Pricing and Hedging Gmwb Riders in a Binomial Framework

Menachem Wenger

A Thesis

in the Department
of
Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Science (Mathematics) at
Concordia University
Montreal, Quebec, Canada

September 2012

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Menachem Wenger**

Entitled: **Pricing and Hedging GMWB Riders in a Binomial Framework**

and submitted in partial fulfillment of the requirements for the degree of

Master of Science (Mathematics)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

______________________________ Examiner
Dr. Patrice Gaillardetz

______________________________ Examiner
Dr. Jose Garrido

______________________________ Thesis Supervisor
Dr. Cody Hyndman

Approved by ______________________________
Chair of Department or Graduate Program Director

____
Dean of Faculty

Date ______________________________

ABSTRACT

Pricing and Hedging GMWB Riders in a Binomial Framework

Menachem Wenger

The guaranteed minimum withdrawal benefit (GMWB) rider guarantees the return of premiums in the form of periodic withdrawals while allowing policyholders to participate fully in any market gains. The product has evolved into a lifetime version (GLWB) and is a vital component of the variable annuity marketplace, representing asset values of $294B as of September 2011.

GMWB riders represent an embedded option on the account value with a fee structure that is different from typical financial derivatives. We present an in-depth study into pricing and hedging the GMWB rider from a financial economic perspective. Our main contributions are twofold. We construct a binomial asset pricing model for GMWBs under optimal policyholder behaviour which results in explicitly formulated perfect hedging strategies in a binomial world. The numerical toolbox for pricing GMWBs in a Black-Scholes world is expanded to include binomial methods.

To motivate our work, we begin with a review of the continuous model and a comprehensive synthesis of results from the literature. Throughout, particular focus is placed on the unique perspectives of the insurer and policyholder and the unifying relationship. We also present an approximation algorithm that significantly improves efficiency of the binomial model while retaining accuracy. Several numerical examples are provided which illustrate both the accuracy and the tractability of the model.

Finally, we explore the effect of deterministic mortality on pricing GMWBs, and run mortality simulations to obtain hedging results which support the diversification principle.

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisor Dr. Cody Hyndman for his patience, guidance, and support throughout my masters studies. Prof. Hyndman encouraged me to select my own research topic, at the same time was always there to provide direction and valuable feedback and suggestions.

I would like to thank Dr. Jose Garrido and Dr. Patrice Gaillardetz for serving on the thesis examining committee and for their insightful comments.

I would also like to thank the Department of Mathematics and Statistics, the faculty members, and graduate students, many of whom have have greatly enhanced my undergraduate and graduate studies at Concordia.

This thesis would not have been possible without the generous financial support of the Natural Science and Engineering Council of Canada, le Fonds de recherche du Québec - Nature et technologies, and Concordia University.

A special mention goes to my former colleagues at Guardian Life Insurance Company of America from whom I learned much about GMWBs. My practical experience with GMWBs was a big asset in writing this thesis.

My deepest gratitude is reserved for my family, and especially my mom, for all their love and encouragement.

Above all, I thank G-d for all the blessings He has granted me and for giving me this opportunity.

In loving memory of my father

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

The variable annuity marketplace has seen tremendous growth in sales since the early 1990's. The growth has corresponded to the increase in product offerings, both in terms of the variable annuity (VA) base contracts and the accompanying riders. Riders are optional add-ons to VAs, providing additional benefits in return for which an additional charge is subtracted annually from the account value (AV).

Variable deferred annuities have two phases: the accumulation period and the annuitization period. During the accumulation period, premiums are deposited with the insurer and can be actively managed by the policyholder to achieve his investment goals by allocating the funds to a selection of investment funds. The policyholder may choose to take partial withdrawals and/or surrender the contract, although the proceeds will likely be subject to contingent deferred sales charges (CDSC), more commonly referred to as surrender charges (SC), and possible tax penalties depending on the age of the policyholder. Upon annuitization the policyholder cedes control over the funds and in return is guaranteed a periodic stream of payments. This phase protects annuitants from longevity risk. The duration of the guaranteed period may range from a fixed number of years (term certain) up to guaranteed for life. The policyholder may choose the payments to be fixed or variable. In the latter case, they will fluctuate based partially on the performance of certain market funds.

The first riders introduced to the VA market were death benefit riders: these guarantee a minimum death benefit to the beneficiaries if the policyholder dies during the accumulation period. Initially offering a simple return of premium, the benefits evolved to offer increasingly rich guarantees in the form of annual roll-ups and highest anniversary values. The next form of riders introduced were guaranteed living benefits (GLBs). The guaranteed minimum accumulation benefit riders (GMABs) guarantee a minimum account value at a specific date (i.e. 10 years from issue date), while the guaranteed minimum income benefit riders (GMIBs) guarantee a minimum annuitization amount by giving policyholders the choice between annuitizing a higher guarantee base at contractually specified annuitization rates or the current account value at the current annuitization rates. The contractual annuitization rates are generally conservative and can be expected to lie below current rates.

Guaranteed minimum withdrawal benefit riders (GMWBs) were introduced in 2002 and guarantee the policyholder will recover at least the total premiums paid into the policy in the form of periodic withdrawals, subject to the annual withdrawals not exceeding a contractual percentage of the premiums. By allowing policyholders to remain in the accumulation phase and retain full control of their investments, policyholders reap the upside potential from equity investments while being protected from downside risk. GMWBs evolved into the guaranteed lifetime withdrawal benefit riders (GLWBs) which guarantee the annual maximal withdrawals for life, thereby introducing a feature of the annuitization phase into the accumulation phase. GMWB and GLWB riders represent embedded financial put options on the account values and techniques from mathematical finance are needed to value these contracts.

The fee structures of these riders add complexity to pricing and risk management processes, relative to the standard financial equity market derivatives where a single upfront premium is charged which has no impact on the future random payoffs. Consistent with the fee structure of VAs, no upfront fees are charged for GMWB riders. Rather, fees are deducted periodically from the AV to pay for the rider where the fees are proportional to the AV. The AV is influenced by the withdrawal behaviour of the policyholder and revenue flow from fees stops in the event of death or surrender. As such there are multiple sources of uncertainty involved in the actual fees to be received. Another subtle impact of the fee structure is that an increase in the fee rate results in higher annual fee income but it also creates a drag on the AV, potentially causing it to reach zero faster which results in earlier termination of fee revenues and increased rider guarantee payouts.

The GLB riders have grown increasingly complex in recent years. Added features range from periodic ratchets and annual roll-ups to specific one-time bonuses if certain criteria are met. While these features were designed to increase the product appeal, they were also designed to entice policyholders to keep their funds in the accounts for longer periods of time to the benefit of the insurer.

Table 1.1 displays the growth figures in annual gross VA sales in the United States over the past two decades. This aligns with the increase in rider offerings. There was a decline in sales following the financial crisis of 2008 but the past two years has seen positive growth figures. A report from LIMRA Retirement Research, November 2011 (LIMRA, 2011), shows an 88% election rate of GLB riders for VAs offering GLB riders for the 3rd quarter of 2011[1]. During the period Jan. 2009 - Sept. 2011, this quarterly election rate ranged from 87% to 90%. Further, 91% of new VA sales in the 3rd quarter of 2011 offered GLB riders. Of the GLBs elected that quarter, 65% were GLWBs. As of September 2011, 55% of all VA assets with GLB elected - both new policies and in-force policies - were GLWBs. This represents an asset value of $294

[1] These percentages are all premium-dollar weighted.

Year	Sales ($)
1992	40
2002	117
2007	184
2008	156
2009	128
2010	140
2011	159

Table 1.1: Variable annuity sales (billions $). Sources: LIMRA (2012)

billion.

It is our belief that the GMWB and GLWB riders are not treated by insurers as a source of direct profit but rather as a tool to drive sales of VAs and their accompanying profits. We will point out in the literature review the consensus among the early papers that these riders were underpriced, supporting this hypothesis that they were only a means to increase VA sales. Indeed, reinsuring all or most of the risk was a popular risk management strategy for the initial GMWB products. Reinsurance premiums increased as reinsurers became more informed of the high risk embedded in these products. Around the time of the financial crisis in 2008 reinsurers stopped offering coverage altogether on GMWB and GLWB riders at which point the importance of internal dynamic hedging programs rose rapidly.

With this in mind, we look at pricing and hedging the GMWB product in a simplified framework consistent with the no-arbitrage principle from financial economics. It is evident that the GLWB riders have come to define the VA market. The GMWBs were the precursor to the GLWBs and as such, a mathematical analysis of the GMWB product is interesting in its own right, even if the GMWB product is no longer a dominant force in the market per se.

1.2 Product Specifications

We introduce the product specifications and notation. At time $t = 0$, a policy (an underlying VA contract plus a GMWB rider) is issued to a policyholder of age x and an initial premium P is received. We assume no subsequent premiums. The premium is invested into a fund which perfectly tracks a risky asset $S = \{S_t; t \geq 0\}$ with no basis risk. One may think of the underlying funds as being deposited in a mutual fund and $\{S_t\}$ as the index tracked by it. The rider fee rate α is applied to the account value $W = \{W_t; t \geq 0\}$. Fees are deducted from the account value (continuously or periodically depending on the model) as long as the contract is in force and the account value is positive.

A guaranteed maximal withdrawal rate g is contractually specified and up to the amount $G := gP$ can be withdrawn annually[2] until P is recovered through cumulative withdrawals (ignoring time value of money), regardless of the evolution of $\{W_t\}$. The policyholder also receives any remaining account value at maturity.

Policyholders have the option of withdrawing any amount provided it does not exceed the remaining account value. If the account value hits zero, then the policyholder receives withdrawals at rate G until the initial premium has been recovered. If annual withdrawals exceed G while the account value is still positive, then a surrender charge is applied to the withdrawals and a reset feature may reduce the guarantee value, i.e. the remaining portion of the initial premium not yet recovered. Policyholders also have the option of surrendering[3] early and receiving the account value less a surrender charge. Any guarantee value is forfeited by surrendering.

Assuming a static withdrawal strategy where G is withdrawn annually (continu-

[2]Contract specifications vary widely by insurer but extra features such as ratchets and rollups may be present which cause potential increases to the balance guaranteed to the policyholder. Consequently, G may increase depending on market performance and withdrawal behaviour but will not decrease. In this case G is a function of $\{A_t\}$ where A is a fictional account representing the GMWB guarantee balance. In our simplified contract where G is constant there is no need to introduce this additional dimension A.

[3]The terminology of lapses and surrenders are used interchangeably.

ously or discretely), we set the maturity $T := 1/g$ since the sum of all withdrawals at T is $TG = G/g = P$. At time T the rider guarantee is worthless and the policyholder receives a terminal payoff of the remaining account value, if it is positive. Essentially, this assumption translates over to a real-world trend of no annuitizations. This assumption is partially justified as VAs are not usually maintained through to annuitization.

1.3 Literature Review

There has been increased research into pricing and hedging GMWB products since the initial paper on the topic by Milevsky and Salisbury (2006). In this section we discuss a few of the more relevant works.

Working with continuous withdrawals and a standard geometric Brownian motion model for $\{S_t\}$, Milevsky and Salisbury (2006) consider two policyholder behaviour strategies. Under a static withdrawal strategy and no lapses the contract is decomposed into a term certain component and a Quanto Asian Put option with the numeraire being a modified account value process. Numerical PDE methods are used to evaluate the ruin probabilities for $\{W_t\}$ and the contract value V_0. A dynamic behaviour strategy is considered where optimal withdrawals occur. A set of linear complementarity equations is derived for this free boundary value problem and solved numerically for V_0. It is found that the optimal strategy reduces to withdrawing G continuously unless W_t exceeds a boundary value depending on the remaining guarantee balance of $P - Gt$, in which case an arbitrarily large withdrawal rate is taken and the policyholder should lapse. Milevsky and Salisbury (2006) conclude that the GMWB riders in effect in 2004 were underpriced relative to the capital markets cost.

The optimal behaviour approach is formalized in Dai et al. (2008) where the contract value process $\{V_t\}$ is formulated as the solution to a singular stochastic control problem with the control variable being the withdrawal rate. Unlike in Milevsky and Salisbury (2006), time dependency and a complete description of the auxiliary conditions are included in this model. To facilitate numerical solutions for the HJB equations a penalty approximation formulation is solved using finite difference methods which converge to the viscosity solution.

Consistent with Milevsky and Salisbury (2006), numerical results provide support that the provision for optimal behaviour is quite valuable and insurers appeared to be underpricing GMWB riders. The optimal strategy consists of withdrawing at rate G (continuously) except for in certain regions of the state space where an infinite withdrawal rate is optimal, which means to "withdraw an appropriate finite amount instantaneously making the equity value of the personal account and guarantee balance to fall to the level that it becomes optimal for him to withdraw [G]" (Dai et al., 2008). However, Dai et al. (2008) allow the policyholder the option of withdrawing any amount of the unrecovered initial premium, even if it exceeds the account value. In other words, if the account value is zero, the policyholder can elect to receive the remaining guarantee balance instantly subject to surrender charges rather than receive G annually. The impact of this assumption is amplified by not including a reset feature in most of their work. The combination of this is the main cause of arriving at optimal strategies differing from Milevsky and Salisbury (2006).

Chen and Forsyth (2008) extend Dai et al. (2008) to an impulse control problem representation where the control set allows for continuous withdrawal rates not exceeding G and instantaneous finite withdrawals. This allows for modeling more realistic but complex product features.

Bauer et al. (2008) develop an extensive and comprehensive framework to price any of the common guarantees available with VAs, assuming that any policyholder events such as surrenders, withdrawals, or death occurs at the end of the year. Deterministic mortality is assumed. Monte-Carlo simulation is used to price the contracts assuming a deterministic behaviour strategy for the policyholders. To price the contracts assuming an optimal withdrawal strategy, a quasi-analytic integral solution is derived and an algorithm is developed by approximating the integrals using a multidimensional discretization approach via a finite mesh. Hence, only a finite subset of all possible strategies are considered. One drawback is that the valuation with optimal behaviour for a single contract could take up to 40 hours (for a 25 year maturity).

Allowing for discrete withdrawals, Bacinello et al. (2011) consider a number of guarantees under a more general financial model with stochastic interest rates and stochastic volatility in addition to stochastic mortality. In particular for GMWBs, a static behaviour strategy (G withdrawn annually and no lapses) is priced using standard Monte Carlo whereas an optimal lapse approach (G withdrawn annually) is priced with a Least Squares Monte Carlo algorithm.

Upper and lower bounds on the price process for the GMWB are derived in Peng et al. (2012) under stochastic interest rates and assuming a static continuous withdrawal strategy of G per year with no lapses. This paper was instrumental to the development of our work because of a tangential result about the relationship between the insured and insurer perspectives.

Ignoring mortality and working with a static withdrawal assumption and no lapses, the primary focus of Liu (2010) is on developing semi-static hedging strategies under both a geometric Brownian motion model and a Heston stochastic volatility model for the underlying asset $\{S_t\}$. However, sufficient attention and detail is paid to pricing the GMWB rider assuming the insured takes constant withdrawals of G/n at the end of each period where there are n time steps per year. Liu (2010) observes that the contract (GMWB plus VA) can be decomposed into a term certain component and a floating strike Asian Call option on a modified process. Both a Monte Carlo approach and a moment-matching lognormal approximation method (based on Levy, 1992) are used to obtain results for increasing n.

1.4 Thesis Overview

In the literature review we pointed out that a range of methods have been applied to price GMWBs under varying policyholder behaviour assumptions. Under a static withdrawal strategy with no lapses the methods include numerical PDE techniques, Monte Carlo simulation, and moment matching analytical approaches. Modeling optimal withdrawal behaviour the methods include more advanced numerical PDE techniques, numerical integration methods, and a Least Squares Monte Carlo approach.

Based on the product specifications listed in Section 1.2, optimal withdrawal behaviour reduces to withdrawing at rate G or lapsing. The rider guarantee represents an intangible and fictional amount. Once the account value is zero, this amount is accessible only through withdrawals at rate G, a product specification adopted by both Milevsky and Salisbury (2006) and Bacinello et al. (2011). The work of Dai et al. (2008) and Chen and Forsyth (2008) do not reflect this and therefore different results are obtained.

Our contributions in this thesis are twofold. In a binomial world we set up an asset pricing model for GMWBs assuming optimal behaviour and construct explicit hedging strategies. In a Black-Scholes world, we expand the numerical toolbox for pricing GMWBs to include binomial tree-based methods. Although in theory the results should converge to those of the continuous withdrawal model with S lognormally distributed; due to the non-recombining nature of the account value the suggested method is found to be numerically expensive. We substantially improve the numerical efficiency without sacrificing significant accuracy of results by adopting an approximation method based on Costabile et al. (2006).

A binomial valuation approach has previously been considered by Bacinello (2005) to price equity-linked life insurance with recurring premiums in the presence of early surrenders. Although the underlying methodology is similar, we deal with the unique features and challenges of modeling GMWB riders for variable annuities. In addition to surrender and mortality, both elements considered by Bacinello (2005), we have an endogenously determined trigger date. The nature of the fees and withdrawals further differentiate our work. Whereas Bacinello (2005) deals exclusively with pricing, we pay equal attention to the hedging constructions in a binomial model, which is facilitated by the consideration of the unique perspectives of the insurer and insured. By focusing on a single product we have the liberty to consider a top-down approach which provides more insight than generic formulations of backward induction schemes.

GMWB and GLWB carriers are exposed to three major types of risk: financial market, mortality, and policyholder behaviour. The two dominant financial market risks are equity market risk, namely poor market performance, and interest rate risk primarily in a low interest rate environment.

A recent quote shows how critical financial market risk is to insurers: "Since interest rates have been low and the stock market volatile, insurers like MetLife and Prudential have lessened their variable annuity business. Sun Life Financial, out of Canada, actually left the variable annuity business altogether"[4]. In this thesis we begin by considering equity risk, then incorporate behaviour risk and finally we consider deterministic mortality models. We do not model the interest rate risk, instead assuming a deterministic rate. The financial aspects of the rider are interesting in their own right and we spend significant time developing and analyzing a model without mortality.

The order of the thesis is briefly outlined. In Chapter 2 we motivate the remainder of the thesis by reviewing the continuous model from Milevsky and Salisbury (2006). The content is largely an integration of results from the literature and in particular we formalize the relationship between the value processes for the GMWB rider from the view of both the insured and the insurer. We present the binomial

[4] http://www.annuityfyi.com/blog/2012/01/not-everyone-is-running-from-variable-annuities/

asset pricing model for GMWBs in Chapter 3. We start with a restricted model but subsequently extend it to allow for surrenders. A dynamic delta hedging strategy is shown to perfectly hedge the GMWB rider. We summarize a binomial approximation algorithm designed to improve numerical efficiency. Numerical results are obtained and compared with results from the literature. The modeling framework is further extended in Chapter 4 to account for diversifiable mortality risk. The effectiveness of diversification is studied with a numerical example by simulating the death times for pools of insured, rapidly growing in size.

1.5 A Discussion on Imperfect Models and Subrational Behaviour

Similar to the models mentioned in Section 1.3, we work with arbitrage-free and complete financial markets and price the rider under the risk-neutral measure. To justify this approach two simplifying assumptions are needed (see Jeanblanc et al., 2009): i) equal borrowing and lending interest rates and ii) a liquid market. This latter assumption means no transaction costs (i.e. the buying price of an asset is equal to its selling price), any amount of shares may be purchased and shortselling is permitted. Policyholders are also assumed to be rational.

Such an approach suffers from serious abstractions from the real-world marketplace. Although the rider is viewed as a complex financial derivative and priced as such, it remains an add-on to the underlying base contract which has its own fees and insurance components. That is, the rider is not available for purchase by itself. Mortality markets are incomplete and the insurance market is not an openly traded liquid market. From the policyholder's perspective there are significant transaction costs in the forms of surrender charges in order to exit a contract. The rider can not be opted out of; the whole contract needs to be surrendered. There may be taxation issues for early surrenders, depending on the age of the insured. From the insurer's view, there are significant entry barriers to the market due to the strict regulatory environment in which insurers operate. This means the behaviour deemed rational by financial economic models working with liquid and frictionless markets may not be an accurate representation.

Modeling policyholder behaviour risk involves two components. Determining the optimal behaviour can be complicated, more so for GMWB products with extra features such as ratchets or rollups which may make it optimal to not withdraw in certain cases. The second component is deciding whether to model optimal behaviour at all. Advocates of assuming sub-optimal behaviour argue that policyholders do not always act in a rational optimizing manner. Charging for optimality places the insurer at a competitive disadvantage but charging too little may prove costly if optimal behaviour is realized. Even if insurers do not charge for optimality in practice it is still of interest to examine optimal behaviour to understand the worst case scenarios.

Knoller et al. (2011) conduct a statistical analysis of the Japanese VA marketplace to learn the extent to which rational lapsation occurs in the real world. The field of behavioural finance helps explain why policyholders may act irrationally. Although the paper concludes with strong support that surrenders are a dynamic reaction to the underlying market performance, it is shown that there is clear heterogeneity among policyholders and some irrationality. The *emergency fund hypothesis* and the need for liquidity help explain irrational surrenders. On the other hand, there are several reasons mentioned in the paper as to why an insured would hold onto the contract rather than optimally surrender it. These include being unable to estimate the optimal strategy, the presence of transaction costs and other heuristics and biases that influence decision making.

Moenig and Bauer (2011) is another paper in this direction which recognizes that contracts are not openly traded. Utilizing a utility-based approach for VAs with

GMWB riders, it is argued that policyholders purchase VAs for investment portfolios and external factors likely play a role in withdrawal and lapse strategies. These factors include the complete retirement portfolio and tax rates. Their results imply that the market prices are indeed fair, contrary to the consensus in Section 1.3.

Given indivisibility of VAs and riders, liquidity constraints, and lack of an open market for GMWB riders, ignoring the VA base contract and calculating the no-arbitrage hedge cost in a risk-neutral framework directly contradicts the assumptions of mathematical finance. Nevertheless, the models have their own merits and the simplifications are necessary to work with a tractable model. To justify calling the fair price the no-arbitrage price, we must assume the existence of a fully liquid secondary market. In this case, optimal behaviour must be assumed. Indeed, if this were not the case the rider would be underpriced and arbitrage situations would arise. While a growing secondary market for payout annuities has been in place for several years, a secondary market for variable annuities has been developing slowly in the past few years. There are companies, such as J.G. Wentworth, which buy back annuities and sell them to investors. However, the market is not openly run and is quite illiquid. In addition, annuities must have non-qualified tax status to be eligible for resale.

There have been legal challenges to this secondary market of late. In 2010, state insurance regulators voted to allow insurers to cancel guaranteed death benefits or living benefits if a policyholder sells the contract[5]. The American Council of Life Insurers argued that "If the institutional investor buys GMWBs en masse, it would eliminate the policy holder behaviour variable, which will cause the GMWB feature for all purchasers across the board to increase"[6].

Notwithstanding all the difficulties with the risk neutral valuation framework we work under similar assumptions to Milevsky and Salisbury (2006) and the related literature. We implicitly assume the existence of a liquid open secondary market,

[5]http://www.lifehealthpro.com/2010/03/08/feature-regulator-group-moves-to-reign-in-secondar
[6]http://www.investmentnews.com/article/20100228/REG/302289992

allowing us to operate in the risk neutral framework and obtain the arbitrage-free hedge cost or fair value of the rider.

Chapter 2

Valuation of GMWBs in a Continuous-Time Framework

In this chapter we review the continuous model constructed by Milevsky and Salisbury (2006) to price GMWBs. By incorporating elements introduced by both Peng et al. (2012) and Liu (2010), this chapter provides a firm and comprehensive synthesis of the theoretical model and motivates the developments in the following chapters. In addition to providing derivations and details that have been omitted in the literature we also contribute new results, in particular on the topic of existence and uniqueness of a fair fee and the formal set-up of the model with lapses.

2.1 Financial Model

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space where $\{B_t^{'}\}_{0 \leq t \leq T}$ is a 1-dimensional standard Brownian motion defined on this space, $B_0^{'} = 0$ a.s., and $T < \infty$. Define $\mathcal{F}_t := \sigma\{B_s^{'}; 0 \leq s \leq t\}$, for all $t \in [0, T]$. Consider the financial market consisting of one risky asset and one riskless asset. The financial market is complete. The unit price of the risky asset $\{S_t^{x,u}\}_{u \le t \le T}$ follows the geometric Brownian motion process

$$dS_t = \mu S_t dt + \sigma S_t dB_t^{'}, \quad t \ge u, \quad S_u = x. \tag{2.1}$$

We write S_t^x instead of $S_t^{x,0}$ and will often write S_t instead of $S_t^{x,u}$ when the initial conditions are easily understood from the context. The unit price of the riskless asset, which is the money market account $\{M_t\}_{0 \le t \le T}$, follows the process

$$dM_t = r(t) M_t dt, \quad t \ge 0, \quad M_0 = 1,$$

where $r(t)$ is the risk-free interest rate at time t. Assuming a constant rate r we have $M_t = e^{rt}$ and $D_t := (M_t)^{-1} = e^{-rt}$ for all $t \ge 0$.

Applying Girsanov's theorem for Brownian motion (see for instance Øksendal, 2003), we have that $\{B_t := B_t^{'} + \theta t\}_{0 \le t \le T}$ is a standard Brownian motion under the unique risk neutral measure $\mathbb{Q}$ equivalent to $\mathbb{P}$ where $\frac{d\mathbb{Q}}{d\mathbb{P}} := N_T$, $N_t = \exp(-\theta B_t^{'} - \frac{1}{2}\theta^2 t)$ for $0 \le t \le T$ and $\theta = \frac{\mu - r}{\sigma}$. Thus $\{S_t^{x,u}\}_{u \le t \le T}$ follows the process:

$$dS_t = r S_t dt + \sigma S_t dB_t, \quad t \ge u, \quad S_u = x. \tag{2.2}$$

We work with the filtered probability space $(\Omega, \mathcal{F}_T, \mathbb{F}, \mathbb{Q})$ where $\mathbb{F} = \{\mathcal{F}_s\}_{0 \le s \le T}$.

2.2 GMWB Valuation

We formulate our assumptions as follows.

Assumption 2.1. *We adopt the financial model from Section 2.1. We assume a static withdrawal strategy where the policyholder takes continuous withdrawals at a rate of $G := gP$ per year. The maturity is $T := \frac{1}{g}$ years. Early lapses are not permitted. We also assume $r > 0$ for reasons to be explained.*

The account value process $\{W_t\}$ is reduced by the instantaneous rider fees $\alpha W_t dt$ and the instantaneous withdrawals $G dt$. By (2.2) the account value $W_t^{P,0}$ follows the

SDE

$$dW_t = (r-\alpha)W_t dt + \sigma W_t dB_t - G dt, \quad 0 \le t \le T, \quad W_0 = P. \tag{2.3}$$

Observe that $\{W_t\}_{t\ge 0}$ is a time-homogeneous diffusion (Markov) process and $W_u^{x,t} \overset{d}{=} W_{u-t}^{x,0}$. However, the price processes will not be time-homogeneous. This SDE can be solved by the method presented in Øksendal (2003, p.79). Define $F_t := e^{-\sigma B_t + \frac{1}{2}\sigma^2 t}$ then we have

$$d(F_t W_t) = F_t\left((r-\alpha)W_t - G\right)dt.$$

Let $H_t := F_t W_t$, then $W_t = H_t/F_t$ and

$$\frac{d(H_t)}{dt} = F_t\left((r-\alpha)H_t/F_t - G\right) = H_t(r-\alpha) - GF_t.$$

This is an ODE in $t \mapsto H_t(\omega)$, for all fixed $\omega \in \Omega$. With the initial condition $H_0 = P$, its solution is

$$H_t(\omega) = Pe^{(r-\alpha)t} - G\int_0^t e^{(r-\alpha)(t-s)+0.5\sigma^2 s - \sigma B_s(\omega)} ds$$

from which it follows that

$$\begin{aligned} W_t &= Pe^{(r-\alpha-0.5\sigma^2)t+\sigma B_t} - G\int_0^t e^{(r-\alpha-0.5\sigma^2)(t-s)+\sigma(B_t-B_s)} ds \\ &= e^{(r-\alpha-0.5\sigma^2)t+\sigma B_t}\left[P - G\int_0^t e^{-(r-\alpha-0.5\sigma^2)s-\sigma B_s} ds\right]. \end{aligned} \tag{2.4}$$

The initial premium P can be factored out of (2.4) because $G = gP = P/T$. Let $\{Z_t\}$ denote the account value process under a no-withdrawal strategy beginning with $Z_0 = 1$. Then Z_t follows the SDE

$$dZ_t = (r-\alpha)Z_t dt + \sigma Z_t dB_t, \quad 0 \le t \le T, \quad Z_0 = 1,$$

with the solution

$$Z_t = e^{(r-\alpha-0.5\sigma^2)t+\sigma B_t}.$$

The α term can be thought of as a continuous dividend payout rate on the asset Z_t.

By (2.4) W_t can be expressed in terms of Z_t:

$$W_t = PZ_t - G\int_0^t \frac{Z_t}{Z_s} ds. \tag{2.5}$$

Milevsky and Salisbury (2006) use a slight variant of this expression involving the inverse of Z.

