Spike Modeling for Interest Rate Derivatives

With an Application to SOFR Caplets

Leif Andersen*and Dominique Bang
Bank of America

First Version: November 2019 This Version: September 2020

Abstract

With the forthcoming introduction of SOFR benchmark rates in the US, market participants will need to adjust their interest rate option models to accommodate a variety of idiosyncrasies of the SOFR rate. The materiality of these changes for quoted options level is currently unknown, and will depend on market sentiment (as expressed in market risk premia, say), regulatory policies, and the rate fixing conventions ultimately available in the market. While we wait for liquidity in SOFR options to build, this paper pre-emptively considers two important characteristics of SOFR derivatives: the backward-looking settlement style of SOFR floating rate payments; and the "jagged" nature of SOFR evolution through time. The latter originates with liquidity conditions in the repo financing markets from which SOFR is constructed, where temporary demand-supply imbalances can result in the formation of short-term spikes of substantial magnitude. We construct a variety of mechanisms that allows us to build rich stochastic models for both "surprising" and anticipated (e.g., year-end) spikes, and demonstrate how to modify existing (smooth) term structure models to capture them. To accommodate high-efficiency pricing of vanilla derivatives in top-down models, we also develop several convenient numerical techniques that allow for efficient pricing of these structures. For instance, a novel scheme merges existing spike-free pricing formulas with a given spike characteristic function in a custom low-dimensional quadrature routine, enabling us to spike-enable standard valuation models (such as SABR) at minimal computational effort. Using SOFR-style caplets for illustration, we numerically demonstrate that the effect of spikes on implied caplet volatility levels and skews can be substantial, even at modest levels of risk premia in the spike model parameters. Besides being useful for the pricing of SOFR derivatives, our paper more broadly establishes a complete mathematical framework for rate spikes, applicable to pricing, scenario generation, and risk management in any rates market where spike phenomena exist.

*Communicating author: leif.andersen@bofa.com. The views in this article are our own, and do not necessarily represent those of our employer. We thank Andy Dickinson, Mark Lake, Vladimir Piterbarg and participants in the Quant Summit Virtual 2020 for helpful comments.

1 Introduction

After significant liquidity problems in the Libor reference rates were laid bare by the Financial Crisis of 2007-2008, US financial regulators created the Alternative Reference Rate Committee (ARRC) with the intent of identifying an alternative reference rate backed by a deep pool of market transactions. In June 2017, the ARRC communicated that its preferred rate was a secured financing rate, named the *Secured Overnight Funding Rate* (SOFR), constructed from the $1 Trillion market for overnight Treasury bond repos[1].

The transition from Libor to SOFR is currently scheduled to take place over the next few years, a monumental undertaking with significant legal, technological, and operational hurdles. From a quantitative perspective, complications are also numerous and range from the relatively mundane (e.g., funding changes and SOFR discount curve construction) to the more complex (e.g, dynamic modeling of SOFR for option products).

A particular challenge pertains to the conversion of legacy Libor-based products to SOFR-based ones; a good discussion of some of the issues that arise for option products can be found in [17]. Complicating matters, in particular, is that SOFR- and Libor-based option products often differ materially, in part because of differences in fixing conventions, and in part because of the differences in the dynamics of Libor and SOFR. Our focus in this paper is on the interaction of fixing conventions with the unique SOFR dynamics, and how proper modeling of stochastic spikes in SOFR can make material differences for some SOFR option products.

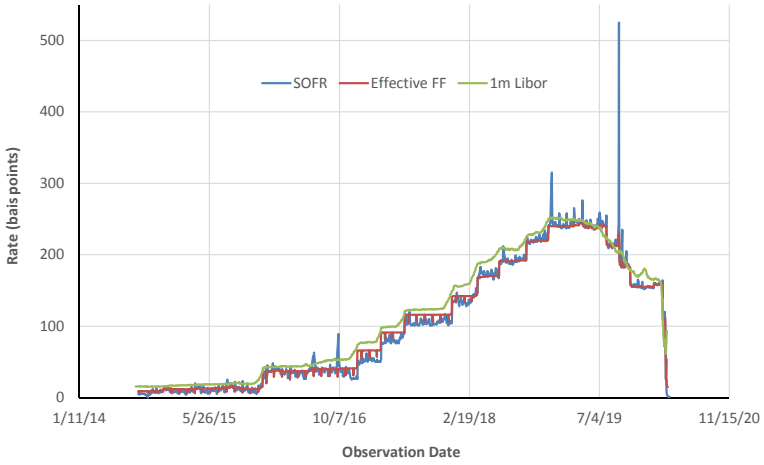
To cast some initial light on the spike phenomena in SOFR markets, consider that SOFR, given its synthesis, will inherit the characteristics of the market for short-term Treasury-backed financing rates. As it turns out, these funding markets have pronounced technical features and can be volatile, often much more so than other benchmark rates such as Fed Funds or Libor. The increased volatility is often especially pronounced at quarter- and year-end when banks' balance sheet adjustments can temporarily decrease their activity levels in Treasury repo markets; this, in turn, can produce sometimes dramatic, but typically short-lived, spikes in funding rates; see Figure 1. Between quarter-ends, unexpected news that affect banks' short-term need for cash or their demand for government securities, or influence the supply of Treasury bonds (e.g., regulatory communications and/or activities) may also cause spikes. A prominent example of this is the infamous 300 basis point spike that formed in mid-September of 2019; see [19], among many other news sources, for the post-mortem analysis of the various events that caused the spike to form.

The data in Figure 1 covers only 2014-2020, which is the period for which the FRB has published SOFR data. For earlier time periods, one can use the primary dealer survey (PDS) repo rate as a good SOFR proxy (see [21] for details on this rate). Of particular interest is how the PDS rate fared during the Financial Crisis of 2007-2008, a period where Treasury repo markets were under considerable stress. In Figure 2 below, we observe that the Treasury repo rate during the crisis exhibited large *downward* spikes, sometimes of many 100s of basis points in magnitude. As explained in detail in [13], these

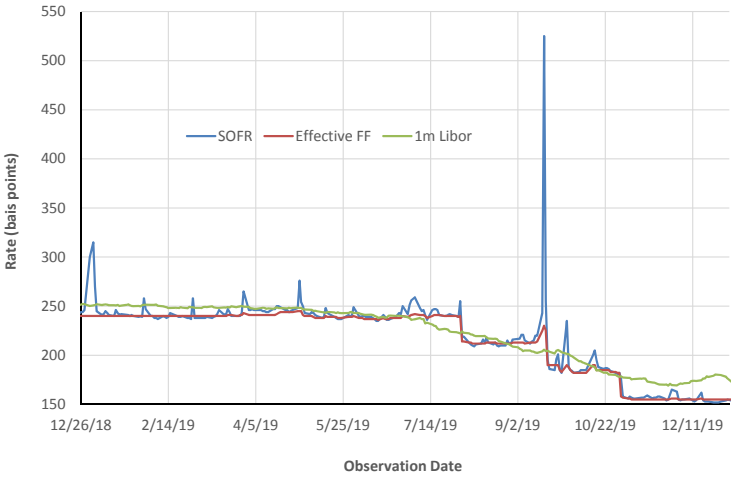
[1] For a primer on SOFR and other post-Libor reference rates around the world, see, for instance, [20].

Figure 1: Time Series of Various Interest Rates

(a) August 2014 - April 2020



(b) 2019 Calendar Year



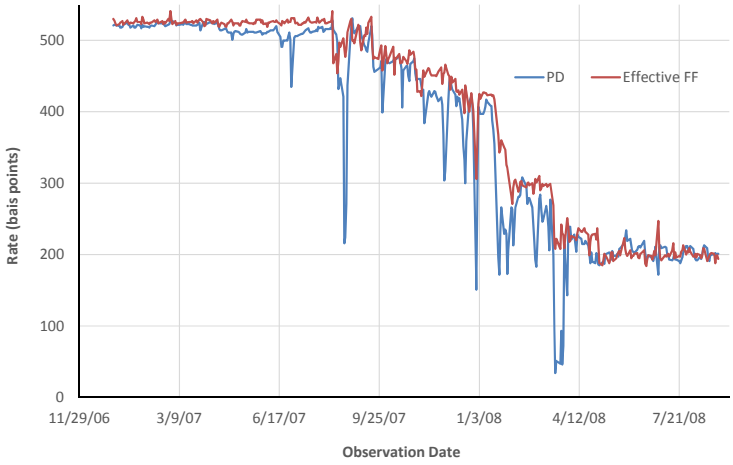
Time series of SOFR, Effective Fed Funds, and 1-month Libor rates. Source: websites of the New York and St. Louis Feds.

spikes are mostly a "flight to quality" phenomenon, with investors hoarding Treasury bonds and forcing recipients of Treasury collateral in repo arrangements to lower their lending rates.

Most of the spikes in Figures 1 and 2 last for only 1-3 days, but some can linger for longer, depending on how pervasive a particular event is. During the Financial Crisis many spike events lasted for a week or more, for instance.

While the presence of spikes in SOFR (and PDS) rates may simply be the way a true transaction-driven market rate behaves, the phenomenon still invites a number of controversies. First, many market participants, especially those in lending markets, have

Figure 2: Crisis Behavior of Treasury Repo Rates



Time series of the primary dealer survey repo rate (PDS) and the Effective Fed Funds rate, January 2007 through August 2008. Source: website of the New York Fed.

expressed unease with the volatility and have voiced doubt that behavior driven by technical demand features of Treasury funding markets should be present in a benchmark rate; see, for instance, [22] or any one of many news articles that were written after the large spike in September 2019. Second, large spikes in SOFR have on several occasions compelled the Federal Reserve to take action to dampen market volatility[2] (see [21]), which have led some market observers to question both the fundamental stability of repo markets and whether SOFR can be argued to be a market-driven rate at all, if regular regulatory intervention is needed to make it a palatable benchmark (e.g., [11]).

In this paper, we stay away from these controversies, and simply accept that the usage of SOFR rates may introduce a new technical complication: how does one modify standard interest rate models to be able to capture spike behavior? There is precedent in the financial model space for spike dynamics, but these efforts have mostly been focused on commodity markets (gas and, in particular, electricity) where demand-supply imbalances regularly cause seasonal spike effects; see [1] for details on some of these models. Porting these models to interest rates is, as we shall see, not without technical challenges.

While our analysis is, we think, of independent mathematical interest with several new results, we can imagine that some market practitioners will voice skepticism about the practical applicability of our modeling approach. First, they may argue, spikes can sometimes be of impressive magnitude but they are infrequent and thin, which might limit their influence on derivatives pricing. Second, they may argue, the SOFR fixing conventions are such[3] that many derivatives involve the formation of *averages* of one

[2]Of particular note is the fact that the FRB in 2013 started conducting daily reverse repo operations, with the aim of preventing large downward moves in repo rates.

[3]See Section 2 for details.

day SOFR rates on the accrual periods, which will tend to reduce the effect of spikes in the first place. And finally, they may say, the Federal Reserve will be so vigilant in the future (e.g., through formal institution of Treasury repo backstops) that spikes of any type will disappear from Treasury funding markets altogether. We cannot say much about the last objection, as the actions of the Federal Reserve and their effects on the market (and the length of time in which these actions are carried out) cannot be predicted at present. For the first objection, however, we note that an order of 5-10 spikes are not uncommon in a given year – enough to make significant impact on some derivatives, as we shall see – but we caution that arguments based on historical measure (a.k.a. $\mathbb{P}$-measure) spike frequencies and durations will fail to consider that strong risk premia will likely form around the duration, frequency, and height of spikes, for the reason we already outlined: market participants generally dislike the presence of spikes in benchmark rates, and are therefore likely to be averse to spike risk. As a consequence, one would expect that spike distribution parameters in the risk-neutral probability measure ($\mathbb{Q}$) – which is the one that matters for option pricing – will be more pronounced than in the historical probability measure ($\mathbb{P}$). In the absence of a liquid SOFR option market, it is presently hard to estimate the size of risk premia, but experience from default markets (credit default swaps), stock markets ("crash-o-phobia"), and commodity markets certainly suggest that the market price demanded for to carry this type of shock risk can be sizable.

As for the notion that spikes and volatility will be damped by averaging, this is a key argument of the ARRC (see [23]) in defense of SOFR. Indeed, as we will show later, there are many payouts for which this argument is reasonable, and for which simple repurposing of diffusion-based Libor models, as in [15], is justifiable. There are, however, also common option payouts where this argument is less convincing, which we shall demonstrate in this paper. In general, the framework we present will allow practitioners to understand the effect and materiality of the micro-structure of SOFR dynamics and volatility for any given payout structure, at modest additional complexity. Indeed, unlike a *top-down* modeling approach where existing models for simple forwards rates are equipped with simple *ad-hoc* extensions of the volatility structure to address the SOFR compounding features, we provide a *bottom-up* approach, with a proper account for physical phenomena at daily time scales[4].

The rest of this paper is organized as follows. In Section 2, we introduce notation and discuss the compounding conventions of SOFR in the context of a simple caplet. In particular, we distinguish between payouts that directly involve a rolling money-market account and those that are effectively written on (one or more) expectations of the money market account. We also elaborate further on the differences between top-down and bottom-up models for caplets. In Section 3, we introduce the basic concepts of spike- and diffusion-contributions to the overnight SOFR rate, and discuss some preliminary modeling candidates for each type of contribution. Our detailed treatment of spike dynamics is broken into several parts. First, in Section 4 we introduce a spike model can-

[4] As discussed later, we also provide convenient ways to extract top-down models from our bottom-up framework, as is often convenient in practice.

didate based on a classical affine jump-diffusion with a large mean-reversion. Based on our analysis of this approach, we progress, in Section 5, to a "compound" two-state continuous Markov chain model with time-dependent intensities and random jump sizes. This is a generally useful model framework for spikes (see [1]), and a number of results for the time integral (and its exponential) of such processes are derived, as needed for interest rate applications,. In Section 6, we consider certain approximations to the results in Section 5, an analysis that provides not only very convenient analytical simplifications, but also allows us, in Section 7, to introduce a series of useful extensions to the frameworks in Sections 4 and 5. Section 8 returns to the pricing of caplets, and lists a variety of computational techniques that allow for efficient pricing of these instruments; some of these techniques are subsequently used in Section 9 to illustrate how inclusion of spikes affects the volatility smiles of caplets. Finally, Section 10 concludes the paper. A series of appendices contain proofs and additional mathematical results.

2 Notation and Preliminaries

2.1 Spot and Forward Rates

Consider a time interval $[T_s, T_p]$ where we interpret T_s as a start date for floating rate fixing observations, and T_p as a payment date. For a short-rate process $r(t)$, consider defining the T_p-measurable quantity

$$R(T_p; T_s) \triangleq \frac{e^{\int_{T_s}^{T_p} r(u)\,du} - 1}{T_p - T_s}, \quad T_s < T_p. \tag{1}$$

We recognize $R(T_p; T_s)$ as a *backward*-looking "effective" rate on the interval $[T_s, T_p]$. R is closely related to the rolling money-market account β, defined as

$$\beta(t) = e^{\int_0^t r(u)\,du},$$

since

$$R(T_p; T_s) = \frac{\beta(T_p)/\beta(T_s) - 1}{T_p - T_s}.$$

While $R(T_p; T_s)$ is only observable at time T_p, its expectation may be defined at any time prior to time T_p. So, for any $t \le T_p$ let us define a forward rate

$$F(t; T_s, T_p) \triangleq \mathrm{E}_t^{T_p}\left(R(T_p; T_s)\right) = \frac{1}{P(t, T_p)}\mathrm{E}_t^{\mathbb{Q}}\left(\frac{\beta(t)}{\beta(T_p)}R(T_p; T_s)\right), \tag{2}$$

where $\mathrm{E}_t^{T_p}$ and $\mathrm{E}_t^{\mathbb{Q}}$ denote time t expectations in the T_p-forward ($\mathbb{Q}^{T_p}$) and risk-neutral ($\mathbb{Q}$) probability measures, respectively; and

$$P(t, T_p) = \mathrm{E}_t^{\mathbb{Q}}\left(\frac{\beta(t)}{\beta(T_p)}\right)$$

is the time t value of a discount bond maturing at time T_p. By definition, our newly defined forward rate $F(t;T_s,T_p)$ is a martingale in the T_p-forward measure.

The forward rate may conveniently be expressed in terms of discount bonds, but we need to differentiate between the case where $t \in [0,T_s]$ and the "stub" case where $t \in (T_s,T_p]$. For the former instance, we notice that

$$\begin{aligned}\mathrm{E}_t^{\mathbb{Q}}\left(\frac{\beta(t)}{\beta(T_p)}R(T_p;T_s)\right) &= \mathrm{E}_t^{\mathbb{Q}}\left(\frac{\beta(t)}{\beta(T_p)}\frac{\beta(T_p)/\beta(T_s)-1}{T_p-T_s}\right)\\ &= \mathrm{E}_t^{\mathbb{Q}}\left(\frac{\beta(t)/\beta(T_s)-\beta(t)/\beta(T_p)}{T_p-T_s}\right) = \frac{P(t,T_s)-P(t,T_p)}{T_p-T_s}\end{aligned}$$

so

$$F(t;T_s,T_p) = \frac{1}{T_p-T_s}\left(\frac{P(t,T_s)}{P(t,T_p)}-1\right), \quad t \in [0,T_s]. \tag{3}$$

For the stub case we get, in the same manner,

$$\begin{aligned}F(t;T_s,T_p) &= \frac{1}{P(t,T_p)}\mathrm{E}_t^{\mathbb{Q}}\left(\frac{\beta(t)/\beta(T_s)-\beta(t)/\beta(T_p)}{T_p-T_s}\right)\\ &= \frac{1}{T_p-T_s}\left(\frac{\beta(t)/\beta(T_s)}{P(t,T_p)}-1\right), \quad t \in (T_s,T_p].\end{aligned} \tag{4}$$

2.2 Caplet Definitions

To illustrate the usage of the above-defined rates in a concrete contract, we consider a K-strike European caplet with fixing at time T_s and payment at time T_p. Two different payout definitions are here reasonable. First, in a classical caplet (**type-I** for our purposes), we define the payout at time T_p, per unit of notional, as written on the T_s-observed forward rate:

$$V_I(T_p) = V_I(T_p;T_s,K) \triangleq (T_p-T_s)\cdot\left(F(T_s;T_s,T_p)-K\right)^+. \tag{5}$$

The definition (5) is reasonable for Libor-based caplets, since a term rate $F(T_s;T_s,T_p)$ is directly observable in the market at time T_s. For SOFR, such term rates do not exist (at least not yet), so it is more natural to define the caplet payout as[5] (**type-II** for our purposes)

$$V_{II}(T_p) = V_{II}(T_p;T_s,K) \triangleq (T_p-T_s)\cdot\left(R(T_p;T_s)-K\right)^+ = (T_p-T_s)\cdot\left(F(T_p;T_s,T_p)-K\right)^+. \tag{6}$$

Irrespective of how we define the payout, the present value of the caplet at some pre-maturity time $t \le T_p$ may be found by discounting the expectation of its terminal payout in measure $\mathbb{Q}^{T_p}$:

$$V_x(t) = P(t,T_p)\cdot\mathrm{E}_t^{T_p}\left(V_x(T_p)\right), \quad x = I, II. \tag{7}$$

[5] In actuality, the compounding will be done daily, rather than continuously. The error introduced by assuming continuous compounding is negligible.

2.3 Modeling Frameworks

2.3.1 Top-Down Caplet Models

When broad applicability is desired, the design of a full-blown interest rate model would proceed through a bottom-up specification of the dynamics of the short rate $r(t)$ and, possibly, that of the entire term structure of instantaneous forward rates; from these dynamics one can then infer the stochastic process for any yield curve quantity, including simply-compounded forward rates such as $F(t; T_s, T_p)$. For the narrow purpose of pricing European caplets, we could, however, dispense with the requirement of having a complete term structure model and adopt a top-down approach where we directly specify the dynamics of each individual caplet forward $F(t; T_s, T_p)$ separately, typically with different model parameters for different values of T_s and T_p. While this approach may not allow for coherent model extensions beyond caplet pricing, there is nevertheless long tradition for the practice of "one-at-a-time" forward or swap rate models when pricing plain-vanilla rates products; see [2] for many examples.

For a type-I caplet on a forward spanning $[T_s, T_p]$, suppose that we wish to use a continuous dynamic model driven by Brownian motion. In a top-down approach we may then, for instance, specify the process of F as a martingale-diffusion of the form

$$dF(t; T_s, T_p) = l\left(F(t; T_s, T_p)\right) \cdot z(t) \cdot dW^{T_p}(t) \tag{8}$$

where $l(\cdot)$ is some time-invariant local volatility function; $z(t)$ is a stochastic volatility process; and W^{T_p} is a $\mathbb{Q}^{T_p}$-Brownian motion. The selection of $l(\cdot)$ and z is in practice guided by the need for analytical tractability or, at least, by the existence of convenient analytical approximations. Popular choices for $l(\cdot)$ and z are those of the SABR family of models; see [9] and, for a more recent version that improves flexibility and remedies certain arbitrage issues, [4]. The SABR SDE is stated explicitly in Section 9.5.

Extending direct specifications such as (8) to type-II caplets will require at minimum an extension to capture how volatility decays away on the interval $[T_s, T_p]$, as information on the averaging ratio $\beta(T_p)/\beta(T_s)$ is gradually revealed. In the absence of a full-blown term structure model, this part must typically be done in an *ad-hoc* manner, as in [15], e.g., by writing

$$dF(t; T_s, T_p) = w(t; T_s, T_p) \cdot l\left(F(t; T_s, T_p)\right) \cdot z(t) \cdot dW^{T_p}(t) \tag{9}$$

where w is an exogenously specified parametric decay function that: decreases monotonically as t goes from T_s to T_p; satisfies $w(T_p; T_s, T_p) = 0$; and has $w(t; T_s, T_p) = 1$ for all $t \leq T_s$.

Given a decay function w, one can then proceed to ponder further how to extract closed-form type-II caplet prices for (9), assuming that such formulas already exist for type-I caplets in the time-invariant specification in (8). Fortunately, a variety of convenient parameter-averaging techniques exist to (approximately) convert (9) into (8), through the extraction of "effective" time-invariant parameters from a time-dependent model; see [16] for the general idea and [10] for the special case of SABR. A closely related approach would be to attempt to map a type-II caplet to a type-I caplet by increasing

either the maturity or the volatility of the type-I caplet, to take into account that volatility persists on the interval $[T_s, T_p]$ for a type-II caplet (see Section 2.3.2 for a simple example of this procedure).

