (t,s,s') \in \mathcal{SP}_N,\ t \in \mathcal{T}_C \tag{75}$$

$$y_t^s \in \mathcal{Y}_t^s \qquad \forall\, t \in \mathcal{T},\ s \in \mathcal{S} \tag{76}$$

$$Z_t^{s,s'} \in \{True, False\} \qquad \forall\, (t,s,s') \in \mathcal{SP}_N,\ t \in \mathcal{T}_C \tag{32}$$

$$z_t^{s,s'} \in \{0,1\} \qquad \forall\, (t,s,s') \in \mathcal{SP}_N,\ t \in \mathcal{T}_C \tag{38}$$

In the simplified Lagrangean relaxation problem, (MSSPS-LD), we remove endogenous constraints (37), (75), (32), and (38), and dualize constraints (17), (19), and (28).

(MSSPS-LD)

$$\min_y \phi_{LD}(\lambda) = \sum_{s\in\mathcal{S}} p^s \sum_{t\in\mathcal{T}} {}^y c_t^s y_t^s + \sum_{(s,s')\in\mathcal{SP}_F} {}^F\lambda_1^{s,s'}\left(y_1^s - y_1^{s'}\right) + \sum_{(t,s,s')\in\mathcal{SP}_X} {}^X\lambda_t^{s,s'}\left(y_{t+1}^s - y_{t+1}^{s'}\right) + \sum_{\substack{(t,s,s')\in\mathcal{SP}_N \\ t\in\mathcal{T}_E}} {}^N\lambda_t^{s,s'}\left(y_{t+1}^s - y_{t+1}^{s'}\right) \tag{77}$$

$$\text{s.t.} \quad \sum_{\tau=1}^{t} {}^y A_{\tau,t}^s y_\tau^s \le a_t^s \qquad \forall\, t \in \mathcal{T},\ s \in \mathcal{S} \tag{74}$$

$$y_t^s \in \mathcal{Y}_t^s \qquad \forall\, t \in \mathcal{T},\ s \in \mathcal{S} \tag{76}$$

Notice that all complicating constraints have now been either removed or dualized, and Equations (74) and (76) apply only to individual scenarios. Further notice, however, that the objective function, Equation (77), still contains variables y_t^s and $y_t^{s'}$, so we cannot yet decompose the problem by scenario. We expand this expression, swap indices s and s' in certain summations, and then simplify in order to rewrite the objective function as Equation (78) (see the supplementary material for further details).

$$\min_y \phi_{LD}(\lambda) = \sum_{s\in\mathcal{S}} \left(p^s \sum_{t\in\mathcal{T}} {}^y c_t^s y_t^s + y_1^s \left(\sum_{(s,s')\in\mathcal{SP}_F} {}^F\lambda_1^{s,s'} - \sum_{(s',s)\in\mathcal{SP}_F} {}^F\lambda_1^{s',s} \right) + \sum_{t\in\mathcal{T}\setminus\{T\}} y_{t+1}^s \left(\sum_{(t,s,s')\in\mathcal{SP}_X} {}^X\lambda_t^{s,s'} - \sum_{(t,s',s)\in\mathcal{SP}_X} {}^X\lambda_t^{s',s} \right) + \sum_{t\in\mathcal{T}_E} y_{t+1}^s \left(\sum_{(t,s,s')\in\mathcal{SP}_N} {}^N\lambda_t^{s,s'} - \sum_{(t,s',s)\in\mathcal{SP}_N} {}^N\lambda_t^{s',s} \right) \right) \tag{78}$$

The variables in the objective function now involve only scenario s, and all other terms are constants. Accordingly, the problem can be decomposed into independent scenario subproblems that can be solved in parallel. This is done in an iterative fashion, as shown in Figure 17 (adapted from Gupta and Grossmann (2011)). In each iteration, we first solve the subproblems with fixed multipliers to obtain a lower bound to the original problem (MSSPS). The lower bound is simply equal to the sum of the subproblem objective function values, and an upper bound is determined by a simple heuristic. In this heuristic, we selectively fix decisions from the subproblems in the original problem to obtain a feasible solution (see the supplementary material for complete details). The solution from the sequential scenario decomposition heuristic may be used as an initial upper bound; however, this is not required. We then

apply the subgradient method (Fisher, 1985) to update the multipliers for the Lagrangean problem,[16] and repeat this process until the difference between the upper bound and lower bound lies within a pre-specified tolerance or until a maximum iteration limit is reached. Note that if we are unable to sufficiently close the gap, it may be necessary to implement a branch-and-bound algorithm such as the one proposed by Goel and Grossmann (2006) and Goel et al. (2006).

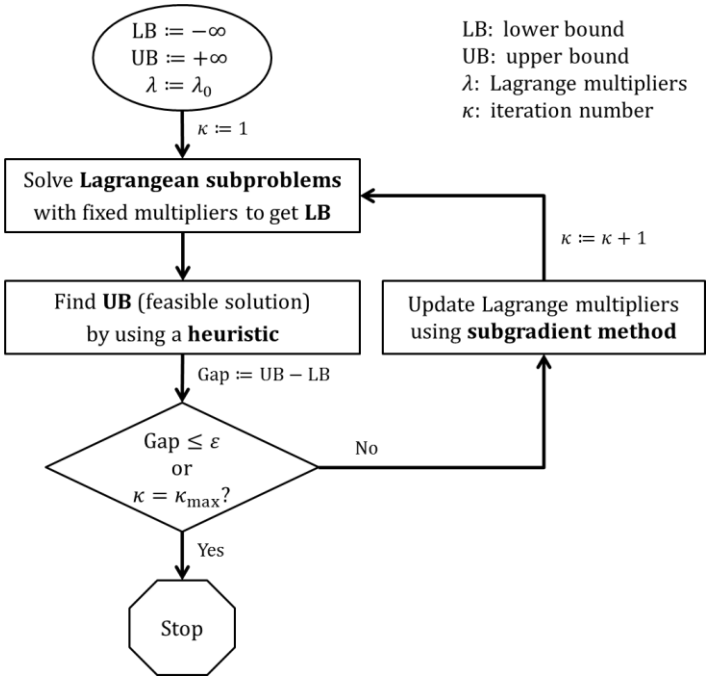


Figure 17. Algorithm for Lagrangean decomposition.

7 Numerical Results

7.1 Motivating Example

Consider the simple process network shown in Figure 18, as adapted from Goel and Grossmann (2006). In this example, a product A is produced in Process III which has an existing capacity of 3 tons/hr and a known yield of 70%. This process requires a feed of chemical B that is currently purchased. The demand of product A is uncertain but must be satisfied for each time period in the planning horizon. If the demand cannot be met by production, product A is purchased from a competitor.

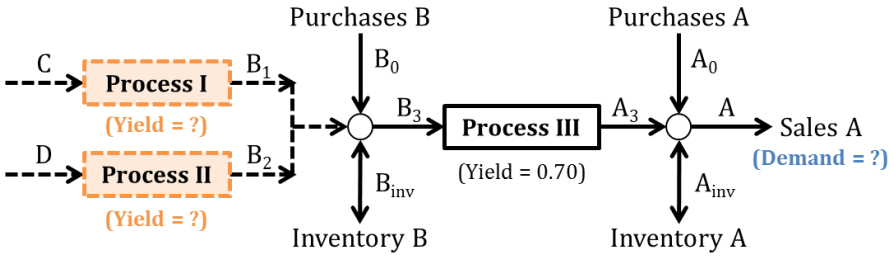


Figure 18. Process network for the motivating example.

[16] There are many alternative multiplier-update procedures. See Escudero, Garín, Pérez, and Unzueta (2013), as well as Oliveira et al. (2013) and the references therein.

Due to the high price of B, it is proposed that some (or all) of this chemical be manufactured from raw material C in a new process, Process I, or from raw material D in a second new process, Process II. These processes are not exclusive, and neither, one, or both may be installed.

The yield of Process I is uncertain, with possible realizations $\{0.69, 0.81\}$, both with an equal probability of 0.5. The yield of Process II is also uncertain, with possible realizations $\{0.62, 0.85\}$, both with equal probabilities. The objective is to determine the optimal investment and operation decisions over a 2-year planning horizon in order to maximize the total expected profit from the sales of A. Over this time horizon, the demand of product A has possible realizations $\{1.10, 3.10\}$ tons/hr in time period 1 and $\{2.25, 4.25\}$ tons/hr in time period 2, each with probability 0.5. We do not provide the remaining problem data here; however, this data is available upon request.

Regarding the *types* of uncertainty, the yields of Process I and Process II represent endogenous parameters (θ_1 and θ_2, respectively), since they are uncertain until the units are installed and operated. For simplicity, we assume that the units are operated immediately after they are installed. The demand of product A is an exogenous parameter (ξ_t), as it is a market value that will be realized automatically in each time period. There are 4 possible combinations of realizations for the endogenous parameters and 4 possible combinations of realizations for the exogenous parameters. This gives rise to a 3-stage, 16-scenario stochastic programming problem. Note that this corresponds to the composite scenario tree previously introduced in Figure 6c. We first use a direct-solution approach to solve the *fullspace model* (i.e., model (MSSP) with no reduction properties applied) and the *reduced model* (i.e., model (MSSP) in its current form, with all reduction properties applied). We then apply the sequential scenario decomposition (SSD) heuristic and Lagrangean decomposition (LD). The corresponding model statistics are provided in Table 1.

Table 1. Model statistics for the motivating example.

Problem Type	Scenarios	Constraints	Continuous Variables	Binary Variables
Fullspace	16	5,985	913	240
Reduced Model	16	1,472	913	120
SSD (Sub 1)	8	784	457	64
SSD (Final)	16	1,355	913	24
LD	4	286	229	24

First, we observe that by applying Properties 1–6 through the set definitions proposed in section 5, we are able to reduce the total number of constraints from 5,985 to 1,472; a 75% reduction, based solely on the removal of redundant NACs. We are also able to eliminate half of the binary variables (specifically, the indistinguishability variables $z_t^{s,s'}$ associated with redundant conditional endogenous NACs). This effect can be even more pronounced in larger problem instances, as will be shown in the next section.

Furthermore, the SSD heuristic requires only one subproblem to fix all of the binary here-and-now decisions. This subproblem consists of scenarios 1, 3, 5, 7, 9, 11, 13, and 15 (see Figure 6c). Recall that we fix the respective binary decisions in these scenarios, and all remaining scenarios, in order to satisfy the corresponding first-period and exogenous non-anticipativity constraints.

Table 1 also indicates that there are still binary variables in the final SSD problem, SSD (Final). This is due to the indistinguishability variables $z_t^{s,s'}$, which are simply calculated quantities given the fixed values of $b_{i,t}^s$. We may choose to either fix these variables prior to generating the model, or allow the solver to perform these calculations. For convenience, we choose the latter option in this case. Note that when there are no other integer variables in the problem, and indistinguishability is determined by Equations (40) and (41), we may obtain the optimal solution of the final SSD problem by solving its LP relaxation. This is because, by these inequalities, $z_t^{s,s'}$ must be 0 or 1 if $b_{i,t}^s$ is also binary.

The problem size reported for Lagrangean decomposition corresponds to the size of each Lagrangean-dual subproblem. For this example, we decompose the problem such that each subproblem corresponds to one subtree (i.e., 4 scenarios with all non-anticipativity constraints intact), rather than one individual scenario. Thus, we must dualize only 3 sets of first-period NACs, which gives 4 independent subproblems of 4 scenarios each.

Note that this Lagrangean-decomposition strategy is inspired by the scenario clustering approach of Escudero et al. (2016); we will use a similar strategy for our LD implementations in the following two sections as well. While, in principle, we may also merge indistinguishable nodes within each subtree such that all non-dualized first-period and exogenous NACs are implicitly enforced (i.e., for each subtree, we may adopt the standard form shown in Figure 2a), this would require significantly more complex notation which we wish to avoid.

