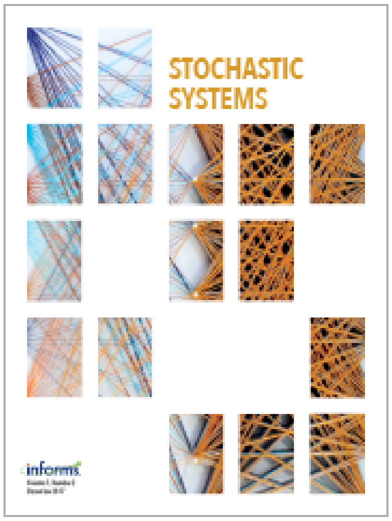


Stochastic Systems

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

A Storage Process with Alternating Lévy Input

Hansjörg Albrecher, Onno Boxma, Offer Kella, Michel Mandjes

To cite this article:
Hansjörg Albrecher, Onno Boxma, Offer Kella, Michel Mandjes (2026) A Storage Process with Alternating Lévy Input. Stochastic Systems

Published online in Articles in Advance 04 Sep 2026

. https://doi.org/10.1287/stsy.2025.0134

informs
https://pubsonline.informs.org/journal/stsy

STOCHASTIC SYSTEMS
Articles in Advance, pp. 1–23
ISSN 1946-5238 (online)

A Storage Process with Alternating Lévy Input

Hansjörg Albrecher,[a,*] Onno Boxma,[b] Offer Kella,[c] Michel Mandjes[d,e]

[a] Faculty of Business and Economics, University of Lausanne and Swiss Finance Institute, 1015 Lausanne, Switzerland; [b] EURANDOM and Department of Mathematics and Computer Science, Eindhoven University of Technology, 5612 AZ Eindhoven, Netherlands; [c] Department of Statistics and Data Science, The Hebrew University of Jerusalem, Jerusalem 9190501, Israel; [d] Mathematical Institute, Leiden University, 2311 EZ Leiden, Netherlands; [e] Korteweg-de Vries Institute for Mathematics, University of Amsterdam, 1012 WP Amsterdam, Netherlands

*Corresponding author

hansjoerg.albrecher@unil.ch (H. Albrecher), o.j.boxma@tue.nl (O. Boxma), offer.kella@gmail.com (O. Kella), m.r.h.mandjes@uva.nl (M. Mandjes)

https://orcid.org/0000-0002-5434-9270 (H. Albrecher), https://orcid.org/0000-0003-4317-5380 (O. Boxma), https://orcid.org/0000-0002-6433-5470 (O. Kella), https://orcid.org/0000-0001-6783-4833 (M. Mandjes)

Received: November 15, 2025
Revised: May 7, 2026
Accepted: July 22, 2026
Published Online in Articles in Advance: September 4, 2026

https://doi.org/10.1287/stsy.2025.0134

Abstract. In this paper, we study a queueing process for which the dynamics are changed once the workload in the queue exceeds a predefined threshold, and these new dynamics stay in force until the queue is emptied, at which point the previous dynamics are again reinstalled. For general spectrally positive Lévy processes in each case, we derive expressions for the workload at an independent exponentially distributed random time horizon and study in detail properties of the workload in stationarity. We work out explicit formulas as well as expressions for the optimal changing threshold for special cases of the underlying process dynamics and a chosen set of involved switching, holding, and service cost functions.

Funding: The research of O. Boxma and M. Mandjes has been partly funded by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek Gravitation [Project Networks, Grant 024.002.003]. The research of O. Kella is partially funded by Israel Science Foundation [Grant 3336/24] and the Vigevani Chair in Statistics.

Keywords: spectrally positive Lévy process • stochastic storage process • Lévy queue workload • alternating Lévy process • optimal switching level

1. Introduction

The stochastic modeling of storage and queueing processes and the derivation of distributional properties of concrete quantities like their workload or the length of their busy periods have a long history in applied probability. Over the years, many model variations were investigated and systematically understood, and nowadays, there is quite a detailed body of tools and techniques for their study available (see, for instance, Asmussen (2003) for a survey). It turned out that for various queueing models, there is an interesting and fruitful duality with models for the surplus process of an insurance portfolio that have been studied in insurance risk theory (Asmussen and Albrecher 2010, Mandjes and Boxma 2023). In particular, model assumptions on the arrival process of customers in a queue (or items in the storage process case, respectively) and the arrival process of claims in the insurance portfolio case have similar motivations and applicability, and an analogous statement holds for the random service requirements of a new customer (size of a newly arrived item in the storage facility, respectively) and the random claim sizes in the insurance model. More than that, the viewpoint of one discipline sometimes also leads to new results or simpler proofs in the other (Albrecher and Boxma 2004, Albrecher et al. 2009). Along that line, whenever some model variation in one field leads to an amenable analysis, there is a good chance that in the other field there is also an interesting analogue or some interpretation and additional insight to gain.

In the present paper, we take up on this entanglement of queueing/storage and risk theory and investigate a storage model that is motivated by a model variant recently studied in insurance risk theory. In the context of dividend payments of an insurance company, Albrecher et al. (2018) studied the ruin probability and expected discounted dividend payments for a situation where dividend payments cannot decrease over time. Concretely, Albrecher et al. (2018) assumed that once during the lifetime of the process (i.e., before the surplus is depleted, which refers to the event of ruin of the company), the rate of dividend payments can be raised permanently and the question was as follows: What is the optimal threshold to do so? Such an assumption is motivated by the fact that shareholders do not like to see a dividend rate decreasing, and it is of interest to see how much that constraint compromises the performance of the company (in terms of overall dividend payments until ruin and the resulting ruin probability). In a queueing setup, the analogous model adaptation naturally is to increase the service speed once the workload has reached a certain threshold and not reducing that speed until the queue is emptied (i.e., the workload has reached level 0 again). The motivation in that case may be that one wants to avoid too large waiting times in the queue, but the switching to and processing of an increased service speed will typically come at a cost, and, depending on the posed objective function concerning costs and benefits, there may be an optimal threshold for such a switch. In a queueing context, Cohen (1976) considered such a switching at a threshold, but in a reversible fashion; that is, once the workload goes again below that threshold, the original service speed is retaken. However, in practice this may not be feasible or efficient, for instance, because of the costs of such a switch.

Motivated by the above considerations, in this paper, we study a considerably more general model: We consider a spectrally positive Lévy process, in which the entire dynamics (the Laplace exponent of the Lévy process; in a queueing setting this could mean service speed, arrival rate of customers and service time distribution) is switched once a threshold is surpassed, and the new dynamics stay in force until the process reaches level zero. We derive results on the resulting workload, both at an exponential time horizon and in stationarity. We establish a number of other new fluctuation results along the way. The model with only the service speed changing at the switching time is contained as a special case in the above results. For the latter variant and the special cases of the two Lévy processes to be Brownian motions or compound Poisson processes with phase-type claims, we work out explicit formulas. For any given holding costs, switching costs, and service costs, we then derive expressions for the aggregate costs associated to each switching level, which allows also to identify the switching level that minimizes these costs. Finally, we illustrate the obtained results in some numerical examples. In the following, we summarize the literature that is most closely related to the results of this paper.

1.1. Related Literature

The queueing literature contains a large number of papers in which arrival rates, service time distributions, and/or service speeds change when certain thresholds are crossed. We refer to Dshalalow (1997) for an extensive survey with 277 references. An important distinction is whether (Category 1) the service speed, and so on, only depends on the present workload level or (Category 2) also on the mode of the system. In the latter category, the system switches from mode 1 to mode 2 when the workload up-crosses some level, and switches back to mode 1 at the first subsequent down-crossing of some other level (notice that if these levels are the same, then we are back in Category 1). Some early references for Category 1 are Cohen (1976) and Gaver and Miller (1962). Gaver and Miller (1962) present a pioneering study on queueing systems in which the workload determines the service speed. Cohen (1976) considers an M/G/1 queue in which the server speed is increased as soon as some workload level b is exceeded and is switched back to the original speed once b is down-crossed again (see also Tijms 1975). Cohen determines the switching level that minimizes a certain weighted sum of holding, switching and fast-serving costs.

Some references for Category 2 are Bae et al. (2003), Barlev and Perry (1993), Bekker (2009), Lee and Kim (2006), and Tijms and van der Duyn Schouten (1978). Tijms and van der Duyn Schouten (1978) consider an M/G/1 queue with finite workload capacity K and two switchover levels m_1 and m_2, with $0 \le m_2 \le m_1 \le K$. If the workload up-crosses level m_1, then the arrival rate, service time distribution, and server speed are adapted. They switch back to their original value when m_2 is subsequently down-crossed. The focus in their paper is on cost minimization. Bae et al. (2003) also study cost minimization for a finite dam. The papers of Barlev and Perry (1993) and Lee and Kim (2006) are devoted to queues in which the service speed is some workload-dependent function $r_i(\cdot)$ in mode i, $i = 1, 2$. The paper that is closest to ours is Bekker (2009). He considers a reflected spectrally positive Lévy process with two modes 1 and 2, with the Laplace exponent of the driving Lévy process being $\varphi_i(\cdot)$ in mode i, $i = 1, 2$. The (m_1, m_2) control rule is the same as specified above. Bekker determines the steady-state workload level in this system; he also devotes some attention to the case of a finite dam, which gives rise to a doubly reflected spectrally positive Lévy process.

In several papers, service speed switching is only allowed at specific instants different from the actual instant at which a workload level is crossed. In Bekker and Boxma (2007), the service speed in an M/G/1 queue is determined by the workload level at customer arrival epochs. In Bekker et al. (2008b), it is determined by the workload level at the arrival instants of an external Poisson observer. In Bekker et al. (2009), this is extended to the case of a spectrally positive Lévy input process. The latter paper also considers the case of instantaneous switching of mode (and hence of Laplace exponent of the Lévy input process) when some workload level is crossed.

We finally mention Bekker et al. (2008a); it considers a reflected spectrally positive Lévy process with two switch-over levels m_1 and $m_2 = 0$. The Laplace exponent of the Lévy process does not instantaneously switch when m_1 is up-crossed, but only after a random delay. If, after that delay, level zero has already been reached, then the Laplace exponent does not change. In Bekker and Boxma (2007) and Bekker et al. (2008a, b, 2009), the steady-state workload distribution is derived.

In the insurance literature, the dual model to switching the Lévy dynamics whenever a threshold level is crossed is the so-called *refraction model*, where dividends are paid at a certain rate above a particular surplus level (the threshold), and no dividends are paid below that level (Gerber and Shiu 2006a, and in full generality, Kyprianou and Loeffen 2010). Concretely, dividend payments reduce the increase of the surplus processes through premiums, so that the difference of the dynamics above and below the threshold is typically only in terms of the drift. In fact, such a refraction strategy is known to be optimal among all admissible dividend payout strategies in the sense that it maximizes the expected discounted dividend payments until ruin (see Jeanblanc-Picqué and Shiryaev (1995) and Asmussen and Taksar (1997) for a proof in the diffusion case; Gerber and Shiu (2006b) for a compound Poisson model; and Loeffen (2008) for the general spectrally one-sided Lévy model). A situation where the entire Lévy dynamics change at a crossing of a threshold does not have an immediate interpretation in terms of the dividend application and has, to the best knowledge of the authors, not been studied in the insurance literature.

The permanent and irreversible increase of the dividend rate at a threshold (as studied for the first time in Albrecher et al. (2018)), which motivates the queueing model of this paper, was further extended to discretely many switching thresholds in Albrecher et al. (2020) for the compound Poisson model and in Albrecher et al. (2022) for a Brownian surplus process (which is referred to as *ratcheting* of dividend payment rates). Using these discrete grids as building blocks, in these papers, the general continuous stochastic control problem of optimal switching thresholds and their rate increases to maximize the expected discounted dividends under this ratcheting constraint was solved (see Gao and Yin (2023), Guan and Xu (2024), Sun and Zhu (2026), and Wang et al. (2025) for recent variants and extensions).

The remainder of the paper is structured as follows. Section 2 introduces the model considered in this paper, together with notation and some preliminaries. Section 3 derives an explicit expression for the workload in this model at an exponential time horizon. Subsequently, Section 4 establishes results for the Laplace transform of the workload distribution in stationarity and gives an interpretation in terms of a decomposition result. Section 5 then translates the formulas into more explicit results for the particular cases where the Lévy processes are Brownian motions and compound Poisson processes with phase-type distributed jumps. Section 6 looks into concrete objective functions for optimizing the threshold level, given a certain structure of switching costs, holding costs, and costs for maintaining the higher service speed. Finally, Section 7 contains some concluding remarks and suggestions for future research directions.

2. Model, Preliminaries, and Notation

In this section we formally define our model, present some helpful preliminaries, and introduce useful notation. Throughout our paper the symbol ~ abbreviates *distributed*, or *distributed like*, and LST stands for *Laplace-Stieltjes Transform*. Also, 1_A denotes the indicator function of the event A. We use a.s. and iff to abbreviate *almost surely* and *if and only if*, respectively.

2.1. Model

The model we study in this paper can be seen as a workload process of a storage system (Prabhu 1998) that is alternatingly driven by the two processes X^+ and X^- (that are such that $X^+(0) = X^-(0) = 0$). The workload process, to be denoted by $Q = \{Q(t) | t \geq 0\}$, is affected by an underlying mode process $J = \{J(t) | t \geq 0\}$ in the following way.

The mode process J can attain two modes: + and −. We start by considering the case that $J(0) = +$; we later discuss the case $J(0) = -$. Let $Q(0) = x \in [0, b]$ for a *threshold* $b > 0$ that is held fixed throughout our analysis. We define Q as the reflection of X^+ at zero until it crosses level b for the first time. This concretely means that we set (Asmussen 2003, section IX.2)

$$Q(t) := X^+(t) + \max\left(-\inf_{0 \leq s \leq t} X^+(s), x\right), \tag{1}$$

for any $t \in [0, \theta_1^+]$, where

$$\theta_1^+ := \inf\{t \geq 0 | Q(t) > b\}.$$

At θ_1^+ the mode becomes $-$, which entails that the process Q switches its behavior to that of X^-, and continues this behavior until it hits zero. More precisely,

$$Q(t) = Q(\theta_1^+) + X^-(t) - X^-(\theta_1^+),$$

for any $t \in [\theta_1^+, \theta_1^-]$, where

$$\theta_1^- := \inf\{t \geq \theta_1^+ \,|\, Q(t) = 0\}.$$

At θ_1^- the mode becomes $+$. From that point, the process regenerates, and the same procedure is repeated indefinitely. This gives rise to the following recursive mechanism that is illustrated by Figure 1.

- While J is in the $+$ mode for the $(k+1)$-st time, we have the following dynamics: The workload process is defined via

$$Q(t) := [X^+(t) - X^+(\theta_k^-)] + \max\left(-\inf_{\theta_k^- \leq s \leq t} [X^+(s) - X^+(\theta_k^-)], Q(\theta_k^-)\right),$$

for $t \in [\theta_k^-, \theta_{k+1}^+)$, with

$$\theta_{k+1}^+ := \inf\{t \geq \theta_k^- \,|\, Q(t) > b\}.$$

At time θ_{k+1}^+ the mode switches from $+$ to $-$: we put $J(\theta_{k+1}^+) := -$.

