On the non-Markovian multiclass queue under risk-sensitive cost

Rami Atar*† Gal Mendelson*†

February 25, 2016; revised July 26, 2016

Abstract

This paper studies a control problem for the multiclass G/G/1 queue for a risk-sensitive cost of the form $n^{-1} \log E \exp \sum_i c_i X_i^n(T)$, where $c_i > 0$ and $T > 0$ are constants, X_i^n denotes the class-i queue length process, and the number of arrivals and service completions per unit time are of order n. The main result is the asymptotic optimality, as $n \to \infty$, of a priority policy, provided that c_i are sufficiently large. Such a result has been known only in the Markovian (M/M/1) case. The index which determines the priority is explicitly computed in the case of Gamma distributed inter-arrival and service times.

AMS subject classifications: 60F10, 60K25, 49N70, 93E20

Keywords: Multi-class G/G/1, risk-sensitive control, large deviations

1 Introduction

We consider the multiclass single server queue, where the arrival and potential service processes are of renewal type. Denote by X_i^n the queue length process of class-i customers in an initially empty system, where the number of arrivals and service completions per unit time scale like n. We seek how to schedule jobs so as to asymptotically minimize the risk-sensitive cost (RSC) $n^{-1} \log E \exp \sum_i c_i X_i^n(T)$, as $n \to \infty$, where $c_i > 0$ and $T > 0$ are constants. Such a cost emphasizes large values of queue length, and so it is of interest when avoiding events such as large buffer overflow or large waiting times is important. This problem has been studied in [2] in the Markovian (M/M/1) setting where it was shown that for c_i sufficiently large, prioritizing service to the classes according to a fixed index is asymptotically optimal (AO). This index is given by $(1 - e^{-c_i})\mu_i$ in case when the service rates are given by $\mu_i n$, and $\mu_i > 0$ are constants. For a broad family of RSC (including the one mentioned above with general constants c_i) it is known by the results of [1] in the Markovian setting, that an AO policy can be identified in terms of a differential game. However, explicit solutions of the differential game are not available in general. The main result of this paper is the extension of the result of [2] to the non-Markovian setting. Namely, we prove that a certain fixed priority policy is AO, assuming c_i are large enough. The index which determines the priority is expressed in terms of the local large deviation (LD) rate functions of the underlying renewal processes alluded to above. In the special case of Gamma distribution,

*Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel

†Research supported in part by the ISF (grant 1315/12)

this index is computed explicitly, and is given by $\theta_i^{-1}(1-e^{-c_i/\kappa_i})$, where the class-$i$ service time is distributed according to Gamma(κ_i,θ_i). The exponential case alluded to above is recovered by setting $\kappa_i = 1$.

Whereas the analysis in the aforementioned works was based on differential games as well as PDE techniques (where the latter refers to [1]), the approach in this paper is to directly estimate the RSC by means of Varadhan's lemma, using LD properties of renewal processes known from the work of Puhalskii and Whitt [6]. Such a direct approach is made possible by identifying an upper and a lower bound on the RSC that asymptotically match one another.

Closely related to this paper is the work by Stolyar and Ramanan [7]. While [7] does not address a RSC, it considers the same model (in a non-Markovian setting) in relation to a LD type cost. The policy studied there, called the *largest weighted delay first* (LWDF) scheduling, prioritizes the classes dynamically by always choosing the customer that has the largest delay, with possible weights for different classes. This policy was shown to asymptotically minimize the decay rate of excessive wait probabilities in stationarity. Thus [7] analyzes a system in steady state (and, in fact, assumes stability conditions), whereas this paper looks at an initially empty queue and provides a finite horizon analysis. With regard to the information required for the scheduling policies to operate, LWDF and the fixed priority policy identified in this paper are on two opposite extremes. LWDF operates without knowing the statistical properties of the stochastic primitives, but requires knowledge of the state of the system at every decision time. The index which determines the priority policy studied in this paper depends on the service time distributions, but does not require knowing the state of the system (besides, of course, which of the buffers are empty at the moment of decision). Thus LWDF is robust to perturbations in the underlying distributions, whereas a fixed priority policy is, in some applications, easier to implement. More importantly, because the priority policy's index depends on the distributions, it also gives significant information on them. Namely, it identifies the class which behaves as the bottleneck with regard to the cost of interest, in the sense that the highest priority class is the one where building up large queues contributes most to the cost. By specifying the index, the result thus indicates which statistical properties govern the bottleneck.

This paper is organized as follows. In Section 2 we introduce the model and the main result, and state an open problem. Section 3 gives an explicit computation of the index in the case of Gamma distribution. The proof of the main result is presented in Section 4, where Subsections 4.1 and, respectively, 4.2 provide matching upper and lower bounds.