We solve the motivating example in GAMS 24.3.3, with CPLEX 12.6.0.1, on a machine with a 2.50 GHz Intel Core i5 CPU and 4 GB of RAM. The optimal solution is to install Process I at the beginning of the first time period with a capacity of 3.704 tons/hr and perform no expansions. The total expected profit is \$5.069 MM. The computational results are given in Table 2. Note that each reported solution time reflects only the solver time and does *not* include the model generation time. Given the complexity of the parameters and sets defined in section 5, it is also worth noting that the generation time for the reduced model is less than one minute for all example problems in this paper.

Table 2. Numerical results for the motivating example.

Problem Type	Total Expected Profit (\$MM)		Optimality Gap	Solution Time (s)
	Lower Bound	Upper Bound		
Fullspace	5.069	5.069	0%	0.08
Reduced Model	5.069	5.069	0%	0.06
SSD	5.069	-	-	0.11
LD	5.069	5.069	0.006%	11.70

We observe that the SSD heuristic obtains the optimal solution as a lower bound, and after 14 iterations, the Lagrangean decomposition algorithm converges to the optimal solution. Figure 19 shows the best bounds obtained by LD at each iteration of the algorithm. Since this is a very simple example, it is faster in this case to directly solve the reduced model than it is to solve subproblems in the alternative solution methods. For larger instances, these alternative methods yield considerable savings in computational time, as will be seen in the next section.

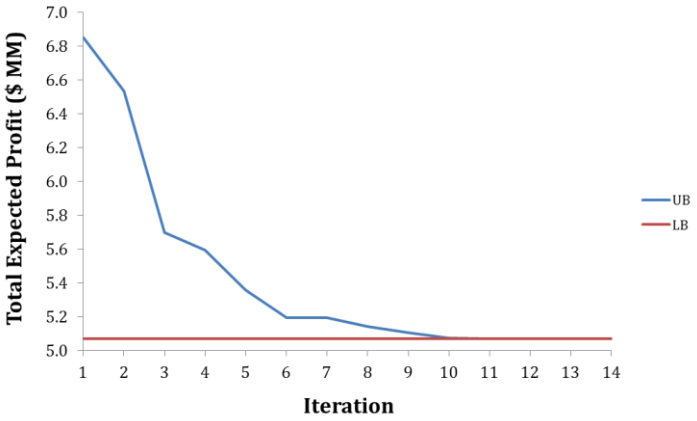


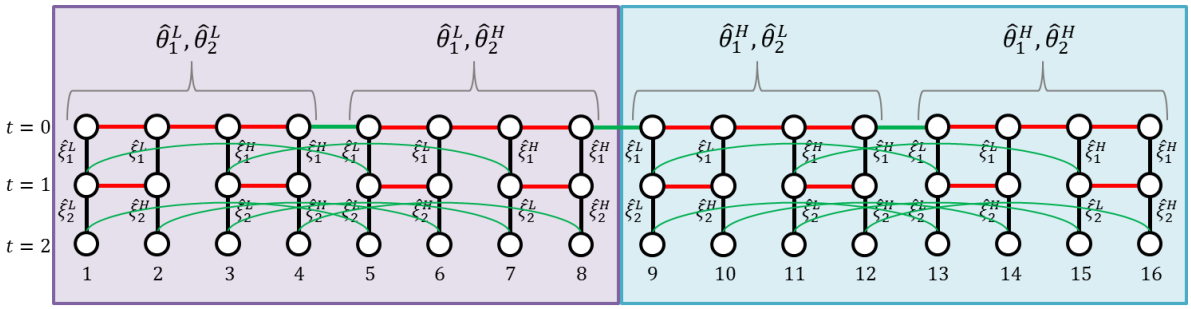
Figure 19. Best bounds on the optimal solution of the motivating example, as obtained by Lagrangean decomposition.

The optimal structure of the composite scenario tree is shown in Figure 20. Notice that, starting from the superstructure form in Figure 6c, the dotted green lines have transitioned into solid green lines for active NACs and have disappeared entirely for inactive NACs. We also show that by taking advantage of the active NACs and the known timing of the endogenous realizations, we are able to recover the standard form of the scenario tree. This form is significantly easier to interpret, as can be seen in Figure 20.

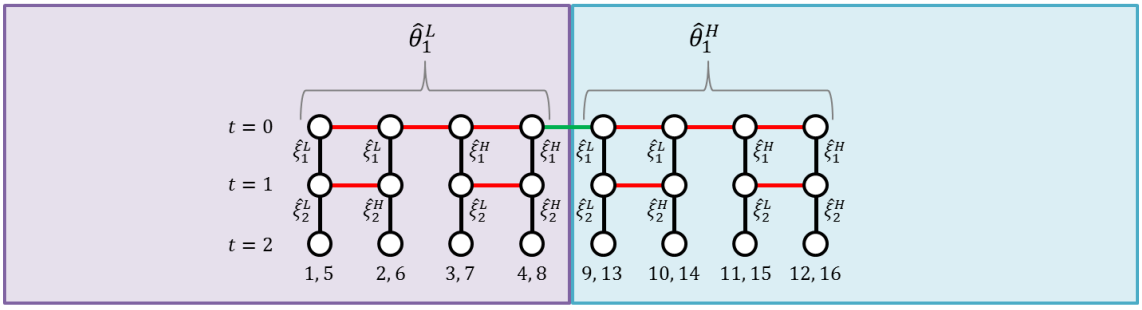
Note that unlike the two-stage case, the value of the stochastic solution (VSS) is not a trivial calculation for multistage problems. We do not perform this calculation here; however, we refer the reader to Escudero et al. (2007) and Maggioni et al. (2014) for further information on this topic.

7.2 Example 1: Capacity Expansion of a Process Network

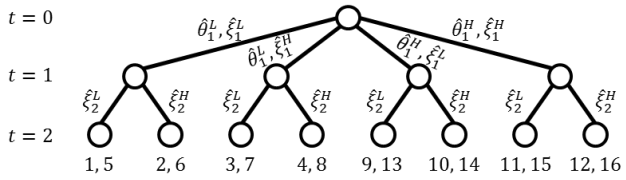
We now consider a larger instance of the motivating example. Specifically, we extend the time horizon to 8 years and consider 2 possible realizations for the demand of product A in each period, as well as 3 possible realizations for the yield of Process I and Process II. This gives 9 possible combinations of realizations for the endogenous parameters, and 256 possible combinations of realizations for the exogenous parameters. The result is a 9-stage stochastic programming problem with 2,304 scenarios. The corresponding model statistics for the fullspace model, the reduced model, and the SSD and LD problems are provided in Table 3. Notice in particular that the fullspace model for this instance has more than *176 million* constraints and approximately *4.8 million* binary variables. Such a model is clearly intractable in its current state. With the application of the reduction properties, we are able to reduce the number of constraints to about 838,000 – a *99.5% reduction*. The number of binary variables is also reduced to about 61,000, which is a 98.7% reduction.



(a) Optimal structure of the composite scenario tree. Notice the symmetry: the yield of Process I is *low* on the left and *high* on the right.



(b) Since the yield of Process II (θ_2) is unrealized during the time horizon, we can merge indistinguishable scenarios based on active NACs.



(c) Noting that the yield of Process I (θ_1) is realized in the first time period, we may now recover the standard form of the scenario tree.

Figure 20. Optimal structure of the composite scenario tree for the motivating example, and the procedure for converting this tree into its equivalent standard form.

Table 3. Model statistics for Example 1.

Problem Type	Scenarios	Constraints	Continuous Variables	Binary Variables
Fullspace	2,304	176,203,009	518,401	4,755,456
Reduced Model	2,304	838,318	518,401	61,416
SSD (Sub 1)	18	9,196	4,051	624
SSD (Sub 2)	18	8,569	4,051	492
SSD (Sub 3)	36	15,943	8,101	828
SSD (Sub 4)	72	29,395	16,201	1,344
SSD (Sub 5)	144	53,611	32,401	2,064
SSD (Sub 6)	288	96,475	64,801	2,880
SSD (Sub 7)	576	170,683	129,601	3,264
SSD (Final)	2,304	771,595	518,401	6,120
LD	256	78,862	57,601	6,144

Because of the longer time horizon, there are now more subproblems required for the sequential scenario decomposition heuristic. Each of these problems is significantly larger than the *one* in the motivating example; however, the model growth is slightly non-intuitive. Specifically, note the decrease in the problem size in the second subproblem, SSD (Sub 2). The reason for this is as follows.

At $\hat{t} = 1$, there are 9 subtrees, each containing 2 exogenous scenario groups. We select *one* scenario from each of these groups. In other words, we consider 18 scenarios in the first subproblem, and at this point, no binary here-and-now decisions have been fixed. In the second subproblem, $\hat{t} = 2$, we first consider 36 scenarios (i.e., 9 subtrees, each containing 4 exogenous scenario groups). The binary decisions have already been fixed in 18 of these scenarios. Accordingly, we neglect these 18 scenarios and consider only the remaining 18. Notice that this is *the same number of scenarios* as the first subproblem (see the supplementary material for further details); however, the binary here-and-now decisions for the previous time periods have already been fixed. This means that there will be fewer constraints and binary variables, as can be seen in Table 3.

In the third subproblem, $\hat{t} = 3$, we first consider 72 scenarios. The binary decisions have already been fixed in 36 of them, so we consider only the remaining 36. Notice that at this point, the number of scenarios in each subproblem begins to double (see the supplementary material). The problem size, however, does *not* double. This can be seen in the corresponding number of binary variables reported in Table 3. Since in each subproblem the binary decisions in all previous time periods have already been fixed, we are able to effectively slow the problem growth.

We emphasize that the SSD subproblems are significantly smaller than the reduced model. At most, we consider 576 scenarios in subproblem 7. This is only 25% of the total number of scenarios. Moreover, this particular subproblem contains only 20% of the constraints of the reduced model and 5% of the binary variables.

In the Lagrangean decomposition algorithm, we again decompose the problem by subtrees (rather than by individual scenarios). This gives 9 subproblems of 256 scenarios each. Like for the SSD heuristic, these subproblems are considerably larger than those in the motivating example.

We solve this problem instance in GAMS 24.3.3, with CPLEX 12.6.0.1, on a machine with a 2.50 GHz Intel Core i5 CPU and 4 GB of RAM. The results are summarized in Table 4.

Table 4. Numerical results for Example 1.

Problem Type	Total Expected Profit ($MM)		Optimality Gap	Solution Time (s)
	Lower Bound	Upper Bound		
Fullspace	-	-	-	-
Reduced Model	142.411	143.828	0.99%	3,670
SSD	142.411	-	-	61
LD	142.411	144.424	1.41%	913

As would be expected, the fullspace model cannot be loaded into memory. After applying the reduction properties, however, we are in fact able to solve this instance to a 0.99% optimality gap in about 1 hour. The best feasible solution obtained from the reduced model is $142.411 MM.

As shown in Table 4, the SSD heuristic provides the same feasible solution as the reduced model in just 61 seconds. We use this value as the initial lower bound for the Lagrangean decomposition algorithm. After 20 iterations, the lower bound does not improve, and we obtain an upper bound of $144.424 MM. This then provides us with bounds on the optimal solution; specifically, within a 1.41% optimality gap. The total time for lower- and upper-bound generation for the alternative solution methods is 974 seconds – a 73% reduction from solving the reduced model directly.

7.3 Example 2: Oilfield Development Planning

We consider a modified form of the MILP described in Gupta and Grossmann (2014a) (see Case (i)) for maximizing the total expected NPV in the development planning of an offshore oilfield. There are 3 oilfields; 3 potential Floating Production, Storage, and Offloading vessels (FPSOs); and 9 possible field-FPSO connections. A total of 30 wells can be drilled over a 5-year planning horizon: 7 for field I, 11 for field II, and 12 for field III. There is also a 3-year lead time for FPSO construction and a 1-year lead time for FPSO expansion. Fields II and III have a known recoverable oil volume (size); however, the size of field I is uncertain. Specifically, there are 2 possible realizations for the size of field I, both with equal probabilities. The oil and gas prices are also uncertain, with 2 possible realizations with equal probabilities in each time period. These prices are assumed to be correlated. The network superstructure for this problem instance is shown in Figure 21.

