Analysis of VIX-linked fee incentives in variable annuities via continuous-time Markov chain approximation

ANNE MACKAY †‡*, MARIE-CLAUDE VACHON‡ and ZHENYU CUI §

†Département de mathématiques, Université de Sherbrooke, Sherbrooke, Canada
‡Département de mathématiques, Université du Québec à Montréal, Montréal, Canada
§School of Business, Stevens Institute of Technology, Hoboken, NJ 07030, USA

(Received 8 August 2022; accepted 5 May 2023; published online 14 June 2023)

We consider the pricing of variable annuities (VAs) with general fee structures under a class of stochastic volatility models which includes the Heston, Hull-White, Scott, α-Hypergeometric, 3/2, and 4/2 models. In particular, we analyze the impact of different VIX-linked fee structures on the optimal surrender strategy of a VA contract with guaranteed minimum maturity benefit (GMMB). Under the assumption that the VA contract can be surrendered before maturity, the pricing of a VA contract corresponds to an optimal stopping problem with an unbounded, time-dependent, and discontinuous payoff function. We develop efficient algorithms for the pricing of VA contracts using a two-layer continuous-time Markov chain approximation for the fund value process. When the contract is kept until maturity and under a general fee structure, we show that the value of the contract can be approximated by a closed-form matrix expression. We also provide a quick and simple way to determine the value of early surrenders via a recursive algorithm and give an easy procedure to approximate the optimal surrender surface. We show numerically that the optimal surrender strategy is more robust to changes in the volatility of the account value when the fee is linked to the VIX index.

Keywords: Variable annuities; Optimal stopping; American options; Stochastic volatility; Heston model; Continuous-time Markov chain

JEL Classifications: C63, G12, G13, G22

1. Introduction

A variable annuity (or segregated fund in Canada) is a hybrid investment instrument mainly used for retirement planning, which offers a life insurance benefit and a financial guarantee. It allows the policyholder to profit from potential gains resulting from an investment in financial markets, while offering protection against losses. The real options embedded in these products are comparable to exotic options, with the following differences: the benefit may depend on the policyholder’s survival (or death), they are long-term investments (generally between 5 and 15 years, or more), and the financial guarantee is funded via a periodic fee (typically set as a percentage of the fund value) as opposed to a premium paid upfront. Different types of protection riders are offered, such as Guaranteed Minimum Maturity Benefit (GMMB), Guaranteed Minimum Death Benefit (GMDB), and Guaranteed Minimum Withdrawal Benefit (GMWB); see Hardy (2003) or Bauer et al. (2008) for details. This paper focuses on the GMMB rider, which guarantees the policyholder a minimum amount at the contract’s maturity. Considering the significant size of the variable annuity market, the management of the risk associated with the guarantees embedded in variable annuities is a major concern for insurance companies, see Niittuinperä (2022). Indeed, variable annuities guarantees

† There are no consensus among practitioners and scientists for these products’ name, and thus, different authors may use different terminologies for the same product.
‡ In United-States, the total variable annuity sales were $125 billion in 2021, representing an increase of 25% with respect to the total VA sales in 2020. Source: LIMRA Secure Retirement Institute, U.S. Individual Annuities survey https://www.limra.com/siteassets/newsroom/fact-tank/sales-data/2021/q4/2012-2021-annuity-sales-updated.pdf.

*Corresponding author. Email: Anne.MacKay@USherbrooke.ca
entail significant risks given their long-term structure and sensitivity to various financial and demographic risks as well as to policyholders’ behavior. For the GMMB rider, this last risk is mostly due to early surrenders.

Surrender risk refers to the uncertainty facing the insurer when a policyholder has the possibility to terminate her contract before its maturity. When she does so, she is entitled to the value accumulated in the variable annuity investment account, subject to a penalty. Kling et al. (2014) show that unexpected lapses can represent a significant risk for insurers. For this reason, surrender risk has raised special attention in the literature (Nitti, 2022, chapter 18). Bacicello et al. (2011) provide a universal pricing framework for various riders and considers different types of surrender behaviors: static, i.e. the contract is never surrendered; or mixed, i.e. the policyholder acts rationally and surrenders the policy as soon as it is optimal from a risk-neutral valuation perspective. Pricing variable annuities under rational surrender behavior is equivalent to solving an optimal stopping problem and corresponds to the worst-case scenarios for insurers, in the sense that it maximizes the risk-neutral value of the contract from the policyholders’ perspective. Under this assumption, Grosen and Jørgensen (1997) study the valuation of interest rate guarantees by assuming that the surrender value will be the same as the benefit value. Milevsky and Salisbury (2001) assume that the policyholder will only get a certain percentage of the fund upon surrender; this hypothesis is more in line with policies seen in practice. Under this assumption, they provide a closed-form analytical solution to the price of a GMDB with surrender in the Black-Scholes framework. In particular, they study the interaction between the surrender charges, the fee rates, and the optimal surrender level. Bernard et al. (2014) study a problem similar to the one of Milevsky and Salisbury (2001), but focus on a GMDB rather than a GMDB. It is well-known that American options with finite maturity generally do not have closed-form solutions. Thus, Bernard et al. (2014) used arbitrage-free techniques in the same vein as Kim (1990) and Carr et al. (1992) in the context of American call and put options to derive an analytical expression for the GMMB rider (without early surrender) under this type of fee structure. Mackay et al. (2017) study how the fee structure and surrender charges affect the surrender region; they also design surrender charges that eliminate surrender incentives for a financially rational policyholder. Other fee designs have been explored in the literature: Delong (2014) considers a general state-dependent fee structure in a Lévy process driven market, whereas Bernard and Moenig (2019) study lapse-and-reentry in variable annuities with time-dependent fee structure. Finally, in a recent study, Wang and Zou (2021) propose a stochastic control approach to determine the optimal fee structure.

Recently, fee structures that are tied to the Chicago Board Options Exchange (CBOE) volatility index, the VIX, have gained attention in the literature, see Cui et al. (2017) and Kouritzin and MacKay (2018). The motivations behind this new fee design come directly from the industry. In 2010, SunAmerica issued two new variable annuities whose fees were tied to the volatility index. More recently, America General Life Insurance Company proposed a fee structure that is linked to the VIX for its Polaris series of variable annuities, see Polaris Platinum O-Series prospectus dated May 3rd, 2021. By allowing the fees to move with the volatility index, the insurer expects to better match the cost of hedging with the premium collected. It also reduces fees for policyholders in low-volatility, rising market environments. The CBOE published two white papers, CBOE (2013a, 2013b), illustrating how VIX-linked fee designs can be advantageous to both variable annuity providers and policyholders. Cui et al. (2017) approach the question from a theoretical perspective by analyzing variable annuities without surrender with a fee structure that is tied to the VIX under a Heston-type stochastic volatility model. They provide a closed-form expression for the GMMB rider and observe that such a structure might help realign fee incomes with the value of the financial guarantee. Kouritzin and MacKay (2018) extend the works of Cui et al. (2017) by applying the VIX fee designs to a GMWB rider and by adding jumps to the underlying index value process.

In this work, we allow fee structures to be as general as possible, i.e. the fee structure may depend on time, the fund value, and also on the latent variance process, making it possible to link the fee to the VIX. In the constant fee case, it is well-known that the misalignment between the fees and the value of the financial guarantee creates an incentive for the risk-neutral, rational policyholder to surrender her policy early (see Milevsky and Salisbury 2001). Since VIX-linked fee structures allow for better alignment of the guaranteed value with the corresponding hedging cost, we expect that such fee designs can also help reduce the insurers’ exposure to surrender risk. For this reason, we numerically study the impact of three different VIX-linked fee designs on the optimal surrender strategy. To do so, we use a two-layer continuous-time Markov chain (CTMC) approximation for the fund dynamics

‡ See Retirement Income Journal available at https://retirement-incomejournal.com/article/sunamerica-links-va-rider-fees-to-volatility-index/.
§ See footnote 6 on p. 9 of the prospectus (the long-form) available at https://aigonlineprospectus.net/AIG/867018103A/index.php?open=POLARIS!5fPLATINUM!5fO-SERIES!5fISP.pdf.
inspired by Cui et al. (2018). Two-layer CTMC approximations have recently been used to price derivatives in stochastic volatility models, see Cui et al. (2018), Cui et al. (2019) and Ma et al. (2021), among others. The methodology proposed by Cui et al. (2018) for approximating two-dimensional diffusions is not only theoretically appealing and applies to most stochastic volatility models, but also is simple to implement for pricing European and American options. Their approach is especially efficient for a short/medium time horizon. However, for derivatives with very long maturities, such as those involved in variable annuities pricing, the methodology proposed by Cui et al. (2018) stretches the computing resource to unacceptable levels. In this paper, we adapt their method to long-maturity cases.

The main contributions of this paper are as follows:

- We extend the work of Cui et al. (2018), done in the context of options pricing, by providing novel efficient algorithms to value options with very long maturities, such as variable annuities, under general stochastic volatility models. More precisely, algorithms 1 and 3 are new to the CTMC literature and allow to accelerate the calculation time considerably. The convergence of the new methodology is also shown theoretically, see proposition 4.3. All the algorithms provided in this paper apply to a general class of stochastic volatility models and can be used for option pricing under other types of bi-dimensional models.

- We propose a methodology to approximate the optimal surrender surface of a VA contract with a GMMDM rider when the underlying index follows a two-dimensional diffusion process. Algorithms 4 and 5, which are based on the Bermudan approximation, can adapt easily to the context of American option pricing to approximate the exercise surface under general stochastic volatility models. This new way of approximating the exercise surface is novel to CTMC literature. The advantage of this method over the current one (see, for instance, Ma et al. 2021) is that the integral representation of the value function† does not need to be derived to obtain the surface approximation.

- The application of CTMC approximation for variable annuity pricing is also new to the literature. Moreover, in this paper, we analyze the optimal surrender strategy for a VA contract with a GMMDM rider under the assumption that the fees are linked to the VIX index. Previously, such an analysis of surrender incentives was performed using a constant fee structure, or in a Black-Scholes framework when the fees are state-dependent. To our knowledge, this is the first time that early surrenders are analyzed jointly with fees depending on the volatility index under a general class of stochastic volatility models.

- We derive a closed-form analytical expression for the VIX index when the variance process follows a continuous-time Markov chain. This expression can also be used to price VIX derivatives.

The remainder of the paper is organized as follows. In section 2, we introduce the market model, the VA contract, and the optimal stopping problem involved in the pricing of variable annuities with surrender. A brief introduction to CTMC approximations for a two-dimensional diffusion process is provided in section 3. In section 4, we apply CTMC approximations to VA contract pricing and provide new efficient algorithms. Section 5 provides the numerical results and discusses how VIX-linked fees affect surrender incentives. Section 6 concludes the paper.

2. Financial setting

2.1. Market model

We consider a filtered probability space \((\Omega, \mathcal{F}, \mathbb{F}, \mathbb{Q})\), where \(\mathbb{F}\) is a complete and right-continuous filtration and where \(\mathbb{Q}\) denotes the pricing measure for our market, see remark 2.1. We consider a risky asset, whose price can be described by the two-dimensional process \((S, V)\) satisfying

\[
\begin{align*}
\text{d}S_t &= r S_t \text{d}t + \sigma_S(V_t) S_t \text{d}W^{(1)}_t, \\
\text{d}V_t &= \mu_V(V_t) \text{d}t + \sigma_V(V_t) \text{d}W^{(2)}_t,
\end{align*}
\]

with \(S_0 = s_0 \in \mathbb{R}_+\) and \(V_0 = v_0 \in \mathbb{S}_V\) where \(\mathbb{S}_V\) denotes the state-space of \(V\) (usually \(\mathbb{R}\) or \(\mathbb{R}_+\) depending on the model, see table 1 for examples), with \(r > 0\) denoting the risk-free rate and with \(W = (W^{(1)}, W^{(2)})_{t \geq 0}\) a two-dimensional correlated Brownian motion with cross-variation \([W^{(1)}, W^{(2)}]_t = \rho t\), where \(\rho \in [-1, 1]\). For simplicity, \(V\) will be referred to as the variance process. We assume that \(\mu_V, \sigma_V : \mathbb{S}_V \mapsto \mathbb{R}\) are continuous and that \(\sigma_S : \mathbb{S}_V \mapsto \mathbb{R}_+\) and \(\sigma_V : \mathbb{S}_V \mapsto \mathbb{R}_+\) are continuously differentiable functions with \(\sigma_S(\cdot) > 0\) and \(\sigma_V(\cdot) > 0\) on the state-space \(\mathbb{S}_V\) of \(V\). Further, we suppose that \(\mu_V, \sigma_V\) and \(\alpha\) are defined such that (1) has a unique-in-law weak solution.

Remark 2.1 In (1), we start directly with the dynamics under the risk-neutral measure, hence the form of the market price of volatility risk is not necessary in our setting. However, as pointed out by Sin (1998), Jourdain (2004) and Cui (2013), a risk-neutral measure may not always exist under stochastic volatility models; additional conditions must be added to the model parameters in order for \(\mathbb{E}^\mathbb{Q}[e^{-\rho t} S_t]_{t \geq 0}\) to be a true martingale under \(\mathbb{Q}\). A list of common SV models is reported in table 1 below, along with conditions for the martingale property to hold under the risk-neutral measure.

2.2. Variable annuity contract

A policyholder enters a variable annuity contract by depositing an initial premium \(F_0\) into a sub-account, which is then invested in a fund tracking the financial market. For

†The integral representation of the value function may be challenging to obtain under general stochastic volatility models unless making some regularity assumptions on the value function as in Ma et al. (2021). Indeed, the smoothness of the value function can be difficult to show under such bi-dimensional models; see Terenzi (2018) and Lamberton and Terenzi (2019).
### Table 1. Examples of stochastic volatility models.

| Model name | Dynamics | Parameters | Cond. for martingale measure |
|------------|----------|------------|-----------------------------|
| Heston (1993) | $dS_t = rS_t dt + \sqrt{\nu_t} S_t dW_t^{(1)}$ | $S_0 > 0$, $\kappa > 0$, $\theta > 0$, $\sigma > 0$ | $\kappa \theta \geq \sigma^2$ |
| Hull and White (1987) | $dS_t = rS_t dt + \sigma_0 S_t dW_t^{(1)}$, $d\nu_t = \kappa (\theta - \nu_t) dt + \sigma \sqrt{\nu_t} dW_t^{(2)}$ | $S_0 > 0$, $\kappa > 0$, $\theta > 0$, $\sigma > 0$ | $\kappa \theta \geq \sigma^2$ |
| Scott (1976) | $dS_t = \sigma_0 S_t dW_t^{(1)}$, $d\nu_t = \kappa (\theta - \nu_t) dt + \sigma \sqrt{\nu_t} dW_t^{(2)}$ | $S_0 > 0$, $\kappa > 0$, $\theta > 0$, $\sigma > 0$ | $\kappa \theta \geq \sigma^2$ |
| Grasselli (2004) | $dS_t = \mu S_t dt + \sigma_0 S_t dW_t^{(1)}$, $d\nu_t = \kappa (\theta - \nu_t) dt + \sigma \sqrt{\nu_t} dW_t^{(2)}$ | $S_0 > 0$, $\kappa > 0$, $\theta > 0$, $\sigma > 0$ | $\kappa \theta \geq \sigma^2$ |
| Heston (1997) | $dS_t = \mu S_t dt + \sigma_0 S_t dW_t^{(1)}$, $d\nu_t = \kappa (\theta - \nu_t) dt + \sigma \sqrt{\nu_t} dW_t^{(2)}$, $d\nu_t = \nu_t dt$ | $S_0 > 0$, $\kappa > 0$, $\theta > 0$, $\sigma > 0$ | $\kappa \theta \geq \sigma^2$ |
| Hall and White (2003) | $dS_t = \mu S_t dt + \sigma_0 S_t dW_t^{(1)}$, $d\nu_t = \kappa (\theta - \nu_t) dt + \sigma \sqrt{\nu_t} dW_t^{(2)}$, $d\nu_t = \nu_t dt$ | $S_0 > 0$, $\kappa > 0$, $\theta > 0$, $\sigma > 0$ | $\kappa \theta \geq \sigma^2$ |
| Jourdain (2004) | $dS_t = \mu S_t dt + \sigma_0 S_t dW_t^{(1)}$, $d\nu_t = \kappa (\theta - \nu_t) dt + \sigma \sqrt{\nu_t} dW_t^{(2)}$, $d\nu_t = \nu_t dt$ | $S_0 > 0$, $\kappa > 0$, $\theta > 0$, $\sigma > 0$ | $\kappa \theta \geq \sigma^2$ |
| Da Fonseca and Martini (2016) | $dS_t = \mu S_t dt + \sigma_0 S_t dW_t^{(1)}$, $d\nu_t = \kappa (\theta - \nu_t) dt + \sigma \sqrt{\nu_t} dW_t^{(2)}$, $d\nu_t = \nu_t dt$ | $S_0 > 0$, $\kappa > 0$, $\theta > 0$, $\sigma > 0$ | $\kappa \theta \geq \sigma^2$ |



### Footnotes:
1. The condition stated in Cui et al. (2013, proposition 2.5.4) is automatically satisfied by requiring the correlation parameter $\rho$ to be non-positive, as pointed out by Dumas (2012).

