Probability in the Engineering and Informational Sciences, **35**, 2021, 730–744.
doi:10.1017/S0269964820000182

STEADY-STATE OPTIMIZATION OF AN EXHAUSTIVE LÉVY STORAGE PROCESS WITH INTERMITTENT OUTPUT AND RANDOM OUTPUT RATE

ROYI JACOBOVIC AND OFFER KELLA

Department of Statistics and Data Science, The Hebrew University of Jerusalem, Jerusalem 9190501, Israel

E-mail: royi.jacobovic@mail.huji.ac.il; offer.kella@gmail.com

Consider a regenerative storage process with a nondecreasing Lévy input (subordinator) such that every cycle may be split into two periods. In the first (off), the output is shut off and the workload accumulates. This continues until some stopping time. In the second (on), the process evolves like a subordinator minus a positive drift (output rate) until it hits the origin. In addition, we assume that the output rate of every on period is a random variable, which is determined at the beginning of this period. For example, at each period, the output rate may depend on the workload level at the beginning of the corresponding busy period. We derive the Laplace–Stieltjes transform of the steady-state distribution of the workload process and then apply this result to solve a steady-state cost minimization problem with holding, setup and output capacity costs. It is shown that the optimal output rate is a nondecreasing deterministic function of the workload level at the beginning of the corresponding on period.

Keywords: Lévy driven queues, optimization with partial information, output rate control, queues with server vacations

1. INTRODUCTION

Motivated by M/G/1 queues with server vacations, [10] considers a regenerative storage process with a nondecreasing Lévy input (subordinator). In particular, the author assumes that every cycle has two periods. In the first period (off), the output is shut off and, hence, the workload accumulates. This continues until some stopping time. Then, at the second period (on), the process evolves like a subordinator minus a positive drift (output rate) until it hits the origin. This paper is about a related model in which the output rate is a random variable which changes between cycles and is determined at the beginning of every on period. For example, the output rate may be a function of the workload level or the number of customers at the beginning of every on period or may be dependent (or independent) on any other information accumulated during the preceding off period. Two possible examples for this model are the following:

1. A machine when once its workload is depleted, it undergoes some maintenance and therefore experiences a down time. During this down time, potential work flows in

© Cambridge University Press 2020 0269-9648/19 $25.00

https://doi.org/10.1017/S0269964820000182 Published online by Cambridge University Press

but is not processed until this down time ends and therefore accumulates in storage. Information about the maintenance process and the arrival of new jobs is gathered during this down time, at the end of which the operator needs to decide at which speed to set the machine. After this, it is left to its own devices until it signals that its buffer is empty again. At this time, the procedure is repeated.

2. A water dam in which water accumulates. When water reaches some level, which remains fixed, an operator must set a release rate which remains fixed until the water reaches some lower level. At that time, this process repeats. The release rate in each cycle can randomly depend on the history of rainfalls and water level since the previous time when the lower level was reached.

In what follows, we derive the Laplace–Stieltjes transform (LST) of the steady-state workload of such a system. Then, a steady-state cost minimization problem is studied, where it is assumed that the workload level at the beginning of every on period is observed. In addition, there are constant holding, setup and output capacity cost rate which create a non-trivial trade-off which has to be balanced by choosing an output rate for the upcoming on period. Once the output rate is chosen, it cannot be changed until the end of the on period. It is shown that there exists an optimal output rate policy, which is a nondecreasing deterministic function of the initial workload level. In addition, a two-phase approach is applied to solve this optimization problem. It is shown that this approach is valid even when the objective functional is not convex. Furthermore, we show that this approach is applicable to a similar problem where the only available information is the number of customers in the system at the beginning of each on period. It is assumed that the service demands of these customers are unknown, and hence, we refer to this problem as an optimization with partial information. The rest of this work is organized as follows: Section 2 is about on finding the steady-state LST of the workload process. In Section 3, we present the general statement of the cost minimization problem along with a two-phase approach for solving it. Section 4 contains two queueing examples including numerical illustrations. In Section 5, we show that the problem with partial information can be similarly solved.

1.1. Related Literature

A summary of recent developments regarding Lévy queues is given by Debicki and Mandjes [7], and a more general treatment of Lévy processes with some applications to queueing theory and other fields is provided by Kyprianou [12]. In addition, recent surveys regarding queueing models with server vacations are given in [10,15]. Related models discuss Lévy queues or more generally a Lévy-driven storage process with an exponent which is adapted over time with respect to some given mechanism. For example, Bekker [2] discusses a Lévy queue with hysteretic control. Bekker et al. [3] provide an analysis of a Lévy-driven model such that whenever the workload crosses some fixed level, a timer with a constant length is activated and when it expires, the exponent is changed, unless the origin had already been hit before. Bekker et al. [4] discuss a Lévy queue with feedback information. In addition, this work considers the case where the exponent might be changed at Poisson arrivals. Another work about a Lévy queue which evolves between exponential timers is [13]. In addition, [1] discusses the case where the exponent is controlled by an external Markovian environment. A special case of the current model is when the output rate is determined as a function of the length of the previous vacation. This makes the current paper related to [5,6] which consider the opposite case, that is, when the vacations lengths depend on the evolution of the workload process during the previous busy period.

https://doi.org/10.1017/S0269964820000182 Published online by Cambridge University Press

2. MODEL SETUP

Consider a probability space $(\Omega, \mathcal{F}, P)$ with a filtration $\mathbb{F} = (\mathcal{F}_t)_{t\geq 0}$ satisfying the usual conditions (right-continuous and augmented). With respect to this filtration, let $J \equiv (J_t)_{t\geq 0}$ be a finite-mean subordinator. Specifically, it is a nondecreasing right-continuous process which starts at the origin and has stationary independent increments. It is well known that $Ee^{-\alpha J_t} = e^{-\eta(\alpha)t}$ for $\alpha, t \geq 0$, where

$$\eta(\alpha) = c\alpha + \int_{(0,\infty)} (1 - e^{-\alpha x})\nu(\mathrm{d}x), \quad \forall \alpha \geq 0, \tag{1}$$