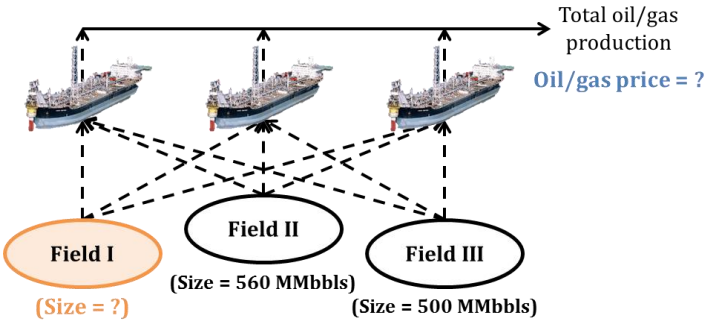


Figure 21. Network superstructure for the oilfield development planning problem. (FPSO images from www.rigzone.com.)

Notice that the size of field I is an endogenous parameter, since this information cannot be realized until we drill the field and begin producing from it. The oil and gas prices are exogenous parameters, as they are market values that will be realized automatically in each time period. We have 2 possible combinations of realizations for the endogenous parameters and 32 possible combinations of realizations for the exogenous parameters. Using the scenario-generation procedure described in section 3, this gives rise to a 6-stage, 64-scenario stochastic programming problem. The corresponding model statistics are shown in Table 5. In particular, notice that the fullspace model consists of 333,249 constraints and 7,360 binary variables. After applying the theoretical reduction properties, there are 124,980 constraints and 7,000 binary variables. This is a 62% reduction in the number of constraints. The number of binary variables is reduced by approximately 5%.

Model statistics for the heuristic and Lagrangean decomposition are also provided in Table 5. Note that for the Lagrangean decomposition algorithm, we again choose not to decompose the problem by individual scenarios. However, rather than decomposing by subtrees, as this leads to very difficult subproblems, we instead consider 32 subproblems of 2 adjacent scenarios each.

Table 5. Model statistics for Example 2.

Problem Type	Scenarios	Constraints	Continuous Variables	Binary Variables
Fullspace	64	333,249	70,465	7,360
Reduced Model	64	124,980	70,465	7,000
SSD (Sub 1)	4	7,589	4,333	440
SSD (Sub 2)	4	7,415	4,333	356
SSD (Sub 3)	8	14,655	8,665	664
SSD (Sub 4)	16	28,687	17,329	1,224
SSD (Final)	64	116,787	69,313	4,504
LD	2	3,776	2,203	218

The problem was modeled in GAMS 24.3.3 and solved with CPLEX 12.6.0.1 on a machine with a 2.93 GHz Intel Core i7 CPU and 12 GB of RAM. Table 6 summarizes the results for the different solution approaches. In the case of solving the reduced model directly, the optimality gap cannot be improved past 50% after more than 11 hours. In contrast, the sequential scenario decomposition heuristic finds a high-quality feasible solution ($7.166 billion) in only *41 seconds*. We initialize the lower bound of the Lagrangean decomposition algorithm to this objective value. After only 14 seconds, the LD algorithm finds a high-quality upper bound ($7.180 billion); the lower bound does not improve. This implies that the SSD solution is within 0.20% of the optimum. Notice that we obtain this information in *less than one minute* of CPU time.

Table 6. Numerical results for Example 2.

Problem Type	Total Expected NPV (10^9)		Optimality Gap	Solution Time (s)
	Lower Bound	Upper Bound		
Reduced Model	6.968	10.495	50.61%	40,562
SSD	7.166	-	-	41
LD	7.166	7.180	0.20%	14

The network structure corresponding to the best feasible solution ($7.166 billion, as obtained by the SSD heuristic) is shown in Figure 22. This solution indicates that we begin installing the necessary infrastructure in the first year. This includes FPSO I and FPSO II, as well as 3 of the 9 possible field-FPSO connections: field I to FPSO I, field II to FPSO I, and field III to FPSO II. Notice that due to the inherent risk in the size of field I, FPSO I is shared among fields I and II rather than devoting a separate FPSO solely to field I.

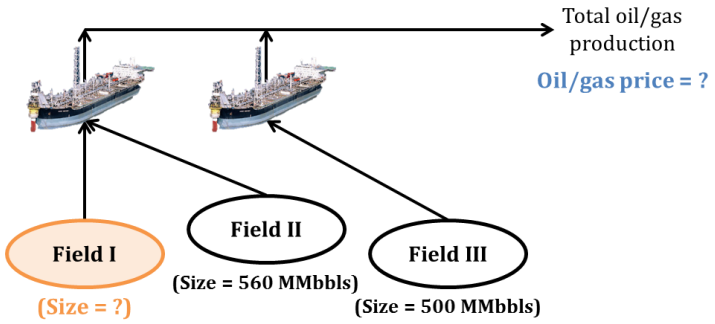


Figure 22. Network structure for the best feasible solution of Example 2. (FPSO images from www.rigzone.com.)

The corresponding drilling schedule is shown in Figure 23. Since it takes 3 years for the FPSOs to be fully operational, drilling cannot begin until the fourth year. For Field II, we drill 10 wells in year 4 and 1 well in year 5. Similarly for Field III, we drill 10 wells in year 4 and 2 wells in year 5. For Field I, however, we wait until year 5 and then drill 7 wells. The strategy here is to drill fields of known size first (as this carries less risk), and then drill the field with an uncertain size. Notice that by the end of the planning horizon, we have drilled the maximum number of wells in all fields.

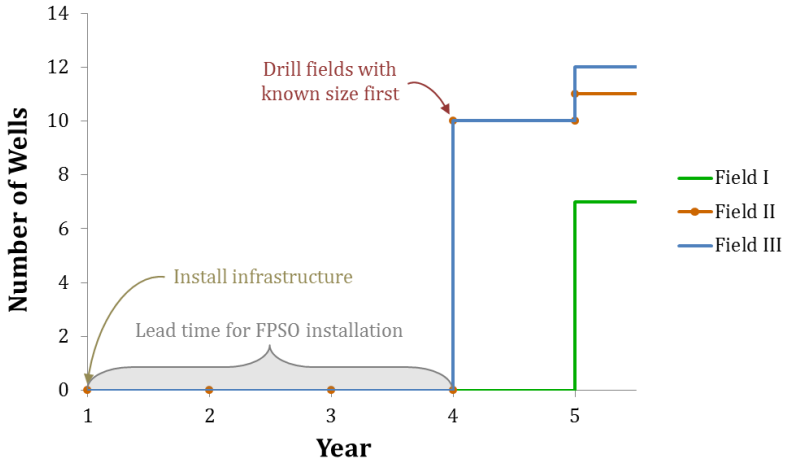


Figure 23. Drilling schedule for the best feasible solution of Example 2.

8 Conclusions

In this paper, we have addressed the general class of multistage stochastic programming problems that involve both endogenous and exogenous uncertain parameters. As little work has been done in this area, we have first provided an extensive review of these two types of uncertainty. Next, we have proposed a superstructure representation for endogenous scenario trees, as well as a composite scenario tree that captures both endogenous and exogenous realizations. This representation serves as the basis for our multistage stochastic programming formulation, model (MSSP).

Using this composite tree, we have also proposed new theoretical properties that can significantly reduce the dimensionality of the model by eliminating all redundant non-anticipativity constraints. Graphical and theoretical proofs have been provided to this effect. The impact of these reduction properties has been demonstrated in a large process network instance, where we have been able to reduce the number of constraints from over 176 million to fewer than 1 million, and the number of binary variables from approximately 4.8 million to 61,000. Put simply, this is the difference between a problem that cannot be loaded into memory and a problem that can be solved efficiently with our alternative solution approaches.

The alternative solution approaches that we have proposed include a novel sequential scenario decomposition heuristic and Lagrangean decomposition. We have applied these techniques to solve two example problems: the capacity expansion of process networks (as previously mentioned), and the development of oilfields. Our numerical results indicate that these solution methods are quite effective at solving problems of this class. In particular, the heuristic can quickly find high-quality feasible solutions, yielding orders-of-magnitude reduction in CPU times. This is especially apparent in the oilfield problem, where the reduced model cannot be solved after more than 11 hours with a direct approach, but can be solved in just 41 seconds with the heuristic.

9 Acknowledgments

We gratefully acknowledge financial support from the Center for Advanced Process Decision-making at Carnegie Mellon University and the ExxonMobil Upstream Research Company.

References

Ahmed, S. (2000). Strategic Planning under Uncertainty: Stochastic integer programming approaches. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Ben-Tal, A., El Ghaoui, L., & Nemirovski, A. (2009). Robust Optimization. New Jersey: Princeton University Press.

Ben-Tal, A., Goryashko, A., Guslitzer, E., & Nemirovski, A. (2004). Adjustable robust solutions of uncertain linear programs. Mathematical Programming, 99 (2, Ser. B), 351-376.

Birge, J. R. (1997). Stochastic Programming Computation and Applications. INFORMS Journal on Computing, 9, 111-133.

Birge, J. R. & Louveaux, F. (2011). Introduction to Stochastic Programming. 2nd Edition. New York, NY: Springer.

Boland, N., Dumitrescu, I., & Froyland, G. (2008). A Multistage Stochastic Programming Approach to Open Pit Mine Production Scheduling with Uncertain Geology. http://www.optimization-online.org/DB_FILE/2008/10/2123.pdf.

Boland, N., Dumitrescu, I., Froyland, G., & Kalinowski, T. (2016). Minimum cardinality non-anticipativity constraint sets for multistage stochastic programming. Mathematical Programming, 157 (1, Ser. B), 69-93.

Bruni, M. E., Beraldi, P., & Conforti, D. (2015). A stochastic programming approach for operating theatre scheduling under uncertainty. IMA Journal of Management Mathematics, 26 (1), 99-119.

Calfa, B. A., Grossmann, I. E., Agarwal, A., Bury, S. J., & Wassick, J. M. (2015). Data-driven individual and joint chance-constrained optimization via kernel smoothing. Computers and Chemical Engineering, 78, 51-69.

Carøe, C. C. & Schultz, R. (1999). Dual decomposition in stochastic integer programming. Operations Research Letters, 24, 37-45.

Choi, J., Realff, M. J., & Lee, J. H. (2004). Dynamic programming in a heuristically confined state space: a stochastic resource-constrained project scheduling application. Computers and Chemical Engineering, 28 (6-7), 1039-1058.

Christian, B. & Cremaschi, S. (2015). Heuristic solution approaches to the pharmaceutical R&D pipeline management problem. Computers and Chemical Engineering, 74, 34-47.

Colvin, M. & Maravelias, C. T. (2008). A stochastic programming approach for clinical trial planning in new drug development. Computers and Chemical Engineering, 32 (11), 2626-2642.

Colvin, M. & Maravelias, C. T. (2009). Scheduling of testing tasks and resource planning in new product development using stochastic programming. Computers and Chemical Engineering, 33 (5), 964-976.

Colvin, M. & Maravelias, C. T. (2010). Modeling methods and a branch and cut algorithm for pharmaceutical clinical trial planning using stochastic programming. European Journal of Operational Research, 203, 205-215.

Colvin, M. & Maravelias, C. T. (2011). R&D pipeline management: Task interdependencies and risk management. European Journal of Operational Research, 215, 616-628.

Dupačová, J. (2006). Optimization under Exogenous and Endogenous Uncertainty. Proceedings of the 24th International Conference on Mathematical Methods in Economics (Ed. Lukáš, L.), 131-136, University of West Bohemia, Pilsen, Czech Republic.

Escudero, L. F., Garín, M. A., Merino, M., & Pérez, G. (2007). The value of the stochastic solution in multistage problems. TOP, 15, 48-64.

Escudero, L. F., Garín, M. A., Merino, M., & Pérez, G. (2009). BFC-MSMIP: an exact branch-and-fix coordination approach for solving multistage stochastic mixed 0-1 problems. TOP, 17, 96-122.