More generally, consider $0 \le t \le u \le T$. By (2.4), with $W_0 = P$ and writing $u = t + (u - t)$ we readily obtain

$$\begin{aligned} W_u &= W_t e^{(r-\alpha-0.5\sigma^2)(u-t)+\sigma(B_u-B_t)} - G\int_t^u e^{(r-\alpha-0.5\sigma^2)(u-s)+\sigma(B_u-B_s)} ds \\ &= W_t \frac{Z_u}{Z_t} - G\int_t^u \frac{Z_u}{Z_s} ds. \end{aligned}$$

We present an alternative form first appearing in Liu (2010). Apply a change of variables $v = t - s$ to (2.5). Then

$$W_t = PZ_t - G\int_0^t \frac{Z_t}{Z_{t-v}} dv. \tag{2.6}$$

By the time-reversibility property of Brownian motion, $\{B_t - B_{t-v}\}_{v\geq 0} \sim \{B_v\}_{v\geq 0}$ under $\mathbb{Q}$ (see Karatzas and Shreve (1991, Lemma 9.4)). Apply this property to (2.6), then

$$W_t \stackrel{d}{=} PZ_t - G\int_0^t Z_v dv. \tag{2.7}$$

From the Markov property for W we have

$$W_u^{x,t} \stackrel{d}{=} xZ_{u-t} - G\int_0^{u-t} Z_v dv. \tag{2.8}$$

In particular (2.7) simplifies for $t = T$ to

$$W_T \stackrel{d}{=} P\left(Z_T - \frac{1}{T}\int_0^T Z_s ds\right). \tag{2.9}$$

This expression is quite familiar from Asian option theory and will be elaborated on in the next section. Liu (2010) works primarily with a discrete-time analogue of

(2.9)[1].

An additional constraint must be included to account for the non-negativity of the account value. That is, $W_t \geq 0$ for all $t \geq 0$. As stated in Milevsky and Salisbury (2006), under this constraint the solution for W_t is:

$$W_t^{P,0} = \max\left[0, \left(Pe^{(r-\alpha-0.5\sigma^2)t+\sigma B_t} - G\int_0^t e^{(r-\alpha-0.5\sigma^2)(t-s)+\sigma(B_t-B_s)}ds\right)\right] \tag{2.10}$$

$$= \max\left[0, PZ_t - G\int_0^t \frac{Z_t}{Z_s}ds\right], \tag{2.11}$$

and more generally

$$W_u^{x,t} = \max\left[0, xe^{(r-\alpha-\frac{1}{2}\sigma^2)(u-t)+\sigma(B_u-B_t)} - G\int_t^u e^{(r-\alpha-0.5\sigma^2)(u-s)+\sigma(B_u-B_s)}ds\right].$$

Equation (2.8) becomes

$$W_u^{x,t} \stackrel{d}{=} \max\left(0, xZ_{u-t} - G\int_0^{u-t} Z_v dv\right). \tag{2.12}$$

Equation (2.10) can be heuristically justified. Relabeling W_t from (2.4) as $\tilde{W}_t$ then $\tilde{W}_u \leq 0$ implies

$$P < G\int_0^u e^{-(r-\alpha-0.5\sigma^2)s-\sigma B_s}ds,$$

and since the integrand is positive, for all $v \geq u$

$$P < G\int_0^v e^{-(r-\alpha-0.5\sigma^2)s-\sigma B_s}ds$$

which gives

$$\tilde{W}_v \leq 0 \text{ for all } v \geq u$$

and (2.10) follows. Once the account value hits zero, it remains at zero.

The next result will be used in Subsection 2.2.3.

Lemma 2.2. *For any fee rate α and guaranteed withdrawal rate g there is a positive*

[1] Liu (2010) justifies the continuous-time equivalence (2.9) using only the independence property of Brownian motion, which is not sufficient to prove the above. The time-reversibility property is needed. We also emphasize that the equivalence is in distribution only, which places limitations on its usage.

probability that the contract matures with a positive account value. That is,

$$\mathbb{Q}(W_T^{P,0} > 0) > 0$$

for all $P > 0$*,* $g > 0$*, and* $\alpha \geq 0$*, where* $W_T^{P,0}$ *is given by (2.10).*

Proof. To see this, observe that

$$W_T^{P,0} > 0 \text{ if and only if } \frac{P}{G} > \int_0^T e^{-(r-\alpha-0.5\sigma^2)s-\sigma B_S} ds.$$

By bounding and removing the deterministic portion from the integrand, we have

$$\frac{P}{G} > \int_0^T e^{-(r-\alpha-0.5\sigma^2)s-\sigma B_S} ds$$

if

$$\frac{P}{G} c^{-1} > \int_0^T e^{-\sigma B_s} ds,$$

where

$$c = \begin{cases} e^{-(r-\alpha-0.5\sigma^2)T} & \text{if } (r-\alpha-0.5\sigma^2) < 0, \\ 1 & \text{otherwise.} \end{cases}$$

To obtain the desired conclusion it is sufficient to show that $\mathbb{Q}(\int_0^T e^{-aB_s} ds < k) > 0$ for all $T,\ a,\ k > 0$. This result is proved in Proposition A.1. □

There are two perspectives from which to view the GMWB rider. A policyholder is likely to view the VA and rider as one combined instrument and would be interested in the total payments received over the duration of the contract. On the other hand, although the rider is embedded into the VA the insurer might want to consider it as a separate instrument. Namely, the insurer is interested in mitigating and hedging the additional risk attributed to the rider.

2.2.1 Policyholder Valuation

The random variable for the time-zero present value of the total payments received by the policyholder over the duration of the contract is

$$\int_0^T Ge^{-rs}ds + e^{-rT}W_T,$$

where W_T is given by (2.10). Referring to standard international actuarial notation (IAN) we write the present value of a continuously paid term-certain annuity as

$$\bar{a}_{\overline{T}|} = \int_0^T e^{-rs}ds = \frac{1-e^{-rT}}{r}.$$

Denote by V_0 the value at $t = 0$ for the complete contract (VA plus GMWB rider). As in Milevsky and Salisbury (2006) we have

$$V_0(P, \alpha, g) = E_{\mathbb{Q}}\left[\int_0^T Ge^{-rs}ds + e^{-rT}W_T\right] = G\bar{a}_{\overline{T}|} + e^{-rT}E_{\mathbb{Q}}[W_T]. \tag{2.13}$$

Recall that $T = 1/g$ and $G = gP$. Since P can be factored out of (2.4) it follows that

$$V_0(P, \alpha, g) = PV_0(\alpha, g), \tag{2.14}$$

where $V_0(\alpha, g) = g\bar{a}_{\overline{T}|} + e^{-rT}E_{\mathbb{Q}}[W_T^{1,0}]$. When (P, α, g) is understood, we drop it from the notation and write V_0.

The value V_0 is an implicit function of the fee rate α. The fair fee rate is defined to be the rate $\alpha^\star$ that satisfies

$$V_0(P, \alpha^\star, g) = P. \tag{2.15}$$

That is, a risk-neutral policyholder expects to receive back exactly the initial premium P. Existence and uniqueness results for $\alpha^\star$ are derived in Subsection 2.2.3. Equation (2.15) does not have a closed form solution and numerical methods must be used to find $\alpha^\star$. From (2.14) observe that $V_0(\alpha^\star, g) = 1$; that is, $\alpha^\star$ is independent of P.

Let $\{V_t\}_{0\leq t\leq T}$ be the process for the evolving value of the contract over time where V_t is the valuation of the contract considering only future cashflows occurring after time t, discounted to time t, and conditional on the information $\mathcal{F}_t$. Then

$$
\begin{aligned}
V_t &= E_{\mathbb{Q}}\left[\int_t^T e^{-r(s-t)} G ds + e^{-r(T-t)} W_T | \mathcal{F}_t\right] \\
&= G\overline{a}_{\overline{T-t|}} + e^{-r(T-t)} E_{\mathbb{Q}}[W_T^{P,0} | \mathcal{F}_t].
\end{aligned}
\tag{2.16}
$$

By the Markov property for W_t (see Øksendal (2003, Theorem 7.1.2)) we have

$$V_t = v(t, W_t),$$

$\mathbb{Q}$-a.s. for all $t \in [0, T]$, where $v : [0, T] \times \mathbb{R}_+ \mapsto \mathbb{R}_+$ is given by

$$v(t, x) = G\overline{a}_{\overline{T-t|}} + e^{-r(T-t)} E_{\mathbb{Q}}[W_T^{x,t}].$$

Alternatively, using (2.11) and (2.12) V_0 can be decomposed into the sum of a term certain annuity component and either a Quanto Asian Put option on Z^{-1} (see Milevsky and Salisbury, 2006) or an Asian Call (floating strike) option on Z (see Liu, 2010). In either formulation the value function v must be a function of both Z_t and some functional $f(\{Z_s; 0 \leq s \leq t\})$ because only the joint process $\{Z_{u,} f(\{Z_s; 0 \leq s \leq u\})$ is Markovian. Therefore we choose to continue working directly with (2.13). However the alternative forms will prove to be useful when exploring different algorithms in Chapter 3.

2.2.2 Insurer Valuation

The alternative viewpoint, applicable to the insurer, is to explicitly consider the embedded guarantee option represented by the rider as a standalone product. We begin by introducing the concept of the trigger time, first defined by Milevsky and Salisbury (2006).

Definition 2.3. The trigger time τ, defined by the stopping time

$$\tau := \inf\{s \in [0, T]; W_s^{P,0} = 0\},$$

is the first hitting time of zero by the account value process. The convention $\inf(\emptyset) = \infty$ is adopted. We have $W_t = 0$ for all $t \geq \tau$.

We define the respective non-decreasing sequences of stopping times $\{\tau_t\}_{t\in[0,T]}$ and $\{\bar{\tau}_t\}_{t\in[0,T]}$ as

$$\tau_t := \tau \vee t = \max(\tau, t)$$

and

$$\bar{\tau}_t := \tau_t \wedge T = \min(\tau_t, T),$$

for all $t \in [0, T]$. For $0 \leq s \leq t \leq T$ and $A \subset [0, T]$, by the Markov property of W_t we have

$$\mathbb{Q}(\bar{\tau}_t \in A | \mathcal{F}_s) = F(s, t, A, W_s), \tag{2.17}$$

$\mathbb{Q}$-a.s. where

$$F(s, t, A, w) := \mathbb{Q}(\bar{\tau}_t^{w,s} \in A)$$

and

$$\bar{\tau}_t^{w,s} = \inf\{u \geq t; W_u^{w,s} = 0\} \wedge T.$$

Remark 2.4. In Lemma 2.2, we showed for any $t > 0$ that $\mathbb{Q}(W_t^{P,0} > 0) > 0$ or equivalently $\mathbb{Q}(\tau \leq t) < 1$. Recall that

$$W_t^{P,0} = 0 \quad \text{if and only if} \quad \frac{P}{G} \leq \int_0^t e^{-(r-\alpha-0.5\sigma^2)s-\sigma B_S} ds.$$

An explicit distribution function for $\int_0^t e^{-(r-\alpha-0.5\sigma^2)s-\sigma B_s} ds$ can be found if $(r - \alpha) < \frac{3}{2}\sigma^2$ (see (A.1) for the formulation based on Borodin and Salminen, 2002). This can be used to calculate ruin probabilities $\mathbb{Q}(\tau \leq t) = \mathbb{Q}(W_t^{P,0} = 0)$. If $(r - \alpha) \geq \frac{3}{2}\sigma^2$ then $e^{-(r-\alpha-0.5\sigma^2)s} < 1$ for all $s > 0$. By removing this deterministic portion from the integrand, an upper bound for $\mathbb{Q}(\tau \leq t)$ can be found by evaluating $\mathbb{Q}(\int_0^t e^{-\sigma B_s} ds \geq \frac{P}{G})$ using (A.1) with $a = 0$. Ruin probabilities are typically calculated under the physical measure. Because of the equivalence of the two measures, the preceding discussion remains unchanged except that r is replaced by μ when switching measures from $\mathbb{Q}$ to $\mathbb{P}$.

If $\tau \leq T$ we say the option is triggered (or exercised) at trigger time τ. Since trigger activity is contingent on the account value hitting zero, this is similar to an American-style put option but one where the exercise date is determined endogenously rather than explicitly by the policyholder.

Let $U = \{U_t; 0 \leq t \leq T\}$ denote the stochastic process for the evolving rider value over time. At time $\bar{\tau}_0$ the rider guarantee entitles the policyholder to receive a term certain annuity for $T - \bar{\tau}_0$ years with an annual payment of G. Fee revenue is received up to time $\bar{\tau}_0$. At time $\bar{\tau}_0$ no uncertainty remains. However, we still consider the policy to be in-force and set the guarantee option value equal to the present value of the remaining guaranteed payments. It is simpler in the model formulation to treat the termination time as T rather than terminating it at time $\bar{\tau}_0$.

This motivates the following definition for U which also appears in Peng et al. (2012). For $t \in [0, T]$ we define

$$U_t := E_{\mathbb{Q}}\left[e^{-r(\bar{\tau}_t - t)} G \bar{a}_{\overline{T-\bar{\tau}_t|}} - \int_t^{\bar{\tau}_t} e^{-r(s-t)} \alpha W_s^{P,0} ds | \mathcal{F}_t\right]. \tag{2.18}$$

The value U_t is the risk-neutral expected discounted difference between future rider payouts and future fee revenues. That is, U_t represents the remaining risk exposure to the insurer in that it is positive when the expected fee revenues fall short of the rider payouts. By the Markov property for $\{W_t\}$ and (2.17) we have

$$U_t = u(t, W_t),$$

$\mathbb{Q}$-a.s. for all $t \in [0, T]$, where $u : [0, T] \times \mathbb{R}_+ \mapsto \mathbb{R}$ is given by

$$u(t, x) = E_{\mathbb{Q}}\left[e^{-r(\bar{\tau}_t^{x,t} - t)} G \bar{a}_{\overline{T-\bar{\tau}_t^{x,t}|}} - \int_t^{\bar{\tau}_t^{x,t}} e^{-r(s-t)} \alpha W_s^{x,t} ds\right]. \tag{2.19}$$

The boundary condition $u(t, 0) = G \bar{a}_{\overline{T-t|}}$ is implied in the above formulation.

2.2.3 Analytic Results

With the goal of arriving at an existence and uniqueness result for $\alpha^\star$, we first prove two basic properties satisfied by V_0.

Lemma 2.5. *V_0, defined by (2.13), is a strictly decreasing and continuous function of α for $\alpha \geq 0$.*

Proof. We fix P and g and omit them from the notation. A monotonicity result is obtained by applying a comparison result for SDEs from Karatzas and Shreve (1991, Proposition 2.18). Since α appears as a negative drift term in the SDE for W_t in (2.3), we have $W_t(\alpha_1) \geq W_t(\alpha_2)$ a.s. for all $t \in [0, T]$ and for all $0 \leq \alpha_1 < \alpha_2$. Thus $E_{\mathbb{Q}}[W_T(\alpha_1)] \geq E_{\mathbb{Q}}[W_T(\alpha_2)]$ which implies $V_0(\alpha_1) \geq V_0(\alpha_2)$.

To obtain the strictly decreasing property, note from Lemma 2.2 that $\mathbb{Q}(A^\alpha) > 0$ for all $\alpha \geq 0$ where $A^\alpha := \{W_T(\alpha) > 0\}$. On the event A^α we have

$$W_T(\alpha) = e^{(r-\alpha-0.5\sigma^2)T+\sigma B_T} \times \left(P - G \int_0^T e^{-(r-\alpha-0.5\sigma^2)s-\sigma B_s} ds \right).$$

Let $0 \leq \alpha_1 < \alpha_2 = \alpha_1 + h$, where h takes an arbitrary positive value. Restricted to the set A^{α_1+h}, we obtain

$$W_T(\alpha_1 + h) \leq e^{-hT} W_T(\alpha_1) < W_T(\alpha_1)$$

implying that $A^{\alpha_1} \supseteq A^{\alpha_1+h}$. It follows that

$$\begin{aligned} V_0(\alpha_1 + h) &= G\bar{a}_{\overline{T}|} + E_Q \left(e^{-rT} W_T(\alpha_1 + h) \mathbf{1}_{\{A^{\alpha_1+h}\}} \right) \\ &< G\bar{a}_{\overline{T}|} + E_Q \left(e^{-rT} W_T(\alpha_1) \mathbf{1}_{\{A^{\alpha_1+h}\}} \right) \\ &\leq V_0(\alpha_1). \end{aligned}$$

To prove continuity fix $\alpha \geq 0$. Let $h > 0$ and denote

$$X_T^h := e^{\sigma B_T} \max \left(0, P - G \int_0^T e^{-(r-\alpha-h-\frac{1}{2}\sigma^2)s-\sigma B_s} ds \right).$$

From (2.10),

$$E_Q(W_T(\alpha + h)) = e^{(r-\alpha-h-\frac{1}{2}\sigma^2)T} E_Q\left(X_T^h\right).$$

Then $X_T^h \geq 0$ for all $h \geq 0$, and $X_T^h \uparrow$ a.s. as $h \downarrow 0$. Applying the Monotone Convergence theorem and by the continuity of the max function,

$$\lim_{h \downarrow 0} E_{\mathbb{Q}}(X_T^h) = E_{\mathbb{Q}}(X_T^{h=0}).$$

The Dominated Convergence theorem was used to interchange the limit and the pathwise Lebesgue-Stieltjes integral. Therefore $\lim_{h\downarrow 0} E_{\mathbb{Q}}(W_T(\alpha+h)) = E_{\mathbb{Q}}(W_T(\alpha))$.

If $\alpha > 0$, then let $h < 0$ and $\lim_{h\uparrow 0} E_{\mathbb{Q}}(W_T(\alpha + h)) = E_{\mathbb{Q}}(W_T(\alpha))$ is obtained using similar arguments. The Monotone Convergence theorem no longer applies; instead the Dominated Convergence theorem justifies interchanging the expectation and limit since $X_T^h \leq Pe^{\sigma B_T}$ and $E_{\mathbb{Q}}(e^{\sigma B_T}) = e^{0.5\sigma^2 T} < \infty$. Therefore the continuity of V_0 follows from (2.13). □

Proposition 2.6. *Under Assumption 2.1 there exists a unique $\alpha^\star$ satisfying*

$$V_0(P, \alpha^\star, g) = P.$$

Proof. The existence of $\alpha^\star$ is obtained by showing that both $V_0(P, 0, g) \geq P$ and $\lim_{\alpha\to\infty} V_0(P, \alpha, g) < P$ and applying the continuity result from Lemma 2.5.

When $\alpha = 0$, the guarantee is offered at no charge and it is obvious that $V_0 \geq P$. More formally, setting $\alpha = 0$ we have from (2.10)

$$W_T \geq \left[Pe^{(r-0.5\sigma^2)T+\sigma B_T} - G \int_0^T e^{(r-0.5\sigma^2)(T-s)+\sigma(B_T - B_s)} ds \right],$$

and since $E_{\mathbb{Q}}[e^{-0.5\sigma^2 t+\sigma B_t}] = 1$, we obtain from (2.13) that

$$\begin{aligned} V_0(P, 0, g) &\geq P + E_{\mathbb{Q}} \left[\int_0^T e^{-rs} G \left(1 - e^{-(0.5\sigma^2)(T-s)+\sigma(B_T - B_s)} \right) ds \right] \\ &= P, \end{aligned}$$

where the expectation on the right evaluates to zero by Fubini's theorem.

As $\alpha \to \infty$, it becomes certain that the embedded GMWB option will be exercised and thus $V_0 = G\bar{a}_{\overline{T}|}$. More formally, for $\alpha > 0$ we have

$$0 \leq W_T(\alpha) \leq Pe^{-\alpha T} e^{(r-0.5\sigma^2)T+\sigma B_T} \leq Pe^{(r-0.5\sigma^2)T+\sigma B_T} \tag{2.20}$$

a.s., and $E_{\mathbb{Q}}[Pe^{(r-0.5\sigma^2)T+\sigma B_T}] = Pe^{rT} < \infty$. The property $B_T < \infty$ a.s. combined with (2.20) gives $\lim_{\alpha\to\infty} W_T(\alpha) = 0$ a.s. Applying the Dominating Convergence theorem,

$$\lim_{\alpha\to\infty} V_0(P, \alpha, g) = G \int_0^T e^{-rs} ds < GT = P,$$

for $r > 0$.

The uniqueness of the solution follows directly from the strictly decreasing property for $V_0(P, \alpha, g)$ from Lemma 2.5. □

Remark 2.7. Assumption 2.1 imposed that $r > 0$. In the case $r = 0$, the optimal solution $\alpha^\star$ must satisfy $W_T(\alpha^\star) = 0$ a.s. By Lemma 2.2, no solution exists.

The next result unifies the insured and insurer perspectives and was first presented in Peng et al. (2012) for the case $t = 0$ under a more general structure with stochastic interest rates. We omit the proof here. In Subsection 2.2.4 we extend this result to the more general case of surrenders and a complete proof will be presented at that time.

Proposition 2.8. *For any $\alpha \geq 0$, the following relation holds for all $t \in [0, T]$ and for all $w > 0$: $v(t, w) = u(t, w) + w$. That is, $V_t = U_t + W_t$ a.s.*

Remark 2.9. By definition of the fair fee rate $\alpha^\star$ we have $U_0(P, \alpha^\star, g) = 0$ as a result of Proposition 2.8. From Lemma 2.5 we have $V_0 < P$ and $U_0 < 0$ for all $\alpha > \alpha^\star$. Likewise, $V_0 > P$ and $U_0 > 0$ for all $\alpha < \alpha^\star$. For any t, we say the contract is in the money (ITM) if $V_t > W_t$ and $U_t > 0$. Similarly, it is out of the money (OTM) if $V_t < W_t$ and $U_t < 0$. It is at the money (ATM) if $V_t = W_t$ and $U_t = 0$.

Remark 2.10. In Section 1.1 we briefly discussed the fund drag created by an increase in the rider fee rate. The strictly decreasing property of V_0 and Proposition 2.8 imply that $U_0 = V_0 - P$ is a strictly decreasing function of α. Thus any increase in expected revenue from an increase in α will always exceed any increase in expected rider payouts.

2.2.4 Extending the Model: Surrenders

We allow the policyholder to surrender the policy prior to time T. In Section 1.5 policyholder behaviour was discussed in some detail in regard to an insurer's risk exposure. Although a policyholder may surrender for a number of reasons, for instance due to an emergency cash crisis, rational behaviour in an economic sense is assumed here. Early surrenders occur only if the proceeds from immediately lapsing the product exceeds the risk-neutral value of keeping the contract in-force.

Upon surrender the policyholder closes out the contract by withdrawing the current account value. The cash proceeds are reduced by a surrender charge on any amount exceeding the annual maximal permitted withdrawal amount specified in the rider contract. Typically, VA contract provisions include contingent deferred sales charge (CDSC) schedules specifying surrender charges as a function of the duration since issue year. An example is an 8-year schedule with a charge of 8% in year 1 and decreasing by 1% each year, followed by no surrender charges after year 9.

We assume the proceeds from surrender charges are invested in the hedging portfolio. Without surrender charges, it would be optimal to surrender the contract when it is OTM. In this case the guarantee has relatively low value in terms of future payouts and the policyholder has an incentive to lapse and avoid paying future annual rider fees. The surrender charges act as a transaction cost and may make it too costly to surrender or even if it is still optimal to surrender, the surrender charge provides the insurer with income to compensate for the loss of future fees.

A surrender option in the context of guaranteed minimum death benefit riders is discussed in Milevsky and Salisbury (2001). It is argued that "when option premiums are paid by installments - even in the presence of complete mortality and financial markets - the ability to 'lapse' de facto creates an incomplete market". The surrender charges complete the market and make the guarantees hedgeable.

To describe the CDSC schedule let $k : [0,T] \mapsto [0,1]$ be a deterministic non-increasing piecewise constant RCLL (right continuous with left limits) function with possible discontinuities at integer time values[2]. For a policy issued at time zero, k_s is the surrender charge applicable at time s. The no-lapse model is easily recovered by setting $k_s = 1$ for all $s \in [0,T)$ and $k_T = 0$ in which case the opportunity to surrender early is worthless. Similarly, we could model a contract which only allows surrenders once a specific duration t_1 is reached, by setting $k_s = 1$ for $s \in [0,t_1)$ and $k_s < 1$ for $s \in [t_1,T]$. However the more common case has $k_s < 1$ for all $s \in [0,T]$. Further, we assume $k_T = 0$ to allow comparison to the no-lapse model where the contract terminates at time T with no surrender charges.

The pricing task becomes an optimal stopping problem. The contract value process for the VA plus GMWB is

$$V_t := \sup_{\eta \in \mathbb{L}_t} V_t^\eta, \tag{2.21}$$

where

$$V_t^\eta = E_{\mathbb{Q}}\left[G\bar{a}_{\overline{\eta-t|}} + e^{-r(\eta-t)}W_\eta(1-k_\eta)|\mathcal{F}_t\right] \tag{2.22}$$

and $\mathbb{L}_t$ is the set of $\mathbb{F}-$adapted stopping times taking values in $[t,T]$. By considering the stopping time $\eta \equiv T$ it is trivial that $V_t \geq V_t^T = V_t^{NL}$, where V_t^{NL} denotes the value process from (2.16). For any $\eta \in \mathbb{L}_t$ define the set $F^\eta := \{\eta \in [\bar{\tau}_t, T)\}$ and consider the modified stopping time η_a, where $\eta_a = \eta$ on $(F^\eta)^c$ and $\eta_a = T$ on F^η. Then $V_t^\eta \leq V_t^{\eta_a}$ and it is sufficient to consider the set $\mathbb{L}_{t,\bar{\tau}_t} \subset \mathbb{L}_t$, where $\mathbb{L}_{t,\bar{\tau}_t}$ contains

[2] The developments hold true for any non-increasing function taking values in $[0,1]$ but we select a function that is an accurate representation of CDSC schedules in products sold in the insurance marketplace.

all $\mathbb{F}-$adapted stopping times taking values in $[t, \bar{\tau}_t) \cup \{T\}$, and $\bar{\tau}_t$ is the trigger time assuming no lapses (Definition 2.3). That is, if the rider is triggered without prior surrender then the future guaranteed payments can not be immediately withdrawn and optimal surrender will naturally occur at maturity time T.

By the Markov property of W_t we have

$$V_t = v(t, W_t)$$

$\mathbb{Q}-$a.s. for all $t \in [0, T]$, where $v : [0, T] \times \mathbb{R}_+ \mapsto \mathbb{R}_+$ is given by

$$v(t, x) = \sup_{\eta \in \mathbb{L}_{t, \bar{\tau}_t^{x,t}}} E_{\mathbb{Q}} \left[G\bar{a}_{\overline{\eta - t}|} + e^{-r(\eta - t)} W_\eta^{x,t} (1 - k_\eta) \right].$$

The fair fee rate remains defined as the rate $\alpha^\star$ such that

$$V_0(P, \alpha^\star, g) = P.$$

Suppose that $k_0 = 0$ and let $\hat{\alpha} := \inf\{\alpha; V_0(P, \alpha, g) = P\}$. Then for all $\alpha \geq \hat{\alpha}$ we have $V_0(P, \alpha, g) = P$, but there will be no buyers as it is optimal to surrender immediately. Insurers will not charge $\alpha < \hat{\alpha}$ because $V_0(P, \alpha, g) > P$. When lapses are permitted but no surrender charges are imposed, there is no unique $\alpha^\star$ and the product is not marketable. To preclude this trivial case, we impose the condition that $k_0 > 0$.

Consider the rider value process given by (2.18). We include lapses by only accounting for any guarantee payouts and rider fee revenues occurring prior to a lapse event. The revenue from the surrender charge is also included. Then

$$U_t := \sup_{\eta \in \mathbb{L}_{t, \bar{\tau}_t}} U_t^\eta, \tag{2.23}$$

where

$$U_t^\eta = E_{\mathbb{Q}} \left[G e^{-r(\bar{\tau}_t - t)} \mathbf{1}_{\{\eta = T\}} \bar{a}_{\overline{T - \bar{\tau}_t}|} - \int_t^\eta \alpha e^{-r(s-t)} W_s ds - e^{-r(\eta - t)} W_\eta k_\eta | \mathcal{F}_t \right].$$

By working with the reduced set $\mathbb{L}_{t, \bar{\tau}_t}$ we only need to condition on $\{\eta = T\}$.

We introduce a value process for the option to surrender and denote it by $L =$

$\{L_t; 0 \leq t \leq T\}$. Let U_t^{NL} be the rider value given by (2.18) in the no-lapse model. Then we define $L_t := U_t - U_t^{NL} \geq 0$ for all $t \in [0, T]$. Since

$$\int_\eta^T \alpha W_s e^{-r(s-t)} ds = -\int_t^\eta \alpha W_s e^{-r(s-t)} ds + \int_t^T \alpha W_s e^{-r(s-t)} ds$$

and

$$-Ge^{-r(\bar{\tau}_t - t)} \mathbf{1}_{\{\eta < \bar{\tau}_t\}} \bar{a}_{\overline{T-\bar{\tau}_t}|} = Ge^{-r(\bar{\tau}_t - t)} \mathbf{1}_{\{\eta = T\}} \bar{a}_{\overline{T-\bar{\tau}_t}|} - Ge^{-r(\bar{\tau}_t - t)} \bar{a}_{\overline{T-\bar{\tau}_t}|},$$

for $\eta \in \mathbb{L}_{t,\bar{\tau}_t}$, it follows that

$$L_t = \sup_{\eta \in \mathbb{L}_{t,\bar{\tau}_t}} L_t^\eta, \tag{2.24}$$

where

$$L_t^\eta = E_{\mathbb{Q}} \left[\int_\eta^T \alpha e^{-r(s-t)} W_s ds - Ge^{-r(\bar{\tau}_t - t)} \mathbf{1}_{\{\eta < \bar{\tau}_t\}} \bar{a}_{\overline{T-\bar{\tau}_t}|} - k_\eta W_\eta e^{-r(\eta - t)} | \mathcal{F}_t \right].$$

This formulation is quite intuitive. For a fixed surrender strategy, the surrender benefit is seen to be the expected value of the fees avoided by early surrender, minus any future benefit payments missed if surrender occurs prior to a trigger time, and minus the surrender charge paid at the time of surrender. It is natural that the insured seeks to optimize this surrender benefit. The Markovian representations for U and L are obvious and are omitted.