$c \geq 0$ and ν is the associated Lévy measure satisfying $\int_{(0,\infty)} x \wedge 1 \nu(\mathrm{d}x) < \infty$. Denote

$$\rho \equiv EJ_1 = \eta'(0) = c + \int_{(0,\infty)} x\nu(\mathrm{d}x) = c + \int_0^\infty \nu((x,\infty))\,\mathrm{d}x \tag{2}$$

and consider a regenerative storage system with the following evolution during the first cycle. The input of the storage system is J which is described above. At time $t = 0$, the system is off, that is, there is no output at that time and the workload accumulates. This continues until some stopping time τ for which $J(\tau) > 0$ with probability one and $E\tau < \infty$. In addition, note that τ is a stopping time with respect to the filtration $\mathbb{F}$ which may be richer than the filtration which is generated by J. Now, at epoch τ, a random rate R is chosen. It is assumed that $R > \rho$ with probability one. Then, the workload process $\{J_t - R(t-\tau); t \geq \tau\}$ continues until $\tau + T^\tau$ which is the first time that it hits the origin. We denote this process by $W^\tau \equiv W^\tau_t$ and an example of its sample-path is shown in the following Figure 1. In particular, the process W^τ restricted to off periods is the same as the Lévy-driven clearing process $Z^\tau \equiv Z^\tau_t$ which was discussed in Section 3 of [10]. In addition, the process W^τ restricted to on periods is denoted by $Y^\tau \equiv Y^\tau_t$. It is possible to apply the results of Sections 3 and 4 of [10] in order to derive the LST of the steady-state distribution of W^τ. Let $\tilde{Z}^\tau$ and $\tilde{Y}^\tau$ be the steady-state distributions of Z^τ and Y^τ, respectively. Thus, we shall deduce that the LST of the steady-state workload process $\tilde{W}^\tau$ is given by the following equation:

$$\tilde{W}^\tau(\alpha) = \frac{E\tau}{E\tau + E\frac{J_\tau}{R-\rho}}\tilde{Z}^\tau(\alpha) + \frac{E\frac{J_\tau}{R-\rho}}{E\tau + E\frac{J_\tau}{R-\rho}}\tilde{Y}^\tau(\alpha), \quad \forall \alpha > 0, \tag{3}$$

where $S(\alpha) = Ee^{-\alpha S}$ is a notation for the LST of a nonnegative random variable S at $\alpha > 0$. Now, the LST of $\tilde{Z}^\tau$ is given by Theorem 3.2 of [10], so it is left to figure the LST of $\tilde{Y}^\tau$. To this end, denote $V = J_\tau$ and whenever $E\frac{V}{R-\rho} < \infty$, let

$$P_{V,R}(A) \equiv \frac{E\frac{V}{R-\rho}1_A}{E\frac{V}{R-\rho}}, \quad \forall A \in \mathcal{F}. \tag{4}$$

We denote the expectation with respect to $P_{V,R}(A)$ by $E_{V,R}$, that is,

$$E_{V,R}S = \frac{E\left(\frac{V}{R-\rho}\cdot S\right)}{E\left(\frac{V}{R-\rho}\right)} \tag{5}$$

for every random variable S for which this expectation exists.

https://doi.org/10.1017/S0269964820000182 Published online by Cambridge University Press

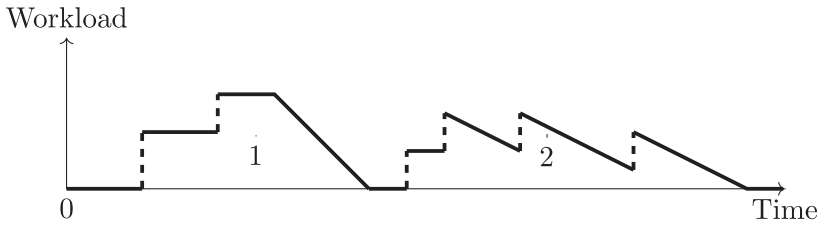


FIGURE **1.** Sample path during the first two cycles. Note: output rates during the first and second cycles are different.

THEOREM 1: *Let $U, N, e_1, e_2, \ldots$ be independent random variables such that $U \sim U[0,1]$, and $\{e_n; n \geq 1\}$ is an i.i.d sequence such that $e_1(\alpha) = \eta(\alpha)/\rho\alpha$ for every $\alpha \geq 0$. In addition, $U, e_1, e_2, \ldots$ are independent of (V, R) and $(N+1)|(V,R) \sim$ Geometric $(1-\rho/R)$. Then,*

$$\tilde{Y}^{\tau}(\alpha) = \frac{E\left[\frac{1-e^{-\alpha V}}{\alpha R - \eta(\alpha)}\right]}{E\left(\frac{V}{R-\rho}\right)} \tag{6}$$

$$= E_{V,R} \exp\left\{-\alpha\left(UV + \sum_{i=1}^{N} e_i\right)\right\}, \quad \forall \alpha \geq 0. \tag{7}$$

PROOF: Theorem 2 of [11] implies that given (V, R)

$$[\alpha R - \eta(\alpha)] \int_0^t e^{-\alpha Y_s^{\tau}}\, \mathrm{d}s + e^{-\alpha V} - 1, \quad \forall t \geq 0$$

is a zero-mean martingale. Therefore, by the optional sampling theorem with the stopping time $T^{\tau} \wedge t$, we have upon taking $t \uparrow \infty$ (monotone convergence)

$$E\left[\int_0^{T^{\tau}} e^{-\alpha Y_s^{\tau}}\, \mathrm{d}s \,\middle|\, V, R\right] = \frac{1 - e^{-\alpha V}}{\alpha R - \eta(\alpha)}, \quad P\text{-a.s.}$$

In addition, it is known (see Equation 2.10 of [10]) that

$$E[T^{\tau} \,|\, (V,R)] = \frac{V}{R-\rho}, \quad P\text{-a.s.}$$

Thus, conditioning and unconditioning with respect to (V, R) imply (6). To prove (7), observe that for every $\alpha > 0$

$$\tilde{Y}^{\tau}(\alpha) = \frac{E\left[\frac{1-e^{-\alpha V}}{\alpha R - \eta(\alpha)}\right]}{E\left(\frac{V}{R-\rho}\right)} = \frac{E\left[\frac{1-e^{-\alpha V}}{\alpha V} \cdot \frac{\alpha(R-\rho)}{\alpha R - \eta(\alpha)} \cdot \frac{V}{R-\rho}\right]}{E\left(\frac{V}{R-\rho}\right)}$$

https://doi.org/10.1017/S0269964820000182 Published online by Cambridge University Press

$$= E_{V,R}\left(\frac{1-e^{-\alpha V}}{\alpha V}\cdot\frac{1-\frac{\rho}{R}}{1-\frac{\rho}{R}\cdot\frac{\eta(\alpha)}{\rho\alpha}}\right)$$
$$= E_{V,R}\left[e^{-\alpha UV}\sum_{n=0}^{\infty}\left(1-\frac{\rho}{R}\right)\left(\frac{\rho}{R}\right)^{n}\left(Ee^{-\alpha e_1}\right)^{n}\right]$$
$$= E_{V,R}e^{-\alpha\left(UV+\sum_{i=1}^{N}e_i\right)} \tag{8}$$

and the proof is complete. ■

The following is immediate.