2.3.2 *Bottom-Up Model Information Extraction: Gaussian Case*

A disadvantage of top-down models that work with exogenously specified parametric forms in the process specification of particular forward rates, is the difficulty of maintaining consistency with a coherent and arbitrage-free term structure model. As a consequence, it can be challenging to incorporate micro-structure effects into pricing dynamics in a granular manner. Given how repo markets, as described, are persistently subjected to local curve effects, this disadvantage can potentially be material for a rate such as SOFR. On the other hand, building a bottom-up model is typically more laborious, computationally costly, and may ultimately not allow for the parameter flexibility of models such as (9).

One potential way of bridging the gap between bottom-up and top-down models is to build a a bottom-up model that captures the physical effects that one desires, and then to extract information from this model and inject it into expressions such as (9) in a reasonable way that preserves tractability. We shall later show (in Section 8.3.2) how this approach, for instance, allows one to capture spike behavior in ad-hoc top-down models without sacrificing much in terms of the computational effort of option pricing expression.

For immediate inspiration, let us here, like other authors before us (see, e.g., [12] and [17]), take a look at a simple one-factor constant-parameter Gaussian term structure model to see how this might inspire a direct specification of the dynamics of a caplet forward rate F. We recall that such a Gaussian model may, for instance, be specified as a short-rate process

$$dr(t) = \kappa\left(\theta(t) - r(t)\right)\, dt + \sigma_r\, dW(t) \tag{10}$$

where $W(t)$ is a $\mathbb{Q}$-Brownian motion; κ is a mean-reversion speed; σ_r is the short-rate (basis point) volatility; and $\theta(t)$ is a deterministic function fully specified by the time 0 term structure of discount bond prices $P(0,\cdot)$.

Appendix A contains a number of useful details for the Gaussian rate model, and it is shown there that the $\mathbb{Q}^{T_p}$-measure process for F consistent with (10) is, with $\Delta \triangleq T_p - T_s$,

$$dF(t;T_s,T_p) = \left(\Delta^{-1} + F(t;T_s,T_p)\right)\sigma_F(t;T_s,T_p)\, dW^{T_p}(t), \tag{11}$$

where

$$\sigma_F(t;T_s,T_p) = \sigma_r \cdot \left(B(t,T_p) - 1_{t\in[0,T_s]}B(t,T_s)\right), \quad B(t,T) \triangleq \frac{1-e^{-\kappa(T-t)}}{\kappa}. \tag{12}$$

(11) is just an ordinary displaced log-normal diffusion process, so a closed-form caplet pricing formula for both type-I and type-II caplets is easily derived, see Appendix A. We are, however, here less interested in such pricing results, as our goal instead is to extract

information from the Gaussian model for use in a top-down approach. For this, consider that we might, for instance, use (11)-(12) to decide that the weight function w in (9) should look like

$$w(t;T_p,T_s) = \begin{cases} \frac{\sigma_F(t;T_s,T_p)}{\sigma_F(T_s;T_s,T_p)} = \frac{1-e^{-\kappa(T_p-t)}}{1-e^{-\kappa\Delta}}, & t \in (T_s,T_p], \\ 1, & t \in [0,T_s]. \end{cases}$$

For small κ, the weight function just becomes a linear decay: $\lim_{\kappa\downarrow 0} w(t;T_p,T_s) = (T_p - t)/\Delta$.

Suppose we instead wanted to use the Gaussian model as a way to determine what "effective" volatility scale $\eta \geq 1$ on σ_r used for a type-I caplet would make it equivalent to a type-II caplet fixing on the interval $[T_s,T_p]$. From the results in Appendix A, we see that type-I and type-II caplets would have the same price at time t, provided that their term variances match:

$$\eta^2 \int_t^{T_s} \sigma_F(u;T_s,T_p)^2\,du = \int_t^{T_p} \sigma_F(u;T_s,T_p)^2\,du, \quad t \in [0,T_s), \tag{13}$$

where again $\Delta = T_p - T_s$. Using the variance expressions in Appendix A, (13) can be solved for η:

$$\begin{aligned} \eta^2 &= \frac{1}{1-e^{-2\kappa\tau_s}}\left(\frac{2(e^{-\kappa\Delta}+\kappa\Delta-1)}{\left(e^{-\kappa\Delta}-1\right)^2} - e^{-2\kappa\tau_s}\right) \\ &= \left(\frac{\Delta}{3(T_s-t)}+1\right) + \kappa\frac{\Delta}{12}\left(\frac{\Delta}{T_s-t}+4\right) + o(\kappa^2), \quad t \in [0,T_s). \end{aligned} \tag{14}$$

When $\kappa = 0$, this replicates the special case in [17].

Not surprisingly, the result in (14) explodes as $t \uparrow T_s$. If we wish to avoid this, we could instead aim to increase the maturity used in type-I option term variance calculations[6] from T_s to $\bar{T}_s$, in the specific sense that

$$\int_t^{\bar{T}_s} \sigma_F(u;\bar{T}_s,\bar{T}_s+\Delta)^2\,du = \int_t^{T_p} \sigma_F(u;T_s,T_s+\Delta)^2\,du, \quad t \in [0,T_s).$$

It can easily be shown that the resulting value of $\bar{T}_s$ satisfies

$$\begin{aligned} \bar{T}_s &= T_s - \frac{1}{2\kappa}\ln\left(1+e^{2\kappa(T_s-\Delta-t)} - \frac{2(e^{-\kappa\Delta}+\kappa\Delta-1)}{e^{-2\kappa(T_s-t)}\left(1-e^{\kappa\Delta}\right)^2}\right) \\ &= T_s + \frac{1}{3}\Delta\left(1+\kappa\left(2(T_s-t)-\frac{17}{12}\Delta\right)\right) + o(\kappa^2), \quad t \in [0,T_s), \end{aligned} \tag{15}$$

which is well-defined even when $t = T_s$.

[6] But not in determining the forward rate $F(t;T_s,T_p)$ itself, obviously.

3 Spike Models: Basic Framework

Whether top-down or bottom-up, industry models for caplets in practice always assume that forward rates are diffusion-type continuous processes. Such models can certainly accommodate *deterministic* spikes at fixed time locations and known magnitudes and durations, by letting the initial term structure of instantaneous forward rates itself have spikes. This is, in fact, completely standard for interest rates dynamics, as curve interpolation artifacts, year-end effects, turns, hikes, etc., naturally tend to produce time 0 forward curves that are discontinuous; classical model dynamics are then used to generate smooth diffusion-style deviations away from the (possibly discontinuous) initial forward curve.

For SOFR, however, we are looking for an entirely new model mechanism that can add spikes to the short-rate process with *stochasticity* in three separate metrics for spikes: occurrence time, height, and duration. As spikes in SOFR rates constitute highly localized shocks on overnight rates, building a spike model is best done in a bottom-up approach where we can focus on the behavior of the short rate $r(t)$.

3.1 Short Rate Process

The approach to spike modeling used in this paper splits the short rate process into two components: a continuous process $r_C(t)$, and a spike process $r_S(t)$, such that

$$r(t) = r_C(t) + r_S(t). \tag{16}$$

For simplicity, the two processes $r_C(t)$ and $r_S(t)$ are assumed independent, which conveniently splits discount bond prices into separate contributions from the two processes:

$$P(t,T) = \mathrm{E}_t^{\mathbb{Q}}\left(e^{-\int_t^T r(u)\,du}\right) = \mathrm{E}_t^{\mathbb{Q}}\left(e^{-\int_t^T r_C(u)\,du}\cdot e^{-\int_t^T r_S(u)\,du}\right) = P_C(t,T)P_S(t,T), \tag{17}$$

where

$$P_x(t,T) \triangleq \mathrm{E}_t^{\mathbb{Q}}\left(e^{-\int_t^T r_x(u)\,du}\right), \quad x = C, S.$$

We define, in the same vein,

$$\beta_x(t) = e^{\int_0^t r_x(u)\,du}, \quad x = C, S. \tag{18}$$

The choice and specification of a continuous process for r_C is a standard problem, covered in detail in, for instance, [2]. Here, we thus focus our attention on $r_S(t)$, and take as our starting hypothesis that $r_S(t)$ can be written as

$$r_S(t) = J(t) - g(t) \tag{19}$$

where $J(t)$ is a spike process, yet to be defined, and $g(t)$ is a deterministic function. With this specification, we have

$$P_S(t,T) = \mathrm{E}_t^{\mathbb{Q}}\left(e^{-\int_t^T J(u)\,du}\right)\cdot e^{\int_t^T g(u)\,du} \triangleq P_J(t,T)\cdot e^{G(T)-G(t)}, \tag{20}$$

where

$$G(t) = \int_0^t g(u)\,du. \tag{21}$$

3.2 Calibration to Initial Term Structure

The role of $g(t)$ is to ensure that the overall rates model is consistent with the time 0 term structure of discount bonds, since, from (17),

$$P_S(0,T) = \frac{P(0,T)}{P_C(0,T)} = P_J(0,T) \cdot e^{\int_0^T g(u)\,du} = P_J(0,T) \cdot e^{G(T)}$$

which is satisfied if

$$G(T) = \ln \frac{P(0,T)}{P_C(0,T)P_J(0,T)}, \quad \text{or} \quad g(T) = \frac{\partial}{\partial T} G(T). \tag{22}$$

If we are given the initial term structure of discount bonds, $P(0,T)$, and have selected a model for $r_C(t)$ that produces discount factors of $P_C(0,T)$, then no matter what process we pick for $J(t)$, the overall rates model will be in calibration to the initial term structure of discount bonds provided that $G(\cdot)$ and $g(\cdot)$ are set as in (22). This, in turn, gives us ample freedom to experiment with spike model specifications for J, without concerns about the calibration to time 0 discount bonds.

3.3 Characteristic Function

Of course, to evaluate the expression (22), and to compute discount factors at any time t, we need a way to compute $P_J(t,T)$ as given by the $\mathbb{Q}$-expectation in (22). The challenge of this calculation shall occupy us shortly, as shall the closely related calculation of the characteristic function

$$\phi_S(k;t,T) \triangleq \mathrm{E}_t^{\mathbb{Q}}\left(e^{ik\int_t^T r_S(u)\,du}\right) = e^{-ik\int_t^T g(u)\,du}\phi_J(k;t,T), \quad \phi_J(k;t,T) \triangleq \mathrm{E}_t^{\mathbb{Q}}\left(e^{ik\int_t^T J(u)\,du}\right), \tag{23}$$

where i is the imaginary unit. Observe that $\phi_S(i;t,T) = P_S(t,T)$ and $\phi_J(i;t,T) = P_J(t,T)$, so in this sense the computation of the characteristic functions amount to an extension of the discount bond calculation. As we show in Section 8, knowledge of $\phi_S(k;t,T)$ is key to the practical computation of type-II caplet prices, and will be the main focus of the next several sections when we introduce a variety of concrete spike models.

4 First Model for J: Affine Jump Spikes

If availability of the characteristic function is important, it is tempting to consider the highly tractable class of affine jump processes, as in [6]. Here, we would write

$$dJ(t) = -\kappa_J J(t)\,dt + dU(t), \quad J(0) = 0, \tag{24}$$

where κ_J is a constant[7] and $U(t)$ is a pure jump process, e.g., a compound Poisson process. To generate something akin to a spike, we would set κ_J to a large number, so that the

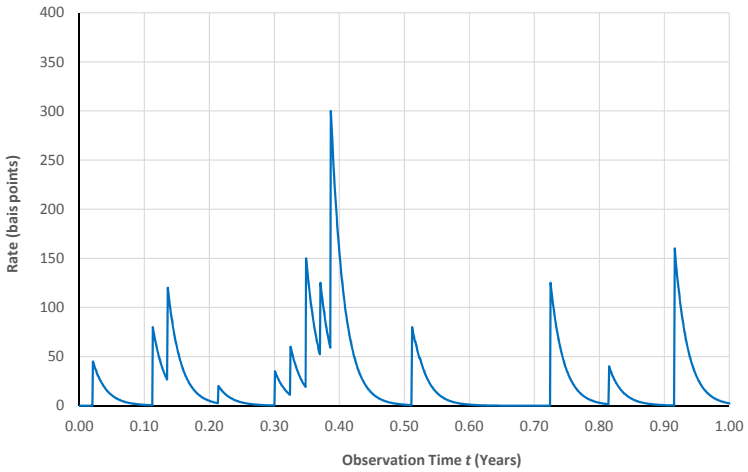
[7] Extensions to deterministic mean reversion speeds are trivial, of course.

effect on J from a jump in U would decay away rapidly, at an exponential rate of κ_J. That is, if U jumps to some (random) value $\tilde{U}$ at time t', $J(t)$ behaves as

$$J(t) = \left(J(t'-) + \tilde{U}\right) \exp^{-\kappa_J(t-t')}, \quad t > t',$$

until the next jump takes place. Schematic sample paths of $J(t)$ are shown in Figure 3 below; notice that all spikes are asymmetrical.

Figure 3: Sample Path of Affine Jump Process



A 1-year sample path for an affine jump process with positive jump magnitudes. The mean reversion speed was set to $\kappa_J = 50$, for a (risk-neutral) spike "width" of 1 week, as computed by the definition (28) .

As indicated, one attraction of (24) is its analytical tractability. To show a specific result, assume that $U(t)$ is compound Poisson with deterministic intensity $\lambda_U(t)$ and jump height $H(t)$. Let $H(t)$ have deterministic density $p_H(t,z)$, to which corresponds a jump height characteristic function of

$$\phi_H(k,t) = \mathrm{E}^{\mathbb{Q}}\left(e^{ik\cdot H(t)}\right) = \int_{-\infty}^{\infty} e^{ik\cdot z} p_H(t,z)\, dz. \tag{25}$$

A canonical example of ϕ_H is that of the *Gaussian* distribution with mean $\mu(t)$ and standard deviation $\sigma(t) > 0$:

$$\phi_H^G(k,t) = \exp\left(ik\cdot\mu(t) - \sigma(t)^2k^2/2\right).$$

If we want certainty that spikes can only be positive, we could, say, instead use an *exponential* distribution with deterministic parameter $\lambda(t) > 0$:

$$\phi_H^E(k,t) = \frac{1}{1 - ik/\lambda(t)}.$$

For the framework above, the characteristic function ϕ_J in (23) can be computed in closed form, as shown in Proposition 1. Since the basic derivation technique shall be useful later, we include a brief proof of the proposition.

Proposition 1 *Let $J(t)$ satisfy (24), with $U(t)$ being a compound Poisson process with intensity $\lambda_U(t)$; let the jump height $H(t)$ have characteristic function $\phi_H(k,t)$, as in (25). Then $\phi_J(k;t,T) = \phi_J(k;t,T,J(t))$, where $\phi_J(k;t,T,J)$ satisfies the backward integro-differential equation*

$$\lambda_U(t)\left(\int_{-\infty}^{\infty}\phi_J(k;t,T,J+z)p_H(t,z)\,dz - \phi_J(k;t,T,J)\right) + \frac{\partial\phi_J(k;t,T,J)}{\partial t} - \kappa_J J\frac{\partial\phi_J(k;t,T,J)}{\partial J} = -ikJ\phi_J(k;t,T,J), \quad (26)$$

subject to $\phi_J(k;T,T,J) = 1$. The solution to (26) is

$$\phi_J(k;t,T,J) = e^{A_J(t,T;k)+ikJ\cdot B_J(t,T)} \quad (27)$$

where

$$B_J(t,T) = \frac{1-e^{-\kappa_J(T-t)}}{\kappa_J}, \quad A_J(t,T;k) = \int_t^T \lambda_U(u)\left(1-\phi_H(k\cdot B_J(u,T),u)\right)\,du.$$

Proof:

Consider the definition of $\phi_J(k;t,T)$ in (23), and freeze $T > t$. Suppressing dependence on k and T, and introducing when necessary dependence on $J(t)$, notice that on $[t,t+dt]$

$$d\phi_J(t) = \phi_J\left(t,J(t)+dU(t)\right) - \phi_J\left(t,J(t)\right) + \frac{\partial\phi_J(t,J(t))}{\partial t}\,dt - \kappa_J J(t)\frac{\partial\phi_J(t,J(t))}{\partial J(t)}\,dt,$$

such that (since $\mathbb{Q}(dU(t)\neq 0) = \lambda_U(t)\,dt$)

$$\frac{\mathrm{E}_t^{\mathbb{Q}}\left(d\phi_J(t)\right)}{dt} = \lambda_U(t)\left(\int_{-\infty}^{\infty}\phi_J\left(t,J(t)+z\right)p_H(t,z)\,dz - \phi_J\left(t\right)\right) + \frac{\partial\phi_J(t)}{\partial t} - \kappa_J J(t)\frac{\partial\phi_J(t)}{\partial J(t)}.$$

But also, by the definition of $\phi_J(k;t,T)$, $\mathrm{E}_t^{\mathbb{Q}}\left(d\phi_J(t)\right)/dt = -ikJ(t)\phi_J(t)$ which combines with the expression above to yield (26). The analytical solution of (26) proceeds by inserting the exponential *ansatz* in (27) into (26), yielding the simple ODE system

$$\frac{\partial B_J(t,T)}{\partial t} - \kappa_J B_J(t,T) = -1, \quad \text{s.t. } B_J(T,T) = 0,$$
$$\lambda_U(t)\left(\phi_H(kB_J(t;T),t)-1\right) + \frac{\partial A_J(t,T;k)}{\partial t} = 0, \quad \text{s.t. } A_J(T,T;k) = 0.$$

The solution to this system is (27). ∎

4.1 Discussion

While (24) is tractable, the approach has some drawbacks. First, we notice that the asymmetric spike realizations (see Figure 3) are rather different in appearance from the historical ones, a consequence of the fact that affine spikes technically persists until perpetuity. Indeed, the effect of a jump in J is not truly a localized spike and the concept of a spike width is not immediately obvious. We could potentially define the "effective width" a as[8]

$$a = \int_{t'}^{\infty} e^{-\kappa_J(u-t')}\, du = \frac{1}{\kappa_J}, \tag{28}$$

but by this definition a material part (close to 40%) of the spike would continue to linger outside the window $[t', t+a]$. Moreover, the process $J(t)$ still would have very significant downward drift everywhere on $[t', t'+a]$, which is perhaps less intuitive for a model of a persistent, albeit short-term, supply-demand shocks.

The issue with persistence of shocks is perhaps particularly noticeable at high levels of jump intensity λ_U (relative to κ_J) as spikes will then start blending into each other and prevent the formation of truly distinct spikes. So, if we, say, increase λ_U locally to model a near-certain quarter- or year-end type spike, absent some type of controlled increase in κ_J the sample paths of J will start containing many overlapping jumps resulting in exaggerated run-up[9] in J. The run-up phenomenon (at a modest scale) can be seen in Figure 3, at around $t = 0.35$.

Besides a certain lack of control over the separation of spikes, the model (24) also lacks any true stochasticity of spike durations. As discussed earlier, spike width can vary from days to weeks, and a flexible spike model should reasonably allow for such variability. We could potentially extend (24) to work with stochastic mean-reversion, but not only would analytical tractability potentially suffer, the resulting effect on spike width would not be particularly transparent, nor would there be much flexibility in picking distributions for spike width. A possible alternative is to work with multiple independent J-processes, all with different mean reversion. As we shall leverage this type of "multi-process" idea later, we expand briefly on it below.

4.2 Mixtures of Multiple Affine Jump Processes

Consider now setting $J(t) = \sum_{i=1}^{N} J_i(t)$ where each process $J_i(t)$ is of the type (24), but with each J_i process having different mean reversion speeds. For generality, we can also extend to allow each J_i to have separate jump intensities $\lambda_{U,i}(t)$ and jump-magnitude characteristic functions $\phi_{H,i}(k,t)$. The characteristic function $\phi_J(k;t,T)$ is then easily

[8] An alternative definition of width would define it as the amount of time it would take for the spike to be reduced below some absolute materiality threshold ε (say 1 basis point), in which case tall spikes would be wider than short ones; this effect is not necessarily consistent with empirical evidence, as large spikes tend to attract a more forceful regulatory response.

[9] For a possible remedy, see Section 7.3.

computed as

$$\phi_J(k;t,T) = \prod_{i=1}^{N} \phi_{J,i}(k;t,T)$$

where each $\phi_{J,i}$ can be computed separately from Proposition 1.

For the process $J(t)$, we can identify a total jump intensity at time t as

$$\lambda_U(t) = \sum_{i=1}^{N} \lambda_{U,i}(t).$$

Moreover, given that a jump in $J(t)$ takes place at time t, we can compute the probability of it originating from the J_i process as

$$p_i(t) = \mathbb{Q}(dU_i(t) \neq 0 | dU(t) \neq 0) = \frac{\lambda_{U,i}(t)}{\lambda_U(t)}, \quad i = 1, \ldots, N.$$

In this sense, we can interpret, say, the mean reversion speed following a jump as having a stochastic mixture distribution with levels $\kappa_{J,i}$ and probabilities $p_i(t)$.

5 Second Model for J: 2-State Compound Markov Chain

In this section, we wish to improve on the dynamic properties of the affine model, by constructing a framework that allows for truly local spikes and for stochasticity in both jump widths and in jump heights. For the purpose of cleanly separating "excited" conditions, where the market is experiencing a demand-supply shock, from "regular" market conditions, we will use a model that explicitly incorporates such distinct, discrete states.

5.1 Basic Idea

To lay the foundation for a simple model for spike phenomena, consider a continuous-time Markov chain $c(t)$ with two discrete states: *base* (state e_1) and *excited* (state e_2). As seen in the risk-neutral measure $\mathbb{Q}$, the transition from state e_1 to state e_2 (resp. e_2 to e_1) is assumed to take place with a deterministic intensity $\lambda_{12}(t) \geq 0$ (resp. $\lambda_{21}(t) \geq 0$). The generator matrix for the chain is therefore

$$Q(t) = \begin{pmatrix} -\lambda_{12}(t) & \lambda_{12}(t) \\ \lambda_{21}(t) & -\lambda_{21}(t) \end{pmatrix}.$$

Any time spent in the excited state e_2 will be used to model the presence of a spike; to keep the spike suitably narrow on average, we would normally have λ_{21} be quite large, to ensure that an excited state resolves itself back to the base level fairly quickly. We use time-dependent intensities $\lambda_{12}(t)$ and $\lambda_{21}(t)$ to ensure that we can capture seasonality (quarter- and year-end) and known future events (e.g. Treasury auctions and FRB hikes) where the likelihood of moving into an excited state is higher than normal.