Escudero, L. F., Garín, M. A., Merino, M., & Pérez, G. (2013). On multistage mixed 0-1 optimization under a mixture of exogenous and endogenous uncertainty in a risk averse environment. XIII International Conference on Stochastic Programming, Bergamo, Italy. http://dinamico2.unibg.it/icsp2013/doc/ms/4%20ICSP_escudero.pdf.

Escudero, L. F., Garín, M. A., Pérez, G., & Unzueta, A. (2013). Scenario Cluster Decomposition of the Lagrangian dual in two-stage stochastic mixed 0-1 optimization. Computers & Operations Research, 40, 362-377.

Escudero, L. F., Garín, M. A., & Unzueta, A. (2016). Cluster Lagrangean decomposition in multistage stochastic optimization. Computers & Operations Research, 67, 48-62.

Fisher, M. L. (1985). An Applications Oriented Guide to Lagrangian Relaxation. Interfaces, 15, 10-21.

Flach, B. (2010). Stochastic Programming with Endogenous Uncertainty: An Application in Humanitarian Logistics. Ph.D. thesis, Pontifical Catholic University of Rio de Janeiro.

Goel, V. & Grossmann, I. E. (2004). A stochastic programming approach to planning of offshore gas field developments under uncertainty in reserves. Computers and Chemical Engineering, 28 (8), 1409-1429.

Goel, V. & Grossmann, I. E. (2006). A class of stochastic programs with decision dependent uncertainty. Mathematical Programming, 108 (2-3, Ser. B), 355-394.

Goel, V., Grossmann, I. E., El-Bakry, A. S., & Mulkay, E. L. (2006). A novel branch and bound algorithm for optimal development of gas fields under uncertainty in reserves. Computers and Chemical Engineering, 30, 1076-1092.

Grossmann, I. E., Apap, R. M., Calfa, B. A., García-Herreros, P., & Zhang, Q. (2016). Recent advances in mathematical programming techniques for the optimization of process systems under uncertainty. Computers and Chemical Engineering, 91, 3-14.

Guignard, M. (2003). Lagrangean Relaxation. TOP, 11 (2), 151-228.

Gupta, V. & Grossmann, I. E. (2011). Solution strategies for multistage stochastic programming with endogenous uncertainties. Computers and Chemical Engineering, 35, 2235-2247.

Gupta, V. & Grossmann, I. E. (2014a). A new decomposition algorithm for multistage stochastic programs with endogenous uncertainties. Computers and Chemical Engineering, 62, 62-79.

Gupta, V. & Grossmann, I. E. (2014b). Multistage stochastic programming approach for offshore oilfield infrastructure planning under production sharing agreements and endogenous uncertainties. Computers and Chemical Engineering, 124, 180-197.

Held, H. & Woodruff, D. L. (2005). Heuristics for Multi-Stage Interdiction of Stochastic Networks. Journal of Heuristics, 11 (5), 483-500.

Hellemo, L. (2016). Managing Uncertainty in Design and Operation of Natural Gas Infrastructure. Ph.D. thesis, Norwegian University of Science and Technology.

Hooshmand Khaligh, F. & MirHassani, S. A. (2016). A mathematical model for vehicle routing problem under endogenous uncertainty. International Journal of Production Research, 54 (2), 579-590.

Hooshmand Khaligh, F. & MirHassani, S. A. (2016). Efficient constraint reduction in multistage stochastic programming problems with endogenous uncertainty. Optimization Methods and Software, 31 (2), 359-376.

Jonsbråten, T. W. (1998). Optimization Models for Petroleum Field Exploitation. Ph.D. thesis, Norwegian School of Economics and Business Administration.

Jonsbråten, T. W., Wets, R. J. B., & Woodruff, D. L. (1998). A class of stochastic programs with decision dependent random elements. Annals of Operations Research, 82, 83-106.

Lappas, N. & Gounaris, C. (2016). Multi-Stage Adjustable Robust Optimization for Process Scheduling under Uncertainty. AIChE Journal, 62 (5), 1646-1667.

Laumanns, M., Prestwich, S., & Kawas, B. (2014). Distribution shaping and scenario bundling for stochastic programs with endogenous uncertainty. Stochastic Programming E-Print Series. http://edoc.hu-berlin.de/series/speps/2014-5/PDF/5.pdf.

Li, P., Arellano-Garcia, H., & Wozny, G. (2008). Chance constrained programming approach to process optimization under uncertainty. Computers and Chemical Engineering, 32 (1-2), 25-45.

Liu, M. L., & Sahinidis, N. V. (1996). Optimization in Process Planning under Uncertainty. Industrial and Engineering Chemistry Research, 35 (11), 4154-4165.

Liu, X., Küçükyavuz, S., & Luedtke, J. (2016). Decomposition algorithms for two-stage chance-constrained programs. Mathematical Programming, 157 (1, Ser. B), 219-243.

Maggioni, F., Allevi, E., & Bertocchi, M. (2014). Bounds in Multistage Linear Stochastic Programming. Journal of Optimization Theory and Applications, 163 (1), 200-229.

Mercier, L. & Van Hentenryck, P. (2011). An anytime multistep anticipatory algorithm for online stochastic combinatorial optimization. Annals of Operations Research, 184 (1), 233-271.

Oliveira, F., Gupta, V., Hamacher, S., & Grossmann, I. E. (2013). A Lagrangean decomposition approach for oil supply chain investment planning under uncertainty with risk considerations. Computers and Chemical Engineering, 50, 184-195.

Peeta, S., Salman, F. S., Gunnec, D., Viswanath, K. (2010). Pre-disaster investment decisions for strengthening a highway network. Computers & Operations Research, 37 (10), 1708-1719.

Pflug, G. Ch. (1990). On-Line Optimization of Simulated Markovian Processes. Mathematics of Operations Research, 15 (3), 381-395.

Powell, W. B. (2011). Approximate Dynamic Programming: Solving the Curses of Dimensionality. 2nd Edition. Hoboken, NJ: John Wiley & Sons.

Raman, R. & Grossmann, I. E. (1991). Relation Between MILP Modelling and Logical Inference for Chemical Process Synthesis. Computers and Chemical Engineering, 15 (2), 73-84.

Rockafellar, R. T. & Wets, R. J.-B. (1991). Scenarios and policy aggregation in optimization under uncertainty. Mathematics of Operations Research, 16, 119-147.

Ruszczyński, A. (1997). Decomposition methods in stochastic programming. Mathematical Programming, 79, 333-353.

Sahinidis, N. V. (2004). Optimization under uncertainty: state-of-the-art and opportunities. Computers and Chemical Engineering, 28 (6-7), 971-983.

Schultz, R. (2003). Stochastic programming with integer variables. Mathematical Programming, 97 (1-2, Ser. B), 285-309.

Solak, S., Clarke, J.-P. B., Johnson, E. L., & Barnes, E. R. (2010). Optimization of R&D project portfolios under endogenous uncertainty. European Journal of Operational Research, 207 (1), 420-433.

Tarhan, B. & Grossmann, I. E. (2008). A multistage stochastic programming approach with strategies for uncertainty reduction in the synthesis of process networks with uncertain yields. Computers and Chemical Engineering, 32, 766-788.

Tarhan, B., Grossmann, I. E., & Goel, V. (2009). Stochastic Programming Approach for the Planning of Offshore Oil or Gas Field Infrastructure under Decision-Dependent Uncertainty. Ind. Eng. Chem. Res., 48 (6), 3078-3097.

Tarhan, B., Grossmann, I. E., & Goel, V. (2013). Computational strategies for non-convex multistage MINLP models with decision-dependent uncertainty and gradual uncertainty resolution. Annals of Operations Research, 203 (1), 141-166.

Terrazas-Moreno, S., Grossmann, I. E., Wassick, J. M., Bury, S. J., & Akiya, N. (2012). An efficient method for optimal design of large-scale integrated chemical production sites with endogenous uncertainty. Computers and Chemical Engineering, 37, 89-103.

Tong, K., Feng, Y., & Rong, G. (2012). Planning under Demand and Yield Uncertainties in an Oil Supply Chain. Industrial and Engineering Chemistry Research, 51 (2), 814-834.

Trespalacios, F. & Grossmann, I. E. (2014). Review of Mixed-Integer Nonlinear and Generalized Disjunctive Programming Methods. Chemie Ingenieur Technik, 86 (7), 991-1012.

Vayanos, P., Kuhn, D., & Rustem, B. (2011). Decision rules for information discovery in multi-stage stochastic programming. Proceedings of the 2011 50th IEEE Conference on Decision and Control and European Control Conference, 7368-7373, Orlando, FL, USA.

Viswanath, K., Peeta, S., & Salman, F. S. (2004). Investing in the Links of a Stochastic Network to Minimize Expected Shortest Path Length. Technical report, Purdue University. https://www.krannert.purdue.edu/programs/phd/working-papers-series/2004/1167.pdf.

Webster, M., Santen, N., Parpas, P. (2012). An approximate dynamic programming framework for modeling global climate policy under decision-dependent uncertainty. Computational Management Science, 9, 339-362.

Williams, H. P. (2013). Model Building in Mathematical Programming. 5th Edition. West Sussex, England: John Wiley & Sons Ltd.

Zhang, Q., Lima, R. M., & Grossmann, I. E. (2016). On the Relation between Flexibility Analysis and Robust Optimization for Linear Systems. AIChE Journal, 62 (9), 3109-3123.

A Appendix

A.1 Proof of Property 1

Consider two indistinguishable scenarios $\hat{s}, \hat{s}' \in \mathcal{S}$ in time period τ, where $\hat{s} < \hat{s}'$. For simplicity, consider only variables $y_\tau^{\hat{s}}$ and $y_\tau^{\hat{s}'}$. By Equation (46), we generate two scenario pairs: $(\hat{s}, \hat{s}')$ and $(\hat{s}', \hat{s})$. Scenario pair $(\hat{s}, \hat{s}')$ corresponds to non-anticipativity constraint $y_\tau^{\hat{s}} = y_\tau^{\hat{s}'}$. Scenario pair $(\hat{s}', \hat{s})$ corresponds to non-anticipativity constraint $y_\tau^{\hat{s}'} = y_\tau^{\hat{s}}$, which is the same equality constraint. By symmetry, we may replace the condition $s \neq s'$ in Equation (46) with $s < s'$. In this case, we only generate the first pair, and we avoid the second, redundant constraint. ■

A.2 Proof of Property 2a

Because *all* scenarios are indistinguishable at the beginning of the first time period, *adjacent* scenarios must also be indistinguishable at that time. Thus, we can enforce non-anticipativity between all scenarios by linking consecutive nodes; e.g., $y_1^1 = y_1^2,\ y_1^2 = y_1^3, \ldots, y_1^{S-1} = y_1^S$. ■

A.3 Proof of Proposition 1

We generate set $\mathcal{SP}_F$ by pairing off all S scenarios in consecutive order. This gives $S - 1$ independent links (i.e., scenario pairs), which is the minimum number of links required to connect S elements. ■

A.4 Proof of Property 2b

As previously stated, exogenous NACs apply only between scenarios s and s' in the same subtree. Because each subtree represents an exogenous scenario tree, non-anticipativity constraints within that tree apply as if the uncertainty were purely exogenous. (Note that in the case of purely-exogenous uncertainty, the adjacent-scenario approach to non-anticipativity is well known (see, for example, Colvin and Maravelias (2011)). However, we provide the rest of the proof for completeness.)

