J. Appl. Prob. **58,** 372–393 (2021)
*doi:*10.1017/jpr.2020.96

WEAK CONVERGENCE OF STOCHASTIC INTEGRALS WITH RESPECT TO THE STATE OCCUPATION MEASURE OF A MARKOV CHAIN

H. M. JANSEN,* *Delft University of Technology*

Abstract

Our aim is to find sufficient conditions for weak convergence of stochastic integrals with respect to the state occupation measure of a Markov chain. First, we study properties of the state indicator function and the state occupation measure of a Markov chain. In particular, we establish weak convergence of the state occupation measure under a scaling of the generator matrix. Then, relying on the connection between the state occupation measure and the Dynkin martingale, we provide sufficient conditions for weak convergence of stochastic integrals with respect to the state occupation measure. We apply our results to derive diffusion limits for the Markov-modulated Erlang loss model and the regime-switching Cox–Ingersoll–Ross process.

Keywords: Markov chain; state occupation measure; stochastic integral; diffusion limit; Markov modulation; regime switching

2010 Mathematics Subject Classification: Primary 60F17
Secondary 60H05

1. Introduction

Stochastic integrals with respect to the state occupation measure of a Markov chain arise naturally in the analysis of queueing systems and diffusions under Markov modulation. We are interested in the weak convergence properties of this type of stochastic integral, as they play a key role in the derivation of scaling limits for such processes.

To make the problem concrete, we introduce some notation. Let $Y \cdot X$ denote the Itô integral of Y with respect to X, and let $\Rightarrow$ denote weak convergence. In addition, let H_n and G_n be stochastic processes satisfying $H_n \Rightarrow H$ and $G_n \Rightarrow G$, with H_n being a suitable integrand and G_n denoting the (scaled and centered) state occupation measure of an irreducible continuous-time Markov chain. We would like to find conditions under which the convergence

$$H_n \cdot G_n \Rightarrow H \cdot G \tag{1.1}$$

holds as well.

Rather remarkably, this case does not seem to be covered by the known results dealing with convergence as in (1.1). To guarantee convergence as in (1.1), it is typically required that G_n is a martingale or that G_n has the so-called P-UT property (cf. [5, 6, 9, 16]). However, neither of these requirements is satisfied if G_n is the state occupation measure of a Markov chain, even though G_n has very nice convergence properties in this important case. An exception is

Received 22 November 2019; revision received 14 September 2020.
* Postal address: Delft Institute of Applied Mathematics, Delft University of Technology, Van Mourik Broekmanweg 6, 2628 XE Delft, the Netherlands. Email address: h.m.jansen@tudelft.nl

[10], which considers a class of Markov-modulated ordinary differential equations that have bounded integrands and feature the state occupation measure as an integrator. The proof there relies on integration by parts under an appropriate differentiability condition, after which the P-UT machinery can be utilized.

The goal of this paper is to formulate practical conditions that guarantee convergence as in (1.1) and can be easily applied to relevant examples such as queueing systems and mean-reverting diffusions under Markov modulation. Because the state occupation measure G_n is a given, we have to impose restrictions on the integrand H_n to obtain convergence as in (1.1). The key insight is that convergence of $H_n \cdot G_n$ is related to the behavior of the total variation process of H_n. Under the condition that this total variation process does not grow too quickly, we prove that (1.1) holds. Relying on tightness arguments, we extend this result and show weak convergence of integral equations of the form

$$Y_n(t) = X_n(t) + \int_0^t H_n(s)\,\mathrm{d}G_n(s) + \int_0^t \Gamma_{\gamma_n}(Y_n)(s)\,\mathrm{d}\tfrac{1}{\sqrt{n}}G_n(s) + \int_0^t \Delta_{\delta_n}(Y_n)(s)\,\mathrm{d}s,$$

where Γ_{γ_n} and Δ_{δ_n} are functions mapping right-continuous paths to right-continuous paths. We demonstrate the relevance of these results by applying them to two examples, in which we derive diffusion limits of the Erlang loss model and the Cox–Ingersoll–Ross (CIR) process under Markov modulation.

The remainder of this paper is organized as follows. In Section 2, we introduce notation and collect a number of basic results needed to prove the main results. In particular, we derive properties of an irreducible, continuous-time Markov chain, its state occupation measure, and its Dynkin martingale. We also establish weak convergence of the state occupation measure. In Section 3, we state and prove the main results in two theorems. The first theorem concerns weak convergence of stochastic integrals with respect to the state occupation measure G_n and provides conditions under which (1.1) holds. The second theorem extends this to a class of stochastic integral equations involving G_n. In Section 4, we apply the main results to derive the diffusion limit for the Markov-modulated Erlang loss model and to establish a small-noise limit for the Markov-modulated CIR process. In Section 5, we draw conclusions and point out some directions for further research. The appendix explores the P-UT property and its relation to G_n in some more detail. It also contains two technical lemmas that are important for proving the main results.

2. Preliminaries

We consider stochastic processes X defined on the interval $[0, \infty)$ and taking values in $\mathbb{R}^p$. We interpret vectors in $\mathbb{R}^p$ as column vectors and equip $\mathbb{R}^p$ with the usual Euclidian norm $\|\cdot\|$. Unless stated otherwise, we assume that X is càdlàg, meaning that its paths are right-continuous and admit finite left-hand limits. We denote the space of càdlàg paths on $[0, \infty)$ with values in $\mathbb{R}^p$ by $\mathbb{D}([0, \infty); \mathbb{R}^p)$, and we assume that $\mathbb{D}([0, \infty); \mathbb{R}^p)$ is equipped with the Skorokhod J_1 topology (cf. [5, Chapter VI]). Weak convergence is denoted by $\Rightarrow$. We refer to uniform convergence on compacts in probability as ucp convergence. The element in $\mathbb{D}([0, \infty); \mathbb{R}^p)$ that is identically equal to 0 is denoted by η_0. We often refer to η_0 as the zero process. We let c be a positive constant that may change from line to line.

Throughout this paper, J denotes a right-continuous, irreducible, continuous-time Markov chain with state space $\{1, \ldots, d\}$ for some $d \in \mathbb{N}$. We denote by $\mathcal{Q}$ the $d \times d$ generator matrix corresponding to J. The state indicator function of J is the $\{0, 1\}^d$-valued process K defined via $K(i; t) = \mathbf{1}_{\{J(t)=i\}}$ for $i \in \{1, \ldots, d\}$ and $t \geq 0$, so it takes values in the set of unit vectors. The

https://doi.org/10.1017/jpr.2020.96 Published online by Cambridge University Press

process K is closely related to the state occupation measure, which is the $\mathbb{R}^d$-valued stochastic process $L(t) = \int_0^t K(s)\,\mathrm{d}s$. On an intuitive level, the state indicator function K registers in which state J is, while the state occupation measure L measures how much time J has spent in each state up to a certain time.

2.1. Basic properties of the deviation matrix

Anticipating upcoming results, we present a number of equalities. Given the irreducible generator matrix $\mathcal{Q}$, we let the $d \times 1$ column vector π denote its stationary distribution, so π is the unique probability vector solving the equation $\pi^\top \mathcal{Q} = 0$. Additionally, D denotes the deviation matrix corresponding to $\mathcal{Q}$; its entries are given by

$$D_{ij} = \int_0^\infty \big(\mathbb{P}(J(s) = j \mid J(0) = i) - \pi_j\big)\,\mathrm{d}s.$$

The integral is well defined, because the irreducibility of $\mathcal{Q}$ implies that the probability $\mathbb{P}(J(t) = j \mid J(0) = i)$ converges exponentially fast to π_j as $t \to \infty$ (cf. [3, p. 356]). Thus, the deviation matrix D provides a measure for how much the Markov chain J deviates from its stationary distribution if it starts in a fixed point.

Following [3], we define the ergodic matrix $\Pi = \mathsf{1}\pi^\top$ and the fundamental matrix $F = D + \Pi$, where $\mathsf{1}$ denotes a $d \times 1$ vector with each entry being 1. Some straightforward arguments (cf. [3]) demonstrate that $\pi^\top D = 0$ and

$$\mathcal{Q}F = F\mathcal{Q} = \Pi - I = D\mathcal{Q} = \mathcal{Q}D. \tag{2.1}$$

Applying these identities, we find that

$$(\mathcal{Q}F)^\top \mathrm{diag}(\pi)F = (\mathcal{Q}D)^\top \mathrm{diag}(\pi)D + (\mathcal{Q}D)^\top \mathrm{diag}(\pi)\Pi,$$

while $(\mathcal{Q}D)^\top \mathrm{diag}(\pi)D = -\mathrm{diag}(\pi)D$ and $(\mathcal{Q}D)^\top \mathrm{diag}(\pi)\Pi = 0$. This leads to the equality

$$F^\top\big(\mathcal{Q}^\top \mathrm{diag}(\pi) + \mathrm{diag}(\pi)\mathcal{Q}\big)F = -\big(\mathrm{diag}(\pi)D + D^\top \mathrm{diag}(\pi)\big). \tag{2.2}$$

Given the irreducible generator matrix $\mathcal{Q}$, the vectors and matrices $\mathsf{1}$, π, Π, F, and D are always as defined above.

2.2. The Dynkin martingale

Markov chains are closely connected to martingales via Dynkin's formula. In the next result (which follows from [1, Lemma 2.6.18] and [1, Lemma 3.8.5]), we define a martingale M that is the Dynkin martingale corresponding to J. Additionally, we note that M is a locally square-integrable martingale. For this class of martingales there are powerful convergence results available, which often depend on the predictable quadratic variation process of such martingales converging in a suitable manner. One of these results is the martingale central limit theorem (MCLT). We would like to invoke it later on, so we present the explicit form of the predictable quadratic variation process of M as well.

Lemma 2.1. *The process M defined via*

$$M(t) = K(t) - K(0) - \int_0^t \mathcal{Q}^\top K(s)\,\mathrm{d}s \tag{2.3}$$

https://doi.org/10.1017/jpr.2020.96 Published online by Cambridge University Press

is a càdlàg, locally square-integrable martingale having predictable quadratic variation process

$$\langle M\rangle(t) = \int_0^t \operatorname{diag}(\mathcal{Q}^\top K(s))\,\mathrm{d}s - \int_0^t \mathcal{Q}^\top \operatorname{diag}(K(s))\,\mathrm{d}s - \int_0^t \operatorname{diag}(K(s))\mathcal{Q}\,\mathrm{d}s.$$

We refer to the process M defined above as the Dynkin martingale associated with the Markov chain J.

2.3. Scaling the Markov chain

In the remainder of this paper we are mainly concerned with the Markov chain J_n, which is a scaled version of J. Formally, we fix $\alpha > 0$ and let J_n denote a continuous-time Markov chain with state space $\{1, \ldots, d\}$ and generator matrix $n^\alpha \mathcal{Q}$, where $n \in \mathbb{N}$. As usual, we assume that J_n has right-continuous paths. Note that we may obtain J_n by applying the time scaling $J_n(t) = J(n^\alpha t)$, so J_n is essentially a sped-up version of J.

