A Heavy Traffic Approach to Modeling Large Life Insurance Portfolios

Jose Blanchet and Henry Lam

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 individual risk model by taking into account the time dimension and the specific policy structure including the premium payments. It is also different from classical risk theory by building the risk process from micro-level contracts and parameters instead of assuming 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.

The study of risk processes is a central topic in actuarial science. Among the literature, the majority focuses on the calculation of ruin probability, 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 (2000) for 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 literature on many-server queues. The basic idea is that under the assumption of large number of policyholders, one can approximate the functionals of these policyholders’ statuses using fluid and diffusion approximations. In the statistics literature, these correspond to stochastic-process version 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 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 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) or 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 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 include the insurer’s cash level, liabilities, and per basis reserve level. These will be discussed in Section 2. We prove 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 ruin probabilities. We model ruin as the situation in which the liabilities surpasses 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 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 focuses 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, He and Tezcan (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 proposes 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 Igehart (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 2 we lay out our model assumptions and define the key quantities that we approximate. Section 3 is devoted to the statement of our main result and its discussion. Section 4 relates to applications, such as ruin probability computations and the identification of the autocorrelation structure of our approximating Gaussian processes. Finally, Section 5 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:

- \( \delta \): 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 (iid))
- \( f(t), F(t), \bar{F}(t) \): 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 \( H(\cdot) \) is non-decreasing and that \( H(0) < 0 \) while \( H(T) > 0 \). Moreover, we assume that \( P(\cdot) \) and \( B(\cdot) \) are continuously differentiable and have bounded first derivatives, 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 aligned with the practice that premiums are paid prior to benefit. 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 \( B(t) = be^{-\delta t} \). This insurance satisfies Assumption 2 under the Equivalence Principle (Bowers et. al. (1997)). Finally, we note that the monotonicity assumption is only needed in the optimization problem introduced in ruin calculation; see Section 3.1.

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, and we note a related paper by Blanchet and Lam (2011) that discusses the scenario of policyholder arrivals.

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

\[
N_n(t) = \sum_{i=1}^{n} I(X_i \leq t).
\] (1)
Similarly, we write $\bar{N}_n(t)$ for the number of surviving policyholders at time $T$, namely

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

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 Cash Process, Reserve Process and Average Cash Process respectively:

**Cash Process** We define the *Cash Process* as the present value at time $t$ of the total accumulated cash generated by all individual accounts, excluding 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.

$$m(t) = E[(P(X_i) - B(X_i))I(X_i \leq t) + P(t)I(X_i > t)] \quad (3)$$

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

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 $m(T) = e^{\delta t} \int_0^T H(s)f(s)ds = 0$.

**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

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

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

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

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 Reserve Process at time $t$, denoted by $\bar{C}_n(t)$, as the sum of the actuarial reserves from all surviving policies. Hence

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

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

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

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

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

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 $T$. We introduce $(W_0(t), 0 \leq t \leq 1)$ as our notation for a Brownian bridge. It turns out that $W_0(t)$ is equal in distribution to $W(t) - tW(0)$, where $W(t)$ is a Brownian motion. It is also the unique Gaussian process with mean 0 and covariance function $Cov(W_0(s), W_0(t)) = s(1-t), s \leq t$. This implies that we can write the identities in distribution (for whole stochastic processes)

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

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 Section 6. Our main result provides a joint approximation to the Cash Process, Reserve Process and Average Cash Process. The proof is given in Section 6.
**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

\[
(C_n(\cdot)/n, \bar{C}_n(\cdot)/n, V_n(\cdot)/n) \Rightarrow (m(\cdot), \bar{m}(\cdot), V(\cdot))
\]

as \(n \to \infty\). Moreover,

\[
\left(\sqrt{n}(C_n(\cdot)/n - m(\cdot)), \sqrt{n}(\bar{C}_n(\cdot)/n - \bar{m}(\cdot)), \sqrt{n}(V_n(\cdot) - V(\cdot))\right)
\]