Accordingly, consider an exogenous scenario tree in its standard form, as shown in Figure 2a. For each time period $t \in \mathcal{T},\ t < T$, each scenario passes through a node that is shared among one or more scenarios. All scenarios that pass through one such node at time t must be indexed consecutively, since they all refer to the same path up until this time (i.e., they have the same history). When we duplicate this node to give each scenario its own respective copy, we create consecutive, indistinguishable nodes that refer to the same state and must be linked together with non-anticipativity constraints (see Figure 2b). One natural approach to enforce non-anticipativity between these indistinguishable scenarios in time period t is to link them together in consecutive order. ■

A.5 Proof of Proposition 2

In each time period, excluding $t = T$, we partition the set of scenarios into exogenous scenario-group subsets ${}^{X}\mathcal{G}_t^k \ \forall\, k \in \mathcal{K}_t$. Since each scenario must be assigned to one group (scenario 1 to group 1, and all others by Equation (50)), the union of all such groups in each time period must give the complete set of scenarios; i.e.,

$$\bigcup_{k \in \mathcal{K}_t} {}^{X}\mathcal{G}_t^k = \mathcal{S} \quad \forall\, t \in \mathcal{T}\backslash\{T\}$$

Thus, we are considering all scenarios in each time period where exogenous NACs apply.

We enforce non-anticipativity between consecutive scenarios in each of the exogenous scenario groups. By Equation (22), the corresponding scenario pairs for $t \in \mathcal{T}\backslash\{T\}$ must be in set $\mathcal{SP}_X$ because they are adjacent (i.e., $(s, s') \in \mathcal{A}$), in the same subtree (i.e., $Sub(s) = Sub(s')$), and indistinguishable (i.e., $Q_t^{s,s'} = True$). Since each group has different realizations for the exogenous parameters and/or different possible realizations for the endogenous parameters, no links between the groups are possible, and such pairs cannot be in set $\mathcal{SP}_X$. Thus, the pairs in each group are the *only* possible exogenous scenario pairs in each time period. It follows that the union of these sets of tuples must be equivalent to set $\mathcal{SP}_X$:

$$\bigcup_{t \in \mathcal{T}\backslash\{T\}} \left[\bigcup_{k \in \mathcal{K}_t} \left\{ (t, s, s'):\ s, s' \in {}^{X}\mathcal{G}_t^k,\ (s, s') \in \mathcal{A} \right\} \right] = \mathcal{SP}_X$$

Now, consider the scenario pairs in each exogenous scenario group in time period $t \in \mathcal{T}\backslash\{T\}$. Because we link consecutive scenarios, this gives $\left| {}^{X}\mathcal{G}_t^k \right| - 1$ scenario pairs in each group, which is the minimum number of links required to connect $\left| {}^{X}\mathcal{G}_t^k \right|$ elements. These pairs cannot be implied through the use of any endogenous scenario pairs, since we generate the exogenous pairs first. Thus, we have the minimum

number of scenario pairs in each group. We have shown that these are the only possible exogenous scenario pairs in each time period, and the union of these sets of tuples is equivalent to set $\mathcal{SP}_X$. Hence, set $\mathcal{SP}_X$ contains the minimum number of exogenous scenario pairs. ■

A.6 Proof of Property 4

By Property 3, endogenous NACs are expressed between scenarios s and s' that differ in the possible realization of a single endogenous parameter $\theta_{i',h}$ and are identical in the realizations of all exogenous parameters in all time periods. Accordingly, for each $i' \in \mathcal{I}$ and $h \in \mathcal{H}_{i'}$, we seek to identify the minimum number of scenario pairs (s, s') that satisfy these conditions.

To this end, in an arbitrary time period $t = \tau$, we partition the set of scenarios into endogenous scenario-group subsets. These subsets are given by ${}^{N}\mathcal{G}^{l}_{i',h}$ and are indexed by $l \in \mathcal{L}_{i',h}$ for each $i' \in \mathcal{I}$ and $h \in \mathcal{H}_{i'}$. By the endogenous scenario-group algorithm, each scenario must be assigned to one such group for each $i' \in \mathcal{I}$ and $h \in \mathcal{H}_{i'}$. In other words, the union of all of these groups must give the complete set of scenarios; i.e.,

$$\bigcup_{l \in \mathcal{L}_{i',h}} {}^{N}\mathcal{G}^{l}_{i',h} = \mathcal{S} \quad \forall\, i' \in \mathcal{I},\ h \in \mathcal{H}_{i'}$$

Thus, we are considering all scenarios in each case where endogenous NACs may apply.

We enforce non-anticipativity between consecutive scenarios in each of the endogenous scenario groups, as indicated by Equation (60). This gives $\left|{}^{N}\mathcal{G}^{l}_{i',h}\right| - 1$ scenario pairs in each group, which is the minimum number of links required to connect $\left|{}^{N}\mathcal{G}^{l}_{i',h}\right|$ elements. Other connections between the scenarios are implied by transitivity.

Furthermore, by Property 3, it is sufficient to consider *only* the pairs formed in each endogenous scenario group. This is because there are no links between groups, other than those that already exist in another group. We can prove this by contradiction.

First, suppose that we have a link between two scenarios, $\hat{s}$ and $\hat{s}'$. By Property 3, these scenarios must differ only in the possible realization of a single endogenous parameter $\theta_{\hat{\imath},\hat{h}}$. Second, assume that this link cannot be formed by pairing two scenarios in the same endogenous scenario group. In other words, these scenarios belong to two separate groups corresponding to parameter $\theta_{\hat{\imath},\hat{h}}$, and by the endogenous scenario-group algorithm, we have $G_N(\hat{\imath}, \hat{h}, \hat{s}) = \hat{l}$ and $G_N(\hat{\imath}, \hat{h}, \hat{s}') = \tilde{l}$. It follows that the respective groups from Equation (59) are ${}^{N}\mathcal{G}^{\hat{l}}_{\hat{\imath},\hat{h}}$ and ${}^{N}\mathcal{G}^{\tilde{l}}_{\hat{\imath},\hat{h}}$. Note that the scenarios in one group must differ from the scenarios in the other group in terms of the possible realization of *at least one* uncertain parameter other than $\theta_{\hat{\imath},\hat{h}}$. (If this were not the case, then ${}^{N}\mathcal{G}^{\hat{l}}_{\hat{\imath},\hat{h}}$ and ${}^{N}\mathcal{G}^{\tilde{l}}_{\hat{\imath},\hat{h}}$ would be a single group.) Since $\hat{s}$ and $\hat{s}'$ differ in the possible realization of $\theta_{\hat{\imath},\hat{h}}$ *and* belong to two separate groups, these scenarios must differ in the possible realizations of *at least two* uncertain parameters. This violates Property 3. Thus, the original

assumption is false and any endogenous scenario pair must be formed between two scenarios in the same endogenous scenario group.

At this point, we have shown that if we consider only scenario pairs (s, s') for which s and s' are consecutive scenarios in an endogenous scenario group: (1) we are able to link all scenarios in each group; and (2) from this linking, we are able to produce all endogenous scenario pairs generated by Property 3 (either explicitly, or implicitly through the use of some explicitly-generated pairs). Note that since the endogenous scenario groups are defined in terms of Property 3, Property 4 must also be *at least* as restrictive as Property 3. In other words, $\mathcal{SP}_{N^4} \subseteq \mathcal{SP}_{N^3}$, and this approach cannot produce any additional scenario pairs that cannot be obtained from Property 3. It follows that Property 4 is a sufficient condition for endogenous scenario-pair generation. ■

A.7 Proof of Proposition 3

This is the case described in Gupta and Grossmann (2011). We approach this proof from a different angle and continue from the proof of Property 4.

So far, we have shown that Property 4 gives the minimum number of pairs among the scenarios in each endogenous scenario group, and that it is sufficient to consider only these pairs. Recall that by Equation (60), we generate one such set of pairs for each group. The complete set of endogenous scenario pairs is given by the union of these sets, as defined in Equation (61). To prove that this resulting set contains the minimum number of pairs, it is only necessary for us to show that the pairs in each group cannot be implied by any other pairs.

First, in the general case, we cannot guarantee that any uncertain parameters will be realized at the same time, since we have assumed that each parameter is associated with a different source. Second, we cannot guarantee that any of the parameters will be unrealized in certain time periods, either, since we have also assumed that there are no initial 'equality' periods. And third, we have assumed that there is no exogenous uncertainty, so there are no exogenous NACs. Thus, all of the endogenous NACs must be applied conditionally, and we cannot use any one to imply any of the others. This case can be seen clearly in Figure 5. It follows that, under these strict assumptions, the complete set of endogenous scenario pairs generated by Property 4, $\mathcal{SP}_{N^4}$, contains the minimum number of pairs. ■

A.8 Proof of Property 5

Consider exogenous scenario group $\hat{k}$ in time period $t = \tau$, where $\tau < T$, which corresponds to a set of scenarios given by ${}^{X}\mathcal{G}_{\tau}^{\hat{k}}$. These scenarios are adjacent, so we may express this set in the general form ${}^{X}\mathcal{G}_{\tau}^{\hat{k}} = \{s:\ s = n, n+1, \dots, N\}$. It follows that the exogenous non-anticipativity constraints between the scenarios in group ${}^{X}\mathcal{G}_{\tau}^{\hat{k}}$ are:

$$y_{\tau}^{n} = y_{\tau}^{n+1},\ y_{\tau}^{n+1} = y_{\tau}^{n+2}, \dots,\ y_{\tau}^{N-1} = y_{\tau}^{N} \qquad \textbf{(79)}$$

where, for simplicity, we consider only variables y_{τ}^{s}. Note that these scenarios are in the same subtree. Let ${}^{X}\mathcal{G}_{\tau}^{\tilde{k}}$ be the corresponding group of scenarios in a different subtree; i.e., all scenarios in the same position in ${}^{X}\mathcal{G}_{\tau}^{\hat{k}}$ and ${}^{X}\mathcal{G}_{\tau}^{\tilde{k}}$ are identical in the realizations of all exogenous parameters but differ in the

possible realization of at least one endogenous parameter. Without loss of generality, we assume that the respective scenarios differ in the possible realization of exactly one endogenous parameter. We may express this set in the general form ${}^{X}\mathcal{G}_{\tau}^{\tilde{k}} = \{s\colon\ s = n^*, n^*+1, \dots, N^*\}$, where $Pos(n) = Pos(n^*)$, $Pos(n+1) = Pos(n^*+1), \dots,\ Pos(N) = Pos(N^*)$. Exogenous NACs between the scenarios in group ${}^{X}\mathcal{G}_{\tau}^{\tilde{k}}$ can be written in the same form as for ${}^{X}\mathcal{G}_{\tau}^{\hat{k}}$:

$$y_{\tau}^{n^*} = y_{\tau}^{n^*+1},\ y_{\tau}^{n^*+1} = y_{\tau}^{n^*+2}, \dots,\ y_{\tau}^{N^*-1} = y_{\tau}^{N^*} \tag{80}$$

Endogenous non-anticipativity constraints apply only between scenarios s and s' in different subtrees, and by Property 3, it is sufficient to consider only s and s' that are identical in all exogenous realizations (i.e., $Pos(s) = Pos(s')$, where $s < s'$). Thus, the endogenous NACs between scenarios $s \in {}^{X}\mathcal{G}_{\tau}^{\hat{k}}$ and $s' \in {}^{X}\mathcal{G}_{\tau}^{\tilde{k}}$ are then:

$$y_{\tau}^{n} = y_{\tau}^{n^*},\ y_{\tau}^{n+1} = y_{\tau}^{n^*+1}, \dots,\ y_{\tau}^{N} = y_{\tau}^{N^*} \tag{81}$$

provided that the scenarios are indistinguishable (i.e., $Z_{\tau}^{s,s'} = True$). Recall that s and s' differ in the possible realization of the same endogenous parameter, and only this one parameter, so the corresponding NACs between these scenarios are all active at the same time or are all ignored at the same time. Furthermore, if $Z_{\tau}^{s,s'} = False$, these non-anticipativity constraints do not apply, so it is only necessary to consider the case where these constraints are active.