The state indicator function of J_n is K_n, while the corresponding state occupation measure is L_n and the corresponding Dynkin martingale is M_n. We let G_n denote a scaled and centered version of the state occupation measure L_n, with

$$G_n(t) = n^{\alpha/2} \int_0^t (K_n(s) - \pi)\,\mathrm{d}s. \tag{2.4}$$

This process is connected to M_n via

$$G_n(t) = n^{-\alpha/2} F^\top M_n(t) - n^{-\alpha/2} F^\top (K_n(t) - K_n(0)), \tag{2.5}$$

which follows from (2.1) and (2.3).

The process G_n in (2.4) is the process that we would like to use as an integrator. Therefore, it is the most important object in this paper. For ease of exposition, we often abuse terminology and refer to G_n as the state occupation measure, leaving out the fact that it is scaled and centered in a specific way.

2.4. Weak convergence of the state occupation measure

A first step towards proving the main result is to derive the weak convergence of the Dynkin martingale M_n and the state occupation measure G_n. We settle this in the next lemma. In particular, it shows that the fluctuations of G_n are well described by a Brownian motion whose predictable quadratic variation process strongly depends on the deviation matrix D of the underlying Markov chain. Its proof relies on a double application of the MCLT.

Lemma 2.2. *For $n \to \infty$, the stochastic process $n^{-\alpha/2} M_n$ converges weakly to a Brownian motion C having predictable quadratic variation process*

$$\langle C\rangle(t) = -\big(\mathcal{Q}^\top \operatorname{diag}(\pi) + \operatorname{diag}(\pi)\mathcal{Q}\big)t.$$

Additionally, for $n \to \infty$, the stochastic process G_n converges weakly to a Brownian motion B having predictable quadratic variation process

$$\langle B\rangle(t) = \big(\operatorname{diag}(\pi)D + D^\top \operatorname{diag}(\pi)\big)t. \tag{2.6}$$

https://doi.org/10.1017/jpr.2020.96 Published online by Cambridge University Press

Proof. We first show how the second statement follows from the first. Suppose that $\hat{M}_n = n^{-\alpha/2} M_n$ converges weakly to the Brownian motion C. In this case, the process $-n^{\alpha/2} \int_0^t \mathcal{Q}^\top K_n(s)\,\mathrm{d}s$ must converge weakly to C as well, due to (2.5). Then the process

$$G_n(t) = n^{\alpha/2} \int_0^t (K_n(s) - \pi)\,\mathrm{d}s = -n^{\alpha/2} \int_0^t F^\top \mathcal{Q}^\top K_n(s)\,\mathrm{d}s$$

converges weakly to the Brownian motion $B = F^\top C$, so

$$\langle B\rangle(t) = F^\top \big(- \big(\mathcal{Q}^\top \mathrm{diag}(\pi) + \mathrm{diag}(\pi)\mathcal{Q}\big)t\big)F = \big(\mathrm{diag}(\pi)D + D^\top \mathrm{diag}(\pi)\big)t.$$

For a justification of the last equality, see (2.2).

In view of the previous considerations, it suffices to prove that $\hat{M}_n$ converges weakly to the Brownian motion C. We would like to invoke the MCLT (cf. [16, Theorem 2.1]) to establish this convergence. To this end, we have to verify several properties: we need ucp convergence of the predictable quadratic variation process $\langle \hat{M}_n\rangle$ to $\langle C\rangle$, together with bounds on the maximum jump sizes of $\hat{M}_n$ and $\langle \hat{M}_n\rangle$.

As a first step, we use Lemma 2.1 to obtain that

$$\langle \hat{M}_n\rangle(t) = \int_0^t \mathrm{diag}(\mathcal{Q}^\top K_n(s))\,\mathrm{d}s - \int_0^t \mathcal{Q}^\top \mathrm{diag}(K_n(s))\,\mathrm{d}s - \int_0^t \mathrm{diag}(K_n(s))\mathcal{Q}\,\mathrm{d}s. \tag{2.7}$$

Clearly, $\langle \hat{M}_n\rangle$ is continuous and the jumps of each entry of $\hat{M}_n$ are bounded by $n^{-\alpha/2}$, so the maximum jump size of $\hat{M}_n$ and $\langle \hat{M}_n\rangle$ converges to 0 as $n \to \infty$. In this case, the MCLT implies that weak convergence of $\hat{M}_n$ to C follows from $\langle \hat{M}_n\rangle$ converging ucp to $\langle C\rangle$.

The key to proving convergence of $\langle \hat{M}_n\rangle$ to $\langle C\rangle$ is the convergence of $n^{-\alpha} M_n$ to the zero process η_0. To establish the latter convergence, we again rely on the MCLT. Clearly, $\langle n^{-\alpha} M_n\rangle = n^{-\alpha}\langle \hat{M}_n\rangle$ and $\langle \hat{M}_n\rangle$ is bounded on compact intervals, so $\langle n^{-\alpha} M_n\rangle$ converges ucp to η_0. Additionally, the maximum jump size of $n^{-\alpha} M_n$ and $\langle n^{-\alpha} M_n\rangle$ converges to 0 as $n \to \infty$, so the MCLT implies that $n^{-\alpha} M_n$ converges ucp to η_0.

Recall that we aim to prove that $\langle \hat{M}_n\rangle$ converges ucp to $\langle C\rangle$. Because we showed that $n^{-\alpha} M_n$ converges ucp to η_0 and K_n is bounded by 1, it follows from the definition of M_n in (2.3) that

$$-\int_0^t F^\top \mathcal{Q}^\top K_n(s)\,\mathrm{d}s \tag{2.8}$$

converges ucp to η_0, too. From the matrix equalities related to the deviation matrix D and the fundamental matrix F we get

$$F^\top \mathcal{Q}^\top K_n(s) = (\mathcal{Q}F)^\top K_n(s) = \big(\pi \mathbf{1}^\top - I\big)K_n(s) = \pi - K_n(s).$$

Combining this with the convergence of the process in (2.8), we conclude that the process $\int_0^t (K_n(s) - \pi)\,\mathrm{d}s$ converges ucp to η_0. This implies that $\langle \hat{M}_n\rangle$ presented in (2.7) converges ucp to

$$\int_0^t \mathrm{diag}(\mathcal{Q}^\top \pi)\,\mathrm{d}s - \int_0^t \mathcal{Q}^\top \mathrm{diag}(\pi)\,\mathrm{d}s - \int_0^t \mathrm{diag}(\pi)\mathcal{Q}\,\mathrm{d}s = -\big(\mathcal{Q}^\top \mathrm{diag}(\pi) + \mathrm{diag}(\pi)\mathcal{Q}\big)t.$$

The last equality is based on the fact that $\pi^\top \mathcal{Q} = 0$. We conclude that $\langle \hat{M}_n\rangle$ converges ucp to $\langle C\rangle$, which establishes weak convergence of $\hat{M}_n$ to C. □

https://doi.org/10.1017/jpr.2020.96 Published online by Cambridge University Press

3. Main results

In this section we state and prove our main results, which we present in two theorems. The first theorem concerns weak convergence of stochastic integrals with respect to the state occupation measure G_n. The second theorem partly relies on the first and concerns weak convergence for a rather general class of stochastic integral equations involving G_n. These results are the key to deriving the diffusion limits for the Markov-modulated Erlang loss model and the regime-switching CIR process, which we focus on in the next section.

3.1. Stochastic integrals with respect to the state occupation measure

Convergence of stochastic integrals with respect to a semimartingale X_n is a delicate subject in general. Even if H_n and X_n are well-behaved deterministic processes converging uniformly to the zero process, the integral $H_n \cdot X_n$ may not converge as $n \to \infty$. Nevertheless, there are two well-known cases in which the analysis simplifies considerably. The first case deals with X_n being a martingale. Then $H_n \cdot X_n$ is typically a martingale, which may be analyzed using tools such as the MCLT. The second case (partly covering the first) deals with X_n being P-UT. Then $(H_n, X_n) \Rightarrow (H, X)$ implies that $H_n \cdot X_n \Rightarrow H \cdot X$ under mild conditions, according to [5, Theorem VI.6.22].

However, if we integrate against the state occupation measure G_n, neither the first nor the second case applies. Indeed, G_n is not a martingale and does not satisfy the P-UT property, as we show in the appendix. We get around this problem by restricting the integrands to be processes of finite variation that converge in a specific way. Under this restriction, we exploit properties of both the Dynkin martingale and the state indicator function to prove weak convergence of stochastic integrals with respect to G_n.

We proceed to develop this idea in the following theorem, which is the first main result of this paper. The statement of the theorem also features an auxiliary process Z_n. It is not relevant for the proof, but its inclusion can be quite useful for applications.

Theorem 3.1. *For fixed $m \in \mathbb{N}$, let $H_{1,n}, \ldots, H_{m,n}$, and Z_n be càdlàg processes, with $H_{1,n}, \ldots, H_{m,n}$ taking values in $\mathbb{R}^{a \times d}$ and Z_n in $\mathbb{R}^b$. Assume that these processes are adapted to some underlying filtration with respect to which the Dynkin martingale M_n is still a martingale. Also assume that each entry of $n^{-\alpha/2} H_{k,n}$ is a finite variation process whose total variation process converges ucp to the zero process η_0. If*

$$(H_{1,n}, \ldots, H_{m,n}, G_n, Z_n) \Rightarrow (H_1, \ldots, H_m, B, Z) \tag{3.1}$$

for $n \to \infty$, then

$$\begin{aligned} &(H_{1,n} \cdot G_n, \ldots, H_{m,n} \cdot G_n, H_{1,n}, \ldots, H_{m,n}, G_n, Z_n) \\ &\quad \Rightarrow (H_1 \cdot B, \ldots, H_m \cdot B, H_1, \ldots, H_m, B, Z) \end{aligned} \tag{3.2}$$

for $n \to \infty$. The process B is a Brownian motion whose predictable quadratic variation process is given by (2.6).

Proof. We first summarize some known results. According to Lemma 2.2, the martingale $n^{-\alpha/2} M_n$ converges weakly to a Brownian motion C and the state occupation measure G_n converges weakly to $B = F^\top C$, which has the predictable quadratic variation process given by (2.6). We also note that $H_{k,n} \cdot G_n = H_{k,n}^- \cdot G_n$, with X^- being the left-hand limit of a càdlàg process X.

https://doi.org/10.1017/jpr.2020.96 Published online by Cambridge University Press

We have established a relation between G_n and $n^{-\alpha/2}M_n$ in (2.5). This relation implies that

$$\big(H^-_{k,n}\cdot G_n\big)(t) = \int_0^t H^-_{k,n}(s)\,\mathrm{d}n^{-\alpha/2}F^\top M_n(s) + R_{k,n}(t),$$

where

$$R_{k,n}(t) = -\int_0^t n^{-\alpha/2}H^-_{k,n}(s)F^\top\,\mathrm{d}K_n(s).$$

We now verify that $R_{k,n}$ converges weakly to η_0. Its (i,j)th entry is given by

$$R_{k,n,i,j}(t) = -\int_0^t \tilde{H}_{k,n,i,j}(s)\,\mathrm{d}\mathbf{1}_{\{J_n(s)=j\}},$$

where $\tilde{H}_{k,n,i,j}$ is the (i,j)th entry of $n^{-\alpha/2}H^-_{k,n}F^\top$. We denote the total variation process of $\tilde{H}_{k,n,i,j}$ by $V_{k,n,i,j}$. The crucial observation here is that the process $\mathbf{1}_{\{J_n(s)=j\}}$ is right-continuous and jumps between 0 and 1. Therefore, we can apply Lemma A.1 to get

$$\sup_{0\le s\le t}\Big|\int_0^s \tilde{H}_{k,n,i,j}(r)\,\mathrm{d}\mathbf{1}_{\{J_n(r)=j\}}\Big| \le V_{k,n,i,j}(t) + \sup_{0\le s\le t}|\tilde{H}_{k,n,i,j}(s)|. \tag{3.3}$$

The process $H_{k,n}$ is of finite variation and converges weakly to H_k, while the total variation process of $n^{-\alpha/2}H_{k,n}$ converges ucp to η_0. As a consequence, both $\tilde{H}^-_{k,n}$ and its total variation process converge ucp to η_0, too. Combining this with the inequality in (3.3), it follows that $R_{k,n}$ converges weakly to η_0.