\[
\Rightarrow \left(e^{\delta t} \left[\int_0^t H(s)dW_0(F(s)) - P(t)W_0(F(t))\right],
\right.
\]

\[
W_0(F(t))e^{\delta t} \left[\int_t^T H(s)f_s(s)ds - P(t)\right],
\]

\[
\frac{e^{\delta t}}{F(t)} \left[\int_0^t H(s)dW_0(F(s)) - P(t)W_0(F(t))\right] + \frac{V(t)}{F(t)}W_0(F(t))
\]

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 Cash Process can be approximated by

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

Simultaneously, we have that the Reserve Process admits the approximation

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

and that

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

The fluctuation around the average in the first two 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 accumulated cash 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 ruin probabilities under different types of life insurance contracts. Also in the next section we will provide closed formulae 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 fluctuates 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.

Figure 1 depicts a joint sample path of the approximating process of \( C_n(t) \) and \( \bar{C}_n(t) \) respectively. Here we use the following assumptions: \( n = 100, \delta = 0.01, T = 1 \), uniform death distribution over \([0, T]\), and a whole life insurance policy with \( b = 1 \); the premium \( p \) is calculated according to the Equivalence Principle.

![Figure 1: Approximating \( C_n(t) \) (in red) and \( \bar{C}_n(t) \) (in blue)](image)

Figure 2 shows \( V(t) \) and an approximating sample path of \( V_n(t) \) given by (11) as a process centered at \( V(t) \). Here we use the same assumptions specified for Figure 1.

We explain how to implement the simulation procedure in Section 6.2 based on the approximations (9), (10) and (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 premiums minus benefits) on individual basis.
However, the overall realized liability of the company is really given by the aggregation of the realized individual present values of the premiums minus the net present values of the paid benefits. 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 overall realized liability process and the asset process. More precisely, when the size of assets are below the statutory 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 prospective method are 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 $U_n(t) + C_n(t) - C_n(t) < 0$ (assuming a constant rate of investment interest). Under a finite-time formulation, the ruin probability is given by

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

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 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 asymptotic. The details of such will be 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, 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
\[ \sqrt{n} \left( \frac{C_n(t) - \bar{C}_n(t)}{n} - (m(t) - \bar{m}(t)) \right) \]
\[ \Rightarrow e^{\delta t} \left[ \int_0^t H(s) dW_0(F(s)) - P(t)W_0(F(t)) \right] - W_0(F(t))e^{\delta t} \left[ \int_t^T H(s) f_1(s) ds - P(t) \right] \]
\[ = e^{\delta t} \left[ \int_0^t H(s) dW_0(F(s)) - W_0(F(t)) \int_t^T H(s) f_1(s) ds \right] \]
(12)

Now the ruin probability is written as
\[ P(\text{ruin}) = P(U_n(t) + C_n(t) - \bar{C}_n(t) < 0 \text{ for some } t \in [0, T]) \]
\[ = P \left( \sup_{0 \leq t \leq T} \frac{\bar{C}_n(t) - C_n(t)}{\sqrt{n}} > ue^{\delta t} \right) \]
\[ = P \left( \sup_{0 \leq t \leq T} X(t) > u \right) (1 + o(1)) \]
(13)
as $n \to \infty$, where
\[ X(t) = \int_0^t H(s) dW_0(F(s)) - W_0(F(t)) \int_t^T H(s) f_1(s) ds \]
(14)
This follows from (12) and the fact that $X(\cdot) \overset{D}{=} -X(\cdot)$.

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 $-U_n(t)$. The approximating net asset corresponds to the simulation output of Figure 1. We use the initial surplus $u = 5$. Note that ruin happens at around time 0.58 in this scenario.