Corollary 1: *Denote*

$$\mu\equiv Ee_1=-\frac{\eta''(0)}{2\rho}=\frac{\int_{(0,\infty)}x^2\nu(\mathrm{d}x)}{2\rho}. \tag{9}$$

Then,

$$E\tilde{Y}^{\tau}=E_{V,R}\left(UV+\sum_{n=1}^{N}e_i\right)=\frac{E\left[\frac{V^2}{2(R-\rho)}+\mu\rho\frac{V}{(R-\rho)^2}\right]}{E\left(\frac{V}{R-\rho}\right)}. \tag{10}$$

We note that the case where $R=\rho+sV$ with $s>0$, Theorem 1 implies that

$$E\tilde{Y}^{\tau}=\frac{EV}{2}+\frac{\mu\rho}{s}EV^{-1}\xrightarrow[s\to\infty]{}\frac{EV}{2}\le\frac{EV^2}{2EV}. \tag{11}$$

The right side of (11) would be the limit if instead of taking $R=\rho+sV$, we took $R=\rho+sEV$. We also note that in both cases $EV/(R-\rho)=1/s$.

3. OPTIMIZATION

Recall the model introduced by Section 2 and note that $V=J(\tau)$. In addition, the following are model assumptions:

1. $EV^2<\infty$.
2. A constant holding cost $h\in(0,\infty)$ per unit of workload per unit of time.
3. A constant setup cost $K\in(0,\infty)$ at the beginning of every busy period.
4. A constant output capacity cost rate $d>0$. That is, if a rate r is chosen for t units of time, then the cost of the operator is dtr. This is to model the cost involved with higher consumption rate of resources as the output rate is increased.
5. There exists a constant $\rho<r<\infty$ such that $\rho<R\le r$, P-a.s. r models the maximal output rate as a consequence of technological constraint. As mentioned in Remark 2, it is required for our analysis to go through.

Our goal is to minimize the steady-state expected cost, where the minimization is overall possible distributions of the random variable R. The general expression for the expected

https://doi.org/10.1017/S0269964820000182 Published online by Cambridge University Press

cost is defined as follows:

$$\begin{aligned} C(R) &\equiv \frac{E\left[K + h\left(\int_0^\tau J_t\,\mathrm{d}t + \int_\tau^{\tau+T^\tau}[J_t - R(t-\tau)]\,\mathrm{d}t\right) + dT^\tau R\right]}{E(\tau + T^\tau)} \\ &= \frac{K + h(E\tau E\tilde{Z}^\tau + ET^\tau E\tilde{Y}^\tau) + dE[RE(T^\tau \mid R, V)]}{E\tau + ET^\tau} \\ &= \frac{hE\tau E\tilde{Z}^\tau + K + dEV + d\rho E\frac{V}{R-\rho} + hE\left[\frac{V^2}{2(R-\rho)} + \mu\rho\frac{V}{(R-\rho)^2}\right]}{E\tau + E\frac{V}{R-\rho}}. \end{aligned} \tag{12}$$

Note that the last equality is justified by Corollary 1 and the identity $E(T^\tau|R,V) = V/(R-\rho)$, P-a.s. To proceed, consider the change of variable

$$X = V\left(\frac{1}{R-\rho} - \frac{1}{r-\rho}\right) \tag{13}$$

and let

$$\begin{aligned} K_1 &\equiv hE\tau E\tilde{Z}^\tau + K + \left[d + \frac{d\rho}{r-\rho} + \frac{h\mu\rho}{(r-\rho)^2}\right]EV + \frac{h}{2(r-\rho)}EV^2, \\ K_2 &\equiv d\rho + 2\frac{h\mu\rho}{r-\rho}, \\ K_3 &\equiv E\tau + \frac{EV}{r-\rho}. \end{aligned} \tag{14}$$

Then, since $EV^2 < \infty$, an equivalent optimization problem is:

$$\begin{aligned} \min: \quad & \frac{K_1 + K_2EX + hE\left(\frac{V}{2}X + \frac{\mu\rho}{V}X^2\right)}{K_3 + EX} \\ \text{s.t.} \quad & X \in \mathcal{F} \\ & X \geq 0, \quad P\text{-a.s.} \end{aligned} \tag{15}$$

where $X \in \mathcal{F}$ means that X is a random variable.

We solve this problem in two phases.

3.1. Phase I

Fix some $\alpha > 0$ and solve the problem under the additional constraint $EX = \alpha$. That is,

$$\begin{aligned} \min: \quad & E\left[\frac{VX}{2} + \mu\rho\frac{X^2}{V}\right] \\ \text{s.t.} \quad & X \in \mathcal{F} \\ & X \geq 0, \quad P\text{-a.s.} \\ & EX = \alpha. \end{aligned} \tag{16}$$

https://doi.org/10.1017/S0269964820000182 Published online by Cambridge University Press

Let $a^+ = \max\{a, 0\}$ and $a \wedge b = \min\{a, b\}$ for every $a, b \in \mathbb{R}$. In addition, for every $v > 0$ and $\lambda \in \mathbb{R}$, define

$$\begin{aligned} x(v,\lambda) &= \frac{v\left(\lambda - \frac{v}{2}\right)^+}{2\mu\rho} = \frac{v\left(\lambda - \frac{v}{2}\right)}{2\mu\rho} 1_{(v/2,\infty)}(\lambda), \\ \xi(\lambda) &= Ex(V,\lambda), \\ f(\alpha) &= \frac{1}{4\mu\rho} EV\left[\lambda_\alpha^2 - \left(\frac{V}{2}\right)^2\right]^+ \end{aligned} \tag{17}$$

where λ_α is defined by the next theorem.