The previous arguments show that $H^-_{k,n}\cdot G_n = H^-_{k,n}\cdot\tilde{M}_n + R_{k,n}$, where $\tilde{M}_n = n^{-\alpha/2}F^\top M_n$. The processes $R_{k,n}$ converge weakly to η_0, so it suffices to prove that

$$(H^-_{1,n}\cdot\tilde{M}_n, \ldots, H^-_{m,n}\cdot\tilde{M}_n, H_{1,n}, \ldots, H_{m,n}, G_n, Z_n) \tag{3.4}$$

converges weakly to the limiting vector of stochastic integrals in (3.2).

We exploit the P-UT framework from [5, Theorem VI.6.22] to derive weak convergence of the processes in (3.4). As a first step, recall that M_n is a locally square-integrable martingale and that $n^{-\alpha/2}M_n$ converges weakly to C, so $\tilde{M}_n = n^{-\alpha/2}F^\top M_n$ is a locally square-integrable martingale that converges weakly to B. Moreover, the jumps of M_n are bounded by 1, so $\tilde{M}_n$ has bounded jumps, too. Then, [5, Corollary VI.6.29] implies that the sequence of martingales $\tilde{M}_n$ has the P-UT property.

For notational convenience, we define $\tilde{M}_{k,n} = \tilde{M}_n$ and $B_k = B$ for $k = 1, \ldots, m$. For an application of [5, Theorem VI.6.22], we have to verify that

$$\big(H_{1,n}, \ldots, H_{m,n}, \tilde{M}_{1,n}, \ldots, \tilde{M}_{m,n}, Z_n\big) \Rightarrow \big(H_1, \ldots, H_m, B_1, \ldots, B_m, Z\big). \tag{3.5}$$

The validity of this weak convergence result follows from (2.5) and (3.1). With $\tilde{M}_n$ being P-UT and having the convergence in (3.5) at our disposal, we invoke [5, Theorem VI.6.22] to obtain the weak convergence of the vector of stochastic integrals in (3.4) to the limit vector of stochastic integrals in (3.2). As argued before, this establishes the weak convergence in (3.2). □

https://doi.org/10.1017/jpr.2020.96 Published online by Cambridge University Press

3.2. Stochastic integral equations involving the state occupation measure

The goal of this paper is to give practical conditions for weak convergence that can be easily applied to relevant examples involving Markov modulation. Therefore, we also introduce the stochastic integral equation

$$Y_n(t) = X_n(t) + \int_0^t H_n(s)\,\mathrm{d}G_n(s) + \int_0^t \Gamma_{\gamma_n}(Y_n)(s)\,\mathrm{d}\bar{G}_n(s) + \int_0^t \Delta_{\delta_n}(Y_n)(s)\,\mathrm{d}s, \tag{3.6}$$

where we define $\bar{G}_n = n^{-\alpha/2}G_n$. The process Y_n takes values in $\mathbb{R}^p$, while the functions Γ_{γ_n} and Δ_{δ_n} map $\mathbb{D}([0,\infty), \mathbb{R}^p)$ into $\mathbb{D}([0,\infty), \mathbb{R}^{p\times d})$ and $\mathbb{D}([0,\infty), \mathbb{R}^p)$, respectively; γ_n and δ_n are parameters. We impose three natural conditions on Γ_{γ_n} and Δ_{δ_n}. First, we assume that these functions are continuous with respect to the Skorokhod J_1 topology. Second, we assume that these functions are uniformly Lipschitz continuous with respect to the supremum norm, meaning that $\sup_{0\le t\le T} \|\Gamma_{\gamma_n}(x)(t) - \Gamma_{\gamma_n}(y)(t)\| \le c \sup_{0\le t\le T} \|x(t) - y(t)\|$ for all possible parameter values γ_n. This implies in particular that (3.6) has a unique solution. Third, we assume that these functions are continuous in their parameters in the sense that $\Gamma_{\gamma_n}(x) - \Gamma_{\gamma}(x) \to \eta_0$ in $\mathbb{D}([0,\infty), \mathbb{R}^{p\times d})$ if $\gamma_n \to \gamma$.

The next theorem, which is the second main result of this paper, shows that Y_n converges weakly to the solution of the stochastic integral equation

$$Y(t) = X(t) + \int_0^t H(s)\,\mathrm{d}B(s) + \int_0^t \Delta_{\delta}(Y)(s)\,\mathrm{d}s, \tag{3.7}$$

provided that $(H_n, G_n, X_n) \Rightarrow (H, B, X)$, $\gamma_n \to \gamma$, $\delta_n \to \delta$, and some additional mild conditions are met. The proof relies on Theorem 3.1 as well as tightness arguments. We give examples of the use of Theorem 3.2 in the next section.

Theorem 3.2. *Impose the conditions of Theorem 3.1. Additionally, let H_n and X_n be càdlàg processes, with H_n taking values in $\mathbb{R}^{p\times d}$ and X_n in $\mathbb{R}^p$. Assume that all processes involved are adapted to some underlying filtration with respect to which the Dynkin martingale M_n is still a martingale. Also assume that each entry of $n^{-\alpha/2}H_n$ is a finite variation process whose total variation process converges ucp to the zero process η_0. If $\gamma_n \to \gamma$, $\delta_n \to \delta$, and $(H_{1,n}, \ldots, H_{m,n}, H_n, G_n, X_n, Z_n) \Rightarrow (H_1, \ldots, H_m, H, B, X, Z)$ for $n \to \infty$, then*

$$\begin{aligned}&(H_{1,n}\cdot G_n, \ldots, H_{m,n}\cdot G_n, H_n\cdot G_n, H_{1,n}, \ldots, H_{m,n}, H_n, G_n, X_n, Y_n, Z_n)\\ &\quad\Rightarrow (H_1\cdot B, \ldots, H_m\cdot B, H\cdot B, H_1, \ldots, H_m, H, B, X, Y, Z),\end{aligned} \tag{3.8}$$

for $n \to \infty$, where Y_n and Y are the unique solutions to (3.6) and (3.7), respectively. The process B is a Brownian motion whose predictable quadratic variation process is given by (2.6).

Proof. It follows from Theorem 3.1 that

$$\begin{aligned}&(H_{1,n}\cdot G_n, \ldots, H_{m,n}\cdot G_n, H_n\cdot G_n, H_{1,n}, \ldots, H_{m,n}, H_n, G_n, X_n, Z_n)\\ &\quad\Rightarrow (H_1\cdot B, \ldots, H_m\cdot B, H\cdot B, H_1, \ldots, H_m, H, B, X, Z),\end{aligned} \tag{3.9}$$

which is just the convergence in (3.8) without the processes Y_n and Y. If we prove weak convergence of Y_n to Y, then joint convergence with (3.9) is a direct consequence of [5, Proposition VI.2.2], because the jumps of Y_n coincide with the jumps of X_n. Therefore, it remains to show that $Y_n \Rightarrow Y$.

https://doi.org/10.1017/jpr.2020.96 Published online by Cambridge University Press

We complete the proof in two steps. In both steps, a crucial role is played by the stochastic process $\hat{Y}_n$ given by

$$\hat{Y}_n(t) = X_n(t) + \int_0^t H_n(s)\,\mathrm{d}G_n(s) + \int_0^t \Gamma_\gamma(\hat{Y}_n)(s)\,\mathrm{d}\bar{G}_n(s) + \int_0^t \Delta_\delta(\hat{Y}_n)(s)\,\mathrm{d}s, \tag{3.10}$$

which is the solution to (3.6) with γ_n replaced by γ and δ_n replaced by δ. In the first step, we show that Y_n is asymptotically equivalent to $\hat{Y}_n$ if $\hat{Y}_n$ converges weakly. In the second step, we show that $\hat{Y}_n \Rightarrow Y$, which implies that $Y_n \Rightarrow Y$ due to the asymptotic equivalence.

For the first step, suppose that $\hat{Y}_n$ converges weakly. To establish the asymptotic equivalence of Y_n and $\hat{Y}_n$, note that

$$\begin{aligned}\left\|Y_n(t) - \hat{Y}_n(t)\right\| \le{}& \left\|\int_0^t \left(\Gamma_{\gamma_n}(Y_n)(s) - \Gamma_{\gamma_n}(\hat{Y}_n)(s)\right)\mathrm{d}\bar{G}_n(s)\right\| \\ &+ \left\|\int_0^t \left(\Gamma_{\gamma_n}(\hat{Y}_n)(s) - \Gamma_{\gamma}(\hat{Y}_n)(s)\right)\mathrm{d}\bar{G}_n(s)\right\| \\ &+ \left\|\int_0^t \left(\Delta_{\delta_n}(Y_n)(s) - \Delta_{\delta_n}(\hat{Y}_n)(s)\right)\mathrm{d}s\right\| \\ &+ \left\|\int_0^t \left(\Delta_{\delta_n}(\hat{Y}_n)(s) - \Delta_{\delta}(\hat{Y}_n)(s)\right)\mathrm{d}s\right\|\end{aligned}$$

and thus

$$\left\|Y_n(t) - \hat{Y}_n(t)\right\| \le I_{0,n}(t) + c\int_0^t \left\|Y_n(s) - \hat{Y}_n(s)\right\|\,\mathrm{d}s$$

for every $t \in [0, T]$, where

$$I_{0,n}(t) = \left\|\int_0^t \left(\Gamma_{\gamma_n}(\hat{Y}_n)(s) - \Gamma_{\gamma}(\hat{Y}_n)(s)\right)\mathrm{d}\bar{G}_n(s)\right\| + \left\|\int_0^t \left(\Delta_{\delta_n}(\hat{Y}_n)(s) - \Delta_{\delta}(\hat{Y}_n)(s)\right)\mathrm{d}s\right\|.$$

The last inequality above is based on the Lipschitz property of Γ_{γ_n} and Δ_{δ_n}. An application of Gronwall's lemma (cf. [8, pp. 287–288]) shows that

$$\sup_{0\le t\le T}\left\|Y_n(t) - \hat{Y}_n(t)\right\| \le \left(\sup_{0\le t\le T} I_{0,n}(t)\right)\mathrm{e}^{cT}.$$

With $\hat{Y}_n$ converging weakly and the functions Γ_{γ_n} and Δ_{δ_n} being continuous in their parameters, it follows that $\sup_{0\le t\le T} I_{0,n}(t)$ converges to 0 in probability, so Y_n and $\hat{Y}_n$ are stochastically equivalent if $\hat{Y}_n$ converges weakly.