The probability (13) typically does not have closed-form solution, although a fair amount is known for sharp asymptotics of the maximum of a Gaussian process (see for instance, Husler and
Figure 3: Approximating $C_n(t) - \bar{C}_n(t)$ (in red) and deterministic trajectory of $-U_n(t)$ (in blue)

Piterbarg (1999) and Dieker (2005)); these approximations depend on the local correlation structure which is obtained in the next section.

Here we will present a formal 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 tails 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 $u_n = o(n^{1/2})$. 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

$$P \left( \sup_{0 \leq t \leq T} X(t) > u \right) \geq \sup_{0 \leq t \leq T} P(X(t) > u) \quad (15)$$

We now analyze the right hand side of (15). Using (6), $X(t)$ can be shown to be equal in distribution
by Ito’s isometry. Hence (see, for example, Adler (1990))

\[
\sigma^2(t) = \int_0^t \left( H(s) - \int_t^T H(v)f_t(v)dv \right)^2 f(s)ds + \left( \int_t^T H(s)f_t(s)ds \right)^2
\]

\[
+ 2 \int_0^t \left( H(s) - \int_t^T H(v)f_t(v)dv \right) f(s)ds \int_t^T H(s)f_t(s)ds
\]

\[
= \int_0^t H(s)^2 f(s)ds - 2 \int_t^T H(s)f_t(s)ds \int_0^t H(s)f(s)ds + \left( \int_t^T H(s)f_t(s)ds \right)^2 F(t)
\]

\[
+ \left( \int_t^T H(s)f_t(s)ds \right)^2 + 2 \int_0^t H(s)f(s)ds \int_t^T H(s)f_t(s)ds - 2 \left( \int_t^T H(s)f_t(s)ds \right)^2 F(t)
\]

\[
= \int_0^t H(s)^2 f(s)ds + F(t) \left( \int_t^T H(s)f_t(s)ds \right)^2
\]

where the second equality comes from the Equivalence Principle, and

\[
P \left( \sup_{0 \leq t \leq T} X(t) > u \right) \geq \sup_{0 \leq t \leq T} P(N(0, \sigma^2(t)) > u)
\]

(17)

On the other hand, the upper bound of the ruin probability is provided by Borell’s inequality (see, for example, Adler (1990))

\[
P \left( \sup_{0 \leq t \leq T} X(t) > u \right) \leq e^{Cu - \frac{1}{2}u^2/\sigma^2(t^*)}
\]

(18)

where \( t^* = \arg\max_{0 \leq t \leq T} \sigma^2(t) \) and \( C \) is a constant depending on \( E \sup_{0 \leq t \leq T} X(t) \).

Together, (17) and (18) give the asymptotic result

\[
\lim_{u \to \infty} \frac{1}{u^2} \log P \left( \sup_{0 \leq t \leq T} X(t) > u \right) = -\frac{1}{2\sigma^2(t^*)}
\]
To find $t^*$, one merely differentiates (16) to get
\[
\frac{d\sigma^2(t)}{dt} = -\frac{f(t)}{F(t)} \left( \int_t^T H(s)f(s)ds \right)^2 - \frac{2}{F(t)} \int_t^T H(s)f(s)dsH(t)f(t) + H(t)^2f(t)
\]
\[
= f(t) \left[ H(t)^2 - 2H(t) \int_t^T H(s)f_i(s)ds - \left( \int_t^T H(s)f_i(s)ds \right)^2 \right]
\]
\[
= f(t) \left[ H(t) - (1 + \sqrt{2}) \int_t^T H(s)f_i(s)ds \right] \left[ H(t) + (\sqrt{2} - 1) \int_t^T H(s)f_i(s)ds \right]
\]
(19)