Since $y_{\tau}^{n} = y_{\tau}^{n+1}$ by Equation (79), and $y_{\tau}^{n^*} = y_{\tau}^{n^*+1}$ by Equation (80), it follows that the first endogenous constraint in Equation (81), $y_{\tau}^{n} = y_{\tau}^{n^*}$, can be restated as $y_{\tau}^{n+1} = y_{\tau}^{n^*+1}$. Notice that this is the second endogenous constraint in Equation (81). This procedure can be continued to produce all of the remaining endogenous constraints in Equation (81). This shows that the exogenous NACs for two groups, along with *one* endogenous NAC linking one scenario from each group, imply all of the other endogenous NACs linking the two groups. Thus, only one endogenous NAC between the groups is sufficient. ■

A.9 Proof of Proposition 4

Starting from Proposition 3, notice that we have relaxed only one assumption; namely, that the problem is purely endogenous. To prove that we have the minimum number of endogenous scenario pairs, it is merely necessary for us to show that after the introduction of exogenous uncertainty, and the application of Property 5, the pairs in each endogenous scenario group cannot be implied by any other pairs.

First, recall that when both endogenous and exogenous uncertain parameters are present in the model, some of the endogenous scenario pairs can be implied by exogenous pairs. All such redundant pairs are eliminated by Property 5.

We may then rely on the remaining arguments in the proof of Proposition 3 to conclude that all endogenous NACs must be applied conditionally, and we cannot use any one to imply any of the others. It follows that, under the stated assumptions, the complete set of endogenous scenario pairs generated by Property 4 and Property 5, $\mathcal{SP}_{N^5}$, contains the minimum number of pairs. ■

A.10 Unique Scenarios Algorithm

For convenience in the algorithm, we partition the set of sources $\mathcal{I}$ into ordered sets $\mathcal{I}_E^t$ and $\mathcal{I}_C^t$ based on the given time period (not to be confused with set $\overline{\mathcal{I}}_t^s$, which was introduced for illustrative purposes in section 2.2). Specifically, in the initial 'equality' time periods $t \in \mathcal{T}_E^{i'}$, the endogenous uncertainty cannot yet be resolved for sources $i' \in \mathcal{I}_E^t$, where $\mathcal{I}_E^t \coloneqq \left\{i' \colon\ i' \in \mathcal{I},\ \ t \in \mathcal{T}_E^{i'}\right\}\ \forall\, t \in \mathcal{T}$. These sets of sources are associated with fixed endogenous NACs (since all uncertain parameters associated with these sources are guaranteed to be unresolved at time t, and thus all scenarios that differ only in the possible realizations of any of these parameters must be indistinguishable at that time). Note that we will always have $\mathcal{I}_E^T \coloneqq \emptyset$, as the initial 'equality' periods should never span the entire time horizon. We do not restrict the definition of $\mathcal{I}_E^t$ to $t \in \mathcal{T}\backslash\{T\}$, however, as this would require us to treat $t < T$ and $t = T$ as two separate cases in the algorithm.

In the remaining 'conditional' time periods $t \in \mathcal{T}_C^{i'}$, the endogenous uncertainty may be resolved for sources $i' \in \mathcal{I}_C^t$, where $\mathcal{I}_C^t \coloneqq \left\{i' \colon\ i' \in \mathcal{I},\ \ t \in \mathcal{T}_C^{i'}\right\}\ \forall\, t \in \mathcal{T}$. These sets of sources are associated with conditional endogenous NACs (since the uncertain parameters associated with these sources are no longer guaranteed to be unresolved at time t, and thus we can only say that the scenarios that differ in the possible realizations of any of these parameters *may* be indistinguishable at that time).

We now present the unique scenarios algorithm, in which we define sets $\mathcal{U}_t^{i',h}$ corresponding to each endogenous parameter $\theta_{i',h}$, for all $i' \in \mathcal{I}$ and $h \in \mathcal{H}_{i'}$, in each time period $t \in \mathcal{T}$.

Unique Scenarios Algorithm

Step 1: For each time period $t \in \mathcal{T}$:

Step 1a: First, consider the sources that are associated with fixed endogenous NACs in this time period. In other words, for each source $i' \in \mathcal{I}_E^t$ (where the sources are considered in ascending numerical order):

i) If i' is the first source in this set (i.e., $i' = \min_{\hat{\imath}'}\left(\hat{\imath}' \in \mathcal{I}_E^t\right)$), initialize the set of unique scenarios to that obtained from Property 5 (e.g., Equation (62)) in order to take advantage of the reductions associated with exogenous scenario grouping:

$$\mathcal{U}_t^{i',1} \coloneqq \widetilde{\mathcal{U}}_t \quad i' = \min_{\hat{\imath}'}\left(\hat{\imath}' \in \mathcal{I}_E^t\right) \tag{82}$$

ii) If there is more than one endogenous parameter associated with source i', define the corresponding set of unique scenarios that can be considered for each of these parameters, as indicated in Equation (83):

$$\mathcal{U}_t^{i',h+1} \coloneqq \mathcal{U}_t^{i',h} \cap \left[\bigcup_{l \in \mathcal{L}_{i',h}} \left\{\min_{\hat{s}}\left(\hat{s} \in {}^{N}\mathcal{G}_{i',h}^{l}\right)\right\}\right] \qquad h = 1, \dots, H_{i'} - 1 \tag{83}$$

iii) If $i' < \max_{\hat{\imath}'}\left(\hat{\imath}' \in \mathcal{I}_E^t\right)$, there is at least one additional source to consider. Accordingly, define the set of unique scenarios for the first endogenous parameter of the *next* source, i'', as indicated in Equation (84):

$$\mathcal{U}_t^{i'',1} \coloneqq \mathcal{U}_t^{i',H_{i'}} \cap \left[\bigcup_{l \in \mathcal{L}_{i',H_{i'}}} \left\{ \min_{\hat{s}} \left(\hat{s} \in {}^{N}\mathcal{G}_{i',H_{i'}}^{l} \right) \right\} \right] \quad i'' = \min_{\hat{\imath}''} \left(\hat{\imath}'' \in \mathcal{I}_E^t,\ \hat{\imath}'' > i' \right) \tag{84}$$

iv) If $i' = \max_{\hat{\imath}'} \left(\hat{\imath}' \in \mathcal{I}_E^t \right)$, this is the last source. Store the current set of unique scenarios in a separate, temporary set, UniqueSet, but in the same manner as Equation (84):

$$\text{UniqueSet} \coloneqq \mathcal{U}_t^{i',H_{i'}} \cap \left[\bigcup_{l \in \mathcal{L}_{i',H_{i'}}} \left\{ \min_{\hat{s}} \left(\hat{s} \in {}^{N}\mathcal{G}_{i',H_{i'}}^{l} \right) \right\} \right] \tag{85}$$

Step 1b: Next, consider the sources that are associated with conditional endogenous NACs in this time period. In other words, for each source $i' \in \mathcal{I}_C^t$ (where the sources are considered in ascending numerical order):

i) If i' is the first source in this set (i.e., $i' = \min_{\hat{\imath}'} \left(\hat{\imath}' \in \mathcal{I}_C^t \right)$), initialize the set of unique scenarios based on the following two conditions:

(1) If $\mathcal{I}_E^t \neq \emptyset$, then t is an initial 'equality' time period for at least one of the sources. Accordingly, initialize the set of unique scenarios to the last-known value, stored in Equation (85), in order to take advantage of the reductions associated with the fixed endogenous NACs:

$$\mathcal{U}_t^{i',1} \coloneqq \text{UniqueSet} \quad i' = \min_{\hat{\imath}'} \left(\hat{\imath}' \in \mathcal{I}_C^t \right) \tag{86}$$

(2) If, however, $\mathcal{I}_E^t = \emptyset$, then there are no fixed endogenous NACs in time period t. Similar to sub-step (i) of Step 1a, initialize the set of unique scenarios to that obtained from Property 5:

$$\mathcal{U}_t^{i',1} \coloneqq \widetilde{\mathcal{U}}_t \quad i' = \min_{\hat{\imath}'} \left(\hat{\imath}' \in \mathcal{I}_C^t \right) \tag{87}$$

In the case where $t = T$, recall that $\widetilde{\mathcal{U}}_T \coloneqq \mathcal{S}$.

ii) Execute sub-step (ii) of Step 1a.

iii) If $i' < \max_{\hat{\imath}'} \left(\hat{\imath}' \in \mathcal{I}_C^t \right)$, define the set of unique scenarios for the first endogenous parameter of the next source, i'', as indicated in Equation (88):

$$\mathcal{U}_t^{i'',1} \coloneqq \mathcal{U}_t^{i',1} \quad i'' = \min_{\hat{\imath}''} \left(\hat{\imath}'' \in \mathcal{I}_C^t,\ \hat{\imath}'' > i' \right) \tag{88}$$

Notice that Step 1a will automatically be skipped for $t = T$, since $\mathcal{I}_E^T = \emptyset$.

Because this algorithm is fairly complex, we next discuss some of the respective expressions in further detail.

Equation (83) addresses the case in which we have multiple endogenous parameters associated with some of the sources of uncertainty. Therefore, this expression updates the set of unique scenarios in time period t when advancing from one endogenous parameter h to the next, $h+1$, for a given source i'. The assumption here is that *all* endogenous parameters associated with the same source must be realized at the same time. This is because an investment *in the source itself* determines the time at which the associated technical information can be realized. For example, once we drill an oilfield and begin producing from it, we assume that we can determine both the size and initial deliverability of the reserves.

Accordingly, for any scenarios s and s' that differ only in the possible realization of an endogenous parameter associated with source i', the corresponding non-anticipativity constraints will all apply at the same time or will all be ignored at the same time. This was previously shown in the discussion surrounding Figure 14. Because it is then sufficient to consider only the case where these constraints are active, when we proceed from the first endogenous parameter of source i' to the second (i.e., $h=1$ to $h=2$), we begin to pair off scenarios between the groups of $\theta_{i',1}$ (we prove this point in Appendix section A.11), and the scenarios in each of those respective groups must be indistinguishable. It is therefore unnecessary to have more than one link between any such groups. The reasoning here is justified in Appendix section A.11.

Thus, in time period t, when advancing from endogenous parameter $\theta_{i',h}$ to the next parameter of the same source, $\theta_{i',h+1}$, we require *at most* only a single 'representative' scenario from each endogenous scenario group corresponding to $\theta_{i',h}$; i.e., $\bigcup_{l\in\mathcal{L}_{i',h}}\left\{\min_{\hat{s}}\left(\hat{s}\in {}^{N}\mathcal{G}^{l}_{i',h}\right)\right\}$. Notice the similarity of this case to our treatment of the exogenous scenario groups in Property 5 (see Equation (62)). We say "*at most*" since some of these representative scenarios may be non-unique based on Property 5 and/or the consideration of other endogenous parameters *before this point* in the unique scenarios algorithm. We successively remove non-unique scenarios in time period t by taking the intersection of the current set of unique scenarios, $\mathcal{U}_t^{i',h}$, and the set of representative scenarios, $\bigcup_{l\in\mathcal{L}_{i',h}}\left\{\min_{\hat{s}}\left(\hat{s}\in {}^{N}\mathcal{G}^{l}_{i',h}\right)\right\}$. The result, as shown in Equation (83), is an *updated* set of unique scenarios that can be considered for the next parameter in the algorithm (i.e., $\theta_{i',h+1}$). It is important to note that this case is evaluated in the same way in all time periods and appears in sub-step (ii) of Step 1a and Step 1b of the algorithm.

Equation (84) addresses the case in which there are endogenous parameters that cannot be realized in some of the initial time periods. Rather than defining an updated set of unique scenarios in time period t for each endogenous parameter of the *same* source i', as in Equation (83), this expression considers the case of advancing from the last endogenous parameter, $H_{i'}$, of source i', to the first endogenous parameter of the next source, i''. The reasoning here is that if the uncertainty in source i' cannot yet be revealed as of time period t (i.e., $i'\in\mathcal{I}_E^t$), then we will have equality constraints corresponding to all scenario pairs (s,s') for which s and s' differ in the possible realization of a parameter associated with i'. This implies that there will then be redundant constraints associated with the parameters of the *next* source, i'', as illustrated in Figure 13. We can thus perform further reduction via the same strategy used in Equation (83) for the case of multiple parameters associated with the same source. Notice, in particular, that the form of Equation (84) is nearly identical to Equation (83).