For simplicity, we will assume that the sub-account is invested in the risky asset $S$. The policyholder often has the right to surrender the contract, or lapse, prior to maturity. This additional flexibility is often called surrender option (or surrender right) in the literature and significantly complicates the valuation of variable annuity contracts. Below we discuss the risk-neutral valuation approach for a variable annuity, under both the assumption that the policyholder makes use of her surrender right or does not.

To do so, we consider a finite time horizon $T \in \mathbb{R}_+$ and let $F = \{F_t\}_{0 \leq t \leq T}$ denote the variable annuity fund (or sub-account) process. Moreover, we let $C : [0, T] \times \mathbb{R}_+ \times \mathcal{S}_V \to \mathbb{R}_+$ denote the fee function and let the continuously compounded fee rate process $(c_t)_{0 \leq t \leq T}$ be defined as

$$
c_t := C(t, F_t, V_t), \quad 0 \leq t \leq T,
$$

where $C$ is assumed to be continuous or bounded and such that (4) has a unique-in-law weak solution. We allow the fee structure to be as general as possible. This setting includes, among others, state-dependent fee structures (see Bernard et al. 2014, Delong 2014, Mackay et al. 2017), VIX-linked fee structures (see Cui et al. 2014, Delong 2014, Mackay et al. 2017), and time-dependent fee structures (see Bernard and Moenig 2019).

We assume that the fees are paid continuously out of the fund at a rate $c_t$, so that the fund value is given by

$$
F_t = S_0 e^{-\int_0^t c_u du}, \quad 0 \leq t \leq T,
$$

with $F_0 = S_0$. Using Itô’s lemma, the dynamics of $F$ under the risk-neutral measure are

$$
dF_t = (r - c_t) F_t dt + \sigma_0 (V_t) F_t dW_t^{(1)},
\quad dV_t = \mu (V_t) dt + \sigma_0 (V_t) dt dW_t^{(2)}. \tag{4}
$$

For simplicity in this paper, we assume that interest rates are constant. However, given the long-term maturity of variable annuity contracts, it may be interesting to allow interest rates to be a deterministic function of time. Such an extension, discussed in appendix B (available online as supplemental material), allows for the exact replication of the term structure of interest rates.

Throughout this paper, $E_{r,y}^x[\cdot]$ is short-hand notation for $E[\cdot | F_t = x, V_t = y]$ and $E_{r}^x[\cdot]$ for $E[\cdot | F_t]$, with $x \in \mathbb{R}_+$, $y \in \mathcal{S}_V$ and $t \in [0, T]$. We also use $E_{r,y}^x[\cdot]$ to denote $E_{0,r,y}^x[\cdot]$.

We focus on a variable annuity with a guaranteed minimum maturity benefit (GMMB) whose payoff at maturity $T$ is $\max(G, F_T)$, where $G \in \mathbb{R}_+$ is a predetermined guaranteed amount. Given $(F_t, V_t) = (x, y)$, the time-$t$ risk-neutral value of the variable annuity assuming that it will not be surrendered early is

$$
v_x(t, x, y) := E_{t,x,y}^r \left[ e^{-r(T-t)} \max(G, F_T) \right]. \tag{5}
$$

On early surrender, the policyholder receives the value of the VA sub-account, reduced by a penalty which, in our setting,
can depend on time and on the value of the variance process \( V \). When no surrender occurs, the maturity benefit is paid at \( T \).

More formally, the VA reward (or gain) function \( \varphi : [0, T] \times \mathbb{R}_+ \times \mathbb{R}_+ \rightarrow \mathbb{R}_+ \) is defined by

\[
\varphi(t, x, y) = \begin{cases} 
g(t, y)x & \text{if } t < T, \\ 
\max(G(x)) & \text{if } t = T,
\end{cases}
\]

where \( g : [0, T] \times \mathbb{R}_+ \rightarrow [0, 1] \) is continuous, non-decreasing in time and satisfies \( \lim_{t \to T^-} g(t, y) = 1 \) \( \forall y \in \mathbb{R}_+ \). In practice, we usually consider the surrender charge (as a percentage of the account value), \( \bar{g} = 1 - g(\cdot, \cdot) \). A common form for the surrender charge function in the literature is \( g(t, y) = e^{-k(T-t)} \) for some constant \( k \geq 0 \), see for example Shen et al. (2016), Mackay et al. (2017) and Kang and Ziveyi (2018). It is the first time, to the best of our knowledge, that variance-dependent surrender charges are considered.

**Remark 2.2** For \( x < G \), the function \( t \mapsto \varphi(t, x, y) \) is discontinuous at \( T \) since

\[
\lim_{t \to T^-} \varphi(t, x, y) = g(t, y)x < x < G = \varphi(T, x, y).
\]

Under the assumption that the policyholder maximizes the risk-neutral value of her VA contract, the time-\( t \) value of the variable annuity policy is given by

\[
v(t, x, y) = \sup_{\tau \in \mathcal{T}_T} \mathbb{E}_{t,x,y}[e^{-r(T-t)} \varphi(t, F_T, V_T)],
\]

where \( \mathcal{T}_T \) is the (admissible) set of all stopping times taking values in the interval \([t, T]\).

Similar to the early exercise premium in the American option literature, the value of the right to surrender, denoted by \( e : [0, T] \times \mathbb{R}_+ \times \mathbb{R}_+ \rightarrow \mathbb{R}_+ \), is defined by

\[
e(t, x, y) := v(t, x, y) - v_e(t, x, y).
\]

### 3. Continuous-time Markov chain approximation

The CTMC framework outlined in this section has been proposed by Cui et al. (2018) for exotic option pricing under stochastic local volatility models. The general idea is to approximate the two-dimensional stock price process by a two-dimensional continuous-time Markov chain. This is done by first approximating the variance process by a CTMC, and then by replacing the variance process by its CTMC approximation in the underlying price process. The resulting regime-switching diffusion process is then further approximated by a CTMC, yielding a two-dimensional CTMC process which converges weakly to the original two-dimensional diffusion process, providing that the generator of the CTMC is chosen correctly.

First, we shall briefly recall the basics of continuous-time Markov chains, following sections 6.9 and 6.10 of Grimmett and Stirzaker (2001). The stochastic process \( \bar{X} = [\bar{X}_t]_{t \geq 0} \) taking values on some discrete state-space \( \mathcal{S}_X \) is called a continuous-time Markov chain if it satisfies the following property (a.k.a. Markov property):

\[
\mathbb{P}(\bar{X}_{t_n} = \bar{x}_i | \bar{X}_{t_{n-1}} = \bar{x}_{i-1}, \ldots, \bar{X}_{t_1} = \bar{x}_1) = \mathbb{P}(\bar{X}_{t} = \bar{x}_i | \bar{X}_{t_{n-1}} = \bar{x}_{i-1})
\]

for all \( \bar{x}_i, \bar{x}_1, \ldots, \bar{x}_{i-1} \in \mathcal{S}_X \) and any time sequence \( t_1 < t_2 < \ldots < t_n \). For \( 0 \leq s \leq t \), we denote the transition probability from state \( \bar{x}_i \in \mathcal{S}_X \) at time \( s \) to state \( \bar{x}_j \in \mathcal{S}_X \) at time \( t \) by \( p_{ij}(s, t) = \mathbb{P}(\bar{X}_s = \bar{x}_j | \bar{X}_s = \bar{x}_i) \). The chain is said to be homogeneous if \( p_{ij}(s, t) = p_{ij}(0, t-s) \) for any \( i, j, s \leq t \). In that case, we use \( p_{ij}(t-s) \) to denote \( p_{ij}(s, t) \).

Going forward, we assume that \( \bar{X} \) is time-homogeneous and \( \mathcal{S}_X \) is finite. The family \( \{P_t := [p_{ij}(t)]_{S_X \times |S_X|}, t \geq 0\} \) of transition probability matrices is referred as the transition semigroup of the Markov Chain.

For an infinitesimal period of length \( h > 0 \), it can be shown that there exist constants \( |q_{ij}|_{1 \leq i, j \leq |S_X|} \), also called transition rates, such that

\[
p_{ij}(h) = \begin{cases} 
q_{ij}h + c(h) & \text{if } i \neq j, \\
1 + q_{ij}h + c(h) & \text{if } i = j,
\end{cases}
\]

where \( c \) is a function satisfying \( \lim_{h \to 0} c(h) = 0 \). From the above, we can conclude that the transition rates must satisfy

\[
\begin{align*}
q_{ij} & \geq 0, & \text{if } i \neq j, \\
q_{ij} & \leq 0, & \text{if } i = j,
\end{align*}
\]

and

\[
\sum_{j=1}^{m} q_{ij} = 0, \quad i = 1, 2, \ldots, m.
\]

The matrix \( Q := [q_{ij}]_{|S_X| \times |S_X|} \) is called the generator of \( \bar{X} \). Under some technical conditions,\(^\dagger\) it can be shown that the transition probability matrix \( P_t \) has the following matrix exponential representation:

\[
P_t = \exp(Qt) = \sum_{k=0}^{\infty} \frac{(Qt)^k}{k!}.
\]

**Assumption 3.1** The fee function defined in (2) is time-independent and denoted by \( c \). That is, \( C(t, x, y) = c(x, y) \) for all \( 0 \leq t \leq T \). Moreover, we only consider functions \( c \) that are continuous or bounded.

Henceforth, we consider that assumption 3.1 holds. That is, we assume that the fee function is time-independent so that the fund process is time-homogeneous. For the CTMC approximation of diffusion processes with time-dependent coefficients, see Ding and Ning (2021).

\(^\dagger\)More precisely, the semigroup \( \{P_t\} \) must be standard – that is, \( p_{ij}(t) \to 1 \) and \( p_{ji}(t) \to 0 \) as \( t \downarrow 0 \) – and uniform – \( \sup_{ij} q_{ij} < \infty \), see Grimmett and Stirzaker (2001), definitions 6.9.4 and 6.10.3, theorems 6.10.1, 6.10.5 and 6.10.6 for details.
3.1. Approximation of the variance process \( \{V_t\}_{t\geq 0} \)

We construct a CTMC \( \{V_t^{(m)}\}_{t\geq 0} \) taking values on a finite state-space \( S_t^{(m)} := \{v_1, v_2, \ldots, v_m\} \), with \( v_i \in \mathcal{S}_V \) and \( m \in \mathbb{N} \), that converges weakly to \( \{V_t\}_{t\geq 0} \) as \( m \to \infty \). Weak convergence of \( V_t^{(m)} \) to \( V_t \) is denoted by \( V_t^{(m)} \to V_t \).

Several approaches are available in the literature to construct the finite state-space \( S_t^{(m)} \), from simple uniform to non-uniform grids (see Tavella and Randall 2000, Mijatović and Pistorius 2013, Lo and Skindilias 2014, for examples of non-uniform grids). The specific grid selected for the numerical analysis performed in this paper is discussed in more details in section 5.

Once the state-space is chosen, the approximating CTMC \( \{V_t^{(m)}\}_{t\geq 0} \) is defined via its generator \( Q^{(m)} = \{q_{ij}\}_{m \times m} \). This generator is constructed so that the first two moments of the transition probability \( \{V_t\}_{t\geq 0} \) and of the approximating CTMC \( \{V_t^{(m)}\}_{t\geq 0} \) coincide; these are the so-called local consistency conditions, see Kushner (1990) and Lo and Skindilias (2014). More precisely, the elements \( q_{ij} \)

\[ 1 \leq i, j \leq m \] of the generator of \( V_t^{(m)} \) are chosen so that for a small time increment \( \delta_t \leq \delta \),

\[
\begin{align*}
\mathbb{E}_t \left[ V_{t+\delta} - V_t \right] & \leq \mathbb{E}_t \left[ V_{t+\delta} - V_t \right] \succeq \mu_V(V_t)\delta_t, \quad \text{and} \\
\mathbb{E}_t \left[ (V_{t+\delta} - V_t)^2 \right] & \geq \mathbb{E}_t \left[ (V_{t+\delta} - V_t)^2 \right] \succeq \sigma_V^2(V_t)\delta_t,
\end{align*}
\]

for all \( t \geq 0 \). To ensure that the local consistency conditions are satisfied, we use the generator proposed by Lo and Skindilias (2014) and given by\footnote{An advised reader will notice some differences between the transition rates stated above, and the ones that appear in Lo and Skindilias (2014). However, one can show that the two rate matrices are equivalent with some simple algebra.}

\[
q_{ij} = \begin{cases} \sigma_V^2(v_j) - \delta_t \mu_V(v_j), & j = i - 1, \\ \delta_t \mu_V(v_i) - q_{i,j} - q_{i,j+1}, & j = i, \\ \sigma_V^2(v_i) + \delta_t \mu_V(v_i), & j = i + 1, \\ \delta_t \mu_V(v_i), & j \neq i, i = 1, i + 1, \\ 0, & \end{cases}
\]

for \( 2 \leq i \leq m - 1 \) and \( 1 \leq j \leq m \) where \( \delta_t = \min_{1 \leq i \leq m-1} \delta_t \). On the borders, we set \( q_{12} = \frac{\mu_V(v_1)}{\delta_t} \), \( q_{12} = -\delta_t, q_{m-1,m} = \frac{\mu_V(v_m)}{\delta_t}, q_{m,m} = -\delta_t, q_{i,j+1}; \) and 0 elsewhere. The transition rates on the boundaries of the state-space are set so that the absolute instantaneous means are maintained at the endpoints. Other schemes could have also been employed (see Chourdakis 2004, Mijatović and Pistorius 2013), but we observed that all of these schemes are equivalent numerically.

To obtain a well-defined \( Q^{(m)} \) matrix, the transition rates in (13) must also satisfy the conditions in (9). Hence, for \( 2 \leq i \leq m - 1 \), we must have

\[
\begin{cases} \delta_t \mu_V(v_i) \leq \frac{\sigma_V^2(v_i)}{\mu_V(v_i)}, & \text{if } \mu_V(v_i) < 0 \\ \delta_t \mu_V(v_i) \geq \frac{\sigma_V^2(v_i)}{\mu_V(v_i)}, & \text{if } \mu_V(v_i) > 0. \end{cases}
\]

If \( \mu_V(v_i) = 0 \), then no additional condition needs to be added.

Remark 3.2 A sufficient condition for (14) to hold is

\[
\max_{1 \leq i \leq m-1} \delta_t \leq \min_{1 \leq i \leq m-1} \frac{\sigma_V^2(v_i)}{\mu_V(v_i)}. \tag{15}
\]

3.2. Approximation of the fund value process \( \{F_t\}_{t\geq 0} \)

The CTMC approximating \( \{F_t\}_{t\geq 0} \) is constructed by first replacing the variance process appearing in the drift and diffusion coefficients by their CTMC approximations, and then by further approximating the resulting regime-switching diffusion process by another CTMC. The resulting two-dimensional regime-switching CTMC can then be mapped to a one-dimensional CTMC on an enlarged state-space.

Lemma 3.3 below allows for the removal of the correlation between the Brownian motions in (4), which is necessary to construct the CTMC approximation of \( \{F_t\}_{t\geq 0} \).

