Insurance: Mathematics and Economics 53 (2013) 237–251

Contents lists available at SciVerse ScienceDirect

Insurance: Mathematics and Economics

journal homepage: www.elsevier.com/locate/ime

A heavy traffic approach to modeling large life insurance portfolios

Jose Blanchet[a], Henry Lam[b,*]

[a] Department of Industrial Engineering and Operations Research, Columbia University, 120th Street, New York, NY 10027, USA

[b] Department of Mathematics and Statistics, Boston University, 111 Cummington Street, Boston, MA 02215, USA

HIGHLIGHTS

- We propose a bottom-up approach to model large life insurance portfolios.
- We use heavy-traffic approximation to derive and justify the structure of the risk processes.
- The risk processes are shown to depend on mortality and insurance contract structure in a tractable manner.
- We formulate and compute ruin probability that takes actuarial reserve into account.
- We identify explicitly the temporal and cross-sectional correlation structure of the derived risk processes.

ARTICLE INFO

Article history:
Received February 2011
Received in revised form
April 2013
Accepted 30 April 2013

ABSTRACT

We explore a new framework to approximate life insurance risk processes in the scenario of plentiful policyholders, via a bottom-up approach. Given the insurance contract structure, we aggregate the balance of individual policy accounts, and derive an approximating Gaussian process with computable correlation structure. The methodology is borrowed from heavy traffic theory in the literature of many-server queues, and involves the so-called fluid and diffusion approximations. Our framework is different from the individual risk model in that it takes into account the time dimension and the specific policy structure including the premium payments. It is also different from classical risk theory in that it builds the risk process from micro-level contracts and parameters instead of assuming aggregated claim and premium processes outright. As a result, our approximating process behaves differently depending on the issued contract structure. We also illustrate the flexibility of our approach by formulating a finite-horizon ruin problem that incorporates actuarial reserve in the consideration.

© 2013 Elsevier B.V. All rights reserved.

The study of risk processes is a central topic in actuarial science. Most of the literature focuses on the calculation of ruin probability and deficits (or overshoots) at the time of ruin, as well as the optimal control of premiums, reinsurance levels, and investment allocation. These questions have been studied under a variety of stochastic settings, from the classical Cramer–Lundberg approximation to diffusion processes. The central theme is that random-walk-type models, with a negatively drifted premium process and a jump process of claims, provide a rich framework to allow plenty of extensions, modifications and problem formulations (see, for example, Asmussen and Albrecher, 2010 for the survey on ruin probability calculations, and Schmidli, 2008 for the counterpart in stochastic control problems).

In this paper, we take a different view from the existing literature. Rather than focusing on the computation of risk-related quantities, we explore the question of the construction of risk process itself. The approach we use is bottom-up: given the structure and parameters of the individual insurance contracts, how does the risk process of the insurer look like on an aggregate scale?

Naturally, the risk process under this framework is the sum of all the individual accounts i.e. the balances of policyholders who entered contract with the insurer over time. For actuaries, this points to the standard one-period individual and collective risk models. However, these standard models do not consider the time dimension. This in turn also restrains the power of such models to capture the specific contract structure involved e.g. the premium payments.

In this regard, our work can be seen as a generalization of the standard risk models to a process-level approximation. Of course, mere summation of all individual accounts might end up getting an unpleasant process that is hardly computable. To tackle this issue, we borrow techniques in so-called heavy traffic theory in the queueing literature. The basic idea is that under the assumption of large number of customers or policyholders, one can approximate the functionals of these policyholders' statuses using fluid and diffusion approximations. In the statistics literature, these correspond

* Corresponding author.
E-mail addresses: jose.blanchet@columbia.edu (J. Blanchet), khhlam@gmail.com, khlam@bu.edu (H. Lam).

0167-6687/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.insmatheco.2013.04.011

to stochastic-process versions of Law of Large Numbers and Central Limit Theorems. With the sheer scale of major insurance companies, the assumption of plentiful policyholders is sensible, and so these approximation techniques can be used. As we will see, these heavy traffic approximation would then lead to a Gaussian process that is as analyzable as many standard processes used in the current risk theory literature. In particular, the correlation structure of this Gaussian process is explicitly computable given the contract structure (see Section 4). To illustrate our argument on tractability, we formulate a finite-horizon ruin problem based on our Gaussian approximation (see Section 3).

We distinguish our contribution from the classical risk theory and standard actuarial risk models in a few ways. First, our model explains how individual insurance policies lead to certain features of the aggregate risk process. The construction of our risk process depends intricately on the premium and benefit structure of single policies. This means that different types of insurance, such as whole life insurance, term life, endowment etc. would lead to different correlation structure of our resulting Gaussian process. This is in sharp contrast to the current model in risk theory, where premium and claim processes are modeled separately, each as a drifted random walk (or its variants) and marked point process. This feature can potentially provide a framework to analyze the effect of contract structure on the firm-wide risk level. Second, our model allows naturally the incorporation of actuarial reserve in our approximation. Indeed, the finite-horizon ruin problem that we formulate in Section 3 will involve the calculation of prospective reserve. Third, since serial correlation is explicitly computable, this provides a way to capture the fluctuation of our approximating process over time, which can be potentially applicable to dynamically monitoring mismatch on the insurer's balance sheet with regard to statistical error.

In a more organized fashion, we summarize our contributions as follows:

(A) Under the assumption of large number of policyholders, we construct the fluid limit and diffusion limit for the aggregate risk processes. (As we mentioned, these correspond to functional Law of Large Numbers and Central Limit Theorem respectively in the statistics community; throughout the paper we mostly use the former terminology to align with the queueing literature, but will also use the latter interchangeably when necessary.) The risk processes that we are interested in include the insurer's cash level, liabilities, and per basis reserve level. These will be discussed in Section 2. We prove and numerically demonstrate that these risk processes can be approximated by Gaussian processes with certain correlation structures.

(B) Using the theory of Gaussian processes, we illustrate how our result can be used to approximate the ruin probabilities. We model ruin as the situation in which the liabilities surpass the assets (plus the initial capital) within a given time horizon (see Section 3.1). This highlights the flexibility of our methodology in incorporating reserve calculation, and also the dependency on the underlying insurance contracts. In particular, we apply our results to several common types of insurance.

(C) Our diffusion approximation shows how, under the Equivalence Principle, the benefit reserve arises as the fluid limit of the empirical cash level per basis at any point in time (see Section 2). These results, we believe, provide a useful perspective into the basic concepts underlying the definition of benefit reserve; see the discussion following Theorem 1.

(D) We compute the correlation structures of our limiting processes, thereby showing their tractability. In particular, we illustrate how our approach allows to evaluate and compare the autocorrelation (as a function of time) of risk processes with different insurance types; see Section 4.

Let us emphasize that our purpose in applications such as (B) and (C) is to illustrate the concepts behind our ideas, and hence the models we are using in this paper are basic. There are certainly many practical considerations to make the model more realistic. We shall list out these generalizations and more realistic extensions that we believe are worth pursuing in Section 5.

In terms of methodology, as aforementioned, we will invoke primarily the machinery in heavy traffic theory i.e. fluid and diffusion approximations in the queueing literature. The ideas date back to Kingman (1961, 1962) for single-server queues, and they still constitute an active research area among the queueing theorists (see the standard surveys of Whitt, 2002 and Billingsley, 1999 for instance). Under fairly mild assumptions, the tools significantly simplify and single out the important elements of the system dynamics of interest, and provide approximate solutions to many important performance measures (in our context, the ruin probability mentioned in (B) constitutes one such example). More precisely, the results in this paper relate to the analysis of so-called many-server queues, which have been substantially studied in recent years. In these queueing systems, customers arrive and elicit service for a random amount of time, as long as there are available servers. When the number of servers is infinite, every customer can start service right at arrival. Connecting to our work, policyholders can be thought of as customers in the queueing system. While the feature of arrivals is not our focus in this paper, the death time of policyholders is analogous to the end of service, and hence the approximation technique is translatable. Some relevant references on the topic include Pang and Whitt (2010) and Decreusefond and Moyal (2008), which focus on infinite-server models, Halfin and Whitt (1981), Kaspi and Ramanan (2010) and Reed (2009), which study finite but large number of servers in different proportion (or so-called regime) to the number of customers, Puhalskii and Reiman (2000) that study queues with multiclass customers, and Dai et al. (2010) on queues with reneging. The common theme of all these work is the heavy traffic technique being applicable to various features of the queues.

Finally, we discuss two papers that use similar approach and highlight our difference. One is a recent working paper by Bensusan and El Karoui (2009), who propose a microstructural approach to model population dynamics to capture mortality/longevity risk. Their motivation is different from ours: instead of building our mortality distribution microstructurally, we make common assumptions on mortality; instead, our focus is on how this mortality assumption, under the interaction with the contract structure, benefit level and premium calculation, leads to a macroscopic fluctuation of total assets, liabilities and other actuarial quantities. Secondly, we note that diffusion approximation has been invoked by Iglehart (1969) in arguing the use of Brownian motion in modeling insurance risk process. However, he maintained a Cramer–Lundberg framework by assuming compound Poisson claims and constantly drifted premium, and showed that under certain scaling their difference converges to a diffusion process. Contract structure, relation between premium and benefit, and actuarial reserve etc. were not considered in his work.

The organization of this paper is as follows. In Section 1 we lay out our model assumptions and define the key quantities that we approximate. Section 2 is devoted to the statement of our main result and its discussion. Section 3 relates to applications in ruin probability computations and shows some examples. Section 4 identifies the autocorrelation structure of our approximating Gaussian processes. Section 5 discusses some extensions. Appendix constitutes an appendix, which is divided into two parts. The first part discusses basic facts about heavy traffic limit theorems and gives the proof of our main result; the second part contains a discussion on the simulation methodology that is used to generate various examples in this paper.

1. Model, assumptions and basic quantities

We consider a portfolio of n independent policyholders at time 0. For simplicity assume that policyholders have the same profile i.e. identical mortality distribution. This assumption is made mainly for simplicity, the extensions of which will be discussed later in the paper. Also we assume a constant rate of interest. We use the following notations throughout the paper:

- δ: constant rate of interest (continuous compounding)
- X_i: the death time of the i-th policyholder (the X_i's are independent and identically distributed (i.i.d.))
- $f(t), F(t), \bar{F}(t)$: density, distribution and survival functions of X_i
- T: upper limit of the support for the death time i.e. $T = \sup\{t > 0 : F(t) < 1\}$
- $P(t)$: accumulated premium payment discounted at time 0 if the policyholder dies at t
- $B(t)$: benefit payment discounted at time 0 if the policyholder dies at t; note that, the case where benefits are paid at times other than death (such as regular bonus prior to death) can be merely redefined as a deduction in accumulated premium payment, and hence is also covered in our framework.

In addition, we make the following technical assumptions that are commonly used:

Assumption 1. We assume that $f(t) > 0$ for all $t \in (0, T)$, with $T < \infty$.

Assumption 2. Define $H(t) := P(t) - B(t)$. We assume that $P(\cdot)$ and $B(\cdot)$ are continuously differentiable and have bounded first derivatives almost everywhere (with respect to the Lebesgue measure), and hence so is $H(\cdot)$.

Assumption 1 is natural in the setting of life insurance, which is the focus of this paper. Assumption 2 is satisfied by all common insurance contracts. For example, in the case of whole life insurance with continuous level premium payment p and benefit b, $P(t) = \int_0^t pe^{-\delta s}ds = p(1 - e^{-\delta t})/\delta$ and $B(t) = be^{-\delta t}$, which clearly satisfy Assumption 2.

We now look at some basic quantities of interest that are related to the n policyholders with assumptions described above. To keep our discussion simple for illustration, throughout the paper we will focus on this setting. There are many natural extensions, such as the arrivals of policyholders over time and multi-profile multi-product business lines. These will be left for future exploration. A companion paper by Blanchet and Lam (2011) discusses the scenario of policyholder arrivals.

