Stochastic Programming Models for Interest-Rate Risk Management

by

Pieter Klaassen

Doctorandus, Erasmus Universiteit Rotterdam
The Netherlands (1989)

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 1994

Author .....
Sloan School of Management
May 17, 1994

Certified by .....
Jeremy F. Shapiro
Professor of Operations Research and Management
Thesis Supervisor

Accepted by .....
Thomas L. Magnanti
Co-director, Operations Research Center

Stochastic Programming Models for Interest-Rate Risk Management

by

Pieter Klaassen

Submitted to the Sloan School of Management on May 17, 1994,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Operations Research

Abstract

We consider investment problems that can be viewed as asset/liability management (ALM) problems, and assume that the investor faces several market imperfections and trading restrictions (transaction costs, restrictions on borrowing, and limits on short sales of assets). It is shown that such problems can be modeled naturally as multistage stochastic programs, and that these models provide substantial flexibility for describing many practical instances. We only consider interest-rate uncertainty and portfolios of fixed-income and derivative securities in this thesis, but the extension to additional sources of uncertainty is discussed.

The study of stochastic programming models for investment problems in the literature has concentrated primarily on the development of efficient solution methods, while little attention has been given to the question of how to describe uncertainty in the model. We show that the optimal solution is very sensitive to this description, and argue that it is both reasonable and important to require that asset prices in the event tree representing the uncertainty are arbitrage-free. We also show how financial term-structure models provide a rigorous way to obtain a description of the uncertainty in future asset prices which satisfies this requirement.

The event tree that follows from a term-structure model, however, is often too large to include fully in a stochastic programming model without losing the ability to solve it. We present state and time aggregation methods that enable a reduction in the size of the event tree while maintaining arbitrage-free prices in the aggregated event tree. These methods furthermore guarantee that the aggregated event tree remains consistent with observed market prices if this was true for the original tree.

We describe how the aggregation methods can also be used as the basis for a novel solution approach to the ALM problem, in which the description of the uncertainty in the stochastic program is iteratively refined. These iterative disaggregations provide additional insights into the effect of uncertainty on the optimal portfolio strategies. The feasibility of decomposition methods to re-optimize the stochastic program in each iteration is discussed. Computational results from the application of this algorithm to a simple ALM problem are presented.

Thesis Supervisor: Jeremy F. Shapiro
Title: Professor of Operations Research and Management

Acknowledgments

I clearly remember the excitement that I felt when I walked along the Charles River on the evening of my first day in Boston in early September, 1989. The glittering skyline of Boston across the river and the lit-up facade of MIT on Killian Court on that warm September evening set the stage for what have been five exciting and enriching years of living in Boston and studying at MIT. Now that the time to leave has come, it is with great pleasure that I remember the many people here who have made these years such an enjoyable and memorable experience.

First of all, I feel very fortunate that I have had the opportunity to work with Jerry Shapiro throughout my studies. I have greatly benefitted from his broad experience, and his continuous enthusiasm and sense of humor have made our frequent conversations very pleasant ones. His many ideas have not only helped me overcome many hurdles on the way to completing this dissertation, but they can easily form the basis for at least ten more years of research.

I am also grateful that Jerry introduced me to Jean-Luc Vila and Stewart Myers in the finance group at the Sloan School, who have pointed out many important issues on the way to blending financial engineering theory and portfolio optimization models. I wish to thank Jean-Luc for his constant interest, his careful reading of many of my manuscripts, and his constructive advice about the proper use of finance theory. I owe thanks to Stewart Myers not only for his suggestions on theoretical finance issues, but also for the practical financial support of part of my research.

My pleasure in studying at MIT can for a large part be attributed to the very friendly and supportive place that the Operations Research Center has been throughout the years, and I want to thank all the people who have made this so: the co-directors, administrative staff, and of course my fellow students. On the practical side of things, I am indebted to Paulette Mosley for carefully keeping track of all my academic requirements and her help in solving many a financial puzzle at the beginning of a new semester.

The contacts with many people I got to know in the ORC have not been been confined to our office space, and I want to thank in particular Kerry Malone (*"Het was fantastisch!"*), Raghu (who taught me how to eat with my hands without feeling uncivilized), Sungsu Ahn, Armann Ingolfsson, Rob Shumsky, and Albert Wagelmans for the many joyful hours we have spent outside the ORC. I sincerely hope that our friendship will extend far beyond our departures from the ORC. I am furthermore

indebted to Sungsu, Armann, Rob, Dave Markowitz and Zhihang Chi for their help and advice on many computer software and hardware problems.

Much of the beauty of Boston and New England I have discovered on trips with the MIT European Club, which has also enabled me to establish many new friendships and preserve some of my ties to the Old World (including the practice of soccer). I want to thank everyone who has been involved in making the Club so active, especially Andreas Kussmaul, Julita Kussmaul-Pomorska, Jari Kinaret, Mike Peterson, Menke Ubbens, Heidi Roth, Robert Kallenberg and Timo Smit.

An important fixed point during all but my first year in Cambridge has been "Triple Vision", the apartment which I have shared with Lars Schade, Keith Beyer and Timo Smit. The almost daily dinners we have had together have truly made it feel like a home away from home. I will long remember many of the topics that came up during our lively dinner discussions, which ranged from the workings of American society and government to the choice of a salad dressing. I hope that we will be able to continue this dinner tradition far into the future, despite the fact that it will necessarily have to be on a much less frequent basis.

I want to thank my parents for their continuous and strong support of my pursuit of a doctoral degree in the US. Their weekly updates on their activities, family life, and Dutch politics have been an invaluable link to the world back home. I also want to thank my brother Lauw for his numerous joyful and personal E-mails, which were always a great pleasure to read.

Last, but certainly not least, I am deeply grateful to Margot for sticking with me in the five long years that are behind us now, and in which the Atlantic Ocean has separated us for most of the time. More than once have I been in doubt as to whether the decision to go to MIT was a wise one, and I am not sure that I know the answer now. Margot's cheerfulness and optimism, expressed in her many long letters and our regular phone calls, have been a constant source of energy to continue on the path that we started, even when the end of the path was moving with us as we went. Now that we finally have the possibility to be together without the prospect of having to part, I look forward to exploring the rest of the world together. *Bedankt, Margot!*

Pieter Klaassen
Cambridge, Massachusetts

Contents

Chapter 1

Problem Definition

The last twenty years have shown a tremendous growth in the markets of derivative financial securities[1]. This growth has been spurred by the development of a theory for the valuation of such securities, which started with the well-known options pricing models of Black and Scholes [6] and Merton [44], and has been extended to a host of other derivative instruments (see Hull [33] for a comprehensive and accessible treatment). Today, this theory is widely used in financial practice.

The payoff patterns of derivative instruments enable investors to better hedge against specific risks and to tailor their portfolios more precisely to their investment objectives than was possible in the past. However, little work has been done on the development of data-driven portfolio optimization models that explicitly consider derivative securities, and which can help investors to determine exactly how these instruments are best included in their portfolios. The formulation and solution of such models is the subject of this thesis, and we will show that financial asset pricing theory plays a central role in this.

Most of the portfolio optimization models that have been used in practice are variants of the mean-variance model which was developed by Markowitz [41]. These models are generally used to construct equity portfolios, and we will show later in this chapter that the assumptions behind this model do not allow for the inclusion of derivative securities or other types of investments. Furthermore, the mean-variance models are static (one-period) models, and thus ignore the dynamic nature of actual portfolio management. Being able to change one's investment portfolio in response

[1] A derivative security is a security whose value depends on one or more underlying variables. Examples are option and future contracts on stocks and bonds. A bond itself can also be viewed as a derivative security, as its value is dependent on interest rates.

to future events may substantially influence the optimal portfolio composition today. Optimal portfolio decisions that follow from a static model may therefore be significantly more expensive than the ones that follow from a model that explicitly includes the possibility for adjustment of the portfolio composition in the future. We therefore consider *dynamic* portfolio optimization models in this thesis.

In the financial economics literature, dynamic models for optimum consumption and portfolio selection have been studied. These models consider the maximization of the expected utility of intertemporal consumption for an individual investor. The analytic solution of these optimization models by (stochastic) dynamic programming necessitates idealized assumptions about the preferences and the behaviour of the investor, the structure of asset prices and the functioning of financial markets. For example, it does not tolerate the inclusion of market frictions such as transaction costs, taxes and position limits, nor does it allow for the consideration of investment constraints that investors face in practice. Although these financial models have been powerful tools in the development of theories about the structure of financial markets, they cannot capture the complexities of realistic portfolio investment problems.

In this thesis, we will study the use of stochastic linear programming models for dynamic portfolio investment problems. Stochastic linear programs were originally formulated by Dantzig [12], and have long been recognized for their ability to model realistic decision problems under uncertainty. Their application in practice has been limited as these models become very large very quickly, and must be solved by numerical optimization methods whose efficiency very much depends on the size of the optimization model. The continuing advances in computer technology, however, together with the development of specialized solution methods for stochastic programs, has renewed interest in the application of these models in recent years.

Several other researchers have considered stochastic programming models for financial applications, many of which will be reviewed in section 1.2. The emphasis in these studies has nearly exclusively been on the devise of efficient solution methods for the resulting models, whereas little attention has been given to the specification of the uncertainty in the model. We will show in this thesis that a "careful" description of the uncertainty in a stochastic programming model for portfolio optimization is crucial to obtain reasonable results from the model. Furthermore, we will show how financial asset pricing theory provides a structured way to obtain such a description.

As indicated before, stochastic programming models can only include a fairly limited description of the uncertainty in the future in order to remain computationally

tractable. Not only is it therefore important *how* one describes the uncertainty in the model, but the optimal solution may also depend significantly on the *level* of uncertainty. This dependency has hardly been studied in the stochastic programming literature. For the portfolio optimization models that are the subject of this thesis we will develop an iterative solution algorithm that gradually increases the level of uncertainty in the stochastic program. This algorithm therefore enables one to monitor exactly how the optimal solution changes with additional uncertainty in the model, and which uncertainty affects the optimal solution most.

The remainder of this chapter is organized as follows. In section 1.1 we will state assumptions for the portfolio optimization problems that we consider in this thesis. Section 1.2 contains a classification and discussion of optimization models for portfolio management under uncertainty that have been proposed in the literature. A summary of the contents of the thesis is given in section 1.3.

1.1 Assumptions

We consider an investor who wants to manage an asset portfolio over time such that the payoffs from this portfolio meet a stream of future target payoffs (liabilities). Both the asset returns and the target payoffs are allowed to be stochastic. A problem with this structure is usually called an *asset/liability management problem*, and we will refer to it as the *ALM problem*.

Depending on the nature of the target payoffs and the length of the investment horizon, this type of problem arises in many contexts in practice. Examples are a pension fund or insurance company that has to meet fixed (pensions) or estimated (insurance claims) liabilities over time. Another example is the selection of a portfolio strategy that provides certain return characteristics over time (e.g., indexing). Hedging is still another application, where the target payoffs equal the expected depreciation in the asset that one wants to be hedged. We do not consider models for the sole purpose of speculation or return maximization. However, obtaining an attractive portfolio return is obviously important to any investor, and this will be reflected in our models.

In this section we will make assumptions about the nature of the stochasticity in the ALM problem, and about the financial markets in which the investor operates.

Assumption 1 *Interest-rate uncertainty is the only source of uncertainty.*

As a consequence we restrict ourselves to portfolios that consist of interest-rate-dependent securities only. Asset/liability management under interest-rate uncertainty is sometimes called *immunization* in the literature, and the asset portfolio that is formed for this purpose a *structured bond portfolio.*

Assumption 1 allows us to focus our discussion and simplify the exposition, but the methodology we develop is not limited to it. In the final chapter of the thesis we will indicate how the methodology can be extended to include additional sources of risk like exchange-rate risk and stock-market risk. Besides, management of interest-rate risk constitutes an important problem in its own right. Special models and methods have been developed just for this purpose, some of which will be reviewed in section 1.2.

Assumption 2 *Markets of interest-rate-dependent securities are dynamically complete and security prices are arbitrage-free under the following conditions:*

1. *There are no transaction costs or taxes.*
2. *Securities are infinitely divisible.*
3. *Interest rates for borrowing and lending are the same.*
4. *Short sales of assets with full use of proceeds are allowed.*

We will often refer to the conditions in this assumption as the *perfect market conditions.*

Loosely stated, market completeness means that the payoffs of any security can be obtained exactly by dynamically managing some portfolio of other securities over time. It should be intuitive that equilibrium in financial markets requires that the price of the security equals the value of such a replicating portfolio at all times; prices are then said to be arbitrage-free. Assumption 2 thus imposes such an equilibrium structure on financial markets. Appendix A contains a more precise description of the main ideas in the theory on asset pricing by arbitrage.

We note that assumption 2 does not imply that *all* investors face perfect market conditions. In fact, we will assume that the situation for the particular investor in our models differs from these conditions. However, assumption 2 *does* require that at least some investors can trade under these conditions, and that they will do so if market prices violate arbitrage relationships. This is not as unrealistic as it may seem at first. There are large investors in today's financial markets whose main objective is to find and take advantage of arbitrage opportunities, and who face conditions that do not deviate much from the perfect market conditions.

The market imperfections confronting the investor in our models are recorded in the following assumption.

Assumption 3 *The individual investor that we consider in our models faces trading restrictions that violate the perfect market conditions in assumption 2.*

This assumption essentially demands sufficient flexibility of a portfolio optimization model to accomodate such trading restrictions. The specific trading restrictions that we will focus on in the sequel are proportional transaction costs, limits on borrowing and short sales of assets, and an interest-rate differential between borrowing and lending rates.

1.2 Optimization Models for Portfolio Management under Uncertainty

This section contains a classification of various optimization models that have been proposed in both the financial and the operations research literature for portfolio management under uncertainty. The emphasis is on the structure of and assumptions behind the different types of models, and we will not limit our discussion to models that satisfy the specific assumptions of the previous section. Instead, we will point out how earlier studies differ in their assumptions from the ones we make in this thesis. We successively discuss static and dynamic portfolio optimization models.

1.2.1 Static Portfolio Optimization Models

Static portfolio optimization models are concerned with the *selection* of a portfolio such that the selected portfolio optimizes certain characteristics among all feasible portfolios. In the discussion below, we make a distinction between one-period models, which include the well-known mean-variance optimization models, and models for duration matching.

One-Period Models

In one-period portfolio optimization models, the objective is typically to find a portfolio that maximizes the investor's expected utility of the portfolio value at some future date. They thus ignore the inherently dynamic nature of portfolio management in general, and asset/liability management in particular.

In practical applications, the maximization of expected utility is mostly replaced by Markowitz's mean-variance analysis [41]. The objective is then to minimize the variance of the portfolio return at a future date, subject to the constraint of a minimum required expected return. It is well-known that this is equivalent to the maximization of expected utility only if investors have quadratic utility, or else if security returns follow a joint normal distribution. Because mean-variance analysis has become quite popular in practice for the selection of equity portfolios, we will discuss in some detail why it is not suitable for the problems we consider in this thesis. First, we will argue that the mean-variance criterium is inappropriate when fixed-income securities and derivative instruments are considered. Second, mean-variance analysis conflicts with equilibrium asset pricing in complete markets (unless all investors have quadratic utility). And third, minimizing the variance of the portfolio return leads to inconsistent strategies in a multiperiod model.

The use of Markowitz's mean-variance models for the construction of equity portfolios is sometimes justified by the claim that the expected rate of return on a stock and its variance tend to be fairly constant over short periods of time, as are the covariances between the rates of return on different stocks. Although this may be a reasonable approximation in the case of equities, it is unrealistic for assets with a fixed maturity date such as derivative instruments and fixed-income securities. For a straight bond, the variance of its return is a nonlinear function of the time to maturity, and decreases to zero close to its maturity date.

In addition, derivative instruments typically have an asymmetric payoff pattern. The popularity of these instruments indicates that investors generally do not value upside and downside risk equally, and thus that the variance as measure of risk is inappropriate. Ritchken [50] shows that a mean-variance investor who invests in a well-diversified mutual fund (which Ritchken takes as a proxy for the market portfolio) can decrease his portfolio's variance of return without sacrificing expected return by writing call options on part of his mutual-fund holdings. The resulting return distribution is negatively skewed, but this is ignored by mean-variance analysis.

The second objection to the use of mean-variance analysis is due to Dybvig and Ingersoll [14]. Assuming perfect markets, and if all investors choose mean-variance efficient portfolios, then the prices of all assets have to follow the Capital Asset Pricing Model (CAPM) in equilibrium. However, Dybvig and Ingersoll show that arbitrage opportunities exist if all assets are priced according to the CAPM and if markets are also complete. Investors will take advantage of these arbitrage opportunities,

thus violating the equilibrium. (The only possible, and unrealistic, exception is when investors have quadratic utility, and all have reached their point of satiation.) The theory of complete markets forms the basis for the valuation of derivative securities, as well as for the optimization models we will develop (see assumption 2). A mean-variance optimality criterium is therefore inappropriate.

As for the third objection, minimization of the variance over a multiperiod investment horizon violates the *conditional weak independence of preferences* (Johnsen and Donaldson [39]). If the variance of the final portfolio value is being minimized, then optimal portfolio decisions at intermediate points in time (when some of the uncertainty will have been resolved) will depend on the portfolio decisions that would have been taken in unrealized states of the world. That is, such portfolio decisions depend on actions that would have been taken if the world would have evolved differently. Preferences are then said to exhibit conditional weak dependence. This is clearly an undesirable property.

Duration Matching

Duration matching models have been specifically designed for asset/liability management under interest-rate uncertainty. Given a stream of future liabilities, they aim to eliminate the interest-rate exposure from the combined asset/liability portfolio by constructing an asset portfolio whose interest-rate sensitivity matches the interest-rate sensitivity of the liabilities. In their simplest form this is accomplished by selecting a portfolio whose value and duration (defined as the first-order derivative of the portfolio value with respect to the interest rate) equal, respectively, the present value and the duration of the future liabilities. To improve the matching, equality of higher-order derivatives is sometimes added as requirement.

Duration matching thus reduces the dynamic nature of asset, liability management to a calculation of present values and the matching of derivatives. This makes the approach structurally simple and computationally attractive, which accounts to a large extent for its popularity. However, it fully ignores the possible mismatch in the *timing* of individual asset cash flows and liabilities, and the associated costs. Furthermore, one continuously has to monitor whether the duration of the asset and liability portfolio remains matched when time proceeds and interest rates change, and adjust the asset portfolio if it does not. With transaction costs present, this can lead to a very expensive strategy over time. We refer to Hiller and Schaack [29] for further discussion of duration matching and comparison with other approaches to

asset/liability management under interest-rate uncertainty.

1.2.2 Dynamic Portfolio Optimization Models

Static models for portfolio optimization are attractive from a computational perspective, but cannot cope with many important costs and considerations that follow from the dynamic nature of portfolio management. The last two decades have shown increased attention to dynamic models, both in theory and practice. In the financial literature, the emphasis has been on models in continuous time, whereas most models in the operations research literature are stated in discrete time. We will discuss both classes of models in turn.

Continuous-Time Models

Continuous-time models for portfolio selection in the financial literature generally consider an investor who wants to maximize his expected utility of consumption and final wealth over some fixed period of time. Given some initial level of wealth, the problem is thus to simultaneously determine an optimal consumption pattern and investment strategy over time as a function of the state of the world. The emphasis in these studies has nearly exclusively been on the *structure* of optimal solutions in an idealized environment and its implications for asset prices and market equilibrium, and not on the solution of realistic instances involving actual market data. We will indicate why continuous-time models are ill-suited for the last purpose.

Continuous-time models for optimum consumption and portfolio selection were introduced by Merton [42, 43] (see also [45, chapters 4 and 5]). It is typically assumed that there is a finite number of state variables that define the state of the world, each of which follows a diffusion process over time. The price process of each security in the economy is also described by a diffusion process, and the parameters of these processes depend on at most the state variables and time. It is furthermore assumed that investors can trade continuously and that perfect market conditions prevail. With some regularity conditions on the utility functions, one can then use stochastic dynamic programming to write the optimal objective function value over time (as a function of the state variables) as the solution to a nonlinear, second-order partial differential equation, and express the optimal consumption and investment strategy in terms of this value function. The explicit solution to this differential equation, however, is only known for particular choices of utility functions and further restrictions on the

behaviour of state variables and securities.

In a few recent studies, some of the perfect market assumptions have been relaxed. Grossman and Vila [21] introduce a borrowing limit as well as a nonnegativity constraint on wealth, but they limit their analysis to an investor with constant relative risk aversion who can only choose between one risky and one riskless asset. He and Pearson [25] consider short-sale constraints. However, most of the assumptions are crucial for the use of stochastic dynamic programming and therefore cannot be relaxed. This and the fact that the approach cannot cope with many additional constraints that would arise in practical applications render continuous-time models unsuitable for our purposes. Besides, most investors do not want to monitor and revise their portfolio in a continuous fashion, but rather only at fixed points in time. We therefore turn our attention to dynamic models in discrete time.

Discrete-Time Models

Multiperiod models in discrete time require that the uncertainty can be approximated by an event tree with a finite number of possible states of the world at each time. They offer substantial flexibility for incorporating market imperfections and constraints that continuous-time models cannot deal with. However, the cost of this flexibility is that such models can not be solved by an analytic solution method, and one has to resort to numerical optimization methods. Efficient application of numerical methods imposes limits on the size of the model, which must be accomplished through restrictions on the level of uncertainty and the number of time periods and assets. Furthermore, additonal structure is often assumed in the models (e.g., linearity) so that relatively efficient numerical methods can be used.

Some authors have applied dynamic programming to solve multiperiod problems. Eppen and Fama [16] calculate the solution of a cash-balance problem with transaction costs, in which the level of cash fluctuates randomly over time, but the returns on the two available assets ("stock" and "bond") are assumed to be known. Edirisinghe, Naik and Uppal [15] consider the replication of options and option portfolios for an investor facing proportional transaction costs as well as trading restrictions in the form of lot-size constraints and position limits. Their model has the structure of an asset/liability management model, but they limit their analysis to one riskless and one risky asset. Dynamic programming is extremely sensitive to the dimensions of the problem as the optimal decisions at each time and in each state have to be determined for every possible portfolio composition. It effectively evaluates every possible

portfolio strategy to find the optimal one, and is therefore impractical for problems of a realistic size and complexity.

Bradley and Crane [7] and Kusy and Ziemba [40] consider stochastic linear programming models for bank asset/liability management. Both models consider proportional transaction costs and taxes, and include policy constraints. The model of Bradley and Crane concentrates on the selection of an optimal bond portfolio investment strategy. Theirs is a truly stochastic model in which the size of the investment portfolio fluctuates over time, as do the returns on classes of bonds with different maturities that they consider. The way in which they define the variables enables them to solve the model by Dantzig-Wolfe decomposition, where the subproblems (corresponding to the inventory balance equations) can be solved in an efficient recursive manner. They apply their approach to models with three time periods, three events per period, and eight asset classes, and claim that it is faster than solving the model as a large linear program.

Kusy and Ziemba only consider uncertainty in the level of future deposits, and assume that the returns on different investments are constant throughout the whole planning horizon. They further simplify their model by letting the portfolio decisions at future dates be independent of the state of the world (i.e., the level of deposits). This gives the model the structure of a stochastic linear program with simple recourse, which they solve using a specialized algorithm of Wets [56]. Although they perform simulations to compare their optimal strategy with the optimal solution from a version of the Bradley and Crane model, they do not explicitly analyze what the effect of their assumption of constant asset returns on the optimal strategy is. Birge [4] shows that neglecting the stochastic nature of the problem can substantially affect the optimal solution.

Hiller and Eckstein [28] consider a stochastic programming model for general asset/liability management under interest-rate uncertainty. To make their model suitable for a massively parallel implementation of Benders decomposition, they exclude rebalancing decisions and assume that any mismatch between portfolio payoffs and future liabilities is resolved by short-term borrowing and lending. The resulting simple structure allows them to solve models with 360 periods (months), 58 different assets and up to 2048 interest-rate scenarios. The implementation on a parallel computer with 16,000 processors gave a speedup factor of less than eight when compared to a serial implementation.

Mulvey and Vladimirou [46] and Cariño et al. [9] describe multiperiod models for

asset allocation that include different classes of bonds, equities and real estate. Both models assume that returns on all asset classes are stochastic, and allow for stochastic liabilities. Mulvey and Vladimirou employ a generalized network structure in their formulation, and use the progressive hedging algorithm of Rockafellar and Wets [52] to solve problems with up to 8 periods, 15 asset classes and 100 scenarios. Cariño et al. formulate the asset/liability management problem of a Japanese insurance company as a multistage stochastic linear program, where the liabilities consist of payments on certain types of insurance policies. They experiment with different solution methods, and conclude that Benders' decomposition (see Birge [3]) is most efficient for large problems (which include 6 periods, 7 asset classes and 1024 economic scenarios).

Despite the use of specialized algorithms and significant increases in computer power, it is clear that stochastic programming models which consider uncertainty in future asset returns and include portfolio rebalancing opportunities can only incorporate a fairly limited description of the true uncertainty. However, few of the papers discussed so far elaborate on how such an approximate description should be obtained. Cariño et al. describe a sampling procedure from return distributions that are estimated from historical data. Hiller and Eckstein are the only ones that use a financial model to describe the uncertainty in future interest rates. However, they ignore its consistency with current market data and thereby introduce arbitrage opportunities in their model. They use the interest-rate model to calculate a theoretical present value for each fixed-income and derivative security, as well as for the stream of future liabilities. The optimal portfolio maximizes the difference in *present value* between the asset portfolio and the liabilities, subject to a budget constraint that limits the *cost* of the asset portfolio as calculated from observed market prices. They thus implicitly assume that the market prices do not reflect the true values of the assets, which causes a strong bias in the optimal portfolio towards assets with high value/cost ratios. Their use of financial asset pricing theory thus violates our assumption 2.

Zipkin [59] also observes some of the missing links and inconsistencies with finance theory in structured bond portfolio models, but he does not discuss the consequences of such inconsistencies or suggest ways to correct them. We will show in chapter 2 that careless specification of asset returns in a necessarily limited description of the uncertainty easily leads to substantial and unwanted biases in the optimal solution.

1.2.3 Conclusion

The static models for mean-variance analysis and duration matching have been popular in practice, mainly because they are conceptually simple and relatively easy to solve. However, it is clear that they ignore most of the dynamics of actual portfolio management.

Continuous-time models for optimum consumption and portfolio selection are primarily used to analyze the structure of optimal investment strategies under idealized conditions, and their implications for market equilibrium. Their solution by stochastic dynamic programming requires specialized assumptions about investors' preferences and the structure of the economy, and quickly becomes impossible when market imperfections and other constraints are added.

Dynamic portfolio optimization models in discrete time offer substantial flexibility to describe a large variety of realistic portfolio investment problems. Their application in practice has been hampered by the fact that the numerical methods used to solve them impose severe limitations on the size of such models. The development of new algorithms, however, particularly in the area of stochastic programming, together with the continuing advances in computer technology have already shown a tremendous increase in the size of the models that can be solved. We will therefore focus our attention on stochastic programming models for the asset/liability management problem that was characterized by the assumptions in section 1.1.

The power of stochastic programming models to describe realistic decision problems under uncertainty has not only been recognized for financial applications, but for a host of other application areas as well. Ermoliev and Wets [17] provides an introduction to stochastic programming models and methods in general, and describes many applications in a variety of areas.

1.3 Thesis Overview

In the next chapter we will present a multistage stochastic programming formulation for the ALM problem that was characterized by the assumptions in section 1.1, and it will become clear that the problem naturally fits in this framework. This formulation will be referred to as the *ALM model.*

As we saw in section 1.2, the specification of the asset-price uncertainty in stochastic programming models for portfolio management has not received much attention in the literature, and its effect on the optimal solution has never been thoroughly

studied. We will show in the second part of chapter 2 that this effect can be very significant. In particular, we will show that if asset prices in the ALM model are not arbitrage-free, then the optimal solution to the ALM model may be biased substantially towards the resulting arbitrage opportunities. This is true even if the investor in the model cannot take advantage of arbitrage opportunities directly because of market frictions and trading restrictions.

Besides being important, the requirement of arbitrage-free asset prices in the model is also a very reasonable one: it is unrealistic to assume that any investor can predict arbitrage opportunities that will arise in the future. We will therefore impose this requirement on the specification of the uncertainty in the ALM model.

In chapter 3 we will discuss how financial models of the term-structure uncertainty can be used to obtain a description of the asset-price uncertainty that satisfies this restriction. The primary purpose of these models has been to value interest-rate-derivative securities in a way that precludes arbitrage opportunities. To obtain accurate security price estimates, however, these models must include a detailed description of the uncertainty in the future term structure of interest rates. We will show that this description is much too detailed to include in a stochastic program without losing the ability to solve it. It is therefore necessary to approximate the description of the uncertainty that follows from a term-structure model before we can use it in the ALM model.

In performing such an approximation, we want the asset prices in the approximate description to remain arbitrage-free. Furthermore, if the security prices that follow from a term-structure model are consistent with observed market prices (it will be assumed that such a term-structure model exists), we want to maintain this consistency in the approximate description. We will show that several intuitive approximations violate one or both of these requirements. The state and time aggregation methods that will be introduced at the end of chapter 3 avoid the pitfalls of these intuitive approximations, and are guaranteed to maintain both properties.

The state and time aggregation methods reduce the number of states, and thereby the number of scenarios, in an event tree that describes the uncertainty in future interest rates and asset prices. By repeated application of these methods one can reduce the number of scenarios to any desired number (in the extreme, to one expected-value scenario) without losing the consistency with observed asset prices or the property that asset prices in the aggregated event tree are arbitrage-free. This enables us to bring the size of the ALM model down to any desired level. The optimal solution

to the ALM model, however, will in general depend on the level of uncertainty that is included in the model. To obtain a robust investment portfolio, it is therefore important to include as much of the relevant uncertainty in the model as possible.

In chapter 4 we will present an iterative solution algorithm in which one gradually increases the level of uncertainty in the ALM model by reversing state and time aggregations that were performed to obtain the initial version of the model. The iterative nature of the algorithm enables one to judge where the uncertainty in the future affects the optimal solution most, and in what fashion. In addition, when more aggregations are reversed, events of decreasing probability are introduced in the model, and this allows for a direct trade-off between the cost of the asset portfolio that hedges against the future liabilities and the probability of events that one wants to be hedged against.

Adding uncertainty to the ALM model by reversing a state or time aggregation in the underlying event tree corresponds simultaneously to a relaxation (addition of variables) and a restriction (addition of constraints) of the stochastic program. This implies that an optimal solution in one iteration of the iterative disaggregation algorithm may not be feasible in the next iteration. We will show, however, that one can always construct a feasible solution to a relaxation of the ALM model. Furthermore, by choosing appropriate parameter values in this relaxed model, we will show that this relaxation has the same optimal solutions as the true model.

In chapter 5 we will discuss decomposition methods for the re-optimization of the ALM model in each iteration of the iterative disaggregation algorithm. The most widely used decomposition method for stochastic programs is (nested) Benders' decomposition. We will show, however, that it is not an efficient method to perform the re-optimizations of the ALM model. This is due to the fact that the ALM model after a state or time disaggregation is not just a restriction of the model before the disaggregation. This invalidates the Benders' cuts that are obtained in one iteration for the ALM model of the next iteration. We will also discuss a different decomposition method, primal-dual decomposition, and show that it has a better ability to use the optimal solution to the ALM model before a disaggregation to find the solution for the model after a disaggregation.

Computational results from the application of the iterative disaggregation algorithm to the solution of a simple asset/liability management problem will be presented in chapter 6. Chapter 7 contains conclusions, and discusses several possible extensions of the models and the solution methodology in this thesis.

Chapter 2

Stochastic Programming Models for Asset/Liability Management

In this chapter, we will formulate the asset/liability management problem under interest-rate uncertainty that was characterized by the assumptions in section 1.1 (the *ALM problem*) as a multistage stochastic linear programming model (the *ALM model*). In the stochastic programming formulation, decisions can only be made at a finite number of discrete points in time, and the uncertainty must be represented as an event tree with a finite number of possible events at each time. In section 2.1 we will introduce terminology and notation for event trees, which will be used throughout this thesis. The ALM model will be presented in section 2.2. We will also mention several possible variations on this formulation, based on models that have been proposed in the literature.

As we indicated in section 1.2, little attention has been given in the literature to the question how the uncertainty in future interest rates and asset prices should be specified in stochastic programming models for optimal portfolio management. We will show in section 2.3 that a careful specification is very important in order to obtain sensible results from the model. Specifically, we will show that if the future asset prices in the model are not arbitrage-free, then the optimal solution to the ALM model can be substantially biased towards the hypothesized arbitrage opportunities in the model. This is the case even if the investor in the model cannot directly take advantage of the arbitrage opportunities because of market frictions and trading restrictions. These results form the main motivation for the use of arbitrage-free term-structure models from the financial literature to specify the uncertainty in the ALM model. This is the topic of the next chapter.

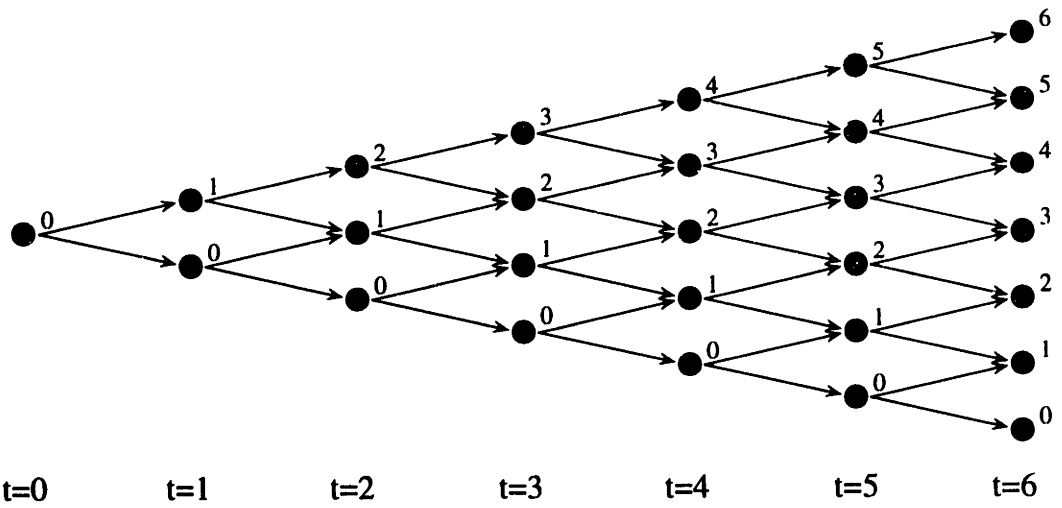

Figure 2-1: A binomial lattice.

2.1 Notation and Terminology

Let the index t denote the points in time at which *events* occur in an event tree. We assume here for simplicity that these points are equally spaced with $t = 0, \ldots, T$. Period t extends from time $t-1$ to time t. The events in an event tree are also called *states*, and correspond to the nodes in a pictorial representation of the event tree. We will refer to a state with the index n.

Figure 2-1 shows an event tree in the form of a binomial lattice, which will be used throughout this section to illustrate the notation and terminology. In a binomial lattice, there are $t+1$ possible states at time t, which are numbered 0 through t. The index number of a state at time t thus equals the number of upward movements in the lattice between time 0 and time t.

A *scenario* at time t corresponds to a *path* in the event tree from time 0 to time t. In a binomial lattice, there are 2^t different scenarios at time t, and figure 2-2 presents a way of numbering the scenarios. We will refer to a scenario by the index s. The set of all scenarios at time t is denoted by $\mathcal{S}_t$, and the set of all scenarios that visit node n at time t by $\mathcal{S}_t^n$. The set $\mathcal{S}_3^2$ thus contains all scenarios that visit node 2 at time 3, which are the scenarios 4 through 6 in figure 2-2. For a scenario s at time t, we denote the corresponding node at time t as $n(s)$.

For each scenario s at time t there is a unique scenario at time $t-1$ that follows the same path in the event tree up to time $t-1$. We will call this scenario at time $t-1$ the *predecessor* of scenario s. For example, the predecessor of scenario 5 at time 3 in

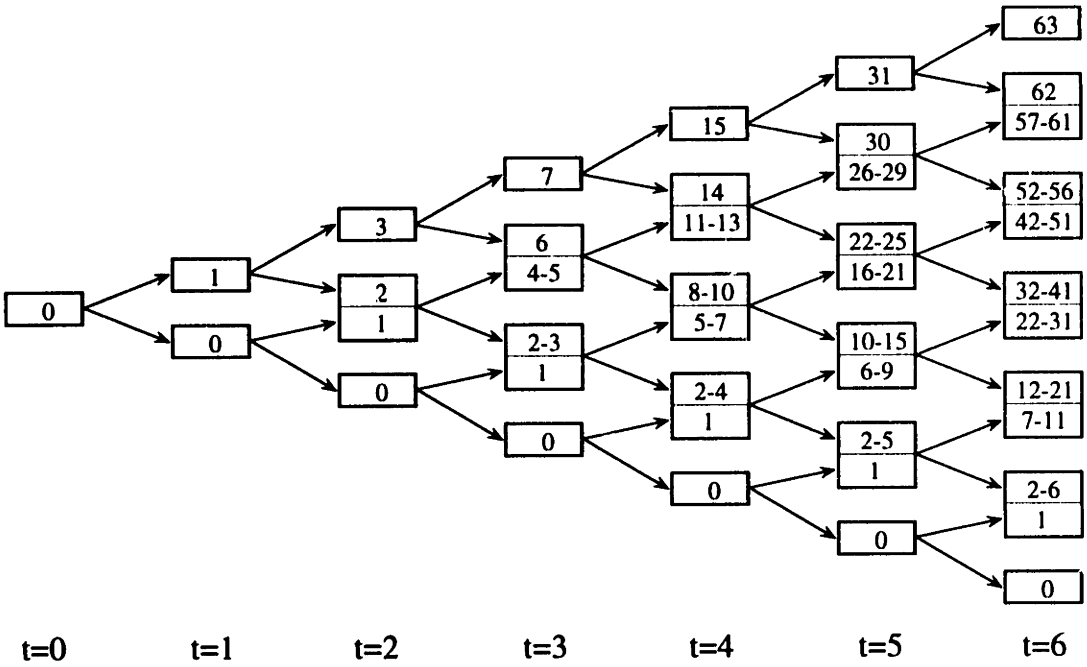

Figure 2-2: Scenario numbers in a binomial lattice.

figure 2-2 is scenario 2 at time 2. We will often denote the predecessor of a scenario s by s^-.

Furthermore, for each scenario s at time t there are several scenarios at time $t+1$ that share scenario s up to time t. We will call these scenarios at time $t+1$ the *successors* of scenario s. Each of these successors will sometimes be denoted as s^+. For example, scenarios 9 and 12 at time 4 are the successors of scenario 5 at time 3 in figure 2-2.

The predecessor and successors of a scenario at time t refer to scenarios at time $t-1$ and $t+1$, respectively. Sometimes we want to relate scenarios at points in time that are more than one period apart. For this purpose, we introduce the terms *ancestor* and *descendant*. A scenario s at time t is the *ancestor* of scenario s' at time $\tau > t$ if these scenarios share the same history up to time t. Vice versa, a scenario s at time t is a *descendant* of scenario s' at time $\tau < t$ if these scenarios share the same history up to time τ. Note that every scenario at time t has exactly one ancestor at each time before time t, but multiple descendants at each time after t. For example, in figure 2-2 scenarios 14, 20, 23 and 27 at time 5 are descendants of scenario 5 at time 3, and this scenario has scenario 1 as its ancestor at time 1.

We will let $\mathcal{D}_\tau(s,t)$ denote the set of all descendants at time $\tau > t$ of scenario s at time t. In a binomial lattice, every scenario at time t has $2^{\tau-t}$ descendants at time

$\tau > t$. For $\tau < t$ we use the notation $\mathcal{A}_\tau(n,t)$ to denote the set of all scenarios at time τ that are ancestors of scenarios that visit node n at time t. That is, $\mathcal{A}_\tau(n,t)$ contains all scenarios at time τ that are ancestors of scenarios in the set $\mathcal{S}_t^n$. In figure 2-2 we have for node 3 at time 4: $\mathcal{A}_3(3,4) = \{4,5,6,7\}$, $\mathcal{A}_2(3,4) = \{1,2,3\}$, $\mathcal{A}_1(3,4) = \{0,1\}$, and $\mathcal{A}_0(3,4) = \{0\}$.

We will use the terms predecessor, successor, ascendant and descendant with respect to nodes in the event tree as well, and their meaning in that case is analogous to the ones just described for scenarios. Notice, however, that a node can have more than one predecessor when the event tree has a lattice structure, whereas a scenario always has just one.

2.2 A Stochastic Programming Formulation for the ALM Problem

In section 1.1 we have made assumptions about the investment environment of the investor in the ALM problem. These assumptions, together with the additional assumption that will be made in section 3.2, form the basis for the model development and solution methodology in this thesis. To actually formulate the ALM problem as a mathematical optimization program, however, we need to make additional assumptions about the specific situation and preferences of the investor. We will do so below, but note that they are by no means crucial to our subsequent development, and primarily serve to make our discussion concrete. After presenting the basic formulation in the section 2.2.1, we will discuss in section 2.2.2 how alternative assumptions about the situation and preferences of the investor can be accommodated in the model formulation.

2.2.1 The ALM Model

In asset/liability management one generally faces a trade-off between the initial cost of the asset portfolio whose payoffs must match the liabilities, and the remaining portfolio value at the end of the model horizon. We will assume here that no shortfalls are allowed in meeting the stream of liabilities, or stated differently, that it is very expensive for the investor not to meet his liabilities at any point in time. The investor has the option, however, to borrow short-term if the asset cash flows fall short of the liabilities. Short-term borrowing creates in effect an extra liability in the next

period, and thus allows for a redistribution of the liabilities over time. The trade-off between the initial investment and the value of the portfolio at the planning horizon is captured in the objective function: the initial portfolio investment is minimized, but any positive or negative final portfolio value is credited to, respectively penalized in, the objective.

Following assumption 3 in section 1.1, the formulation includes proportional transaction costs, limits on borrowing and short sales of assets, and a difference (*spread*) between interest rates for short-term borrowing and lending. However, the formulation does not include side constraints that could be imposed in practical applications, and would depend on the application context. Examples are legal restrictions and constraints that reflect management policy. It should become clear that many constraints of this type can be added to the formulation without changing its basic structure or impairing the application of the solution methods that will be the topic of later chapters. Other constraints, however, that are imposed in simpler models to control the riskiness of the solution should be much less necessary in a stochastic programming model as risk (uncertainty) is included explicitly in the model.

Let H denote the planning horizon of the investor (in years). The investor can revise his portfolio at the beginning of each of T periods, and we assume here for simplicity that the length of every period is the same, denoted as $\Delta \equiv H/T$. Let L_t^n denote the liability that is due at the end of period t if state n occurs. Let D_t^n be the vector of dividends paid (comprising coupon, principal and other payments) on all securities at the end of period t if state n occurs, and S_t^n the vector of ex-dividend security prices in state n at time t. (We assume that liabilities are due and dividends are paid at the end of a period only.) The riskless one-period interest rate (continuously compounded and annualized) in state n at time t is denoted as r_t^n, and we define the discount factor $P_t^n \equiv \exp(-r_t^n \Delta)$; P_t^n can be interpreted as the price in state n at time t of a riskless one-period zero-coupon bond that pays one dollar at the end of period t. The interest-rate spread (continuously compounded and annualized) between the investor's one-period borrowing rate and r_t^n is assumed to be constant through time and denoted by ρ. The upper bound on one-period borrowing for the investor in state n at time $t < T$ is written as $\bar{Z}_t^n$, while $\bar{Z}_T^n$ denotes the upper bound on a negative final portfolio value in state n at time T. The proportional transaction cost rate c is assumed to apply to both purchases and sales of securities, but not to one-period borrowing or lending.

Although we have assumed that the liabilities as well as the asset prices and

dividends are state dependent and not path dependent, trading strategies will in general be path dependent. To illustrate this, consider an event tree in the form of a binomial lattice, and an investor who chooses to invest in one-period discount bonds only. If the sequence of events is an "upstate" at $t = 1$ followed by a "downstate" at $t = 2$, then his wealth at $t = 2$ will be equal to $1/(P_0 P_1^1)$ for every dollar invested at $t = 0$. However, if the sequence of events would have been a "downstate" followed by an "upstate" (resulting in the same state at $t = 2$ in the lattice) then his wealth would have been equal to $1/(P_0 P_1^0)$ at $t = 2$ per dollar invested at $t = 0$. As $P_1^0 \neq P_1^1$ in general, the investor's wealth at time 2 depends on the sequence of events leading up to that time.

In the special case that markets are dynamically complete and frictionless, and when the investor's objective is to minimize the initial cost of a self-financing trading strategy whose payoffs match or exceed the liabilities (i.e., final portfolio values are ignored in the objective function), then the arguments in appendix A show that an optimal trading strategy exists that is path *in*dependent. (Notice that this problem is analogous to the valuation of a security by arbitrage, where the liabilities correspond to the security payoffs.) Cox and Huang [10] have shown that, under the same market conditions, this path independence property also holds in optimal consumption/investment problems in which the investor has a fixed investment budget, and maximizes state-dependent utility of intertemporal consumption and final wealth. They solve the optimal consumption/investment problem in two steps. First, the investor's optimal consumption pattern over time is determined, given his initial budget. In the second step, a self-financing trading strategy is found that finances this consumption pattern, and this step thus resembles the earlier cost minimization problem. He and Pearson [24, 25] have shown that this approach can be extended to the case of incomplete markets and short-sale constraints. If transaction costs exist, however, this two-step solution procedure can no longer be applied, and the trading strategies will become path dependent. Edirisinghe, Naik and Uppal [15, pg.123] illustrate this with a simple example for an investor facing proportional transaction costs.

The variables in the optimization model, corresponding to the initial and future portfolio decisions, are therefore path dependent (also called *scenario* dependent) instead of just state dependent. We use the following variables:

xb_t^s = vector of asset purchases in scenario s at time t.

xs_t^s = vector of asset sales in scenario s at time t.

xh_t^s = vector of asset holdings in scenario s at time t *after* rebalancing (i.e., the portfolio holdings during period $t+1$).

y_t^s = if $t < T$, amount available at time $t+1$ from one-period lending in scenario s at time t;
if $t = T$, value of final portfolio if positive (0 otherwise).

z_t^s = if $t < T$, amount due at time $t+1$ from one-period borrowing in scenario s at time t;
if $t = T$, value of final portfolio if negative (0 otherwise).

The variables xb_t^s, xs_t^s, y_t^s and z_t^s are nonnegative by definition. If no short sales are allowed, xh_t^s must also be nonnegative.

Notice that the variables y_t^s and z_t^s do not equal the amount of short-term lending, resp. borrowing, in scenario s at time t, but the amounts that will be received, resp. have to be paid back, in the successor scenarios at time $t+1$. Thus, the actual amounts of short-term lending and borrowing in scenario s at time t are $P_t^{n(s)} y_t^s$ and $(e^{-\rho\Delta} P_t^{n(s)}) z_t^s$, respectively. This definition of y_t^s and z_t^s will simplify the analysis in later chapters.

The objective function of the investor that was discussed earlier can now be stated mathematically as

$$v = (1+c) S_0 xh_0 + P_0 y_0 - \left(e^{-\rho\Delta} P_0\right) z_0 - \lambda_1 \sum_{s \in S_T} q_T^s (y_T^s - \lambda_2 z_T^s) \tag{2.1}$$

which has to be minimized. The first three terms capture the cost of the initial asset portfolio, where it is assumed that the investor does not start with an initial portfolio (i.e., $xh_0 = xb_0$). The first term represents the cost, including transaction costs, of investing in the available assets, while the second term is the amount of short-term lending. The investor is allowed to borrow at time 0 (to be paid back at time 1), and the third term in the objective equals the borrowed amount. Short-term borrowing is limited by the constraint $z_0 \leq \bar{Z}_0$.

The last term in the objective function represents the expected present value of the final portfolio, with λ_1 and λ_2 as weights. The coefficient q_T^s is a probability weighted present-value factor for the final portfolio value in scenario s at time T. We postpone a discussion of the specific definition of q_T^s to section 2.3.1. The parameter $\lambda_1 \leq 1$ represents a (subjective) weight on the total final portfolio value, and $\lambda_2 \geq 1$ an extra weight on negative final portfolio values.

At every point in time $t = 1, \ldots, T-1$, we impose the following constraints for each scenario $s \in \mathcal{S}_t$:

$$D_t^{n(s)} xh_{t-1}^{s^-} + y_{t-1}^{s^-} - z_{t-1}^{s^-} + (1-c) S_t^{n(s)} xs_t^s - (1+c) S_t^{n(s)} xb_t^s - P_t^{n(s)} y_t^s + \left(e^{-\rho\Delta} P_t^{n(s)}\right) z_t^s = L_t^{n(s)} \quad (2.2)$$

$$xh_{t-1}^s - xs_t^s + xb_t^s - xh_t^s = 0 \quad (2.3)$$

$$z_t^s \leq \bar{Z}_t^{n(s)} \quad (2.4)$$

The first constraint makes sure that sufficient cash flow is generated to meet the liability. This constraint will be referred to as the *cash-balance* constraint. The first three terms in this constraint represent the net cash flow from the portfolio in the previous period: dividend payments on the assets plus the return on short-term lending, and minus the required payment for short-term borrowing. The next two terms reflect rebalancing of the asset portfolio: revenues are generated by selling assets, and money can be invested by buying assets, where both are adjusted for transaction costs. The final two terms on the left-hand side equal the amounts of short-term lending and borrowing, respectively.

The second constraint (2.3) represents a set of constraints, one for each asset, that links the portfolio holdings in the previous and the current period (i.e., before and after rebalancing). These wil be called *portfolio-balance* constraints. The last constraint (2.4) imposes an upper bound on short-term borrowing, where we note that the upper bound is specified on the amount that has to be paid back in the next period. This constraint will be referred to as the *borrowing* constraint.

For every scenario s at the planning horizon the final portfolio value $(y_T^s - z_T^s)$ is determined through the constraints:

$$(D_T^{n(s)} + S_T^{n(s)}) xh_{T-1}^{s^-} + y_{T-1}^{s^-} - z_{T-1}^{s^-} - y_T^s + z_T^s = L_T^{n(s)} \quad (2.5)$$

$$z_T^s \leq \bar{Z}_T^{n(s)} \quad (2.6)$$

The first three terms in the first constraint, the cash-balance constraint, determine the final portfolio value before meeting the liability. The portfolio holdings are converted at the current market prices, the return on short-term lending is added and the required payment on short-term borrowing subtracted. If this value is higher than the liability $L_T^{n(s)}$, then $y_T^s \geq 0$ and $z_T^{n(s)} = 0$; otherwise $y_T^{n(s)} = 0$ and $z_T^{n(s)} \geq 0$ (note that we need $\lambda_2 > 1$ in the objective function to prevent both y_T^s and z_T^s from being positive in an optimal solution). Constraint (2.6) imposes an upper bound on the

negative final portfolio value z_T^s.

This completes the mathematical formulation of the ALM problem. We have assumed that both the objective function and the constraints are linear, so the formulation constitutes a linear program. Because the model describes a multiperiod problem in which every two successive periods are linked, the linear program is generally called a *multistage* linear program. Furthermore, the constraints at time t have the same structural form for every scenario s, and differ only in the values of the coefficients and the right-hand sides. These coefficients and right-hand sides are random variables whose values were assumed to depend on the (random) state of nature at time t. The formulation for the ALM problem is therefore called a multistage *stochastic* linear program.

Although we have assumed that all data (asset prices, dividends, and liabilities) are state dependent, it is clear that path-dependent data would not change the formulation as constraints are included for every single scenario. We further note that the stochastic program treats all scenarios equally in the sense that it forces the liabilities to be met exactly in every scenario, *irrespective* of their relative likelihood of occurrence. The only place in the formulation where the scenario probabilities can appear is as part of the probability-weighted discount factors q_T^s for the final portfolio values in the objective.

Finally note that this formulation of the ALM problem always has a feasible solution. As no upper bound is imposed on the initial investment, it is possible to meet every future liability by investing enough in a short-term lending strategy.

2.2.2 Variations on the Formulation

This section discusses several possible modifications to the formulation of the ALM problem in the previous section, reflecting alternative assumptions about the specific situation and preferences of the investor.

Edirisinghe, Naik and Uppal [15] consider a special case of the formulation in which $\lambda_1 = 0$ and $\bar{Z}_T^s = 0$ for all scenarios s at the terminal date. They thus require that the final portfolio value after meeting the liabilities is nonnegative in all scenarios, but assume that the investor places no value on any positive value.

If the investor owns an initial portfolio of securities, then constraints (2.2)–(2.4) can be used at time 0 with $\bar{x}_{-1}$, y_{-1} and z_{-1} representing the existing portfolio (which are thus constants). Minimization of the total cost of the portfolio as in (2.1) then comes down to the minimization of the additionally required investment

$$(1+c)S_0 xb_0 - (1-c)S_0 xs_0 + P_0 y_0 - \left(e^{-\rho\Delta} P_0\right) z_0$$

This formulation is considered by Hiller and Shapiro [30].

Many consumption/investment problems assume that the investor has an initial budget W available for investment at time 0, and that he wants to maximize his expected utility of final wealth. The objective would thus be to maximize $E\{U(\tilde{y}_T - \tilde{z}_T)\}$ where $U(\cdot)$ is the investor's objective function of final wealth, and the expectation is taken with respect to the investor's probability beliefs about the states of nature at time T. $(\tilde{y}_T - \tilde{z}_T)$ represents the random value of the final portfolio. The constraints at time 0 would be:

$$(1+c)S_0 xh_0 + P_0 y_0 - e^{-\rho\Delta} P_0 z_0 = W$$
$$z_0 \leq \bar{Z}_0$$

Typically, it is assumed that the investor is risk averse, and thus that $U(\cdot)$ is a concave function.

Mulvey and Vladimirou [46] discuss the use of nonlinear utility functions in their stochastic network programming models. To keep the models linear, one can approximate the utility function by a piecewise linear function. Bradley and Crane [7] just maximize the expected value of the portfolio at the terminal date, and incorporate the risk attitudes of the investor through additional (linear) constraints.

We have assumed that the liabilities have to be met exactly at each point in time and in all scenarios. However, short-term borrowing and lending is allowed so that cash shortages at one point in time can be offset by surpl:sses at other points. Hiller and Schaack [29] and Zipkin [59] also follow this approach. Kusy and Ziemba [40] and Cariño et al. [9] use shortfall and surplus variables instead. Any shortfall in the matching between assets and liabilities is directly penalized in the objective, and surplusses are credited to the objective. The main difference with our approach is thus that shortfalls and surplusses appear in the objective function directly, while they get incorporated in the initial cost of the portfolio strategy or the final portfolio value in our approach. Shortfall and surplus variables can be used in other types of constraints as well (e.g., policy and regulatory constraints), and the associated penalties can reflect the likelihood of the scenarios. Although estimation of the penalties may not be straightforward, it should be clear that they can be easily accommodated in our formulation.

Specification of the model horizon H is somewhat arbitrary in many applications

as there is often no natural terminal date for the asset/liability management problem. In the choice for H one has to trade off the additional realism of the model when H is large with the increase in computational complexity. Grinold [20] shows for general multistage decision problems that reducing the length of the planning horizon in the model can significantly influence the optimal solution, and compares several ways to deal with this problem of so-called "end effects". His favorite method, the dual equilibrium technique, adds an additional period to the model for which the constraints and the variables represent a discounted weighted average of the constraints and variables of all periods after H that were not included in the model. Cariño et al. [9] discuss the application of this technique in their stochastic programming model for asset/liability management. We handle possible end effects in our formulation by including the final portfolio value in the objective function with a certain weight (corresponding to the "salvage" technique of Grinold).

Bradley and Crane [7] and Kusy and Ziemba [40] include capital gains taxes in their models. As these taxes are levied on the difference between the sales price and the original purchase price of an asset, a separate variable must be defined for every possible holding period of an asset. This significantly increases the number of variables in the model. Furthermore, it implies that the constraints would not only link the variables of a scenario s at time t to the variables of its predecessor scenario s^- at time $t-1$, but to the variables of all of its ascendent scenarios before time $t-1$ as well. Although we perform our analysis in this and the next chapters for the ALM model of section 2.2.1, it should not be difficult to extend the results to this alternative formulation.

2.3 Specification of the Data in the ALM Model

In section 1.1 we made the assumption that asset prices in financial markets are arbitrage-free, i.e., they do not enable investors to make riskless, unlimited profits (see appendix A for a precise definition of arbitrage-free asset prices). Although this assumption may not always be fully satisfied in practice, we argued that it is a reasonable approximation of reality as there are many firms in today's financial markets which have as their main business to look for, and take advantage of, violations of arbitrage-pricing relationships.

For the specification of the asset-price uncertainty in the ALM model we will require that *future* asset prices in the model are arbitrage-free as well. This requirement

is even more reasonable than the earlier assumption, as it is unrealistic to assume that any investor can forecast future mispricings of assets that would enable riskless arbitrage profits. Besides being reasonable, we will show in section 2.3.2 that it is also an important requirement in order to obtain sensible solutions from the ALM model. Specifically, we will show that the optimal solution to the model can be substantially biased towards the hypothesized arbitrage opportunities if asset prices are not arbitrage-free in the model. This is the case even if the investor is unable to take advantage of the arbitrage opportunities directly because of market frictions and trading restrictions. We will first, however, give a mathematical characterization of arbitrage-free asset prices, and show in section 2.3.1 that this property together with the assumption of dynamically complete markets implies a natural definition of the discount factors q_T^s in the ALM model from the previous section.

From theorem A.1 in appendix A we know that the property of arbitrage-free asset prices implies that there must exist a probability measure on the event tree (not necessarily representing the investor's probability beliefs) such that the expected one-period return on all assets, calculated with respect to this probability measure, equals the riskless one-period return in all states of the event tree. Let π_t^s denote the unconditional probability of scenario s at time t according to this *risk-neutral* probability measure, and $\bar{\pi}_t^s$ its conditional probability, given its predecessor scenario s^- at time $t-1$. That is, $\bar{\pi}_t^s \equiv \pi_t^s / \pi_{t-1}^{s^-}$. We note that this conditional probability only depends on the state of scenario s at time t, and thus $\bar{\pi}_t^{n(s)} = \bar{\pi}_t^s$. The fact that asset prices are arbitrage-free in the event tree can now be formulated mathematically as:

$$S_{i,t}^n = P_t^n \left(\sum_{n^+} \hat{\pi}_{t+1}^{n^+} \left(S_{i,t+1}^{n^+} + D_{i,t+1}^{n^+} \right) \right) \tag{2.7}$$

where the summation is over all successor nodes n^+ of node n at time t. This relation must hold for all assets $i = 1, \ldots, I$, and in every node n at each time $t = 0, \ldots, T-1$ in the event tree.

2.3.1 Definition of the Present-Value Factors

In addition to the assumption of arbitrage-free asset prices, we have also made the assumption in section 1.1 that financial markets are dynamically complete. This implies that the risk-neutral probabilities π_t^s are unique (see proposition A.2 in appendix A). We will argue in this section that both assumptions imply the following natural definition for the present-value factors q_T^s in the ALM model:

$$q_T^s = \pi_T^s \left(\prod_{t=0}^{T-1} P_t^{n(s^-)} \right) \tag{2.8}$$

The term between brackets is the discount factor, which equals the product of the one-period risk-free discount factors along the path in the event tree that corresponds to scenario s. A final portfolio value (positive or negative) is thus discounted to time 0 using the one-period interest rates along the scenario path in the event tree, and weighted by its risk-neutral probability of occurrence.