Proposition 2.8 formalized the precise relationship between $\{U_t\}$ and $\{V_t\}$ in the no-lapse model. The next proposition generalizes that relationship to the current model and is an extension of a result proved by Peng et al. (2012) for no lapses. The complete contract V consists of three parts (i) the account value itself, (ii) the benefit net of fees derived from the equity floor guarantee without the option of surrendering and (iii) the additional benefit derived from the added option of surrendering.

Proposition 2.11. *Let V_t, U_t^{NL}, L_t, U_t be defined by (2.21), (2.18), (2.24) and (2.23) respectively. Then for all $\alpha \geq 0$ and for all $t \in [0, T]$, we have*

$$V_t = W_t + U_t^{NL} + L_t \quad a.s., \tag{2.25}$$

and equivalently

$$V_t = W_t + U_t \quad a.s. \tag{2.26}$$

Proof. Fix $t \in [0, T]$. Applying the product rule to the term $(e^{-r(s-t)} W_s)$ for any $s \in [t, T]$,

$$\begin{aligned} d(e^{-r(s-t)} W_s) &= -re^{-r(s-t)} W_s ds + e^{-r(s-t)} dW_s \\ &= -re^{-r(s-t)} W_s ds + e^{-r(s-t)}[(r-\alpha) W_s ds + \sigma W_s dB_s - G ds] \\ &= -\alpha e^{-r(s-t)} W_s ds + e^{-r(s-t)} \sigma W_s dB_s - e^{-r(s-t)} G ds. \end{aligned}$$

Fix $\eta \in \mathbb{L}_{t,\bar{\tau}_t}$. Integrating over the interval $[t, \eta \wedge \bar{\tau}_t]$, and observing that $W_{s \wedge \bar{\tau}_t} = W_s$ for all $s \in [t, T]$, we obtain

$$e^{-r(\eta-t)} W_\eta - W_t = -\int_t^\eta \alpha W_s e^{-r(s-t)} ds - G \bar{a}_{\overline{\eta \wedge \bar{\tau}_t - t}|} + \int_t^\eta e^{-r(s-t)} \sigma W_s dB_s.$$

Note that $G\bar{a}_{\overline{\eta - t}|} = G\bar{a}_{\overline{\eta \wedge \bar{\tau}_t - t}|} + Ge^{-r(\bar{\tau}_t - t)} \bar{a}_{\overline{\eta \vee \bar{\tau}_t - \bar{\tau}_t}|}$. Having fixed $\eta \in \mathbb{L}_{t,\bar{\tau}_t}$ we have $\bar{a}_{\overline{\eta \vee \bar{\tau}_t - \bar{\tau}_t}|} = \mathbf{1}_{\{\eta = T\}} \bar{a}_{\overline{T - \bar{\tau}_t}|}$. Then

$$\begin{aligned} &e^{-r(\eta-t)} W_\eta + G\bar{a}_{\overline{\eta - t}|} = \\ &\qquad W_t + Ge^{-r(\bar{\tau}_t - t)} \mathbf{1}_{\{\eta = T\}} \bar{a}_{\overline{T - \bar{\tau}_t}|} - \int_t^\eta \alpha W_s e^{-r(s-t)} ds + \int_t^\eta e^{-r(s-t)} \sigma W_s dB_s. \end{aligned}$$

We have that

$$E_{\mathbb{Q}}\left[\int_u^v (W_s)^2 ds\right] < E_{\mathbb{Q}}\left[\int_u^v P e^{2(r-\alpha-0.5\sigma^2)s + 2\sigma B_s} ds\right] < \infty,$$

thus by a standard result the above Itô integral term is a martingale (see Øksendal (2003, Corollary 3.2.6)) and $E_{\mathbb{Q}}[\int_t^\eta e^{-r(s-t)} \sigma W_s dB_s | \mathcal{F}_t] = 0$. Subtracting $e^{-r(\eta-t)} W_\eta k_\eta$ from both sides and taking conditional expectations w.r.t. $\mathcal{F}_t$, we obtain

$$V_t^\eta = W_t + U_t^\eta.$$

This holds for any η and remains true when taking the supremum. Therefore

$$V_t = W_t + U_t.$$

□

Remark 2.12. For $\alpha^\star$, such that $V_0 = P$, we have that $U_0(\alpha^\star) = 0$ and $L_0(\alpha^\star) = -U_0^{NL}(\alpha^\star)$. For any $\alpha \geq 0$, Proposition 2.8 and Proposition 2.11 imply

$$\begin{aligned} L_t &= V_t - V_t^{NL} \\ &= \sup_{\eta \in \mathbb{L}_{t,\bar{\tau}_t}} E_{\mathbb{Q}} \left[e^{-r(\eta-t)} W_\eta (1-k_\eta) - e^{-r(T-t)} W_T - G e^{-r\eta} \bar{a}_{\overline{T-\eta|}} | \mathcal{F}_t \right]. \end{aligned}$$

This expression is interpreted as the insured selecting the surrender time to maximize the tradeoff between receiving the account value (less surrender charges) today, rather than at maturity, and foregoing the rights to any future withdrawals.

Rather than presenting a PDE approach, we have defined the price processes in this chapter in terms of risk-neutral expectations. This was done partially to motivate the developments in the following chapters of the thesis. In both the no-lapse and the lapse model the PDEs for the processes can be explicitly written. In the latter case, we obtain the linear complementarity formulation.

Beginning with the HJB equations for the more general stochastic control problem, Dai et al. (2008) reduce it to a linear complementarity formulation. Milevsky and Salisbury (2006), Dai et al. (2008), and Chen and Forsyth (2008) work with an additional dimension representing the guarantee balance because the control is the withdrawal process. By considering a constant withdrawal rate and eliminating the guarantee balance process, the linear complementarity formulation from Dai et al. (2008) reduces to the PDE expression obtained in the optimal stopping problem set up in this section.

Chapter 3

Valuation of GMWBs in a Binomial Asset Pricing Model

The discrete-time binomial asset pricing model was introduced in the seminal paper by Cox, Ross and Rubinstein (1979) and has had a major impact on the financial literature. The model can be treated as either the true underlying model in a binomial world or as an approximating model of a true underlying continuous model.

Binomial models have a number of appealing properties. They are intuitive to understand and utilize elementary mathematics. Indeed, binomial models have become the standard pedagogical tool used to introduce students to dynamic pricing theory. The binomial model converges to the Black and Scholes (1973) model and yields good approximations for more complex financial options with no analytic solutions in the continuous time pricing models. Due to the discrete time and finite state space nature, lattice-based binomial methods can be quite valuable to observing and deriving results which can then be studied in a more complex framework. Through dynamic programming and backward induction algorithms, binomial pricing models can easily be implemented in any standard programming environment (e.g. C++ or Python although our tool of choice is Matlab).

In contrast to standard Monte Carlo simulation methods, the binomial approach works for American-style options with early exercise capability. More importantly an explicit exact hedging strategy can be prescribed. A thorough comparison of binomial and finite difference methods is provided in Geske and Shastri (1985). Although binomial methods can be seen to be a special case of finite difference methods there are fundamental differences between the two general methods. Finite difference methods approximate the PDE whereas binomial methods approximate the underlying stochastic process directly.

Binomial models are ideally suited for non path-dependent products. In such a setting, aside from enabling a simple theoretical framework, it is computationally efficient to obtain reliable numerical results. As we discuss in this chapter, the GMWB product is path-dependent. From a theoretical viewpoint, a formally defined binomial asset pricing model for the GMWB is of significant value, both as a tool for better understanding the product and exploring new results, and as a pedagogical tool.

There are several textbooks treating binomial pricing theory at length. Our primary reference is Shreve (2004a) and a secondary reference is Duffie (2001). In the following sections, we generalize the approach presented in Shreve (2004a).

3.1 A General Framework

We assume the existence of a financial market consisting of one risky asset S and one riskless asset, the money market B. Let n be the number of timesteps per year then $N = T \times n$ is the total number of timesteps modeled and $\delta t = 1/n$ is the length of each timestep. For $i \in \mathcal{I}_N^+ := \{1, \ldots, N-1, N\}$, write S_i and B_i for the respective asset values at time $i\delta t$. Assuming a constant continuously compounded interest rate r we have $B_i = B_{i-1}e^{r\delta t}$ with $B_0 = 1$. Given S_{i-1}, the asset value S_i takes one of two values: $S_{i-1}u$ or $S_{i-1}d$. The value u represents an up-movement in the asset value and d represents a down-movement in the asset value. For all i this random asset growth factor should be independent of S_{i-1}. To rule out arbitrage opportunities and the trivial case of no randomness, u and d must satisfy (see Shreve, 2004a)

$$0 < d < e^{r\delta t} < u. \tag{3.1}$$

Consider a sequence of N coin tosses. Let $\Omega = \Omega_N := \{H, T\}^N$ and $\mathcal{F} := 2^\Omega$. That is, Ω is the N-ary Cartesian product of the set $\{H, T\}$ and contains all possible sequences of the N coin tosses. Denote a sample point of Ω by $\bar{\omega}_N := \omega_1 \dots \omega_N := (\omega_1, \dots, \omega_N)$. Consider the stochastic process $\xi = (\xi_i)_{1 \le i \le N}$, where $\xi_i : \Omega \mapsto \{u, d\}$ is

$$\xi_i(\bar{\omega}_N) = \begin{cases} u & \text{if } \omega_i = H, \\ d & \text{if } \omega_i = T. \end{cases}$$

Then for any fixed $\bar{\omega}_N$, $\xi_i(\bar{\omega}_N)$ maps i to the growth factor of S in period i. The natural filtration is $\mathcal{F}_i = \sigma(\xi_j;\ j \le i)$. We work with the probability measure $\mathbb{P}$ on the finite discrete probability space where for any set $A \in \mathcal{F}_N$

$$\mathbb{P}(A) := \sum_{\bar{\omega}_N \in A} \tilde{p}^{\{\# \text{ of } H \text{ in } \bar{\omega}_N\}} (1 - \tilde{p})^{\{\# \text{ of } T \text{ in } \bar{\omega}_N\}}$$

and $\tilde{p} > 0$ is the physical or real-world probability of observing a H at any particular coin toss or correspondingly observing a u at any particular time step. This completes the construction of the probability space $(\Omega, \mathcal{F}_N, \mathbb{F} = \{\mathcal{F}_i\}_{0 \le i \le N}, \mathbb{P})$.

The process $\{S_i\}$ follows $S_i = S_0 \times \prod_{j=1}^{i} \xi_j$ where S_0 is the initial value of the risky asset. Then $S_i \in \mathcal{F}_i$ and is dependent on only the first n components of any random path $\bar{\omega}_N \in \Omega$. We write $\bar{\omega}_i = \omega_1 \dots \omega_i$ to refer to the specific path evolution up to time i. For any $j \le i$, we write

$$\xi_j(\bar{\omega}_i) = \begin{cases} u & \text{if } \omega_j = H, \\ d & \text{if } \omega_j = T. \end{cases}$$

Notation 3.1. Replace H and T with u and d respectively when defining Ω, therefore the sample path $\bar{\omega}_N$ refers directly to the evolution of the underlying asset S where each $\omega_j \in \{u, d\}$. Then for any $\bar{\omega}_i$,

$$S_i = S_0 \prod_{j=1}^{i} \xi_j(\bar{\omega}_i) = S_0 \prod_{j=1}^{i} \omega_j = S_0 u^{\{\# \text{ of } u \text{ in } \bar{\omega}_i\}} d^{\{\# \text{ of } d \text{ in } \bar{\omega}_i\}}. \tag{3.2}$$

The financial market is complete with a unique risk-neutral measure $\mathbb{Q}$ defined by

$$\mathbb{Q}(A) := \sum_{\bar{\omega}_N \in A} p^{\{\# \text{ of } u \text{ in } \bar{\omega}_N\}} q^{N - \{\# \text{ of } u \text{ in } \bar{\omega}_N\}}$$

for any set $A \in \mathcal{F}_N$, where

$$p := \frac{e^{r\delta t} - d}{u - d} \tag{3.3}$$

and $q := 1 - p$ (see Cox et al. (1979) for derivation of p). Note that $p \in (0, 1)$ by (3.1) and there are no $(\mathbb{Q}, \mathcal{F}_N)$-negligible sets and so all results hold surely.

If σ is the variance of the continuously compounded rate of return of S, then following the Cox, Ross, and Rubinstein (CRR) parametrization for u and d we set

$$u = e^{\sigma\sqrt{\delta t}},$$

$$d = e^{-\sigma\sqrt{\delta t}}.$$

We present two results justifying the validity of this parametrization.

Proposition 3.2. *(Cox et al., 1979) Consider a single risky asset S. The continuously compounded rate of return of S over the time period $[0, T]$ is denoted $r^s_T = ln\left(\frac{S_T}{S_0}\right)$. Suppose we have the empirical mean and variance of r^s_T, denoted by $\hat{\mu}T$ and $\hat{\sigma}^2 T$ respectively. Consider the binomial model with n timesteps per year, $\delta t = 1/n$, and maturity T. If the binomial model parameters u, d, and $\tilde{p}$ are set equal to:*

$$u = e^{\hat{\sigma}\sqrt{\delta t}},$$

$$d = e^{-\hat{\sigma}\sqrt{\delta t}},$$

$$\tilde{p} = \frac{1}{2} + \frac{1}{2}\frac{\hat{\mu}}{\hat{\sigma}}\sqrt{\delta t},$$

then as $n \to \infty$, the mean and variance of r_T^s under the binomial model converges to $\hat{\mu}T$ and $\hat{\sigma}^2 T$ respectively.

Proposition 3.3. *Assume the existence of $(\Omega, \mathcal{F}_T, \mathbb{F}, \mathbb{Q})$ on which S_t follows the geometric Brownian motion process, $dS_t = rS_t dt + \sigma S_t dW_t$, where W_t is a $\mathbb{Q}$-Brownian motion. Consider the binomial model for S_i^n with n timesteps per year, $\delta t = 1/n$, and maturity T, on the space $(\Omega^b, \mathcal{F}_t^b, \mathbb{F}^b, \mathbb{Q}^b)$ constructed in this section. If the binomial model parameters u, d, and p are set equal to:*

$$u = e^{\sigma\sqrt{\delta t}},$$

$$d = e^{-\sigma\sqrt{\delta t}},$$

$$p = \frac{e^{r\delta t} - d}{u - d},$$

then for all $t \in [0, T]$, as $n \to \infty$, S_{nt}^n converges in distribution to S_t, where nt is an integer and S_{nt}^n is the random asset value at time t.

Proof. See Cox et al. (1979) or Shreve (2004b, Exercise 3.8). □

3.2 Valuation without Surrenders

3.2.1 The Account Value

We specify the underlying assumptions for this section.

Assumption 3.4. *We assume the existence of the space $(\Omega, \mathcal{F}_N, \mathbb{F} = \{\mathcal{F}_i\}_{0 \le i \le N}, \mathbb{Q})$ constructed in Section 3.1. Early surrenders are not allowed. Under the static withdrawal strategy the policyholder receives $G = gP\delta t$ each time period. We set $T := 1/g$.*[1] *At the end of each period the pro-rated rider fee is first deducted and then the*

[1] The theoretical developments are presented assuming T to be an integer. For values of g such that T is not an integer, the algorithms can be adapted to incorporate the final fractional period. Set $N = \lfloor T \cdot n \rfloor + 1$ and the final period has time length of $T - \left(\frac{N-1}{n}\right)$years. All the parameters, including the periodic payment G, need to be scaled for the terminal period to reflect the shortened duration. This is the approach we use to obtain $\alpha^\star$ when T is not an integer.

periodic withdrawal is subtracted. We restrict $r > 0$.

Notation 3.5. For conciseness we omit the δt-dependence from the notation for $G,\ r,$ and α. Denote $\bar{r} := r\delta t$ and $\bar{\alpha} := \alpha\delta t$.

Beginning with $S_0 = P$, the binomial tree for $\{S_i\}$ is constructed forward in time. For $i \in \mathcal{I}_N^+$, set

$$S_i = \xi_i S_{i-1}.$$

We define another binomial tree which contains two values at each node, W_{i^-} and W_i. The first is the account value after adjusting for market movements but before fees are deducted or withdrawals are made and the latter is the account value after adjusting for fees and withdrawals. We have

$$\begin{aligned} W_0 &= P, \\ W_{i^-} &= \frac{S_i}{S_{i-1}} W_{i-1} = \xi_i W_{i-1}, \\ W_i &= \max\left\{W_{i^-} e^{-\bar{\alpha}} - G, 0\right\}, \end{aligned}$$

for $i \in \mathcal{I}_N^+$. Although the tree for the underlying asset $\{S_i\}$ is recombining, the tree for the account value $\{W_i\}$ is non-recombining. For any i there are $i+1$ nodes for S_i but 2^i nodes for W_i on the respective trees. The subtraction of the periodic withdrawals imposes a path dependency on the model.

Example 3.6. Figure 3.1 provides an example of a binomial tree for $\{W_i\}$[2] with the parameters: $r = 5\%$, $\sigma = 20\%$, $g = 25\%$ and $\delta t = 1$. Therefore $p = 0.5775$ and $\alpha^\star = 3.07\%$. The withdrawal rate was selected to be unrealistically high to limit the contract to 4 years, thus the tree has only 16 nodes in the final period.

[2]constructed with the software *Tree Diagram Generator*, version 1.0

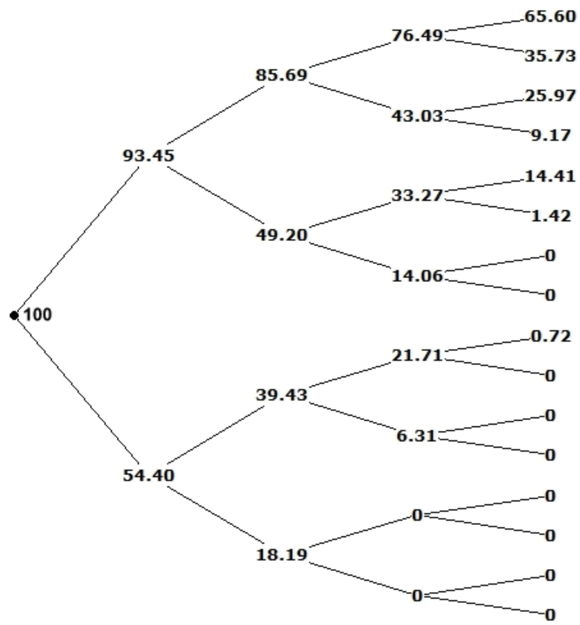


Figure 3.1: Sample binomial tree for account value process

3.2.2 Policyholder Perspective

The discrete-time counterpart to (2.16) is

$$\begin{aligned} V_N &= W_N, \\ V_i &= E_Q\left[\sum_{m=i+1}^{N} G e^{-\bar{r}(m-i)} + e^{-\bar{r}(N-i)} W_N | \mathcal{F}_i\right] \\ &= G a_{\overline{N-i|}} + e^{-\bar{r}(N-i)} E_Q[W_N|\mathcal{F}_i] \end{aligned} \tag{3.4}$$

for $i \in \mathcal{I}_{N-1}$, with $\mathcal{I}_{N-1} := [0, 1, \ldots, N-1]$ and $a_{\overline{m|}} = \frac{1-e^{-\bar{r}\,m}}{e^{\bar{r}}-1}$. For $i = 0$ this reduces to

$$V_0 = G a_{\overline{N|}} + e^{-\bar{r}N} E_Q[W_N]. \tag{3.5}$$

The process $\{V_i\}$ represents the value of the combined annuity plus GMWB rider contract at each timepoint just after the deduction of fees and withdrawals. By the Markov property we have

$$V_i = v(i, W_i),$$

where $v : \mathcal{I}_N \times \mathbb{R}_+ \mapsto \mathbb{R}_+$ is

$$v(i,x) = \begin{cases} x & i = N, \\ [G + pv(i+1, w(xu)) + qv(i+1, w(xd))]e^{-\bar{r}} & i < N, \end{cases} \tag{3.6}$$

and $w : \mathbb{R}_+ \mapsto \mathbb{R}_+$ is given by

$$w(x) = \max\{xe^{-\bar{\alpha}} - G, 0\}. \tag{3.7}$$

We remark that $\{e^{-\bar{r}i}V_i + Ga_{\overline{i}|}\}_{0 \le i \le N}$ is a $(\mathbb{Q}, \mathbb{F})$ martingale for all α.

We restate the definition of $\alpha^\star$ from Chapter 2. The fair fee rate $\alpha^\star$ satisfies

$$V_0(P, \alpha^\star, g) = P. \tag{3.8}$$

There is no closed form solution for $\alpha^\star$. In our numerical results, we iteratively solved for the fair fee using the bisection method.

Due to the path-dependent nature of the account value process, one practical drawback of the backward induction approach for V is the necessity of storing large arrays of data. To obtain V_0 an array of size 2^N is needed to record V_N for all nodes in the final period. In contrast, for recombining trees the array size needed is only $N+1$. When early surrenders are not permitted, we can directly calculate $v(i,x)$ without using trees and avoid any strain on memory capacity.

We first express W_N analogously to (2.12). Since only the terminal values of W_N are required, the *max* condition seen in the tree for W_i can be disregarded prior to period N. It follows that

$$\begin{aligned} W_N &= \max\left[\xi_N e^{-\bar{\alpha}}\left(\xi_{N-1}e^{-\bar{\alpha}}\left(...\left(\xi_2 e^{-\bar{\alpha}}\left(P\xi_1 e^{-\bar{\alpha}} - G\right) - G\right)...\right) - G\right) - G, 0\right] \\ &= \max\left[0, Pe^{-\bar{\alpha}N}\prod_{i=1}^{N}\xi_i - G\sum_{i=0}^{N-1} e^{-\bar{\alpha}i}\prod_{j=N-(i-1)}^{N}\xi_j\right] \\ &= \max\left[0, W_M e^{-\bar{\alpha}(N-M)}\prod_{i=M+1}^{N}\xi_i - G\sum_{i=0}^{N-M-1} e^{-\bar{\alpha}i}\prod_{j=N-(i-1)}^{N}\xi_j\right], \end{aligned} \tag{3.9}$$

where $M < N$ and the convention $\prod_{N+1}^{N}(\cdot) = 1$ is used. We apply the reversal technique from Liu (2010) (see also (2.7)), which is justified by the exchangeability property of the sequence $\{\xi_i\}_{i=1}^{N}$, and consider the reversed sequence which is equal in distribution. Conditioned on $W_M = x$, we obtain

$$W_N^{x,M} \stackrel{d}{=} \max\left[0, xe^{-\bar{\alpha}(N-M)} \prod_{i=1}^{N-M} \xi_i - G \sum_{i=0}^{N-M-1} e^{-\bar{\alpha}i} \prod_{j=1}^{i} \xi_j\right].$$

Let $\{Z_i\}$ be the account value process when there are no withdrawals, beginning with $Z_0 = 1$. Then

$$W_N^{x,M} \stackrel{d}{=} \max\left[0, xZ_{N-M} - G \sum_{i=0}^{N-M-1} Z_i\right]$$

and in particular, with $M = 0$, $x = P$, and $G = P/N$ we obtain

$$W_N \stackrel{d}{=} P \max\left[0, Z_N - \frac{1}{N} \sum_{i=0}^{N-1} Z_i\right]. \tag{3.10}$$

As pointed out by Liu (2010), V_0 can be expressed as an Asian (floating strike) Call option on $\{Z_i\}$ plus a term certain component.

Many of the terminal nodes in the tree for $\{W_i\}$ will be zero as a result of the periodic withdrawals, fees, and possible negative returns on S. Consider the recombining tree for $\{Z_i\}$ with $N + 1$ nodes for period N. At each node, for each path leading to it the average must be computed to calculate W_N. Suppose that for some $i \leq N$ we have $W_N = 0$ on all paths with i jumps of u and $N - i$ jumps of d. Then $W_N = 0$ for all paths with less than i jumps of u. Consequently, once we reach a node on the tree for Z such that $W_N = 0$ for all paths, no further paths need be considered.

There is an efficient permutation function in C++, *next_permutation*, which quickly loops through all distinct paths having i jumps of u and $N - i$ jumps of d. By looping through each node and all the paths at each node we can avoid the exponential growth in memory storage although we show in our numerical results that the run-time will increase significantly. By (3.4), with $\zeta := N - m$ we can write

$$v(m,x) = Ga_{\overline{\zeta}|} + e^{-\bar{r}\zeta} \sum_{k=0}^{a} p^{\zeta-k} q^k \sum_{\Xi_{\zeta,k}} \max\left(xe^{-\bar{\alpha}\zeta} u^{\zeta-k} d^k \right. \\ \left. - G \left[\sum_{i=0}^{\zeta-1} e^{-\bar{\alpha} i} \left[\prod_{j=1}^{i} \omega_j \right] \right], 0 \right), \tag{3.11}$$

where $\Xi_{\zeta,k}$ is the set of $\binom{\zeta}{k}$ unique permutations of a path with $\zeta - k$ up and k down movements and a is the first value of k for which the summand produces zero.

The continuity and monotonicity properties for V as a function of α presented in Lemma 2.5 and the resulting existence and uniqueness of $\alpha^\star$ remain true in the discrete binomial framework. However, in a finite probability space $\mathbb{Q}(W_N > 0) = 0$ for sufficiently large α. Consequently, strict monotonicity holds only on a bounded interval for α.

Lemma 3.7. *For all fixed $(i,x) \in \mathcal{I}_{N-1} \times \mathbb{R}_{++}$, the contract value function $v(i,x)$, defined by (3.6), as a function of α is continuous for $\alpha \geq 0$ and strictly decreasing on $[0, b^{x,i})$ where*

$$b^{x,i} := \min\{\alpha \geq 0 : W_N^{x,i} = 0 \text{ a.s.}\} < \infty.$$

Further, if (i,x) satisfies

$$x > G \sum_{j=1}^{N-i} d^j \tag{3.12}$$

then $b^{x,i} > 0$, otherwise $b^{x,i} = 0$. For $\alpha \geq b^{x,i}$, $v(x,i) = Ga_{\overline{N-i}|}$.

Proof. From the equivalent expression for $v(i,x)$ in (3.11), the continuity result is immediate. The maximum possible value for $W_N^{x,i}$ is obtained by the path corresponding to $\omega_j = u$ for all $j = i+1, \ldots, N$. Thus

$$b^{x,i} = \min\{\alpha \geq 0 : W_N^{x,i}(uu \ldots u) = 0\}.$$

From (3.9), $W_N^{x,i}(uu\dots u) = 0$ if and only if

$$f(\alpha) := \left(x(e^{-\bar{\alpha}}u)^{N-i} - G \sum_{j=0}^{N-i-1} (e^{-\bar{\alpha}}u)^j \right) \leq 0.$$

But $f \in C^\infty$ and $\lim_{\alpha \to \infty} f(\alpha) = -G < 0$. We have $f(0) > 0$ if and only if (3.12) holds. If $f(0) > 0$ then there exists $0 < b^{x,i} < \infty$. If $f(0) \leq 0$, then $b^{x,i} = 0$. The remaining part of this proof is similar to Lemma 2.5. Assume (i, x) is such that $b^{x,i} > 0$. Let

$$A^\alpha := \{W_N^{x,i}(\alpha) > 0\}.$$

Then $A^\alpha \neq \emptyset$ for $\alpha < b^{x,i}$. Fix $\alpha \in [0, b^{x,i})$ and consider $\alpha^{(1)}$ such that $\alpha < \alpha^{(1)} < b^{x,i}$. When restricted to the set $A^{\alpha^{(1)}}$, (3.9) implies

$$0 < W_N^{x,i}(\alpha^{(1)}) < W_N^{x,i}(\alpha),$$

which in turn implies $A^{\alpha^{(1)}} \subseteq A^\alpha$. We conclude that $v(i, x; \alpha^{(1)}) < v(i, x; \alpha)$. □

In particular, (3.12) holds for $(i, x) = (0, P)$ since $G = P/N$ and $d < 1$. The existence and uniqueness of $\alpha^\star$ is discussed in the next proposition. The proof is deferred to Subsection 3.2.3.