- Likewise, while J is in the $-$ mode for the kth time, the workload dynamics are defined as follows:

$$Q(t) = Q(\theta_k^+) + X^-(t) - X^-(\theta_k^+),$$

for $t \in [\theta_k^+, \theta_k^-)$, with

$$\theta_k^- := \inf\{t \geq \theta_k^+ \,|\, Q(t) = 0\}.$$

At time θ_k^-, the mode switches from $-$ to $+$: We put $J(\theta_k^-) := +$.

It is now also clear how to define the process in case $J(0) = -$ and $Q(0) \geq 0$: In that case, the process Q first behaves as X^- until it hits level 0 and then alternates to the $+$ mode where it behaves as the reflected version of X^+, and so on. Clearly, the process $(Q, J) = \{(Q(t), J(t)) \,|\, t \geq 0\}$ is Markovian.

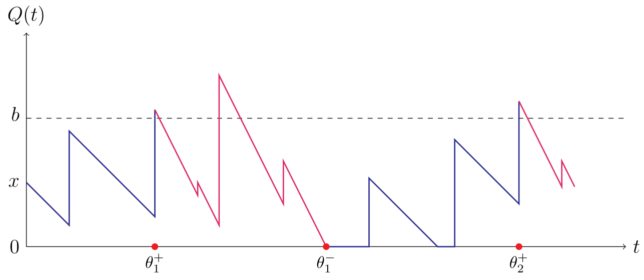
Thus far, we have not specified the driving processes X^+ and X^- yet. In this paper, these will be assumed to be *spectrally positive Lévy processes*—that is, Lévy processes with no negative jumps. We exclude the case in which these processes are *subordinators* (i.e., almost surely nondecreasing), because in that setting there would be no reflection in mode $+$. Also, we assume that X^+ is not a nonincreasing linear drift, because this would lead to $\theta_1^+ = \infty$.

Let us first briefly introduce this class of processes and recall several useful known results.

2.2. Preliminaries on Spectrally Positive Lévy Processes

In this section, we (i) define the class of Lévy processes that we use in this paper, (ii) discuss the corresponding reflected versions to be interpreted as Lévy-driven workload processes, (iii) present a number of useful fluctuation-theoretic identities, and (iv) recall a result on the exceedance time of a Lévy-driven workload process.

Figure 1. Sample Path of the Workload Process Q with $J(0) = +$ and $Q(0) = x \in [0, b]$



Notes. The red dots indicate mode transitions: at times θ_k^+ the mode switches from $+$ to $-$, at times θ_k^- from $-$ to $+$. While in the $+$ mode, the workload process behaves as the reflected version of X^+ (blue segments), whereas while in the $-$ mode, the workload process behaves as X^- (magenta segments).

2.2.1. Spectrally Positive Lévy Process. Assume that $X = \{X(t) | t \geq 0\}$ is a spectrally positive Lévy process, that is, a Lévy process with no negative jumps (Bertoin 1998, Kyprianou 2013). Throughout, we assume $X(0) = 0$. The Lévy process is characterized via its *Laplace exponent* $\varphi(\cdot)$. Indeed, for each $\alpha, t \geq 0$ we have that $\mathbb{E}e^{-\alpha X(t)} = e^{\varphi(\alpha)t}$, where, for constants c and $\sigma^2 \geq 0$, and a (Lévy) measure ν on $(0, \infty)$ such that $\int_{(0,\infty)} \min(1, x^2)\, \nu(\mathrm{d}x) < \infty$,

$$\varphi(\alpha) = -c\alpha + \frac{\sigma^2}{2}\alpha^2 + \int_{(0,\infty)} (e^{-\alpha x} - 1 + \alpha x 1_{(0,1]}(x))\nu(\mathrm{d}x). \tag{2}$$

It is noted that $\varphi(\cdot)$ is a convex (therefore continuous) function on $[0, \infty)$ with $\varphi(0) = 0$. X is not a subordinator (i.e., a nondecreasing Lévy process) iff $\varphi(\alpha) \to \infty$. Also, X is a subordinator iff $\sigma^2 = 0$, $\int_{(0,1]} x\nu(\mathrm{d}x) < \infty$ and $\tilde{c} := c - \int_{(0,1]} x\nu(\mathrm{d}x) \geq 0$. In this case,

$$\varphi(\alpha) = -\left(\tilde{c}\alpha + \int_{(0,\infty)} (1 - e^{-\alpha x})\nu(\mathrm{d}x)\right). \tag{3}$$

Therefore, when X is not a subordinator, which in this section is assumed without further mention, then either $\varphi(\cdot)$ is strictly increasing on $[0, \infty)$ or it first decreases to a negative value and then increases to infinity. For each $\beta \geq 0$, we denote by $\psi(\cdot)$ the *right-inverse* of $\varphi(\cdot)$:

$$\psi(\beta) = \inf\{\alpha | \varphi(\alpha) > \beta\}, \tag{4}$$

that is, $\psi(\beta)$ is the largest nonnegative root of $\varphi(\alpha) - \beta$.

2.2.2. Lévy-Driven Workload Process. We proceed by introducing the concept of a *Lévy-driven queue*, which can be seen as the Lévy process reflected at zero. To this end, we define (Asmussen 2003, section IX.2)

$$L(t) := \max\left(-\inf_{0\leq s\leq t} X(s), W(0)\right), \quad W(t) := X(t) + L(t), \tag{5}$$

with $W(0)$ the initial workload level; see (1). Here the nondecreasing process $L = \{L(t) | t \geq 0\}$ is often referred to as the *local time process of W at zero* and the nonnegative process $W = \{W(t) | t \geq 0\}$ as the *reflected process* associated with X (or, equivalently, the *workload process* with net input process X). It is well known that L is the unique nondecreasing and (because X has no negative jumps) continuous process with $L(0) = 0$, satisfying the (Skorokhod) conditions: $W(t) \geq 0$ and $\int_{[0,\infty)} W(s)L(\mathrm{d}s) = 0$. Evidently, in case $W(0) = 0$, there is no need to maximize with zero.

2.2.3. Fluctuation-Theoretic Identities. Let T_β be an exponentially distributed random variable with mean β^{-1}, independent of the Lévy process X. Denote by $\mathbb{E}_x$ the expected value when it is assumed that $W(0) = x \geq 0$. Now consider the following three stopping times:

$$\tau_b := \inf\{t | W(t) > b\}, \quad \sigma_0 := \inf\{t | W(t) = 0\}, \quad \sigma(x) := \inf\{t | X(t) \leq -x\}. \tag{6}$$

Define

$$\eta(\alpha, \beta) := \frac{\beta}{\beta - \varphi(\alpha)}, \quad \xi(\alpha, \beta) := \frac{\alpha}{\psi(\beta)}\, \eta(\alpha, \beta).$$

Directly from the definition of the Laplace exponent, we have

$$\mathbb{E}e^{-\alpha X(T_\beta)} = \int_0^\infty \beta e^{-\beta t}\, e^{\varphi(\alpha)t} \mathrm{d}t = \eta(\alpha, \beta). \tag{7}$$

Now consider the hitting time of level 0, by (6) written as σ_0, as a function of the initial value $W(0) = x$. As can be found in Kyprianou (2013, section VIII.1), $\{\sigma(x) | x \geq 0\}$ is a subordinator with Laplace exponent $\psi(\beta)$, so that

$$\mathbb{E}_x e^{-\beta\sigma_0} = \mathbb{E}e^{-\beta\sigma(x)} = e^{-\psi(\beta)x}, \tag{8}$$

where the first equality is evident. Another key result is that the LST of $W(T_\beta)$ allows an explicit expression, which is a mixture of two functions that are exponential in the initial level x. Indeed,

$$\mathbb{E}_x e^{-\alpha W(T_\beta)} = \eta(\alpha, \beta)\, e^{-\alpha x} - \xi(\alpha, \beta)\, e^{-\psi(\beta)x}; \tag{9}$$

see, for example, Dębicki and Mandjes (2015, section IV.1). This result is sometimes referred to as the *time-dependent generalized Pollaczek-Khinchine formula* for spectrally positive Lévy processes. We also note that when $\varphi'(0) > 0$, then the workload process converges in distribution to $M \sim \sup_{s\geq 0} X(s)$, regardless of the value of the initial condition $W(0)$, with LST

$$\mathbb{E}e^{-\alpha M} = \lim_{\beta\downarrow 0} \mathbb{E}_x e^{-\alpha W(T_\beta)} = \frac{\varphi'(0)\,\alpha}{\varphi(\alpha)}; \tag{10}$$

See, for example, Asmussen (2003, corollary IX.3.4). This result can be considered as the *generalized Pollaczek-Khinchine formula*.

2.2.4. First Passage Time of a Lévy-Driven Workload Process. We finally discuss an LST involving the reflected process at the first passage time τ_b. It identifies the joint LST of the value of the reflected process the first time it exceeds b, jointly with the time that this happens. The object of interest is thus

$$\Lambda_x(\alpha,\beta) := \mathbb{E}_x e^{-\alpha W(\tau_b)-\beta\tau_b} = \mathbb{E}_x e^{-\alpha W(\tau_b)} 1_{\{T_\beta>\tau_b\}}; \tag{11}$$

Here, the last equality can be verified by direct computation. In Avram et al. (2004), this object was expressed in terms of a general class of scale functions (see also Kyprianou 2013, theorem VIII.10). The precise form is not relevant in the context of the present paper; later in this work, we argue that in special cases it can be evaluated in explicit terms.

2.3. Notation

Throughout our analysis, we assume that the processes $X^\pm$ are spectrally positive Lévy processes, which are not subordinators and start at zero, and X^+ is not a deterministic nonincreasing drift, having Laplace exponents $\varphi^\pm(\cdot)$ with right-inverses $\psi^\pm(\cdot)$. We also put

$$\eta^\pm(\alpha,\beta) := \frac{\beta}{\beta-\varphi^\pm(\alpha)}, \quad \xi^\pm(\alpha,\beta) := \frac{\alpha}{\psi^\pm(\beta)}\,\eta^\pm(\alpha,\beta).$$

We define W^+, $\Lambda_x^+(\alpha,\beta)$, and τ_b^+ as W, $\Lambda_x(\alpha,\beta)$, and τ_b, respectively, but then with respect to X^+ rather than X. Likewise, we define σ_0^- and $\sigma^-(x)$ as σ_0 and $\sigma(x)$, respectively, but then with respect to X^- rather than X.

We will employ $\mathbb{P}_x^\pm$ and $\mathbb{E}_x^\pm$ to denote the probability and expectation conditional on $Q(0)=x$ and $J(0)=\pm$ (where $x\in[0,b]$ if $J(0)=+$ and $x\geq 0$ if $J(0)=-$).

We will use the notations $\mathbb{P}_x^\pm$ and $\mathbb{E}_x^\pm$ only when Q appears, and otherwise, the qualifiers $\pm$ will not appear. For example $\mathbb{E}_x^\pm e^{-\alpha Q(T_\beta)}$ or $\mathbb{P}_x^\pm(Q(t)\in \mathrm{d}x)$ as opposed to $\mathbb{E}_x e^{-\alpha W^+(\tau_b^+)-\beta\tau_b^+}$ or $\mathbb{E}_x e^{-\beta\sigma_0^-}$. Also, when in addition the starting point is not relevant or we discuss some general result, then, as already done earlier, we will write $\mathbb{E}$ without any upper or lower index.

3. Workload at an Exponentially Distributed Time

The objective of this section is to identify the LST of the workload at an independently sampled exponentially distributed time, for each of the two possible initial modes, starting at workload level x. Indeed, we succeed in expressing

$$\kappa_x^\pm(\alpha,\beta) := \mathbb{E}_x^\pm e^{-\alpha Q(T_\beta)}$$

in terms of the model primitives. Here T_β is an exponentially distributed time (with mean β^{-1}), independent of the driving processes X^+ and X^-. In our analysis, we extensively use the fluctuation-theoretic results for spectrally positive Lévy processes as given in Section 2.2, in particular, the LST of first hitting times of the driving Lévy processes (8), the LST of the workload at an exponentially distributed time (9), and the LST of the first hitting time of the workload process (11).

It is noted that the LST $\kappa_x^\pm(\alpha,\beta)$ is effectively a *double* transform because it can be written as

$$\kappa_x^\pm(\alpha,\beta) = \int_0^\infty\int_0^\infty \beta e^{-\alpha y-\beta t}\,\mathbb{P}_x^\pm(Q(t)\in \mathrm{d}y)\,\mathrm{d}t.$$

By Laplace inversion (Abate and Whitt 1995), one can thus numerically evaluate the density of $Q(t)$ when $\kappa_x^\pm(\alpha,\beta)$ has been determined. In particular the techniques developed in Den Iseger (2006) lend themselves to multidimensional Laplace inversion; see Asghari et al. (2014) and Den Iseger et al. (2013) for more specific implementation guidelines.

The following straightforward fact is the starting point of our analysis.

Lemma 1. *Let Y be a regenerative process with first regeneration time τ and let Y_d be a delayed regenerative process with first regeneration time τ_d, such that $\{Y_d(\tau_d+t)|t\geq 0\}\sim Y$. Then, for T_β being exponentially distributed with mean β^{-1}, independent of everything else, we have that*

$$\mathbb{E}e^{-\alpha Y_d(T_\beta)} = \mathbb{E}e^{-\alpha Y_d(T_\beta)}1_{\{T_\beta\leq\tau_d\}} + \mathbb{E}e^{-\beta\tau_d}\,\mathbb{E}e^{-\alpha Y(T_\beta)}, \tag{12}$$

$$\mathbb{E}e^{-\alpha Y(T_\beta)} = \frac{\mathbb{E}e^{-\alpha Y(T_\beta)}1_{\{T_\beta\leq\tau\}}}{1-\mathbb{E}e^{-\beta\tau}} = \frac{\mathbb{E}e^{-\alpha Y(T_\beta)}1_{\{T_\beta\leq\tau\}}}{\mathbb{P}(T_\beta\leq\tau)}. \tag{13}$$

Proof. The proof of this lemma is elementary. The second term on the right of the top equation represents $\mathbb{E}e^{-\alpha Y_d(T_\beta)}1_{\{T_\beta>\tau_d\}}$, which is seen to equal $\mathbb{E}e^{-\beta\tau_d}\,\mathbb{E}e^{-\alpha Y(T_\beta)}$ upon observing that

$$\mathbb{P}(T_\beta>\tau_d) = \mathbb{E}\mathrm{e}^{-\beta\tau_d},\quad [Y_d(T_\beta)|T_\beta>\tau_d]\sim Y(T_\beta).$$

It is also directly seen that the top equation becomes the bottom one when $(Y_d,\tau_d)=(Y,\tau)$. The second equality in (13) is a direct consequence of the identity $\mathbb{E}e^{-\beta\tau}=\mathbb{P}(T_\beta>\tau)$. □

We note that Lemma 1 remains valid when τ or τ_d are not a.s. finite and the definitions of *regenerative* and *delayed regenerative* are extended. That is, only on $\tau<\infty$ ($\tau_d<\infty$, respectively) the process regenerates at time τ (τ_d, respectively,) and its future is independent of the past.