We emphasize in particular that even if we briefly increase $\lambda_{12}(t)$ to very high levels during a small time period (to model a near-certain event), the model will not suffer the "run-up" problem discussed in the context of affine models, as every jump into an excited state will automatically be completely extinguished (by a jump into the base-state), before another spike can form. Further, if $\lambda_{21}(t)$ is kept at reasonable levels during any (brief) period where $\lambda_{12}(t)$ is very high, we will typically also only see a single spike form, as the Markov chain will stay in state e_2 during the period where $\lambda_{12}(t)$ intensities are elevated. The ability to keep spikes crisply distinct and to have control over spike count and duration is a significant advantage of the Markov chain approach.

Now, in the same spirit as for the affine model, we associate each transition from e_1 into e_2 with a random draw of a variable H with time-dependent density $p_H(t,z)$ and characteristic function $\phi_H(k,t)$. Specifically, our spike process $J(t)$ is set to zero whenever $c(t)$ is in the base state e_1, but if, say, c moves from state e_1 to e_2 at time t', then J is assumed to jump by a random amount $H(t')$, with deterministic density of $p_H(t',\cdot)$. At the next transition from state e_2 to state e_1, J will then jump back to zero.

We assume that H is independent of c and that the probability of $H = 0$ is zero, as would be the case if the density was bounded in a region around zero, say. This implies that if we observe that $J(t) = 0$ ($J(t) \neq 0$), we know a.s. that $c(t) = e_1$ ($c(t) = e_2$). As a consequence, bond price expectations P_J may be written as deterministic functions of J only:

$$P_J(t,T) = P_J(t,T,J(t)). \tag{29}$$

5.2 Some Technical Results for the Markov Chain

As foundation for later work, we now list some results – some straightforward, others less so – for the Markov chain $c(t)$.

5.2.1 Probability Distribution of $c(t)$

Starting from a known state $c(t)$, let us consider the probability distribution of state $c(s)$, $s \geq t$. Define therefore, in the risk-neutral measure

$$p_{ij}(t,s) = \mathbb{Q}\left(c(s) = e_j | c(t) = e_i\right), \quad s \geq t,$$

and assemble these quantities in a 2×2 matrix $p(t,s)$. Elementary properties of the Poisson process shows that, for instance,

$$\begin{aligned} p_{11}(t,s+ds) &= p_{11}(t,s)(1-\lambda_{12}(s)\,ds) + p_{12}(t,s)\lambda_{21}(s)\,ds \\ &= p_{11}(t,s)(1+Q_{11}(s)\,ds) + p_{12}(t,s)Q_{21}(s)\,ds. \end{aligned} \tag{30}$$

Proceeding in the same fashion for the remaining i,j, we get $p(t,s+ds) = p(t,s)\,(I+Q(s))$ or, for $s \geq t$,

$$\frac{\partial p(t,s)}{\partial s} = p(t,s)Q(s), \quad p(t,t) = I, \tag{31}$$

where I is the 2×2 identity matrix. This is, of course, just the forward Kolmogorov equation for our inhomogenous Markov chain.

The matrix $p(t,s)$ is sometimes known as the (time-ordered) *exponential matrix* of $Q(s)$; it can be solved directly from (31) by a standard ODE solver, such as Runge-Kutta. Alternatively, its components can be solved in closed-form, as an expression involving time integrals of the Markov chain intensities:

Lemma 1 *Define* $\bar{p}(t,s) \triangleq e^{-\int_t^s (\lambda_{12}(u)+\lambda_{21}(u))\,du}$. *Then*

$$p_{11}(t,s) = \bar{p}(t,s)\left(1 + \int_t^s \frac{\lambda_{21}(u)}{\bar{p}(t,u)}\,du\right), \quad p_{12}(t,s) = 1 - p_{11}(t,s),$$
$$p_{22}(t,s) = \bar{p}(t,s)\left(1 + \int_t^s \frac{\lambda_{12}(u)}{\bar{p}(t,u)}\,du\right), \quad p_{21}(t,s) = 1 - p_{22}(t,s).$$

Proof: We only show the result for p_{11}, the results for the remaining p_{ij} are proven the same way. First, observe that $p_{12}(t,s) = 1 - p_{11}(t,s)$ which together with (30) yields

$$\frac{\partial p_{11}(t,s)}{\partial s} = -\lambda_{12}(s)p_{11}(t,s) + (1 - p_{11}(t,s))\lambda_{21}(s) = -\left(\lambda_{12}(s) + \lambda_{21}(s)\right)p_{11}(t,s) + \lambda_{21}(s),$$

subject to $p_{11}(t,t) = 1$. The solution to this standard initial value ODE problem is as given in the Lemma. ∎

We note in passing that we may write $\bar{p}(t,s) = \bar{p}_1(t,s)\bar{p}_2(t,s)$, where

$$\bar{p}_1(t,s) \triangleq e^{-\int_t^s \lambda_{12}(u)\,du}, \quad \bar{p}_2(t,s) \triangleq e^{-\int_t^s \lambda_{21}(u)\,du}. \tag{32}$$

The quantities $\bar{p}_i(t,s)$ are here the probabilities of starting in state e_i at time t and never transitioning away from state e_i on the interval $[t,s]$.

5.2.2 Transition Counts

Let us now examine the number of transitions between the two states of the Markov chain. For instance, consider defining $N_{up}(t,s)$ as the (random) number of transitions from state e_1 to state e_2 on the time interval $[t,s]$; we can loosely think of this as the number of spikes produced on the interval. We can define probabilities

$$\alpha_{ij}^n(t,s) \triangleq \mathbb{Q}\left(N_{up}(t,s) = n, c(s) = e_j | c(t) = e_i\right), \quad i,j = 1,2. \tag{33}$$

Given the nature of spike modeling, we are primarily interested in the case[10] $c(t) = e_1$, where we notice that:

$$\alpha_{12}^0(t,s) = 0 \text{ for all } s \geq t; \tag{34}$$
$$\alpha_{11}^0(t,t) = 1; \tag{35}$$
$$\alpha_{1j}^n(t,t) = 0 \text{ for all } n > 0 \text{ and } j = 1,2. \tag{36}$$

[10]The case where $c(t) = e_2$ can be established in identical manner and, in the interest of brevity, is left to the reader.

By the same recurrence technique we used above, we see that

$$\alpha_{11}^n(t, s+ds) = \alpha_{11}^n(t,s)\,(1-\lambda_{12}(s)\,ds) + \alpha_{12}^n(t,s)\lambda_{21}(s)\,ds, \quad n \geq 0,$$
$$\alpha_{12}^n(t, s+ds) = \alpha_{12}^n(t,s)\,(1-\lambda_{21}(s)\,ds) + \alpha_{11}^{n-1}(t,s)\lambda_{12}(s)\,ds, \quad n > 0,$$

which, after a slight realignment, results in the following Lemma.

Lemma 2 *With α_{11} and α_{21} defined as in (33), then*

$$\frac{\partial \alpha_{11}^n(t,s)}{\partial s} = -\lambda_{12}(s)\alpha_{11}^n(t,s) + \lambda_{21}(s)\alpha_{12}^n(t,s), \quad n \geq 0,$$
$$\frac{\partial \alpha_{12}^n(t,s)}{\partial s} = -\lambda_{21}(s)\alpha_{12}^n(t,s) + \lambda_{12}(s)\alpha_{11}^{n-1}(t,s), \quad n > 0.$$

The boundary conditions are listed in (34)-(36).

We can iteratively solve the equations in Lemma 2, using standard ODE results. The result, which involves nested integrals, is omitted for brevity.

In some circumstances, one would expect that a compact closed-form result for the transition count distribution exists. Proposition 2 shows one result, which requires that the ratio $\lambda_{12}(t)/\lambda_{21}(t)$ is constant. Figure 4 in Section 6.1 contains some numerical results using the formulas in Proposition 2.

Proposition 2 *Consider the differential system established in Lemma 2, and define the ordinary generating functions (OGFs) for the sums over α_{11}^n and α_{12}^n:*

$$\mathcal{G}_{11}(t,s;\tau) \triangleq \sum_{n=0}^{+\infty} \alpha_{11}^n(t,s)\tau^n, \quad \mathcal{G}_{12}(t,s;\tau) \triangleq \sum_{n=0}^{+\infty} \alpha_{12}^n(t,s)\tau^n, \quad \mathcal{G}(t,s;\tau) \triangleq \begin{pmatrix} \mathcal{G}_{11}t,s;\tau) \\ \mathcal{G}_{12}(t,s;\tau) \end{pmatrix}.$$

Assuming that $\nu \triangleq \lambda_{12}(t)/\lambda_{21}(t)$ is constant, then we have

$$\mathcal{G}(t,s;\tau) = e^{-\frac{1+\nu}{2}\Lambda_{21}(t,s)} \left[\frac{\sinh\left(\delta(\tau)\Lambda_{21}(t,s)\right)}{\delta(\tau)} \begin{pmatrix} \frac{1-\nu}{2} \\ \nu\tau \end{pmatrix} + \cosh\left(\delta(\tau)\Lambda_{21}(t,s)\right) \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right] \tag{37}$$

where we denoted

$$\delta(\tau) \triangleq \sqrt{\left(\frac{1-\nu}{2}\right)^2 + \nu\tau}, \quad \Lambda_{21}(t,s) \triangleq \int_t^s \lambda_{21}(u)\,du = -\ln \bar{p}_2(t,s). \tag{38}$$

Proof: See Appendix B. ∎

Remark 1 *The assumption of constant ν in Proposition 2 can be relaxed to allow ν to be piece-wise constant. The solution for $\mathcal{G}(t,s;\tau)$ would then require a recursive integration on each sub-interval where ν is constant (updating at each time the boundary conditions).*

The result (37) allows us to determine $\alpha^n_{11}(t,s)$ and $\alpha^n_{12}(t,s)$ (and their sum) by a straightforward Taylor expansion of (37) in τ at order n.

In applications, one may wish to maintain a constant λ_{21} to keep the spike *width* distribution time-homogenous, even when λ_{12} varies through time; this violates the conditions of Proposition 2. We can get around this issue in several ways, for instance by pragmatically replacing ν with time averages (e.g., $\Lambda_{12}(t,s)/\Lambda_{21}(t,s)$ where $\Lambda_{12}(t,s)$ is defined as $-\ln \bar{p}_1(t,s)$). Another approach would be to split the Markov chain into the sum of multiple Markov chains each of which satisfy the conditions of Proposition 2; see Sections 4.2 and 6.5 for the basic idea.

5.3 Compound Markov Chain

5.3.1 Density for J

Let us first contemplate the density of $J(s)$, given a starting state at time $t \leq s$. The following result is the relevant one:

Proposition 3 *Define*

$$p^1_J(t,s,y)\,dy \triangleq \mathbb{Q}\left(J(s) \in [y, y+dy] | c(t) = e_1\right),$$
$$p^2_J(t,s,x,y)\,dy \triangleq \mathbb{Q}\left(J(s) \in [y, y+dy] | c(t) = e_2, J(t) = x\right).$$

Then, with $\delta(y)$ being Dirac's delta-function in $y = 0$,

$$p^1_J(t,s,y) = p_{11}(t,s)\delta(y) + \bar{p}_2(t,s)\int_t^s \frac{p_{11}(t,v)\lambda_{12}(v)p_H(v,y)}{\bar{p}_2(t,v)}\,dv, \tag{39}$$

$$p^2_J(t,s,x,y) = p_{21}(t,s)\delta(y) + \bar{p}_2(t,s)\delta(y-x) + \bar{p}_2(t,s)\int_t^s \frac{p_{21}(t,v)\lambda_{12}(v)p_H(v,y)}{\bar{p}_2(t,v)}\,dv. \tag{40}$$

Proof: (sketch)

We can prove the proposition in several ways, e.g. by writing down evolutionary ODEs from fundamental principles and then solving them analytically. For a more intuitive approach, consider, for instance, the case where $c(t) = e_1$ and note that $J(s)$ reaching some level $y \neq 0$ will require: 1) c is at level e_1 at some time $v \geq t$, and then jumps to level e_2 at time $v+$; 2) $H(v)$ is drawn at level y; and 3) c stays in level e_2 to time s. The probability of steps 1-3 for any fixed value of v is

$$p_{11}(t,v)\lambda_{12}(v)dv \cdot p_H(v,y) \cdot \bar{p}_2(v,s) = p_{11}(t,v)\lambda_{12}(v)dv \cdot p_H(v,y) \cdot \frac{\bar{p}_2(t,s)}{\bar{p}_2(t,v)}.$$

If we integrate over all values of $v \in (t,s)$ we recover the probability density of $J(s)$ for those states where $J(s) \neq 0$. Adding to this expression a Dirac delta-function with mass $p_{11}(t,s)$ will capture those outcomes where $J(s) = 0$, and recovers the result (39). The result (40) can be produced in similar fashion; notice that we here also need a delta-function to capture those scenarios where the Markov chain stays in state e_2 (and J therefore stays at x) for the entirety of the interval $[t,s]$, the probability of which is $\bar{p}_2(t,s)$. ∎

5.3.2 *Backward Equation for Characteristic Function of* $\int J(u)\,du$

For derivatives pricing, the density of $J(t)$ itself is typically less interesting than the density – or, more or less equivalently, the characteristic function ϕ_J – of the *time integral* of $J(t)$, as we discussed in Section 3. For this, we first consider establishing a backward equation for the quantities

$$\phi_J^1(k;t,s) = \mathrm{E}^{\mathbb{Q}}\left(e^{ik\int_t^s J(u)\,du}|c(t)=e_1\right), \tag{41}$$

$$\phi_J^2(k;t,s,y) = \mathrm{E}^{\mathbb{Q}}\left(e^{ik\int_t^s J(u)\,du}|c(t)=e_2, J(t)=y\right). \tag{42}$$

As mentioned before, setting $k = i$ would recover the discount bond pricing results needed for, say, model calibration; see Section 3.2.

The derivation of a backward equation can proceed along similar lines as the proof of Proposition 1, although some care must be taken to explicitly incorporate the dynamics of transitions between states in the Markov chain.

Lemma 3 *The quantities $\phi_J^1(k;t,s)$ and $\phi_J^2(k;t,s,y)$ in (41)-(42) satisfy the following backward equation:*

$$\frac{\partial\phi_J^1(k;t,s)}{\partial t} = \lambda_{12}(t)\phi_J^1(k;t,s) - \lambda_{12}(t)\int_{-\infty}^{\infty}\phi_J^2(k;t,s,y)p_H(y,t)\,dy, \quad \phi_J^1(k;s,s)=1, \tag{43}$$

$$\frac{\partial\phi_J^2(k;t,s,y)}{\partial t} = (-iky+\lambda_{21}(t))\,\phi_J^2(k;t,s,y) - \lambda_{21}(t)\,\phi_J^1(k;t,s), \quad \phi_J^2(k;s,s,y)=1. \tag{44}$$

Proof: See Appendix B. ∎

The result in Lemma 3 expresses ϕ_J^2 and ϕ_J^1 as a coupled system of *integro-differential* equations. We can decouple the equations by first solving (44) to relate ϕ_J^2 and ϕ_J^1:

$$\phi_J^2(k;t,s,y) = e^{iky(s-t)}e^{-\int_t^s\lambda_{21}(u)\,du} + \int_t^s e^{iky(v-t)}e^{-\int_t^v\lambda_{21}(u)\,du}\lambda_{21}(v)\phi_J^1(k;v,s)\,dv. \tag{45}$$

Inserting this expression into (43), and using the definition (25) for ϕ_H, a few re-arrangements leads to a single integro-differential equation for ϕ_J^1:

$$\begin{aligned}\frac{\partial\phi_J^1(k;t,s)}{\partial t} &= \lambda_{12}(t)\phi_J^1(k;t,s) - \lambda_{12}(t)\,e^{-\int_t^s\lambda_{21}(u)\,du}\phi_H(s-t,t)\\ &\quad - \lambda_{12}(t)\int_t^s\phi_H(v-t,t)\,e^{-\int_t^v\lambda_{21}(u)\,du}\lambda_{21}(v)\phi_J^1(k;v,s)\,dv,\end{aligned} \tag{46}$$

subject to $\phi_J^1(k;s,s)=1$.

Whether we choose to solve (43)-(44) or (45)-(46), numerical methods are generally needed[11]. However, convenient approximation methods exist (see Section 6), and in special cases more convenient results are possible. For instance, assuming time-homogeneity, (46) can be solved using a Laplace transform formalism (where the integral component can be seen as a product of convolution, easily handled in the Laplace domain). If in addition to time homogeneity we also suppose that H has a discrete distribution, we get an even simpler result:

Proposition 4 *Consider the integro-differential problem (46), and assume that ϕ_H is discrete and time-homogenous, as in*

$$\phi_H(k,t) = \phi_H(k) = \sum_{p=1}^{N} \pi_p e^{ikz_p},$$

for some probability weights π_p and nodes z_p. Also assume that $\lambda_{12}(t) = \lambda_{12}$, $\lambda_{21}(t) = \lambda_{21}$ for constants λ_{12}, λ_{21}. Then, the solution to (46), $n(t) \triangleq \phi_J^1(k;t,s)$ satisfies the following order $N+1$ ordinary differential equation:

$$\sum_{p=0}^{N} \alpha_p n^{(p+1)}(t) = \lambda_{12}\left[\lambda_{21}\sum_{p=0}^{N}\alpha_p \sum_{i=0}^{p-1} n^{(i)}(t) + \sum_{p=0}^{N}\alpha_p n^{(p)}(t)\right] \tag{47}$$

subject to the boundary conditions $n(s) = 1$, $n'(s) = 0$, and

$$n^{(p+1)}(s) = \lambda_{12}\left[\lambda_{21}\sum_{i=0}^{p-1} n^{(i)}(s) + n^{(p)}(s) - \bar{\phi}_H^{(p)}(0)\right], \quad 1 \le p \le N-1.$$

Here, α_p and $\bar{\phi}_H$ are defined via

$$\prod_{p=1}^{N}(\omega + iz_p - \lambda_{21}) = \sum_{p=0}^{N}\alpha_p \omega^p, \quad \bar{\phi}_H(k) \triangleq e^{\lambda_{21}k}\phi_H(-k) = \sum_{p=1}^{N}\pi_p e^{(\lambda_{21}-iz_p)k}.$$

Proof: See Appendix B. ∎

In Proposition 4 the linear ODE system (47) can be solved by standard methods, e.g., using matrix exponentiation. We notice that the result of Proposition 4 can easily be extended to cover piece-wise constant functions $\phi_H(k,t)$, $\lambda_{12}(t)$ and $\lambda_{21}(t)$ by recursively integrating the problem on an adapted time subdivision.

Finally, we observe that the choice made for ϕ_H in Proposition 4 corresponds to a discrete distribution $p_H(z) = \sum_{p=1}^{N}\pi_p\delta(z-z_p)$. Weights π_p and nodes z_p can, for instance, be inspired by a Gaussian quadrature to best mimic a desired (continuous) distribution.

[11] Solution of (46) can be done by combining a numerical ODE solver with a trapezoid integration scheme for the integral component; see [18].

5.3.3 *Forward Equation for Characteristic Function of* $\int J(u)\,du$

Should one want to rely on numerical methods to establish the characteristic function ϕ_J in (104), it will often be more efficient to rely on a *forward* equation, rather than the backward equation of Lemma 3. In applications we normally stand at a fixed time t (say, $t = 0$ for the calibration problem in Section 3.2) and wish to numerically solve for $\phi_J(k; t, s)$ in a single sweep for multiple values of s, starting with $s = t$ and incrementally moving s *forward* while t is fixed. The backward equations, however, work the "opposite way": here we start with $t = s$ and incrementally move t backwards, while s is fixed.

While often less convenient in numerical work, backwards equations are typically easy to derive and general in nature. In contrast, forward equations are generally reserved only for quantities directly associated with the density or with discounted densities, also known as *Arrow-Debreu state prices* or the *Green's function*. So, to work out forward equations, consider the quantities

$$p^1_{AD}(k; t, s, y) = \mathrm{E}^{\mathbb{Q}}\left(e^{ik\int_t^s J(u)\,du}\delta(J(s) - y))\middle| c(t) = e_1\right), \tag{48}$$

$$p^2_{AD}(k; t, s, x, y) = \mathrm{E}^{\mathbb{Q}}\left(e^{ik\int_t^s J(u)\,du}\delta(J(s) - y))\middle| c(t) = e_2, J(t) = x\right). \tag{49}$$

We note that then, for instance,

$$\phi^1_J(k; s, t) = \mathrm{E}^{\mathbb{Q}}\left(e^{ik\int_t^s J(u)\,du}\middle| c(t) = e_1\right) = \int_{-\infty}^{\infty} p^1_{AD}(k; t, s, y)\,dy; \tag{50}$$

these expectations are ultimately what we are mostly interested in and can be done by numerical integration.

For brevity, we focus only on p^1_{AD} (p^2_{AD} may be trivially done by the same techniques) and define the intermediate quantities

$$z^{11}_{AD}(k; t, s) = \mathrm{E}^{\mathbb{Q}}\left(e^{ik\int_t^s J(u)\,du}1_{c(s)=e_1}\middle| c(t) = e_1\right),$$

$$z^{12}_{AD}(k; t, s, y) = \mathrm{E}^{\mathbb{Q}}\left(e^{ik\int_t^s J(u)\,du}1_{c(s)=e_2}\delta(J(s) - y)\middle| c(t) = e_1\right).$$

Then we have the following proposition:

Proposition 5 *We may write*

$$p^1_{AD}(k; t, s, y) = z^{11}_{AD}(k; t, s)\delta(y) + z^{12}_{AD}(k; t, s, y),$$

where, subject to initial conditions $z^{12}_{AD}(k; t, t, y) = 0$ *and* $z^{11}_{AD}(k; t, t) = 1$,

$$\frac{\partial z^{12}_{AD}(k; t, s, y)}{\partial s} = (ik \cdot y - \lambda_{21}(s))\, z^{12}_{AD}(k; t, s, y) + z^{11}_{AD}(k; t, s)p_H(s, y)\lambda_{12}(s), \tag{51}$$

$$\frac{\partial z^{11}_{AD}(k; t, s)}{\partial s} = -z^{11}_{AD}(k; t, s)\lambda_{12}(s) + \lambda_{21}(s)\int_{-\infty}^{\infty} z^{12}_{AD}(k; t, s, y)\,dy. \tag{52}$$

Proof: See Appendix B. ∎

The result in Proposition 5 is, like that of Lemma 3, a coupled system of integro-differential equations. We may detangle these equations in the same manner as (45)-(46), yielding integro-differential equations that, under assumption of time-homogeneity, may be solved via Laplace transform methods. As was the case for the backward equation (see Proposition 4), assuming further that the distribution of H is discrete yields results similar to those in Proposition 4. We omit details for brevity.