For the second step, we define

$$I_{1,n}(t) = \int_0^t \Gamma_\gamma(\hat{Y}_n)(s)\,\mathrm{d}\bar{G}_n(s), \qquad I_{2,n}(t) = \int_0^t \Delta_\delta(\hat{Y}_n)(s)\,\mathrm{d}s$$

to ease notation. We aim to show that $\hat{Y}_n \Rightarrow Y$, which implies that $Y_n \Rightarrow Y$ in view of the asymptotic equivalence of Y_n and $\hat{Y}_n$. We take the classical tightness approach to prove that $\hat{Y}_n \Rightarrow Y$. First, we establish that $\hat{Y}_n$ is stochastically bounded (meaning that $\sup_{0\le t\le T}\|\hat{Y}_n(t)\|$

https://doi.org/10.1017/jpr.2020.96 Published online by Cambridge University Press

is tight for every $T > 0$). This implies stochastic boundedness of $I_{1,n}$ and $I_{2,n}$. Using this, we argue that $I_{1,n}$ and $I_{2,n}$ are C-tight, from which we derive the tightness of $\hat{Y}_n$. Finally, we prove that every converging subsequence of $\hat{Y}_n$ converges weakly to Y, demonstrating that $\hat{Y}_n \Rightarrow Y$.

We start by establishing stochastic boundedness of $\hat{Y}_n$. The Lipschitz property of Γ_γ and Δ_δ implies that

$$\begin{aligned}\|\hat{Y}_n(t)\| &\le \|X_n(t)\| + \Big\| \int_0^t H_n(s)\,\mathrm{d}G_n(s)\Big\| + \int_0^t \big\|\Gamma_\gamma(\hat{Y}_n)(s)(K_n(s)-\pi)\big\|\,\mathrm{d}s \\ &\quad + \int_0^t \big\|\Delta_\delta(\hat{Y}_n)(s)\big\|\,\mathrm{d}s \\ &\le \|X_n(t)\| + \left\| \int_0^t H_n(s)\,\mathrm{d}G_n(s)\right\| + c\int_0^t \left(1 + \sup_{0\le u\le s}\big\|\hat{Y}_n(u)\big\|\right)\mathrm{d}s.\end{aligned}$$

As before, an application of Gronwall's lemma then leads to the inequality

$$\sup_{0\le t\le T}\big\|\hat{Y}_n(t)\big\| \le \left(\sup_{0\le t\le T}\|X_n(t)\| + \sup_{0\le t\le T}\left\|\int_0^t H_n(s)\,\mathrm{d}G_n(s)\right\| + cT\right)\mathrm{e}^{cT},$$

so

$$\mathbb{P}\left(\sup_{0\le t\le T}\big\|\hat{Y}_n(t)\big\| > a\right) \le \mathbb{P}\left(\left(\sup_{0\le t\le T}\|X_n(t)\| + \sup_{0\le t\le T}\left\|\int_0^t H_n(s)\,\mathrm{d}G_n(s)\right\| + cT\right)\mathrm{e}^{cT} > a\right).$$

The probability on the right-hand side can be made arbitrarily small uniformly in n by taking a large enough, because X_n and $H_n{\cdot}G_n$ converge weakly and are therefore tight. Consequently, $\sup_{0\le t\le T}\big\|\hat{Y}_n(t)\big\|$ is tight and thus $\hat{Y}_n$ is stochastically bounded. The Lipschitz property of Γ_γ and Δ_δ implies that $\Gamma_\gamma(\hat{Y}_n)$ and $\Delta_\delta(\hat{Y}_n)$ are stochastically bounded as well.

It also follows from the previous arguments that the processes $I_{1,n}$ and $I_{2,n}$ are stochastically bounded. We now argue that these processes are C-tight. For $\epsilon > 0$, note that

$$\begin{aligned}\mathbb{P}\left(\sup_{\substack{t_1,t_2\in[0,T]\\ 0<t_2-t_1<\epsilon}}\|I_{1,n}(t_2) - I_{1,n}(t_1)\| > a\right) &\le \mathbb{P}\left(\sup_{0\le t\le T}\big\|\Gamma_\gamma(\hat{Y}_n)(t)\big\| > b\right) \\ &\quad + \mathbb{P}\left(\sup_{\substack{t_1,t_2\in[0,T]\\ 0<t_2-t_1<\epsilon}}\int_{t_1}^{t_2}\big\|\Gamma_\gamma(\hat{Y}_n)(s)(K_n(s)-\pi)\big\|\,\mathrm{d}s > a;\ \sup_{0\le t\le T}\big\|\Gamma_\gamma(\hat{Y}_n)(t)\big\| \le b\right).\end{aligned}$$

Since $\Gamma_\gamma(\hat{Y}_n)$ is stochastically bounded, the first term on the right-hand side can be made arbitrarily small uniformly in n by choosing b large enough. For fixed b, the second term equals zero for each n for small enough ϵ. Consequently, the term on the left-hand side can be made arbitrarily small uniformly in n by choosing ϵ small enough. Together with $I_{1,n}$ being stochastically bounded, this means that $I_{1,n}$ is C-tight (cf. [5, Proposition VI.3.26]). Analogous arguments show that $I_{2,n}$ is C-tight.

The next step is to derive the tightness of $\hat{Y}_n$. The processes X_n and $H_n{\cdot}G_n$ converge weakly to X and $H{\cdot}B$, so X_n and $H_n{\cdot}G_n$ are tight. Since $H_n{\cdot}G_n$ has a continuous limit by Theorem 3.1, we know that $H_n{\cdot}G_n$ is also C-tight. With X_n being tight and $H_n{\cdot}G_n$, $I_{1,n}$, and $I_{2,n}$ being C-tight, it follows from [5, Lemma VI.3.32] that $\hat{Y}_n = X_n + (H_n{\cdot}G_n) + I_{1,n} + I_{2,n}$ is tight with respect to the Skorokhod J_1 topology.

https://doi.org/10.1017/jpr.2020.96 Published online by Cambridge University Press

Knowing that $\hat{Y}_n$ is tight, it remains to verify that Y is the unique limit point of $\hat{Y}_n$. We take an arbitrary weakly converging subsequence of $\hat{Y}_n$ (which we also denote by $\hat{Y}_n$ for simplicity) having limit point $\tilde{Y}$. Now consider the terms on the right-hand side of (3.10). By the J_1 continuity of Γ_γ and Δ_δ, the processes $\Gamma_\gamma(\hat{Y}_n)$ and $\Delta_\delta(\hat{Y}_n)$ converge weakly to $\Gamma_\gamma(\tilde{Y})$ and $\Delta_\delta(\tilde{Y})$, which implies that $I_{2,n}$ converges weakly to $\int_0^t \Delta_\delta(\tilde{Y})(s)\,\mathrm{d}s$, while $I_{1,n}$ converges weakly to η_0 by Lemma A.2. Consequently, the right-hand side of (3.10) converges to the right-hand side of (3.7) with Y replaced by $\tilde{Y}$, which implies that the limit point $\tilde{Y}$ satisfies (3.7). Thus, every limit point of $\hat{Y}_n$ satisfies (3.7). Because the integral equation (3.7) has a unique solution, we conclude that $\hat{Y}_n$ converges weakly to the unique solution Y of (3.7). □

4. Applications

In this section we present two applications of our main results as formulated in Theorems 3.1 and 3.2. The purpose of these examples is to demonstrate that the main results can be applied to a wide range of models. The first example establishes diffusion limits for the Markov-modulated Erlang loss model as well as related models, which are finite-variation processes with a reflecting boundary. The second example establishes a small-noise limit for the Markov-modulated Cox–Ingersoll–Ross (CIR) process, which is not a finite-variation process. Our main results are instrumental in proving both diffusion limits: each example requires weak convergence of a stochastic integral $H_n \cdot G_n$ to $H \cdot G$, as well as weak convergence of the solution of (3.6) to the solution of (3.7). The assumptions of the main results are readily verified in both examples.

We use the following notation and conventions throughout this section. Given a function $\lambda\colon \{1, \dots, d\} \to \mathbb{R}$, we identify it with a d-dimensional column vector that we also denote by λ. Additionally, we define $\lambda^\pi = \lambda^\top \pi$. This quantity may be interpreted as a time-averaged version of λ, because $\int_0^t \lambda(J_n(s))\,\mathrm{d}s$ converges to $\lambda^\pi t$ by Lemma 2.2.

4.1. Markov-modulated many-server queues with finite waiting room

We are interested in a class of many-server queues with a finite or infinite waiting room. An important example is the Erlang loss model, which is the special case in which there is no waiting room. We study such systems under Markov modulation, meaning that the parameters depend on an independently evolving Markov chain (also referred to as the background process). The background process represents an external environment to which the system reacts, for instance by having an extremely large arrival rate if the environment is in some emergency state.

We now describe the model in more detail. We consider a queueing system with $n \in \mathbb{N}$ servers and a waiting room of size $m_n \in \{0\} \cup \mathbb{N} \cup \{\infty\}$. We focus on the scenario in which the parameters of the queueing system are influenced by an independent background process J_n, where J_n is the usual Markov chain with state space $\{1, \dots, d\}$, irreducible generator matrix $n\mathcal{Q}$, and stationary distribution π. While the background process is in state i, jobs arrive at the system according to a Poisson process with rate $\lambda_n(i)$ and servers work at speed $\mu(i)$. Each job has an independent service requirement that has an exponential distribution with unit mean. If a job arrives and there are less than n jobs in service, then it goes into service immediately. If all servers are busy when a new job arrives, then there are two possible cases. In the first case, there are less than m_n jobs waiting and the new job enters the system to wait for service. In the second case, there are already m_n jobs waiting and the new job is rejected from the system. Once a job finishes service, it leaves the system. If there are jobs waiting for service when a

https://doi.org/10.1017/jpr.2020.96 Published online by Cambridge University Press

job finishes service, then one of those jobs is sent to the corresponding server on a first-come, first-served basis.

We denote the number of jobs in the system with n servers at time t by $Q_n(t)$ and represent it as

$$Q_n(t) = Q_n(0) + A\left(\int_0^t \lambda_n(J_n(s))\,\mathrm{d}s \right) - S\left(\int_0^t \mu(J_n(s))(Q_n(s) \wedge n)\,\mathrm{d}s \right) - U_n(t).$$

Here, $Q_n(0)$ is an independent random variable denoting the initial number of jobs in the system, while A and S are independent, unit-rate Poisson processes. The loss process U_n records the number of arrivals if there are m_n jobs waiting for service. It may be interpreted as the downward-reflecting barrier at $n + m_n$ for Q_n, meaning that U_n is the unique, nonnegative, nondecreasing stochastic process such that $Q_n(t) \le n + m_n$ and $\int_0^\infty \mathbf{1}_{\{Q_n(s) < n+m_n\}}\,\mathrm{d}U_n(s) = 0$ (cf. [12]).