It is directly seen that the system regenerates at times when the mode switches from $-$ to $+$, which brings us into the setting of the above lemma. In the sequel, we simply write $\kappa(\alpha,\beta)=\kappa_0^+(\alpha,\beta)$ for the LST for the (nondelayed) regenerative process, that is, the process starting at time 0 at workload level 0 and being in mode +. In the remainder of this section, we determine these quantities. We break up the underlying computations into the following two steps.

(i) The simplest case is the one in which we start from some workload level $x>0$ in the minus mode. Then until the beginning of a new regeneration epoch the process behaves like X^- starting from x, after which it behaves like Q starting from zero (which corresponds to the nondelayed case). Therefore, by (12) in Lemma 1,

$$\begin{aligned}\kappa_x^-(\alpha,\beta) &= \mathbb{E}_x^- e^{-\alpha Q(T_\beta)}1_{\{T_\beta\leq\sigma_0^-\}} + \mathbb{E}_x e^{-\beta\sigma_0^-}\,\kappa(\alpha,\beta)\\ &= \mathbb{E}e^{-\alpha(x+X^-(T_\beta))}1_{\{T_\beta\leq\sigma^-(x)\}} + \mathbb{E}_x e^{-\beta\sigma_0^-}\,\kappa(\alpha,\beta)\\ &= e^{-\alpha x}\,\mathbb{E}e^{-\alpha X^-(T_\beta)} - e^{-\alpha x}\,\mathbb{E}e^{-\alpha X^-(T_\beta)}1_{\{T_\beta>\sigma^-(x)\}} + \mathbb{E}_x e^{-\beta\sigma_0^-}\,\kappa(\alpha,\beta),\end{aligned} \tag{14}$$

where the second equality reflects that on $\{T_\beta\leq\sigma_0^-\}$ the process Q behaves as $x+X^-$. By (7), we have that the first term on the right-hand side equals $\eta^-(\alpha,\beta)e^{-\alpha x}$, and by (8), the third term equals $\kappa(\alpha,\beta)e^{-\psi^-(\beta)x}$, so we are left with evaluating the second term. Also, because $X^-(\sigma^-(x))=-x$ by the spectrally positive nature of the process X^-, it follows by the memoryless property of T_β and the strong Markov property for X^- that

$$e^{-\alpha x}\,\mathbb{E}e^{-\alpha X^-(T_\beta)}1_{\{T_\beta>\sigma^-(x)\}} = \mathbb{E}_x e^{-\beta\sigma_0^-}\,\mathbb{E}e^{-\alpha X^-(T_\beta)} = \eta^-(\alpha,\beta)\,e^{-\psi^-(\beta)x}. \tag{15}$$

Upon combining the above findings, we conclude that, for $\alpha\neq\psi^-(\beta)$,

$$\kappa_x^-(\alpha,\beta) = \eta^-(\alpha,\beta)(e^{-\alpha x}-e^{-\psi^-(\beta)x}) + \kappa(\alpha,\beta)e^{-\psi^-(\beta)x}, \tag{16}$$

whereas for $\alpha=\psi^-(\beta)$, Bernoulli-l'Hôpital's rule implies that the first term on the right-hand side becomes $\beta x e^{-\psi^-(\beta)x}/(\varphi^-)'(\psi^-(\beta))$.

(ii) Consider now the case in which we start from some workload level $x\in[0,b]$ in the + mode. Then, because Q behaves like W^+ until time τ_b^+, by the memoryless property for T_β and the strong Markov property of the process (Q,J), we have that

$$\begin{aligned}\kappa_x^+(\alpha,\beta) &= \mathbb{E}_x^+ e^{-\alpha Q(T_\beta)}1_{\{T_\beta\leq\tau_b^+\}} + \mathbb{E}_x^+ e^{-\alpha Q(T_\beta)}1_{\{T_\beta>\tau_b^+\}}\\ &= \mathbb{E}_x e^{-\alpha W^+(T_\beta)}1_{\{T_\beta\leq\tau_b^+\}} + \mathbb{E}_x[e^{-\beta\tau_b^+}\kappa^-_{W^+(\tau_b^+)}(\alpha,\beta)].\end{aligned} \tag{17}$$

To identify the terms on the right-hand side, we start by observing that the first term can be rewritten as

$$\mathbb{E}_x e^{-\alpha W^+(T_\beta)}1_{\{T_\beta\leq\tau_b^+\}} = \mathbb{E}_x e^{-\alpha W^+(T_\beta)} - \mathbb{E}_x e^{-\alpha W^+(T_\beta)}1_{\{T_\beta>\tau_b^+\}}. \tag{18}$$

From (9), the first term on the right-hand side of (18) is

$$\eta^+(\alpha,\beta)\,e^{-\alpha x} - \xi^+(\alpha,\beta)\,e^{-\psi^+(\beta)x}. \tag{19}$$

Again using the fact that T_β is exponentially distributed and relying on (9) and (11), the second term on the right-hand side of (18) is seen to equal

$$\mathbb{E}_x e^{-\beta\tau_b^+}\mathbb{E}_{W^+(\tau_b^+)}e^{-\alpha W^+(\tau_b^+ + T_\beta)} \tag{20}$$

$$= \mathbb{E}_x[e^{-\beta\tau_b^+}(\eta^+(\alpha,\beta)e^{-\alpha W^+(\tau_b^+)} - \xi^+(\alpha,\beta)e^{-\psi^+(\beta)W^+(\tau_b^+)})]$$

$$= \eta^+(\alpha,\beta)\Lambda_x^+(\alpha,\beta) - \xi^+(\alpha,\beta)\Lambda_x^+(\psi^+(\beta),\beta). \tag{21}$$

We are left with evaluating the second term on the right-hand side of (17). By (16), it equals

$$\begin{aligned}&\mathbb{E}_x[e^{-\beta\tau_b^+}(\eta^-(\alpha,\beta)(e^{-\alpha W^+(\tau_b^+)} - e^{-\psi^-(\beta)W^+(\tau_b^+)}) + \kappa(\alpha,\beta)\,e^{-\psi^-(\beta)W^+(\tau_b^+)})]\\ &\quad = \eta^-(\alpha,\beta)(\Lambda_x^+(\alpha,\beta) - \Lambda_x^+(\psi^-(\beta),\beta)) + \Lambda_x^+(\psi^-(\beta),\beta)\,\kappa(\alpha,\beta).\end{aligned} \tag{22}$$

Now inserting all derived expressions into (17), we arrive at

$$\begin{aligned}\kappa_x^+(\alpha,\beta) &= \eta^+(\alpha,\beta)\,e^{-\alpha x} - \xi^+(\alpha,\beta)\,e^{-\psi^+(\beta)x} - \eta^+(\alpha,\beta)\,\Lambda_x^+(\alpha,\beta) + \xi^+(\alpha,\beta)\,\Lambda_x^+(\psi^+(\beta),\beta)\\ &\quad + \eta^-(\alpha,\beta)(\Lambda_x^+(\alpha,\beta) - \Lambda_x^+(\psi^-(\beta),\beta)) + \Lambda_x^+(\psi^-(\beta),\beta)\,\kappa(\alpha,\beta).\end{aligned} \tag{23}$$

The last step is to identify the LST $\kappa(\alpha,\beta)$ that corresponds to the nondelayed regenerative process Q (i.e., the one starting from workload level zero and the plus mode). Inserting $x = 0$ in (22), the left-hand side of (23) is $\kappa(\alpha,\beta)$, which can then be solved from the equation. This eventually leads to the following expression: For $\alpha \notin \{\psi^+(\beta),\psi^-(\beta)\}$, with $\zeta(\beta) := 1 - \Lambda_0^+(\psi^-(\beta),\beta)$,

$$\begin{aligned}\kappa(\alpha,\beta) &= \frac{1}{\zeta(\beta)}\left(1 - \Lambda_0^+(\alpha,\beta) - \frac{\alpha}{\psi^+(\beta)}(1 - \Lambda_0^+(\psi^+(\beta),\beta))\right)\frac{\beta}{\beta - \varphi^+(\alpha)}\\ &\quad + \frac{1}{\zeta(\beta)}(\Lambda_0^+(\alpha,\beta) - \Lambda_0^+(\psi^-(\beta),\beta))\frac{\beta}{\beta - \varphi^-(\alpha)}.\end{aligned} \tag{24}$$

It is noted that inserting $\varphi^-(\cdot) = \varphi^+(\cdot) = \varphi(\cdot)$ into (24) provides a nice sanity check: We obtain the well-known expression

$$\kappa(\alpha,\beta) = \frac{\beta}{\psi(\beta)}\frac{\psi(\beta) - \alpha}{\beta - \varphi(\alpha)};$$

see, for example, Kyprianou (2013, section VI.5.2).

Remark 1. We noticed that the process (Q, J) regenerates at subsequent epochs where the mode process J switches from $-$ to $+$. The above computations implicitly reveal the LST of the time until a new regenerative cycle begins, starting from $x \in [0, b]$ in the plus mode:

$$\mathbb{E}_x[e^{-\beta\tau_b^+}\,\mathbb{E}[e^{-\beta\sigma^-(W^+(\tau_b^+))}\,|\,W^+(\tau_b^+)]] = \mathbb{E}_x e^{-\beta\tau_b^+ - \psi^-(\beta)W^+(\tau_b^+)} = \Lambda_x^+(\psi^-(\beta),\beta).$$

In particular, $\Lambda_0^+(\psi^-(\beta),\beta)$ denotes the LST of a regeneration cycle. This means that $\zeta(\beta) = 1 - \Lambda_0^+(\psi^-(\beta),\beta)$ can be interpreted as the probability that such a regeneration cycle is larger than the exponentially distributed random variable T_β. Note that, thus far, there was no need for an assumption that ensures that the time until regeneration is a.s. finite.

We now state the conclusion of the above lines of computations.

Theorem 1. *Assume that X^+ is not a subordinator. Then $\kappa(\alpha,\beta)$ is given by (24), for any $\alpha \notin \{\psi^+(\beta),\psi^-(\beta)\}$,*

$$\kappa_x^+(\alpha,\beta) = \left(e^{-\alpha x} - \Lambda_x^+(\alpha,\beta) - \frac{\alpha}{\psi^+(\beta)}(e^{-\psi^+(\beta)x} - \Lambda_x^+(\psi^+(\beta),\beta))\right)\frac{\beta}{\beta-\varphi^+(\alpha)}$$
$$+ (\Lambda_x^+(\alpha,\beta) - \Lambda_x^+(\psi^-(\beta),\beta))\frac{\beta}{\beta-\varphi^-(\alpha)} + \Lambda_x^+(\psi^-(\beta),\beta)\,\kappa(\alpha,\beta), \tag{25}$$

and

$$\kappa_x^-(\alpha,\beta) = (e^{-\alpha x} - e^{-\psi^-(\beta)x})\frac{\beta}{\beta-\varphi^-(\alpha)} + \kappa(\alpha,\beta)\,e^{-\psi^-(\beta)x}. \tag{26}$$

Observe that for $\alpha \in \{\psi^+(\beta),\psi^-(\beta)\}$, we can identify $\kappa(\alpha,\beta)$ by a straightforward application of the Bernoulli-l'Hôpital rule; we omit the details.

Remark 2. Upon inspection of the proof underlying Theorem 1, it is readily verified that it is possible to derive more refined results. For instance, one can provide an expression for (starting at workload level $x \in [0,b]$ and the mode being +) the joint LST of (i) the value of Q at the first time J switches from + to −, (ii) this first time J switches from + to −, and (iii) the length of the subsequent interval in which the mode was −. With the parameters α, β, and γ corresponding to these random variables (i), (ii), and (iii), respectively, this LST can be written as

$$\mathbb{E}_x[e^{-\alpha W^+(\tau_b^+)-\beta\tau_b^+}\mathbb{E}[e^{-\gamma\sigma^-(W^+(\tau_b^+))} \mid W^+(\tau_b^+)]] = \mathbb{E}_x[e^{-\alpha W^+(\tau_b^+)-\beta\tau_b^+}e^{-\psi^-(\gamma)W^+(\tau_b^+)}]$$
$$= \Lambda_x^+(\alpha+\psi^-(\gamma),\beta). \tag{27}$$

Remark 3. We have ruled out the case that X^+ is a subordinator. However, it can be observed that our analysis extends naturally to this setting. The only two necessary modifications are as follows.

i. Replace the expression for $\Lambda^+(\alpha,\beta)$ found in Avram et al. (2004) by its analogue in which the scale functions are substituted by the corresponding potential (renewal) measures; see, for instance, Kyprianou (2013, exercise 5.7) (with the parameters α and β being swapped).

ii. In order to obtain the counterparts of (19) and (21), apply (7) rather than (9). As a consequence, in the counterpart (19), we find only the first term, that is, the one proportional to $\eta^+(\alpha,\beta)$; the same applies to the counterpart of (21).

We also refer to Kella (1998) for a related study that considers, within this context of subordinators, decomposition results in a somewhat more general framework; notably, it allows for general positive finite-mean stopping times, rather than just τ_b^+.

4. The Workload in Stationarity

Whereas in Section 3 we focused on the LST of the workload at an exponentially distributed time, we now focus on its counterpart for the corresponding ergodic distribution (being the long-run average limiting distribution). In Section 4.1, we derive the LST of the stationary workload, whereas in Section 4.2, we provide an insightful interpretation of the obtained expressions.

From here on, it proves useful to denote $m^{\pm} := (\varphi^{\pm})'(0)$.

4.1. LST of Stationary Workload

For any regenerative process Y with first regeneration epoch τ having a finite mean, such a stationary distribution always exists, and if Y^* is a random variable having this distribution then

$$\mathbb{E}f(Y^*) = \frac{1}{\mathbb{E}\tau}\mathbb{E}\int_0^\tau f(Y(s))\mathrm{d}s, \tag{28}$$

for sufficiently nice functions $f(\cdot)$; in particular, (28) holds for bounded continuous (or even just Borel) functions. Hence, when taking $f(x) = e^{-\alpha x}$, the LST of Y^* is given by

$$\mathbb{E}e^{-\alpha Y^*} = \frac{1}{\mathbb{E}\tau}\mathbb{E}\int_0^\tau e^{-\alpha Y(s)}\,\mathrm{d}s. \tag{29}$$

We concentrate on the *long-run average* limiting distribution, thus avoiding the tedious case in which τ has an arithmetic distribution. Indeed, if τ does not have an arithmetic distribution, then $Y(t)$ has a limiting distribution as $t \to \infty$.