6 Approximations to Markov Chain Model

Although we have provided sufficient numerical and analytical results to make the 2-state compound Markov chain model operable in practice, for the purpose of type-II caplet pricing (which we imagine will be a high-volume trading business), we would like to make the framework even more convenient to use. For this, we first focus on deriving compact expressions for the case where λ_{21} is large relative to λ_{12} (i.e., when spikes are thin and when the exact spike timing is mostly a surprise). As it turns out, the resulting approximation idea, when supplemented with techniques to handle spikes with near-certain timing, can form the basis for a very flexible spike model framework in its own right, a topic we return to in Section 7.

6.1 Approximation to $N_{up}(t,s)$

We first revisit previous results, in Section 5.2.2, on the distribution of the number of up-transitions in the Markov chain c. If $\lambda_{21}(t)$ is much smaller than $\lambda_{12}(t)$, we might perhaps be able to ignore the contribution of down-jumps and simply approximate the number of jumps in our Markov chain with the number of jumps in a (inhomogeneous) Poisson process with intensity $\lambda_{12}(t)$. With the notation in Section 5.2.2, this amounts to

$$\alpha_{11}^n(t,s) + \alpha_{12}^n(t,s) \approx \frac{(-\ln \bar{p}_1(t,s))^n \, \bar{p}_1(t,s)}{n!}. \tag{53}$$

(53) can be improved in various ways to take into account the width of the spikes. For instance, we could write

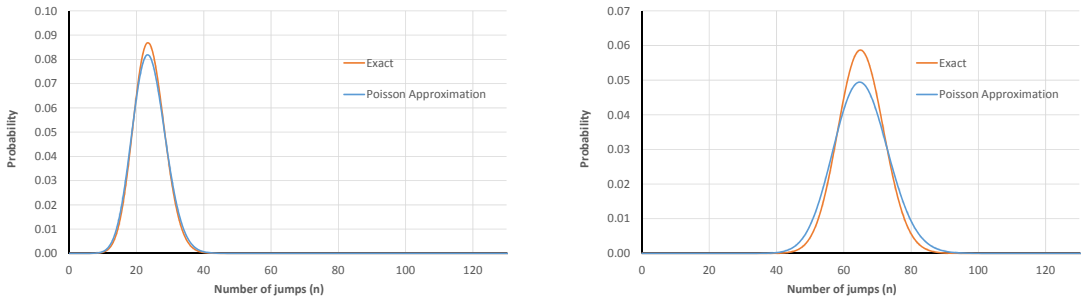
$$\alpha_{11}^n(t,s) + \alpha_{12}^n(t,s) \approx \frac{\left(\int_t^s \tilde{\lambda}(u)\,du\right)^n e^{-\int_t^s \tilde{\lambda}(u)\,du}}{n!}, \tag{54}$$

where $\tilde{\lambda}$ scales down λ_{12} by the ratio $\lambda_{21}/(\lambda_{12}+\lambda_{21})$. More precisely,

$$\tilde{\lambda}(u) = \tilde{\lambda}(u;t,s) = \lambda_{12}(u)\frac{\int_t^s \lambda_{21}(u)\,du}{\int_t^s \lambda_{12}(u)\,du + \int_t^s \lambda_{21}(u)\,du} = \lambda_{12}(u)\frac{\ln \bar{p}_2(t,s)}{\ln \bar{p}_1(t,s) + \ln \bar{p}_2(t,s)}. \tag{55}$$

Figure 4: Poisson Approximation of Jump Count Distribution

(a) $\lambda_{12} = 5, \lambda_{21} = 100$ (b) $\lambda_{12} = 15, \lambda_{21} = 100$



Test of the Approximation (54) for $s - t = 5$yrs. Intensity parameters were constant at the levels listed in the panels above. The "Exact" numbers in the graphs were computed from Proposition 2.

Figure 4 below illustrates the quality of this approximation for different parameters. Roughly speaking, the accuracy is good if λ_{21} is more than 5-10 times larger than λ_{12}.

Next, define a t-indexed process

$$I(t; t_0) \triangleq \int_{t_0}^{t} J(u)\, du,$$

and assume (as would most often be the case in practice) that the chain c is in state e_1 at time t_0; extensions to $c(t_0) = e_2$ are straightforward. We note that the sample paths of $I(t; t_0)$ are flat as a function of t, except at excursion intervals where c jumps into state e_1 and then reverts to state e_2. For instance, assume that c jumps to state e_2 at time $t' > t_0$ and reverts back to state e_1 at time t''. Also assume that the jump variable H takes the value $H(t')$. Then in the vicinity of t', $I(t; t_0)$ is

$$I(t; t_0) = I(t'; t_0) + (t - t')H(t'), \quad t \in [t', t'']. \tag{56}$$

Truncating the step heights to reflect some horizon end time T, we can more accurately associate[12] the "up-jump" time t' with the quantity

$$\Lambda(t'; T) = H(t')(\min(t'', T) - t').$$

With this truncation, we note that if there are up-jump times $t_1, t_2, t_3, \ldots, t_k$ on the interval $[t_0, T]$, then

$$I(T; t_0) = \sum_{i=1}^{k} \Lambda(t_i; T). \tag{57}$$

In light of the result in Section 6.1, this representation is reminiscent of compound Poisson processes with random jumps of height $\Lambda(t_i; T)$, an observation that we will proceed to exploit.

[12] We will colloquially use the term "up-jump" ("down-jump") to indicate transitions from e_1 to e_2 (resp. e_2 to e_1), even though spikes actually can be both downward and upward in our framework.

6.2 Time-homogenous Excursions

Since compound Poisson processes are particularly tractable when jump distributions are time-invariant, let us first assume that λ_{21} is constant and H has time-independent density (we will relax this below). Under these assumptions, we can define a characteristic function for the individual spike contributions to $I(t_0; T)$:

$$\begin{aligned}\phi_\Lambda(T-t', k) &\triangleq \mathrm{E}^{\mathbb{Q}}_{t'}\left(e^{ik\cdot\Lambda(t';T)}\right)\\ &= \int_{\mathbb{R}^+} \mathrm{E}^{\mathbb{Q}}_{t'}\left(e^{ik\cdot H\cdot(\min(t'',T)-t')}|t''-t'=x\right)\lambda_{21}e^{-\lambda_{21}x}\,dx\\ &= \lambda_{21}\int_0^{T-t'}\phi_H(kx)e^{-\lambda_{21}x}\,dx\; + e^{-\lambda_{21}(s-t')}\phi_H\left(k\cdot(T-t')\right),\end{aligned}$$

where ϕ_H was defined in (25) and where the integral may sometimes be available in closed form, but is generally easy to compute numerically.

6.2.1 *Examples of $\phi_\Lambda(h, k)$.*

To give examples, observe that if ϕ_H is that of a Gaussian with mean μ and standard deviation σ, then[13]

$$\phi^G_\Lambda(h,k) = \lambda_{21}\int_0^h \exp\left((ik\cdot\mu-\lambda_{21})\,x-\sigma^2k^2x^2/2\right)\,dx + e^{-\lambda_{21}h}\exp\left(ik\cdot\mu h-\sigma^2k^2h^2/2\right). \tag{58}$$

If ϕ_H is Exponential with parameter λ, then we have

$$\phi^E_\Lambda(h,k) = \lambda_{21}\int_0^h \frac{e^{-\lambda_{21}x}}{1-ik\cdot x/\lambda}\,dx + e^{-\lambda_{21}h}\frac{1}{1-ik\cdot h/\lambda}. \tag{59}$$

The integrals in (58) and (59) may be computed in closed form by this lemma:

Lemma 4 *Denote $\mathcal{I}(h) \triangleq \int_0^h \exp\left(ax-b^2x^2/2\right)\,dx$ and $\mathcal{J}(h) \triangleq \int_0^h (1+bx)^{-1}\exp\left(-ax\right)\,dx$. Then we have:*

$$\mathcal{I}(h) = \frac{\sqrt{\pi}}{b\sqrt{2}}\exp\left(\frac{a^2}{2b^2}\right)\left[\mathrm{erf}\left(\frac{-a/b+bh}{\sqrt{2}}\right)-\mathrm{erf}\left(-\frac{a/b}{\sqrt{2}}\right)\right], \tag{60}$$

$$\mathcal{J}(h) = \frac{e^{a/b}}{b}\left(\mathrm{E}_1\left(a/b\right)-\mathrm{E}_1\left(a/b+ah\right)\right), \tag{61}$$

where the error function *is defined as* $\mathrm{erf}(z) \triangleq \frac{2}{\sqrt{\pi}}\int_0^z \exp\left(-v^2\right)\,dv$ *and the* exponential integral *is defined as* $\mathrm{E}_1(z) \triangleq \int_z^\infty v^{-1}\exp\left(-v\right)\,dv$.

[13] Notice that the integral here diverges for all imaginary values of k if the upper limit goes to infinity. Without the truncation at some finite s, the product of an exponential and a Gaussian has a tail that is too fat to allow for a moment generating function. Of course, if k is real the integral will still be defined. We also note that the product of an exponential and a Gaussian also has a density that is infinite at the origin, although this singularity is integrable.

Proof: As $ax - b^2x^2/2 = \frac{1}{2}a^2/b^2 - \frac{1}{2}b^2(x - a/b^2)^2$, we can use the change of variable $v = \frac{1}{\sqrt{2}}b(x - a/b^2)$ in $\mathcal{I}(h)$. This results in

$$\mathcal{I}(h) = \frac{\sqrt{2}}{b}\exp\left(\frac{a^2}{2b^2}\right)\int_{-\frac{a}{b\sqrt{2}}}^{-\frac{a}{b\sqrt{2}}+\frac{bh}{\sqrt{2}}}\exp\left(-v^2\right)\,dv$$

which proves (60). As for $\mathcal{J}(h)$, we perform the substitution $v = a/b + ax$, such that

$$\int_0^h \frac{e^{-ax}}{1+bx}\,dx = \frac{e^{a/b}}{b}\int_{\frac{a}{b}}^{\frac{a}{b}+ah}\frac{e^{-v}}{v}\,dv.$$

which is (61). ∎

Remark 2 *For the purpose of evaluating $e^x\mathrm{E}_1(x)$ for large x, we can use the asymptotic result*

$$e^x\mathrm{E}_1(x) \approx x^{-1}A(x), \quad A(x) = \frac{x^2 + k_1x + k_2}{x^2 + k_3x + k_4}$$

where $k_1 = 4.03640$, $k_2 = 1.15198$, $k_3 = 5.03637$, $k_4 = 4.19160$.

6.3 Approximation of Characteristic Function ϕ_J

We are now ready to attack the characteristic function ϕ_J of the entire integral $I(T;t_0)$. Let us start by writing (with $\Lambda(t_0;T) \equiv 0$)

$$\begin{aligned}\phi_J^1(k;t_0,T) &= \mathrm{E}_{t_0}^{\mathbb{Q}}\left(e^{ik\cdot I(T;t_0)}|c(t_0)=e_1\right) = \mathrm{E}_{t_0}^{\mathbb{Q}}\left(e^{ik\sum_{n=0}^{Nup(t_0,T)}\Lambda(t_n;T)}|c(t_0)=e_1\right)\\ &= \sum_{n=0}^{\infty}\mathbb{Q}\left(N_{up}(t_0,T)=n|c(t_0)=e_1\right)\mathrm{E}^{\mathbb{Q}}\left(e^{ik\sum_{i=0}^{n}\Lambda(t_i;T)}|N_{up}(t_0,T)=n,c(t_0)=e_1\right)\\ &= \sum_{n=0}^{\infty}\left(\alpha_{11}^n(t_0,T)+\alpha_{12}^n(t_0,T)\right)\mathrm{E}^{\mathbb{Q}}\left(e^{\sum_{i=0}^{n}\Lambda(t_i;T)}|N_{up}(t_0,T)=n,c(t_0)=e_1\right) \qquad (62)\end{aligned}$$

where $\alpha_{11}^n(t,T)$, $\alpha_{12}^n(t,T)$ were defined in Section 5.2.2.

As written, (62) is exact, but not necessarily easy to compute. To proceed, we can make the following simplifying assumptions:

Assumption 1 *Besides assuming that λ_{21} is constant and H has time-independent distribution, assume that:*

- *Conditional on $N_{up}(t_0,T)$ (and $c(t_0)=e_1$) the $\Lambda(t_i;T)$ are approximately independent.*
- *Conditional on $N_{up}(t_0,T)=n$ (and $c(t_0)=e_1$), the $\Lambda(t_i;T)$ all approximately have characteristic function of $\phi_\Lambda^E(h_n,k)$ where $h_1,h_2,..$ are given constants.*

Lemma 5 *Under Assumption 1, we have*

$$\begin{aligned}\mathrm{E}^{\mathbb{Q}}_{t_0}\left(e^{ik\cdot I(T;t_0)}|c(t_0)=e_1\right) &\approx \sum_{n=0}^{\infty}\left(\alpha^n_{11}(t_0,T)+\alpha^n_{12}(t_0,T)\right)\prod_{i=0}^{n}\mathrm{E}^{\mathbb{Q}}\left(e^{ik\cdot\Lambda(t_i;T)}|N_{up}(t_0,T)=n,c(t_0)=e_1\right)\\ &\approx \sum_{n=0}^{\infty}\left(\alpha^n_{11}(t_0,T)+\alpha^n_{12}(t_0,T)\right)\phi_\Lambda(h_n,k)^n. \qquad (63)\end{aligned}$$

Remark 3 *In Lemma 5, we can imagine setting $h_n = (T-t_0)/n$, for instance.*

While (63) is straightforward to compute, we can use the result of Section 6.1 to make it even simpler by making the following additional assumptions.

Assumption 2 *In addition to Assumption 1, we also assume that:*

- *$\alpha^n_{11}(t,T)+\alpha^n_{12}(t,T)$ may be approximated as in (54).*
- *h_n may be treated as approximately constant, at some level $\bar{h}$, independent of n.*

This leads to the following very simple expression for the characteristic function of $I(T;t_0)$.

Proposition 6 *Under Approximation 2,*

$$\phi_J(k;t_0,T)=\mathrm{E}^{\mathbb{Q}}_{t_0}\left(e^{ik\cdot I(T;t_0)}\right)\approx\mathrm{E}^{\mathbb{Q}}_{t_0}\left(e^{ik\cdot I(s;t_0)}|c(t_0)=e_1\right)\approx e^{(\phi_\Lambda(\bar{h},k)-1)\int_{t_0}^{T}\tilde{\lambda}(u;t_0,T)\,du},$$

where $\tilde{\lambda}(u;t_0,T)$ is given in (55).

Proof: With Assumption 2, we have

$$\begin{aligned}\mathrm{E}^{\mathbb{Q}}_{t_0}\left(e^{ik\cdot I(T;t_0)}|c(t_0)=e_1\right) &\approx \sum_{n=0}^{\infty}\frac{\left(\int_{t_0}^{T}\tilde{\lambda}(u)\right)^n\phi_\Lambda(\bar{h},k)^n e^{-\int_{t_0}^{T}\tilde{\lambda}(u)\,du}}{n!}\\ &=\sum_{n=0}^{\infty}\frac{\left(\phi_\Lambda(\bar{h},k)\int_{t_0}^{T}\tilde{\lambda}(u)\right)^n}{n!}e^{-\int_{t_0}^{T}\tilde{\lambda}(u)\,du}\\ &=e^{\phi_\Lambda(\bar{h},k)\int_{t_0}^{T}\tilde{\lambda}(u)}e^{-\int_{t_0}^{T}\tilde{\lambda}(u)\,du}=e^{(\phi_\Lambda(\bar{h},k)-1)\int_{t_0}^{T}\tilde{\lambda}(u)}.\end{aligned}$$

If spikes are thin, then also

$$\mathrm{E}^{\mathbb{Q}}_{t_0}\left(e^{ik\cdot I(T;t_0)}\right)\approx\mathrm{E}^{\mathbb{Q}}_{t_0}\left(e^{ik\cdot I(T;t_0)}|c(t_0)=e_1\right)$$

and we are done. ∎

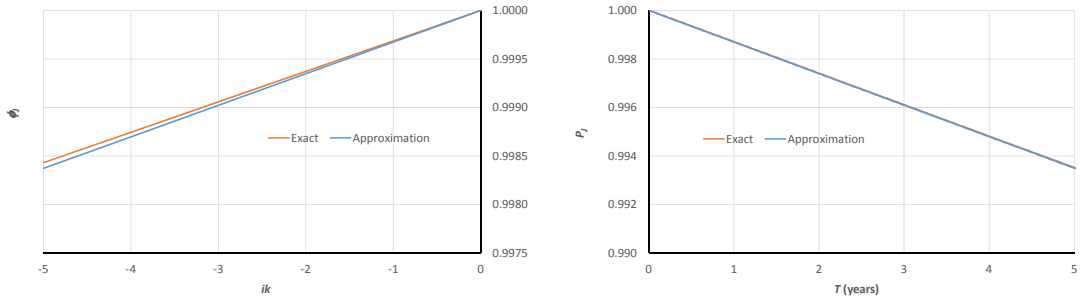
The set of assumptions behind Proposition 6 can be expected to be accurate in the limit of high λ_{21} (relative to λ_{12}), where spike width is narrow. The numerical example below examines precision for some practically relevant parameter magnitudes.

6.4 Numerical Test

Let us consider a model where risk-neutral intensities are $\lambda_{12} = 15$ and $\lambda_{21} = 100$, corresponding to an average spike width of $1/100$ years (about 2.5 business days), and an average of $1/15$ years (about 0.8 months) time spent between spikes. We use an exponential distribution for jump heights, with a parameter of $\lambda = 100$, for an average rate jump height of $1/100$ or 1%. Panel (a) in Figure 5 shows the exact function $\phi_J(k; 0, 0.25)$ versus the approximate one computed by Proposition 6, using Lemma 4 and (59) in Section 6.2.1 to establish ϕ_Λ. To keep the characteristic function real, we used purely imaginary values of k, ranging from $0 \cdot i$ to $5 \cdot i$. The value of $\phi_J(k; 0, 0.25)$ at $k = i$ (approximately 0.9997) is $P_J(0, 0.25)$. Given the practical importance of the case $k = i$, where ϕ_J becomes equal to P_J, Panel (b) in the figure shows $\phi_J(i; 0, T) = P_J(0; T)$ for various values of T

Figure 5: Approximation of ϕ_J and P_J

(a) $\phi_J(k; 0, 0.25)$ (b) $P_J(0, T)$



Panel (a): $\phi_J(k; 0, 0.25)$ for exponentially distributed jumps, as a function of ik. Panel (b): $P_J(0, T)$ (i.e., $\phi_J(i; 0, T)$), as a function of T. Model parameters were: $\lambda_{12} = 15$, $\lambda_{21} = 100$, and $\lambda = 100$. The results in the figure were insensitive to $\bar{h}$; we concretely used $\bar{h} = 1$.

For practical applications, the precision of the approximation appears more than adequate (indeed, the approximation error on P_J is not discernible at the resolution level of the graph).

6.5 Time-Dependent Excursion Distributions

In practice, there may be the need for having time-dependence in both the height and duration of jumps. We can handle this the same way as we did in Section 4.2 for the affine jump model, by splitting the Markov chain into multiple independent chains $c_1, c_2, \ldots, c_m$, with chain j having time-dependent up-jump intensity $\lambda^j_{12}(t)$ and constant down-jump intensity λ^j_{21}. We would also assume that each chain is associated with independent jump magnitudes $H_1, H_2, \ldots, H_m$, all with different (but time-homogenous) distributions.

While certainly not necessary, we could insist that for any given value of t, only one of the $\lambda^j_{12}(t)$ functions is different from zero – this way, the multiple chains are kept

(mostly) separate, and the overall setup could closely mimic a single chain with "up" intensity of

$$\lambda_{12}(t) = \sum_{j=1}^{m} \lambda_{12}^{j}(t)$$

and m different levels of "down" intensity, which we, say, could use to model any piecewise flat $\lambda_{21}(t)$. Also, we would be able to model m different jump distributions.

The above idea effectively writes the jump integral on $[t_0, T]$ as

$$I(T; t_0) = \sum_{j=1}^{m} I^j(T; t_0)$$

for m independent sub-integrals that are all of the simple type discussed in the previous section. This allows for easy extension of the approximations we developed earlier; for instance, under the terms of Assumption 2, we would have

$$\phi_J(k; t_0, T) = \mathrm{E}_{t_0}^{\mathbb{Q}}\left(e^{ik\cdot I(T;t_0)}\right) = \prod_{j=1}^{m} \mathrm{E}_{t_0}^{\mathbb{Q}}\left(e^{ik\cdot I^j(T;t_0)}\right) \approx \prod_{j=1}^{m} e^{(\phi_\Lambda^j(\bar{h}_j,k)-1)\int_{t_0}^{T}\tilde{\lambda}_j(u)\,du}, \tag{64}$$

where the definition of $\phi_\Lambda^j(\bar{h}, k)$ and $\tilde{\lambda}_j(u)$ should be obvious.

7 Third Model for J: Dirac Delta Spikes

7.1 Basic Specification

The ideas in Section 6 above essentially amount to approximating the dynamics of the time-integral I of jumps in the 2-state inhomogeneous compound Markov chain framework with a finite set of compound Poisson processes with time-dependent jump intensities. The intuition is straightforward: if spikes are infrequent and thin (i.e. λ_{21} is large relative to λ_{12}), the sample path for I across a spike in the interval $[t', t'']$ closely resembles a step function, with random step height of $(t'' - t')H(t')$. By modeling the spike as a step in I, we effectively approximate "thin" spikes in $J(t)$ as Dirac delta-functions (or *impulses*) with a random mass equal to the area of the spikes being approximated. The parameters of the approximating compound Poisson process are extracted bottom-up from the parameters of the original Markov chain model.

We could, however, flip the above procedure on its head, and outright *define* the process for $I(t; t_0)$ as the sum of m independent (inhomogeneous) compound Poisson processes with jump-intensities $\tilde{\lambda}_j(t)$ and random jump magnitudes Λ_j, independent of each other. Each Λ_j would then have exogenously specified characteristic functions of $\phi_\Lambda^j(k)$. With this framework, the expression (64) becomes exact:

$$\phi_J(k; t_0, T) = \mathrm{E}_{t_0}^{\mathbb{Q}}\left(e^{ik\cdot I(T;t_0)}\right) = \prod_{j=1}^{m} e^{(\phi_\Lambda^j(k)-1)\int_{t_0}^{T}\tilde{\lambda}_j(u)\,du} \tag{65}$$

and the approximations in Section 6 would simply amount to a way of a) intuitively interpreting the model; and b) parameterizing the model in practice.