One subtle difference here is that the sources in set $\mathcal{I}_E^t$ may be nonconsecutively indexed, which means that we cannot use index $i' + 1$ to access the next element in the set (as we do with $h + 1$ to access the next parameter in Equation (83)). Instead, we use a strategy similar to that previously introduced in Equation (60) and state $i'' = \min_{\hat{\imath}''}\left(\hat{\imath}'' \in \mathcal{I}_E^t,\ \hat{\imath}'' > i'\right)$. This expression simply allows us to advance from one source, i', to the next-lowest-indexed source, i'', in an ordered manner.

Equation (85) is a special case of Equation (84) that is evaluated only for the last source in set $\mathcal{I}_E^t$. The corresponding set of unique scenarios is stored in a temporary set, UniqueSet, which may be used for initialization in sub-step (i) of Step 1b.

In Step 1b, notice that time period t is not an initial 'equality' period for source i'. Here we consider sources $i' \in \mathcal{I}_C^t$, and we can no longer guarantee that we will have equality constraints corresponding to the scenario pairs (s, s') for which s and s' differ in the possible realization of a parameter associated with i'. It follows that since all of these constraints are conditional, the only reduction that we can perform is for multiple parameters associated with the same source i' (via sub-step (ii)); we cannot make any further assumptions to eliminate constraints corresponding to the next source, i''. We thus use Equation (88) in which the definition of the set of unique scenarios is unchanged from one source to the next. We state $\mathcal{U}_t^{i'',1} \coloneqq \mathcal{U}_t^{i',1}$ in this equation, rather than $\mathcal{U}_t^{i'',1} \coloneqq \mathcal{U}_t^{i',H_{i'}}$, because the reduction from sub-step (ii) cannot be carried over to the next source as it does in Step 1a. Also, note that the sources in set $\mathcal{I}_C^t$ may be nonconsecutively indexed, so like the treatment of set $\mathcal{I}_E^t$ in Equation (84), we use $i'' = \min_{\hat{\imath}''}\left(\hat{\imath}'' \in \mathcal{I}_C^t,\ \hat{\imath}'' > i'\right)$ to access the next-lowest-indexed source.

As a brief example of how this algorithm is applied, consider Figure 13 and assume that only these 4 scenarios are under consideration. We start at $t = 1$. Since there is a single endogenous parameter associated with each of the two sources, we will drop the h index to simplify the notation.

It is clear that we are starting with unique scenarios $\widetilde{\mathcal{U}}_1 = \{1, 5, 9, 13\}$ from Property 5. It is also clear that we have endogenous scenario groups ${}^N\mathcal{G}_1^1 = \{1, 9\}$ and ${}^N\mathcal{G}_1^2 = \{5, 13\}$ corresponding to θ_1, and ${}^N\mathcal{G}_2^1 = \{1, 5\}$ and ${}^N\mathcal{G}_2^2 = \{9, 13\}$ corresponding to θ_2. Notice that $t = 1$ is an initial 'equality' period only for θ_2 (i.e., $\mathcal{T}_E^1 = \emptyset$ and $\mathcal{T}_E^2 = \{1\}$), so $\mathcal{I}_E^1 = \{2\}$ and $\mathcal{I}_C^1 = \{1\}$.

We start with Step 1a for the sources associated with fixed endogenous NACs in the first time period and must consider $i' \in \{2\}$. We first initialize the corresponding set of unique scenarios (i.e., $\mathcal{U}_1^2$) in sub-step (i), as indicated in Equation (82): $\mathcal{U}_1^2 \coloneqq \widetilde{\mathcal{U}}_1 = \{1, 5, 9, 13\}$.

Notice that because there is only one endogenous parameter associated with source 2, we skip sub-step (ii). Since $i' = \max_{\hat{\imath}'}(\hat{\imath}' \in \{2\}) = 2$, we also skip sub-step (iii) and proceed to (iv). This yields the following, by Equation (85):

$$\text{UniqueSet} \coloneqq \mathcal{U}_1^2 \cap \left[\bigcup_{l \in \mathcal{L}_2} \left\{\min_{\hat{s}}\left(\hat{s} \in {}^N\mathcal{G}_2^l\right)\right\}\right] = \{1, 5, 9, 13\} \cap \left[\left\{\min_{\hat{s}}(\hat{s} \in \{1, 5\})\right\} \cup \left\{\min_{\hat{s}}(\hat{s} \in \{9, 13\})\right\}\right]$$

which simplifies to UniqueSet $\coloneqq \{1,5,9,13\} \cap \{1,9\} = \{1,9\}$.

We then continue to Step 1b for the sources associated with conditional endogenous NACs in the first time period. Here, we must consider $i' \in \{1\}$. We first initialize the set of unique scenarios to the last-known value, given by the temporary set UniqueSet, in sub-step (i), condition (1). Specifically, by Equation (86), $\mathcal{U}_1^1 \coloneqq \{1,9\}$.

Because there is only one endogenous parameter associated with source 1, we skip sub-step (ii). We also skip sub-step (iii) since $i' = \max_{i'}(i' \in \{1\}) = 1$. In practice, we would then continue to $t = 2$. To summarize, $\mathcal{U}_1^1 \coloneqq \{1,9\}$ and $\mathcal{U}_1^2 \coloneqq \{1,5,9,13\}$.

Equation (33) requires us to form pairs among consecutive scenarios in sets ${}^{N}\mathcal{G}_1^1 \cap \mathcal{U}_1^1$, ${}^{N}\mathcal{G}_1^2 \cap \mathcal{U}_1^1$, ${}^{N}\mathcal{G}_2^1 \cap \mathcal{U}_1^2$, and ${}^{N}\mathcal{G}_2^2 \cap \mathcal{U}_1^2$ for $t = 1$. These sets are given by $\{1,9\} \cap \{1,9\}$, $\{5,13\} \cap \{1,9\}$, $\{1,5\} \cap \{1,5,9,13\}$, and $\{9,13\} \cap \{1,5,9,13\}$, respectively, which reduce to $\{1,9\}$, $\emptyset$, $\{1,5\}$, and $\{9,13\}$, respectively. Pairing off consecutive scenarios in these sets yields pairs $(1,9)$, $(1,5)$, and $(9,13)$, as shown in Figure 13. Notice that none of the remaining pairs can be implied by any of the others.

It is worth noting that repeating this procedure for a case such as Figure 14 will lead to slightly different scenario pairs than pictured. This is due to the order in which we consider the endogenous parameters. Specifically, by considering pairs among scenarios that differ in the possible realization of $\theta_{i,2}$ first, and then those for $\theta_{i,1}$, we obtain scenario pairs $(\hat{s}, \hat{s}')$, $(\hat{s}, \hat{s}'')$, and $(\hat{s}'', \hat{s}''')$ in Figure 14. By the unique scenarios algorithm, however, we consider the parameters in numerical order (i.e., $\theta_{i,1}$, followed by $\theta_{i,2}$) and instead obtain $(\hat{s}, \hat{s}')$, $(\hat{s}, \hat{s}'')$, and $(\hat{s}', \hat{s}''')$. Although the first set may appear to be more "natural" based on the appearance of Figure 14, both sets are equally valid since only 3 pairs are required to link the 4 scenarios.

A.11 Proof of Property 6

Consider the endogenous scenario groups $l = 1, 2, \ldots, \left|\mathcal{L}_{i,\hat{h}}\right|$ corresponding to endogenous parameter $\theta_{i,\hat{h}}$. The scenarios in each of these respective group differ *only* in the possible realization of $\theta_{i,\hat{h}}$. Given that there are $M_{i,\hat{h}}$ (or $\left|\Theta_{i,\hat{h}}\right|$) possible realizations for $\theta_{i,\hat{h}}$, and in each respective group, each scenario must have a different possible realization for this endogenous parameter, it follows that there can only be $M_{i,\hat{h}}$ scenarios in each of these groups (i.e., one for each possible realization of $\theta_{i,\hat{h}}$).

Furthermore, the lowest-indexed scenario in each of these groups must have the lowest realization for $\theta_{i,\hat{h}}$. This is simply a consequence of the ordering on the set of realizations $\Theta_{i,\hat{h}}$ (i.e., $\hat{\theta}_{i,\hat{h}}^1 < \hat{\theta}_{i,\hat{h}}^2 < \cdots < \hat{\theta}_{i,\hat{h}}^{M_{i,\hat{h}}}$) and the lexicographical ordering on the Cartesian products used in the scenario-generation process (see section 3). Specifically, as can be seen in Equation (5) and even more clearly in Figure 6c, we must exhaust all possible combinations of realizations for the uncertain parameters that occur *after* $\theta_{i,\hat{h}}$ in the Cartesian product before the realization of $\theta_{i,\hat{h}}$ can be incremented to the next possible value. This means that a scenario $\hat{s}$ defined with a low realization for $\theta_{i,\hat{h}}$ will come before a scenario $\hat{s}'$ with a high realization for $\theta_{i,\hat{h}}$ and all of the same possible realizations for the other uncertain parameters.

Accordingly, in an endogenous scenario group, it follows that the lowest-indexed scenario must have the lowest realization for $\theta_{\hat{\imath},\hat{h}}$, the next-lowest-indexed scenario must have the next-lowest realization for $\theta_{\hat{\imath},\hat{h}}$, and so forth, until we reach the highest-indexed scenario in the group, which must have the highest realization for $\theta_{\hat{\imath},\hat{h}}$. For example, if $\theta_{\hat{\imath},\hat{h}}$ were defined with 3 possible realizations (*low* (L), *medium* (M), or *high* (H)), there would be 3 scenarios in each of the corresponding (ordered) endogenous scenario groups, with realizations of the following form: $(\dots, \hat{\theta}^{L}_{\hat{\imath},\hat{h}}, \dots), (\dots, \hat{\theta}^{M}_{\hat{\imath},\hat{h}}, \dots), (\dots, \hat{\theta}^{H}_{\hat{\imath},\hat{h}}, \dots)$. Note that this is the case depicted in Figure 10 for Property 4.

Consider two scenarios $\hat{s}$ and $\hat{s}'$ from one arbitrary group $\hat{l}$ corresponding to $\theta_{\hat{\imath},\hat{h}}$ (i.e., ${}^{N}\mathcal{G}^{\hat{l}}_{\hat{\imath},\hat{h}}$). Recall that this means that $\hat{s}$ and $\hat{s}'$ must have the same possible realizations for all uncertain parameters except $\theta_{\hat{\imath},\hat{h}}$. Because *all* scenarios must be placed in an endogenous scenario group corresponding to each endogenous parameter (by the endogenous scenario-group algorithm), both of these scenarios will also be placed into groups for a different endogenous parameter, $\theta_{\tilde{\imath},\tilde{h}}$. The two scenarios cannot be placed in the same endogenous scenario group in this case, however, since they have different possible realizations for $\theta_{\hat{\imath},\hat{h}}$ and thus would differ in the possible realizations of both $\theta_{\hat{\imath},\hat{h}}$ *and* $\theta_{\tilde{\imath},\tilde{h}}$ (i.e., 2 parameters). This would violate Property 3. It follows, then, that two scenarios in the same endogenous scenario group cannot appear together in *any* other endogenous scenario group, for *any* endogenous parameter. This is a fairly obvious conclusion since $\hat{s}$ and $\hat{s}'$ differ in the possible realization of only one endogenous parameter, and in any arbitrary time period, we would expect scenario pair $(\hat{s}, \hat{s}')$ to appear only once.