2 Model and main result

The multiclass G/G/1 model considered has a single server and $I \geq 2$ buffers with infinite room, where each buffer is dedicated to a class of customers. Customers that arrive into the system are queued in the corresponding buffers. Within each class, service is provided in the order of arrival, where the server may only serve the customer at the head of each line. Processor sharing is allowed, and so the server is capable of serving up to I customers (of distinct classes) simultaneously. It is assumed that the system starts empty. Arrivals occur according to independent renewal processes, and service times are independent and identically distributed for each class. Let parameters $\lambda_i > 0, i \in \mathcal{I} := \{1, 2, \ldots, I\}$ be given, representing the *reciprocal mean inter-arrival times* of class-i customers. Let $\{IA_i(l) : l \in \mathbb{N}\}_{i\in\mathcal{I}}$ be independent sequences of strictly positive i.i.d. random variables with mean $\mathbb{E}[IA_i(1)] = 1/\lambda_i$, $i \in \mathcal{I}$. With $\sum_1^0 = 0$, the number of arrivals of class-i

customers up to time t is given by

$$A_i(t) = \sup\Big\{l \ge 0 : \sum_{k=1}^{l} IA_i(k) \le t\Big\}, \quad t \ge 0.$$

Similarly, let parameters $\mu_i > 0, i \in \mathcal{I}$ be given, representing *reciprocal mean service times*. Let independent sequences $\{ST_i(l) : l \in \mathbb{N}\}_{i\in\mathcal{I}}$ of strictly positive i.i.d. random variables (independent of the sequences $\{IA_i\}$) with mean $\mathbb{E}[ST_i(1)] = 1/\mu_i$. The time required to complete the service of the l-th class-i customer is given by $ST_i(l)$, and the *potential service time* processes are defined as

$$S_i(t) = \sup\Big\{l \ge 0 : \sum_{k=1}^{l} ST_i(k) \le t\Big\}, \quad t \ge 0.$$

Let $A = (A_i)_{i\in\mathcal{I}}$ and $S = (S_i)_{i\in\mathcal{I}}$.

For $i \in \mathcal{I}$, let X_i represent the number of class-i customers in the system, and write $X = (X_i)_{i\in\mathcal{I}}$. Let B be a process taking values in $\mathbb{U} := \{u \in \mathbb{R}_+^I : \sum_{i\in\mathcal{I}} u_i \le 1\}$, representing the fraction of effort devoted by the server to the various customer classes. The number of service completions of class-i jobs during the time interval $[0, t]$ is then given by

$$D_i(t) := S_i(T_i(t)), \tag{1}$$

where

$$T_i(t) = \int_0^t B_i(s)ds. \tag{2}$$

We thus have

$$X_i(t) = A_i(t) - D_i(t) = A_i(t) - S_i(T_i(t)), \qquad t \ge 0. \tag{3}$$

Note that, by construction, the arrival and potential service processes have RCLL paths, and accordingly, so do D and X. It is also assumed that B has RCLL paths.

The process B is regarded a control, that is determined based on observations from the past (and present) events in the system. A precise definition is as follows. The process B is said to be an *admissible control* if

- It is adapted to the filtration

$$\sigma\{A_i(s), S_i(T_i(s)), i \in \mathcal{I}, s \le t\},$$

  where T_i are given by (2);

- For $i \in \mathcal{I}$ and $t \ge 0$, one has

$$X_i(t) = 0 \quad \text{implies} \quad B_i(t) = 0, \tag{4}$$

  where X_i are given by (3).

Denote the class of all admissible control processes B by $\mathcal{B}$. Note that this class depends on the processes A and S, but we consider these processes to be fixed.

Denote by $\hat{\ell}_i(x) = \log \mathbb{E}[e^{xIA_i(1)}]$ and $\hat{k}_i(x) = \log \mathbb{E}[e^{xST_i(1)}]$ the cumulant generating functions for the interarrival and service time distributions. Our main assumptions on these distributions are as follows.

Assumption 2.1. *i. For every* $\gamma \in \mathbb{R}$, $\limsup_{t\to\infty} t^{-1} \log \mathbb{E}[e^{\gamma A_i(t)}] < \infty$*,* $i \in \mathcal{I}$*.*
ii. $\hat{\ell}_i(x) < \infty$ *and* $\hat{k}_i(x) < \infty$ *for some* $x > 0$ *and all* $i \in \mathcal{I}$*.*

Remark 2.1. *A sufficient condition for Assumption 2.1(i) is the existence of a constant* $c > 0$ *such that* $\mathbb{P}(IA_i(1) < \alpha) \le c\alpha$ *for all* $\alpha \ge 0$*.*