7.2 Extensions

Besides improving the analytical tractability of the model considerably, we notice that the compound Poisson framework allows us to mimic a much broader set of distributions of spike duration, beyond the exponential distribution inherent in the 2-state Markov chain setup. For instance, we may introduce bounded distributions for spike width if we believe that spike durations are restricted to finite intervals (e.g., longer than a day, less than 10 business days), something that would also get around certain technical issues with exponential distributions (see footnote 13). Generally, if the jth compound Poisson process is associated with spikes with random duration D_j and random magnitude of H_j, then we can define $\Lambda_j = D_j H_j$ and parameterize $\phi^j_\Lambda(m)$ as

$$\phi^j_\Lambda(k) = \mathrm{E}^{\mathbb{Q}}_{t_0}\left(e^{ik\cdot D_j H_j}\right) = \int_{\mathbb{R}^+} \mathrm{E}^{\mathbb{Q}}_{t_0}\left(e^{ik\cdot xH_j}\right) d_j(x)\,dx = \int_{\mathbb{R}^+} \phi^j_H(kx) d_j(x)\,dx,$$

where $d_j(x)$ is the density of the jth duration D_j.

For instance, if spike durations of the jth compound Poisson process are uniformly distributed on $[a_j, b_j]$, then simply

$$\phi^j_\Lambda(k) = \frac{1}{b_j - a_j}\int_{a_j}^{b_j} \phi^j_H(kx)\,dx. \tag{66}$$

Even simpler, one could pragmatically assume that spikes can only last multiple of full days and that the duration is a *discrete* uniform probability on the interval of integers $[n_j, m_j]$. Then:

$$\phi^j_\Lambda(k) = \frac{1}{m_j - n_j + 1}\sum_{p=n_j}^{m_j} \phi^j_H(kp\delta). \tag{67}$$

where δ is 1 business day.

7.3 Year- and Quarter-End Spike Modeling

As mentioned, the approach in Section 7.2 above originated from expressions (in Section 6.5) derived under the assumption that $1/\tilde{\lambda}$ is significantly larger than the average spike duration. This is, by the very nature of spikes, an assumption that one would certainly expect to hold for a typical spike process, but there are situations where it may not hold *locally*. For instance, some types of spikes can have very little uncertainty about their timing, which, if modeled in a Poisson framework, may imply very high levels of $\tilde{\lambda}$ in a narrow interval in time. For instance, the earlier mentioned year- and quarter-end spikes – *YQ spikes* for short – are characterized by near-certainty that the short-rate will spike significantly on a narrow interval (often a single day) in the future. We emphasize that the exact height of the short-rate on the interval in question remains uncertain, but its location in time is known with a high degree of certainty. If these spikes are directional in nature (which they typically are), they will manifest themselves as outright spikes in the 1-day forward curve of interest rates, on the dates in question (see Section 9.4).

Suppose that we know with high certainty that a YQ spike will take place at some narrow time interval $[t', t' + \epsilon]$ in the future. Modeling of this spike can be done straightforwardly in the Markov chain framework by increasing significantly the up-jump intensity $\lambda_{12}(t)$ to some large constant λ'_{12} for $t \in [t', t' + \epsilon]$. This way, the probability p of at least one up-jump on $[t', t' + \epsilon]$ will be (assuming that the chain is in state e_1 at time $t'-$)

$$p = 1 - \exp\left(-\lambda'_{12} \cdot \epsilon\right) \tag{68}$$

which we can use to dimension λ'_{12}. For instance, if we want a 90% chance of an up-jump, we set $\lambda'_{12} = -\ln 0.1/\epsilon$. As usual, $\lambda_{21}(t)$ for $t \in [t', t' + \epsilon]$ could be set based on the average duration for YQ spikes, for instance, and would typically be smaller than λ'_{12}.

The approach above would work perfectly well for the Markov chain, but if we attempt to port it over to a setting with Dirac delta functions, we risk the same type of "run-up" phenomenon we described for the affine jump process (see Section 4). Specifically, if we set $\tilde{\lambda}$ in (65) locally to a very high number – e.g. by using the dimensioning technique in (68)– the probability of more than a single jump in $[t', t' + \epsilon]$ would be

$$1 - \exp\left(-\lambda'_{12} \cdot \epsilon\right) - \lambda'_{12} \cdot \epsilon \exp\left(-\lambda'_{12} \cdot \epsilon\right) = p + (1 - p)\ln(1 - p),$$

which can be verified to be much larger than zero if p is close to 1 (it approaches 1 when p approaches 1). We will thus have a high likelihood of many more than a single jump taking place in $[t', t' + \epsilon]$ – and with each jump adding a full Dirac delta-function to the rates process. As discussed earlier this undesirable behavior is not shared by the full Markov Chain approach, since the chain must jump down to state e_1 before a second jump can takes place; this makes the chain effectively wait out the period $[t', t' + \epsilon]$ where the jump intensity λ_{12} is elevated without producing more than a single up-jump.

To accommodate YQ-type spikes in Dirac delta models[14], we could consider adding to the model in Section 7.2 (where jump timing is a surprise) a new type of point process where the location of all future jumps are essentially known in advance. Let $t'_1, t'_2, \ldots$ denote these times[15], and assume, for a slightly richer model, that they are associated with probabilities $p_i, p_2, \ldots$ of a jump actually occurring. The random variable determining the mass of the (delta-function) spike at time t'_i is denoted $\Lambda'(t'_i)$, to which corresponds a running integral

$$I'(T; t_0) = \sum_{t'_i \in [t_0, T]} \Lambda'(t'_i) 1_{u_i < p_i} \tag{69}$$

where u_i is a sequence of independent random variables uniformly distributed on $[0, 1]$. If we assume that the ith spike mass $\Lambda'(t'_i)$ has characteristic function $\phi'_{\Lambda,i}(k)$ then

$$\phi'_J(k; t_0, T) = \mathrm{E}^{\mathbb{Q}}_{t_0}\left(e^{ik \cdot I'(s;t_0)}\right) = \prod_{t'_i \in [t_0, T]} \mathrm{E}^{\mathbb{Q}}_{t_0}\left(e^{ik \cdot \Lambda'(t'_i) 1_{u_i < p_i}}\right) = \prod_{t'_i \in [t_0, T]} \left((1 - p_i) + p_i \phi'_{\Lambda,i}(k)\right).$$

[14] The same idea can be used to enhance affine jump models.

[15] Of course, to avoid jumps effectively blending into each other, it might be most reasonable to ensure that the t'_i points are spaced further apart than the spike duration implied by the mass of the delta functions.

This characteristic function is perfectly well-behaved, unlike the limit of (65) for large values of $\tilde{\lambda}$.

In practice, we would normally combine YQ spikes with "surprise" spikes, which would lead to characteristic functions of the type

$$\phi_J(k;t_0,T) = e^{(\phi_\Lambda(k)-1)\int_{t_0}^{T}\tilde{\lambda}_J(u)\,du} \prod_{t_i'\in[t_0,T]} \left((1-p_i)+p_i\phi'_{\Lambda,i}(k)\right). \tag{70}$$

Of course, we could have multiple different YQ spike processes existing simultaneously, along with multiple regular spike processes (as in Section 6.5). We leave these straightforward extensions to the reader.

7.4 Other Point Processes

With (69) and (70), we have supplemented our unpredictable Poisson point process for spike timing, with a completely predictable process with static timing events. While the combination of both types of processes (as in (70)) should allow us to cover a wide range of practical applications, we can obviously consider introducing other types of processes as well. For instance, it is not difficult to introduce default-style models where only a single spike can take place on a given time intervals $[t_a, t_b]$. Whether such extensions are necessary remains to be seen, but they are, in any case, straightforward.

7.5 Some Implications

The results in Proposition 6 and in equations (64), (65), and (70) have been derived under the effective assumption that spikes are infinitely thin Dirac delta-functions. As result, bond price expectations (and characteristic functions) on $[t_0, T]$ are a.s. *deterministic* functions of t_0 (and T). This is in contrast to the original two-state Markov model, where these expectations would depend on $J(t_0)$, as we noted in (29). We can loosely understand the loss of the extra state-variable $J(\cdot)$ as being a consequence of the probability of $J(\cdot) \neq 0$ being negligible if spikes are sufficiently thin. Indeed, the bond price expectations in a two-state Markov model will jump by a finite amount as J jumps from 0 to some other state, but will revert back a small period later; in the limit, the jump in expectations will take place on a time-interval set with measure 0, and will therefore effectively vanish.

For the purpose of bond-type calculations, we also notice that

$$P_J(0,T) \triangleq \mathrm{E}^{\mathbb{Q}}\left(e^{-\int_0^T J(u)\,du}\right) = \mathrm{E}^{\mathbb{Q}}\left(e^{-\int_0^t J(u)\,du}\right)\cdot \mathrm{E}_t^{\mathbb{Q}}\left(e^{-\int_t^T J(u)\,du}\right)$$

which follows from the independence of $I(t;0)$ and $I(T;t)$, and from the fact that $P_J(t,T)$ as argued is a deterministic function of t and T. Thus, as one might expect,

$$P_J(t,T) = \frac{P_J(0,T)}{P_J(0,t)}, \tag{71}$$

where $P_J(0,T)$ and $P_J(0,t)$ can be computed from (65) by setting $k = i$. The absence of randomness in bond price expectations – despite profound randomness in J – is, again, a consequence of spikes being infinitesimally thin. This, of course, also means that options (such as type-I caplets) written on forward rates extracted from bond price expectations will not see additional volatility[16] from a delta-style spike model. This, however, is *not* true for type-II caplets, where the options are written directly on the exponentiated integral of the spikes, rather than on its expectation.

8 Option Pricing

The top-down spike models discussed in previous chapters are straightforward to use in applications via Monte Carlo methods. Indeed, simulation of affine dynamics (see 24) is a well-understood problem, and the dynamic simulation of spike locations, heights, and durations in the 2-state Markov chain only involve sampling from pre-defined distributions. Specifically, to simulate spikes in the Markov chain model, we sequentially:

1. Draw each up-jump time (time at which c transitions from state e_1 to state e_2) from an exponential distribution with deterministic intensity function λ_{12}.

2. Draw the jump magnitude from a specified distribution with deterministic density p_H.

3. Draw the spike duration (time at which c transitions back to state e_1) from an exponential distribution with deterministic intensity λ_{21}.

Steps 1-3 are repeated until we exceed the horizon we are interested in, after which a path of $J(t)$ has been created. The path of $r_S(t)$ is then constructed from (19). Should we at any point t need to compute bond prices $P_S(t,\cdot)$, we can rely on the numerical and/or analytical methods in Sections 5, 6, and 7.

For the models in Section 7, we can proceed in the same manner, except that the spike duration in Step 3 i) will be applied to the mass of a delta-function on $J(t)$, and ii) may be drawn from a larger class of distributions than an exponential one.

While Monte Carlo simulation is a reasonable choice for a range of exotic interest rate products, it is often an inconvenient technique for simple flow products that are traded in high volume and require computational efficiency. In the next sections, we develop high-efficiency methods for caplets.

8.1 Type-I Caplet

For simplicity, let us focus on delta-function jump modeling of the type in Section 7, and let us start by considering the pricing of a type-I caplet. Anchoring calendar time at

[16] There may, however, be some subtle level effects, as we describe in Section 8.1.

$t = 0$ for simplicity of notation, we are interested in establishing

$$\begin{aligned} V_I(0,K) &= \mathrm{E}^{\mathbb{Q}}\left(e^{-\int_0^{T_p} r(u)\,du}\left(\left(\frac{1}{P(T_s,T_p)}-1\right)-K\Delta\right)^+\right) \\ &= \mathrm{E}^{\mathbb{Q}}\left(e^{-\int_0^{T_p}(r_C(u)+J(u)-g(u))\,du}\left(\left(\frac{1}{P_C(T_s,T_p)P_S(T_s,T_p)}-1\right)-K\Delta\right)^+\right) \\ &= P_S(0,T_s)\mathrm{E}^{\mathbb{Q}}\left(e^{-\int_0^{T_p} r_C(u)\,du}\left(\left(\frac{1}{P_C(T_s,T_p)}-1\right)-\tilde{K}\Delta\right)^+\right) \end{aligned}$$

where we use the notation of Section 3.1; where

$$\tilde{K}\Delta \triangleq (1+K\Delta)\,\frac{P_S(0,T_p)}{P_S(0,T_s)} - 1;$$

and where the third equality follows from the deterministic nature of P_S (see Section 7) combined with the independence of r_C and J. Thus we have established that the price of the caplet can simply be computed as $P_S(0,T_s)$ times the price of a caplet on the continuous forward rate with modified strike $\tilde{K}$.

When working with idealized Dirac delta spikes, we thus conclude that the impact on type-I caplet prices from spikes is essentially a scaling effect on the option's price and strike. This scaling originates with the fact that the continuous model is generally no longer calibrated to the full discount curve $P(0,\cdot)$, but instead to $P_C(0,\cdot) = P(0,\cdot)/P_S(0,\cdot)$. In the specific case where the spike model is calibrated such that $P_S(0,T) = 1$ for all T, then introducing a Dirac delta spike component will *not* affect type-I caplets.

As we shall see shortly, the conclusion that caplet prices are unaffected by spikes does *not* carry over to type-II caplets. It also does not hold strictly for type-I caplets when the affine or the full 2-state Markov chain models are used, since then $J(t)$ here becomes a full stochastic state-variable of $P(t,T)$. Still, if spikes are narrow, we would expect the conclusion to hold approximately for these models, in the sense that type-I caplet prices are only weakly affected by spikes.

8.2 Type-II Caplet

We now consider a type-II caplet paying

$$V_{II}(T_p,K) = \left(\frac{\beta(T_p)}{\beta(T_s)} - 1 - K\Delta\right)^+ = \left(\frac{\beta_C(T_p)}{\beta_C(T_s)}\cdot\frac{\beta_S(T_p)}{\beta_S(T_s)} - K^*\right)^+,$$

where $K^* \triangleq 1+K\Delta$. The time t value of the option is, as in (7), $V_{II}(t,K) = P(t,T_p)\mathrm{E}_t^{T_p}(V_{II}(T_p,K))$ where expectations in the T_p-forward measure $\mathbb{Q}^{T_p}$ is characterized by (with $X(T_p)$ being some T_p-measurable random variable)

$$\mathrm{E}_t^{T_p}\left(X(T_p)\right) = P(t,T_p)^{-1}\mathrm{E}_t^{\mathbb{Q}}\left(\frac{\beta(t)}{\beta(T_p)}X(T_p)\right). \tag{72}$$

To simplify notation, let us define

$$Z_S(T_p) \triangleq \ln \beta_S(T_p) - \ln \beta_S(T_s), \quad Z_C(T_p) \triangleq \ln \beta_C(T_p) - \ln \beta_C(T_s),$$

such that, at some time $t < T_p$,

$$V_{II}(t,K) = P(t,T_p)\mathrm{E}_t^{T_p}\left(\left(e^{Z(T_p)} - K^*\right)^+\right), \quad Z(T_p) \triangleq Z_C(T_p) + Z_S(T_p). \tag{73}$$

As in Section 3.2, we define bond prices $P_x(t,T)$, $x = S, C$, such that $P(t,T_p)$ in (73) satisfies $P(t,T_p) = P_S(t,T_p)P_C(t,T_p)$. Like before, we may assume that at time t, the function $G(\cdot)$ in (21) has been calibrated such that, for any $T > t$, $P_S(t,T) = P(t,T)/P_C(t,T)$, where

$$P_S(t,T) = \mathrm{E}_t^{\mathbb{Q}}\left(\frac{\beta_S(t)}{\beta_S(T)}\right) = \mathrm{E}_t^{\mathbb{Q}}\left(\frac{\beta_J(t)}{\beta_J(T)}\right)e^{G(T)-G(t)},$$

and where $P(t,T)$ is observable at time t. $P_C(t,T)$ can be freely specified.

8.3 Type-II Caplet Pricing by Spectral Methods

With (73), we have written type-II caplet prices in a well-known form that is, for instance, conducive to Fourier methods and other schemes based on characteristic functions. For these techniques, we first need to fundamentally establish the $\mathbb{Q}^{T_p}$-measure characteristic function

$$\varphi_Z(k;t) = \mathrm{E}_t^{T_p}\left(e^{ik\cdot Z(T_p)}\right) = \mathrm{E}_t^{T_p}\left(e^{ik\cdot Z_S(T_p)}\right)\mathrm{E}_t^{T_p}\left(e^{ik\cdot Z_C(T_p)}\right) \triangleq \varphi_S(k;t)\cdot\varphi_C(k;t), \tag{74}$$

which we will proceed to consider next. Subsequently, we then show how to use the characteristic functions for efficient caplet pricing.

Assume first that we are provided with the log-characteristic function for the continuous-time rates process in "its own" forward measure $\mathbb{Q}^{T_p,C}$:

$$\tilde{\varphi}_C(k;t) = \mathrm{E}_t^{T_p,C}\left(e^{ik\cdot Z_2(T_p)}\right) \triangleq P_C(t,T_p)^{-1}\mathrm{E}_t^{\mathbb{Q}}\left(\frac{\beta_C(t)}{\beta_C(T_p)}e^{ik\cdot\left(\ln\beta_C(T_p)-\ln\beta_C(T_s)\right)}\right).$$

Appendix A shows, for instance, how to derive $\tilde{\varphi}_C(k;t)$ for the Gaussian model used earlier. With our assumptions of independence, $\tilde{\varphi}_C$ and φ_C are conveniently identical:

$$\begin{aligned}
\varphi_C(k;t) &= P(t,T_p)^{-1}\mathrm{E}_t^{\mathbb{Q}}\left(\frac{\beta(t)}{\beta(T_p)}e^{ik\cdot\left(\ln\beta_C(T_p)-\ln\beta_C(T_s)\right)}\right)\\
&= P(t,T_p)^{-1}\mathrm{E}_t^{\mathbb{Q}}\left(\frac{\beta_S(t)}{\beta_S(T_p)}\right)\mathrm{E}_t^{\mathbb{Q}}\left(\frac{\beta_C(t)}{\beta_C(T_p)}e^{ik\cdot(\ln\beta_C(T)-\ln\beta_C(T-\Delta))}\right)\\
&= \frac{P_S(t,T_p)}{P(t,T_p)}P_C(t,T_p)\tilde{\varphi}_C(k;t) = \tilde{\varphi}_C(k;t).
\end{aligned} \tag{75}$$

It remains for us to establish the log-characteristic function $\varphi_S(k;t)$. For this, we recall that we already defined (in Section 3.3) $\phi_S(k;t,T)$ as

$$\phi_S(k;t,T) = \mathrm{E}_t^{\mathbb{Q}}\left(e^{ik\cdot(\ln\beta_S(T)-\ln\beta_S(t))}\right) = \phi_J(k;t,T)e^{-ik\cdot(G(T)-G(t))}$$

where G was defined in (21) and a variety of results for $\phi_J(k;t,T)$ were derived in previous sections. For instance, $\phi_J(k;t,T)$ was given outright in (65) as part of the spike model specification. Fortunately, we can often relate φ_S to ϕ_S (or ϕ_J), either exactly or approximately. For instance, we have:

Lemma 6 *When spikes are Dirac delta-functions, as for the models in Section 7, we have*

$$\varphi_S(k;t) = \begin{cases} \frac{P_S(t,T_s)}{P_S(t,T_p)}\phi_S(k+i;T_s,T_p), & t\in[0,T_s], \\ \frac{(\beta_S(t)/\beta_S(T_s))^{ik}}{P_S(t,T_p)}\phi_S(k+i;t,T_p), & t\in(T_s,T_p]. \end{cases} \tag{76}$$

Remark 4 *Since we can compute $P_S(t,T_s) = \phi_S(i;t,T_s)$ and $P_S(t,T_p) = \phi_S(i;t,T_p)$, it follows that φ_S is completely determined by ϕ_S.*

Proof: Assume first that $t\in[0,T_s]$, such that

$$\begin{aligned} \varphi_S(k;t) &= P(t,T_p)^{-1}\mathrm{E}_t^{\mathbb{Q}}\left(\frac{\beta(t)}{\beta(T_p)}e^{ik\cdot\left(\ln\beta_S(T_p)-\ln\beta_S(T_s)\right)}\right) \\ &= \frac{P_C(t,T_p)}{P(t,T_p)}\mathrm{E}_t^{\mathbb{Q}}\left(e^{\ln\beta_S(t)-\ln\beta_S(T_p)}e^{ik\cdot\left(\ln\beta_S(T_p)-\ln\beta_S(T_s)\right)}\right) \\ &= \frac{P_C(t,T_p)}{P(t,T_p)}\mathrm{E}_t^{\mathbb{Q}}\left(e^{\ln\beta_S(t)-\ln\beta_S(T_s)}e^{(ik-1)\cdot\left(\ln\beta_S(T_p)-\ln\beta_S(T_s)\right)}\right) \\ &= P_S(t,T_p)^{-1}\mathrm{E}_t^{\mathbb{Q}}\left(e^{\ln\beta_S(t)-\ln\beta_S(T_s)}\right)\mathrm{E}_t^{\mathbb{Q}}\left(\mathrm{E}_{T_s}^{\mathbb{Q}}\left(e^{(ik-1)\cdot\left(\ln\beta_S(T_p)-\ln\beta_S(T_s)\right)}\right)\right) \\ &= \frac{P_S(t,T_s)}{P_S(t,T_p)}\phi_S(k+i;T_s,T_p), \end{aligned} \tag{77}$$

where the last step follows from the fact that $\phi_S(k+i;T_s,T_p)$ is deterministic (see Section 7.5).