To argue that this is a reasonable definition, first consider the set of scenarios at time T in which the final portfolio value is positive (i.e., all $s \in \mathcal{S}_T$ for which $y_T^s > 0$). Because of the assumption of dynamically complete markets, it is possible in principle to construct a self-financing trading strategy in the available securities that provides payoffs of y_T^s at time T in all scenarios $s \in \mathcal{S}_T$ for which $y_T^s > 0$, and zero in all other scenarios. It can thus be viewed as if the investor owns a marketed security with random payoffs y_T^s at time T.

As security prices are assumed to be arbitrage-free, a natural candidate for the discount factor q_T^s is the weight on y_T^s in the arbitrage-free value of this hypothetical security at time 0. By a recursive application of relation (2.7), it is not difficult to see that q_T^s in (2.8) is exactly this weight. As the arbitrage-free value of this hypothetical security at time 0 may not reflect the actual value that the investor assigns to the random surplus portfolio value at the terminal date, the parameter λ_1 is used for adjustment.

The argument that deficits z_T^s at time T should also be discounted with the discount factor q_T^s in equation (2.8) is similar. It can be viewed as if the investor is *short* a hypothetical security that pays z_T^s in scenario $s \in \mathcal{S}_T$. In dynamically complete and arbitrage-free markets it is natural to value this hypothetical security by arbitrage, which leads to the discount factor of equation (2.8). To correct for the fact that the investor may attach a value to the random portfolio deficit at time T that is different from its arbitrage-free value, we have included the parameter λ_2 in the formulation.

The choice of values for the parameters λ_1 and λ_2 is investor dependent. As we have assumed that the investor's main goal is the construction of a portfolio strategy in order to meet his future liabilities, it is reasonable to assume $\lambda_1 \leq 1$ and $(\lambda_1 \lambda_2) \geq 1$. An interpretation in terms of the hypothetical securities that we discussed earlier is that the investor would not be able to sell his (scenario-dependent) final portfolio surplus at time T at more than its arbitrage-free value at time 0, whereas he would have to pay more at time 0 than the arbitrage-free value of his deficits at time T to

cover this future short position.

The choice of λ_1 and λ_2 as simple constants is a particularly simple one, and can be easily generalized. A straightforward generalization is to make them state or scenario dependent. Instead of being a linear operator, λ_1 could also be replaced by a piecewise linear, increasing and concave function that is evaluated for the surplus value in every single scenario. This would reflect a situation in which differences in the portfolio surplus across scenarios negatively influences the value that the investor attaches (at time 0) to a portfolio surplus at time T. Similarly, the product $(\lambda_1\lambda_2)$ could be replaced by a piecewise linear, increasing and convex function with an analogous interpretation. The use of such piecewise linear functions is similar in spirit to calculating the expected utility of the final portfolio value with respect to a concave utility function.

2.3.2 Effects on the Optimal Solution

In this section we will show that a violation of the no-arbitrage condition (2.7) by the asset prices in the ALM model can lead to a substantial bias in its optimal solution. This bias is caused by spurious profit opportunities that are introduced in the model by a violation of condition (2.7). To perform our analysis, we will first make some simplifying assumptions about the parameters in the ALM model, and then discuss the generalization of the results when these assumptions are relaxed. The following lemma will play an important role.

Lemma 2.1 *If $\lambda_1 = 1$ in the ALM model, and if q_T^s is defined by* (2.8)*, then the only possible value for the dual variable φ_t^s on the cash-balance constraint for scenario s at time t in an optimal solution to the ALM model is given by:*

$$\varphi_t^s = q_t^s \qquad \textit{with} \qquad q_t^s \equiv \pi_t^s \left(\prod_{\tau=0}^{t-1} P_\tau^{n(s^-)} \right) \tag{2.9}$$

for all $t = 1, \ldots, T$ and $s \in \mathcal{S}_t$.

PROOF: The result follows from the constraints in the dual formulation of the ALM model that correspond to the variables y_t^s for all $t = 0, \ldots, T$ and $s \in \mathcal{S}_t$. These dual constraints are (the corresponding primal variable is listed at the beginning of each constraint):

$$y_0 \;:\; \sum_{s \in \mathcal{S}_1} \varphi_1^s \leq P_0 \tag{2.10}$$

$$y_t^s \;:\; -\varphi_t^s\, P_t^{n(s)} + \sum_{s^+} \varphi_{t+1}^{s^+} \le 0 \tag{2.11}$$

$$y_T^s \;:\; -\varphi_T^s \le -q_T^s \tag{2.12}$$

where we have used $\lambda_1 = 1$ in the last constraint. We will first prove that $\varphi_t^s \ge q_t^s$ for all $t = 1, \ldots, T$ and $s \in \mathcal{S}_t$, and subsequently that equality must hold.

Constraint (2.12) states $\varphi_t^s \ge q_t^s$ for $t = T$. We will prove that it also holds for $t < T$ by induction. Assume the inequality holds for all $t = t^* + 1, \ldots, T$ and $s \in \mathcal{S}_t$. To show that it also holds for $t = t^*$, note that constraint (2.11) for $t = t^*$ implies:

$$\varphi_{t^*}^s\, P_{t^*}^{n(s)} \ge \sum_{s^+} \varphi_{t^*+1}^{s^+} \tag{2.13}$$

where the summation is over all successor scenarios s^+ of s. By induction, $\varphi_{t^*+1}^{s^+} \ge q_{t^*+1}^{s^+}$, and thus

$$\varphi_{t^*}^s\, P_{t^*}^{n(s)} \ge \sum_{s^+} q_{t^*+1}^{s^+} \tag{2.14}$$

By definition, $q_{t^*+1}^{s^+} = q_{t^*}^s P_{t^*}^{n(s)} \hat{\pi}_{t^*+1}^{n(s^+)}$ for each successor s^+ of s. Substituting this in (2.14), and noting that $\sum_{s^+} \hat{\pi}_{t^*+1}^{n(s^+)} = 1$, we obtain $\varphi_{t^*}^s \ge q_{t^*}^s$. Thus $\varphi_t^s \ge q_t^s$ for all $t = 1, \ldots, T$ and $s \in \mathcal{S}_t$.

We will now show that equality holds, again by induction. Because $\varphi_1^s \ge q_1^s$ for all $s \in \mathcal{S}_1$, it follows from constraint (2.10) that:

$$\sum_{s \in \mathcal{S}_1} q_1^s \le \sum_{s \in \mathcal{S}_1} \varphi_1^s \le P_0 \tag{2.15}$$

By definition, $q_1^s = P_0 \pi_1^s$. Because $\sum_{s \in \mathcal{S}_1} \pi_1^s = 1$, it follows that equality must hold throughout in (2.15), and thus $\varphi_1^s = q_1^s$ for all $s \in \mathcal{S}_1$. This establishes (2.9) for $t = 1$.

To show that the equality also holds for all $t > 1$, we will prove that $\varphi_{t+1}^s = q_{t+1}^s$ must hold for each $s \in \mathcal{S}_{t+1}$ if $\varphi_t^s = q_t^s$ for all $s \in \mathcal{S}_t$. If the latter equality holds, then constraint (2.11) implies for an arbitrary scenario $s \in \mathcal{S}_t$:

$$\sum_{s^+} q_{t+1}^{s^+} \le \sum_{s^+} \varphi_{t+1}^{s^+} \le P_t^{n(s)} \varphi_t^s = P_t^{n(s)} q_t^s \tag{2.16}$$

where we have used the earlier result that $\varphi_{t+1}^{s^+} \ge q_{t+1}^{s^+}$ for all successors s^+ of s. Because $q_{t+1}^{s^+} = q_t^s P_t^{n(s)} \hat{\pi}_{t+1}^{n(s^+)}$ for each s^+ by definition, and as $\sum_{s^+} \hat{\pi}_{t+1}^{n(s^+)} = 1$, we see that equality must hold throughout in (2.16). This implies $\varphi_{t+1}^{s^+} = q_{t+1}^{s^+}$ for each s^+. Because scenario $s \in \mathcal{S}_t$ was chosen arbitrarily, it follows that $\varphi_{t+1}^s = q_{t+1}^s$ for all

$s \in \mathcal{S}_{t+1}$ if $\varphi_t^s = q_t^s$ for all $s \in \mathcal{S}_t$. By induction therefore $\varphi_t^s = q_t^s$ for all $t = 1, \ldots, T$ and $s \in \mathcal{S}_t$.

$\mathcal{QED}$

The interpretation of the quantity q_t^s in (2.9) is analogous to the interpretation of q_T^s in (2.8), namely the arbitrage-free value at time 0 of a dollar in scenario s at time t. Notice that this lemma does not require any assumptions about the ALM model beyond the ones that were explicitly stated. That is, the result is true for any level of the transactions cost rate c and the interest-rate differential ρ, for any choice of λ_2, and for arbitrary upper bounds on short-term borrowing. The assumption $\lambda_1 = 1$ corresponds to an investor who has no time preference for money, i.e., he is indifferent between one dollar now and a random payment at time T whose arbitrage-free value at time 0 equals one dollar.

Notice also that the lemma restricts the *possible* value for φ_t^s in an optimal solution to the dual of the ALM model. That is, the dual may not have an optimal solution at all, implying that the ALM model itself is unbounded. An unbounded solution to the ALM model is obviously an undesirable outcome, as it means that the investor would generate unlimited *benefits* (as measured by the objective function) from meeting his liabilities.

One way in which the ALM model will have an unbounded solution is if $\lambda_1 > 1$, which follows directly from the proof of the lemma. We will show below that an unbounded solution can also occur if $\lambda_1 \leq 1$ and if the asset prices in the ALM model do not satisfy the no-arbitrage condition (2.7). We will first consider the case that $\lambda_1 = 1$ in proposition 2.1, where we also assume that the investor has no possibility for short-term borrowing. This assumption, together with the assumption that he cannot short sell assets, implies that he cannot directly take advantage of an arbitrage opportunity in the model by forming an arbitrage portfolio (in the sense of definition A.1 in appendix A). In proposition 2.2 we will allow for short-term borrowing, which enables us to relax the assumption $\lambda_1 = 1$.

A Special Version of the ALM Model

We will assume for now that the transaction cost rate $c = 0$, and generalize the results later to the case $c > 0$. We also impose the restriction that the final portfolio value has to be positive in all scenarios, and define the present-value factors q_T^s by (2.8). The assumption of no transaction costs allows us to eliminate the variables for asset

purchases and sales (xb_t^s and xs_t^s) from the ALM model, and write it in terms of the portfolio holdings xh_t^s only. (To see this, we note that $c = 0$ implies that xb_t^s and xs_t^s have the same coefficient in the cash-balance equation. Using the portfolio-balance equations, one can therefore substitute $(xh_{t-1}^{s^-} - xh_t^s)$ for $(xs_t^s - xb_t^s)$ in the cash-balance equation.) Instead of the vector notation that was used in section 2.2.1 we will rewrite the model in terms of individual securities and let $D_{i,t}^n$ and $S_{i,t}^n$ denote the dividend payment and ex-dividend price of security i in state n at time t. Under these assumptions, and with $\lambda_1 = 1$, the ALM problem can be written mathematically as:

$$\min \sum_{i=1}^{I} S_{i,0} xh_{i,0} + P_0 y_0 - \sum_{s \in \mathcal{S}_T} \pi_T^s \left(\prod_{t=0}^{T-1} P_t^{n(s^-)} \right) y_T^s \tag{2.17}$$

$$\text{s.t.} \quad \sum_{i=1}^{I} \left(D_{i,t}^{n(s)} + S_{i,t}^{n(s)} \right) xh_{i,t-1}^{s^-} + y_{t-1}^{s^-} - \sum_{i=1}^{I} S_{i,t}^{n(s)} xh_{i,t}^s - P_t^{n(s)} y_t^s = L_t^{n(s)}$$

$$\forall \ s \in \mathcal{S}_t, t = 1, \ldots, T-1$$

$$\sum_{i=1}^{I} \left(D_{i,T}^{n(s)} + S_{i,T}^{n(s)} \right) xh_{i,T-1}^{s^-} + y_{T-1}^{s^-} - y_T^s = L_T^{n(s)} \qquad \forall \ s \in \mathcal{S}_T$$

$$xh_{i,t}^s \geq 0, \ y_t^s \geq 0 \qquad \forall \ s \in \mathcal{S}_t, \ t = 0, \ldots, T.$$

To interpret this formulation, we note that in the case of zero transaction costs, it is sufficient to know the *value* of the portfolio in a scenario instead of the individual asset holdings because it is costless to change the portfolio composition. This is reflected in the cash-balance constraints. The first term in each cash-balance constraint represents the value of the portfolio that is carried over from the previous period, and the second term the return on short-term lending from that period. Their sum is the total available wealth, which can be used to construct a new portfolio after the liability is met. When $t = T$, the difference between the available wealth and the liability equals the final portfolio value. The objective function is the same as in the original ALM model, except that a term for short-term borrowing is not included here.

The following proposition states that the solution of the ALM model (2.17) is unbounded if there is a security and a state in the event tree in which the expected one-period return on the security (calculated with respect to the risk-neutral probability measure) exceeds the riskless one-period return, i.e., violates relation (2.7).[1]

[1] Actually, the proof does not depend on the fact that the probability π_T^s in the expression for q_T^s (note that this is the only place in the formulation where a probability appears) represents the

Proposition 2.1 *The solution to* (2.17) *is bounded if and only if the following inequality holds for every asset i in all states n at every time $t = 0, \ldots, T-1$ in the event tree:*

$$S_{i,t}^n \geq P_t^n \left(\sum_{n^+} \hat{\pi}_{t+1}^{n^+} \left(S_{i,t+1}^{n^+} + D_{i,t+1}^{n^+} \right) \right) \tag{2.18}$$

where the summation is over all successor states n^+ of state n in the event tree.

PROOF: We prove the proposition by considering the dual of (2.17). By linear programming duality, (2.17) has a bounded solution if and only if its dual is feasible. Let φ_t^s denote the dual variable for the cash-balance constraint of scenario s at time t. These are the only variables in the dual of (2.17), and because $\lambda_1 = 1$, it was shown in lemma 2.1 that $\varphi_t^s = q_t^s$ is the only possible solution to this dual problem. This was derived from the dual constraints that correspond to the variables for short-term lending y_t^s.

Now consider the dual constraints that correspond to the asset holding variables:

$$xh_{i,0} \quad : \quad \sum_{s \in \mathcal{S}_1} \varphi_1^s (D_{i,1}^{n(s)} + S_{i,1}^{n(s)}) \leq S_{i,0} \tag{2.19}$$

$$xh_{i,t}^s \quad : \quad \sum_{s^+} \varphi_{t+1}^{s^+} (D_{i,t+1}^{n(s^+)} + S_{i,t+1}^{n(s^+)}) - \varphi_t^s S_{i,t}^{n(s)} \leq 0 \tag{2.20}$$

where the corresponding primal variables are listed at the beginning of each constraint. We will show that these constraints are feasible in $\varphi_t^s = q_t^s$ if and only if relation (2.18) is satisfied.

By substitution of $\varphi_1^s = q_1^s = P_0 \pi_1^s$ in (2.19) we obtain for each $i = 1, \ldots, I$:

$$S_{i,0} \geq P_0 \sum_{s \in \mathcal{S}_1} \pi_1^s (D_{i,1}^{n(s)} + S_{i,1}^{n(s)}) \tag{2.21}$$

As every scenario $s \in \mathcal{S}_1$ corresponds to a different node in the event tree, and $\pi_1^s = \hat{\pi}_1^s$, this is precisely relation (2.18).

For $t > 0$, substitute $\varphi_t^s = q_t^s$ and $\varphi_{t+1}^{s^+} = q_{t+1}^{s^+} = q_t^s P_t^{n(s)} \hat{\pi}_{t+1}^{n(s^+)}$ for each s^+ in (2.20) to obtain

$$S_{i,t}^{n(s)} \geq P_t^{n(s)} \sum_{s^+} \hat{\pi}_{t+1}^{n(s^+)} (D_{i,t+1}^{n(s^+)} + S_{i,t+1}^{n(s^+)}) \tag{2.22}$$

---

risk-neutral probability of scenario s at time T. If π_T^s would be the probability that follows from some other probability measure on the event tree, then the proposition holds with respect to this other measure. However, as we argued in section 2.3.1, it is natural to choose π_T^s in the expression for q_T^s as the risk-neutral probability given the assumptions we have made about financial markets.

Because every successor scenario s^+ of s corresponds to a different node in the event tree, this is relation (2.18).

$\mathcal{QED}$

We note that neither the proof of lemma 2.1 nor the proof of this proposition depends on the stream of liabilities. The results are thus valid for any multiperiod investment problem that is formulated as a stochastic program.

The following corollaries immediately follow from the lemma and the proposition.

Corollary 2.1 *If inequality* (2.18) *is satisfied for all assets* $i = 1, \ldots, I$ *in every state in the event tree, then the optimal solution value of* (2.17) *equals the arbitrage-free value at time 0 of a security whose payoffs exactly match the stream of future liabilities.*

PROOF: Because $\lambda_1 = 1$ in (2.17), we know from lemma 2.1 that $\varphi_t^s = q_t^s$ is the only *possible* solution to the dual problem of (2.17). As condition (2.18) is satisfied, the previous proposition tells that this *must* be the dual solution. The value of the objective function in the dual problem is therefore equal to

$$\sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \varphi_t^s L_t^{n(s)} = \sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} \pi_t^s \left(\prod_{\tau=0}^{t-1} P_\tau^{n(s^-)} \right) L_t^{n(s)}$$

which is the arbitrage-free value at time 0 of the stream of liabilities.

$\mathcal{QED}$

Corollary 2.2 *If inequality* (2.18) *holds with equality for all assets* $i = 1, \ldots, I$ *in every state in the event tree, then every feasible solution in* (2.17) *is optimal.*

PROOF: If equality holds in (2.18) for all assets $i = 1, \ldots, I$ in every state in the event tree, then it follows from the proof of proposition 2.1 that all constraints in the dual of (2.17) will be satisfied with equality. Thus, for every feasible solution in (2.17) complementary slackness holds, and thus this solution must be optimal.

$\mathcal{QED}$

We note that the condition in this corollary is precisely the condition that asset prices are arbitrage-free.

Before qualifying these results with respect to the assumptions that were made in formulation (2.17), we will present a trading strategy that would lead to an unbounded

solution in this simplified formulation when inequality (2.18) is violated. Let the expected one-period return on asset i^* (calculated with respect to the risk-neutral probability measure) exceed the riskless one-period return in state n^* at time t^*, and consider the following self-financing trading strategy:

1. Take \$1 for investment at time 0, and invest it at the riskless one-period interest rate until time t^*.

2. For all scenarios $s \in \mathcal{S}_{t^*}^{n^*}$, invest the accumulated money in asset i^* during period $t^* + 1$. For all other scenarios, lend the accumulated money at the riskless one-period rate during period $t^* + 1$.

3. For all scenarios at time $t^* + 1$, take the accumulated money and roll it over at the riskless one-period interest rate until time T.

Through a somewhat tedious but otherwise straightforward argument, one can show that this trading strategy strictly improves the objective function of (2.17). That is, the expected present value of the payoffs from this strategy at time T, which is credited to the objective, exceeds the initial cost of this strategy (one dollar). As no limit has been imposed on how much the investor can invest at time 0, an arbitrarily large amount can be spent on this trading strategy, causing an unbounded solution value for the stochastic program. Note that this trading strategy is independent of the stream of liabilities.

Relaxation and Modification of the Assumptions

It is clear from the discussion above that a violation of condition (2.18) leads to a solution in the ALM model (2.17) that is unbounded because there is no restriction on the initial investment. It should also be clear, however, that the optimal solution will still be strongly affected by violations of condition (2.18) when a constraint on the initial budget is added to (2.17), despite the fact that it can no longer be unbounded in that case.

To derive condition (2.18) we have assumed that the transaction cost rate $c = 0$. When $c > 0$, we can use a similar argument as in the proof of proposition 2.1 to obtain the following sufficient condition on the expected one-period asset returns which prevents an unbounded solution to the ALM model:

$$(1+c)S_{i,t}^{n} \geq P_t^n \left(\sum_{n^+} \hat{\pi}_{t+1}^{n^+} \left((1-c)S_{i,t+1}^{n^+} + D_{i,t+1}^{n^+} \right) \right) \tag{2.23}$$

This condition has to be satisfied for all assets $i = 1, \ldots, I$ in every state n at each time $t = 0, \ldots, T-1$ in the event tree. Notice that it reduces to condition (2.18) for $c = 0$. However, when $c > 0$ this is only a sufficient condition for the ALM model to be bounded, while it is both sufficient and necessary when $c = 0$. This can be seen intuitively from the trading strategy, described earlier, which leads to an unbounded solution if condition (2.18) is violated for some asset i^* in some state n^* of the event tree when $c = 0$. If condition (2.23) is violated for asset i^* in state n^* when $c > 0$, then exactly the same trading strategy can be followed to obtain an unbounded solution. Notice that this trading strategy only involves a one-period investment in asset i^*, and that condition (2.23) states that the expected one-period excess return (i.e., in excess of the risk-free return) must compensate for the transaction costs incurred by buying the asset at the beginning of the period and selling it at the end. However, if condition (2.18) is violated for asset i^* in several states in the event tree, while (2.23) is satisfied in all states, it could be possible to construct a self-financing trading strategy that involves an investment in asset i^* over multiple periods, and which would lead to an unbounded solution to the ALM model.

We can strenghten our results by relaxing the assumption that no short sales of assets are allowed. If short selling is allowed, then condition (2.18) in proposition 2.1 must be satisfied with equality to prevent an unbounded solution to the ALM model in (2.17), and thus reduces to the no-arbitrage condition (2.7). If the transaction cost rate $c > 0$, then condition (2.23) together with the condition

$$(1-c)S_{i,t}^{n} \leq P_t^n \left(\sum_{n^+} \hat{\pi}_{t+1}^{n^+} \left((1+c)S_{i,t+1}^{n^+} + D_{i,t+1}^{n^+} \right) \right) \tag{2.24}$$

are sufficient (but not necessary) to prevent an unbounded solution to the ALM model.

The assumption $\lambda_1 = 1$ has been crucial in our analysis sofar. We can derive similar conditions on the asset returns for arbitrary $\lambda_1 \in (0, 1]$, however, if we allow for short-term borrowing in the ALM model. The possibility of short-term borrowing essentially enables the investor to take advantage of an arbitrage opportunity in the model if it exists. To show this formally, consider the modification of the ALM model in (2.17) where we include the possibility for unlimited short-term borrowing, while allowing an arbitrary value for λ_1:

$$\min \sum_{i=1}^{I} S_{i,0} xh_{i,0} + P_0 y_0 - \left(e^{-\rho\Delta} P_0\right) z_0 - \lambda_1 \sum_{s \in \mathcal{S}_T} \pi_T^s \left(\prod_{t=0}^{T-1} P_t^{n(s^-)}\right) y_T^s \tag{2.25}$$

$$\begin{aligned}
\text{s.t.} \quad & \sum_{i=1}^{I} \left(D_{i,t}^{n(s)} + S_{i,t}^{n(s)}\right) xh_{i,t-1}^{s^-} + y_{t-1}^{s^-} - z_{t-1}^{s^-} - \sum_{i=1}^{I} S_{i,t}^{n(s)} xh_{i,t}^{s} \\
& \qquad -P_t^{n(s)} y_t^s + \left(e^{-\rho\Delta} P_t^{n(s)}\right) z_t^s = L_t^{n(s)} \qquad \forall\ s \in \mathcal{S}_t, t = 1, \ldots, T-1 \\
& \sum_{i=1}^{I} \left(D_{i,T}^{n(s)} + S_{i,T}^{n(s)}\right) xh_{i,T-1}^{s^-} + y_{T-1}^{s^-} - z_{T-1}^{s^-} - y_T^s = L_T^{n(s)} \qquad \forall\ s \in \mathcal{S}_T \\
& xh_{i,t}^s \geq 0,\ y_t^s \geq 0,\ z_t^s \geq 0 \qquad \forall\ s \in \mathcal{S}_t,\ t = 0, \ldots, T.
\end{aligned}$$

For this version of the ALM model, we have the following result:

Proposition 2.2 *If $0 < \lambda_1 \leq 1$, then the solution to (2.25) is unbounded if the following inequality is violated for any asset i in at least one state n in the event tree:*

$$S_{i,t}^n \geq e^{-\rho\Delta} P_t^n \left(\sum_{n^+} \hat{\pi}_{t+1}^{n^+} \left(S_{i,t+1}^{n^+} + D_{i,t+1}^{n^+}\right)\right) \tag{2.26}$$

where the summation is over all successor states n^+ of state n in the event tree.

PROOF: We prove the proposition by showing that the dual of (2.25) is only feasible, and thus (2.25) itself bounded, if condition (2.26) is satisfied for every asset i in each state n of the event tree.

Through an analogous induction argument as in the first part of the proof of lemma 2.1, but with the inclusion of the parameter λ_1, we find that the dual constraints with respect to the lending variables y_t^s imply that $\varphi_t^s \geq \lambda_1 q_t^s$ for all $s \in \mathcal{S}_t$ and $t = 1, \ldots, T$.

The dual constraints that correspond to the borrowing variables z_t^s are:

$$z_0 \ : \ -\sum_{s \in \mathcal{S}_1} \varphi_1^s \leq -e^{-\rho\Delta} P_0 \tag{2.27}$$

$$z_t^s \ : \ \varphi_t^s (e^{-\rho\Delta} P_t^{n(s)}) - \sum_{s^+} \varphi_{t+1}^{s^+} \leq 0 \tag{2.28}$$

where the corresponding primal variable is listed at the beginning of each constraint.

First consider $t = 0$. If we divide the dual constraint (2.27) through by $e^{-\rho\Delta} P_0$, and the dual constraint (2.19) that corresponds to $xh_{i,0}$ by $S_{i,0}$, and then add both constraints, we obtain:

$$\sum_{s \in \mathcal{S}_1} \varphi_1^s \left(\frac{D_{i,1}^{n(s)} + S_{i,1}^{n(s)}}{S_{i,0}} - \frac{1}{e^{-\rho\Delta} P_0} \right) \leq 0$$

Using $\varphi_1^s \geq \lambda_1 q_1^s = \lambda_1 P_0 \pi_1^s$, it follows that it must also be true that

$$\lambda_1 P_0 \sum_{s \in \mathcal{S}_1} \pi_1^s \left(\frac{D_{i,1}^{n(s)} + S_{i,1}^{n(s)}}{S_{i,0}} - \frac{1}{e^{-\rho\Delta} P_0} \right) \leq 0$$

For $\lambda_1 > 0$, this inequality will only be satisfied if the sum in this expression is nonpositive. Noting that every scenario $s \in \mathcal{S}_1$ corresponds to a different node in the event tree, and $\pi_1^s = \hat{\pi}_1^s$, this will be true only if relation (2.26) holds for $t = 0$

The proof for $t > 0$ is identical, and therefore omitted.

QED

Notice that condition (2.26) in this proposition is only a sufficient condition for unboundedness of the ALM model, and not a necessary condition, unlike condition (2.18) in proposition 2.1.

If there are limits on short-term borrowing, then a violation of condition (2.26) does not lead to an unbounded solution, but the optimal solution will certainly be biased towards the corresponding arbitrage opportunity. If the transaction cost rate $c > 0$, then condition (2.26) can be modified in a similar way as was done earlier for the case $\lambda_1 = 1$.

In conclusion, we have shown that a specification of asset prices in the ALM model that is not arbitrage-free can lead to substantial biases in the optimal solution. The requirement of arbitrage-free asset prices in the model is therefore not only logically reasonable, but also important to obtain sensible solutions from the model. The conditions in propositions 2.1 and 2.2 were obtained independent of the stream of future liabilities in the model, and therefore apply to a stochastic programming model for any dynamic portfolio management problem. Even if the assumptions behind these propositions are not satisfied (i.e., if $\lambda_1 < 1$ and short-term borrowing is not possible) then a violation of condition (2.18) in proposition 2.1 may still lead to biases in the optimal solution towards hypothesized arbitrage opportunities. In that case, however, whether a bias is actually present will not only depend on the size of the violation and the parameter values in the model (c, λ_1, λ_2), but also on the correlation between the payoffs from the security for which the violation occurs and the required liability payments.

Chapter 3

Using Term-Structure Models to Describe the Uncertainty in the ALM Model

In the previous chapter we imposed the restriction on a description of the asset-price uncertainty in the ALM model that asset prices cannot admit arbitrage opportunities. It was shown that this is a crucial restriction in order to obtain reasonable results from the model. In this chapter we will discuss how financial term-structure models can be used to obtain a description of the asset-price uncertainty that satisfies this restriction.

Financial term-structure models aim to describe the uncertainty in the future term structure of interest rates, and have primarily been used to value interest-rate-derivative securities. In section 3.1 we will give a brief overview of the literature on arbitrage-free term-structure models, and present the model of Ho and Lee [31] in some detail as most of our numerical results in this and later chapters are obtained with this model.

Security prices that are calculated from a term-structure model have the important property that they are arbitrage-free. This makes these models good candidates to provide a description of the uncertainty for the ALM model. To obtain accurate security price estimates, however, these models allow the term structure to change either in a continuous fashion, or in short time increments if a description in discrete time is used. We will show in section 3.2 that a stochastic programming model can only incorporate a relatively limited description of the uncertainty to remain computationally tractable. It is therefore necessary to approximate the description

of the uncertainty that follows from a term-structure model before we can use it in the ALM model.

In performing such an approximation, we want the asset prices to remain arbitrage-free. Furthermore, if the security prices that follow from a term-structure model are consistent with observed market prices, we want to maintain this consistency in the approximate description. We will show in section 3.2 that several intuitive ways to obtain an approximate description from a term-structure model violate one or both of these requirements. With the insights obtained from this analysis, section 3.3.2 will then present state and time aggregation methods which do maintain both properties. A reformulation of the ALM model in which the description of the uncertainty is based on an aggregated term-structure model, obtained from an initial term-structure model through multiple state and time aggregations, closes this chapter.

3.1 Arbitrage-Free Models of the Term-Structure Uncertainty

The primary purpose of arbitrage-free term-structure models that have been proposed in the financial literature is the valuation of interest-rate-derivative securities. Most models assume that the prices of such securities, as well as the term structure of interest rates, are fully determined by the process of the instantaneous (i.e., very short-term) interest rate (also called the short rate) and are thus one-factor models. To calculate the prices, they assume that no arbitrage opportunities can exist (see appendix A for an explanation of asset pricing by arbitrage).

To specify a process for the short rate, two different approaches have been taken. In the first approach, the process is stated in terms of some unknown parameters, and values for these parameters are chosen such that the prices of a set of securities that are implied by the interest-rate model match their market prices as closely as possible. The continuous-time models of Vasicek [55] and Cox, Ingersoll and Ross [11] are examples of this approach.

The second approach takes the observed term structure of interest rates as given, and constructs a process for the short rate that is consistent with this term structure, and does not permit arbitrage opportunities. This approach guarantees that the prices of default-free zero-coupon bonds (also called discount bonds) that follow from the model equal their counterparts in the market. This second approach is adopted in the models of Ho and Lee [31] and Black, Derman and Toy [5]. Hull and White [34]

describe how the models of Vasicek [55] and Cox, Ingersoll and Ross [11] can be extended to match the current term structure of interest rates. These extended models, as well as the model of Black et al., can in addition be fitted to any given term structure of yield volatilities on zero-coupon bonds. Hull and White [36] describe a general procedure for the construction of interest-rate models in discrete-time that fit the term structure of interest rates as well as the term structure of discount-bond yield volatilities.

The stochastic programming formulation for the ALM problem requires a description of the interest-rate uncertainty in the form of an event tree. The models of Ho and Lee [31] and Black, Derman and Toy [5] are discrete-time models that specify a binomial lattice for the possible movements in the short rate, while Hull and White [36] assume a trinomial lattice, and these models can therefore be used directly. The Vasicek [55] and Cox, Ingersoll and Ross [11] models are continuous-time models that specify a diffusion process for the instantaneous interest rate, and therefore have to be approximated by a discrete-time model before they can be used in a stochastic programming formulation. Hull and White [35] describe a general method to accomplish this in a way that preserves the drift and the variance of the diffusion process in the discrete-time approximation, and which guarantees that prices of derivative securities that are calculated from the discretized model converge to the prices in the continuous-time model when the length of the time step approaches zero. As in Hull and White [36], the discretized interest-rate process is described by a trinomial lattice.

Although it may seem at first that one-factor models are quite restrictive, they are in fact able to accommodate many different shapes of the term structure. They do, however, imply that instantaneous price changes of all interest-rate-dependent securities are perfectly correlated. Brennan and Schwartz [8] and Heath, Jarrow and Morton [26, 27] have proposed multifactor models for the term structure that do not carry this implication, but the valuation of derivative securities according to these models is computationally much more demanding. Furthermore, they have not been proven superior in the explanation of observed security prices.

There is no agreement or conclusive empirical evidence about which of the proposed models of the term structure is the most realistic one and fits market data best. Clearly, the ability to match the observed term structure of interest rates is an attractive feature. The models of Vasicek and Cox. et al. assume mean reversion of the short rate, which seems to be supported by empirical data. A drawback of the

Vasicek model and the Ho and Lee model is that interest rates can become negative, which is not possible in the other models.

Although not necessarily the most realistic model, we have chosen to describe the term-structure model of Ho and Lee in detail in the next section because it is a relatively simple model, it highlights important characteristics of general one-factor term-structure models, and it has been used widely in practice. We have also used this model for numerical computations in this and subsequent chapters.

3.1.1 The Ho and Lee Model

In the term-structure model of Ho and Lee [31], the short rate is the single determinant of the prices of interest-rate-derivative securities. Instead of specifying the process for the short rate directly, however, Ho and Lee describe the evolution of the prices of default-free zero-coupon bonds with different maturities, and derive from that the process of the short rate. We will follow their development in this section[1].

Consider a multiperiod economy with a finite horizon H, equally spaced trading dates $t = 0, \ldots, T$ $(T = H)$ and a finite number of possible states at each time. Let $P_t^n(\tau)$ denote the price in state n at time t of a default-free zero-coupon bond that has τ time periods left to maturity $(\tau = 0, \ldots, T - t)$. $P_t^n(\tau)$ as a function of τ is called the *discount function*, and Ho and Lee model the changes of this discount function over time. They assume that the complete discount function at time 0 is known[2].

In a world of certainty (only one possible state at each trading date), it must be true that $P_{t+1}(\tau) = P_t(\tau + 1)/P_t(1)$ to prevent arbitrage opportunities (note that $P_{t+1}(\tau)$ and $P_t(\tau + 1)$ represent prices of the same bond, but at different points in time). Under uncertainty, Ho and Lee assume that the discount function can change in two directions in each period. They describe the possible changes for each discount

[1] For general one-factor term-structure models, Hull and White [36] show the relationship between the specification of price processes for default-free zero-coupon bonds, processes for the instantaneous forward rates, and the process for the short rate.

[2] This is equivalent to knowing the complete term structure of interest rates at time 0. If $r_t^n(\tau)$ denotes the continuously compounded yield in state n at time t on a zero-coupon bond with τ periods left to maturity, then $r_t^n(\tau)$ is defined by

$$r_t^n(\tau) = \frac{-\ln P_t^n(\tau)}{\tau \Delta}$$

and $r_t^n(\tau)$ as a function of τ is called the term structure of interest rates (or yield curve) in state n at time t. In their paper, Ho and Lee assume $\Delta = 1$, i.e. that interest rates are defined per period.

bond as perturbations from the required change in a world of certainty:

$$\text{Upstate:} \quad P_{t+1}^{n+1}(\tau) = \frac{P_t^n(\tau+1)}{P_t^n(1)} h(\tau) \tag{3.1}$$

$$\text{Downstate:} \quad P_{t+1}^{n}(\tau) = \frac{P_t^n(\tau+1)}{P_t^n(1)} h^*(\tau) \tag{3.2}$$

The perturbation functions $h(\cdot)$ and $h^*(\cdot)$ are assumed to depend only on the remaining time to maturity of the bonds, and satisfy $h(0) = h^*(0) = 1$ and $h(\tau) > 1$, $h^*(\tau) < 1$ for $\tau > 0$.

The first restriction that Ho and Lee impose on the perturbation functions is that the discount bond prices in the event tree do not admit arbitrage opportunities. They show that this is equivalent to the requirement that

$$\pi h(\tau) + (1-\pi) h^*(\tau) = 1 \tag{3.3}$$

for some constant π, independent of τ. Using the definition of the perturbation functions, this can be rewritten as $P_t^n(\tau+1) = P_t^n(1)\left[\pi P_{t+1}^{n+1}(\tau) + (1-\pi) P_{t+1}^n(\tau)\right]$, and the constant π is therefore referred to as the *implied* (or *risk-neutral*) binomial probability.

As the second restriction on the perturbation functions, Ho and Lee require that the price processes are path independent. That is, if the price of a discount bond in state n at time t follows an "upstate" and a "downstate" in the next two periods, respectively, then its price at time $t+2$ must be the same as when it would have followed a "downstate" and an "upstate" move, respectively. The number of different states at time t is therefore limited to $(t+1)$, which are indexed as $n = 0, \ldots, t$, and the changes in the discount function over time can be represented by a binomial lattice (see figure 2-1). Ho and Lee show that the path-independence condition, together with condition (3.3), leads to the following expressions for the perturbation functions:

$$h(\tau) = \frac{1}{\pi + (1-\pi)\delta^\tau} \qquad \text{and} \qquad h^*(\tau) = \delta^\tau \cdot h(\tau) \qquad \forall \tau \geq 1 \tag{3.4}$$

where δ is some constant between 0 and 1 ($\delta = 1$ is the certainty case). The parameters π and δ, together with the initial discount function, thus completely define the term-structure model of Ho and Lee.

The short rate r_t^n in state n at time t is defined as the continuously compounded interest rate on the zero-coupon bond with one period left to maturity: $r_t^n = -\ln P_t^n(1)$. Equations (3.1) and (3.2) enable us to write:

$$P_t^n(1) = \frac{P_0(t+1)}{P_0(t)} \frac{\delta^{t-n}}{\pi + (1-\pi)\delta^t} \quad , n \leq t.$$

The process for the short rate can thus be written as

$$\begin{aligned} r_t^0 &= \ln\left[\frac{P_0(t)}{P_0(t+1)}\right] + \ln\left[\pi\delta^{-n} + (1-\pi)\right], \\ r_t^n &= r_t^0 + n\ln\delta \qquad \text{for } n \leq t. \end{aligned}$$

We note that $r_t^{n+1} - r_t^n = \ln\delta$, independent of n and t. Thus, the volatility of the short rate is independent of the state at a given time. Furthermore, to prevent the short rate from being negative anywhere in the binomial lattice, we need that $r_t^t \geq 0$ for all t (note that $\ln\delta \leq 0$), or equivalently,

$$\delta > \left[\frac{\frac{P_0(t+1)}{P_0(t)} - \pi}{1-\pi}\right]^{\frac{1}{t}} \quad \forall\, t = 0, \ldots, T. \tag{3.5}$$

So far, no explicit assumptions were made about probability beliefs of investors about changes in the term structure. If an investor assigns a probability q $(0 < q < 1)$ to every "upstate" movement in the binomial lattice, then the expected short rate and its variance (evaluated at time 0) can respectively be written as

$$\mu_t \equiv E_0\{\tilde{r}_t\} = \ln\left[\frac{P_0(t)}{P_0(t+1)}\right] + \ln\left[\pi\delta^{-t} + (1-\pi)\right] + tq\ln\delta \tag{3.6}$$

$$\sigma_t^2 \equiv E_0\{(\tilde{r}_t - \mu_t)^2\} = tq(1-q)(\ln\delta)^2 \tag{3.7}$$

The *term premium* on a τ-period discount bond is defined as the difference between the expected one-period return on this bond and the riskless one-period return. Ho and Lee prove that this term premium in state n at time t equals:

$$\frac{1}{P_t^n(1)}\left(\left[\frac{q + (1-q)\delta^\tau}{\pi + (1-\pi)\delta^\tau}\right] - 1\right)$$

where q is defined as before. If all term premia are zero (i.e., $q = \pi$), then the *(local) expectations hypothesis* is said to hold.

3.1.2 Asset Prices in the Ho and Lee Model

One of the explicit requirements in the construction of the Ho and Lee model was that the prices of the discount bonds are arbitrage-free. This section shows how

the model can be used to determine arbitrage-free prices of general interest-rate-derivative securities. We will also illustrate how these arbitrage-free prices depend on the number of time periods (T) within the model horizon H.

Consider an interest-rate-contingent claim that matures at the end of period T (i.e. the model horizon), and which pays a dividend D_t^n to its holder in state n at time t ($t = 1, \ldots, T; n = 0, \ldots, t$). To avoid that riskless arbitrage profits can be made by dynamic trading in the discount bonds and the contingent claim, Ho and Lee prove that the following relation must hold between the ex-dividend prices of the contingent claim in the event tree for all $t = 0, \ldots, T-1$ and $n = 0, \ldots, t$:

$$S_t^n = P_t^n(1) \cdot \left[\pi(S_{t+1}^{n+1} + D_{t+1}^{n+1}) + (1-\pi)(S_{t+1}^n + D_{t+1}^n)\right] \tag{3.8}$$

where $S_T^n = 0$ for all $n = 0, \ldots, T$. The proof follows directly from the theory of asset pricing by arbitrage, which is explained in appendix A. Relation (3.8) specifies a recursive relationship between the contingent-claim prices, which is very convenient for computation. It is clear from this recursive relation that one only needs to know the short rate in each state of the event tree to calculate arbitrage-free asset prices, which shows that the Ho and Lee model is a one-factor model. Notice further that the pricing formula only involves the *implied* binomial probability π, and not the *subjective* probability q; that is, the arbitrage-free prices do not depend on individual probability beliefs.

We have assumed that the dividends on the contingent claim depend only on the state in the event tree, and not on the history of states. This excludes some classes of interest-rate-contingent claims such as floating-rate bonds and mortgage-backed securities. It is conceptually straightforward to extend the pricing relation to such securities with path-dependent payoffs: we just need to redefine the states at each trading date as the sequence of states up to that trading date. However, this will cause the number of states at trading date t to grow from $t+1$ to 2^t, and thus the computational effort to value securities with path-dependent payoffs will increase exponentially with T instead of quadratically.

To use the Ho and Lee model, one has to decide in how many time steps T to divide a given horizon H. Increasing the number of time steps enables a more accurate valuation of securities, but at an increasing computational cost. For a set of European call options on a discount bond, we will illustrate the dependence of their arbitrage-free prices at time 0 (calculated using equation (3.8)) on the number of time steps.

We consider four European call options that are written on a default-free zero-coupon bond with a maturity of 5 years and face value of 100. The time to maturity for the options is 1, 2, 3, and 4 years, and we will refer to them as options 1 through 4 respectively. For the construction of the Ho and Lee model, we assume that the yield curve is flat with a yield of 8% for all maturities. We have chosen $\pi = 0.5$ as risk-neutral binomial probability. The strike price of each option equals the expected bond price at the option's maturity, calculated with respect to the risk-neutral probability measure. We have varied the number of time steps in the model between 5 and 120 to calculate arbitrage-free option prices at time 0. For each number of time steps in this range, the value of the parameter δ is chosen such that the volatility of the short rate equals 0.31% per year (again with respect to the risk-neutral probability measure)[3].

Figure 3-1 depicts the arbitrage-free time-0 prices for the options as a function of the number of time steps in the Ho and Lee model[4]. The prices converge with an increasing number of time steps in the model, but more than 100 time steps are needed before all prices have converged to within a precision of at least two decimal digits (corresponding to prices in dollar cents). We note that 100 time steps in the model (5 years) corresponds to 20 periods before the maturity of option 1, 40 periods before the maturity of option 2, and so forth. Thus, the security with the shortest maturity will often determine the minimum number of time steps that is required for the convergence of *all* prices to within a certain precision.

We have performed the same calculations with different assumptions about the shape of the initial term structure, the exercise prices of the options, and the volatility of the short rate, and the results are very similar. That is, when the interest-rate model is used to value interest-rate-contingent claims with various maturities, a substantial number of time steps may be needed to make sure that the implied arbitrage-free prices at time 0 of all securities have converged to a realistic level of precision. As we will show in the next section, this complicates the use of term-structure models to describe the uncertainty in the ALM model.

---

[3] This level of volatility guarantees that interest rates in the model are nonnegative at all times for any number of time steps in the given range. Inequality (3.5) specifies a lower bound for the value of the parameter δ, which implies an upper bound on the volatility of the short rate through equation (3.7) (note that $\delta \leq 1$). Under our assumption of a flat yield curve, it is not difficult to see that the lower bound for δ will be highest for $t = T$ and $T = 120$. The volatility of 0.31% is just below the corresponding upper bound on the volatility.

[4] The time-0 price of the underlying discount bond is independent of the number of time steps as the Ho and Lee model is constructed such that the implied prices of discount-bonds match their actual prices at time 0 (see the previous section).

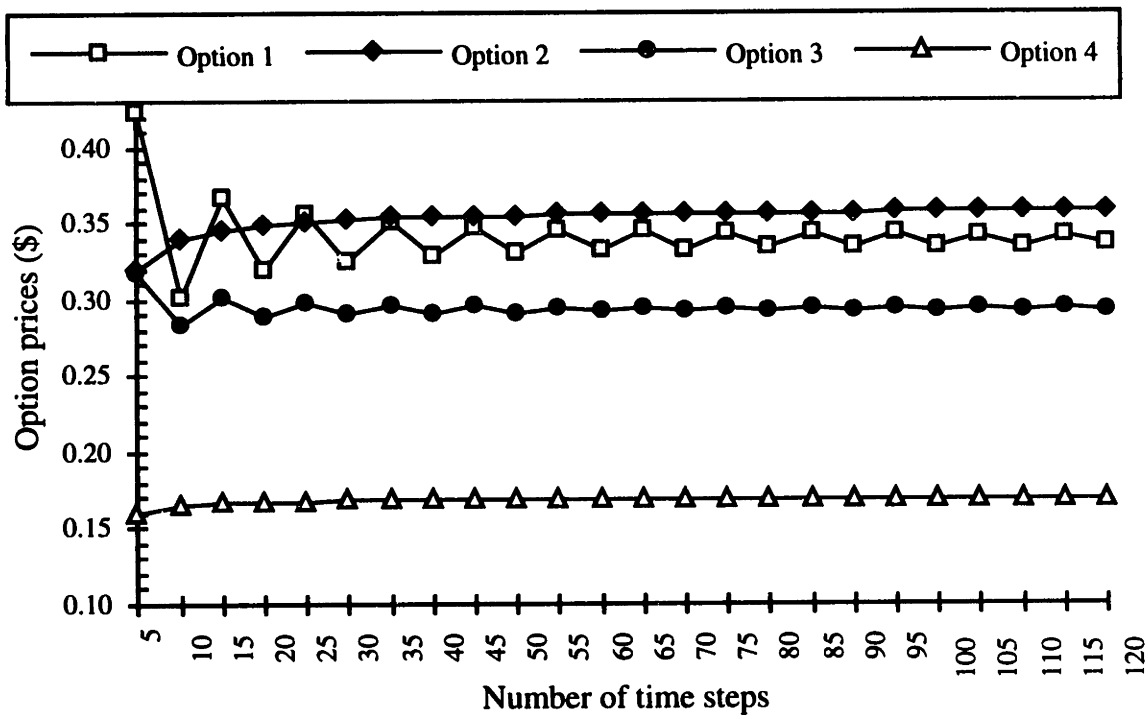

Figure 3-1: Arbitrage-free option prices at time 0 from Ho and Lee model as a function of the number of time steps.

3.2 Combining Stochastic Programs and Term-Structure Models

Discrete-time term-structure models can in principle be used directly to specify the uncertainty in asset prices and interest rates in the stochastic programming formulation for the ALM problem: the event tree that describes the uncertainty in the term-structure models is used to define the (interest-rate) scenarios in the stochastic program, and the arbitrage-free security prices at each node in the event tree serve as estimates of future security prices. Of course, one has to be able to specify the liabilities and the security dividends (such as interest and principal payments on bonds) as a function of the states in the event tree of the term-structure model.

An important issue is the consistency of the arbitrage-free security prices at time 0 that follow from the model with observed market prices. As mentioned before, there is no conclusive empirical evidence about which of the proposed term-structure models explains observed security prices best. We make the following additional assumption

for our models.

Assumption 4 *There exists a term-structure model whose implied arbitrage-free security prices at time 0 equal the observed market prices.*

Section 3.1.2 has illustrated for the Ho and Lee model what is true for discrete-time term-structure models in general: the number of time steps needed to obtain convergence of arbitrage-free security prices that are calculated from the model in at least two decimal digits (i.e., dollar cents) can be substantial, especially if securities have different maturities. Although the calculation of prices for securities with state-dependent payoffs from a term-structure model with several hundred time steps is still relatively fast if the event tree has a lattice structure[5], the ALM model as presented in section 2.2.1 cannot take advantage of a lattice structure and includes separate variables and constraints for every scenario. As the number of scenarios increases exponentially with the number of time steps, the ALM model imposes a strict limit on this number of time steps.

To illustrate this, let I denote the number of assets available for trading at each trading date $t = 0, \ldots, T$. If the interest-rate uncertainty is described by a binomial lattice, as in the Ho and Lee model, then the number of variables in the formulation of the ALM problem of section 2.2 equals[6] $(3I + 4)2^T - (5I + 2)$ and the number of constraints[7] (excluding the upper bounds on short-term borrowing) $(I+2)2^T - 2(I+1)$. The number of variables and constraints thus roughly doubles with *each* extra time step, and increases linearly with the number of assets. Table 3.1 lists the number of variables and constraints for different values of I and T. If a trinomial instead of a binomial lattice is used to describe the interest-rate uncertainty, as in the models of Hull and White [35, 36], then the size of the models explodes even faster as is illustrated in table 3.2.

This "curse of dimensionality" forces us to seriously limit the number of time steps in the ALM model, and we have indicated that this number will generally be much smaller than the minimally required number of time steps in a discrete-time term-structure model that satisfies assumption 4. The next section will discuss methods

[5]The computational effort is proportional to the number of nodes in the event tree, which is a quadratic function of the number of time steps if the event tree is a binomial or trinomial lattice.

[6]There are 2^t scenarios at each time t, $I+2$ variables at time 0, $(3I+2)$ variables for each scenario at times 1 through $T-1$, and 2 variables for each scenario at time T.

[7]One cash-balance constraint and I portfolio-balance constraints for each scenario at $t = 1, \ldots, T-1$, and a cash-balance constraint for each scenario at time T.

T	$I = 10$		$I = 20$		$I = 50$	
	Variables	Constraints	Variables	Constraints	Variables	Constraints
2	84	26	154	46	364	106
3	220	74	410	134	980	314
4	492	170	922	310	2,212	730
5	1,036	362	1,946	662	4,676	1,562
6	2,124	746	3,994	1,366	9,604	3,226
7	4,300	1,514	8,090	2,774	19,460	6,554
8	8,652	3,050	16,282	5,590	39,172	13,210
9	17,356	6,122	32,666	11,222	78,596	26,522
10	34,764	12,266	65,434	22,486	157,444	53,146

Table 3.1: Number of variables and constraints in the ALM model of section 2.2.1 for different numbers of traded securities (I) and time periods (T), and when the interest-rate process follows a binomial lattice.

T	$I = 10$		$I = 20$		$I = 50$	
	Variables	Constraints	Variables	Constraints	Variables	Constraints
2	126	42	226	72	526	162
3	450	159	820	279	1,930	639
4	1,422	510	2,602	900	6,142	2,070
5	4,338	1,563	7,948	2,763	18,778	6,363
6	13,086	4,722	23,986	8,352	56,686	19,242
7	39,330	14,199	72,100	25,119	170,410	57,879
8	118,062	42,630	216,442	75,420	511,582	173,790
9	354,258	127,923	649,468	226,323	1,535,098	521,523
10	1,062,846	383,802	1,948,546	679,032	4,605,646	1,564,722

Table 3.2: Number of variables and constraints in the ALM model of section 2.2.1 for different numbers of traded securities (I) and time periods (T), and when the interest-rate process follows a trinomial lattice.

that can be used to approximate the description of the interest-rate uncertainty that follows from a term-structure model.

3.3 Approximation of the Interest-Rate Uncertainty

The approximation methods that we will present in this section all approximate the description of the uncertainty in interest rates and asset prices that follows from a discrete-time term-structure model. We will assume that a discrete-time model of the term-structure uncertainty is known that satisfies assumption 4, but which implies a number of interest-rate scenarios that is too large to include in the ALM model. This model will be referred to as the *fully consistent* (term-structure) model.

As mentioned in the introduction to this chapter, we want an approximation method to maintain two properties of the asset prices. First, we want the asset prices to be arbitrage-free in the approximate description. Besides being a reasonable requirement, we have shown in the previous chapter that a violation of this property can have a very significant and undesirable effect on the optimal solution to the ALM model. Second, we want the approximate description to satisfy assumption 4, i.e., maintain consistency with the observed market prices. This is important because we want to be able to use observed market prices in the ALM model. If we use market prices in the ALM model while the approximate description of the uncertainty is inconsistent with them, then we essentially create an arbitrage opportunity in the model.

In the first part of this section we will describe three somewhat intuitive ways to reduce the number of scenarios in the fully consistent term-structure model, and for each of them we will indicate how the approximated asset prices will violate one or both of these required properties. The analysis of the errors in these approximation methods forms a warming-up for the state aggregation and time aggregation methods in the second part of this section. These aggregation methods combine states and time periods in the event tree from a discrete-time term-structure model such that the two properties in the previous paragraph remain satisfied.

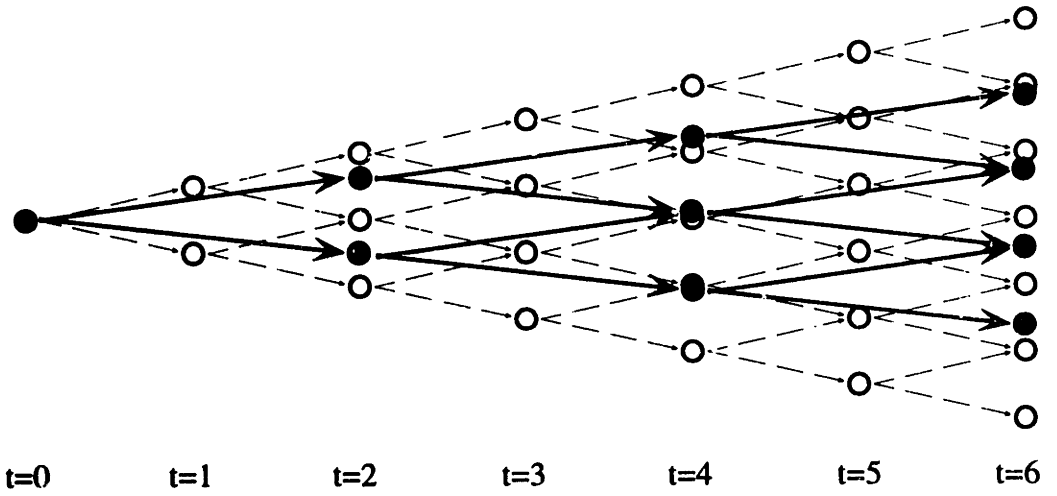

Figure 3-2: Coarse tree approximation: every time step in the coarse tree (solid lines and filled nodes) has the length of two time steps in the fully consistent tree (dotted lines and open nodes).

3.3.1 Inconsistent Approximations

As noted earlier, financial models of the term-structure uncertainty have hardly been considered in the literature as a means of describing the uncertainty in future interest rates and asset prices in stochastic programming models for fixed-income portfolio management, and our approximations therefore do not reflect common practice. However, much of the literature does not discuss at all how this uncertainty must be specified, and it should be clear that an arbitrary specification has an even nigher risk of introducing opportunities for riskless arbitrage profits in an optimization model than the ones considered here.

Coarse tree approximation

In the coarse tree approximation we use the data that form the basis for the fully consistent term-structure model to construct a version of the model with a longer time step, and thus fewer states and implied scenarios. For example, if the fully consistent model is the Ho and Lee model with a certain number of time steps that is constructed from a given yield curve and short-term interest-rate volatility, then the "coarse tree" is a version of the Ho and Lee model that is constructed from the same market data, but with a smaller number of time steps. This situation is depicted in figure 3-2, where every time step in the coarse tree has the length of two time steps in the fully consistent tree.

It is clear that the asset prices within the coarse-tree approximation are arbitrage-free as the coarse tree is itself a term-structure model. Furthermore, because the same yield curve is used as input for the coarse tree and the fully consistent tree, the implied arbitrage-free prices at time 0 for discount bonds of all maturities are the same in both models. However, this is not true for general interest-rate derivative securities. This has been illustrated in figure 3-1 (section 3.1.2) for the prices of European call options on a discount bond that follow from the Ho and Lee model when the number of time steps changes, but the same underlying market data are used. If the Ho and Lee model with 120 time steps is taken as the fully consistent model in this figure (i.e., the arbitrage-free security prices from this model are assumed to match the observed market prices), then it is clear that the arbitrage-free values for the options that follow from a coarse tree approximation with a significantly smaller number of time steps (e.g., 5 or 10) substantially differ from their market prices. Using both the coarse tree approximation and the current market prices in a stochastic programming model thus leads to an inconsistency between current prices and (arbitrage-free) values of securities, and an optimal portfolio will most likely be biased towards assets that seem underpriced. To what extent this will happen may depend on other model parameters as well, as was discussed in section 2.3.2.

Subtree approximation

The subtree approximation takes a subset of states from the fully consistent term-structure model as approximate description of the uncertainty for the stochastic programming model. To illustrate this, suppose that the fully consistent term-structure model uses a binomial lattice to describe the uncertainty in interest rates (e.g., the Ho and Lee model). Consider the subtree (or properly, sub*lattice*) of this lattice that only includes states at the end of every second time step, and of those states only every second state. If the states in the full lattice are numbered as in figure 2-1, and if node n at time t is included in the sublattice, then this node has nodes n and $n+2$ at time $t+2$ as its successors in the sublattice. This subtree approximation is depicted in figure 3-3.