The endogenous scenario groups corresponding to $\theta_{\hat{\imath},\hat{h}}$ will have the following form: ${}^{N}\mathcal{G}^{\hat{l}}_{\hat{\imath},\hat{h}} \coloneqq \left\{s\colon\ s = \alpha_1,\, \alpha_2, \dots, \alpha_{M_{\hat{\imath},\hat{h}}}\right\}$, ${}^{N}\mathcal{G}^{\hat{l}'}_{\hat{\imath},\hat{h}} \coloneqq \left\{s\colon\ s = \beta_1,\, \beta_2, \dots, \beta_{M_{\hat{\imath},\hat{h}}}\right\}$, ${}^{N}\mathcal{G}^{\hat{l}''}_{\hat{\imath},\hat{h}} \coloneqq \left\{s\colon\ s = \eta_1,\, \eta_2, \dots, \eta_{M_{\hat{\imath},\hat{h}}}\right\}$, etc. By Property 4, we pair off consecutive scenarios in each of these groups. Note that these scenarios may be nonconsecutively indexed, and this necessitates the use of a different naming convention than used previously in the proof of Property 5. The associated endogenous NACs for an arbitrary time period $t = \tau$ are then:

$$y_\tau^{\alpha_1} = y_\tau^{\alpha_2}, \dots, y_\tau^{\alpha_{M_{\hat{\imath},\hat{h}}-1}} = y_\tau^{\alpha_{M_{\hat{\imath},\hat{h}}}} \quad \textbf{(89)}$$

$$y_\tau^{\beta_1} = y_\tau^{\beta_2}, \dots, y_\tau^{\beta_{M_{\hat{\imath},\hat{h}}-1}} = y_\tau^{\beta_{M_{\hat{\imath},\hat{h}}}} \quad \textbf{(90)}$$

$$y_\tau^{\eta_1} = y_\tau^{\eta_2}, \dots, y_\tau^{\eta_{M_{\hat{\imath},\hat{h}}-1}} = y_\tau^{\eta_{M_{\hat{\imath},\hat{h}}}} \quad \textbf{(91)}$$

provided that $\theta_{\hat{\imath},\hat{h}}$ has not yet been realized (i.e., the scenarios are indistinguishable). If $\theta_{\hat{\imath},\hat{h}}$ *has* been realized, then the scenarios are distinguishable and the NACs do not apply, so it is only necessary for us to consider the former case where these constraints are active.

At this point, recall that for any endogenous parameter, every scenario in $\mathcal{S}$ can be accounted for as a member of one of the endogenous scenario groups corresponding to that parameter. Further recall that none of the scenarios in those respective groups can appear together in any other group. This means that *all* other endogenous scenario groups can be produced, respectively, by selecting one scenario from

different groups defined for $\theta_{\hat{\imath},\hat{h}}$. Accordingly, we will continue to use the same naming convention in the definitions of other endogenous scenario groups in this proof (i.e., we will use α_m, β_m, η_m, etc. for scenarios, where $m = 1, 2, \dots, M_{\hat{\imath},\hat{h}}$).

Now, consider two scenarios $\hat{s}$ and $\hat{s}''$ from two separate groups $\hat{l}$ and $\hat{l}'$ corresponding to $\theta_{\hat{\imath},\hat{h}}$ (i.e., ${}^{N}\mathcal{G}_{\hat{\imath},\hat{h}}^{\hat{l}}$ and ${}^{N}\mathcal{G}_{\hat{\imath},\hat{h}}^{\hat{l}'}$). In this case, the scenarios *may* differ in the possible realization of $\theta_{\hat{\imath},\hat{h}}$ (depending on their respective positions in the two groups), and *must* differ in the possible realization of at least one endogenous parameter other than $\theta_{\hat{\imath},\hat{h}}$ (since they belong to two separate groups corresponding to $\theta_{\hat{\imath},\hat{h}}$). Notice that the only way for scenarios $\hat{s}$ and $\hat{s}''$ to have the same possible realization for $\theta_{\hat{\imath},\hat{h}}$ is if they have the same *position* in both groups (not to be confused with parameter $Pos(s)$). For example, if they both have the lowest realization for $\theta_{\hat{\imath},\hat{h}}$, they must be the lowest-indexed scenarios in their respective groups; if they both have the highest realization for $\theta_{\hat{\imath},\hat{h}}$, they must be the highest-indexed scenarios in their respective groups. If, then, $\hat{s}$ and $\hat{s}''$ have the same possible realization for $\theta_{\hat{\imath},\hat{h}}$ (i.e., the same position in both of their groups) and differ only in the possible realization of endogenous parameter $\theta_{\tilde{\imath},\tilde{h}}$, they will be placed in the same group corresponding to $\theta_{\tilde{\imath},\tilde{h}}$ by the endogenous scenario-group algorithm. (Note that if $\hat{s}$ and $\hat{s}''$ instead differ in other possible parameter realizations, they will be placed in different groups, and the discussion that follows would apply for the specific endogenous parameter for which this condition does apply.)

The endogenous scenario groups corresponding to $\theta_{\tilde{\imath},\tilde{h}}$ will then have the following form: ${}^{N}\mathcal{G}_{\tilde{\imath},\tilde{h}}^{\tilde{l}} \coloneqq \{s\colon\ s = \alpha_1, \beta_1, \eta_1, \dots\}$, ${}^{N}\mathcal{G}_{\tilde{\imath},\tilde{h}}^{\tilde{l}'} \coloneqq \{s\colon\ s = \alpha_2, \beta_2, \eta_2, \dots\}, \dots,$ ${}^{N}\mathcal{G}_{\tilde{\imath},\tilde{h}}^{\tilde{l}''} \coloneqq \left\{s\colon\ s = \alpha_{M_{\hat{\imath},\hat{h}}}, \beta_{M_{\hat{\imath},\hat{h}}}, \eta_{M_{\hat{\imath},\hat{h}}}, \dots\right\}$, etc. Notice that in ${}^{N}\mathcal{G}_{\tilde{\imath},\tilde{h}}^{\tilde{l}}$, the lowest-indexed scenario from group ${}^{N}\mathcal{G}_{\hat{\imath},\hat{h}}^{\hat{l}}$ has been grouped with the lowest-indexed scenarios from groups ${}^{N}\mathcal{G}_{\hat{\imath},\hat{h}}^{\hat{l}'}$ and ${}^{N}\mathcal{G}_{\hat{\imath},\hat{h}}^{\hat{l}''}$, the second-lowest-indexed scenarios have been grouped in ${}^{N}\mathcal{G}_{\tilde{\imath},\tilde{h}}^{\tilde{l}'}$, and so forth. In general, the remaining groups corresponding to $\theta_{\tilde{\imath},\tilde{h}}$ would be generated from all other groups corresponding to $\theta_{\hat{\imath},\hat{h}}$, in the same manner, and would consist of scenarios other than α_m, β_m, and η_m. Note that the same general approach also applies for the endogenous scenario groups of all other endogenous parameters, with the respective scenarios selected from different groups corresponding to $\theta_{\hat{\imath},\hat{h}}$. As before, the associated NACs for time period $t = \tau$ are:

$$y_\tau^{\alpha_1} = y_\tau^{\beta_1},\ y_\tau^{\beta_1} = y_\tau^{\eta_1}, \dots \tag{92}$$

$$y_\tau^{\alpha_2} = y_\tau^{\beta_2},\ y_\tau^{\beta_2} = y_\tau^{\eta_2}, \dots \tag{93}$$

$$y_\tau^{\alpha_{M_{\hat{\imath},\hat{h}}}} = y_\tau^{\beta_{M_{\hat{\imath},\hat{h}}}},\ y_\tau^{\beta_{M_{\hat{\imath},\hat{h}}}} = y_\tau^{\eta_{M_{\hat{\imath},\hat{h}}}}, \dots \tag{94}$$

provided that $\theta_{\tilde{\imath},\tilde{h}}$ has not yet been realized.

Notice that because $y_\tau^{\alpha_1} = y_\tau^{\alpha_2}$ by Equation (89), and $y_\tau^{\beta_1} = y_\tau^{\beta_2}$ by Equation (90), we can rewrite the *first* endogenous constraint in Equation (92), $y_\tau^{\alpha_1} = y_\tau^{\beta_1}$, as $y_\tau^{\alpha_2} = y_\tau^{\beta_2}$. Notice that this is the *first* endogenous constraint in Equation (93).

Since $y_\tau^{\eta_1} = y_\tau^{\eta_2}$ by Equation (91), we can use this constraint with Equation (90) to rewrite the *second* endogenous constraint in Equation (92), $y_\tau^{\beta_1} = y_\tau^{\eta_1}$, as $y_\tau^{\beta_2} = y_\tau^{\eta_2}$. Notice that this is the *second* endogenous constraint in Equation (93).

Given any remaining scenarios that differ from α_2, β_2, and η_2 in the possible realization of only $\theta_{\tilde{\imath},\tilde{h}}$, this process can be continued to produce all of the remaining NACs corresponding to group ${}^{N}\mathcal{G}_{\tilde{\imath},\tilde{h}}^{\tilde{\imath}'}$ in Equation (93). In fact, if we consider only scenarios α_m, β_m, and η_m (where $m = 1, 2, \dots, M_{\tilde{\imath},\tilde{h}}$), it is not difficult to see that by using the first, second, third, etc. NACs from Equations (89), (90), and (91), along with Equation (92), we can imply *all* of the NACs corresponding to the groups for $\theta_{\tilde{\imath},\tilde{h}}$ that involve these scenarios. This includes the NACs for ${}^{N}\mathcal{G}_{\tilde{\imath},\tilde{h}}^{\tilde{\imath}''}$, as shown in Equation (94).

To summarize the results in a general sense, first recall that we begin with the endogenous scenario groups corresponding to an arbitrary parameter $\theta_{\tilde{\imath},\tilde{h}}$. We will refer to these groups as our "base" groups. We assume that the corresponding NACs apply as equality constraints (i.e., fixed endogenous NACs).

Further recall that the groups for all other endogenous parameters can be produced by selecting scenarios from different base groups, where in each case, the scenarios have *the same position* in their respective base groups (e.g., the lowest indexed, the second-lowest indexed, etc.). However, for an arbitrary parameter $\theta_{\tilde{\imath},\tilde{h}}$, we have just shown that the groups produced from the *second-lowest-indexed* scenarios, the *third-lowest-indexed* scenarios, etc. result in redundant NACs. This is the case regardless of whether these constraints are conditional or fixed.

The reasoning here, in general, is that *all* NACs associated with an arbitrary endogenous parameter $\theta_{\tilde{\imath},\tilde{h}}$ can be implied by the base-group NACs and the NACs derived from pairing off the *lowest-indexed* scenarios from those base groups. (Note that our choice to use the lowest-indexed scenarios (rather than, for example, the highest) is arbitrary, and we have made this selection for convenience.)

It then follows that, for generating endogenous scenario pairs, it is sufficient to consider only scenarios s and s' that are in the *first position* of their respective base groups (i.e., the *lowest-indexed* scenarios from these groups), excluding all scenarios eliminated by Property 5. We refer to these scenarios as "unique." We may then proceed to another endogenous parameter for which the associated NACs apply as equality constraints, and consider new base groups, where we allow only the *new* lowest-indexed scenarios that were members of the previous set of unique scenarios. Because the introduction of new equality constraints may allow us to imply other existing endogenous constraints, we may be able to remove additional scenarios from the pairing process each time we update our set of unique scenarios. In time periods where we do not have fixed endogenous NACs, a similar strategy can still be used for the case of multiple parameters associated with the same source, although the set of unique scenarios can only be updated in the context of that particular source.

The strategy outlined here is the basis for the unique scenarios algorithm. ■

A.12 Proof of Proposition 5

Starting from Proposition 4, notice that we have relaxed the two remaining assumptions; specifically, that there are no initial 'equality' periods and that there is only one endogenous parameter associated with each source of uncertainty. To prove that we have the minimum number of endogenous scenario pairs in this general case, it is only necessary for us to show that after relaxing the two assumptions, and then introducing Property 6, the pairs in each endogenous scenario group cannot be implied by any other pairs.

As previously discussed, there will be redundant scenario pairs in the model when we consider initial 'equality' periods and multiple endogenous parameters associated with some of the sources of uncertainty. These redundant pairs are eliminated by Property 6.

It follows that although there are both fixed endogenous NACs and conditional endogenous NACs, none of the remaining pairs can be used to imply any of the others. Thus, the complete set of endogenous scenario pairs, $\mathcal{SP}_N$, generated by Property 4, Property 5, and Property 6 contains the minimum number of pairs. ■