Next, assume that $t\in(T_s,T_p]$, in which case

$$\begin{aligned} \varphi_S(k;t) &= P(t,T_p)^{-1}\mathrm{E}_t^{\mathbb{Q}}\left(\frac{\beta(t)}{\beta(T_p)}e^{ik\cdot\left(\ln\beta_S(T_p)-\ln\beta_S(T_s)\right)}\right) \\ &= \frac{P_C(t,T_p)}{P(t,T_p)}\mathrm{E}_t^{\mathbb{Q}}\left(e^{ik\cdot(\ln\beta_S(t)-\ln\beta_S(T_s))}e^{(ik-1)\cdot\left(\ln\beta_S(T_p)-\ln\beta_S(t)\right)}\right) \\ &= \frac{e^{ik\cdot(\ln\beta_S(t)-\ln\beta_S(T_s))}}{P_S(t,T_p)}\phi_S(k+i;t,T_p), \end{aligned}$$

which concludes the proof. ∎

It follows from the proof that the result in Lemma 6 holds in all generality for $t \in (T_s, T_p]$, but *not* for $t \leq T_S$. Indeed, for both the affine and full 2-state Markov chain, step (77) in the proof will not hold as the T_s-expectation becomes a function of $J(T_s)$. Of course, if spikes are thin, Lemma 6 may nevertheless still be used as a good approximation for the full 2-dimensional Markov chain model; see Section 9.2 for some numerical results.

Below, we list the exact results for $\varphi_S(k;t)$ when $t \leq T_s$, for both the affine model in Section 4 and the Markov chain model in Section 5. For both results, we only list expressions for φ_J as we can then easily recover φ_S from the relation

$$\varphi_S(k;t) = \varphi_J(k;t)e^{-ik(G(T_p)-G(T_s))} \tag{78}$$

where G was defined in (21).

Proposition 7 *Consider the affine model of Section 4, and let notation be as in Proposition 1. For $t \in (T_s, T_p]$, the result in Lemma 6 still holds. For $t \in [0, T_s]$ we have*

$$\varphi_J(k;t) = \frac{e^{A_J(T_s,T_p;k+i)}}{P_J(t,T_p)} e^{\mathcal{A}((ik-1)B_J(T_s,T_p);t,T_s)-J(t)\cdot\mathcal{B}((ik-1)B_J(T_s,T_p);t,T_s)}, \quad t \in [0,T_s],$$

with (x being any complex number)

$$\begin{aligned}\mathcal{B}(x;t,T) &= B_J(t,T) - xe^{-\kappa_J(T-t)},\\ \mathcal{A}(x;t,T) &= \int_t^T \lambda_J(u)\left(1-\phi_H\left(i\mathcal{B}(x;u,T),u\right)\right)du.\end{aligned}$$

Proof: See Appendix B. ∎

Proposition 8 *We adopt the notations from Section 5 for the Markov chain model, and assume that $J(t) = 0$. For $t \in (T_s, T_p]$, the result in Lemma 6 still holds. For $t \in [0, T_s]$ we have*

$$\varphi_J(k;t) = \frac{\omega(k+i;t,T_s,T_p)}{P_J(t,T_p)}, \quad t \in [0,T_s], \tag{79}$$

where

$$\begin{aligned}\omega(k;t,T_s,T_p) &\triangleq \mathrm{E}_t^{\mathbb{Q}}\left(e^{-\int_t^{T_s} J(u)\,du}\phi_J(k;T_s,T_p,J(T_s))|J(t)=0\right)\\ &= z_{AD}^{11}(i;t,T_s)\cdot\phi_J^1(k;T_s,T_p) + \int_{-\infty}^{\infty} z_{AD}^{12}(i;t,T_s,x)\cdot\phi_J^2(k;T_s,T_p,x)\,dx.\end{aligned} \tag{80}$$

Remark 5 *In Proposition 8, see (41)-(42) and Lemma 3 for definitions and results for ϕ_J^1 and ϕ_J^2; and see (51)-(52) and Proposition 5 for z_{AD}^{11} and z_{AD}^{12}. Since ϕ_J^2 can be written as an integral equation involving ϕ_J^1 (see (45)), we may also rewrite Proposition 8 to involve only ϕ_J^1.*

Remark 6 *For brevity, Proposition 8 only considers the case where $J(t) = 0$. Results for $J(t) \neq 0$ are easily derived using similar methods.*

Proof: See Appendix B. ∎

The numerical implementation of the integro-differential equations in Proposition 8 are reasonably involved, so the simpler approximations that can be extracted from Lemma 6 are often quite convenient in practice. Of course, tractability for z_{AD}^{11}, z_{AD}^{12} and ϕ_J^2 will benefit from assumptions of discreteness and time-homogeneity for the random variable H, in a similar fashion as for ϕ_J^1 (see Proposition 4). This will allow one to re-cast integro-differential equations into standard ODEs, easily solvable by standard means.

8.3.1 Type-II Caplet Pricing by Fourier Integration

In Section 8.3 above, we described how to compute the characteristic function $\varphi_Z(k;t)$ for the variable Z in the option payout expectation (73). A well-established body of results then allows us to compute the option value by (damped) Fourier integration. The following Proposition summarizes these results, using notation adapted from [3].

Proposition 9 *Define*

$$F^*(t) \triangleq \mathrm{E}_t^{T_p}\left(e^{Z(T_p)}\right) = \varphi_Z(-i;t), \quad \varphi_Z^*(k;t) \triangleq \varphi_Z(k;t)e^{-ik\cdot\ln F^*(t)}$$

and set $\omega^* \triangleq \ln(F^*(t)/K^*)$, $K^* = 1 + K\Delta$. Then, for an arbitrary damping constant $\alpha \in (-1, 0)$,

$$\frac{V_{II}(t,K)}{P(t,T_p)} = F^*(t) - \frac{F^*(t)e^{\alpha\omega^*}}{\pi}\int_0^\infty \mathrm{Re}\left\{e^{i\omega^* x}\, Q(x - i\alpha)\right\} dx, \quad Q(z) \triangleq \frac{\varphi_Z^*(z-i;t)}{z(z-i)}. \quad (81)$$

Remark 7 *It is possible to extend the range of α beyond the interval $(-1, 0)$, although care must be taken to i) include residue terms from the poles of Q; and ii) not venture beyond the singularities of φ_Z^*. See [14] for details.*

Evaluation of the indefinite integral in (81) can sometimes be challenging, e.g. for large values of ω^* where the Fourier term $e^{i\omega^* x}$ in the integrand may oscillate rapidly. In [3] it is discussed how to address this, and other challenges, from a computational efficiency standpoint. Some of the convergence remedies involve careful choice of α, combined with a deformation of the integration contour in the complex plane. These subtle issues will not be discussed here or in our numerical tests in Section 9, but will benefit an industrial-level implementation.

8.3.2 *Type-II Caplet Pricing by Quadrature*

The technique in Section 8.3.1 requires knowledge of the T_p-measure characteristic function φ_C for the continuous process. As we saw in Appendix A, φ_C is directly available for Gaussian models and, more broadly, for the affine bottom-up models of the type considered in, e.g., [6]. However, in a number of practically important top-down models (see Section 2.3.1), φ_C might not be readily available, even when closed-form caplet pricing expressions (or approximations) are. Rather than attempting to back out φ_C from caplet prices, it would be more convenient if we could use the continuous-process caplet pricing expressions directly to establish $V_{II}(t,K)$ through a "mixing" integral against a spike kernel. To state a result along these lines, let us define the continuous-model caplet price as

$$V_{II}^C(t,K) = P_C(t,T_p)\mathrm{E}_t^{T_p}\left(\left(e^{Z_C(T_p)} - K^*\right)^+\right), \quad K^* = 1 + K\Delta,$$

a quantity that we, as mentioned, now assume is known in closed-form for any value of K (or K^*). We notice that the full caplet price (taking into account also spikes) is then, by the property of conditional expectations,

$$\begin{aligned}\frac{V_{II}(t,K)}{P(t,T_p)} &= \mathrm{E}_t^{T_p}\left(e^{Z_S(T_p)}\left(e^{Z_C(T_p)} - \frac{K^*}{e^{Z_S(T_p)}}\right)^+\right)\\ &= \mathrm{E}_t^{T_p}\left(e^{Z_S(T_p)}\frac{V_{II}^C\left(t,(e^{-Z_S(T_p)}K^* - 1)/\Delta\right)}{P_C(t,T_p)}\right)\\ &\triangleq \mathrm{E}_t^{T_p}\left(e^{Z_S(T_p)}\Theta\left(e^{-Z_S(T_p)}\right)\right)\end{aligned} \tag{82}$$

where the function Θ by assumption is known in closed-form (since $V_{II}^C(t,\cdot)$ is).

With (82), we have written the option price as a problem of computing the expectation of a one-dimensional function of $Z_S(T_p)$ (or $e^{Z_S(T_p)}$), given knowledge of the characteristic function of $Z_S(T_p)$. Several computational approaches are conceivable here, ranging from naive integration over the density of $Z_S(T_p)$ (which would require inversion of the characteristic function φ_S, *a la* [7]) to moment-based polynomial expansions and quadrature methods. Density integration is unlikely to be computationally efficient in general, so we focus on the latter two methods, giving results below that can form the basis for efficient computational implementation.

Proposition 10 *Assume that*

$$\Theta(\xi) \approx \sum_{n=0}^{N} \gamma_n \xi^n \tag{83}$$

then

$$V_{II}(t,K) \approx P(t,T_p)\sum_{n=0}^{N}\gamma_n\varphi_S\left(i(n-1);t\right). \tag{84}$$

Proof: By definition of the characteristic function φ_S, we have

$$\mathrm{E}_t^{T_p}\left(\left(e^{-Z(T)}\right)^n\right) = \mathrm{E}_t^{T_p}\left(e^{-nZ(T)}\right) = \varphi_S(i\cdot n; t)$$

and the result follows from noting that

$$\frac{V_{II}(t,K)}{P(t,T_p)} \approx \mathrm{E}_t^{T_p}\left(e^{Z_S(T_p)}\sum_{n=0}^{N}\gamma_n e^{-nZ_S(T_p)}\right) = \sum_{n=0}^{N}\gamma_n \mathrm{E}_t^{T_p}\left(e^{(1-n)Z_S(T_p)}\right).$$

∎

To establish the expansion (83), we can, for instance, rely on Chebyshev interpolation, in which case the resulting sum[17] (84) amounts to N-point *Clenshaw-Curtis quadrature*.

One possible drawback of the result in Proposition 10 is that the interpolation weights γ_n must be computed whenever Θ changes. An alternative is to work out a quadrature rule that is independent of Θ, but will integrate any polynomial up to a certain degree exactly. We are then in the realm of *Gaussian quadratures*, see [18] for a readable introduction and for the steps needed to establish a new quadrature rule. The following results fill in these steps for our application.

First we introduce a scalar product, which shall form the base for the quadrature method:

Definition 1 (Scalar product) *Let $\rho_S(z_S)$ be the $\mathbb{Q}^{T_p}$-density of $Z_S(T_p)$, and define a measure $d\mu(z_S) = e^{z_S}\rho_S(z_S)\,dz_S$. Let also $\varsigma > 0$ be a constant and introduce a scalar product on $\mathbb{R}[X]\times\mathbb{R}[X]$ via:*

$$\langle U, V\rangle_\varsigma \triangleq \int_{-\infty}^{+\infty} U(e^{-\varsigma z_S})V(e^{-\varsigma z_S})\,d\mu(z_S) = \sum_{m=0}^{+\infty}\varphi_S(i(\varsigma m - 1); t)\sum_{p+q=m} U_p V_q, \tag{85}$$

where $U(z) = \sum_{m\geq 0} U_m z^m$ and $V(z) = \sum_{m\geq 0} V_m z^m$ are polynomials. Further, denote $(q_n)_{n\geq 0}$ the $\langle,\rangle_\varsigma$−orthogonal family of monic polynomials, with $\deg(q_n) = n$.

The polynomial q_n can, for instance, be constructed using the so-called *three-term recurrence relation* (see [18]), valid for $n \geq 0$,

$$q_{n+1}(x) = (x-\alpha_n)q_n(x) - \beta_n q_{n-1}(x),$$

where $q_{-1}(x) = 0$, $q_0(x) = 1$. The coefficients α_n and β_n are given as

$$\alpha_n = \frac{\langle q_n, xq_n\rangle_\varsigma}{\langle q_n, q_n\rangle_\varsigma}, \quad \beta_n = \frac{\langle q_n, q_n\rangle_\varsigma}{\langle q_{n-1}, q_{n-1}\rangle_\varsigma},$$

which may be computed directly from (85).

Equipped with a scalar product and associated orthogonal polynomials, we can draw directly from standard results for Gaussian quadrature to arrive at the following result:

[17] In a performant scheme, the sum would be re-arranged as a sum over Chebyshev polynomials.

Proposition 11 *Let $\varsigma > 0$ and $N \geq 1$ be given. Equation (82) can be approximated to effective order $\varsigma(2N-1)$ by:*

$$\frac{V_{II}(t,K)}{P(t,T_p)} \approx \sum_{j=1}^{N} w_j \Theta(\xi_j^{\frac{1}{\varsigma}}) \tag{86}$$

where the nodes $(\xi_j)_{1\leq j\leq N}$ are the N real and isolated roots[18] of polynomial q_N in Definition 1, and the weights $(w_j)_{1\leq j\leq N}$ satisfy the Vandermonde linear system

$$\sum_{j=1}^{N} w_j \xi_j^m = \varphi_S(i(\varsigma m - 1); t), \quad m = 0, \cdots, N-1.$$

Proof: (**sketch**) From (82), we have:

$$\frac{V_{II}(t,K)}{P(t,T_p)} = \int_{-\infty}^{+\infty} \Theta\left((e^{-\varsigma z_S})^{\frac{1}{\varsigma}}\right) d\mu(z_S) \triangleq \int_{-\infty}^{+\infty} \Theta_\varsigma\left(e^{-\varsigma z_S}\right) d\mu(z_S).$$

The theory of Gaussian quadratures states that

$$\int_{-\infty}^{+\infty} \Theta_\varsigma\left(e^{-\varsigma z_S}\right) d\mu(z_S) \approx \sum_{j=1}^{N} w_j \Theta_\varsigma(\xi_j) = \sum_{j=1}^{N} w_j \Theta(\xi_j^{\frac{1}{\varsigma}}).$$

where weights and nodes are as stated in the proposition, which is (86).

Moreover, the approximation is exact whenever Θ_ς is replaced by z^m for $0 \leq m \leq 2N-1$, yielding an exact value for $\int_{-\infty}^{+\infty} e^{-xz_S}\, d\mu(z_S)$ for any $x \in \{0, \varsigma, \cdots, \varsigma(2N-1)\}$. Thus, if ς is a well chosen integer greater than 1, e.g. 5, the approach will be exact at any multiple of 5 less than $5(2N-1)$ and, in practice, very accurate at any intermediate points. This is what we observe in practice and is a consequence of the smoothness of the characteristic function of the measure μ combined with the exactness of the method at bracketing points. ∎

Designing, sizing, and documenting a truly performant implementation of the quadrature scheme above is a computational problem beyond the scope of the current paper, and is left for future work. We can, however, still offer a few remarks. First, the three-term recurrence relation and the Vandermonde system may be formulated as an eigenvalues/eigenvectors problem for improved numerical stability, especially for a large number of nodes (N); see [8] for details. Second, the choice of ς plays an important role in the trade-off between the accuracy of the method and the number of nodes (N) used. For instance, for the examples we consider in Section 9, a little fine-tuning of ς allowed us to achieve excellent accuracy with N as low as 3 to 10.

[18] As the roots of q_N bracket those of q_{N-1} (and so forth), these roots can be found iteratively by a secant method or similar.

9 Some Numerical Results for Type-II Caplets

9.1 Basic Setup

For our numerical tests, we focus on type-II caplets, and throughout will assume that the continuous forward curve at the time t_0 of pricing is flat at 1%, such that

$$P_C(t_0, T) = e^{-0.01\cdot(T-t_0)}, \quad T \geq t_0.$$

In addition, whenever $t_0 > T$ (i.e., we need past rate sets, as when $t_0 \in (T_s, T_p)$), we will assume that

$$\beta(t_0)/\beta(T) = e^{-0.01\cdot(T-t_0)}, \quad T \leq t_0. \tag{87}$$

In our first three tests (Sections 9.2-9.4), the continuous time rate dynamics originate from a Gaussian short-rate model (see Appendix A for details), with parameters $\sigma_r = 0.005$ and $\kappa = 0.02$. As we add a variety of spike types to these basic continuous-time dynamics, the total rates model will, of course, have a non-Gaussian distribution and will return different type-II caplet prices than will the pure continuous model. To capture distribution changes from spikes in a succinct manner, we define an *implied basis point volatility* σ_r^{imp} as the value of σ_r that we must use in a purely continuous Gaussian model (keeping $\kappa = 0.02$) to reproduce a given caplet price. By its definition, σ_r^{imp} will then depend on the caplet characteristics, i.e. the strike and maturity; we will refer to the graph of σ_r^{imp} against strike, for any fixed caplet maturity, as the (basis point) *volatility smile*.

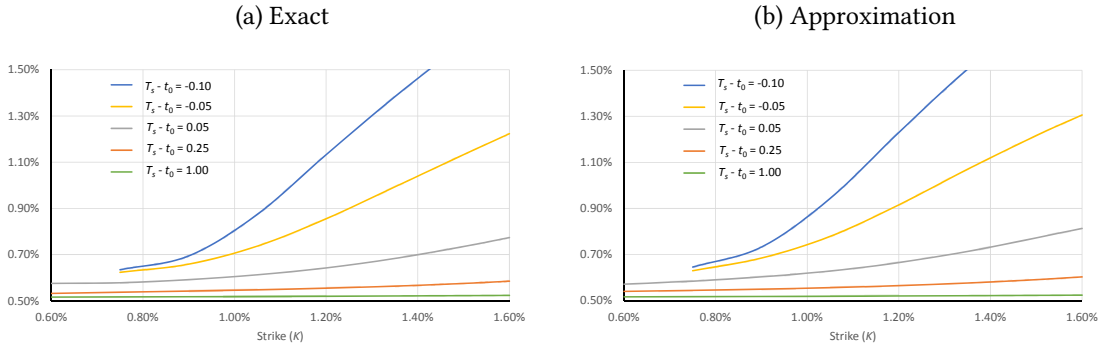
Our fourth, and final, test is for a spike-enhanced SABR model. We will again rely on the concept of an implied volatility smile to gauge the impact of spikes, but we will use a more conventional top-down definition of the smile (see Section 9.5 for the details).

9.2 Test 1: Compound Markov Chain with Exponential Spikes

In our first test, we use the Markov chain model with exponential jumps, parameterized as in Section 6.4 (so: $\lambda = \lambda_{21} = 100$, $\lambda_{12} = 15$). We assume that $P_S(t_0, T) = 1$ for all T, which fixes the calibration function $g(\cdot)$ in Section 3.2 accordingly. We assume that $\Delta = T_p - T_s = 0.25$ (which is the most common configuration in the US), and in Figure 6 graph the time t_0 value of σ_r^{imp} as a function of $T_p - t_0$ and the caplet strike K. We show both the true values of σ_r^{imp}, as well as those computed by applying the Fourier technique in Section 8.3.1 to the characteristic function approximation from Section 6 (see Proposition 6).

As we can see, the approximations in Section 6 generate slightly more extreme skews than the true Markov chain, but overall work well. Both panels a) and b) in Figure 6 show that the inclusion of spikes result in a pronounced upwards skew of σ_r^{imp} when graphed against caplet strike. For the parameters used in the figure, this skew effect is modest for $T_p - t_0$ beyond, say, a year or so, but then increases very significantly as t_0 gets closer to the caplet maturity. This temporal behavior is not unexpected: compared to continuous

Figure 6: σ_r^{imp} for Markov Chain Model with Exponential Jumps



Time t_0 implied basis point volatility for Markov chain spike model, as function of strike K and $T_s - t_0$; negative values of $T_s - t_0$ signify that the pricing time is inside the rate observation period $[T_s, T_p]$, where $\Delta = T_p - T_s = 0.25$. The continuous dynamics are Gaussian, as described in Section 9.1; and the Markov chain is configured as in Figure 5. The "Exact" numbers in the graphs were computed by Monte Carlo simulation; the "Approximation" numbers were computed by Proposition 9, after deployment of the approximations from Proposition 6 (and aided by Lemma 4 and Remark 2).

models, jump-type dynamics tend to produce very significant volatility skews for short maturities, with the skew decaying away fairly quickly as maturity is increased.

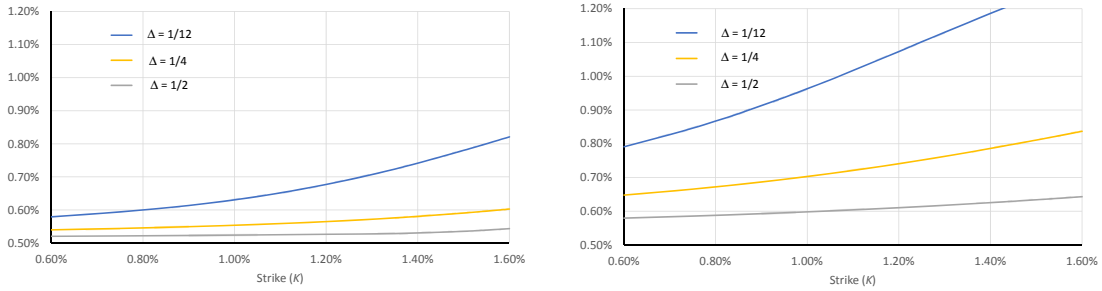
To offer some intuition for the shape of the graphs in Figure 6, consider that the calibration condition $P_S(t_0, T) = 1$ effectively implies that the spike model will add two types of scenarios to the distribution of the caplet forward rate: i) a small negative offset (from the calibration function g); and ii) a low-probability – but more significant – positive increase (from random, but always positive, spikes). Together, these effects add volatility to the distribution of rates, but also create kurtosis with a fatter upper tail. On a relative basis, the kurtosis effect will be largest when maturity is short and the likelihood of significant continuous moves in the rate on $[T_s, T_p]$ is low. Since σ_r^{imp} increasing in strike is basically an indicator of a fat upper tail, the intuition behind Figure 6 should now be clear.

To further ponder what might affect the materiality of the volatility skew in Figure 6, it is clear that increasing λ_{12} should increase the size of the skew, and increasing[19] λ and λ_{21} should decrease it. Beyond these obvious effects, one might expect that the influence of spikes on volatility would grow in significance for tight intervals $[T_s, T_p]$ where the averaging effects discussed in Section 1 are less pronounced. Figure 7 contains illustrative examples.

[19] Recall that the expected jump magnitude is $1/\lambda$.

Figure 7: σ_r^{imp} for Markov Chain Model with Exponential Jumps

(a) $\lambda = \lambda_{21} = 100$, $\lambda_{12} = 15$ (b) $\lambda = \lambda_{21} = 75$, $\lambda_{12} = 25$



Time t_0 implied basis point volatility for Markov chain spike model, as function of strike K and Δ. In all graphs $T_s - t_0 = 0.25$. The continuous dynamics are Gaussian, as described in Section 9.1; and the Markov chain is configured as indicated above the graphs. Option prices were computed by Proposition 9, after deployment of the approximations from Proposition 6 (and aided by Lemma 4 and Remark 2).