Proposition 3.8. *Under Assumption 3.4 there exists a unique $\alpha^\star \in [0, b^{P,0})$ such that $V_0(P, \alpha^\star, g) = P$.*

Remark 3.9. For $r = 0$ we have $V_0(P, \alpha, g) = P$ for all $\alpha \geq b^{P,0}$. Thus $r > 0$ is a necessary condition to ensure uniqueness of $\alpha^\star$. When uniqueness fails to hold it makes sense to define the fair fee as $\alpha^\star = \inf\{\alpha : V_0(P, \alpha, g) = P\}$. When $r = 0$, $\alpha^\star = b^{0,P}$ although it should make no difference to charge any higher rate since in any case $W_N = 0$ a.s. In Chapter 2 it was shown in the trivial case of the model with surrenders and $k_0 = 0$ that we can have non-uniqueness even when $r > 0$ and $\mathbb{Q}(W_N = 0) < 1$.

3.2.3 Insurer Perspective

We begin by defining the discrete-time analogues of the trigger time from the continuous model.

Definition 3.10. In the binomial model, the trigger time τ is defined as the stopping time

$$\tau(\omega_1 \dots \omega_N) := \inf\{i \geq 1; W_i(\omega_1 \dots \omega_i) = 0\},$$

where $\inf(\emptyset) = \infty$. For any fixed sequence $\bar{\omega}_i$ and for any $k \leq i$ we write $\tau(\bar{\omega}_i) \leq k$ if $(\bar{\omega}_i \omega_{i+1} \dots \omega_N) \in \{\tau \leq k\}$ for all possible paths $(\bar{\omega}_i \omega_{i+1} \dots \omega_N)$, where $\omega_j \in \{u, d\}$ for all $i + 1 \leq j \leq N$.

It is convenient to define the respective non-decreasing sequences of stopping times $\{\tau_i\}_{i=0,1,\dots,N}$ and $\{\bar{\tau}_i\}_{i=0,1,\dots,N}$ with $\tau_i := \tau \vee i$ and $\bar{\tau}_i := \tau_i \wedge N$ for $i \in \mathcal{I}_N$. For $0 \leq j \leq i \leq N$ and $k \in \{i, i+1, \dots, N\} \cup \{\infty\}$, by the Markov property of $\{W_i\}$ we have

$$\mathbb{Q}(\tau_j = k | \mathcal{F}_i) = H(i, j, k, W_{i^-}), \tag{3.13}$$

where

$$H(k \wedge N, j, k, x) = \begin{cases} \mathbf{1}_{\{x>0, w(x)=0\}} & k \leq N, \\ \mathbf{1}_{\{w(x)>0\}} & k = \infty, \end{cases}$$

and for $i \vee 1 \leq l < k \wedge N$

$$H(l, j, k, x) = \begin{cases} pH(l+1, j, k, w(x)u) + qH(l+1, j, k, w(x)d) & x > 0, \\ 0 & x = 0. \end{cases}$$

For $i = 0$, we have

$$H(0, 0, k, x) = pH(1, 0, k, xu) + qH(1, 0, k, xd).$$

If $\tau = \infty$ the contract matures with a positive account value at time $N\delta t = T$ and the option is not exercised. Similar to the comment made in Subsection 2.2.2 (on page

24), if $\tau < \infty$ then at time τ no uncertainty remains and any hedging portfolios can be liquidated at that time and the amount $Ga_{\overline{N-\tau}|}$ paid to the insured. It is simpler to model the contract until period N even if trigger occurs earlier although this will no longer be the case once mortality is introduced.

Since the value processes at each timepoint are ex-fees and ex-withdrawals, the component $(G - W_{\tau^-} e^{-\bar{\alpha}}) \geq 0$ is the rider payment made immediately at trigger time. For any period i, the net rider payout at time $i\delta t$ is $(G - W_{i^-} e^{-\bar{\alpha}})^+ - W_{i^-}(1 - e^{-\bar{\alpha}})$. Then the discrete-time version of (2.18) is

$$
\begin{aligned}
U_i &= E_Q\left[\sum_{j=i+1}^{N} e^{-\bar{r}(j-i)} \left[\left(G - W_{j^-} e^{-\bar{\alpha}}\right)^+ - W_{j^-}\left(1 - e^{-\bar{\alpha}}\right)\right] |\mathcal{F}_i\right] \\
&= E_Q\Big[\left(G - W_{\bar{\tau}_i^-} e^{-\bar{\alpha}}\right) e^{-\bar{r}(\bar{\tau}_i - i)} \mathbf{1}_{\{i+1 \leq \bar{\tau}_i\}} + \sum_{m=\bar{\tau}_i+1}^{N} G e^{-\bar{r}(m-i)} \\
&\quad - \sum_{m=i+1}^{\bar{\tau}_i} e^{-\bar{r}(m-i)} W_{m^-}\left(1 - e^{-\bar{\alpha}}\right) |\mathcal{F}_i\Big]
\end{aligned}
\tag{3.14}
$$

for $i \in \mathcal{I}_{N-1}$. The terminal value is $U_N = 0$.

By the Markov property for $\{W_i\}$ we have

$$U_i = u(i, W_i),$$

where $u : \mathcal{I}_N \times \mathbb{R}_+ \mapsto \mathbb{R}$ is defined by[3]

$$
u(i,x) = \begin{cases} 0 & i = N, \\ e^{-\bar{r}}[pu^-(i+1, xu) + qu^-(i+1, xd)] & 0 \leq i < N, \end{cases}
\tag{3.15}
$$

$u^- : \mathcal{I}_N^+ \times \mathbb{R}_+ \mapsto \mathbb{R}$ is defined by

$$
u^-(i,x) = u(i, w(x)) + (G - xe^{-\bar{\alpha}})^+ - x(1 - e^{-\bar{\alpha}}),
\tag{3.16}
$$

and $w(x)$ is provided by (3.7). The function $u^-(i,x)$ represents the rider value at time-

[3]There is some abuse of notation with u referring both to the up-movement in the binomial model and to the rider value function. However, it is always clear from the context whether we are referring to the constant value or to the rider value function.

point i cum-fees and cum-withdrawals, where x is the AV before fees and withdrawals are deducted.

Since the processes (3.4) and (3.14) are the respective discrete-time versions of (2.16) and (2.18) it is expected that the $V = U + W$ relationship proved in Proposition 2.8 holds in the binomial model. Indeed this can be shown directly from (3.4) and (3.14). We provide an alternative proof applying backward induction to the functions $v(i, x)$ and $u(i, x)$.

Proposition 3.11. *Under Assumption 3.4, for all* $\alpha \geq 0$ *we have*

$$V_i = U_i + W_i$$

for all $i = 0, 1, \ldots, N$.

Proof. We apply backward induction and show that $v(i, x) = u(i, x) + x$ for all $(i, x) \in \mathcal{I}_N \times \mathbb{R}_+$. By definition $v(N, x) = u(N, x) + x$ for all $x \in \mathbb{R}_+$. Assume $v(i, x) = u(i, x) + x$ holds for all $x \in \mathbb{R}_+$ for some $1 \leq i \leq N$. We need to show that $v(i - 1, y) = u(i - 1, y) + y$ for all $y \in \mathbb{R}_+$. Applying the induction hypothesis,

$$\begin{aligned} v(i-1, y) &= e^{-\bar{r}}[G + pv(i, w(uy)) + qv(i, w(dy))] \\ &= e^{-\bar{r}}[pu(i, w(uy)) + qu(i, w(dy)) + p(w(uy) + G) + q(w(dy) + G)]. \end{aligned}$$

From (3.15) and (3.16) we have

$$\begin{aligned} u(i-1, y) = e^{-\bar{r}}\{&pu(i, w(uy)) + qu(i, w(dy)) + p[(G - uye^{-\bar{\alpha}})^+ - uy(1 - e^{-\bar{\alpha}})] \\ &+ q[(G - dye^{-\bar{\alpha}})^+ - dy(1 - e^{-\bar{\alpha}})]\}. \end{aligned}$$

Observe

$$w(y) - (G - ye^{-\bar{\alpha}})^+ = ye^{-\bar{\alpha}} - G.$$

Then

$$w(y) + G - (G - ye^{-\bar{\alpha}})^+ + y(1 - e^{-\bar{\alpha}}) = y,$$

therefore

$$
\begin{aligned}
v(i-1,y) - u(i-1,y) &= e^{-\bar{r}}[puy + qdy] \\
&= y
\end{aligned}
$$

since $pu + qd = e^{\bar{r}}$ by the definition of the risk-neutral probabilities (3.3).

Therefore

$$v(i-1,y) = u(i-1,y) + y$$

for all $y \in \mathbb{R}_+$ and the result holds. □

Proof of Proposition 3.8. From Lemma 3.7, for $\alpha \geq b^{P,0} > 0$, we have $V_0(P,\alpha,g) = Ga_{\overline{N}|} < P$ for $r > 0$. By the definition of U in (3.14) we have $U \geq 0$ for $\alpha = 0$. By Proposition 3.11,

$$V_0(P,\alpha=0,g) = U_0(P,\alpha=0,g) + P \geq P.$$

By the continuity and strictly decreasing property from Lemma 3.7, there exists a unique $\alpha^\star \in [0, b^{P,0})$. □

3.2.4 Hedging

In the binomial asset pricing model with one risky asset and the money market account, a contingent claim can be perfectly hedged through discrete-time rebalancing because there are only two possible movements in the underlying asset each period.

Consider first the no-hedging strategy whereby the fee revenues are placed in the money market and at time τ, if $\tau < \infty$, the rider payoff is paid from this account. The $\mathcal{F}_{\bar{\tau}_0}$-measurable random variable $\mathcal{C}_{\bar{\tau}_0}$ measures the total cost of the rider to the insurer over the contract lifespan, discounted to time zero, when hedging is not used. We introduce notation for the periodic fees, with $F_i := W_{i^-}(1 - e^{-\bar{\alpha}})$ for $i \in \mathcal{I}_N^+$ and

$F_0 = 0$. We have

$$\mathcal{C}_{\bar{\tau}_0} = e^{-\bar{r}\bar{\tau}_0}\left[\left(G - (W_{\bar{\tau}_0^-})e^{-\bar{\alpha}}\right)^+ + Ga_{\overline{N-\bar{\tau}_0}|} - \sum_{i=1}^{\bar{\tau}_0} F_i \times e^{\bar{r}(\bar{\tau}_0 - i)}\right].$$

Note that $U_0 = E_{\mathbb{Q}}[\mathcal{C}_{\bar{\tau}_0}]$, but we are concerned with the pathwise results of $\mathcal{C}_{\bar{\tau}_0}$ in relation to the outcomes resulting from a dynamic hedging strategy.

To dynamically hedge the rider the insurer establishes a hedging portfolio, which attempts to replicate the rider so that any rider claims can be fully paid out by the portfolio. The party managing the rider risk does not have access to the account value funds to mitigate any risk, rather the only sources of revenue are the rider fees. Denoting the replicating portfolio by $\{X_i\}$, the objective is to have $X_i = U_i$ for all i in a pathwise manner.

Following Shreve (2004a), we define the adapted portfolio process $\{\Delta_i\}_{0 \le i \le N-1}$. On each time interval $[i\delta t, (i+1)\delta t)$ until maturity and for all outcomes the replicating portfolio should maintain a position of $\Delta_i(\omega_1 \ldots \omega_i)$ units in S. Using the Markov property of $\{W_i\}$ we define

$$\Delta_i := \Delta(i, W_i, S_i),$$

where $\Delta : \mathcal{I}_{N-1} \times \mathbb{R}_+ \times \mathbb{R}_+ \mapsto \mathbb{R}$ is given by

$$\Delta(i, x, y) = \frac{u^-(i+1, ux) - u^-(i+1, dx)}{uy - dy}. \tag{3.17}$$

This indicates that $\Delta_i = 0$ for $\tau \le i \le N-1$. Hedging with S is no longer required after the trigger time as no uncertainty remains. By the nature of the rider as an embedded put-like option, Δ will always take non-positive values corresponding to short positions in S. Any positive (negative) portfolio cash balance is invested in (borrowed from) the money market.

Beginning with initial capital $X_0 = x_0 \in \mathbb{R}$, the replicating portfolio $\{X_i\}$ follows

$$X_i = (X_{i-1} - \Delta_{i-1}S_{i-1})\, e^{\bar{r}} + \Delta_{i-1}S_i + F_i - (G - W_{i^-}e^{-\bar{\alpha}})^+ \tag{3.18}$$

for $i \in \mathcal{I}_N^+$. Over any period the change in the portfolio value of $X_i - X_{i-1}$ consists of the sum of four components: a) the return in the money market earned on both the prior portfolio balance and the proceeds from the shorted stock $(X_{i-1} - \Delta_{i-1} S_{i-1})(e^{\bar{r}} - 1)$; b) the capital gain or loss on the shorted stock $(S_i - S_{i-1})\Delta_{i-1}$; c) the end of period rider fees F_i; and d) the negative of that period's rider claim (if any), paid at the end of the period and given by $(G - W_{i^-} e^{-\bar{\alpha}})^+$. Note that if the static hedging strategy $\Delta \equiv 0$ is used then $X_N e^{-\bar{r}N} = -\mathcal{C}_{\bar{\tau}0}$. That is, we just obtain the result from no-hedging.

This next theorem is similar to Shreve (2004a, Theorem 2.4.8), and the proof follows from there. As such, we omit the proof and provide it in Section 3.3 when we generalize the result for lapses.

Theorem 3.12. *Under Assumption 3.4, if the fee α is charged and the initial capital is $x_0 = U_0(P, \alpha, g)$, then an insurer who maintains the replicating portfolio X_i by following the portfolio process prescribed by (3.17) will be fully hedged. That is,*

$$X_i = U_i$$

for $i \in \mathcal{I}_N$.

Remark 3.13. In particular, if $\tau \leq N$ then $X_\tau = G \times a_{\overline{N-\tau|}}$. When $\alpha^\star$ is charged we have $U_0 = 0$ and no initial capital is required for the replicating portfolio. The rider is different from the standard financial options in that there is no upfront cost to finance the hedge but rather it is self-financed through periodic contingent fees. If the fee charged is not the fair fee ($\alpha \neq \alpha^\star$), then the insurer must make an initial deposit to the hedging portfolio if $\alpha < \alpha^\star$ or may consume from the portfolio at time zero if $\alpha > \alpha^\star$. The insurer can justify a lower fee by either depositing capital into the portfolio and selling the policy at a loss or by charging an initial fee per unit premium at time zero to the insured. Likewise, charging a higher fee results in a time zero profit of U_0.

3.3 Extending the Model: Surrenders

We extend the binomial pricing model to include the possibility for early surrenders. See the beginning of Subsection 2.2.4 for a general discussion on surrenders.

Assumption 3.14. *We modify Assumption 3.4 by allowing for early surrenders. Surrenders occur at the end of any time period, after the fees and withdrawals have been deducted. For valuation purposes, the end of period time point is considered ex-post fees and withdrawals but ex-ante surrenders.*

Let $k^a : \{0, 1, \dots T\} \mapsto [0, 1]$ be the non-increasing function describing the surrender charge schedule, satisfying $k^a_0 > 0$ and $k^a_T = 0$. The surrender charge rate k^a_i is applied for surrenders during time $[i, i + 1)$. We denote the corresponding function for the surrender charge rate upon surrender at the end of period i by $k : \{0, 1, \dots, N\} \to [0, 1]$. Then $k_i = k^a_{\lfloor i\delta t \rfloor}$. By the discussion in Subsection 2.2.4, for all $i \in \mathcal{I}_N$ we have

$$V_i = \max_{\eta \in \mathbb{L}_i} V_i^\eta = \max_{\eta \in \mathbb{L}_{i,\bar{\tau}_i}} V_i^\eta, \tag{3.19}$$

where

$$V_i^\eta = E_Q \left[Ga_{\overline{\eta - i|}} + W_\eta (1 - k_\eta) e^{-\bar{r}(\eta - i)} | \mathcal{F}_i \right] \tag{3.20}$$

and $\mathbb{L}_{i,\bar{\tau}_i}$ is the set of $\mathbb{F}-$adapted stopping times taking values in $\{i, i + 1, \dots, N\}$ subject to the constraint $\eta < \bar{\tau}_i$ or $\eta = N$. Recall that $\bar{\tau}_i$ is the trigger time assuming no lapses. The fair rider fee satisfies $V_0(P, \alpha^\star, g) = P$.

With the objective of classifying the optimal surrender policy we introduce some notation. For any $0 \le i \le N$, define a rescaled filtration $\mathbb{F}^i = \{\mathcal{F}^i_j := \mathcal{F}_{j+i}; 0 \le j \le N - i\}$. For any $\eta \in \mathbb{L}_i$ define

$$Y^{\eta,i} := \left\{ Y_j^{\eta,i} = e^{-\bar{r}((j+i)\wedge\eta)} V^\eta_{(j+i)\wedge\eta} + Ga_{\overline{(j+i)\wedge\eta|}} \right\}_{0 \le j \le N-i}, \tag{3.21}$$

then $Y^{\eta,i}$ is a $(\mathbb{Q}, \mathbb{F}^i)$ martingale. Define

$$\tilde{\eta}_i := \min\{j \geq i;\ V_j = W_j(1-k_j)\} \leq N, \tag{3.22}$$

then it is a well-known result from American contingent claims theory that $\tilde{\eta}_i$ is optimal in the sense that $V_i = V_i^{\tilde{\eta}_i}$ (proving this in our context is straightforward based on Duffie (2001, p.35) but requires (3.21)). That is, $\tilde{\eta}_i$ is an optimal lapsation policy for the insured to follow going forth from time $i\delta t$, given the current market state and no prior surrender.

The backward induction scheme is constructed to evaluate V on a binomial tree. By the Markov property for $\{W_i\}$ we have

$$V_i = v(i, W_i),$$

where $v : \mathcal{I}_N \times \mathbb{R}_+ \mapsto \mathbb{R}_+$ is given recursively as

$$\begin{cases} v(N,x) = x(1-k_N) = x, \\ v(i,x) = \max\{(G + pv(i+1, w(ux)) + qv(i+1, w(dx)))e^{-\bar{r}}, x(1-k_i)\}. \end{cases}$$

When solving for $\alpha^\star$ we may write $v(0,P) = [G + pv(1, w(uP) + qv(1, w(dP))]e^{-\bar{r}}$, since $k_0 > 0$.

We turn towards the rider value U and option to surrender value L. Naturally the discrete-time version of (2.23) is given by

$$U_i := \max_{\eta \in \mathbb{L}_{i,\bar{\tau}_i}} E_{\mathbb{Q}}[\sum_{j=i+1}^{\eta} e^{-\bar{r}(j-i)}[(G - W_{j^-}e^{-\bar{\alpha}})^+ - W_{j^-}(1-e^{-\bar{\alpha}})] - e^{-\bar{r}(\eta-i)}k_\eta W_\eta | \mathcal{F}_i] \tag{3.23}$$

where $\sum_{j=i+1}^{i}(\cdot) = 0$. Recall that $L_i := U_i - U_i^{NL} \geq 0$, where U_i^{NL} is the rider value in the no-lapse case (3.14). Then

$$L_i = \max_{\eta \in \mathbb{L}_{i,\bar{\tau}_i}} L_i^\eta, \tag{3.24}$$

where

$$L_i^\eta = E_{\mathbb{Q}}\left[\sum_{j=\eta+1}^{N} e^{-\bar{r}(j-i)}\left[W_{j^-}\left(1-e^{-\bar{\alpha}}\right)-\left(G-W_{j^-}e^{-\bar{\alpha}}\right)^+\right]-e^{-\bar{r}(\eta-i)}k_\eta W_\eta|\mathcal{F}_i\right].$$

We write

$$U_i = u(i, W_i),$$

where u:$\mathcal{I}_N \times \mathbb{R}_+ \mapsto \mathbb{R}$ is recursively defined by

$$\begin{cases} u(N,x) = -k_N x = 0, \\ u(i,x) = \max\{e^{-\bar{r}}[pu^-(i+1,ux)+qu^-(i+1,dx)], -k_i x\}, \end{cases}$$

and $u^- : \mathcal{I}_N^+ \times \mathbb{R}_+ \mapsto \mathbb{R}$ follows

$$u^-(i,x) = u(i,w(x)) + (G - xe^{-\bar{\alpha}})^+ - x(1-e^{-\bar{\alpha}}). \tag{3.25}$$

Denoting the rider value function in the no-lapse model from (3.15) by $u^{NL}(i,x)$, we have $L_i = l(i, W_i)$, where $l : \mathcal{I}_N \times \mathbb{R}_+ \mapsto \mathbb{R}_+$ is given by

$$\begin{cases} l(N,x) = -k_N x = 0, \\ l(i,x) = \max\{e^{-\bar{r}}(pl(i+1,w(ux))+ql(i+1,w(dx))), -u^{NL}(i,x) - k_i x\}. \end{cases}$$

This definition of l satisfies $l = u - u^{NL}$ as can be shown using backwards induction. Note that $u^{NL}(i,0) \geq 0$ which implies the boundary condition $l(i,0) = 0$. Once the rider is triggered, early surrender is suboptimal since the account value is zero and any remaining guarantee is forfeited upon surrender.

Proposition 3.15. *Under Assumption 3.14, for all $\alpha \geq 0$ and for all $i \in \mathcal{I}_N$, we have*

$$V_i = U_i + W_i, \tag{3.26}$$

or equivalently

$$V_i = L_i + U_i^{NL} + W_i. \tag{3.27}$$

Proof. Equation (3.26) can be proved using backward induction on the recursive functions v and u, similar to Proposition 3.11. We omit the details. □

Remark 3.16. From (3.26) we also have $\hat{\eta}_i = \min\{j \geq i;\ U_j = -k_j W_j\}$.

The adapted portfolio process $(\Delta_i)_{0\leq i<N}$ is defined similarly to (3.17). We have

$$\Delta_i = \Delta(i, W_i, S_i),$$

where $\Delta : \mathcal{I}_N \times \mathbb{R}_+ \times \mathbb{R}_+ \mapsto \mathbb{R}$ is defined as

$$\Delta(i, x, y) = \frac{u^-(i+1, ux) - u^-(i+1, dx)}{uy - dy}, \tag{3.28}$$

and $u^-(i, x)$ is given by (3.25).

We define a sequence of stopping times which classify suboptimal behaviour. Recall $\tilde{\eta}_i$ from (3.22). Let $\tilde{\eta}^0 := \tilde{\eta}_0$ and for $1 \leq j \leq m$ we denote

$$\tilde{\eta}^j = \tilde{\eta}_{z_j},$$

where $z_0 = 0$, $z_j = (\tilde{\eta}^{j-1} + 1) \wedge N$, and $m = \min\{i; \tilde{\eta}^i = N \text{ a.s.}\}$.

We introduce a consumption process $C = \{C_i\}_{0\leq i<N}$ where $C_i := c(i, W_i)$ and this process is linked to the suboptimal behaviour. It represents the additional cash flow received each time a policyholder behaves suboptimally by not surrendering. The function $c : \mathcal{I}_{N-1} \times \mathbb{R}_+ \mapsto \mathbb{R}_+$ is defined by

$$c(i, x) := v(i, x) - [pv(i+1, w(ux)) + qv(i+1, w(dx)) + G]e^{-\bar{r}} \geq 0. \tag{3.29}$$

Note that we can characterize, in terms of $\{\tilde{\eta}^j\}$, precisely when C will be strictly positive. We have $C_{\tilde{\eta}^j} > 0$ for all $0 \leq j < M := \min\{b; z_b = N\} \leq m$, where M is a random variable. Otherwise $C_i = 0$.

There is a fine distinction between $C_{\tilde{\eta}^j}$ and $L_{\tilde{\eta}^j}$ for all $j < M$. Consider the two surrender strategies of $\tilde{\eta}^{j+1}$ and $\eta = N$. The first strategy corresponds to surrendering at the next best time after $\tilde{\eta}^j$ and the latter strategy is equivalent to never surrendering early. Then $C_{\tilde{\eta}^j} = V_{\tilde{\eta}^j} - V_{\tilde{\eta}^j}^{\tilde{\eta}^{j+1}}$ but $L_{\tilde{\eta}^j} = V_{\tilde{\eta}^j} - V_{\tilde{\eta}^j}^{NL}$. At any time when it is optimal to surrender immediately, C provides the marginal value from surrendering now instead of at the next optimal time, whereas L is the marginal value from acting now instead of at maturity.

By Proposition 3.11 and Proposition 3.15 it is true that $V_{\tilde{\eta}^j}^{\tilde{\eta}^{j+1}} = U_{\tilde{\eta}^j}^{\tilde{\eta}^{j+1}} + W_{\tilde{\eta}^j}$ and $V_{\tilde{\eta}^j} = U_{\tilde{\eta}^j} + W_{\tilde{\eta}^j}$. Therefore C can be written in terms of U as

$$c(i,x) = u(i,x) - [pu^{-}(i+1,ux) + qu^{-}(i+1,dx)]e^{-\bar{r}}. \tag{3.30}$$

The next theorem extends the hedging results presented in Shreve (2004a, Theorem 4.4.4) by incorporating the complication of the periodic revenues and rider claims, and shows that the insurer can perfectly hedge the rider risk by maintaining a replicating portfolio following (3.28). Furthermore, the insurer may have positive consumption under suboptimal surrender behaviour. Beginning with $X_0 = x_0$, the replicating portfolio is constructed forward recursively. For all $i \in \mathcal{I}_N^+$ we have

$$X_i = [X_{i-1} - \Delta_{i-1}S_{i-1} - C_{i-1}]\, e^{\bar{r}} + \Delta_{i-1}S_i + F_i - (G - W_{i^-}e^{-\bar{\alpha}})^+. \tag{3.31}$$

Theorem 3.17. *Under Assumption 3.14, if the fee α is charged and the initial capital is $x_0 = U_0$, then an insurer who maintains the replicating portfolio X_i defined by (3.31) by following the portfolio process (3.28), depositing any fee revenue into the portfolio, consuming when permitted, paying any rider claims as they come due and liquidating the portfolio either upon early surrender (if any) or at timepoint N will be fully hedged throughout the contract lifespan. More generally, for all $i \in \mathcal{I}_N$ and all surrender strategies, we have*

$$X_i = U_i.$$

Proof. Following the approach presented in Shreve (2004a), we proceed by induction. By assumption we have that $X_0 = U_0$. Assume for some $0 \leq i < N$ that $X_i = U_i$.

We need to show that for all $\bar{\omega}_i$,

$$X_{i+1}(\bar{\omega}_i u) = U_{i+1}(\bar{\omega}_i u),$$

$$X_{i+1}(\bar{\omega}_i d) = U_{i+1}(\bar{\omega}_i d).$$

We omit the $\bar{\omega}_i$ notation for conciseness. Substituting U_i for X_i in (3.31), using (3.28), (3.30), and the fact $q = \frac{u-e^{\bar{r}}}{u-d}$ we obtain

$$\begin{aligned} X_{i+1}(u) &= \Delta_i S_i(u - e^{\bar{r}}) + (U_i - C_i)e^{\bar{r}} + F_{i+1}(u) - (G - W_{i+1^-}(u)e^{-\bar{\alpha}})^+ \\ &= q[u^-(i+1, uW_i) - u^-(i+1, dW_i)] + (pu^-(i+1, uW_i) + qu^-(i+1, dW_i) \\ &\quad + F_{i+1}(u) - (G - W_i ue^{-\bar{\alpha}})^+ \\ &= u^-(i+1, uW_i) + F_{i+1}(u) - (G - W_i ue^{-\bar{\alpha}})^+ \\ &= u(i+1, w(uW_i)) \\ &= U_{i+1}(u). \end{aligned}$$

A similar argument shows that $X_{i+1}(d) = U_{i+1}(d)$. Since $\bar{\omega}_i$ was arbitrary we have $X_{i+1} = U_{i+1}$ and the result holds. □

Remark 3.18. Remark 3.13 remains true. Assuming the insured follows the optimal surrender strategy $\tilde{\eta}_0$, then $X_{\tilde{\eta}_0} = U_{\tilde{\eta}_0}$ and on $\{\tilde{\eta}_0 < \bar{\tau}_0\}$ we have that $X_{\tilde{\eta}_0} = U_{\tilde{\eta}_0} = -k_{\tilde{\eta}_0} W_{\tilde{\eta}_0}$, whereas $X_N = U_N = 0$ on $\{\tilde{\eta}_0 = N\}$. There is no consumption. If the insured allows the first optimal surrender time $\{\tilde{\eta}_0 < \bar{\tau}_0\}$ to elapse, then the insurer will consume $C_{\tilde{\eta}_0}$ and the remaining portfolio is still sufficient to hedge the contract over the remaining lifespan. If the insured allows the next optimal surrender time $\{\tilde{\eta}_0 < \tilde{\eta}^1 < \bar{\tau}_0\}$ to elapse, if it exists, then the insurer consumes an additional $C_{\tilde{\eta}^1}$ and this continues until the earlier of trigger or timepoint N.

Finally suppose the insured surrenders at a suboptimal time. For a given path $\bar{\omega}_N$, surrender occurs at a timepoint $i \neq \tilde{\eta}^j$ for all $0 \leq j \leq M(\bar{\omega}_N)$. Then the insured receives $W_i(1-k_i)$ and in turn foregoes $V_i - W_i(1-k_i) > 0$ of value. The insurer's portfolio value is $X_i + k_i W_i > 0$ and the insurer has a positive consumption. Indeed by (3.26) we have $V_i - W_i(1-k_i) = U_i + W_i k_i > 0$, but $X_i = U_i$.