THEOREM 2: *$\xi(\cdot)$ is continuous on $[0,\infty)$, there exists $\lambda_\alpha > 0$ such that $\xi(\lambda_\alpha) = \alpha$, and $X = x(V,\lambda_\alpha)$ is a solution of (16). $f(\alpha)$ is the optimal value of (16), which is finite, convex, nondecreasing and continuous on $[0,\infty)$.*

We note that from (13) and (17), it follows that for every $\alpha, v \in [0,\infty)$ the optimal output rate

$$R = R(v,\alpha) = \rho + \left[\frac{1}{r-\rho} + \frac{\left(\lambda_\alpha - \frac{v}{2}\right)^+}{2\mu\rho}\right]^{-1} \tag{18}$$

which is nondecreasing in v. In particular, this holds for the optimal α resulting from Subsection 3. We also note that since $x(v,\lambda)$ is nondecreasing in λ, then λ_α is nondecreasing in α implying that $R(v,\alpha)$ is nonincreasing in α.

PROOF: Note that $\xi(0) = 0$. Let $\lambda \geq 0$ and notice that

$$x(V,\lambda) = \frac{V\left(\lambda - \frac{V}{2}\right)^+}{2\mu\rho} \leq \frac{\lambda^2}{\mu\rho} < \infty. \tag{19}$$

Thus, dominated convergence implies that $\xi(\cdot)$ is continuous and finite on $[0,\infty)$. In addition, monotone convergence implies that $\xi(\lambda) \to \infty$ as $\lambda \to \infty$, and hence, there exists $\lambda_\alpha > 0$ such that $\xi(\lambda_\alpha) = \alpha$.

The optimization (16) is a special case of optimizing quadratic function with random coefficients. That is,

$$\begin{aligned} \min: \quad & E[AX^2 + BX + C] \\ \text{s.t.} \quad & X \in \mathcal{F} \\ & X \geq 0, \quad P\text{-a.s.} \\ & EX = \alpha \end{aligned} \tag{20}$$

with $A = \mu\rho/V$, $B = V/2$ and $C = 0$. Thus, the fact that $X = x(V,\lambda_\alpha)$ is an optimum follows directly from Corollary 1 of [8] (see also Section 4 there). Note that $f(\alpha)$ resulting by a substitution of $x(V,\lambda_\alpha)$ into the objective function of (20) with the appropriate A, B and C. Thus,

$$f(\alpha) = \frac{1}{4\mu\rho} EV\left[\lambda_\alpha^2 - \left(\frac{V}{2}\right)^2\right]^+ \leq \frac{\lambda_\alpha}{\mu\rho} < \infty. \tag{21}$$

Moreover, observe that for $X \geq 0$, P-a.s., $0 \leq VX/2 + \mu\rho(X^2/V)$, P-a.s. Therefore, all other results regarding $f(\cdot)$ follow directly by Proposition 1 of [8]. ■

https://doi.org/10.1017/S0269964820000182 Published online by Cambridge University Press

REMARK 1: *If there exists a minimal output rate, that is, $\rho < r_{\min} \le R$, P-a.s., then it can be checked that $X \le r_0 V$, P-a.s. where $r_0 = (r_{\min} - \rho)^{-1} - (r - \rho)^{-1} > 0$. Thus, it is enough to consider $\alpha \in (0, r_0 EV]$. In addition, the results of Section 4 of [8] can be applied directly to show that there exists $\lambda_\alpha \ge 0$ such that $Ex(V, \lambda_\alpha) \wedge (r_0 V) = \alpha$, and $x(V, \lambda_\alpha) \wedge (r_0 V)$ is an optimum. In addition, Proposition 1 of [8] can be used to show that $f(\cdot)$ satisfies the same properties stated in Theorem 2.*

REMARK 2: *The proof of Theorem 2 does not require the assumption $EV^2 < \infty$. In fact, it requires only that V is such that $P(V > 0) > 0$. Therefore, if $d = 0$ and $r = \infty$, then and, hence, the results of Theorem 2 remain valid with weaker assumptions on the distribution of V. Especially, note that $r = \infty$ models a situation where at the end of an off period, the server can either choose any finite output rate and starts an on period with this rate or to pick an infinite output rate, that is to perform a clearing and then to go back on a new off period.*

3.2. Phase II

Recall that for every $\alpha \ge 0$, the optimal value $f(\alpha)$ of phase I is finite and $f(0) = 0$. The second phase is the optimization problem

$$\begin{aligned} \min: \quad & \frac{K_1 + K_2\alpha + hf(\alpha)}{K_3 + \alpha} \\ \text{s.t.} \quad & \alpha \ge 0 \end{aligned} \tag{22}$$

In addition, for every $v \ge 0$ and $\lambda \ge 0$, let

$$\begin{aligned} a(\lambda, v) &\equiv v\left(\lambda - \frac{v}{2}\right)^+, \\ b(\lambda, v) &\equiv v\left[\lambda^2 - \left(\frac{v}{2}\right)^2\right]^+, \\ A(\lambda) &\equiv Ea(\lambda, V), \\ B(\lambda) &= Eb(\lambda, V), \\ \tilde{a}(\lambda) &\equiv EV1_{[0,2\lambda]}(V). \end{aligned} \tag{23}$$

and define

$$G(\lambda) \equiv \frac{K_1 + K_2\xi(\lambda) + hf[\xi(\lambda)]}{K_3 + \xi(\lambda)} = \frac{K_1 + \frac{K_2}{2\mu\rho}A(\lambda) + \frac{h}{4\mu\rho}B(\lambda)}{K_3 + \frac{A(\lambda)}{2\mu\rho}}. \tag{24}$$

Since $\xi(\cdot)$ is a nondecreasing continuous function such that $\xi(0) = 0$ and $\xi(\lambda) \to \infty$ as $\lambda \to \infty$, then, in order to solve (15), it suffices to minimize $G(\cdot)$ on $[0, \infty)$. The following theorem guarantees that there exists a minimizer of $G(\cdot)$ which can be derived by standard line-search techniques.