We consider this system in the quality-and-efficiency-driven (QED) or Halfin–Whitt regime (cf. [12]), suitably modified to incorporate the Markov modulation (cf. [2, 7, 11]). More specifically, we impose the following condition.

Condition 4.1. *As $n \to \infty$, the initial condition $\sqrt{n}\big(\frac{1}{n}Q_n(0) - 1\big)$ converges in distribution to a random variable $X(0)$. Additionally, $\frac{\lambda_n(i)}{n} \to \bar{\lambda}(i)$ for every $i \in \{1, \ldots, d\}$, and*

$$\sqrt{n} \sum_{i=1}^{d} \left(\mu(i) - \frac{\lambda_n(i)}{n} \right) \pi(i) \to \gamma \mu^\pi, \tag{4.1}$$

where $\gamma \in \mathbb{R}$ is fixed. The waiting room m_n satisfies $\frac{m_n}{\sqrt{n}} \to \kappa$ for some $\kappa \in [0, \infty]$.

This condition reduces to the standard QED regime if there is no modulation and thus $d = 1$. Indeed, in that case, (4.1) states that $\sqrt{n}\big(\mu - \frac{\lambda_n}{n}\big) \to \gamma\mu$ for certain real-valued variables λ_n, μ, and γ, which implies in particular that $\bar{\lambda} = \mu$.

The convergence in (4.1) trivially holds if the system operates in the standard QED regime for any state of the background process, meaning that $\sqrt{n}\big(\mu(i) - \frac{\lambda_n(i)}{n}\big)$ converges to a constant for every $i \in \{1, \ldots c, d\}$. However, (4.1) may also hold if the system does not operate in the standard QED regime for certain states of the background process. This is the most interesting scenario, because the system switches between QED behavior and non-QED behavior.

We now derive the diffusion limit for the scaled and centered queue content process $\hat{Q}_n = \sqrt{n}\big(\frac{1}{n}Q_n - 1\big)$ in the QED regime formulated in Condition 4.1, utilizing the main results. The limit coincides with the usual diffusion limit for nonmodulated many-server queues in the QED regime (cf. [12]) if the system operates in the standard QED regime for any state of the background process. However, if the system does not operate in the standard QED regime for a certain state of the background process, then the limit includes an additional Brownian term capturing the extra variability introduced by the modulation.

Theorem 4.1. *Under Condition 4.1, the process $\hat{Q}_n$ converges weakly to the solution of the stochastic integral equation*

$$\hat{Q}(t) = X(0) - \gamma\mu^\pi t + \sqrt{\bar{\lambda}^\pi + \mu^\pi}\,W(t) + \int_0^t (\bar{\lambda} - \mu)^\top\,\mathrm{d}B(s) - \int_0^t \mu^\pi(\hat{Q} \wedge 0)\,\mathrm{d}s - \hat{U}(t),$$

where $\hat{U}$ is the downward-reflecting barrier at κ for $\hat{Q}$. The processes B and W are independent Brownian motions, where W is a standard Brownian motion and the predictable quadratic variation process of B is given by (2.6).

https://doi.org/10.1017/jpr.2020.96 Published online by Cambridge University Press

Proof. The first step is to rewrite $\hat{Q}_n$ in the form required for an application of Theorem 3.2. Observe that

$$\begin{aligned}\sqrt{n}\left(\frac{1}{n}Q_n(t)-1\right) &= \sqrt{n}\left(\frac{1}{n}Q_n(0)-1\right)\\ &\quad+\sqrt{n}\left(\frac{1}{n}A\left(n\int_0^t\frac{\lambda_n(J_n(s))}{n}\,\mathrm{d}s\right)-\int_0^t\frac{\lambda_n(J_n(s))}{n}\,\mathrm{d}s\right)\\ &\quad+\sqrt{n}\left(\int_0^t\frac{\lambda_n^\top}{n}K_n(s)\,\mathrm{d}s-\int_0^t\frac{\lambda_n^\top}{n}\pi\,\mathrm{d}s\right)+\sqrt{n}\int_0^t\frac{\lambda_n^\top}{n}\pi\,\mathrm{d}s\\ &\quad-\sqrt{n}\left(\frac{1}{n}S\left(n\int_0^t\mu(J_n(s))\left(\frac{1}{n}Q_n(s)\wedge 1\right)\mathrm{d}s\right)\right.\\ &\qquad\left.-\int_0^t\mu(J_n(s))\left(\frac{1}{n}Q_n(s)\wedge 1\right)\mathrm{d}s\right)\\ &\quad-\sqrt{n}\left(\int_0^t\mu(J_n(s))\left(\frac{1}{n}Q_n(s)\wedge 1\right)\mathrm{d}s-\int_0^t\mu(J_n(s))\,\mathrm{d}s\right)\\ &\quad-\sqrt{n}\left(\int_0^t\mu^\top K_n(s)\,\mathrm{d}s-\int_0^t\mu^\top\pi\,\mathrm{d}s\right)\\ &\quad-\sqrt{n}\int_0^t\mu^\top\pi\,\mathrm{d}s-\frac{1}{\sqrt{n}}U_n(t),\end{aligned}$$

so

$$\begin{aligned}\hat{Q}_n(t) &= X_n(t)+\int_0^t\left(\frac{\lambda_n}{n}-\mu\right)^\top\mathrm{d}G_n(s)-\int_0^t(\hat{Q}_n(s)\wedge 0)\mu^\top\,\mathrm{d}\bar{G}_n(s)\\ &\quad-\int_0^t\mu^\pi(\hat{Q}_n(s)\wedge 0)\,\mathrm{d}s-\frac{1}{\sqrt{n}}U_n(t)\end{aligned}$$

with

$$X_n(t)=\hat{Q}_n(0)+\hat{A}_n(\tau_{1,n}(t))-\hat{S}_n(\tau_{2,n}(t))+\sqrt{n}\left(\frac{\lambda_n}{n}-\mu\right)^\top\pi t.$$

Here, we denote $\hat{A}_n(t)=\sqrt{n}\left(\frac{1}{n}A(nt)-t\right)$ and $\hat{S}_n(t)=\sqrt{n}\left(\frac{1}{n}S(nt)-t\right)$. The random time changes $\tau_{1,n}$ and $\tau_{2,n}$ are given by $\tau_{1,n}(t)=\int_0^t\frac{\lambda_n(J_n(s))}{n}\,\mathrm{d}s$ and $\tau_{2,n}(t)=\int_0^t\mu(J_n(s))\left(\frac{1}{n}Q_n(s)\wedge 1\right)\mathrm{d}s$.

Theorem 3.2 is not directly applicable to $\hat{Q}_n$, due to the presence of the process $\frac{1}{\sqrt{n}}U_n$. We get around this issue via the application of a standard method (cf. [12, 14]). The key observation here is that $\frac{1}{\sqrt{n}}U_n$ is the downward-reflecting barrier at $\hat{\kappa}_n=\frac{m_n}{\sqrt{n}}$ for $\hat{Q}_n$, since U_n is the downward-reflecting barrier at $n+m_n$ for Q_n.

Define the functions Ψ_δ and Φ_δ mapping $\mathbb{D}([0,\infty),\mathbb{R})$ into $\mathbb{D}([0,\infty),\mathbb{R})$ via $\Psi_\delta(x)(t)=\sup_{0\le s\le t}((x(s)-\delta)\vee 0)$ and $\Phi_\delta(x)(t)=x(t)-\Psi_\delta(x)(t)$. Both Ψ_δ and Φ_δ are Lipschitz

https://doi.org/10.1017/jpr.2020.96 Published online by Cambridge University Press

continuous in the supremum norm and in the J_1 metric for a fixed boundary level δ (cf. [15, Theorem 13.5.1]). A minor variation on the arguments in [14] establishes that

$$Y_n(t) = X_n(t) + \int_0^t \left(\frac{\lambda_n}{n} - \mu\right)^\top \mathrm{d}G_n(s) - \int_0^t (\Phi_{\hat{\kappa}_n}(Y_n)(s) \wedge 0)\mu^\top \mathrm{d}\bar{G}_n(s) - \int_0^t \mu^\pi(\Phi_{\hat{\kappa}_n}(Y_n)(s) \wedge 0)\,\mathrm{d}s$$

is a well-defined stochastic process and that $\Phi_{\hat{\kappa}_n}(Y_n) = \hat{Q}_n$. If Y_n converges weakly to some limiting process Y, then $\Phi_{\hat{\kappa}_n}(Y_n)$ converges weakly to $\Phi_\kappa(Y)$, since $\hat{\kappa}_n \to \kappa$ and the map $\Phi_\delta(x)$ is continuous both in δ and in x. Consequently, to prove weak convergence of $\hat{Q}_n$, it suffices to prove weak convergence of Y_n.

The process Y_n has exactly the form required for an application of Theorem 3.2. Clearly, Y_n satisfies (3.6) with $H_n = \left(\frac{\lambda_n}{n} - \mu\right)^\top$, $\Gamma_{\gamma_n} = \Gamma_{\hat{\kappa}_n} = (\Phi_{\hat{\kappa}_n} \wedge 0)\mu^\top$, and $\Delta_{\delta_n} = \Delta_{\hat{\kappa}_n} = \mu^\pi(\Phi_{\hat{\kappa}_n} \wedge 0)$. The continuity properties of Γ_{γ_n} and Δ_{δ_n} follow from [15, Chapter 13], so it remains to verify weak convergence of (H_n, G_n, X_n). Since H_n converges to the constant $\bar{\lambda} - \mu$ by Condition 4.1, we only have to show weak convergence of (G_n, X_n).

We prove the required weak convergence of (G_n, X_n) as follows. The intial condition $\hat{Q}_n(0)$ is independent and converges to some random variable $X(0)$, while $\sqrt{n}\left(\frac{\lambda_n}{n} - \mu\right)^\top \pi$ converges to a constant. Therefore, weak convergence of (G_n, X_n) follows from weak converges of $(\hat{A}_n \circ \tau_{1,n}, \hat{S}_n \circ \tau_{2,n}, G_n)$, which in turn follows from weak convergence of $(\hat{A}_n, \tau_{1,n}, \hat{S}_n, \tau_{2,n}, G_n)$ by the continuous mapping theorem (CMT) if $\hat{A}_n$ and $\hat{S}_n$ converge to continuous processes (cf. [15, Theorem 13.2.2]).

The processes $\hat{A}_n$ and $\hat{S}_n$ are independent, scaled, and centered standard Poisson processes, so they converge jointly to two independent, standard Brownian motions W_1 and W_2. Additionally, G_n converges to a Brownian motion B that is independent of W_1 and W_2. Therefore, the required convergence of $(\hat{A}_n, \tau_{1,n}, \hat{S}_n, \tau_{2,n}, G_n)$ follows if $\tau_{1,n}$ and $\tau_{2,n}$ both converge ucp to a deterministic limit.