Lemma 3.3 (lemma 1 of Cui et al. (2018)) Let \( F \) and \( V \) be defined as in (4). Define \( \gamma(x) := \int_x^\infty \frac{\sigma_2(y)}{\sigma_1(y)} \mathrm{d}u \) and \( X_t := \ln(F_t - \rho V_t), \ t \in [0,T] \). Then \( X_t \) satisfies

\[
\begin{align*}
dX_t &= \mu_X(X_t, V_t) \, \mathrm{d}t + \sigma_X(X_t, V_t) \, \mathrm{d}W_t^*, \\
\sigma_X(X_t, V_t) W_t^* &= \sqrt{1 - \rho^2} \sigma_2(Y_t) \, \frac{\sigma_2(y)}{\sigma_1(y)} \, \mathrm{d}u \\
\psi(Y_t) := \frac{\psi(Y_t)}{\sigma_1(Y_t)} &= \mu_Y(Y_t) Y_t \, \frac{1}{2} \sigma_2^2(Y_t) \, \frac{\sigma_2(y)}{\sigma_1(y)} \, \mathrm{d}u \\
\psi(Y_t) &= \psi(Y_t) + \frac{1}{2} \left[ \sigma_Y(Y_t) \sigma_2(Y_t) - \sigma_Y(Y_t) \sigma_2(Y_t) \right] \\
&= \mu_Y(Y_t) \sigma_2^2(Y_t) \, \frac{\sigma_2(y)}{\sigma_1(y)} \, \mathrm{d}u + \frac{1}{2} \sigma_2^2(Y_t) \, \frac{\sigma_2(y)}{\sigma_1(y)} \, \mathrm{d}u,
\end{align*}
\]

for \( x \in \mathbb{R}, y \in \mathcal{S}_V \). The proof relies on the multidimensional Itô formula (see Lemma 1 of Cui et al. (2018) for details). Given the CTMC approximation of the process \( V^{(m)} \) and its generator \( Q^{(m)} \), the diffusion process in (16) can now be approximated by a regime-switching diffusion process \( \{X_t^{(m)}\}_{t \geq 0} \):

\[
dX_t^{(m)} = \mu_X(X_t^{(m)}, V_t^{(m)}) \, \mathrm{d}t + \sigma_X(X_t^{(m)}, V_t^{(m)}) \, \mathrm{d}W_t^*, \tag{17}
\]

where regimes are determined by the states of the approximated variance process, \( \{v_1, v_2, \ldots, v_m\} \). To construct a regime-switching CTMC \( \{X_t^{(m,N)}, V_t^{(m)}\} \) approximating the regime-switching diffusion \( \{X_t^{(m)}, V_t^{(m)}\} \), we fix a state for the variance process \( V_t^{(m)} \) (or equivalently a regime) and construct a CTMC approximation for \( X_t^{(m)} \) given that \( V_t^{(m)} \) is in that state. This is done using the procedure described in section 3.1 for a one-dimensional diffusion process. The procedure is then repeated for each state in \( S_{V_t}^{(m)} \), and the approximating CTMCs are combined with \( V^{(m)} \) to obtain the final regime-switching CTMC.

More precisely, let \( X_t^{(m,N)} \) be the CTMC approximation of \( X_t^{(m)} \) taking values on a finite state-space \( S_{X_t}^{(m,N)} = \{x_1, x_2, \ldots, x_N\}, N \in \mathbb{N} \). For each \( v_i \in S_{V_t}^{(m)} \), we define the
generator $G_{i}^{(N)} = [\lambda_{ij}^{(N)}]_{N \times N}$ of $X_{t}^{(N)}$, given that the variance process is in state $v_{i}$ at time $t \geq 0$ by

$$
\lambda_{ij}^{(N)} = \begin{cases}
\frac{\sigma_{k}^{2}(v_{i}) - \delta_{ij}^{0} \mu_{X}(x_{i}, v_{i})}{\lambda_{i,j}^{0} - \lambda_{i,j+1}} & j = i - 1 \\
-\lambda_{ij}^{0} & j = i \\
\frac{\sigma_{k}^{2}(v_{i}) + \delta_{ij}^{0} \mu_{X}(x_{i}, v_{i})}{\lambda_{i,j-1}^{0}} & j = i + 1 \\
0 & j \neq i, i - 1, i + 1,
\end{cases}
$$

for $2 \leq i \leq N - 1$ and $1 \leq j \leq N$, where $\delta_{ij} = x_{i+1} - x_{i}$, $i = 1, 2, \ldots, N - 1$. On the boundaries, we set $\lambda_{12}^{(N)} = \frac{\mu_{X}(x_{1}, v_{0})}{R_{1}^{0}}$, $\lambda_{N,N-1}^{(N)} = -\lambda_{N,N}^{(N)} = -\lambda_{N,N-1}^{(N)}$ and $0$ elsewhere.

Using $V^{(m)}$ and the relation presented in lemma 3.3, the approximated fund process $F^{(m,N)}$, which approximates $F$, is defined by

$$
F_{t}^{(m,N)} := \exp \left\{ X_{t}^{(m)} + \rho \gamma(V_{t}) \right\}, \quad 0 \leq t \leq T.
$$

**Remark 3.4 (Convergence of the approximation)** Such a construction of the regime-switching CTMC ensures that the two-dimensional process $(X_{t}^{(m,N)}, V_{t}^{(m)})$ converges weakly to $(X_{t}, V_{t})$ as $m, N \to \infty$. The main idea is to show that the generator of $(X_{t}^{(m,N)}, V_{t}^{(m)})$ is uniformly close to the infinitesimal generator of $(X_{t}, V_{t})$ as $m, N \to \infty$, to then conclude that $(X_{t}^{(m,N)}, V_{t}^{(m)}) \Rightarrow (X_{t}, V_{t})$ using the results of Ether and Kurtz (2005) which relies on semi-group theory. Moreover, since the function $h : \mathbb{R} \to \mathbb{S}_{V} \to \mathbb{R}$, defined by $h(x, y) = e^{\gamma V_{t}}(x, y)$ is continuous, we have that $F^{(m,N)} \Rightarrow F$ by the continuous mapping Theorem, see Billingsley (1999, theorem 2.7).

For one-dimensional processes, intuition and detailed explanations of the proof can be found in Mijatović and Pistorius (2013, section 5) (or in the unabridged version of the paper Mijatović and Pistorius 2009, section 6); for stochastic volatility models, see Cui (2018, section 2.4).

The last step is to convert the regime-switching CTMC $(X_{t}^{(m,N)}, V_{t}^{(m)})$ into a one-dimensional CTMC process $Y_{t}$ on an enlarged state-space $S_{X}^{(m,N)} := \{1, 2, \ldots, m N\}$. This is done in theorems 1 of Cui et al. (2019), below.

**Proposition 3.5 (Theorem 1 of Cui et al. (2019))** Consider a regime-switching CTMC $(X_{t}^{(m,N)}, V_{t}^{(m)})$ taking values in $S_{X}^{(m,N)} \times S_{V}^{(m)}$ where $S_{X}^{(m,N)} := \{x_{1}, x_{2}, \ldots, x_{m}\}$ and $S_{V}^{(m)} := \{v_{1}, v_{2}, \ldots, v_{m}\}$; and another one-dimensional CTMC, $(Y_{t}^{(m,N)}, t \geq T)$, taking values in $S_{Y}^{(m,N)} := \{1, 2, \ldots, m N\}$ and its transition rate matrix $Q^{(m,N)}$ given by

$$
\begin{pmatrix}
q_{11}^{(N)} & q_{12}^{(N)} & \cdots & q_{1m}^{(N)} \\
q_{21}^{(N)} & q_{22}^{(N)} & \cdots & q_{2m}^{(N)} \\
\vdots & \vdots & \ddots & \vdots \\
q_{m1}^{(N)} & q_{m2}^{(N)} & \cdots & q_{mm}^{(N)} + G_{m}
\end{pmatrix},
$$

where $I_{N}$ is the $N \times N$ identity matrix, $G_{l}^{(N)} = [\lambda_{ij}^{(N)}]_{N \times N}$, $l = 1, 2, \ldots, m$ and $Q^{(m)} = [\phi_{ij}^{(m)}]_{m \times m}$ are the generators defined in (18) and (13), respectively. Define the function $\psi : S_{X}^{(m,N)} \times S_{V}^{(m)} \to S_{Y}^{(m,N)}$ by $\psi(x_{i}, v_{j}) = (l - 1)N + n$ and its inverse $\psi^{-1} : S_{Y}^{(m,N)} \to S_{X}^{(m,N)} \times S_{V}^{(m)}$ by $\psi^{-1}(n, v) = (x_{i}, v_{j})$ for $n, v_{j} \in S_{Y}^{(m,N)}$, where $n \leq N$ is the unique integer such that $n_{i} = (l - 1)N + n$ for some $l \in \{1, 2, \ldots, m\}$. Then, we have

$$
E \left[ \Psi(X_{t}^{(m,N)}, V_{t}^{(m)}) \big| X_{0}^{(m,N)} = x_{i}, V_{0}^{(m)} = v_{j} \right] = E \left[ \Psi^{-1}(\Psi^{-1}(X_{t}^{(m,N)})) \big| Y_{0} = (k - 1)N + 1 \right].
$$

for any path-dependent function $\Psi$ such that the expectation on the left-hand side is finite.

### 4. Variable annuity pricing via CTMC approximation

In this section, we use the CTMC approximation of the fund value process to price variable annuities under different surrender strategies. We provide a simple way to approximate the optimal surrender surface, which is the extension in three dimensions of the exercise boundary for two-dimensional processes. The present section extends to variable annuity pricing the work of Cui et al. (2018) and Cui et al. (2019), in which a CTMC approximation is used for option pricing.

Recall that $(X_{t}^{(m,N)}, V_{t}^{(m)})$ is the regime-switching CTMC approximation of $(X_{t}, V_{t})$ (see lemma 3.3) taking values in a finite state-space $S_{X}^{(m,N)} \times S_{V}^{(m)}$ where $S_{X}^{(m,N)} = [x_{1}, x_{2}, \ldots, x_{m}]$, $S_{V}^{(m)} = [v_{1}, v_{2}, \ldots, v_{m}]$, $N \in \mathbb{N}$; and $Y_{t}^{(m,N)} = [y_{1}, y_{2}, \ldots, y_{m}]$, $m \in \mathbb{N}$. We have also defined $F^{(m,N)}$ in terms of $(X_{t}^{(m,N)}, V_{t}^{(m)})$ in (19).

Throughout this section, we denote by $[e_{k}^{(m,N)}]_{k=1}^{m}$ the standard basis in $\mathbb{R}^{m,N}$, i.e. $e_{k}$ represents a row vector of size $1 \times mN$ having a value of $1$ in the $(k - 1)N + i$-th entry and $0$ elsewhere.

#### 4.1. Variable annuity without early surrenders

Consider an initial premium $F_{0} > 0$ and let $V_{0} \in \mathbb{S}_{V}$. The risk-neutral value of a variable annuity contract assuming no early surrenders can be approximated by

$$
v_{c}(F_{0}, V_{0}) = E \left[ e^{-rT} \max(G, F_{T}) | F_{0}, V_{0} \right] = e^{-rT} \max(G, F_{T}^{(m,N)}) | F_{0}, V_{0} = v_{c}.
$$

Here, we assume $x_{i} \in S_{X}^{(m,N)}$ and $v_{k} \in S_{V}^{(m)}$, with $x_{i} = \ln(F_{0}) - \rho \gamma(V_{0})$.

**Proposition 4.1** Let $F_{0} > 0$ be the initial premium, with $x_{0}^{(m,N)} = \ln(F_{0}) - \rho \gamma(V_{0}) = x_{i} \in S_{X}^{(m,N)}$ and $0 = v_{k} \in \mathbb{S}_{V}$. The risk-neutral value at time $0$ of a variable annuity contract held until maturity $T$ and with guaranteed amount $G > 0$ can be approximated by

$$
v_{c}^{(m,N)}(F_{0}, V_{0}) = E \left[ e^{-rT} \max(G, F_{T}^{(m,N)}) | F_{0} = F_{0}, V_{0} = V_{0} \right] = e^{-rT} \max(G, F_{T}^{(m,N)}) | F_{0} = F_{0}, V_{0} = V_{0}.
$$

If $X_{t}^{(m,N)}$ and $V_{t}^{(m)}$ are not part of their respective grids, then the two points can be added to the grids, or the option price must be linearly interpolated between grid points, see remark 5.1 for details.
where $G^{(m,N)}$ is defined in (20) and $H$ is a column vector of size $mN \times 1$ whose $(l-1)N + n$-th entry is given by

$$h_{(l-1)N+n} = \max (G, e^{h_{t+\rho}(\psi)}) , \quad 1 \leq l \leq m, 1 \leq n \leq N. \tag{23}$$

The proof follows by noticing that (22) is the matrix representation of the conditional expectation of a (function of $a$) discrete one-dimensional random variable whose conditional probability mass function is given by the transitional probabilities $p_{(k-1)N+j}(T), 1 \leq j \leq mN$, with $P(T) = \{p_{(0)}(T)\}_{k \in \mathbb{N} \times \mathbb{N}} = \exp([TG]^{(m)})$ as per (11).

**Remark 4.2** (Convergence of VA prices without early surrenders) From remark 3.4, we know that $F^{(m,N)} \Rightarrow F$, as $m, N \rightarrow \infty$. From the continuous mapping theorem, we also have that $\psi(T, F^{(m,N)}, V^{(m)}) \Rightarrow \psi(T, F, V_T)$. For derivatives with a continuous and bounded payoff, convergence of the prices follows directly from the definition of weak convergence, see for example Billingsley (1995, theorem 25.8, (i) and (ii)). However, convergence of the prices for derivatives with unbounded payoffs is not as clear. We note that if $F^{(m,N)}$ is a $\mathbb{Q}$-martingale for each $m, N$, then $\{\psi(T, F^{(m,N)}, V^{(m)})\}_m$ is uniformly integrable and the convergence of the prices follows. Showing this property is however out of the scope of this paper. In the one-dimensional setting, convergence of the price approximations is discussed in Zhang and Li (2021). Detailed error and convergence analysis of the two-dimensional CTMC approximation for European call options is performed in Ma et al. (2022).

### 4.1.1. Fast algorithm

When considering a typical time horizon of 10, 15 or 20 years as is often the case in VA pricing, the probability that the fund or the volatility processes reaches high (resp. low) value is higher than when shorter maturities are concerned, and so the grid’s upper (resp. lower) bound must be set to a higher (resp. lower) value. This complicates the pricing of VAs compared to financial options (generally written for short or medium time-horizon), since more discretization points $m$ and $N$ are needed in order to capture the distribution of the variance and the fund value process correctly. Theoretically, this is not a problem; however, several numerical issues can occur when implementing the pricing formula numerically. First, the generator matrix $G^{(m,N)}$ can become very large and thus require a large amount of storage space, which may cause memory problems. Second, calculating the exponential of a large sparse $mN \times mN$ matrix over a long time horizon is time-consuming.

Using the tower property of conditional expectations and an approximation based on the assumption that the variance process remains constant over small time periods (see proposition 4.3), we propose a new algorithm that speeds up the pricing of the VA contract without early surrenders (see algorithm 1).

The approximation used in algorithm 1 follows from proposition 4.3 presented below.

**Proposition 4.3** Let $h > 0$ with $h \ll T$ and $0 \leq t \leq T - h$. For any function $\phi$ such that the expectation on the left-hand side of (24) is finite, we have that

$$\mathbb{E} \left[ \phi \left( t + h, X^{(m,N)}_{t+\rho h}, V^{(m)}_{t+\rho h} \right) | X^{(m,N)}_t = x^i, V^{(m)}_t = v^j \right] = \sum_{j=1}^m \mathbb{E} \left[ \phi \left( t + h, X^{(m,N)}_{t+\rho h}, V^{(m)}_{t+\rho h} \right) | V^{(m)}_{t+\rho h} = v^j, \right] X^{(m,N)}_t = x^i \right], \tag{24}$$

where $c(h)$ is a function satisfying $\lim_{h \rightarrow 0} \frac{c(h)}{h} = 0$.