THEOREM 3: *Let $\lambda^* \equiv (K_1 - K_2K_3)^+/K_3h$. Then, $G(\cdot)$ is absolutely-continuous unimodal function on $[0, \infty)$ with a minimizer in $[0, \lambda^*]$.*

https://doi.org/10.1017/S0269964820000182 Published online by Cambridge University Press

PROOF: For every $a, b \geq 0$, define

$$U(a,b) \equiv \frac{K_1 + \frac{K_2}{2\mu\rho}a + \frac{h}{4\mu\rho}b}{K_3 + \frac{a}{2\mu\rho}} \tag{25}$$

and notice that

$$G(\lambda) = U[A(\lambda), B(\lambda)], \quad \lambda \geq 0. \tag{26}$$

In addition, observe that $a(\cdot)$ and $b(\cdot)$ are convex with respect to λ and hence so are $A(\cdot)$ and $B(\cdot)$. Now, since $A(\cdot)$ and $B(\cdot)$ are convex, they are both functions of bounded variation. Therefore, since $U(\cdot)$ is differentiable on $[0,\infty)^2$, then Theorem 33 of page 81 in [14] implies that $G(\cdot)$ has a density which equals to

$$g(\lambda) = U_a[A(\lambda), B(\lambda)]D^+A(\lambda) + U_b[A(\lambda), B(\lambda)]D^+B(\lambda), \quad \forall \lambda \geq 0 \tag{27}$$

where U_a, U_b are the partial derivatives of U with respect to a and b, and D^+A, D^+B are the right derivatives of $A(\cdot)$ and $B(\cdot)$ with respect to $\lambda(\cdot)$. Note that from this, it follows that $G(\cdot)$ is absolutely continuous on $[0,\infty)$. Now, for every $v \geq 0, \lambda \geq 0$, the right derivatives of $a(\cdot)$ and $b(\cdot)$ with respect to λ are given by the following equation:

$$\begin{aligned} D^+a(\lambda, v) &= v1_{[0,2\lambda]}(v), \\ D^+b(\lambda, v) &= 2\lambda v1_{[0,2\lambda]}(v). \end{aligned} \tag{28}$$

Therefore, recalling that $EV^2 < \infty$, hence $EV < \infty$, and we have by dominated convergence that, for $\lambda > 0$,

$$\tilde{a}(\lambda) = D^+A(\lambda) = \frac{D^+B(\lambda)}{2\lambda}. \tag{29}$$

Therefore, for every $\lambda > 0$, $g(\lambda)$ is given by

$$g(\lambda) = \frac{\frac{\tilde{a}(\lambda)}{2\mu\rho}\left[K_3K_2 - K_1 + \lambda K_3 h + \frac{h}{4\mu\rho}[2\lambda A(\lambda) + B(\lambda)]\right]}{\left[K_3 + \frac{A(\lambda)}{2\mu\rho}\right]^2}. \tag{30}$$

Now, observe that $\tilde{a}(\lambda)$ and $2\lambda A(\lambda) + B(\lambda)$ are nonnegative and nondecreasing on $[0,\infty)$. Therefore, the numerator of $g(\cdot)$ can change its sign at most once. Clearly, if $K_1 \leq K_2K_3$, then $G(\cdot)$ is nondecreasing and nonnegative in which zero is a minimizer. On the other hand, if $K_1 > K_2K_3$, since the numerator converges to infinity as $\lambda \to \infty$, then there exists a minimizer which is bounded from above by λ^*, the solution of

$$K_3K_2 - K_1 + \lambda K_3 h = 0 \tag{31}$$

and the result follows. ■

REMARK 3: *From (30), it can be seen that for every $\lambda \in (0, \lambda^*)$, $g(\lambda)$ is bounded by*

$$M \equiv \frac{\tilde{a}(\lambda^*)}{2K_3^2\mu\rho}\left[K_3K_2 + K_1 + \lambda^*K_3h + \frac{h}{4\mu\rho}[2\lambda^*A(\lambda^*) + B(\lambda^*)]\right]$$

and hence, $G(\cdot)$ is Lipschitz on $[0,\lambda^*]$ with a constant M. This can be useful for the purpose of error bounds.

REMARK 4: *It is easy to see that when the distribution of V has no atoms, then $G(\cdot)$ is continuously differentiable with $G' = g$ (see (30)).*

https://doi.org/10.1017/S0269964820000182 Published online by Cambridge University Press

4. TWO QUEUEING EXAMPLES

4.1. Example 1

Consider an $M/M/1$ system such that at the beginning of every busy period, the server observes the service demand of the first customer and determines a service rate for the rest of this cycle in order to minimize (15). Let $\nu \in (0,\infty)$ be the arrival rate and assume that service demands are exponentially distributed with rate $\theta \in (0,\infty)$. Thus, in such a case, $V \sim \exp(\theta)$, $\tau \sim \exp(\nu)$, $\tilde{Z}_t = 0$, $\forall t \geq 0$, $\rho = \nu/\theta$ and $\mu = \frac{1}{2}(1 + \mathrm{Var}(V)/EV) = (1+\theta^{-1})/2$. For simplicity, assuming $r = \rho + 1$, the resulting parameters of the model are

$$\begin{aligned} K_1 &= K + \frac{d}{\theta}\left(1+\frac{\nu}{\theta}\right) + \frac{h}{\theta^2}(\nu+2), \\ K_2 &= \frac{1}{\theta}\left(d\nu + \frac{2h\nu}{\theta}\right), \\ K_3 &= \frac{1}{\nu} + \frac{1}{\theta}. \end{aligned} \tag{32}$$

In particular, since $V \sim \exp(\theta)$, then for every $\lambda \geq 0$

$$\begin{aligned} A(\lambda) &= \frac{\lambda}{\theta}\psi(2\lambda,2) - \frac{1}{\theta^2}\psi(2\lambda,3), \\ B(\lambda) &= \frac{\lambda^2}{\theta}\psi(2\lambda,2) - \frac{3}{2\theta^3}\psi(2\lambda,4), \end{aligned} \tag{33}$$

where $\psi(\cdot,n)$ is the cumulative distribution function of Erlang distribution with shape n and rate θ. Figure 2 illustrates the results of Theorem 3 regarding the graph of $G(\cdot)$ as a function of λ when $\theta = \nu = 1$. For example, the left plot presents a negative effect of setup cost on the optimal output rate. Namely, we consider $K = h = d = 1$ (blue), $K = 75,\ h = d = 1$ (red) and $K = 200,\ h = d = 1$ (green). The central plot presents a positive effect of