Note that $f(t) > 0$ for all $t \in (0, T)$. Since $H(t)$ is non-decreasing, we have $\int_t^T H(s)f_i(s)ds \geq H(t)$ and so $H(t) - (1 + \sqrt{2}) \int_t^T H(s)f_i(s)ds < 0$ for all $t \in [0, T)$. Also, by the Equivalence Principle and that $H(0) < 0$ and is continuous, we have $H(t) + (\sqrt{2} - 1) \int_t^T H(s)f_i(s)ds < 0$ for a punctured neighborhood of $t = 0$ i.e. $t \in (0, \epsilon)$ for some $\epsilon > 0$. On the other hand, since $\int_t^T H(s)f_i(s)ds \geq H(t)$ and that $H(T) > 0$ and is continuous, we have $H(t) + (\sqrt{2} - 1) \int_t^T H(s)f_i(s)ds > 0$ 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 $H(t) + (\sqrt{2} - 1) \int_t^T H(s)f_i(s)ds$. Then the global maximizer is either the zero or the discontinuous point of $d\sigma^2(t)/dt$ that gives the highest value of $\sigma^2(t)$.

**Example 1:** Suppose $X_i$ follows uniform distribution on $[0, T]$. Also 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
\[
H(t) + (\sqrt{2} - 1) \int_t^T H(s)f_i(s)ds
\]
\[
= \frac{p}{\delta} - \left( \frac{p}{\delta} + b \right) e^{-\delta t} + (\sqrt{2} - 1) \left[ \frac{p}{\delta} - \frac{1}{\delta(T - t)} \left( \frac{p}{\delta} + b \right) (e^{-\delta t} - e^{-\delta T}) \right]
\]
We assume the values $T = 1, b = 1$ and $\delta = 0.01$. Then $t^* = 0.414$. Note that
\[
\sigma^2(t)
\]
\[
= \frac{1}{T} \left[ \frac{p^2}{\delta^2} - \frac{2p}{\delta^2} \left( \frac{p}{\delta} + b \right) (1 - e^{-\delta t}) + \frac{1}{2\delta} \left( \frac{p}{\delta} + b \right)^2 (1 - e^{-2\delta t}) \right]
\]
\[- \left( 1 - \frac{t}{T} \right) \left[ \frac{p}{\delta} - \frac{1}{\delta(T - t)} \left( \frac{p}{\delta} + b \right) (e^{-\delta t} - e^{-\delta T}) \right]^2
\]
Plugging in $t^*$, we have $\sigma(t^*) = 0.256$.

**Example 2:** Suppose $X_i$ follows the same distribution as above, with same values of $T$ and $b$. However, let us consider an increasing premium rate $pe^{\mu t}$ where $\mu < \delta$. Let $\mu = 0.05$. Then the
Equivalence Principle gives
\[ p = \frac{b(\delta - \mu)^2}{\delta - \mu} \frac{1 - e^{-\delta t}}{1 + e^{-(\delta - \mu)t}} \]

In this case
\[ H(t) = \frac{p}{\delta - \mu} - \frac{p}{\delta - \mu} e^{-(\delta - \mu)t} - be^{-\delta t} \]
and
\[ H(t) + (\sqrt{2} - 1) \int_t^T H(s)f_t(s)ds \]
\[ = \frac{p}{\delta - \mu} - \frac{p}{\delta - \mu} e^{-(\delta - \mu)t} - be^{-\delta t} \]
\[ + (\sqrt{2} - 1) \left[ \frac{p}{\delta - \mu} - \frac{p}{(\delta - \mu)^2} \frac{1}{T-t} (e^{-(\delta - \mu)t} - e^{-(\delta - \mu)T}) - \frac{b}{\delta(T-t)} (e^{-\delta t} - e^{-\delta T}) \right] \]