The proof of proposition 4.3 is reported in appendix A.1. The above proposition allows the separation of the matrices $[G^{(N)}_{N}]_{j=1}^N$ and $Q^{(m)}$, so that the new algorithm now requires $m$ times the calculation of the exponential of $N \times N$ matrices and one time the exponential of a $m \times m$ matrix over small time-intervals. Hence, by reducing the size of the matrix in the matrix exponential and the length of the time interval over which the exponential is calculated leads to a significant reduction in computation time and to more effective management of the memory space. Note also that the added cost of computing $m + 1$ matrix exponentials (rather than only one) is counterbalanced by the reduced cost of computing the exponential of smaller matrices over a short time interval $h$.

The following notation is used in algorithm 1 below.

1. We use $M \in \mathbb{N}$ time steps of length $\Delta M = T/M$.
2. $B = [b_{m,n}]_{m,n=1}^N$ denotes a matrix of size $m \times N$, containing the value of the VA contract. More precisely, the matrix $B$ is updated at each time step, so that after the first iteration, $b_{m,n} \approx \mathbb{E} [e^{-\Delta M} \psi(T, e^{\Delta M \gamma} X^{(m,N)}_T, V^{(m)}_T) | X^{(m,N)}_T = x^i, V^{(m)}_T = v^j]$; after the second iteration, $b_{m,n} \approx \mathbb{E} [e^{-2\Delta M} \psi(T, e^{2\Delta M \gamma} X^{(m,N)}_T, V^{(m)}_T) | X^{(m,N)}_T = x^i, V^{(m)}_T = v^j]$; and so on.
3. $B_{n \cdot} = \{b_{m,n}\}_{m=1}^N$ denotes the $n$-th column of $B$, $n = 1, 2, \ldots, N$.
4. $B_{\cdot j}$ denotes the $j$-th row of $B$, $j = 1, 2, \ldots, m$.
5. The symbol $\top$ indicates the matrix (vector) transpose operation.

The computational gain of using algorithm 1 over the previous algorithm comes at the cost of a loss of accuracy since the conditional expectations are approximated over small time intervals (refer to proposition 4.3). Numerical experiments below show that highly accurate results are obtained in seconds when the time step is small, but the algorithm can perform poorly when the time step is not small enough. This new algorithm can be easily adapted to a wide range of payoff functions and extends the work of Cui et al. (2018) by accelerating the calculation time considerably. In particular, it allows for efficient pricing of long-maturity derivatives.

**Remark 4.4** Since at each time step, algorithm 1, or the ‘Fast Algorithm’, takes advantage of the tower property of conditional expectations over short time periods of the same length $\Delta M$, the transition probability matrix can be pre-computed at the beginning of the procedure and stored, accelerating the numerical process greatly.

### 4.2. Variable annuity with early surrenders

We approximate the value of the VA contract (including the right to surrender) by its Bermudan counterpart for a large
number of monitoring dates. Bermudan options can be exercised early, but only at predetermined dates $R \subset [0, T]$. Thus, Bermudan options are similar to American options, but the region of the permitted exercise times is a subset of $[0, T]$ containing a finite number of exercise dates, $\{t_0, t_1, \ldots, t_M\}$ with $t_z \in [0, T]$ for some $M \in \mathbb{N}$.

In this paper, we use the term Bermudan (resp. American) contract to refer to a variable annuity under which the policyholder has the right to surrender her contract prior to maturity on predetermined dates (resp. at any time prior to maturity). In the same vein, a variable annuity without surrender rights is also called a European contract. Note that these terms do not refer to existing contracts, and they are used to simplify explanations. Naturally, as $M \to \infty$, we expect the price of the Bermudan contract to converge to the one of a variable annuity with surrender rights as defined in (7). The latter is formalized in the following.

Let $\Delta M = T/M$ for some $M \in \mathbb{N}$ and define the set $\mathcal{M} = \{t_0, t_1, \ldots, t_M\}$ with $t_z \in [0, T]$, $z = 0, 1, \ldots, M$ such that $t_0 = 0$ and $t_M = T$. The time-$t$ risk-neutral value of the Bermudan contract with permitted exercise dates $\Delta M$ is

$$ b_M(t, x, y) = \sup_{\tau \in T_{\Delta M}^+} \mathbb{E}_x[e^{-r(T-\tau)}\varphi(\tau, F_T, V_T)] $$

where $T_{\Delta M}^+$ is the set of stopping times taking values in $\mathcal{M}$. Proposition 4.5 below shows that $b_M(t, x, y) \to v(t, x, y)$ as $M \to \infty$.

**Proposition 4.5** As $M \to \infty$, the value function of the Bermudan variable annuity contract (25) converges to its American counterpart (7), that is,

$$ \lim_{M \to \infty} b_M(t, x, y) = v(t, x, y). $$

The proof can be found in appendix A.2.

**Remark 4.6** In the practical context of variable annuities, surrenderers are often only allowed at specific times (such as on the policy anniversary dates). In these cases, the Bermudan contract may be more realistic than its American counterpart presented in (7).

We denote the Bermudan contract value process by

$$ B := \{B_z := b_M(t_z, F_{t_z}, V_{t_z})\}_{0 \leq z \leq M}. $$

From the principle of dynamic programming (see example Lamberton (1998, theorem 10.1.3)), it is well-known that the discretized problem admits the following representation:

$$ \begin{align*}
B_M &= \psi(T, F_T, V_T) \\
B_z &= \max (\psi(t_z, F_{t_z}, V_{t_z}), e^{-r\Delta M} \mathbb{E}_{t_z}[B_{z+1}]), & 0 \leq z \leq M - 1.
\end{align*} \tag{26} $$

Using CTMCs, we can define an approximation for the time-$t$ risk-neutral value of the Bermudan contract by

$$ b_M^{(m,N)}(t, x, y) = \sup_{\tau \in T_{\Delta M}^+} \mathbb{E}_x[e^{-r(T-\tau)}\varphi(\tau, F_T^{(m,N)}, V_T^{(m,N)}) | F_T^{(m,N)} = x, V_T^{(m,N)} = y]. $$

The approximation of the Bermudan contract value process, denoted by $B^{(m,N)} := \{B_z^{(m,N)} := b_M^{(m,N)}(t_z, F_{t_z}^{(m,N)}, V_{t_z}^{(m,N)})\}_{0 \leq z \leq M}$, is thus given by

$$ \begin{align*}
B_M^{(m,N)} &= \psi(T, F_T^{(m,N)}, V_T^{(m)}) \\
B_z^{(m,N)} &= \max (\psi(t_z, F_{t_z}^{(m,N)}, V_{t_z}^{(m,N)}), e^{-r\Delta M} \mathbb{E}_{t_z}[B_{z+1}^{(m,N)}]), & 0 \leq z \leq M - 1.
\end{align*} \tag{27} $$

**Algorithm 1:** Variable annuity without early surrenders via CTMC approximation – fast algorithm

**Input:** Initialize $Q^{(m)}$ as in (13) and $G^{(N)}$ for $j = 1, 2, \ldots, m$, as in (18)

M $\in \mathbb{N}$, the number of time steps,

$\Delta_M \leftarrow T/M$, the size of a time step

1. Set $B_{j\ast} \leftarrow [\varphi(T, e^{\alpha_i + r\gamma_{j\ast}}), \psi_{j\ast}]_{l=1}^M$ for $j = 1, 2, \ldots, m$

2. for $j = 1, 2, \ldots, m$

3. \hspace{1em} /* Calculate the transition probability matrices */

4. \hspace{2.5em} /* Adjusted transition probability matrix of $X^{(m,N)}$ given $V^{(m)} = \nu_j$ over a period of length $\Delta_M$ */

5. \hspace{3.5em} $P^{\nu_j} \leftarrow e^{G^{(N)}\Delta_M e^{-r\Delta_M}}$

6. \hspace{2.5em} /* Transition probability matrix of $V^{(m)}$ over a period of length $\Delta_M$ */

7. \hspace{3.5em} $P^{\nu} \leftarrow e^{Q^{(m)}\Delta_M}$

8. for $z = M - 1, \ldots, 0$

9. \hspace{1em} for $j = 1, 2, \ldots, m$

10. \hspace{2em} /* VA valuation */

11. \hspace{2.5em} $\bar{\Pi}_{t_z} \leftarrow P^{\nu} H_{t_z}$

12. \hspace{1em} for $n = 1, 2, \ldots, N$ do

13. \hspace{2em} /* Initialize */

14. \hspace{3.5em} $T_{z,n} \leftarrow e^{Q^{(m)}\Delta_M e^{-r\Delta_M}}$

15. \hspace{3.5em} $M_{z,n} \leftarrow e^{Q^{(m)}\Delta_M e^{-r\Delta_M}}$

16. \hspace{3.5em} $\bar{\Pi}_{t_z} \leftarrow P^{\nu} H_{t_z}$

17. return $b_M$
Or equivalently in terms of the process \( Y^{(m,N)} \), we have that

\[
\begin{align*}
B^{(m,N)}_M &= \psi(T, \psi^{-1}(Y^{(m,N)}_T)), \\
B^{(m,N)}_z &= \max \left( \psi(t, \psi^{-1}(Y^{(m,N)}_t)), e^{-r\Delta t} \mathbb{E}_z[B^{(m,N)}_{z+1}] \right), \\
0 \leq z \leq M - 1,
\end{align*}
\]

(28)

where \( \psi^{-1} \) is defined in Proposition 3.5.†

Finally, we have \( B_M(0, F_0, V_0) = B_0 \approx B^{(m,N)}_0 = B^{(m,N)}(0, F_0, V_0) \). Based on the above, an approximation for the value of the Bermudan contract can be obtained as described in the proposition below.

**Proposition 4.7** Let \( F_0 > 0, V_0 \in S_V \) and \( G^{(m,N)} \) be the generator defined in (20). The risk-neutral value of a variable annuity with maturity \( T > 0 \) and guaranteed amount \( G > 0 \) can be approximated recursively by

\[
\begin{align*}
B^{(m,N)}_M &= H^{(1)}, \\
B^{(m,N)}_z &= \max(\mathbf{H}^{(2)}; e^{-r\Delta t} \exp(\Delta_M G^{(m,N)}) \mathbf{B}^{(m,N)}_{z+1}), \\
0 \leq z \leq M - 1,
\end{align*}
\]

(29)

for \( M \in \mathbb{N} \) sufficiently large and where the maximum is taken element by element (also known as the parallel maxima). \( H^{(1)} \) and \( H^{(2)}; z = 0, 1, \ldots, M - 1 \) are column vectors of size \( mN \times 1 \) whose \( (l - 1)N + n \)-th entries, \( h^{(1)}_{l-1}N+n \) and \( h^{(2)}_{l-1}N+n \) are respectively given by

\[
\begin{align*}
h^{(1)}_{l-1}N+n &= \max(G, e^{r \gamma V^{(m,N)}_l}), \quad \text{and} \\
h^{(2)}_{l-1}N+n &= g(t, \gamma V^{(m,N)}_l),
\end{align*}
\]

(30)

\( 1 \leq l \leq m \) and \( 1 \leq n \leq N \).

Specifically, given \( X^{(m,N)}_0 = x \in \mathbb{S}_V \) (with \( V^{(m)}_0 = V_0 = v_k \)), the approximated value of the Bermudan contract is given by

\[
b^{(m,N)}(0, F_0, V_0) = c^* \mathbf{B}^{(m,N)}_0.
\]

Hence, based on the last proposition, algorithm 2 below provides a CTMC approximation for the value of a variable annuity contract (including early surrenders).

**Remark 4.8** (Convergence of VA prices with early surrenders)

Recall, from remark 3.4 that \( F^{(m,N)} \Rightarrow F \) as \( m, N \to \infty \). The convergence of the price of the Bermudan contract written on \( F^{(m,N)} \) to the price of the Bermudan contract written on \( F \), that is \( B^{(m,N)} \Rightarrow B_0 \) as \( m, N \to \infty \), follows from Song et al. (2013, theorem 9), and the results of Palczewski and Stettner (2010, theorem 3.5), on the alternative continuous reward representation of the value function \( v \).†

Finally, the convergence of the price of the Bermudan contract to its American counterpart as \( M \) goes to infinity follows from proposition 4.5.

To our knowledge, detailed error and convergence analysis for the two-layers CTMC approximation of early-exercise options have not yet been performed in the literature. However, Cui et al. (2018) demonstrate the accuracy of the approximation numerically in the context of American put option pricing.

### 4.2.1. Fast algorithm.

Similar to the no surrender case, the efficiency of algorithm 2 can be improved significantly by using an approximation based on the assumption that the variance process remains constant over small time periods (see proposition 4.3).

Let \( \varphi(t) := \left[ \varphi(t, e^{r \gamma V^{(m,N)}_l}), V^{(m,N)}_l \right]_{l=1}^m \) be a \( m \times N \) matrix representing the payoff at time \( t \) for each state in \( \mathbb{S}^{(m)}_V \times \mathbb{S}^{(N)}_X \), and \( \varphi_{j-1} \) be the \( j \)-th row of \( \varphi(t) \). We denote the matrix (vector) transpose operation by \( \top \).

The Fast Algorithm to value VA with surrender rights is given in algorithm 3.

The Fast Algorithms to price VA contracts with and without early surrenders are very similar. The only difference is the additional line 11 in algorithm 3. In fact, at a given time \( t_i \) (that is, we fix one \( z \) in the loop line 6 to 11), we can observe that at the end of the inner loop (line 9 and 10), the matrix \( B \) contains the continuation value of the Bermudan contract at \( t_i \). Since Bermudan contracts can be surrendered at any time in \( T_M \), we simply need to calculate the maximum between the continuation value and the payoff at \( t_i \) to obtain the value of the Bermudan contract at that time (line 11). Therefore, the only difference between the Fast Algorithms for VA pricing with and without early surrenders stems from the fact that the latter contract cannot be surrendered prior to maturity, and thus, only the continuation value needs to be calculated at each time step (that is, line 11 is not used to price VA contracts without early surrenders).

The computational effort in algorithms 1 and 3 resides in the calculation of the matrix exponentials (line 3 to line 5). Hence, once they are (pre-)computed, one can price variable annuity contracts with and without surrender rights simultaneously at almost no additional cost. This also holds true for any other VA contracts with different guaranteed amounts, that is, a large variety of contracts with different guarantee structures and surrender rights can be priced simultaneously for almost the same computational effort as a single contract. Numerical experiments below demonstrate the efficiency and the accuracy of the Fast Algorithm.

The new algorithm can be easily adapted to a wide range of payoff functions. Thus, it extends the previous work of Cui et al. (2018), done in the context of option pricing, by significantly decreasing calculation time and allowing for efficient valuation of long-maturity derivatives.