3.4 Binomial Asian Approximation Method

The contract (VA plus GMWB) can be decomposed into an Asian-type option, as discussed in Chapter 2 (p. 22). Hull and White (1993) developed an approximation method to value path-dependent financial options on a binomial lattice in a more efficient manner. The key idea is to use only a representative set of averages at each node and apply linear interpolation in the backwards induction scheme. A summary of this method and related papers is provided in Costabile et al. (2006).

The following drawbacks of the Hull and White (1993) method are discussed in Costabile et al. (2006). It is highly sensitive to a parameter h which controls the number of representative averages considered at each node. Further, for any given timestep the same set of averages is applied for each node. Finally, the set of representative averages do not correspond to actual averages from the original non-recombining lattice. It has been shown that the method in Hull and White (1993) does not converge, since only a fixed h is considered.

Costabile et al. (2006) propose an approximation method which addresses these issues and can be used for pricing European and American Asian (fixed strike) calls. The method can be easily modified for any option payoff which depends on a valid function of the asset price path. A proof of convergence is not provided but numerical results show convergence for European Asian calls while American Asian calls do not perform as well and appear to converge at a much slower rate. The options considered by Costabile et al. (2006) have significantly shorter maturities compared to the GMWB riders. The method reduces the number of contract values considered in the backwards induction scheme from $O(2^N)$ to $O(N^4)$. In our work, memory constraints limited the number of time steps in the binomial trees to $N = 28$ but with this method we can consider up to $N = 128$ timesteps. We briefly describe the approximation method applied to GMWBs with lapses but refer the reader to Costabile et al. (2006) for more details.

Using (3.9)[4] we can rewrite (3.19) as

$$V_0 = \max_{\eta \in \mathbb{L}_0} E_Q \left[G a_{\overline{\eta}|} + P \max\left(Z_\eta \left(1 - \frac{1}{N} \sum_{i=1}^{\eta} \frac{1}{Z_i} \right), 0 \right) (1 - k_\eta)\, e^{-\bar{r}\eta} \right], \qquad (3.32)$$

where

$$Z_n = \prod_{i=1}^{n} e^{-\bar{\alpha}} \xi_i = e^{-\bar{\alpha}\, n} \frac{S_n}{S_0}.$$

We have $V_i = v(i, Z_i, \sum_{j=1}^{i} Z_j^{-1})$, where $v : \mathcal{I}_N \times \mathbb{R}_+ \times \mathbb{R}_+ \mapsto \mathbb{R}_+$ is recursively defined as follows. The backward induction scheme begins with $i = N$ and

$$v(N, x, y) = P \max\left(x \left(1 - \frac{1}{N} y \right), 0 \right).$$

For $0 \le i < N$,

$$\begin{aligned} v(i, x, y) = \ \max \Bigg[\bigg[G + p v\Big(i+1, xue^{-\bar{\alpha}}, y + \left(xue^{-\bar{\alpha}}\right)^{-1}\Big) \\ + q v\Big(i+1, xde^{-\bar{\alpha}}, y + \left(xde^{-\bar{\alpha}}\right)^{-1}\Big) \bigg] e^{-\bar{r}}, x \left(1 - \frac{1}{N} y\right) (1 - k_i) \Bigg]. \end{aligned}$$

Let (i, j) denote the node reached by j up-movements and $(i-j)$ down-movements in the recombining tree for Z. We write $z(i, j)$ for the value of Z at node (i, j). For each node, we construct a set of $j(i-j)+1$ representative averages[5] which is a subset of the complete set of $\binom{i}{j}$ averages for the paths at that node. Denote the first (and lowest) element by $A(i, j, 1)$ where

$$A(i, j, 1) = \sum_{h=0}^{j} \left(ue^{-\bar{\alpha}}\right)^{-h} + \left(ue^{-\bar{\alpha}}\right)^{-j} \sum_{h=1}^{i-j} \left(de^{-\bar{\alpha}}\right)^{-h}.$$

This average is taken along the path beginning with j up-movements of u and followed

[4] For the no-lapse case, (3.10) may also be used. It cannot be used to model lapses since it is equivalent in distribution only and not pathwise.

[5] We keep the terminology of *average* even though we do not divide by $i + 1$.

by $(i-j)$ down-movements of d. Excluding the initial point and terminal point we find the highest point of $\{S_i\}$ along the path (if there are more than one such points, select the first one) and substitute that node with the node directly below it in the $\{Z_i\}$ tree to obtain a new path and take its average. This is repeated $j(i-j)$ times to obtain the set $A(i,j) = \{A(i,j,k); 1 \leq k \leq j(i-j)+1\}$. The final path considered will be the one with $(i-j)$ down-movements followed by j up-movements. None of the previous paths are allowed to be below this path.

When working with the function v on the tree for Z and applying backward induction, linear interpolation must be used whenever the computed average is not in the representative set for that node. Consider a node (i,j) where $i < N$ and $j \leq i$ and select any $A(i,j,k) \in A(i,j)$. Denote $A^u := A(i,j,k) + z(i+1,j+1)^{-1}$ and $A^d := A(i,j,k) + z(i+1,j)^{-1}$. To compute $v(i, z(i,j), A(i,j,k))$, the values $v(i+1, z(i+1,j+1), A^u)$ and $v(i+1, z(i+1,j), A^d)$ are needed. Suppose that $A^u \notin A(i+1,j+1)$, then write A^u_l for the highest element of $A(i+1,j+1)$ lower than A^u. Similarly, A^u_h is the lowest element higher than A^u. We obtain $v(i+1, z(i+1,j+1), A^u)$ by applying linear interpolation to the corresponding contract values for A^u_l and A^u_h. Similar steps are followed for A^d. The scheme from Costabile et al. (2006) has the benefit that linear interpolation is not needed for many of the computations of v.

For the framework in Costabile et al. (2006), whether the algorithm begins with the path giving the highest average, selects paths in the described manner, and stops when the path giving the lowest average is obtained, or vice versa, the same set of averages are obtained. This symmetry is a result of the underlying asset changing by factors of u and d, where $ud = 1$. However, this symmetry does not hold in our model because the process Z changes by factors of $ue^{-\alpha}$ and $de^{-\alpha}$. For example, an up-move followed by a down-move does not return Z to its initial value. The downward trend of the Z-tree complicates the approximation algorithm. Consequently, the sets $A(i,j)$ will change depending on whether the lowest or highest path is initially considered.

3.5 Numerical Results

The computational applications of the binomial model for the GMWB rider are limited for two reasons. The binomial tree for the account value process is non-recombining and the riders have significantly longer durations in contrast to the usual European and American equity options which typically have durations not exceeding one year. The withdrawal rate g can be expected to range from 5% to 10% corresponding to maturities of 10 to 20 years. Clearly δt must be significantly smaller than one if the value processes in the binomial world are to provide an accurate approximation of the value processes in the Black-Scholes world established in Chapter 2.

For $g = 5\%$ the binomial tree will contain $2^{20} > 10^6$ nodes in the final period with just one timestep per year. The backward induction (tree) algorithm (Method A) requires too much memory for small values of δt. In the no-lapse model we saw that (3.11) allows a direct approach (Method B), where the algorithm loops through each path and minimal memory is required. We will see shortly that Method B is significantly slower than Method A. Although Method B enables using marginally smaller δt values, we quickly run into time constraints as the number of paths grows at $O(2^N)$.

Beginning with the no-lapse case, we provide numerical results comparing our model to previous results in the literature and find that even with large values for δt our simple model is a reasonable approximation of more complex models. Moreover, within a binomial world, it allows us to analyze the hedging results and the effect of the parameters on the losses when hedging is not implemented.

3.5.1 Bisection Algorithm

The bisection algorithm is used to solve for $\alpha^\star$. Define $f : \mathbb{R}_+ \mapsto \mathbb{R}_+$ by $f(\alpha) = V_0(P, \alpha, g) - P$. Then $f(\alpha^\star) = 0$ by (3.8). Choose an initial pair (α^u, α^l) such that

$f(\alpha^u) < 0$ and $f(\alpha^l) > 0$. Determine an acceptable error tolerance $\epsilon^\star$ and stop iterations when $|f(\alpha)| < \epsilon^\star$. Beginning with iteration $i = 1$, set $\alpha_i = \frac{\alpha^u+\alpha^l}{2}$. If $|f(\alpha_i)| < \epsilon^\star$ then α_i is an acceptable estimate for $\alpha^\star$. If $|f(\alpha_i)| \geq \epsilon^\star$ then set $\alpha^u = \alpha_i$ ($\alpha^l = \alpha_i$) if $f(\alpha_i) < 0$ ($f(\alpha_i) > 0$) and iterate again by incrementing i. We use $P = 100$ and $\epsilon^\star \leq 0.001$ in all our results achieving accuracy of 1×10^{-5} for a unit premium.

3.5.2 The Fair Rider Fee

Milevsky and Salisbury (2006) use numerical PDE techniques to solve for V_0, as defined by (2.13), and present the fair fees for various (g, σ) combinations. In Liu (2010), a discrete-time model is developed (see Section 1.3) and the contract values are estimated using Monte Carlo simulation with a geometric mean strike Asian call option as a control variate. Both papers assume S is lognormally distributed. In theory we expect convergence of results for both models and our binomial model. However Liu (2010) obtains results significantly lower than those of Milevsky and Salisbury (2006), from which it is concluded that Milevsky and Salisbury's results are on average 28% too high.

Table 3.1 provides a comparison between the two cited papers and the binomial model. In the discrete models δt = 1/timesteps. The parameters are: $P = 100$, $g = 10\%$, $r = 5\%$, $\sigma = 20\%$, $T = 1/g = 10$. For $\delta t = 1$, results from the binomial model and Liu (2010) are sufficiently close. We reach three timesteps per year under Method B, and observe that our model supports Liu's results, albeit to a limited degree.

For the same parameters Table 3.2 displays sample run-times (in seconds) to calculate V_0 for a single value of α. The differences may seem small for $n < 3$ and external factors also affect the run-times; however being that C++ is far more efficient to run for identical code we see that Method B is significantly slower to run. Under

	M&S (2006)	Liu (2010)			Binomial		
timesteps/year	continuous	1	12	4000	1	2	3
$\alpha^\star$(bps)	140	92.41	96.65	97.28	92.20	94.55	95.35

Table 3.1: Comparison of results for $\alpha^\star$: $g = 10\%,\ r = 5\%,\ \sigma = 20\%$

Timesteps	Method A (Trees, Matlab)	Method B (Loop, C++)
$n = 1$	7.7×10^{-4}	3×10^{-3}
$n = 2$	0.80	2.5
$n = 3$		3×10^{3}

Table 3.2: Computational time comparison (in seconds)

Method B with $n = 3$ and $\alpha = 95.35$bps, it is seen that $W_N = 0$ for all paths with less than 11 up-moves and therefore the bottom 10 nodes in the recombining tree for Z do not need to be evaluated. This does not prevent the run-time from rapidly growing.

While the binomial model is a valuable theoretical framework for viewing the GMWB rider, it is the Asian approximation method which reveals the practical value of such a model. Implementing the Asian approximation method, we attain results up to $n = 10$. Monthly timesteps should be attainable with more efficient programming and superior hardware. The results in Table 3.3 strongly imply convergence to the $\alpha^\star$ computed by Liu (2010).

Table 3.4 contains additional results for different g and σ values. The fair fees are increasing in both g and σ and is quite sensitive to the latter. Sensitivity results have been discussed at length in the literature (see Chen et al., 2008). The return

n	1	2	3	5	7	9	10
$\alpha^\star$(bps)	92.30	94.64	95.40	96.05	96.33	96.48	96.54
$V_0(\alpha = 97.3)$(\$)	99.767	99.880	99.917	99.945	99.958	99.965	99.967

Table 3.3: Asian approximation results

($\alpha^\star$, bps)		$\sigma = 20\%$			$\sigma = 30\%$		
$g\%$	T	MS[a]	L[b]	B[c]	MS[a]	L[b]	B[c]
5	20	37	28.5	27.1(1)	90	76.5	74.8(1)
6	16.67	54	40.6	38.7(1)	123	103.7	101.5(1)
7	14.29	73	53.8	51.3(1)	158	132.3	129.4(1)
8	12.5	94	n/a	64.6(1)	194	n/a	158.3(1)
9	11.11	117	n/a	80.1(2)	232	n/a	189.3(2)
10	10	140	96.7	94.6(2)	271	221.2	219.1(2)

[a] Milevsky and Salisbury (2006) [b] Liu (2010) with $n = 12$
[c] Binomial with n in parentheses

Table 3.4: Comparison with previous results for $\alpha^\star$, ($r = 5\%$)

of premium guaranteed by the GMWB does not include time value of money and as g increases, the maturity decreases and V_0 increases in value for any fixed α because of the interest rate effect. Consequently $\alpha^\star$ must increase. Our results consistently support Liu (2010) at the expense of Milevsky and Salisbury (2006).

In Figure 3.2, V_0 is plotted against α for different T values. The parameters are: $P = 100$, $r = 5\%$, $\sigma = 20\%$, $\delta t = 1$, and $g = \frac{1}{T}$. The fair fee is the point of intersection between the horizontal line $V_0 = 100$ and the curves. When the curves are plotted over the wider range [0,0.05] the linearity resemblance seen on $[0, 0.01]$ disappears and the curves have a more pronounced convex shape. As α increases, the likelihood of trigger rises but the decrease in the expected discounted terminal account value is less sensitive for sufficiently large α.

It is important to consider the sensitivity of V_0 to α in a neighbourhood around $\alpha^\star$, for a given set of parameters. Figure 3.2 reflects the changing sensitivity for different values of T. For the parameters in Table 3.1, the binomial method with $\delta t = 2$ gives $V_0(100, 140 \text{ bps}, 10\%) = 98.02$ and it can be deceptive to only look at $\alpha^\star$. The objective is to solve for the fair fee and in our pricing framework, charging a different fee leads to arbitrage no matter the size of $|\alpha - \alpha^\star|$. However in the real world with the constraints of Section 1.5, mispricing may not lead to arbitrage and

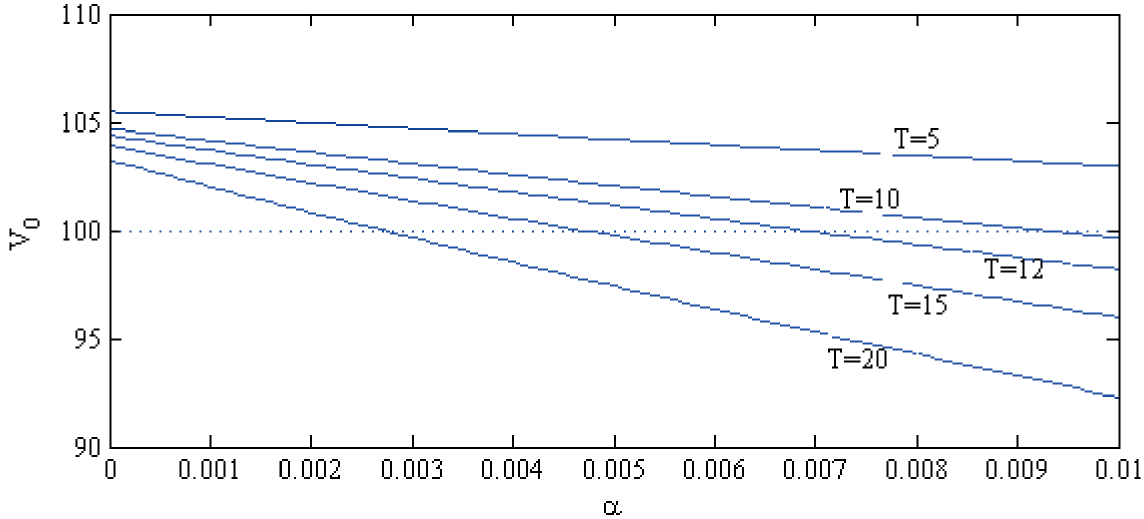


Figure 3.2: Plotting V_0 as a function of α for varying T. Parameters are: $r = 5\%$, $\sigma = 20\%$, and $g = 1/T$.

it is crucial to look at this sensitivity in addition to $\alpha^\star$.

3.5.3 Distribution of Trigger

Milevsky and Salisbury (2006) numerically solve the Kolmogorov backward equation for $\mathbb{P}(\tau \leq T)$ and provide results for different combinations of (μ, σ) with the parameters $g = 7\%$ and $\alpha = 40$bps. In Table 3.5 these results are compared with those obtained from our binomial model. To avoid fractional years, we set $T = 14$ and $g = 7.14\%$.

In Milevsky and Salisbury (2006), S_t is modeled by geometric Brownian motion

$$dS_t = \mu S_t dt + \sigma S_t dB'_t,$$

where B'_t is $\mathbb{P}$-Brownian motion. Referring to the parametrization from Proposition 3.2 we have

$$r^s_T := \ln\left(\frac{S_T}{S_0}\right) = \left(\mu - \frac{1}{2}\sigma^2\right) T + \sigma B'_T,$$

therefore $\mathbb{E}_{\mathbb{P}}[r^s_T] = (\mu - \frac{1}{2}\sigma^2)T$ and $Var_{\mathbb{P}}[r^s_T] = \sigma^2 T$. We set

$$u = e^{\sigma\sqrt{\delta t}},$$
$$d = e^{-\sigma\sqrt{\delta t}},$$
$$\tilde{p} = \frac{1}{2} + \frac{1}{2}\left(\mu - \frac{1}{2}\sigma^2\right)\frac{1}{\sigma}\sqrt{\delta t}.$$

Note that $\tilde{p} < 1$ holds only if $\mu < \frac{1}{2}\sigma^2 + \sigma\frac{1}{\sqrt{\delta t}}$. For $\delta t = 1$ this condition is violated for $\sigma = 10\%$ and $\mu = 12\%$.

In general, the probability mass function of τ w.r.t. $\mathbb{P}$ can be calculated in the binomial model with (3.13), where

$$\mathbb{P}(\tau = i) = H(0, 0, i, P)$$

for $i \in \{1, 2, \ldots, N, \infty\}$. Of course, p must be replaced with $\tilde{p}$.

The level of accuracy in Table 3.5 varies by parameters. For fixed μ, as σ increases the binomial model results shift from underestimating the continuous model to overestimating it.

Remark 3.19. Applying (3.13) with two timesteps a year, 2^{28} paths need to be evaluated and we run into capacity issues in both Matlab and C++. For $\delta t = 0.5$, we use the approach in (3.11) except that rather than working with $e^{-rT}W_T$, we use the indicator function $\mathbf{1}_{\{W_T=0\}}$ remembering to take account of the probabilities for the lower nodes with more than a down movements.

Remark 3.20. We stated the exact distribution function for τ in Remark 2.4 subject to the constraint $(\mu - \alpha) < \frac{3}{2}\sigma^2$. In Table 3.5 this constraint only holds for

$$(\mu, \sigma) \in \{(0.04, 0.18), (0.04, 0.25), (0.06, 0.25), (0.08, 0.25)\}.$$

However, by Remark 2.4 upper bounds can be obtained through (A.1).

	$\sigma = 10\%$			$\sigma = 15\%$			$\sigma = 18\%$			$\sigma = 25\%$		
	M&S	Binomial		M&S	Binomial		M&S	Binomial		M&S	Binomial	
		$\delta t = 0.5$	$\delta t = 1$		$\delta t = 0.5$	$\delta t = 1$		$\delta t = 0.5$	$\delta t = 1$		$\delta t = 0.5$	$\delta t = 1$
$\mu = 4\%$	19.0%	16.0%	15.2%	31.4%	31.1%	30.9%	37.8%	38.2%	38.2%	49.9%	50.8%	50.6%
$\mu = 6\%$	7.0%	4.5%	3.6%	18.5%	17.8%	16.9%	25.5%	25.3%	25.0%	39.6%	40.5%	40.2%
$\mu = 8\%$	1.7%	0.7%	0.3%	9.3%	8.2%	7.4%	15.5%	15.0%	14.5%	30.5%	30.8%	30.4%
$\mu = 10\%$	0.3%	0.0%	0.0%	4.1%	3.1%	2.3%	8.6%	7.8%	7.1%	22.2%	22.2%	21.7%
$\mu = 12\%$	0.04%	0.0%	-	1.6%	0.9%	0.4%	4.4%	3.5%	2.7%	15.5%	15.2%	14.4%

Table 3.5: $\mathbb{P}(\tau < \infty)$: comparing binomial model to continuous time model from Milevsky and Salisbury (2006)

3.5.4 Comparison of Hedging and No Hedging

We investigate the impact of volatility on the fees, triggers and losses. The parameters are: $g = 10\%$, $T = 10$, $P = 100$, and $\delta t = 1$. The risk free rate r is 5% and the drift term μ of the underlying asset is 7.5%. We consider $\sigma = 15\%$ and $\sigma = 30\%$. The respective fair fees $\alpha^\star$ are 41.8bps and 216.7bps. The probability mass function for τ under the physical measure is displayed in Figure 3.3. Recall that $\tau = \infty$ when $W_T > 0$. The two σ values were selected to magnify the interaction between volatility, the trigger time distribution and consequently the rider payouts. Higher volatility implies more adverse market returns and a greater likelihood of early trigger. An additional effect on trigger comes from the rider fee. The fee rate is very sensitive to volatility and the fees drag down the account value further, resulting in more frequent early trigger times.

We consider the strategies of no hedging and dynamic delta hedging prescribed in Subsection 3.2.4. Define $\Pi := e^{-\bar{r}N} X_N$ to be the discounted profit. When Δ follows the prescribed portfolio process (3.17) we obtain the hedging profit, Π^H. If $\Delta \equiv 0$ we obtain the profit under no hedging, Π^{NH}. The superscripts are omitted when it is clear which profits we are analyzing. Figure 3.4 plots both $-\Pi^H$ and $-\Pi^{NH}$ against τ_0 for the complete set of outcomes ($2^{10} = 1024$ paths). The values are per \$100 initial premium.

The dynamic delta hedging strategy results in no losses. Without hedging, the range of potential losses by each random trigger time has a decreasing trend because a later trigger time implies additional periods of fee revenue and fewer periods of any rider guarantee payout. The effect of the volatility σ is particularly visible for those pathwise outcomes where $\tau = \infty$. When $\sigma = 15\%$ there is a 87% probability of a positive terminal account value but the gains are small. On the other hand, there is only a 50% probability that $\tau = \infty$ when $\sigma = 30\%$ but the potential profits are large due to the high fees. Figure 3.5 shows the cumulative distribution function of

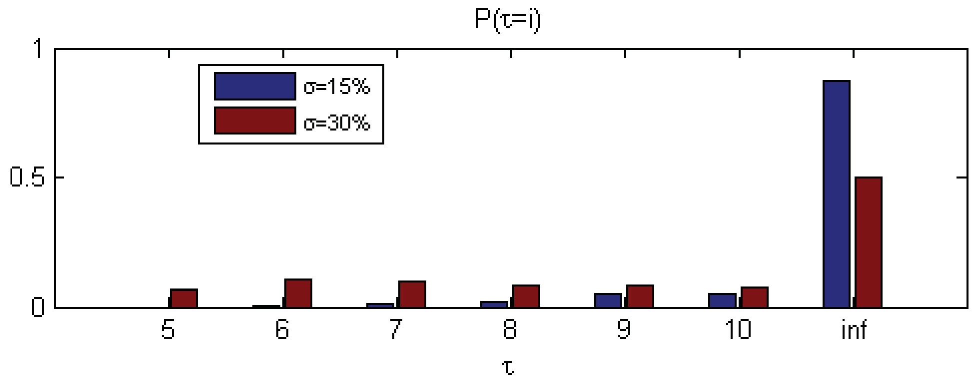


Figure 3.3: Probability mass function - $\mathbb{P}$

Values per \$100	$\sigma = 15\%$	$\sigma = 30\%$
$E_{\mathbb{P}}(\Pi^{NH})$	1.84	4.19
$\text{SD}_{\mathbb{P}}(\Pi^{NH})$	4.28	21.34
$TVaR_{0.10}(\Pi^{NH})$	9.30	32.60

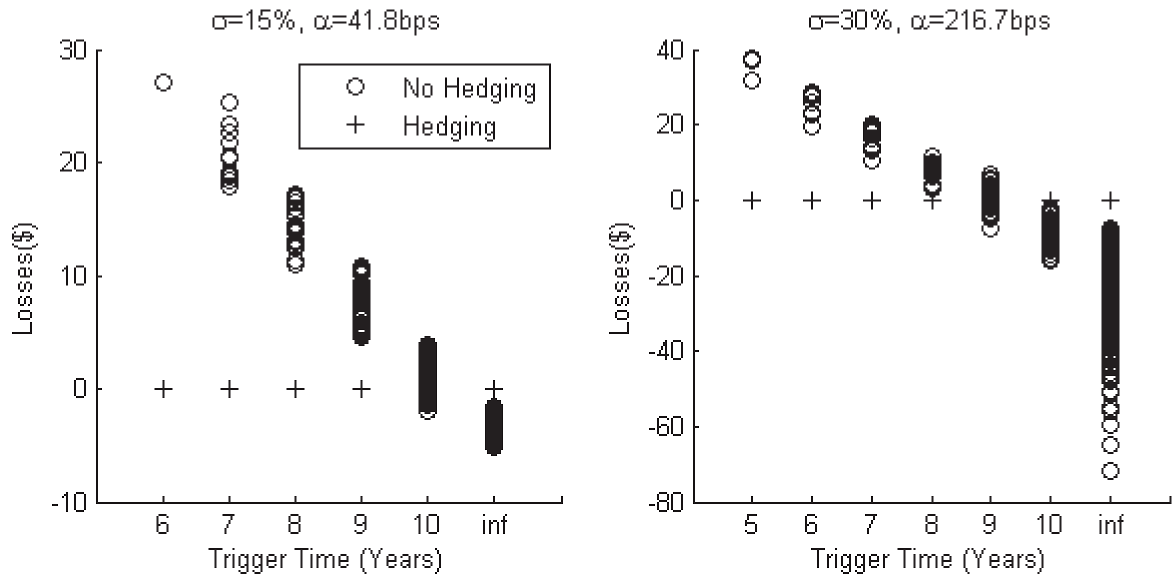
Table 3.6: Profit metrics for no hedging (no lapses)

the profits when there is no hedging.

We present several metrics for Π^{NH} under $\mathbb{P}$. The standard deviation is denoted $SD(\Pi)$. The tail value at risk is $TVaR_\gamma(\Pi) := E_{\mathbb{P}}[-\Pi | \Pi \leq -VaR_\gamma(\Pi)]$ where $VaR_\gamma(\Pi) = -\inf\{x : \mathbb{P}(\Pi \leq x) > \gamma\}$. Table 3.6 shows the values for this sensitivity analysis of σ. It only amplifies the effect of σ on the insurer's risk and highlights the importance of a thorough hedging scheme.