We get that \( t^* = 0.416 \). Note that
\[ \sigma^2(t) = \frac{1}{T} \left[ \frac{p^2 t}{(\delta - \mu)^2} + \frac{p^2}{2(\delta - \mu)^3} (1 - e^{-2(\delta - \mu)t}) + \frac{b^2}{2\delta} (1 - e^{-2\delta t}) - \frac{2p^2}{(\delta - \mu)^3} (1 - e^{-(\delta - \mu)t}) \right. \]
\[ - \left. \frac{2pb}{\delta(\delta - \mu)} (1 - e^{-\delta t}) + \frac{2pb}{(\delta - \mu)(2\delta - \mu)} (1 - e^{-(2\delta - \mu)t}) \right] + \left( 1 - \frac{t}{T} \right) \]
\[ \left[ \frac{p}{\delta - \mu} - \frac{p}{(\delta - \mu)^2} \frac{1}{T-t} (e^{-(\delta - \mu)t} - e^{-(\delta - \mu)T}) - \frac{b}{\delta(T-t)} (e^{-\delta t} - e^{-\delta T}) \right]^2 \]

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

**Example 3:** Assume the same setting as the last two examples, but this time with term life insurance with tenor \( l < T \). Let \( l = 0.5 \). Then
\[ P(t) = \begin{cases} \frac{p(1-e^{-\delta t})}{p(1-e^{-\delta l})} & \text{for } t \leq l \\ \frac{1}{\delta} & \text{for } t > l \end{cases} \]
and
\[ B(t) = \begin{cases} be^{-\delta t} & \text{for } t \leq l \\ 0 & \text{for } t > l \end{cases} \]
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
\[ H(t) = \begin{cases} \frac{p}{\delta} - \left( \frac{p}{\delta} + b \right) e^{-\delta t} & \text{for } t \leq l \\ \frac{p}{\delta}(1-e^{-\delta l}) & \text{for } t > l \end{cases} \]
and
\[ H(t) + (\sqrt{2} - 1) \int_t^T H(s)f_t(s)ds \]
\[ = \begin{cases} \frac{p}{\delta} - \left( \frac{p}{\delta} + b \right) e^{-\delta t} + (\sqrt{2} - 1) \left[ \frac{p}{\delta} \frac{l-t}{T-t} - \left( \frac{p}{\delta} + b \right) \frac{1}{T-t} (e^{-\delta t} - e^{-\delta l}) + \frac{T-l}{T-t} \frac{b}{\delta} (1 - e^{-\delta l}) \right] & \text{for } t \leq l \\ \sqrt{2} \frac{p}{\delta}(1-e^{-\delta l}) & \text{for } t > l \end{cases} \]
We get that \( t^* = 0.5 \). Note that
\[
\sigma^2(t) = \begin{cases} 
\frac{1}{T} \left[ \frac{\mu t}{\alpha} - \frac{2\mu}{\alpha} \left( \frac{\alpha}{\beta} + b \right) (1 - e^{-\delta t}) + \left( \frac{\alpha}{\beta} + b \right)^2 \frac{1}{2\alpha^2} (1 - e^{-2\delta t}) \right] \\
\quad + \left( 1 - \frac{t}{T} \right) \left[ \frac{\mu}{T - t} \left( \frac{\alpha}{\beta} + b \right) (1 - e^{-\delta t}) + \left( \frac{\alpha}{\beta} + b \right)^2 \frac{1}{2\alpha^2} (1 - e^{-2\delta t}) + \frac{T - t}{T - \delta} (1 - e^{-\delta t}) \right]^2 \\
\frac{1}{T} \left[ \frac{\mu t}{\alpha} - \frac{2\mu}{\alpha} \left( \frac{\alpha}{\beta} + b \right) (1 - e^{-\delta t}) + \left( \frac{\alpha}{\beta} + b \right)^2 \frac{1}{2\alpha^2} (1 - e^{-2\delta t}) + (t - l)^2 \frac{1}{2\alpha^2} (1 - e^{-\delta t})^2 \right] \\
\quad + \left( 1 - \frac{t}{T} \right) \frac{\mu^2}{\alpha^2} (1 - e^{-\delta t})^2 
\end{cases}
\]
for \( t \leq l \) and \( t > l \).

Plugging in \( t^* \), we get \( \sigma(t^*) = 0.679 \).