Let $N_n(t)$ be the number of deaths before time t. With the notation above, we write

$$N_n(t) = \sum_{i=1}^{n} I(X_i \leq t). \tag{1}$$

Similarly, we write $\bar{N}_n(t)$ for the number of surviving policyholders at time T, namely

$$\bar{N}_n(t) = N_n(T) - N_n(t) = n - N_n(t). \tag{2}$$

Our results involve the following three basic quantities of interest, all of which can be expressed in terms of (1) and (2) above. For convenience, we name these quantities as Total Cash Process, Total Reserve Process and Average Cash Process respectively:

Total Cash Process. We define the *Total Cash Process* as the present value at time t of the total accumulated cash generated by all individual accounts, excluding the initial surplus. We denote it by $C_n(t)$:

$$C_n(t) := e^{\delta t} \sum_{i=1}^{n} [(P(X_i) - B(X_i))I(X_i \leq t) + P(t)I(X_i > t)].$$

Observe that we can write more neatly as

$$C_n(t) = e^{\delta t} \left[\int_0^t H(s)dN_n(s) + P(t)\bar{N}_n(t) \right].$$

We also define $m(t)$ to be the mean of the cash contribution from an individual account over time i.e.

$$\begin{aligned} m(t) &= E[(P(X_i) - B(X_i))I(X_i \leq t) + P(t)I(X_i > t)] \\ &= e^{\delta t} \left[\int_0^t H(s)f(s)ds + P(t)\bar{F}(t) \right]. \end{aligned} \tag{3}$$

The Equivalence Principle indicates that one should select the premium level in such a way that the total (i.e. up to the end of the time horizon) actuarial net present value of the premiums is equal to that of the benefits paid (see Bowers et al., 1997). In our notation, assuming the validity of the Equivalence Principle amounts to saying that $m(T) = e^{\delta T} \int_0^T H(s)f(s)ds = 0$.

Total Reserve Process. The actuarial reserve at time t of a given contract is the amount of capital that the insurance company should set aside for future contingencies, defined by the expected present value of the contract's future net cost. In other words, it is the difference of the actuarial net present value at time t of the benefits to be paid and the premiums to be earned. (This definition is used under the prospective method Bowers et al., 1997.) We denote $V(t)$ as the actuarial reserve. In mathematical terms, this is

$$\begin{aligned} V(t) &:= e^{\delta t} \int_t^T (B(s) - (P(s) - P(t)))f_t(s)ds \\ &= e^{\delta t} \left[P(t) - \int_t^T H(s)f_t(s)ds \right] \end{aligned}$$

where $f_t(s) = f(s|X_i > t) = f(s)/\bar{F}(t)$. If the Equivalence Principle holds, one can also compute $V(t)$ using the retrospective method (Bowers et al., 1997), thereby obtaining

$$V(t) = \frac{e^{\delta t} \left[\int_0^t H(s)f(s)ds + P(t)\bar{F}(t) \right]}{\bar{F}(t)}. \tag{4}$$

If the Equivalence Principle is used, we also call $V(t)$ the benefit reserve.

Insurance company must reflect the total reserves in their balance sheets as liability. We define the *Total Reserve Process* at time t, denoted by $\bar{C}_n(t)$, as the sum of the actuarial reserves from all surviving policies. Hence

$$\begin{aligned} \bar{C}_n(t) &:= \bar{N}_n(t)V(t) \\ &= \bar{N}_n(t)e^{\delta t} \int_t^T (B(s) - (P(s) - P(t)))f_t(s)ds \\ &= \bar{N}_n(t)e^{\delta t} \left[P(t) - \int_t^T H(s)f_t(s)ds \right]. \end{aligned}$$

We also define the related quantity $\bar{m}(t)$ as

$$\bar{m}(t) = e^{\delta t} \left[P(t)\bar{F}(t) - \int_t^T H(s)f(s)ds \right], \tag{5}$$

which as we shall see is the fluid limit of $\bar{C}_n(\cdot)$ as $n \to \infty$.

Average Cash Process. As mentioned earlier, at time t, insurance company must recognize the liabilities reflected by the total reserves of the surviving policyholders. Those liabilities are to be faced, ideally, with the generated cash from the past. This motivates associating an *Average Cash Process* to each surviving policyholder, which we denote by $V_n(t)$. This quantity divides up the accumulated cash equally among the current survivors. In mathematical terms, it is

$$V_n(t) := \frac{e^{\delta t}\left[\int_0^t H(s)dN_n(s) + P(t)\bar{N}_n(t)\right]}{\bar{N}_n(t)} = \frac{C_n(t)}{\bar{N}_n(t)}.$$

As we shall study, under the Equivalence Principle, the process $V_n(\cdot)$ fluctuates around $V(\cdot)$.

In the next section, we will describe our main results involving limit theorems and approximations to these key quantities.

2. Main result

In order to describe our results we need to recall the definition of Brownian bridge, an important process obtained out of conditioning the value of Brownian motion at time 1. We introduce $(W_0(t), 0 \le t \le 1)$ as our notation for a Brownian bridge. It turns out that $W_0(t)$ is equal in distribution to $W(t) - tW(1)$, where $W(t)$ is a Brownian motion. It is also the unique Gaussian process with mean 0 and covariance function $\mathrm{Cov}(W_0(s), W_0(t)) = s(1-t), s \le t$. This implies that we can write the identities in distribution (for whole stochastic processes)

$$W_0(F(\cdot)) \stackrel{\mathcal{D}}{=} W(F(\cdot)) - F(\cdot)W(1) \stackrel{\mathcal{D}}{=} \int_0^{\cdot} \sqrt{f(s)}dW(s) - F(\cdot)\int_0^T \sqrt{f(s)}dW(s). \quad (6)$$

See, for example, Steele (2001) and Karatzas and Shreve (2008).

We are now ready to state and discuss our results. They are formulated in terms of weak convergence in a useful topology on spaces of functions, called the Skorokhod topology. The discussion of this topology and its preliminary theorems will be discussed in the Appendix. Our main result provides a joint approximation to the Total Cash Process, Total Reserve Process and Average Cash Process. The proof is given in the Appendix.

Theorem 1. *Assume that the Equivalence Principle holds and therefore that the identity (4) is in force. Regarding $(C_n(\cdot), \bar{C}_n(\cdot), V_n(\cdot))$ as elements in $D[0,T] \times D[0,T] \times D[0, T-\epsilon]$ for any $\epsilon \in (0,T)$ equipped with Skorokhod product topology, we have that*

$$\left(C_n(\cdot)/n, \bar{C}_n(\cdot)/n, V_n(\cdot)\right) \Rightarrow (m(\cdot), \bar{m}(\cdot), V(\cdot)) \quad (7)$$

as $n \to \infty$. Moreover,

$$\begin{aligned}&\left(\sqrt{n}\left(C_n(\cdot)/n - m(\cdot)\right), \sqrt{n}\left(\bar{C}_n(\cdot)/n - \bar{m}(\cdot)\right), \sqrt{n}(V_n(\cdot) - V(\cdot))\right) \\ &\Rightarrow \Bigg(e^{\delta t}\left[\int_0^t H(s)dW_0(F(s)) - P(t)W_0(F(t))\right], \\ &W_0(F(t))e^{\delta t}\left[\int_t^T H(s)f_t(s)ds - P(t)\right], \\ &\frac{e^{\delta t}}{\bar{F}(t)}\left[\int_0^t H(s)dW_0(F(s)) - P(t)W_0(F(t))\right] \\ &+ \frac{V(t)}{\bar{F}(t)}W_0(F(t))\Bigg) \end{aligned} \quad (8)$$

as $n \to \infty$.

The $\epsilon > 0$ in the theorem is to avoid zero divider at time T. The approximation in (8) suggests that when n is large, the Total Cash Process can be approximated by

$$C_n(t) \approx nm(t) + \sqrt{n}e^{\delta t}\left[\int_0^t H(s)dW_0(F(s)) - P(t)W_0(F(t))\right]. \quad (9)$$

Simultaneously, we have that the Total Reserve Process admits the approximation

$$\bar{C}_n(t) \approx n\bar{m}(t) + \sqrt{n}W_0(F(t))e^{\delta t}\left[\int_t^T H(s)f_t(s)ds - P(t)\right], \quad (10)$$

and that the Average Cash Process is approximated by

$$V_n(t) \approx V(t) + \frac{1}{\sqrt{n}}\left\{\frac{e^{\delta t}}{\bar{F}(t)}\left[\int_0^t H(s)dW_0(F(s)) - P(t)W_0(F(t))\right] + \frac{V(t)}{\bar{F}(t)}W_0(F(t))\right\}. \quad (11)$$

The first two processes can be interpreted as the insurer's total asset and total liability respectively. The fluctuation around the average in the these processes is smallest at the two ends of the time horizon, namely, at time 0 and at T, since we know for sure that there are 0 and n decrements respectively; the fluctuations become larger in the middle of the time range. The maximum fluctuation of the net asset process, obtained as the difference of the Total Cash Process and the Total Reserve Process, will occur at a time t^* which is characterized in Section 3.1.

The approximation (8) is joint in function space, so thanks to the continuous mapping principle (Theorem 2 in the Appendix), we can approximate the distribution of a whole (continuous) functional of the sample paths $C_n(\cdot)$ and $\bar{C}_n(\cdot)$. This is precisely the significance of the previous result. As a particular application, we will show in the next section how to exploit the continuous mapping principle to estimate the ruin probabilities under different types of life insurance contracts. In Section 4 we will provide closed-form formulas for the joint correlation of the limiting Gaussian processes in the right hand side of (8); thereby fully characterizing the whole asymptotic distribution of assets and liabilities across time.

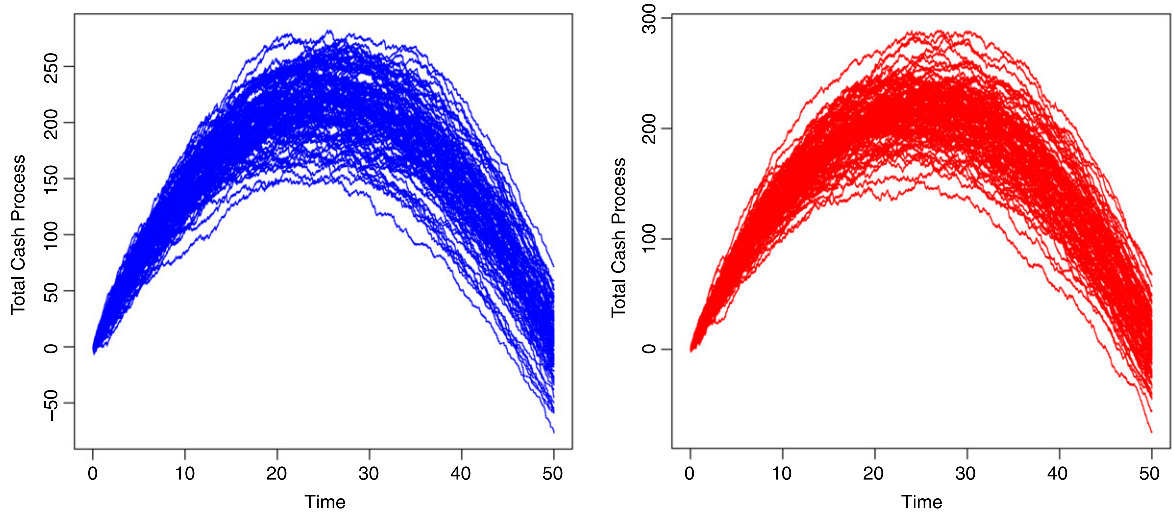
The approximation dictated by the third component, namely $V_n(\cdot)$, provides a link between our stochastic formulation for a large pool of policyholders and the classical reserve evaluation $V(t)$. It also provides support for the use of the Equivalence Principle from a micro-structural perspective. In particular, we show that under the Equivalence Principle the individual cash accounts fluctuate around the benefit reserve as the number of policyholders increases. Moreover, the result provides a Central Limit Theorem correction. We envision that our results in this section are potentially useful in evaluating in practice whether the difference between assets and liabilities on the balance sheet is within normal statistical error, although such application certainly requires being able to include other stylized features (such as investments in risky assets and so forth), which we plan to investigate in the future.