Recall that the process (Q, J) regenerates when the jump process has a transition from $-$ to $+$, where it is noted that at these times the workload level is necessarily zero. We impose from now on the condition $m^- > 0$: The driving process X^- has a negative mean. This guarantees that the regeneration time of Q starting from the plus mode from workload level zero has a finite expected value provided that $\mathbb{E}_0 W^+(\tau_b^+) < \infty$. This follows by recalling that x/m^- is the mean time it takes the process W^- to hit zero when starting from level x, so that the mean regeneration time equals

$$\mathbb{E}_0 \tau_b^+ + \frac{\mathbb{E}_0 W^+(\tau_b^+)}{m^-}, \tag{30}$$

where it also noted that, unless X^+ is a nonincreasing drift, $\mathbb{E}_0 \tau_b^+$ is always finite (Appendix A, Lemma A.1). Also, $\mathbb{E}_0 W(\tau_b^+)$ is finite provided that $\int_{(1,\infty)} x\nu^+(\mathrm{d}x) < \infty$, where ν^+ is the Lévy measure associated with the process X^+ (Appendix A, Lemma A.2).

The goal of this section is to identify the LST of the stationary workload, denoted by $\kappa(\alpha)$ in the sequel. This LST could be found from the expression given in Theorem 1 by letting $\beta \downarrow 0$ in $\kappa_x^{\pm}(\alpha, \beta)$. It is somewhat simpler, however, to follow this limiting procedure applied to $\kappa(\alpha, \beta)$ as given in (24). We start by defining

$$\ell_1(\alpha, \beta) := -\frac{\partial}{\partial \alpha} \Lambda_0^+(\alpha, \beta), \quad \ell_2(\alpha, \beta) := -\frac{\partial}{\partial \beta} \Lambda_0^+(\alpha, \beta).$$

Observing that $\psi^-(0) = 0$ due to $m^- > 0$, and in addition, recalling that $\zeta(\beta) = 1 - \Lambda_0^+(\psi^-(\beta), \beta)$, we thus find that

$$\kappa(\alpha) = \frac{1}{\zeta'(0)} \left(\frac{1 - \Lambda_0^+(\alpha, 0)}{\varphi^-(\alpha)} - \frac{1 - \Lambda_0^+(\alpha, 0)}{\varphi^+(\alpha)} \right) + \frac{1}{\zeta'(0)} \frac{\alpha}{\varphi^+(\alpha)} \lim_{\beta \downarrow 0} \left(\frac{1 - \Lambda_0^+(\psi^+(\beta), \beta)}{\psi^+(\beta)} \right).$$

We have to distinguish between the cases $\psi^+(0) = 0$ and $\psi^+(0) > 0$. In both cases, we obtain an expression for $\kappa(\alpha)$, where it is noted that the former case requires an elementary application of Bernoulli-l'Hôpital's rule, recalling that $(\psi^+)'(0) = 1/m^+$. The following theorem states the result. Define

$$\mathscr{L}(\alpha) := \ell_1(\alpha, 0) + \ell_2(\alpha, 0)\, m^+.$$

It is readily verified that, using the compact notation $\ell_i := \ell_i(0,0)$ for $i = 1, 2$,

$$\zeta'(0) = \ell_1\, (\psi^-)'(0) + \ell_2. \tag{31}$$

Observe that $\ell_1 = \mathbb{E}_0 W^+(\tau_b^+)$ and $\ell_2 = \mathbb{E}_0 \tau_b^+$.

Theorem 2. *Assume that $X^{\pm}$ are not subordinators, X^+ is not a nonincreasing drift, and $m^- > 0$. Then, for any $\alpha \ge 0$, if $\psi^+(0) = 0$ (which is the case iff $m^+ \ge 0$), then the LST of the stationary workload is given by*

$$\kappa(\alpha) = \frac{1}{\zeta'(0)} \left(\frac{1 - \Lambda_0^+(\alpha, 0)}{\varphi^-(\alpha)} - \frac{1 - \Lambda_0^+(\alpha, 0)}{\varphi^+(\alpha)} \right) + \frac{1}{\zeta'(0)} \frac{\alpha}{\varphi^+(\alpha)} \mathscr{L}(0),$$

and if $\psi^+(0) > 0$ (which is the case iff $m^+ < 0$), then

$$\kappa(\alpha) = \frac{1}{\zeta'(0)} \left(\frac{1 - \Lambda_0^+(\alpha, 0)}{\varphi^-(\alpha)} - \frac{1 - \Lambda_0^+(\alpha, 0)}{\varphi^+(\alpha)} \right) + \frac{1}{\zeta'(0)} \frac{\alpha}{\varphi^+(\alpha)} \frac{1 - \Lambda_0^+(\psi^+(0), 0))}{\psi^+(0)}.$$

There is an alternative, insightful manner to arrive at the result of Theorem 2. It separately identifies the LST $\kappa^+(\alpha)$ that corresponds to the stationary workload conditioned on being in the + mode and $\kappa^-(\alpha)$ that is its counterpart for the $-$ mode.

We first analyze $\kappa^+(\alpha)$. The starting point is the following evident identity:

$$\mathscr{F}(\alpha, \beta) := \mathbb{E}_0 \int_0^{\tau_b^+} e^{-\alpha W^+(s) - \beta s}\, \mathrm{d}s = \mathbb{E}_0 \int_0^{\infty} e^{-\alpha W^+(s) - \beta s}\, 1_{\{\tau_b^+ \ge s\}} \mathrm{d}s$$
$$= \frac{1}{\beta}\, \mathbb{E}_0 [e^{-\alpha W^+(T_\beta)} 1_{\{\tau_b^+ \ge T_\beta\}}].$$

Using this relation, an application of (21) yields that

$$\mathscr{F}(\alpha,\beta) = \left(1 - \Lambda_0^+(\alpha,\beta) - \frac{\alpha}{\psi^+(\beta)}(1 - \Lambda_0^+(\psi^+(\beta),\beta))\right)\frac{1}{\beta - \varphi^+(\alpha)}. \tag{32}$$

Inserting $\alpha = 0$ gives

$$\mathscr{F}(0,\beta) = \mathbb{E}_0 \int_0^{\tau_b^+} e^{-\beta s}\mathrm{d}s = \frac{1 - \mathbb{E}_0 e^{-\beta\tau_b^+}}{\beta} = \frac{1 - \Lambda_0^+(0,\beta)}{\beta}, \tag{33}$$

which we can also infer from Remark 2, upon taking $x = \alpha = \gamma = 0$, observing that because $m^- > 0$, we have $\psi^-(0) = 0$. Therefore, by a straightforward computation,

$$\frac{\mathscr{F}(\alpha,\beta)}{\mathscr{F}(0,\beta)} = \left(\frac{1 - \Lambda_0^+(\alpha,\beta)}{1 - \Lambda_0^+(0,\beta)} - \frac{\alpha}{\psi^+(\beta)}\frac{1 - \Lambda_0^+(\psi^+(\beta),\beta)}{1 - \Lambda_0^+(0,\beta)}\right)\frac{\beta}{\beta - \varphi^+(\alpha)}. \tag{34}$$

Now $\kappa^+(\alpha)$ follows by sending β to zero in (34). In the case that $\psi^+(0) = 0$, we have

$$\kappa^+(\alpha) = \left(\mathscr{L}(0) - \frac{1 - \Lambda_0^+(\alpha,0)}{\alpha}\right)\frac{\alpha}{\ell_2\,\varphi^+(\alpha)}, \tag{35}$$

$$\kappa^+(\alpha) = \left(\frac{1 - \Lambda_0^+(\psi^+(0),0)}{\psi^+(0)} - \frac{1 - \Lambda_0^+(\alpha,0)}{\alpha}\right)\frac{\alpha}{\ell_2\,\varphi^+(\alpha)}. \tag{36}$$

We continue by analyzing $\kappa^-(\alpha)$. To this end, we first consider our spectrally positive negative mean Lévy process such that whenever the process hits zero it jumps by an independent (from the Lévy process, that is) random nonnegative amount distributed like some V. For such a process, the steady-state distribution is that of a convolution of two distributions. The first is the steady-state distribution of a reflected Lévy process and the other is the stationary excess lifetime distribution associated with V (Kella and Whitt 1991, theorem 4.3; Ivanovs and Kella 2013, theorem 1). Therefore, recognizing that setup, the random amount that the minus mode starts with is distributed like $W^+(\tau_b^+)$, the LST of the process conditioned on being in the second phase is given by

$$\kappa^-(\alpha) = \frac{m^-\,\alpha}{\varphi^-(\alpha)}\,\frac{1 - \mathbb{E}_0 e^{-\alpha W^+(\tau_b^+)}}{\alpha\,\mathbb{E}_0 W^+(\tau_b^+)} = \frac{m^-\,\alpha}{\varphi^-(\alpha)}\,\frac{1 - \Lambda_0^+(\alpha,0)}{\alpha\,\ell_1}. \tag{37}$$

Noting that the expected time consecutively spent in the + mode is $\mathbb{E}_0\tau_b^+ = \ell_2$, and its counterpart in the − mode is

$$\mathbb{E}_0 W^+(\tau_b^+)/m^- = \ell_1/m^- = \ell_1\,(\psi^-)'(0),$$

we can compute the long-run fractions of time spent in each of the modes. We finally have that the steady-state LST of Q is given by

$$\kappa(\alpha) = \frac{\ell_2}{\ell_1\,(\psi^-)'(0) + \ell_2}\kappa^+(\alpha) + \frac{\ell_1\,(\psi^-)'(0)}{\ell_1\,(\psi^-)'(0) + \ell_2}\kappa^-(\alpha). \tag{38}$$

Because by (31) the denominators are $\zeta'(0)$, it is consistent with the result found in Theorem 2.

Remark 4. Bekker (2009) considers the same model, except that his process switches from the + mode to the − mode when a threshold m_1 is up-crossed, and it switches back to the + mode when a threshold $m_2 \le m_1$ is down-crossed (extending Cohen's model (Cohen 1976) where $m_1 = m_2$). He only considers the stationary workload distribution. Using a martingale approach, he expresses that distribution, for both modes, into scale functions.

4.2. Interpretation via a Decomposition Result

In this section, we provide an interpretation of the form of $\kappa^+(\alpha)$ in the case that $m^+ > 0$, as given in (35). It follows from a fairly general decomposition result, which we present below. In the setup considered in this section, we let W be a reflected Lévy process associated with a nondeterministic spectrally positive Lévy process X having a Laplace exponent $\varphi(\cdot)$ that satisfies $\varphi'(0) > 0$.

Defining $\overline{c} := \int_{(1,\infty)} x\,\nu(\mathrm{d}x)$, because $\varphi'(0) = -c - \overline{c}$, we have that $\varphi'(0) > 0$ implies that $\overline{c} < -c$, so that $\overline{c}$ cannot be infinite and c is negative. Therefore, here, *nondeterministic* X means that the case $X(t) = ct$ for $c < 0$ and all $t \ge 0$

is excluded. It thus follows that Lemmas A.1 and A.2 imply that $\mathbb{E}\tau_b$, $\mathbb{E}W(\tau_b)$ are both finite (and clearly positive), where $\tau_b := \inf\{t \mid W(t) > b\}$.

Suppose that

◦ M has the stationary distribution associated with W (see (10)),

◦ M_b has the stationary distribution associated with the regenerative process that behaves like W until time τ_b and then restarts from zero at time τ_b,

◦ Y_b has the excess lifetime distribution of $W(\tau_b)$, that is, it has the density

$$f_{Y_b}(t) = \frac{\mathbb{P}(W(\tau_b) > t)}{\mathbb{E}W(\tau_b)} 1_{[0,\infty)}(t),$$

and LST,

$$\mathbb{E}e^{-\alpha Y_b} = \frac{1 - \mathbb{E}e^{-\alpha W(\tau_b)}}{\alpha\, \mathbb{E}W(\tau_b)}. \tag{39}$$

◦ V_b has the stationary distribution of the waiting time and virtual waiting time of an M/G/1 queue with service times distributed like $W(\tau_b)$ and arrival rate

$$\lambda_b := \frac{1}{\varphi'(0)\,\mathbb{E}\tau_b + \mathbb{E}W(\tau_b)} < \frac{1}{\mathbb{E}W(\tau_b)}. \tag{40}$$

Hence, by the Pollaczek-Khinchine formula (see (10)),

$$\mathbb{E}e^{-\alpha V_b} = \frac{1 - \rho_b}{1 - \rho_b\, \mathbb{E}e^{-\alpha Y_b}}, \tag{41}$$

where

$$\rho_b := \lambda_b\, \mathbb{E}W(\tau_b) = \frac{\mathbb{E}W(\tau_b)}{\varphi'(0)\,\mathbb{E}\tau_b + \mathbb{E}W(\tau_b)}. \tag{42}$$

We are now ready to state the following decomposition result.

Theorem 3. *For every $b \in (0,\infty)$, with M_b and V_b being independent,*

$$M \sim M_b + V_b. \tag{43}$$

Proof. We define a regenerative cycle in the following way. We start at a time that the workload level is zero, and we end at the first time *after* τ_b at which the workload is zero again. This cycle can be broken into two parts: a part until time τ_b and a part from time τ_b until the workload hits zero, having expected lengths $\mathbb{E}\tau_b$ and $\mathbb{E}W(\tau_b)/\varphi'(0)$.

By regenerative theory, recalling (42), the long run fraction of time W spends at a pre-τ_b part is $1 - \rho_b$.

◦ The ergodic distribution of the regenerative process restricted to pre-τ_b parts is that of M_b.

◦ The ergodic distribution of the regenerative process restricted to post-τ_b parts is the distribution of a Lévy process with additional independent and identically distributed jump inputs distributed as $W(\tau_b)$ whenever the process hits zero. It is distributed as $M + Y_b$, where M and Y_b are independent (Kella and Whitt 1991, theorem 4.2).

The ergodic distribution of the regenerative process W in its first regenerative cycle is just the ergodic distribution of W, which is that of M. Therefore,

$$\mathbb{E}e^{-\alpha M} = (1 - \rho_b)\,\mathbb{E}e^{-\alpha M_b} + \rho_b\, \mathbb{E}e^{-\alpha M}\, \mathbb{E}e^{-\alpha Y_b}, \tag{44}$$

which is equivalent to

$$\mathbb{E}e^{-\alpha M} = \mathbb{E}e^{-\alpha M_b} \frac{1 - \rho_b}{1 - \rho_b\, \mathbb{E}e^{-\alpha Y_b}} = \mathbb{E}e^{-\alpha M_b}\, \mathbb{E}e^{-\alpha V_b}, \tag{45}$$

and the proof is complete. □

From (45) and (39), we have the following.

Corollary 1. *For any* $b > 0$,

$$\mathbb{E}e^{-\alpha M_b} = \frac{\varphi'(0)\,\alpha}{\varphi(\alpha)}\left(1 + \frac{\mathbb{E}e^{-\alpha W(\tau_b)} - 1 + \alpha\mathbb{E}W(\tau_b)}{\varphi'(0)\,\alpha\,\mathbb{E}\tau_b}\right). \tag{46}$$

We conclude this section by interpreting (35), which provides an expression for $\kappa^+(\alpha)$ in the case where $m^+ > 0$. Representation (35) basically is a translation of $\mathbb{E}e^{-\alpha M_b} = \mathbb{E}e^{-\alpha M}/\mathbb{E}e^{-\alpha V_b}$ (see Theorem 3 and (45)). More specifically, it is an immediate consequence of (46) upon identifying $\mathbb{E}e^{-\alpha M_b} = \kappa^+(\alpha)$, $W = W^+$, $\tau_b = \tau_b^+$, $\varphi(\cdot) = \varphi^+(\cdot)$ and dividing the term between large brackets in (35) by $\ell_2 m^+$.