We will show that asset prices in this sublattice are not in general arbitrage-free when the prices at a node in the sublattice are copied from the full lattice. Consider a security S with dividend D_t^n and ex-dividend price S_t^n in node n at time t in the full lattice. For simplicity, assume that this security does not pay dividends at time $t+1$, and that the risk-neutral conditional probability of un upstate move in the

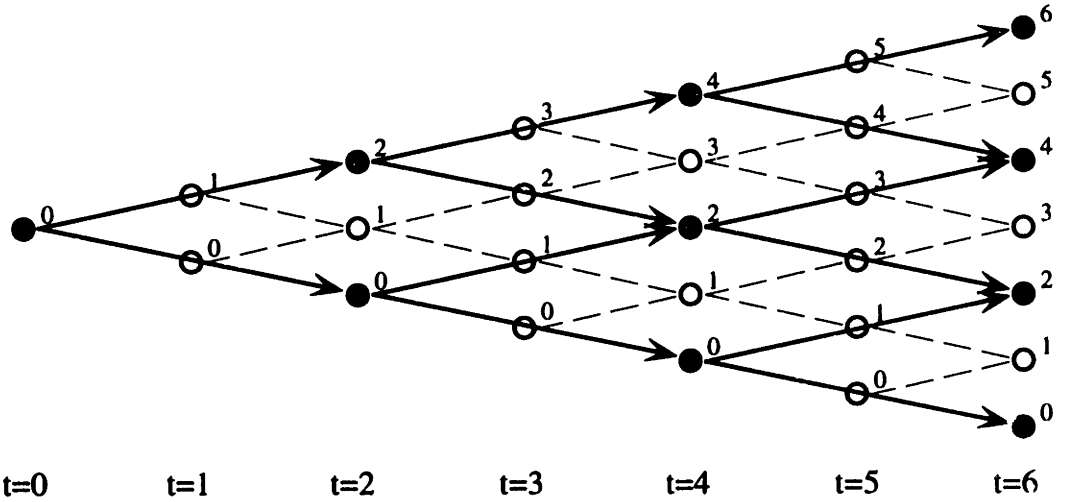

Figure 3-3: Subtree approximation: both the subtree (solid lines and filled nodes) and the fully consistent tree are assumed to be binomial lattices.

full lattice equals π in all states. Using the arbitrage-free pricing relation (3.8), S_t^n can be expressed as follows as a function of the prices of the security at time $t+2$:

$$\begin{aligned} S_t^n = P_t^n \big((1-\pi) \quad & P_{t+1}^n \quad \left[\pi(S_{t+2}^{n+1} + D_{t+2}^{n+1}) + (1-\pi)(S_{t+2}^n + D_{t+2}^n)\right] \\ +\pi \quad & P_{t+1}^{n+1} \quad \left[\pi(S_{t+2}^{n+2} + D_{t+2}^{n+2}) + (1-\pi)(S_{t+2}^{n+1} + D_{t+2}^{n+1})\right]\big) \end{aligned} \tag{3.9}$$

For the sublattice, let $\bar{\pi}$ denote the risk-neutral conditional probability of an upstate move, $\bar{P}_t^n$ the price in state n at time t of a riskless investment that pays one dollar at time $t+2$, and $\bar{S}_t^n$ the arbitrage-free ex-dividend price of security S in state n at time t. Assuming $\bar{S}_{t+2}^{n'} = S_{t+2}^{n'}$ for $n' = n, n+1, n+2$ and using pricing relation (3.8) in the sublattice, we have

$$\bar{S}_t^n = \bar{P}_t^n \left(\bar{\pi}(S_{t+2}^{n+2} + D_{t+2}^{n+2}) + (1-\bar{\pi})(S_{t+2}^n + D_{t+2}^n)\right) \tag{3.10}$$

To show that no arbitrage opportunities exist in the sublattice, we need to find expressions for $\bar{\pi}$ and $\bar{P}_t^n$, *independent of the security*, such that $S_t^n = \bar{S}_t^n$ (theorem A.1 in appendix A). We will see that this is not possible in general. Suppose S is a discount bond that matures at time $t+2$ (i.e., $D_{t+2}^n = D_{t+2}^{n+1} = D_{t+2}^{n+2}$ and $S_{t+2}^n = S_{t+2}^{n+1} = S_{t+2}^{n+2} = 0$). Then for $S_t^n = \bar{S}_t^n$ it must be true that

$$\bar{P}_t^n = P_t^n \left(\pi P_{t+1}^{n+1} + (1-\pi) P_{t+1}^n\right)$$

Substituting this in (3.10), we can solve for the value of $\bar{\pi}$ that makes $\bar{S}_t^n$ equal to S_t^n for an arbitrary security S. If we denote cum-dividend prices as $\hat{S}_t^n \equiv (S_t^n + D_t^n)$,

then the value for $\bar{\pi}$ must satisfy

$$\bar{\pi}(\hat{S}^{n+2}_{t+2} - \hat{S}^{n}_{t+2}) = \pi\left(\frac{\pi P^{n+1}_{t+1}(\hat{S}^{n+2}_{t+2} - \hat{S}^{n+1}_{t+2}) + [P^{n+1}_{t+1} + (1-\pi)P^{n}_{t+1}](\hat{S}^{n+1}_{t+2} - \hat{S}^{n}_{t+2})}{\pi P^{n+1}_{t+1} + (1-\pi)P^{n}_{t+1}}\right)$$

If $\hat{S}^{n+2}_{t+2} = \hat{S}^{n}_{t+2}$, then it is easy to see that this equality can only be satisfied if $\hat{S}^{n+1}_{t+2} = \hat{S}^{n+2}_{t+2} = \hat{S}^{n}_{t+2}$, and $\bar{\pi}$ is undetermined in this case. If $\hat{S}^{n+2}_{t+2} \neq \hat{S}^{n}_{t+2}$, then there is a unique solution for $\bar{\pi}$, but it is clear that it depends on the cum-dividend prices of security S at time $t+2$, and thus will be different for different securities. Furthermore, nothing guarantees that $\bar{\pi}$ is a probability: it is possible to choose values for $\hat{S}^{n+2}_{t+2}$, $\hat{S}^{n+1}_{t+2}$ and $\hat{S}^{n}_{t+2}$ such that $\bar{\pi} < 0$ or $\bar{\pi} > 1$.

For other choices of subtrees, it can be shown in a similar way that prices in the subtree will in general not be arbitrage-free if these prices are copied from the corresponding nodes in the fully consistent tree. In the full tree, the security price S^n_t can be written as a function of the prices in all the states at any time $t + \tau > t$ that can be reached from state n at time t. In the subtree, however, we want to write the same price S^n_t as a function of only a subset of the prices at time $t + \tau$, where the form of the function is not allowed to depend on the security prices themselves. It should be clear that this is not possible in general.

In contrast to the coarse tree approximation, which leads to arbitrage opportunities in the first period only when used as description of the uncertainty in the ALM model, the subtree approximation will cause arbitrage opportunities at any point in time. To illustrate that the difference between the implied arbitrage-free value of a security, calculated from the subtree, and its actual price at time 0 can be very substantial, we compare these numbers in figure 3-4 for European call options on a discount bond. The characteristics of the call options and the data for the Ho and Lee model on which the calculations are based are the same as in section 3.1.2. The subtrees were chosen similarly to the one in figure 3-3, but with an increasing number of time steps in the fully consistent tree (assumed to be the Ho and Lee model with 120 time steps) making up one time step in the subtree. That is, if τ time steps in the fully consistent tree span one period in the subtree, than a node n at time t in the subtree has nodes n and $n + \tau$ as its successors at time $t + \tau$. To calculate the arbitrage-free security values in the subtree, we have assumed that the conditional probability of an upward move in the subtree is the same as in the full tree (i.e., $\bar{\pi} = \pi = 1/2$). The increase in the arbitrage-free option values that follow from a

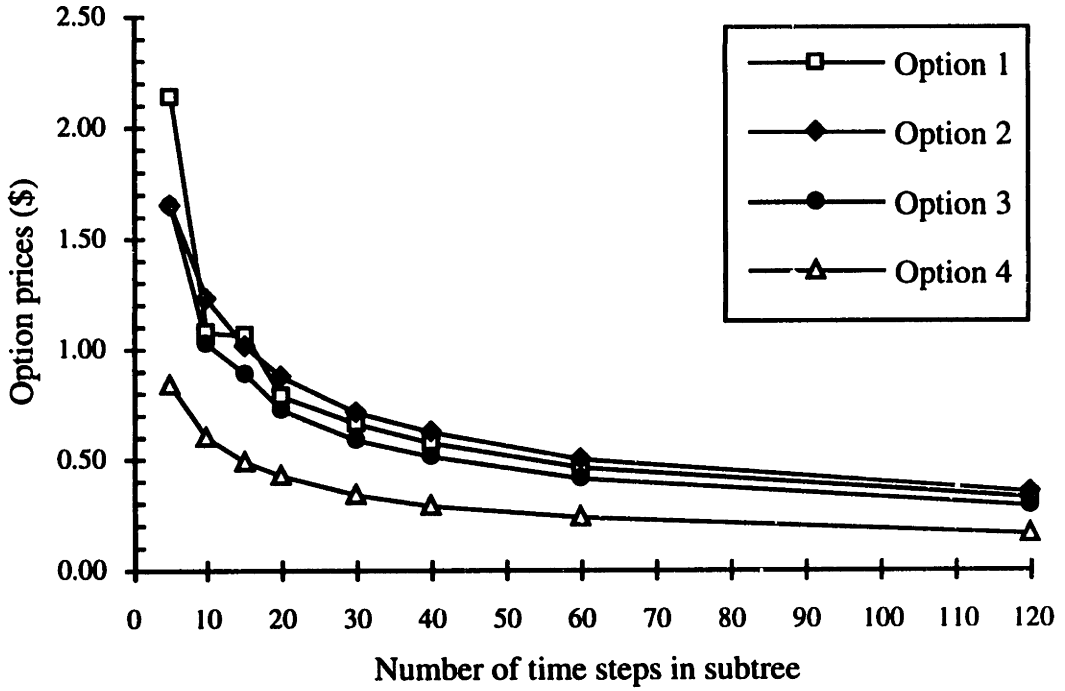

Figure 3-4: Arbitrage-free values at time 0 for European call options, as a function of the number of time steps in a subtree.

subtree when the number of periods in the subtree decreases can to a large extent be attributed to the fact that the interest-rate volatility is higher if the number of periods in the subtree is smaller.

Sampling approximation

Another way to reduce the number of scenarios in the stochastic programming formulation is by sampling scenarios from the fully consistent term-structure model. (Hiller and Eckstein [28] apply sampling from a term-structure model in their solution approach, but they ignore the inconsistency of their term-structure model with observed market prices; see section 1.2). The risk-neutral probability measure on the event tree in the fully consistent model serves as the sampling distribution, and each sampled path is assigned equal probability in the stochastic programming model. We will show that security prices in the sampled set of scenarios easily admit arbitrage opportunities.

Consider state n at time t in the fully consistent event tree, and let n^+ denote a successor state at time $t+1$. Furthermore, let $\hat{\pi}_t^{n^+}$ denote the conditional risk-neutral probability of state n^+ at time $t+1$, given state n at time t. As prices in the fully consistent tree are arbitrage-free, the price S_t^n of any security must satisfy

$$S_t^n = P_t^n \sum_{n^+} \hat{\pi}_t^{n^+} \left(S_{t+1}^{n^+} + D_{t+1}^{n^+} \right).$$

Let M denote the number of sampled scenarios from the fully consistent tree that visit state n at time t, and let M^+ be the number of scenarios from these M scenarios that visit the successor state n^+ at time $t+1$. As all sampled scenarios have equal probability in the stochastic programming model, the necessary condition for arbitrage-free prices within the sampled set of scenarios is that the M^+ are such that $(M^+/M) = \hat{\pi}_t^{n^+}$. It is clear that exact equality will only happen by chance. Although it may be approximately satisfied for t close to 0 when the sample size is large, severe violations will almost certainly occur in states close to the planning horizon, which are only included by few sampled scenarios.

Dantzig and Infanger [13] use importance sampling in their solution approach to multistage stochastic linear programs for portfolio optimization. Importance sampling is often used for the computation of multiple integrals or sums, and is a variant of Monte Carlo sampling. It aims to reduce the variance of the estimate for the value of the integral or sum (as compared to ordinary Monte Carlo sampling) by changing the sampling distribution in a way that gives a higher probability to events with a large impact on the value of the estimate. In the context of the ALM problem, importance sampling would increase the sampling probability of scenarios that have a relatively large effect on the objective function. Changing the sampling distribution from the risk-neutral probability measure, however, will increase the likelihood of arbitrage opportunities within the sampled set of scenarios even further.

3.3.2 State and Time Aggregation

In contrast to the approximation methods of the previous section, the state aggregation and time aggregation methods that will be presented in this section enable us to approximate a description of the interest-rate uncertainty from a term-structure model that satisfies assumption 4, but which is too large to include in the ALM model, in a way that preserves the consistency with observed market prices and guarantees that asset prices are arbitrage-free within the approximate description.

State Aggregation

Let N_t denote the number of states at time t ($t = 0, \ldots, T$) in the event tree that describes the interest-rate uncertainty in the original (i.e., unaggregated) term-structure model that satisfies assumption 4. We number the states at each time t from $n = 0$ to $n = N_t - 1$. We say that *state aggregation* is performed in state n at time t if all the successor states of state n in the event tree are combined into one (aggregate) state. In an aggregated event tree, a state is characterized by the triplet of numbers (t, n, k): the time t, node number n and aggregation level k. The *aggregation level* of a state indicates how many state aggregations were performed to obtain the state. (In the original event tree, all states have aggregation level 0.) We impose the restriction that all successor states of a state in an aggregated event tree must have the same aggregation level. This implies that we always aggregate states with the same aggregation level when we perform state aggregation. If state aggregation is performed in state (t, n, k), and if all its successor states have aggregation level k', then the aggregated successor state will have aggregation level $k + 1$ and it will be assigned the node number n.

To illustrate this, consider the (unaggregated) event tree in figure 3-5[8]. If state aggregation is performed in state 1 at time 2, then the resulting tree is depicted in figure 3-6(a). Figure 3-6(b) shows the aggregated tree if state aggregation is instead performed in state 0 at time 1. For the aggregated event tree in figure 3-6(a), the restriction from the previous paragraph implies that we cannot perform state aggregation in state 1 at time 1, as its aggregated successor state at time 2 would have successor states at time 3 with different aggregation levels. One would first need to perform state aggregation in states 2 and 3 at time 2 before state aggregation in state 1 at time 1 is possible. Figure 3-7(a) shows the result of these aggregations. If state aggregation is performed in state (0,1) at time 2 in the aggregated tree of figure 3-6(b), then the resulting tree is depicted in figure 3-7(b). Note that the aggregation level of the successors of state (0,1) at time 2 was zero before the aggregation, whereas its single aggregated successor has aggregation level 2.

It follows from our definitions that state n at time t with aggregation level k in an aggregated event tree is the result of the aggregation of all states at time t in the original (i.e., unaggregated) event tree that had state n at time $t - k$ as

[8]For states with aggregation level 0 we only denote the node number n, otherwise we will denote the pair (n, k).

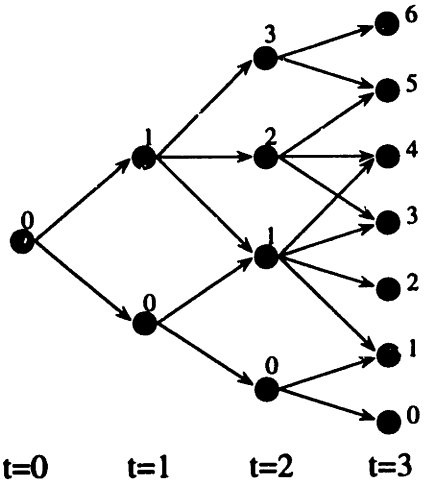

Figure 3-5: An unaggregated event tree.

their ancestor. The number of states with aggregation level k at time t is thus at most N_{t-k} ($k \leq t$). Furthermore, descendant states at time $t+\tau$ of state (t,n,k) can at most have aggregation level $k+\tau$. For state (t,n,k), we will denote the collection of node numbers of its successors in an aggregated event tree as $\mathcal{N}_t^{(n,k)}(k')$, where $k' \leq k+1$ is the aggregation level of its successors. For the event tree in figure 3-5, $\mathcal{N}_2^{(1,0)}(0) = \{1,2,3,4\}$, $\mathcal{N}_2^{(1,0)}(1) = \{1\}$ (see figure 3-6(a)), and $\mathcal{N}_2^{(1,1)}(1) = \{1,2,3\}$ (see figure 3-7(a)). The way in which we assign node numbers to aggregated states implies that $\mathcal{N}_t^{(n,k)}(k+1) = \{n\}$ and $\mathcal{N}_t^{(n,k)}(k') = \mathcal{N}_{t+1}^{(n,k+1)}(k'+1)$ with $k' \leq k+1$.

We will now describe how to define interest rates and asset prices in an aggregated event tree such that the asset prices are arbitrage-free, and the arbitrage-free prices at time 0 equal the ones in the original tree. To show that the prices in the aggregated tree are arbitrage-free, we use theorem A.1 in appendix A and construct an equivalent martingale measure on the aggregated event tree. Let $P_t^{(n,k)}$ represent the price in state (t,n,k) of a riskless investment that pays one dollar at time $t+1$.[9] Furthermore, let $D_t^{(n,k)}$ denote the dividend payment on a security S in state (t,n,k), and $S_t^{(n,k)}$ its ex-dividend price in that state. The risk-neutral conditional probability of visiting state $(t+1,n',k')$ in the aggregated event tree, given state (t,n,k), is

[9]The continuously compounded one-period riskless interest rate $r_t^{(n,k)}$ in state (t,n,k) thus equals $-(1/\Delta)\ln P_t^{(n,k)}$, where Δ is the length of a time step in the event tree.

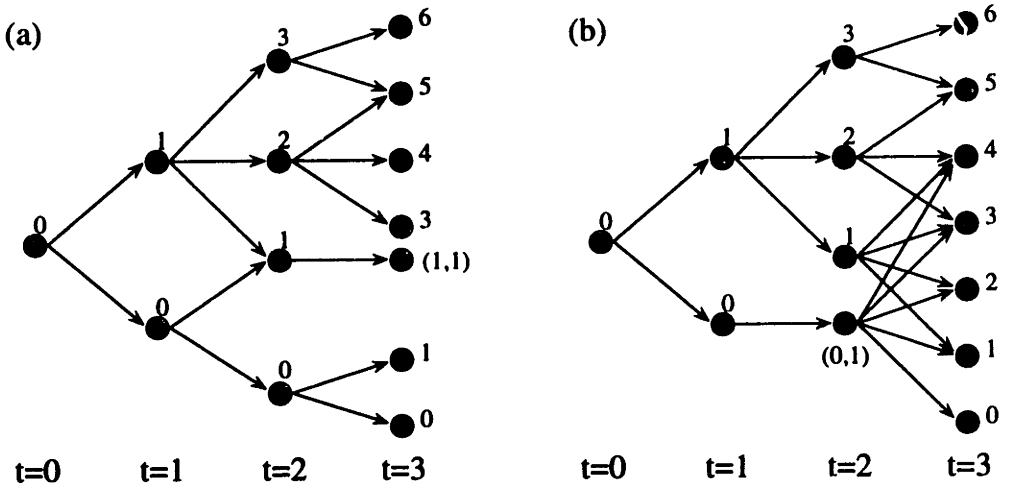

Figure 3-6: (a) Result of state aggregation in state 1 at time 2 in the event tree of figure 3-5. (b) Result of state aggregation in state 0 at time 1 in the event tree of figure 3-5.

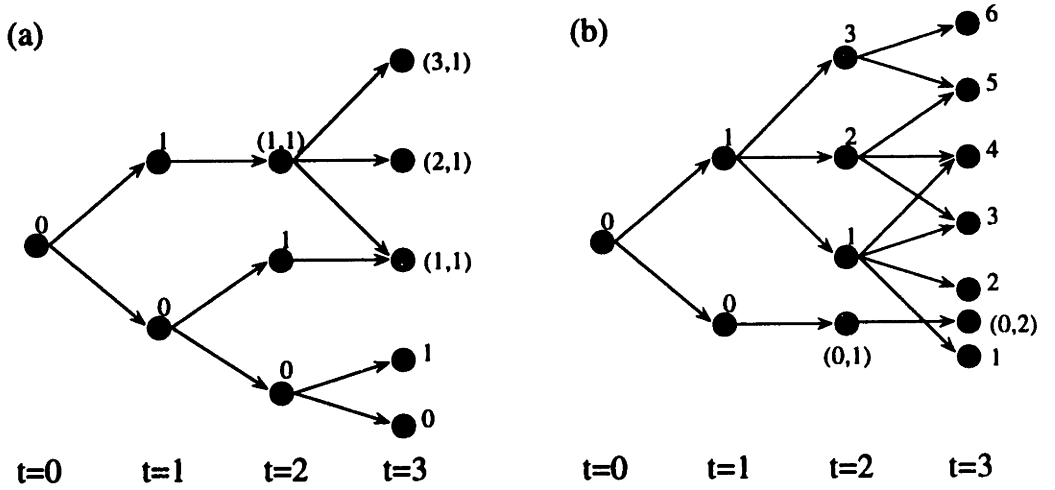

Figure 3-7: (a) Result of state aggregation in states 2 and 3 at time 2, followed by state aggregation in state 1 at time 1, in the event tree of figure 3-6(a). (b) Result of state aggregation in state (0,1) at time 2 in the event tree of figure 3-6(b).

written as $\hat{\pi}_{t/t+1}^{(n,k)/(n',k')}$, with $k' \leq k+1$. We will use the following definitions:

$$P_t^{(n,k)} = \begin{cases} P_t^n & \text{if } k=0 \\ \sum_{n' \in \mathcal{N}_{t-1}^{(n,k-1)}(k-1)} \hat{\pi}_{t-1/t}^{(n,k-1)/(n',k-1)} \cdot P_t^{(n',k-1)} & \text{if } k=1,\ldots,t \end{cases} \quad (3.11)$$

$$D_t^{(n,k)} = \begin{cases} D_t^n & \text{if } k=0 \\ \sum_{n' \in \mathcal{N}_{t-1}^{(n,k-1)}(k-1)} \hat{\pi}_{t-1/t}^{(n,k-1)/(n',k-1)} \cdot D_t^{(n',k-1)} & \text{if } k=1,\ldots,t \end{cases} \quad (3.12)$$

$$S_t^{(n,k)} = \begin{cases} S_t^n & \text{if } k=0 \\ \sum_{n' \in \mathcal{N}_{t-1}^{(n,k-1)}(k-1)} \hat{\pi}_{t-1/t}^{(n,k-1)/(n',k-1)} \cdot S_t^{(n',k-1)} & \text{if } k=1,\ldots,t \end{cases} \quad (3.13)$$

$$\hat{\pi}_{t/t+1}^{(n,k)/(n',k)} = \begin{cases} \hat{\pi}_{t/t+1}^{n/n'} & \text{if } k=0 \\ \hat{\pi}_{t-1/t}^{(n,k-1)/(n',k-1)} \left(\dfrac{P_t^{(n',k-1)}}{P_t^{(n,k)}} \right) & \text{if } k=1,\ldots,t \end{cases} \quad (3.14)$$

where $n \leq N_{t-k}$; $\hat{\pi}_{t/t+1}^{n/n'}$ denotes the risk-neutral conditional probability in the original event tree. Note that it follows from definition (3.14) that

$$\sum_{n' \in \mathcal{N}_t^{(n,k)}(k)} \hat{\pi}_{t/t+1}^{(n,k)/(n',k)} = \sum_{n' \in \mathcal{N}_{t-1}^{(n,k-1)}(k-1)} \hat{\pi}_{t-1/t}^{(n,k-1)/(n',k-1)} \left(\frac{P_t^{(n',k-1)}}{P_t^{(n,k)}} \right) = 1$$

where we have used $\mathcal{N}_{t-1}^{(n,k-1)}(k-1) = \mathcal{N}_t^{(n,k)}(k)$ in the first equality, and definition (3.11) in the second equality. This shows that the $\hat{\pi}$'s as defined in (3.14) are indeed probabilities. The aggregated quantities in (3.11)–(3.13) for $k \geq 1$ are therefore in effect weighted averages of the quantities with aggregation level $k-1$, where the weights are the $\hat{\pi}$'s of (3.14).

For a state (t,n,k) with successors of aggregation level k at time $t+1$, the following proposition states that the conditional probabilities in (3.14) define a risk-neutral probability measure on the aggregated event tree if interest rates, asset prices and dividends are calculated according to the formulas in (3.11)–(3.13). As a consequence, the aggregated asset prices in the tree do not admit arbitrage opportunities. Furthermore, because asset prices at time 0 cannot be aggregated, they are the same in the aggregated and the original (i.e., unaggregated) event tree.

Proposition 3.1 *If security prices are arbitrage-free in the unaggregated event tree, then definitions* (3.11)–(3.14) *imply for all* $t = 0,\ldots,T-1$, $k = 0,\ldots,t$ *and* $n =$

$0, \ldots, N_{t-k}$

$$S_t^{(n,k)} = P_t^{(n,k)} \cdot \left(\sum_{n' \in \mathcal{N}_t^{(n,k)}(k)} \hat{\pi}_{t/t+1}^{(n,k)/(n',k)} \cdot \left(S_{t+1}^{(n',k)} + D_{t+1}^{(n',k)} \right) \right) \tag{3.15}$$

$$= P_t^{(n,k)} \cdot \left(S_{t+1}^{(n,k+1)} + D_{t+1}^{(n,k+1)} \right) \tag{3.16}$$

PROOF: If equation (3.15) is true, then equation (3.16) follows directly from definitions (3.13) and (3.12). We will prove equation (3.15) by induction on k.

For $k = 0$, equation (3.15) is the arbitrage-free pricing relation in the unaggregated event tree (compare with equation (3.8) for the Ho and Lee model), which holds by assumption. Suppose that equation (3.15) holds for $k = \bar{k}$. We will show that it also holds for $k = \bar{k} + 1$.

From definition (3.13):

$$S_t^{(n,\bar{k}+1)} = \sum_{n' \in \mathcal{N}_{t-1}^{(n,\bar{k})}(\bar{k})} \hat{\pi}_{t-1/t}^{(n,\bar{k})/(n',\bar{k})} \cdot S_t^{(n',\bar{k})} \tag{3.17}$$

From the induction hypothesis:

$$S_t^{(n',\bar{k})} = P_t^{(n',\bar{k})} \cdot \left(S_{t+1}^{(n',\bar{k}+1)} + D_{t+1}^{(n',\bar{k}+1)} \right)$$

By substituting this equation in (3.17), and using $\mathcal{N}_{t-1}^{(n,\bar{k})}(\bar{k}) = \mathcal{N}_t^{(n,\bar{k}+1)}(\bar{k}+1)$ and definition (3.14), we get:

$$S_t^{(n,\bar{k}+1)} = \sum_{n' \in \mathcal{N}_{t-1}^{(n,\bar{k})}(\bar{k})} \hat{\pi}_{t-1/t}^{(n,\bar{k})/(n',\bar{k})} \cdot P_t^{(n',\bar{k})} \cdot \left(S_{t+1}^{(n',\bar{k}+1)} + D_{t+1}^{(n',\bar{k}+1)} \right)$$

$$= P_t^{(n,\bar{k}+1)} \cdot \left(\sum_{n' \in \mathcal{N}_t^{(n,\bar{k}+1)}(\bar{k}+1)} \hat{\pi}_{t-1/t}^{(n,\bar{k})/(n',\bar{k})} \left(\frac{P_t^{(n',\bar{k})}}{P_t^{(n,\bar{k}+1)}} \right) \left(S_{t+1}^{(n',\bar{k}+1)} + D_{t+1}^{(n',\bar{k}+1)} \right) \right)$$

$$= P_t^{(n,\bar{k}+1)} \cdot \left(\sum_{n' \in \mathcal{N}_t^{(n,\bar{k}+1)}(\bar{k}+1)} \hat{\pi}_{t/t+1}^{(n,\bar{k}+1)/(n',\bar{k}+1)} \cdot \left(S_{t+1}^{(n',\bar{k}+1)} + D_{t+1}^{(n',\bar{k}+1)} \right) \right),$$

which had to be proved.

$\mathcal{QED}$.

Equations (3.15) and (3.16) in proposition 3.1 imply that asset prices in an aggregated tree, as defined in (3.13), are arbitrage-free when the aggregation level of a state is *lower than or equal to* the aggregation level of its successors. However, by

substituting definitions (3.13) and (3.12) recursively into (3.15), the aggregated asset prices in a state can be written as a function of the prices in its successor states in case the aggregation level of the state is *higher* than that of its successors. This recursive substitution directly indicates the appropriate risk-neutral conditional probabilities $\hat{\pi}_{t/t+1}^{(n,k)/(n',k')}$ that guarantee that the asset prices are also arbitrage-free in this case, where $k' < k$ is the aggregation level of the successors of state (t, n, k) (note that relation (3.14) only defines the risk-neutral conditional probabilities for the case that the successors of an aggregated state have the same aggregation level as the state itself). We thus have the following result:

Corollary 3.1 *Asset prices in an aggregated event tree that is obtained after one or more state aggregations have been performed in the original event tree, are arbitrage-free if asset prices in the original tree are arbitrage-free, and if interest rates, asset prices and dividends in the aggregated tree are calculated according to the formulas* (3.11)–(3.13).

Although aggregated asset prices thus remain arbitrage-free when state aggregations are performed, they may no longer satisfy certain relationships that were satisfied in the original event tree. To see this, suppose that state n at time t in the unaggregated event tree has two successor nodes at time $t+1$, denoted here as n_1 and n_2. Let the risk-neutral conditional probability of state n_1 at time $t+1$, given state n at time t, be equal to $1/2$, and suppose there is a bond with a price of 95 in state n_1 and 105 in state n_2. Consider a call option on this bond which expires at time $t+1$, and has an exercise price of 100. By definition, its dividends in states n_1 and n_2 are respectively 0 and 5. If we perform state aggregation in node n at time t, then the aggregated bond price at time $t+1$ will be equal to 100, while the aggregated option dividend becomes 2.5. The payoff on the option in the aggregated tree thus violates its definition as the positive part of the difference between the bond price and the exercise price[10].

Time Aggregation

Time aggregation involves the merging of time steps in an event tree. Specifically, we say that time aggregation is performed in state (t, n, k) in an (aggregated) event tree

[10]Such violations will occur for any derivative security with payoffs that are a nonlinear function of the value of the underlying security. When the payoff pattern is convex or concave, Jensen's inequality can be used to formally show this.

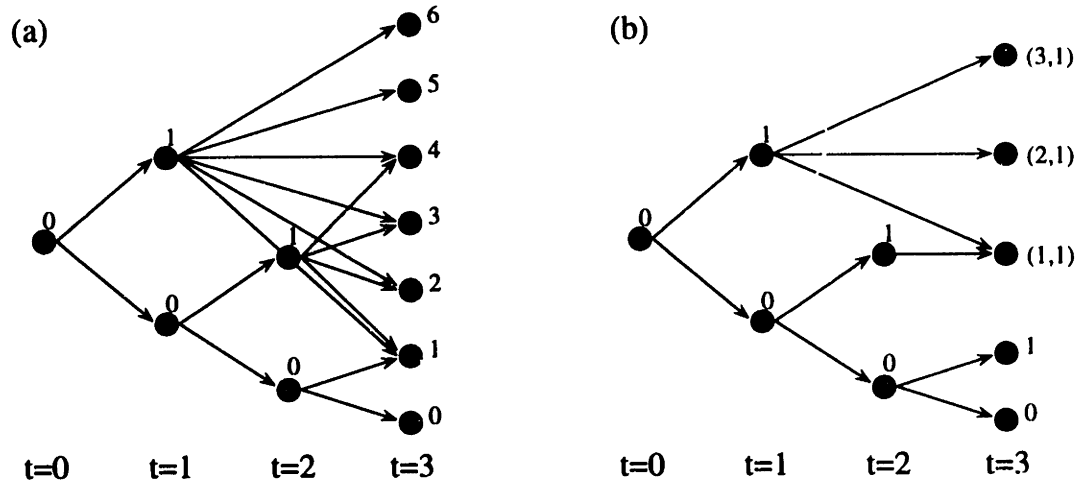

Figure 3-8: (a) Result of time aggregation in state 1 at time 1 in the event tree of figure 3-5. (b) Result of time aggregation in state 1 at time 1 in the event tree of figure 3-7(a).

if we replace the transitions from state (t, n, k) to its successors by direct transitions from state (t, n, k) to the successors of its successors. We have illustrated this in figure 3-8. Figure 3-8(a) depicts the event tree after time aggregation is performed in state 1 at time 1 in the event tree of figure 3-5. Figure 3-8(b) shows the result if time aggregation is performed in state 1 at time 1 in the aggregated event tree of figure 3-7(a).

We impose two restrictions on time aggregation. First, we require that all successors of a state occur at the same point in time in the event tree. This implies that we cannot perform time aggregation in state 0 at time 0 in the event tree of figure 3-8(a), as it would cause this state to have successors at both time 2 and time 3. The second restriction is the one we also imposed for state aggregation, namely that all successor states of a state in the event tree have the same aggregation level. This restriction prohibits time aggregation in state 0 at time 1 in the event tree of figure 3-8(b).

If state (t, n, k) in an aggregated event tree has its successors, all of aggregation level k', at time $t + \tau$ $(\tau \geq 1)$ then we denote the set of node numbers of these successors by $\mathcal{N}_{t/t+\tau}^{(n,k)}(k')$. (From the discussion in the previous section, $k' \leq k + \tau$.) Thus, $\mathcal{N}_{1/3}^{(1,0)}(0) = \{1, 2, 3, 4, 5, 6\}$ in figure 3-8(a), and $\mathcal{N}_{1/3}^{(1,0)}(1) = \{1, 2, 3\}$ in figure 3-8(b). It is not difficult to see that $\mathcal{N}_{t/t+\tau}^{(n,k)}(k') = \mathcal{N}_{t/t+\tau+\bar{\tau}}^{(n,k)}(k' + \bar{\tau})$ and $\mathcal{N}_{t/t+\tau}^{(n,k)}(k') = \mathcal{N}_{t-\bar{\tau}/t+\tau}^{(n,k-\bar{\tau}}(k')$ with $k \leq t$ and $k' \leq t + \tau$. Furthermore, $\mathcal{N}_{t/t+\tau}^{(n,k)}(k + \tau) = \{n\}$.

Because time aggregation in state (t, n, k) increases the length of the period fol-

lowing that state in the event tree, the definition of the "short-term" riskless interest rate in that state needs to be adjusted to account for this longer period. Let $r^{(n,k)}_{t\to t+\tau}$ denote the (continuously compounded) riskless interest rate in state (t,n,k) that is valid between time t and $t+\tau$, and define $P^{(n,k)}_{t\to t+\tau} \equiv \exp(-r^{(n,k)}_{t\to t+\tau}\cdot\Delta\tau)$ (i.e., $P^{(n,k)}_{t\to t+\tau}$ can be interpreted as the price in state (t,n,k) of a riskless investment that pays \$1 at time $t+\tau$). The value of $P^{(n,k)}_{t\to t+\tau}$ (and implicitly $r^{(n,k)}_{t\to t+\tau}$) is defined by:

$$P^{(n,k)}_{t\to t+\tau} = \prod_{j=0}^{\tau-1} P^{(n,k+j)}_{t+j} \tag{3.18}$$

Our next concern is how to account for dividend payments in states that are eliminated from the event tree due to time aggregation. In principle, we have two options (besides ignoring them): bring them forward in time to the state in which time aggregation is performed, or postpone them to the successor states of the state that is eliminated. It is easy to see that this last method introduces path-dependency of asset payoffs in the aggregated event tree if the original tree has a lattice structure, and we therefore choose the first option. To be precise, if time aggregation is performed in state (t,n,k), then we assume that the *arbitrage-free value* of the dividends in its successor states is paid out in state (t,n,k). Let $D^{(n,k)}_{t\to t+\tau}$ denote the arbitrage-free value in state (t,n,k) of all dividends paid out in its descendant states between time t and $t+\tau$ (with the dividends at times t and $t+\tau$ excluded). From proposition 3.1, and using (3.18):

$$D^{(n,k)}_{t\to t+\tau} = \begin{cases} 0 & \text{for} \quad \tau = 1 \\ \displaystyle\sum_{j=1}^{\tau-1} P^{(n,k)}_{t\to t+j}\cdot D^{(n,k+j)}_{t+j} & \text{for} \quad \tau > 1, \end{cases} \tag{3.19}$$

where $D^{(n,k+j)}_{t+j}$ is defined by (3.12) for all j.

In an aggregated event tree, suppose that state (n,k) at time t has successors at time $t+\tau$ ($\tau \geq 1$), all of aggregation level $k+\tau-1$. We define the risk-neutral conditional probability $\hat{\pi}^{(n,k)/(n',k+\tau-1)}_{t/t+\tau}$ of visiting the successor state $(t+\tau, n', k+\tau-1)$ as:

$$\hat{\pi}^{(n,k)/(n',k+\tau-1)}_{t/t+\tau} = \hat{\pi}^{(n,k+\tau-1)/(n',k+\tau-1)}_{t+\tau-1/t+\tau} \tag{3.20}$$

The next proposition generalizes proposition 3.1 to event trees in which both state and time aggregations have been performed. It relates the aggregated asset prices in a state (t,n,k) to the prices in its successor states at some time $t+\tau$ ($\tau \geq 1$) when all successors have aggregation level $k+\tau-1$.

Proposition 3.2 *If security prices are arbitrage-free in the unaggregated event tree, then definitions* (3.11)-(3.14) *and* (3.18)-(3.20) *imply the following relation for all* $t = 0, \ldots, T-1$, $k = 0, \ldots, t$, $n = 0, \ldots, N_{t-k}$ *and* $\tau = 1, \ldots, T - t$:

$$S_t^{(n,k)} = D_{t\to t+\tau}^{(n,k)} + P_{t\to t+\tau}^{(n,k)} \cdot \left(\sum_{n' \in \mathcal{N}_{t/t+\tau}^{(n,k)}(k+\tau-1)} \hat{\pi}_{t/t+\tau}^{(n,k)/(n',k+\tau-1)} \cdot \left(S_{t+\tau}^{(n',k+\tau-1)} + D_{t+\tau}^{(n',k+\tau-1)} \right) \right) \quad (3.21)$$

$$= D_{t\to t+\tau}^{(n,k)} + P_{t\to t+\tau}^{(n,k)} \left(S_{t+\tau}^{(n,k+\tau)} + D_{t+\tau}^{(n,k+\tau)} \right) \quad (3.22)$$

PROOF: Equation (3.22) follows directly from equation (3.21) because of the identities $\mathcal{N}_{t/t+\tau}^{(n,k)}(k+\tau-1) = \mathcal{N}_{t+\tau-1/t+\tau}^{(n,k+\tau-1)}(k+\tau-1)$ and $\hat{\pi}_{t/t+\tau}^{(n,k)/(n',k+\tau-1)} = \hat{\pi}_{t+\tau-1/t+\tau}^{(n,k+\tau-1)/(n',k+\tau-1)}$ (definition (3.20)), and substitution of definitions (3.12) and (3.13). We will prove equation (3.21) by induction on τ and k.

For $\tau = 1$, proposition 3.2 reduces to proposition 3.1. Suppose that equation (3.21) holds for some $\tau = \bar{\tau} \geq 1$ and $k = \bar{k}$, with $0 \leq \bar{k} \leq t$. The proof that it is also true for $\tau = \bar{\tau}$ and $k = \bar{k}+1$ when $\bar{k} < t$ is analogous to the proof of proposition 3.1, and is therefore omitted here. We show that the equation will also hold for $\tau = \bar{\tau}+1$ and $k = \bar{k}$.

By hypothesis it is true that:

$$S_t^{(n,\bar{k})} = D_{t\to t+\bar{\tau}}^{(n,\bar{k})} + P_{t\to t+\bar{\tau}}^{(n,\bar{k})} \cdot \left(\sum_{n' \in \mathcal{N}_{t/t+\bar{\tau}}^{(n,\bar{k})}(\bar{k}+\bar{\tau}-1)} \hat{\pi}_{t+\bar{\tau}-1/t+\bar{\tau}}^{(n,\bar{k}+\bar{\tau}-1)/(n',\bar{k}+\bar{\tau}-1)} \left(S_{t+\bar{\tau}}^{(n',\bar{k}+\bar{\tau}-1)} + D_{t+\bar{\tau}}^{(n',\bar{k}+\bar{\tau}-1)} \right) \right) \quad (3.23)$$

where we have used definition (3.20). As $\mathcal{N}_{t/t+\bar{\tau}}^{(n,\bar{k})}(\bar{k}+\bar{\tau}-1) = \mathcal{N}_{t+\bar{\tau}-1/t+\bar{\tau}}^{(n,\bar{k}+\bar{\tau}-1)}(\bar{k}+\bar{\tau}-1)$, we have from definition (3.12):

$$\sum_{n' \in \mathcal{N}_{t/t+\bar{\tau}}^{(n,\bar{k})}(\bar{k}+\bar{\tau}-1)} \hat{\pi}_{t+\bar{\tau}-1/t+\bar{\tau}}^{(n,\bar{k}+\bar{\tau}-1)/(n',\bar{k}+\bar{\tau}-1)} \cdot D_{t+\bar{\tau}}^{(n',\bar{k}+\bar{\tau}-1)} = D_{t\to\bar{\tau}}^{(n,\bar{k}+\bar{\tau})}$$

Noting further that $D_{t\to t+\bar{\tau}+1}^{(n,\bar{k})} = D_{t\to t+\bar{\tau}}^{(n,\bar{k})} + P_{t\to t+\bar{\tau}}^{(n,\bar{k})} \cdot D_{t+\bar{\tau}}^{(n,\bar{k}+\bar{\tau})}$, we can rewrite (3.23) as:

$$S_t^{(n,\bar{k})} = D_{t\to t+\bar{\tau}+1}^{(n,\bar{k})} + P_{t\to t+\bar{\tau}}^{(n,\bar{k})} \cdot \left(\sum_{n' \in \mathcal{N}_{t/t+\bar{\tau}}^{(n,\bar{k})}(\bar{k}+\bar{\tau}-1)} \hat{\pi}_{t+\bar{\tau}-1/t+\bar{\tau}}^{(n,\bar{k}+\bar{\tau}-1)/(n',\bar{k}+\bar{\tau}-1)} \cdot S_{t+\bar{\tau}}^{(n',\bar{k}+\bar{\tau}-1)} \right) \quad (3.24)$$

From proposition 3.1:

$$S_{t+\bar{\tau}}^{(n',\bar{k}+\bar{\tau}-1)} = P_{t+\bar{\tau}}^{(n',\bar{k}+\bar{\tau}-1)} \cdot \left(S_{t+\bar{\tau}+1}^{(n',\bar{k}+\bar{\tau})} + D_{t+\bar{\tau}+1}^{(n',\bar{k}+\bar{\tau})}\right).$$

Substituting this in (3.24), using definition (3.18) to write $P_{t\to t+\bar{\tau}}^{(n,\bar{k})} \cdot P_{t+\bar{\tau}}^{(n,\bar{k}+\bar{\tau})} = P_{t\to t+\bar{\tau}+1}^{(n,\bar{k})}$, and definition (3.14) to write:

$$\hat{\pi}_{t+\bar{\tau}-1/t+\bar{\tau}}^{(n,\bar{k}+\bar{\tau}-1)/(n',\bar{k}+\bar{\tau}-1)} \left(\frac{P_{t+\bar{\tau}}^{(n',\bar{k}+\bar{\tau}-1)}}{P_{t+\bar{\tau}}^{(n,\bar{k}+\bar{\tau})}}\right) = \hat{\pi}_{t+\bar{\tau}/t+\bar{\tau}+1}^{(n,\bar{k}+\bar{\tau})/(n',\bar{k}+\bar{\tau})}$$

we get:

$$\begin{aligned} S_t^{(n,\bar{k})} &= D_{t\to t+\bar{\tau}+1}^{(n,\bar{k})} + P_{t\to t+\bar{\tau}+1}^{(n,\bar{k})} \cdot \\ &\quad \left(\sum_{n'\in\mathcal{N}_{t/t+\bar{\tau}}^{(n,\bar{k})}(\bar{k}+\bar{\tau}-1)} \hat{\pi}_{t+\bar{\tau}/t+\bar{\tau}+1}^{(n,\bar{k}+\bar{\tau})/(n',\bar{k}+\bar{\tau})} \cdot \left(S_{t+\bar{\tau}+1}^{(n',\bar{k}+\bar{\tau})} + D_{t+\bar{\tau}+1}^{(n',\bar{k}+\bar{\tau})}\right)\right) \\ &= D_{t\to t+\bar{\tau}+1}^{(n,\bar{k})} + P_{t\to t+\bar{\tau}+1}^{(n,\bar{k})} \cdot \\ &\quad \left(\sum_{n'\in\mathcal{N}_{t/t+\bar{\tau}+1}^{(n,\bar{k})}(\bar{k}+\bar{\tau})} \hat{\pi}_{t/t+\bar{\tau}+1}^{(n,\bar{k})/(n',\bar{k}+\bar{\tau})} \cdot \left(S_{t+\bar{\tau}+1}^{(n',\bar{k}+\bar{\tau})} + D_{t+\bar{\tau}+1}^{(n',\bar{k}+\bar{\tau})}\right)\right) \end{aligned}$$

where definition (3.20) and the identity $\mathcal{N}_{t/t+\bar{\tau}}^{(n,\bar{k})}(\bar{k}+\bar{\tau}-1) = \mathcal{N}_{t/t+\bar{\tau}+1}^{(n,\bar{k})}(\bar{k}+\bar{\tau})$ are used in the last equality. This completes the proof.

$\mathcal{QED}$.

By substituting relations (3.12) and (3.13) into equation (3.21), one can write $\left(S_t^{(n,k)} - D_{t\to t+\tau}^{(n,k)}\right)$ as a discounted weighted average of aggregated asset prices and dividends in the successors of state (t,n,k) at time $t+\tau$ when the aggregation level of these successors is lower than $k+\tau-1$.

The added dividend term $D_{t\to t+\tau}^{(n,k)}$ in equation (3.21) somewhat complicates its interpretation as compared to the interpretation of equation (3.15) in proposition 3.1, which was recorded in corollary 3.1. If we view $D_{t\to t+\tau}^{(n,k)}$ as a dividend on the security that is paid out directly after time t, then equation (3.21) can be interpreted as an arbitrage pricing relation. We can also rewrite equation (3.21) as:

$$\begin{aligned} S_t^{(n,k)} &= P_{t\to t+\tau}^{(n,k)} \cdot \\ &\left(\sum_{n'\in\mathcal{N}_{t/t+\tau}^{(n,k)}(k+\tau-1)} \hat{\pi}_{t/t+\tau}^{(n,k)/(n',k+\tau-1)} \left(S_{t+\tau}^{(n',k+\tau-1)} + D_{t+\tau}^{(n',k+\tau-1)} + \left(\frac{D_{t\to t+\tau}^{(n,k)}}{P_{t\to t+\tau}^{(n,k)}}\right)\right)\right) \end{aligned}$$

in which case $\left(D_{t\to t+\tau}^{(n,k)}/P_{t\to t+\tau}^{(n,k)}\right)$ can be viewed as a postponed dividend that is paid out in each successor state of state (t,n,k) in the aggregated tree. Again, equation (3.21) can be interpreted as an arbitrage pricing relation. As noted before, however, postponing the interim dividends to a later point in time makes them predecessor-dependent in the successor states if the original event tree had a lattice structure.

3.4 The Aggregated ALM Model

In this section we present the formulation of the ALM model when it is based on an aggregated event tree, obtained from some original event tree by multiple state and time aggregations. This model will be called the *aggregated ALM model.* For its formulation, we need to define what the liabilities and the upper bounds on short-term borrowing (which we assumed to be state dependent) are in the states of such an aggregated event tree. Not surprisingly, we define these quantities in a similar way as the aggregated interest rates, asset prices and dividends in the previous section.

Let $L_t^{(n,k)}$ denote the liability and $\bar{Z}_t^{(n,k)}$ the upper bound on short-term borrowing in state (t,n,k) of an event tree in which only state aggregations have been performed. $L_t^{(n,k)}$ and $\bar{Z}_t^{(n,k)}$ are defined by the recursive relations:

$$L_t^{(n,k)} = \begin{cases} L_t^n & \text{if} \quad k=0 \\ \sum\limits_{n'\in\mathcal{N}_{t-1/t}^{(n,k-1)}(k-1)} \hat{\pi}_{t-1/t}^{(n,k-1)/(n',k-1)}\, L_t^{(n',k-1)} & \text{if} \quad k=1,\ldots,t. \end{cases} \tag{3.25}$$

$$\bar{Z}_t^{(n,k)} = \begin{cases} \bar{Z}_t^n & \text{if} \quad k=0 \\ \sum\limits_{n'\in\mathcal{N}_{t-1/t}^{(n,k-1)}(k-1)} \hat{\pi}_{t-1/t}^{(n,k-1)/(n',k-1)}\, \bar{Z}_t^{(n',k-1)} & \text{if} \quad k=1,\ldots,t. \end{cases} \tag{3.26}$$

When time aggregation is performed in the event tree and states are eliminated in which liabilities are due, then the assumption is made that the arbitrage-free value of the liabilities has to be paid in the predecessor state of the state at which they were due originally. (This is analogous to the assumption in section 3.3.2 about the prepayment of dividends in states that are eliminated from the event tree.) If $L_{t\to t+\tau}^{(n,k)}$ denotes the arbitrage-free value (in state (t,n,k)) of the liabilities that have to be paid in state (t,n,k) and all its descendant states between time t and $t+\tau$ (with the liabilities at time $t+\tau$ excluded), then $L_{t\to t+\tau}^{(n,k)}$ equals:

$$L_{t\to t+\tau}^{(n,k)} = \begin{cases} L_t^{(n,k)} & \text{if} \quad \tau = 1 \\ L_t^{(n,k)} + \sum_{j=1}^{\tau-1} P_{t\to t+j}^{(n,k)} \cdot L_{t+j}^{(n,k+j)} & \text{if} \quad \tau > 1, \end{cases} \tag{3.27}$$

which follows from proposition 3.1 and definitions (3.18) and (3.25).

The elimination of states due to time aggregation also requires an extra definition of a short-term borrowing limit that extends over multiple periods. Define:

$$\bar{Z}_{t\to t+\tau}^{(n,k)} = \min_{j=0,\ldots,\tau-1} \left\{ \frac{\bar{Z}_{t+j}^{(n,k+j)}}{e^{-\rho\Delta(\tau-j)} P_{t+j\to t+\tau}^{(n,k+j)}} \right\} \tag{3.28}$$

This upper bound makes sure that the amount of short-term borrowing will not exceed any of the (aggregated) upper bounds between time t and $t+\tau$ if the amount that is borrowed in state (n,k) at time t is rolled over from period to period until time $t+\tau$.

The trading dates in the aggregated event tree are written as $t_0, t_1, \ldots, t_T$, with $t_0 = 0$ and $t_T = H$. Furthermore, period j in the aggregated event tree (i.e., the time between t_{j-1} and t_j) consists of $\tau_j \equiv t_j - t_{j-1}$ time steps in the original event tree (each of which had length Δ). In analogy with the definition of q_t^s in section 2.3, the discount factors in the aggregated ALM model are defined as:

$$q_{t_j}^s = \begin{cases} 1 & \text{if} \quad j = 0 \\ q_{t_{j-1}}^{s^-} \left(P_{t_{j-1}\to t_j}^{n(s^-)} \, \hat{\pi}_{t_{j-1}/t_j}^{n(s^-)/n(s)} \right) & \text{if} \quad j = 1, \ldots, T. \end{cases} \tag{3.29}$$

The aggregated ALM model can now be formulated as:

$$\begin{aligned} v = \min \quad & (S_{t_0} - D_{t_0\to t_1}) \, x\!h_{t_0} + P_{t_0\to t_1} y_{t_0} - \left(e^{-\rho\Delta\tau_1} P_{t_0\to t_1} \right) z_{t_0} \\ & \qquad - \lambda_1 \sum_{s\in\mathcal{S}_{t_T}} q_{t_T}^s \left(y_{t_T}^s - \lambda_2 z_{t_T}^s \right) + L_{t_0\to t_1} \\ \text{s.t.} \quad & D_{t_j}^{n(s)} x\!h_{t_{j-1}}^{s^-} + y_{t_{j-1}}^{s^-} - z_{t_{j-1}}^{s^-} + (1-c) S_{t_j}^{n(s)} x\!s_{t_j}^s - (1+c) S_{t_j}^{n(s)} x\!b_{t_j}^s \\ & \qquad + D_{t_j\to t_{j+1}}^{n(s)} x\!h_{t_j}^s - P_{t_j\to t_{j+1}}^{n(s)} y_{t_j}^s + \left(e^{-\rho\Delta\tau_{j+1}} P_{t_j\to t_{j+1}}^{n(s)} \right) z_{t_j}^s \\ & \qquad = L_{t_j\to t_{j+1}}^{n(s)} \qquad \forall\, s \in \mathcal{S}_{t_j},\ j = 1, \ldots, T-1 \\ & x\!h_{t_{j-1}}^{s^-} - x\!s_{t_j}^s + x\!b_{t_j}^s - x\!h_{t_j}^s = 0 \qquad \forall\, s \in \mathcal{S}_{t_j},\ j = 1, \ldots, T-1 \\ & z_{t_j}^s \le \bar{Z}_{t_j\to t_{j+1}}^{n(s)} \qquad \forall\, s \in \mathcal{S}_{t_j},\ j = 1, \ldots, T-1 \\ & \left(D_{t_T}^{n(s)} + S_{t_T}^{n(s)} \right) x\!h_{t_{T-1}}^{s^-} + y_{t_{T-1}}^{s^-} - z_{t_{T-1}}^{s^-} - y_{t_T}^s + z_{t_T}^s = L_{t_T}^{n(s)} \quad \forall\, s \in \mathcal{S}_{t_T} \\ & z_{t_T}^s \le \bar{Z}_{t_T}^{n(s)} \qquad \forall\, s \in \mathcal{S}_{t_T} \end{aligned} \tag{3.30}$$

To simplify notation, this formulation assumes that all states at time t_j have successors at time t_{j+1}. Furthermore, we have not explicitly mentioned the aggregation level of quantities, but subsumed this in the index $n(s)$.

The main difference between the structure of this aggregated ALM model and the original model of section 2.2.1 stems from the prepayment of intermediate dividends. This is reflected in the extra terms $D^{n(s)}_{t_j \to t_{j+1}} x h^s_{t_j}$ in the cash-balance constraints, and the adjustment in the coefficient for $x h_{t_0}$ in the objective function. In addition, the objective function includes the constant $L_{t_0 \to t_1}$, the present value of the liabilities before time t_1.

The state aggregation and time aggregation methods thus enable us to use financial models of the term-structure uncertainty to build stochastic programming models for the ALM problem with a realistic and internally consistent description of the uncertainty in future interest rates and asset prices, and at the same time control the size of the optimization models. The next chapter will show how the aggregation methods can in addition be used as the basis for a flexible, iterative solution algorithm to solve these optimization models.

Chapter 4

Solving the ALM Problem by Iterative Disaggregation

In the previous chapter we have shown how the state and time aggregation methods can be used to reduce the number of states, and thereby the number of scenarios, in an event tree that describes the uncertainty in future interest rates and asset prices. We have seen that such a reduction is necessary if one wants to base the description of the uncertainty in the ALM model on a discrete-time term-structure model that is consistent with observed market prices. We proved the important result that asset prices in an aggregated event tree, resulting after the application of state and time aggregation, are arbitrage-free and consistent with observed market prices if this is true for the prices in the original event tree (which was assumed to be the case in assumption 4 of section 3.2). As was shown in section 2.3.2, this property is crucial in order to prevent unwanted biases in the optimal solution to the ALM model.

However, the optimal solution *will* in general depend on the level of uncertainty that is included in the ALM model. To obtain a robust investment portfolio, it is therefore important to include as much of the relevant uncertainty in the model as possible. In this chapter we present an iterative solution algorithm that gradually increases the level of uncertainty in the ALM model by reversing state and time aggregations that were performed to obtain the initial version of the model. The iterative nature of the algorithm allows us to judge where the uncertainty in the future affects the optimal solution most, and in what fashion. In addition, when more aggregations are reversed, events of decreasing probability are introduced in the model, and this enables a direct trade-off between the cost of the asset portfolio that hedges against the future liabilities and the probability of events that one wants to

be hedged against. As noted before, a stochastic programming model itself treats all scenarios in the model equally in the sense that it forces the constraints to be satisfied for each individual scenario, irrespective of its probability of occurrence.

Our analysis makes use of Zipkin's work [58, 57] on the aggregation of variables (*columns*) and constraints (*rows*) in general linear programs, and the relevant results of his work are summarized in section 4.1. Section 4.2 will show how both state and time aggregation correspond to the aggregation of variables and constraints in the stochastic programming formulation of the ALM problem. Because this implies that the ALM model after a state or time disaggregation is at the same time a relaxation (addition of variables) and a restriction (addition of constraints) of the ALM model before the disaggregation, it is not obvious how to recover a feasible solution for the disaggregated ALM model from an optimal solution to the aggregate ALM model. In section 4.3 we will show how one can always construct a feasible solution to a relaxation of the disaggregated ALM model. By choosing appropriate parameter values, we will furthermore show that this relaxation has the same optimal solutions as the true model.

4.1 Aggregation of Variables and Constraints in Linear Programs

This section follows Zipkin [58, 57]. Consider a linear program in the general form

$$\begin{array}{lll} v^* = & \min & cx \\ & \text{subject to} & Ax \geq b \\ & & x \geq 0 \end{array} \tag{4.1}$$

where $c = (c_j)$ is an n-vector, $b = (b_i)$ is an m-vector, $A = (a_{ij})$ is an $m \times n$ matrix, and x is an n-vector of variables. We will refer to this problem as the *original problem.*

Let $\rho = \{R_k : k = 1, \ldots, K\}$ be a partition of the set $\{1, \ldots, m\}$ and $\sigma = \{S_l : l = 1, \ldots, L\}$ a partition of $\{1, \ldots, n\}$, where $\mid R_k \mid = m_k$ and $\mid S_l \mid = n_l$. In an aggregation of (4.1), one replaces all rows in each set R_k by a single row, and all columns in each set S_l by a single column.

We assume that rows and columns are aggregated by taking weighted sums. For each k and l, let f^k be a nonnegative m_k-vector and g^l a nonnegative n_l-vector. These

vectors are called *weighting vectors*[1]. Define

$$\begin{array}{lllllllll} c^l & \equiv & (c_j)_{j\in S_l}, & \tilde{c}_l & \equiv & c^l g^l, & \text{and} & \tilde{c} & \equiv & (\tilde{c}_l) \\ b^k & \equiv & (b_i)_{i\in R_k}, & \tilde{b}_k & \equiv & b^k f^k, & \text{and} & \tilde{b} & \equiv & (\tilde{b}_k) \\ A^l_k & \equiv & (a_{ij})_{i\in R_k, j\in S_l}, & \tilde{a}_{kl} & \equiv & f^k A^l_k g^l, & \text{and} & \tilde{A} & \equiv & (\tilde{a}_{kl}) \end{array}$$

The *aggregate problem* is:

$$\begin{array}{lll} \tilde{v} = & \min & \tilde{c}X \\ & \text{subject to} & \tilde{A}X \geq \tilde{b} \\ & & X \geq 0 \end{array} \tag{4.2}$$

which has K constraints and an L-vector of variables X. We assume that both (4.1) and (4.2) have finite optimal primal and dual solutions.

Let $(\tilde{X}, \tilde{U})$ denote a pair of optimal primal and dual solutions to (4.2), and define the following solution to (4.1):

$$\begin{array}{lll} \tilde{x}^l & = & g^l \tilde{X}_l, \qquad l = 1, \ldots, L \\ \tilde{u}^k & = & \tilde{U}_k f^k, \qquad k = 1, \ldots, K \end{array}$$

If $\tilde{x}$ denotes the sequence of $\tilde{x}^l$'s in proper order for (4.1), and similarly for $\tilde{u}$, then $(\tilde{x}, \tilde{u})$ is called a *fixed-weight solution* to (4.1) and its dual, derived from $(\tilde{X}, \tilde{U})$. The following result is easy to see.

Proposition 4.1 $c\tilde{x} = \tilde{u}b = \tilde{v}$.