To illustrate Theorem 1 and the approximations (9)–(11), consider a batch of $n = 1000$ policyholders, each with a mortality distribution following a mixture of uniform distribution. Such mixture distribution arises as linear interpolation of the life table (see Bowers et al., 1997). For illustration, consider $T = 50$ i.e. a maximum life span of 50 years from present, and a monthly precision of the life table $k = 50 \times 12$. More specifically, we have the density of death random variable given by

$$f(t) = \sum_{j=1}^{k} p(j)\frac{k}{T}I\left(\frac{T}{k}(j-1) \le t < \frac{T}{k}j\right).$$

We further assume that $p(j)$ follows a discretized Gompertz's law given by

$$p(j) \propto e^{-e^{0.001(j-1)}} - e^{-e^{0.001j}}, \quad j = 1, \ldots, k.$$



(a) Actual dynamic.

(b) Approximation using Gaussian process.

Fig. 1. 100 sample paths of Total Cash Process by generating the actual dynamic versus via approximation.

Moreover, we assume that $\delta = 0.01$. The following graphs compare each of our approximations (9)–(11) to the actual dynamic. Fig. 1(a) shows 100 sample paths of $C_n(t)$ i.e. the Total Cash Process generated by $n = 1000$ policyholders with the aforementioned mortality distribution. Fig. 1(b) shows 100 sample paths of our Gaussian approximation, namely the right hand side of (9), by generating a sequence of Gaussian random variables. As shown in the graphs, the sample paths behave very similarly between the actual and our approximate processes.

Along the same line, Fig. 2(a) shows 100 sample paths of $\bar{C}_n(t)$ i.e. the Total Reserve Process by 1000 policyholders, whereas Fig. 2(b) shows the counterpart using our approximation in the right hand side of (10). Again, the sample paths behave similarly between the actual and approximate processes.

Fig. 3(a) shows 100 sample paths of $V_n(t)$, the Average Case Process, and also the value of benefit reserve $V(t)$. The sample paths of $V_n(t)$ center around $V(t)$, which is guaranteed by the Law of Large Numbers (7) in Theorem 1. Moreover, we can approximate $V_n(t)$ by our Gaussian approximation in (11). This is illustrated by Fig. 3(b), which again shows 100 sample paths of the right hand side of (3(b)) using Gaussian random variables.

From both Fig. 3(a) and (b) we see that the fluctuation of the Average Cash Process around the deterministic benefit reserve grows as time approaches T. In both figures, the Average Cash Process is close to the reserve only before $t = 30$. This phenomenon occurs because the Average Cash Process involves dividing by the number of survivals in the portfolio, and as time goes on, this number decreases to 0 almost surely. As a result, the process can attain very small or large values. It is worth noting that this phenomenon has nothing to do with our approximation i.e. it is an intrinsic property of the Average Cash Process itself from its definition laid out in Section 1 (as can be seen by the similar behavior of both Fig. 3(a) and (b)). Moreover, Theorem 1 makes clear that the Functional Law of Large Numbers and Central Limit Theorem works for the Average Cash Process only on $[0, T - \epsilon]$ for some prefixed $\epsilon > 0$, hence excluding the period of time close to T.

We explain how to implement the simulation procedure in Appendix A.2 to generate the approximations (9)–(11).

3. Applications and examples

Prevailing insurance practice calculates reserve based on the mathematical expectation of cash flows (i.e. the actuarial net present value of future benefits minus premiums) on individual basis. The aggregation of these individual reserves forms the liability for the insurer. Considering the process (9) we derived in the previous section as the fluctuation of overall assets, an interesting problem would be to analyze the mismatch between the liability process and the asset process. More precisely, when the size of assets are below the net premium reserve requirement, we say that a ruin occurs. Because the heavy traffic limit is Gaussian, and the theory of Gaussian processes is well developed, one can approximate such ruin probability easily.

3.1. Ruin probabilities

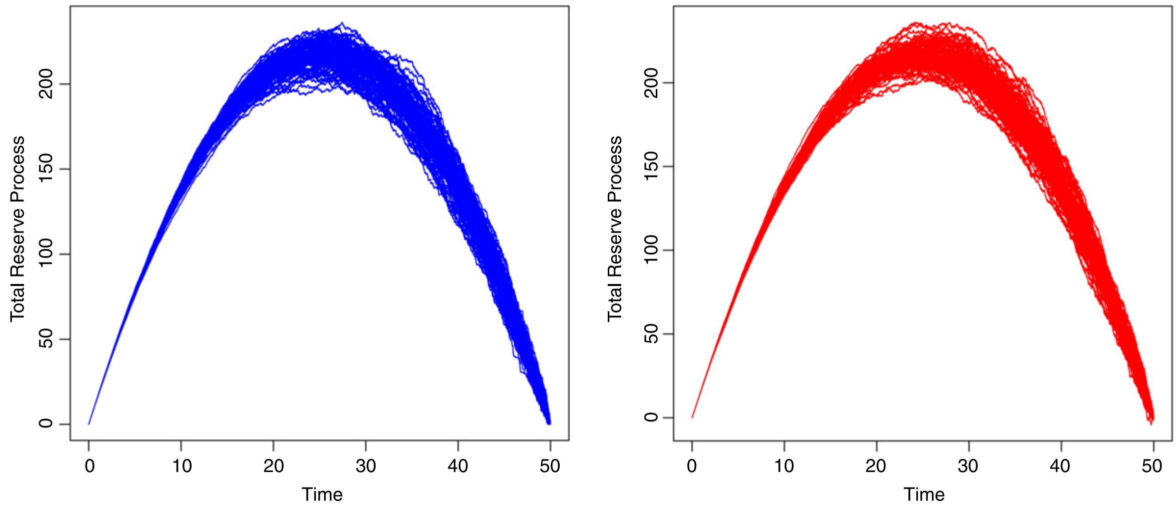
Here we formulate a ruin problem based on reserve requirement. Suppose that the prospective method is employed to set up required reserve on the balance sheet. Bankruptcy then occurs whenever the total asset falls short of the liability, plus initial surplus. More precisely, let the initial surplus be U_n that is scaled with n. The interpretation of the scaling is natural as a company with large number of policyholders in the system will naturally start with a large initial amount of capital requirement; this is precisely the initial surplus. We define $U_n(t) = U_n e^{\delta t}$ to be the value at time t of the initial surplus.

Ruin occurs if $U_n(t) + C_n(t) - \bar{C}_n(t) < 0$ (assuming a constant rate of investment interest). Under a finite-time formulation, the ruin probability is given by

$$P(U_n(t) + C_n(t) - \bar{C}_n(t) < 0 \quad \text{for some } t \in [0, T]).$$

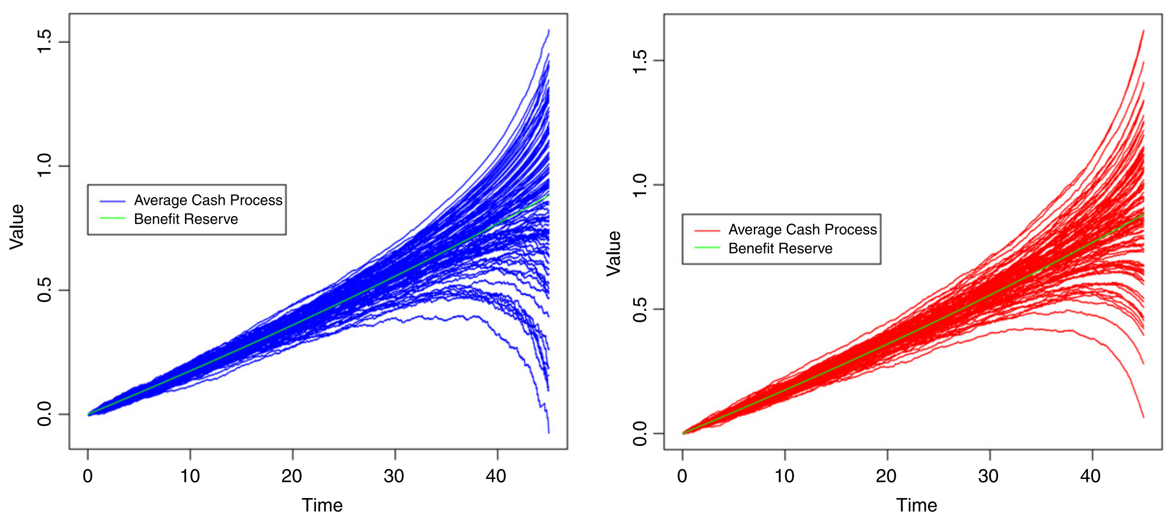
This formulation differs from the classical setting mainly in two aspects: (1) the risk processes $C_n(t)$ and $\bar{C}_n(t)$ depend on the structure of the insurance contracts rather than separate modeling of premium and claim processes; (2) the per-basis reserve that resembles the actual practice of the insurance company can be incorporated naturally into our framework.

We make two main assumptions in our formulation. First, we assume the Equivalence Principle for calculating premiums, due to market competition. Under this assumption the process $C_n(t) - \bar{C}_n(t)$ is essentially centered. Second, we assume that the surplus is scaled as $U_n = u\sqrt{n}$ for some $u > 0$. Note that other scaling of U_n would lead to different approximations. For example, if U_n is of order n, then rather than using our diffusion-type approximation in the previous sections, one would have to turn to large deviations



(a) Actual dynamic. (b) Approximation using Gaussian process.

Fig. 2. 100 sample paths of Total Reserve Process by generating the actual dynamic versus via approximation.



(a) Actual dynamic. (b) Approximation using Gaussian process.

Fig. 3. 100 sample paths of Average Cash Process by generating the actual dynamic versus via approximation.

asymptotic. The details of such are reported in Blanchet and Lam (2011), which also presents asymptotics and simulation design for estimating ruin problem under policyholder arrivals.

By Theorems 1 and 2 in the Appendix, we have

$$\frac{C_n(t)-\bar{C}_n(t)}{n} \Rightarrow m(t)-\bar{m}(t) = e^{\delta t}\left[\int_0^t H(s)f(s)ds + \int_t^T H(s)f(s)ds\right].$$

Note that, under the Equivalence Principle this process will be identically zero. To find the fluctuation of this process, again we use Theorem 2, scale by $\sqrt{n}$ and get

$$\sqrt{n}\left(\frac{C_n(t)-\bar{C}_n(t)}{n}-(m(t)-\bar{m}(t))\right) \Rightarrow e^{\delta t}\left[\int_0^t H(s)dW_0(F(s)) - P(t)W_0(F(t))\right] - W_0(F(t))e^{\delta t}\left[\int_t^T H(s)f_t(s)ds - P(t)\right]$$

$$= e^{\delta t}\left[\int_0^t H(s)dW_0(F(s)) - W_0(F(t))\int_t^T H(s)f_t(s)ds\right]. \quad (12)$$

Now the ruin probability is written as

$$\begin{aligned} P(\text{ruin}) &= P(U_n(t)+C_n(t)-\bar{C}_n(t) < 0 \text{ for some } t \in [0,T]) \\ &= P\left(\sup_{0\le t\le T}\frac{\bar{C}_n(t)-C_n(t)}{\sqrt{n}} > ue^{\delta t}\right) \\ &= P\left(\sup_{0\le t\le T} X(t) > u\right)(1+o(1)) \quad (13) \end{aligned}$$

as $n \to \infty$, where

$$X(t) = \int_0^t H(s)dW_0(F(s)) - W_0(F(t))\int_t^T H(s)f_t(s)ds. \quad (14)$$

This follows from (12) and the fact that $X(\cdot) \stackrel{\mathcal{D}}{=} -X(\cdot)$.