5. Explicit Evaluation

In our expressions, a key component is $\Lambda_x^+(\alpha,\beta)$—the joint LST of the random variables $W^+(\tau_b^+)$ and τ_b^+, starting from x. As mentioned earlier, in Avram et al. (2004), this LST can be expressed in terms of scale functions, which are defined via their Laplace transforms. Although scale functions are explicitly known for certain classes of spectrally one-sided Lévy processes, in general, they cannot be written in closed form in terms of the model primitives (Kyprianou and Rivero 2008, Hubalek and Kyprianou 2011, Kuznetsov et al. 2012).

In Section 5.1, we consider X^+ to belong to a class of spectrally positive Lévy processes that allow for a relatively explicit analysis. Specifically, we focus on compound Poisson (CP) processes with phase-type (upward) jumps, perturbed by Brownian motion (BM). This class is particularly relevant for two reasons: (i) phase-type random variables can approximate any nonnegative random variable arbitrarily closely (Asmussen 2003, theorem III.4.2), and (ii) any spectrally positive Lévy process can be approximated arbitrarily well by CP processes perturbed by BM (Asmussen and Rosiński 2001). In Section 5.2, we consider two special cases: the case where X^+ corresponds to just BM and the case where X^+ corresponds to just CP with exponentially distributed jumps (so that W^+ behaves as the workload in an M/M/1 queue).

5.1. Superimposed CP with Phase-Type Jumps and BM

In the case where X^+ is a compound Poisson process with phase-type upward jumps, perturbed by a BM, we are able to evaluate $\Lambda_x^+(\alpha,\beta)$. The analysis relies on an application of the optional sampling theorem to the following zero-mean martingale $\mathscr{M}$, defined for each $\alpha,\beta \ge 0$. Let $L^+(\cdot)$ denote the local time of $W^+(\cdot)$ at zero (recall (5)) and assume $W^+(0) = x$. Now insert $Y_t := L^+(t) + (\beta/\alpha)\,t$ into Kella and Yor (2017, corolloary 1), to have

$$\mathscr{M}(t) := (\varphi^+(\alpha) - \beta)\int_0^t e^{-\alpha W^+(s) - \beta s}\,\mathrm{d}s + e^{-\alpha x} - e^{-\alpha W^+(t) - \beta t} - \alpha\int_0^t e^{-\beta s}\,\mathrm{d}L^+(s); \tag{47}$$

see also Asmussen (2014, section 5).

Let $\boldsymbol{s} \in \mathbb{R}^d$ be a probability row vector, $\mathbf{1} \in \mathbb{R}^d$ a column vector of ones and $T \in \mathbb{R}^{d\times d}$ the phase generator (a substochastic matrix with spectral radius strictly less than one). Then the tail of the phase-type distribution function associated with the upward jumps is given by $\boldsymbol{s}\,e^{Tt}\,\mathbf{1}$ for $t \ge 0$. When in addition the arrival rate of the phase-type distributed jumps is λ^+ and the perturbing BM has drift c^+ and variance coefficient $(\sigma^2)^+$, then, with $\boldsymbol{t} = -T\mathbf{1}$,

$$\varphi^+(\alpha) = -c^+\alpha - \lambda^+(1 - \boldsymbol{s}(\alpha I - T)^{-1}\boldsymbol{t}) + \frac{1}{2}(\sigma^2)^+\alpha^2.$$

The main idea is to apply an optional sampling theorem at the stopping time τ_b^+ (complemented by an application of the dominated and monotone convergence theorems) to obtain the identity $0 = \mathscr{M}(0) = \mathbb{E}_x\,\mathscr{M}(\tau_\ell^+)$. We conclude that

$$0 = (\varphi^+(\alpha) - \beta)\,\mathbb{E}_x\left[\int_0^{\tau_b^+} e^{-\alpha W^+(s) - \beta s}\mathrm{d}s\right] + e^{-\alpha x} - \Lambda_x^+(\alpha,\beta) - \alpha\,K(\beta), \tag{48}$$

with

$$K(\beta) := \mathbb{E}_x\left[\int_0^{\tau_b^+} e^{-\beta s}\mathrm{d}L^+(s)\right].$$

With $\boldsymbol{e}_i$ denoting the ith unit vector, let $\chi_i(\alpha) := \boldsymbol{e}_i\,(\alpha I - T)^{-1}\,\boldsymbol{t}$ be the LST of our phase-type distribution when $\boldsymbol{s} = \boldsymbol{e}_i$. Define $\mathscr{E}_i$ as the event that b is exceeded due to a jump and that the phase of this jump at exceedance is $i \in \{1,\ldots,d\}$, whereas $\mathscr{E}_0$ denotes the event that b is exceeded due to *creeping* (i.e., due to the driving process' Brownian component); observe that in the former case, there is an overshoot, whereas in the latter case there is not. Let $k_i(\beta)$ be the LST of τ_b^+ on $\mathscr{E}_i$:

$$k_i(\beta) := \mathbb{E}_x[e^{-\beta\tau_b^+} 1_{\mathscr{E}_i}],$$

with $i \in \{0,\dots,d\}$. Now observe that we can write

$$\Lambda_x^+(\alpha,\beta) = e^{-\alpha b}\left(\sum_{i=1}^{d} \chi_i(\alpha)\, k_i(\beta) + k_0(\beta)\right).$$

We are thus left with identifying the $d+2$ functions $K(\beta), k_0(\beta),\dots,k_d(\beta)$. Note that $\boldsymbol{s}(\alpha I - T)^{-1}\boldsymbol{t}$ can be written as the ratio of two polynomials (in α), with the degree of the denominator being d. As a consequence, $\varphi^+(\alpha)$ can be written as the ratio of two polynomials, with the numerator being of degree $d+2$ and the denominator of degree d. In the case where the $d+2$ zeroes in the complex plane of the resulting equation $\varphi^+(\alpha) = \beta$ are distinct, let us say $\alpha_1,\dots,\alpha_{d+2}$, we thus obtain the following equations (which are linear in the $d+2$ unknown functions) from (48):

$$0 = e^{-\alpha_j x} - e^{-\alpha_j b}\left(\sum_{i=1}^{d} \chi_i(\alpha_j)\, k_i(\beta) + k_0(\beta)\right) - \alpha_j\, K(\beta), \tag{49}$$

for $j = 1,\dots,d+2$.

Although we above dealt with the case $(\sigma^2)^+ > 0$, we complete this section by focusing on $(\sigma^2)^+ = 0$. We also impose the requirement that $c^+ < 0$ to make sure that X^+ is not a subordinator, implying that b is exceeded due to a jump. It is readily checked that now $\varphi^+(\alpha)$ can be written as the ratio of two polynomials, with the numerator being of degree $d+1$ and the denominator of degree d. In this case, because of the absence of the Brownian component, we have $k_0(\beta) = 0$. Hence, if $\varphi^+(\alpha) = \beta$ has $d+1$ distinct zeroes $\alpha_1,\dots,\alpha_{d+1}$ in the complex plane, we are to solve the system of linear equations

$$0 = e^{-\alpha_j x} - e^{-\alpha_j b}\sum_{i=1}^{d} \chi_i(\alpha_j)\, k_i(\beta) - \alpha_j\, K(\beta), \tag{50}$$

for $j = 1,\dots,d+1$.

5.2. Special Cases

As announced, in this section, we consider the cases of X^+ corresponding to BM and to CP with exponentially distributed upward jumps (so that W^+ is the workload of an M/M/1 system).

5.2.1. BM. In the terminology of Section 5.1, this model instance corresponds to $d = 0$; the equation $\varphi^+(\alpha) = \beta$ has two roots. Let $\alpha_1(\beta)$ denote the negative root and $\alpha_2(\beta)$ the positive root. Following the procedure outlined in Section 5.1, the linear system (49) needs to be solved; it can be written as

$$e^{-\alpha_1(\beta)b}\,\mathbb{E}_x e^{-\beta\tau_b^+} + \alpha_1(\beta)\, K(\beta) = e^{-\alpha_1(\beta)x},$$
$$e^{-\alpha_2(\beta)b}\,\mathbb{E}_x e^{-\beta\tau_b^+} + \alpha_2(\beta)\, K(\beta) = e^{-\alpha_2(\beta)x}.$$

Solving these two linear equations with respect to the unknowns $\mathbb{E}_x e^{-\beta\tau_b^+}$ and $K(\beta)$, we conclude that, with $\mathscr{G}(\alpha,x) := e^{-\alpha x}/\alpha$,

$$\mathbb{E}_x e^{-\beta\tau_b^+} = \frac{\mathscr{G}(\alpha_1(\beta),x) - \mathscr{G}(\alpha_2(\beta),x)}{\mathscr{G}(\alpha_1(\beta),b) - \mathscr{G}(\alpha_2(\beta),b)}. \tag{51}$$

Locally abbreviating $\sigma^2 = (\sigma^2)^+ > 0$ and $c = c^+$, the roots of $\varphi^+(\alpha) - \beta = \frac{1}{2}\sigma^2\alpha^2 - c\alpha - \beta$ are, for $i = 1,2$,

$$\alpha_i(\beta) = \frac{c + (-1)^i\sqrt{c^2 + 2\sigma^2\beta}}{\sigma^2}. \tag{52}$$

We thus arrive at, with $\mathbb{E}_x e^{-\beta\tau_b^+}$ given by (51),

$$\Lambda_x^+(\alpha,\beta) = \mathbb{E}_x e^{-\alpha W^+(\tau_b^+) - \beta\tau_b^+} = e^{-\alpha b}\,\mathbb{E}_x e^{-\beta\tau_b^+}. \tag{53}$$

We mention that (51) is consistent with Mayerhofer (2019, theorem 1) and, for the case $\mu = 0$ and $\sigma^2 = 1$, with Borodin and Salminen (2015, 2.0.1, p. 361).

5.2.2. CP with Exponentially Distributed Jumps and a Negative Drift. As mentioned, in the case X^+ corresponds to CP with exponentially distributed jumps and (negative) drift $-c$, where $c>0$, the process $W^+(t/c)$ behaves as the workload of an M/M/1 queue. We let the arrival rate equal λ and the mean jump size μ^{-1}. As a consequence, according to the procedure of Section 5.1, we are to solve the linear system (50):

$$e^{-\alpha_1(\beta)b}\,\frac{\mu}{\mu+\alpha_1(\beta)}\,\mathbb{E}_x e^{-\beta\tau_b^+}+\alpha_1(\beta)\,K(\beta)=e^{-\alpha_1(\beta)x}$$
$$e^{-\alpha_2(\beta)b}\,\frac{\mu}{\mu+\alpha_2(\beta)}\,\mathbb{E}_x e^{-\beta\tau_b^+}+\alpha_2(\beta)\,K(\beta)=e^{-\alpha_2(\beta)x}.$$

This system of equations was found by applying the approach presented in Section 5.1 with $d=1$, $c^+=c$, $(\sigma^2)^+=0$, $\lambda^+=\lambda>0$, and $\chi_1(\alpha)=\chi(\alpha):=\mu/(\mu+\alpha)$. We obtain that

$$\mathbb{E}_x e^{-\beta\tau_b^+}=\frac{\mathscr{G}(\alpha_1(\beta),x)-\mathscr{G}(\alpha_2(\beta),x)}{\mathscr{G}(\alpha_1(\beta),b)\,\chi(\alpha_1(\beta))-\mathscr{G}(\alpha_2(\beta),b)\,\chi(\alpha_2(\beta))}. \tag{54}$$

The two roots of

$$\varphi^+(\alpha)-\beta=c\alpha-\lambda\left(1-\frac{\mu}{\mu+\alpha}\right)-\beta=\frac{c\alpha^2-(\lambda+\beta-c\mu)\alpha-\mu\beta}{\mu+\alpha} \tag{55}$$

are, for $i=1,2$,

$$\alpha_i(\beta)=\frac{\lambda+\beta-c\mu+(-1)^i\sqrt{(\lambda+\beta-c\mu)^2+4c\mu\beta}}{2c}.$$

We conclude, with $\mathbb{E}_x e^{-\beta\tau_b^+}$ given by (54),

$$\Lambda_x^+(\alpha,\beta)=\mathbb{E}_x e^{-\alpha W^+(\tau_b^+)-\beta\tau_b^+}=e^{-\alpha b}\,\frac{\mu}{\mu+\alpha}\,\mathbb{E}_x e^{-\beta\tau_b^+}. \tag{56}$$

Remark 5. By Identity (48), we note that inserting $\beta=0$ leads to

$$\varphi^+(\alpha)\mathbb{E}_x\int_0^{\tau_b^+}e^{-\alpha W^+(s)}\mathrm{d}s=\Lambda_x^+(\alpha,0)-e^{-\alpha x}+\alpha\,\mathbb{E}_x L^+(\tau_b^+), \tag{57}$$

where

$$\mathbb{E}_x L^+(\tau_b^+)=\mathbb{E}_x W^+(\tau_b^+)-\mathbb{E}_x X^+(\tau_b^+)=\mathbb{E}_x W^+(\tau_b^+)+m^+\,\mathbb{E}_x\tau_b^+. \tag{58}$$

Therefore, when $m^+<0$, we have that $\psi^+(0)>0$, which upon substituting $\alpha=\psi^+(0)$ in (48) gives an alternative way of deriving our expression for $\kappa^+(\alpha)$ for this case. Indeed, either from taking $\alpha=\psi^+(0)$ and $x=0$ in (57), or from (58), we obtain that $\mathbb{E}_0 L^+(\tau_b^+)=[1-\Lambda_0^+(\psi^+(0),0)]/\psi^+(0)$.

6. Optimizing the Threshold

In this section, we consider, in the context of our storage process with alternating input, a threshold optimization problem. Given certain switching costs, costs for operating in the − mode, and holding costs, we pose the following question: What is the switching level b_{opt} for which a particular weighted sum of the long-run average costs is minimized? We assume that $m^->0$, to make sure that the process returns to the + mode. Throughout our analysis, we intensively apply the results that we obtained for the two special cases dealt with in Section 5.2. We first consider the case that both X^+ and X^- are BMs and then the case that both are CP processes with deterministic drifts. In both cases, we assume that X^+ and X^- are probabilistically identical, except that they have different deterministic drifts (to be interpreted as service rates).

There are switching costs K_s involved with changing the service speed, costs K_f per time unit for working fast, and holding costs K_h per time unit for each unit of work in the system. Denote the length of an arbitrary − period by θ_b, which is distributed like $\sigma^-(W^+(\tau_b^+))$. By averaging over one regenerative cycle with mean $\mathbb{E}_0\tau_b^++\mathbb{E}\theta_b$, we arrive at the following objective function representing the mean costs per time unit:

$$C(b):=\frac{K_s+K_f\,\mathbb{E}\theta_b+K_h\left[\mathbb{E}_0\int_0^{\tau_b^+}W^+(s)\,\mathrm{d}s+\mathbb{E}_{W^+(\tau_b^+)}\int_0^{\theta_b}W^-(s)\,\mathrm{d}s\right]}{\mathbb{E}_0\tau_b^++\mathbb{E}\theta_b}, \tag{59}$$

which we would like to minimize over b. We remark in passing that a similar optimization problem has been discussed for the M/G/1 queue in Cohen (1976), for the case that the server instantaneously switches back to the mode 1 speed when level b is first down-crossed again.