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 \( \sigma^2_P(t^*) \). In this case, we have to solve the problem of minimizing \( \max_{0 \leq t \leq T} \sigma^2_P(t) \) over a fixed family of insurance policies \( P \). This question will be explored in future work.

**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. \( \sigma^2(t^*) \) 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.

### 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 \( \tilde{C}_n(t) \) 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:**

\[
\begin{align*}
\text{Cov}(C_n(t), C_n(t')) &= \frac{e^{\delta(t+t')}}{\sqrt{n}} \left[ \frac{T}{0} (H(s) - P(t))(H(s) - P(t')) f(s) ds - \int_0^t (H(s) - P(t)) f_1(s) ds \int_0^{t'} (H(s) - P(t')) f_1(s) ds \right] \\
\text{Cov}(\tilde{C}_n(t), \tilde{C}_n(t')) &= \frac{e^{\delta(t+t')}}{\sqrt{n}} F(t \wedge t') \int_0^T (H(s) - P(t)) f_2(s) ds \int_0^{t'} (H(s) - P(t')) f_2(s) ds
\end{align*}
\]
Cross temporal covariance between Cash Process and Reserve Process:

\[
\frac{\text{Cov}(C_n(t), \bar{C}_n(t'))}{\sqrt{n}} = e^{\delta(t+t')} \left[ \bar{F}(t') \int_0^t (H(s) - P(t))f(s)ds - \int_0^t (H(s) - P(t))f(s)dsI(t > t') \right] \int_t^T (H(s) - P(t'))f(t')ds
\]

These in particular give:

Variance for Cash Process and Reserve Process:

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

\[
\frac{\text{Var}(\bar{C}_n(t))}{\sqrt{n}} = e^{2\delta t} F(t) \bar{F}(t) \left[ \int_t^T (H(s) - P(t))f(t)ds \right]^2
\]

Cross covariance between Cash Process and Reserve Process:

\[
\frac{\text{Cov}(C_n(t), \bar{C}_n(t'))}{\sqrt{n}} = e^{2\delta t} \bar{F}(t) \int_0^t (H(s) - P(t))f(s)ds \int_t^T (H(s) - P(t))f(t')ds
\]

From these one can calculate the temporal and cross correlations

\[
\text{Corr}(C_n(t), C_n(t')) = \frac{\text{Cov}(C_n(t), C_n(t'))}{\sqrt{\text{Var}(C_n(t))\text{Var}(C_n(t'))}}
\]

\[
\text{Corr}(\bar{C}_n(t), \bar{C}_n(t')) = \frac{\text{Cov}(\bar{C}_n(t), \bar{C}_n(t'))}{\sqrt{\text{Var}(\bar{C}_n(t))\text{Var}(\bar{C}_n(t'))}}
\]

\[
\text{Corr}(C_n(t), \bar{C}_n(t')) = \frac{\text{Cov}(C_n(t), \bar{C}_n(t'))}{\sqrt{\text{Var}(C_n(t))\text{Var}(\bar{C}_n(t'))}}
\]