If only columns are aggregated (i.e., $K = m$ and $R_k = \{k\}$), then $\tilde{x}$ is feasible in (4.1) and thus $\tilde{v} \geq v^*$. However, if both columns and rows have been aggregated, then $\tilde{x}$ need not be feasible in (4.1), nor $\tilde{u}$ in the dual of (4.1), and it is therefore not clear whether $\tilde{v} \geq v^*$ or $\tilde{v} \leq v^*$. The fixed-weight solution $(\tilde{x}, \tilde{u})$ still satisfies an aggregate form of complementary slackness, namely $(\tilde{u}A - c)\tilde{x} = \tilde{u}(b - A\tilde{x}) = 0$.

To derive bounds on the deviation of $\tilde{v}$ from v^*, Zipkin assumes that (generalized) upper bounds are known for the values of primal and dual variables in an optimal solution to the original problem. Let $\rho' = \{R'_k : k = 1, \ldots, K'\}$ be a partition of $\{1, \ldots, m\}$ and $\sigma' = \{S'_l : l = 1, \ldots, L'\}$ a partition of $\{1, \ldots, n\}$. It is assumed that positive numbers $\{d_1, \ldots, d_n\}$ and nonnegative numbers $\{p_1, \ldots, p_{L'}\}$ are known such that some optimal solution x^* to (4.1) satisfies:

[1] Zipkin [57] assumes that columns and rows are aggregated by taking weighted *averages*, i.e., that the elements of each weighting vector f_k and g_l sum to one. We do not make that assumption here.

$$\sum_{j \in S'_l} d_j x^*_j \leq p_l, \qquad l = 1, \ldots, L' \tag{4.3}$$

as well as positive numbers $\{e_1, \ldots, e_m\}$ and nonnegative numbers $\{q_1, \ldots, q_{K'}\}$ such that some optimal solution u^* to the dual of (4.1) satisfies:

$$\sum_{i \in R'_k} u^*_i e_i \leq q_k, \qquad k = 1, \ldots, K'. \tag{4.4}$$

Let $A_{i.}$ denote the ith row of the matrix A, $A_{.j}$ the jth column, and $[y]^+$ the positive part of y. Zipkin [57] proves the following proposition.

Proposition 4.2 *If upper bounds* (4.3) *and* (4.4) *are known, then*

$$\tilde{v} - \epsilon^- \leq v^* \leq \tilde{v} + \epsilon^+ \tag{4.5}$$

where

$$\epsilon^+ \equiv \sum_{k=1}^{K'} \left[\max_{i \in R'_k} \{(b_i - A_{i.}\tilde{x})/e_i\} \right]^+ q_k$$

$$\epsilon^- \equiv \sum_{l=1}^{L'} \left[\max_{j \in S'_l} \{(\tilde{u}A_{.j} - c_j)/d_j\} \right]^+ p_l$$

PROOF: For the upper bound:

$$\begin{aligned} v^* = u^*b &\leq u^*b + (c - u^*A)\tilde{x} \\ &= c\tilde{x} + u^*(b - A\tilde{x}) \\ &= \tilde{v} + \sum_{k=1}^{K'} \sum_{i \in R'_k} u^*_i e_i \left(\frac{b_i - A_{i.}\tilde{x}}{e_i} \right) \\ &\leq \tilde{v} + \sum_{k=1}^{K'} \max_{i \in R'_k} \left\{ \frac{b_i - A_{i.}\tilde{x}}{e_i} \right\} \sum_{i \in R'_k} u^*_i e_i \end{aligned}$$

The upper bound on v^* now follows from (4.4). The lower bound on v^* is proved in a similar manner.

QED

These bounds on v^* thus provide a measure of the error that is introduced by solving the aggregate problem (4.2) instead of the original problem (4.1). Note that ϵ^+ is a function of the infeasibilities $[b_i - A_{i.}\tilde{x}]^+$ in (4.1) with respect to the fixed-weight solution $\tilde{x}$, and ϵ^- of the infeasibilities in the dual[2] of (4.1) with respect to $\tilde{u}$.

[2]It is tempting to think of the infeasibilities in the dual problem as corresponding to negative reduced costs of the associated primal variables, but we note that reduced costs are only defined

In the preceding discussion, we have assumed that the aggregate linear problem is solved to optimality before the bounds in proposition 4.2 are calculated. They also apply, however, if only a suboptimal solution to this problem is known. As primal suboptimality is equivalent to dual infeasibility, the value of ϵ^- will probably be higher in this case.

4.2 Aggregation in the ALM Model

In this section we show how both state and time aggregation correspond to the aggregation of variables and constraints in the ALM model. We can therefore use the results from the previous section to estimate the loss of accuracy when the ALM model with an aggregated description of the uncertainty is being solved. These results will be especially useful in the iterative disaggregation algorithm, which is the topic of section 4.3.

To apply state and time aggregation, we imposed the restriction in section 3.3.2 that all successors of a state in an aggregated event tree occur at the same point in time and have the same aggregation level. For expositional convenience, we will make the following additional assumption in this section: if the successors of a state (t, n, k) occur at time $t+\tau$ ($\tau \geq 1$), then the aggregation level of the successors is either $k+\tau$ (in which case there is only one successor) or $k+\tau-1$. (We note that this is the situation which was explicitly considered in propositions 3.1 and 3.2.) As was emphasized in section 3.3.2, the state and time aggregation are not limited to situations that conform to these assumptions, and this is true for the analysis in this section as well.

For the remainder of this section we furthermore assume that transaction costs are zero ($c = 0$), as positive transaction costs would only complicate the notation but not change the results. Our starting point is the aggregated ALM model of section 3.4.

4.2.1 State Aggregation

To see how state aggregation in the event tree corresponds to the aggregation of both variables and constraints in the stochastic programming formulation of the ALM problem, we consider the state aggregation that is depicted in figure 4-1. State aggregation is performed in state (n, k_0) at time t_j. We have defined $k_1 \equiv k_0 + \tau_{j+1}$,

with respect to a basis in the primal problem, and the fixed-weight solution may not define a basis. Furthermore, even if $\tilde{x}$ does define a basis, $\tilde{u}$ may not be the corresponding dual vector.

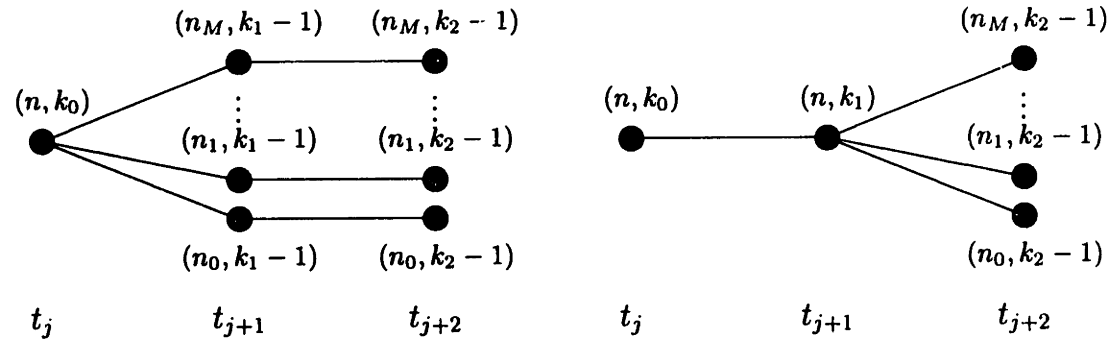

Figure 4-1: Basic state aggregation ($k_1 \equiv k_0 + \tau_{j+1}$; $k_2 \equiv k_1 + \tau_{j+2}$).

and $k_2 \equiv k_1 + \tau_{j+2}$, with $\tau_{j+1} \geq 1$ and $\tau_{j+2} \geq 1$. Before the aggregation, the successor states at time t_{j+1} of state (n, k_0) at time t_j have aggregation level $k_1 - 1$, and are identified by the node numbers $n_0, n_1, \ldots, n_M$. State aggregation in a situation as depicted in figure 4-1 will be called *basic state aggregation.*

Consider a scenario s^- in state (t_j, n, k_0) of figure 4-1(a). Let s_l denote its successor scenario in node $(n_l, k_1 - 1)$ at time t_{j+1}, and s_l^+ the descendant scenario of s^- in node $(n_l, k_2 - 1)$ at time t_{j+2} $(l = 0, 1, \ldots, M)$. Assume for now that $t_{j+2} < t_T = H$. The constraints in the aggregated ALM model that stem from scenario s^- in state (t_j, n, k_0) and correspond to the arcs in figure 4-1(a) are (for $l = 0, 1, \ldots, M$):

$$
\begin{aligned}
& D_{t_{j+1}}^{(n_l,k_1-1)} xh_{t_j}^{s^-} + y_{t_j}^{s^-} - z_{t_j}^{s^-} + S_{t_{j+1}}^{(n_l,k_1-1)} xs_{t_{j+1}}^{s_l} - S_{t_{j+1}}^{(n_l,k_1-1)} xb_{t_{j+1}}^{s_l} \\
& \qquad + D_{t_{j+1}\to t_{j+2}}^{(n_l,k_1-1)} xh_{t_{j+1}}^{s_l} - P_{t_{j+1}\to t_{j+2}}^{(n_l,k_1-1)} y_{t_{j+1}}^{s_l} + \left(e^{-\rho\Delta\tau_{j+2}} P_{t_{j+1}\to t_{j+2}}^{(n_l,k_1-1)}\right) z_{t_{j+1}}^{s_l} \\
& \qquad = L_{t_{j+1}\to t_{j+2}}^{(n_l,k_1-1)} \qquad (4.6) \\
& xh_{t_j}^{s^-} - xs_{t_{j+1}}^{s_l} + xb_{t_{j+1}}^{s_l} - xh_{t_{j+1}}^{s_l} = 0 \qquad (4.7) \\
& z_{t_{j+1}}^{s_l} \leq \bar{Z}_{t_{j+1}\to t_{j+2}}^{(n_l,k_1-1)} \qquad (4.8) \\
& D_{t_{j+2}}^{(n_l,k_2-1)} xh_{t_{j+1}}^{s_l} + y_{t_{j+1}}^{s_l} - z_{t_{j+1}}^{s_l} + S_{t_{j+1}}^{(n_l,k_2-1)} xs_{t_{j+2}}^{s_l^+} - S_{t_{j+1}}^{(n_l,k_2-1)} xb_{t_{j+2}}^{s_l^+} \\
& \qquad + D_{t_{j+2}\to t_{j+3}}^{(n_l,k_2-1)} xh_{t_{j+2}}^{s_l^+} - P_{t_{j+2}\to t_{j+3}}^{(n_l,k_2-1)} y_{t_{j+2}}^{s_l^+} + \left(e^{-\rho\Delta\tau_{j+3}} P_{t_{j+2}\to t_{j+3}}^{(n_l,k_2-1)}\right) z_{t_{j+2}}^{s_l^+} \\
& \qquad = L_{t_{j+2}\to t_{j+3}}^{(n_l,k_2-1)} \qquad (4.9) \\
& xh_{t_{j+1}}^{s_l} - xs_{t_{j+2}}^{s_l^+} + xb_{t_{j+2}}^{s_l^+} - xh_{t_{j+2}}^{s_l^+} = 0 \qquad (4.10) \\
& z_{t_{j+2}}^{s_l^+} \leq \bar{Z}_{t_{j+2}\to t_{j+3}}^{(n_l,k_2-1)} \qquad (4.11)
\end{aligned}
$$

where constraints (4.6)-(4.8) correspond to the successor scenarios at time t_{j+1}, and

constraints (4.9)-(4.11) to the descendant scenarios at time t_{j+2}

We will show how these constraints, and the variables in these constraints, can be aggregated to obtain the set of constraints in the ALM model that corresponds to figure 4-1(b), i.e., after the state aggregation in state (t_j, n, k_0). First, we multiply constraints (4.6)–(4.8) for each $l = 0, 1, \ldots, M$ by the risk-neutral conditional probability $\hat{\pi}^{n_l}_{t_{j+1}}$, and then sum these constraints over all l. Constraints (4.6)–(4.8) for all $l = 0, 1, \ldots, M$ are thus replaced by the three aggregated constraints:

$$\begin{aligned}
&\left(\sum_{l=0}^{M} \hat{\pi}^{n_l}_{t_{j+1}} D^{(n_l,k_1-1)}_{t_{j+1}}\right) xh^{s^-}_{t_j} + y^{s^-}_{t_j} - z^{s^-}_{t_j} + \sum_{l=0}^{M} \left(\hat{\pi}^{n_l}_{t_{j+1}} S^{(n_l,k_1-1)}_{t_{j+1}}\right) xs^{s_l}_{t_{j+1}} \\
&\quad - \sum_{l=0}^{M} \left(\hat{\pi}^{n_l}_{t_{j+1}} S^{(n_l,k_1-1)}_{t_{j+1}}\right) xb^{s_l}_{t_{j+1}} + \sum_{l=0}^{M} \left(\hat{\pi}^{n_l}_{t_{j+1}} D^{(n_l,k_1-1)}_{t_{j+1}\to t_{j+2}}\right) xh^{s_l}_{t_{j+1}} \\
&\quad - \sum_{l=0}^{M} \left(\hat{\pi}^{n_l}_{t_{j+1}} P^{(n_l,k_1-1)}_{t_{j+1}\to t_{j+2}}\right) y^{s_l}_{t_{j+1}} + e^{-\rho\Delta\tau_{j+2}} \sum_{l=0}^{M} \left(\hat{\pi}^{n_l}_{t_{j+1}} P^{(n_l,k_1-1)}_{t_{j+1}\to t_{j+2}}\right) z^{s_l}_{t_{j+1}} \\
&\quad = \sum_{l=0}^{M} \hat{\pi}^{n_l}_{t_{j+1}} L^{(n_l,k_1-1)}_{t_{j+1}\to t_{j+2}} \qquad (4.12)
\end{aligned}$$

$$xh^{s^-}_{t_j} - \sum_{l=0}^{M} \hat{\pi}^{n_l}_{t_{j+1}} \cdot xs^{s_l}_{t_{j+1}} + \sum_{l=0}^{M} \hat{\pi}^{n_l}_{t_{j+1}} \cdot xb^{s_l}_{t_{j+1}} - \sum_{l=0}^{M} \hat{\pi}^{n_l}_{t_{j+1}} \cdot xh^{s_l}_{t_{j+1}} = 0 \qquad (4.13)$$

$$\sum_{l=0}^{M} \hat{\pi}^{n_l}_{t_{j+1}} \cdot z^{s_l}_{t_{j+1}} \le \sum_{l=0}^{M} \hat{\pi}^{n_l}_{t_{j+1}} \bar{Z}^{(n_l,k_1-1)}_{t_{j+1}\to t_{j+2}} \qquad (4.14)$$

where we have used that $\sum_{l=0}^{M} \hat{\pi}^{n_l}_{t_{j+1}} = 1$. The second step is to aggregate the variables $xs^{s_l}_{t_{j+1}}$ over $l = 0, 1, \ldots, M$ by taking an unweighted sum (that is, add the constraint coefficients of all $xs^{s_l}_{t_{j+1}}$ in each constraint), and do the same for the variables $xb^{s_l}_{t_{j+1}}$, the variables $xh^{s_l}_{t_{j+1}}$, the variables $y^{s_l}_{t_{j+1}}$ and the variables $z^{s_l}_{t_{j+1}}$. We will denote the corresponding aggregated variables as $xs^{s}_{t_{j+1}}$, $xb^{s}_{t_{j+1}}$, $xh^{s}_{t_{j+1}}$, $y^{s}_{t_{j+1}}$ and $z^{s}_{t_{j+1}}$, respectively. Using the fact that $\sum_{l=0}^{M} \hat{\pi}^{n_l}_{t_{j+1}} D^{(n_l,k_1-1)}_{t_{j+1}} = D^{(n,k_1)}_{t_{j+1}}$, and analogous relations for the other data, the variable aggregations reduce the constraints (4.12)–(4.14) and (4.9)–(4.11) to:

$$\begin{aligned}
&D^{(n,k_1)}_{t_{j+1}} xh^{s^-}_{t_j} + y^{s^-}_{t_j} - z^{s^-}_{t_j} + S^{(n,k_1)}_{t_{j+1}} xs^{s}_{t_{j+1}} - S^{(n,k_1)}_{t_{j+1}} xb^{s}_{t_{j+1}} \\
&\quad + D^{(n,k_1)}_{t_{j+1}\to t_{j+2}} xh^{s}_{t_{j+1}} - P^{(n,k_1)}_{t_{j+1}\to t_{j+2}} y^{s}_{t_{j+1}} + \left(e^{-\rho\Delta\tau_{j+2}} P^{(n,k_1)}_{t_{j+1}\to t_{j+2}}\right) z^{s}_{t_{j+1}} \\
&\quad = L^{(n,k_1)}_{t_{j+1}\to t_{j+2}} \qquad (4.15)
\end{aligned}$$

$$xh^{s^-}_{t_j} - xs^{s}_{t_{j+1}} + xb^{s}_{t_{j+1}} - xh^{s}_{t_{j+1}} = 0 \qquad (4.16)$$

$$z^{s}_{t_{j+1}} \le \bar{Z}^{(n,k_1)}_{t_{j+1}\to t_{j+2}} \qquad (4.17)$$

$$D^{(n_l,k_2-1)}_{t_{j+2}} xh^s_{t_{j+1}} + y^s_{t_{j+1}} - z^s_{t_{j+1}} + S^{(n_l,k_2-1)}_{t_{j+1}} xs^{s_l^+}_{t_{j+2}} - S^{(n_l,k_2-1)}_{t_{j+1}} xb^{s_l^+}_{t_{j+2}}$$
$$+D^{(n_l,k_2-1)}_{t_{j+2}\to t_{j+3}} xh^{s_l^+}_{t_{j+2}} - P^{(n_l,k_2-1)}_{t_{j+2}\to t_{j+3}} y^{s_l^+}_{t_{j+2}} + \left(e^{-\rho\Delta r_{j+3}} P^{(n_l,k_2-1)}_{t_{j+2}\to t_{j+3}}\right) z^{s_l^+}_{t_{j+2}}$$
$$= L^{(n_l,k_2-1)}_{t_{j+2}\to t_{j+3}} \tag{4.18}$$

$$xh^s_{t_{j+1}} - xs^{s_l^+}_{t_{j+2}} + xb^{s_l^+}_{t_{j+2}} - xh^{s_l^+}_{t_{j+2}} = 0 \tag{4.19}$$

$$z^{s_l^+}_{t_{j+2}} \le \bar{Z}^{(n_l,k_2-1)}_{t_{j+2}\to t_{j+3}} \tag{4.20}$$

where constraints (4.18)–(4.20) are present for each $l = 0, 1, \ldots, M$. The set of constraints (4.15)–(4.20) corresponds precisely to the situation in figure 4-1(b), witho s denoting the single successor scenarios at time t_{j+1} of scenario s^- at time t_j after the state aggregation.

When $t_{j+2} = t_T$ or $t_{j+1} = t_T$, the row and column operations in the ALM model that correspond to state aggregation in state (t_j, n, k_0) are identical to the ones just described, although the constraints themselves will slightly differ from (4.6)–(4.11). State aggregation in more complicated situations than the one depicted in figure 4-1, for example when the aggregation level of the successors of state (t_j, n, k_0) is lower than $k_1 - 1$, or when the states at time t_{j+1} have more than one successor at time t_{j+2} (whose aggregation level is therefore lower than $k_2 - 1$), correspond to row and column operations in the ALM model that follow the same pattern as the operations for basic state aggregation: constraints relating to states that are aggregated at time t_{j+1} are combined by taking a weighted average, with the weights equal to the conditional probabilities, and variables in these constraints are combined by unweighted aggregation.

When $t_{j+1} = t_T$, the variable aggregations will also affect the objective function. Before the state aggregation, the term in the objective value that corresponds to the scenarios s_l is

$$\lambda_1 \sum_{l=0}^{M} q^{s_l}_{t_T} \left(y^{s_l}_{t_T} - \lambda_2 z^{s_l}_{t_T}\right)$$

Because definition (3.29) implies

$$\sum_{l=0}^{M} q^{s_l}_{t_T} = q^{s^-}_{t_{T-1}} P^n_{t_{T-1}\to t_T} \sum_{l=0}^{M} \hat{\pi}^{n_l}_{t_{j+1}} = q^{s^-}_{t_{T-1}} P^n_{t_{T-1}\to t_T} = q^s_{t_T}$$

we see that the aggregated objective function has precisely the form of the objective function in the ALM model after the state aggregation.

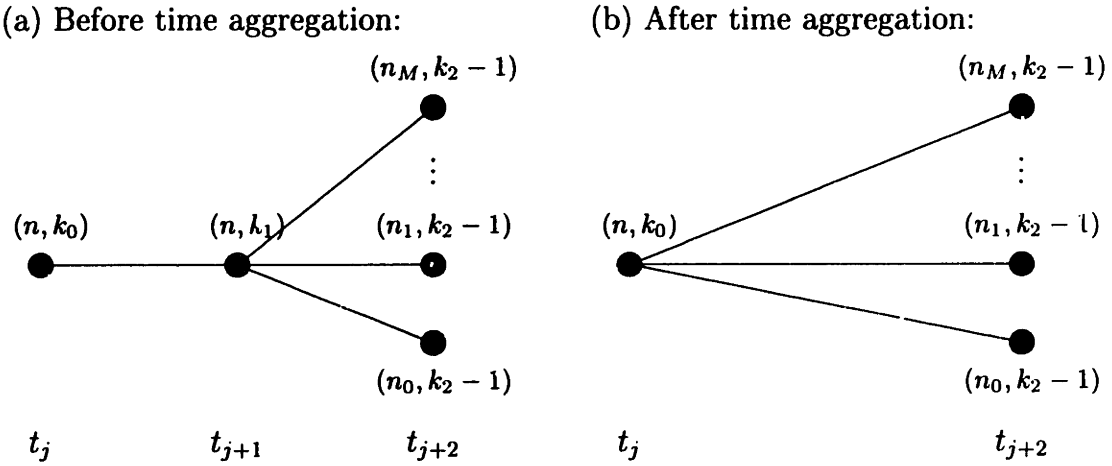

Figure 4-2: Basic time aggregation ($k_1 \equiv k_0 + \tau_{j+1}$; $k_2 \equiv k_1 + \tau_{j+2}$).

4.2.2 Time Aggregation

To show how time aggregation affects the variables and constraints in the ALM model, we consider the time aggregation of figure 4-2, which will be referred to as *basic time aggregation*. Time aggregation is performed in state (t_j, n, k_0) of figure 4-1(a). We note that the situation in figure 4-2(a) is identical to figure 4-1(b), and we employ the same notation as used there.

Assuming $t_{j+2} < t_T$, the constraints in the ALM model that stem from a scenario s^- in state (t_j, n, k_0) and correspond to the arcs in figure 4-2(a) are (4.15)–(4.20), where s denotes the successor of s^- at time t_{j+1} and s_l^+ its descendant in node $(n_l, k_2 - 1)$ at time t_{j+2} $(l = 0, 1, \ldots, M)$. To derive the constraints for the ALM model that corresponds to figure 4-2(b) by aggregating variables and constraints, we also use the constraints that link scenario s^- at time t_j to its predecessor (assume $t_j > 0$). Let s^{--} denote this predecessor scenario at time t_{j-1}. The constraints that link scenarios s^- and s^{--} are:

$$\begin{aligned} D_{t_j}^{(n,k_0)} xh_{t_{j-1}}^{s^{--}} + y_{t_{j-1}}^{s^{--}} - z_{t_{j-1}}^{s^{--}} + S_{t_j}^{(n,k_0)} xs_{t_j}^{s^-} - S_{t_j}^{(n,k_0)} xb_{t_j}^{s^-} \\ + D_{t_j \to t_{j+1}}^{(n,k_0)} xh_{t_j}^{s^-} - P_{t_j \to t_{j+1}}^{(n,k_0)} y_{t_j}^{s^-} + \left(e^{-\rho \Delta \tau_{j+1}} P_{t_j \to t_{j+1}}^{(n,k_0)}\right) z_{t_j}^{s^-} \\ = L_{t_j \to t_{j+1}}^{(n,k_0)} \end{aligned} \tag{4.21}$$

$$xh_{t_{j-1}}^{s^{--}} - xs_{t_j}^{s^-} + xb_{t_j}^{s^-} - xh_{t_j}^{s^-} = 0 \tag{4.22}$$

$$z_{t_j}^{s^-} \leq \bar{Z}_{t_j \to t_{j+1}}^{(n,k_0)} \tag{4.23}$$

Time aggregation in state (t_j, n, k_0) implies $xs^s_{t_{j+1}} = 0$ and $xb^s_{t_{j+1}} = 0$, and we substitute these constraints in the cash-balance and portfolio balance-constraints at time t_{j+1} (equations (4.15) and (4.16))[3]. The portfolio-balance constraint (4.16) is now $xh^{s^-}_{t_j} = xh^s_{t_{j+1}}$. If we substitute $xh^{s^-}_{t_j}$ for $xh^s_{t_{j+1}}$ in all constraints (or equivalently, aggregate these variables and call the resulting variable $xh^{s^-}_{t_j}$), then constraints (4.15) (4.20) become (note that the portfolio-balance constraint (4.16) has become a vacuous constraint):

$$\begin{aligned}
&\left(D^{(n,k_1)}_{t_{j+1}} + D^{(n,k_1)}_{t_{j+1}\to t_{j+2}}\right) xh^{s^-}_{t_j} + y^{s^-}_{t_j} - z^{s^-}_{t_j} - P^{(n,k_1)}_{t_{j+1}\to t_{j+2}} y^s_{t_{j+1}} \\
&\qquad + \left(e^{-\rho\Delta\tau_{j+2}} P^{(n,k_1)}_{t_{j+1}\to t_{j+2}}\right) z^s_{t_{j+1}} = L^{(n,k_1)}_{t_{j+1}\to t_{j+2}} && (4.24)\\
&z^s_{t_{j+1}} \le \bar{Z}^{(n,k_1)}_{t_{j+1}\to t_{j+2}} && (4.25)\\
&D^{(n_l,k_2-1)}_{t_{j+2}} xh^{s^-}_{t_j} + y^s_{t_{j+1}} - z^s_{t_{j+1}} + S^{(n_l,k_2-1)}_{t_{j+1}} xs^{s_l^+}_{t_{j+2}} - S^{(n_l,k_2-1)}_{t_{j+1}} xb^{s_l^+}_{t_{j+2}} \\
&\qquad + D^{(n_l,k_2-1)}_{t_{j+2}\to t_{j+3}} xh^{s_l^+}_{t_{j+2}} - P^{(n_l,k_2-1)}_{t_{j+2}\to t_{j+3}} y^{s_l^+}_{t_{j+2}} + \left(e^{-\rho\Delta\tau_{j+3}} P^{(n_l,k_2-1)}_{t_{j+2}\to t_{j+3}}\right) z^{s_l^+}_{t_{j+2}} \\
&\qquad = L^{(n_l,k_2-1)}_{t_{j+2}\to t_{j+3}} && (4.26)\\
&xh^{s^-}_{t_j} - xs^{s_l^+}_{t_{j+2}} + xb^{s_l^+}_{t_{j+2}} - xh^{s_l^+}_{t_{j+2}} = 0 && (4.27)\\
&z^{s_l^+}_{t_{j+2}} \le \bar{Z}^{(n_l,k_2-1)}_{t_{j+2}\to t_{j+3}} && (4.28)
\end{aligned}$$

where constraints (4.26)–(4.28) are present for each $l = 0, 1, \ldots, M$.

Next, we multiply the cash-balance constraint (4.24) of scenario s at time t_{j+1} by $P^{(n,k_0)}_{t_j\to t_{j+1}}$, and add it to the cash-balance constraint (4.21) of its predecessor s^- at time t_j. Using

$$\begin{aligned}
&P^{(n,k_0)}_{t_j\to t_{j+1}} P^{(n,k_1)}_{t_{j+1}\to t_{j+2}} = P^{(n,k_0)}_{t_j\to t_{j+2}} && \text{(see definition (3.18))}\\
&D^{(n,k_0)}_{t_j\to t_{j+1}} + P^{(n,k_0)}_{t_j\to t_{j+1}} \left(D^{(n,k_1)}_{t_{j+1}} + D^{(n,k_1)}_{t_{j+1}\to t_{j+2}}\right) = D^{(n,k_0)}_{t_j\to t_{j+2}} && \text{(see definition (3.19))}\\
&L^{(n,k_0)}_{t_j\to t_{j+1}} + P^{(n,k_C)}_{t_j\to t_{j+1}} L^{(n,k_1)}_{t_{j+1}\to t_{j+2}} = L^{(n,k_0)}_{t_j\to t_{j+2}} && \text{(see definition (3.27))}
\end{aligned}$$

the resulting cash-balance constraint of scenario s^- at time t_j is

$$\begin{aligned}
&D^{(n,k_0)}_{t_j} xh^{s^{--}}_{t_{j-1}} + y^{s^{--}}_{t_{j-1}} - z^{s^{--}}_{t_{j-1}} + S^{(n,k_0)}_{t_j} xs^{s^-}_{t_j} - S^{(n,k_0)}_{t_j} xb^{s^-}_{t_j} + D^{(n,k_0)}_{t_j\to t_{j+2}} xh^{s^-}_{t_j} \\
&\qquad - P^{(n,k_0)}_{t_j\to t_{j+2}} y^s_{t_{j+1}} + \left(\left(e^{-\rho\Delta\tau_{j+1}} - 1\right) P^{(n,k_0)}_{t_j\to t_{j+1}}\right) z^{s^-}_{t_j} \\
&\qquad + \left(e^{-\rho\Delta\tau_{j+2}} P^{(n,k_0)}_{t_j\to t_{j+2}}\right) z^s_{t_{j+1}} = L^{(n,k_0)}_{t_j\to t_{j+2}} && (4.29)
\end{aligned}$$

[3] In terms of row operations, this is equivalent to subtracting $S^{(n,k_1)}_{t_{j+1}}$ times $xs^s_{t_{j+1}} = 0$ from and adding $S^{(n,k_1)}_{t_{j+1}}$ times $xb^s_{t_{j+1}} = 0$ to the cash-balance constraint (4.15), and adding $xs^s_{t_{j+1}} = 0$ to and subtracting $xb^s_{t_{j+1}} = 0$ from the portfolio-balance constraint (4.16).

Finally, we add $\left(e^{-\rho\Delta\tau_{j+2}}P^{(n,k_1)}_{t_{j+1}\to t_{j+2}}\right)$ times the column of $z^{s^-}_{t_j}$ to the column of $z^{s}_{t_{j+1}}$, and use $z^{s^-}_{t_j}$ to denote the associated aggregated variable. If we also rename the variable $y^{s}_{t_{j+1}}$ to $y^{s^-}_{t_j}$, then the complete set of (aggregated) constraints is:

$$\begin{aligned}
&D^{(n,k_0)}_{t_j}\not{x}h^{s^{--}}_{t_{j-1}} + y^{s^{--}}_{t_{j-1}} - z^{s^{--}}_{t_{j-1}} + S^{(n,k_0)}_{t_j}\not{x}s^{s^-}_{t_j} - S^{(n,k_0)}_{t_j}\not{x}b^{s^-}_{t_j} \\
&\qquad + D^{(n,k_0)}_{t_j\to t_{j+2}}\not{x}h^{s^-}_{t_j} - P^{(n,k_0)}_{t_j\to t_{j+2}}y^{s^-}_{t_j} + \left(e^{-\rho\Delta(\tau_{j+1}+\tau_{j+2})}P^{(n,k_0)}_{t_j\to t_{j+2}}\right)z^{s^-}_{t_j} \\
&\qquad = L^{(n,k_0)}_{t_j\to t_{j+2}} && (4.30)\\
&\not{x}h^{s^{--}}_{t_{j-1}} - \not{x}s^{s^-}_{t_j} + \not{x}b^{s^-}_{t_j} - \not{x}h^{s^-}_{t_j} = 0 && (4.31)\\
&\left(e^{-\rho\Delta\tau_{j+2}}P^{(n,k_1)}_{t_{j+1}\to t_{j+2}}\right)z^{s^-}_{t_j} \le \bar{Z}^{(n,k_0)}_{t_j\to t_{j+1}} && (4.32)\\
&z^{s^-}_{t_j} \le \bar{Z}^{(n,k_1)}_{t_{j+1}\to t_{j+2}} && (4.33)\\
&D^{(n_l,k_2-1)}_{t_{j+2}}\not{x}h^{s^-}_{t_j} + y^{s^-}_{t_j} - z^{s^-}_{t_j} + S^{(n_l,k_2-1)}_{t_{j+1}}\not{x}s^{s_l^+}_{t_{j+2}} - S^{(n_l,k_2-1)}_{t_{j+1}}\not{x}b^{s_l^+}_{t_{j+2}} \\
&\qquad + D^{(n_l,k_2-1)}_{t_{j+2}\to t_{j+3}}\not{x}h^{s_l^+}_{t_{j+2}} - P^{(n_l,k_2-1)}_{t_{j+2}\to t_{j+3}}y^{s_l^+}_{t_{j+2}} + \left(e^{-\rho\Delta\tau_{j+3}}P^{(n_l,k_2-1)}_{t_{j+2}\to t_{j+3}}\right)z^{s_l^+}_{t_{j+2}} \\
&\qquad = L^{(n_l,k_2-1)}_{t_{j+2}\to t_{j+3}} && (4.34)\\
&\not{x}h^{s^-}_{t_j} - \not{x}s^{s_l^+}_{t_{j+2}} + \not{x}b^{s_l^+}_{t_{j+2}} - \not{x}h^{s_l^+}_{t_{j+2}} = 0 && (4.35)\\
&z^{s_l^+}_{t_{j+2}} \le \bar{Z}^{(n_l,k_2-1)}_{t_{j+2}\to t_{j+3}} && (4.36)
\end{aligned}$$

Except for the two upper bounds on $z^{s^-}_{t_j}$, this set of constraints corresponds exactly to the situation in figure 4-2(b). By noting that definition (3.28) implies

$$\bar{Z}^{(n,k_0)}_{t_j\to t_{j+2}} = \min\left\{\frac{\bar{Z}^{(n,k_0)}_{t_j\to t_{j+1}}}{e^{-\rho_i\Delta\tau_{j+2}}P^{(n,k_1)}_{t_{j+1}\to t_{j+2}}}, \bar{Z}^{(n,k_1)}_{t_{j+1}\to t_{j+2}}\right\}$$

we can replace the two upper bounds on $z^{s^-}_{t_j}$ by the single upper bound $z^{s^-}_{t_j} \le \bar{Z}^{(n,k_0)}_{t_j\to t_{j+2}}$.

We have assumed that $t_{j+2} < t_T$ and $t_j > t_0$. When $t_{j+2} = t_T$, the variable and constraint aggregations remain exactly the same as they don't involve the variables or constraints for scenarios at time t_{j+2}. When time aggregation is performed at time 0 ($t_j = t_0$), the aggregations change in the following manner. As there is no budget constraint at time 0 in the ALM model, we cannot add $P_{t_0\to t_1}$ times the cash-balance constraint (4.24) at time t_1 to that budget constraint, and instead we add it to the objective function. The objective function then becomes:

$$\begin{aligned}
v \;=\; & \left(S_{t_0} - D_{t_0\to t_2}\right)\not{x}h_{t_0} + P_{t_0\to t_2}y^s_{t_1} - \left(\left(e^{-\rho\Delta\tau_1}-1\right)P_{t_0\to t_1}\right)z_{t_0} \\
& + \left(e^{-\rho\Delta\tau_2}P_{t_0\to t_2}\right)z^s_{t_1} + L_{t_0\to t_2} - \lambda_1\sum_{s\in\mathcal{S}_{t_T}} q^s_{t_T}\left(y^s_{t_T} - \lambda_2 z^s_{t_T}\right)
\end{aligned}$$

The subsequent aggregation of the variables $z^s_{t_1}$ and z_{t_0} is then performed analogous to the aggregation of the variables $z^s_{t_{j+1}}$ and $z^{s^-}_{t_j}$ described earlier, while $y^s_{t_1}$ is renamed to y_{t_0}.

4.3 The Iterative Disaggregation Algorithm

We assume that a discrete-time model of the term-structure uncertainty is known that satisfies assumption 4 in section 3.2, and that state and time aggregations have been performed in the corresponding event tree to obtain a set of interest-rate scenarios that is sufficiently small to serve as description of the uncertainty in the ALM model.[4] We furthermore assume that this initial ALM model has been solved to optimality, for example by one of the special solution methods for stochastic programs that are discussed in the next chapter, or possibly as a straight linear program. This is the point at which the iterative disaggregation algorithm starts. An iteration of the algorithm consists of the following steps:

1. Perform a *disaggregation* (i.e. reverse one or more aggregations) in the aggregated event tree.
2. Find a feasible solution to the disaggregated ALM model, based on the optimal solution from the previous iteration.
3. Re-optimize the disaggregated ALM model.

We will discuss steps 1 and 2 of the algorithm in sections 4.3.2 and 4.3.1 below. A discussion of the re-optimization in step 3 is postponed to the next chapter, where we will present a new decomposition method for multistage stochastic linear programs. It will be shown that this decomposition method can take full advantage of a feasible solution that is constructed in step 2, and thus of the information that has been obtained from previous iterations.

Our analysis will partly be based on the dual of the aggregated ALM model (3.30), and we therefore state this dual formulation here. For scenario s at time t, let φ^s_t denote the dual variable that is associated with the cash-balance constraint in the ALM model, $\mu^s_t = (\mu^s_{1,t}, \ldots, \mu^s_{I,t})$ the vector of dual variables for the portfolio-balance constraints, and $-\xi^s_t$ the dual variable for the borrowing constraint. The index $i =$

[4]The state aggregation method allows us in principle to reduce the complete event tree to a single "expected value" scenario. In this expected-value scenario, the one-period return on all assets would equal the riskless one-period return.

$1, \ldots, I$ is used to refer to the individual assets.The dual of the aggregated ALM model is:

$$v = \max \sum_{j=1}^{T-1} \sum_{s \in \mathcal{S}_{t_j}} \left(\varphi^s_{t_j} L^{n(s)}_{t_j \to t_{j+1}} - \xi^s_{t_j} \bar{Z}^{n(s)}_{t_j \to t_{j+1}} \right) + \sum_{s \in \mathcal{S}_{t_T}} \left(\varphi^s_{t_T} L^{n(s)}_{t_T} - \xi^s_{t_T} \bar{Z}^{n(s)}_{t_T} \right) \quad (4.37)$$

subject to

$$\begin{aligned}
xh_{i,t_0} \quad &: \quad \sum_{s \in \mathcal{S}_{t_1}} \left(\varphi^s_{t_1} D^{n(s)}_{i,t_1} + \mu^s_{i,t_1} \right) \leq (1+c) S_{i,t_0} - D_{i,t_0 \to t_1} && \text{(i)} \\
y_{t_0} \quad &: \quad \sum_{s \in \mathcal{S}_{t_1}} \varphi^s_{t_1} \leq P_{t_0 \to t_1} && \text{(ii)} \\
z_{t_0} \quad &: - \sum_{s \in \mathcal{S}_{t_1}} \varphi^s_{t_1} - \xi_{t_0} \leq -e^{-\rho \Delta \tau_1} P_{t_0 \to t_1} && \text{(iii)} \\
xs^s_{i,t_j} \quad &: \quad \varphi^s_{t_j} (1-c) S^{n(s)}_{i,t_j} - \mu^s_{i,t_j} \leq 0 && \text{(iv)} \\
xb^s_{i,t_j} \quad &: - \varphi^s_{t_j} (1+c) S^{n(s)}_{i,t_j} + \mu^s_{i,t_j} \leq 0 && \text{(v)} \\
xh^s_{i,t_j} \quad &: \quad \varphi^s_{t_j} D^{n(s)}_{i,t_j \to t_{j+1}} - \mu^s_{i,t_j} + \sum_{s^+} \left(\varphi^{s^+}_{t_{j+1}} D^{n(s^+)}_{i,t_{j+1}} + \mu^{s^+}_{i,t_{j+1}} \right) \leq 0 && \text{(vi)} \\
xh^s_{i,t_{T-1}} \quad &: \quad \varphi^s_{t_{T-1}} D^{n(s)}_{i,t_{T-1} \to t_T} - \mu^s_{i,t_{T-1}} + \sum_{s^+} \varphi^{s^+}_{t_T} \left(D^{n(s^+)}_{i,t_T} + S^{n(s^+)}_{i,t_T} \right) \leq 0 && \text{(vii)} \\
y^s_{t_j} \quad &: - \varphi^s_{t_j} P^{n(s)}_{t_j \to t_{j+1}} + \sum_{s^+} \varphi^{s^+}_{t_{j+1}} \leq 0 && \text{(viii)} \\
z^s_{t_j} \quad &: \quad \varphi^s_{t_j} \left(e^{-\rho \Delta \tau_{j+1}} P^{n(s)}_{t_j \to t_{j+1}} \right) - \xi^s_{t_j} - \sum_{s^+} \varphi^{s^+}_{t_{j+1}} \leq 0 && \text{(ix)} \\
y^s_{t_T} \quad &: - \varphi^s_{t_T} \leq -\lambda_1 q^s_{t_T} && \text{(x)} \\
z^s_{t_T} \quad &: \quad \varphi^s_{t_T} - \xi^s_{t_T} \leq (\lambda_1 \lambda_2) q^s_{t_T} && \text{(xi)} \\
& \quad \xi^s_{t_j} \geq 0 \qquad \forall\, s \in \mathcal{S}_{t_j},\ j = 0, \ldots, T
\end{aligned}$$

The associated primal variables have been listed at the beginning of the constraints. We have not denoted the scope of each constraint, as this follows directly from the indices of the associated primal variables.

4.3.1 Constructing a Feasible Solution

It follows from the previous section that a state or time disaggregation introduces new variables as well as constraints in the ALM model, and the disaggregated ALM model is thus at the same time a relaxation and a restriction of the model before the

disaggregation. It is therefore not obvious how to construct a feasible solution to the disaggregated model from an optimal solution to the aggregated model. A natural starting point is the fixed-weight solution that is derived from this optimal solution, (see section 4.1). We will analyze below what infeasibilities the fixed-weight solution that corresponds to state and time disaggregation causes in the disaggregated ALM model. We will then show how the fixed-weight solutions can be modified so that they are feasible in a relaxation of the ALM model, and indicate how this relaxed version of the model can be used in the iterative disaggregation algorithm to solve the unrelaxed model.

Fixed-Weight State Disaggregation

Suppose that we know an optimal solution to the ALM model in which the underlying (aggregated) event tree includes the situation of figure 4-1(b). Let $\hat{x}_t^s \equiv (\hat{xs}_t^s, \hat{xb}_t^s, \hat{xh}_t^s, \hat{y}_t^s, \hat{z}_t^s)$ denote the optimal primal solution in this model for scenario s at time t, and $\hat{u}_t^s \equiv (\hat{\varphi}_t^s, \hat{\mu}_t^s, \hat{\xi}_t^s)$ the corresponding optimal dual solution. If we reverse the state aggregation of figure 4-1, then the fixed-weight solution for each scenario s^- in state (t_j, n, k_0), its successor s_l in state $(t_{j+1}, n_l, k_1 - 1)$, and its descendant s_l^+ in state $(t_{j+2}, n_l, k_2 - 1)$ is:

$$
\begin{array}{lll}
\text{Primal}: \tilde{x}_{t_j}^{s^-} = \hat{x}_{t_j}^{s^-} & & \\
\qquad \tilde{x}_{t_{j+1}}^{s_l} = \hat{x}_{t_{j+1}}^{s} & , l = 0, \ldots, M & \\
\qquad \tilde{x}_{t_{j+2}}^{s_l^+} = \hat{x}_{t_{j+2}}^{s_l^+} & , l = 0, \ldots, M & \\
\text{Dual} \quad : \tilde{u}_{t_j}^{s^-} = \hat{u}_{t_j}^{s^-} & & \\
\qquad \tilde{u}_{t_{j+1}}^{s_l} = \hat{\pi}_{t_{j+1}}^{n_l} \hat{u}_{t_{j+1}}^{s} & , l = 0, \ldots, M & \\
\qquad \tilde{u}_{t_{j+2}}^{s_l^+} = \hat{u}_{t_{j+2}}^{s_l^+} & , l = 0, \ldots, M &
\end{array}
\tag{4.38}
$$

One can verify by substitution that this fixed-weight solution will satisfy all constraints for the scenarios at times t_j and t_{j+2} in the disaggregated ALM model, as well as the portfolio-balance constraints for all scenarios at time t_{j+1}. However, it may violate the cash-balance and borrowing constraint for the scenarios at time t_{j+1}. By subtracting the cash-balance constraint in scenario s at time t_{j+1} before the disaggregation (with $\hat{x}$ as the optimal, and thus feasible, solution) from the cash-balance constraint in scenario s_l at time t_{j+1} (with $\hat{x}$ as proposed solution), we can see that this last cash-balance constraint is only satisfied by $\hat{x}$ if

$$
\left(D_{t_{j+1}}^{(n_l,k_1-1)} - D_{t_{j+1}}^{(n_l,k_1)}\right) \hat{xh}_{t_j}^{s^-} + (1-c)\left(S_{t_{j+1}}^{(n_l,k_1-1)} - S_{t_{j+1}}^{(n_l,k_1)}\right) \hat{xs}_{t_{j+1}}^{s}
$$

$$
\begin{aligned}
&-(1+c)\left(S_{t_{j+1}}^{(n_l,k_1-1)} - S_{t_{j+1}}^{(n_l,k_1)}\right)\hat{xb}_{t_{j+1}}^{s} + \left(D_{t_{j+1}\to t_{j+2}}^{(n_l,k_1-1)} - D_{t_{j+1}\to t_{j+2}}^{(n_l,k_1)}\right)\hat{xh}_{t_{j+1}}^{s} \\
&- \left(P_{t_{j+1}\to t_{j+2}}^{(n_l,k_1-1)} - P_{t_{j+1}\to t_{j+2}}^{(n_l,k_1)}\right)\hat{y}_{t_{j+1}}^{s} + e^{-\rho\Delta\tau_{j+2}}\left(P_{t_{j+1}\to t_{j+2}}^{(n_l,k_1-1)} - P_{t_{j+1}\to t_{j+2}}^{(n_l,k_1)}\right)\hat{z}_{t_{j+1}}^{s} \\
&= L_{t_{j+1}\to t_{j+2}}^{(n_l,k_1-1)} - L_{t_{j+1}\to t_{j+2}}^{(n_l,k_1)} \qquad (4.39)
\end{aligned}
$$

The borrowing constraint for scenario s_l at time t_{j+1} is only satisfied if

$$
\hat{z}_{t_{j+1}}^{s} \leq \bar{Z}_{t_{j+1}\to t_{j+2}}^{(n_l,k_1-1)} \qquad (4.40)
$$

Although it is very unlikely that these two constraints will be satisfied by all scenarios s_l, we note that they hold *on average*, that is, if we sum the constraints over all l and weigh constraint l with the conditional probability of the corresponding state. This implies that if the borrowing constraint is violated for some of the scenarios at t_{j+1}, then it must be satisfied by others. In the special case that the upper bound on short-term borrowing is state independent, (4.40) will be satisfied for all $l = 0, 1, \ldots, M$.

If we substitute the dual variables of the fixed-weight solution in the dual of the ALM model after the disaggregation, we find that the constraints that correspond to the primal variables for the scenarios s_l at time t_{j+1} may be violated[5], while the constraints that correspond to the primal variables for scenarios at time t_j and t_{j+2} will be satisfied. Furthermore, if the dual constraint that corresponds to $xs_{i,t_{j+1}}^{s_l}$ for some asset i and scenario s_l is violated, then the dual constraint with respect to $xb_{i,t_{j+1}}^{s_l}$ will be satisfied, and vice versa. Similarly, if the dual constraint that corresponds to $y_{t_{j+1}}^{s_l}$ ($z_{t_{j+1}}^{s_l}$) is violated for some scenario s_l, then the dual constraint with respect to $z_{t_{j+1}}^{s_l}$ ($y_{t_{j+1}}^{s_l}$) will be satisfied. Also, the fixed-weight solution can only violate the dual constraint that corresponds to $xs_{j,t_{j+1}}^{s_l}$ ($xb_{j,t_{j+1}}^{s_l}$) if $S_{t_{j+1}}^{(n_l,k_1-1)}$ is greater (smaller) than $S_{t_{j+1}}^{(n_l,k_1)}$.

Fixed-Weight Time Disaggregation

We consider reversing the time aggregation in figure 4-2. Let $\hat{x}_t^s \equiv (\hat{xs}_t^s, \hat{xb}_t^s, \hat{xh}_t^s, \hat{y}_t^s, \hat{z}_t^s)$ denote the optimal primal solution for scenario s at time t in the ALM model before the time disaggregation (figure 4-2(b)), and $\hat{u}_t^s \equiv (\hat{\varphi}_t^s, \hat{\mu}_t^s, \hat{\xi}_t^s)$ the associated optimal dual solution. The fixed-weight solution for the ALM model after the time disaggre-

[5] As the weighted average of these constraints gives precisely the set of constraints for scenario s at time t_{j+1} before the disaggregation with $\hat{u}_{t_{j+1}}^s$ as solution, the fixed-weight solution can again be said to satisfy the constraints on average.

gation (figure 4-2(a)) that is derived from this optimal solution is:

$$\begin{aligned}
\text{Primal}: \tilde{x}_{t_j}^{s^-} &= \hat{x}_{t_j}^{s^-} \text{ except } \tilde{y}_{t_j}^{s^-} = P_{t_j \to t_{j+1}}^{(n,k_0)} \hat{y}_{t_j}^{s^-} \qquad (4.41)\\
&\qquad \text{and } \tilde{z}_{t_j}^{s^-} = \left(e^{-\rho \Delta \tau_{j+1}} P_{t_j \to t_{j+1}}^{(n,k_0)}\right) \hat{z}_{t_j}^{s^-}\\
\tilde{xs}_{t_{j+1}}^{s} &= \tilde{xb}_{t_{j+1}}^{s} = 0\\
\tilde{xh}_{t_{j+1}}^{s} &= \hat{xh}_{t_j}^{s^-}\\
\tilde{y}_{t_{j+1}}^{s} &= \hat{y}_{t_j}^{s^-}\\
\tilde{z}_{t_{j+1}}^{s} &= \hat{z}_{t_j}^{s^-}\\
\tilde{x}_{t_{j+2}}^{s_l^+} &= \hat{x}_{t_{j+2}}^{s_l^+} \qquad , l = 0, \ldots, M\\
\text{Dual}: \tilde{\varphi}_{t_j}^{s^-} &= \hat{\varphi}_{t_j}^{s^-} \qquad (4.42)\\
\tilde{\mu}_{t_j}^{s^-} &= \hat{\mu}_{t_j}^{s^-}\\
\tilde{\xi}_{t_j}^{s^-} &= \begin{cases} \dfrac{\hat{\xi}_{t_j}^{s^-}}{e^{-\rho \Delta \tau_{j+2}} P_{t_{j+1} \to t_{j+2}}^{(n,k_1)}} & \text{if } \bar{Z}_{t_j \to t_{j+2}}^{(n,k_0)} = \dfrac{\bar{Z}_{t_j \to t_{j+1}}^{(n,k_0)}}{e^{-\rho \Delta \tau_{j+2}} P_{t_{j+1} \to t_{j+2}}^{(n,k_1)}} \\ 0 & \text{otherwise} \end{cases}\\
\tilde{\varphi}_{t_{j+1}}^{s} &= P_{t_j \to t_{j+1}}^{(n,k_0)} \hat{\varphi}_{t_j}^{s^-}\\
\tilde{\mu}_{t_{j+1}}^{s} &= P_{t_j \to t_{j+1}}^{(n,k_0)} \hat{\mu}_{t_j}^{s^-}\\
\tilde{\xi}_{t_{j+1}}^{s} &= \begin{cases} \hat{\xi}_{t_j}^{s^-} & \text{if } \bar{Z}_{t_j \to t_{j+2}}^{(n,k_0)} = \bar{Z}_{t_{j+1} \to t_{j+2}}^{(n,k_1)} \\ 0 & \text{otherwise} \end{cases}\\
\tilde{u}_{t_{j+2}}^{s_l^+} &= \hat{u}_{t_{j+2}}^{s_l^+} \qquad , l = 0, \ldots, M
\end{aligned}$$

It is straightforward to verify by substitution that the primal fixed-weight solution (4.41) satisfies all constraints in the disaggregated ALM model with the possible exception of the cash-balance constraint for scenario s^- at time t_j and scenario s at time t_{j+1}. The cash-balance constraint for scenario s^- at time t_j will only be satisfied if

$$\left(D_{t_j \to t_{j+1}}^{(n,k_0)} - D_{t_j \to t_{j+2}}^{(n,k_0)}\right) \hat{xh}_{t_j}^{s^-} = L_{t_j \to t_{j+1}}^{(n,k_0)} - L_{t_j \to t_{j+2}}^{(n,k_0)}$$

and the cash-balance constraint for scenario s at time t_{j+1} only if

$$\left(D_{t_{j+1}}^{(n,k_1)} + D_{t_{j+1} \to t_{j+2}}^{(n,k_1)}\right) \hat{xh}_{t_j}^{s^-} = L_{t_{j+1} \to t_{j+2}}^{(n,k_1)}$$

Notice that these equations can only be violated if prepayment of dividends and/or liabilities occurs in state (n, k_0) at time t_j before the disaggregation. Thus, if no dividends are paid on the assets and if no liabilities are due between time t_j and t_{j+2}, then the primal fixed-weight solution after the time disaggregation constitutes a feasible solution.

When substituting the dual fixed-weight solution (4.42) in the dual of the ALM model after the time disaggregation, one finds that the constraints that correspond to the primal variables $xh_{t_j}^{s^-}$, $y_{t_j}^{s^-}$, $z_{t_j}^{s^-}$ and $z_{t_{j+1}}^{s}$ may be violated.

Modifying the Fixed-Weight Solutions to Obtain Feasibility

We have seen that fixed-weight disaggregation may cause violations in the cash-balance constraints for some scenarios in the ALM model, and possibly in the borrowing constraints if state disaggregation is performed. If the cash-balance constraint for a scenario s at time t is violated after a fixed-weight disaggregation, an obvious way to make it feasible is by increasing the amount of short-term lending in case of a cash surplus, and the amount of short-term borrowing in case of a cash deficit. In the last case, however, this may lead to a violation of the borrowing constraint in the scenario. Furthermore, additional short-term lending or borrowing in scenario s at time t also increases the amounts of short-term lending, respectively borrowing, in its descendant scenarios if we correct for violations in the cash-balance constraints in these scenarios in the same manner. Thus, even if additional borrowing in scenario s at time t does not violate the borrowing constraint in that scenario, it may cause a violation of the borrowing constraint in one of its descendants.

It follows that the upper bounds on short-term borrowing may obstruct the construction of a feasible solution from the fixed-weight solution that only involves adjustments in the amounts of short-term borrowing and lending. We therefore propose to relax the ALM model by including for all $t_j = t_0, \ldots, t_{T-1}$ and each $s \in \mathcal{S}_{t_j}$ an additional variable for short-term borrowing, $v_{t_j}^s$, in the ALM model on which there is no upper bound, but for which the interest-rate differential with the short-term lending rate, ϱ, is greater than ρ. Furthermore, an additional variable $v_{t_T}^s$ is introduced for each $s \in \mathcal{S}_{t_T}$ to take care of a negative final portfolio value that exceeds $\bar{Z}_{t_T}^s$, and the weight λ_3 on this quantity in the objective function is chosen to exceed λ_2.

In proposition 4.3 below we will show that values for ρ and λ_2 can be chosen such that an optimal solution to the ALM model will not involve short-term borrowing or cause a negative final portfolio value in any scenario. If we therefore set ϱ and λ_3 equal to these values, we are guaranteed that an optimal solution to the relaxation of the ALM model will satisfy $v_t^s = 0$ for all t and s, and will thus be an optimal solution to the true model. Furthermore, as the relaxation and the true model have the same set of optimal solutions in this case, the objective value of the feasible solution in the relaxed ALM model provides an upper bound on the optimum objective value of the

true model.

We will need the upper bounds on the dual variables $\varphi^s_{t_j}$ from the following lemma to prove proposition 4.3.

Lemma 4.1 *If asset prices in the aggregated event tree are arbitrage-free, then any feasible solution in the dual* (4.37) *of the aggregate ALM problem* (3.30) *satisfies:*

$$\lambda_1 q^s_{t_j} \leq \varphi^s_{t_j} \leq \lambda_1 q^s_{t_j} + \beta^s_{t_j} \tag{4.43}$$

for all $s \in \mathcal{S}_{t_j}$ *and* $j = 1, \ldots, T$, *where*

$$\beta^s_{t_1} \equiv \min\left\{(1-\lambda_1)P_{t_0 \to t_1},\ \gamma^s_{t_1}\right\} \qquad \textit{with} \tag{4.44}$$

$$\gamma^s_{t_1} \equiv \min_{i \in \mathcal{I}^{n(s)}_{t_1}} \left\{ \frac{((1+c)S_{i,t_0} - D_{i,t_0 \to t_1}) - \lambda_1(1-c)\,(S_{i,t_0} - D_{i,t_0 \to t_1})}{(1-c)S^{n(s)}_{i,t_1} + D^{n(s)}_{i,t_1}} \right\}$$

and

$$\beta^s_{t_j} \equiv \min\left\{\beta^{s^-}_{t_{j-1}} P^{n(s^-)}_{t_{j-1} \to t_j},\ \gamma^s_{t_j}\right\} \qquad \textit{for } j = 2, \ldots, T, \qquad \textit{with} \tag{4.45}$$

$$\gamma^s_{t_j} \equiv \min_{i \in \mathcal{I}^{n(s)}_{t_j}} \left\{ \frac{\beta^{s^-}_{t_{j-1}} \left((1+c)S^{n(s^-)}_{i,t_{j-1}} - D^{n(s^-)}_{i,t_{j-1} \to t_j}\right) + \lambda_1 q^{s^-}_{t_{j-1}} c \left(2S^{n(s^-)}_{i,t_{j-1}} - D^{n(s^-)}_{i,t_{j-1} \to t_j}\right)}{(1-c)S^{n(s)}_{i,t_j} + D^{n(s)}_{i,t_j}} \right\}$$

while the subset of assets $\mathcal{I}^{n(s)}_{t_j}$ *is defined as* $\left\{i \in \{1, \ldots, I\} : S^{n(s)}_{i,t_j} + D^{n(s)}_{i,t_j} > 0\right\}$.

PROOF: We will first establish the lower bounds on $\varphi^s_{t_j}$. Constraint (x) in (4.37) states the lower bound for $j = T$. Consider a scenario s at time t_j ($j < T$), and suppose that the lower bound holds for each of its successors s^+ at time t_{j+1}, i.e., $\varphi^{s^+}_{t_{j+1}} \geq \lambda_1 q^{s^+}_{t_{j+1}}$. We will show that this implies $\varphi^s_{t_j} \geq \lambda_1 q^s_{t_j}$, and by induction it follows that $\varphi^s_{t_j}$ for all $j = 1, \ldots, T$ must satisfy the lower bound in (4.43).

Using the induction hypothesis, constraint (viii) in (4.37) implies

$$\varphi^s_{t_j} P^{n(s)}_{t_j \to t_{j+1}} \geq \sum_{s^+} \varphi^{s^+}_{t_{j+1}} \geq \lambda_1 \sum_{s^+} q^{s^+}_{t_{j+1}}$$

Because $q^{s^+}_{t_{j+1}} = q^s_{t_j} P^{n(s)}_{t_j \to t_{j+1}} \hat{\pi}^{n(s^+)}_{t_{j+1}}$ for all successor scenarios s^+ of s, we have

$$\varphi^s_{t_j} P^{n(s)}_{t_j \to t_{j+1}} \geq \lambda_1 q^s_{t_j} P^{n(s)}_{t_j \to t_{j+1}} \sum_{s^+} \hat{\pi}^{n(s^+)}_{t_{j+1}}$$

As $\sum_{s^+} \hat{\pi}^{n(s^+)}_{t_{j+1}} = 1$, the lower bound for $\varphi^s_{t_j}$ directly follows.