In this section, we first derive in Section 6.1 a couple of identities that are helpful in evaluating $C(b)$. Then, in the next two sections, we provide closed-form expressions for all quantities appearing on the right-hand side of (59) for the two special cases mentioned above. The section concludes with an analysis of the optimizing threshold b_{opt}.

Remark 6. We remark that, in Feinberg and Kella (2002), an old conjecture that D-policies are optimal for the average cost per unit time criterion in an M/G/1 queue with a removable server is proved. The cost structure there consisted of switching costs, running costs, and holding costs per unit time that is a nonnegative nondecreasing right-continuous function of a current workload in the system. This criterion means that there is an optimal policy that either runs the server all the time or switches the server off when the system becomes empty and switches it on when the workload reaches or exceeds some threshold D.

A similar result might be expected to hold under the more general setup considered in the current study. Namely, that the optimal policy considered in this section is optimal among a much larger class of policies. However, this is beyond the scope of the current paper and is left for future research.

Remark 7. As a reviewer suggested, instead of moving from the − to the + regime only when the inventory is depleted, one might consider performing this change when the inventory hits some level $a \in [0, b]$ from above. This setup would of course result in a richer optimization problem where one would then want to optimize with respect to both a and b. The mathematics involved in computing various descriptive results can be expected to be of a similar nature as the ones carried out in the previous sections. However, this paper was motivated by, and was supposed to be the queueing counterpart of, the ratcheting model in a risk setup as presented in Albrecher et al. (2018). Moreover, we do not expect that the attractive simple decomposition result described in Theorem 3 would continue to hold.

6.1. Some Identities

In this section, we develop a set of general identities that we later use to evaluate the cost $C(b)$. Let X be a spectrally positive Lévy process with Laplace exponent $\varphi(\cdot)$. The processes W and L denote the associated workload and the local time at zero, respectively. We recall that, using the notation of (2),

$$-\varphi'(0) = c + \int_{(1,\infty)} x\nu(\mathrm{d}x), \quad \varphi''(0) = \sigma^2 + \int_{(0,\infty)} x^2\nu(\mathrm{d}x), \ -\varphi'''(0) = \int_{(0,\infty)} x^3\nu(\mathrm{d}x). \tag{60}$$

We also note that (48) and (57) remain valid with the plus signs removed, with any τ such that $\mathbb{E}_x\tau < \infty$ replacing τ_b^+ and with $\mathbb{E}_x e^{-\alpha W(\tau)-\beta\tau}$ replacing $\Lambda_x^+(\alpha,\beta)$. Namely,

$$(\varphi(\alpha) - \beta)\mathbb{E}_x \int_0^\tau e^{-\alpha W(s)-\beta s}\,\mathrm{d}s = \mathbb{E}_x e^{-\alpha W(\tau)-\beta\tau} - e^{-\alpha x} + \alpha\, \mathbb{E}_x \int_0^\tau e^{-\beta s}\,\mathrm{d}L(s),$$

and (for $\beta = 0$),

$$\varphi(\alpha)\mathbb{E}_x \int_0^\tau e^{-\alpha W(s)}\,\mathrm{d}s = \mathbb{E}_x e^{-\alpha W(\tau)} - e^{-\alpha x} + \alpha\, \mathbb{E}_x L(\tau). \tag{61}$$

When $\mathbb{E}_x W(\tau) < \infty$ and $\mathbb{E}_x\tau < \infty$, we have, as a direct application of $\mathbb{E}_x X(\tau) = x - \varphi'(0)\mathbb{E}_x\tau$ and $L(\tau) = W(\tau) - X(\tau)$, that $\mathbb{E}_x L(\tau) = \mathbb{E}_x W(\tau) - x + \varphi'(0)\mathbb{E}_x\tau$.

We start by considering the case when $\varphi'(0) < 0$, implying that there is a positive root of $\varphi(\alpha) = 0$. When $\varphi'(0) > 0$ and for some $-\infty \le \alpha^* < 0$, we have that $\varphi(\alpha)$ is finite on $(\alpha^*, 0)$ and converges to infinity as $\alpha \downarrow \alpha^*$, then there is a negative root of $\varphi(\alpha) = 0$. With a slight abuse of notations, in both cases, we denote this root by $\psi(0)$, remarking that here we deviate from the definition of $\psi(\cdot)$ as the right-inverse of $\varphi(\cdot)$. Inserting this root in (61) gives

$$0 = \mathbb{E}_x e^{-\psi(0)W(\tau)} - e^{-\psi(0)x} + \psi(0)(\mathbb{E}_x W(\tau) - x + \varphi'(0)\,\mathbb{E}_x\tau). \tag{62}$$

Therefore, in this case, whenever there are simple expressions for both $\mathbb{E}_x e^{-\alpha W(\tau)}$ and $\mathbb{E}_x W(\tau)$, we can compute $\mathbb{E}_x\tau$ from

$$\mathbb{E}_x\tau = \frac{\mathbb{E}_x(e^{-\psi(0)W(\tau)} - 1 + \psi(0)\,W(\tau)) - (e^{-\psi(0)x} - 1 + \psi(0)\,x)}{-\varphi'(0)\psi(0)}; \tag{63}$$

observe that (i) $-\varphi'(0)\psi(0) > 0$ and (ii) whenever $W(\tau) > x$ a.s., the numerator is positive, because $e^z - 1 - z$ and $e^{-z} - 1 + z$ are increasing on $[0, \infty)$.

We continue by considering the case when $\varphi'(0) = 0$. Supposing that $\mathbb{E}_x W^2(\tau) < \infty$ can be computed, then we differentiate both sides of (61) twice, let $\alpha \downarrow 0$, and subsequentially divide by $\varphi''(0)$. This yields

$$\mathbb{E}_x\tau = \frac{\mathbb{E}_x W^2(\tau) - x^2}{\varphi''(0)}. \tag{64}$$

Recalling the expression for $\varphi''(0)$ from (60), we conclude that $\varphi''(0)$ is positive (excluding the case where X is a negative drift) and finite. Therefore, whenever there is a simple expression for $\mathbb{E}_x W^2(\tau)$ and $W(\tau) > x$ a.s., we have an expression for $\mathbb{E}_x\tau$.

It now becomes easy to compute $\mathbb{E}_x \int_0^\tau W(s)\,\mathrm{d}s$ via (61). Whenever $\varphi'(0) \neq 0$, we differentiate both sides twice and set $\alpha = 0$ to obtain

$$\varphi''(0)\mathbb{E}_x\tau - 2\varphi'(0)\,\mathbb{E}_x \int_0^\tau W(s)\,\mathrm{d}s = \mathbb{E}_x W^2(\tau) - x^2. \tag{65}$$

When $\varphi'(0) = 0$, we differentiate three times and set $\alpha = 0$ to arrive at

$$\varphi'''(0)\mathbb{E}_x\tau - 3\varphi''(0)\,\mathbb{E}_x \int_0^\tau W(s)\,\mathrm{d}s = -\mathbb{E}_x W^3(\tau) + x^3. \tag{66}$$

For the BM case, ν is zero, so that the three numbers $\varphi'(0)$, $\varphi''(0)$, and $\varphi'''(0)$ become c, σ^2, and 0, respectively. For the CP case with jump rate λ, a possible drift c and jumps distributed like a positive random variable Y, these become $c + \lambda\mathbb{E}Y$, $\lambda\mathbb{E}Y^2$, and $\lambda\mathbb{E}Y^3$, respectively. In particular, if the jump size Y is exponentially distributed with mean μ^{-1}, then

$$\mathbb{E}Y = \frac{1}{\mu}, \quad \mathbb{E}Y^2 = \frac{2}{\mu^2}, \quad \mathbb{E}Y^3 = \frac{6}{\mu^3}. \tag{67}$$

In particular, it will be useful to note for what follows that, for any $b > 0$,

$$\mathbb{E}(b+Y) = b + \frac{1}{\mu}, \quad \mathbb{E}(b+Y)^2 = b^2 + \frac{2b}{\mu} + \frac{2}{\mu^2}, \quad \mathbb{E}(b+Y)^3 = b^3 + \frac{3b^2}{\mu} + \frac{6b}{\mu^2} + \frac{6}{\mu^3}. \tag{68}$$

6.2. BM

In this section, we consider the case of BM with drift c^+ in mode + and $c^- < 0$ in mode −. We thus have

$$\varphi^+(\alpha) = -c^+\alpha + \frac{\sigma^2}{2}\alpha^2, \quad \varphi^-(\alpha) = -c^-\alpha + \frac{\sigma^2}{2}\alpha^2.$$

We continue by determining the various components of the objective function given in (59).

i. $\mathbb{E}_0\tau_b^+$. This quantity can be obtained from (63) for $c^+ \neq 0$ and from (64) for $c^+ = 0$, where we work with the Lévy process X^+ and plug in the specific stopping time $\tau := \tau_b^+$. For $c^+ \neq 0$, we use that $\psi^+(0) = 2c^+/\sigma^2$ (with $\psi^+(0)$ in the meaning introduced in Section 6.1), $m^+ = -c^+$, and $W^+(\tau_b^+) = b$, to obtain

$$\mathbb{E}_0\tau_b^+ = \left(\exp\left(-\frac{2c^+b}{\sigma^2}\right) - 1 + \frac{2c^+b}{\sigma^2}\right) \Big/ \frac{2(c^+)^2}{\sigma^2}. \tag{69}$$

For $c^+ = 0$, by using (64) with $(\varphi^+)''(0) = \sigma^2$ (or by letting $c^+ \downarrow 0$ in (69)), we readily obtain

$$\mathbb{E}_0\tau_b^+ = \frac{b^2}{\sigma^2}. \tag{70}$$

An alternative approach would be to obtain $\mathbb{E}_0\tau_b^+$ by differentiating $\mathbb{E}_0 e^{-\beta\tau_b^+}$ with respect to β, multiplying by −1, and inserting $\beta = 0$, but this procedure gives rise to rather involved calculations.

ii. $\mathbb{E}\theta_b$. Trivially,

$$\mathbb{E}\theta_b = \frac{b}{(\varphi^-)'(0)} = \frac{b}{-c^-}. \tag{71}$$

iii. $\mathbb{E}_0 \int_0^{\tau_b^+} W^+(s)\,\mathrm{d}s$. Using (65) and (66), we obtain for $c^+ \neq 0$ that

$$\mathbb{E}_0 \int_0^{\tau_b^+} W^+(s)\,\mathrm{d}s = \frac{b^2}{2c^+} - \frac{\sigma^2}{2c^+}\mathbb{E}_0\tau_b^+, \tag{72}$$

whereas for $c^+ = 0$, we have

$$\mathbb{E}_0 \int_0^{\tau_b^+} W^+(s)\,\mathrm{d}s = \frac{b^3}{3\sigma^2}. \tag{73}$$

iv. $\mathbb{E}_{W^+(\tau_b^+)} \int_0^{\theta_b} W^-(s)\,\mathrm{d}s$. We have (see (37)),

$$\frac{\mathbb{E}_{W^+(\tau_b^+)} \int_0^{\theta_b} e^{-\alpha W^-(s)}\,\mathrm{d}s}{\mathbb{E}\theta_b} = \frac{-c^-}{\frac{\sigma^2}{2}\alpha - c^-} \frac{1 - \mathrm{e}^{-\alpha b}}{\alpha b}, \tag{74}$$

which yields

$$\mathbb{E}_0 \int_0^{\theta_b} W^-(s)\,\mathrm{d}s = \frac{b^2}{-2c^-} + \frac{b\sigma^2}{2(c^-)^2}. \tag{75}$$

6.3. CP with Exponential Jumps

In this section, we consider an M/M/1 queue with arrival rate λ, mean service requirement $1/\mu$, and service speed $r^+ = -c^+$ in mode + and $r^- = -c^-$ in mode −. Define $\varrho^\pm := \lambda/(\mu r^\pm)$. We assume $\varrho^- < 1$, so that it is guaranteed that zero is reached in mode −. We have

$$\varphi^+(\alpha) = r^+\alpha - \lambda\left(1 - \frac{\mu}{\mu + \alpha}\right), \quad \varphi^-(\alpha) = r^-\alpha - \lambda\left(1 - \frac{\mu}{\mu + \alpha}\right).$$

We proceed by determining the various components of the objective function for this M/M/1 system with switching depletion rate.

i. $\mathbb{E}_0 \tau_b^+$. First of all, observe that, because of the memoryless property of the exponential distribution, $W^+(\tau_b^+) \sim b + Y$ with Y being exponentially distributed with mean μ^{-1}. For $\varrho^+ \neq 1$, we obtain the mean length of a mode + period from (63), and using that the above expression for $\varphi^+(\alpha)$ implies that $\psi^+(0) = \mu(\varrho^+ - 1)$ and $m^+ = r^+(1 - \varrho^+)$:

$$\begin{aligned}
\mathbb{E}_0 \tau_b^+ &= -\frac{1}{m^+\psi^+(0)}\left(\mathbb{E}_0[e^{-\psi^+(0)W^+(\tau_b^+)} - 1 + \psi^+(0)W^+(\tau_b^+)]\right) \\
&= -\frac{1}{m^+\psi^+(0)}\left(e^{-\mu(\varrho^+-1)b}\frac{\mu}{\mu + \psi^+(0)} - 1 + \psi^+(0)\left(b + \frac{1}{\mu}\right)\right) \\
&= \frac{1}{r^+(\varrho^+ - 1)}\left(b + \frac{1}{\mu}\right) + \frac{e^{-\mu(\varrho^+-1)b}}{\lambda(\varrho^+ - 1)^2} - \frac{1}{r^+\mu(\varrho^+ - 1)^2}.
\end{aligned} \tag{76}$$