To prove this convergence of $\tau_{1,n}$ and $\tau_{2,n}$, we apply Theorem 3.2 to the process $\bar{Y}_n = \frac{1}{\sqrt{n}} Y_n$. Writing $\bar{\kappa}_n = \frac{1}{\sqrt{n}}\hat{\kappa}_n$, we get

$$\bar{Y}_n(t) = \frac{1}{\sqrt{n}}\hat{Q}_n(0) + \frac{1}{\sqrt{n}}\hat{A}_n(\tau_{1,n}(t)) - \frac{1}{\sqrt{n}}\hat{S}_n(\tau_{2,n}(t)) + \left(\frac{\lambda_n}{n} - \mu\right)^\top \pi t + \int_0^t \frac{1}{\sqrt{n}}\left(\frac{\lambda_n}{n} - \mu\right)^\top \mathrm{d}G_n(s) - \int_0^t (\Phi_{\bar{\kappa}_n}(\bar{Y}_n)(s) \wedge 0)\mu^\top \mathrm{d}\bar{G}_n(s) - \int_0^t \mu^\pi(\Phi_{\bar{\kappa}_n}(\bar{Y}_n) \wedge 0)\,\mathrm{d}s.$$

The processes $\frac{1}{\sqrt{n}}\hat{A}_n$ and $\frac{1}{\sqrt{n}}\hat{S}_n$ both converge ucp to η_0. With $\tau_{1,n}$ and $\tau_{2,n}$ being bounded on compact intervals, it follows that $\frac{1}{\sqrt{n}}\hat{A}_n \circ \tau_{1,n}$ and $\frac{1}{\sqrt{n}}\hat{S}_n \circ \tau_{2,n}$ converge ucp to η_0, too. Condition 4.1 implies that $\frac{1}{\sqrt{n}}\hat{Q}_n(0)$ converges to 0 in probability, and that both $\left(\frac{\lambda_n}{n} - \mu\right)^\top \pi$ and $\frac{1}{\sqrt{n}}\left(\frac{\lambda_n}{n} - \mu\right)$ converge to 0. Also, $\bar{\kappa}_n$ converges to 0. Consequently, Theorem 3.2 guarantees weak convergence of $\bar{Y}_n$ to the unique process $\bar{Y}$ satisfying $\bar{Y}(t) = -\int_0^t \mu^\pi(\Phi_0(\bar{Y}) \wedge 0)\,\mathrm{d}s$. The zero process is the unique solution of this equation, so $\bar{Y} = \eta_0$ and $\bar{Y}_n$ converges ucp to η_0.

https://doi.org/10.1017/jpr.2020.96 Published online by Cambridge University Press

Recall that we aim to prove that $\tau_{1,n}$ and $\tau_{2,n}$ both converge ucp to a deterministic limit in order to obtain weak convergence of $(\hat{A}_n, \tau_{1,n}, \hat{S}_n, \tau_{2,n}, G_n)$. By Condition 4.1 and Lemma 2.2, the process $\tau_{1,n}$ converges ucp to the deterministic function τ_1^π with $\tau_1^\pi(t) = \bar{\lambda}^\pi t$. It follows from Lemma 2.2 and $\bar{Y}_n$ converging ucp to η_0 that the process $\tau_{2,n}$ converges ucp to the deterministic function τ_2^π with $\tau_2^\pi(t) = \mu^\pi t$.

We conclude that $(\hat{A}_n, \tau_{1,n}, \hat{S}_n, \tau_{2,n}, G_n)$ converges weakly to $(W_1, \tau_1^\pi, W_2, \tau_2^\pi, B)$. This implies weak convergence of $(\hat{A}_n \circ \tau_{1,n}, \hat{S}_n \circ \tau_{2,n}, G_n)$ to $(W_1 \circ \tau_1^\pi, W_2 \circ \tau_2^\pi, B)$ by the CMT, where we note that $W_1 \circ \tau_1^\pi$ and $W_2 \circ \tau_2^\pi$ have the same law as $\sqrt{\bar{\lambda}^\pi} W_1$ and $\sqrt{\mu^\pi} W_2$. As argued before, this proves weak convergence of (G_n, X_n) to (B, X), where the process X is given by $X(t) = X(0) + \sqrt{\bar{\lambda}^\pi} W_1 - \sqrt{\mu^\pi} W_2 - \gamma \mu^\pi t$. Then, Theorem 3.2 implies that Y_n converges weakly to the process Y satisfying

$$Y(t) = X(t) + \int_0^t (\bar{\lambda} - \mu)^\top \, \mathrm{d}B(s) - \int_0^t \mu^\pi (\Phi_\kappa(Y) \wedge 0) \, \mathrm{d}s.$$

Here, the process B is the Brownian motion given in the theorem and $X(0)$, W_1, W_2, and B are independent.

Deriving the weak convergence of the scaled and centered queue content process $\hat{Q}_n = \Phi_{\kappa_n}(Y_n)$ is now a simple matter of applying the CMT. It follows that $\hat{Q}_n$ converges weakly to the process $\hat{Q} = \Phi_\kappa(Y)$, so $\hat{Q}$ satisfies the stochastic integral equation

$$\hat{Q}(t) = X(t) + \int_0^t (\bar{\lambda} - \mu)^\top \, \mathrm{d}B(s) - \int_0^t \mu^\pi (\hat{Q} \wedge 0) \, \mathrm{d}s - \hat{U}(t),$$

with $\hat{U}$ being the downward-reflecting barrier at κ for $\hat{Q}$. □

4.2. The Markov-modulated CIR process

The previous example concerns the Markov-modulated Erlang loss model, which is a finite-variation stochastic process with reflection. In the next example, we focus on a process that does not have sample paths of finite variation, namely the CIR process under Markov modulation.

In interest rate models, the CIR process is often used to model the short rate. The CIR process R is defined via the stochastic integral equation

$$R(t) = x + \lambda t - \mu \int_0^t R(s) \, \mathrm{d}s + \sigma \int_0^t \sqrt{R(s)} \, \mathrm{d}W(s),$$

where λ and μ are positive constants and σ is some real number. The process W is a standard Brownian motion. This conventional CIR process with fixed parameters may be enhanced with a modulating process that makes the parameters change stochastically over time. In a financial context, this is often referred to as regime switching. An example is switching from a bull market (good economic conditions) to a bear market (bad economic conditions), which may influence the volatility of the short rate, for instance. Another example may be an influential

https://doi.org/10.1017/jpr.2020.96 Published online by Cambridge University Press

person tweeting messages at random: parameters change if a tweet is posted, but go back to their original values once the tweet loses its effect.

In this example, we consider the following Markov-modulated CIR process with small noise and study its scaling limit. Let λ, μ, and σ be real-valued functions on $\{1, \dots, d\}$, with λ and μ taking positive values. Let $x \geq 0$ be the initial condition and fix $\alpha > 0$. We are interested in the process R_n defined via

$$R_n(t) = x + \int_0^t (\lambda(J_n(s)) - \mu(J_n(s))R_n(s))\,\mathrm{d}s + \frac{1}{\sqrt{n}} \int_0^t \sigma(J_n(s))\sqrt{R_n(s)}\,\mathrm{d}W(s),$$

where J_n is the usual Markov chain with state space $\{1, \dots, d\}$, irreducible generator matrix $n^\alpha \mathcal{Q}$, and stationary distribution π. A well-known property of the nonmodulated CIR process is that it is nonnegative if it starts from a nonnegative position (cf. [4]). Clearly, this property carries over to its Markov-modulated version, so R_n is nonnegative.

We use the parameter α to reflect that the background process may operate on a different time scale than the CIR dynamics. For instance, switches between a bull market and a bear market occur on a much slower time scale than fluctuations in the interest rate, which can be modeled by taking $\alpha < 1$. The value of α has a significant influence on the behavior of R_n. Roughly speaking, the fluctuations of R_n are dominated by the dynamics on the slowest time scale. If $\alpha < 1$, then the background process operates on the slowest time scale and the fluctuations of R_n are dominated by the fluctuations of J_n. If $\alpha > 1$, then the CIR dynamics operates on the slowest time scale and dominates the fluctuations of R_n, with the background process averaging out. The boundary case $\alpha = 1$ incorporates the effects of both the CIR dynamics and the background process, leading to the most complicated behavior.

As mentioned, our aim is to find a scaling limit for the Markov-modulated CIR process R_n. We consider the case in which n becomes large, so the noise term becomes small and the background process switches states relatively rapidly. The next theorem presents the corresponding diffusion limit for R_n. In its proof, we first show that R_n converges ucp to the unique solution r of the integral equation

$$r(t) = x + \lambda^\pi t - \mu^\pi \int_0^t r(s)\,\mathrm{d}s.$$

We then proceed by studying the fluctuations of R_n around this limit and prove via an application of the main results that $\hat{R}_n = n^\beta(R_n - r)$ converges weakly to a diffusion process, where $\beta = \min\{1/2, \alpha/2\}$.

We apply the scaling factor n^β instead of the usual $\sqrt{n}$ to account for the influence of the background process. As indicated earlier, the time scale of the background process is relatively slow compared to the time scale of the CIR dynamics if $\alpha < 1$, in which case the fluctuations of R_n are dominated by the fluctuations of the background process. Because the fluctuations of the background process are of order $n^{-\alpha/2}$, we have to use the scaling n^β to obtain a nondegenerate limit.

The limiting diffusion of $\hat{R}_n$ depends explicitly on properties of the background process and on the value of α. If $\alpha < 1$, then the small-noise term disappears and the diffusion part of the limiting process is completely determined by the fluctuations of the background process. If $\alpha > 1$, then the background process averages out and the diffusion part arises from the small-noise term. As explained earlier, the boundary case $\alpha = 1$ incorporates both effects.

https://doi.org/10.1017/jpr.2020.96 Published online by Cambridge University Press

Theorem 4.2. *The process $\hat{R}_n$ converges weakly to the Ornstein–Uhlenbeck process Y satisfying*

$$\begin{aligned} Y(t) &= \mathbf{1}_{\{\alpha\le 1\}} \int_0^t (\lambda - r(s)\mu)^\top \,\mathrm{d}B(s) \\ &\quad + \mathbf{1}_{\{\alpha\ge 1\}} \int_0^t \sqrt{\sigma^\top \operatorname{diag}(\pi)\sigma}\sqrt{r(s)}\,\mathrm{d}W(s) - \mu^\pi \int_0^t Y(s)\,\mathrm{d}s, \end{aligned} \tag{4.2}$$

The processes B and W are independent Brownian motions, where W is a standard Brownian motion and the predictable quadratic variation process of B is given by (2.6).