3.5.4.1 Hedging in a Continuous Model

In the binomial model a perfect hedge is attainable. Suppose instead the underlying asset follows the geometric Brownian motion process given by (2.1). A perfect hedge in this case entails continuously rebalancing the hedging positions by taking a position at any time t of $\frac{W}{S}\frac{\partial U}{\partial W}$ units of S (see Chen et al., 2008). In practice, the positions will be rebalanced only a finite number of times each year which introduces hedging errors. We model the fees and withdrawals to occur only at year-end in order to contrast with the previous result in the binomial model for $\delta t = 1$. This differs from the continuous



(a) (b)

Figure 3.4: Hedging and no-hedging losses, with $r = 5\%$ and $g = 10\%$

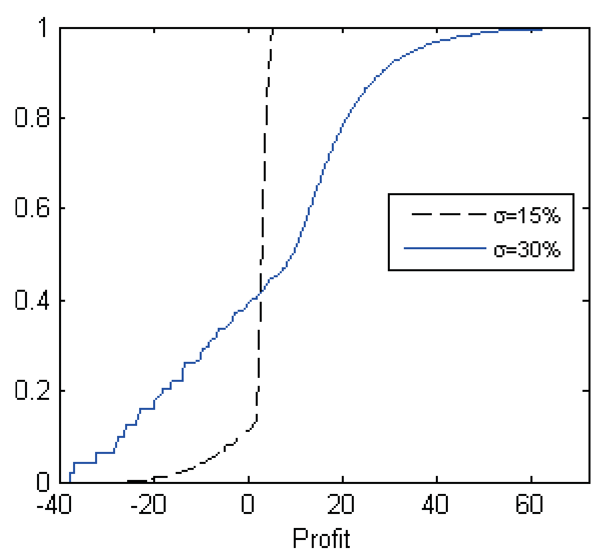


Figure 3.5: CDF of Π^{NH} w.r.t. $\mathbb{P}$

model of Chapter 2 where fees and withdrawals are deducted continuously. For all $k \in I = \{1, \ldots, T\}$ and for all $s \in (k-1, k]$ we have

$$W_s = W_{k-1^+} e^{(r-\frac{1}{2}\sigma^2)(s-(k-1))+\sigma(B_s - B_{k-1})},$$
$$W_{k^+} = \max(W_k e^{-\alpha} - G, 0),$$

where $G = Pg$. The risk-neutral value of the annuity plus GMWB rider is

$$V_0(P, \alpha, g) = \sum_{i=1}^{T} Ge^{-ri} + e^{-rT} E_{\mathbb{Q}}(W_{T^+}).$$

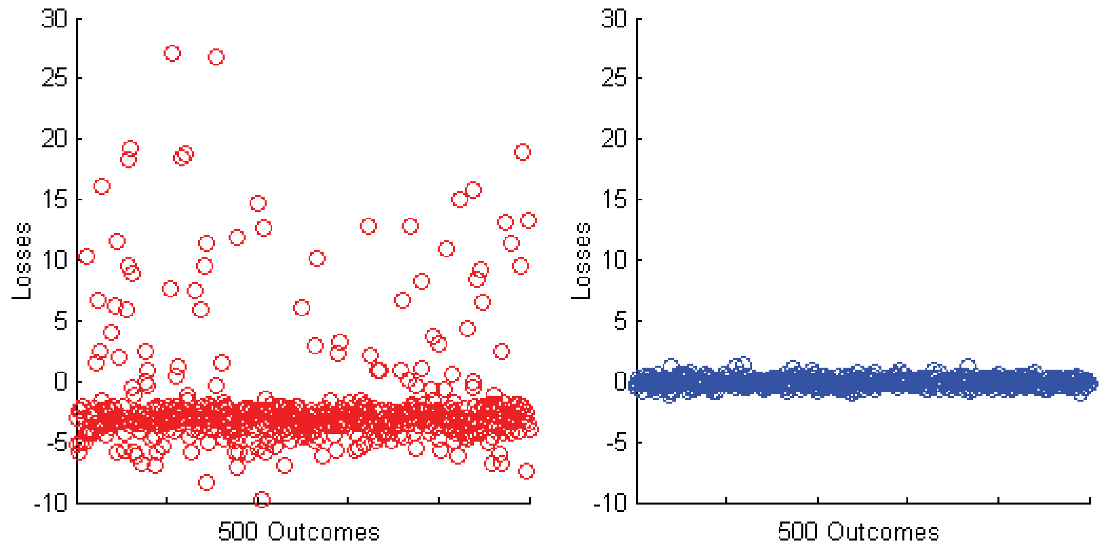
The parameters used are $P = 100$, $g = 10\%$, $r = 5\%$, $\mu = 7.5\%$, $\sigma = 15\%$, and $T = 10$. We used Monte Carlo simulation to obtain $\alpha^\star \approx 45$bps (50,000 paths were simulated). By the Markov property of $\{W_t\}$ the value of the embedded rider is $U_k = u(k, W_{k^+})$ for $k \in \{0, 1, \ldots, T\}$, where

$$u(k, x) = E_{\mathbb{Q}}\left[\sum_{i=\tau^x \wedge 1}^{T-k} (G - W_i^x e^{-\alpha})^+ e^{-ri} 1_{\{\tau^x \leq T-k\}} - \sum_{i=1}^{\tau^x \wedge (T-k)} e^{-ri} W_i^x (1 - e^{-\alpha})\right]$$

and $\tau^x = \inf\{s \geq 0;\ W_s^x = 0\}$. For any non-integer $t \in [0, T]$ we have $U_t = u(t, W_t)$, where

$$u(t, x) = e^{-r(\lceil t \rceil - t)} E_{\mathbb{Q}}[u(\lceil t \rceil, (W^x_{\lceil t \rceil - t} e^{-\alpha} - G)^+) + (G - W^x_{\lceil t \rceil - t} e^{-\alpha})^+ - W^x_{\lceil t \rceil - t}(1 - e^{-\alpha})].$$

We analyzed the effectiveness of a dynamic hedging strategy with weekly rebalancing for 500 path outcomes generated under $\mathbb{P}$. For $t \in \{0, \frac{1}{52}, \frac{2}{52}, \ldots, \frac{519}{52}, 10\}$ and $w \in \mathbb{R}_+$, Monte Carlo simulations (using 1000 paths) yielded $U_t(w-1)$ and $U_t(w+1)$. We approximated $\frac{\partial U}{\partial W}$ with $\Delta_t(W_t) = \frac{U_t(W_t+1) - U_t(W_t-1)}{2}$ where the same set of generated paths was used to obtain both values in the numerator. Using the same paths and taking the central difference has been shown to reduce variability of results (Glasserman, 2004). Figure 3.6 displays the discounted losses for no hedging and for weekly hedging for each generated path. Based on the simulations, $\mathbb{P}(\tau = \infty) = 84.4\%$. As supported by Table 3.7, the weekly hedging considerably mitigates the equity risk.



(a) No hedging (b) Weekly hedging

Figure 3.6: Continuous model with g =10%, r =5%, μ = 7.5% , σ=15%, α=45bps

Values per \$100	No Hedging	Hedging (Weekly)
$E_{\mathbb{P}}[\Pi]$	1.86	0.07
$\mathrm{SD}_{\mathbb{P}}[\Pi]$	4.63	0.36
$TVaR_{0.10}(\Pi)$	10.15	0.61

Table 3.7: Profit metrics for continuous model with weekly hedging and no hedging

There are negative hedging errors in contrast to the case when the underlying model is binomial.

3.5.5 The Fair Rider Fee with Surrenders

We compare our results for $\alpha^\star$ with those in the literature. For the parameter set of $g = 7\%$, $r = 5\%$, and $k_i = 1\%$ for all i, Table 3.8 compares the binomial model with $\delta t = 1$ to Milevsky and Salisbury (2006). Although the results are proportionally closer, as compared to Table 3.1, it is inconclusive if the differences are mostly due to $\delta t = 1$ or if the results presented by Milevsky and Salisbury (2006) in the lapse case suffer from the same inaccuracies as in the no-lapse case.

We apply the Asian approximation method with the parameters $g = 10\%$, $r =$

$\sigma(\%)$	15	18	20	25	30
Milevsky and Salisbury (2006)	97	136	160	320	565
Binomial ($\delta t = 1$)	33	89	138	283	455

Table 3.8: Comparison of $\alpha^\star$ to previous results; with $g = 7\%$, $r = 5\%$, and $k = 1\%$.

n	$\alpha^\star$(bps)	$V_0(\alpha{=}146.4)$($)	$\alpha^\star$(actual)
1	131.00	99.689	130.54
2	141.98	99.933	141.75
3	143.37	99.949	
4	146.04	99.994	
5	146.40	100	
6	146.70	100.005	

Table 3.9: Asian approximation results - lapses

5%, $\sigma = 20\%$, and $k = 3\%$ in Table 3.9. The convergence is slower than in the no-lapse case, but that is a result of the early surrender decisions which are being approximated. This is consistent with the findings of Costabile et al. (2006). The rightmost column shows $\alpha^\star$ under the original binomial model. The increase in $\alpha^\star$ when n is increased from one to two suggests that a sizeable portion of the differences in Table 3.8 can be attributed to the low value of n in the binomial model.

We set r equal to the instantaneous risk-free rate long term mean and σ equal to the variance long term mean used in the stochastic interest rate and volatility processes in Bacinello et al. (2011). We found that comparing V_0 for varying α, in the no lapse case the binomial model provides close estimates even for $\delta t = 0.5$. In Table 3.10 we list the difference in the contract value between the two methods for varying α and $P = 100$, $g = 10\%$, $r = 3\%$, $\sigma = 20\%$, and $k = 3\%$. The models have fundamental differences and we do not expect to attain exact results in the limit.

Sensitivity results for g, r, and σ are shown in Table 3.11. The baseline case is set to $g = 10\%$, $r = 5\%$, $\sigma = 20\%$, and a CDSC of $k = 3\%$. The fair fee $\alpha^\star$ is increasing with g and σ but decreasing with r. It is most sensitive to r. This is due to the long duration of the contract. Clearly a stochastic interest rate approach is well-justified.

$\alpha(\%)$	1	2	3	4	5
$V_0^B(\alpha) - V_0^{BMOP}(\alpha)^{a,b}$: (no lapse)	-0.186	-0.113	-0.035	0.05	0.096
$V_0^B(\alpha) - V_0^{BMOP}(\alpha)$: (lapse)	0.153	0.546	0.75	0.78	1.04

[a] V_0^B refers to the binomial method, with $\delta t = 0.5$.
[b] V_0^{BMOP} refers to Bacinello et al. (2011).

Table 3.10: Comparison of V_0 with previous results: $g = 10\%$, $P = 100$, $r = 3\%$, $\sigma = 20\%$, and $k = 3\%$.

$g\%$	$\alpha^\star$ (bps)	$V_0(\alpha_1)$ (\$)	$\sigma\%$	$\alpha^\star$	$V_0(\alpha_1)$	$r\%$	$\alpha^\star$	$V_0(\alpha_1)$
5	30	97.21	10	10	97	1	1199	108.21
6	47	97.87	15	44	97.84	2	673	105.54
7	68	98.44	18	87	99.08	3	397	103.29
8	90	98.95	20	142	100	4	244	101.43
9	110	99.38	25	318	102.46	5	142	100
10	142	100	30	562	105.12	6	77	98.87

[a] Baseline case is $g = 10\%$, $r = 5\%$, $\sigma = 20\%$, $k = 3\%$, $\alpha_1 = 142$bps.
[b] For the first column, $\delta t = 1$ for $g \leq 9\%$. All other values use $\delta t = 2$.

Table 3.11: Sensitivity results for $\alpha^\star$

Under the parameters of $g = 10\%$, $r = 5\%$, $\sigma = 25\%$, and $\delta t = 1$, the impact of the CDSC schedule on $\alpha^\star$ is shown in Table 3.12. Allowing surrenders with no penalties, the fair fee will be exorbitant to compensate for this option. As the penalties increase, the fee approaches the corresponding fee in the no-lapse model. For sufficiently high penalties, the option to surrender yields no marginal value.

3.5.6 Hedging and No Hedging with Surrenders

We consider the parameters: $P = 100$, $g = 10\%$, $r = 5\%$, $\sigma = 25\%$, and $\delta t = 1$. The drift of S is $\mu = 7.5\%$. The surrender charge schedule applied is $k_i = \max(.09 - .01i, 0)$ for $i = 1 \ldots 10$. Figure 3.7 plots the aggregate losses, discounted to time zero, for the set of all outcomes for both the no-surrender model and the model with early surrenders. The respective fair fees are charged. In Figure 3.7b the no-hedging results are denoted by L and T: the former are outcomes where it is optimal to lapse while

Description of Schedule	$\alpha^\star$(bps)
No-Lapse Model	**152**
$k_i = 0$ for $i = 1, \ldots, 9$	491
$k_i = 1\%$ for $i = 1, \ldots, 9$	430
$k_i = 3\%$ for $i = 1, \ldots, 9$	309
$k_i = 5\%$ for $i = 1, \ldots, 9$	217
$k_i = 7\%$ for $i = 1, \ldots, 9$	169
$k_i = 8\%$ for $i = 1, \ldots, 9$	155
$k_i \geq 8.38\%$ for $i = 1, \ldots, 9$	**152**
$k_i = (10 - i)\%$, for $i = 1, \ldots, 9$	171
$k_i = (9 - i)\%$, for $i = 1, \ldots, 9$	188

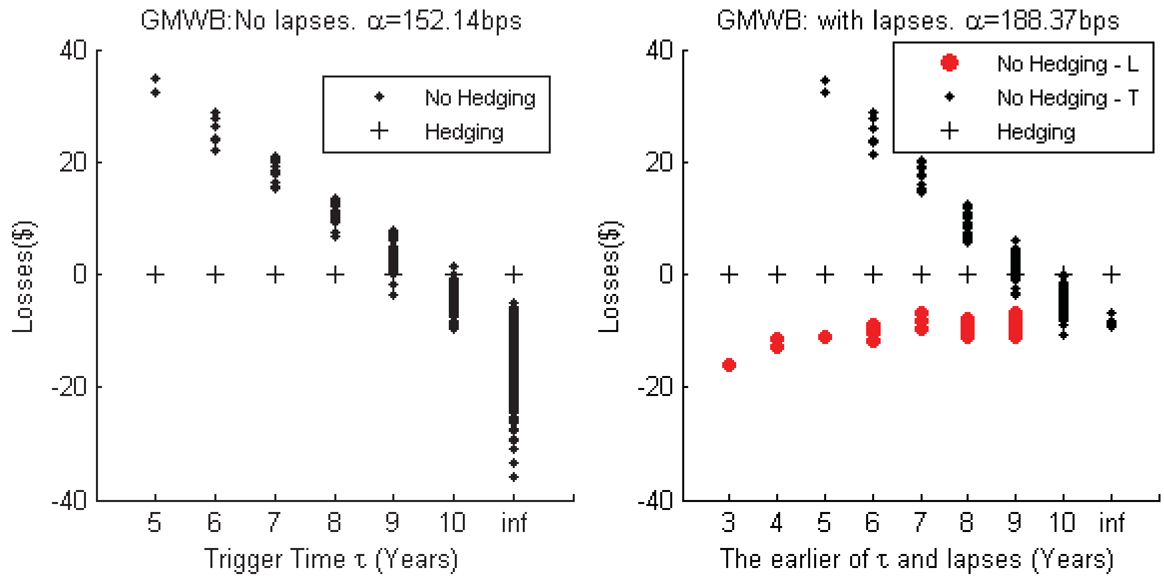
Table 3.12: Impact of k on $\alpha^\star$

the latter are those for which no lapse occurs.

Table 3.13 shows the $\mathbb{P}-$distribution of trigger times and surrender times, where $\eta^\star$ denotes an optimal early surrender. Note that $\mathbb{P}(\tau = \infty) \approx 60\%$ when surrenders are not allowed, but this reduces to $\mathbb{P}(\tau = \infty) \approx 0.65\%$ when surrenders are permitted. Allowing lapses causes a shift as it becomes preferable in many outcomes when the market is doing well for the policyholder to lapse rather than face the likelihood of the rider maturing without being triggered.

For the outcomes where it is optimal to lapse, the profits to the insurer are decreasing for years 3-7. This is due to the design of k. The higher surrender charge in earlier years outweighs the additional fees received when lapses occur later.

We look at L_0 in Figure 3.8. When α is small, there is little incentive to surrender early and $L_0 \approx 0$. For greater values of α, there is incentive to surrender and avoid paying future fees. This is reflected in the growth of L_0.



(a) (b) SC begins at 8% and decreases 1% per annum

Figure 3.7: Hedging and no hedging, with and without lapses: $g = 10\%$, $r = 5\%$, $\sigma = 25\%$.

	No Lapses	Model with Lapses	
i	$\mathbb{P}(\tau = i)$	$\mathbb{P}(\tau = i)$	$\mathbb{P}(\eta^\star = i)$
3	0	0%	20.28%
4	0	0%	16.73%
5	2.90%	2.90%	4.91%
6	5.80%	5.80%	8.11%
7	7.83%	7.83%	3.57%
8	6.29%	9.08%	4.42%
9	9.48%	7.37%	2.37%
10	7.23%	5.98%	0
∞	60.47%	.65%	0
Sum	1	39.61%	60.39%

Table 3.13: Probability distribution of τ and lapses for Figure 3.7

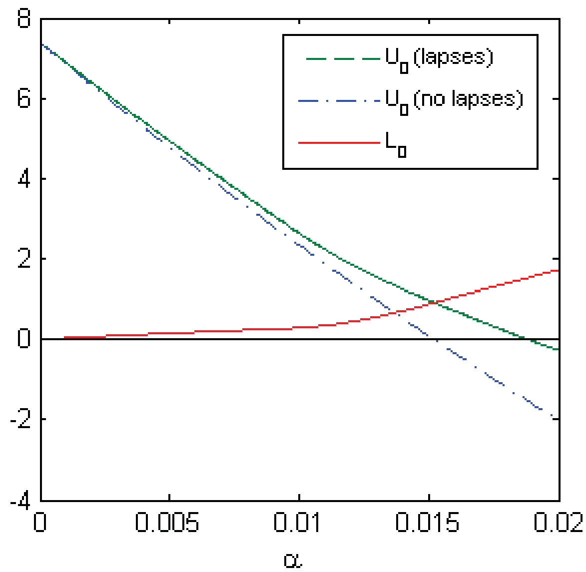


Figure 3.8: Value of L_0: $g = 10\%$, $r = 5\%$, $\sigma = 25\%$, $\delta t = 1$, and a declining SC schedule

Chapter 4

Extending the Model: Including Mortality

The simplification of disregarding mortality was used in several papers for GMWBs including Milevsky and Salisbury (2006) and Dai et al. (2008). Mortality factors do need to be considered in practice. Depending on the goal of the analysis, the level of preciseness attained by including mortality may not justify the added complexity and dimensionality of the model. In particular, in the papers mentioned the focus was on studying the optimal policyholder behaviour strategy and including mortality only detracts from the presentation of the results.

Mortality risk is typically assumed to be independent of financial risk. Further, under the assumption of independent lives and deterministic forces of mortality (hazard rates) a simple application of the strong law of large numbers justifies the claim that mortality risk is diversifiable. By issuing a sufficiently large portfolio of homogeneous policies the insurer can completely account for the mortality risk by taking the expected value of claim payments under the appropriate mortality probability distribution (Boyle and Schwartz, 1977). Therefore under these assumptions mortality risk is not priced by capital markets in an economic equilibrium (no-arbitrage) approach and there is no difference between the physical and risk-neutral measures (Milevsky et al., 2006). In a stochastic mortality framework the non-diversifiable component of mortality risk must be priced into the contract.

Milevsky et al. (2006) list capacity constraints in immediate annuity markets as one of several industry trends which justify charging for mortality risk. We remark that in variable annuity markets, both finite demand and regulatory limits on capital at risk lend support to modeling capacity constraints in order to determine whether there is a non-negligible impact.

The effect of mortality for GMWBs clearly depends on the death benefits (DBs). When benefit payments are similar for both death and survival, there is minimal impact. Indeed, Bacinello et al. (2011) found that guaranteed minimum death benefit (GMDB) riders add little value to the contract in the presence of other living benefit riders and a relatively short maturity.

There are several possibilities for the contract specifications in the event of death. The trivial case is the return of the current account value (without any surrender charges deducted) while the default option is often a return of premium (ROP) clause with the payoff being the maximum of the current account value and the total premiums reduced by withdrawals. More complex options may have the death benefit increase over time through ratchets or rollups. GMWB riders usually include the ROP death benefit but allow the policyholder the option of adding richer death benefit riders.

We extend the model from Chapter 3 to include mortality under the independence of lives assumption and deterministic forces of mortality. It is straightforward to obtain the price processes V and U, which for each insured are dependent on the survival status. The rider fee is obtained assuming diversifiable mortality risk, as is the hedging portfolio; however, we consider a numerical simulation to emphasize that under capacity constraints and finite number of policies there is mortality risk and the product is not fully hedged.

4.1 Mortality Framework

In this section we establish a mortality framework. The classical actuarial theory and notation used follows that of Bowers et al. (1997). In addition, the measure-theoretic aspects and inclusion of counting processes follows closely the frameworks of Møller (1998) and Wang (2008).

Assumption 4.1. *Homogeneous policies are issued to a pool of l_x policyholders, each of age x. Measured from issue date, the random times of death, denoted by $\{T_j^x; j = 1, \ldots, l_x\}$ where T_j^x is the time of death for policyholder j, are absolutely continuous, independent and identically distributed, and lie on a probability space $(\Omega^M, \mathcal{F}^M, \mathbb{P}^M)$.*

Consider a representative random variable T^x where $T^x \stackrel{d}{=} T_j^x$. The support of T^x is $[0, T^\star)$ where $T^\star \leq \infty$ is the maximum remaining lifetime for a person age x. Corresponding to the binomial model with $\delta t = 1/n$ and $n \in \mathbb{N}_+$, let K^x denote the period in which death occurs. Then $K^x = \lceil T^x/\delta t \rceil$. In other words, $K^x = i$ is equivalent to $(i-1)\delta t < T^x \leq i\delta t$. For $j = 1, \ldots, l_x$, define the counting processes

$$D^{x,j} = \{D_i^{x,j} := \mathbf{1}_{\{K_j^x \leq i\}}; i = 1, \ldots, N\}.$$

We work with the filtration generated by $\{D^{x,j}\}_{1 \leq j \leq l_x}$. The filtration is $\mathbb{F}^{M,x} := \{\mathcal{F}_i^{M,\{x,l_x\}}\}_{1 \leq i \leq N}$ where $\mathcal{F}_i^{M,\{x,l_x\}} := \mathcal{F}_i^{M,x,1} \vee \cdots \vee \mathcal{F}_i^{M,x,l_x}$ and $\mathcal{F}_i^{M,x,j} = \sigma(D_l^{x,j}; l = 1, \ldots, i)$. We work with the resulting filtered probability space $(\Omega^M, \mathcal{F}_N^{M,\{x,l_x\}}, \mathbb{F}^{M,x}, \mathbb{P}^M)$.

Remark 4.2. The notation $\mathcal{G} \vee \mathcal{H}$, where $\mathcal{G}$ and $\mathcal{H}$ are σ-algebras, means the σ-algebra generated by $\mathcal{G} \cup \mathcal{H}$.

We define the process which produces 1 while the insured j is still alive by $A_i^{x,j} := 1 - D_i^{x,j}$ for $i \in \mathcal{I}_N$.

By Assumption 4.1, T^x has a density function f_{T^x}. Its cdf is denoted $F_{T^x}(t) := \mathbb{P}(T^x \leq t)$. The deterministic force of mortality, $\mu_x(t)$, is defined as the conditional probability density function of T^x at time t, given survival to that time. Then

$$\mu_x(t) := \frac{f_{T^x}(t)}{1 - F_{T^x}(t)}. \tag{4.1}$$

We introduce some additional actuarial notation:

$$\begin{aligned} {}_jp_{x+i} &:= \mathbb{P}(K^x > i + j | K^x > i) = \mathbb{P}(T^x > (i+j)\delta t | T^x > i\delta t), \\ {}_{j|l}q_{x+i} &:= \mathbb{P}(i + j < K^x \leq i + j + l | K^x > i), \end{aligned}$$

and we write p_{x+i} for ${}_1p_{x+i}$, ${}_jq_{x+i}$ for ${}_{0|j}q_{x+i}$, and q_{x+i} for ${}_1q_{x+i}$. It follows that ${}_iq_x = F_{T^x}(i\delta t)$, ${}_ip_x = 1 - {}_iq_x$, and ${}_{j|l}q_{x+i} = {}_{j+l}q_{x+i} - {}_jq_{x+i}$. From (4.1) we have $f_{T^x}(i\delta t) = \mu_x(i\delta t)_{i\delta t}p_x$ and ${}_jp_{x+i} = e^{-\int_0^{j\delta t} \mu_{x+i\delta t}(u)du}$ (see Bowers et al. (1997) for details). Note that F_{T^x}, f_{T^x}, and μ_x are defined on the reals, while ${}_jp_{x+i}$ and ${}_{j|l}q_{x+i}$ are defined on the integers.

Bowers et al. (1997) provide several analytical laws of mortality.

Definition 4.3. Under the Makeham law

$$\mu_x(t) := A + Bc^{x+t}$$

where $B > 0$, $A \geq -B$, $c > 1$ and $x + t \geq 0$.

As a result, under the Makeham law:

$${}_ip_x = \exp\left(-i\delta tA - \frac{B}{\ln(c)}(c^{x+i\delta t} - c^x)\right).$$

Example 4.4. The parameters used to develop the illustrative life table under the Makeham law in Bowers et al. (1997) are: $A = 0.7 \times 10^{-3}$, $B = 0.05 \times 10^{-3}$ and $c = 10^{0.04}$. Figure 4.1 plots both $f_{T^x}(t)$ and $\mathbb{P}(T^x > t)$ for $x = 60$ and $t \in [0, 50]$.

We state one additional useful result from Wang (2008). For $i \leq j$,

$$\mathbb{P}(T^x > j\delta t | \mathcal{F}_i^{M,x}) = (1 - D_i^x)_{j-i}p_{x+i}$$

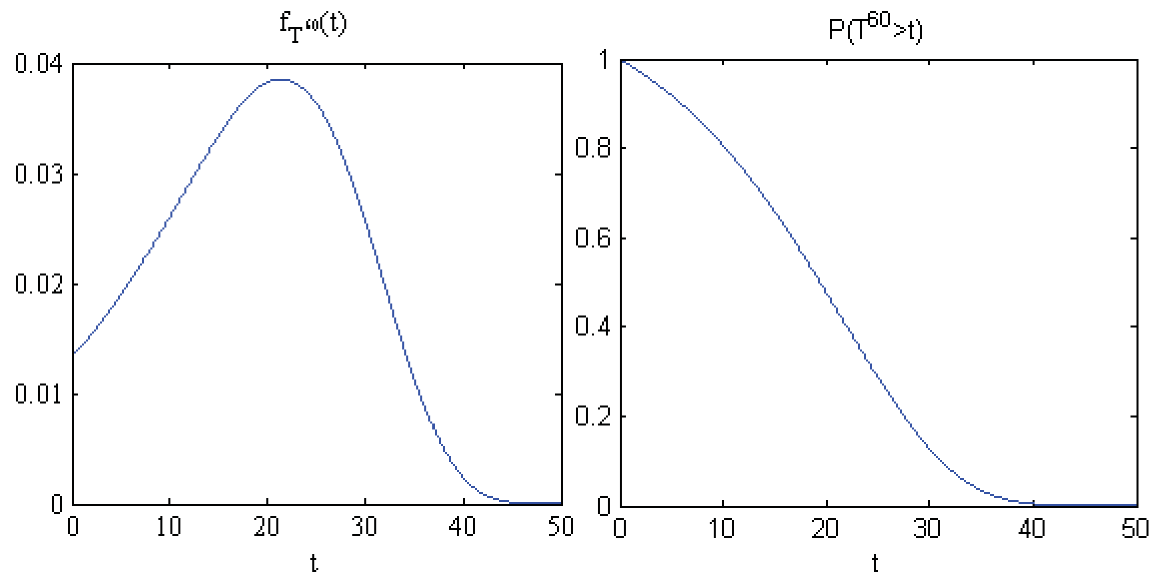


Figure 4.1: Using the Makeham law, with $A = 0.7 \times 10^{-3},\ B = 0.05 \times 10^{-3}$ and $c = 10^{0.04}$

and

$$\mathbb{P}(i\delta t < T^x \leq j\delta t | \mathcal{F}_i^{M,x}) = (1 - D_i^x)_{j-i}q_{x+i}.$$

4.2 Death Benefit Design

We consider both the ratchet DB and the return of premium DB. The ratchet DB has the feature that on each ratchet date, the death benefit base will increase to the current account value, provided the account value is higher. Let

$$0 \leq t_1 < t_2 < \cdots < t_m \leq T$$

represent the set of ratchet dates prior to maturity. Then the rescaled set, in terms of binomial time periods, is

$$I = \left\{ \frac{t_1}{\delta t}, \frac{t_2}{\delta t}, \ldots, \frac{t_m}{\delta t} \right\} \subset \mathcal{I}_N.$$

The GMWB and GMDB are treated as one rider with the aim of solving for the fair fee $\alpha^\star$ as before. Alternatively, one could separate the two and specify the GMDB rider fee exogenously. Let DB_i be the death benefit guarantee base at timepoint i, with $DB_0 = P$. Then $DB_i = db(i, W_{i^-}, DB_{i-1})$, where $db : \mathcal{I}_N \times \mathbb{R}_+ \times \mathbb{R}_+ \mapsto \mathbb{R}_+$ is defined as

$$\begin{cases} db(0, x, y) = x, \\ db(i, x, y) = \max\left(w(x)\mathbf{1}_{\{i \in I\}}, \frac{w(x)}{xe^{-\alpha}} y\right). \end{cases} \tag{4.2}$$

If $I = \emptyset$, then the ratchet DB reduces to a simple return of premium DB.

Note that $DB_i = 0$ for $i \geq \tau$. However we assume that conditional on survival to the trigger date, the guaranteed payments are paid regardless of life status; that is, the present value of the remaining payments is paid upon death if trigger has previously occurred. The death benefit of $\max(DB_i, W_{i+1^-})$ is paid at time $(i+1)\delta t$, if death occurs during the $(i+1)$th period but prior to trigger time. In the limit as $\delta t \to 0$ this corresponds to the death benefit being paid at the instantaneous time of death.

The death benefit base in (4.2) is reduced by withdrawals in a pro-rata manner, meaning it is reduced by the same proportion as the account value. Another method is called dollar-for-dollar withdrawal adjusted. Assume a policyholder holds a deep in the money GMDB, with $DB_i \gg W_i$ (where $x \gg y$ means y is much less than x). By withdrawing $0.9W_i$ and ignoring surrender charges, under the dollar-for-dollar reduction method the policyholder holds a GMDB with only 10% of the previous account value but a death benefit base of $DB_i - 0.9W_i \gg 0$. Under the pro-rata method, the new death benefit base is $0.1DB_i \ll DB_i - 0.9W_i$.

4.3 Pricing and Hedging

A key underlying assumption for the remainder of our work is stated.

Assumption 4.5. *There is independence between biometric and financial risks. Let* $(\Omega^S, \mathcal{F}_N^S, \mathbb{F}^S, \mathbb{Q}^S)$ *and* $(\Omega^M, \mathcal{F}_N^{M,\{x,l_x\}}, \mathbb{F}^{M,x}, \mathbb{P}^M)$ *be the filtered probability spaces con-*

structed in Section 3.1 and Section 4.1 respectively. We work with the product space $(\Omega, \mathcal{F}_N, \mathbb{F}, \mathbb{Q})$ *where* $\Omega := \Omega^M \times \Omega^S$, $\mathbb{F} := \{\mathcal{F}_i\}_{i=0}^N$, $\mathcal{F}_i := \mathcal{F}_i^{M,\{x,l_x\}} \times \mathcal{F}_i^S := \sigma(\{A \times B : A \in \mathcal{F}_i^{M,\{x,l_x\}},\ B \in \mathcal{F}_i^S\})$ *and* $\mathbb{Q} := \mathbb{P}^M \times \mathbb{Q}^S$.