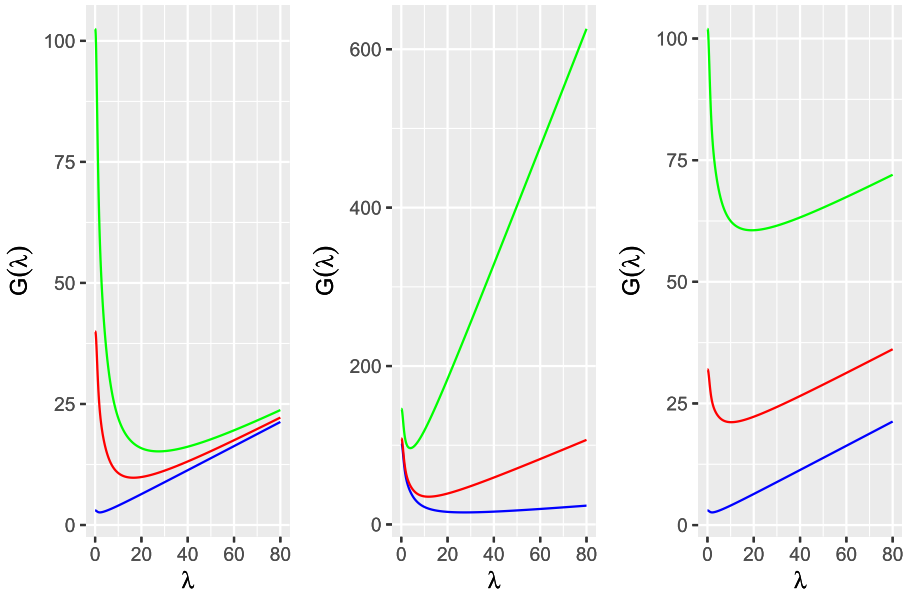


FIGURE 2. $G(\lambda)$ as a function of λ in an $M/M/1$ system.

https://doi.org/10.1017/S0269964820000182 Published online by Cambridge University Press

holding cost on the optimal output rate. That is, we consider $K = 200,\ h = d = 1$ (blue), $K = 200,\ h = 5, d = 1$ (red) and $K = 200,\ h = 30,\ d = 1$ (green). The right plot presents a negative effect of output capacity cost on the optimal output $g(\cdot)$. That is, we consider $K = h = d = 1$ (blue), $K = 1 = h,\ d = 30$ (red) and $K = h = 1,\ d = 100$ (green).

4.2. Example 2

Consider an $M/G/1$ system with a server who determines the service rate at the beginning of every busy period in order to minimize (15). Let $\nu = \frac{1}{2}$ be the arrival rate and assume that service demands are uniformly distributed on $[0, 1]$. Thus, in such a case, $V \sim U[0, 1]$, $\tau \sim \exp(\frac{1}{2})$, $\tilde{Z}_t = 0,\ \forall t \geq 0$, $\rho = 1$ and $\mu = \frac{1}{2}(1 + \mathrm{Var}(V)/EV) = \frac{7}{12}$. In addition, for simplicity, we assume that $r = \rho + 1$. Thus,

$$\begin{aligned} K_1 &= K + d + \frac{5h}{8}, \\ K_2 &= d + \frac{7h}{6}, \\ K_3 &= 2.5. \end{aligned} \tag{34}$$

Since $V \sim U[0, 1]$, then for every $0 < \lambda \leq \frac{1}{2}$, we have that $V|\{V \leq 2\lambda\} \sim U[0, 2\lambda]$. Therefore,

$$A(\lambda) = P(V \leq 2\lambda) E\left[V\left(\lambda - \frac{V}{2}\right) \middle| V \leq 2\lambda \right] = \frac{2\lambda^3}{3} \tag{35}$$

and similarly $B(\lambda) = \lambda^4$. In addition, for every $\lambda > \frac{1}{2}$,

$$A(\lambda) = EV\left(\lambda - \frac{V}{2}\right) = \frac{\lambda}{2} - \frac{1}{6} \tag{36}$$

and

$$B(\lambda) = E\left(\lambda^2 V - \frac{V^3}{4}\right) = \frac{\lambda^2}{2} - \frac{1}{16}. \tag{37}$$

An insertion of these results into the definition of $g(\cdot)$ implies that

$$G(\lambda) = \begin{cases} \dfrac{K_1 + \frac{K_2}{3\mu\rho}\lambda^3 + \frac{h}{4\mu\rho}\lambda^4}{K_3 + \frac{\lambda^3}{3\mu\rho}} & 0 \leq \lambda \leq \frac{1}{2} \\ \dfrac{K_1 + \frac{K_2}{2\mu\rho}\left(\frac{\lambda}{2} - \frac{1}{6}\right) + \frac{h}{4\mu\rho}\left(\frac{\lambda^2}{2} - \frac{1}{16}\right)}{K_3 + (2\mu\rho)^{-1}\left(\frac{\lambda}{2} - \frac{1}{6}\right)} & \frac{1}{2} < \lambda < \infty \end{cases}.$$

Figure 3 illustrates the graph of $G(\cdot)$ as a function of λ when the costs are varying. The left plot presents a negative effect of setup cost on the optimal output rate. Namely, we consider $K = h = d = 1$ (blue), $K = 75,\ h = d = 1$ (red) and $K = 200,\ h = d = 1$ (green). The central plot presents a positive effect of holding cost on the optimal output rate. That is, we consider $K = 200,\ h = d = 1$ (blue), $K = 200,\ h = 5, d = 1$ (red) and $K = 200,\ h = 10,\ d = 1$ (green). The right plot presents a positive effect of the output capacity cost rate on the optimal output $G(\cdot)$. That is, we consider $K = 200,\ h = d = 1$ (blue), $K = 200,\ h = 1,\ d = 50$ (red) and $K = 200,\ h = 1,\ d = 100$ (green). Interestingly, unlike the case presented by Figure 2, Figure 3 presents a case when the output capacity cost rate d has a positive effect on the optimal output rate.