For $\varrho^+ = 1$, we have $(\varphi^+)'(0) = m^+ = 0$ but $(\varphi^+)''(0) = 2\lambda/\mu^2 \neq 0$, and (64) in combination with (68) yields

$$\mathbb{E}_0 \tau_b^+ = \frac{\mathbb{E}_0(W^+(\tau_b^+))^2}{(\varphi^+)''(0)} = \left(b^2 + 2\frac{b}{\mu} + \frac{2}{\mu^2}\right)\Big/\frac{2\lambda}{\mu^2}. \tag{77}$$

ii. $\mathbb{E}\theta_b$. The mean length of the mode − period is just the mean length of an M/M/1 busy period that starts with an (exceptional) service time with mean $b + \frac{1}{\mu}$, and hence

$$\mathbb{E}\theta_b = \frac{1}{r^-(1 - \varrho^-)}\left(b + \frac{1}{\mu}\right). \tag{78}$$

iii. $\mathbb{E}_0 \int_0^{\tau_b^+} W^+(s)\,\mathrm{d}s$. Using (65) and (66), we obtain for $\varrho^+ \neq 1$ that

$$\begin{aligned}
\mathbb{E}_0 \int_0^{\tau_b^+} W^+(s)\,\mathrm{d}s &= \frac{1}{2m^+}((\varphi^+)''(0)\,\mathbb{E}_0\tau_b^+ - \mathbb{E}_0 W^2(\tau_b^+)) \\
&= \frac{1}{r^+(\varrho^+ - 1)}\left(\frac{b^2}{2} + \frac{b}{\mu} + \frac{1}{\mu^2}\right) + \frac{\varrho^+}{\mu(1 - \varrho^+)}\mathbb{E}_0\tau_b^+,
\end{aligned} \tag{79}$$

whereas for $\varrho^+ = 1$,

$$\mathbb{E}_0 \int_0^{\tau_b^+} W^+(s)\,\mathrm{d}s = \frac{\mu^2}{6\lambda}\left(b^3 + 3\frac{b^2}{\mu} + 6\frac{b}{\mu^2} + \frac{6}{\mu^3}\right) - \frac{\mu}{2\lambda}\left(b^2 + 2\frac{b}{\mu} + \frac{2}{\mu^2}\right). \tag{80}$$

Notice that, for $\varrho^+ < 1$, the factor in front of $\mathbb{E}_0\tau_b^+$ in the last line of (79) can be interpreted as the mean steady-state workload of a conventional M/M/1 queue (without any switching, that is) operating with service speed r^+.

iv. $\mathbb{E}_{W^+(\tau_b^+)}\int_0^{\theta_b} W^-(s)\,\mathrm{d}s$. We have (see (37)),

$$\frac{\mathbb{E}_{W^+(\tau_b^+)}\int_0^{\theta_b} e^{-\alpha W^-(s)}\,\mathrm{d}s}{\mathbb{E}\theta_b} = \frac{1-\varrho^-}{1-\varrho^-\mathscr{E}(\alpha)}\,\frac{1-e^{-\alpha b}E(\alpha)}{\alpha\left(b+\frac{1}{\mu}\right)}, \tag{81}$$

with $\mathscr{E}(\alpha) := \mu/(\mu+\alpha)$ denoting the LST of an exponentially distributed random variable with mean μ^{-1}. Noting that this expression is the product of the LST of the M/M/1 waiting time or workload distribution and that of the residual of $W^+(\tau_b^+)$, using (78), we obtain

$$\mathbb{E}_{W^+(\tau_b^+)}\int_0^{\theta_b} W^-(s)\,\mathrm{d}s = \frac{1}{r^-(1-\varrho^-)}\left(\frac{b^2}{2}+\frac{b}{\mu}+\frac{1}{\mu^2}\right) + \frac{\varrho^-}{\mu r^-(1-\varrho^-)^2}\left(b+\frac{1}{\mu}\right). \tag{82}$$

6.4. Optimal Switching Threshold

Now that we have all components of the objective function $C(b)$, we turn to the question of determining b_{opt} that minimizes $C(b)$.

The easiest optimization problem arises for BM, in the case that $c^+ = 0$. The objective function now becomes

$$C(b) = \frac{K_s + K_f\frac{b}{-c^-} + K_h\frac{b^3}{3\sigma^2}}{\frac{b^2}{\sigma^2}+\frac{b}{-c^-}} + K_h\,\frac{\sigma^2}{-2c^-}, \tag{83}$$

and all terms are positive (recalling that we imposed the assumption that $c^- < 0$). As expected, for small b, the switching costs K_s dominate, and $C(b)$ tends to infinity essentially proportionally to $1/b$. For large b, the holding costs K_h dominate, and $C(b)$ tends to infinity essentially linearly in b. It is noted that $C(b)$ is, in general, not convex, and the derivative of $C(b)$ has four roots, of which at least one is positive. The value of b that minimizes $C(b)$ can be determined in a routine manner (Figure 2(a)).

If $K_s = 0$, then one can divide numerator and denominator of $C(b)$ by b, and it is readily seen that a unique minimum occurs for

$$b_{\text{opt}} = -\frac{\sigma^2}{-c^-} + \sqrt{\left(\frac{\sigma^2}{-c^-}\right)^2 + 3\,\frac{K_f}{K_h}\cdot\frac{\sigma^2}{-c^-}}. \tag{84}$$

In this expression, the effects of large or small cost coefficients K_f and K_h are evident. First, a large K_h leads to a small b_{opt}, as expected. Second, a large K_f results in a large b_{opt}. Apparently, even though the costs of operating in the $-$ mode are high in this case, the fact that the mean duration of a mode $+$ period is an order b larger than that of a mode $-$ period makes it profitable to choose a large value of b (Figure 3(a)).

In the CP case with exponentially distributed jumps, where our process corresponds to an M/M/1-type storage system with switching service rate, the easiest optimization problem arises when $\varrho^+ = 1$. With Y denoting an exponentially distributed random variable with mean μ^{-1}, we introduce $m_k(b) := \mathbb{E}(b+Y)^k$ to make our

Figure 2. Cost Function (83) (a) and (86) (b) as a Function of the Switching Threshold b for $\sigma = 0.2$, $c^- = -1$, $K_s = 0.2$, $K_f = 0.05$, $K_h = 0.2$

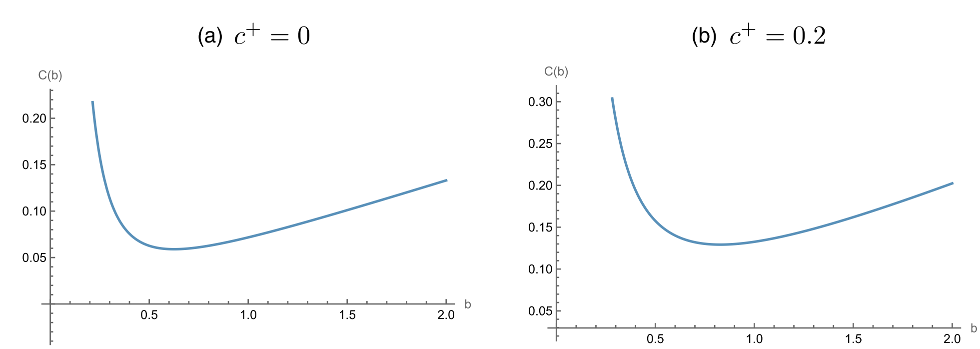
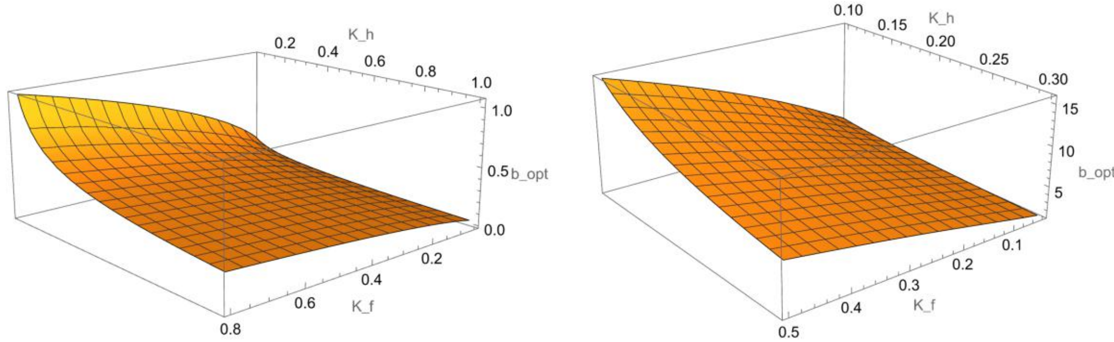


Figure 3. Optimal Switching Threshold b_{opt} as a Function of Costs K_h and K_f for the Brownian Motion Model (Left) and the Compound Poisson Model with Exponential Jumps (Right)

(a) $\sigma = 0.2,\ c^+ = 0,\ c^- = -1,\ K_s = 0$ (b) $\mu = 1,\ \lambda = 1,\ \varrho^+ = 1,\ \varrho^- = 0.95,\ K_s = 0.2$



expressions more compact. The objective function now becomes

$$C(b) = \frac{K_s + K_f \frac{m_1(b)}{r^-(1-\varrho^-)} + K_h\left[\frac{\mu^2}{6\lambda}m_3(b) - \frac{\mu}{2\lambda}m_2(b)\right]}{\frac{2\mu^2}{\lambda}m_2(b) + \frac{m_1(b)}{r^-(1-\varrho^-)}} + K_h\frac{\varrho^-}{\mu r^-(1-\varrho^-)}, \tag{85}$$

where, just as we did in (83), a term that is proportional to K_h and that does not involve b has been taken out of the fraction. For large b, the holding costs K_h (again) dominate and $C(b)$ tends to infinity, essentially linearly in b. Like in the case of the driving process being BM, the derivative of $C(b)$ has four roots, of which at least one is positive because the derivative of $C(b)$ at $b = 0$ is negative. The cost-minimizing threshold b_{opt} can be determined in the standard way (see Figure 4(b) for an illustration of the cost function $C(b)$ for a particular choice of costs K_s, K_f, K_h). Figure 3(b) depicts the optimal level of b minimizing the cost function (85) as a function of costs K_h and K_f, for $K_s = 0.2$ (for exponential service times, (85) is a rational function of degree 3 whose minimum can be obtained numerically for each combination of K_h and K_f).

In the case of BM with $c^+ \neq 0$, the objective function becomes

$$C(b) = \frac{K_s + K_f\frac{b}{-c^-} + K_h\left[\frac{b^2}{2}\left(\frac{1}{c^+} - \frac{1}{c^-}\right) - \frac{\sigma^2}{2c^+}\mathbb{E}_0\tau_b^+ + \frac{b\sigma^2}{2(c^-)^2}\right]}{\mathbb{E}_0\tau_b^+ + \frac{b}{-c^-}}, \tag{86}$$

with $\mathbb{E}_0\tau_b^+$ being given in (69). For small b, the switching costs again dominate, and $C(b)$ tends to infinity inversely proportionally to b. For large b, the behavior of $C(b)$ strongly depends on the sign of c^+: (i) if $c^+ > 0$, then $C(b)$ grows linearly in b for large b, but (ii) if $c^+ < 0$ (corresponding to downward drift in the + mode), then $\mathbb{E}_0\tau_b^+$ grows exponentially in b, and hence $C(b)$ effectively behaves as $K_h\sigma^2/(-2c^+)$ for large b. In either case, b_{opt} can be determined using standard numerical software (Figure 2(b)).

We conclude by considering the case of CP with exponentially distributed jumps and $\varrho^+ \neq 1$. The objective function becomes

$$C(b) = \frac{K_s + K_f\frac{m_1(b)}{r^-(1-\varrho^-)} + K_hC_h(b)}{\mathbb{E}_0\tau_b^+ + \frac{m_1(b)}{r^-(1-\varrho^-)}}, \tag{87}$$

Figure 4. Cost Function (85) (b) and (87) ((a) and (c)) as a Function of the Switching Threshold b for $\mu = 1$, $\lambda = 1$, $\rho^- = 0.95$, $K_s = 0.2, K_f = 0.05, K_h = 0.2$

(a) $\rho^+ = 0.97$ (b) $\rho^+ = 1$ (c) $\rho^+ = 1.1$

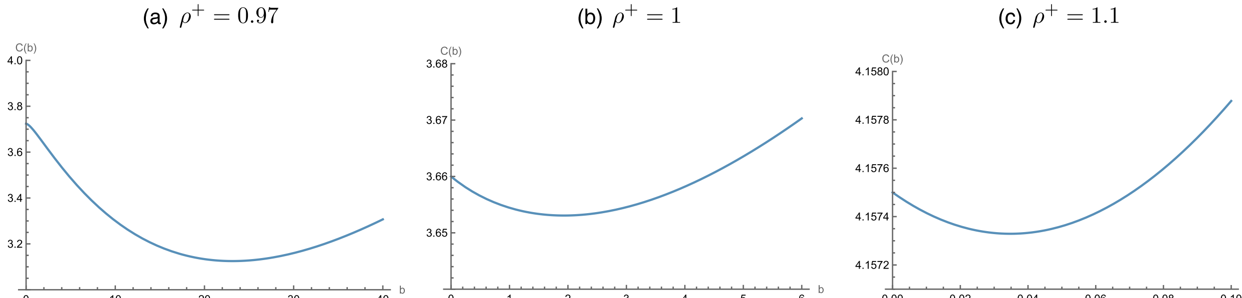
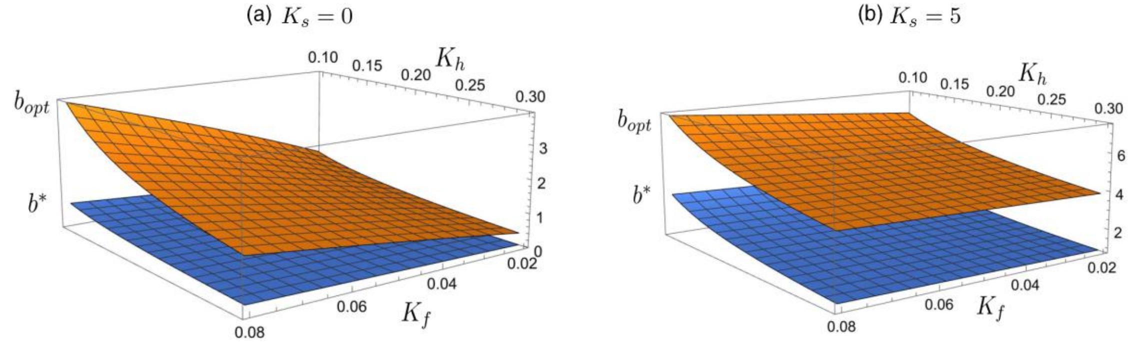


Figure 5. Comparison of b_{opt} (Orange, Above) and b^* from (89) (Blue, Below) as a Function of Costs K_h and K_f, for the Compound Poisson Model with Exp(1) Jumps: $\lambda = 1$, $\varrho^+ = 1$, $\varrho^- = 0.95$



with

$$C_h(b) := \frac{\varrho^+ \, \mathbb{E}_0 \tau_b^+}{\mu(1-\varrho^+)} - \frac{m_2(b)}{r^+(1-\varrho^+)} + \frac{m_2(b)}{r^-(1-\varrho^-)} + \frac{\varrho^- \, m_1(b)}{\mu r^-(1-\varrho^-)^2}, \tag{88}$$

and with $\mathbb{E}_0 \tau_b^+$ being given in (76).

For large b, the behavior of $C(b)$ strongly depends on the value of ϱ^+. If $\varrho^+ > 1$, then $C(b)$ grows essentially linearly in b for large b. If $\varrho^+ < 1$ (downward drift in the + mode), then $\mathbb{E}_0 \tau_b^+$ grows exponentially for large b, so that $C(b)$ behaves as $K_h \varrho^+/(\mu(1-\varrho^+))$ for large b. The minimum of $C(b)$ can be determined numerically relying on standard computational packages (Figure 4, (a) and (c)). Note how dramatically the location of the minimum b_{opt} changes in Figure 4 as ρ^+ transits through the value 1.