Denote $x_i^* = \sup\{x : \hat{\ell}_i(x) < \infty\}$ and $x_i^\# = \sup\{x : \hat{k}_i(x) < \infty\}$, and note that $x_i^* > 0$ and $x_i^\# > 0$. Let

$$\ell_i(y) = \sup_{x<x_i^*} (x - y\hat{\ell}_i(x)), \qquad k_i(y) = \sup_{x<x_i^\#} (x - y\hat{k}_i(x)). \tag{5}$$

Throughout, let $T \in (0,\infty)$ be fixed. Let $\mathcal{AC}$ denote the class of absolutely continuous functions mapping $[0,T] \to \mathbb{R}$, $\mathcal{AC}_0 = \{a \in \mathcal{AC} : a(0) = 0\}$, and

$$\mathbb{L}_i(a) = \begin{cases} \displaystyle\int_0^T \ell_i(\dot{a})dt, & \text{if } a \in \mathcal{AC}_0, \\ +\infty, & \text{otherwise,} \end{cases} \qquad \mathbb{K}_i(s) = \begin{cases} \displaystyle\int_0^T k_i(\dot{s})dt, & \text{if } s \in \mathcal{AC}_0, \\ +\infty, & \text{otherwise.} \end{cases} \tag{6}$$

Let rescaled versions of the arrival and service processes be defined by

$$A^n(t) = \frac{1}{n}A(nt), \qquad S^n(t) = \frac{1}{n}S(nt), \qquad t \in [0,T].$$

Then it is known by [6] (Theorem 6.1) that, for each i, the processes A_i^n [resp., S_i^n] satisfy the Large Deviations Principle (LDP) in $\mathbb{D} = \mathbb{D}([0,T] : \mathbb{R})$ with the J_1 topology, with the good rate function $\mathbb{L}_i$ [resp., $\mathbb{K}_i$] (see [4] for the terminology; in particular, the term 'good' refers to having compact sublevel sets).

We have already introduced rescaled versions of the processes A and S, and we now let

$$X^n(t) = \frac{1}{n}X(nt), \qquad T^n(t) = \frac{1}{n}T(nt), \qquad t \in [0,T]. \tag{7}$$

By (3),

$$X_i^n(t) = A_i^n(t) - S_i^n(T_i^n(t)). \tag{8}$$

Fix $c \in (0,\infty)^I$. For each $n \in \mathbb{N}$ consider the RS cost and the corresponding value, given by

$$J^n(B) = \frac{1}{n}\log \mathbb{E}\Big[e^{nc\cdot X^n(T)}\Big], \qquad B \in \mathcal{B}, \qquad V^n = \inf_{B\in\mathcal{B}} J^n(B). \tag{9}$$

We are interested in the asymptotics

$$\overline{V} = \limsup_n V^n, \qquad \underline{V} = \liminf_n V^n.$$

Our main result is the asymptotic optimality of a fixed priority policy. By this we mean that we fix an ordering and apply preemptive-resume prioritization according to

$$B_1 = 1_{\{X_1>0\}}, \qquad B_i = 1_{\{\sum_{j=1}^{i-1} X_j=0,\, X_i>0\}}, \qquad i \ge 2. \tag{10}$$

This relation defines uniquely the processes B, X, since (1), (2), (3) and (10) have a unique solution (which can be argued by induction on the jump times), which moreover satisfies the definition of an admissible control. We will denote the control process thus defined by B^*.

For $i \in \mathcal{I}$ denote $C_i^* = \sup_{z\ge 0}(c_iz - \ell_i(z))$ and $C_i^\# = \inf_{z\ge 0}(c_iz + k_i(z))$.

Theorem 2.1. *Let Assumption 2.1 hold. Assume also that $C_i^* > C_i^{\#}$ for every i. Let the classes be labeled in such a way that $C_1^{\#} \geq C_2^{\#} \geq \cdots \geq C_I^{\#}$. Consider the priority policy introduced above, and the corresponding admissible control B^* of* (10). *Then*

$$\lim_n J^n(B^*) = \overline{V} = \underline{V} = V := T\Big[\sum_i C_i^* - C_1^{\#}\Big].$$

Remark 2.2. *Notice that, when the distributions of $IA_i(1)$ and $ST_i(1)$ all have unbounded support, one has $C^* > C^{\#}$ whenever the constants c_i are large enough. Indeed, the functions $\hat{\ell}_i$ and $\hat{k}_i$ are then superlinear in this case, and therefore ℓ_i and k_i are finite, by which*

$$C^* - C^{\#} = \sup_{z\geq 0, \hat{z}\geq 0} (c(z-\hat{z}) - \ell(z) - k(\hat{z})) \geq c(z_1 - z_2) - \ell(z_1) - k(z_2),$$

for some fixed $z_1 > z_2$. Obviously, this argument is still valid when ℓ_i and k_i are only finite on a common interval.