https://doi.org/10.1017/S0269964820000182 Published online by Cambridge University Press

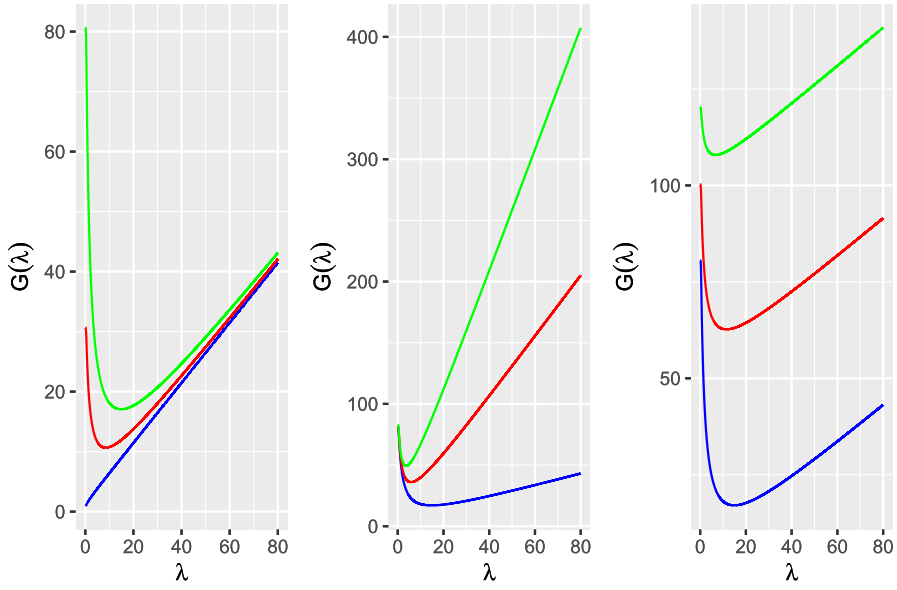


FIGURE **3.** $G(\lambda)$ as a function of λ in an $M/G/1$ system with uniform service demands.

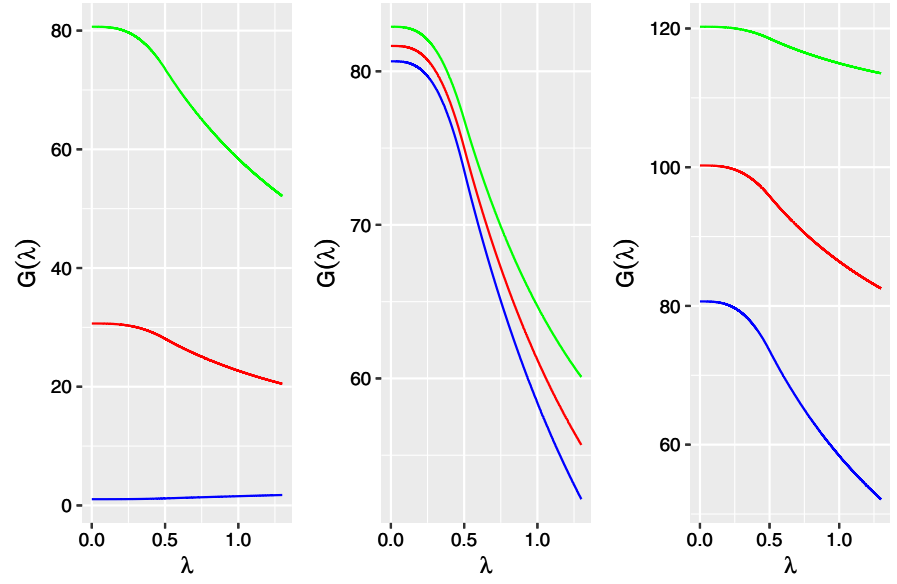


FIGURE **4.** $G(\lambda)$ as a function of λ in an $M/G/1$ system with uniform service demands. Note that the graph is neither convex nor concave.

Figure 4 presents the same graphs which are sketched in Figure 3 on the interval $[0, 1.3]$. Especially, notice that all the graphs are smooth at 0.5. In addition, Figure 4 demonstrates the fact that the graph of $G(\cdot)$ can be neither convex nor concave.

https://doi.org/10.1017/S0269964820000182 Published online by Cambridge University Press

5. OPTIMIZATION WITH PARTIAL INFORMATION

So far, the assumption was that the decision maker knows the exact value of V when she has to pick the output rate. To see a slightly different point of view regarding this model, consider the case where $V = \sum_{k=1}^{N} \tilde{V}_i$ such that $\tilde{V}_1, \tilde{V}_2, \ldots$ are i.i.d nonnegative random variables, and N is some discrete random variable which is independent of $\{\tilde{V}_i\}_{i=1}^{\infty}$. In particular, we assume that $E\tilde{V}_1 = \delta$ and $\mathrm{Var}(\tilde{V}_1) = \sigma^2$. In the context of queueing systems, N is interpreted as the number of customers in the system when the service rate is chosen. Now, in this part, the assumption is that only N is observed (and not the service times) when the output rate is picked. That is, we assume that (N, R) and $\tilde{V}_1, \tilde{V}_2, \ldots$ are independent. Observe that the laws of total expectation and variance imply that

$$\begin{aligned}
&\frac{1}{2}E\frac{V^2}{R-\rho} + \mu\rho E\frac{V}{(R-\rho)^2} \\
&\quad = \frac{1}{2}\left\{\mathrm{Var}\left(\frac{V}{\sqrt{R-\rho}}\right) + \left(E\frac{V}{\sqrt{R-\rho}}\right)^2\right\} \\
&\qquad + \mu\rho\delta E\frac{N}{(R-\rho)^2} \\
&\quad = \frac{\sigma^2}{2}E\frac{N}{R-\rho} + \frac{\delta^2}{2}E\frac{N^2}{R-\rho} + \mu\rho\delta E\frac{N}{(R-\rho)^2}. \qquad \textbf{(38)}
\end{aligned}$$

Thus, an insertion of (38) in (12) provides an expression with numerator

$$\begin{aligned}
&hE\tau E\tilde{Z}^{\tau} + K + d\delta EN \\
&\quad + \left(d\rho\delta + \frac{h\sigma^2}{2}\right)E\frac{N}{R-\rho} + h\left(\frac{\delta^2}{2}E\frac{N^2}{R-\rho} + \mu\rho\delta E\frac{N}{(R-\rho)^2}\right) \qquad \textbf{(39)}
\end{aligned}$$

and denominator $E\tau + \delta E(N/(R-\rho))$. Thus, an appropriate change of variable implies that this minimization is analogous to a special case of (15) with V which has a discrete distribution with a support which is a subset of $\{1, 2, \ldots\}$.