### 4.3. Optimal surrender surface

In this section, we provide an algorithm to approximate the optimal surrender strategy for variable annuities with a general fee structure depending on the fund value and the variance.
Algorithm 2: Variable annuity with early surrenders via CTMC approximation

Input: Initialize $G^{(m,N)}$ as in (20), $H^{(1)}$ and $H^{(2)}_c$, for $z = 0, 1, \ldots, M - 1$, as in (30)
$M \in \mathbb{N}$, the number of time steps,
$\Delta_M \leftarrow T/M$, the size of a time step
1 Set $B^{(m,N)}_M \leftarrow H^{(1)}$ and $A^{\Delta_M} \leftarrow \exp(\Delta_M G^{(m,N)})e^{-r \Delta_M}$
2 for $z = M - 1, M - 2, \ldots, 0$ do
3 $\left[ B^{(m,N)}_z \right] \leftarrow \max[H^{(2)}_c, A^{\Delta_M} B^{(m,N)}_{z+1}]$
4 $b^{(m,N)}_M(0, F_0, V_0) \leftarrow e_d b^{(m,N)}_0$
5 return $b^{(m,N)}_M(0, F_0, V_0)$

Algorithm 3: Variable annuity with early surrenders via CTMC approximation – fast algorithm

Input: Initialize $Q^{(m)}$ as in (13) and $G^N_j$ for $j = 1, 2, \ldots, m$, as in (18)
$M \in \mathbb{N}$, the number of time steps,
$\Delta_M \leftarrow T/M$, the size of a time step
1 Set $\phi(t_c) \leftarrow [\phi(t_c, e^{x+\rho y(t_c)}), v_j]_{j=1}^{m_N}$ for $z = 0, 1, \ldots, M$
2 Set $B_{s,z} \leftarrow \phi_{s,z}(t_M)$ for $j = 1, 2, \ldots, m$
3 for $j = 1, 2, \ldots, m$ do
4 /* Calculate the transition probability matrices */
5 /* Adjusted transition probability matrix of $X^{(m,N)}$ given $V^{(m)} = v_j$ over a period of length $\Delta_M$ */
6 $p^X_{j} \leftarrow e^{G^N_j \Delta_M} e^{-r \Delta_M}$
7 /* Transition probability matrix of $V^{(m)}$ over a period of length $\Delta_M$ */
8 $P^{V} \leftarrow e^{Q^{(m)} \Delta_M}$
9 /* VA valuation */
10 for $z = M - 1, \ldots, 0$ do
11 for $j = 1, 2, \ldots, m$ do
12 $E_{j,j} \leftarrow P^X_{j} B^T_{j,j}$
13 for $n = 1, 2, \ldots, N$ do
14 $B_{s,n} \leftarrow P^V E_{n,s}$
15 $B = \max(B, \phi(t_c))$
16 return $b_{hi}$

process. Policyholder behavior may significantly impact pricing and hedging of variable annuities, Kling et al. (2014). Thus, analyzing optimal surrender behavior is crucial for insurers when developing risk management strategies for variable annuities, Bauer et al. (2017), Niittuinerä (2022). Optimal surrender strategies have been studied in the literature in different contexts, see for instance Mackay (2014), Bernard et al. (2014), Bernard and MacKay (2015), Shen et al. (2016) and Kang and Ziveyi (2018).

The algorithms provided in this section are based on the CTMC Bermudan approximation and can thus be easily adapted to a wide range of payoff functions. In the context of American put option pricing, Ma et al. (2021) use the integral representation of the value function to derive a CTMC approximation for the optimal exercise surface. However, the method used by Ma et al. (2021) to derive such a representation requires the value function to be smooth enough, which can be difficult to prove for certain payoff functions and under general stochastic volatility models, see Lamberton and Terenzi (2019) for instance. The Bermudan approximation of the optimal surrender surface presented in this section does not require any specific regularity properties on the value function and thus extends the work of Ma et al. (2021) to more general payoffs and value functions.

The goal of this section is to approximate the (optimal) surrender surface using the CTMC approximation. To this end, we first introduce additional definitions and notations.

Definition 4.9 Let $E = [0, T] \times \mathbb{R}_+ \times \mathcal{S}_V$. The continuation region $C \subset E$ is defined as

$$C = \{(t, x, y) \in E : v(t, x, y) > \phi(t, x, y)\},$$

and the surrender region $D \subseteq E$, as

$$D = \{(t, x, y) \in E : v(t, x, y) = \phi(t, x, y)\}.$$

Remark 4.10 It follows from definition 4.9 that $E = C \cup D$, since $v(t, x, y) \geq \phi(t, x, y)$ for all $(t, x, y) \in E$. If the value function $v$ is continuous, then $C$ is an open set and $D$ is closed; and the optimal surrender surface is the boundary $\partial C$ of $C$.

Definition 4.9 provides a simple way of approximating the optimal surrender surface via CTMC approximation. To
do so, denote by \( f_{nl} \) the approximated fund process associated to \((x_n, v_i) \in S_{i}^{(N)} \times S_{i}^{(m)} \) such that \( f_{nl} = e \lambda^+ \gamma(y_{1:n}) \) and let \( S^{(m,N)}_{f} = [f_{nl}]_{n=1}^{N,m} \) be the state-space of \( F^{(m,N)} \). We also denote by \( S^{(m,N)}_{f, \star} = [f_{nl}]_{n=1}^{N} \) the l-section of \( S^{(m,N)}_{f} \). Moreover, let \( \hat{H}_{M} = \{t_0, t_1, \ldots t_{m-1}\} \), that is \( \hat{H}_{M} = H_{M} \setminus t_{M} \), and \( \hat{E}_{l} = \hat{H}_{M} \times \omega_{l}^{(m,N)} \times [v_{l}] \). Using the previous definitions, the l-section, \( l \in \{1, 2, \ldots, m\} \), of the continuation and the surrender regions can be approximated using the CTMC processes via

\[
C^{(m,N)}_{l} = \left\{ (t, z, f_{nl}, v_i) \in \hat{E}_{l} \mid b^{(m,N)}_{l}(t, z, f_{nl}, v_i) > g(t, z, v_i)f_{nl} \right\},
\]

and

\[
D^{(m,N)}_{l} = \left\{ (t, z, f_{nl}, v_i) \in \hat{E}_{l} \mid b^{(m,N)}_{l}(t, z, f_{nl}, v_i) = g(t, z, v_i)f_{nl} \right\}
\cup \{T\} \times S_{f, \star} \times [v_{l}],
\]

respectively. Hence, the approximated continuation and surrender regions are given by

\[
C^{(m,N)} = \bigcup_{l=1}^{m} C^{(m,N)}_{l}, \quad D^{(m,N)} = \bigcup_{l=1}^{m} D^{(m,N)}_{l}.
\]

We use the notation of proposition 4.7, with \( b^{(i-1)N} \) denoting the \((i-1)N + r\)-th entry of \( B^{(m)}_{r} \). For \((t, f_{nl}, v_{i})\), \( 0 \leq z \leq M - 1 \), \( 1 \leq n \leq N \) and \( 1 \leq l \leq m \), we set \((t, f_{nl}, v_i) \in C^{(m,N)} \) if \( b^{(i-1)N+m} < g(t, v_i)f_{nl} \) and \((t, f_{nl}, v_i) \in D^{(m,N)} \) otherwise. The approximated optimal surrender surface can then be obtained by analyzing the shape of \( C^{(m,N)} \).

We are now interested in studying the shape of the surrender region. To do so, we fix \( t \in [0, T) \) and \( y \in S_{Y} \) and consider the set of points \( D_{\gamma} \subseteq \mathbb{R}_{+} \) for which it is optimal to surrender the contract. More precisely, we define \( D_{\gamma} \) by

\[
D_{\gamma} = \{ f \in \mathbb{R}_{+} \mid (t, f, y) \in D \}.
\]

Suppose that, for all couples \((t, y) \in [0, T) \times S_{Y} \), the set \( D_{\gamma} \) is of the form \( f^{\star}(t, y, \infty) \) for some \( f^{\star}(t, y) \in \mathbb{R}_{+} \). That is, \( f^{\star}(t, y) \) is the smallest fund value for which it is optimal to surrender the contract at time \( t \) for a volatility level \( y \), and for any fund value greater than \( f^{\star}(t, y) \), it is also optimal to surrender. Mathematically, this may be expressed by

\[
f^{\star}(t, y) := \inf \{ f \in \mathbb{R}_{+} \mid f \in D_{\gamma} \} = \inf[D_{\gamma}]. \tag{31}
\]

Under this assumption, the continuation and the surrender regions can be expressed as

\[
C = \left\{ (t, f, y) \in E \mid f < f^{\star}(t, y) \right\},
\]

and

\[
D = \left\{ (t, f, y) \in E \mid f \geq f^{\star}(t, y) \right\} \cup \{T\} \times \mathbb{R}_{+} \times S_{Y},
\]

respectively.

Hence, under this assumption, the optimal surrender surface \( f^{\star} \) splits \( E \) in two regions: at or above the surface is the surrender region, and below, the continuation region. That is, the set \( D_{\gamma} \) is connected.† In this paper, we say that the surrender region is of ‘threshold type’ if for any \((t, y) \in [0, T) \times S_{Y} \), the set \( D_{\gamma} \) is connected. There is a financial interpretation for such a form for the surrender region. As explained in Milevsky and Salisbury (2001), it is optimal for the policyholder to hold on to the contract when the fund value is low since there is a higher chance that the guarantee will be triggered at maturity.

**Remark 4.11** The surrender region can take any shape; see for examples Mackay et al. (2017, figures 4 and B.1). However, for specific fee and surrender charge structures, it can be shown that the surrender region is of threshold type, see for instance Mackay (2014, appendix 2.A), when the index value process is modeled by a geometric Brownian motion. Other authors take this form for the surrender region as an initial assumption, see for example Kang and Ziveyi (2018). In the context of financial derivative pricing, Jacka (1991, proposition 2.1), shows that the continuation region of American put options is of threshold type under the Black-Scholes setting whereas Touzi (1999, section 2), proves it for some stochastic volatility models, and De Angelis and Stabile (2019, proposition 4.1), in a very general setting.

When the surrender region is of threshold type, a simple algorithm can be developed to approximate the optimal surrender surface. The idea is based on the definition of \( f^{\star}(t, y) \) in (31) above: for each \( t \in H_{M} \) and \( v_{i} \in S_{i}^{(m)} \), we identify the smallest fund value \( f^{(m,N)}(t, v_{i}) \) for which it is optimal to surrender. Algorithm 4 returns the approximated optimal surrender surface \( f^{(m,N)}(t, v_{i}) \) (under the assumption that the surrender region is of threshold type) and the approximated value of a variable annuity with early surrenders given \( X^{(m,N)}_{0} = x_{i} = \ln(F_{0}) - \rho y(V_{0}) \) and \( V^{(m)}_{0} = V_{0} = v_{i} \).

We note that the derivation of the optimal surrender surface is not mandatory to obtain the value of the Bermudan contract, as observed from algorithm 2 or 3 in the previous subsection. Similarly as above, algorithm 5 is the fast version of algorithm 4. Recall that \( \varphi(t) := \{ \varphi(t, e^{\lambda^+ \gamma(y_{1:n})}) \}_{n=1}^{N} \) is a \( m \times N \) matrix representing the payoff at time \( t \) for each state in \( S_{i}^{(N)} \times \Delta X_{i} \), and \( \varphi_{j,n}(t) \) (resp \( B_{j,n} \)) is the \( j \)-th row of \( \varphi(t) \) (resp \( B \)). We also denote by \( b_{j,n} \), the \( (j,n) \)-entry of the matrix \( B \).

**Remark 4.12** Algorithms 2 and 3 do not require the specification of any particular form for the surrender region, which is not the case for many of the numerical procedures presented in the literature (Bernard et al. 2014, Shen et al. 2016, Kang and Ziveyi 2018). Hence, their scope is more general.

The accuracy of the approximated surrender boundary is demonstrated numerically in appendix B (available online as supplemental material).

### 4.4. CTMC approximation of the VIX

In section 5, we analyze numerically the impact of various VIX-linked fee structures on the optimal surrender strategy. Since analytical formulas for the VIX are not always known for all models listed in table 1, we use a CTMC approach...


Algorithm 4: Optimal surrender surface (of threshold type) via CTMC approximation

Input: Initialize $G^{(m,N)}$ as in (20), $H^{(1)}$ and $H^{(2)}$, for $z = 0, 1, \ldots, M - 1$, as in (30)

$M \in \mathbb{N}$, the number of time steps,

$\Delta M \leftarrow T/M$, the size of a time step

1. Set $B^{(m,N)}_{-M} \leftarrow H^{(1)}$ and $A_{\Delta M} \leftarrow \exp[\Delta M G^{(m,N)}] e^{-r \Delta M}$

2. for $z = M - 1, M - 2, \ldots, 0$ do

   3. $B^{(m,N)}_{z+1} \leftarrow \max\{H^{(2)}, A_{\Delta M} B^{(m,N)}_{z+1}\}$

   4. for $l = 1, 2, \ldots, m$ do

      5. $n \leftarrow 1$

      6. while $(b_{z,l}(t_{n+1}) > g(t_{n}, v_{l}) e^{\sigma_{s}^{2} \rho_{g}(t_{n})})$ and $(n < N)$ do

         7. $n \leftarrow n + 1$

         8. $f^{(m,N)}(t_{n}, v_{l}) \leftarrow e^{\sigma_{s}^{2} \rho_{g}(t_{n})}$

   9. $b^{(m,N)}_{m}(0, F_{0}, V_{0}) \leftarrow e_{d} B^{(m,N)}_{d}$

10. return $f^{(m,N)}$ and $b^{(m,N)}_{m}(0, F_{0}, V_{0})$

Algorithm 5: Optimal surrender surface (of threshold type) via CTMC approximation – fast algorithm

Input: Initialize $Q^{(m)}$ as in (13) and $G^{(m,N)}_{j}$ for $j = 1, 2, \ldots, m$, as in (18)

$M \in \mathbb{N}$, the number of time steps,

$\Delta M \leftarrow T/M$, the size of a time step

1. Set $\phi(t_{0}) \leftarrow [\phi(t_{0}, e^{\sigma_{s}^{2} \rho_{g}(t_{0})}, v_{j})]^{m_{N}}_{j=1}$ for $z = 0, 1, \ldots, M$

2. Set $B_{j+1} \leftarrow \phi_{j+1}(t_{0})$ for $j = 1, 2, \ldots, m$

/* Calculate the transition probability matrices */

3. for $j = 1, 2, \ldots, m$ do

   4. /* Adjusted transition probability matrix of $X^{(m,N)}$ given $V^{(m)} = v_{j}$ over a period of length $\Delta M$ */

   5. $p_{X}^{j} \leftarrow e^{G^{(m,N)}_{j} \Delta M} e^{-r \Delta M}$

   6. /* Transition probability matrix of $V^{(m)}$ over a period of length $\Delta M$ */

   7. $P^{V} \leftarrow e^{Q^{(m,N)} \Delta M}$

   8. /* VA valuation */

6. for $z = M, M - 1, \ldots, 0$ do

7. for $j = 1, 2, \ldots, m$ do

   8. $E_{j} \leftarrow P_{j}^{X} B_{j+1}^{\phi}$

   9. for $n = 1, 2, \ldots, N$ do

   10. $B_{n} \leftarrow P^{V} E_{n}$

   11. $B = \max(B, \phi(t_{0}))$

12. for $j = 1, 2, \ldots, m$ do

   13. $n \leftarrow 1$

   14. while $(b_{j,n} > g(t_{n}, v_{j}) e^{\sigma_{s}^{2} \rho_{g}(t_{n})})$ and $(n < N)$ do

         15. $n \leftarrow n + 1$

         16. $f^{(m,N)}(t_{n}, v_{j}) \leftarrow e^{\sigma_{s}^{2} \rho_{g}(t_{n})}$

17. return $f^{(m,N)}$ and $b^{(m,N)}_{j}$


To approximate the value of the volatility index. This is the case for the numerical experiments performed under the 3/2 model whose results are available online in appendix B.† In this section, we propose an approximation for the VIX when the variance process is approximated by a CTMC.

Let $\{V^{2}_{t}\}_{t \geq 0}$ be the process representing the square of the VIX, defined by

$$VIX^{2}_{t} = E_{0} \left[ \frac{1}{\tau} \int_{t}^{t+\tau} \sigma_{v}^{2}(V_{s}) \, ds \right],$$

with $\tau = 30/365$, see Cui et al. (2021) equation (6) for details.

Recall that $V^{(m)}$ is the CTMC approximation of $V$ taking values on a finite state-space $S^{(m)}_{V} := \{v_{1}, v_{2}, \ldots, v_{m}\}$.

† For the 3/2 model, a closed-form expression for the VIX may be found in Carr and Sun (2007, theorem 4). However, as pointed out by Drimus (2012), the integral that appears in the analytical formula is difficult to implement and is not suited for fast and accurate numerical methods. For this reason, the CTMC approximation of the VIX is used in the numerical examples under the 3/2 model.
When the variance process is a CTMC, an integral expression can be obtained for the value of the volatility index. The CTMC approximation of the VIX, denoted by $VIX_{(m)} = \{VIX_{(m)}^{(n)}\}_{n=0}^{\infty}$, is given in the proposition below.

**Proposition 4.13** Given $V_{(m)}^{(n)} = v_k$, the square of the VIX index at time $t$ can be approximated by

$$
(VIX_{(m)}^{(n), k})^2 := \mathbb{E} \left[ 1 \int_{t}^{t+\tau} \sigma_{t}^2(V_{(m)}^{(n)}) \, ds \big| V_{(m)}^{(n)} = v_k \right]
$$

$$
= \frac{1}{\tau} \int_{0}^{\tau} e^{-H_{m}(v_{j})} \, ds,
$$

(32)

where $\tau = 30/365$, $Q_{(m)}$ is the generator of $V_{(m)}$ defined in (13), $e_k$ is the $k^{th}$ canonical basis vector of $\mathbb{R}^m$ and $H$ is a $m \times 1$ vector whose $j$-th entry $h_j$ is given by $h_j = \sigma_{j}^2(v_j)$, $j = 1, 2, \ldots, m$.