Proof. We start by showing that R_n converges ucp to r. Given R_n, it is convenient to define

$$\begin{aligned} R_n^*(t) &= x + \int_0^t (\lambda(J_n(s)) - \mu(J_n(s))R_n(s))\,\mathrm{d}s, \\ R_n^\dagger(t) &= \frac{1}{\sqrt{n}} \int_0^t \sigma(J_n(s))\sqrt{R_n(s)}\,\mathrm{d}W(s), \end{aligned}$$

so R_n may be represented as $R_n = R_n^* + R_n^\dagger$. We also define $\bar{R}_n = R_n - r$. Then, for fixed $T > 0$, some straightforward calculations lead to the inequality

$$\mathbb{E}\sup_{0\le s\le t} |\bar{R}_n(s)| \le \mathbb{E}\bar{I}_n^{(1)}(T) + \mathbb{E}\bar{I}_n^{(2)}(T) + \mathbb{E}\sup_{0\le s\le t} |R_n^\dagger(s)| + c\int_0^t \mathbb{E}\sup_{0\le u\le s} |\bar{R}_n(u)|\,\mathrm{d}s \tag{4.3}$$

for each $t \in [0, T]$, where

$$\begin{aligned} \bar{I}_n^{(1)}(t) &= \sup_{0\le s\le t} \left| \int_0^s (\lambda(J_n(u)) - \lambda^\pi)\,\mathrm{d}u \right|, \\ \bar{I}_n^{(2)}(t) &= \sup_{0\le s\le t} \left| \int_0^s r(u)(\mu(J_n(u)) - \mu^\pi)\,\mathrm{d}u \right|. \end{aligned}$$

The expectations $\mathbb{E}\bar{I}_n^{(1)}(T)$ and $\mathbb{E}\bar{I}_n^{(2)}(T)$ converge to 0, due to Theorem 3.1 and the fact that both random variables are bounded. We use here that r is of finite variation.

We would like to get a bound on $\mathbb{E}\sup_{0\le s\le t} |R_n^\dagger(s)|$, so that we can apply Gronwall's lemma (cf. [8, pp. 287–288]) to the inequality in (4.3). To this end, we rely on the Burkholder–Davis–Gundy inequalities (cf. [8, Theorem 3.3.28]) as well as Jensen's inequality to obtain that

$$\begin{aligned} \mathbb{E}\sup_{0\le s\le t} |R_n^\dagger(s)| &\le \frac{c}{\sqrt{n}} \sqrt{\int_0^t r(s) + \mathbb{E}\int_0^t |\bar{R}_n(s)|\,\mathrm{d}s} \\ &\le \frac{c}{\sqrt{n}} \left(1 + \int_0^t r(s) + \int_0^t \mathbb{E}\sup_{0\le u\le s} |\bar{R}_n(u)|\,\mathrm{d}s\right). \end{aligned}$$

Plugging this in into (4.3) and applying Gronwall's lemma, we conclude that the expectation $\mathbb{E}\sup_{0\le s\le t} |\bar{R}_n(s)| = \mathbb{E}\sup_{0\le s\le t} |R_n(s) - r(s)|$ converges to 0 as $n \to \infty$. This implies in particular that R_n converges ucp to r.

The next step is to study the fluctuations of the Markov-modulated CIR process R_n around its limit r. More precisely, we would like to characterize the asymptotic behavior of $\hat{R}_n =$

https://doi.org/10.1017/jpr.2020.96 Published online by Cambridge University Press

$n^\beta(R_n - r)$. Recall that the state occupation measure G_n related to the background process J_n converges weakly to a Brownian motion B with predictable quadratic variation process $\langle B\rangle$ given by (2.6). Because the background process and the Brownian motion W are independent, we may assume that B and W are independent.

To study the fluctuations of R_n around r, we consider the process $\hat{R}_n = n^\beta(R_n - r)$, which satisfies

$$\hat{R}_n(t) = n^{\beta-1/2}\int_0^t \sigma(J_n(s))\sqrt{R_n(s)}\,\mathrm{d}W(s) + n^{\beta-\alpha/2}\int_0^t \big(\lambda - r(s)\mu\big)^\top \mathrm{d}G_n(s) - \int_0^t \hat{R}_n(s)\mu^\top \mathrm{d}\bar{G}_n(s) - \mu^\pi \int_0^t \hat{R}_n(s)\,\mathrm{d}s.$$

To be able to apply Theorem 3.2 to $\hat{R}_n$, it suffices to show weak convergence of (H_n, G_n, X_n), where

$$H_n(t) = n^{\beta-\alpha/2}(\lambda - r(t)\mu)^\top, \qquad X_n(t) = n^{\beta-1/2}\int_0^t \sigma(J_n(s))\sqrt{R_n(s)}\,\mathrm{d}W(s).$$

We first observe that H_n is a deterministic process of finite variation and converges uniformly on compacts to the process H given by $H(t) = \mathbf{1}_{\{\alpha\le 1\}}(\lambda - r(t)\mu)^\top$. Therefore, we only have to prove weak convergence of (X_n, G_n), which follows from a straightforward application of Theorem 3.1 combined with the MCLT, as we demonstrate next.

We know from (2.5) that G_n is equal to the locally square-integrable martingale $n^{-1/2}F^\top M_n$ plus a term that converges uniformly to η_0, so it suffices to show that $(X_n, n^{-1/2}F^\top M_n)$ converges weakly. This is a local martingale whose maximum jump size converges to 0, so weak convergence of $(X_n, n^{-1/2}F^\top M_n)$ follows from its predictable quadratic covariation process converging ucp to a deterministic function (cf. [16]). Since X_n is a stochastic integral with respect to W and the processes W and $n^{-1/2}F^\top M_n$ are independent, we know that $\langle X_n, n^{-1/2}F^\top M_n\rangle = \eta_0$, so we only have to show convergence of $\langle X_n\rangle$ and $\langle n^{-1/2}F^\top M_n\rangle$. In Lemma 2.2 we established ucp convergence of $\langle n^{-1/2}F^\top M_n\rangle$ to $\langle B\rangle$. It remains to prove convergence of $\langle X_n\rangle$.

Note that X_n is a continuous local martingale with $\langle X_n\rangle$ given by

$$\begin{aligned}\langle X_n\rangle(t) &= n^{2\beta-1}\int_0^t \sigma^\top \mathrm{diag}(K_n(s))\sigma R_n(s)\,\mathrm{d}s \\ &= n^{2\beta-1}\int_0^t \sigma^\top \mathrm{diag}(K_n(s))\sigma(R_n(s) - r(s))\,\mathrm{d}s + n^{2\beta-1}\int_0^t \sigma^\top \mathrm{diag}(K_n(s))\sigma r(s)\,\mathrm{d}s.\end{aligned}$$

The penultimate integral above converges ucp to η_0, due to R_n converging ucp to r. The last integral above converges ucp to

$$\mathbf{1}_{\{\alpha\ge 1\}}\int_0^t \sigma^\top \mathrm{diag}(\pi)\sigma r(s)\,\mathrm{d}s, \tag{4.4}$$

due to Theorem 3.1 and $2\beta - 1$ being equal to $\min\{0, (\alpha-1)/2\}$. We use here that r is of finite variation. Consequently, $\langle X_n\rangle$ converges ucp to the process in (4.4), too. The MCLT then implies that X_n converges weakly to a Brownian motion whose predictable quadratic variation process is given by (4.4).

The previous arguments establish weak convergence of (X_n, G_n). Now applying Theorem 3.2 to $\hat{R}_n$, we conclude that $\hat{R}_n$ converges weakly to the process Y given by (4.2). □

https://doi.org/10.1017/jpr.2020.96 Published online by Cambridge University Press

5. Summary and concluding remarks

We investigated weak convergence of stochastic integrals with respect to the state occupation measure of a Markov chain. The motivation behind this was that standard results do not apply to this elementary yet important case. Indeed, the state occupation measure is not a martingale nor has the P-UT property. One of the underlying problems turned out to be that the total variation of the integrand may grow too quickly. Relying on this insight, we formulated a condition for the total variation of the integrand. In the first main result, we proved that stochastic integrals with respect to the state occupation measure converge weakly under this condition. We extended this to a class of stochastic integral equations in the second main result.

We demonstrated the relevance of these results by applying them to two examples. The first concerned a finite-variation process with a reflecting boundary, whereas the second concerned a diffusion process whose sample paths were not of finite variation. Clearly, the main results can also be used to investigate a large class of related models involving Markov modulation, such as single-server queues, networks of many-server queues, and multidimensional diffusion processes.

There are several other possible directions for further research. An interesting question is whether it is possible to derive similar results in the Skorokhod M_1 topology, which is weaker than the Skorokhod J_1 topology. The results in this paper require, for instance, that the arrival process for the Erlang loss model converges weakly in the J_1 topology, but convergence in the M_1 topology is more natural for certain applications (cf. [13]). This appears to be a little-explored area and may necessitate a different approach. Finally, we remark that the main results are only valid for finite-dimensional processes. Since many models feature infinite-dimensional processes, it would also be interesting to see whether the main results can be extended to that setting.

Appendix A. Auxiliary results

Weak convergence of stochastic integrals $H_n \cdot G_n$ is the central problem of this paper. As indicated earlier, the so-called P-UT property is often the key to establishing such convergence results. In this appendix we explore the P-UT property and its relation to the problem at hand. First, we give a formal definition of the P-UT property. Second, we sketch an example showing that G_n does not have the P-UT property. This example also indicates why G_n does not have the P-UT property. Third, we derive a bound for a class of Lebesgue–Stieltjes integrals that are closely connected to integrals with respect to G_n. This bound provides another perspective on the reason why G_n does not have the P-UT property. Moreover, it suggests what conditions we have to impose on the integrand H_n to guarantee weak convergence of $H_n \cdot G_n$. We end this appendix with a continuity result for state occupation measures.

A.1. The state occupation measure and the P-UT property

Let X_n be a sequence of one-dimensional semimartingales relative to a filtration $\mathbb{F}$. We say that X_n has the P-UT property, or simply that X_n is P-UT, if the collection $\{|(H_n \cdot X_n)(t)| : n \in \mathbb{N}, H_n \text{ is } \mathbb{F}\text{-predictable with } |H_n| \leq b\}$ is tight for all $t > 0$ and $b > 0$. We say that a sequence X_n of d-dimensional semimartingales is P-UT if each of its components is P-UT (cf. [5, p. 377] and [5, p. 381]). The acronym P-UT stands for 'predictably uniformly tight'; see [5, 6, 9] for more details.

https://doi.org/10.1017/jpr.2020.96 Published online by Cambridge University Press

The main reason for introducing the P-UT property can be found in [5, Theorem VI.6.22]. Loosely speaking, this result states that $H_n \cdot X_n \Rightarrow H \cdot X$ if $(H_n, X_n) \Rightarrow (H, X)$ and X_n is P-UT. This is exactly the type of result we are interested in, with the semimartingale X_n being the state occupation measure G_n. However, G_n is not P-UT, so this result is not applicable if we integrate against G_n.

The following arguments demonstrate that the state occupation measure G_n is not P-UT. Recall the connection between G_n and $n^{-\alpha/2}M_n$ established in (2.5). The process $n^{-\alpha/2}M_n$ is a locally square-integrable martingale with bounded jumps and converges weakly to a Brownian motion by Lemma 2.2, so $n^{-\alpha/2}M_n$ is P-UT according to [5, Corollary VI.6.29]. In turn, this implies that G_n is P-UT if and only if $n^{-\alpha/2}K_n$ is P-UT (cf. [5, p. 377]).