We present the more general model allowing for early surrenders and as in Section 3.3 optimal policyholder behaviour is assumed. The no-lapse model is obtained under the following assumption.

Assumption 4.6. *(No-lapse model) The surrender charges satisfy* $k_i = 1$ *for all* $i < N$ *and* $k_N = 0$. *This implies that the set of admissible lapse strategies is* $\mathbb{L}_0 = \{N\}$.

Without loss of generality, from now until after Theorem 4.8 we let $l_x = 1$. The value process $\{V_i^M\}_{0 \le i \le N}$ is defined as

$$\begin{aligned} V_i^M = A_i^x \max_{\eta \in \mathbb{L}_{i,\bar{\tau}_i}} E_{\mathbb{Q}}\Big[& D_{\bar{\tau}_i \wedge \eta}^x \left(\max(DB_{K^x - 1}, W_{K^x -}) e^{-\bar{\mathrm{r}}(K^x - i)} + Ga_{\overline{K^x - 1 - i|}} \right) \\ & + A_{\bar{\tau}_i \wedge \eta}^x \left(Ga_{\overline{\eta - i|}} + W_\eta (1 - k_\eta) e^{-\bar{\mathrm{r}}(\eta - i)} \right) | \mathcal{F}_i \Big]. \end{aligned}$$

Observe that all $\eta \in \mathbb{L}_i$ are $\mathbb{F}^S$-stopping times and are independent of the mortality probability measure. Any lapse strategy η is only exercised if the insured is still alive. It remains true that the optimal lapse strategy must lie in $\mathbb{L}_{i,\bar{\tau}_i} \subset \mathbb{L}_i$.

Conditioning on the time of death and taking the expectation w.r.t. $\mathbb{P}^M$ (justified by the independence of $\mathbb{Q}^S$ and $\mathbb{P}^M$) we obtain

$$V_i^M = A_i^x V_i,$$

where

$$V_i = \max_{\eta \in \mathbb{L}_{i,\bar{\tau}_i}} V_i^\eta$$

and

$$V_i^\eta = E_{\mathbb{Q}^S}\left[\sum_{j=i}^{\bar{\tau}_i\wedge\eta-1} {}_{j-i|}q_{x+i}\left(\max\left(DB_j, W_{j+1^-}\right)e^{-\bar{\mathrm{r}}(j+1-i)} + Ga_{\overline{j-i|}}\right)\right. \tag{4.3}$$

$$\left. + {}_{\bar{\tau}_i\wedge\eta-i}p_{x+i}\left(Ga_{\overline{\eta-i|}} + W_\eta(1-k_\eta)e^{-\bar{\mathrm{r}}(\eta-i)}\right)\Big|\mathcal{F}_i^S\right].$$

The definition for the fair fee rate $\alpha^\star$ remains unchanged and it satisfies $V_0^M = P$. Select any $\eta \in \mathbb{L}_0$. Denote ${}^R\tilde{V}_i^\eta$ to be the total contract payouts up to timepoint i under this surrender strategy and discounted to $t = 0$. Then

$${}^R\tilde{V}_i^\eta = \sum_{j=0}^{\tau\wedge\eta\wedge i-1}\left(A_j^x - A_{j+1}^x\right)\left[\max(DB_j, W_{j+1^-})e^{-\bar{\mathrm{r}}(j+1)} + Ga_{\overline{j|}}\right] + A^x_{\tau\wedge\eta\wedge i}Ga_{\overline{\eta\wedge i|}}.$$

Let ${}^RV_i^\eta := E_{\mathbb{P}^M}[{}^R\tilde{V}_i^\eta]$. Then we have

$${}^RV_i^\eta = \sum_{j=0}^{\tau\wedge\eta\wedge i-1} {}_{j|}q_x\left[\max(DB_j, W_{j+1^-})e^{-\bar{\mathrm{r}}(j+1)} + Ga_{\overline{j|}}\right] + {}_{\tau\wedge\eta\wedge i}p_x Ga_{\overline{\eta\wedge i|}}.$$

For any $0 \le i \le N$, define a rescaled filtration $\mathbb{F}^{S,i} = \{\mathcal{F}_j^{S,i} := \mathcal{F}_{j+i}^S; 0 \le j \le N-i\}$. Then the process

$$Y^{\eta,i} = \left\{Y_j^{\eta,i} = e^{-\bar{\mathrm{r}}((j+i)\wedge\eta)}\,{}_{(j+i)\wedge\eta}p_x V^\eta_{(j+i)\wedge\eta} + {}^RV^\eta_{(j+i)\wedge\eta}\right\}_{0\le j\le N-i} \tag{4.4}$$

is a $(\mathbb{Q}^S, \mathbb{F}^{S,i})$ martingale. The optimal surrender strategy, $\hat{\eta}_i$, is given by (3.22) (the proof is similar and uses the martingale (4.4)).

Since $\{W_i, DB_i\}_{i=0,1,\ldots N}$ is a 2-dimensional Markov process we have

$$V_i^M = A_i^x v(i, W_i, DB_i),$$

where $v : \mathcal{I}_N \times \mathbb{R}_+ \times \mathbb{R}_+ \mapsto \mathbb{R}_+$ is recursively defined by

$$v(N, x, y) = x$$

and for $0 \leq i \leq N-1$

$$
\begin{aligned}
v(i,x,y) = \max\{ & e^{-\bar{r}}[p_{x+i}(G + pv(i+1, w(ux), db(i+1, ux, y)) \\
& + qv(i+1, w(dx), db(i+1, dx, y))) \\
& + q_{x+i}((p\max(y, ux) + q\max(y, dx))\mathbf{1}_{\{x>0\}} + \mathbf{1}_{\{x=0\}} G a_{\overline{N-i}|})], x(1-k_i)\}.
\end{aligned}
$$

This implies the boundary condition $v(i,0,y) = G a_{\overline{N-i}|}$.

The rider value process must account for the following cash flow components. The rider fee is paid prior to trigger while the insured is alive and has not surrendered. If surrender occurs prior to trigger time then no cost is incurred for the GMWB rider. In the event that no surrender occurs and the insured is alive at trigger time, the periodic GMWB guarantee is paid out until maturity regardless of death. If death occurs prior to the earlier of trigger time or surrender time, then any excess of the death benefit over the current account value is a cost incurred by the rider. Putting this together, we have

$$
\begin{aligned}
U_i^M = A_i^x \max_{\eta \in \mathbb{L}_{i,\bar{\tau}_i}} E_{\mathbb{Q}}\Bigg[& \sum_{j=i+1}^{\eta} e^{-\bar{r}(j-i)} \Big[A_{\bar{\tau}_i}^x \left(G - W_{j^-} e^{-\bar{\alpha}}\right)^+ - A_j^x W_{j^-}\left(1 - e^{-\bar{\alpha}}\right) \\
& - k_\eta W_\eta e^{-\bar{r}(\eta - i)} A_\eta^x \Big] + D_\eta^x \left(DB_{K^x - 1} - W_{K^x -}\right)^+ e^{-\bar{r}(K^x - i)} | \mathcal{F}_i \Bigg].
\end{aligned}
\tag{4.5}
$$

Then $U_i^M = A_i^x U_i = A_i^x u(i, W_i, DB_i)$, where $u : \mathcal{I}_N \times \mathbb{R}_+ \times \mathbb{R}_+ \mapsto \mathbb{R}$ is described by

$$
\begin{cases}
u(N,x,y) = 0, \\
u(i,x,y) = \max\{e^{-\bar{r}}(pu^-(i+1, ux, y) + qu^-(i+1, dx, y)), -k_i x\},
\end{cases}
\tag{4.6}
$$

and $u^- : \mathcal{I}_N^+ \times \mathbb{R}_+ \times \mathbb{R}_+ \mapsto \mathbb{R}$ is given by

$$
\begin{aligned}
u^-(i,0,y) &= G\ddot{a}_{\overline{N-i+1}|}, \\
u^-(i,x,y) &= p_{x+i-1}[(G - xe^{-\bar{\alpha}})^+ - x(1 - e^{-\bar{\alpha}}) + u(i, w(x), db(i,x,y))] \\
&\quad + q_{x+i-1}(y - x)^+.
\end{aligned}
\tag{4.7}
$$

The notation $\ddot{a}_{\overline{i+1}|} = 1 + a_{\overline{i}|}$ is an annuity due. Under Assumption 4.6 it is easy to check that the term $-k_i x$ is never binding. Note that $A^x_{i-1} u^-(i, W_{i^-}, DB_{i-1})$ is measurable w.r.t. $\mathcal{F}^S_i \times \mathcal{F}^{M,\{x,l_x\}}_{i-1}$. It is the rider value at timepoint i evaluated once the market movement for the past period is known, but prior to any transactions occurring (i.e. fees, withdrawals or death benefits). That is, the insurer knows the exact market growth in the funds over the past period but is waiting to find out about the status of the policyholder.

We denote $\{U^{M,NL}_i\}$ to refer to (4.5) when Assumption 4.6 is in place. The marginal rider value from the option to surrender is $L^M_i := U^M_i - U^{M,NL}_i \geq 0$ and can be written as

$$\begin{aligned} L^M_i = A^x_i \max_{\eta \in \mathbb{L}_{i,\bar{\tau}_i}} E_{\mathbb{Q}} \Bigg[& \sum_{j=\eta+1}^{N} e^{-\bar{\mathrm{r}}(j-i)} \left[A^x_j W_{j^-} \left(1 - e^{-\bar{\alpha}}\right) - A^x_{\bar{\tau}_i} \left(G - W_{j^-} e^{-\bar{\alpha}}\right)^+ \right] \\ & - A^x_\eta \left[k_\eta W_\eta e^{-\bar{\mathrm{r}}(\eta - i)} + D^x_N \left(DB_{K^x - 1} - W_{K^x -}\right)^+ e^{-\bar{\mathrm{r}}(K^x - i)} \right] |\mathcal{F}_i \Bigg]. \end{aligned} \tag{4.8}$$

Then $L^M_i = A^x_i l(i, W_i, DB_i)$, where $l : \mathcal{I}_N \times \mathbb{R}_+ \times \mathbb{R}_+ \mapsto \mathbb{R}_+$ is given by

$$\begin{aligned} l(N, x, y) &= 0, \\ l(i, x, y) &= \max\{p_{x+i} e^{-\bar{r}} (pl(i+1, w(ux), db(i+1, ux, y)) \\ &\qquad + ql(i+1, w(dx), db(i+1, dx, y))), -u^{NL}(i, x, y) - k_i x\}. \end{aligned}$$

Backward induction verifies that $l(i, x, y) = u(i, x, y) - u^{NL}(i, x, y)$.

Proposition 4.7. *For any $\alpha > 0$ we have*

$$V^M_i = U^M_i + A^x_i W_i \tag{4.9}$$

or equivalently

$$V^M_i = U^{M,NL}_i + L^M_i + A^x_i W_i \tag{4.10}$$

$\mathbb{Q}$-*a.s. for all* $0 \leq i \leq N$.

Proof. The equality (4.9) can be proved either directly from (4.3) and (4.5) or through backward induction applied to the functions v, u, and u^-. The procedure is similar to the proof of Proposition 3.11. We omit the details. □

The $\mathbb{F}^s$-adapted portfolio process $\{\Delta_i\}$ is defined by $\Delta_i = \Delta(i, S_i, W_i, DB_i)$, where $\Delta : \mathcal{I}_{N-1} \times \mathbb{R}_+^3 \mapsto \mathbb{R}$ is given by

$$\Delta(i, w, x, y) = \frac{u^-(i+1, ux, y) - u^-(i+1, dx, y)}{wu - wd}. \tag{4.11}$$

Note that $\Delta(i, w, 0, y) = 0$. For a given policy, the insurer follows $\{\Delta_i\}$ only up until the death of the policyholder or the surrender of the policy.

Similar to Section 3.3, we define a consumption process $\{C_i\}_{0 \le i \le N-1}$ where $C_i = c(i, W_i, DB_i)$ and $c : \mathcal{I}_N \times \mathbb{R}_+ \times \mathbb{R}_+ \mapsto \mathbb{R}_+$ is defined as

$$\begin{aligned} c(i, x, y) :&= v(i, x, y) - e^{-\bar{\mathrm{r}}}[p_{x+i}(G + pv(i+1, w(ux), db(i+1, ux, y)) \\ &\quad + qv(i+1, w(dx), db(i+1, dx, y)))] \\ &\quad + q_{x+i}((p \max(y, ux) + q \max(y, dx))\mathbf{1}_{\{x>0\}} + \mathbf{1}_{\{x=0\}} G a_{\overline{N-i|}})] \\ &= u(i, x, y) - e^{-\bar{\mathrm{r}}}[pu^-(i+1, ux, y) + qu^-(i+1, dx, y)]. \end{aligned} \tag{4.12}$$

The second equality can be verified using Proposition 4.7, similar to (3.30). Under Assumption 4.6 we have $C \equiv 0$.

Construct the replicating portfolio by starting with initial capital $X_0 = x_0$ and following the portfolio process $\{\Delta_i\}$. For $i \in \mathcal{I}_N^+$ we have

$$\begin{aligned} X_i = \left(X_{i-1} - A^x_{i-1}(\Delta_{i-1} S_{i-1} + C_{i-1})\right) e^{\bar{\mathrm{r}}} + A^x_{i-1}\Delta_{i-1} S_i + A^x_i \left[F_i - \left(G - W_{i^-} e^{-\bar{\alpha}}\right)^+\right] \\ - \left(A^x_{i-1} - A^x_i\right) \left[(DB_{i-1} - W_{i^-})^+ \mathbf{1}_{\{\tau \ge i\}} + G \ddot{a}_{\overline{N-i+1|}} \mathbf{1}_{\{\tau < i\}}\right]. \end{aligned} \tag{4.13}$$

The fees, payouts, portfolio process, and consumption process have all been defined in $\mathbb{F}^S$. Of course they are only applicable while the policy is in force (prior to death or surrender). For that reason, the terms are accompanied by A^x_i factors in (4.13). Given a surrender strategy $\eta \in \mathbb{L}_0$, the insurer will close out its position at timepoint η and the process of interest is $\{X_{i \wedge \eta}\}_{0 \le i \le N}$. The time zero profit is $\Pi = e^{-\bar{\mathrm{r}}\eta} X_\eta$, since if death occurs prior to η then the portfolio remains unchanged for all periods between death and η, aside from interest accumulation.

Although we no longer have almost sure equivalence of U^M and X with respect to the product measure $\mathbb{Q}$, an analogous result holds by considering the conditional expectation with respect to $\mathbb{P}^M$.

Theorem 4.8. *Suppose the fee rate α is charged and the initial capital is $x_0 = U_0^M$. Then the following relation holds between X_i, described by (4.13), and U_i^M, given by (4.5):*

$$\mathbb{Q}^S(\mathbb{E}_{\mathbb{P}^M}[X_i - U_i^M] = 0) = 1$$

for all $i \in \mathcal{I}_N$.

Proof. We proceed by induction. By assumption we have that $X_0 = U_0^M$. Suppose that $E_{\mathbb{P}^M}[X_i] = E_{\mathbb{P}^M}[U_i^M]$ $\mathbb{Q}^S$-a.s. for some $i \in \mathcal{I}_{N-1}$. For a process H_i we write $H_i(\bar{\omega}_i; j)$ for its value at time i for the specific path $\bar{\omega}_i\omega_{i+1}\dots\omega_N \in \Omega^S$ (where ω_j can take any value in $\{u, d\}$ for all $j > i$) and the specific set $(K^x)^{-1}(j) \in \mathcal{F}_N^{M,\{x,1\}}$. For any fixed $\bar{\omega}_i$ we need to show that

$$\begin{cases} E_{\mathbb{P}^M}[X_{i+1}(\bar{\omega}_i u; K^x)] = E_{\mathbb{P}^M}[U_{i+1}^M(\bar{\omega}_i u; K^x)], \\ E_{\mathbb{P}^M}[X_{i+1}(\bar{\omega}_i d; K^x)] = E_{\mathbb{P}^M}[U_{i+1}^M(\bar{\omega}_i d; K^x)]. \end{cases}$$

We prove the first equality, the second one is shown in an identical manner. For conciseness, we omit $\bar{\omega}_i$.

Observe that $E_{\mathbb{P}^M}[U_{i+1}^M(u; K^x)] = {}_{i+1}p_x U_{i+1}(u)$. Also $X_{i+1}(u; j) = X_{i+1}(u; K^x > i+1)$ for all $j > i+1$, since $X_{i+1} \in \mathcal{F}_{i+1}$. From (4.13) we have $X_{i+1}(u; j) = X_i(; j)e^{\bar{r}}$ for all $j \leq i$. Therefore

$$\begin{aligned} E_{\mathbb{P}^M}[X_{i+1}(u, K^x)] &= \sum_{j=1}^{N} {}_{j-1|}q_x X_{i+1}(u, j) + {}_N p_x X_{i+1}(u, K^x > N) \\ &= \sum_{j=1}^{i} {}_{j-1|}q_x X_i(; j)e^{\bar{r}} + {}_{i|}q_x X_{i+1}(u, i+1) + {}_{i+1}p_x X_{i+1}(u, K^x > i+1). \end{aligned}$$

After applying (4.13) to $X_{i+1}(u, i+1)$ and $X_{i+1}(u; K^x > i+1)$, we obtain

$$E_{\mathbb{P}^M}[X_{i+1}(u, K^x)] = E_{\mathbb{P}^M}[X_i(; K^x)]e^{\bar{r}} +_i p_x[\Delta_i S_i(u - e^{\bar{r}}) - C_i e^{\bar{r}} \\ - p_{x+i}((G - W_i u e^{-\bar{\alpha}})^+ - W_i u(1 - e^{-\bar{\alpha}})) \\ - q_{x+i}((DB_i - W_i u)^+ \mathbf{1}_{\{\tau > i\}} + G\ddot{a}_{\overline{N-i|}} \mathbf{1}_{\{\tau \le i\}})]. \tag{4.14}$$

By the induction hypothesis,

$$E_{\mathbb{P}^M}[X_i(; K^x)]e^{\bar{r}} = {}_i p_x U_i e^{\bar{r}}.$$

Then substituting (4.11) and (4.6) and applying (4.7) (in the form U^-_{i+1}, but conditioning on $\tau > i$), we have

$$\begin{aligned} E_{\mathbb{P}^M}[X_{i+1}(u, K^x)] &= {}_i p_x[(U_i - C_i)e^{\bar{r}} + (U^-_{i+1}(u) - U^-_{i+1}(d))q - G p_{x+i} \mathbf{1}_{\{\tau \le i\}} \\ &\quad + \mathbf{1}_{\{\tau > i\}}(p_{x+i} U_{i+1}(u) - U^-_{i+1}(u)) - q_{x+i}(G\ddot{a}_{\overline{N-i|}} \mathbf{1}_{\{\tau \le i\}})] \\ &= {}_i p_x[U^-_{i+1}(u) \mathbf{1}_{\{\tau \le i\}} - G p_{x+i} \mathbf{1}_{\{\tau \le i\}} + \mathbf{1}_{\{\tau > i\}} p_{x+i} U_{i+1}(u) \\ &\quad - q_{x+i} G\ddot{a}_{\overline{N-i|}} \mathbf{1}_{\{\tau \le i\}}] \\ &= {}_{i+1} p_x[\mathbf{1}_{\{\tau > i\}} U_{i+1}(u) + G a_{\overline{N-(i+1)|}} \mathbf{1}_{\{\tau \le i\}} \\ &= {}_{i+1} p_x U_{i+1}(u). \end{aligned}$$

This completes the proof. □

Suppose homogeneous policies are sold to l_x independent policyholders aged x, each with an initial premium of P and the fair rider fee $\alpha^\star$ is charged. For the pool of l_x insureds, the number of deaths between time $i\delta t$ and $(i+1)\delta t$ is

$$\mathcal{D}_i^{l_x,x} := \sum_{j=1}^{l_x} \left(A_i^{x,j} - A_{i+1}^{x,j} \right)$$

for $i \in \mathcal{I}_{N-1}$. The number of members alive at time i is

$$\mathcal{A}_i^{l_x,x} = \sum_{j=1}^{l_x} A_i^{x,j} = l_x - \sum_{j=1}^{i-1} \mathcal{D}_j^{l_x,x}.$$

By the strong law of large numbers (SLLN), as $l_x \to \infty$,

$$\frac{\mathcal{D}_i^{l_x,x}}{l_x} \to {}_{i|}q_x \quad \text{and} \quad \frac{\mathcal{A}_i^{l_x,x}}{l_x} \to {}_ip_x$$

$\mathbb{P}^M$-a.s., for all $i \in \mathcal{I}_N$.

The aggregate replicating portfolio process is the sum of the individual replicating portfolio processes given by (4.13):

$$X_i^{\{l_x\}} = \sum_{j=1}^{l_x} X_i^j,$$

where $X_i^j \in \mathcal{F}_i^S \times \mathcal{F}_i^{M,x,j}$ for $1 \le j \le l_x$ and $1 \le i \le N$. The aggregate rider value process is

$$U_i^{M,\{l_x\}} = \sum_{j=1}^{l_x} U_i^{M,j} = \mathcal{A}_i^{l_x,x} U_i,$$

since $U_i^j = 0$ if $A_i^{x,j} = 0$. We define two processes $\{X_i^\star = E_{\mathbb{P}^M}[X_i^{\{1\}}]\}_{i=0,1,\ldots,N}$ and $\{U_i^\star = E_{\mathbb{P}^M}[U_i^{M,\{1\}}]\}_{i=0,1,\ldots,N}$, both of which lie in $(\Omega^S, \mathcal{F}_N^S, \mathbb{F}^S, \mathbb{Q}^S)$. Then by the SLLN we have

$$\left\{\frac{X_i^{\{l_x\}}}{l_x}\right\} \to \{X_i^\star\} \quad \text{and} \quad \left\{\frac{U_i^{M,\{l_x\}}}{l_x}\right\} \to \{U_i^\star\}$$

$\mathbb{P}^M$-a.s., as $l_x \to \infty$. Beginning with $X_0^\star = 0$, from (4.14) we have

$$\begin{aligned} X_i^\star = X_{i-1}^\star e^{\bar{r}} + {}_{i-1}p_x \Big[&\Delta_{i-1}(S_i - S_{i-1}e^{\bar{r}}) - C_{i-1}e^{\bar{r}} + p_{x+i-1}\left[F_i - \left(G - W_{i^-}e^{-\bar{\alpha}}\right)^+\right] \\ &- q_{x+i-1}\left[(DB_{i-1} - W_{i^-})^+ \mathbf{1}_{\{\tau \ge i\}} + G\ddot{a}_{\overline{N-i+1|}} \mathbf{1}_{\{\tau < i\}}\right] \Big] \end{aligned}$$

for $i \in \mathcal{I}_N^+$. It is immediate that $U_i^\star = {}_ip_x U_i$. Finally, from Theorem 4.8 we have

$$X_i^\star = U_i^\star$$

$\mathbb{Q}^S$-a.s., for $i \in \mathcal{I}_N$.

Mortality risk diversification is attained in the limit as $l_x \to \infty$, and we have perfect hedging. The fair fee was determined assuming optimal surrender behaviour on the part of each policyholder, given survival. If policyholders act irrationally then the insurer can consume from each portfolio at each occurrence of this irrationality. The limiting aggregate portfolio process for the pool is constructed on the basis of homogeneous behaviour of all policyholders, whether or not they act rationally.

Remark 4.9. The limiting process was obtained assuming homogeneous policies. This assumption can be weakened to allow for varying initial premiums P by policy, although each policy must have an issue age of x and a common rider fee α. This is true since P can be scaled out of all the processes and the rider fee is independent of the premium P. Let the premium for policy i be P_i. Suppose $\{P_i; i \geq 1\}$ satisfies $\sum_{i=1}^{n} P_i \to \infty$ as $n \to \infty$. Further assume that $\{P_i\}$ is monotonically increasing and satisfies $\sup_{n\geq 1} \frac{nP_n}{\sum_{i=1}^{n} P_i} < \infty$ or that $\{P_i\}$ is monotonically decreasing in which case no condition is needed. From Theorem 1 in Etemadi (2006), as $l_x \to \infty$, we have

$$\frac{\sum_{j=1}^{l_x} P_j A_i^{x,j}}{\sum_{j=1}^{l_x} P_j} \to {}_i p_x$$

$\mathbb{P}^M$-a.s. for all $i \in \mathcal{I}_N$. Therefore

$$\left\{ \frac{X_i^{\{l_x\}}}{\sum_{i=1}^{l_x} P_i} \right\} \to \{X_i^\star\},$$

with a similar result for $U^\star$. The average is taken on a per premium dollar basis and both $X^\star$and $U^\star$ have $P = 1$.

4.4 Numerical Results

We consider two examples. The mortality is modeled using Example 4.4.

Example 4.10. Figure 4.2 plots the fair rider fee $\alpha^\star$ against the issue age x for a GMWB with a return of premium DB and an annual ratchet DB without lapses. The parameters are: $g = 7.14\%$, $T = 14$, $r = 5\%$, $\sigma = 20\%$, and $\delta t = 1$. The ratchet adds considerably more value to the contract. The figure on the right zooms in on the ages 40-70. The GMWB plus return of premium DB rider is largely insensitive to x. The

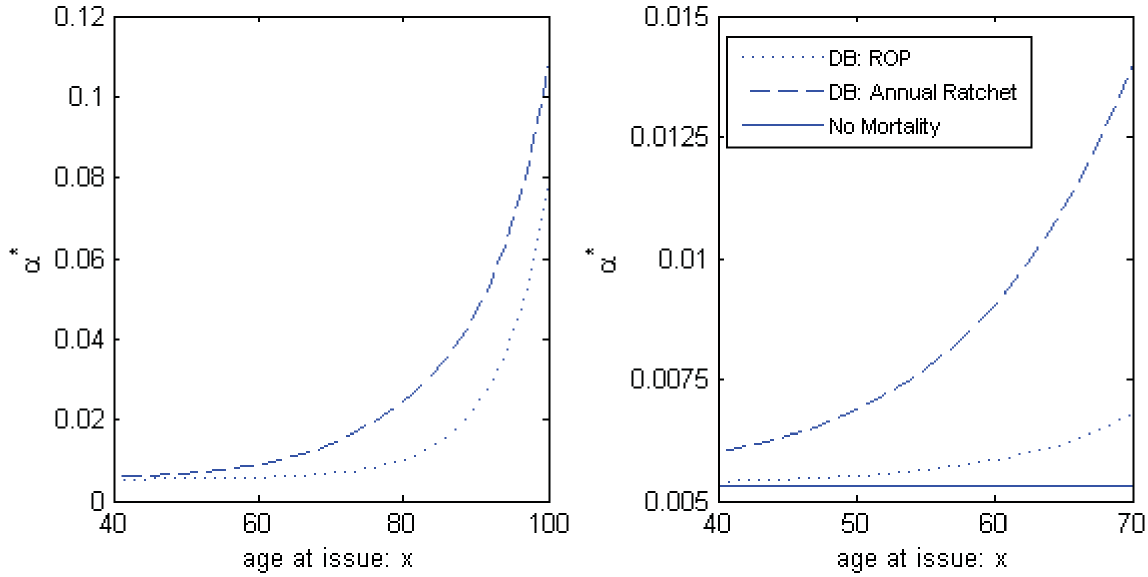


Figure 4.2: $\alpha^{\star}$ as a function of issue age x

payouts upon death or survival are fairly similar in this instance. Under the model from Section 3.2 without mortality, we have $\alpha^{\star} = 53\text{bps}$ or $V_0(100, 53\text{bps}) = 100$. For the return of premium DB with $x = 60$, we have $\alpha^{\star} = 58\text{bps}$ and $V_0^M(100, 53\text{bps}) = 100.35$. Depending on the product specifications and parameters, mortality may have only a small effect.

Example 4.11. The diversifiable mortality risk assumption is often quick to be used in the literature. Given the prescribed portfolio process (4.11) which assumes the risk is diversifiable, we consider the hedging losses when there are only a finite number of policies sold. For $l_x \in \{10, 1000, 100000\}$ we simulated the time of deaths for each policy to obtain $\{\widehat{T}_j^x\}_{1 \le j \le l_x}$, and computed the average losses per policy per \$100 premium for each path in the binomial model. The parameters used are: $x = 60$, $g = 10\%$, $T = 10$, $r = 5\%$, $\sigma = 15\%$, $\delta t = 1$, and $P = 100$. Surrenders are not allowed.

For the GMWB with an annual ratchet DB, Figure 4.3 shows the convergence of the hedging losses to zero under the delta hedging strategy as l_x increases. The values are time-zero present values and the losses under no hedging are also displayed.