The proof is a direct consequence of Fubini’s Theorem and the CTMC representation for conditional expectations.

**Remark 4.14** Since $V_{(m)}$ is time-homogeneous, the approximation does not depend on $t$ and thus, for each $k \in \{1, \ldots, m\}$ it needs to be calculated only once for all $t \geq 0$.

The integral part in (32) can be approximated via a quadrature rule. In the numerical example section, this is done by dividing the interval $[0, \tau]$ into $n > 0$ equidistant sub-intervals for $z = 0, 1, 2, \ldots, n$ with

$$
t_z := z\Delta_n, \quad \Delta_n := \tau/n.
$$

The approximation then becomes

$$
(VIX_{(m)}^{(n), k})^2 \approx \frac{\Delta_n}{\tau} e_k \sum_{i=1}^{n} e^{Q_{(m)} s_i} H.
$$

(33)

The approximation in (33) can be implemented in a straightforward manner. However, it requires calculating the exponential of a matrix $n$ times, which can be computationally inefficient. By making use of the tower property of conditional expectations, algorithm 6 speeds up the calculation of (33).

We remark that other quadrature rules could be used in for the numerical calculation of (32). Using schemes that have a faster rate of convergence may reduce the number matrix calculations needed.

The vector $\theta_{m \times 1}$ represents the null column vector of size $m$. Note that algorithm 6 requires the calculation of a matrix exponential only once at the beginning of the procedure, which makes the algorithm very efficient.

Numerical experiments in the next section also demonstrate the accuracy and the efficiency of algorithm 6 empirically. This new VIX CTMC approximation can also be used for approximating the price of VIX derivatives. This is left as future research.

## 5. Numerical analysis

In the constant fee case, the misalignment between the fees and the value of the financial guarantee creates an incentive for the policyholder to surrender her policy prematurely (see Milevsky and Salisbury 2001 for details). Indeed, when the fund value is high, the amount of the fees paid is also high, but the put option embedded in a GMMB is out-of-the-money and worth very little since the probability for the guarantee to be triggered at maturity is low. Thus, the policyholder pays high fees for a financial guarantee that has a low value. This is clearly an incentive for a policyholder to surrender her policy prior to maturity. Mackay et al. (2017) considered state-dependent fee structures where the fee is paid when the fund value is under a certain level, and showed that this particular type of fee structure reduces insurers’ exposure to policyholder behavior under risk-neutral value maximization assumption. Fees that are tied to the S&P volatility index, the VIX, are also studied in the literature, Cui et al. (2017) and Kouritzin and MacKay (2018). Since the volatility is negatively correlated with the stock price (see for instance Rebonato 2004), we expect VIX-linked fees to be low when the fund value is high and to be higher when the fund value is low. Cui et al. (2017) and Kouritzin and MacKay (2018) showed numerically that linking the fee to the volatility index VIX may help realign revenues with variable annuity liabilities (for VA without surrender rights). Hence, it is reasonable to believe that linking the fees to the VIX may help to reduce insurers’ exposure to surrender risk. This will be explored in greater detail in the numerical experiments conducted below.

This section first discusses the market, the VA, and the CTMC parameters used in all numerical experiments performed below. Then, we investigate the efficiency of the Fast Algorithms (algorithms 1 and 3). In the third subsection, we discuss different structures for the VIX-linked fee, that is, different ways to link the fee rate to the VIX index. Finally, we analyze the impact of different VIX-linked fee structures on the value of VAs and their optimal surrender strategy. We restrict our analysis to the classical Heston model. Using our framework, any model listed in table 1 could have been used.

As supplemental material (available online), we investigate the numerical accuracy of the approximated optimal surrender surface derived in algorithm 4 (or equivalently 5), and the CTMC approximation of the VIX (algorithm 6). We also analyze numerically the impact of time-dependent risk-free rates on the variable annuity values and the optimal surrender surface. Finally, we explore numerically the impact of VIX-linked fee structures under the 3/2 model.

### 5.1. Market, VA and CTMC parameters

We consider a market under regular conditions: low initial variance $\theta_0$, low long-term variance $\theta$, moderate volatility of volatility $\sigma$, and moderate speed reversion $\kappa$.

† Bloomberg provides historical Heston calibrated parameters to market data on a daily basis via its Option Pricing template (OVME). These parameters are often used in practice for over-the-counter option pricing. Bloomberg’s Heston calibrated speed reversion parameter is $\kappa = 3.6881$ as of December 31, 2019, $\kappa = 5$ as of March 31, 2020 and $\kappa = 1.1397$ as of September 30, 2020. The parameter selected for our numerical experiments falls approximately in the middle of those of December 2019 and September 2020. In the financial literature, Aït-Sahalia and Kimmel (2007) obtain $\kappa = 5.07$ whereas Garcia et al. (2011) obtain $\kappa = 0.173$, and again our values fall between these two values.
value of the variance is set to \( V_0 = 0.03 \), the correlation to \( \rho = -0.75 \) and the risk-free rate to \( r = 0.03 \). The selected market parameters are summarized in table 2 and the model dynamic is given in table 1.

The variable annuity parameters are set to \( F_0 = S_0 = 100 \), \( T = 10 \) (years), and \( G = 100 \). We assume that the payoff when the contract is surrendered early is given by \( g(t,V_t)F_t \), with

\[
g(t,y) = e^{-k(T-t)}, \quad y \in S_V,
\]

and \( k = 0.2\% \). The choices for the fee function \( c(x,y) \) are discussed in greater detail in the next section.

Note that a numerical analysis under the Heston model with \( G = F_0e^{\beta T} \), \( \delta = 2\% \) (rather than \( G = 100 \)) is also performed in appendix B, available online as supplemental material. We also investigate the impact of time-dependent risk-free rates on the value of variable annuity contracts and the optimal surrender surface in section B.5.

For all numerical examples in this paper, we use the non-uniform grid proposed by Tavella and Randall (2000, chapter 5.3). For example, suppose that \( \tilde{X} \) is a one-dimensional diffusion process approximated by a continuous-time Markov chain \( X^{(m)} \) taking values on a finite state-space \( S_\tilde{X} = \{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n\} \), \( n \in \mathbb{N} \). The state-space of the approximated process can be determined as follows:

\[
\tilde{x}_i = \tilde{X}_0 + \tilde{\alpha} \sinh \left( c_2 \frac{i}{n} + c_1 \left[ 1 - \frac{i}{n} \right] \right),
\]

\[
i = 2, \ldots, n-1,
\]

where

\[
c_1 = \sinh^{-1} \left( \frac{\tilde{x}_1 - \tilde{X}_0}{\tilde{\alpha}} \right), \quad \text{and} \quad c_2 = \sinh^{-1} \left( \frac{\tilde{x}_n - \tilde{X}_0}{\tilde{\alpha}} \right).
\]

Here the constant, \( \tilde{\alpha} \geq 0 \), controls the degree of non-uniformity of the grid. The choices for the two boundary states and the non-uniformity parameter \( \tilde{\alpha}_o \) (resp. \( \tilde{\alpha}_x \)) for \( V^{(m)} \) (resp. \( X^{(m,N)} \)) are discussed in more details below.

**Remark 5.1** When the initial values of the auxiliary and variance processes are not in the grid, they can be inserted (see for example Cui et al. 2019, section 2.3 for details), or the value of the variable annuity must be interpolated between the appropriate grid points.

Non-uniform schemes have been used frequently in the literature for options pricing via CTMC approximation methods, see Mijatović and Pistorius (2013), Lo and Skindilias (2014), Cai et al. (2015), Kirkby et al. (2017), Cui et al. (2018), Cui et al. (2019), Leitao Rodriguez et al. (2021) and Ma et al. (2021), among others. For a deep analysis of grid designs and how they can affect convergence, the reader is referred to Zhang and Li (2019).

Unless stated otherwise, all numerical experiments are performed using the CTMC parameters listed in table 3. Recall that \( m \) is the number of grid points for the variance process whereas \( N \) represents the number of grid points of the fund process, \( \tilde{\sigma}_x \) (resp. \( \tilde{\alpha}_x \)) is the grid non-uniformity parameter of the variance (resp. the auxiliary) process. The grid’s upper and lower bounds are respectively \( v_1 \) and \( v_m \) for the variance process and \( x_1 \) and \( x_N \) for the auxiliary process with \( X_0 = \ln(F_0) - \rho y(V_0) \). The values of \( V_0 \) and \( X_0 \) are inserted in their respective grid as per remark 5.1. Finally, we use \( M = 500 \times 10 \) steps, which corresponds to a computing frequency of approximately twice daily (since there are approximately 250 trading days per year).

Note that, under the Heston model, good approximations of the transition density of the variance process can be obtained with a small number of grid points, see for example Cui et al. (2019, figure 3).

All the numerical experiments are carried out with Matlab R2015a on a Core i7 desktop with 16GM RAM and speed 2.40 GHz.

### 5.2. Efficiency of the fast algorithms

The valuation of options (or variable annuities) using CTMC requires the calculation of a matrix exponential to obtain the transition probability matrix. In algorithm 2, the two-dimensional process is mapped onto a one-dimensional process resulting in a generator of size \( mN \times mN \). Thus, we need to calculate the exponential of a \( mN \times mN \) matrix to obtain the transition probability matrix. For large values of \( mN \), this procedure might stretch computing resources to unacceptable levels. Algorithms 1 and 3, proposed in section 4, require \( m \) times the calculation of the exponential of a \( N \times N \) matrix and

---

**Algorithm 6:** Efficient algorithm for the calculation of the VIX using CTMC approximation

**Input:** Initialize \( Q^{(m)} \) as in (13) and \( H \) as in proposition 4.3

\( n \in \mathbb{N} \), the number of time steps,

\( \Delta_n \leftarrow \tau/n \), the size of a time step

1. Set \( A_{\Delta_n} \leftarrow \exp[\Delta_n Q^{(m)}] \), \( S \leftarrow 0_{m \times 1} \) and \( E \leftarrow H \)

2. for \( z = n, n-1, \ldots, 1 \) do

3. \( E \leftarrow A_{\Delta_n}E \)

4. \( S \leftarrow S + E \)

5. \( \text{VIX}^{(m),k} \leftarrow \sqrt{e_t S} \frac{\Delta_n}{T} \)

---

### Table 2. Market parameters.

| Parameter | Value |
|-----------|-------|
| \( V_0 \)  | 0.03  |
| \( \kappa \) | 2.00  |
| \( \theta \) | 0.04  |
| \( \sigma \) | 0.20  |
| \( \rho \)  | -0.75 |
| \( r \)    | 0.03  |
When the size of the exponent in the matrix exponential is greater than 200 × 200, we observe that the function fastExpm for Matlab, see Mentink-Vigier (2023),† which is designed for the fast calculation of matrix exponentials of large sparse matrices, can further speed up the calculation. Combining the function fastExpm and the Fast Algorithms can speed up the code by up to 100 times (for European and Bermudan contracts). For the European contract, the Fast Algorithm can reduce the computation time by up to 12 times whereas the function fastExpm, by up to 7 times. For the Bermudan contract, the computation time is reduced by up to 4 times with the Fast Algorithm and by up to 40 times with the function fastExpm. When \( m = 50 \) and \( N = 100 \) the running time is approximately 6 seconds for both the European and the Bermudan contracts, confirming the high efficiency of the new algorithms.

Figure 1 illustrates the computation time in seconds of the ‘Fast Algorithm’, algorithm 1 for the European contract and algorithm 3 for the Bermudan contract, combined with the function fastExpm of Mentink-Vigier (2023) (when the size of the generator in the matrix exponential is greater than 200 × 200), and the computation time of the ‘Regular Algorithm’ for the European contract (22) and algorithm 2 for the Bermudan one. The run times in figure 1 are recorded using the market, VA, and CTMC parameters of subsection 5.1, except for the number of grid points for the auxiliary process \( X \) which is set to \( N = 100, 200, 300 \) and 500, respectively. Moreover, we use a constant fee structure, that is we fix \( c(x, y) = 1.5338\% \) for all \((x, y) \in \mathbb{R}_+ \times S_Y\).

We also compare the accuracy of the new algorithms. For comparison, we use the regular algorithms for the European contract and for the Bermudan contract. The absolute difference in the VA prices between the regular and the fast algorithms is around \( 10^{-5} \) for both European and Bermudan contracts, whereas the relative difference is around \( 10^{-5} \), confirming the accuracy of the fast algorithms. See also appendix (B) available online for more details.

**Remark 5.2.** When valuing a European contract, the use of the Expokit of Sidje (1998) based on Krylov subspace projection methods accelerates the computation time considerably. Moreover, software packages in Matlab and Fortran can be downloaded for free at https://www.maths.uq.edu.au/expokit. When \( N = 500 \), we observe that the function expm in the Expokit can accelerate the running time by up to 7 times (compared to the fast algorithm).

However, since the function takes advantage of the product of a matrix exponential with a vector, the fast algorithm is more efficient when valuing Bermudan contracts. Indeed, when using the fast algorithm, matrix exponentials are calculated only once at the beginning of the procedure; whereas when using the Expokit, it needs to be calculated at each time step (since we need to calculate the matrix exponential multiplied by a vector in order to make use of the procedure) which slows down the execution considerably. Hence, when valuing a Bermudan contract, algorithm 3 is up to 9 times faster than the regular algorithm using Expokit procedures of Sidje (1998).

As mentioned previously, the computational effort in algorithms 1 and 3 resides in the calculation of the matrix exponentials at the beginning of the two procedures. Hence, once they are (pre-)computed, one can obtain the value of variable annuities with and without surrender rights simultaneously at almost no additional cost. For instance, when \( N = 500 \), we simultaneously obtain the prices of variable annuities with and without surrender rights in 270 seconds; whereas the values of the Bermudan and the European contracts can be obtained separately in approximately 250 seconds for each. This also holds true for any other VA contracts with different guaranteed amounts; that is, a variety of contracts can be priced for almost the same computational effort as a single contract.

### 5.3. Fee structures and fair fee parameters

First, recall from subsection 4.4 that

\[
\text{VIX}_t^2 = \mathbb{E}_t \left[ \frac{1}{\tau} \int_{t}^{t+\tau} \sigma_x^2(V_s) \, ds \right], \quad \text{with } \tau = 30/365.
\]

For all the numerical experiments conducted below, we consider three types of VIX-linked fee structures. As in Cui et al. (2017), we use an uncapped VIX\(^2\)-linked fee structure (the ‘Uncapped VIX\(^2\)’ fee structure) of the form

\[
c_t = \tilde{c} + \tilde{m}\text{VIX}_t^2, \quad 0 \leq t \leq T,
\]

with \( \tilde{c}, \tilde{m} \geq 0 \).

When the volatility is high (e.g., financial turmoil), the Uncapped VIX\(^2\) fee rate can become excessive (see table 5 for examples). This motivates our choice of imposing a cap to this type of fee. Thus, for the second fee structure, we consider capped VIX\(^2\)-linked fees (or simply the ‘Capped VIX\(^2\)’ fee structure or Capped fee structure) of the form

\[
c_t = \min(\tilde{c} + \tilde{m}\text{VIX}_t^2, K), \quad 0 \leq t \leq T,
\]

where \( \tilde{c}, \tilde{m} \geq 0 \) and \( K > 0 \).

---

†The function fastExpm is based on Hogben et al. (2011) and Kuprov (2011).
Finally, we look at an uncapped fee structure linked to the VIX (rather than the VIX squared) given by
\[ c_t = \tilde{c} + \tilde{m} V_t, \quad 0 \leq t \leq T, \]
with \( \tilde{c}, \tilde{m} \geq 0 \), also called the ‘Uncapped VIX’ fee structure. Such a structure can help to keep the fee rates reasonable during high volatility periods.