Knowing this, we aim to show that $n^{-\alpha/2}K_n$ (and thus G_n) is not P-UT by finding an integral $H_n^\top \cdot n^{-\alpha/2}K_n$ that grows without bound as $n \to \infty$, even though H_n is a bounded and predictable process as in the definition of the P-UT property. Define H_n as the left-continuous version of $\mathbf{1} - K_n$, so $H_n(0) = \mathbf{1} - K_n(0)$ and $H_n(t) = \mathbf{1} - K_n(t-)$ for $t > 0$. Because H_n is bounded and predictable, the family $\{(H_n^\top \cdot n^{-\alpha/2}K_n)(t) \mid n \in \mathbb{N}\}$ must be tight for each $t > 0$ if $n^{-\alpha/2}K_n$ is P-UT. However, the random variable $(H_n^\top \cdot K_n)(1)$ counts the number of jumps that J_n makes in the time interval $[0, 1]$. Because J_n is a Markov chain with generator matrix $n^\alpha \mathcal{Q}$, its number of jumps in $[0, 1]$ is of order n^α. Consequently, $(H_n^\top \cdot n^{-\alpha/2}K_n)(1)$ is of order $n^{\alpha/2}$, so it does not converge and $\{(H_n^\top \cdot n^{-\alpha/2}K_n)(1) \mid n \in \mathbb{N}\}$ is not tight. We conclude that G_n cannot be P-UT.

The underlying problem here is that a Lebesgue–Stieltjes integral may not converge if the total variation of the integrand or the integrator grows without bound. Because the total variation of the integrator G_n is a given, this suggests that we have to put restrictions on the total variation of the integrand H_n if we want the stochastic integral $H_n \cdot G_n$ to converge.

A.2. A bound for Lebesgue–Stieltjes integrals

Here, we derive an upper bound for a class of Lebesgue–Stieltjes integrals. This result is closely related to the previous insight, which connects G_n not being P-UT to the behavior of Lebesgue–Stieltjes integrals. The upper bound is important in two ways. First, it indicates what type of restrictions we should impose on the supremum and on the total variation of an integrand H_n if we would like the stochastic integral $H_n \cdot G_n$ to converge. Second, we use this result to prove Theorem 3.1, which concerns weak convergence of stochastic integrals with respect to G_n.

Lemma A.1. *Let $y\colon [0, \infty) \to \mathbb{R}$ be a function of bounded variation and let $x\colon [0, \infty) \to \{0, 1\}$ be a right-continuous function. Then the Lebesgue–Stieltjes integral $y \cdot x$ satisfies*

$$\sup_{0 \le t \le T} \left| \int_0^t y(s)\, \mathrm{d}x(s) \right| \le v_y(T) + \sup_{0 \le t \le T} |y(t)|$$

for every fixed $T > 0$, where v_y denotes the total variation function of y.

Proof. To prove the lemma, it suffices to show that

$$\left| \int_0^t y(s)\, \mathrm{d}x(s) \right| \le v_y(t) + \sup_{0 \le s \le t} |y(s)| \tag{A.1}$$

for every fixed $t \ge 0$. Note that x has alternating jumps of size $+1$ and -1, and is constant between jumps. Thus, if x has at most one jump in $[0, t]$, then (A.1) holds trivially.

https://doi.org/10.1017/jpr.2020.96 Published online by Cambridge University Press

Suppose that x has exactly $2m$ jumps in $[0, t]$ (where $m \in \mathbb{N}$) and denote the jump times by $0 < s_1 < \ldots < s_{2m} \leq t$. If the first jump equals $+1$, then $\int_0^t y(s)\,\mathrm{d}x(s) = \sum_{k=0}^{m-1} (y(s_{2k+1}) - y(s_{2k+2}))$, so

$$\left| \int_0^t y(s)\,\mathrm{d}x(s) \right| \leq \sum_{k=0}^{m-1} |y(s_{2k+2}) - y(s_{2k+1})| \leq v_y(t). \tag{A.2}$$

If the first jump equals -1, then $\int_0^t y(s)\,\mathrm{d}x(s) = \sum_{k=0}^{m-1} (-y(s_{2k+1}) + y(s_{2k+2}))$, so (A.2) holds in this case, too. Hence, (A.1) holds if x has an even number of jumps.

Suppose that x has exactly $2m+1$ jumps in $[0, t]$ (where $m \in \mathbb{N}$) and denote the jump times by $0 < s_1 < \ldots < s_{2m} < s_{2m+1} \leq t$. Taking $\delta = (s_{2m+1} - s_{2m})/2$, we get $\int_0^t y(s)\,\mathrm{d}x(s) = \int_0^{s_{2m}+\delta} y(s)\,\mathrm{d}x(s) + \int_{s_{2m}+\delta}^t y(s)\,\mathrm{d}x(s)$ and

$$\left| \int_0^{s_{2m}+\delta} y(s)\,\mathrm{d}x(s) + \int_{s_{2m}+\delta}^t y(s)\,\mathrm{d}x(s) \right| \leq v_y(s_{2m}) + |y(s_{2m+1})| \leq v_y(t) + \sup_{0\leq s\leq t} |y(s)|.$$

Consequently, (A.1) also holds if x has an odd number of jumps. □

A.3. A convergence result for deterministic state occupation measures

Here, we state an elementary convergence result for deterministic state occupation measures and outline its proof, which is closely related to well-known results for absolutely continuous functions. We denote by $\mathcal{V} \subset \mathbb{D}([0, \infty); \mathbb{R})$ the collection of all absolutely continuous functions f that admit a representation $f(t) = \int_0^t g(s)\,\mathrm{d}s$, where $g \in \mathcal{W}$. The collection $\mathcal{W}$ comprises all functions $g \in \mathbb{D}([0, \infty); \mathbb{R})$ with $g(t) \in [-1, 1]$. To each $f \in \mathcal{V}$ corresponds a unique $g \in \mathcal{W}$ such that $f(t) = \int_0^t g(s)\,\mathrm{d}s$, and we denote this unique function g by $\dot{f}$. In the special case that $\dot{f}$ takes values in $\{0, 1\}$, we may interpret $\dot{f}$ as a state indicator function and f as the corresponding state occupation measure.

Lemma A.2. *Let $f, f_1, f_2, \ldots \in \mathcal{V}$ and $h, h_1, h_2 \in \mathbb{D}([0, \infty); \mathbb{R})$. If $f_n \to f$ in $\mathcal{V}$ and $h_n \to h$ in $\mathbb{D}([0, \infty); \mathbb{R})$ in the Skorokhod J_1 topology as $n \to \infty$, then*

$$\sup_{0\leq t\leq T} \left| \int_0^t \dot{f}_n(s) h_n(s)\,\mathrm{d}s - \int_0^t \dot{f}(s) h(s)\,\mathrm{d}s \right| \to 0 \tag{A.3}$$

as $n \to \infty$ for each $T > 0$.

Proof. Since the functions $f, f_1, f_2, \ldots$ are continuous, convergence of $f_n \to f$ in the Skorokhod J_1 topology is equivalent to $f_n \to f$ under the supremum norm. In turn, this implies that (A.2) holds if $h_n = h$ and h is a step function.

If $h_n = h$ but h is not a step function, then there exist step functions $\tilde{h}_m$ that converge uniformly to h for $m \to \infty$. Decomposing $\dot{f}_n h - \dot{f} h = (\dot{f}_n - \dot{f})(h - \tilde{h}_m) + (\dot{f}_n - \dot{f})\tilde{h}_m$ and using that $(\dot{f}_n - \dot{f})(h - \tilde{h}_m)$ converges uniformly to 0 as $m \to \infty$, it follows that (A.3) also holds if $h_n = h$ and h is not necessarily a step function.

Finally, in the general case that $h_n \to h$ in $\mathbb{D}([0, \infty); \mathbb{R})$, we note that $\dot{f}_n h_n - \dot{f} h = \dot{f}_n(h_n - h) + (\dot{f}_n h - \dot{f} h)$. Since $\dot{f}_n(h_n - h)$ converges pointwise to 0 at all continuity points of h and is bounded uniformly in n on compact sets, we may use the previous considerations to conclude that (A.3) is also valid in this case. □

https://doi.org/10.1017/jpr.2020.96 Published online by Cambridge University Press

This lemma is useful if we have a sequence of stochastic processes $Y_n \Rightarrow Y$ in $\mathbb{D}([0, \infty); \mathbb{R})$ and a sequence of state occupation measures $X_n \Rightarrow X$ in $\mathcal{V}$, where X is a deterministic limit. In this case, the CMT and the lemma together imply that $Y_n \cdot X_n \Rightarrow Y \cdot X$.

Acknowledgements

Part of this research was performed while the author was affiliated with the Centre for Applications in Natural Resource Mathematics (CARM), School of Mathematics and Physics, The University of Queensland, Australia. The author was funded by Australian Research Council (ARC) Discovery Project DP180101602 and by the NWO Gravitation Programme NETWORKS, grant number 024.002.003.

References

[1] Aggoun, L. and Elliott, R. (2004). *Measure Theory and Filtering*. Cambridge University Press.
[2] Arapostathis, A., Das, A., Pang, G. and Zheng, Y. (2019). Optimal control of Markov-modulated multiclass many-server queues. *Stoch. Sys.* **9**, 155–181.
[3] Coolen-Schrijner, P. and van Doorn, E. A. (2002). The deviation matrix of a continuous-time Markov chain. *Prob. Eng. Inf. Sci.* **16**, 351–366.
[4] Cox, J. C., Ingersoll, Jr., J. E. and Ross, S. A. (1985). A theory of the term structure of interest rates. *Econometrica* **53**, 385–407.
[5] Jacod, J. and Shiryaev, A. N. (2003). *Limit Theorems for Stochastic Processes*, 2nd ed. Springer, Berlin.
[6] Jakubowski, A., Mémin, J. and Pages, G. (1989). Convergence en loi des suites d'intégrales stochastiques sur l'espace $\mathbb{D}^1$ de Skorokhod. *Prob. Theory Relat. Fields* **81**, 111–137.
[7] Jansen, H. M. (2018). Scaling limits for modulated infinite-server queues and related stochastic processes. PhD thesis, University of Amsterdam and Ghent University.
[8] Karatzas, I. and Shreve, S. E. (1998). *Brownian Motion and Stochastic Calculus*, 2nd ed. Springer, New York.
[9] Kurtz, T. G. and Protter, P. (1991). Weak limit theorems for stochastic integrals and stochastic differential equations. *Ann. Prob.* **19**, 1035–1070.
[10] Kurtz, T. G. and Protter, P. (1991). Wong–Zakai corrections, random evolutions, and simulation schemes for SDE's. In *Stochastic Analysis: Liber Amicorum for Moshe Zakai*. ed. E. Meyer-Wolf, E. Merzbach, and A. Shwartz. Academic Press, Boston, pp. 331–346.
[11] Mandjes, M., Taylor, P. G. and De Turck, K. (2017). The Markov-modulated Erlang loss system. *Performance Evaluation* **116**, 53–69.
[12] Pang, G., Talreja, R. and Whitt, W. (2007). Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Prob. Surv.* **4**, 193–267.
[13] Pang, G. and Whitt, W. (2010). Continuity of a queueing integral representation in the M_1 topology. *Ann. Appl. Prob.* **20**, 214–237.
[14] Reed, J., Ward, A. R. and Zhan, D. (2013). On the generalized drift Skorokhod problem in one dimension. *J. Appl. Prob.* **50**, 16–28.
[15] Whitt, W. (2002). *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer, New York.
[16] Whitt, W. (2007). Proofs of the martingale FCLT. *Prob. Surv.* **4**, 268–302.