The next figure depicts a sample path of the approximating net asset process $C_n(t)-\bar{C}_n(t)$ and the deterministic trajectory of $-U_n(t)$. We use $n = 100$, $\delta = 0.01$, whole life insurance with $b =$

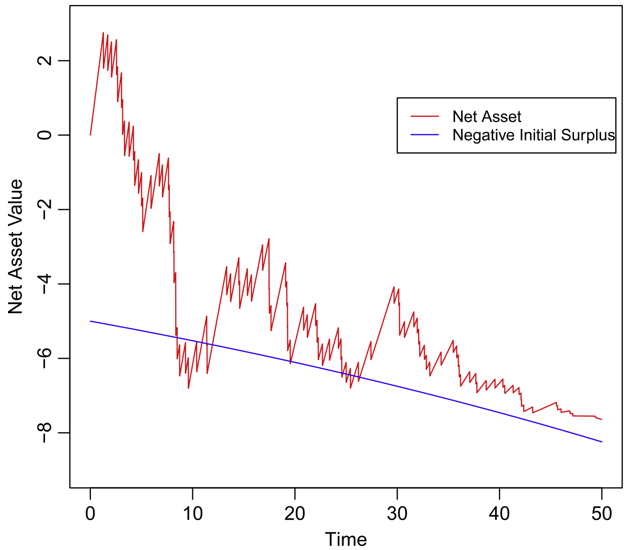


Fig. 4. Net asset process $C_n(t) - \bar{C}_n(t)$ (in red) and deterministic trajectory of $-U_n(t)$ (in blue). (For interpretation of the references to colour in this figure's legend, the reader is referred to the web version of this article.)

Table 1
Sample mean and 95% confidence interval for ruin probability using 10,000 samples of actual dynamic for different n and u.

Value of u	Value of n	Estimated probability	95% confidence interval
0.1	10	0.687	(0.658, 0.716)
0.1	100	0.784	(0.758, 0.810)
0.1	1000	0.829	(0.810, 0.852)
0.5	10	0.311	(0.282, 0.340)
0.5	100	0.342	(0.313, 0.371)
0.5	1000	0.353	(0.323, 0.383)
1.0	10	0.059	(0.044, 0.074)
1.0	100	0.066	(0.051, 0.081)
1.0	1000	0.081	(0.064, 0.098)

Table 2
Sample mean and 95% confidence interval for ruin probability using 10,000 samples of approximating Gaussian process for different u.

Value of u	Estimated probability	95% confidence interval
0.1	0.839	(0.831, 0.846)
0.5	0.353	(0.344, 0.362)
1.0	0.062	(0.057, 0.067)

1 and continuous level premium calculated under the Equivalence Principle, with mortality distribution as the interpolation of Gompertz's law as specified in generating all the figures in Section 2 (the level premium rate is calculated to be $p = 0.038$). For the specific sample path depicted in Fig. 4, ruin happens at around time 8.5.

Next, we investigate the precision of the approximation (13). To do so, we compare our approximation with the actual probability by running simulation for both quantities, using different portfolio size n and initial surplus $U_n = u\sqrt{n}$. Namely, for each $n = 10, 100, 1000$ and $u = 0.1, 0.5, 1$ (note that the initial surplus U_n is based at u and scaled in $\sqrt{n}$), we run simulation with 10,000 samples to estimate the probability. We use $\delta = 0.01$ and whole life insurance with $b = 1$, and the interpolated Gompertz's law as in the example in Section 2. Table 1 shows the estimated probabilities and 95% confidence intervals for the simulation that uses the actual dynamics i.e. by generating the portfolio of policyholders.

To compare, we also estimate the approximation $P(\sup_{0\le t\le T} X(t) > u)$ in (13) by simulating the process $X(t)$ defined in (14). Table 2 shows the simulation result of our approximating process for different values of u, each using 10,000 samples.

The approximation (13) predicts that the values in Table 2 should be close to those in Table 1 for equal values of u, when n is large enough. This is indeed the case as demonstrated, with the approximation being better for larger n. For $n = 1000$, the 95% confidence interval for the actual dynamic and the Gaussian approximation overlaps for all three values of u. As a side observation, the 95% confidence intervals for the Gaussian approximation are consistently smaller than those of the actual dynamic when using 10,000 samples for each.

For u larger than 1, it typically takes a long time to simulate the probability using either the actual process or our approximating Gaussian process (note that for $n = 1000, u = 1$ means the initial surplus is $u\sqrt{n} = 31.6$). This is where our approximation (13) can be useful in tackling the problem. While the probability (13) typically does not have closed-form solution, a fair amount is known for sharp asymptotics of the maximum of a Gaussian process (see for instance, Husler and Piterbarg, 1999 and Dieker, 2005); these approximations depend on the local correlation structure which is obtained in the next section. Moreover, designing efficient Monte Carlo method, such as importance sampling, for Gaussian process is well-studied (see for example Blanchet and Li, 2011) and can be easier to handle than the actual dynamic.

Here we will present an analytical approximation that is popular in the context of Gaussian queues (see Chapter 5 in Mandjes, 2007) and is easy to develop. The approximation works well for large values of u. Of course, one has to be careful when we use this approximation because our diffusion limit is established assuming that u is $O(1)$. The Gaussian approximation, however, still remains valid for the tail if u is allowed to grow as $n \to \infty$ at a sufficiently slow speed. In the presence of a large deviations result, which can be derived in our current setting, it suffices to let $u := u_n \to \infty$ in such a way that $u_n = o\left(n^{1/2}\right)$. This is what is known as moderate deviations scaling (see Chapter 8 in Ganesh et al., 2004). This is the type of asymptotic environment that we have in mind when we use our Gaussian approximation for tail probabilities.

We have

$$P\left(\sup_{0\le t\le T} X(t) > u\right) \ge \sup_{0\le t\le T} P(X(t) > u). \tag{15}$$

We now analyze the right hand side of (15). Using (6), $X(t)$ can be shown to be equal in distribution to

$$\begin{aligned}
&\int_0^t H(s)\sqrt{f(s)}dW(s) - \int_0^t H(s)f(s)ds \\
&\quad \times \int_0^T \sqrt{f(s)}dW(s) - \int_0^t \sqrt{f(s)}dW(s) \\
&\quad \times \int_t^T H(s)f_t(s)ds + F(t)\int_0^T \sqrt{f(s)}dW(s) \\
&\quad \times \int_t^T H(s)f_t(s)ds \\
&= \int_0^t \left(H(s) - \int_t^T H(s)f_t(s)ds\right) dW(F(s)) \\
&\quad + \left(F(t)\int_t^T H(s)f_t(s)ds - \int_0^t H(s)f(s)ds\right) W(1) \\
&= \int_0^t \left(H(s) - \int_t^T H(s)f_t(s)ds\right) dW(F(s)) \\
&\quad + W(1)\int_t^T H(s)f_t(s)ds \\
&\overset{\mathcal{D}}{=} N(0, \sigma^2(t))
\end{aligned}$$

where the second equality comes from the Equivalence Principle, and

$$\begin{aligned}\sigma^2(t) &= \int_0^t \left(H(s) - \int_t^T H(v)f_t(v)dv\right)^2 f(s)ds \\ &\quad + \left(\int_t^T H(s)f_t(s)ds\right)^2 \\ &\quad + 2\int_0^t \left(H(s) - \int_t^T H(v)f_t(v)dv\right) f(s)ds \\ &\quad \times \int_t^T H(s)f_t(s)ds \\ &= \int_0^t H(s)^2 f(s)ds - 2\int_t^T H(s)f_t(s)ds \\ &\quad \times \int_0^t H(s)f(s)ds + \left(\int_t^T H(s)f_t(s)ds\right)^2 F(t) \\ &\quad + \left(\int_t^T H(s)f_t(s)ds\right)^2 + 2\int_0^t H(s)f(s)ds \\ &\quad \times \int_t^T H(s)f_t(s)ds - 2\left(\int_t^T H(s)f_t(s)ds\right)^2 F(t) \\ &= \int_0^t H(s)^2 f(s)ds + \bar{F}(t)\left(\int_t^T H(s)f_t(s)ds\right)^2 \end{aligned} \tag{16}$$

by Ito's isometry. Obviously, then,

$$P\left(\sup_{0\le t\le T} X(t) > u\right) \ge \sup_{0\le t\le T} P(N(0,\sigma^2(t)) > u). \tag{17}$$

On the other hand, the upper bound of the ruin probability is provided by Borell's inequality, a standard result in the theory of Gaussian process (see, for example, Adler, 1990)

$$P\left(\sup_{0\le t\le T} X(t) > u\right) \le e^{Cu - \frac{1}{2}u^2/\sigma^2(t^*)} \tag{18}$$

where $t^* = \operatorname{argmax}_{0\le t\le T}\sigma^2(t)$ and C is a constant depending on $E\sup_{0\le t\le T} X(t)$.

With (17) and (18), we obtain that

$$\frac{1}{\sqrt{2\pi}\sigma(t^*)} e^{-\frac{1}{2}u^2/\sigma^2(t^*)} \le P\left(\sup_{0\le t\le T} X(t) > u\right) \le e^{Cu - \frac{1}{2}u^2/\sigma^2(t^*)}.$$

Taking logarithms and dividing by u^2, we get

$$\begin{aligned}\frac{1}{u^2}\log\left(\frac{1}{\sqrt{2\pi}\sigma(t^*)}\right) - \frac{1}{2\sigma^2(t^*)} &\le \frac{1}{u^2}\log P\left(\sup_{0\le t\le T} X(t) > u\right) \\ &\le \frac{C}{u} - \frac{1}{2\sigma^2(t^*)}.\end{aligned}$$

Letting $u \to \infty$ and using a sandwich argument, we have the asymptotic result

$$\lim_{u\to\infty} \frac{1}{u^2}\log P\left(\sup_{0\le t\le T} X(t) > u\right) = -\frac{1}{2\sigma^2(t^*)}. \tag{19}$$

In fact, we can strengthen (19) to obtain computable exact bounds for $P\left(\sup_{0\le t\le T} X(t) > u\right)$. The obvious lower bound is given in (17). For the upper bound, note that in general the bound in (18) depends on the constant C that is hard to obtain (and so (18) can only serve as guidance to the logarithmic asymptotic in (19)). Hence we use a result from Piterbarg (1996), which applies to our setting as follows. Let a be a constant such that $P(\sup_{0\le t\le T} X(t) > a) \le 1/2$. Then for any $u > a$, we have

$$P\left(\sup_{0\le t\le T} X(t) > u\right) \le 2\Phi\left(\frac{u-a}{\sigma(t^*)}\right) \tag{20}$$

where $\Phi(x) = P(N(0,1) > x)$ for a standard Gaussian variable $N(0,1)$. In the next subsections we will demonstrate through examples how one can use these bounds in practice.

Lastly, to simplify the calculations needed to characterize $\sigma^2(t^*)$, we impose the following additional regularity assumption on $H(\cdot)$, which will be satisfied by all the examples in the next subsection. We stress that our main result, Theorem 1, is not subject to this assumption.

Assumption 3. The function $H(\cdot)$ is non-decreasing and that $H(0) < 0$ while $H(T) > 0$.

This assumption facilitates the search for the value of t^*. Namely, to find t^*, one merely differentiates (16) to get

$$\begin{aligned}\frac{d\sigma^2(t)}{dt} &= -\frac{f(t)}{\bar{F}(t)}\left(\int_t^T H(s)f(s)ds\right)^2 \\ &\quad - \frac{2}{\bar{F}(t)}\int_t^T H(s)f(s)ds H(t)f(t) + H(t)^2 f(t) \\ &= f(t)\left[H(t)^2 - 2H(t)\int_t^T H(s)f_t(s)ds \right. \\ &\quad \left. - \left(\int_t^T H(s)f_t(s)ds\right)^2\right] \\ &= f(t)\left[H(t) - (1+\sqrt{2})\int_t^T H(s)f_t(s)ds\right] \\ &\quad \times \left[H(t) + (\sqrt{2}-1)\int_t^T H(s)f_t(s)ds\right].\end{aligned} \tag{21}$$

Note that $f(t) > 0$ for all $t \in (0,T)$. Since $H(t)$ is non-decreasing, we have $\int_t^T H(s)f_t(s)ds \ge H(t)$ and so $H(t) - (1+\sqrt{2})\int_t^T H(s)f_t(s)ds < 0$ for all $t \in [0,T)$. Also, by the Equivalence Principle and that $H(0) < 0$ and is continuous, we have $H(t) + (\sqrt{2}-1)\int_t^T H(s)f_t(s)ds < 0$ for a punctured neighborhood of $t = 0$ i.e. $t \in (0,\epsilon)$ for some $\epsilon > 0$. On the other hand, since $\int_t^T H(s)f_t(s)ds \ge H(t)$ and that $H(T) > 0$ and is continuous, we have $H(t) + (\sqrt{2}-1)\int_t^T H(s)f_t(s)ds > 0$ for a punctured neighborhood of $t = T$ i.e. $t \in (T-\epsilon, T)$ for some $\epsilon > 0$. These lead to the conclusion that there is a global maximum in the interior of the domain i.e. $(0,T)$.