Since the volatility process is Markovian, the volatility index at \( t \) will depend only on \( V_t \). Therefore, the three fee functions are of the form \( c(y), y \in S_V \).

The fee structure parameters \( \tilde{c} \) and \( \tilde{m} \) are set in such a way that the contract is fair at inception. That is, the initial amount invested by the policyholder, \( F_0 \), is equal to the expected discounted value of the future benefit (without early surrenders). Those parameters are called the fair parameters and are henceforth denoted by a star: \((\tilde{c}^*, \tilde{m}^*)\). We set the fee in this manner to calculate the value added by the right to surrender. To identify the fair fee vector \((\tilde{c}^*, \tilde{m}^*)\), we first fix a multiplier \( \tilde{m}^* \), and then solve for the corresponding fair base fee \( \tilde{c}^* \). Note that such a fair base fee \( \tilde{c}^* \) does not exist for all values of \( \tilde{m}^* > 0 \).

When \( c_t = \tilde{c}^* + \tilde{m}^* V_t \), the fair parameters are obtained using the exact formula of Cui et al. (2017). For the other fee structures, the fair parameters are calibrated using a CTMC approximation with \( N = 100 \) (all other CTMC parameters are the same as in table 3), to reduce the computation time. Note that very accurate VA prices are obtained extremely fast with \( N = 100 \) under the Heston model (see appendix B, available online, for numerical details). Table 4 presents the fair fee vectors \((\tilde{c}^*, \tilde{m}^*)\), and table 5 shows examples of fair fee rates produced by each fair fee vector at different volatility levels \((\sqrt{V_t})\).

The uncapped VIX\(^2\)-linked fee rate can get very high as the volatility increases, reaching levels as high as 7.4653% when the volatility is 42.772%. During the last COVID-19 financial crisis, volatility reached levels as high as 80% in March 2020.† This motivates the two other fee structures we propose. We note that the introduction of a cap does not significantly affect the calibrated fair fee parameters. We also observe, from table 5 that the second and the third fee structures allow to keep the fees at reasonable levels during high volatility periods compared to the Uncapped VIX\(^2\) fee structure.

### 5.4. Effect of VIX-linked fees on surrender incentives

Recall that under the Heston (1993) model, the price of the risky asset satisfies
\[
\begin{align*}
\text{d}S_t &= r S_t \text{d}t + \sqrt{V_t} S_t \text{d}W_t^{(1)}, \\
\text{d}V_t &= \kappa (\theta - V_t) \text{d}t + \sigma \sqrt{V_t} \text{d}W_t^{(2)},
\end{align*}
\]  
(35)

with \( S_0 \) and \( V_0 \) are deterministic, and where \( W = (W^{(1)}, W^{(2)})^T \) is a bi-dimensional correlated Brownian motion under \( \mathbb{Q} \) and such that \( [W^{(1)}, W^{(2)}]_t = pt \) with \( p \in [-1, 1] \), the speed of the mean-reversion \( \kappa > 0 \), the long term variance \( \theta > 0 \) and the volatility of the variance \( \sigma > 0 \) (also called the volatility of the volatility).

Moreover, when the market is modeled by (35), the VIX has a closed-form expression given by
\[ VIX_t^2 = B + AV_t \]  
(36)

with \( A = \frac{1-e^{-\frac{t}{\tau}}}{\frac{t}{\tau}} \) and \( B = \frac{\theta e^{-\frac{t}{\tau}}}{\frac{t}{\tau}} + \frac{\eta}{\sigma} \), see Zhu and Zhang (2007) for details. The three fee structures exposed in subsection 5.3 can thus be obtained explicitly in terms of the current volatility using (36) as shown in table 6.

Now from lemma 3.3, we find that \( r(x) = x/\sigma \). Thus, given a certain fee process \( c_t \) (listed in table 6), the dynamics of the auxiliary process can be derived as
\[
\begin{align*}
\text{d}X_t &= \mu_X (X_t, V_t) \text{d}t + \sigma_X (V_t) \text{d}W_t^X, \\
\text{d}V_t &= \mu_V (V_t) \text{d}t + \sigma_V (V_t) \text{d}W_t^V,
\end{align*}
\]  
(37)

where \( \mu_X (X_t, V_t) = r - \frac{\mu_0}{\sigma} - c_t + V_t (\frac{\theta}{\sigma} - \frac{1}{2}) \), and \( \sigma_X (V_t) = \sqrt{(1 - \rho^2)V_t}, 0 \leq t \leq T \).

† See VIX historical data at https://www.cboe.com/tradable_products/vix/vix_historical_data/.
Table 4. Fair fee vectors \((\tilde{c}^*, \tilde{m}^*)\).

| \(\tilde{c}^*\) | \(\tilde{m}^*\) |
|----------------|----------------|
| 1.5338%        | 0.0000         |
| 1.0036%        | 0.1500         |
| 0.4741%        | 0.3000         |
| 0.0000%        | 0.4345         |

\[ c_t = \tilde{c}^* + \tilde{m}^* \sqrt{V_{i,t}} \]

\[ c_t = \min(\tilde{c}^* + \tilde{m}^* \sqrt{V_{i,t}}^2, K), K = 2\% \]

| \(\tilde{c}^*\) | \(\tilde{m}^*\) |
|----------------|----------------|
| 1.5338%        | 0.0000         |
| 1.0112%        | 0.1500         |
| 0.5415%        | 0.3000         |
| 0.0000%        | 0.4927         |

\[ c_t = \tilde{c}^* + \tilde{m}^* \sqrt{V_{i,t}} \]

\[ c_t = \min(\tilde{c}^* + \tilde{m}^* \sqrt{V_{i,t}}^2, K), K = 2\% \]

| \(\tilde{c}^*\) | \(\tilde{m}^*\) |
|----------------|----------------|
| 1.5338%        | 0.0000         |
| 1.0750%        | 0.0250         |
| 0.6164%        | 0.0500         |
| 0.0000%        | 0.0836         |

Table 5. Fair fee rates in %.

\[ a) c_t = \tilde{c}^* + \tilde{m}^* \sqrt{V_{i,t}} \]

| \(\sqrt{V_{i,t}}\) | \(\tilde{m}^*\) |
|-------------------|----------------|
| 8.702             | 1.5338         |
| 13.882            | 1.5338         |
| 30.434            | 1.5338         |
| 42.772            | 1.5338         |

\[ b) c_t = \min(\tilde{c}^* + \tilde{m}^* \sqrt{V_{i,t}}^2, K), K = 2\% \]

| \(\sqrt{V_{i,t}}\) | \(\tilde{m}^*\) |
|-------------------|----------------|
| 8.702             | 1.5338         |
| 13.882            | 1.5338         |
| 30.434            | 1.5338         |
| 42.772            | 1.5338         |

\[ c) c_t = \tilde{c}^* + \tilde{m}^* \sqrt{V_{i,t}} \]

| \(\sqrt{Y_{i,t}}\) | \(\tilde{m}^*\) |
|-------------------|----------------|
| 8.702             | 1.5338         |
| 13.882            | 1.5338         |
| 30.434            | 1.5338         |
| 42.772            | 1.5338         |

Table 6. Fair fee process under the Heston model.

| Fee Structure | \(c_t, 0 \leq t \leq T\) |
|---------------|--------------------------|
| Uncapped VIX\(^2\) | \(\tilde{c}^* + \tilde{m}^* (A + BV_t)\) |
| Capped VIX\(^2\) | \(\min(K, \tilde{c}^* + \tilde{m}^* (A + BV_t))\) |
| Uncapped VIX | \(\tilde{c}^* + \tilde{m}^* \sqrt{(A + BV_t)}\) |

Using the CTMC technique outlined in section 3 and the market, VA, and CTMC parameters of subsection 5.1, we perform the valuation of a variable annuity with and without early surrenders (‘VA with ES’ and ‘VA without ES’, respectively). The results are reported in table 7 below.\(^\dagger\)

\(^\dagger\) Numerical experiments under the Heston model have been performed using equation (22), and algorithms 2 and 4. Note however that similar results are obtained when using the Fast Algorithms, see appendix B, available online for details.

First, we observe that the fair value of the variable annuity without early surrenders is approximately \(F_0 = 100\) for all fair fee vectors. This is because fair fee parameters are calibrated such that the value of the VA without surrender rights at time \(t = 0\) is equal to the initial premium (\(F_0 = 100\)). Moreover, under the Uncapped VIX\(^2\)-linked structure, fair fee parameters are obtained using the exact pricing formula of Cui et al. (2017), confirming the accuracy of the approximated model. The absolute error is around \(10^{-4}\) for all fee vectors for this fee structure. An accuracy of around \(10^{-3}\) can be obtained for each value (the value of the VA with and without surrender rights) with fewer grid points with significantly less computational effort. Detailed results with \(N = 100\) and \(N = 1000\) are given in appendix B (available online as supplemental material) with their respective computation time. The value of the surrender right is calculated as the difference between the values of the variable annuity with and without early surrenders.

As \(\tilde{m}^*\) increases, the risk-neutral value of early surrenders remains very close for all fee structures. One might expect...
the VIX-linked fee structure to reduce the risk-neutral value of the surrender rights since this type of fee structure realigns income and liability (Cui et al. 2017), but this is not what is observed here. However, VIX-linked fee structures have an impact on optimal surrender strategies, as shown below.

In figure 2, the shape of the approximated optimal surrender surface associated with each of the VIX-linked fee structures is illustrated for different values of the fair multiplier. We observe that the surrender region is of threshold type for all $\tilde{m}^\ast$. We also note that, regardless of the value of $\tilde{m}^\ast$, the boundary is a concave function of time. It slowly increases to a maximum and decreases rapidly to the guarantee level $G$. This is consistent with earlier findings of Bernard et al. (2014) and Kang and Ziveyi (2018).

For fixed $t \in [0, T]$, we observe, in figures 2 and 3, that the $t$ section of $f^{(m,N)}$, the function $y \mapsto f^{(m,N)}(t,y)$, denoted by $f_{t}^{(m,N)}$, is increasing for all $\tilde{m}^\ast$. Hence, when the volatility is high, variable annuities are surrendered at higher fund values than when the volatility is low. This is in line with the findings of Kang and Ziveyi (2018). However, as $\tilde{m}^\ast$ increases, the function $f^{(m,N)}$ increases at a slower rate (and particularly for the uncapped structures, panel (a) and (c) of figures 2 and 3). For instance, fix $\tilde{m}^\ast \in [0, 0.15, 0.3, 0.4345]$, $y \in \mathbb{R}^{(m)}$ and note, from figure 3 (panel (a) or (c)), that the $y$ section of $f^{(m,N)}$, the function $t \mapsto f^{(m,N)}(t,y)$, denoted by $f_{y}^{(m,N)}$, is pushed upwards as $y$ increases. However, the difference between the low and the high volatility $y$ section is less significant as $\tilde{m}^\ast$ grows. This means that the optimal surrender decision for uncapped VIX-linked fees is less sensitive to volatility fluctuations when fees are tied to the volatility index. In the case of the capped structure (panel (b) of figures 2 and 3), we also note that approximated optimal surrender surfaces are gradually increasing as volatility grows; however, they now increase at a similar pace for all $\tilde{m}^\ast$. This can be interpreted from a financial perspective as pointed out below.

In figure 4, we fix a volatility level $y$ and we compare the $y$ sections of $f^{(m,N)}$, $f^{(m,N)}$, for different fair multipliers $\tilde{m}^\ast$. When the volatility is low ($\sqrt{\gamma} = 8.702\%$), see for instance the first graph of figure 4(a), $f^{(m,N)}$ is pushed upwards as $\tilde{m}^\ast$ increases. This means that a variable annuity contract with a fully dependent uncapped VIX2-linked fee structure ($\tilde{m}^\ast = 0.4345$) is surrendered at higher fund values than the constant fee one ($\tilde{m}^\ast = 0$). Indeed, when the volatility is low, the fee paid under a variable annuity contract with a VIX2-linked fee structure is also low (see table 5 for examples), making VIX-linked fee contracts more attractive than the constant fee ones. However, as the volatility rises, VIX2-linked fees also rise and so, the relation between the optimal surrender decision and $\tilde{m}^\ast$ reverts. The second graph of figure 4(a) shows that variable annuity contracts are surrendered at almost all the same fund value levels when the volatility equals to 13.882%, regardless of $\tilde{m}^\ast$. The latter may be explained by the fact that fee rates are all around the same level when $\sqrt{\gamma} = 13.882\%$, that is $\pm 1\%$ as per table 5. However, when the volatility increases (see for instance the third and the last graphs of figure 4(a)), VIX2-linked fees are also high (refer again to table 5 for examples), making VIX2-linked fee variable annuity contracts less attractive than the constant fee ones. And so, when volatility is high, we observe that $f^{(m,N)}$ is pushed downward with increasing values of $\tilde{m}^\ast$. In other words, for high volatility levels, variable annuity contracts with VIX2-linked fee structures are surrendered at lower fund values than contracts with constant fee structures. This is financially intuitive, as pointed out by Bernard et al. (2014) under the Black-Scholes setting with a constant fee function, since when the fee gets higher, the policyholder has to pay more for the guarantee, and so, the mismatch between the premium for guarantee and its value is even greater; resulting in earlier exercise time. The analysis above shows that the findings of Bernard et al. (2014) also extend to stochastic fee structures. Similar conclusions can be drawn for the Uncapped VIX-linked fees, figure 4(c).
Figure 2. Approximated optimal surrender surface of VIX-linked fees VAs for different values of fair multiplier $\tilde{m}^*$ under the Heston model. The x-axis represents the time and the y-axis the variance. (a) $c_t = \tilde{c}^* + \tilde{m}^* \mathrm{VIX}_t^2$ (b) $c_t = \min(\tilde{c}^* + \tilde{m}^* \mathrm{VIX}_t^2, K)$, $K = 2\%$ (c) $c_t = \tilde{c}^* + \tilde{m}^* \mathrm{VIX}_t$.

Figure 3. The y section of the approximated optimal surrender surface, $f_y^{(m,N)}$, for different volatility levels $\sqrt{y}$ and fair multipliers $\tilde{m}^*$. (a) $c_t = \tilde{c}^* + \tilde{m}^* \mathrm{VIX}_t^2$ (b) $c_t = \min(\tilde{c}^* + \tilde{m}^* \mathrm{VIX}_t^2, K)$, $K = 2\%$ (c) $c_t = \tilde{c}^* + \tilde{m}^* \mathrm{VIX}_t$.
We observe in figure 4(b) that when the volatility is low, capped VIX\(^2\)-linked fee VA contracts are surrendered at lower fund values than constant fee ones, like for the two other fee structures. However, for high enough volatility levels, the cap is reached, and thus, the fee paid under the capped structure is the same for all \(m^* > 0\) (see table 5); capped VIX\(^2\)-linked fee VA contracts are thus all surrendered at similar fund value level (figure 4(b) graphs 3 and 4 when \(m^* > 0\)). This illustrates again the relation that exists between fees and optimal surrender decisions. VA contracts with higher fee rates are surrendered at lower fund values than contracts with lower fee rates.

Similar conclusions can be drawn under the Heston model when \(\bar{g} = F_0 e^{\tilde{g}T}\) with \(\tilde{g} = 2\%\) or when the interest rates depend on time; and under the 3/2 stochastic volatility model, see appendix B, available online as supplemental material, for details.

6. Conclusion

In this paper, we provide a framework based on CTMC approximations to analyze the surrender incentives resulting from VIX-linked fees in variable annuities under general stochastic volatility models. Under the assumption that the policyholder maximizes the risk-neutral value of her variable annuity, the pricing of a variable annuity is an optimal stopping problem with a time-discontinuous reward function. Under general fee and surrender charge structures, we develop efficient numerical algorithms based on a two-layers CTMC approximation to price variable annuities with and without early surrenders. We derive a closed-form analytical formula for the value of a variable annuity without surrender rights and provide a quick and simple way of determining early surrenders value via a recursive algorithm. We also present an easy procedure to approximate the optimal surrender surface under the hypothesis that the surrender region is of threshold type. Finally, we observe numerically that VIX-linked fees do not significantly affect the value of early surrenders under the selected set of Heston parameters. However, numerical examples also reveal that the optimal surrender decision is impacted by VIX-linked fee structures. In particular, we observe that the optimal surrender strategy is more stable with respect to volatility changes when the fees are linked to the volatility index.