We will now prove the upper bound for $\varphi^s_{t_1}$. Constraint (ii) in (4.37) allows us to write

$$\varphi_{t_1}^s \leq P_{t_0 \to t_1} - \sum_{s' \in \mathcal{S}_{t_1}, s' \neq s} \varphi_{t_1}^{s'}$$

Using the previously established lower bounds for $\varphi_{t_1}^{s'}$, and noting that $q_{t_1}^s = P_{t_0 \to t_1} \hat{\pi}_{t_1}^{n(s)}$, we get

$$\begin{aligned}
\varphi_{t_1}^s &\leq P_{t_0 \to t_1} - \lambda_1 P_{t_0 \to t_1} \sum_{s' \in \mathcal{S}_{t_1}, s' \neq s} \hat{\pi}_{t_1}^{n(s')} \\
&= P_{t_0 \to t_1} - \lambda_1 P_{t_0 \to t_1} \left(1 - \hat{\pi}_{t_1}^{n(s)}\right) \\
&= \lambda_1 q_{t_1}^s + (1 - \lambda_1) P_{t_0 \to t_1}
\end{aligned}$$

This proves the first part of the upper bound for $\varphi_{t_1}^s$ in (4.43).

For the second part we consider constraint (i) in (4.37). Constraint (iv) specifies a lower bound on μ_{i,t_1}^s, and using this in constraint (i) enables us to write

$$\sum_{s \in \mathcal{S}_{t_1}} \varphi_{t_1}^s \left(D_{i,t_1}^{n(s)} + (1-c) S_{i,t_1}^{n(s)}\right) \leq (1+c) S_{i,t_0} - D_{i,t_0 \to t_1}$$

Using the same strategy as before, we can derive an upper bound on $\varphi_{t_1}^s$ for a specific scenario $s \in \mathcal{S}_{t_1}$ by rewriting this as:

$$\begin{aligned}
&\varphi_{t_1}^s \left(D_{i,t_1}^{n(s)} + (1-c) S_{i,t_1}^{n(s)}\right) \\
&\leq (1+c) S_{i,t_0} - D_{i,t_0 \to t_1} - \sum_{s' \in \mathcal{S}_{t_1}, s' \neq s} \varphi_{t_1}^{s'} \left(D_{i,t_1}^{n(s')} + (1-c) S_{i,t_1}^{n(s')}\right) \\
&\leq (1+c) S_{i,t_0} - D_{i,t_0 \to t_1} - \lambda_1 P_{t_0 \to t_1} \sum_{s' \in \mathcal{S}_{t_1}, s' \neq s} \hat{\pi}_{t_1}^{n(s')} \left(D_{i,t_1}^{n(s')} + (1-c) S_{i,t_1}^{n(s')}\right) \\
&\leq (1+c) S_{i,t_0} - D_{i,t_0 \to t_1} - \\
&\quad \lambda_1 P_{t_0 \to t_1} \left(\sum_{s' \in \mathcal{S}_{t_1}} \hat{\pi}_{t_1}^{n(s')} \left(D_{i,t_1}^{n(s')} + (1-c) S_{i,t_1}^{n(s')}\right) - \hat{\pi}_{t_1}^{n(s)} \left(D_{i,t_1}^{n(s)} + (1-c) S_{i,t_1}^{n(s)}\right) \right) \\
&\leq (1+c) S_{i,t_0} - D_{i,t_0 \to t_1} - \lambda_1 (1-c) \left(S_{i,t_0} - D_{i,t_0 \to t_1}\right) + \\
&\quad \lambda_1 P_{t_0 \to t_1} \hat{\pi}_{t_1}^{n(s)} \left(D_{i,t_1}^{n(s)} + (1-c) S_{i,t_1}^{n(s)}\right)
\end{aligned}$$

where the last inequality follows from the fact that asset prices in the aggregated event tree are arbitrage-free, i.e. (see proposition 3.2),

$$P_{t_0 \to t_1} \sum_{s' \in \mathcal{S}_{t_1}} \hat{\pi}_{t_1}^{n(s')} \left(D_{i,t_1}^{n(s')} + S_{i,t_1}^{n(s')}\right) = S_{i,t_0} - D_{i,t_0 \to t_1}$$

If $D_{i,t_1}^{n(s)} + S_{i,t_1}^{n(s)} = 0$, then the inequality above does not imply an upper bound on $\varphi_{t_1}^s$. Otherwise, we have (use again $q_{t_1}^s = P_{t_0 \to t_1} \hat{\pi}_{t_1}^{n(s)}$):

$$\varphi^s_{t_1} \leq \lambda_1 q^s_{t_1} + \left(\frac{((1+c)S_{i,t_0} - D_{i,t_0 \to t_1}) - \lambda_1 (1-c)\,(S_{i,t_0} - D_{i,t_0 \to t_1})}{(1-c)S^{n(s)}_{i,t_1} + D^{n(s)}_{i,t_1}} \right)$$

Because this inequality holds for all assets i with $D^{n(s)}_{i,t_1} + S^{n(s)}_{i,t_1} > 0$, we obtain the upper bound on $\varphi^s_{t_1}$ that is stated in the lemma.

The upper bound on $\varphi^s_{t_j}$ when $j > 1$ is derived in a similar fashion from constraints (viii) and (vi) in (4.37). Consider a scenario s at time t_j ($j \geq 1$) and suppose that $\varphi^s_{t_j}$ satisfies the upper bound in the lemma. We will show that this implies that $\varphi^{s^+}_{t_{j+1}}$ for all successors s^+ of s must satisfy the upper bound in the lemma as well. This proves the lemma by induction.

Let scenario s^* at time t_{j+1} be an arbitrary successor scenario of scenario s at time t_j. From constraint (viii) in (4.37)

$$\varphi^{s^*}_{t_{j+1}} \leq \varphi^s_{t_j} P^{n(s)}_{t_j \to t_{j+1}} - \sum_{s^+ \neq s^*} \varphi^{s^+}_{t_{j+1}}$$

The induction hypothesis and the lower bounds $\varphi^{s^+}_{t_{j+1}} \geq \lambda_1 q^{s^+}_{t_{j+1}} = \lambda_1 q^s_{t_j} P^{n(s)}_{t_j \to t_{j+1}} \hat{\pi}^{n(s^+)}_{t_{j+1}}$ allow us to write

$$\begin{aligned}
\varphi^{s^*}_{t_{j+1}} &\leq \left(\lambda_1 q^s_{t_j} + \beta^s_{t_j}\right) P^{n(s)}_{t_j \to t_{j+1}} - \lambda_1 q^s_{t_j} P^{n(s)}_{t_j \to t_{j+1}} \sum_{s^+ \neq s^*} \hat{\pi}^{n(s^+)}_{t_{j+1}} \\
&= \left(\lambda_1 q^s_{t_j} + \beta^s_{t_j}\right) P^{n(s)}_{t_j \to t_{j+1}} - \lambda_1 q^s_{t_j} P^{n(s)}_{t_j \to t_{j+1}} \left(1 - \hat{\pi}^{n(s^*)}_{t_{j+1}}\right) \\
&= \beta^s_{t_j} P^{n(s)}_{t_j \to t_{j+1}} + \lambda_1 q^{s^*}_{t_{j+1}}
\end{aligned}$$

This proves the first part of the upper bound for $\varphi^{s^*}_{t_{j+1}}$ in (4.45).

The second part follows from constraint (vi) in (4.37) if $j+1 < T$, and from constraint (vii) if $j+1 = T$. When $j+1 < T$, constraints (iv) and (v) enable us to derive the following inequality from constraint (vi):

$$-\varphi^s_{t_j}\left((1+c)S^{n(s)}_{i,t_j} - D^{n(s)}_{i,t_j \to t_{j+1}}\right) + \sum_{s^+} \varphi^{s^+}_{t_{j+1}}\left(D^{n(s^+)}_{i,t_{j+1}} + (1-c)S^{n(s^+)}_{i,t_j}\right) \leq 0$$

For an arbitrary successor s^* of scenario s we can rewrite this inequality as:

$$\begin{aligned}
&\varphi^{s^*}_{t_{j+1}}\left(D^{n(s^*)}_{i,t_{j+1}} + (1-c)S^{n(s^*)}_{i,t_j}\right) \\
&\leq \varphi^s_{t_j}\left((1+c)S^{n(s)}_{i,t_j} - D^{n(s)}_{i,t_j \to t_{j+1}}\right) - \sum_{s^+ \neq s^*} \varphi^{s^+}_{t_{j+1}}\left(D^{n(s^+)}_{i,t_{j+1}} + (1-c)S^{n(s^+)}_{i,t_j}\right) \\
&\leq \left(\lambda_1 q^s_{t_j} + \beta^s_{t_j}\right)\left((1+c)S^{n(s)}_{i,t_j} - D^{n(s)}_{i,t_j \to t_{j+1}}\right) \\
&\qquad -\lambda_1 q^s_{t_j} P^{n(s)}_{t_j \to t_{j+1}} \sum_{s^+ \neq s^*} \hat{\pi}^{n(s^+)}_{t_{j+1}}\left(D^{n(s^+)}_{i,t_{j+1}} + (1-c)S^{n(s^+)}_{i,t_j}\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \left(\lambda_1 q^s_{t_j} + \beta^s_{t_j}\right)\left((1+c)S^{n(s)}_{i,t_j} - D^{n(s)}_{i,t_j\to t_{j+1}}\right) - \lambda_1 q^s_{t_j}(1-c)\left(S^{n(s)}_{i,t_j} - D^{n(s)}_{i,t_j\to t_{j+1}}\right) \\
&\qquad +\lambda_1 q^s_{t_j} P^{n(s)}_{t_j\to t_{j+1}} \hat{\pi}^{n(s^*)}_{t_{j+1}}\left(D^{n(s^*)}_{i,t_{j+1}} + (1-c)S^{n(s^*)}_{i,t_j}\right) \\
&= \left(\lambda_1 q^s_{t_j} + \beta^s_{t_j}\right)\left((1+c)S^{n(s)}_{i,t_j} - D^{n(s)}_{i,t_j\to t_{j+1}}\right) - \lambda_1 q^s_{t_j}(1-c)\left(S^{n(s)}_{i,t_j} - D^{n(s)}_{i,t_j\to t_{j+1}}\right) \\
&\qquad +\lambda_1 q^{s^*}_{t_{j+1}}\left(D^{n(s^*)}_{i,t_{j+1}} + (1-c)S^{n(s^*)}_{i,t_j}\right)
\end{aligned}
$$

where we have used the lower bounds on $\varphi^{s^+}_{t_{j+1}}$ and the induction hypothesis in the second inequality, and proposition 3.2 in the third inequality. As such an inequality holds for all assets i, and only implies an upper bound for $\varphi^{s^*}_{t_{j+1}}$ when $S^{n(s^*)}_{i,t_j} + D^{n(s^*)}_{i,t_{j+1}} > 0$, the upper bound (4.45) follows.

When $j+1 = T$, we use constraint (vii) in (4.37) instead of constraint (vi) to find the upper bound on $\varphi^s_{t_T}$. The derivation is analagous, and therefore omitted here.

$\mathcal{QED}$

It is instructive to consider the special case of zero transaction costs ($c = 0$). The quantities $\beta^s_{t_j}$ then reduce to:

$$
\begin{aligned}
\beta^s_{t_1} &\equiv (1-\lambda_1)\cdot \min\left\{P_{t_0\to t_1}, \min_{i\in\mathcal{I}^{n(s)}_{t_1}}\left\{\frac{S_{i,t_0} - D_{i,t_0\to t_1}}{S^{n(s)}_{i,t_1} + D^{n(s)}_{i,t_1}}\right\}\right\} \\
\beta^s_{t_j} &\equiv \beta^{s^-}_{t_{j-1}}\cdot \min\left\{P^{n(s^-)}_{t_{j-1}\to t_j}, \min_{i\in\mathcal{I}^{n(s)}_{t_j}}\left\{\frac{S^{n(s^-)}_{i,t_{j-1}} - D^{n(s^-)}_{i,t_{j-1}\to t_j}}{S^{n(s)}_{i,t_j} + D^{n(s)}_{i,t_j}}\right\}\right\} \qquad ,j = 2,\ldots,T.
\end{aligned}
$$

This shows clearly that the upper bound on $\varphi^s_{t_j}$ is a function of the realized asset returns in state $n(s)$ at time t_j. The higher these realized returns are, the lower the upper bound is. As $\varphi^s_{t_j}$ is the dual variable on the cash-balance constraint, this agrees with the economic intuition that an extra dollar in a state with high asset returns is worth less than an extra dollar in a state with low asset returns.

If the parameter λ_1 is one, then $\beta^s_{t_1} = 0$, and thus $\beta^s_{t_j} = 0$ for all $j = 1,\ldots,T$. This implies that the upper bound on $\varphi^s_{t_j}$ coincides with its lower bound in lemma 4.1, and thus $\varphi^s_{t_j} = q^s_{t_j}$ as only possible solution. This result was also found in lemma 2.1 in section 2.3.2.

The bounds of lemma 4.1 are used in the following proposition to derive lower bounds on the interest-rate differential ρ and the parameter λ_2 so that an optimal solution to the aggregated ALM model will not involve short-term borrowing if ρ exceeds this lower bound, and exclude negative final portfolio values if λ_2 is greater than its lower bound.

Proposition 4.3 *If asset prices in the aggregated event tree are arbitrage-free, if* $\lambda_1 > 0$, *and if the aggregated ALM model* (3.30) *is feasible, then an optimal solution satisfies:*

$$z_0 = 0 \qquad \text{if} \quad \rho > \frac{1}{\Delta\tau_1} \ln\left[\frac{1}{\lambda_1}\right] \tag{4.46}$$

$$z^s_{t_j} = 0 \qquad \text{if} \quad \rho > \frac{1}{\Delta\tau_{j+1}} \ln\left[1 + \frac{\beta^s_{t_j}}{\lambda_1 q^s_{t_j}}\right] \qquad , s \in \mathcal{S}_{t_j},\ j = 1, \ldots, T-1 \tag{4.47}$$

$$z^s_{t_T} = 0 \qquad \text{if} \quad \lambda_2 > 1 + \frac{\beta^s_{t_T}}{\lambda_1 q^s_{t_T}} \qquad , s \in \mathcal{S}_{t_T} \tag{4.48}$$

where $\beta^s_{t_j}$ *is defined in lemma 4.1.*

PROOF: Without loss of generality, we consider the aggregated ALM model (3.30) without short-term borrowing constraints. This implies that the variables $\xi^s_{t_j}$ disappear from the dual problem (4.37).

Constraint (ii) in (4.37) states that any feasible solution to this dual problem must satisfy

$$e^{\rho\Delta\tau_1} \geq \frac{P_{t_0 \to t_1}}{\sum_{s \in \mathcal{S}_{t_1}} \varphi^s_{t_1}}$$

Using the lower bounds from lemma 4.1, we can bound the fraction on the right by

$$\frac{P_{t_0 \to t_1}}{\sum_{s \in \mathcal{S}_{t_1}} \varphi^s_{t_1}} \leq \frac{P_{t_0 \to t_1}}{\sum_{s \in \mathcal{S}_{t_1}} \lambda_1 q^s_{t_1}} = \frac{P_{t_0 \to t_1}}{\lambda_1 P_{t_0 \to t_1}} = \frac{1}{\lambda_1}$$

where we have used $q^s_{t_1} = P_{t_0 \to t_1} \hat{\pi}^{n(s)}_{t_1}$. It follows that if ρ is such that $e^{\rho\Delta\tau_1} > 1/\lambda_1$, then any feasible solution to (4.37) will satisfy constraint (ii) with strict inequality, and by complementary slackness, any optimal solution to (3.30) must have $z_0 = 0$. This proves relation (4.46).

To prove relation (4.47), note that constraint (ix) in (4.37) implies that any feasible solution to the dual ALM problem must satisfy

$$e^{\rho\Delta\tau_{j+1}} \geq \frac{P^{n(\varepsilon)}_{t_j \to t_{j+1}} \varphi^s_{t_j}}{\sum_{s^+} \varphi^{s^+}_{t_{j+1}}}$$

Using lemma 4.1, we can derive an upper bound for the quantity on the right:

$$\frac{\varphi^s_{t_j} P^{n(s)}_{t_j \to t_{j+1}}}{\sum_{s^+} \varphi^{s^+}_{t_{j+1}}} \leq \frac{P^{n(s)}_{t_j \to t_{j+1}} \left(\lambda_1 q^s_{t_j} + \beta^s_{t_j}\right)}{\sum_{s^+} \varphi^{s^+}_{t_{j+1}}} \leq \frac{P^{n(s)}_{t_j \to t_{j+1}} \left(\lambda_1 q^s_{t_j} + \beta^s_{t_j}\right)}{\lambda_1 q^s_{t_j} P^{n(s)}_{t_j \to t_{j+1}} \sum_{s^+} \hat{\pi}^{n(s^+)}_{t_{j+1}}} = 1 + \frac{\beta^s_{t_j}}{\lambda_1 q^s_{t_j}}$$

If we choose ρ such that $e^{\rho\Delta\tau_{j+1}}$ is greater than this upper bound, then constraint (ix) will be satisfied with strict inequality for any feasible solution to (4.37), and complementary slackness implies that $z^s_{t_j} = 0$ in every optimal solution to (3.30).

The condition on λ_2 in (4.48) for $z^s_{t_T} = 0$ follows directly from constraint (xi) in (4.37), the upper bound on $\varphi^s_{t_T}$ from lemma 4.1, and complementary slackness.

$\mathcal{QED}$

We noted before that $\beta^s_{t_j} = 0$ for all s and $j \geq 1$ if $\lambda_1 = 1$. In that case any $\rho > 0$ will prevent short-term borrowing in an optimal solution, and any $\lambda_2 > 1$ a negative final portfolio value.

4.3.2 Choosing a Disaggregation

In step 1 of the iterative disaggregation algorithm one has to choose a state in an aggregated event tree in which to perform a state or time disaggregation. We will discuss two ways of making this choice. The first method determines bounds on the possible change in the optimum objective value of the ALM model, and is based directly on proposition 4.2. The second method uses sensitivity analysis to estimate the sensitivity of the objective value to a disaggregation.

Bounds on the Change in Optimum Objective Value

In the previous section we have shown how to construct a feasible solution to a relaxation of the ALM model from the fixed-weight solution after a state or time disaggregation. The objective value of this feasible solution in the relaxed model forms an upper bound on the optimum objective value of the true model if the parameters for the additional borrowing variables in the relaxed model satisfy the bounds in proposition 4.3.

It should be noted that the arbitrage-free value of all liabilities always forms a lower bound on the objective value of the ALM model, aggregated or not. This follows directly from the definition of the liabilities in an aggregated event tree and the fact that the asset prices in the ALM model are arbitrage-free.

We will show here how proposition 4.2 can be used to derive bounds on the change in optimum objective value of the ALM model after one state or time disaggregation is performed in the underlying event tree. These bounds may be tighter than the bounds just mentioned.

Remember that the bounds of proposition 4.2 are based on the size of the constraint violations of the fixed-weight solution in the primal and dual problem, together with (generalized) upper bounds on the dual and primal variables. For basic state and time disaggregation in the ALM model, we analyzed the constraint violations of the fixed-weight solutions in section 4.3.1. We will show in proposition 4.4 below how upper bounds on the primal variables can be found if an upper bound on the initial investment is known, and state upper bounds on the relevant dual variables in proposition 4.5.

To choose a disaggregation in the ALM model, we propose to calculate the bounds on the objective value change for each single state and time disaggregation in the underlying event tree. The state with the largest difference between the upper and lower bound ($(\epsilon^+ + \epsilon^-)$ in proposition 4.2) is chosen for the actual disaggregation.

Proposition 4.2 can also be used to calculate bounds if one wants to perform several state and/or time disaggregations before re-optimizing the ALM model. However, the sheer number of possible combinations makes the calculation of the bounds a daunting task in that case. Instead, one can use the bounds that apply to individual state and time disaggregations as a guide for choosing multiple disaggregations.

The next proposition states bounds on the primal variables that can be used in proposition 4.2 to calculate a lower bound on the objective value of the ALM model after a disaggregation.

Proposition 4.4 *If there is a maximum $\bar{W}_0$ on the initial investment, and if short-term borrowing is limited ($z^s_{t_j} \leq \bar{Z}^{n(s)}_{t_j} < \infty$) at all times, then an optimal solution in the aggregated ALM model* (3.30) *satisfies:*

$$(S_{t_0} - D_{t_0 \to t_1})\, xh_{t_0} + P_{t_0 \to t_1} y_{t_0} \leq \bar{W}_0 + \left(e^{-\rho \Delta \tau_1} P_{t_0 \to t_1}\right) \bar{Z}_{t_0} \quad , \tag{4.49}$$

for all $s \in \mathcal{S}_{t_j}$ and $j = 1, \ldots, T-1$:

$$\begin{aligned}
\left(S^{n(s)}_{t_j} - D^{n(s)}_{t_j \to t_{j+1}}\right) xh^s_{t_j} + P^{n(s)}_{t_j \to t_{j+1}} y^s_{t_j} &\leq \bar{W}^s_{t_j} + \left(e^{-\rho \Delta \tau_{j+1}} P^{n(s)}_{t_j \to t_{j+1}}\right) \bar{Z}^{n(s)}_{t_j} \\
\left(S^{n(s)}_{t_j} - D^{n(s)}_{t_j \to t_{j+1}}\right) xb^s_{t_j} &\leq \bar{W}^s_{t_j} + \left(e^{-\rho \Delta \tau_{j+1}} P^{n(s)}_{t_j \to t_{j+1}}\right) \bar{Z}^{n(s)}_{t_j} \\
\left(S^{n(s^-)}_{t_{j-1}} - D^{n(s^-)}_{t_{j-1} \to t_j}\right) xs^s_{t_j} &\leq \bar{W}^{s^-}_{t_{j-1}} + \left(e^{-\rho \Delta \tau_j} P^{n(s^-)}_{t_{j-1} \to t_j}\right) \bar{Z}^{n(s^-)}_{t_{j-1}}
\end{aligned} \tag{4.50}$$

and for all $s \in \mathcal{S}_{t_T}$:

$$y^s_{t_T} \leq \bar{W}^s_{t_T} \tag{4.51}$$

where for each $s \in \mathcal{S}_{t_j}$ *and* $j = 1, \ldots, T$*:*

$$\bar{W}^s_{t_j} \equiv \bar{W}^{s^-}_{t_{j-1}} \cdot \max\left\{\frac{1}{P^{n(s^-)}_{t_{j-1}\to t_j}}, \max_{i=1,\ldots,I}\left\{\frac{S^{n(s)}_{i,t_j} + D^{n(s)}_{i,t_j}}{S^{n(s^-)}_{i,t_{j-1}} - D^{n(s^-)}_{i,t_{j-1}\to t_j}}\right\}\right\} \tag{4.52}$$
$$+ \bar{Z}^{n(s^-)}_{t_{j-1}} \cdot \max\left\{0, \max_{i=1,\ldots,I}\left\{\frac{S^{n(s)}_{i,t_j} + D^{n(s)}_{i,t_j}}{S^{n(s^-)}_{i,t_{j-1}} - D^{n(s^-)}_{i,t_{j-1}\to t_j}}\right\} e^{-\rho\Delta\tau_j} P^{n(s^-)}_{t_{j-1}\to t_j} - 1\right\} - L^{n(s)}_{t_j}$$

PROOF: We derive the bounds by setting the transaction cost rate c equal to zero. Bounds on the portfolio holdings that apply in the absence of transaction costs are certainly valid when transaction costs are positive.

Inequality (4.49) follows immediately from the assumption that the initial portfolio investment is restricted by $\bar{W}_0$ and $z_0 \le \bar{Z}_0$.

When $c = 0$, the cash-balance constraint for scenario s at time t_j can be written as:

$$\left(D^{n(s)}_{t_j} + S^{n(s)}_{t_j}\right) \not{x}h^{s^-}_{t_{j-1}} + y^{s^-}_{t_{j-1}} - z^{s^-}_{t_{j-1}} = \tag{4.53}$$
$$\left(S^{n(s)}_{t_j} - D^{n(s)}_{t_j\to t_{j+1}}\right) \not{x}h^s_{t_j} + P^{n(s)}_{t_j\to t_{j+1}} y^s_{t_j} - \left(e^{-\rho\Delta\tau_{j+1}} P^{n(s)}_{t_j\to t_{j+1}}\right) z^s_{t_j} + L^{n(s)}_{t_j}$$

The left-hand side of this equation equals the value in scenario s at time t_j of the investment portfolio that has been constructed in its predecessor s^- at time t_{j-1}. We will show how to determine an upper bound on this value, given a maximum on the investment at time t_{j-1}. This upper bound in turn implies an upper bound on the values of the variables $\not{x}h^s_{t_j}$ and $y^s_{t_j}$ in the right-hand side of equation (4.53).

Suppose $\bar{W}^{s^-}_{t_{j-1}}$ is an upper bound on the total investment in scenario s^- at time t_{j-1}, i.e.,

$$\left(S^{n(s^-)}_{t_{j-1}} - D^{n(s^-)}_{t_{j-1}\to t_j}\right) \not{x}h^{s^-}_{t_{j-1}} + P^{n(s^-)}_{t_{j-1}\to t_j} y^{s^-}_{t_{j-1}} - \left(e^{-\rho\Delta\tau_j} P^{n(s^-)}_{t_{j-1}\to t_j}\right) z^{s^-}_{t_{j-1}} \le \bar{W}^{s^-}_{t_{j-1}}$$

while $z^{s^-}_{t_{j-1}} \le \bar{Z}^{n(s^-)}_{t_{j-1}}$. Furthermore, let i^* be the index of the asset with the highest return between states $n(s^-)$ and $n(s)$, i.e.,

$$R^{n(s)}_{i^*,t_j} \equiv \frac{S^{n(s)}_{i^*,t_j} + D^{n(s)}_{i^*,t_j}}{S^{n(s^-)}_{i^*,t_{j-1}} - D^{n(s^-)}_{i^*,t_{j-1}\to t_j}} = \max_{i=1,\ldots,I}\left\{\frac{S^{n(s)}_{i,t_j} + D^{n(s)}_{i,t_j}}{S^{n(s^-)}_{i,t_{j-1}} - D^{n(s^-)}_{i,t_{j-1}\to t_j}}\right\}.$$

To determine the highest possible portfolio value in scenario s at time t_j from investments at time t_{j-1}, we consider three possible situations. First assume $R^{n(s)}_{i^*,t_j} > \left(e^{\rho\Delta\tau_j}/P^{n(s^-)}_{t_{j-1}\to t_j}\right)$. In that case, the maximum portfolio value would be obtained if

money is borrowed up to the limit at time t_{j-1}, and everything is invested in asset i^*. Thus in this situation

$$\left(D_{t_j}^{n(s)} + S_{t_j}^{n(s)}\right) x h_{t_{j-1}}^{s^-} + y_{t_{j-1}}^{s^-} - z_{t_{j-1}}^{s^-} \leq \left(\bar{W}_{t_{j-1}}^{s^-} + \left(e^{-\rho\Delta\tau_j} P_{t_{j-1}\to t_j}^{n(s^-)}\right) \bar{Z}_{t_{j-1}}^{n(s^-)}\right) R_{i^*,t_j}^{n(s)} - \bar{Z}_{t_{j-1}}^{n(s^-)} \tag{4.54}$$

The second possibility is $\left(e^{\rho\Delta\tau_j}/P_{t_{j-1}\to t_j}^{n(s^-)}\right) \geq R_{i^*,t_j}^{n(s)} > \left(1/P_{t_{j-1}\to t_j}^{n(s^-)}\right)$. In this case, the highest portfolio value in scenario s results if no money is borrowed at time t_{j-1}, and all available funds are invested in asset i^*, implying:

$$\left(D_{t_j}^{n(s)} + S_{t_j}^{n(s)}\right) x h_{t_{j-1}}^{s^-} + y_{t_{j-1}}^{s^-} - z_{t_{j-1}}^{s^-} \leq \bar{W}_{t_{j-1}}^{s^-} R_{i^*,t_j}^{n(s)} \tag{4.55}$$

Alternatively, $\left(1/P_{t_{j-1}\to t_j}^{n(s^-)}\right) \geq R_{i^*,t_j}^{n(s)}$. The most profitable strategy would now be to invest all available funds in the riskless asset at time t_{j-1}, and not borrow money. In this situation therefore

$$\left(D_{t_j}^{n(s)} + S_{t_j}^{n(s)}\right) x h_{t_{j-1}}^{s^-} + y_{t_{j-1}}^{s^-} - z_{t_{j-1}}^{s^-} \leq \bar{W}_{t_{j-1}}^{s^-} \left(\frac{1}{P_{t_{j-1}\to t_j}^{n(s^-)}}\right) \tag{4.56}$$

Combining the three possible situations concerning the structure of asset returns between state $n(s^-)$ at time t_{j-1} and state $n(s)$ at time t_j, we obtain from (4.54), (4.55) and (4.56)

$$\begin{aligned}
&\left(D_{t_j}^{n(s)} + S_{t_j}^{n(s)}\right) x h_{t_{j-1}}^{s^-} + y_{t_{j-1}}^{s^-} - z_{t_{j-1}}^{s^-} \\
&\leq \max\left\{\left(\bar{W}_{t_{j-1}}^{s^-} + e^{-\rho\Delta\tau_j} P_{t_{j-1}\to t_j}^{n(s^-)} \bar{Z}_{t_{j-1}}^{n(s^-)}\right) R_{i^*,t_j}^{n(s)} - \bar{Z}_{t_{j-1}}^{n(s^-)}, \bar{W}_{t_{j-1}}^{s^-} R_{i^*,t_j}^{n(s)}, \frac{\bar{W}_{t_{j-1}}^{s^-}}{P_{t_{j-1}\to t_j}^{n(s^-)}}\right\} \\
&= \bar{W}_{t_{j-1}}^{s^-} \max\left\{R_{i^*,t_j}^{n(s)}, \frac{1}{P_{t_{j-1}\to t_j}^{n(s^-)}}\right\} + \bar{Z}_{t_{j-1}}^{n(s^-)} \left[\left(e^{-\rho\Delta\tau_j} P_{t_{j-1}\to t_j}^{n(s^-)}\right) R_{i^*,t_j}^{n(s)} - 1\right]^+
\end{aligned}$$

where $[x]^+$ denotes the positive part of x. Using this inequality and $z_{t_j}^s \leq \bar{Z}_{t_j}^s$ in the cash-balance constraint (4.53), we get

$$\begin{aligned}
\left(S_{t_j}^{n(s)} - D_{t_j\to t_{j+1}}^{n(s)}\right) x h_{t_j}^s + P_{t_j\to t_{j+1}}^{n(s)} y_{t_j}^s \leq{}& \left(e^{-\rho\Delta\tau_{j+1}} P_{t_j\to t_{j+1}}^{n(s)}\right) \bar{Z}_{t_j}^s - L_{t_j}^{n(s)} + \\
& \bar{W}_{t_{j-1}}^{s^-} \max\left\{R_{i^*,t_j}^{n(s)}, \frac{1}{P_{t_{j-1}\to t_j}^{n(s^-)}}\right\} + \bar{Z}_{t_{j-1}}^{n(s^-)} \left[\left(e^{-\rho\Delta\tau_j} P_{t_{j-1}\to t_j}^{n(s^-)}\right) R_{i^*,t_j}^{n(s)} - 1\right]^+
\end{aligned}$$

By substitution of definition (4.52) we obtain the relation in (4.50), and by induction this relation must hold for all $j = 1, \ldots, T-1$.

For $j = T$ and $s \in \mathcal{S}_{t_T}$, the cash-balance constraint is

$$\left(D^{n(s)}_{t_T} + S^{n(s)}_{t_T}\right) \mathit{xh}^{s^-}_{t_{T-1}} + y^{s^-}_{t_{T-1}} - z^{s^-}_{t_{T-1}} = y^{s}_{t_T} - z^{s}_{t_T} + L^{n(s)}_{t_T}$$

and following an analogous derivation as above, we obtain

$$y^s_{t_T} - z^s_{t_T} \leq \bar{W}^s_{t_T}$$

Because $\lambda_2 \geq 1$, $y^s_{t_T}$ and $z^s_{t_T}$ will not simultaneously be greater than zero in an optimal solution to (3.30), which establishes the upper bound (4.51).

To obtain the relation in (4.50) that involves the vector $\mathit{xb}^s_{t_j}$, we note that for all assets i, $\mathit{xb}^s_{i,t_j} \leq \mathit{xh}^s_{i,t_j}$, and thus $\left(S^{n(s)}_{i,t_j} - D^{n(s)}_{i,t_j \to t_{j+1}}\right) \mathit{xb}^s_{i,t_j} \leq \left(S^{n(s)}_{i,t_j} - D^{n(s)}_{i,t_j \to t_{j+1}}\right) \mathit{xh}^s_{i,t_j}$. Summing over all i gives:

$$\begin{aligned}
\left(S^{n(s)}_{t_j} - D^{n(s)}_{t_j \to t_{j+1}}\right) \mathit{xb}^s_{t_j} &\leq \left(S^{n(s)}_{t_j} - D^{n(s)}_{t_j \to t_{j+1}}\right) \mathit{xh}^s_{t_j} \\
&\leq \left(S^{n(s)}_{t_j} - D^{n(s)}_{t_j \to t_{j+1}}\right) \mathit{xh}^s_{t_j} + P^{n(s)}_{t_j \to t_{j+1}} y^s_{t_j} \\
&\leq \bar{W}^s_{t_j} + \left(e^{-\rho \Delta \tau_{j+1}} P^{n(s)}_{t_j \to t_{j+1}}\right) \bar{Z}^{n(s)}_{t_j}
\end{aligned}$$

where we have used the previous result for the last inequality.

The inequality in (4.50) that involves the vector $\mathit{xs}^s_{t_j}$ can be derived in an analogous manner by noting that for all assets i, $\mathit{xs}^s_{i,t_j} \leq \mathit{xh}^{s^-}_{i,t_{j-1}}$.

$\mathcal{QED}$

To calculate the upper bound on the objective value according to proposition 4.2, restrictions on the size of the dual variables are needed.[6] In section 4.3.1 we have seen that the fixed-weight solutions after a state or time disaggregation can only violate cash-balance and borrowing constraints, and it is therefore sufficient to state bounds on the dual variables that correspond to these constraints ($\varphi^s_{t_j}$ and $\xi^s_{t_j}$). In lemma 4.1 bounds were stated on the variables $\varphi^s_{t_j}$, and the next proposition uses these bounds to derive upper bounds on the variables $\xi^s_{t_j}$.

[6] These dual variable bounds are used in conjunction with constraint violations in the primal problem when the fixed-weight solution is implemented. In section 4.1, the constraints in the primal problem were all stated as greater-than-or-equal-to constraints, whereas the cash-balance and portfolio-balance constraints in the ALM model have been formulated as equality constraints. However, it is easy to see that they could have been written as greater-than-or-equal-to constraints instead of equalities without changing the set of optimal solutions to the ALM model. It follows that the upper bound of proposition 4.2 when applied to the ALM model will only incorporate constraint violations in the cash-balance and portfolio-balance constraints with respect to this inequality formulation.

Proposition 4.5 *In any optimal solution to the dual* (4.37) *of the aggregated ALM model:*

$$\xi_{t_0} \leq P_{t_0 \to t_1} \left[\left(e^{-\rho \Delta \tau_1} - \lambda_1 \right) \right]^+ \tag{4.57}$$

$$\xi^s_{t_j} \leq P^{n(s)}_{t_j \to t_{j+1}} \left[e^{-\rho \Delta \tau_{j+1}} \left(\lambda_1 q^s_{t_j} + \beta^s_{t_j} \right) - \lambda_1 q^s_{t_j} \right]^+ \quad , s \in \mathcal{S}_{t_j}, \ \ j = 1, \ldots, T-1 \tag{4.58}$$

$$\xi^s_{t_T} \leq \left[\beta^s_{t_T} - (\lambda_2 - 1) \lambda_1 q^s_{t_j} \right]^+ \quad , s \in \mathcal{S}_{t_T} \tag{4.59}$$

where $[x]^+$ *denotes the positive part of* x.

PROOF: To prove (4.57), we first note that ξ_{t_0} only appears in constraint (iii) of the dual problem (4.37). Furthermore, an increase in ξ_{t_0} leads to a deterioration of the objective function, and ξ_{t_0} will therefore be chosen as small as possible in an optimal solution. Thus from constraint (iii) and the nonnegativity constraint:

$$\xi_{t_0} = \max \left\{ 0, e^{-\rho \Delta \tau_1} P_{t_0 \to t_1} - \sum_{s \in \mathcal{S}_{t_1}} \varphi^s_{t_1} \right\}$$

Using lemma 4.1 and the fact that $q^s_{t_1} = P_{t_0 \to t_1} \hat{\pi}^{n(s)}_{t_1}$, we can bound the quantity on the right by

$$e^{-\rho \Delta \tau_1} P_{t_0 \to t_1} - \sum_{s \in \mathcal{S}_{t_1}} \varphi^s_{t_1} \leq e^{-\rho \Delta \tau_1} P_{t_0 \to t_1} - \lambda_1 \sum_{s \in \mathcal{S}_{t_1}} q^s_{t_1} = P_{t_0 \to t_1} \left(e^{-\rho \Delta \tau_1} - \lambda_1 \right)$$

from which the upper bound (4.57) follows.

The bounds on $\xi^s_{t_j}$ and $\xi^s_{t_T}$ can be proved in an analogous fashion from the bounds of lemma 4.1 and constraints (ix), respectively (xi), in (4.37).

QED

The quality of the bounds on the change in optimal objective value after a disaggregation depends to an important extent on the quality of the upper bounds on the variables. The bounds on the dual variables $\varphi^s_{t_j}$ in lemma 4.1 are calculated by forward recursion, and could therefore be fairly weak when t_j is large. In the special case that $\xi^s_{t_j} = 0$ for all $s \in \mathcal{S}_{t_j}$ and $j = 0, \ldots, T$ (i.e., there are no borrowing constraints, or they are not binding in an optimal solution), we can use constraints (xi) and (ix) in the dual (4.37) of the aggregated ALM problem to obtain the following upper bounds on $\varphi^s_{t_j}$:

$$\varphi^s_{t_j} \leq \lambda_1 \lambda_2 e^{\rho \Delta (t_T - t_j)} q^s_{t_j} \quad , s \in \mathcal{S}_{t_j}, \ \ j = 1, \ldots, T. \tag{4.60}$$

These upper bounds are calculated by backward recursion, which implies that they may be tighter than the upper bounds in lemma 4.1 at points in time close to the planning horizon.

Even if one isn't sure that the borrowing constraints are not binding in an optimal solution, the upper bounds in (4.60) can be used to calculate an *approximate* bound on the change in objective value.

Sensitivity Analysis

Instead of computing *bounds* on the change in the objective value of the ALM model after a disaggregation, sensitivity analysis can be used to provide an *estimate* of this change. A computational advantage of sensitivity analysis is that it saves the calculation of the quantities $\beta_{t_j}^s$ and $\bar{W}_{t_j}^s$ for each scenario in the upper bounds on the primal and dual variables.

Let $\hat{x}_t^s \equiv (\hat{xs}_t^s, \hat{xb}_t^s, \hat{xh}_t^s, \hat{y}_t^s, \hat{z}_t^s)$ denote an optimal solution for scenario s at time t in the ALM model before a disaggregation, and $\hat{u}_t^s \equiv (\hat{\varphi}_t^s, \hat{\mu}_t^s, \hat{\xi}_t^s)$ the corresponding dual solution. We will show how to use sensitivity analysis when a basic state or time disaggregation is performed in the aggregated event tree that underlies the ALM model. Our starting point in each case is the fixed-weight solution.

For the basic state disaggregation that corresponds to figure 4-1, we have seen earlier that the fixed-weight solution as defined in (4.38) may violate the cash-balance and borrowing constraints in states at time t_{j+1}. For scenario s_l in state $(n_l, k_1 - 1)$ at time t_{j+1}, define $U_{t_{j+1}}^{s_l}$ as the violation of the cash-balance constraint:

$$\begin{aligned} U_{t_{j+1}}^{s_l} = L_{t_{j+1}\to t_{j+2}}^{(n_l,k_1-1)} - \\ \Big[D_{t_{j+1}}^{(n_l,k_1-1)} \hat{xh}_{t_j}^{s^-} + \hat{y}_{t_j}^{s^-} - \hat{z}_{t_j}^{s^-} + (1-c) S_{t_{j+1}}^{(n_l,k_1-1)} \hat{xs}_{t_{j+1}}^{s} \\ - (1+c) S_{t_{j+1}}^{(n_l,k_1-1)} \hat{xb}_{t_{j+1}}^{s} + D_{t_{j+1}\to t_{j+2}}^{(n_l,k_1-1)} \hat{xh}_{t_{j+1}}^{s} - P_{t_{j+1}\to t_{j+2}}^{(n_l,k_1-1)} \hat{y}_{t_{j+1}}^{s} \\ + \left(e^{-\rho \Delta\tau_{j+2}} P_{t_{j+1}\to t_{j+2}}^{(n_l,k_1-1)}\right) \hat{z}_{t_{j+1}}^{s} \Big] \end{aligned} \tag{4.61}$$

and $V_{t_{j+1}}^{s_l}$ as the violation of the borrowing constraint:

$$V_{t_{j+1}}^{s_l} = \left[\hat{z}_{t_{j+1}}^{s} - \bar{Z}_{t_{j+1}\to t_{j+2}}^{(n_l,k_1-1)}\right]^+ \tag{4.62}$$

where scenario s at time t_{j+1} is the single successor of scenario s^- at time t_j before the state disaggregation.

If we associate the fixed-weight dual solution (i.e., $\tilde{\varphi}_{t_{j+1}}^{s_l} = \hat{\pi}_{t_{j+1}}^{n_l} \hat{\varphi}_{t_{j+1}}^{s}$ and $\tilde{\xi}_{t_{j+1}}^{s_l} = \hat{\pi}_{t_{j+1}}^{n_l} \hat{\xi}_{t_{j+1}}^{s}$; $\hat{\pi}_{t_{j+1}}^{n_l}$ is the risk-neutral conditional probability of state $(n_l, k_1 - 1)$ at

time t_{j+1}, given state (n, k_0) at time t_j) with these constraints, we have

$$\sum_{l=0}^{M} \tilde{\varphi}^{s_l}_{t_{j+1}} U^{s_l}_{t_{j+1}} = \hat{\varphi}^{s}_{t_{j+1}} \sum_{l=0}^{M} \hat{\pi}^{n_l}_{t_{j+1}} U^{s_l}_{t_{j+1}} = 0$$

because the sum in the last expression is zero (the cash-balance constraints are satisfied *on average*; see also section 4.3.1). Thus, the "normal" way of performing sensitivity analysis with respect to the cash-balance constraints (and on the basis of the fixed-weight solution) predicts that the objective value will not change after the state disaggregation.

Instead, we propose to measure the *sensitivity* of the objective function to the state disaggregation in state (n, k_0) at time t_j by only considering the scenarios s_l at time t_{j+1} for which $U^{s_l}_{t_{j+1}}$ is positive[7]. The sensitivity measure is then defined as:

$$\begin{aligned} \varepsilon &\equiv \sum_{s \in \mathcal{S}^{(n,k_1)}_{t_{j+1}}} \sum_{l=0}^{M} \left(\tilde{\varphi}^{s_l}_{t_{j+1}} \left[U^{s_l}_{t_{j+1}} \right]^+ + \tilde{\xi}^{s_l}_{t_{j+1}} V^{s_l}_{t_{j+1}} \right) \\ &= \sum_{s \in \mathcal{S}^{(n,k_1)}_{t_{j+1}}} \left(\hat{\varphi}^{s}_{t_{j+1}} \sum_{l=0}^{M} \hat{\pi}^{n_l}_{t_{j+1}} \left[U^{s_l}_{t_{j+1}} \right]^+ + \hat{\xi}^{s}_{t_{j+1}} \sum_{l=0}^{M} \hat{\pi}^{n_l}_{t_{j+1}} V^{s_l}_{t_{j+1}} \right) \end{aligned} \tag{4.63}$$

Note that the sum is taken over all scenarios s in state (n, k_1) at time t_{j+1}, as state disaggregation in state (n, k_0) at time t_j splits each of these scenarios in $M + 1$ new scenarios.

A similar sensitivity measure can be defined for basic time disaggregation, corresponding to the reversal of the aggregation in figure 4-2. It was found in section 4.3.1 that the fixed-weight primal solution satisfies all constraints in the disaggregated ALM model with the possible exception of the cash-balance constraints for each scenario s^- in state (n, k_0) at time t_j, and its successor scenario s at time t_{j+1}. Let $U^{s^-}_{t_j}$ and $U^{s}_{t_{j+1}}$ equal the discrepancies in the respective cash-balance constraints (defined as in (4.61)). The fixed-weight dual solution has $\tilde{\varphi}^{s^-}_{t_j} = \hat{\varphi}^{s^-}_{t_j}$ and $\tilde{\varphi}^{s}_{t_{j+1}} = P^{(n,k_0)}_{t_j \to t_{j+1}} \hat{\varphi}^{s^-}_{t_j}$ (see (4.42)), and it is easy to see that $\tilde{\varphi}^{s^-}_{t_j} U^{s^-}_{t_j} + \tilde{\varphi}^{s}_{t_{j+1}} U^{s}_{t_{j+1}} = 0$. That is, the estimated effect of the time disaggregation on the objective value is zero. To measure the sensitivity of the objective value to the time disaggregation, we only consider the

[7] This corresponds to writing the cash-balance constraints as greater-than-or-equal-to constraints instead of equalities in the ALM model, and defining U^s_t as the violation by the fixed-weight solution of the cash-balance constraint in inequality form. See also footnote 6 in this chapter.

scenarios in which U_t^s is positive, and define

$$\eta \equiv \sum_{s^- \in \mathcal{S}_{t_j}^{(n,k_0)}} \left(\tilde{\varphi}_{t_j}^{s^-} \left[U_{t_j}^{s^-} \right]^+ + \tilde{\varphi}_{t_{j+1}}^{s} \left[U_{t_{j+1}}^{s} \right]^+ \right) \tag{4.64}$$

$$= \sum_{s^- \in \mathcal{S}_{t_j}^{(n,k_0)}} \hat{\varphi}_{t_j}^{s^-} \left(\left[U_{t_j}^{s^-} \right]^+ + P_{t_j \to t_{j+1}}^{(n,k_0)} \left[U_{t_{j+1}}^{s} \right]^+ \right)$$

As we noted in section 4.3.1, the fixed-weight primal solution will satisfy the cash-balance constraints after the time disaggregation if no dividends are paid and no liability payments are due between time t_j and time t_{j+2} (i.e., if no prepayment of dividends and liabilities takes place in state (n, k_0) before the time disaggregation). In that case, the sensitivity measure η will be zero.

It is clear from this discussion that the sensitivity measure ε for a state disaggregation may very often dominate the sensitivity measure η for a time disaggregation. If one would therefore base the disaggregation solely on the values of ε and η, a new trading date may only occasionally be added to the aggregated ALM model. If that is the case, one may want to use a different criterium to decide when to perform a time disaggregation. An example is to impose the somewhat arbitrary restriction that the number of successors of a state in an aggregated event tree can never exceed a certain maximum. If a state is chosen for a state disaggregation according to its value of ε, but if this state disaggregation would lead to a number of successor states that exceeds this maximum, then a time instead of a state disaggregation is performed.

Obviously, many other criteria, less rigorous than the ones discussed here, can be devised to perform the disaggregations in an aggregated event tree. The efficiency of such rules can only be judged through computational experiments, and may be problem specific. In chapter 7 we report on some results for a simple asset/liability management problem.

4.3.3 Terminating the Algorithm

For the decision when to terminate the iterative disaggregation algorithm, we would like to have a measure of how close the current solution is to the solution of the unaggregated ALM model (i.e., the ALM model that is based on the unaggregated event tree). Because an aggregated ALM model is neither a restriction nor a relaxation of the unaggregated model, it is impossible to tell precisely how the optimal solution to the aggregated model relates to the solution for the unaggregated model. To

choose the disaggregations in each iteration of the iterative disaggregation algorithm, we have suggested the use of proposition 4.2 to calculate bounds on the difference in optimum objective value between the aggregated ALM models in successive iterations. However, these bounds are most likely too weak to be meaningfull (if at all possible to calculate) if one would try to translate the optimal solution to an aggregated model to a solution for the unaggregated model by fixed-weight disaggregation, because the unaggregated model is typically very much larger than the aggregated models that are solved in the algorithm.

Instead, a decision to terminate the algorithm will have to be based on the results in past iterations, and a trade-off has to be made between the value of extra detail in the ALM model and the cost of re-optimizing the resulting model. In practical applications, the ALM model will generally be updated and re-solved when new information becomes available over time, and an investor will primarily be interested in the optimal portfolio decisions at time 0. One can therefore decide to terminate the algorithm if the optimal portfolio decisions at time 0 have remained stable in recent iterations.

An extra test for the robustness of the optimal portfolio strategy could be performed as follows. Define a version of the ALM model which is significantly less aggregated than the last model that was optimized in the iterative disaggregation algorithm, and construct a feasible solution to this model (or a relaxation of it; see section 4.3.1) from the optimal strategy by fixed-weight disaggregation. If this fixed-weight solution does not lead to large violations of the cash-balance constraint, particularly at times close to time 0, one can conclude that the obtained portfolio strategy forms a good hedge against additional uncertainty that was not explicitly considered in the model. Otherwise, one may decide to continue the algorithm.

Obviously, many different criteria can be employed for the decision when to stop, depending on the nature of the specific problem on hand and the investor's objectives. These may also influence the strategy that is followed for choosing the disaggregations in each iteration. The iterative disaggregation algorithm provides the flexibility for variations on these points, and enables the investor to see the corresponding effects.

Chapter 5

Decomposition Methods for the Optimization of the ALM Model

In chapters 2 and 3 we have formulated the ALM problem as a multistage stochastic linear program, and in chapter 4 we proposed to solve this ALM model by the iterative disaggregation algorithm. The different parts of this algorithm were discussed in detail there, except for the re-optimization of the ALM model in each iteration. That is the topic of this chapter.

Probably the best known and most widely used solution method for stochastic programs is Benders' decomposition. The method was originally developed by Benders [1] for mixed-integer programming problems, and adapted to two-stage stochastic programs by Van Slyke and Wets [54]. Birge [3] extended the method to multistage stochastic programs. It has been successfully applied to a variety of practical stochastic programming problems (see, for example, Ermoliev and Wets [17]), and in a recent comparison of different solution methods, Cariño et al. [9] concluded that Benders' decomposition was the method of choice for large instances of their asset/liability management model, which they formulated as a multistage stochastic linear program. We will show, however, that Benders' decomposition is not well suited to perform the re-optimizations of the ALM model in the iterative disaggregation algorithm as it has to discard most of the information from previous iterations.

We will also present a different decomposition method, primal-dual decomposition, and show that it can make full use of a previous solution to the ALM model to perform the re-optimizations in the iterative disaggregation algorithm. The version of the method that we will discuss is directly based on the primal-dual method which Grinold [19] describes for two-stage linear programs. He shows that this method is

equivalent to a steepest-ascent algorithm for linear programs. Grinold also discusses an application of his two-stage method to solve multistage linear programs, which he calls trajectory optimization. Our proposed extension of the method to multistage stochastic programs differs from his, however, and we will point out the differences in the sequel.

To convey the basic idea of both Benders' and primal-dual decomposition, we will introduce both methods in section 5.1 for two-stage stochastic linear programs. It will become clear that both decomposition methods essentially consider the same reformulation of a stochastic program, although they differ in how they make use of it. Their extension to multistage stochastic linear programs will be discussed in section 5.2. In section 5.3 we will study the applicability of each of the decomposition methods for the re-optimization of the ALM model in the iterative disaggregation algorithm.

We only describe the basic version of Benders' decomposition in this chapter. In the literature, several enhancements have been described for the method, many of which are based on the assumption that the stochasticity in the stochastic program is restricted to the right-hand-side vector (we note that this is not the case in the ALM model). Infanger [37] provides a recent overview, and contains many references. Other decomposition methods for stochastic programming have been proposed in the literature as well. We will not discuss these methods here, but refer to Ermoliev and Wets [17] for a relatively complete overview and many references.

5.1 Decomposition Methods for Two-Stage Stochastic Linear Programs

We view the two-stage stochastic program as a sequential decision problem under uncertainty over two periods (stages). Uncertainty exists about the state of the world that will occur at the beginning of the second period. The problem is to determine a first-stage decision as well as a set of second-stage decisions, one for each possible state of the world in the second period, so as to optimize some objective. This problem can be stated mathematically as:

$$
\begin{array}{lllllll}
\text{(SLP):}\ v^* = \min & c_0x_0 & +c_1^1x_1^1 & +c_1^2x_1^2 & +\ldots+c_1^Kx_1^K & & \\
\text{s.t.} & A_0x_0 & & & & = b_0 & \\
& F_1^1x_0 & +A_1^1x_1^1 & & & = b_1^1 & \\
& F_1^2x_0 & & +A_1^2x_1^2 & & = b_1^2 & \\
& \vdots & & & \ddots & \vdots & \\
& F_1^Kx_0 & & & +A_1^Kx_1^K & = b_1^K & \\
& x_0 \geq 0, x_1^1 \geq 0, x_1^2 \geq 0, & \ldots, x_1^K \geq 0 & & & &
\end{array} \tag{5.1}
$$

where x_0 is a n_0-vector of first-stage decisions, x_1^k a n_1-vector of second-stage decisions if state of the world k occurs at the end of the first stage, the right-hand-side vectors b_0 and b_1^k have length m_0 and m_1 respectively, and all other vectors and matrices are dimensioned accordingly. We assume that this problem is feasible, and has a finite optimal solution.[1]

The vector of objective coefficients c_1^k has the form $\pi_k q_1^k$ in a stochastic program, where π_k denotes the probability that state of the world k occurs at the end of the first stage, and q_1^k is a n_1-vector. We note that the formulation of (SLP) allows that the right-hand-side vectors b_1^k, constraint matrices F_1^k and A_1^k, and the vectors of objective coeeficients q_1^k all depend on the particular state of the world k after the first stage.

We will now derive a different formulation for (SLP) that forms the basis of the decomposition methods that we are going to discuss. Given some vector of first-stage decisions x_0, define a separate *subproblem* for each of the possible states of the world k at the beginning of the second stage:

$$
\begin{array}{lll}
(\text{SUB}^k): \ h^k(x_0) = \min & c_1^kx_1^k & \\
\text{s.t.} & A_1^kx_1^k = b_1^k - F_1^kx_0 & \\
& x_1^k \geq 0 &
\end{array} \tag{5.2}
$$

This allows us to rewrite the stochastic program as:

$$
\begin{array}{lll}
\text{(SLP):}\ v^* = \min & c_0x_0 + \sum_{k=1}^K h^k(x_0) & \\
\text{s.t.} & A_0x_0 = b_0 & \\
& x_0 \geq 0 &
\end{array} \tag{5.3}
$$

Consider the dual of the subproblem (5.2), where u_1^k is the m_1-vector of dual variables:

[1] The assumption of feasibility is necessary for the primal-dual method, whereas the assumption of a finite optimal solution is needed to apply Benders' decomposition.

$$(\text{DSUB}^k): \quad h^k(x_0) = \max \quad u_1^k (b_1^k - F_1^k x_0) \qquad (5.4)$$
$$\text{s.t.} \quad u_1^k A_1^k \leq c_1^k$$

We will denote the set of feasible solutions of this dual subproblem by U^k, i.e., $U^k = \{u | u A_1^k \leq c_1^k\}$. Because (SLP) is assumed to have a bounded solution, (DSUB^k) must be feasible, and thus U^k is non-empty. Note that the feasible set U^k does not depend on the vector of first-stage decisions x_0.

U^k is a polyhedron in $\mathbf{R}^{m_1}$, and we denote its set of extreme points by $\mathcal{P}^k$, and its set of extreme rays by $\mathcal{R}^k$. The set of extreme rays $\mathcal{R}^k$ can be characterized as the set of extreme points of the set $Y = \{y | y A_1^k \leq 0, ey = 1\}$, where e is a vector with all elements equal to one (see Murty [47]).

It is well-known that if a linear program has an optimal solution, then there exists an extreme point that is optimal (see Murty [47, theorem 3.3]). This property allows us to write the optimal solution to (DSUB^k) as:

$$h^k(x_0) = \begin{cases} \max\limits_{p \in \mathcal{P}^k} p (b_1^k - F_1^k x_0) & \text{if } r(b_1^k - F_1^k x_0) \leq 0 \quad \forall\, r \in \mathcal{R}^k \\ \infty & \text{otherwise} \end{cases} \qquad (5.5)$$

The condition $r(b_1^k - F_1^k x_0) \leq 0$ for all $r \in \mathcal{R}^k$ states that there is no extreme ray of U^k that strictly improves the objective function of (DSUB^k) for the current value of x_0, and thus would lead to an unbounded solution. By linear programming duality, an unbounded solution to the dual subproblem (DSUB^k) implies infeasibility of the primal subproblem (SUB^k) for the given vector of first-stage decisions x_0, and thus infeasibility of (SLP). The vector of first-stage decisions x_0 should therefore be chosen such that for all $k = 1, \ldots, K$:

$$(r F_1^k)\, x_0 \geq r b_1^k \qquad \forall\, r \in \mathcal{R}^k \qquad (5.6)$$

These constraints will be called *feasibility cuts.* If x_0 satisfies all feasibility cuts, then (DSUB^k) can be rewritten as

$$(\text{DSUB}^k): \quad h^k(x_0) = \min \quad \theta^k \qquad (5.7)$$
$$\text{s.t.} \quad \theta^k \geq p (b_1^k - F_1^k x_0) \qquad \forall\, p \in \mathcal{P}^k$$

where we have used the characterization of $h^k(x_0)$ in (5.5). The constraints in (5.7) are called *optimality cuts.*

Two properties of $h^k(x_0)$ follow directly from its characterization in (5.5). First, $h^k(x_0)$ is a piecewise linear and convex function when it is finite. Second, if $\tilde{x}_0$ is such that $h^k(\tilde{x}_0)$ is finite, and if $h^k(\tilde{x}_0) = \tilde{p}(b_1^k - F_1^k \tilde{x}_0)$ for some $\tilde{p} \in \mathcal{P}^k$, then $(-\tilde{p} F_1^k)$ is a subgradient of $h^k(x_0)$ in $x_0 = \tilde{x}_0$. This second property follows directly from

$$h^k(x_0) \geq \tilde{p}(b_1^k - F_1^k x_0) = \tilde{p}(b_1^k - F_1^k[\tilde{x}_0 + x_0 - \tilde{x}_0]) = h^k(\tilde{x}_0) - \tilde{p}F_1^k(x_0 - \tilde{x}_0) \quad (5.8)$$

and plays an important role in the decomposition methods.