Remark 8. It is interesting to compare the optimal threshold with the one from Cohen (1976), where the switching to the − mode is reversible at the next down-crossing of b, so that several switches can take place during a busy period. Most of the results in Cohen (1976) are derived for the case of no switching costs ($K_s = 0$); only for the case $\rho^+ = 1$, there is also a formula for $K_s \geq 0$. We therefore focus on the comparison for $\rho^+ = 1$. Cohen (1976, equation (4.21)) gives, for Exp(1)-distributed service times, the explicit formula

$$b^* = -\frac{1}{1-\rho^-} + \sqrt{\frac{1}{(1-\rho^-)^2} + \frac{2}{K_h}\left(K_f \frac{\rho^-}{1-\rho^-} + K_s\right)} \tag{89}$$

for the optimal switching threshold b^* in the reversible case (which for $K_s = 0$ coincides also with the solution of Cohen (1976, equation (4.11)), derived by different techniques). Note that for the sake of comparison, we assume that in Cohen's model switching costs only apply when switching from the + to the − mode. Figure 5(a) compares this optimal b^*, as a function of costs K_f and K_h, with the optimal threshold b_{opt} that minimizes (85) for this case, without and with considerable switching costs. As expected, the optimal threshold b_{opt} is always larger than b^*, as without the possibility of switching back at that level one needs to be more prudent to invoke the higher costs of faster speed (with cost K_f per time unit). The larger the holding costs K_h, the smaller that threshold gets for both models.

7. Concluding Remarks

In this paper, we studied a queueing process with Lévy input that is modified whenever the workload process hits a prespecified threshold for the first time during a busy period, and after the queue is empty, it restarts with the original dynamics. We obtained expressions for the workload at an independent exponential time and studied the stationary workload in more detail. For a particular specification of costs for switching, holding, and increased service speed, we also derived formulas for optimal switching thresholds in the case of an underlying Brownian motion or compound Poisson process with positive jumps as Lévy input.

There are various directions in which further research could be of interest. In the paper, we assumed the holding costs to be independent of the current workload. It may be practically relevant to remove that assumption. From a mathematical perspective, this will lead to a more complicated form of (59), where the function $K_h(s)$ will appear within the integrals, and it will then depend on the concrete form of the latter function whether the analysis still remains amenable. Note that Cohen (1976) allowed such costs to be a different constant below and above the chosen threshold, but the goal here would be to choose the optimal threshold given a concrete exogenous form of $K_h(s)$.

Another direction of interest may be to jointly study the choice of the threshold b and the increased service rate r^- after upcrossing b, so that the degree of additional service becomes part of the optimization procedure. Furthermore, it could be practically relevant to only inspect the level of the workload at independent exponentially distributed epochs and switch the process only at those times in case one is in the + mode above b. One could then study the efficiency loss of not continuously observing the process.

In risk theory, the problem of increasing the dividend rate once over time was later generalized to identify the optimal schedule of when to raise the dividends continuously and by how much in order to maximize the expected dividend payments (Albrecher et al. 2020, 2022). Such a stochastic control problem could also be of interest in the present queueing application by determining the optimal schedule of raising the service speed gradually as a function of current workload relative to a given cost function. However, the mathematical techniques needed for such an approach will be quite different from the ones used in this paper. A first step in that direction could be to allow for multiple thresholds in the analysis and investigate the sensitivity of the cost function.

Appendix A. Auxiliary Results

In the following two lemmas, X is a Lévy process that is not the negative of a subordinator, W is the associated reflected process, and, for $b > 0$, $\tau_b := \inf\{t \mid W(t) > b\}$. Note that the deterministic function ct for $c \le 0$ also counts as *the negative of a subordinator* (which has been ruled out).

Lemma A.1. *There exists some $0 < \alpha^* \le \infty$ such that for all $\alpha \in (0, \alpha^*)$, $\mathbb{E}e^{\alpha\tau_b} < \infty$. In particular, τ_b has finite moments of all orders.*

Proof. If X is not the negative of a subordinator, then, for any $s > 0$, $\mathbb{P}(X(s) \le 0) < 1$ (Sato 2005, theorem 24.11). Therefore, for some $x > 0$, we have that $\mathbb{P}(X(s) > x) > 0$. Take $m \ge 1$ such that $mx \ge b$. Then

$$\begin{aligned}\mathbb{P}(X(ms) > b) &\ge \mathbb{P}(X(ms) > mx) \\ &\ge \mathbb{P}\left(\bigcap_{k=1}^{m}\{X(ks) - X((k-1)s) > x\}\right) \\ &= \mathbb{P}(X(s) > x)^m > 0. \end{aligned} \tag{A.1}$$

Take $t = ms$ (so that $\mathbb{P}(X(t) > b) > 0$) and denote

$$N := \inf\{n \mid X(nt) - X((n-1)t) > b\}. \tag{A.2}$$

Clearly,

$$W(Nt) \ge W((N-1)t) + X(Nt) - X((N-1)t) \ge X(Nt) - X((N-1)t) > b, \tag{A.3}$$

which implies that $\tau_b \le Nt$. Because N is geometrically distributed with success probability $p := \mathbb{P}(X(t) > b) > 0$, we have for any $0 < \alpha < \alpha^* \equiv -\frac{1}{t}\log(1-p)$ (where $\alpha^* = \infty$ in case $p = 1$) that

$$\mathbb{E}e^{\alpha\tau_b} \le \mathbb{E}e^{\alpha t N} = \frac{\alpha t e^{\alpha t}}{1 - (1-p)e^{\alpha t}} < \infty, \tag{A.4}$$

and we are done. □

Lemma A.2. *When, in addition, X is spectrally positive, $\mathbb{E}W(\tau_b) < \infty$ iff $\int_{(1,\infty)} x\nu(\mathrm{d}x) < \infty$.*

Proof. It is clear that $\mathbb{E}W(\tau_b) < \infty$ iff

$$\mathbb{E}(W(\tau_b) - b)1_{\{W(\tau_b) - b > 1\}} < \infty.$$

Now multiply the displayed equation in Kyprianou (2013, theorem VII.7) by u and integrate with respect to u on the interval $(1, \infty)$, as well as the rest of the variables. The integration with respect to u (for fixed v) results in

$$\int_{(1,\infty)} u\,\nu(v + \mathrm{d}u) = \int_{(1+v,\infty)} (u - v)\,\nu(\mathrm{d}u) = \int_{(1+v,\infty)} u\,\nu(\mathrm{d}u) - v\,\nu(1+v, \infty), \tag{A.5}$$

where the right-hand side is bounded above by $\int_{(1,\infty)} u\,\nu(\mathrm{d}u)$ and below by

$$\int_{(1+b,\infty)} u\,\nu(\mathrm{d}u) - b\nu(1,\infty) = \int_{(1,\infty)} u\,\nu(\mathrm{d}u) - \int_{[1,1+b)} u\,\nu(\mathrm{d}u) - b\nu(1,\infty). \tag{A.6}$$

Both upper and lower bounds are finite iff $\int_{(1,\infty)} u\,\nu(\mathrm{d}u)$ and hence the equivalence follows. □

References

Abate J, Whitt W (1995) Numerical inversion of Laplace transforms of probability distributions. *ORSA J. Comput.* 7:36–43.

Albrecher H, Azcue P, Muler N (2020) Optimal ratcheting of dividends in insurance. *SIAM J. Control Optim.* 58:1822–1845.

Albrecher H, Azcue P, Muler N (2022) Optimal ratcheting of dividends in a Brownian risk model. *SIAM J. Financial Math.* 13:657–701.

Albrecher H, Bäuerle N, Bladt M (2018) Dividends: From refracting to ratcheting. *Insurance Math. Econom.* 83:45–58.

Albrecher H, Borst S, Boxma O, Resing J (2009) The tax identity in risk theory: A simple proof and an extension. *Insurance Math. Econom.* 44:304–306.

Asghari N, den Iseger P, Mandjes M (2014) Numerical techniques in Lévy fluctuation theory. *Methodol. Comput. Appl. Probab.* 16:31–52.

Asmussen S (2003) *Applied Probability and Queues* (Springer, New York).

Asmussen S (2014) Lévy processes, phase-type distributions, and martingales. *Stochastic Model.* 30:443–468.

Asmussen S, Albrecher H (2010) *Ruin Probabilities*, 2nd ed. (World Scientific, Singapore).

Asmussen S, Rosiński J (2001) Approximations for small jumps of Lévy processes with a view towards simulation. *J. Appl. Probab.* 38:482–493.

Asmussen S, Taksar M (1997) Controlled diffusion models for optimal dividend pay-out. *Insurance Math. Econom.* 20:1–15.

Avram F, Kyprianou AE, Pistorius MR (2004) Exit problems for spectrally negative Lévy processes and applications to (Canadized) Russian options. *Ann. Appl. Probab.* 14:215–235.

Bae J, Kim S, Lee EY (2003) Average cost under the $P^M_{\lambda,\tau}$ policy in a finite dam with compound Poisson inputs. *J. Appl. Probab.* 40:519–526.

Barlev S, Perry D (1993) Two-stage release rule procedure in a regenerative dam. *Probab. Engrg. Inform. Sci.* 7:571–588.

Bekker R (2009) Queues with Lévy input and hysteretic control. *Queueing Systems* 63:281–299.

Bekker R, Boxma OJ (2007) An $M/G/1$ queue with adaptable service speed. *Stochastic Model.* 23:373–396.

Bekker R, Boxma OJ, Kella O (2008a) Queues with delays in two-state strategies and Lévy input. *J. Appl. Probab.* 45:314–332.

Bekker R, Boxma OJ, Resing JAC (2008b) Queues with service speed adaptations. *Statist. Neerlandica* 62:441–457.

Bekker R, Boxma OJ, Resing JAC (2009) Lévy processes with adaptable exponent. *Adv. Appl. Probab.* 41:177–205.

Bertoin J (1998) *Lévy Processes* (Cambridge University Press, Cambridge, UK).

Borodin A, Salminen P (2015) *Handbook of Brownian Motion—Facts and Formulae*, 2nd ed. (Birkhäuser, Basel).

Cohen JW (1976) On the optimal switching level for an M/G/1 queueing system. *Stochastic Processes Appl.* 4:297–316.

Dębicki K, Mandjes M (2015) *Queues and Lévy Fluctuation Theory* (Springer, Berlin).

Den Iseger P (2006) Numerical transform inversion using Gaussian quadrature. *Probab. Engrg. Inform. Sci.* 20:1–44.

Den Iseger P, Gruntjes P, Mandjes M (2013) A Wiener–Hopf based approach to numerical computations in fluctuation theory for Lévy processes. *Math. Methods Oper. Res.* 78:101–118.

Dshalalow JH (1997) Queueing systems with state dependent parameters. Dshalalow JH, ed. *Frontiers in Queueing* (CRC Press, Boca Raton, FL), 61–116.

Feinberg E, Kella O (2002) Optimality of D-policies for an M/G/1 queue with a removable server. *Queueing Systems* 42:355–376.

Gao H, Yin C (2023) A Lévy risk model with ratcheting and barrier dividend strategies. *Math. Foundations Comput.* 6:268–279.

Gaver DP, Miller RG (1962) Limiting distributions for some storage problems. Arrow KJ, Karlin S, Scarf H, eds. *Studies in Applied Probability and Management Science* (Stanford University Press, Stanford, CA), 110–126.

Gerber HU, Shiu ES (2006a) On optimal dividends: From reflection to refraction. *J. Comput. Appl. Math.* 186:4–22.

Gerber HU, Shiu ES (2006b) On optimal dividend strategies in the compound Poisson model. *North Amer. Actuarial J.* 10:76–93.

Guan C, Xu ZQ (2024) Optimal ratcheting of dividend payout under Brownian motion surplus. *SIAM J. Control Optim.* 62:2590–2620.

Hubalek F, Kyprianou AE (2011) Old and new examples of scale functions for spectrally negative Lévy processes. Dalang RC, Dozzi M, Russo F, eds. *Proc. Seminar on Stochastic Analysis, Random Fields and Applications VI: Centro Stefano Franscini* (Springer, Berlin), 119–145.

lbrecher H, Boxma OJ (2004) A ruin model with dependence between claim sizes and claim intervals. *Insurance Math. Econom.* 35:245–254

Ivanovs J, Kella O (2013) Another look into decomposition results. *Queueing Systems* 75:19–28.

Jeanblanc-Picqué M, Shiryaev AN (1995) Optimization of the flow of dividends. *Russian Math. Surveys* 50:257–277.

Kella O (1998) An exhaustive Lévy storage process with intermittent output. *Stochastic Model.* 14:979–992.

Kella O, Whitt W (1991) Queues with server vacations and Lévy processes with secondary jump input. *Ann. Appl. Probab.* 1:104–117.

Kella O, Yor M (2017) Unifying the Dynkin and Lebesgue-Stieltjes formulae. *J. Appl. Probab.* 54:252–266.

Kuznetsov A, Kyprianou AE, Rivero V (2012) The theory of scale functions for spectrally negative Lévy processes. Cohen SN, Madan DB, Winkel M, eds. *Lévy Matters II: Recent Progress in Theory and Applications: Fractional Lévy Fields, and Scale Functions* (Springer, Berlin), 97–186

Kyprianou AE (2013) *Fluctuations of Lévy Processes with Applications*, 2nd ed. (Springer, Berlin).

Kyprianou AE, Loeffen RL (2010) Refracted Lévy processes. *Ann. IHP Probab. Statist.* 46:24–44.

Kyprianou AE, Rivero V (2008) Special, conjugate and complete scale functions for spectrally negative Lévy processes. *Electronic J. Probab.* 13:1672–1701.

Lee J, Kim J (2006) A workload-dependent M/G/1 queue under a two-stage service policy. *Oper. Res. Lett.* 34:531–538.

Loeffen RL (2008) On optimality of the barrier strategy in de Finetti's dividend problem for spectrally negative Lévy processes. *Ann. Appl. Probab.* 18:1669–1680.

Mandjes M, Boxma O (2023) *The Cramér-Lundberg Model and Its Variants* (Springer, Cham, Switzerland).

Mayerhofer E (2019) Three essays on stopping. *Risks* 7:1–10.

Prabhu NU (1998) *Stochastic Storage Processes: Queues, Insurance Risk, Dams, and Data Communication* (Springer, Berlin).

Sato K (2005) *Lévy Processes and Infinitely Divisible Distributions* (Cambridge University Press, Cambridge, UK).

Sun F, Zhu D (2026) Spectrally negative Lévy risk model under ratcheting dividend strategy and capital injections with transaction costs. *Communications in Statistics-Theory and Methods* 55(4):1246–1267.

Tijms HC (1975) On a switch-over policy for controlling the workload in a queueing system with two constant service rates and fixed switch-over costs. *Zeitschrift Oper. Res.* 21:19–32.

Tijms HC, van der Duyn Schouten FA (1978) Inventory control with two switch-over levels for a class of M/G/1 queueing systems with variable arrival and service rate. *Stochastic Processes Appl.* 6:213–222.

Wang W, Xu R, Yan K (2025) Optimal ratcheting of dividends with capital injection. *Math. Oper. Res.* 50:2073–2111.