To solve for the global maximum, one can numerically solve for the zeros of $H(t) + (\sqrt{2}-1)\int_t^T H(s)f_t(s)ds$. Then the global maximizer is either the zero or the discontinuous point of $d\sigma^2(t)/dt$ that gives the highest value of $\sigma^2(t)$.

3.2. Examples of calculation of $\sigma^2(t^*)$

In this subsection we will discuss some examples on calculating the value of $\sigma^2(t^*)$ in the Gaussian approximation we described. The first example assumes uniform mortality distribution and unit time horizon, as a simple illustration of our method. Then we will consider a more realistic distribution that comes from interpolation of the life table, and show our analysis proceeds equally handily in this case.

Example 1 (*Uniform Mortality*). Suppose X_i follows uniform distribution on $[0,T]$. We consider three different contracts: whole life insurance with level premium, increasing premium, and term life insurance.

Whole life insurance with level premium. Assume whole life insurance with continuous level premium p and benefit b i.e. $H(t) = p(1-e^{-\delta t})/\delta - be^{-\delta t}$. By the Equivalence Principle we can calculate

$$p = \frac{b\delta(1-e^{-\delta T})}{\delta T - 1 + e^{-\delta T}}$$

and so

$$\begin{aligned} H(t) + (\sqrt{2}-1)\int_t^T H(s)f_t(s)ds &= \frac{p}{\delta} - \left(\frac{p}{\delta}+b\right)e^{-\delta t} + (\sqrt{2}-1) \\ &\quad \times \left[\frac{p}{\delta} - \frac{1}{\delta(T-t)}\left(\frac{p}{\delta}+b\right)(e^{-\delta t}-e^{-\delta T})\right]. \end{aligned}$$

Assuming the inputs $T = 1, b = 1$ and $\delta = 0.01$, we have $t^* = 0.414$. Note that

$$\begin{aligned} \sigma^2(t) &= \frac{1}{T}\Bigg[\frac{p^2 t}{\delta^2} - \frac{2p}{\delta^2}\left(\frac{p}{\delta}+b\right)(1-e^{-\delta t}) \\ &\quad + \frac{1}{2\delta}\left(\frac{p}{\delta}+b\right)^2(1-e^{-2\delta t})\Bigg] - \left(1-\frac{t}{T}\right) \\ &\quad \times \left[\frac{p}{\delta} - \frac{1}{\delta(T-t)}\left(\frac{p}{\delta}+b\right)(e^{-\delta t}-e^{-\delta T})\right]^2. \end{aligned}$$

Plugging in t^*, we have $\sigma(t^*) = 0.256$.

Whole life insurance with increasing premium. Suppose the same benefit amount $b = 1$ is guaranteed. However, let us consider an increasing premium rate $pe^{\mu t}$. Let $\mu = 0.05$. Then the Equivalence Principle gives

$$p = \frac{b(\delta-\mu)^2}{\delta}\frac{1-e^{-\delta T}}{\delta-\mu-1+e^{-(\delta-\mu)T}}.$$

In this case

$$H(t) = \frac{p}{\delta-\mu} - \frac{p}{\delta-\mu}e^{-(\delta-\mu)t} - be^{-\delta t}$$

and

$$\begin{aligned} H(t) + (\sqrt{2}-1)\int_t^T H(s)f_t(s)ds &= \frac{p}{\delta-\mu} - \frac{p}{\delta-\mu}e^{-(\delta-\mu)t} - be^{-\delta t} + (\sqrt{2}-1) \\ &\quad \times \Bigg[\frac{p}{\delta-\mu} - \frac{p}{(\delta-\mu)^2}\frac{1}{T-t}(e^{-(\delta-\mu)t}-e^{-(\delta-\mu)T}) \\ &\quad - \frac{b}{\delta(T-t)}(e^{-\delta t}-e^{-\delta T})\Bigg]. \end{aligned}$$

We get that $t^* = 0.416$. Note that

$$\begin{aligned} \sigma^2(t) &= \frac{1}{T}\Bigg[\frac{p^2 t}{(\delta-\mu)^2} + \frac{p^2}{2(\delta-\mu)^3}(1-e^{-2(\delta-\mu)t}) \\ &\quad + \frac{b^2}{2\delta}(1-e^{-2\delta t}) - \frac{2p^2}{(\delta-\mu)^3}(1-e^{-(\delta-\mu)t}) \\ &\quad - \frac{2pb}{\delta(\delta-\mu)}(1-e^{-\delta t}) \\ &\quad + \frac{2pb}{(\delta-\mu)(2\delta-\mu)}(1-e^{-(2\delta-\mu)t})\Bigg] + \left(1-\frac{t}{T}\right) \\ &\quad \times \Bigg[\frac{p}{\delta-\mu} - \frac{p}{(\delta-\mu)^2}\frac{1}{T-t}(e^{-(\delta-\mu)t}-e^{-(\delta-\mu)T}) \\ &\quad - \frac{b}{\delta(T-t)}(e^{-\delta t}-e^{-\delta T})\Bigg]^2. \end{aligned}$$

Plugging in t^*, we get $\sigma(t^*) = 0.519$.

Term life insurance. Assume now a term life insurance with tenor $l < T$. Let $l = 0.5$. Then

$$P(t) = \begin{cases} \dfrac{p(1-e^{-\delta t})}{\delta} & \text{for } t \le l \\ \dfrac{p(1-e^{-\delta l})}{\delta} & \text{for } t > l \end{cases}$$

and

$$B(t) = \begin{cases} be^{-\delta t} & \text{for } t \le l \\ 0 & \text{for } t > l. \end{cases}$$

The Equivalence Principle gives

$$p = \frac{b\delta(1-e^{-\delta l})}{l\delta - 1 + e^{-\delta l} + (T-l)\delta(1-e^{-\delta l})}.$$

Note that

$$H(t) = \begin{cases} \dfrac{p}{\delta} - \left(\dfrac{p}{\delta}+b\right)e^{-\delta t} & \text{for } t \le l \\ \dfrac{p}{\delta}(1-e^{-\delta l}) & \text{for } t > l \end{cases}$$

and

$$H(t) + (\sqrt{2}-1)\int_t^T H(s)f_t(s)ds = \begin{cases} \dfrac{p}{\delta} - \left(\dfrac{p}{\delta}+b\right)e^{-\delta t} + (\sqrt{2}-1) \\ \quad \times \left[\dfrac{p}{\delta}\dfrac{l-t}{T-t} - \left(\dfrac{p}{\delta}+b\right)\dfrac{1}{\delta(T-t)}(e^{-\delta t}-e^{-\delta l}) \right. \\ \quad \left. + \dfrac{T-l}{T-t}\dfrac{p}{\delta}(1-e^{-\delta l})\right] & \text{for } t \le l \\ \sqrt{2}\dfrac{p}{\delta}(1-e^{-\delta l}) & \text{for } t > l \end{cases}$$

which has a zero at $t = 0.5$. Note that

$$\sigma^2(t) = \begin{cases} \dfrac{1}{T}\left[\dfrac{p^2 t}{\delta^2} - \dfrac{2p}{\delta^2}\left(\dfrac{p}{\delta}+b\right)(1-e^{-\delta t}) + \left(\dfrac{p}{\delta}+b\right)^2 \right. \\ \quad \left. \times \dfrac{1}{2\delta}(1-e^{-2\delta t})\right] + \left(1-\dfrac{t}{T}\right)\left[\dfrac{p}{\delta}\dfrac{l-t}{T-t} \right. \\ \quad - \left(\dfrac{p}{\delta}+b\right)\dfrac{1}{\delta(T-t)}(e^{-\delta t}-e^{-\delta l}) \\ \quad \left. + \dfrac{T-l}{T-t}\dfrac{p}{\delta}(1-e^{-\delta l})\right]^2 & \text{for } t \le l \\ \dfrac{1}{T}\left[\dfrac{p^2 l}{\delta^2} - \dfrac{2p}{\delta^2}\left(\dfrac{p}{\delta}+b\right)(1-e^{-\delta l}) + \left(\dfrac{p}{\delta}+b\right)^2 \right. \\ \quad \left. \times \dfrac{1}{2\delta}(1-e^{-2\delta l}) + (t-l)\dfrac{p^2}{\delta^2}(1-e^{-\delta l})^2\right] \\ \quad + \left(1-\dfrac{t}{T}\right)\dfrac{p^2}{\delta^2}(1-e^{-\delta l})^2 & \text{for } t > l. \end{cases}$$

In fact, in this particular case there are more than one t^*, namely any value in $[0.5, 1)$. They all give $\sigma(t^*) = 0.679$.

We see that both increasing premium structure with rate 0.05 and term life at time 0.5 have $\sigma^2(t^*)$ larger than whole life insurance. In other words, implementing whole life insurance gives the insurer a better risk profile. An interesting question would be the type of insurance policy, say P, that minimizes $\sigma_P^2(t^*)$. In this case, we have to solve the problem of minimizing $\max_{0\le t\le T}\sigma_P^2(t)$ over a fixed family of insurance policies P. The general solution to this question will be explored in future work.

Example 2 (*Interpolation of Life Table*). Let us now consider the mortality assumption we use in our simulation in Section 2 i.e. a linear interpolation of life table with $T = 50$, $k = 50 \times 12$, and the mortality's density

$$f(t) = q(j)\frac{k}{T}I\left(\frac{T}{k}(j-1) \le t < \frac{T}{k}j\right)$$

where $q(j), j = 1, \dots, k$ is the probability that X_i falls into the interval between $(T/k)(j-1)$ and $(T/k)j$. Moreover,

$$q(j) \propto e^{-e^{0.001(j-1)}} - e^{-e^{0.001j}}$$

follows a discretized Gompertz's law.

Whole life insurance with level premium. Consider whole life insurance with continuous level premium p, benefit $b = 1$, and assume interest rate $\delta = 0.1$. As in Example 1, $H(t) = p(1 - e^{-\delta t})/\delta - be^{-\delta t} = p/\delta - (p/\delta + b)e^{-\delta t}$. Note that

$$\bar{F}(t) = p(j)\frac{k}{T}\left(\frac{T}{k}j - t\right) + \sum_{r=j+1}^{k} p(r)I(j < k)$$

for $(T/k)(j-1) \le t < (T/k)j$. For convenience, let us also define

$$G(t,\lambda) = \int_t^T e^{-\lambda s}f(s)ds = p(j)\frac{k}{T\lambda}(e^{-\lambda t} - e^{-\lambda(T/k)j}) + \sum_{r=j+1}^{k}(e^{-\lambda(T/k)(j-1)} - e^{-\lambda(T/k)j})I(j < k)$$

for $(T/k)(j-1) \le t < (T/k)j$. From the Equivalence Principle the premium is

$$p = \frac{bE[e^{-\delta X_i}]}{E[(1/\delta)(1 - e^{-\delta X_i})]} = \frac{b\delta G(0,\delta)}{1 - G(0,\delta)}.$$

We have

$$\begin{aligned} H(t) &+ (\sqrt{2}-1)\int_t^T H(s)f_t(s)ds \\ &= \frac{p}{\delta} - \left(\frac{p}{\delta} + b\right)e^{-\delta t} + (\sqrt{2}-1) \\ &\quad\times \frac{(p/\delta)\bar{F}(t) - (p/\delta + b)G(t,\delta)}{\bar{F}(t)} \end{aligned}$$

which has a zero at $t = 19.03$. Now

$$\begin{aligned} \sigma^2(t) &= \int_0^t \left(\frac{p}{\delta} - \left(\frac{p}{\delta} + b\right)e^{-\delta t}\right)^2 f(s)ds \\ &\quad + \bar{F}(t)\left(\int_t^T \left(\frac{p}{\delta} - \left(\frac{p}{\delta} + b\right)e^{-\delta t}\right)f_t(s)ds\right)^2 \\ &= \frac{p^2}{\delta^2}F(t) - 2\frac{p}{\delta}\left(\frac{p}{\delta} + b\right)(G(0,\delta) - G(t,\delta)) \\ &\quad + \left(\frac{p}{\delta} + b\right)^2 (G(0,2\delta) - G(t,2\delta)) \\ &\quad + \frac{1}{\bar{F}(t)}\left(\frac{p}{\delta}\bar{F}(t) - \left(\frac{p}{\delta} + b\right)G(t,\delta)\right)^2. \end{aligned}$$

Comparing the value of $\sigma(t)$ at 19.03 with all the discontinuous points of $f(t)$ i.e. $(T/k)j, j = 0, \dots, m$, we see that 19.03 maximizes $\sigma(t)$. This gives $t^* = 19.03$ and $\sigma(t^*) = 0.492$.