Now consider the net asset process under the Equivalence Principle given by \(X(t)\sqrt{n}\) defined in
(14). We have
\[ \text{Cov}(X(t), X(t')) = \int_0^{t\wedge t'} \left( H(s) - \int_t^T H(v)f_t(v)dv \right) \left( H(s) - \int_{t'}^T H(v)f_{t'}(v)dv \right) f(s)ds \]
\[ + \int_t^T H(s)f_t(s)ds \int_0^{t'} \left( H(s) - \int_t^T H(v)f_t(v)dv \right) f(s)ds \]
\[ + \int_0^{t'} H(s)f_{t'}(s)ds \int_t^T \left( H(s) - \int_{t'}^T H(v)f_{t'}(v)dv \right) f(s)ds \]
\[ + \int_t^{t\wedge t'} H(s)f_t(s)ds \int_{t'}^{t\wedge t'} H(s)f_{t'}(s)ds \]
\[ = \int_0^{t\wedge t'} H(s)^2f(s)ds - \int_{t\wedge t'}^T H(s)f_{t\wedge t'}(s)ds \int_0^{t\wedge t'} H(s)f(s)ds + \int_{t\wedge t'}^T H(s)f_{t\wedge t'}(s)ds \int_0^{t\wedge t'} H(s)f(s)ds \]
\[ + \bar F(t \vee t') \int_t^T H(s)f_t(s)ds \int_{t'}^T H(s)f_{t'}(s)ds \]
\[ = \int_0^{t\wedge t'} H(s)^2f(s)ds - \int_{t\wedge t'}^T H(s)f_{t\wedge t'}(s)ds \int_0^{t\wedge t'} H(s)f(s)ds + 2\bar F(t \vee t') \int_t^T H(s)f_t(s)ds \int_{t'}^T H(s)f_{t'}(s)ds \]

by the Equivalence Principle in the last equality, and so in particular
\[ \text{Var}(X(t)) = \int_0^t H(s)^2f(s)ds + \bar F(t) \left( \int_t^T H(s)f_t(s)ds \right)^2 \]

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 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 heavy traffic and weak convergence theory. The second part of this appendix concerns the implementation of the simulations shown in the paper.

6.1 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.

6.1.1 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,

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

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

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

(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 $d_\infty$ 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

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

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 Cash Process and 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 $h : D[0, T] \to S$ 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 $d_\infty$ 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 $d_\infty$ 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 $X_i, i = 1, \ldots, n$ with distribution $F(x)$ supported on $[0, T]$, define

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

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

$$F_n(t) \Rightarrow F(t) \quad (20)$$

and

$$\sqrt{n}(F_n(t) - F(t)) \Rightarrow W_0(F(t)) \quad (21)$$

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

The limits in (20) and (21) 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.

6.1.2 Proofs of Main Results

Proof of Theorem 1. Our strategy is to show that processes $\sqrt{n}(C_n(\cdot)/n - m(\cdot))$, $\sqrt{n}(\bar{C}_n(\cdot)/n - \bar{m}(\cdot))$, 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 $\sqrt{n}(C_n(\cdot)/n - m(\cdot))$.

By Theorem 3 and (1), we have $N_n(t)/n \Rightarrow F(t)$ and

$$Z_n(t) := \sqrt{n} \left( \frac{N_n(t)}{n} - F(t) \right) \Rightarrow W_0(F(t)) \quad (22)$$
on $D[0, T]$. Note that (22) describes the process for the fraction of deaths over time as $n \to \infty$. Next consider the integral
\[ \int_0^t H(s)dr(s). \] (23)
Note that by Assumption 2 $H(s)$ has a continuous and bounded first derivative, so for $r \in C[0, T]$ with bounded quadratic variation, integration by parts gives
\[ \int_0^t H(s)dr(s) = H(t)r(t) - H(0)r(0) - \int_0^t H'(s)r(s)ds \] (24)
We can extend the domain of definition of (23) to the whole of $C[0, T]$ by (24). To ease notation denote $\| \cdot \|$ as the uniform norm so that $d_\infty(x, y) = \|x - y\|$. Note that for $r_1, r_2 \in C[0, T]$,
\[ \left\| \int_0^t H(s)dr_1(s) - \int_0^t H(s)dr_2(s) \right\| \leq (2\|H\| + \int_0^T |H'(s)|ds) \|r_1 - r_2\| \]
which shows continuity of $\int_0^t H(s)dr(s)$ on $r \in C[0, T]$. Now write
\[ \frac{C_n(t)}{n} = e^{\delta t} \left[ \int_0^t H(s)d \left( \frac{N_n(s)}{n} + P(t) \left( \frac{N_n(T)}{n} - \frac{N_n(t)}{n} \right) \right) \right] \]
Continuity of integral (23) and the fact that elementary operations are continuous on $C[0, T]$ yields 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 $C[0, T]$, Theorem 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 process processes $C_n(\cdot)$ is entirely similar. First, it is clear from Theorem 3 that
\[ \frac{C_n(t)}{n} = \left( \frac{N_n(T)}{n} - \frac{N_n(t)}{n} \right) e^{\delta t} \left[ P(t) - \int_t^T H(s)f(t)ds \right] \]
\[ \Rightarrow \bar{F}(t)e^{\delta t} \left[ P(t) - \int_t^T H(s)f(t)ds \right] = \bar{m}(t). \]