By adding the feasibility cuts (5.6) to (5.3) for each k, and using the characterization of $h^k(x_0)$ in (5.7), we obtain the following formulation of (SLP):

$$\begin{array}{lllll} \text{(SLP):} \ v^* = \min & c_0 x_0 + \sum_{k=1}^K \theta^k & & & \quad (5.9) \\ \text{s.t.} & A_0 x_0 & = b_0 & & \\ & (rF_1^k)\, x_0 & \geq r b_1^k & \forall\, r \in \mathcal{R}^k, & k = 1, \ldots, K \\ & (pF_1^k)\, x_0 + \theta^k & \geq p b_1^k & \forall\, p \in \mathcal{P}^k, & k = 1, \ldots, K \\ & x_0 \geq 0 & & & \end{array}$$

This formulation of (SLP) forms the basis of both Benders' and primal-dual decomposition.[2]

5.1.1 Benders' Decomposition

The number of extreme points and extreme rays of the feasible region U^k in a dual subproblem (DSUB^k) could be very large, leading to a large number of constraints in (5.9). In an optimal solution to (5.9), however, only few of the feasibility and optimality cuts will typically be binding, and knowledge of a small subset of the extreme points and extreme rays of U^k for each k would therefore in principle be sufficient to determine a solution to (SLP). Benders' decomposition aims to find these subsets in an iterative manner without enumerating all extreme points and rays of each U^k.

Suppose that only a subset $\tilde{\mathcal{P}}^k \subset \mathcal{P}^k$ of the extreme points and a subset $\tilde{\mathcal{R}}^k \subset \mathcal{R}^k$ of the extreme rays of the feasible region U^k are known in some iteration of the method. (At the start of Benders' decomposition both $\tilde{\mathcal{P}}^k$ and $\tilde{\mathcal{R}}^k$ may be empty.) By only including the cuts that correspond to these subsets, one obtains the following relaxation of (5.9):

$$\begin{array}{lllll} \tilde{v} = \min & c_0 x_0 + \sum_{k=1}^K \theta^k & & & \quad (5.10) \\ \text{s.t.} & A_0 x_0 & = b_0 & & \\ & (rF_1^k)\, x_0 & \geq r b_1^k & \forall\, r \in \tilde{\mathcal{R}}^k \subset \mathcal{R}^k, & k = 1, \ldots, K \\ & (pF_1^k)\, x_0 + \theta^k & \geq p b_1^k & \forall\, p \in \tilde{\mathcal{P}}^k \subset \mathcal{P}^k, & k = 1, \ldots, K \\ & x_0 \geq 0 & & & \end{array}$$

[2]This formulation of (SLP) can also be derived by first taking the dual of (SLP) in (5.1), and then applying Lagrangean relaxation to the constraints in this dual formulation that correspond to the vector of first-stage decisions x_0 (the *coupling* constraints). This is essentially what Grinold [19] does.

This problem is called the *master* problem. As long as no elements of $\mathcal{P}^k$ are known (i.e., $\tilde{\mathcal{P}}^k = \emptyset$), the variable θ^k is set equal to zero; this corresponds to the situation that no feasible solution to (SLP) has yet been found. We will now describe an iteration in Benders' decomposition.

Suppose $(\tilde{x}_0, \tilde{\theta}^1, \ldots, \tilde{\theta}^K)$ is an optimal solution to the master problem (5.10). Solve each of the K subproblems for $x_0 = \tilde{x}_0$. The optimal solution to the k-th subproblem will satisfy one of the following:

1. $h^k(\tilde{x}_0) = \infty$:
   This corresponds to the situation that (SUB^k) is infeasible, and (DSUB^k) is unbounded. We thus find an extreme ray $\tilde{r}$ of U^k with $\tilde{r}(b_1^k - F_1^k\tilde{x}_0) > 0$. Equivalently, $(\tilde{r}F_1^k)\tilde{x}_0 < \tilde{r}b_1^k$, and $\tilde{x}_0$ therefore violates the feasibility cut in (5.9) that corresponds to $\tilde{r}$. That is, $\tilde{r} \notin \tilde{\mathcal{R}}^k$, and the set $\tilde{\mathcal{R}}^k$ is increased with this new extreme ray.

2. $\infty > h^k(\tilde{x}_0) > \tilde{\theta}^k$:
   Let $\tilde{p}$ denote the optimum extreme point solution to (DSUB^k). Then $h^k(\tilde{x}_0) = \tilde{p}(b_1^k - F_1^k\tilde{x}_0) > \tilde{\theta}^k$, or equivalently $(\tilde{p}F_1^k)\tilde{x}_0 + \tilde{\theta}^k < \tilde{p}b_1^k$. Thus $(\tilde{x}_0, \tilde{\theta}^k)$ violates the optimality cut in (5.9) that corresponds to the extreme point $\tilde{p} \in \mathcal{P}^k$, implying $\tilde{p} \notin \tilde{\mathcal{P}}^k$. The new extreme point $\tilde{p}$ is therefore added to the set $\tilde{\mathcal{P}}^k$.

3. $h^k(\tilde{x}_0) = \tilde{\theta}^k$:
   In this case, $(\tilde{x}_0, \tilde{\theta}^k)$ satisfies the feasibility and optimality cuts for all $r \in \mathcal{R}^k$ and $p \in \mathcal{P}^k$, respectively.

When situation 1 or 2 occurs for a subproblem k, a new feasibility or optimality cut is added to the master problem. After the appropriate cuts are added for all subproblems, the master problem can be re-optimized. This re-optimization can be done efficiently using the dual simplex method, as the added cuts are the only constraints that are violated by the current solution $(\tilde{x}_0, \tilde{\theta}^1, \ldots, \tilde{\theta}^K)$.

If situation 3 occurs for all K subproblems, then we have

$$\tilde{v} = c_0\tilde{x}_0 + \sum_{k=1}^{K} \tilde{\theta}^k = c_0\tilde{x}_0 + \sum_{k=1}^{K} h^k(\tilde{x}_0) \geq v^*$$

where the inequality follows from the formulation of (SLP) in (5.3). On the other hand, $\tilde{v} \leq v^*$ because the master problem is a relaxation of (SLP). Thus $\tilde{v} = v^*$, and $\tilde{x}_0$ is an optimal first-stage solution. The optimal second-stage decisions follow from the solutions to the subproblems.

Because the number of extreme points and extreme rays is finite for each polyhedral set U^k, situations 1 and 2 can only occur a finite number of times for each subproblem. This proves the finite convergence of the Benders' decomposition method. In practical applications, however, Benders' decomposition is usually terminated before absolute convergence, based on the bounds on v^* that it provides in each iteration.

When $\tilde{\mathcal{P}}^k$ is nonempty for every subproblem $k = 1, \ldots, K$, the solution to the master problem $\tilde{v}$ is a lower bound on v^* as this master problem is a relaxation of the true problem (5.9). Because a constraint is added to the master problem in each iteration, the last solution to the master problem always gives the highest lower bound.

An upper bound on v^* is obtained whenever situation 2 or 3 occurs for each subproblem in an iteration, i.e., all dual subproblems are bounded. This implies that the primal subproblem (SUB^k) has an optimal solution, say $\tilde{x}_1^k$, for each k, and the solution $(\tilde{x}_0, \tilde{x}_1^1, \ldots, \tilde{x}_1^K)$ constitutes a feasible solution to (SLP). Thus $c_0\tilde{x}_0 + \sum_{k=1}^K c_1^k \tilde{x}_1^k$ forms an upper bound on v^*. This upper bound, however, does not necessarily decrease with every new feasible solution that is found.

5.1.2 Primal-Dual Decomposition

In contrast to Benders' decomposition, the primal-dual decomposition method finds a *better* feasible solution (i.e., with lower objective value) to (SLP) in each iteration, given some initial feasible solution. However, it does not supply a bound on the deviation of the corresponding solution value from the optimum solution v^*.

For the primal-dual decomposition method it is assumed that an initial feasible first-stage solution $\tilde{x}_0$ is known such that all subproblems (SUB^k) are feasible (i.e., $h^k(\tilde{x}_0) < \infty$ for all k). It is clear from the preceding paragraph that any subsequent first-stage solution that is generated by the primal-dual method will also satisfy this assumption. Given this assumption, $(\tilde{x}_0, \tilde{\theta}^1, \ldots, \tilde{\theta}^K)$ with $\tilde{\theta}^k = h^k(\tilde{x}_0)$ for each $k = 1, \ldots, K$ is a feasible solution to (SLP) in (5.9). What follows is an outline of an iteration in the primal-dual decomposition method. The individual parts in the iteration will be discussed in detail after that.

Each iteration passes through two steps, but possibly several times, in order to find a new and improved solution to (SLP) in (5.9). In the first step, a *direction-finding problem* is solved that either establishes the optimality of the current solution $\tilde{x}_0$, or supplies a descent direction $(d, \sigma^1, \ldots, \sigma^K)$ from $(\tilde{x}_0, \tilde{\theta}^1, \ldots, \tilde{\theta}^K)$. In the second step, the maximum stepsize α is determined that can be taken from $(\tilde{x}_0, \tilde{\theta}^1, \ldots, \tilde{\theta}^K)$

in this direction $(d, \sigma^1, \ldots, \sigma^K)$ while maintaining feasibility in (SLP). If $\alpha = 0$, then the descent direction $(d, \sigma^1, \ldots, \sigma^K)$ is not a feasible direction. We will show, however, that this gives us information that can be used to update the direction-finding problem, and we return to the first step. If $\alpha > 0$, then $(d, \sigma^1, \ldots, \sigma^K)$ is a feasible direction, and $(\tilde{x}_0 + \alpha d, \tilde{\theta}^1 + \alpha\sigma^1, \ldots, \tilde{\theta}^K + \alpha\sigma^K)$ an improved solution to (SLP) in (5.9). A new iteration is then started.

Our description below of the individual steps in each iteration largely follows Grinold [19], although we will present a more efficient way to update the direction-finding problem if a descent direction turns out to be an infeasible direction. A description of Grinold's method can be found in Shapiro [53, section 6.5] as well, who also provides a convergence proof of the method.

The Direction-Finding Problem

Let $\mathcal{P}^k(\tilde{x}_0)$ denote the set of all *optimal* extreme points in the dual subproblem (DSUB^k) when $x_0 = \tilde{x}_0$, and $\mathcal{R}^k(\tilde{x}_0)$ the set of all *binding* extreme rays. That is,

$$\mathcal{P}^k(\tilde{x}_0) = \{p \in \mathcal{P}^k | p(b_1^k - F_1^k \tilde{x}_0) = h^k(\tilde{x}_0)\}$$
$$\mathcal{R}^k(\tilde{x}_0) = \{r \in \mathcal{R}^k | r(b_1^k - F_1^k \tilde{x}_0) = 0\}$$

Notice that these sets correspond to the binding optimality and feasibility cuts in the current solution $(\tilde{x}_0, \tilde{\theta}^1, \ldots, \tilde{\theta}^K)$ for (SLP) in (5.9); all other cuts are satisfied by strict inequality.

The *direction-finding problem* aims to determine a feasible direction $(d, \sigma^1, \ldots, \sigma^K)$ in which to change $(\tilde{x}_0, \tilde{\theta}^1, \ldots, \tilde{\theta}^K)$ so as to maximize the decrease in the objective function value of (SLP) in (5.9). We will first assume that we have complete knowledge of the sets $\mathcal{P}^k(\tilde{x}_0)$ and $\mathcal{R}^k(\tilde{x}_0)$. An optimal direction in the direction-finding problem is then guaranteed to be a *feasible* descent direction. We will then relax this assumption, and consider the case that only partial knowledge of $\mathcal{P}^k(\tilde{x}_0)$ and $\mathcal{R}^k(\tilde{x}_0)$ exists.

Assuming that the sets $\mathcal{P}^k(\tilde{x}_0)$ and $\mathcal{R}^k(\tilde{x}_0)$ are known completely, the direction-finding problem is:

$$
\begin{aligned}
\text{(DIR):}\ \delta^* = \min\ & c_0 d + \sum_{k=1}^{K} \sigma^k && && (5.11)\\
\text{s.t.}\ & A_0 d && = 0 && \\
& (rF_1^k)\, d && \geq 0 && \forall\, r \in \mathcal{R}^k(\tilde{x}_0),\ \ k = 1, \ldots, K \\
& (pF_1^k)\, d + \sigma^k && \geq 0 && \forall\, p \in \mathcal{P}^k(\tilde{x}_0),\ \ k = 1, \ldots, K \\
& d_i && \geq 0 && \text{if } \tilde{x}_{0i} = 0 \\
& -e \leq d \leq e && &&
\end{aligned}
$$

The constraints $-e \leq d \leq e$ are imposed to normalize d. The other constraints make sure that d is a feasible direction, which follows directly from the formulation of (SLP) in (5.9) and the definition of the sets $\mathcal{P}^k(\tilde{x}_0)$ and $\mathcal{R}^k(\tilde{x}_0)$. Grinold [19] shows that an optimal solution to this direction-finding problem is the *steepest-descent* direction among all feasible directions. Notice that $(d, \sigma^1, \ldots, \sigma^K) = (0, 0, \ldots, 0)$ is a feasible solution, and thus $\delta^* \leq 0$. An optimal value $\delta^* = 0$ implies that no descent direction can be found, and thus the current solution $\tilde{x}_0$ is optimal. If $\delta^* < 0$, then we can strictly improve the current objective value of (SLP) by changing $\tilde{x}_0$ in the direction d^*; the value for σ^k equals the corresponding change in $h^k(x_0)$.

To see this interpretation for σ^k, suppose that d^* is an optimal solution for d in (DIR). The value of σ^k is then completely defined by the constraints:

$$\sigma^k \geq (-pF_1^k)\, d^* \qquad \forall\, p \in \mathcal{P}^k(\tilde{x}_0)$$

As σ^k is being minimized in (DIR), we have

$$\sigma^k = \max_{p \in \mathcal{P}^k(\tilde{x}_0)} (-pF_1^k)\, d^*$$

We have seen in (5.8) that $(-pF_1^k)$ is a subgradient of $h^k(x_0)$ in $x_0 = \tilde{x}_0$ for each $p \in \mathcal{P}^k(\tilde{x}_0)$. Furthermore, the subdifferential (the set of all subgradients) of $h^k(x_0)$ in $x_0 = \tilde{x}_0$ consists of all convex combinations of the vectors $(-pF_1^k)$, $p \in \mathcal{P}^k(\tilde{x}_0)$ (see, for example, Nemhauser and Wolsey [48, section I.2.4]). Rockafellar [51, section 23] shows that σ^k therefore equals the directional derivative of $h^k(\tilde{x}_0)$ in the direction d^*. Because $h^k(x_0)$ is a piecewise linear function, this implies that there exists a constant $\bar{\alpha} > 0$ such that

$$h^k(\tilde{x}_0 + \alpha d^*) = h^k(\tilde{x}_0) + \alpha\sigma^k \qquad \forall\, \alpha \in [0, \bar{\alpha}]$$

That is, for a small enough stepsize, σ^k equals the change in $h^k(x_0)$ when x_0 makes a step in the direction d^* from its current value $\tilde{x}_0$.Thus, if $\delta^* < 0$ in (DIR), then the optimal solution d^* for d is a descent direction from $\tilde{x}_0$, and we are guaranteed that a

positive step can be taken in this direction. The new solution to (SLP) will therefore have a strictly lower objective value.

We now consider the relaxation of (DIR) that arises if only subsets $\tilde{\mathcal{P}}^k(\tilde{x}_0)$ of $\mathcal{P}^k(\tilde{x}_0)$ and $\tilde{\mathcal{R}}^k(\tilde{x}_0)$ of $\mathcal{R}^k(\tilde{x}_0)$ are known for each $k = 1, \ldots, K$. This *approximate direction-finding problem* is:

$$
\begin{aligned}
(\widetilde{\text{DIR}}): \quad \tilde{\delta} = \min \; & c_0\, d + \sum_{k=1}^{K} \sigma^k && && (5.12)\\
\text{s.t.} \; & A_0 d && = 0 && \\
& (r F_1^k)\, d && \geq 0 && \forall\, r \in \tilde{\mathcal{R}}^k(\tilde{x}_0),\;\; k = 1, \ldots, K \\
& (p F_1^k)\, d + \sigma^k && \geq 0 && \forall\, p \in \tilde{\mathcal{P}}^k(\tilde{x}_0),\;\; k = 1, \ldots, K \\
& d_i && \geq 0 && \text{if } \tilde{x}_{0i} = 0 \\
& -e \leq d \leq e &&
\end{aligned}
$$

In solving this approximate problem, we sacrifice the guarantee that an optimal direction $(\tilde{d}, \tilde{\sigma}^1, \ldots, \tilde{\sigma}^K)$ is a *feasible* descent direction, and thus that we can take a positive step in that direction.

To find out if a solution $(\tilde{d}, \tilde{\sigma}^1, \ldots, \tilde{\sigma}^K)$ to $(\widetilde{\text{DIR}})$ is a feasible descent direction, Grinold [19] suggests to solve the following *feasibility problem* for each of the K subproblems:

$$
\begin{aligned}
f^k(\tilde{d}) = \max \; & -u_1^k (F_1^k \tilde{d}) && (5.13)\\
\text{s.t.} \; & u_1^k A_1^k \leq c_1^k \\
& u_1^k \,(b_1^k - F_1^k \tilde{x}_0) = h^k(\tilde{x}_0)
\end{aligned}
$$

The last constraint imposes that only solutions $u_1^k \in U^k$ can be considered that are optimal in the dual subproblem when $x_0 = \tilde{x}_0$. If $f^k(\tilde{d})$ is unbounded, then we must have found an extreme ray $r \in \mathcal{R}^k(\tilde{x}_0)$ with $-r(F_1^k \tilde{d}) > 0$, and thus $r \notin \tilde{\mathcal{R}}^k(\tilde{x}_0)$. Alternatively, if $f^k(\tilde{d}) > \tilde{\sigma}^k$, then $-\tilde{u}_1^k (F_1^k \tilde{d}) > \tilde{\sigma}^k$ for the optimal solution $\tilde{u}_1^k$, and thus $\tilde{u}_1^k \notin \tilde{\mathcal{P}}^k(\tilde{x}_0)$. In these two cases, we have found a constraint in (DIR) that is violated by $\tilde{d}$, and $\tilde{d}$ is therefore not a feasible direction from $\tilde{x}_0$. The violated constraint should be added to $(\widetilde{\text{DIR}})$, which can then be re-optimized to obtain a new direction. If $f^k(\tilde{d}) = \tilde{\sigma}^k$ for all subproblems k, then $\tilde{d}$ satisfies all constraints in (DIR), and thus is a feasible descent direction.

We will show below that we do not need to solve these separate feasibility problems to update the sets $\tilde{\mathcal{P}}^k(\tilde{x}_0)$ and $\tilde{\mathcal{R}}^k(\tilde{x}_0)$ in $(\widetilde{\text{DIR}})$, but that the procedure that is used to determine the maximum stepsize supplies the necessary information. Obviously, this maximum stepsize will be zero if $(\tilde{d}, \tilde{\sigma}^1, \ldots, \tilde{\sigma}^K)$ is not a feasible direction.

If $(\tilde{d}, \tilde{\sigma}^1, \ldots, \tilde{\sigma}^K)$ is a feasible descent direction, then the stepsize procedure finds some maximum stepsize $\alpha > 0$ that can be taken in this direction without losing feasibility in (SLP). Let $\hat{x}_0 \equiv \tilde{x}_0 + \alpha\tilde{d}$ denote the new first-stage solution after this step. The sets $\tilde{\mathcal{P}}^k(\hat{x}_0)$ and $\tilde{\mathcal{R}}^k(\hat{x}_0)$ can now be initialized for each subproblem k with all elements in $\tilde{\mathcal{P}}^k(\tilde{x}_0)$ and $\tilde{\mathcal{R}}^k(\tilde{x}_0)$ that correspond to binding constraints in ($\widetilde{\text{DIR}}$). This follows directly from the fact that

$$h^k(\hat{x}_0 + \alpha\tilde{d}) = h^k(\tilde{x}_0) + \alpha\tilde{\sigma}^k = p\,(b_1^k - F_1^k\tilde{x}_0) - \alpha(pF_1^k)\,\tilde{d} = p\,(b_1^k - F_1^k(\tilde{x}_0 + \alpha\tilde{d}))$$

for all $p \in \mathcal{P}^k(\tilde{x}_0)$ such that $\tilde{\sigma}^k = -pF_1^k\tilde{d}$, and

$$r\,(b_1^k - F_1^k(\tilde{x}_0 + \alpha\tilde{d})) = r\,(b_1^k - F_1^k\tilde{x}_0) - \alpha(rF_1^k)\,\tilde{d} = 0$$

for all $r \in \mathcal{R}^k(\tilde{x}_0)$ such that $rF_1^k\tilde{d} = 0$. Note that $\tilde{\mathcal{P}}^k(\hat{x}_0)$ will contain at least one element.

Finally, because ($\widetilde{\text{DIR}}$) is a relaxation of (DIR), we can conclude that the current first-stage solution $\tilde{x}_0$ is optimal whenever $\tilde{\delta} = 0$ in ($\widetilde{\text{DIR}}$).

Stepsize Determination

Assume that the approximate direction-finding problem ($\widetilde{\text{DIR}}$) has been solved, and that the optimal solution $(\tilde{d}, \tilde{\sigma}^1, \ldots, \tilde{\sigma}^K)$ is such that $\tilde{\delta} < 0$. The question is now how to determine the maximum stepsize $\alpha \geq 0$ that can be taken in this direction from the current solution $(\tilde{x}_0, \tilde{\theta}^1, \ldots, \tilde{\theta}^K)$ to (SLP) without losing feasibility in (5.9). Note that we allow the direction to be infeasible, in which case $\alpha = 0$.

As $A_0\tilde{d} = 0$, the constraints $A_0(\tilde{x}_0 + \alpha\tilde{d}) = b_0$ are satisfied for any α. The first restriction on α stems from the nonnegativity of the first-stage solution, that is, $\tilde{x}_0 + \alpha\tilde{d} \geq 0$. Thus, α must satisfy:

$$\alpha \leq \alpha_0 \equiv \begin{cases} \infty & \text{if } \tilde{d} \geq 0 \\ \min\limits_{i:\tilde{d}_i<0} \left\{ \dfrac{-\tilde{x}_{0i}}{\tilde{d}_i} \right\} & \text{otherwise.} \end{cases} \tag{5.14}$$

Because we have imposed that $\tilde{d}_i \geq 0$ if $\tilde{x}_{0i} = 0$ in ($\widetilde{\text{DIR}}$), $\alpha_0 > 0$.

The second restriction on the stepsize is that all feasibility and optimality cuts remain satisfied in (5.9). Shapiro [53, section 6.5] suggests a line-search procedure to determine this maximum stepsize. We will show that it can be obtained efficiently by parametric linear programming in the subproblems (SUB^k). If the stepsize turns out to be zero in a subproblem, then the corresponding parametric linear program

will indicate a constraint in (DIR) that is violated by the current solution to ($\widetilde{\text{DIR}}$). Grinold [19] also uses parametric linear programming, but only after he has ascertained that a descent direction is feasible (and thus the stepsize is guaranteed to be positive) by solving the feasibility problems that were discussed earlier.

To simplify notation, we will omit the superscript k and describe the procedure for the generic subproblem

$$\begin{aligned} (\text{SUB}): \quad h(x_0) = \min \quad & c_1 x_1 \\ \text{s.t.} \quad & A_1 x_1 = b_1 - F_1 x_0 \\ & x_1 \geq 0 \end{aligned} \tag{5.15}$$

Let $\tilde{x}_1$ denote an optimal basic feasible solution in this subproblem for $x_0 = \tilde{x}_0$. We will write this subproblem in canonical form with respect to this optimal solution to perform parametric linear programming.

Let $\tilde{x}_1^B$ denote the vector of basic variables in $\tilde{x}_1$, and $\tilde{x}_1^N$ the vector of nonbasic variables. Let B represent the optimal basis (i.e., the columns of the constraint matrix A_1 that correspond to the basic variables) and A^N the columns of A_1 that correspond to the nonbasic variables. The cost vector c_1 is partitioned correspondingly in c^B and c^N. For $x_0 = \tilde{x}_0 + \alpha\tilde{d}$, we can now write the subproblem (SUB) in canonical form as:

$$\begin{aligned} h(\tilde{x}_0 + \alpha\tilde{d}) = \min \quad & (c^N - c^B B^{-1} A^N)\tilde{x}_1^N + c^B B^{-1}(b_1 - F_1(\tilde{x}_0 + \alpha\tilde{d})) \\ \text{s.t.} \quad & \tilde{x}_1^B + (B^{-1}A^N)\tilde{x}_1^N = B^{-1}(b_1 - F_1(\tilde{x}_0 + \alpha\tilde{d})) \\ & \tilde{x}_1^B, \tilde{x}_1^N \geq 0 \end{aligned} \tag{5.16}$$

The dual solution that corresponds to $\tilde{x}_1$ is $\tilde{u}_1 \equiv (c^B B^{-1})$, which is an optimal dual solution if $\alpha = 0$, and thus $\tilde{u}_1 \in \mathcal{P}(\tilde{x}_0)$. We will assume $\tilde{u}_1 \in \tilde{\mathcal{P}}(\tilde{x}_0)$ as well. Notice that the second term in the objective is a constant, which can be rewritten as:

$$c^B B^{-1}(b_1 - F_1(\tilde{x}_0 + \alpha\tilde{d})) = h(\tilde{x}_0) - \alpha\tilde{u}_1 F_1 \tilde{d}$$

To simplify notation further, define

$$\begin{aligned} \bar{c}^B &\equiv c^B - c^B B^{-1} B = 0 \\ \bar{c}^N &\equiv c^N - c^B B^{-1} A^N \\ \bar{A}^N &\equiv B^{-1} A^N \\ \bar{b} &\equiv B^{-1}(b_1 - F_1 \tilde{x}_0) \\ \bar{d} &\equiv -B^{-1} F_1 \tilde{d} \end{aligned} \tag{5.17}$$

so that the subproblem in (5.16) can be written concisely as

$$\begin{aligned} h(\tilde{x}_0 + \alpha\tilde{d}) = h(\tilde{x}_0) - \alpha\tilde{u}_1 F_1 \tilde{d} \\ + \min \quad & \bar{c}^N \tilde{x}_1^N \\ \text{s.t.} \quad & \tilde{x}_1^B + \bar{A}^N \tilde{x}_1^N = \bar{b} + \alpha\bar{d} \\ & \tilde{x}_1^B, \tilde{x}_1^N \geq 0 \end{aligned} \tag{5.18}$$

The vector $\bar{c}^N$ is an $(n_1 - m_1)$-vector with individual elements $\bar{c}_j^N$, $\bar{b}$ and $\bar{d}$ are m_1-vectors with individual elements $\bar{b}_i$ and $\bar{d}_i$, and $\bar{A}^N$ is an $m_1 \times (n_1 - m_1)$ matrix with elements $\bar{a}_{ij}^N$. The basis B is an invertible $m_1 \times m_1$ matrix, and an element of its *inverse* B^{-1} is denoted as β_{ij}. For any matrix D, the i-th row is denoted by $D_{i.}$, and the j-th column by $D_{.j}$.

We have the following result:

Proposition 5.1 *If $\alpha_1 \geq 0$ is the largest value for α in (SUB) so that the current basis B remains feasible, then $(\tilde{x}_0 + \alpha\tilde{d}, \tilde{\theta} + \alpha\tilde{\sigma})$ will satisfy all feasibility and optimality cuts of this subproblem in (5.9) when $\alpha \leq \alpha_1$. Furthermore, if the current solution is dual nondegenerate (i.e., $\bar{c}^N > 0$), then at least one cut will be violated if $\alpha > \alpha_1$.*

PROOF: We first note that the definitions of the reduced-cost coefficients $\bar{c}^B$ and $\bar{c}^N$ do not involve α, so that B remains optimal as long as it remains feasible. Notice further that the corresponding dual solution $\tilde{u}_1$ does not depend on α either, and $\tilde{u}_1$ therefore remains optimal in the dual subproblem as long as the basis B is optimal in the primal subproblem.

Assume $\alpha \leq \alpha_1$. Then B is feasible in (SUB), and thus there cannot exist an extreme ray r in the dual subproblem such that $r\,(b_1 - F_1(\tilde{x}_0 + \alpha\tilde{d})) > 0$. This implies that all feasibility cuts for this subproblem in (5.9) are satisfied in $\tilde{x}_0 + \alpha\tilde{d}$.

Furthermore, $\tilde{u}_1$ is still optimal in the dual subproblem, and thus for an arbitrary extreme point p of the dual feasible region:

$$p\,(b_1 - F_1(\tilde{x}_0 + \alpha\tilde{d})) \leq \tilde{u}_1\,(b_1 - F_1(\tilde{x}_0 + \alpha\tilde{d})) = h(\tilde{x}_0) - \alpha\tilde{u}_1 F_1 \tilde{d} \leq \tilde{\theta} + \alpha\tilde{\sigma}$$

because $\tilde{\theta} = h(\tilde{x}_0)$, and $-\tilde{u}_1 F_1 \tilde{d} \leq \tilde{\sigma}$ from the solution to ($\widetilde{\text{DIR}}$). This implies that all optimality cuts are satisfied in $(\tilde{x}_0 + \alpha\tilde{d}, \tilde{\theta} + \alpha\tilde{\sigma})$.

Now assume $\alpha > \alpha_1$. This implies that the basis B is no longer feasible in (SUB), and (because of the nondegeneracy assumption) $\tilde{u}_1$ is no longer optimal in its dual. Two cases can occur. The first is that the subproblem (SUB) becomes infeasible

altogether, in which case there must be an extreme ray r in the dual subproblem with $r\,(b_1 - F_1(\tilde{x}_0 + \alpha\tilde{d})) > 0$. This corresponds to a feasibility cut that is violated in $x_0 = \tilde{x}_0 + \alpha\tilde{d}$.

Alternatively, (SUB) remains feasible, and the dual subproblem has a new optimal extreme-point solution, say $\hat{u}_1$. Note that there must be at least one element in $\tilde{\mathcal{P}}(\tilde{x}_0)$, say $\tilde{\hat{u}}_1$, for which $\tilde{\sigma} = -\tilde{\hat{u}}_1 F_1 \tilde{d}$ in $(\widetilde{\mathrm{DIR}})$. Thus

$$\begin{aligned}\hat{u}_1\,(b_1 - F_1(\tilde{x}_0 + \alpha\tilde{d})) > \tilde{u}_1\,(b_1 - F_1(\tilde{x}_0 + \alpha\tilde{d})) &= \tilde{\hat{u}}_1\,(b_1 - F_1(\tilde{x}_0 + \alpha\tilde{d})) \\ &= h(\tilde{x}_0) - \alpha\tilde{\hat{u}}_1 F_1\tilde{d} = \tilde{\theta} + \alpha\tilde{\sigma}\end{aligned}$$

which corresponds to an optimality cut that is violated by $(\tilde{x}_0 + \alpha\tilde{d}, \tilde{\theta} + \alpha\tilde{\sigma})$.

$\mathcal{QED}$

Although this proposition tells us that, barring degeneracy, a cut will be violated if $\alpha > \alpha_1$, it does not tell us which one. We will show shortly that this violated cut can be identified by one dual simplex pivot in the subproblem. If the current solution is dual degenerate, and if the dual simplex pivot is a degenerate pivot (i.e., resulting in a change of basis but not in a change of the solution), then the current direction $\tilde{d}$ may still be a feasible descent direction from $\tilde{x}_0$, and a new value for α_1 can be calculated by parametric linear programming with respect to the new basis.

We distinguish three different possibilities for the value of α_1 that is obtained by parametric linear programming in (SUB):

1. $\alpha_1 = \infty$:
   This situation occurs if $\bar{d} \geq 0$.

2. $0 < \alpha_1 < \infty$:
   This happens if $\bar{d}_i \geq 0$ for all rows i in (5.18) with $\bar{b}_i = 0$, and there is at least one row i such that $\bar{b}_i > 0$ and $\bar{d}_i < 0$. Then α_1 is equal to

$$\alpha_1 = \left(\frac{-\bar{b}_r}{\bar{d}_r}\right) \qquad \text{with} \qquad r \equiv \arg\min_{i:\bar{d}_i<0}\left\{\frac{-\bar{b}_i}{\bar{d}_i}\right\}. \tag{5.19}$$

3. $\alpha_1 = 0$:
   This can only be the case if there is a row r in (5.18) with $\bar{b}_r = 0$ and $\bar{d}_r < 0$, i.e., the current solution must be (primal) degenerate.

Suppose that we have performed the parametric linear programming in each of the K subproblems (SUB^k), and let α_1^k denote the upper bound on α that is obtained from

subproblem k. When $\alpha_1^k > 0$ for all k (i.e., situation 1 or 2 occurs for each subproblem), then the actual stepsize equals

$$\alpha = \min\{\alpha_0, \alpha_1^1, \ldots, \alpha_1^K\}, \tag{5.20}$$

where α_0 was defined in (5.14). If $\alpha = \infty$, then (SLP) is unbounded. Otherwise, the new feasible solution for (SLP) in (5.9) is $(\tilde{x}_0 + \alpha\tilde{d}, \tilde{\theta}^1 + \alpha\tilde{\sigma}^1, \ldots, \tilde{\theta}^K + \alpha\tilde{\sigma}^K)$. We will refer to this new solution as $(\hat{x}_0, \hat{\theta}^1, \ldots, \hat{\theta}^K)$.

Before we solve the approximate direction-finding problem ($\widetilde{\text{DIR}}$) for this new solution, the constraints in this problem have to be updated. We noted before that the sets $\tilde{\mathcal{P}}^k(\hat{x}_0)$ and $\tilde{\mathcal{R}}^k(\hat{x}_0)$ for each subproblem k can be initialized with the elements in $\tilde{\mathcal{P}}^k(\tilde{x}_0)$ and $\tilde{\mathcal{R}}^k(\tilde{x}_0)$ which corresponded to binding constraints in ($\widetilde{\text{DIR}}$) in the solution $(\tilde{d}, \tilde{\sigma}^1, \ldots, \tilde{\sigma}^K)$. (Note that each $\tilde{\mathcal{P}}^k(\hat{x}_0)$ will contain at least one element.) Furthermore, if $\alpha = \alpha_0 < \infty$ in (5.20), then one of the elements in the new first-stage solution $\hat{x}_0$ has become zero, say the k-th element, and we must add the constraint $d_k \geq 0$ to ($\widetilde{\text{DIR}}$).

Alternatively, if $\alpha = \alpha_1^k$ in (5.20) for some subproblem k, then the solution to this subproblem becomes degenerate in the new first-stage solution $x_0 = \hat{x}_0$. We will show below that a dual simplex pivot in this subproblem with row r as the pivot row (r as defined in (5.19)) supplies us with either a new optimal extreme point $p \in \mathcal{P}^k(\hat{x}_0) \setminus \tilde{\mathcal{P}}^k(\hat{x}_0)$, or a new binding extreme ray $r \in \mathcal{R}^k(\hat{x}_0) \setminus \tilde{\mathcal{R}}^k(\hat{x}_0)$. The appropriate set $\tilde{\mathcal{P}}^k(\hat{x}_0)$ or $\tilde{\mathcal{R}}^k(\hat{x}_0)$ should be enlarged with the new element, and the corresponding constraint added to ($\widetilde{\text{DIR}}$).

We note that this situation of (primal) degeneracy in the subproblem is precisely what causes $\alpha_1 = 0$ in case 3 above. If this happens for at least one subproblem k, then the actual stepsize $\alpha = 0$, and we can resolve this situation in exactly the same manner by performing a dual simplex pivot in this subproblem.

Finding New Optimal Extreme Points and Binding Extreme Rays

Let $\tilde{x}_0$ denote the current first-stage solution, and $(\tilde{d}, \tilde{\sigma}^1, \ldots, \tilde{\sigma}^K)$ the solution to the approximate direction-finding problem ($\widetilde{\text{DIR}}$) for $x_0 = \tilde{x}_0$. Consider a subproblem k, and let it be represented by the formulation in (5.18). (We will omit the superscript k in what follows.) We assume that row r in this formulation of the subproblem has $\bar{b}_r = 0$ and $\bar{d}_r < 0$. That is, the current optimal basic feasible solution to the subproblem is primal degenerate, and becomes infeasible if $\tilde{x}_0$ changes to $\tilde{x}_0 + \alpha\tilde{d}$ for any positive α. If we choose row r as the pivot row for a dual simplex pivot in this subproblem, two

situations can occur:

1. $\bar{A}^N_{r.} \geq 0$:
   This implies that a dual simplex pivot is not possible in row r. The following proposition shows that the vector $(-B^{-1}_{r.})$ is then an extreme ray in the dual subproblem which is binding in $x_0 = \tilde{x}_0$, and for which the constraint in (DIR) is violated by the current solution to ($\widetilde{\text{DIR}}$). Thus $(-B^{-1}_{r.}) \notin \tilde{\mathcal{R}}(\tilde{x}_0)$, and we should add $(-B^{-1}_{r.})$ to the set $\tilde{\mathcal{R}}(\tilde{x}_0)$, add the corresponding constraint to ($\widetilde{\text{DIR}}$), and re-optimize this direction-finding problem to obtain a different proposal for a descent direction.

   Proposition 5.2 *Under the conditions stated above,* $(-B^{-1}_{r.}) \in \mathcal{R}(\tilde{x}_0)$, $-B^{-1}_{r.}F_1\tilde{d} < 0$, *and thus* $(-B^{-1}_{r.}) \notin \tilde{\mathcal{R}}(\tilde{x}_0)$.

   PROOF: The set of feasible solutions in the dual subproblem is $U \equiv \{u \mid uA_1 \leq c_1\}$. Because $B^{-1}_{r.}A^N \equiv \bar{A}^N_{r.} \geq 0$ by assumption, while $B^{-1}_{r.}B$ is a unit vector with a one in the r-th position, it follows that $(-B^{-1}_{r.})$ is a ray of the set U. That $(-B^{-1}_{r.})$ is an *extreme* ray of U follows from the fact that $(-B^{-1}_{r.})$ is an extreme point of the set $\{u \mid uA_1 \leq 0, ue = \sum_{j=1}^{m_1} \beta_{rj}\}$ (see Murty [47, sections 3.4 and 3.7]).

   Using the definition of $\bar{b} =$ in (5.17) and the fact that $\bar{b}_r = 0$, we verify that $(-B^{-1}_{r.})$ is a binding extreme ray in (SUB) when $x_0 = \tilde{x}_0$:

   $$-B^{-1}_{r.}(b_1 - F_1\tilde{x}_0) = -\bar{b}_r = 0$$

   Thus $(-B^{-1}_{r.}) \in \mathcal{R}(\tilde{x}_0)$. Finally, $(-B^{-1}_{r.}) \notin \tilde{\mathcal{R}}(\tilde{x}_0)$ because

   $$-B^{-1}_{r.}F_1\tilde{d} = \bar{d}_r < 0$$

   which violates the constraint in ($\widetilde{\text{DIR}}$).

   $\mathcal{QED}$

2. $\bar{a}^N_{rj} < 0$ for at least one nonbasic variable j:
   Let the column index s refer to the column that achieves the dual simplex minimum ratio[3]:

[3]If the minimum ratio is achieved by more than one column, an anti-cycling rule should be used to choose the index s.

$$s \equiv \arg \min_{j:\bar{a}^N_{rj}<0} \left\{ \frac{-\bar{c}^N_j}{\bar{a}^N_{rj}} \right\}. \tag{5.21}$$

Perform a dual simplex pivot on column s with pivot element $\bar{a}^N_{rs}$, and let $\hat{B}$ denote the new optimum basis. From elementary row operations in the revised simplex method it follows that the optimal dual solution $\hat{u}_1$ that corresponds to the new basis $\hat{B}$ equals:

$$\hat{u}_1 = \tilde{u}_1 - (\bar{c}^N_s / \bar{a}^N_{rs})\, B^{-1}_{r\cdot} \tag{5.22}$$

If $\bar{c}^N_s = 0$, then $\hat{u}_1 = \tilde{u}_1$ and the simplex pivot is dual degenerate. However, the updated direction with respect to the basis $\hat{B}$ is $\hat{d} \equiv -\hat{B}^{-1} F_1 \tilde{d}$, with $\hat{d}_r = (\bar{d}_r / \bar{a}^N_{rs}) > 0$. Thus, α_1 is no longer restricted by row r to a value of 0, and a new stepsize determination should be performed in the subproblem.

If $\bar{c}^N_s > 0$, and thus $\hat{u}_1 \neq \tilde{u}_1$, the next proposition states that $\hat{u}_1$ is an optimal extreme point in the dual subproblem when $x_0 = \tilde{x}_0$, and the corresponding constraint in (DIR) is violated by the current solution to ($\widetilde{\text{DIR}}$) if $-\tilde{u}_1 F_1 \tilde{d} = \tilde{\sigma}$ (i.e., the constraint in ($\widetilde{\text{DIR}}$) with respect to $\tilde{u}_1$ is binding). We then add this constraint to ($\widetilde{\text{DIR}}$) and re-optimize the direction-finding problem to find a new descent direction. We note that the condition $-\tilde{u}_1 F_1 \tilde{d} = \tilde{\sigma}$ is satisfied if $\tilde{u}_1$ was the most recent addition to the set $\tilde{\mathcal{P}}(\tilde{x}_0)$ when ($\widetilde{\text{DIR}}$) was solved, and thus will be satisfied in the course of the primal-dual method.

Proposition 5.3 *Under the conditions stated above, $\hat{u}_1 \in \mathcal{P}(\tilde{x}_0)$. Furthermore, if $-\tilde{u}_1 F_1 \tilde{d} = \tilde{\sigma}$, then $\hat{u}_1 \notin \tilde{\mathcal{P}}(\tilde{x}_0)$.*

PROOF: From the fact that $\hat{u}_1$ is defined with respect to the basis $\hat{B}$ it follows that $\hat{u}_1$ is an extreme point of the dual feasible region. Furthermore:

$$\begin{aligned}
\hat{u}_1(b_1 - F_1\tilde{x}_0) &= \tilde{u}_1(b_1 - F_1\tilde{x}_0) - (\bar{c}^N_s/\bar{a}^N_{rs})\, B^{-1}_{r\cdot}(b_1 - F_1\tilde{x}_0) \\
&= h(\tilde{x}_0) - (\bar{c}^N_s/\bar{a}^N_{rs})\bar{b}_r \\
&= h(\tilde{x}_0)
\end{aligned}$$

Thus $\hat{u}_1 \in \mathcal{P}(\tilde{x}_0)$. To prove that $\hat{u}_1 \notin \tilde{\mathcal{P}}(\tilde{x}_0)$ if $-\tilde{u}_1 F_1 \tilde{d} = \tilde{\sigma}$, we write

$$\begin{aligned}
\hat{u}_1 F_1 \tilde{d} &= \tilde{u}_1 F_1 \tilde{d} - (\bar{c}^N_s/\bar{a}^N_{rs}) B^{-1}_{r\cdot} F_1 \tilde{d} \\
&= -\tilde{\sigma} - (\bar{c}^N_s/\bar{a}^N_{rs})\bar{d}_r
\end{aligned}$$

Because $\bar{a}_{rs}^N < 0$, $\bar{d}_r < 0$ and $\bar{c}_s^N > 0$, we have $\hat{u}_1 F_1 \tilde{d} + \tilde{\sigma} < 0$. Thus, $\hat{u}_1$ cannot have been an element of $\tilde{\mathcal{P}}(\tilde{x}_0)$ when ($\widetilde{\text{DIR}}$) was solved.

$\mathcal{QED}$

If there is more than one row r in the formulation of the subproblem in (5.18) that has $\bar{b}_r = 0$ and $\bar{d}_r < 0$, then we must choose a row index for the dual simplex pivot using some anti-cycling rule.

The sets $\mathcal{P}^k(x_0)$ and $\mathcal{R}^k(x_0)$ in each subproblem k are finite for any feasible first-stage solution x_0, and the situation that $\alpha_1^k = 0$ in the stepsize determination can therefore only occur a finite number of times before we find a descent direction in ($\widetilde{\text{DIR}}$) that is feasible, and therefore allows a strictly positive stepsize. Because each of the functions $h^k(x_0)$ is piecewise linear and has a finite number of segments, the primal-dual method converges to an optimal solution of (SLP) in a finite number of iterations.

5.2 Decomposition Methods for Multistage Stochastic Linear Programs

We consider the multistage stochastic linear program as a sequential decision problem under uncertainty, where each stage corresponds to a time period, and uncertainty exists about the state of the world at the beginning of each period. It is assumed that this uncertainty can be represented by an event tree, and we will make use of the terminology and notation for event trees that was introduced in section 2.1. In this section we will describe the extension of both Benders' decomposition and primal-dual decomposition to multistage stochastic programs.

A multistage stochastic linear program can be stated mathematically as

$$\text{(MSLP)}: \quad v^* = \min \quad c_0 x_0 + \sum_{s \in \mathcal{S}_1} c_1^s x_1^s + \sum_{s \in \mathcal{S}_2} c_2^s x_2^s + \ldots + \sum_{s \in \mathcal{S}_T} c_T^s x_T^s \tag{5.23}$$

$$\begin{aligned}
\text{s.t.} \quad & A_0 x_0 && = b_0 && \\
& F_1^s x_0 + A_1^s x_1^s && = b_1^s && \forall\, s \in \mathcal{S}_1 \\
& F_2^s x_1^{s^-} + A_2^s x_2^s && = b_2^s && \forall\, s \in \mathcal{S}_2 \\
& \qquad \ddots && \vdots && \\
& F_T^s x_{T-1}^{s^-} + A_T^s x_T^s && = b_T^s && \forall\, s \in \mathcal{S}_T \\
& x_0 \geq 0,\ x_t^s \geq 0 \ \forall\, s \in \mathcal{S}_t \quad , t = 1, \ldots, T &&
\end{aligned}$$

where x_0 is a n_0-vector of first-stage decisions, and x_t^s a n_t-vector of decisions at time t if scenario s occurs (remember that a scenario at time t was defined as a *sequence of events*, or *path*, in the event tree from time 0 to time t). The right-hand-side vectors b_0 and b_t^s have length m_0 and m_t, respectively, and the constraint matrices are dimensioned accordingly. As in the two-stage stochastic program, the vector of objective coefficients c_t^s has the form $\pi_t^s q_t^s$, where π_t^s denotes the (unconditional) probability of scenario s at time t, and q_t^s is a n_t-vector. Note that the formulation of (MSLP) allows that all data are scenario dependent. We assume that this problem is feasible, and has a finite optimal solution.[4]

If the stochastic program is decomposed by stage (time period), then we obtain a separate subproblem for each scenario at the beginning of a stage. The subproblem for a scenario s at time T, given some decision vector $x_{T-1}^{s^-}$ from its predecessor scenario s^- at time $T-1$, is

$$\begin{aligned}
\text{(SUB}_T^s\text{)}: \quad h_T^s(x_{T-1}^{s^-}) = \min \quad & c_T^s x_T^s \\
\text{s.t.} \quad & A_T^s x_T^s = b_T^s - F_T^s x_{T-1}^{s^-} \\
& x_T^s \geq 0
\end{aligned} \tag{5.24}$$

At each time $t = 1, \ldots, T-1$, the subproblem for a scenario s, given some vector of decisions $x_{t-1}^{s^-}$ from its predecessor scenario, can be written as:

[4] See footnote 1.

$$
\begin{aligned}
(\mathrm{SUB}_t^s): \quad h_t^s(x_{t-1}^{s^-}) = \min \quad & c_t^s x_t^s + \sum_{s^+} h_{t+1}^{s^+}(x_t^s) \\
\text{s.t.} \quad & A_t^s x_t^s = b_t^s - F_t^s x_{t-1}^{s^-} \\
& x_t^s \geq 0
\end{aligned}
\tag{5.25}
$$

where the summation in the objective is over all successor scenarios s^+ of s. The problem (MSLP) can now be written in decomposed form as

$$
\begin{aligned}
(\mathrm{MSLP}): \quad v^* = \min \quad & c_0 x_0 + \sum_{s \in \mathcal{S}_1} h_1^s(x_0) \\
\text{s.t.} \quad & A_0 x_0 = b_0 \\
& x_0 \geq 0
\end{aligned}
\tag{5.26}
$$

Similar to what was done for two-stage stochastic programs, we will describe the value of each function $h_t^s(x_{t-1}^{s^-})$ in terms of feasibility and optimality cuts.

We first note that the subproblem for a scenario s at time $T-1$ is a two-stage stochastic linear program, and we can thus use our results from section 5.1 to reformulate this subproblem as:

$$
\begin{aligned}
&(\mathrm{SUB}_{T-1}^s): \quad h_{T-1}^s(x_{T-2}^{s^-}) = \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (5.27)\\
&\min \quad c_{T-1}^s x_{T-1}^s + \sum_{s^+} \theta_T^{s^+} \\
&\text{s.t.} \quad A_{T-1}^s x_{T-1}^s = b_{T-1}^s - F_{T-1}^s x_{T-2}^{s^-} \\
&\qquad (rF_T^{s^+})\, x_{T-1}^s \geq r b_T^{s^+} \qquad \forall\, r \in \mathcal{R}_T^{s^+},\ s^+ \in \mathcal{D}_T(s, T-1) \\
&\qquad (pF_T^{s^+})\, x_{T-1}^s + \theta_T^{s^+} \geq p b_T^{s^+} \qquad \forall\, p \in \mathcal{P}_T^{s^+},\ s^+ \in \mathcal{D}_T(s, T-1) \\
&\qquad x_{T-1}^s \geq 0
\end{aligned}
$$

where $\theta_T^{s^+}$ represents the value of $h_T^{s^+}(x_{T-1}^s)$, and the sets $\mathcal{P}_T^{s^+}$ and $\mathcal{R}_T^{s^+}$ contain all extreme points and extreme rays, respectively, of the feasible region in the dual of $(\mathrm{SUB}_T^{s^+})$.

To describe $h_{T-1}^s(x_{T-2}^{s^-})$ in terms of feasibility and optimality cuts, consider the dual of (SUB_{T-1}^s) in (5.27), where u_{T-1}^s is the dual vector for the set of constraints $A_{T-1}^s x_{T-1}^s = b_{T-1}^s - F_{T-1}^s x_{T-2}^{s^-}$, and $\gamma_T^{s^+,p}$ ($\nu_T^{s^+,r}$) the dual variable for the optimality cut (feasibility cut) that corresponds to the extreme point p (extreme ray r) in the set $\mathcal{P}_T^{s^+}$ ($\mathcal{R}_T^{s^+}$):

$$
\begin{aligned}
(\mathrm{DSUB}^s_{T-1}): \quad & h^s_{T-1}(x^{s^-}_{T-2}) = \qquad (5.28)\\
\max \quad & u^s_{T-1}\,(b^s_{T-1} - F^s_{T-1}x^{s^-}_{T-2}) + \sum_{s^+}\left(\sum_{r\in\mathcal{R}^{s^+}_T}\nu^{s^+,r}_T r + \sum_{p\in\mathcal{P}^{s^+}_T}\gamma^{s^+,p}_T p\right) b^{s^+}_T\\
\text{s.t.} \quad & u^s_{T-1}A^s_{T-1} + \sum_{s^+}\left(\sum_{r\in\mathcal{R}^{s^+}_T}\nu^{s^+,r}_T r + \sum_{p\in\mathcal{P}^{s^+}_T}\gamma^{s^+,p}_T p\right) F^{s^+}_T \le c^s_{T-1}\\
& \sum_{p\in\mathcal{P}^{s^+}_T}\gamma^{s^+,p}_T = 1 \qquad \forall\, s^+\in\mathcal{D}_T(s,T-1)\\
& \nu^{s^+,r}_T,\ \gamma^{s^+,p}_T \ge 0 \qquad \forall\, r\in\mathcal{R}^{s^+}_T,\ p\in\mathcal{P}^{s^+}_T,\ s^+\in\mathcal{D}_T(s,T-1)
\end{aligned}
$$

Because the sets $\mathcal{P}^{s^+}_T$ and $\mathcal{R}^{s^+}_T$ are independent of x^s_{T-1}, it follows that the feasible region of (DSUB^S_{T-1}) is independent of $x^{s^-}_{T-2}$.

To interpret this formulation, let $U^{s^+}_T$ denote the feasible region in the dual of the successor subproblem $(\mathrm{SUB}^{s^+}_T)$. It is well-known that any point in $U^{s^+}_T$ can be written as the sum of a convex combination of the extreme points and a nonnegative combination of the extreme rays of $U^{s^+}_T$ (see, for example, Murty [47, section 3.7]), i.e., as

$$
\sum_{r\in\mathcal{R}^{s^+}_T}\nu^{s^+,r}_T r + \sum_{p\in\mathcal{P}^{s^+}_T}\gamma^{s^+,p}_T p \quad \text{with} \sum_{p\in\mathcal{P}^{s^+}_T}\gamma^{s^+,p}_T = 1 \qquad (5.29)
$$
$$
\text{and } \nu^{s^+,r}_T, \gamma^{s^+,p}_T \ge 0 \ \forall\, r\in\mathcal{R}^{s^+}_T, p\in\mathcal{P}^{s^+}_T
$$

The formulation of (DSUB^S_{T-1}) in (5.28) has therefore in effect replaced each dual vector $u^{s^+}_T$ and its associated constraints $(u^{s^+}_T A^{s^+}_T \le c^{s^+}_T)$ by this equivalent description.

We will denote the feasible region of (DSUB^S_{T-1}) in (5.28) as U^s_{T-1}, its set of extreme points as $\mathcal{P}^s_{T-1}$ and the set of all extreme rays as $\mathcal{R}^s_{T-1}$. Each of these extreme points and extreme rays consists of a value for the vector u^s_{T-1} as well as values for $\nu^{s^+,r}_T$ and $\gamma^{s^+,p}_T$ for all $r\in\mathcal{R}^{s^+}_T$, $p\in\mathcal{P}^{s^+}_T$ and every $s^+\in\mathcal{D}_T(s,T-1)$. The collection of these values will be written concisely as $(u^s_{T-1}, \{\nu^{s^+,r}_T\}, \{\gamma^{s^+,p}_T\})$.

To formulate the feasibility and optimality cuts that follow from the formulation of (DSUB^S_{T-1}) in (5.28), we simplify notation and define

$$
\begin{aligned}
&\text{For } r \in \mathcal{R}_T^s: \quad G_T^{s,r} \equiv rF_T^s \quad \text{and} \quad g_T^{s,r} \equiv rb_T^s \\
&\text{For } p \in \mathcal{P}_T^s: \quad E_T^{s,p} \equiv pF_T^s \quad \text{and} \quad e_T^{s,p} \equiv pb_T^s \\
&\text{For } r \in \mathcal{R}_{T-1}^s: G_{T-1}^{s,r} \equiv \bar{u}_{T-1}^s F_{T-1}^s \quad \text{and} \\
&\qquad g_{T-1}^{s,r} \equiv \bar{u}_{T-1}^s b_{T-1}^s + \sum_{s^+} \left(\sum_{r' \in \mathcal{R}_T^{s^+}} \bar{\nu}_T^{s^+,r'} g_T^{s^+,r'} + \sum_{p' \in \mathcal{P}_T^{s^+}} \bar{\gamma}_T^{s^+,p'} e_T^{s^+,p'} \right) \\
&\qquad \text{where } r = (\bar{u}_{T-1}^s, \{\bar{\nu}_T^{s^+,r'}\}, \{\bar{\gamma}_T^{s^+,p'}\}) \\
&\text{For } p \in \mathcal{P}_{T-1}^s: E_{T-1}^{s,p} \equiv \hat{u}_{T-1}^s F_{T-1}^s \quad \text{and} \\
&\qquad e_{T-1}^{s,p} \equiv \hat{u}_{T-1}^s b_{T-1}^s + \sum_{s^+} \left(\sum_{r' \in \mathcal{R}_T^{s^+}} \hat{\nu}_T^{s^+,r'} g_T^{s^+,r'} + \sum_{p' \in \mathcal{P}_T^{s^+}} \hat{\gamma}_T^{s^+,p'} e_T^{s^+,p'} \right) \\
&\qquad \text{where } p = (\hat{u}_{T-1}^s, \{\hat{\nu}_T^{s^+,r'}\}, \{\hat{\gamma}_T^{s^+,p'}\})
\end{aligned}
\tag{5.30}
$$

Now the feasibility cuts for $x_{T-2}^{s^-}$ can be written as:

$$
G_{T-1}^{s,r} x_{T-2}^{s^-} \geq g_{T-1}^{s,r} \qquad \forall\ r \in \mathcal{R}_{T-1}^s \tag{5.31}
$$

and the optimality cuts as

$$
E_{T-1}^{s,p} x_{T-2}^{s^-} + \theta_{T-1}^s \geq e_{T-1}^{s,p} \qquad \forall\ p \in \mathcal{P}_{T-1}^s \tag{5.32}
$$

where θ_{T-1}^s represents the value of the function $h_{T-1}^s(x_{T-2}^{s^-})$. These optimality cuts are derived from the fact that $-E_{T-1}^{s,p} = -\hat{u}_{T-1}^s F_{T-1}^s$ is a subgradient of $h_{T-1}^s(x_{T-2}^{s^-})$ in $x_{T-2}^{s^-} = \tilde{x}_{T-2}^{s^-}$ if $p = (\hat{u}_{T-1}^s, \{\hat{\nu}_T^{s^+,r'}\}, \{\hat{\gamma}_T^{s^+,p'}\}) \in \mathcal{P}_{T-1}^s$ is an optimal dual solution for $h_{T-1}^s(\tilde{x}_{T-2}^{s^-})$.

By adding these cuts to $(\text{SUB}_{T-2}^{s^-})$ for each successor subproblem (SUB_{T-1}^s) of s^-, and replacing $h_{T-1}^s(x_{T-2}^{s^-})$ by θ_{T-1}^s, we obtain a reformulation of $(\text{SUB}_{T-2}^{s^-})$ that is similar in form to (5.27). In general, by repeating the procedure recursively for subproblems at times $T-2, \ldots, 1$, we obtain the following formulation for the subproblem of a scenario s at time t $(t = 1, \ldots, T-1)$:

$$
\begin{aligned}
&(\text{SUB}_t^s): \quad h_t^s(x_{t-1}^{s^-}) = \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (5.33) \\
&\quad \min \quad c_t^s x_t^s + \sum_{s^+} \theta_{t+1}^{s^+} \\
&\quad \text{s.t.} \quad A_t^s x_t^s = b_t^s - F_t^s x_{t-1}^{s^-} \\
&\qquad\quad G_{t+1}^{s^+,r} x_t^s \geq g_{t+1}^{s^+,r} \qquad \forall\, r \in \mathcal{R}_{t+1}^{s^+},\ s^+ \in \mathcal{D}_{t+1}(s,t) \\
&\qquad\quad E_{t+1}^{s^+,p} x_t^s + \theta_{t+1}^{s^+} \geq e_{t+1}^{s^+,p} \qquad \forall\, p \in \mathcal{P}_{t+1}^{s^+},\ s^+ \in \mathcal{D}_{t+1}(s,t) \\
&\qquad\quad x_t^s \geq 0
\end{aligned}
$$

while the the full problem (MSLP) can be written as

$$
\begin{aligned}
(\text{MSLP}): \quad v^* = \min \quad & c_0 x_0 + \sum_{s \in \mathcal{S}_1} \theta_1^s \\
\text{s.t.} \quad & A_0 x_0 = b_0 \\
& G_1^{s,r} x_0 \geq g_1^{s,r} \qquad \forall\, r \in \mathcal{R}_1^s, \;\; s \in \mathcal{S}_1 \\
& E_1^{s,p} x_0 + \theta_1^s \geq e_1^{s,p} \qquad \forall\, p \in \mathcal{P}_1^s, \;\; s \in \mathcal{S}_1 \\
& x_0 \geq 0
\end{aligned}
\tag{5.34}
$$

The matrices $G_t^{s,r}$ and $E_t^{s,p}$ and the constants $g_t^{s,r}$ and $e_t^{s,p}$ were defined in (5.30) for $t = T$ and $t = T-1$. Their definitions for $t < T-1$ are analogous to the definitions for $t = T-1$, and are given below for completeness:

$$
\begin{aligned}
&\text{For } r \in \mathcal{R}_t^s : G_t^{s,r} \equiv \bar{u}_t^s F_t^s \qquad \text{and} \\
&\qquad g_t^{s,r} \equiv \bar{u}_t^s b_t^s + \sum_{s^+} \left(\sum_{r' \in \mathcal{R}_{t+1}^{s^+}} \bar{\nu}_{t+1}^{s^+,r'} g_{t+1}^{s^+,r'} + \sum_{p' \in \mathcal{P}_{t+1}^{s^+}} \bar{\gamma}_{t+1}^{s^+,p'} e_{t+1}^{s^+,p'} \right) \\
&\qquad \text{where } r = (\bar{u}_t^s, \{\bar{\nu}_{t+1}^{s^+,r'}\}, \{\bar{\gamma}_{t+1}^{s^+,p'}\}) \\
&\text{For } p \in \mathcal{P}_t^s : E_t^{s,p} \equiv \hat{u}_t^s F_t^s \qquad \text{and} \\
&\qquad e_t^{s,p} \equiv \hat{u}_t^s b_t^s + \sum_{s^+} \left(\sum_{r' \in \mathcal{R}_{t+1}^{s^+}} \hat{\nu}_{t+1}^{s^+,r'} g_{t+1}^{s^+,r'} + \sum_{p' \in \mathcal{P}_{t+1}^{s^+}} \hat{\gamma}_{t+1}^{s^+,p'} e_{t+1}^{s^+,p'} \right) \\
&\qquad \text{where } p = (\hat{u}_t^s, \{\hat{\nu}_{t+1}^{s^+,r'}\}, \{\hat{\gamma}_{t+1}^{s^+,p'}\})
\end{aligned}
\tag{5.35}
$$

It should be clear that the cuts in the subproblem for a scenario s at time $t > 0$ reflect all necessary information from its descendant subproblems at times $t+1, \ldots, T$. By extension, the cuts in the formulation of (MSLP) in (5.34) represent all necessary information in the original formulation from time 1 onwards.