Whole life insurance with increasing premium. As in Example 1, consider now an increasing premium $pe^{\mu t}$. By the Equivalence Principle, we have

$$p = \frac{b/\delta G(0,\delta)}{(1/(\delta-\mu))(1 - G(0,\delta-\mu))}.$$

Also, $H(t) = p/(\delta-\mu) - (p/(\delta-\mu))e^{-(\delta-\mu)t} - be^{-\delta t}$, and

$$\begin{aligned} H(t) &+ (\sqrt{2}-1)\int_t^T H(s)f_t(s)ds \\ &= \frac{p}{\delta-\mu} - \frac{p}{\delta-\mu}e^{-(\delta-\mu)t} - be^{-\delta t} + (\sqrt{2}-1)\frac{1}{\bar{F}(t)} \\ &\quad\times \left(\frac{p}{\delta-\mu}\bar{F}(t) - \frac{p}{\delta-\mu}G(t,\delta-\mu) - bG(t,\delta)\right). \end{aligned}$$

Putting $\mu = 0.05$, there is a zero at $t = 22.39$. Now

$$\begin{aligned} \sigma^2(t) = &\frac{p^2}{(\delta-\mu)^2}F(t) + \frac{p^2}{(\delta-\mu)^2}(G(0,2(\delta-\mu)) \\ &- G(t,2(\delta-\mu))) + b^2(G(0,2\delta) - G(t,2\delta)) \\ &- \frac{2p^2}{(\delta-\mu)^2}(G(0,\delta-\mu) - G(t,\delta-\mu)) \\ &- \frac{2pb}{\delta-\mu}(G(0,\delta) - G(t,\delta)) \\ &+ \frac{2pb}{\delta-\mu}(G(0,2\delta-\mu) - G(t,2\delta-\mu)) + \frac{1}{\bar{F}(t)} \\ &\times \left(\frac{p}{\delta-\mu}\bar{F}(t) - \frac{p}{\delta-\mu}G(t,\delta-\mu) - bG(t,\delta)\right)^2. \end{aligned}$$

Again, comparing with all the discontinuous points of $f(t)$, 22.39 dominates and so $t^* = 22.39$, with $\sigma(t^*) = 0.645$.

Term life insurance. Consider now a term life insurance with tenor l. By the Equivalence Principle,

$$p = \frac{b\delta(G(0,\delta) - G(l,\delta))}{1 - (G(0,\delta) - G(l,\delta)) - e^{-\delta l}\bar{F}(l)}.$$

So

$$H(t) + (\sqrt{2}-1)\int_t^T H(s)f_t(s)ds$$

$$= \begin{cases} \frac{p}{\delta} - \left(\frac{p}{\delta} + b\right)e^{-\delta t} + (\sqrt{2}-1)\frac{1}{\bar{F}(t)}\left(\frac{p}{\delta}\bar{F}(t) - \left(\frac{p}{\delta} + b\right) \times (G(t,\delta) - G(l,\delta)) - \frac{p}{\delta}G(l,\delta)\right) & \text{for } t \le l \\ \frac{p}{\delta}(1 - e^{-\delta l}) + (\sqrt{2}-1)\frac{p}{\delta}(1 - e^{-\delta l}) & \text{for } t > l. \end{cases}$$

Putting $l = T/2 = 25$, this gives a zero at $t = 25$. Now

$\sigma^2(t)$

$$= \begin{cases} \frac{p^2}{\delta^2}F(t) - 2\frac{p}{\delta}\left(\frac{p}{\delta}+b\right)(G(0,\delta)-G(t,\delta)) \\ \quad + \left(\frac{p}{\delta}+b\right)^2 (G(0,2\delta)-G(t,2\delta)) \\ \quad + \frac{1}{\bar{F}(t)}\left(\frac{p}{\delta}\bar{F}(t) - \left(\frac{p}{\delta}+b\right)(G(t,\delta)-G(l,\delta)) \right. \\ \quad \left. - \frac{p}{\delta}G(l,\delta)\right)^2 \quad \text{for } t \le l \\ \frac{p^2}{\delta^2}F(t) - 2\frac{p}{\delta}\left(\frac{p}{\delta}+b\right)(G(0,\delta)-G(l,\delta)) \\ \quad + \left(\frac{p}{\delta}+b\right)^2 (G(0,2\delta)-G(l,2\delta)) \\ \quad + \frac{p^2}{\delta^2}(1-e^{-\delta l})^2(F(t)-F(l)) + \frac{1}{\bar{F}(t)} \\ \quad \frac{p^2}{\delta^2}(1-e^{-\delta l})^2 \quad \text{for } t > l. \end{cases}$$

In this case t^* is any value in $[25, 50)$, which all give $\sigma(t^*) = 0.619$.

Comparing the $\sigma(t^*)$ for all three cases, again whole life insurance appears to attain the smallest $\sigma(t^*)$ and hence the best risk profile among the three.

Remark 1. Suppose we relax the identical profile assumption for policyholders (as discussed in Section 1)and we replace this assumption by a distribution of policyholder types. If the distribution is discrete, then all our results in this paper still hold, except that rather than using Gaussian process driven by one Brownian motion, the Gaussian process will be a mixture of Gaussian processes each driven by an independent Brownian motion. The quantity $\sigma^2(t^*)$ can be found similarly but the optimization problem will be less linear. On the other hand, if the type of distribution is continuous, then the limiting process will involve Brownian sheet. This seems to introduce further technicalities that are therefore left to future work.

3.3. Obtaining exact analytic bounds for ruin probabilities

In this subsection, we will use our calculation of $\sigma^2(t^*)$ to obtain exact bounds for the ruin probability $P(\max_{0\le t\le T} X(t) > u)$, via the lower bound (17) and the upper bound (20). To illustrate our exact bounds, we continue our calculation in Table 2 in Section 3.1. In particular, we test our bounds for larger values of u, from 1.1 up to 1.5.

We use $\sigma^2(t^*) = 0.492$, obtained from the example on whole life insurance with level premium and discretized Gompertz mortality in Section 3.2. For the upper bound, we take $a = 0.41$. This choice is made by testing sequentially on different values of a and picking an a such that the 95% upper confidence bound is less than or close to 0.5. (our upper confidence bound of the estimate of $P(\max_{0\le t\le T} X(t) > u)$ is (0.439, 0.501)). For this choice of a, we obtain our exact bounds for $u = 1.1$ up to 1.5. Moreover, we compute the mid-point between the lower and upper bound, as a quick summarizing rule.

We compare our analytic bounds to simulation outputs. First, we run 1000 samples for the corresponding Gaussian processes and compute the 95% confidence intervals of our outputs. Second, we take $n = 100$ and run the actual dynamic of the portfolio for 1000 times and also output the estimates and the confidence intervals. The results are shown in Tables 3–5.

We make a few comments on the comparisons of these results. First, in all tested values of u, the analytical lower and upper bounds contain the sample means of both the simulations of the Gaussian processes and the actual dynamics, and the mid-points appear to capture the same magnitudes as the simulation

Table 3
Exact lower and upper bounds for the Gaussian processes for different hit levels.

Value of u	Exact lower bound from (17)	Exact upper bound from (20)	Mid-point
1.1	0.0127	0.161	0.0867
1.2	0.00736	0.108	0.0579
1.3	0.00412	0.0705	0.0373
1.4	0.00222	0.0442	0.0232
1.5	0.00115	0.0267	0.0139

Table 4
Simulation estimates for the Gaussian processes.

Value of u	Simulation estimate	95% confidence interval of simulation estimate
1.1	0.039	(0.0270, 0.0510)
1.2	0.031	(0.0202, 0.0418)
1.3	0.016	(0.00821, 0.0238)
1.4	0.014	(0.00671, 0.0213)
1.5	0.005	(0.000621, 0.00938)

Table 5
Simulation estimates for the actual dynamic.

Value of u	Simulation estimate	95% confidence interval of simulation estimate
1.1	0.054	(0.0400, 0.0680)
1.2	0.033	(0.0219, 0.0441)
1.3	0.022	(0.0129, 0.0311)
1.4	0.004	$(8.12 \times 10^{-05}, 0.00792)$
1.5	0.002	$(-0.000774, 0.00477)$

results. The lower bounds appear to be closer than the simulation outputs, but the upper bounds are getting tighter as u increases. Asymptotically, the lower and upper bounds should converge to the same logarithmic limit, as in (19). Next, the simulation outputs for the actual dynamic are for $n = 100$. In general, running simulation for the actual dynamic is computationally intensive, with running time from a few hours to several days as n becomes as large as thousands. We would expect our bounds as well as the Gaussian simulation to perform better for larger n; but as we can already see, these approximations are reasonably accurate for the relatively small $n = 100$. Third, simulating the Gaussian process, in our experiments, took place in magnitudes from minutes to hours. For small enough values of u, it therefore provides a feasible alternative to simulating the actual dynamic. We caution, however, that simulating Gaussian processes involves discretization that introduces extra bias. The choices between using the analytical bounds and Gaussian simulation should be decided by users depending on situations. Generally speaking, we believe that analytical bounds will get more and more beneficial as u grows.