5.1. When V Has a Discrete Distribution

As explained in the previous part of this section, now V models the number of customers in the system at the beginning of an on period. Therefore, we assume that V is a random variable which receives values on the set $\{1, 2, \ldots\}$ with probabilities

$$p_i \equiv P(V = i), \quad \forall i = 1, 2, \ldots \qquad \textbf{(40)}$$

such that $\sum_{i=1}^{\infty} p_i = 1$. In such a case, for every $\lambda \geq 0$

$$A(\lambda) = \sum_{i=1}^{\infty} p_i i\left(\lambda - \frac{i}{2}\right)^+ = \sum_{i=1}^{\lfloor 2\lambda \rfloor} p_i i\left(\lambda - \frac{i}{2}\right) \qquad \textbf{(41)}$$

and similarly

$$B(\lambda) = \sum_{i=1}^{\lfloor 2\lambda \rfloor} p_i i\left(\lambda^2 - \frac{i^2}{4}\right). \qquad \textbf{(42)}$$

https://doi.org/10.1017/S0269964820000182 Published online by Cambridge University Press

Therefore, for every $\lambda \geq 0$

$$G(\lambda) = \frac{K_1 + \sum_{i=1}^{\lfloor 2\lambda \rfloor} p_i \left[\frac{K_2}{2\mu\rho} i \left(\lambda - \frac{i}{2} \right) + \frac{hi}{4\mu\rho} \left(\lambda^2 - \frac{i^2}{4} \right) \right]}{K_3 + (2\mu\rho)^{-1} \sum_{i=1}^{\lfloor 2\lambda \rfloor} p_i i \left(\lambda - \frac{i}{2} \right)}. \tag{43}$$

Now, by Theorem 3, an efficient line search method can be used in order to find $i \in \{\frac{1}{2}, \frac{2}{2}, \frac{3}{2}, \ldots\}$ such that the optimal λ belongs to $I = [i - \frac{1}{2}, i)$. Then, for every $\lambda \in I$,

$$G(\lambda) = \frac{S_i \lambda^2 + T_i \lambda + U_i}{Q_i + W_i \lambda} \equiv G_i(\lambda) \tag{44}$$

where

$$\begin{aligned} S_i &= \frac{h}{4\mu\rho} \sum_{j=1}^{2i-1} p_j j, \\ T_i &= \frac{K_2}{2\mu\rho} \sum_{j=1}^{2i-1} p_j j, \\ U_i &= K_1 - \sum_{j=1}^{2i-1} j^2 \left(\frac{K_2}{4\mu\rho} + \frac{hj}{16\mu\rho} \right), \\ Q_i &= K_3 - \frac{1}{4\mu\rho} \sum_{j=1}^{2i-1} p_j j^2, \\ W_i &= (2\mu\rho)^{-1} \sum_{j=1}^{2i-1} p_j j \end{aligned} \tag{45}$$

with the convention that an empty sum equals zero. Finally, it is left to minimize $G_i(\cdot)$ on I. This can be done analytically, and hence, unlike the general case, when only partial information is available, the optimal output rate can be calculated efficiently with no error at all.

Acknowledgment

This work was supported by Grant No. 1647/17 from the Israel Science Foundation and the Vigevani Chair in Statistics.

References

1. Asmussen, S. & Kella, O. (2000). A multi-dimensional martingale for Markov additive processes and its applications. *Advances in Applied Probability* 32(2): 376–393.
2. Bekker, R. (2009). Queues with Lévy input and hysteretic control. *Queueing Systems* 63(1–4): 281.
3. Bekker, R., Boxma, O.J., & Kella, O. (2008). Queues with delays in two-state strategies and Lévy input. *Journal of Applied Probability* 45(2): 314–332.
4. Bekker, R., Boxma, O.J., & Resing, J.A.C. (2009). Lévy processes with adaptable exponent. *Advances in Applied Probability* 41(1): 177–205.
5. Boxma, O., Mandjes, M., & Kella, O. (2008). On a queuing model with service interruptions. *Probability in the Engineering and Informational Sciences* 22(4): 537–555.
6. Boxma, O., Kella, O., & Mandjes, M. (2010). On a generic class of Lévy-driven vacation models. *Probability in the Engineering and Informational Sciences* 24(1): 1–12.
7. Debicki, K. & Mandjes, M (2015). *Queues and Lévy fluctuation theory*. Springer.
8. Jacobovic, R. & Kella, O. (2019). Minimizing a stochastic convex function subject to stochastic constraints and some applications. arXiv:1906.09604.

https://doi.org/10.1017/S0269964820000182 Published online by Cambridge University Press

9. Ke, J.C., Wu, C.H., & Zhang, Z.G. (2010). Recent developments in vacation queueing models: A short survey. *International Journal of Operations Research* 7(4): 3–8.
10. Kella, O. (1998). An exhaustive Lévy storage process with intermittent output. *Stochastic Models* 14(4): 979–992.
11. Kella, O. & Whitt, W. (1992). Useful martingales for stochastic storage processes with Lévy input. *Journal of Applied Probability* 29(2): 396–403.
12. Kyprianou, A.E (2006). *Introductory lectures on fluctuations of Lévy processes with applications*. Springer Science and Business Media.
13. Palmowski, Z. & Vlasiou, M. (2011). A Lévy input model with additional state-dependent services. *Stochastic Processes and Their Applications* 121(7): 1546–1564.
14. Protter, P.E. (2004). *Stochastic integration and differential equations*, 2nd ed. Springer.
15. Tian, N. & Zhang, Z.G (2006). *Vacation queueing models: theory and applications*, vol. 93. Springer Science & Business Media.

https://doi.org/10.1017/S0269964820000182 Published online by Cambridge University Press