Both Benders' decomposition and primal-dual decomposition are based on this reformulation of (MSLP), and each of them iteratively generates cuts from the solutions to (approximate) formulations of the subproblems. One should note, however, that the cuts which were described above assumed a *complete* knowledge of the set of extreme points and extreme rays of the feasible region in the dual subproblems. Specifically, the cuts that were derived from the subproblem of a scenario s at some time $t < T$ assumed knowledge of the complete sets $\mathcal{P}_{t+1}^{s^+}$ and $\mathcal{R}_{t+1}^{s^+}$ for all successors s^+ of scenario s. It is clear that this knowledge will generally not be available. In fact, the whole purpose of the decomposition methods is to gather this information in the course of the algorithm, and only to the extent needed. We will show below

that cuts which are derived from the subproblem of a scenario s at time $t = T-1$ are still valid, although weaker, if they are based on only partial knowledge of the sets of extreme points $\mathcal{P}_T^{s^+}$ and extreme rays $\mathcal{R}_T^{s^+}$ for all its successors s^+. The argument is analogous when $t < T-1$.

Consider the formulation of (DSUB_{T-1}^s) in (5.28), and suppose that only subsets $\tilde{\mathcal{P}}_T^{s^+}$ and $\tilde{\mathcal{R}}_T^{s^+}$ of $\mathcal{P}_T^{s^+}$ and $\mathcal{R}_T^{s^+}$ are known for each successor scenario s^+ of s, where each $\tilde{\mathcal{P}}_T^{s^+}$ is assumed to contain at least one element. We will refer to this approximate formulation as $(\widetilde{\text{DSUB}}_{T-1}^s)$, and to the corresponding optimal objective value as $\tilde{h}_{T-1}^s(\hat{x}_{T-2}^{s^-})$, with $\hat{x}_{T-2}^{s^-}$ a given vector. It is obvious from (5.28) that partial knowledge of the sets $\mathcal{P}_T^{s^+}$ and $\mathcal{R}_T^{s^+}$ restricts the feasible region, and thus $h_{T-1}^s(x_{T-2}^{s^-}) \geq \tilde{h}_{T-1}^s(x_{T-2}^{s^-})$ for any vector $x_{T-2}^{s^-}$. We note, however, that every extreme point (ray) of the feasible region in $(\widetilde{\text{DSUB}}_{T-1}^s)$ when $\tilde{\mathcal{P}}_T^{s^+} \neq \emptyset$ is also an extreme point (ray) of the feasible region in (DSUB_{T-1}^s); this follows directly from the characterization of extreme points in Murty [47, section 3.4].

Assume first that $(\widetilde{\text{DSUB}}_{T-1}^s)$ is bounded, and let $\hat{p} = (\hat{u}_{T-1}^s, \{\hat{\nu}_T^{s^+,r'}\}, \{\hat{\gamma}_T^{s^+,p'}\})$ be the optimal solution. Then $-\hat{u}_{T-1}^s F_{T-1}^s$ is a subgradient of $\tilde{h}_{T-1}^s(x_{T-2}^{s^-})$ in $x_{T-2}^{s^-} = \hat{x}_{T-2}^{s^-}$, and thus for all $x_{T-2}^{s^-}$

$$\tilde{h}_{T-1}^s(x_{T-2}^{s^-}) \geq \tilde{e}_{T-1}^{s,\hat{p}} - \tilde{E}_{T-1}^{s,\hat{p}} x_{T-2}^{s^-}$$

where

$$\tilde{E}_{T-1}^{s,\hat{p}} \equiv \hat{u}_{T-1}^s F_{T-1}^s \qquad \text{and}$$

$$\tilde{e}_{T-1}^{s,\hat{p}} \equiv \hat{u}_{T-1}^s b_{T-1}^s + \sum_{s^+} \left(\sum_{r' \in \tilde{\mathcal{R}}_T^{s^+}} \hat{\nu}_T^{s^+,r'} g_T^{s^+,r'} + \sum_{p' \in \tilde{\mathcal{P}}_T^{s^+}} \hat{\gamma}_T^{s^+,p'} e_T^{s^+,p'} \right)$$

Because $h_{T-1}^s(x_{T-2}^{s^-}) \geq \tilde{h}_{T-1}^s(x_{T-2}^{s^-})$, it follows that

$$\tilde{E}_{T-1}^{s,\hat{p}} x_{T-2}^{s^-} + \theta_{T-1}^s \geq \tilde{e}_{T-1}^{s,\hat{p}}$$

is a valid optimality cut in $(\text{SUB}_{T-2}^{s^-})$.

If $(\widetilde{\text{DSUB}}_{T-1}^s)$ is unbounded, it must have a ray $\bar{r} = (\bar{u}_{T-1}^s, \{\bar{\nu}_T^{s^+,r'}\}, \{\bar{\gamma}_T^{s^+,p'}\})$ such that

$$\tilde{g}_{T-1}^{s,\bar{r}} - \tilde{G}_{T-1}^{s,\bar{r}} x_{T-2}^{s^-} > 0$$

where

$$\tilde{G}_{T-1}^{s,\bar{r}} \equiv \bar{u}_{T-1}^s F_{T-1}^s \qquad \text{and}$$

$$\tilde{g}_{T-1}^{s,\bar{r}} \equiv \bar{u}_{T-1}^{s} b_{T-1}^{s} + \sum_{s^+} \left(\sum_{r' \in \tilde{\mathcal{R}}_T^{s^+}} \bar{\nu}_T^{s^+,r'} g_T^{s^+,r'} + \sum_{p' \in \tilde{\mathcal{P}}_T^{s^+}} \bar{\gamma}_T^{s^+,p'} e_T^{s^+,p'} \right)$$

By setting $\nu_T^{s^+,r'} = 0$ for all $r' \in \mathcal{R}_T^{s^+} \setminus \tilde{\mathcal{R}}_T^{s^+}$ and $\gamma_T^{s^+,p'} = 0$ for all $p' \in \mathcal{P}_T^{s^+} \setminus \tilde{\mathcal{P}}_T^{s^+}$, it is clear that $\bar{r}$ also defines a ray in (DSUB_{T-1}^{s}) that leads to an unbounded solution when $x_{T-2}^{s^-} = \hat{x}_{T-2}^{s^-}$. Thus

$$\tilde{G}_{T-1}^{s,\bar{r}} x_{T-2}^{s^-} \geq \tilde{g}_{T-1}^{s,\bar{r}}$$

is a valid feasibility cut for $(\text{SUB}_{T-2}^{s^-})$.

A third possibility is that $(\widetilde{\text{DSUB}}_{T-1}^{s})$ is infeasible. We note, however, that (DSUB_{T-1}^{s}) is feasible because of the assumption that (MSLP) is bounded. The infeasibility in $(\widetilde{\text{DSUB}}_{T-1}^{s})$ must therefore be due to the fact that only the partial sets $\tilde{\mathcal{P}}_T^{s^+}$ and $\tilde{\mathcal{R}}_T^{s^+}$ are included in its formulation. More elements of $\mathcal{P}_T^{s^+}$ and $\mathcal{R}_T^{s^+}$ should then be generated to make $(\widetilde{\text{DSUB}}_{T-1}^{s})$ feasible.

An analogous argument establishes the validity of the cuts that are generated from subproblems at times $t < T-1$ when only subsets of all extreme points and rays are known of the dual subproblems at time $t+1$.

5.2.1 Nested Benders' Decomposition

In nested Benders' decomposition, information between subproblems is passed forward as well as backward in time. On the one hand, the subproblem for a scenario s at time t $(1 \leq t \leq T-1)$ serves as a master problem for its successor subproblems at time $t+1$, to which it passes values of the decision vector x_t^s. On the other hand, it is used to generate feasibility and optimality cuts that are added to the (sub)problem of its predecessor scenario at time $t-1$. Although different strategies are possible for the order in which subproblems are solved and information is exchanged, we will describe the so-called *fast-forward-fast-backward* method. This strategy was found superior among the strategies that Gassmann [18] tested on several multistage stochastic linear programming models.

In the fast-forward-fast-backward method, one makes alternatingly a *forward pass* and a *backward pass* through time. We will briefly describe each of these passes, and assume that in each subproblem only a subset of the complete set of cuts is known. If a subproblem is solved that contains no optimality cuts from one or more of its successor subproblems (this can only be the case during the first forward pass), then the corresponding variables θ are assigned a value of zero.

A Forward Pass

In a forward pass, the aim is to construct a feasible solution to (MSLP) by passing decision vectors from subproblems to successor subproblems. At the start of a forward pass, the master problem at time 0 is solved, and the optimal solution $\tilde{x}_0$ is passed to the subproblems at time 1. If any of these subproblems is infeasible, a feasibility cut is found that is violated by $\tilde{x}_0$, this cuts is added to the time-0 master problem, and a new solution for x_0 must be found. Alternatively, if all subproblems at time 1 are feasible, then their optimal solutions $\tilde{x}_1^s$ are passed to the successor subproblems at time 2. Each of these subproblems at time 2 is solved, and if any of them is infeasible, say ($\text{SUB}_2^{s^+}$), then a feasibility cut is added to its predecessor subproblem (SUB_1^s) at time 1. If (SUB_1^s) becomes infeasible with the additional cut and for the current value of $x_0 = \tilde{x}_0$, then a new feasibility cut is added to the master problem at time 0, and a new solution for x_0 must be obtained. Otherwise, the new solution $\tilde{x}_1^s$ to (SUB_1^s) is passed to *all* its successor subproblems, and these subproblems are re-optimized.

In general, the subproblem (SUB_t^s) of a scenario s at some time $t = 1, \ldots, T-1$ is solved in the course of a forward pass when either a solution $\tilde{x}_{t-1}^{s^-}$ is passed from its predecessor subproblem, or a feasibility cut is added from one of its successor subproblems. If (SUB_t^s) has an optimal solution $\tilde{x}_t^s$, then this solution is passed to each of its successor subproblems, and these successor subproblems are solved next. Otherwise, (SUB_t^s) is infeasible, and a feasibility cut is added to its predecessor subproblem, which needs to be re-optimized.

A forward pass ends when a solution has been obtained to all subproblems at time T. At that point, $\tilde{x}_0$, the solution to the master problem at time 0, and all subproblem solutions $\tilde{x}_t^s$ constitute a feasible solution to (MSLP). The corresponding objective value is thus an *upper bound* on v^*.

A Backward Pass

In a backward pass, an optimality cut is generated from each subproblem and added to its predecessor subproblem. First, for each subproblem at time T an optimality cut is constructed from the solution that was found at the end of a forward pass, and these cuts are added to the formulation of the respective predecessor subproblems at time $T-1$. (Note that the added optimality cuts can never cause infeasibility of these subproblems, as the variables θ_T^s are unrestricted in sign.) A new solution is then determined for each of these subproblems at time $T-1$, optimality cuts

are constructed from these solutions, and added to the predecessor subproblems at time $T-2$.

This procedure continues, moving backward in time, until new optimality cuts have been added to the master problem at time 0. The solution value of the master problem after these cuts have been added provides a new *lower bound* on the optimum value v^* of (MSLP). At that point, a new forward pass can be started.

One obtains a better lower bound on v^* after each backward pass, and a feasible solution in each forward pass. The decomposition algorithm can be terminated if the solution value of the best feasible solution that has been found is deemed close enough to the last lower bound. Obviously, if upper and lower bound are equal to each other, an optimal solution has been found.

5.2.2 Primal-Dual Decomposition

The extension of primal-dual decomposition from two-stage to multistage stochastic programs is not possible in the same way as was discussed for Benders' decomposition. That is, it is not possible to gradually increase the sets of extreme points and extreme rays of each dual subproblem by making multiple forward and backward passes. This is due to the fact that one cannot ensure the feasibility of a descent direction for the decision vector x_t^s at some time $t \leq T-2$ without knowing the complete sets of extreme points and extreme rays in the dual subproblems at time $t+2$.

To see this, suppose we solve the approximate direction-finding problem for a scenario s at time $T-2$, where $\tilde{x}_{T-2}^s$ is the current value of x_{T-2}^s:

$$
\begin{aligned}
&(\widetilde{\mathrm{DIR}}_{T-2}^s): \quad \tilde{\delta} = \qquad\qquad (5.36)\\
&\min \; c_{T-2}^s\, d + \sum_{s^+} \sigma_{T-1}^{s^+}\\
&\text{s.t.} \quad A_{T-2}^s\, d \qquad\qquad = 0\\
&\qquad G_{T-1}^{s^+,r}\, d \qquad\qquad \geq 0 \quad \forall\, r \in \tilde{\mathcal{R}}_{T-1}^{s^+}(\tilde{x}_{T-2}^s),\;\; s^+ \in \mathcal{D}_{T-1}(s, T-2)\\
&\qquad E_{T-1}^{s^+,p}\, d + \sigma_{T-1}^{s^+} \quad \geq 0 \quad \forall\, p \in \tilde{\mathcal{P}}_{T-1}^{s^+}(\tilde{x}_{T-2}^s),\;\; s^+ \in \mathcal{D}_{T-1}(s, T-2)\\
&\qquad d_i \geq 0 \qquad\qquad \text{if } \tilde{x}_{T-2,i}^s = 0\\
&\qquad -e \leq d \leq e
\end{aligned}
$$

The sets $\tilde{\mathcal{P}}_{T-1}^{s^+}(\tilde{x}_{T-2}^s)$ and $\tilde{\mathcal{R}}_{T-1}^{s^+}(\tilde{x}_{T-2}^s)$ are subsets of the sets of optimal extreme points $\mathcal{P}_{T-1}^{s^+}(\tilde{x}_{T-2}^s)$ and binding extreme rays $\mathcal{R}_{T-1}^{s^+}(\tilde{x}_{T-2}^s)$, respectively, in the dual

subproblem of successor scenario s^+ at time $T-1$ when $x^s_{T-2} = \tilde{x}^s_{T-2}$. Let $\tilde{d}$ denote the optimal solution for d in $(\widetilde{\mathrm{DIR}}^s_{T-2})$, and $\tilde{\sigma}^{s^+}_{T-1}$ the optimal value for $\sigma^{s^+}_{T-1}$. To check the feasibility of $\tilde{d}$ as a descent direction, we could solve the feasibility problem for each succesor scenario s^+ of s:

$$
\begin{aligned}
f^{s^+}_{T-1}(\tilde{d}) = \quad & \max \; -u^{s^+}_{T-1} F^{s^+}_{T-1} \tilde{d} && (5.37)\\
\text{s.t. } & u^{s^+}_{T-1} A^{s^+}_{T-1} + \sum_{s' \in \mathcal{D}_T(s^+, T-1)} \left(\sum_{r \in \mathcal{R}^{s'}_T} \nu^{s',r}_T g^{s',r}_T + \sum_{p \in \mathcal{P}^{s'}_T} \gamma^{s',p}_T e^{s',p}_T \right) \le c^{s^+}_{T-1} \\
& \textstyle\sum_{p \in \mathcal{P}^{s'}_T} \gamma^{s',p}_T = 1 \qquad \forall\, s' \in \mathcal{D}_T(s^+, T-1) \\
& u^{s^+}_{T-1} (b^{s^+}_{T-1} - F^{s^+}_{T-1} \tilde{x}^s_{T-2}) \\
& \qquad + \sum_{s' \in \mathcal{D}_T(s^+, T-1)} \left(\sum_{r \in \mathcal{R}^{s'}_T} \nu^{s',r}_T g^{s',r}_T + \sum_{p \in \mathcal{P}^{s'}_T} \gamma^{s',p}_T e^{s',p}_T \right) = h^{s^+}_{T-1}(\tilde{x}^s_{T-2}) \\
& \nu^{s',r}_T,\ \gamma^{s',p}_T \ge 0 \qquad \forall\, r \in \mathcal{R}^{s'}_T,\ p \in \mathcal{P}^{s'}_T,\ s' \in \mathcal{D}_T(s^+, T-1)
\end{aligned}
$$

The constraints in this problem are the constraints of the dual subproblem $(\mathrm{DSUB}^{s^+}_{T-1})$ plus an additional constraint that says that attention must be limited to feasible solutions which are optimal for $x^s_{T-2} = \tilde{x}^s_{T-2}$.

As in the two-stage case, if $f^{s^+}_{T-1}(\tilde{d}) > \tilde{\sigma}^{s^+}_{T-1}$, then $\tilde{d}$ is not a feasible direction, and we either find a new binding extreme ray (if $f^{s^+}_{T-1}(\tilde{d}) = \infty$) or optimal extreme point (if $\tilde{\sigma}^{s^+}_{T-1} < f^{s^+}_{T-1}(\tilde{d}) < \infty$) for which we should add a constraint to $(\widetilde{\mathrm{DIR}}^s_{T-2})$. If $f^{s^+}_{T-1}(\tilde{d}) = \tilde{\sigma}^{s^+}_{T-1}$, then $\tilde{d}$ is a feasible direction from $\tilde{x}^s_{T-2}$ as far as the subproblem for successor scenario s^+ is concerned.

Now suppose only a subset $\tilde{\mathcal{P}}^{s'}_T \neq \emptyset$ of $\mathcal{P}^{s'}_T$ and a subset $\tilde{\mathcal{R}}^{s'}_T$ of $\mathcal{R}^{s'}_T$ is known for each scenario $s' \in \mathcal{D}_T(s^+, T-1)$ in the formulation of the feasibility problem. This *approximate* feasibility problem is thus a restriction of the true feasibility problem, and its objective value $\tilde{f}^{s^+}_{T-1}(\tilde{d}) \le f^{s^+}_{T-1}(\tilde{d})$. If $\tilde{f}^{s^+}_{T-1}(\tilde{d}) = \tilde{\sigma}^{s^+}_{T-1}$ while $f^{s^+}_{T-1}(\tilde{d}) > \tilde{\sigma}^{s^+}_{T-1}$, then the approximate feasibility problem suggests that $\tilde{d}$ is a feasible direction, while in reality it is not. If a step $\alpha > 0$ would therefore be taken from $\tilde{x}^s_{T-2}$ in the direction $\tilde{d}$, the new solution $\tilde{x}^s_{T-2} + \alpha\tilde{d}$ and $\tilde{\theta}^{s^+}_{T-1} + \alpha\tilde{\sigma}^{s^+}_{T-1}$ for all $s^+ \in \mathcal{D}_{T-1}(s, T-2)$ is not feasible in (SUB^s_{T-2}). As it is impractical to collect all extreme points and extreme rays of the dual subproblems, we therefore need a different decomposition approach.

An Alternative Decomposition Approach

We can apply the primal-dual decomposition method to multistage stochastic linear programs by essentially solving a sequence of two-stage programs. Suppose a feasible solution for (MSLP) is known, which we will denote as $\tilde{x}_0$ and $\tilde{x}_t^s$ for all $s \in \mathcal{S}_t$ and $t = 1, \ldots, T$, or concisely as $(\tilde{x}_0, \{\tilde{x}_t^s\})$. Given the solution $\tilde{x}_{T-2}^s$ for each $s \in \mathcal{S}_{T-2}$, we can solve the subproblems (SUB_{T-1}^s) for all $s \in \mathcal{S}_{T-1}$ using the primal-dual decomposition method for two-stage stochastic programs that was described in section 5.1.2. Once the optimal solution to (SUB_{T-1}^s) for each $s \in \mathcal{S}_{T-1}$ is found, we move backward in time one period, and solve (SUB_{T-2}^s) for each $s \in \mathcal{S}_{T-2}$, given the solution $\tilde{x}_{T-3}^{s^-}$ from its predecessor subproblem $(\text{SUB}_{T-3}^{s^-})$. This optimization can again be performed by the primal-dual decomposition method for two-stage stochastic programs, where the second-stage problems are the subproblems at time $T-1$. This procedure is continued, moving backwards in time, until the problem at time 0 has been solved. At that point, we have an optimal solution to (MSLP).

One point in this approach needs elaboration. Suppose we have solved the subproblem for a scenario s at time $T-1$ by the primal-dual decomposition method for two-stage stochastic programs as described in section 5.1.2. If parametric linear programming was used to determine the maximum stepsize in a descent direction from x_{T-1}^s, then we know an optimal extreme-point solution to $(\text{SUB}_T^{s^+})$ for each successor scenario s^+ of s in each iteration of the algorithm, and thus also when the optimum value of x_{T-1}^s is found. Let $\hat{x}_{T-1}^s$ denote this optimum value, and $\hat{x}_T^{s^+}$ the optimum extreme-point solution to $(\text{SUB}_T^{s^+})$ for each successor s^+. Thus, $(\hat{x}_{T-1}^s, \{\hat{x}_T^{s^+}\})$ constitutes an optimal solution to (SUB_{T-1}^s), but it is not necessarily an optimal *extreme-point* solution. Before we can apply the primal-dual decomposition to solve its predecessor subproblem at time $T-2$, however, we need such an extreme point solution in order to perform the parametric linear programming. We thus have to convert the optimal solution $(\hat{x}_{T-1}^s, \{\hat{x}_T^{s^+}\})$ to an extreme point solution before we can proceed. Obviously, this will be not only be the case if we move backward from time $T-1$ to time $T-2$, but for any step backward in time in the algorithm.

A drawback of this decomposition approach is that all subproblems have to be solved to optimality before one can move a stage backward. This is necessary to determine whether a descent direction is feasible, and to make sure that a positive stepsize maintains feasibility in all subproblems. These properties in turn are necessary to prove finite convergence of the method.

It is clear that the subproblems increase in size when we move backward in time.

The computational effort that is involved in performing the parametric programming pivots in the subproblems thus increases when we approach time 0. We note, however, that the direction-finding problems remain approximately the same in size at any stage, as their size is determined only by the *number* of successor subproblems.

Trajectory Optimization

Grinold [19] describes an alternative approach for solving multistage *deterministic* linear problems by primal-dual decomposition, which is easily generalized to multistage stochastic linear programs. He essentially transforms the multistage problem into a two-stage problem, and solves it by the primal-dual decomposition method of section 5.1.2.

For each scenario s at every time $t = 1, \ldots, T$, a new vector z_t^s of length m_t is defined by $z_t^s = F_t^s x_{t-1}^{s^-}$, where s^- is the predecessor scenario of scenario s. The collection of all these vectors z_t^s is called a *trajectory*, and will be denoted by z. Given a value $\tilde{z}$ for the trajectory, the multistage stochastic program decomposes into separate subproblems, one for each scenario s at every time $t = 0, \ldots, T$:

$$
\begin{aligned}
h_t^s(\tilde{z}) = \min \quad & c_t^s x_t^s \\
\text{s.t.} \quad & A_t^s x_t^s = b_t^s - \tilde{z}_t^s \\
& F_{t+1}^{s^+} x_t^s = \tilde{z}_{t+1}^{s^+} \qquad \forall\, s^+ \in \mathcal{D}_{t+1}(s,t) \\
& x_t^s \geq 0
\end{aligned}
$$

where the subproblems at time 0 and time T are obvious specializations of this formulation. After a solution has been obtained for each of these subproblems, a direction-finding problem is solved to find a descent direction for the trajectory $\tilde{z}$, or conclude that it is an optimal trajectory. This direction-finding problem includes cuts from all the subproblems. Grinold shows formally that the combined solutions to the subproblems constitute an optimal solution to (MSLP) if $\tilde{z}$ is optimal.

In terms of a two-stage stochastic program, the variables in the trajectory z form the first-stage variables, and the decision vectors x_t^s the second stage-variables. We note that the direction-finding problem may be quite large as it includes a variable for each variable in the trajectory, and cuts from all subproblems. However, if a subproblem has a unique dual optimum solution given the current trajectory, implying only one optimality cut in the direction-finding problem, then the direction-finding problem can be split into disjoint problems.

5.3 Re-Optimizations in the Iterative Disaggregation Algorithm

It is straightforward to write the ALM model of section 2.2.1 in matrix notation such that it has the form of (MSLP) in (5.23). To convert all constraints in the ALM model to equality constraints, we need to introduce a slack variable w_t^s in the borrowing constraint for each scenario s at every time t. All variables in the ALM model were assumed to be nonnegative, which agrees with the assumption in (MSLP). The decision vector x_t^s in the formulation of (MSLP) corresponds to the set of all variables for scenario s at time t in the ALM model:

$$x_t^s \equiv (xs_t^s, xb_t^s, xh_t^s, y_t^s, z_t^s, w_t^s)$$

In the relaxation of the ALM model which is being solved during the iterative disaggregation algorithm (see section 4.3), the decision vector x_t^s also includes the extra borrowing variable v_t^s.

The definition of the vectors of objective coefficients c_t^s, the constraint matrices A_t^s and F_t^s, and the right-hand-side vectors b_t^s for the ALM model follows directly from this definition of the decision vectors x_t^s. In the ALM model, the matrices A_t^s and F_t^s and the vectors b_t^s were assumed to depend on the states in the event tree, but to be the same for all scenarios that correspond to the same state. Thus, if there are multiple scenarios in a state, which is the case if the event tree has a lattice structure, then the ALM model is a special case of (MSLP).

As the ALM model has the form of (MSLP) in (5.23), we can in principle use any of the decomposition methods that have been described in this chapter to optimize the ALM model. If no information about the solution itself is available, then Benders' decomposition is probably the most efficient method. Computational experience has indicated that for many problems Benders' decomposition finds a "good" approximate solution fairly quickly, although convergence to the optimum solution often slows down considerably in the course of the algorithm. A potential problem with Benders' decomposition is that degeneracy in the dual subproblems can lead to the generation of many redundant cuts (see Birge [3]).

The primal-dual method is a more rigid method as it maintains feasibility of the solution in each iteration. This requires more work per iteration of the method, and limits the changes in the solution between iterations. A starting solution that is far from the optimum may therefore require a prohibitively large number of primal-dual

steps before the solution approaches the optimum. Furthermore, as a local-search method, the primal-dual method is not able to provide bounds on the proximity to the optimum.

We will show in section 5.3.1, however, that Benders' decomposition is not an efficient method to re-optimize the ALM model in each iteration of the iterative disaggregation algorithm. In particular, we will show that many cuts which are generated during the solution of an aggregated version of the ALM model can no longer be used after a disaggregation is performed. As a consequence, Benders' decomposition has to discard much of the information about the solution that was obtained in previous iterations.

The primal-dual decomposition method, in contrast, has a much better ability to use this information. We will elaborate on this in section 5.3.2.

5.3.1 Benders' Decomposition

Suppose that an aggregated ALM model has been optimized by Benders' decomposition, and that a state or time disaggregation is performed in a node of the aggregated event tree. We will show below that this disaggregation will destroy the validity of the feasibility and optimality cuts in the scenario subproblems at that node. These cuts can therefore not be used in the subsequent re-optimization of the ALM model. Furthermore, as these cuts have been used in the construction of the optimality and feasibility cuts in the subproblems of all ascendant scenarios as well, the disaggregation also destroys the validity of the cuts in these subproblems. A fortiori, none of the cuts in the master problem at time 0 remains necessarily a valid cut after a state or time disaggregation. Thus, the disaggregation necessitates the generation of a completely new set of cuts in each of these subproblems as well as the time-0 master problem.

State Disaggregation

Consider the state disaggregation that is the reversal of the basic state aggregation in figure 4-1. Using the notation of the current chapter, the subproblem for scenario s^* in state (n, k_1) at time t_j before the disaggregation (i.e., corresponding to figure 4-1(b)) can be written as:

$$(\widehat{\mathrm{SUB}}_{t_j}^{s^*}): \quad \hat{h}_{t_j}^{s^*}(x_{t_{j-1}}^{s^-}) = \min \quad c_{t_j}^{s^*} x_{t_j}^{s^*} + \hat{h}_{t_{j+1}}^{s}(x_{t_j}^{s^*}) \tag{5.38}$$
$$\text{s.t.} \quad A_{t_j}^{(n,k_0)} x_{t_j}^{s^*} = b_{t_j}^{(n,k_0)} - F_{t_j}^{(n,k_0)} x_{t_{j-1}}^{s^-}$$
$$x_{t_j}^{s^*} \geq 0$$

In this formulation, scenario s^- at time t_{j-1} represents the predecessor scenario of scenario s^*, and $x_{t_{j-1}}^{s^-}$ is the associated decision vector. Scenario s at time t_{j+1} denotes the single successor scenario of scenario s^* before the state disaggregation, corresponding to node (n, k_2) in figure 4-1(b). The objective value function $\hat{h}_{t_{j+1}}^{s}(x_{t_j}^{s^*})$ of its subproblem $(\mathrm{SUB}_{t_{j+1}}^{s})$ is characterized by the optimality cuts:

$$\hat{h}_{t_{j+1}}^{s}(x_{t_j}^{s^*}) \geq e_{t_{j+1}}^{s,p} - E_{t_{j+1}}^{s,p} x_{t_j}^{s^*} \qquad \forall\ p \in \mathcal{P}_{t_{j+1}}^{s} \tag{5.39}$$

where it is assumed that $x_{t_j}^{s^*}$ satisfies all feasibility cuts. The state disaggregation splits node (n, k_2) at time t_{j+1} in $M+1$ new nodes, and we denote the successor scenario of scenario s^* in the new node $(n_l, k_2 - 1)$ as s_l for $l = 0, \ldots, M$. The subproblem for scenario s^* at time t_j after the state disaggregation is then:

$$(\widetilde{\mathrm{SUB}}_{t_j}^{s^*}): \quad \tilde{h}_{t_j}^{s^*}(x_{t_{j-1}}^{s^-}) = \min \quad c_{t_j}^{s^*} x_{t_j}^{s^*} + \sum_{l=0}^{M} \tilde{h}_{t_{j+1}}^{s_l}(x_{t_j}^{s^*}) \tag{5.40}$$
$$\text{s.t.} \quad A_{t_j}^{(n,k_0)} x_{t_j}^{s^*} = b_{t_j}^{(n,k_0)} - F_{t_j}^{(n,k_0)} x_{t_{j-1}}^{s^-}$$
$$x_{t_j}^{s^*} \geq 0$$

We have shown in section 3.3.2 that $(\widehat{\mathrm{SUB}}_{t_j}^{s^*})$ can be obtained from $(\widetilde{\mathrm{SUB}}_{t_j}^{s^*})$ through column and row aggregations, and one is therefore neither a restriction nor a relaxation of the other. Thus we don't know whether $\tilde{h}_{t_j}^{s^*}(x_{t_{j-1}}^{s^-}) \geq \hat{h}_{t_j}^{s^*}(x_{t_{j-1}}^{s^-})$, or, equivalently, whether $\sum_{l=0}^{M} \tilde{h}_{t_{j+1}}^{s_l}(x_{t_j}^{s^*}) \geq \hat{h}_{t_{j+1}}^{s}(x_{t_j}^{s^*})$. This implies that the optimality cut

$$\sum_{l=0}^{M} \tilde{h}_{t_{j+1}}^{s_l}(x_{t_j}^{s^*}) \geq e_{t_{j+1}}^{s,p} - E_{t_{j+1}}^{s,p} x_{t_j}^{s^*} \qquad \forall\ p \in \mathcal{P}_{t_{j+1}}^{s} \tag{5.41}$$

may not be a valid cut, and we therefore cannot use it in $(\widetilde{\mathrm{SUB}}_{t_j}^{s^*})$.

To see that the feasibility cuts for $x_{t_j}^{s^*}$ in $(\widehat{\mathrm{SUB}}_{t_j}^{s^*})$ are not necessarily valid in $(\widetilde{\mathrm{SUB}}_{t_j}^{s^*})$, it suffices to note that an unbounded extreme ray in the dual subproblem of scenario s before the disaggregation is not necessarily an unbounded ray in any of the dual subproblems of the scenarios s_l after the disaggregation. This follows directly from the fact that the feasible region and the objective function in the dual subproblems are state dependent.

Time Disaggregation

To illustrate the problems with the Benders' cuts in case of time disaggregation, we consider the reversal of the basic time aggregation in figure 4-2. Before the time disaggregation, the subproblem for a scenario s^* in node (n, k_0) at time t_j can be written as:

$$\begin{aligned}(\widehat{\mathrm{SUB}}^{s^*}_{t_j}): \quad \hat{h}^{s^*}_{t_j}(x^{s^-}_{t_{j-1}}) = \min \quad & c^{s^*}_{t_j} x^{s^*}_{t_j} + \sum_{l=0}^{M} \hat{h}^{s_l^+}_{t_{j+2}}(x^{s^*}_{t_j}) \\ \text{s.t.} \quad & A^{(n,k_0)}_{t_j} x^{s^*}_{t_j} = b^{(n,k_0)}_{t_j} - F^{(n,k_0)}_{t_j} x^{s^-}_{t_{j-1}} \\ & x^{s^*}_{t_j} \geq 0\end{aligned} \tag{5.42}$$

As before, scenario s^- at time t_{j-1} is the predecessor scenario of scenario s^*, and $x^{s^-}_{t_{j-1}}$ is the associated decision vector. Scenario s_l^+ at time t_{j+2} denotes the successor scenario of scenario s^* that corresponds to node $(n_l, k_2 - 1)$ in figure 4-2(b), with $l = 0, \ldots, M$. The function value $\hat{h}^{s_l^+}_{t_{j+2}}(x^{s^*}_{t_j})$ for each $l = 0, \ldots, M$ in $(\widehat{\mathrm{SUB}}^{s^*}_{t_j})$ is defined by optimality cuts that are derived from the corresponding subproblem at time t_{j+2}, while $x^{s^*}_{t_j}$ is restricted by feasibility cuts from all of these subproblems.

After the time disaggregation, a new node (n, k_1) is introduced at time t_{j+1}, and the successor scenario of scenario s^* in this node is denoted by s. The subproblem for this successor scenario is:

$$\begin{aligned}(\widetilde{\mathrm{SUB}}^{s}_{t_{j+1}}): \quad \tilde{h}^{s}_{t_{j+1}}(x^{s^*}_{t_j}) = \min \quad & c^{s}_{t_{j+1}} x^{s}_{t_{j+1}} + \sum_{l=0}^{M} \tilde{h}^{s_l^+}_{t_{j+2}}(x^{s}_{t_{j+1}}) \\ \text{s.t.} \quad & A^{(n,k_1)}_{t_{j+1}} x^{s}_{t_{j+1}} = b^{(n,k_1)}_{t_{j+1}} - F^{(n,k_1)}_{t_{j+1}} x^{s^*}_{t_j} \\ & x^{s}_{t_{j+1}} \geq 0\end{aligned} \tag{5.43}$$

Because the time disaggregation does not change the subproblems of the scenarios at time t_{j+2}, we can characterize the functions $\tilde{h}^{s_l^+}_{t_{j+2}}(x^{s}_{t_{j+1}})$ in $(\widetilde{\mathrm{SUB}}^{s}_{t_{j+1}})$ by the same optimality cuts that were used in $(\widehat{\mathrm{SUB}}^{s^*}_{t_j})$, except that $x^{s^*}_{t_j}$ is replaced by $x^{s}_{t_{j+1}}$. In a similar manner, the feasibility cuts that were used to restrict $x^{s^*}_{t_j}$ in $(\widehat{\mathrm{SUB}}^{s^*}_{t_j})$ can now be used in $(\widetilde{\mathrm{SUB}}^{s}_{t_{j+1}})$ to restrict $x^{s}_{t_{j+1}}$.

The subproblem for scenario s^* at time t_j after the time disaggregation is

$$\begin{aligned}(\widetilde{\mathrm{SUB}}^{s^*}_{t_j}): \quad \tilde{h}^{s^*}_{t_j}(x^{s^-}_{t_{j-1}}) = \min \quad & c^{s^*}_{t_j} x^{s^*}_{t_j} + \tilde{h}^{s}_{t_{j+1}}(x^{s^*}_{t_j}) \\ \text{s.t.} \quad & A^{(n,k_0)}_{t_j} x^{s^*}_{t_j} = b^{(n,k_0)}_{t_j} - F^{(n,k_0)}_{t_j} x^{s^-}_{t_{j-1}} \\ & x^{s^*}_{t_j} \geq 0\end{aligned} \tag{5.44}$$

As $\tilde{h}^{s}_{t_{j+1}}(x^{s^*}_{t_j})$ is the objective value function of the new subproblem (5.43), we do not know any optimality or feasibility cuts from it. Furthermore, because the cuts in the subproblems of all ascendant scenarios of s^* were derived from the old cuts in $(\widetilde{\mathrm{SUB}}^{s^*}_{t_j})$, which do not apply to $(\widetilde{\mathrm{SUB}}^{s^*}_{t_j})$, new cuts need to be derived for all these ascendent subproblems as well.

5.3.2 Primal-Dual Decomposition

In order to start the primal-dual decomposition method, one needs an initial feasible solution, but any feasible solution will do. We have shown in section 4.3 how a feasible solution for a relaxation of the ALM model can be constructed after a disaggregation is performed in the underlying event tree. This feasible solution is directly derived from the optimum solution in the previous iteration of the algorithm. Furthermore, if certain conditions on some parameter values are satisfied, then the relaxation of the ALM model is guaranteed to have the same solution as the true model. We can thus apply the primal-dual decomposition method to this relaxation of the ALM model, starting from the constructed feasible solution, to find an optimal solution to the ALM model.

We mentioned before that the primal-dual method is most likely not a very efficient method if the initial solution is far from the optimum solution. However, each disaggregation in the iterative disaggregation algorithm makes only a relatively minor modification to the ALM model, and the optimum solution to the previous model should therefore provide a good starting solution. Furthermore, we note that Benders' decomposition constructs a *global* piecewise-linear approximation of the objective function in the course of the method to find the optimum solution. Primal-dual decomposition, in contrast, only builds a *local* piecewise-linear approximation around the current solution before a step in a descent direction is taken (where the linear pieces correspond to the set of *optimal* extreme points and *binding* extreme rays). Close to an optimal solution, this may cause the primal-dual decomposition method to converge faster to the optimum solution than Benders' decomposition.

In the discussion of the primal-dual decomposition method for multistage stochastic linear programs in section 5.2.2, we mentioned as drawbacks of the method that each subproblem has to be solved to complete optimality before we can can move backward one stage, and that the subproblems continue to increase in size when moving backward in time. The parametric linear programming pivots in these subproblems, needed to establish feasibility of a descent direction or find violated cuts, thus become

increasingly more computational intensive. However, the direction-finding problem remains relatively small as its size is dependent only on the *number* of successor subproblems. We note that once we are solving the first-stage problem, i.e., improving the solution of x_0, it is not necessary to find its optimal value before starting the next iteration in the iterative disaggregation algorithm.

If we would apply the trajectory optimization method of Grinold to re-optimize the ALM model, the situation is basically reversed. Trajectory optimization essentially solves the ALM model as a large two-stage stochastic program, in which there are many small subproblems, but a very large direction-finding problem. Due to the many subproblems, it is likely that the approximate direction-finding problem has to be updated with cuts and re-optimized many times before a descent direction is found that is feasible in all subproblems. Because of its size, re-optimizing the direction-finding problem may be very costly.

Chapter 6

A Computational Example

In this chapter we will present the results from the application of the iterative disaggregation algorithm to a simple asset/liability management problem. The problem we consider is small and highly structured, but it enables us to illustrate many features of the algorithm as well as the state and time aggregation methods of chapter 3. Furthermore, we will show that the solution of the problem by iterative disaggregation provides useful information about the sensitivity of the optimal solution to changes in the description of the uncertainty in the model.

The purpose of this computational exercise is to show the feasibility and value of solving asset/liability management problems by the iterative disaggregation algorithm. The emphasis is not on computational efficiency or the size of the problems that can be solved, and we have therefore not implemented any of the decomposition methods that were discussed in chapter 5.

6.1 A Simple Asset/Liability Management Problem

Interest-rate options on treasury bonds that are traded in the market have a time to maturity of only a few months. They can therefore be used to hedge against short-term interest-rate exposure, but not directly for long-term hedging purposes. We will consider the problem of an investor whose exposure to interest-rate variability is long-term ("long" in the sense that it exceeds the maturity of the traded options), and who wants to devise a dynamic trading strategy in the short-term options that guarantees him the payoffs of a long-term option.

This problem is similar to the problem that Hamilton [22] considers for the replication of a long-term option on a stock. He does not, however, construct a dynamic trading strategy, but compares several "myopic" approaches. In the first approach, he constructs a portfolio of options that are traded today such that the sensitivity of the portfolio value with respect to changes in the price of the stock (as measured by the first and second derivatives, the so-called *delta* and *gamma* measures, which can be readily calculated from the Black-Scholes option-pricing formula) equals the sensitivity of the option that is being replicated. This approach is thus very similar to duration matching for asset/liability management under interest-rate uncertainty (see section 1.2).

In the second approach, he uses a linear programming model to construct a portfolio of options that are traded today such that the portfolio value at the time when the first options in the portfolio expire exceeds the theoretical value of the replicated option at that point in time for a number of different stock-price scenarios. Hamilton concludes that the second approach provides a better hedging portfolio. We note, however, that neither of the approaches can give a guarantee that one ends up with sufficient funds at the time when the payoffs from the replicated option are needed. Furthermore, these approaches cannot take transaction costs into account which are incurred when the portfolio is rebalanced (i.e., new short-term options are bought at future points in time), and therefore ignore the effect of these costs on the optimal portfolio composition.

6.2 Problem Statement

We consider an investor who owns a zero-coupon treasury bond that matures in two years from the current date, and assume that he plans to sell the bond after one year. However, he wants to be guaranteed a minimum return on the bond, for example, to meet a fixed liability at that date. That is, he wants to protect himself against rising interest rates (and thus falling bond prices). One way to achieve this is to enter into a forward contract that allows him to sell the bond after one year at a specified price. Although this protects him against rising interest rates, it also prevents him from making a profit if interest rates fall. We assume that the investor wants to be able to benefit from a fall in interest rates, and therefore in effect wants to buy a one-year put option on the bond. Options that are traded in the market, however, are assumed to have a maturity of at most four months. The problem is thus to construct a dynamic

trading strategy, including the short-term options, whose payoffs match the payoffs of the hypothetical one-year put option (the *replicated* option). We assume that the investor does not allow for shortfalls, and the final portfolio value must therefore be nonnegative in all scenarios.

The face value of the two-year zero-coupon bond is \$1000. We assume that the yield curve is flat with a yield of 8% for all maturities[1] and the current price of the bond is therefore \$852.14. The forward price[2] of the bond for delivery after one year is \$923.12, and we assume that the investor wants to generate a cash flow from the sale of the bond of at least \$932.35 (101% of the forward price). That is, the strike price of the one-year put option that he wants to replicate is \$932.35.

We assume that traded option contracts on the two-year bond are initiated at the beginning of every two-month period, and that the options have an initial maturity of four months. Furthermore, for every option maturity, three call options and three put options are traded, which differ only in their exercise prices (respectively 99.5%, 100% and 100.5% of the forward bond price on the maturity date of the options). This implies that at every point in time, the investor can trade in six different put options and six different call options on the two-year bond.

It is furthermore assumed that the prices of the options that are traded at time 0 are consistent with a version of the Ho and Lee model that incorporates 120 time steps, in which the implied binomial probability is 1/2, and the volatility of the short-term interest rate 0.7% per year. (This number of time steps is large enough so that the calculated option values have converged to at least two-decimal precision, and the volatility level prevents negative interest rates in the model at any point in time; see also section 3.1.1.) The theoretical value of the replicated put option at time 0 according to this model is \$8.73. The data for the traded options, including their price at time 0 if applicable, are listed in table 6.1. The initiation and expiration dates are specified in terms of the time steps in the Ho and Lee model. Because options 10, 11 and 12 were initiated before time 0, no initiation date is specified for them. The strike price of each option is both given as absolute number and as percentage of the forward bond price (between brackets).

---

[1] As before, we assume that the yields are given as continuously compounded yields.

[2] If $P_0(\tau)$ denotes the current price of a zero-coupon bond with maturity τ, then the forward price of a zero-coupon bond with maturity T for some time $t < T$ is equal to $(P_0(T)/P_0(t))$ times the face value of the bond. This is the arbitrage-free delivery price in a forward contract for delivery of the bond with maturity T at time t.

Option number	Initiation date	Expiration date	Strike price	Price at time 0: put	call
10		20	\$859.26 (99.5%)	\$0.40	\$4.66
11		20	\$863.58 (100.0%)	\$1.76	\$1.76
12		20	\$867.90 (100.5%)	\$4.66	\$0.40
20	0	40	\$870.80 (99.5%)	\$0.76	\$5.03
21	0	40	\$875.17 (100.0%)	\$2.28	\$2.28
22	0	40	\$879.55 (100.5%)	\$5.03	\$0.77
30	20	60	\$882.49 (99.5%)	-	-
31	20	60	\$886.92 (100.0%)	-	-
32	20	60	\$891.36 (100.5%)	-	-
40	40	80	\$894.33 (99.5%)	-	-
41	40	80	\$898.83 (100.0%)	-	-
42	40	80	\$903.32 (100.5%)	-	-
50	60	100	\$906.34 (99.5%)	-	-
51	60	100	\$910.89 (100.0%)	-	-
52	60	100	\$915.44 (100.5%)	-	-
60	80	120	\$918.50 (99.5%)	-	-
61	80	120	\$923.12 (100.0%)	-	-
62	80	120	\$927.73 (100.5%)	-	-

Table 6.1: Data for traded options on the two-year zero-coupon bond.

6.3 Disaggregation Strategy

To start the iterative disaggregation algorithm, we have aggregated states and time steps in the Ho and Lee model of the previous section to obtain the aggregated event tree for iteration 0 in figure 6-3. The interest rate (as a percentage per year) ranges between 8.128% in the lowest state to 7.872% in the highest state, and the corresponding liability (payoffs on the replicated put option) between \$10.55 and \$8.38. Because trading dates are included in this aggregated event tree for all points in time at which dividends are paid (i.e., options expire) and liabilities are due ($t = 120$), no adjustment for the prepayment of dividends and liabilities is necessary in the ALM model (compare with the aggregated ALM model in section 3.4). Note that the initial event tree has only four different scenarios at $t = 120$, and the corresponding ALM model is thus small and easy to solve.

To choose a state in the aggregated event tree in which to perform a disaggregation, we have used the sensitivity analysis that was described in the second part of

section 4.3.2. In the specific example that we consider here, however, it is not possible to base the disaggregation directly on the value of the sensitivity measures that were introduced in that section. To see this, we note that it only makes sense to calculate the sensitivity measures for a state disaggregation in states that have successor states in which a liability payment is due. In the problem at hand, these are the states at the beginning of the last period in the aggregated event tree. A disaggregation strategy that is solely based on the level of this sensitivity measure would therefore only perform state disaggregations in these states. Furthermore, it was shown in section 4.3.2 that the sensitivity measure for a time disaggregation will be zero in every state of the aggregated event tree from the previous paragraph as there are no prepaid dividends and liabilities. We have therefore modified the disaggregation strategy in the following way.

First, we have imposed the restriction that a state in the aggregated event tree can never have more than two successor states.[3] As a consequence, if the sensitivity measure indicates that a state disaggregation should be performed in a state which already has two successor states, then we perform a time disaggregation in that state instead.

Second, we identify a *critical scenario* in each state for which the sensitivity measure is calculated. The critical scenario in a state is the scenario that gives the highest single contribution to the sensitivity measure ε in that state (remember that the sensitivity measure in a state is a summation of the sensitivities over all scenarios in that state). If a state is selected for a state disaggregation based on its value for ε, then a disaggregation is performed along the path in the event tree that corresponds to the critical scenario in that state (the *critical path*). If possible, a state disaggregation is performed somewhere along this critical path. When there are multiple possibilities, then the state disaggregation is performed at the earliest point in time at which it is possible. If no state disaggregation is possible, then a time disaggregation is performed in the state at the beginning of the longest period on the critical path (i.e., comprising the largest number of time steps in the original Ho and Lee model). If there is more than one possibility, then the time disaggregation is performed as early as possible in the tree.

We have also used the critical path to define the way in which a new state after a

[3]In addition, we have imposed the restrictions that have been imposed throughout this thesis, namely that all successor states of a state must occur at the same point in time, and that they must all have the same aggregation level.

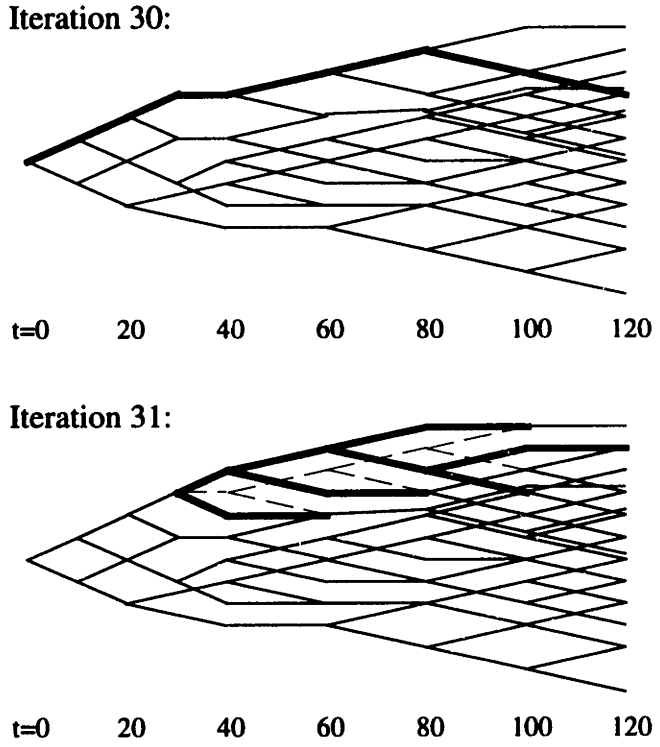

Figure 6-1: Change in event tree after a state disaggregation on the critical path.

state disaggregation is connected to the existing event tree. After a state disaggregation is performed in some state along the critical path, then the state disaggregation is basically pushed forward along the path until all new states are connected to the existing tree, or the end of the tree is reached. This is illustrated in figure 6-1 for the state disaggregation in iteration 31 of the iterative disaggregation algorithm when applied to the base-case problem that will be discussed in the next section. The critical path in the event tree of iteration 30 is indicated by the fat line. In iteration 31, a state disaggregation is performed in the state on the path at time 30. The new arcs in the event tree of iteration 31 after the state disaggregation are indicated in bold, and the ones that have disappeared from the event tree are drawn as dashed lines.

A time disaggregation in our implementation consists of a basic time disaggregation (i.e., the reversal of figure 4-2), followed by a basic state disaggregation (the reversal of figure 4-1) in the same state. This is depicted in figure 6-2. If a time disaggregation adds a state between time t and $t + \tau$ in the aggregated event tree, then this state is added in the middle between these points in time, i.e., at time $t + \tau/2$.

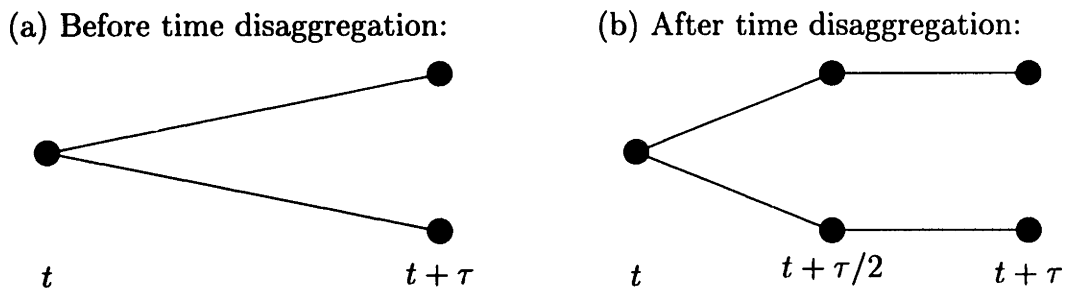

Figure 6-2: A time disaggregation that comprises one *basic* time disaggregation followed by a basic state disaggregation.

6.4 Computational Results

We have coded the iterative disaggregation algorithm for this problem on a Sun 10 workstation with 32 MB of internal memory (RAM) in the C programming language. The CPLEX callable library for linear programming was used to perform the optimizations in each iteration. The ALM model has been re-optimized in each iteration as a large linear program, where the optimal basis from the previous iteration was used to define a starting basis. The optimal basis columns from one iteration typically do not define a full basis for the model in the next iteration, but CPLEX allows the specification of an incomplete basis, and will complement it with additional columns to construct an initial basis.

We have defined a base-case problem in which the transaction cost rate is 1%, and the final-portfolio weight $\lambda_1 = 0.9$. The investor has the possibility to borrow up to 10$ in each state at the riskless one-period interest rate plus one basis point (one hundreth of a percent), and he faces a 1% borrowing spread for any amount in excess of that. We have assumed that the investor can only take long positions in the short-term options.

We will first analyze the development of the event tree in the course of the algorithm, and then study the optimal solution for this base-case problem. Subsequently, we will show how this optimal solution changes under different assumptions about the transaction costs and the weight λ_1 on the final portfolio value in the objective function.

Iteration	Event tree		ALM model		
	states	scenarios	rows	columns	nonzeros
0	18	23	131	485	1007
5	27	60	312	1018	2368
10	28	99	423	1405	3259
15	28	120	492	1611	3788
20	45	140	596	1963	4570
25	43	186	780	2564	5999
30	57	206	926	3033	6990
35	58	253	1111	3624	8418
40	68	342	1572	5120	11770

Table 6.2: Size of the event tree and the corresponding ALM model in selected iterations of the iterative disaggregation algorithm when applied to the base-case problem.

6.4.1 Results for the Base-Case Problem

Figures 6-3 and 6-4 depict the development of the aggregated event tree in the course of the algorithm. Time disaggregations are performed in iterations 17, 26, 30 and 38, and state disaggregations in all other iterations. In the final event tree (iteration 40), the interest rate at time 120 deceases from 8.384% at the bottom of the tree to 7.617% at the top. The corresponding payoffs from the replicated put option (the liabilities) range from \$12.82 to \$6.33.

Table 6.2 lists the number of nodes and scenarios in these event trees, as well as the corresponding dimensions of the ALM model. The complete run of 40 iterations only takes a few minutes in real time. The number of simplex pivots that is needed to re-optimize the ALM model in each iteration varies between less than ten and a few hundred. We have run the algorithm for more than 40 iterations as well, but this had no substantial effect on the optimal solution. Furthermore, after an additional 20 to 30 iterations CPLEX is no longer able to optimize the resulting ALM model with the available computer memory because of the large size of the linear program at that point (approximately 18,000 columns and 6,000 rows, corresponding to an event tree with over 1200 scenarios). This clearly illustrates the need for the use of decomposition methods when one wants to solve large stochastic programming models.

The development of the event tree in the course of the algorithm is partly determined by the value of the sensitivity measure in states at the beginning of the last

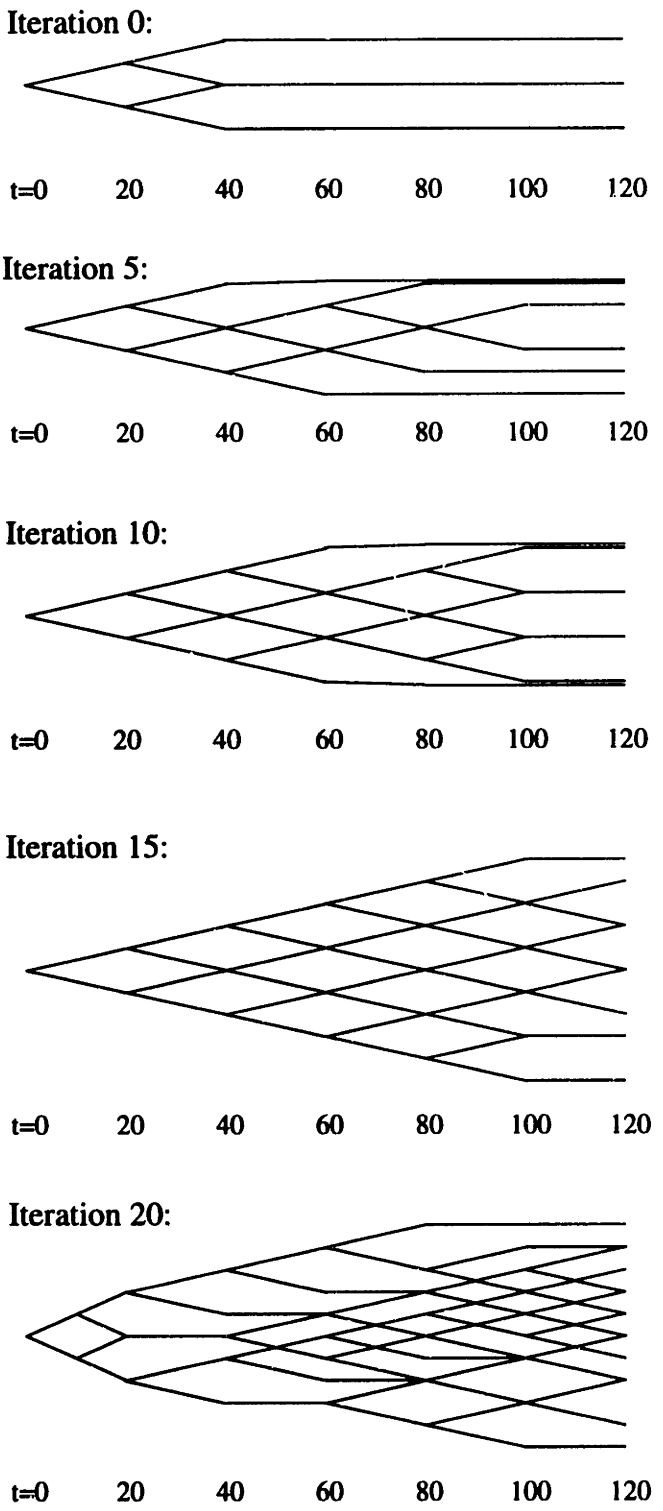

Figure 6-3: Changes in the event tree during the iterative disaggregation algorithm for the base-case model: iterations 0 through 20.