4. Correlation structure

Our model also provides a framework for studying temporal correlations of the risk processes. The Gaussian nature of the limits we have discussed allows easy computation. As an illustration, consider the processes $C_n(t)$ and $\bar{C}_n(t)$ in (8). As aforementioned, they can be interpreted as the assets and liabilities of the insurance company. Their variances as well as temporal and cross correlations can be found easily as follows:

Temporal covariances for Cash Process and Reserve Process:

$$\frac{\mathrm{Cov}(C_n(t), C_n(t'))}{n} \to e^{\delta(t+t')}\left[\int_0^{t\wedge t'} (H(s)-P(t))(H(s)-P(t'))f(s)ds\right.$$

$$- \int_0^t (H(s) - P(t))f(s)ds \int_0^{t'} (H(s) - P(t'))f(s)ds \Bigg]$$

$$\frac{\mathrm{Cov}(\bar{C}_n(t), \bar{C}_n(t'))}{n} \to e^{\delta(t+t')} F(t \wedge t')\bar{F}(t \vee t') \int_t^T (H(s) - P(t))f_t(s)ds \times \int_{t'}^T (H(s) - P(t'))f_{t'}(s)ds.$$

Cross temporal covariance between Cash Process and Reserve Process:

$$\frac{\mathrm{Cov}(C_n(t), \bar{C}_n(t'))}{n} \to e^{\delta(t+t')} \left[\bar{F}(t') \int_0^t (H(s) - P(t))f(s)ds - \int_{t'}^t (H(s) - P(t))f(s)ds I(t > t') \right] \times \int_{t'}^T (H(s) - P(t'))f_{t'}(s)ds.$$

These in particular give:

Variance for Cash Process and Reserve Process:

$$\frac{\mathrm{Var}(C_n(t))}{n} \to e^{2\delta t} \left[\int_0^t (H(s) - P(t))^2 f(s)ds - \left(\int_0^t (H(s) - P(t))f(s)ds \right)^2 \right]$$

$$\frac{\mathrm{Var}(\bar{C}_n(t))}{n} \to e^{2\delta t} F(t)\bar{F}(t) \left[\int_t^T (H(s) - P(t))f_t(s)ds \right]^2.$$

Cross covariance between Cash Process and Reserve Process:

$$\frac{\mathrm{Cov}(C_n(t), \bar{C}_n(t))}{n} \to e^{2\delta t} \bar{F}(t) \int_0^t (H(s) - P(t))f(s)ds \times \int_t^T (H(s) - P(t))f_t(s)ds.$$

From these one can calculate the temporal and cross correlations

$$\mathrm{Corr}(C_n(t), C_n(t')) = \frac{\mathrm{Cov}(C_n(t), C_n(t'))}{\sqrt{\mathrm{Var}(C_n(t))\mathrm{Var}(C_n(t'))}}$$

$$\mathrm{Corr}(\bar{C}_n(t), \bar{C}_n(t')) = \frac{\mathrm{Cov}(\bar{C}_n(t), \bar{C}_n(t'))}{\sqrt{\mathrm{Var}(\bar{C}_n(t))\mathrm{Var}(\bar{C}_n(t'))}}$$

$$\mathrm{Corr}(C_n(t), \bar{C}_n(t')) = \frac{\mathrm{Cov}(C_n(t), \bar{C}_n(t'))}{\sqrt{\mathrm{Var}(C_n(t))\mathrm{Var}(\bar{C}_n(t'))}}.$$

Now consider the net asset process under the Equivalence Principle approximated by $X(t)\sqrt{n}$ defined in (14). We have

$$\begin{aligned}
\mathrm{Cov}(X(t), X(t')) &= \int_0^{t\wedge t'} \left(H(s) - \int_t^T H(v)f_t(v)dv \right) \times \left(H(s) - \int_{t'}^T H(v)f_{t'}(v) \right) f(s)ds \\
&\quad + \int_t^T H(s)f_t(s)ds \times \int_0^{t'} \left(H(s) - \int_{t'}^T H(v)f_{t'}(v)dv \right) f(s)ds \\
&\quad + \int_{t'}^T H(s)f_{t'}(s)ds \times \int_0^t \left(H(s) - \int_t^T H(v)f_t(v)dv \right) f(s)ds \\
&\quad + \int_t^T H(s)f_t(s)ds \int_{t'}^T H(s)f_{t'}(s)ds \\
&= \int_0^{t\wedge t'} H(s)^2 f(s)ds - \int_{t\wedge t'}^T H(s)f_{t\wedge t'}(s)ds \times \int_0^{t\wedge t'} H(s)f(s)ds + \int_{t\wedge t'}^T H(s)f_{t\wedge t'}(s)ds \times \int_0^{t\vee t'} H(s)f(s)ds \\
&\quad + \bar{F}(t\vee t') \int_t^T H(s)f_t(s)ds \times \int_{t'}^T H(s)f_{t'}(s)ds \\
&= \int_0^{t\wedge t'} H(s)^2 f(s)ds - \int_{t\wedge t'}^T H(s)f_{t\wedge t'}(s)ds \times \int_0^{t\wedge t'} H(s)f(s)ds + 2\bar{F}(t\vee t') \times \int_t^T H(s)f_t(s)ds \int_{t'}^T H(s)f_{t'}(s)ds
\end{aligned}$$

by the Equivalence Principle in the last equality, and so in particular

$$\mathrm{Var}(X(t)) = \int_0^t H(s)^2 f(s)ds + \bar{F}(t) \left(\int_t^T H(s)f_t(s)ds \right)^2$$

which recovers the value of $\sigma^2(t)$ in (16).

5. Extensions

We emphasize that the current work serves as a first attempt to introduce the heavy traffic approach in modeling large life insurance portfolios on the sample path level. Regarding the stochastic component, especially in modeling ruin, the biggest limitation of the current work is the ignorance of the dynamic arrival process of policyholders. When such arrivals are present, the risk process will be a functional of an underlying infinite-server queue, in which the service times are the death times of the arriving policyholders. Such consideration will be one of our key future research directions.

As we discussed in the previous section, another important relaxation is the identical profile assumption. Whereas a discrete mixture of policyholders is straightforward, technicality arises when the mixture is continuous. Besides, several other directions of extensions can be pursued. A few possible and important extensions are: (1) exploring more complicated policy structures e.g. unit-linked products (2) modeling the interest rate as a market risk and stochastically changing (3) incorporating operational cost and other expenses (4) allowing time-varying correlation among policyholders e.g. Markov-modulated arrival rate and death distribution (5) relaxing the Equivalence Principle assumption and allowing safety loading etc.

Appendix. Technical development and Monte Carlo simulation methodology

This appendix is divided into two parts. We first provide the proofs of our main results, which require a quick review of some basic facts on the heavy traffic and weak convergence theory. The second part of this appendix concerns the implementation of the simulations shown in the paper.

A.1. Review of weak convergence and proofs of main results

Before we provide the proof of our main results, let us review some results on weak convergence theory in function spaces. Then we proceed with our proofs.

A.1.1. Review of weak convergence results

Define $C[0,T]$ as the set of all continuous functions on $[0,T]$ equipped with the uniform metric, denoted by $d_\infty(\cdot)$. That is,

$$d_\infty(x,y)=\sup_{0\le t\le T}|x(t)-y(t)|.$$

We also define $D[0,T]$ to be the set of all cadlag (left continuous with right limit) functions on $[0,T]$, and we equip the space with the standard Skorokhod metric, which we shall denote by $d_J(\cdot)$. In particular, if we let $\mathcal{A}$ be the set of all strictly increasing continuous functions that map $[0,T]$ into itself we have that

$$d_J(x,y)=\inf_{\lambda\in\mathcal{A}}\sup_{0\le t\le T}|(x\circ\lambda)(t)-y(t)|$$

(see, for example, Billingsley, 1999). In order to quickly have a grasp of the Skorokhod topology, it is easy to show that $x_n\to x$ in the d_J metric if and only if there exists a sequence of elements $\lambda_n\in\mathcal{A}$ such that $x_n\circ\lambda_n\to x$ in the d_∞ metric.

The uniform or Skorokhod topologies in a product space such as $C[0,T]\times C[0,T]$ (or $D[0,T]\times D[0,T]$) are defined as the sum of the corresponding metrics in each projection. In particular, for instance, if (x_1,x_2) and (y_1,y_2) are elements in $C[0,T]\times C[0,T]$, then we define the product uniform metric as

$$d_\infty^{\Pi}((x_1,x_2),(y_1,y_2))=d_\infty(x_1,y_1)+d_\infty(x_2,y_2).$$

Entirely analogous considerations and definitions apply to the Skorokhod product metric.

We denote "$\Rightarrow$" for weak convergence of probability measure. A useful characterization of weak convergence is that for a sequence of probability measures P_n, $P_n\Rightarrow P$ on the space $C[0,T]$ (respectively $D[0,T]$) if and only if $\int gdP_n\to\int gdP$ for any bounded, continuous function g on $C[0,T]$ (respectively $D[0,T]$). The uniform and Skorokhod topologies are set to define continuity for g. Equivalently, we say that a sequence of stochastic processes on $C[0,T]$ (respectively $D[0,T]$), $Y_n\Rightarrow Y$, if and only if $Eg(Y_n)\to Eg(Y)$ for any bounded continuous function g defined on $C[0,T]$ (respectively $D[0,T]$). This characterization will be useful for our development.

The main reason for developing weak convergence results in spaces of functions is given by the continuous mapping principle, which allows to derive further approximation results by expressing quantities of interest (such as the Total Cash Process and the Total Reserve Process defined in Section 1) as functions of a suitable process. A statement of the continuous mapping theorem is given next (see, for example, Billingsley, 1999):

Theorem 2. *Let $h:D[0,T]\to\mathcal{S}$ be measurable and D_h be the set of its discontinuities in Skorokhod topology. If the sequence of stochastic processes in $D[0,T]$, $Y_n\Rightarrow Y$, and $P(Y\in D_h)=0$, then $h(Y_n)\Rightarrow h(Y)$. In particular, if $Y\in C[0,T]$, the same conditions and results hold by defining D_h to be the set of discontinuity in the uniform topology. Moreover, the theorem holds for product of $D[0,T]$ spaces (respectively $C[0,T]$).*

By carefully choosing the continuous functionals (such as maximal, integral etc.) we will be able to obtain handy convergence results. Although the Skorokhod metric is rather explicit, the reduction to checking continuity in the d_∞ metric when Y is continuous (which is the second statement of the theorem) makes some of our calculations easier. This reduction comes from the fact that if $x_n\to x$ in the d_J metric and if x is in $C[0,T]$ then $x_n\to x$ in the d_∞ metric.

We shall also use the following standard result of the weak convergence of empirical process into Brownian bridge (see, for example, Billingsley, 1999 and Dudley, 1999). This can be summarized as:

Theorem 3. *For any i.i.d. random variables X_i, $i=1,\dots,n$ with distribution $F(x)$ supported on $[0,T]$, define*

$$F_n(t)=\frac{1}{n}\sum_{i=1}^{n}I(X_i\le t)$$

and regarding $F_n(t)$ as elements of $D[0,T]$ equipped with the Skorokhod topology, we have

$$F_n(t)\Rightarrow F(t) \tag{22}$$

and

$$\sqrt{n}(F_n(t)-F(t))\Rightarrow W_0(F(t)) \tag{23}$$

where $W_0(t)$ is standard Brownian bridge on $[0,1]$ as defined in the discussion prior to (6).

The limits in (22) and (23) are well-known in statistics and probability (see for example Dudley, 1999). In the queueing literature, these types of results are known as fluid and diffusion limits. Fluid limit refers to approximation by deterministic trajectory and hence its name. Diffusion limit refers to approximation by diffusion process, in this case driven by Brownian bridge. These convergence results and their various extensions serve as building blocks of other more complicated limit approximations.

A.1.2. Proofs of main results

Proof of Theorem 1. Our strategy is to show that processes $\sqrt{n}(C_n(\cdot)/n-m(\cdot))$, $\sqrt{n}\left(\bar{C}_n(\cdot)/n-\bar{m}(\cdot)\right)$, and $\sqrt{n}(V_n(\cdot)-V(\cdot))$ can be expressed each as continuous functions of the same underlying process. Therefore, because of the form of the product metrics, we will have joint convergence in the product topology. We then can treat each of the three processes separately. We first concentrate on $\sqrt{n}(C_n(\cdot)/n-m(\cdot))$.

First, $N_n(t)/n$ is the fraction of deaths over time, and is the empirical process of the n death random variables. By Theorem 3 and (1), we have $N_n(t)/n\Rightarrow F(t)$ and

$$Z_n(t):=\sqrt{n}\left(\frac{N_n(t)}{n}-F(t)\right)\Rightarrow W_0(F(t)) \tag{24}$$

on $D[0,T]$. Here (24) captures the fluctuation of the fraction of deaths over time, centered around the deterministic mean process $F(t)$, as $n\nearrow\infty$. Next consider the integral

$$\int_0^t H(s)dr(s). \tag{25}$$

Note that by Assumption 2 $H(s)$ has a continuous and bounded first derivative almost everywhere, so for any $r\in D[0,T]$, integration by parts gives

$$\int_0^t H(s)dr(s)=H(t)r(t)-H(0)r(0)-\int_0^t H'(s)r(s)ds. \tag{26}$$

To ease notation denote $\|\cdot\|$ as the uniform norm so that $d_\infty(x,y)=\|x-y\|$. Note that for $r_1,r_2\in C[0,T]$,

$$\left\|\int_0^t H(s)dr_1(s)-\int_0^t H(s)dr_2(s)\right\|$$
$$\le\left(2\|H\|+\int_0^T|H'(s)|ds\right)\|r_1-r_2\|$$

which shows continuity of $\int_0^t H(s)dr(s)$ on $r \in C[0,T]$. Now write

$$\frac{C_n(t)}{n} = e^{\delta t}\left[\int_0^t H(s)d\left(\frac{N_n(s)}{n}\right) + P(t)\left(\frac{N_n(T)}{n} - \frac{N_n(t)}{n}\right)\right].$$

Continuity of integral (25) and the fact that elementary operations are continuous on $C[0,T]$ yield that

$$e^{\delta t}\left[\int_0^t H(s)dr(s) + P(t)(r(T) - r(t))\right]$$

is continuous on $r \in C[0,T]$. Since $F(\cdot)$ is continuous and lies in $C[0,T]$, Theorems 2 and 3 concludes the first component of (7).