9.3 Test 2: Gaussian Spikes with Uniformly Distributed Duration

As we demonstrated above, the usage of strictly positive, exponentially distributed jump heights will add an upward skew to the implied volatility smile. U-shaped smile effects from spikes are, however, also possible, and can arise in several ways. For instance, if we allow for both positive and negative spikes, we would expect fat tails in both the upper and lower parts of the rate density, which in turn would produce more of a U-shape.

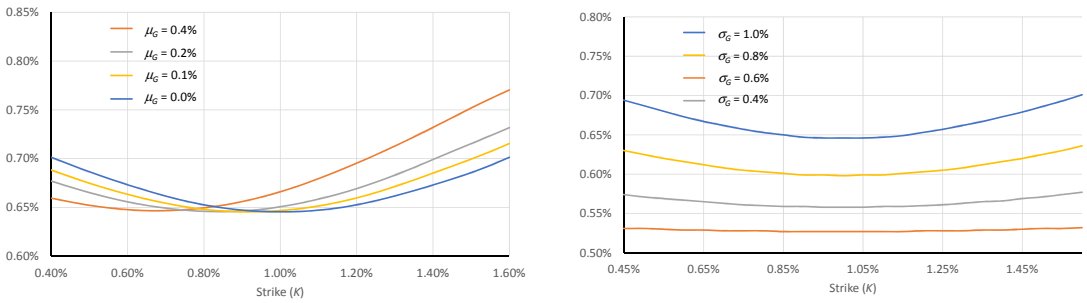
Adding negative jumps to the framework in Section 9.2 could be done, e.g., by using the principles of Section 6.5, by including another independent spike process, but one where the jump heights are set equal to the *negative* of an exponential distribution. For variation, we will here instead use Gaussian spike heights, but now we associate it with a spike duration uniformly distributed on the interval [1, 10 days], as in[20] (66). We are thus in the delta-spike framework, and will use an intensity of[21] $\tilde{\lambda} = 13.4$ to model the spike occurrence, with the $g(\cdot)$ function again calibrated to ensure that $P_S(t_0, T) = 1$ for all T.

Figure 8 confirms that the model setup described above is now able to produce U-shaped volatility smiles. The curvature and low-point location of the smile depend in an obvious way on the volatility (σ_G) and mean (μ_G) of the Gaussian spike height distribution. Notice that we, for variation, used the quadrature scheme of Proposition 11 to compute option prices.

[20] We computed the required integral by a simple numerical integration scheme.
[21] This is the same value as is used in the approximation in panel b) of Figure 6.

Figure 8: σ_r^{imp} for Delta-Style Spikes with Gaussian Height and Uniform Width

(a) $\sigma_G = 0.01$ (b) $\mu_G = 0$



Time t_0 implied basis point volatility for delta-style spike model with Gaussian spike height and uniformly distributed spike width. In all graphs $T_s - t_0 = 0$, and $\Delta = 0.25$. The continuous dynamics are Gaussian, as described in Section 9.1, and spikes are modeled as Dirac delta-functions, as in Section 7. The mean and volatility parameters of the Gaussian height distribution (μ_G and σ_G) are indicated on the graphs; the spike duration is uniformly distributed on [1, 10 days]. Option prices were computed by Proposition 11.

9.4 Test 3: Exponential Year-end Spikes of Constant Duration

In Sections 9.2 and 9.3, we focused on surprise-type spikes. Now, we use the results of Section 7.3 to turn our attention to a model that contains a single year-end (YE) spike. In practice, we would, of course, always combine a multiple of such models together to capture all spikes with (near-) certain location; we would also most likely add a surprise-style spike model, like those considered in our previous test cases above.

For simplicity (and variety) we consider an exponential distribution with parameter λ for the jump height H, and a *constant* jump duration of D business days; this is a special case of (67). In that case, the characteristic function of $D \cdot H$ is simply:

$$\phi_\Lambda(k) = \phi_H(Dk) = \frac{1}{1 - i\frac{Dk}{\lambda}} \tag{88}$$

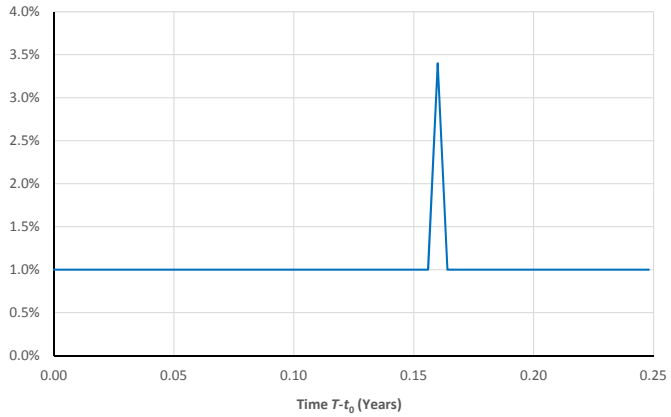
Applying (70) with a jump probability p, we get:

$$\phi_J(k; t_0, T) = (1 - p) + p\phi_\Lambda(k) = (1 - p) + \frac{p}{1 - i\frac{Dk}{\lambda}}. \tag{89}$$

As in previous examples, we use a 1% forward rate for the continuous discount function $P_C(t_0, T)$. This time, however, we do *not* force $P_S(t_0, T)$ to 1 for all T, but simply set the calibration function g to 0 everywhere. As a result, $P_S(t_0, T)$ will be strictly less than 1 in the narrow window where the YE spike exists, given rise to a spike in the short-dated forward curve as seen at time t_0; see Figure 9. We can think of these spikes in the short-dated forward curve as the time t_0 expectation of a future stochastic spike in spot rates.

Narrow spikes in the forward curve around quarter- and year-end dates are a hallmark of real SOFR forward curves, but are in practice not associated with any real stochasticity – instead, the spikes are essentially just deterministic "hills" of fixed magnitude that a continuous process climbs over. As we have discussed, there is in practice considerable uncertainty about the size and duration of these hills, so a rational model would associate them with a stochastic spike process, as we do here.

Figure 9: Overnight (1-day) Forward Rate in YE Spike Model



The overnight forward rate $f(t_0, T, T + 1\text{ bday})$ for YE Spike Model. The YE spike is modeled as a delta-spike, located 60 days from t_0. Its effective width is $D = 2$ business days, its probability of taking place is $p = 0.8$, and its height is exponentially distributed with parameter $\lambda = 100$. The calibration function g was set to zero.

As in previous tests, the impact of stochasticity in our YE spike on implied volatility smiles[22] will depend on time to maturity, spike height and width distributions, and so forth. In Figure 10, we show in particular how increasing spike duration D causes the spike impact (here: an upward tilt in the skew) to increase.

9.5 Test 4: SABR model with Exponential Spikes

A common specification of the SABR model, adapted to the backward-looking compounding convention, would write, top-down style,

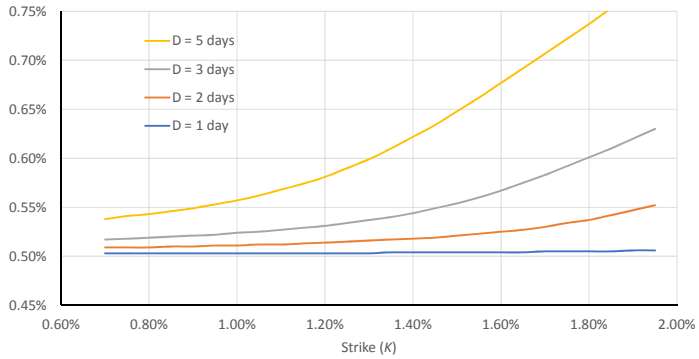
$$dF(t; T_s, T_p) = w(t; T_s, T_p)\sigma(t)\left(\gamma + F(t; T_s, T_p)\right)^{\beta} dW^{T_p}(t), \tag{90}$$

$$d\sigma(t) = \sigma(t)\alpha\, dZ(t), \quad \sigma(t_0) = \sigma_0, \tag{91}$$

where W^{T_p} and Z are Brownian motions in the T_p-forward measure, satisfying $dZ(t) \cdot dW^{T_p}(t) = \rho\, dt$. Relative to the standard SABR model of [10], there are two changes to

[22] When computing implied volatilities here, it is important that the spike-free continuous Gaussian model used when solving for σ_r^{imp} is calibrated to the total yield curve $P(t_0, T) = P_C(t_0, T)P_S(t_0, T)$. As we explained (see Figure 9), for our test case here it is no longer the case that $P(t_0, T) = P_C(t_0, T)$.

Figure 10: σ_r^{imp} for Exponential YE Spike and Constant Width



Time t_0 implied basis point volatility for YE spike model, as function of caplet strike K and spike duration D. In all graphs $T_s - t_0 = 0$, and $\Delta = 0.25$. The continuous dynamics are Gaussian, as described in Section 9.1; and the YE spike model was configured as in Figure 9. Option prices were computed by Proposition 11, using (89).

these dynamics: a) we follow the ideas of Section 2.3.1 to use a weight function $w(t; T_s, T_p)$ to handle the decay of volatility on $[T_s, T_p]$; and b) we have introduced a shift parameter $\gamma > 0$ to allow the forward to become negative, as is normally needed in recent interest rate markets. As we discussed in Section 2.3.1, there are various parameter averaging techniques available to us that can translate (90)-(91) into a time-homogenous system with (approximately) unchanged distribution of $F(T_p; T_s, T_p)$. Let the deployment of such techniques approximate (90)-(91) with

$$dF(t; T_s, T_p) = \sigma(t) \left(\gamma^* + F(t; T_s, T_p)\right)^{\beta^*} dW^{T_p}(t), \tag{92}$$

$$d\sigma(t) = \sigma(t)\alpha^* dZ(t), \quad \sigma(t_0) = \sigma_0^*, \tag{93}$$

with $dW^{T_p} \cdot dZ(t) = \rho^* dt$. Now, $\sigma_0^*, \alpha^*, \beta^*, \rho^*, \gamma^*$ are new "effective" parameters that encapsulate the effect of $w(t)$. For our numerical tests, we simply assume that these effective parameters are given, found either by outright calibration to type-II caplets, or by a parameter averaging technique deployed to parameters extracted from type-I caplets.

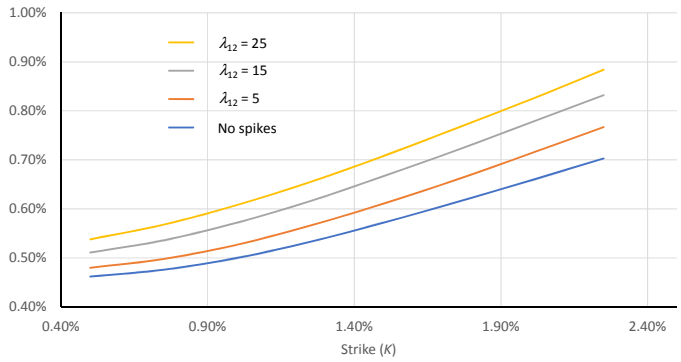
To add spikes to the above SABR model, we will use the Markov chain specification from Section 9.2, approximated as in Panel b) of Figure 6 and then use the quadrature method in Proposition 11 to combine it with the well-known option pricing formulas for the SDEs (92)-(93). As in Sections 9.2 and 9.2, we calibrate the time t_0 discount function $P_S(t_0, T)$ to equal 1 for all $T \geq t_0$. For the purpose of stating the volatility smile for the combined model, we change our convention slightly from the previous sections, and now define the implied SABR volatility smile by solving for the value of σ_F^{imp} that makes

the model

$$dF(t;T_s,T_p) = \sigma_F^{imp}\, dW^{T_p}(t)$$

produce the same option price as the full spike-enable SABR model. Results are in Figure 11.

Figure 11: σ_F^{imp} for SABR Model + Markov Chain with Exponential Spikes



Time t_0 implied basis point volatility σ_F^{imp} for delta-style spike model added to a SABR model. The basic SABR model parameters were: $\alpha^* = 50\%$, $\beta^* = 1/2$, $\rho^* = 0.3$, $\gamma^* = 2\%$, and $\sigma_0^* = 2.87\%$. The spike model was as in Figure 6, but with λ_{12} varying as indicated on the graph. We assumed $\Delta = 0.25$, $F(t_0;T_s,T_p) = 1\%$, and $T_s - t_0 = 0.0$. Option prices were computed by Proposition 11, using the same spike process approximation as was used to create panel b) of Figure 6.

10 Conclusion

In this paper, we introduced a comprehensive framework for the inclusion of random spikes in dynamic interest rate models. The framework allows for incorporation into pricing and risk management the types of liquidity phenomena that routinely affect financing rates, such as repo rates and the new SOFR benchmark rate. Covering both spikes with predictable timing as well as those that are "surprises", we gave technical and computational results sufficient for model calibration, bottom-up simulation, and the efficient pricing of vanilla derivatives by spectral methods. We demonstrated numerically and theoretically how the inclusion of spikes can materially affect the implied volatility smiles for those contracts, such as SOFR-style caplets, that settle on a backward-looking compounded rate.

As discussed in detail, there are considerable uncertainties still prevailing in the market for SOFR derivatives, so details around, say, the most suitable specialization of our framework for accurate and parsimonious SOFR market calibration can only be exhaustively examined once the market gains liquidity, and a variety of institutional and regulatory questions get settled. We obviously must leave this for future work. Also left for

future work is a more detailed exposition and examination of the numerical methods used in the paper; in the interest of time and space, we kept our description of these matters at a fairly high level, leaving aside details needed for a robust implementation that can deal with stressed parameters. Some of the principles of the numerical schemes we developed (e.g., the quadrature methods) appear to be a useful tool for the broader objective of efficiently adding new process components (not just spikes) to models with known closed-form option pricing. This topic is also left for future research.

A Appendix: Gaussian One-Factor Model

A.1 Forward Rate Process

If rates follow (10), we know that (see, e.g., [2]), in the T-forward measure,

$$dP(t,T)/P(t,T) = O(dt) - \sigma_r B(t,T)\, dW^T(t),$$

where $W^T(t)$ is a $\mathbb{Q}^T$-Brownian motion, and $B(t,T)$ is given in (12). Therefore, when $t \in [0, T_p]$ we can use the martingale property, Ito's lemma, and (3) to see that (with $\Delta = T_p - T_s$)

$$\begin{aligned} dF(t;T_s,T_p) &= \Delta^{-1} \cdot d\left(\frac{P(t,T_s)}{P(t,T_p)}\right) = \Delta^{-1} \cdot \frac{P(t,T_s)}{P(t,T_p)} \sigma_r \left(B(t,T_p) - B(t,T_s)\right)\, dW^{T_p}(t) \\ &= \left(\Delta^{-1} + F(t;T_s,T_p)\right) \sigma_r \left(B(t,T_p) - B(t,T_s)\right)\, dW^{T_p}(t), \quad t \in [0,T_s]. \end{aligned} \tag{94}$$

Similarly, for $t \in (T_s, T_p]$,

$$\begin{aligned} dF(t;T_s,T_p) &= \Delta^{-1} \cdot d\left(\frac{\beta(t)/\beta(T_s)}{P(t,T_p)}\right) = \frac{1}{\Delta} \frac{\beta(t)/\beta(T_s)}{P(t,T_p)} \sigma_r B(t,T_p)\, dW^{T_p}(t) \\ &= \left(\Delta^{-1} + F(t;T_s,T_p)\right) \sigma_r B(t,T_p)\, dW^{T_p}(t), \quad t \in (T_s,T_p]. \end{aligned} \tag{95}$$

In totality, we have verified (11)-(12).

A.2 Term Variance and Option Pricing

With $\sigma_F(u;T_s,T_p)$ defined as in (12), and with $\tau_p \triangleq T_p - t$ and $\tau_s \triangleq T_s - t$, we may now define a term variance v_{II} for a type-II caplet as

$$\begin{aligned} v_{II}(t;T_s,T_p) &\triangleq \int_t^{T_p} \sigma_F(u;T_s,T_p)^2\, du \\ &= \frac{\sigma_r^2}{2\kappa^3} \times \begin{cases} 4e^{-\kappa\tau_p} - e^{-2\kappa\tau_p} + 2\kappa\tau_p - 3, & t \in (T_s,T_p], \\ 2(e^{-\kappa\Delta} + \kappa\Delta - 1) - \left(e^{\kappa\Delta} - 1\right)^2 e^{-2\kappa\tau_p}, & t \in [0,T_s]. \end{cases} \end{aligned} \tag{96}$$

In the Ho-Lee limit ($\kappa \downarrow 0$), this becomes

$$\lim_{\kappa\downarrow 0} v_{II}(t;T_s,T_p) = \sigma_r^2 \times \begin{cases} \frac{1}{3}\tau_p^3, & t \in (T_s,T_p], \\ \Delta^2\tau_s + \frac{1}{3}\Delta^3, & t \in [0,T_s]. \end{cases} . \tag{97}$$

Given (96) (and (97)), we can use known results for displaced log-normal diffusion to get a type-II option price of

$$V_{II}(t;K,T_s,T_p) = P(t,T_p)\left\{(1+F(t;T_s,T_p)\Delta)N(d_+) - (1+K\Delta)N(d_-)\right\}, \tag{98}$$

where

$$d_\pm = \frac{\ln\left(\frac{1+F(t;T_s,T_p)\Delta}{1+K\Delta}\right) \pm \frac{1}{2}v_{II}(t;T_s,T_p)}{\sqrt{v_{II}(t;T_s,T_p)}}.$$

To contrast the result above with the payout of a type-I caplet paying at time T_p and fixing at time T_s, we notice that the term variance that applies for such a caplet would be simply, for $t \in [0,T_s]$,

$$v_I(t;T_s,T_p) = \sigma_r^2 \int_t^{T_s} \left(B(u,T_p) - B(u,T_s)\right)^2 du = \sigma_r^2 \frac{\left(e^{-\kappa\Delta}-1\right)^2}{2\kappa^3}\left(1-e^{-2\kappa\tau_s}\right). \tag{99}$$

A.3 Characteristic Function for Money Market Account

For a type-II caplet, the payout is determined by the (exponential of) the accrual integral

$$\int_{T_s}^{T_p} r(u)\,du = \ln\beta(T_p) - \ln\beta(T_s).$$

We wish to establish the T_p-forward measure characteristic function

$$\tilde{\varphi}_G(k;t) = \mathrm{E}_t^{T_p}\left(e^{ik\cdot\left(\ln\beta(T_p)-\ln\beta(T_s)\right)}\right),$$

where $k \in \mathbb{R}$ and i is the imaginary unit. From (1) and (2) we see that

$$ik\cdot\left(\ln\beta(T_p)-\ln\beta(T_s)\right) = ik\cdot\ln\left(1+\Delta R(T_s,T_p)\right) = ik\cdot\ln\left(1+\Delta F(T_p;T_s,T_p)\right).$$

Given (11), an application of Ito's lemma to $q(t) = \exp\left(ik\cdot\ln\left(1+\Delta F(t;T_s,T_p)\right)\right)$ yields

$$dq(t)/q(t) = O\left(dW^{T_p}(t)\right) - \frac{1}{2}ik(ik-1)\sigma_F(t;T_s,T_p)^2\,dt$$

such that

$$\tilde{\varphi}_G(k;t) = \mathrm{E}_t^{T_p}\left(q(T_p)\right) = e^{ik\cdot\ln\left(1+\Delta F(t;T_s,T_p)\right)}e^{\frac{1}{2}ik(ik-1)v_{II}(t;T_s,T_p)}, \tag{100}$$

where v_{II} was defined in (96).

The result in (100) shows that $\ln\beta(T_p) - \ln\beta(T_s)$, as one would expect, is a Gaussian random variable in measure $\mathbb{Q}^{T_p}$, with mean $\ln(1+\Delta F(t;T_s,T_p)) - \frac{1}{2}v_{II}(t;T_s,T_p)$ and variance $v_{II}(t;T_s,T_p)$.