Moreover, a straightforward application of the continuous mapping principle and Theorem 3 (following similar continuity arguments as those given earlier) yields
\[ \sqrt{n} \left( \frac{C_n(t)}{n} - \bar{m}(t) \right) = (Z_n(T) - Z_n(t))e^{\delta t} \left[ P(t) - \int_t^T H(s)f(t)ds \right] \Rightarrow -W_0(F(t))e^{\delta t} \left[ P(t) - \int_t^T H(s)f(t)ds \right]. \]
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

\[
\sqrt{n}(V_n(t) - V(t)) = \sqrt{n} \left( V_n(t) - \frac{nV(t)F(t)}{N_n(t)} \right) + \sqrt{n} \left( \frac{nV(t)F(t)}{N_n(t)} - V(t) \right)
\]

\[
= \frac{\sqrt{n}e^{\delta t}}{N_n(t)/n} \left( \int_0^t H(s) d \left( \frac{N_n(s)}{n} \right) + P(t) \frac{\tilde{N}_n(t)}{n} - \int_0^t H(s)dF(s) - P(t)\tilde{F}(t) \right)
\]

\[
= \frac{V(t)}{N_n(T)/n - N_n(t)/n} \left( \int_0^t H(s)dZ_n(s) + P(t)(Z_n(T) - Z_n(t)) \right) - \frac{V(t)}{N_n(T)/n - N_n(t)/n} \left( Z_n(T) - Z_n(t) \right)
\]

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

\[
\left( Z_n(t), \frac{N_n(t)}{n} \right) \Rightarrow (W_0(F(t)), \tilde{F}(t))
\]

follows jointly in \( D[0, T] \times D[0, T] \). Since the limiting processes are continuous, it suffices to check that (25) is a continuous mapping from \((W_0(F(t)), \tilde{F}(t))\) 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.

\[ \square \]

### 6.2 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 advocates our use of heavy traffic approximation. Note that all the processes we introduced so far are elementary functions of

\[
\left( \int_0^t H(s)dW_0(F(s)), W_0(F(t)) \right)
\]

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

\[
Y_i = \int_{t_{i-1}}^{t_i} H(s)dW_0(F(s)), \quad i = 1, \ldots, m
\]

Note that

\[
Y_i = Y_{i-1} + \int_{t_{i-1}}^{t_i} H(s)dW_0(F(s)) = Y_{i-1} + \int_{t_{i-1}}^{t_i} H(s)dW(F(s)) - \int_{t_{i-1}}^{t_i} H(s)f(s)dsW(1), \quad i = 1, \ldots, m
\]

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

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

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

$$\int_{t_{i-1}}^{t_i} H(s)dW(F(s)) = \int_{t_{i-1}}^{t_i} H(s)f(s)ds \frac{x_i - x_{i-1}}{F(t) - F(t_{i-1})} + \int_{t_{i-1}}^{t_i} H(s)d\tilde{W}(F(s)) - \int_{t_{i-1}}^{t_i} H(s)f(s)ds \frac{\tilde{W}(F(t_i) - F(t_{i-1}))}{F(t) - F(t_{i-1})}$$

We then have

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

where

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

and

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

Therefore, to simulate (26), we first output $x_i, i = 1, \ldots, m$, and then conditional on $x_i, i = 1, \ldots, m$,

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

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

References