Values per \$100	Hedging			No Hedging			
l_x	10	1000	100000	10	1000	100000	∞
GMWB + Ratchet DB							
$E_{\mathbb{Q}}[\Pi \mid \{\widehat{T}_j^x\}_{1\leq j\leq l_x}]$	0.122	0.030	0.004	0.122	0.030	0.004	0
$SD_{\mathbb{Q}}[\Pi \mid \{\widehat{T}_j^x\}_{1\leq j\leq l_x}]$	0.768	0.175	0.008	5.631	5.787	5.860	5.860
GMWB + Return of Premium DB							
$E_{\mathbb{Q}}[\Pi \mid \{\widehat{T}_j^x\}_{1\leq j\leq l_x}]$	0.261	0.054	0.001	0.261	0.054	0.001	0
$SD_{\mathbb{Q}}[\Pi \mid \{\widehat{T}_j^x\}_{1\leq j\leq l_x}]$	0.446	0.091	0.004	5.560	5.736	5.776	5.777

Table 4.1: Profit metrics with and without hedging, with GMDBs

Figure 4.4 plots the losses for the limiting portfolio $X^\star$. Table 4.1 provides the profit metrics $E_{\mathbb{Q}}[\Pi|\{\widehat{T}_j^x\}_{1\leq j\leq l_x}]$ and $SD_{\mathbb{Q}}[\Pi|\{\widehat{T}_j^x\}_{1\leq j\leq l_x}]$ for hedging and no hedging where Π is the average profit per policy discounted to $t = 0$. The results are also given when the DB rider is a ROP. The results for both DBs were obtained using the same sets of simulated death times. The column with $l_x = \infty$ represents the results for $X^\star$. The fair fee with the ratchet is 57bps and with the ROP is 44bps. The metrics were calculated using the exact binomial distribution under $\mathbb{Q}$ for the financial risk and the simulated deaths for the mortality risk. For the purpose of examining convergence w.r.t. l_x, we assume no market price of risk (i.e. $\mu = r$).

Selling a limited number of policies or facing capacity constraints does not impose a significant risk to the insurer in this case because the payouts are similar upon death or survival and diversification occurs rapidly. The average hedging profits are higher with the ROP, but the profits (losses) have more volatility with the ratchet since it pays higher benefits and has higher fees. Under $\mathbb{Q}$ the expected profits are equal under hedging and no hedging. It is the variance that is reduced by hedging.

Without mortality risk, each policy in the pool is subject to a common equity risk and in the binomial world the correct hedging strategy works for any number of policies. Mortality risk introduces incompleteness into the model. Under the

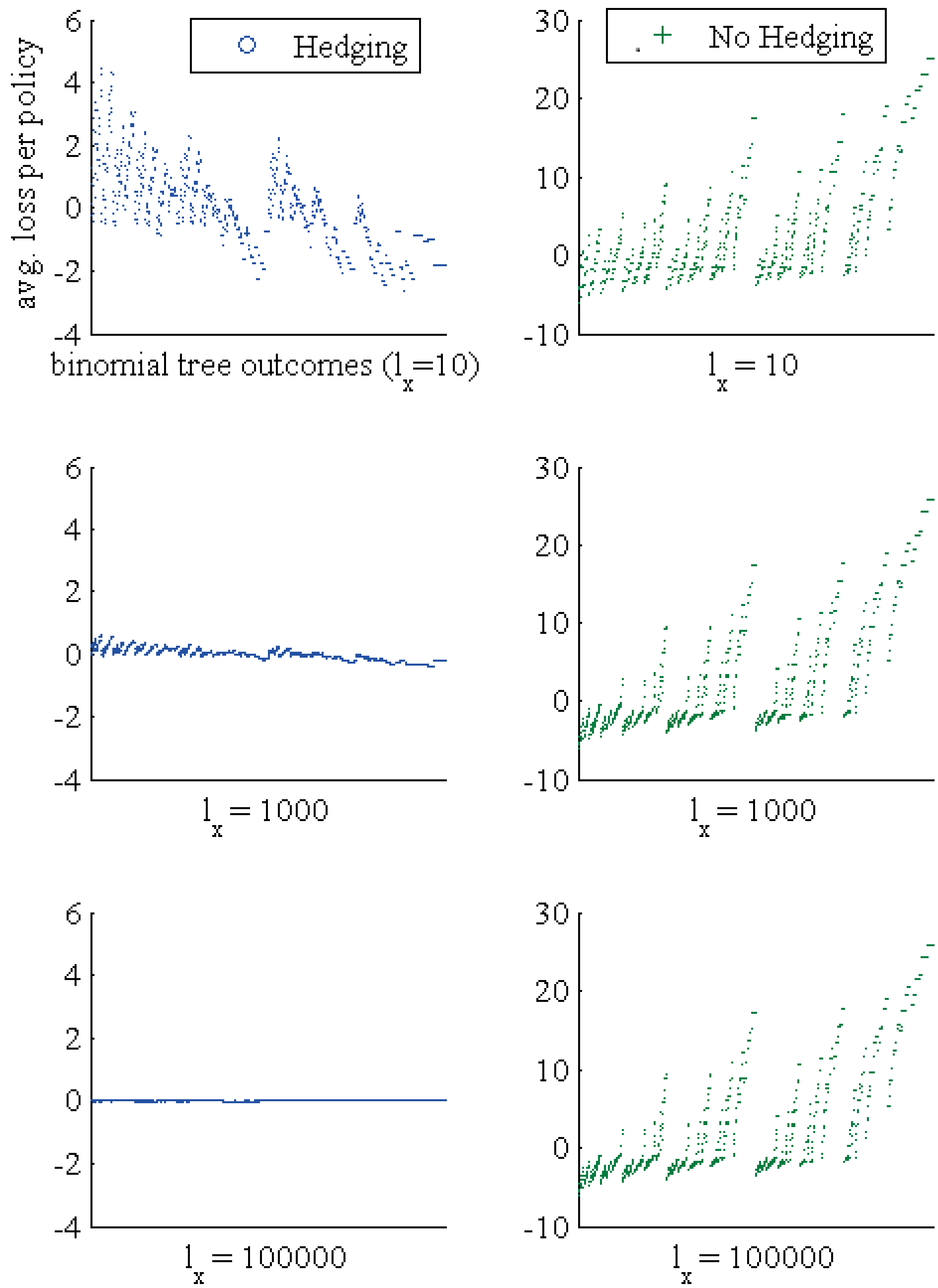


Figure 4.3: Convergence of losses for GMWB plus ratchet DB as $l_x \to \infty$ where the average losses per policy under simulated mortality are shown for each market outcome.

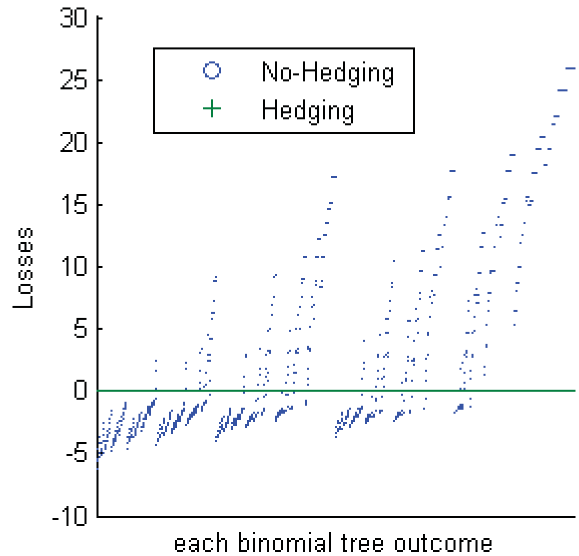


Figure 4.4: Losses for GMWB plus ratchet DB with complete diversification ($X^{\star}$)

assumption of mortality risk diversification the market regains completeness. This occurs in the limit by selling sufficiently large pools of relatively small contract sizes.

Aside from risk pooling and diversification, other risk-management options are reinsurance and longevity bonds. Additionally the typical large life insurer with significant amounts of underwritten business in life insurance and annuities has a degree of natural risk reduction since these instruments have partially offsetting risks. Assuming none of these option are available - there are no re-insurers, longevity bonds do not exist and the insurer only sells annuities - the insurer's main tool for mitigating its risk exposure is by selling a large number of policies of relatively small amounts, thus reducing fluctuations in the realized mortality rates around the expected mortality rates.

Conclusion and Future Work

Based on the continuous model from Milevsky and Salisbury (2006), we constructed a binomial asset pricing model for the GMWB rider incorporating optimal policyholder surrender behaviour. The ability to model early surrenders using the basic tools is one distinct advantage over Monte Carlo methods. The other advantage was demonstrated by easily obtaining an explicit hedging strategy in a binomial (CRR) world that was proved to perfectly hedge the product. A drawback of this model is the $O(2^N)$ growth of the non-recombining binomial trees. Nevertheless, by the tractability of the model and its finite nature, it is straightforward to obtain numerical results concerning any aspect of the product, provided that δt is manageable. The qualitative conclusions drawn from such an analysis are expected to hold true in the more general continuous model. Indeed, our numerical results are consistent with those presented in more complex models.

When treating the binomial model as an approximation of an underlying continuous model, solving for the fair fee with just two time periods a year produced results close to those obtained with Monte Carlo methods. Although currently we face memory capacity constraints, future technological advancements may increase the applicability of the binomial model as a practical pricing and hedging method. At the very least, it is a useful machinery to obtain preliminary results before resorting to more powerful tools.

With a little more programming expertise, the Asian approximation method (Section 3.4) could produce results for monthly timesteps. It remains to study the effectiveness of hedging strategies based on the Asian approximation method. As a result of the rider fees, the tree for Z trends downwards which has implications on the Asian approximation method (see Section 3.4, closing paragraph). The effect of this asymmetry on the results needs to be examined and quantified.

The diversification argument for mortality risk is sometimes abused in the literature. After applying diversification arguments to obtain the fair fee and hedging results, we imposed capacity constraints by considering finite pools and saw that diversification occurs fairly rapidly. The results support the common claim that insurers are able to diversify mortality risk.

This work was presented for a basic GMWB rider. Although including additional features such as rollups and ratchets requires adding the dimension of A (the guarantee balance), it remains straightforward to model the product in a binomial framework assuming static withdrawals and optimal lapses. However, the optimal strategy is no longer restricted to withdrawing G or lapsing. It may be optimal to not withdraw but still keep the policy in force. The methods presented in this thesis are ill-suited to deal with this and more general mesh techniques, similar to Bauer et al. (2008) but simplified to binomial movements, would be necessary.

The framework developed in this thesis models a withdrawal at the end of each binomial period. This was done to approximate the continuous model from Milevsky and Salisbury (2006). In practice, withdrawals occur a finite number of times per year (i.e. annually as in Bauer et al. (2008) or even more frequently such as monthly). The binomial model can be generalized by considering n periods per year but $k \leq n$ withdrawals, where $k/n \in \mathbb{N}$. When $k = n$, this reduces to the model developed in this thesis. Between withdrawal dates, the binomial tree will have recombining branches and computational results could be obtained for larger n. Rather than growing at $O(2^{nT})$, the tree grows at a reduced rate of $O(n/k + 1)^{kT}$.

Suboptimal Behaviour

Modern insurance products allow for much optionality and decision making on the part of the policyholder. Future research will focus on further developing tractable models to address policyholder behaviour risk, but which reflect the unique characteristics of the insurance markets. We have considered the two extremes in terms of policyholder lapse behaviour: no lapses and optimal lapses. It is clear that neither of these extremes are observed in practice. Otherwise a pool of homogeneous policies would instantaneously surrender at the moment it is optimal to do so. Pricing under optimal lapses is justified by the argument that policyholders have the right to act optimally. However, in Section 1.5 we pointed out vital differences between the GMWB product and standard financial options, of which the lack of a liquid secondary market is paramount.

Several recent papers, including De Giovanni (2010) and Li and Szimayer (2010), are based on rational expectation approaches that address the *emergency fund hypothesis* and *interest rate hypothesis* and more accurately price unit-linked life insurance products. These papers can be linked back to Stanton (1995) where a model was developed for rational mortgage prepayments for Mortgage-Backed Securities (MBS) for pools of mortgages with heterogeneous transaction cost structures. There are several modeling differences though. Namely, the recent papers deal with a single representative policy but Stanton (1995) explicitly treats heterogeneous pools. There is a phenomena known as the burnout effect which refers to the changing demographics of a pool over time. Stanton (1995) models this and also includes randomized decision times to address the observed lack of uniformity in behaviour for a set of homogeneous holders.

We have considered new modeling approaches for GMWBs based on both Stanton (1995) and Li and Szimayer (2010) but only discuss the former here. We model a pool of policyholders with heterogeneous behaviour, owing to either different levels of rationality or imperfect financial competence. In this case, burnout reflects the higher proportion of subrational policyholders over time as more rational policyholder lapse when optimal. Our problem differs from Stanton (1995) because in a pool of homogeneous policies every insured faces identical surrender transaction costs.

As a first step, we consider that policyholders may be unable to perfectly determine optimal behaviour and will only become aware of the correct decision once the contract is significantly out of the money (OTM). A population can be subdivided into cohorts by the levels of financial ability and rationality. Initially we focus on single cohorts only. We introduce a fictional penalty term which is added to the surrender costs to reflect the imperfect understanding of the policyholder. This penalty term is fictional in the sense that upon surrender the penalty term does not influence the cash flow received. However it modifies the optimal surrender decision and allows us to model suboptimal behaviour in the typical optimal behaviour framework.

Let θ denote the penalty term. It represents the degree of out-of-moneyness needed to convince the policyholder to surrender. Denote the optimal surrender strategy by

$$\eta^\star = \arg\max_{\eta \in \mathbb{L}_0} E_{\mathbb{Q}} \left[G a_{\overline{\eta}\rceil} + e^{-\bar{r}\eta} W_\eta (1 - k_\eta)(1 - \theta) \right].$$

Then the value is

$$V_0^\theta = E_{\mathbb{Q}} \left[G a_{\overline{\eta^\star}\rceil} + e^{-\bar{r}\eta^\star} W_{\eta^\star} (1 - k_{\eta^\star}) \right].$$

We introduce an intensity rate ρ to model the absence of instantaneous reaction of a group of homogeneous policyholders with identical contracts. This intensity rate can be thought of as driving a randomized decision time, similar to Stanton (1995). A base lapse rate λ is incorporated to reflect the *emergency fund hypothesis*.

The surrender intensity rate γ_i is expressed as

$$\gamma_i = \begin{cases} \lambda & V_i^c > W_i(1 - k_i)(1 - \theta), \\ \lambda + \rho & V_i^c \leq W_i(1 - k_i)(1 - \theta), \end{cases}$$

where V_i^c is the continuation value of the contract. Let $q_a^R = 1 - e^{-\rho}$ denote the annual probability of lapsing when it is optimal to do so - given the penalty term. Then $q_{\delta t}^R = 1 - (1 - q_a^R)^{\delta t}$ is the probability for a period of length δt. We denote the total probability of lapsing in any period i by

$$\begin{cases} q_e = 1 - e^{-\lambda\delta t} & V_i^c > W_i(1-k_i)(1-\theta), \\ q_r = 1 - e^{-(\lambda+\rho)\delta t} & V_i^c \leq W_i(1-k_i)(1-\theta). \end{cases}$$

A backward induction scheme is formulated for $V_i^\theta = v(i, W_i)$. Beginning with period N, we have $v(N, x) = x$ for all $x \in \mathbb{R}_+$. For $0 \leq i < N$ and $x \in \mathbb{R}_+$,

$$v^c(i,x) = (pv(i+1, w(ux) + qv(i+1, w(dx)) + G)e^{-r}$$

is the continuation value. Taking into account the different lapse probabilities:

$$v(i,x) = \begin{cases} v^c(i,x), & \text{if } x = 0; \\ q_e x(1-k_i) + (1-q_e)v^c(i,x), & \text{if } v^c(i,x) > x(1-k_i); \\ v^c(i,x), & \text{if } x(1-k_i)(1-\theta) < v^c(i,x) \leq x(1-k_i); \\ q_r x(1-k_i)(1-\theta) \\ \quad + (1-q_r)v^c(i,x), & \text{if } v^c(i,x) \leq x(1-k_i)(1-\theta). \end{cases}$$

In the third case $v(i,x) = v^c(i,x)$ because even though $x(1-k_i)$ is received upon surrender, $x(1-k_i) - v^c(i,x)$ must be attributed to the penalty term which is computed separately. The cash flows associated with the penalty term are described by $\psi_i = \psi(i,x)$ where $\psi_N \equiv 0$. The continuation value ψ^c is defined by

$$\psi^c(i,x) = (p\psi(i+1, w(ux)) + q\psi(i+1, w(dx)))e^{-r}.$$

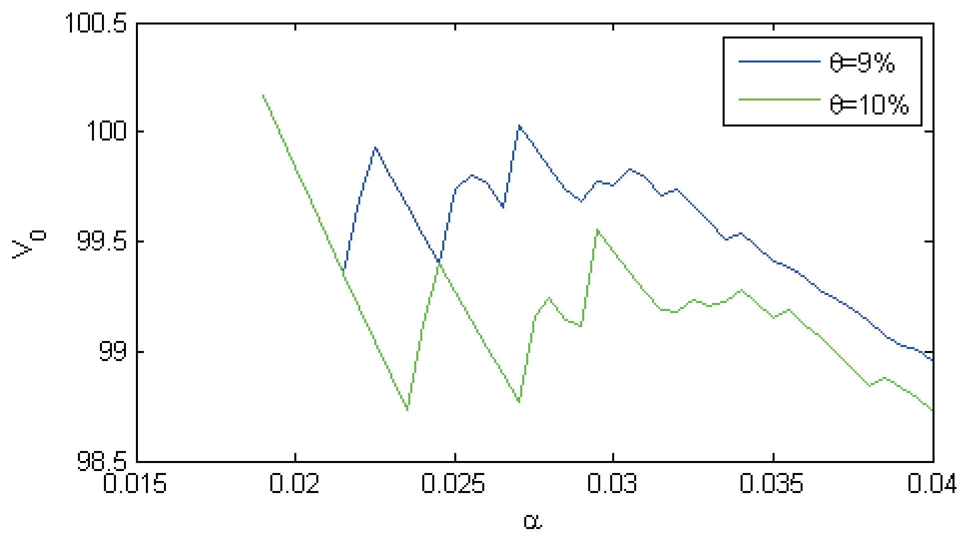
The value function ψ is given by

$$\psi(i,x) = \begin{cases} 0, & \text{if } x = 0; \\ \psi^c(i,x)(1-q_e), & \text{if } v^c(i,x) > x(1-k_i); \\ \psi^c(i,x)(1-q_e) \\ \quad + q_e(x(1-k_i) - v^c(i,x)), & \text{if } x(1-k_i)(1-\theta) < v^c(i,x) \leq x(1-k_i); \\ \psi^c(i,x)(1-q_r) + q_r x(1-k_i)\theta, & \text{if } v^c(i,x) \leq x(1-k_i)(1-\theta). \end{cases}$$

Then we have $V_0 = v(0,P) + \psi(0,P)$.

Preliminary Results There is a flaw with this model. V_0 behaves unpredictably and is no longer a strictly decreasing function of α. In fact there are jumps and non-unique solutions to $\alpha^\star$. This is seen in the provided (α, V_0) plot for $\theta = 9\%$ and $\theta = 10\%$. Consider a fee rate α_1 and the path $\bar{\omega}_i$ such that W_i falls between the surrender boundary with the penalty term and the surrender boundary without it. As α increases, the account value drops and will fall below the lower boundary which causes a surrender. But since the penalty term is fully recovered, this bumps the contract value up to the surrender boundary without the penalty.

Despite this flaw, we obtained numerous comparative statics that are intuitive with a subrational approach. Future work will try to address this issue. As a final comment, we believe the simple yet elegant result used repeatedly in this thesis which unifies the perspectives of the insurer and policyholder in the general case of optimal behaviour will play a key role in exploring subrational models, although additional terms may need to be added to maintain the relationship.



V_0 as a function of α with $\rho = \infty$ and $\lambda = 0$.

References

Bacinello, A.R. (2005). Endogenous model of surrender conditions in equity-linked life insurance. *Insurance: Mathematics and Economics* 37(2):270–296.

Bacinello, A.R., Millossovich, P., Olivieri, A., and Pitacco, E. (2011). Variable annuities: A unifying valuation approach. *Insurance: Mathematics and Economics* 49(3):285–297.

Bauer, D., Kling, A., and Russ, J. (2008). A universal pricing framework for guaranteed minimum benefits in variable annuities. *ASTIN Bulletin* 38(2):621–651.

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy* 81:637–654.

Borodin, A.N. and Salminen, P. (2002). *Handbook of Brownian Motion: Facts and Formulae*, 2nd edition. Birkhäuser, Basel, Boston, Berlin.

Bowers, Gerber, Hickman, Jones, and Nesbitt (1997). *Actuarial Mathematics*, 2nd edition. Society of Actuaries, Schaumburg, Illinois.

Boyle, P. and Schwartz, E.S. (1977). Equilibrium prices of guarantees under equity-linked contracts. *Journal of Risk and Insurance* 44(4):639–660.

Chen, Z. and Forsyth, P.A. (2008). A numerical scheme for the impulse control formulation for pricing variable annuities with a guaranteed minimum withdrawal benefit (GMWB). *Numerische Mathematik* 109(4):535–569.

Chen, Z., Vetzal, K., and Forsyth, P.A. (2008). The effect of modelling parameters on the value of GMWB guarantees. *Insurance: Mathematics and Economics* 43(1):165–173.

Costabile, M., Massabo, I., and Russo, E. (2006). An adjusted binomial model for pricing Asian options. *Review of Quantitative Finance and Accounting* 27(3):285–296.

Cox, J.C., Ross, S.A., and Rubinstein, M. (1979). Option pricing: A simplified approach. *Journal of Financial Economics* 7(3):229–263.

Dai, M., Kwok, Y.K., and Zong, J. (2008). Guaranteed minimum withdrawal benefit in variable annuities. *Mathematical Finance* 18(4):595–611.

De Giovanni, D. (2010). Lapse rate modeling: a rational expectation approach. *Scandinavian Actuarial Journal* 2010(1):56–67.

Duffie, D. (2001). *Dynamic Asset Pricing Theory*, 3rd edition. Princeton University Press, Princeton, NJ.

Etemadi, N. (2006). Convergence of weighted averages of random variables revisited. *Proceedings of the American Mathematical Society* 134(9):2739–2744.

Geske, R. and Shastri, K. (1985). Valuation by approximation: A comparison of alternative option valuation techniques. *The Journal of Financial and Quantitative Analysis* 20(1):45–71.

Glasserman, P. (2004). *Monte Carlo Methods in Financial Engineering*. Springer-Verlag, New York.

Hull, J. and White, A. (1993). Efficient procedures for valuing European and American path-dependent options. *Journal of Derivatives* 1(1):21–31.

Jeanblanc, M., Yor, M., and Chesney, M. (2009). *Mathematical Methods for Financial Markets*. Springer Finance, London.

Karatzas, I. and Shreve, S.E. (1991). *Brownian Motion and Stochastic Calculus*, 2nd edition. Springer-Verlag, New York.

Knoller, C., Kraut, G., and Schoenmaekers, P. (2011). On the propensity to surrender a variable annuity contract - an empirical analysis of dynamic policyholder behaviour. *Munich Risk and Insurance Center Working Paper No. 7.* Available at SSRN: http://ssrn.com/abstract=1976203.

Levy, E. (1992). Pricing European average rate currency options. *Journal of International Money and Finance* 11(5):474–491.

Li, J. and Szimayer, A. (2010). The effect of policyholders' rationality on unit-linked life insurance contracts with surrender guarantees. *Working paper.* Available at SSRN: http://ssrn.com/abstract=1725769.

LIMRA (2011). LIMRA retirement research, November 2011 - 3Q2011GLBSales. http://www.limra.com/newscenter/databank/3Q2011GLBSales.pdf. Accessed 23 May 2012.

LIMRA (2012). Annuity sales estimates 2002-2011. http://www.limra.com/PDFs/NewsCenter/DataBank/AnnuitySales02-11.pdf. Accessed 23 May 2012.

Liu, Y. (2010). *Pricing and Hedging the Guaranteed Minimum Withdrawal Benefits in Variable Annuities*. Ph.D. thesis, University of Waterloo.

Milevsky, M.A., Promislow, S.D., and Young, V.R. (2006). Killing the law of large numbers: Mortality risk premiums and the Sharpe ratio. *Journal of Risk and Insurance* 73(4):673–686.

Milevsky, M.A. and Salisbury, T.S. (2001). The real option to lapse a variable annuity: Can surrender charges complete the market? *Conference Proceedings of the 11th Annual International AFIR Colloquium* v2:537–561.

Milevsky, M.A. and Salisbury, T.S. (2006). Financial valuation of guaranteed minimum withdrawal benefits. *Insurance: Mathematics and Economics* 38(1):21–38.

Moenig, T. and Bauer, D. (2011). Revisiting the risk-neutral approach to optimal policyholder behavior: A study of withdrawal guarantees in variable annuities. *Working paper.* Available online at `http://www.uni-ulm.de/fileadmin/website_uni_ulm/mawi.inst.140/Vortr%C3%A4ge/2011_MoenigBauer_RNVtoPHbehavior.pdf`.

Møller, T. (1998). Risk-minimizing hedging strategies for unit-linked life insurance contracts. *ASTIN Bulletin* 28(1):17–47.

Øksendal, B.K. (2003). *Stochastic Differential Equations: An Introduction with Applications*, 6th edition. Springer Verlag, Berlin, Heidelberg, and New York.

Peng, J., Leung, K.S., and Kwok, Y.K. (2012). Pricing guaranteed minimum withdrawal benefits under stochastic interest rates. *Quantitative Finance* 12(6):933–941.

Shreve, S.E. (2004a). *Stochastic Calculus for Finance I: The Binomial Asset Pricing Model*. Springer Finance, New York.

Shreve, S.E. (2004b). *Stochastic Calculus for Finance II: Continuous-Time Models*. Springer Finance, New York.

Stanton, R. (1995). Rational prepayment and the valuation of mortgage-backed securities. *The Review of Financial Studies* 8(3):677–708.

Wang, S. (2008). *Longevity Risks: Modelling and Financial Engineering*. Ph.D. thesis, Ulm University.

Appendix A

Additional Proofs and Results

In this section some additional details and proofs are provided.

Proposition A.1. *For any* $T, a, k > 0$ *we have* $\mathbb{P}(\int_0^T e^{-aB_s} ds < k) > 0$*, where* B_s *is a standard* $\mathbb{P}$*-Brownian motion process*[1].

Proof. Write $\int_0^T e^{-aB_s} ds = \int_0^u e^{-aB_s} ds + \int_u^T e^{-aB_s} ds$. Define $u := \frac{k}{2e^a}$ and condition on the events $A = \{B_s > -1; \forall s \in [0, u]\}$ and $C = \{B_s > M; \forall s \in [u, T]\}$, where M satisfies $e^{-aM} = \frac{k}{2(T-u)}$. Then

$$\mathbb{P}(\int_0^T e^{-aB_s} ds < k) \geq \mathbb{P}(\int_0^u e^{-aB_s} ds + \int_u^T e^{-aB_s} ds < k,\ A \cap C),$$

and A implies $\int_0^u e^{-aB_s} ds < ue^a = \frac{k}{2}$. Likewise B implies $\int_u^T e^{-aB_s} ds < (T-u)e^{-aM} = \frac{k}{2}$. By Borodin and Salminen (2002, formula 1.1.2.4)

$$\mathbb{P}_x(\inf_{0 \leq s \leq t} B_s > y) = 2\Phi(\frac{x-y}{\sqrt{t}}) - 1, \quad y \leq x,$$

where Φ is the cdf of the standard normal distribution. To see $\mathbb{P}(C \mid A) > 0$, condition further on those ω-paths where $B_u > M + \epsilon$ for any $\epsilon > 0$. Thus $\mathbb{P}(A \cap C) = \mathbb{P}(A)\mathbb{P}(C \mid A) > 0$. □

Fact A.2. *Let* B_s *be a* $\mathbb{P}$*-Brownian motion process. The distribution for* $\int_0^T e^{-as-bB_s} ds$,

[1] I thank Dr. Anthony Quas, University of Victoria, for assistance with this result.

for any $T \geq 0,\ a, b \in \mathbb{R},\ b \neq 0,\ a < b^2$ *is provided in Borodin and Salminen (2002, formula 2.1.10.4):*

$$\mathbb{P}(\int_0^T e^{-as-bB_s} ds \leq k) = \int_0^k \frac{|b| \left(\frac{b^2 y}{4}\right)^{-\frac{a}{b^2}}}{\sqrt{2y}} e^{-\frac{a^2 T}{2b^2} - \frac{1}{b^2 y}} m_{\frac{b^2 T}{2}} \left(-\frac{a}{b^2} - \frac{1}{2}, \frac{1}{b^2 y}\right) dy, \quad \text{(A.1)}$$

for any $k \geq 0$. *For* $Re(\mu) > -\frac{3}{2}$ *and* $Re(z) > 0$,

$$m_y(\mu, z) = \frac{8z^{\frac{3}{2}} \Gamma(\mu + \frac{3}{2}) e^{\frac{\pi^2}{4y}}}{\pi \sqrt{2\pi y}} \int_0^\infty e^{-z\cosh(2u) - \frac{u^2}{y}} M(-\mu, \frac{3}{2}, 2z \sinh^2(u)) \sinh(2u) \sin(\frac{\pi u}{y}) du$$

and Kummer's function (of the first kind) is defined as

$$M(a, b, x) := 1 + \sum_{k=1}^{\infty} \frac{a(a+1)\cdots(a+k-1)x^k}{b(b+1)\cdots(b+k-1)k!}.$$