For the first component of (8), write

$$\sqrt{n}\left(\frac{C_n(t)}{n} - m(t)\right) = e^{\delta t}\left[\int_0^t H(s)dZ_n(s) + P(t)(Z_n(T) - Z_n(t))\right]$$

and (8) follows by similar argument, noting that $Z_n(T) = Z(T) = 0$.

The treatment of the processes $\bar{C}_n(\cdot)$ is entirely similar. First, it is clear from Theorem 3 that

$$\frac{\bar{C}_n(t)}{n} = \left(\frac{N_n(T)}{n} - \frac{N_n(t)}{n}\right)e^{\delta t}\left[P(t) - \int_t^T H(s)f_t(s)ds\right] \Rightarrow \bar{F}(t)e^{\delta t}\left[P(t) - \int_t^T H(s)f_t(s)ds\right] = \bar{m}(t).$$

Moreover, a straightforward application of the continuous mapping principle and Theorem 3 (following similar continuity arguments as those given earlier) yield

$$\sqrt{n}\left(\frac{\bar{C}_n(t)}{n} - \bar{m}(t)\right) = (Z_n(T) - Z_n(t))e^{\delta t}\left[P(t) - \int_t^T H(s)f_t(s)ds\right] \Rightarrow -W_0(F(t))e^{\delta t}\left[P(t) - \int_t^T H(s)f_t(s)ds\right].$$

Finally, for $V_n(\cdot)$ note that

$$V_n(t) = e^{\delta t}\left[\frac{\int_0^t H(s)d(N_n(s)/n)}{N_n(T)/n - N_n(t)/n} + P(t)\right].$$

A direct application of Theorem 2 allows us to conclude the third component of (7). Next write

$$\begin{aligned}
&\sqrt{n}(V_n(t) - V(t)) \\
&= \sqrt{n}\left(V_n(t) - \frac{nV(t)\bar{F}(t)}{\bar{N}_n(t)}\right) + \sqrt{n}\left(\frac{nV(t)\bar{F}(t)}{\bar{N}_n(t)} - V(t)\right) \\
&= \frac{\sqrt{n}e^{\delta t}}{\bar{N}_n(t)/n}\left(\int_0^t H(s)d\left(\frac{N_n(s)}{n}\right) + P(t)\frac{\bar{N}_n(t)}{n} - \int_0^t H(s)dF(s) - P(t)\bar{F}(t)\right) - \frac{\sqrt{n}V(t)}{\bar{N}_n(t)/n}\left(\frac{\bar{N}_n(t)}{n} - \bar{F}(t)\right) \\
&= \frac{e^{\delta t}}{N_n(T)/n - N_n(t)/n} \times \left(\int_0^t H(s)dZ_n(s) + P(t)(Z_n(T) - Z_n(t))\right) - \frac{V(t)}{N_n(T)/n - N_n(t)/n}(Z_n(T) - Z_n(t)).
\end{aligned} \tag{27}$$

Since $\bar{F}(t)$ is deterministic it follows that the following weak convergence result

$$\left(Z_n(t), \frac{N_n(t)}{n}\right) \Rightarrow (W_0(F(t)), \bar{F}(t))$$

follows jointly in $D[0,T] \times D[0,T]$. Since the limiting processes are continuous, it suffices to check that (27) is a continuous mapping from $(W_0(F(t)), \bar{F}(t))$ on $C[0, T-\epsilon]^2$ to $\mathbb{R}$, by Theorem 2. The argument proceeds as in the analysis of $C_n(\cdot)$ given earlier and we conclude our result. □

A.2. Simulation methodology

We lay out the simulation methodology we use to generate the graphs in this paper. The few and elementary steps in the methodology advocate our use of heavy traffic approximation. Note that, all the processes we introduced so far are elementary functions of

$$\left(\int_0^t H(s)dW_0(F(s)), W_0(F(t))\right) \tag{28}$$

in the pointwise sense. So we will discuss how to generate a path of these quantities. More precisely, we will generate a discretized version of this path at time points $t_0 = 0, t_1, \ldots, t_m = T$. Define $Y_0 = 0$ and

$$Y_i = \int_0^{t_i} H(s)dW_0(F(s)), \quad i = 1, \ldots, m.$$

Note that

$$\begin{aligned}
Y_i &= Y_{i-1} + \int_{t_{i-1}}^{t_i} H(s)dW_0(F(s)) \\
&= Y_{i-1} + \int_{t_{i-1}}^{t_i} H(s)dW(F(s)) - \int_{t_{i-1}}^{t_i} H(s)f(s)dsW(1), \\
&\quad i = 1, \ldots, m
\end{aligned}$$

by (6). Our simulation algorithm is then as follows. First generate $W(F(t_1)), \ldots, W(F(t_m))$, where

$$W(F(t_i)) = W(F(t_{i-1})) + N(0, F(t_i) - F(t_{i-1})), \quad i = 1, \ldots, m.$$

For convenience let the realizations be $W(F(t_i)) = x_i$. We have immediately that $W_0(F(t_i)) = x_i - F(t_i)x_m$ for $i = 1, \ldots, m$.

Using the interpretation of Brownian bridge as the conditional process given the end points of standard Brownian motion, we have, given $W(F(t_{i-1})) = x_{i-1}$ and $W(F(t_i)) = x_i$, $W(F(t))$, $t \in [t_{i-1}, t_i]$ is equal in distribution to

$$x_{i-1} + \frac{F(t) - F(t_{i-1})}{F(t_i) - F(t_{i-1})}(x_i - x_{i-1}) + \tilde{W}(F(t) - F(t_{i-1})) - \frac{F(t) - F(t_{i-1})}{F(t_i) - F(t_{i-1})}\tilde{W}(F(t_i) - F(t_{i-1}))$$

where $\tilde{W}(\cdot)$ is a standard Brownian motion. Moreover, given the values of $W(F(t_i)) = x_i, i = 1, \ldots, m$, $\{W(F(t))\}_{t_{i-1} \le t \le t_i}$ are independent portions of sample paths, and hence

$$\begin{aligned}
\int_{t_{i-1}}^{t_i} H(s)dW(F(s)) &= \int_{t_{i-1}}^{t_i} H(s)f(s)ds\frac{x_i - x_{i-1}}{F(t_i) - F(t_{i-1})} \\
&\quad + \int_{t_{i-1}}^{t_i} H(s)d\tilde{W}(F(s)) \\
&\quad - \int_{t_{i-1}}^{t_i} H(s)f(s)ds\frac{\tilde{W}(F(t_i) - F(t_{i-1}))}{F(t_i) - F(t_{i-1})}.
\end{aligned}$$

We then have

$$\int_{t_{i-1}}^{t_i} H(s)dW(F(s)) \sim R_i := N(\mu_i, \sigma_i^2)$$

where

$$\mu_i = \int_{t_{i-1}}^{t_i} H(s)f(s)ds \frac{x_i - x_{i-1}}{F(t_i) - F(t_{i-1})}$$

and

$$\sigma_i^2 = \int_{t_{i-1}}^{t_i} H(s)^2 f(s)ds - \left(\int_{t_{i-1}}^{t_i} H(s)f(s)ds\right)^2 \frac{1}{F(t_i) - F(t_{i-1})}.$$

Therefore, to simulate (28), we first output x_i, $i = 1, \ldots, m$, and then conditional on x_i, $i = 1, \ldots, m$,

$$(Y_i, W(F(t_i))) = \left(Y_{i-1} + R_i - x_m \int_{t_{i-1}}^{t_i} H(s)f(s)ds, \right.$$
$$\left. x_i - F(t_i)x_m \right)$$

for $i = 1, \ldots, m$.

References

Adler, R., 1990. An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes. In: Institute of Mathematical Statistics Lecture Notes-Monograph Series, vol. 12.

Asmussen, S., Albrecher, H., 2010. Ruin Probabilities. World Scientific.

Bensusan, H., El Karoui, N., 2009. Microscopic modelling of population dynamic: an analysis of longevity risk. Preprint.

Billingsley, P., 1999. Convergence of Probability Measures. Wiley.

Blanchet, J., Lam, H., 2011. Importance sampling for actuarial cost analysis under a heavy traffic model. In: Proceedings of the Winter Simulation Conference, WSC. pp. 3812–3823.

Blanchet, J., Li, C., 2011. Efficient simulation for the maximum of infinite horizon Gaussian processes. Journal of Applied Probability 48 (2), 467–489.

Bowers, N., Gerber, H., Hickman, J., Jones, D., Nesbitt, C., 1997. Actuarial Mathematics. Society of Actuaries.

Dai, J.G., He, S., Tezcan, T., 2010. Many-server diffusion limits for $G/Ph/n + GI$ queues. The Annals of Applied Probability 20 (5), 1854–1890.

Decreusefond, L., Moyal, P., 2008. A functional central limit theorem for the $M/GI/\infty$ queue. The Annals of Applied Probability 18 (6), 2156–2178.

Dieker, A.B., 2005. Extremes of Gaussian processes over an infinite horizon. Stochastic Processes and their Applications 115, 207–248.

Dudley, R., 1999. Uniform Central Limit Theorems. In: Cambridge Studies in Advanced Mathematics.

Ganesh, A., O'Connell, N., Wischik, D., 2004. Big Queues. Springer.

Halfin, S., Whitt, W., 1981. Heavy-traffic limits for queues with many exponential servers. Operations Research 29 (3), 567–588.

Husler, J., Piterbarg, V., 1999. Extremes of a certain class of Gaussian processes. Stochastic Processes and their Applications 83 (2), 257–271.

Iglehart, D., 1969. Diffusion approximations in collective risk theory. Journal of Applied Probability 6, 285–292.

Karatzas, I., Shreve, S.E., 2008. Brownian Motion and Stochastic Calculus, second ed. Springer.

Kaspi, H., Ramanan, K., 2010. SPDE limits of many-server queues. Preprint.

Kingman, J.F.C., 1961. The single server queue in heavy traffic. Mathematical Proceedings of the Cambridge Philosophical Society 57, 902–904.

Kingman, J.F.C., 1962. On queues in heavy traffic. Journal of the Royal Statistical Society: Series B 24 (2), 383–392.

Mandjes, M., 2007. Large Deviations for Gaussian Queues. Wiley.

Pang, G., Whitt, W., 2010. Two-parameter heavy-traffic limits for infinite-server queues. Queueing Systems: Theory and Applications 65 (4), 325–364.

Piterbarg, V.I., 1996. Asymptotic Methods in the Theory of Gaussian Processes and Fields. American Mathematical Society.

Puhalskii, A.A., Reiman, M.I., 2000. The multiclass $GI/PH/N$ queue in the Halfin–Whitt regime. Advances in Applied Probability 32 (2), 564–595.

Reed, J., 2009. The G/GI/N queue in the Halfin–Whitt regime. The Annals of Applied Probability 19 (6), 2211–2269.

Schmidli, H., 2008. Stochastic Control in Insurance. Springer.

Steele, J.M., 2001. Stochastic Calculus and Financial Applications. Springer.

Whitt, W., 2002. Stochastic-Process Limits. Springer.