B Appendix: Proofs

B.1 Proof of Proposition 2

First, we recast the problem in compact form, amenable to analytical resolution. Multiplying the equations in Lemma 2 by τ^n and summing up the first (resp. the second) equation for $n \geq 0$ (resp. $n \geq 1$) yields

$$\begin{aligned}\frac{\partial \mathcal{G}_{11}(t,s;\tau)}{\partial s} &= -\lambda_{12}(s)\mathcal{G}_{11}(t,s;\tau) + \lambda_{21}(s)\mathcal{G}_{12}(t,s;\tau),\\ \frac{\partial \mathcal{G}_{12}(t,s;\tau)}{\partial s} &= -\lambda_{21}(s)\mathcal{G}_{12}(t,s;\tau) + \tau\lambda_{12}(s)\mathcal{G}_{11}(t,s;\tau),\end{aligned}$$

where we used $\alpha_{12}^0(t,s) = 0$. Moreover, the boundary conditions (34)-(36) nicely reduce to $\mathcal{G}_{11}(t,t;\tau) = \sum_{n=0}^{+\infty} \alpha_{11}^n(t,t)\tau^n = 1$ and $\mathcal{G}_{12}(t,t;\tau) = \sum_{n=0}^{+\infty} \alpha_{12}^n(t,t)\tau^n = 0$, yielding:

$$\frac{\partial \mathcal{G}(t,s;\tau)}{\partial s} = \Omega(s;\tau)\cdot\mathcal{G}(t,s;\tau), \quad \Omega(s;\tau) \triangleq \begin{pmatrix} -\lambda_{12}(s) & \lambda_{21}(s) \\ \tau\lambda_{12}(s) & -\lambda_{21}(s)\end{pmatrix}, \tag{101}$$

and subject to the initial condition $\mathcal{G}(t,t;\tau) = (1,0)^\top$. While the generic case of (101) may be solved by, e.g., the so-called *Magnus expansion*, a particularly favorable configuration (among possible others) emerges when $\nu = \lambda_{12}(s)/\lambda_{21}(s)$ is a constant, as Ω then simplifies to a separable form:

$$\Omega(s;\tau) = \lambda_{21}(s)\omega(\tau), \quad \omega(\tau) \triangleq \begin{pmatrix} -\nu & 1 \\ \nu\tau & -1\end{pmatrix}.$$

The solution to (101) then becomes:

$$\mathcal{G}(t,s;\tau) = \exp\left(\Lambda_{21}(t,s)\omega(\tau)\right)\cdot\begin{pmatrix}1\\0\end{pmatrix}, \tag{102}$$

where Λ_{21} was defined in (38). It is now a matter of computing the matrix exponential above, which we achieve by standard means of matrix diagonalization. Solving the characteristic equation for $\omega(\tau)$, we find after a few manipulations that the eigenvalues of $\Lambda_{21}(t,s)\omega(\tau)$ are:

$$\lambda_\pm(\tau) \triangleq \Lambda_{21}(t,s)\left(-\frac{1}{2}(1+\nu) \pm \delta(\tau)\right),$$

where $\delta(\tau)$ is as in (38). Thus, diagonalizing $\Lambda_{21}(t,s)\omega(\tau)$ and exponentiating the resulting equality yields

$$\exp\left(\Lambda_{21}(t,s)\omega(\tau)\right) = \mathcal{P}\begin{pmatrix}\exp(\lambda_+(\tau)) & 0\\ 0 & \exp(\lambda_-(\tau))\end{pmatrix}\mathcal{P}^{-1}, \tag{103}$$

where $\mathcal{P}$ is a 2×2 matrix containing the eigenvectors of $\Lambda_{21}(t,s)\omega(\tau)$. Consider now a and b satisfying:

$$\exp(\lambda_+(\tau)) = a\lambda_+(\tau) + b, \quad \exp(\lambda_-(\tau)) = a\lambda_-(\tau) + b.$$

From (103) it then follows that

$$\exp\left(\Lambda_{21}(t,s)\omega(\tau)\right) = a\Lambda_{21}(t,s)\omega(\tau) + bI,$$

where I is the identity matrix. Solving for a and b, we get

$$\begin{aligned}\exp\left(\Lambda_{21}(t,s)\omega(\tau)\right) &= \frac{e^{\lambda_+(\tau)} - e^{\lambda_-(\tau)}}{\lambda_+(\tau) - \lambda_-(\tau)}\Lambda_{21}(t,s)\omega(\tau) + \frac{\lambda_+ e^{\lambda_-(\tau)} - \lambda_- e^{\lambda_+(\tau)}}{\lambda_+(\tau) - \lambda_-(\tau)} I \\ &= e^{-\frac{1+\nu}{2}\Lambda_{21}(t,s)} \frac{\sinh\left(\delta(\tau)\Lambda_{21}(t,s)\right)}{\delta(\tau)}\omega(\tau) \\ &+ e^{-\frac{1+\nu}{2}\Lambda_{21}(t,s)} \left(\cosh(\delta(\tau)\Lambda_{21}(t,s)) + \frac{1+\nu}{2} \cdot \frac{\sinh\left(\delta(\tau)\Lambda_{21}(t,s)\right)}{\delta(\tau)}\right) I.\end{aligned}$$

The result (37) in Proposition 37 then follows from (102) after a few manipulations. ∎

B.2 Proof of Lemma 3

Considering s fixed, we first define a t-indexed process (with some slight abuse of notation)

$$\phi_J(t) = \phi_J(k;t,s) = 1_{c(t)=e_1}\phi_J^1(k;t,s) + 1_{c(t)=e_2}\phi_J^2(k;t,s,J(t)), \tag{104}$$

and we wish estimate the expected growth of $\phi_J(t)$ over the interval $[t, t+dt]$. Suppressing dependence on arguments k and s for clarity, notice first that if $c(t) = e_1$, then $\phi_J(t) = \phi_J^1(t)$ and

$$\phi_J(t+dt) - \phi_J(t) = 1_{c(t+dt)=e_1}\phi_J^1(t+dt) + 1_{c(t+dt)=e_2}\phi_J^2(t+dt, J(t+dt)) - \phi_J^1(t)$$

such that

$$\begin{aligned}\mathrm{E}^{\mathbb{Q}}\left(d\phi_J(t)|c(t)=e_1\right) &= \left(1-\lambda_{12}(t)\,dt\right)\phi_J^1(t+dt) - \phi_J^1(t) \qquad (105)\\ &\quad + \lambda_{12}(t)\,dt \cdot \mathrm{E}^{\mathbb{Q}}\left(\phi_J^2(t+dt, J(t+dt))|c(t)=e_1\right) \\ &= \frac{\partial\phi_J^1(t)}{\partial t}\,dt - \lambda_{12}(t)\phi_J^1(t)\,dt + \lambda_{12}(t)\,dt\int_{-\infty}^{\infty}\phi_J^2(t,y)p_H(y,t)\,dy. \qquad (106)\end{aligned}$$

Proceeding in similar fashion for the case where $c(t) = e_2$ and $J(t) = y$, we get

$$\mathrm{E}^{\mathbb{Q}}\left(d\phi_J(t)|c(t)=e_2, J(t)=y\right) = \lambda_{21}(t)\,dt\cdot\phi_J^1(t) + \frac{\partial\phi_J^2(t,y)}{\partial t}\,dt - \lambda_{21}(t)\,dt\cdot\phi_J^2(t,s,y). \tag{107}$$

It is clear from the definition of $\phi_J^1(k;t,s)$ and $\phi_J^2(k;t,s,y)$ that $\mathrm{E}^{\mathbb{Q}}\left(d\phi_J(t)|c(t)=e_1\right) = 0$ and $\mathrm{E}^{\mathbb{Q}}\left(d\phi_J(t)|c(t)=e_2, J(t)=y\right) = -iky\phi_J^2(k;t,s,y)\,dt$. Combining this with (106) and (107) yield the result in the Lemma. ∎

B.3 Proof of Proposition 4

First, using time-homogeneity and the notations introduced, equation (46) writes

$$n'(t) = \lambda_{12} n(t) - \lambda_{12} \bar{\phi}_H(t-s) - \lambda_{12}\lambda_{21} \int_t^s \bar{\phi}_H(t-v) n(v)\, dv$$

subject to $n(s) = 1$. By successive differentiation with respect to t, we get, for $p \geq 0$:

$$n^{(p+1)}(t) = \lambda_{12}\left(\lambda_{21} \sum_{k=0}^{p-1} n^{(k)}(t) + n^{(p)}(t) - \bar{\phi}_H^{(p)}(t-s) - \lambda_{21} \int_t^s \bar{\phi}_H^{(p)}(t-v) n(v)\, dv\right). \quad (108)$$

Thus, applying $t = s$ yields a recursive expression for the boundary conditions (given $n(s) = 1$):

$$n^{(p+1)}(s) = \lambda_{12}\left(\lambda_{21} \sum_{k=0}^{p-1} n^{(k)}(s) + n^{(p)}(s) - \bar{\phi}_H^{(p)}(0)\right),$$

as stated in the proposition.

Next, we introduce the differential operator $L \triangleq \prod_{p=1}^{N} (\partial_t + (\lambda_{21} - iz_p)I)$. Per construction, we have $L(\bar{\phi}_H) = 0$, that is:

$$\sum_{p=0}^{N} \alpha_p \bar{\phi}_H^{(p)} = 0$$

Multiplying (108) by α_p and summing-up for $p = 0, \cdots, N$ yields:

$$\begin{aligned}
\sum_{p=0}^{N} \alpha_p n^{(p+1)}(t) &= \lambda_{12}\left(\lambda_{21} \sum_{p=0}^{N} \alpha_p \sum_{k=0}^{p-1} n^{(k)}(t) + n^{(p)}(t)\right) \\
&\quad - \lambda_{12} L(\bar{\phi}_H)(t-s) - \lambda_{12}\lambda_{21} \int_t^s L(\bar{\phi}_H)(t-v) n(v)\, dv \\
&= \lambda_{12}\left(\lambda_{21} \sum_{p=0}^{N} \alpha_p \sum_{k=0}^{p-1} n^{(k)}(t) + n^{(p)}(t)\right)
\end{aligned}$$

since $L(\bar{\phi}_H) = 0$. This proves the proposition. ∎

B.4 Proof of Proposition 5

From the definition of z^{12}_{AD} and the fact that $1 = 1_{c(s)=e_1} + 1_{c(s)=e_2}$, we see that

$$\begin{aligned}
&z^{12}_{AD}(k;t,s+ds,y) \\
&\quad = \mathrm{E}^{\mathbb{Q}}\left((1+ik\cdot J(s)ds)e^{ik\int_t^s J(u)\,du}\left(1_{c(s)=e_1}+1_{c(s)=e_2}\right)1_{c(s+ds)=e_2}\delta\left(J(s+ds)-y\right)\middle| c(t)=e_1\right) \\
&\quad = \mathrm{E}^{\mathbb{Q}}\left(e^{ik\int_t^s J(u)\,du}1_{c(s)=e_1,c(s+ds)=e_2}\delta\left(H(s)-y\right)\middle| c(t)=e_1\right) \\
&\qquad\qquad + \mathrm{E}^{\mathbb{Q}}\left((1+ik\cdot J(s)\,ds)e^{ik\int_t^s J(u)\,du}1_{c(s)=e_2,c(s+ds)=e_2}\delta\left(J(s)-y\right)\middle| c(t)=e_1\right) \\
&\quad = z^{11}_{AD}(k;t,s)p_H(s,y)\lambda_{12}(s)\,ds + z^{12}_{AD}(k;t,s,y)\left(1-\lambda_{21}(s)\,ds\right)(1+ik\cdot y\,ds),
\end{aligned}$$

where the last equality follows from elementary properties of the Poisson process. Equation (51) follows. In similar manner,

$$\begin{aligned}
z^{11}_{AD}(k;t,s+ds) &= \mathrm{E}^{\mathbb{Q}}\left((1+ik\cdot J(s)\,ds)\,e^{ik\int_t^s J(u)\,du}\left(1_{c(s)=e_1,c(s+ds)=e_1}+1_{c(s)=e_2,c(s+ds)=e_1}\right)\middle| c(t)=e_1\right) \\
&= z^{11}_{AD}(k;t,s)\left(1-\lambda_{12}(s)\,ds\right)+\mathrm{E}^{\mathbb{Q}}\left((1+ik\cdot J(s)\,ds)\;e^{ik\int_t^s J(u)\,du}1_{c(s)=e_2}\middle| c(t)=e_1\right)\lambda_{21}(s)\,ds \\
&= z^{11}_{AD}(k;t,s)\left(1-\lambda_{12}(s)\,ds\right)+\mathrm{E}^{\mathbb{Q}}\left(\,e^{ik\int_t^s J(u)\,du}1_{c(s)=e_2}\middle| c(t)=e_1\right)\lambda_{21}(s)\,ds \\
&= z^{11}_{AD}(k;t,s)\left(1-\lambda_{12}(s)\,ds\right)+\int_{-\infty}^{\infty} z^{12}_{AD}(k;t,s,y)\,dy\cdot\lambda_{21}(s)\,ds,
\end{aligned}$$

which is (52). ∎

B.5 Proof of Proposition 7

We start out with the following useful lemma:

Lemma 7 *For any complex number α, define*

$$\phi^{\alpha}_J(t,T,J(t)) \triangleq \mathrm{E}^{\mathbb{Q}}_t\left(e^{\alpha J(T)-\int_t^T J(u)\,du}\right).$$

Then

$$\phi^{\alpha}_J(t,T,J) = e^{\mathcal{A}(\alpha;t,T)-J\cdot\mathcal{B}(\alpha;t,T)} \tag{109}$$

where

$$\begin{aligned}
\mathcal{B}(\alpha;t,T) &= \frac{1-e^{-\kappa_J(T-t)}}{\kappa_J}-\alpha e^{-\kappa_J(T-t)} = B_J(t,T)-\alpha e^{-\kappa_J(T-t)}, \\
\mathcal{A}(\alpha;t,T) &= \int_t^T \lambda_J(u)\left(1-\phi_H\left(i\mathcal{B}(\alpha;u,T),u\right)\right)\,du.
\end{aligned}$$

Proof: (sketch): The proof of the lemma is nearly identical to that of Proposition 1. Specifically, we make the exponential ansatz (109) and then write down the ODEs for $\mathcal{A}$ and $\mathcal{B}$ necessary to make $\mathrm{E}_t^{\mathbb{Q}}(d\phi_J^\alpha(t)) = J(t)\phi_J^\alpha(t)\,dt$. The resulting ODE system is

$$\begin{aligned}\frac{\partial \mathcal{B}(\alpha;t,T)}{\partial t} &= -1 + \kappa_J(t)\mathcal{B}(\alpha;t,T),\\ \frac{\partial \mathcal{A}(\alpha;t,T)}{\partial t} &= \lambda_J(t)\left(1 - \phi_H\left(i\mathcal{B}(\alpha;t,T),t\right)\right),\end{aligned}$$

where the terminal conditions are now easily seen to be $\mathcal{A}(\alpha;T,T) = 0$ and $\mathcal{B}(\alpha;T,T) = -\alpha$. Solving this ODE system proves the proposition. ∎

Equipped with the above lemma, the proof of Proposition 7 is now straightforward. Indeed, proceeding in the same fashion as in the proof of Lemma 6, we see that, for $t \le T_s$,

$$\begin{aligned}\varphi_J(k;t) &= \mathrm{E}_t^{T_p}\left(e^{ik\int_{T_s}^{T_p} J(u)\,du}\right)\\ &= \frac{1}{P_J(t,T)}\mathrm{E}_t^{\mathbb{Q}}\left(e^{ik\int_{T_s}^{T_p} J(u)\,du}e^{-\int_t^{T_p} J(u)\,du}\right)\\ &= \frac{1}{P_J(t,T)}\mathrm{E}_t^{\mathbb{Q}}\left(e^{-\int_t^{T_s} J(u)\,du}\mathrm{E}_{T_s}^{\mathbb{Q}}\left(e^{(ik-1)\int_{T_s}^{T_p} J(u)\,du}\right)\right), \quad t \le T_s.\end{aligned}$$

From Proposition 1, we know that

$$\mathrm{E}_{T_s}\left(e^{(ik-1)\int_{T_s}^{T_p} J(u)\,du}\right) = \phi_J\left(k+i;T_s,T_p\right) = e^{A_J(T_s,T_p;k+i)+(ik-1)J(T_s)\cdot B_J(T_s,T_p)}$$

so

$$\begin{aligned}\varphi_J(k;t) &= \frac{e^{A_J(T_s,T_p;k+i)}}{P_J(t,T_p)}\mathrm{E}_t^{\mathbb{Q}}\left(e^{-\int_t^{T_s} J(u)\,du}e^{(ik-1)J(T_s)\cdot B_J(T_s,T_p)}\right)\\ &= \frac{e^{A_J(T_s,T_p;k+i)}}{P_J(t,T_p)}e^{\mathcal{A}((ik-1)B_J(T_s,T_p);t,T_s)-J(t)\cdot\mathcal{B}((ik-1)B_J(T_s,T_p);t,T_s)}, \quad t \le T_s,\end{aligned}$$

where the second equality follows from Lemma 7 with $\alpha = (ik-1)B_J(T_s,T_p)$.

The case where $t \in (T_s,T_p]$ is proven in the same fashion; we omit the details. ∎

B.6 Proof of Proposition 8

By definition,

$$\begin{aligned}\varphi_J(k;t) &= \frac{1}{P_J(t,T_p)}\mathrm{E}_t^{\mathbb{Q}}\left(e^{-\int_t^{T_p} J(u)\,du} e^{ik\int_{T_s}^{T_p} J(u)\,du}\right)\\ &= \frac{1}{P_J(t,T_p)}\mathrm{E}_t^{\mathbb{Q}}\left(e^{-\int_t^{T_s} J(u)\,du} e^{(ik-1)\int_{T_s}^{T_p} J(u)\,du}\right)\\ &= \frac{1}{P_J(t,T_p)}\mathrm{E}_t^{\mathbb{Q}}\left(e^{-\int_t^{T_s} J(u)\,du}\phi_J(k+i;T_s,T_p,J(T_s))\right)\end{aligned}$$

where the last expression follows by conditional expectations. This establishes (79), given our assumption that $J(t) = 0$. To prove (80), we note that (see Lemma 3)

$$\phi_J\left(k+i;T_s,T_p,J(T_s)\right) = 1_{J(T_s)=0}\phi_J^1(k;T_s,T_p) + 1_{J(T_s)\neq 0}\phi_J^2(k;T_s,T_p,J(s))$$

such that

$$\begin{aligned}\omega(t;T_s,T_p) &= \mathrm{E}_t^{\mathbb{Q}}\left(e^{-\int_t^{T_s} J(u)\,du}1_{J(T_s)=0}\phi_J^1(k;T_s,T_p)|J(t)=0\right)\\ &\qquad + \mathrm{E}_t^{\mathbb{Q}}\left(e^{-\int_t^{T_s} J(u)\,du}1_{J(T_s)\neq 0}\phi_J^2(k;T_s,T_p,J(T_s))|J(t)=0\right)\\ &\triangleq \omega_1(t;T_s,T_p) + \omega_2(t;T_s,T_p).\end{aligned}$$

Here $\phi_J^1(k;T_s,T_p)$ is deterministic, so, by definition of z_{AD}^{11},

$$\omega_1(t;T_s,T_p) = \mathrm{E}_t^{\mathbb{Q}}\left(e^{-\int_t^{T_s} J(u)\,du}1_{J(T_s)=0}|J(t)=0\right)\cdot\phi_J^1(k;T_s,T_p) = z_{AD}^{11}(i;t,T_s)\cdot\phi_J^1(k;T_s,T_p).$$

Similarly,

$$\begin{aligned}\omega_2(t;s,T) &= \mathrm{E}_t^{\mathbb{Q}}\left(e^{-\int_t^{T_s} J(u)\,du}1_{J(T_s)\neq 0}\phi_J^2(k;T_s,T_p,J(T_s))|J(t)=0\right)\\ &= \int_{-\infty}^{\infty}\mathrm{E}_t^{\mathbb{Q}}\left(e^{-\int_t^{T_s} J(u)\,du}1_{J(T_s)\neq 0}\delta(J(T_s)-x)\cdot\phi_J^2(k;T_s,T_p,x)|J(t)=0\right)dx\\ &= \int_{-\infty}^{\infty}\mathrm{E}_t^{\mathbb{Q}}\left(e^{-\int_t^{s} J(u)\,du}1_{J(s)\neq 0}\delta(J(s)-x)|J(t)=0\right)\phi_J^2(k;T_s,T_p,x)\,dx\\ &= \int_{-\infty}^{\infty} z_{AD}^{12}(i;t,T_s,x)\cdot\phi_J^2(k;T_s,T_p,x).\end{aligned}$$

References

[1] Andersen, L. (2008), "Markov Models for Commodity Futures: Theory and Practice," *Quantitative Finance*, 10(8), 831-854.

[2] Andersen, L. and V. Piterbarg (2010), *Interest Rate Modeling*, Atlantic Financial Press.

[3] Andersen, L. and M. Lake (2019), "Robust High-Precision Option Pricing by Fourier Transforms: Contour Deformations and Double-Exponential Quadrature," *Wilmott Magazine*, July, 22-41.

[4] Bang, D. (2018), "Local Stochastic Volatility: Shaken, not Stirred,", *Risk Magazine*, December, 136-141.

[5] Carr, P. and D. Madan (1999), "Option Valuation using the fast Fourier Transform," *Journal of Computational Finance*, 2 (4), 61-73.

[6] Duffie, D., J. Pan, and K. Singleton (2000), "Transform Analysis and Asset Pricing for Affine Jump-Diffusions," *Econometrica*, 68 (6), 1343-1376.

[7] Gil-Pelaez, J. (1951), "A Note on the Inversion Theorem," *Biometrika*, 38, 481-482.

[8] Gautschi, W. (1996), "Orthogonal polynomials: applications and computation," *Acta Numerica*, 5, 45-119

[9] Hagan, P., D. Kumar, A. Lesniewski, and D. Woodward (2002), "Managing Smile Risk," *Wilmott Magazine*, September, 84-108.

[10] Hagan, P., Lesniewski, A., and D. Woodward (2018), "Managing Vol Surfaces," *Wilmott Magazine*, January, 24-43.

[11] Heltman, J. (2019), "What ongoing repo turmoil means for banks," *American Banker,* https://www.americanbanker.com/list/what-ongoing-repo-turmoil-means-for-banks .

[12] Henrard, M. (2019), "A Quant Perspective on IBOR Fallback Consultation Results," Woorking Paper, muRisQ Advisory.

[13] Hordahl, P. and M. King (2008), "Developments in Repo Markets during the Financial Turmoil," *BIS Quarterly Review*, December. https://www.bis.org/publ/qtrpdf/r_qt0812e.pdf .

[14] Lee, R. W. (2004), "Option Pricing by Transform Methods: Extensions, Unification, and Error Control," *Journal of Computational Finance*, 7(3), 51-86.

[15] Lyashenko, A. and F. Mercurio (2019), "Libor Replacement: a Modelling Framework for In-Arrears Term Rates," *Risk Magazine*, July, 72-77.

[16] Piterbarg, V. (2007), "Markovian Projection for Volatility Calibration," *Risk Magazine*, April.

[17] Piterbarg, V. (2020), "Interest Rates Benchmark Reform and Options Markets," Working Paper, Nat West Markets. https://ssrn.com/abstract=3537925 or http://dx.doi.org/10.2139/ssrn.3537925 .

[18] Press, W., S. Teukolsky, W. Vetterling, and B. Flannery (2007), *Numerical Recipes in C++: The Art of Scientific Computing*, Cambridge University Press.

[19] Schneir, J., W. Martinez, and J. Woytash (2019), "Repo Rate Spike: A 'Tail' Of Low Liquidity," PIMCO Blog, September. https://blog.pimco.com/en/2019/09/repo-rate-spike-a-tail-of-low-liquidity.

[20] Schrimpf, A. and V. Sushko (2019), "Beyond LIBOR: a Primer on the New Reference Rates," *Bank of International Settlements*. https://www.bis.org/publ/qtrpdf/r_qt1903e.pdf .

[21] Schulhofer-Wohl, S. (2019), "Understanding Recent Fluctuations in Short-Term Interest Rates," *Chicago Fed Letter*, No. 423, September. https://www.chicagofed.org/publications/chicago-fed-letter/2019/423.

[22] S&P Global Market Intelligence (2019), "After repo rates spike, leveraged loan investors raise concern about SOFR volatility," September. https://www.spglobal.com/marketintelligence/en/news-insights/latest-news-headlines/leveraged-loan-news/after-repo-rates-spike-leveraged-loan-investors-raise-concern-about-sofr-volatility .

[23] Wipf, T. (2019), "Wave Goodbye to Libor. Welcome its Successor, SOFR," Bloomberg Opinion. https://www.bloomberg.com/opinion/articles/2019-12-06/wave-goodbye-to-libor-welcome-its-successor-sofr .