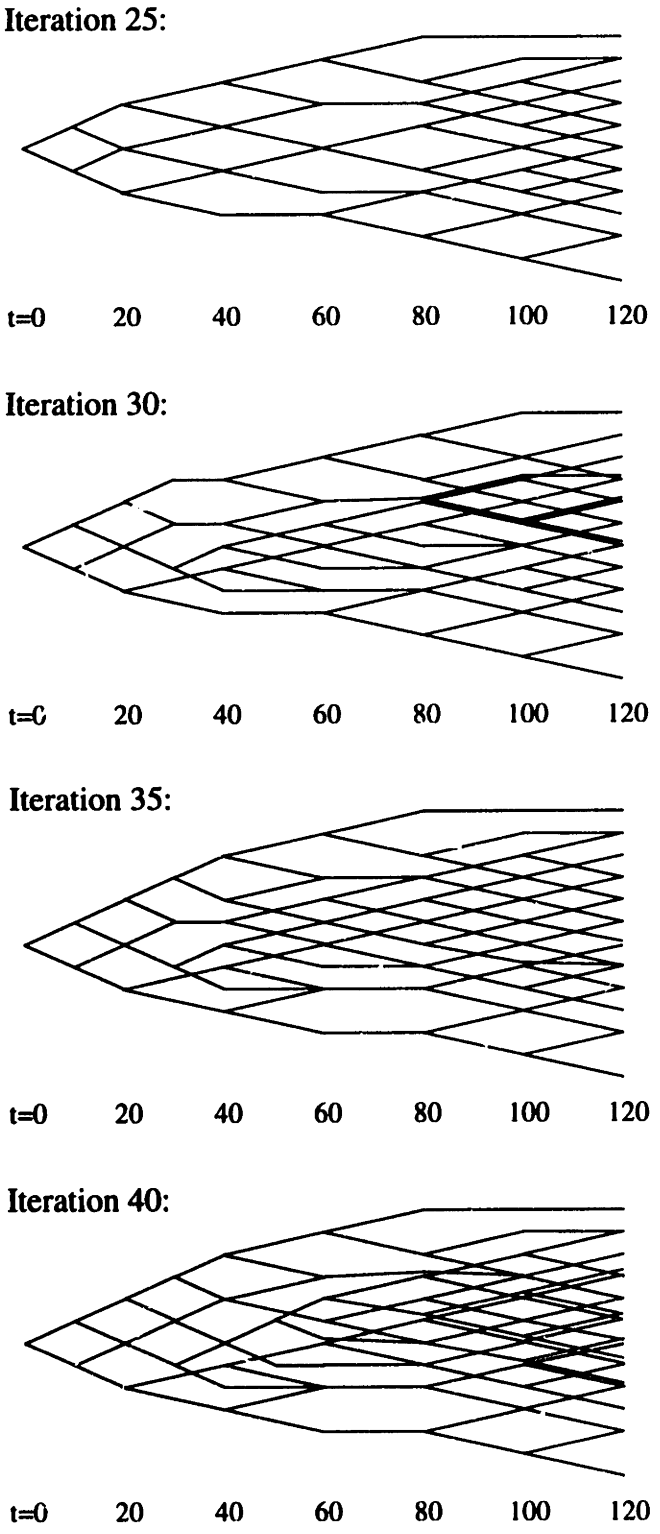

Figure 6-4: Changes in the event tree during the iterative disaggregation algorithm for the base-case model: iterations 25 through 40.

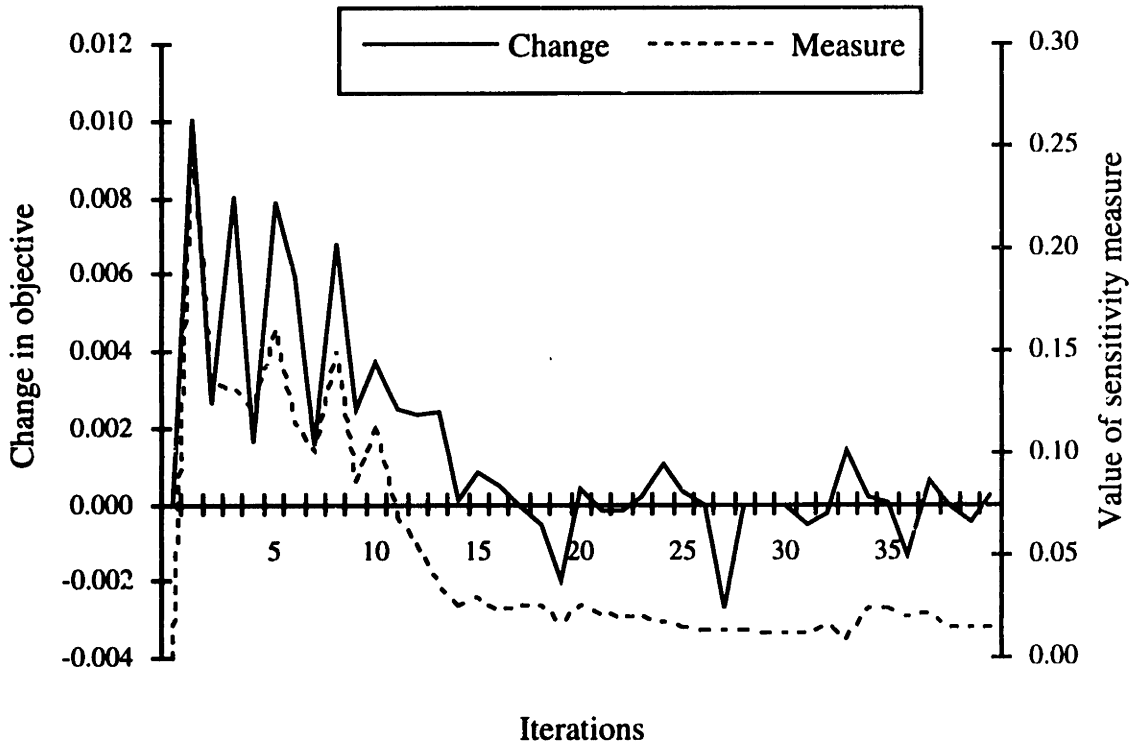

Figure 6-5: Change in the optimal objective value and the value of the sensitivity measure as a function of the iterations in the iterative disaggregation algorithm (the base-case problem).

period in the tree (the only states for which the measure is calculated), and partly by the critical scenario in that state. Figure 6-5 compares the value of the sensitivity measure with the actual change in the optimum objective value of the ALM model in each iteration. The correlation between the two is high in the early iterations, but decreases in later iterations.

A study of the sensitivity measure across different states in each iteration shows that the number of scenarios in a state (and correspondingly, its probability of occurrence) is an important determinant of its value. The sensitivity measure therefore exhibits a bias towards states in the center of the event tree. As the variance in the number of scenarios per state increases with the growth of the event tree, this bias becomes stronger in the course of the algorithm. This is clear from the development of the event tree in figures 6-3 and 6-4. An implication is that the critical scenario in the state with the largest value of the sensitivity measure plays an increasingly

important role for the actual disaggregation that is performed in the event tree.

Figures 6-3 and 6-4 show that relatively many disaggregations occur in the upper half of the event tree. This may seem unintuitive at first, as states in that part of the event tree are most favorable to the investor. Analysis of the critical path in each iteration, however, reveals the reason for this. Typically, the interest rates along the critical path decrease initially, and increase after a certain point in time. (In the event tree, this corresponds to paths that move upward in the event tree at first, and downward later on.) That is, the interest-rate path is initially favorable for the investor, but turns unfavorable after a while. It is not surprising that this kind of scenarios are the most difficult to hedge against.

We will continue with a discussion of the optimal solution to the base-case problem. The optimal value of the objective function in the course of the algorithm is depicted in figure 6-6, together with the cost (including transaction costs) of the optimal portfolio at time 0. The objective value remains relatively stable after iteration 15, and the cost of the initial portfolio after iteration 20. The difference between the two lines represents the value of the excess portfolio at the terminal date that is credited to the objective. The average excess in a scenario at time 120 is \$0.11 with a standard deviation of \$0.28. The standard deviation is high because the excesses almost exclusively occur in the upper part of the event tree (corresponding to the states with low liabilities).

In each iteration, the optimal portfolio at time 0 only consists of the put options 10 and 20 and short-term lending. Because positions in the options are restricted to long positions, and as the investor wants to replicate a long put option, it is clear that the call options have an unattractive payoff pattern, and are therefore not included in the portfolio. Of the short-term put options, the selected options are the ones that are most out-of-the-money (i.e., with the lowest strike price). These options provide the investor with the largest *relative* difference in payoff in different states of the world per option bought. As a consequence, to obtain a certain *absolute* difference in payoffs between different states of the world in the future, the out-of-the-money options require the smallest investment in dollars. This is attractive as we have assumed a transaction cost rate that applies to the dollar investments in the options, but not to short-term lending.

Figure 6-7 shows the optimal portfolio composition as a function of the iteration number. Notice that this figure depicts the *number* of options bought (the vertical axis on the left) versus the *dollar amount* of short-term lending (the vertical axis on

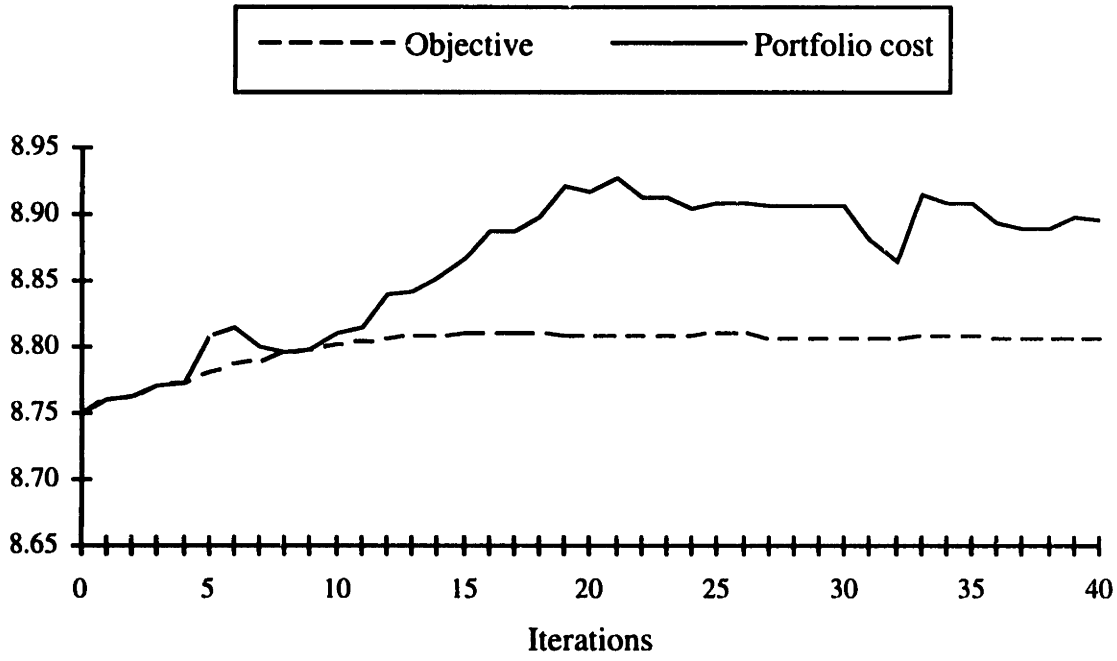

Figure 6-6: Optimal objective value and initial portfolio cost as a function of the number of iterations in the iterative disaggregation algorithm (the base-case problem).

the right). Convergence in the optimal portfolio composition is not nearly as clear as the convergence in the optimal objective value in figure 6-6. In fact, a small change in the event tree sometimes has a substantial effect on the option holdings in the optimal portfolio at time 0. However, the dollar amount invested in the options as fraction of the portfolio cost remains approximately constant (a little below 20%). These observations strongly suggest that there are multiple optimal solutions to the problem, each of which contains approximately the same amount in short-term lending, but they differ in the division of the remaining part of the portfolio investment among the short-term put options. As is clear from a comparison of figures 6-6 and 6-7, different divisions can correspond to a different trade-off in the objective function between the initial portfolio cost and the final portfolio value, although the objective value itself remains approximately the same.

To provide an idea of the portfolio rebalancing that takes place after time 0, figure 6-8 shows the optimal portfolio composition in the scenarios up to time 20 in the final ALM model (i.e., the model in iteration 40). The "wealth" in each scenario in this figure equals the portfolio value, calculated from the corresponding option

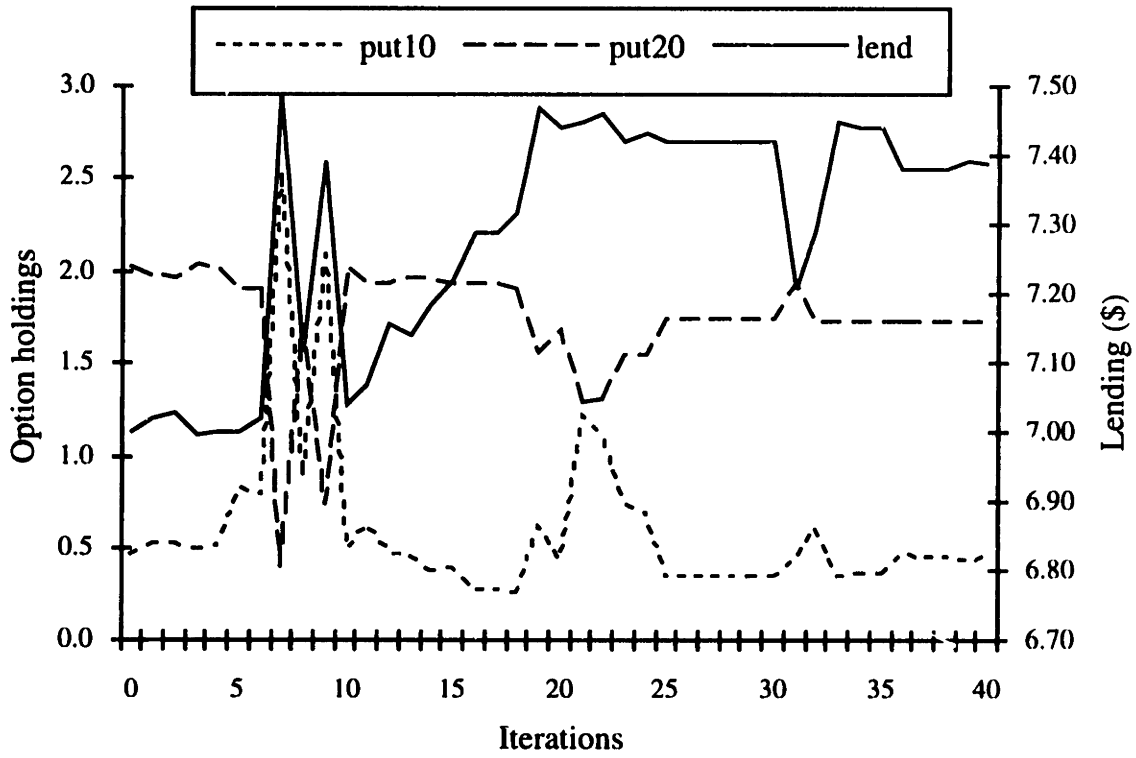

Figure 6-7: Optimal portfolio composition as a function of the number of iterations in the iterative disaggregation algorithm (the base-case problem).

prices in the event tree. Put option 10 expires at time 20, while put option 30 is initiated at that time. This is reflected in the portfolio composition.

Notice that additional options are bought whenever the interest rate decreases (corresponding to an "upstate" in the event tree), whereas the option holdings remain unchanged when the interest rate increases. That is, additional options are bought whenever the value of the options in the portfolio decreases. The total dollar value of the options in the portfolio, however, is lower in states that correspond to a lower interest rate, as is the investor's wealth. This is to be expected, as his liabilities at time 120 are lower when the interest rate is lower.

6.4.2 Variations in the Transaction Cost Rate

To see what the effect of the transaction cost rate is on the optimal portfolio at time 0 and the portfolio rebalancing strategy after time 0, we have varied the transaction cost rate c between 0.1% and 2%, all with the final-portfolio weight $\lambda_1 = 0.9$. Table 6.3

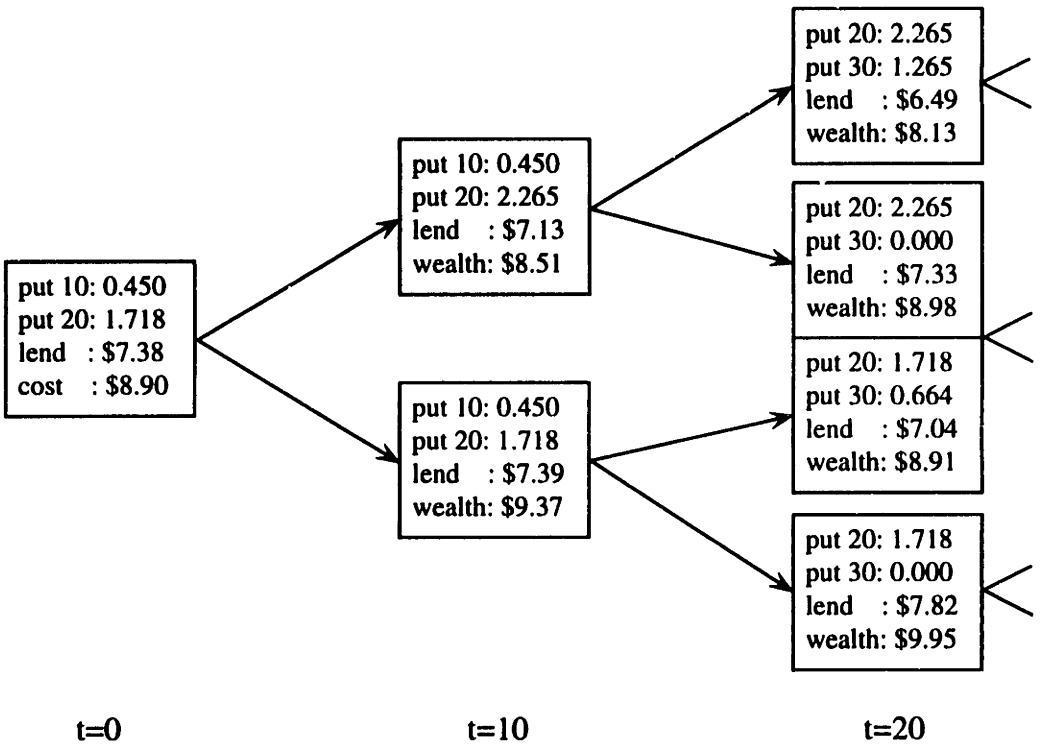

Figure 6-8: Portfolio rebalancing in the base-case problem.

compares the optimal portfolio composition at time 0 and the corresponding objective values. When transaction costs increase, more of the initial portfolio is invested in short-term lending (on which no transaction costs are paid) and less in the options. A consequence of the changing portfolio composition is that the asset cash flows match the liabilities less precisely when transaction costs increase. This is reflected in the expected value of the cash surplus at the model horizon (the expected final excess in table 6.3).

The variations in the relative option holdings in the optimal portfolio for different transaction cost rates is another indication for the existence of multiple optimal solutions, as discussed earlier. Analysis of the changes in the optimal portfolio composition in the course of the algorithm for the different transaction cost rates shows a similar pattern as in figure 6-7: the division of the initial portfolio investment between short-term lending and option purchases is approximately constant, but the individual option holdings can show fairly large swings in successive iterations.

Transaction costs have a significant effect on portfolio rebalancing after time 0. To illustrate this, figure 6-9 shows the optimal portfolio composition in the scenarios up to time 20 of the final event tree (corresponding to iteration 40) when the transaction

$c =$	Objective	Portfolio cost time 0	Expected final excess	Portfolio composition		
				put 10	put 20	lending
0.1%	\$8.74	\$8.74	\$0.00	0.872	1.713	\$7.08
0.5%	\$8.77	\$8.78	\$0.00	0.000	2.403	\$6.93
1.0%	\$8.81	\$8.89	\$0.11	0.450	1.718	\$7.39
1.5%	\$8.84	\$9.02	\$0.22	0.000	2.006	\$7.46
2.0%	\$8.85	\$9.18	\$0.40	0.118	1.711	\$7.80

Table 6.3: Optimal solution and portfolio composition at time 0 under different transaction cost rates.

cost rate is 2.0%. In this case, no changes are made to the optimal option holdings in any of the scenarios except the top scenario at time 20, where some investment is made in put option 30. For a transaction cost rate of 1.0%, it was seen in figure 6-8 that additional investments in the options were made after every upward move in the event tree.

6.4.3 Variations in the Final-Portfolio Weight

To see the effect of changes in the final portfolio weight λ_1 on the optimal solution, we have varied this parameter between the values 0.8 and 0.98. The transaction cost rate was kept constant at 1%, as in the base-case problem. Table 6.4 displays the optimization results. When λ_1 increases, it becomes less important to match the liabilities exactly. As a consequence, the transaction cost rate increases in relative importance, and a larger part of the initial portfolio investment is spent on short-term lending. A higher value of λ_1 also decreases the need for active portfolio rebalancing. Although the optimal objective value decreases slightly when λ_1 increases, the initial portfolio cost increases relatively steeply. Correspondingly, the expected final portfolio value increases significantly with λ_1.

6.5 Concluding Remarks

The application of the iterative disaggregation algorithm to the solution of the simple asset/liability management problem in this chapter has illustrated several useful features of the algorithm. Most importantly, the solutions that are generated in the course of the algorithm provide a good insight into the sensitivity of the optimal

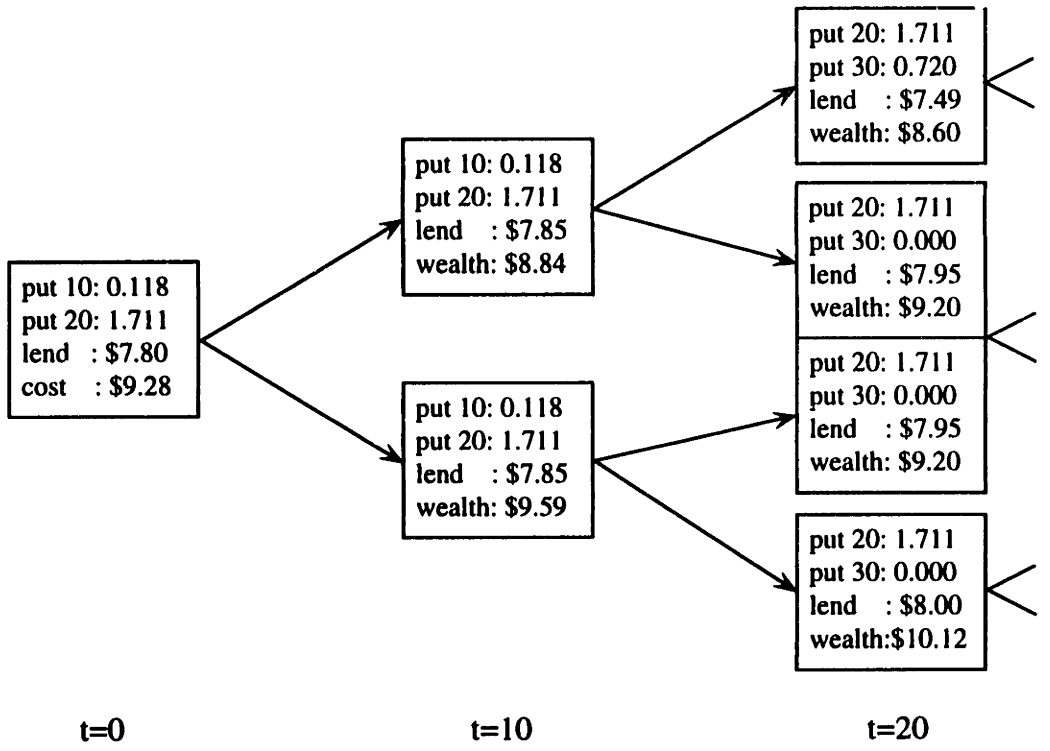

Figure 6-9: Portfolio rebalancing when the transaction cost rate equals 2.0%.

solution to an increasing level of uncertainty in the stochastic programming model. For the particular problem in this chapter, the optimal objective value was shown to converge fairly rapidly. The optimal portfolio composition exhibited a more volatile behaviour, but the general structure of the optimal solution was stable, with approximately fixed proportions of the optimal portfolio at time 0 invested in riskless assets (short-term lending) and risky assets (short-term options). Furthermore, the selected options were the same in each iteration, only the holdings in these selected options varied in successive iterations. As these variations did not correspond to changes in the objective value, this suggests that multiple optimal (or very close to optimal) solutions exist. As many of these solutions correspond to a somewhat different trade-off between the cost of the initial portfolio and the final portfolio value, the choice of the actual portfolio may be based on this trade-off.

Judging from the results for the particular problem in this chapter, therefore, the solutions that are obtained in the course of the algorithm give strong suggestions about the general structure of the optimal portfolio (often referred to as *asset allocation* in the literature) as well as the individual assets that should be included in this portfolio. It was also shown that the selected assets were independent of the

$\lambda_1 =$	Objective	Portfolio cost time 0	Expected final excess	Portfolio composition		
				put 10	put 20	lending
0.80	\$8.81	\$8.81	\$0.00	0.000	2.423	\$6.94
0.85	\$8.81	\$8.82	\$0.01	0.846	1.723	\$7.15
0.90	\$8.81	\$8.89	\$0.11	0.450	1.718	\$7.39
0.95	\$8.79	\$9.07	\$0.31	0.266	1.704	\$7.65
0.98	\$8.77	\$10.25	\$1.64	0.000	0.084	\$10.19

Table 6.4: Optimal solution and portfolio composition at time 0 for different values of the final-portfolio weight in the objective.

values for the transaction cost rate and the final-portfolio weight, while these parameters did influence the proportion of risky versus riskless assets in the optimal portfolio. Whether these results hold in general can only be determined by further experimentation on different problems.

In our implementation of the iterative disaggregation algorithm, we have made some particular choices for the disaggregation strategy, and an open question is how the results are affected if different choices are made. The sensitivity measure that we have used to select states for a state disaggregation was shown to have a bias towards states in the center of the event tree, where the number of scenarios per state is large (and thus the probability of occurrence high). The resulting portfolio strategy can therefore be viewed as a hedge against the most likely course of events. In many problems, one may be primarily interested in hedging against extreme scenarios, each of which has only a small probability of occurring, but a possibly large effect on the investor's wealth. In such cases, one should modify the disaggregation strategy so that it has a tendency to select these extreme scenarios.

Furthermore, we have restricted the number of successor states in the event tree to a maximum of two, and a time disaggregation was performed if a state was chosen for a disaggregation that already had two successor states. An obvious modification is to allow for more than two successors in the event tree before a time disaggregation is performed. This would also enable us to avoid states in the event tree with only one successor state, corresponding to a period with no uncertainty, which is a somewhat undesirable situation in an event tree. For example, if a maximum of three successor states is allowed in the event tree, then one could perform time disaggregations as in figure 6-10. With an increase in the maximum number of successor states, however,

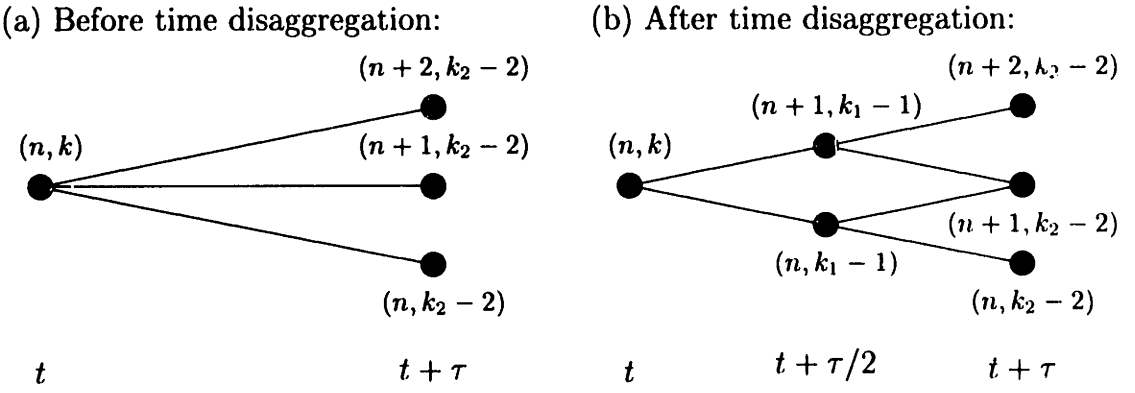

Figure 6-10: Time disaggregation when the maximum number of successor states in the event tree equals three ($k_1 \equiv k + \tau/2$; $k_2 \equiv k + \tau$).

the event tree will grow faster in the course of the iterative disaggregation algorithm, thereby limiting the number of iterations.

The sensitivity measure that we have used to select states for a state disaggregation in the event tree appeared to be a good indicator of the size of the change in the objective function, especially in the earlier iterations of the algorithm. Whether it is superior to the calculation of bounds on this change, which was discussed in chapter 4 as an alternative method for choosing the disaggregation, can only be verified by more computational testing.

In our implementation of the iterative disaggregation algorithm, the stochastic program was re-optimized as a large linear program in each iteration. As we have indicated, this puts a heavy demand on the available computer memory, which turned out to be the decisive factor for the size of the models that could be solved in our case (rather than computer time). This illustrates the need for the use of decomposition methods when large stochastic programming models have to be solved.

Chapter 7

Conclusions and Directions for Further Research

In this thesis we have shown that financial asset pricing theory plays an important role in the formulation of stochastic programming models for asset/liability management under interest-rate uncertainty. Specifically, we have shown that if the description of the asset-price uncertainty in the stochastic program is not arbitrage-free, then the optimal solution to the model may be substantially biased towards the arbitrage opportunities that are present in this description. This is the case even if the investor in the model cannot directly take advantage of these arbitrage opportunities because of market frictions and trading restrictions. As it is unrealistic to assume that any investor can predict arbitrage opportunities that will arise in the future, this also is a very reasonable restriction to impose on the description of the uncertainty.

To obtain a description that satisfies this restriction, we have shown that financial models of the term-structure uncertainty provide a very useful starting point. Prices of interest-rate-derivative securities that are calculated from such models have the desired property that they are arbitrage-free. However, to obtain accurate security price estimates, these models have to include a level of detail about the term-structure uncertainty that is much too large to include in a stochastic program without losing its computational tractability. To resolve this, we have presented state and time aggregation methods which aggregate states in the term-structure model, thereby reducing the number of interest-rate scenarios that is implied by the model, without losing the consistency of security prices in the model. That is, the security prices in the aggregated model remain arbitrage-free, and the calculated prices at time 0 equal the prices from the original model.

Although the state and time aggregation methods enable a reduction of any size in the number of interest-rate scenarios, the optimal solution to the stochastic programming model will generally be sensitive to the level of uncertainty in the model. To be able to explicitly study this sensitivity, we have shown how the state and time aggregation methods can be used as the basis for an iterative disaggregation algorithm to solve the ALM model. In this algorithm, the aggregation methods are initially used to construct an optimization model with only few scenarios, which is therefore easy to solve. In subsequent iterations, additional uncertainty is introduced in the model by reversing earlier state and time aggregations (called disaggregations). We have shown that the optimal solution in each iteration provides information as to where additional uncertainty in the model will affect the solution most, and this information can therefore be used to choose the disaggregations. Furthermore, we have shown that the optimal solution in one iteration can provide a starting point for the optimization in the next iteration, despite the fact that the stochastic programming models in subsequent iterations are neither restrictions nor relaxations of each other.

We have reported on the application of the iterative disaggregation algorithm to the solution of a simple asset/liability management problem. It was shown that the history of solutions in the course of the algorithm provides useful information about the optimal portfolio composition and the sensitivity of the objective function with respect to additional uncertainty in the model. Both the objective function value and the structure of the portfolio composition converged fairly quickly with the number of iterations in this specific example.

In the implementation of the algorithm, however, we have imposed several restrictions on the structure of the aggregated event tree and the disaggregation strategy. Only further experimentation with the algorithm on a variety of problems, and with different and possibly less restrictive assumptions, can give us more insights in the dependence of the observed solution behaviour on the particular assumptions as well as the problem itself. Furthermore, we have re-optimized the stochastic programming model as a large linear program in each iteration. The associated memory requirements on the computer severely limited the size of the models that could be solved, and therefore the number of iterations in the algorithm. This shows the need for alternative optimization methods, in particular decomposition methods, to re-optimize the stochastic program in the course of the algorithm.

The most commonly used decomposition method for stochastic programs is (nested) Benders' decomposition. We have shown in this thesis, however, that this decompo-

sition method cannot take advantage of the optimal solution in one iteration of the algorithm to perform the re-optimization in the next. We have also discussed a primal-dual decomposition method which does not have this disadvantage, and can make full use of the previous optimal solution. As was shown, this method is essentially a local-search method in which the optimal solution from the previous iteration can act as the starting point for the local-search procedure. Because this starting point will be relatively close to the new optimum, especially in later iterations of the algorithm when the stochastic program model is large and individual disaggregations have a relatively minor effect on the model formulation, this approach has intuitive appeal.

As described in chapter 5, however, the method requires that the stochastic program is essentially fully optimized in every iteration of the algorithm in order to apply it in a decomposed fashion. Furthermore, the convergence proof of the method requires that feasibility is maintained in every step of the method, which makes the determination of a descent direction a somewhat cumbersome and possibly time-consuming effort, and may have the effect that only small changes in the solution can be realized for every descent direction that is found. The practical efficiency of the primal-dual decomposition method as it was described in chapter 5 is therefore questionable. Because a local-search method seems a sensible approach in the context of the iterative disaggregation algorithm, however, many of the concepts in the primal-dual method may prove useful for the design of heuristic methods in which the above requirements are relaxed. Research in this direction, both on an algorithmic and a computational level, is needed.

We will discuss below how the model formulation that we have considered in this thesis can be generalized by relaxing some of the underlying assumptions. Key assumptions in our model development, however, are that security prices do not admit arbitrage opportunities, and that a term-structure model exists that is consistent with observed market prices of securities. Satisfaction of this last assumption is clearly too much to ask for in reality. Although it may be possible to choose the parameter values in a term structure model such that differences between implied arbitrage-free values and market prices are small, a perfect match will seldom be possible. The question is then which prices to use in a portfolio optimization model, the observed prices or the arbitrage-free values according to the term-structure model. We have shown that arbitrage opportunities in the description of the uncertainty, which would be created if market prices are used that differ from the arbitrage-free values, can

have a significant effect on the optimal solution to a portfolio optimization model. Furthermore, the primary purpose of a model for asset/liability management is to obtain a portfolio strategy that forms a reliable hedge against the future liabilities. Such a strategy should not be influenced by hypothesized arbitrage opportunities. We therefore suggest the use of the arbitrage-free values instead of the market prices in a portfolio optimization model if there is a difference between the two. Of course, one can always optimize the model twice, once with the arbitrage-free values and once with the market prices, and analyze what the differences are.

Extensions of the Model Formulation

In the formulation of the ALM model we have assumed that all variables are continuous, and all constraints and the objective function linear. This is attractive from a computational viewpoint as it enables the use of efficient linear programming methods to optimize the model. The formulation, however, is certainly not limited to these assumptions.

Integer variables (in particular, binary variables) would for example be necessary if one wants to model a transaction cost rate that decreases with the traded dollar amount or when transaction costs are discrete, and if one wants to impose lower bounds on the trading size when trading takes place. In a different problem context, Bienstock and Shapiro [2] show that integer variables can be embedded in Benders' decomposition method to solve the resulting stochastic program. All integer variables must then be included in the master problem at time 0.

A nonlinear objective function would arise if one chooses to use a nonlinear utility function instead of the simple scalars λ_1 and λ_2 to evaluate the final portfolio value in the objective function. Mulvey and Vladimirou [46] include utility functions in their stochastic programming models for portfolio optimization, and they apply the progressive hedging algorithm of Rockafellar and Wets [52] with apparent success to solve these models. One could also approximate the utility function by a piecewise linear function, in which case the model can still be solved by linear programming methods.

Another alternative for the evaluation of the final portfolio value in the objective function is to view the scalars λ_1 and λ_2 as Lagrange multipliers on constraints that specify a desired final portfolio value in each scenario[1] (the weights would then ob-

[1] J.F. Shapiro, personal communication.

viously be scenario dependent). Lagrangian relaxation techniques could be used to determine values for these multipliers, which then provide an indication of the trade-off in cost of realizing the desired portfolio value in different scenarios. This would in essence create a nested solution approach to the ALM model, in which values for the multipliers are determined and iteratively updated in the outer loop, and the ALM model is solved inside the loop.

In our model development we have made the simplifying assumption that interest-rate uncertainty is the only source of uncertainty, and we have only discussed the one-factor term-structure model of Ho and Lee [31] in some detail. The structure of the ALM model does not change, however, if term-structure models with more than one factor are considered (on which we briefly commented in chapter 3), or other sources of uncertainty are included. The only real restriction is that the uncertainty can be represented as an event tree in discrete time with a finite number of states at each time.

Grabbe [49] describes an extension of the Black-Scholes option-pricing model to the valuation of options on foreign exchange. His model includes uncertainty in both the interest rate and the foreign-exchange rate, and could therefore be used to add currency risk to the ALM model. Although Grabbe's model is a continuous-time model, the method of Hull and White [35] can be applied to obtain a discretized version that can be used in the ALM model (see also Javaheri [38]).

A challenging research question is how to add equity investments, and thus stock-market uncertainty, to the ALM model. To our knowledge, no models have been proposed in the literature that combine a description of interest-rate and stock-market uncertainty. A major difficulty in the construction of a model that describes the uncertainty in stock prices is the fact that there is no single factor that determines stock prices as much as the interest rate does in the case of fixed-income securities. Any reasonable description is therefore likely to involve several (if not many) different factors. As the size of an event tree tends to increase exponentially with the number of stochastic factors, a large number of such factors easily leads to event trees of unmanageable dimensions.

As a first approximation, one could therefore model the risk in some stock-market *index* instead of individual stocks. In recent years, large markets for trading in options and futures contracts on stock market indices have been established, and these derivative securities could therefore be included in the resulting ALM model. A solution to the ALM model would give an indication of the optimal division in the asset

portfolio between fixed-income and equity investments. Different models may then be used to select individual stocks within the equity component of the portfolio.

Extension of the Solution Approach

The iterative disaggregation algorithm that we have proposed to solve the ALM problem is a novel approach to the solution of stochastic programs in general. In most cases, stochastic programming models are solved for some particular and approximate description of the uncertainty only. Sometimes bounds can be calculated on the deviation of the obtained solution from the "true" optimum solution (corresponding to the description of the uncertainty from which the approximate description in the stochastic program is derived). These bounds may be fairly loose in practice, however, and only provide information on the objective value, and not on the sensitivity of the optimal solution. We have shown that the iterative disaggregation algorithm supplies information on both.

Although we have described the algorithm and the aggregation methods on which it is based specifically for the asset/liability management model that is the subject of study in this thesis, the underlying ideas can be applied to any stochastic program. We have shown that the aggregations and disaggregations have to be performed carefully in the ALM model in order to prevent arbitrage opportunities, which can substantially bias the optimal solution. In other applications this may not be a concern, in which case more freedom exists in how to perform the disaggregations. Furthermore, if the stochasticity in the stochastic program is restricted to the right-hand-side vector, then Benders' decomposition method can be used efficiently to perform the reoptimizations in the course of the algorithm, as the Benders' cuts from one iteration remain valid for the next iteration in that case.

Appendix A

Asset Valuation by Arbitrage

This appendix explains the basic concepts of the theory on asset valuation by arbitrage. This theory started with the option pricing models of Black and Scholes [6] and Merton [44], but has since developed in a complete theory on the structure of financial markets. Our overview here is adapted from Huang and Litzenberger [32], to which we also refer for proofs of the theorems and propositions. We assume throughout that perfect market conditions prevail: there are no transaction costs or taxes, securities are infinitely finely divisible, interest rates for borrowing and lending are the same, and short sales of assets with full use of proceeds are allowed.

A.1 Market Equilibrium and Arbitrage

We consider a multiperiod economy in which one of a finite number of *scenarios*, collectively referred to as the scenario space Ω, will occur. Over time, information becomes available as to which scenario will occur, and this process of information revelation is formally denoted by the *information structure* $\mathbf{F} = \{\mathcal{F}_t; t = 0, 1, \ldots, T\}$. We assume that $\mathcal{F}_0 = \{\Omega\}$, and that the exact scenario will not be known until the final date T. The information structure can be represented in the form of an *event tree*, as in figure A-1 for a two-period economy with five possible scenarios ($\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$). In this example, $\mathcal{F}_0 = \{\Omega\}, \mathcal{F}_1 = \{\{\omega_1, \omega_2\}, \{\omega_3, \omega_4, \omega_5\}\}$, and $\mathcal{F}_2 = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}$. We will refer to the elements of $\mathcal{F}_t$ for each t as *states* or *events*.

We assume that prices and dividends for all securities in the economy are *adapted* to the information structure, i.e., that they can be written as functions of the states in the event tree. We furthermore assume that all investors in the economy agree on

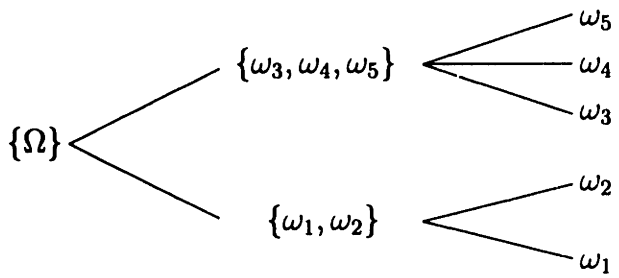

Figure A-1: Information structure in a 2-period economy.

these mappings, and that each of them assigns a strictly positive probability to each possible state. However, they may disagree on the magnitude of the probabilities and thus the likelihood of occurrence of each scenario.

An important question is whether the system of state-dependent security prices, as agreed upon by all investors, is a "reasonable" one. Central to its answer is the notion of an arbitrage opportunity.

Definition A.1 (Arbitrage opportunity) *An arbitrage opportunity is present in the economy if there exists a self-financing trading strategy*[1] *whose payoffs are non-negative everywhere and strictly positive in at least one state $a_t \in \mathcal{F}_t$ for some t, and whose initial investment is nonpositive.*

It is clear that investors would engage in such a trading strategy as much as possible (assuming they always prefer more to less), and such a trading opportunity therefore cannot exist if markets are in equilibrium. We will henceforth assume that markets are in equilibrium, and thus that arbitrage opportunities cannot be present. The following theorem, due to Harrison and Kreps [23], provides an important characterization of the absence of arbitrage opportunities in the economy.

Theorem A.1 *Asset prices in the economy do not admit arbitrage opportunities if and only if there exists an equivalent probability measure*[2] *on $\mathbf{F}$ such that the one-period expected return in any period with respect to this probability measure is identical for all assets.*

[1] A trading strategy is self-financing if it does not require any cash inflows after time 0.

[2] Two probability measures are equivalent if they have the same sets of nonzero probability.

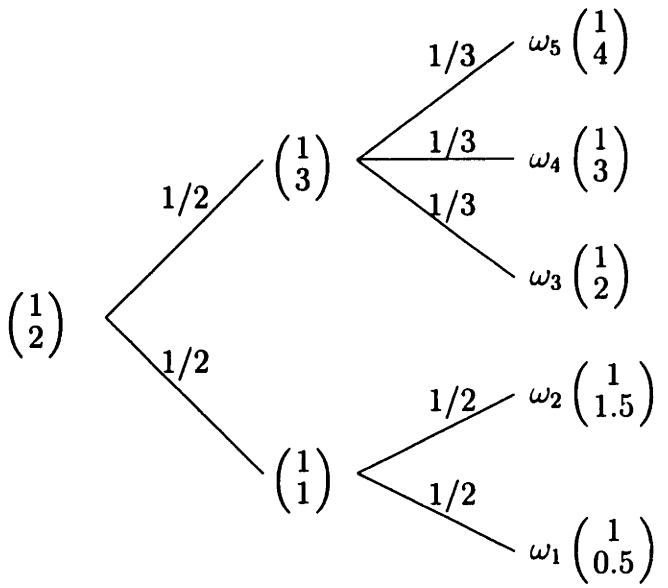

Figure A-2: A two-period economy with two securities.

This theorem can be restated as saying that arbitrage opportunities are excluded if and only if there exists an equivalent probability measure on $\mathbf{F}$ such that *relative* asset prices are martingales with respect to this probability measure. This is why the equivalent probability measure of A.1 is often called the *equivalent martingale measure.* It is also sometimes referred to as the risk-neutral probability measure, as it could represent the probability beliefs of a risk-neutral investor in equilibrium.

Figure A-2 depicts the cum-dividend prices (i.e., prices plus accumulated dividends) of two securities in the economy of figure A-1. The numbers on the arcs represent an equivalent martingale measure, and thus theorem A.1 tells us that the security prices are arbitrage-free.

A.2 Asset Pricing by Arbitrage

In addition to being a proof that no arbitrage opportunities exist, the equivalent martingale measure also simplifies the calculation of arbitrage-free security prices in the event tree. This section shows how.

Consider a security S^* in the economy whose dividend process is adapted to $\mathbf{F}$. Security S^* is said to be *marketed* if there exists a self-financing trading strategy in the other securities whose payoffs exactly match the dividends of S^*. (This trading

strategy is said to *finance* security S^*, and is also called a *replicating* trading strategy for security S^*.) We note that all investors will agree on whether S^* is marketed, and that this does not depend on investors' individual probability beliefs. It should be clear that if no arbitrage opportunities are allowed, the price of security S^* at time 0 must be equal to the initial cost of the replicating portfolio, and its price at later dates to the corresponding value of the replicating portfolio.

Consider again the two-period economy of figure A-2 with two securities, and let S^* be a new security which pays dividends of 2, 4, 2, and 1 in states ω_1, ω_2, ω_3, and ω_4 at time 2, respectively, and nothing in all other states. S^* is marketed as a self-financing trading strategy in the two existing securities can be constructed that exactly provides the payoffs from S^*. If α and β represent the portfolio holdings of the two securities in this replicating trading strategy, then $(\alpha, \beta) = (4, -1)$ at time 0 and in state $\{\omega_3, \omega_4, \omega_5\}$ at time 1, and $(\alpha, \beta) = (1, 2)$ in state $\{\omega_1, \omega_2\}$ at time 1 (negative holdings represent short sales). The price of security S^* must therefore be 2 at time 0, 3 in state $\{\omega_1, \omega_2\}$ at time 1 and 1 in state $\{\omega_3, \omega_4, \omega_5\}$ at time 1 to prevent arbitrage opportunities.

Alternatively, we can calculate the arbitrage-free prices of security S^* through the equivalent martingale measure. Notice that the one-period conditional expected return on the two existing securities in figure A-2 with respect to the equivalent martingale measure is zero in all states. By theorem A.1, this must also be true for S^* to prevent arbitrage opportunities. Thus the price of S^* must equal 3 in state $\{\omega_1, \omega_2\}$ at time 1 and 1 in state $\{\omega_3, \omega_4, \omega_5\}$ at time 1. Using these time 1 prices, it follows that the price at time 0 must be equal to 2.

This way of valuing securities is called *asset pricing by arbitrage*. We note that it depends crucially on whether securities are marketed; if a security is not marketed, then its price cannot be uniquely determined. The next section presents a sufficient condition for *all* securities to be marketed.

A.3 Complete Markets

The ability to replicate any state-dependent payoff pattern is captured by the definition of dynamically complete markets.

Definition A.2 (Dynamically complete markets) *Markets are dynamically complete if every state-dependent payoff pattern can be obtained exactly by a self-financing trading strategy in the available securities.*

Markets that are not dynamically complete are called *incomplete.* In the economy of figure A-2, it is not possible to construct a self-financing trading strategy in the existing two assets that gives a payoff of 1 in state ω_5 at time 2, and nothing in all other states. The security markets in this economy are therefore incomplete. The next proposition gives a necessary and sufficient condition for markets to be dynamically complete.

Proposition A.1 *Security markets in a multiperiod economy are dynamically complete if and only if at every node in the event tree that represents the information structure, the number of traded securities having linearly independent random returns over the next period equals the number of branches leaving the node.*

It is clear that this condition is violated in state $\{\omega_3, \omega_4, \omega_5\}$ at time 1 in figure A-2. We would need one more security, whose returns in states ω_3, ω_4 and ω_5 at time 2 are linearly independent of the returns on the two existing securities, to complete the markets in this example.

If the price system in an economy does not admit arbitrage opportunities, then the following proposition gives another characterization of dynamically complete markets.

Proposition A.2 *If an equivalent martingale measure exists in a securities markets economy, it is unique if and only if markets are dynamically complete.*

Figure A-3 depicts the two-period economy of figure A-2 with an additional security that completes the markets. Moreover, the cum-dividend prices of this third security in the event tree are such that the equivalent martingale measure of figure A-2 still is an equivalent martingale measure. From proposition A.2, this equivalent martingale measure is therefore unique.

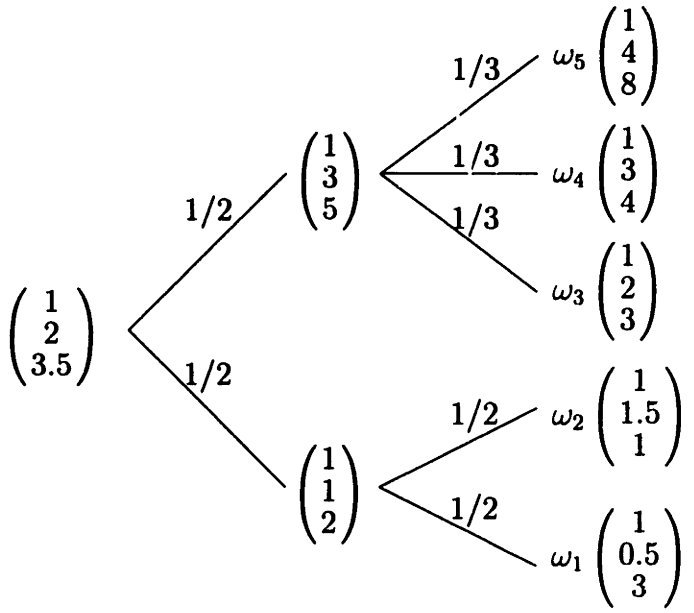

Figure A-3: A two-period economy with three securities.

Bibliography

[1] J.F. Benders. Partitioning procedures for solving mixed-variables programming problems. *Numerische Mathematik*, 4:238–252, 1962.

[2] D. Bienstock and J.F. Shapiro. Optimizing resource acquisition decisions by stochastic programming. *Management Science*, 34(2):215–228, February 1988.

[3] J.R. Birge. Decomposition and partitioning methods for multi-stage stochastic linear programs. *Operations Research*, 33(5):989–1007, Sept.-Oct. 1982.

[4] J.R. Birge. The value of the stochastic solution in stochastic linear programs with fixed recourse. *Mathematical Programming*, 24:314–325, 1982.

[5] F. Black, E. Derman, and W. Toy. A one-factor model of interest rates and its application to treasury bond options. *Financial Analysts Journal*, pages 33–39, January-February 1990.

[6] F. Black and M. Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81:637–659, May-June 1973.

[7] S.P. Bradley and D.B. Crane. A dynamic model for bond portfolio management. *Management Science*, 19(2):139–151, October 1972.

[8] M.J. Brennan and E.S. Schwartz. An equilibrium model of bond pricing and a test of market efficiency. *Journal of Financial and Quantitative Analysis*, 17:301–330, September 1982.

[9] D.R. Cariño, T. Kent, D.H. Myers, C. Stacy, M. Sylvanus, A.L. Turner, K. Watanabe, and W.T. Ziemba. Russell-yasuda kasai model: An asset/liability model for a japanese insurance company using multistage stochastic programming. to appear in *Interfaces*, 1993.

[10] J.C. Cox and C.-f. Huang. Optimal consumption and portfolio policies when asset prices follow a diffusion process. *Journal of Economic Theory*, 49:33–83, 1989.

[11] J.C. Cox, J.E. Ingersoll, and S.A. Ross. A theory of the term structure of interest rates. *Econometrica*, 53:385–408, 1985.

[12] G.B. Dantzig. Linear programming under uncertainty. *Management Science*, 1:197–206, 1955.

[13] G.B. Dantzig and G. Infanger. Multi-stage stochastic linear programs for portfolio optimization. Technical Report SOL 91-11, Systems Optimization Laboratory, Stanford University, Stanford CA 94305, September 1991.

[14] P.H. Dybvig and J.E. Ingersoll, Jr. Mean-variance theory in complete markets. *Journal of Business*, 55(2):233–251, 1982.

[15] C. Edirisinghe, V. Naik, and R. Uppal. Optimal replication of options with transaction costs and trading restrictions. *Journal of Financial and Quantitative Analysis*, 28(1):117–138, March 1993.

[16] G.D. Eppen and E.F. Fama. Three asset cash balance and dynamic portfolio problems. *Management Science*, 17(5):311–319, January 1971.

[17] Yu. Ermoliev and R. J.-B. Wets, editors. *Numerical Techniques for Stochastic Optimization*. Springer Verlag, 1988.

[18] H.I. Gassmann. MSLiP: A computer code for the multistage stochastic linear programming problem. *Mathematical Programming*, 47:407–423, 1990.

[19] R.C. Grinold. Steepest ascent for large scale linear programs. *SIAM Review*, 14(3):447–464, July 1972.

[20] R.C. Grinold. Model building techniques for the correction of end effects in multistage convex programs. *Operations Research*, 31(3):407–431, May-June 1983.

[21] S.J. Grossman and J.-L. Vila. Optimal dynamic trading with leverage constraints. *Journal of Financial and Quantitative Analysis*, 27(2):151–168, June 1992.

[22] M.A. Hamilton. A comparison of quantitative methods for engineering synthetic options. Master's thesis, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1993.

[23] M.J. Harrison and D.M. Kreps. Martingales and arbitrage in multiperiod securities markets. *Journal of Economic Theory*, 20:381–408, 1979.

[24] H. He and N.D. Pearson. Consumption and portfolio policies with incomplete markets and short-sale constraints: The finite-dimensional case. Technical report, Institution of Business and Economic Research, University of California at Berkeley, September 1989.

[25] H. He and N.D. Pearson. Consumption and portfolio policies with incomplete markets and short-sale constraints: The infinite-dimensional case. *Journal of Economic Theory*, 54:259–304, August 1991.

[26] D. Heath, R. Jarrow, and A. Morton. Bond pricing and the term structure of interest rates: A new methodology for contingent claims valuation. Technical report, Cornell University, 1987.

[27] D. Heath, R. Jarrow, and A. Morton. Bond pricing and the term structure of interest rates: A discrete time approximation. *Journal of Financial and Quantitative Analysis*, 25(4):259–304, December 1990.

[28] R.S. Hiller and J. Eckstein. Stochastic dedication: Designing fixed income portfolios using massively parallel benders decomposition. *Management Science*, 39(11):1422–1438, November 1993.

[29] R.S. Hiller and C. Schaack. A classification of structured bond portfolio modeling techniques. *Journal of Portfolio Management*, pages 37–48, Fall 1990.

[30] R.S. Hiller and J.F. Shapiro. Stochastic programming models for asset/liability management problems. Technical report, International Financial Services Research Center, Sloan School of Management, M.I.T., August 1989.

[31] T. Ho and S.-B. Lee. Term structure movements and pricing interest rate contingent claims. *Journal of Finance*, 41(5):1011–1029, December 1986.

[32] C.-f. Huang and R.H. Litzenberger. *Foundations for Financial Economics.* North-Holland, 1988.

[33] J. Hull. *Options, Futures and Other Derivative Securities.* Prentice-Hall, 1989.

[34] J. Hull and A. White. Pricing interest-rate-derivative securities. *The Review of Financial Studies*, 3(4):573–592, 1990.

[35] J. Hull and A. White. Valuing derivative securities using the explicit finite difference method. *Journal of Financial and Quantitative Analysis*, 25(1):87–99, March 1990.

[36] J. Hull and A. White. One-factor interest-rate models and the valuation of interest-rate derivative securities. *Journal of Financial and Quantitative Analysis*, 28(2):235–254, 1993.

[37] G. Infanger. *Planning Under Uncertainty: Solving Large-Scale Stochastic Linear Programs.* Boyd & Fraser, 1994.

[38] A. Javaheri. Application of the explicit finite difference method to the pricing of call options on foreign exchange. Master's thesis, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1994. in preparation.

[39] T.H. Johnsen and J.B. Donaldson. The structure of intertemporal preferences under uncertainty and time-consistent planning. *Econometrica*, 53(6):1451–1458, November 1985.

[40] M.I. Kusy and W.T. Ziemba. A bank asset and liability management model. *Operations Research*, 34(3):356–376, May-June 1986.

[41] H.M. Markowitz. *Portfolio Selection.* Basil Blackwell, second edition, 1991.

[42] R.C. Merton. Lifetime portfolio selection under uncertainty: The continuous-time case. *Review of Economics and Statistics*, 51:247–257, August 1969.

[43] R.C. Merton. Optimum consumption and portfolio rules in a continuous-time model. *Journal of Economic Theory*, 3:373–413, December 1971.

[44] R.C. Merton. Theory of rational option pricing. *Bell Journal of Economics and Management Science*, 4:141–183, Spring 1973.

[45] R.C. Merton. *Continuous-Time Finance.* Basil Blackwell, revised edition, 1992.

[46] J.M. Mulvey and H. Vladimirou. Stochastic network programming for financial planning problems. *Management Science*, 38(11):1642–1664, November 1992.

[47] K.G. Murty. *Linear Programming.* John Wiley & Sons, 1983.

[48] G.L. Nemhauser and L.A. Wolsey. *Integer and Combinatorial Optimization.* Wiley-Interscience, 1988.

[49] J. Orlin Grabbe. The pricing of call and put options on foreign exchange. *Journal of International Money and Finance*, 2:239–253, 1983.

[50] P.H. Ritchken. Enhancing mean-variance analysis with options. *The Journal of Portfolio Management*, pages 67–71, Spring 1985.

[51] R.T. Rockafellar. *Convex Analysis.* Princeton University Press, 1970.

[52] R.T. Rockafellar and R. J.-B. Wets. Scenarios and policy aggregation in optimization under uncertainty. *Mathematics of Operations Research*, 16(1):119–147, February 1991.

[53] J.F. Shapiro. *Mathematical Programming: Structures and Algorithms.* John Wiley & Sons, 1979.

[54] R. Van Slyke and R. J.-B. Wets. L-shaped linear programs with applications to optimal control and stochastic programming. *SIAM Journal of Applied Mathematics*, 17(4):638–663, July 1969.

[55] O. Vasicek. An equilibrium characterization of the term structure. *Journal of Financial Economics*, 5:177–188, 1977.

[56] R. J.-B. Wets. Solving stochastic programs with simple recourse. *Stochastics*, 10:219–242, 1984.

[57] P. Zipkin. Bounds for row-aggregation in linear programs. *Operations Research*, 28(4):903–916, July-August 1980.

[58] P. Zipkin. Bounds on the effect of aggregating variables in linear programs. *Operations Research*, 28(2):403–418, March-April 1980.

[59] P. Zipkin. The structure of structured bond portfolio models. *Operations Research*, 40(Supplement 1):S157–S169, January-February 1992.