All algorithms and results of this paper can easily be adapted to incorporate the term structure of interest rates by modeling the risk-free rate as a deterministic function of time. However, extension to stochastic interest rate models can present some numerical challenges. Indeed, adding a third dimension to the problem necessitates increasing the size of the generator, which can cause numerical challenges specifically for the valuation of long-term derivatives such as variable annuities. Algorithms 1 and 3 of this paper address the numerical challenge encountered when trying to value long-term derivatives under general two-dimensional models. The extension to higher dimension models is not straightforward and is left as future research. Moreover, short-rate models that allow to reproduce the term structure of interest rates are time-inhomogeneous, adding to the numerical difficulty of this extension for long-maturity derivatives since matrix exponential of the time-dependent generator now needs to be calculated at each time step to obtain the transition
probability matrices. Finally, the CTMC approximation of the VIX presented in section 4.4 can also be used to price exotic path-dependent options on the VIX, which is left as future research.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The first author acknowledges support from the National Science and Engineering Research Council of Canada [grant number RGPIN-2016-04869] and from the Fonds de recherche du Québec – Nature et Technologies [grant number 253180]. The second author was supported by Fonds de recherche du Québec – Nature et Technologies [grant number RGPIN-2016-04869] and from the Natural Sciences and Engineering Research Council of Canada. The first author acknowledges support from the National Natural Science Foundation of China [grant number 257320]. The third author acknowledges partial support from recherche du Québec – Nature et Technologies [grant number 3310365].

Supplemental data for this article can be accessed online at Supplemental data [grant number CN2-2113906].

Supplemental data

Supplemental data for this article can be accessed online at http://dx.doi.org/10.1080/14697688.2023.2215278.

References

Aït-Sahalia, Y. and Kimmel, R., Maximum likelihood estimation of stochastic volatility models. *J. Financ. Econ.*, 2007, **83**(2), 413–452.
Alonso-García, J., Sherris, M., Thirurajah, S. and Ziveyi, J., Taxation and policyholder behavior: The case of guaranteed minimum accumulation benefits. Working Paper, 2022. Available at SSRN 3629101.
Bacinello, A.R., Millossovich, P., Olivieri, A. and Pitacco, E., Variable annuities: A unifying valuation approach. *Insurance: Math. Econ.*, 2011, **49**(3), 285–297.
Bassan, B. and Ceci, C., Optimal stopping problems with discontinuous reward: Regularity of the value function and viscosity solutions. *Stoch. Stochastic Rep.*, 2002, **72**(1-2), 55–77.
Bauer, D., Kling, A. and Russ, J., A universal pricing framework for guaranteed minimum benefits in variable annuities. *ASTIN Bulletin: J. IAA*, 2008, **38**(2), 621–651.
Bauer, D., Gao, J., Moenig, T., Ulm, E.R. and Zhu, N., Policyholder exercise behavior in life insurance: The state of affairs. *N. Am. Actuar. J.*, 2017, **21**(4), 485–501.
Bernard, C. and MacKay, A., Reducing surrender incentives through fee structure in variable annuities. In *Innovations in Quantitative Risk Management*, edited by K. Glaü, M. Scherer, and R. Zagst, pp. 209–223, 2015 (Springer: Cham).
Bernard, C. and Moenig, T., Where less is more: Reducing variable annuity fees to benefit policyholder and insurer. *J. Risk Insurance*, 2019, **86**(3), 761–782.
Bernard, C., Hardy, M. and MacKay, A., State-dependent fees for variable annuity guarantees. *ASTIN Bulletin: J. IAA*, 2014, **44**(3), 559–585.
Billingsley, P., *Probability and Measure*, 3rd ed., 1995 (John Wiley & Sons, Inc.: New York).
Billingsley, P., *Convergence of Probability Measures*, 2nd ed., 1999 (John Wiley & Sons, Inc.: New York).
Cai, N., Song, Y. and Kou, S., A general framework for pricing Asian options under Markov processes. *Oper. Res.*, 2015, **63**(3), 540–554.
Cai, N., Kou, S. and Song, Y., A unified framework for regime-switching models. Working Paper, 2019. Available at SSRN 3310365.
Carr, P. and Sun, J., A new approach for option pricing under stochastic volatility. *Rev. Derivatives Res.*, 2007, **10**, 87–150.
Carr, P., Jarrow, R. and Myneri, R., Alternative characterizations of American put options. *Math. Finance*, 1992, 2(2), 87–106.
CBOE, VIX for variable annuities: A study considering the advantages of tying a variable annuity fee to VIX. White Paper, 2013a.
CBOE, VIX for variable annuities – part II: A study considering the advantages of tying a variable annuity fee to VIX. White Paper, 2013b.
Chourdakis, K., Non-affine option pricing. *J. Derivatives*, 2004, **11**(3), 10–25.
Cui, Z., Martingale property and pricing for time-homogeneous diffusion models in Finance. PhD Thesis, University of Waterloo, 2013.
Cui, Z., Feng, R. and MacKay, A., Variable annuities with VIX-linked fee structure under a heston-type stochastic volatility model. *N. Am. Actuar. J.*, 2017, **21**(3), 458–483.
Cui, Z., Kirkby, J.L. and Nguyen, D., A general valuation framework for SABR and stochastic local volatility models. *SIAM J. Financial Math.*, 2018, **9**(2), 520–563.
Cui, Z., Kirkby, J.L. and Nguyen, D., Continuous-time Markov chain and regime switching approximations with applications to options pricing. In *Modeling, Stochastic Control, Optimization, and Applications*, edited by G. Yin and Q. Zhang, pp. 115–146, 2019 (Springer: Hoboken).
Cui, Z., Lee, C. and Liu, M., Valuation of VIX derivatives through combined Ito-Taylor expansion and Markov chain approximation. Working Paper, Stevens Institute of Technology, 2021.
Da Fonseca, J. and Martini, C., The a-hypergeometric stochastic volatility model. *Stoch. Process. Their Appl.*, 2019, **129**(5), 1472–1502.
De Angelis, T. and Stabile, G., On lipschitz continuous optimal stopping boundaries. *SIAM J. Control Optim.*, 2019, **57**(1), 402–436.
Delong, L., Pricing and hedging of variable annuities with state-dependent fees. *Insurance: Math. Econom.*, 2014, **58**, 24–33.
Ding, K. and Ning, N., Markov chain approximation and measure change for time-inhomogeneous stochastic processes. *Appl. Math. Comput.*, 2021, **392**, 125732.
Drimus, G.G., Options on realized variance by transform methods: A non-affine stochastic volatility model. *Quant. Finance*, 2012, **12**(11), 1679–1694.
Ethier, S.N. and Kurtz, T.G., *Markov Processes: Characterization and Convergence*, 2005 (John Wiley & Sons Inc.: Hoboken).
Garcia, R., Lewis, M.-A., Pastorrello, S. and Renault, É., Estimation of subjective and risk-neutral distributions based on moments of integrated volatility. *J. Econom.*, 2011, **160**(1), 22–32.
Grasselli, M., The 2/2 stochastic volatility model: A unified approach for the Heston and the 3/2 model. *Math. Finance*, 2017, **27**(4), 1013–1034.

Supplemental data for this article can be accessed online at http://dx.doi.org/10.1080/14697688.2023.2215278.

ORCID

Anne Mackay © http://orcid.org/0000-0002-8495-1983
Zhenya Cui © http://orcid.org/0000-0001-9429-7121

A. Mackay et al.
A.1. Proof of proposition 4.3

First, recall that

\[ P(X_{t+h}^{(mN)} = x_i | Y_{t+h}^{(mN)} = v_j | Y_t^{(mN)} = x_i, Y_t^{(mN)} = v_k) = \begin{cases} \frac{1}{N} & \text{if } j = k \text{ and } x_i = v_j, \\ 0 & \text{otherwise.} \end{cases} \]

Mentink-Vigier, F., Fast exponential matrix for Matlab (full/sparse), fastExpm. Available online at: Github https://github.com/fmentink/fastExpm (accessed 3 February 2023).

Mijatović, A. and Pistorius, M., Continuously monitored barrier options under Markov processes. Available at SSRN 1462822.

Mijatović, A. and Pistorius, M., Continuously monitored barrier options under Markov processes. *Math. Finance*, 2013, 23(1), 1–38.

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

Niiutiuni Perei, J., Chapter 18 – policyholder behavior and management actions. In *IAA Risk Book*, 2022 (International Actuarial Association: Ottawa).

Palczewski, J. and Stettner, L., Finite horizon optimal stopping of time-discontinuous functionals with applications to impulse control with delay. *SIAM J. Control Optim.*, 2010, 48(8), 4874–4909.

Rebonato, R., *Volatility and Correlation: The Perfect Hedges and the Fox*, 2nd ed., 2004 (John Wiley & Sons Ltd: Chichester).

Schweizer, M., On Bermudan options. In *Advances in Finance and Stochastics*, edited by K. Sandmann and P. J. Schönbucher, pp. 257–270, 2002 (Springer: Heidelberg).

Scott, L.O., Option pricing when the variance changes randomly: Theory, estimation, and an application. *J. Financial Quant. Anal.*, 1987, 22(4), 419–438.

Shen, Y., Sherris, M. and Ziveyi, J., Valuation of guaranteed minimum maturity benefits in variable annuities with surrender options. *Insurance: Math. Econom.*, 2016, 69, 127–137.

Sidje, R.B., Expokit: A software package for computing matrix exponentials. *ACM Trans. Math. Softw.*, 1998, 24(1), 130–156.

Sin, C.A., Complications with stochastic volatility models. *Adv. Appl. Probab.*, 1998, 30(1), 256–268.

Song, Q., Yin, G. and Zhang, Q., Weak convergence methods for approximation of the evaluation of path-dependent functionals. *SIAM J. Control Optim.*, 2013, 51(5), 4189–4210.

Tavella, D. and Randall, C., *Pricing Financial Instruments: The Finite Difference Method*, 2000 (John Willey & Sons Inc.: New York).

Terenzi, G., Option prices in stochastic volatility models. PhD Thesis, Université Paris-Est Marne-La-Vallée, 2018.

Touzi, N., American options exercise boundary when the volatility changes randomly. *Appl. Math. Optim.*, 1999, 39, 411–422.

Wang, G. and Zou, B., Optimal fee structure of variable annuities. *Insurance: Math. Econom.*, 2011, 51(4), 419–4210.

Williams, D., *Probability with Martingales*, 1991 (Cambridge University Press: Cambridge).

Zhang, G. and Li, L., A general approach for parisian stopping times under Markov processes. Working Paper, 2021. Available at arXiv:2107.06605.

Zhu, Y. and Zhang, J.E., Variance term structure and VIX futures pricing. *Int. J. Theoretical Appl. Finance*, 2007, 10(1), 111–127.
By inspection of the matrix $G^{[m,N]}$ and since $h$ is small, we have by (8) that
\[
\mathbb{P}(X_{t+h}^{(m,N)} = x_i, V_{t+h}^{(m)} = v_j \mid X_t^{(m,N)} = x_i, V_t^{(m)} = v_j) = \begin{cases} 
q_{ik} + c^j_k(h) & \text{if } i = l, j \neq k \\
q_{ik} + \lambda_{ij}^k h + c^j_k(h) & \text{if } i = l, j = k \\
1 + (q_{ik} + \lambda_{ij}^k h + c^j_k(h)) & \text{if } i \neq l, j \neq k.
\end{cases}
\]

where the functions $\{c^j_k\}$ satisfy $\lim_{h \to 0} c^j_k(h) = 0$ for $i, l \in \{1, 2, \ldots, N\}$ and $k, j \in \{1, 2, \ldots, m\}$.

From the last equality, we observe that the regime-switching CTMC cannot change regime $(V_t)$ and state $(X_t^{(m,N)})$ simultaneously over small time intervals. It follows that
\[
\mathbb{E}_t \left[ \phi(t + h, X_t^{(m,N)}, V_t^{(m)}) \mid X_t^{(m,N)} = x_i, V_t^{(m)} = v_k \right] = \sum_{j=1}^m \phi(t + h, x_i, v_j)(q_{jk} + c^j_k(h)) \\
+ \sum_{l=1}^N \sum_{j=1}^m \phi(t + h, x_i, v_l)(\lambda_{ij}^l h + c^j_l(h)) \\
+ \phi(t + h, x_i, v_k)(1 + \lambda_{ik}^k h + c^k_k(h)) \\
= \sum_{j=1}^m \phi(t + h, x_i, v_j) + \sum_{l=1}^N \phi(t + h, x_i, v_l)q_{lk} + \sum_{j=1}^m \phi(t + h, x_i, v_l)(\lambda_{ij}^l h + c^j_l(h)) \\
+ \lambda_{ik}^k h + c^k_k(h) + \sum_{j=1}^m \phi(t + h, x_i, v_j)q_{jk}h \\
+ \sum_{l=1}^N \phi(t + h, x_i, v_l)q_{lk}h.
\]

The results follows from setting $c(h) = c_m(h) - c_N(h)$.

### A.2. Proof of proposition 4.5

In order to prove proposition 4.5, we first need to introduce some additional notation. Suppose that $t = 0$. For $0 \leq s \leq T$, define the discounted reward process of the original and the modified contracts by

\[ Z_s = e^{-r_s} \psi(s, F_s, V_s), \quad \text{and} \quad \tilde{Z}^{(M)}_s = Z_s \mathbf{1}_{\{s \in T_M\}}, \]

respectively. Let the processes $\{J_t^{(M)}\}_{0 \leq t \leq T}$ and $\{\tilde{J}_t^{(M)}\}_{0 \leq t \leq T}$ be defined by

\[
J_t^{(M)} = \text{ess sup}_{t \in T_M} \mathbb{E}_t[Z_t], \quad \text{and} \quad \tilde{J}_t^{(M)} = \text{ess sup}_{t \in T_M} \mathbb{E}_t[\tilde{Z}_t^{(M)}], \quad (A1)
\]

respectively. As explained in Schwarz (2002), since the Bermudan contract cannot be exercised outside the region of permitted exercise times and $Z$ is non-negative, we expect that the value of the Bermudan contract with the payoff process $Z$ to be equal to an American contract with the payoff process $\tilde{Z}^{(M)}$. That is, we expect $J_t^{(M)}$ to be equal to $\tilde{J}_t^{(M)}$, almost surely. This idea is formalized in Schwarz (2002, proposition 3) which is restated in the following lemma.

**Lemma A.1** (Schwarz 2002, proposition 3) Fix $M \in \mathbb{N}$ and let $J_t^{(M)}$ and $\tilde{J}_t^{(M)}$ be defined as in (A1), then $J_t^{(M)} = \tilde{J}_t^{(M)}$, almost surely.

We can now prove proposition 4.5.

**Proof of proposition 4.5.** The following is inspired by the proof of proposition 6 of Bassan and Ceci (2002). Without loss of generality, suppose that $t = 0$. From lemma A.1, we have that $J_t^{(M)} = \tilde{J}_t^{(M)}$, almost surely. Hence, the Bermudan contract in (25) may be expressed as follows:

\[
b_M(0, x, y) = \sup_{\tau \in T_M} \mathbb{E}_0[\tilde{Z}_\tau^{(M)}].
\]

Now notice that for each $t \geq 0$, $\tilde{Z}_t^{(M)}(\omega) \uparrow Z_t(\omega)$ for all $\omega \in \Omega$. Thus, given $F_0 = x$ and $V_0 = y$ and by using the monotone convergence theorem (Williams 1991, theorem 5.3), we obtain

\[
\nu(0, x, y) = \sup_{\tau \in T_M} \mathbb{E}_0[Z_\tau] = \sup_{\tau \in T_M} \mathbb{E}_0[\tilde{Z}_\tau^{(M)}] \\
= \sup_{M \in \mathbb{N}} \sup_{\tau \in T_M} \mathbb{E}[\tilde{Z}_\tau^{(M)}] \\
= \sup_{M \in \mathbb{N}} \mathbb{E}[\tilde{Z}_\tau^{(M)}] = \sup_{M \in \mathbb{N}} b_M(0, x, y) \\
= \lim_{M \to \infty} b_M(0, x, y).
\]