The condition $C_i^* > C_i^{\#}$, $i \in \mathcal{I}$ plays an important role in the proof of the result, as elaborated in Remark 4.3 below. It is natural to ask whether this condition can be relaxed. We leave this as an open question (this question has been resolved in the Markovian case; see Section 3).

Problem 2.1. *Does there exist an AO policy of fixed priority type for general c_i? If so, can the index be computed?*

3 Gamma distributed service time

In this section we evaluate the priority index for Gamma distributed service times, extending the case of exponential service times known from [2]. Modeling-wise, the significance of this distribution is that it includes as a special case the Erlang distribution, which corresponds to the service time of a job that takes a fixed number of steps to complete, where each step is exponentially distributed.

Thus, let the i-class service time be distributed according to $Gamma(\kappa_i, \theta_i)$, by which the density function is given by $\Gamma(\kappa_i)^{-1}\theta_i^{-\kappa_i}x^{\kappa_i - 1}e^{-x/\theta_i}$, $x > 0$. In what follows, we drop the subscript i for simplicity.

The log moment generating function can be computed to give

$$\hat{k}(x) = \log Ee^{xST} = -\kappa \log(1-\theta x) \quad x < x^* := \frac{1}{\theta},$$

$\hat{k}(x) = \infty$, $x \geq x^*$. Calculating k by the formula $k(z) = \sup_{x<x^*}\{x - z\hat{k}(x)\}$ gives, for $z \geq 0$,

$$k(z) = \frac{1}{\theta} - \kappa z + \kappa z \log(\theta\kappa z). \tag{11}$$

To calculate the index $C^{\#} = \inf_{z\geq 0}(cz + k(z))$, note that $k(z)$, $z \geq 0$, is differentiable and its derivative is invertible. Namely, $k'(z) = \kappa\log(\kappa z\theta)$, and

$$k'^{-1}(z) = \frac{1}{\kappa\theta}e^{z/\kappa}, \quad z \in \mathbb{R}.$$

Therefore, the minimizing z in the expression for $C^{\#}$ is the solution of $c + k'(z) = 0$, given by $k'^{-1}(-c)$. Thus

$$
\begin{aligned}
C^{\#} &= ck'^{-1}(-c) + k(k'^{-1}(-c)) \\
&= c\frac{1}{\kappa\theta}e^{-c/\kappa} + k\big(\frac{1}{\kappa\theta}e^{-c/\kappa}\big) \\
&= c\frac{1}{\kappa\theta}e^{-c/\kappa} + \frac{1}{\theta} - \kappa\frac{1}{\kappa\theta}e^{-c/\kappa} + \kappa\frac{1}{\kappa\theta}e^{-c/\kappa}\log\big(\theta\kappa\frac{1}{\kappa\theta}e^{-c/\kappa}\big) \\
&= c\frac{1}{\kappa\theta}e^{-c/\kappa} + \frac{1}{\theta} - \frac{1}{\theta}e^{-c/\kappa} - \frac{1}{\theta}e^{-c/\kappa}\frac{c}{\kappa} \\
&= \frac{1}{\theta}\big(1 - e^{-\frac{c}{\kappa}}\big).
\end{aligned} \tag{12}
$$

As a special case, take $\kappa = 1$ and recover the case of an exponential with parameter $\mu = 1/\theta$, namely $\mu(1 - e^{-c})$ (see [2]). One may contrast this index with the well known $c\mu$ index, that is known to be optimal for *risk neutral* queue length cost with weights c_i, where the mean service times are given by μ_i^{-1}. In (12), both c and the parameters of the distribution enter nonlinearly.

The optimality obtained in [2] of the index $\mu(1 - e^{-c})$ in the Markovian case has been proved there under the assumption $c > \log(\mu/\lambda)$. As shown below, in this case, this assumption coincides with the assumption $C_i^* > C_i^{\#}$, $i \in \mathcal{I}$. Recently, Anup Biswas [3] has settled Problem 2.1 above in the Markovian case, by showing that the result is valid for any set of parameters c_i, thus extending the validity of the main result of [2] beyond the assumption $c > \log(\mu/\lambda)$.

To further discuss the main result, let us assume that the interarrivals are also modeled as Gamma distributed, and let $\kappa_{i,a}$ and $\theta_{i,a}$ denote their parameters. Also, in what follows, write $\kappa_{i,s}$ and $\theta_{i,s}$ for the Gamma distribution parameters of the service times. Note by Remark 2.1, that Assumption 2.1(i) holds provided $\kappa_{i,a} \geq 1$, and by the above calculation, so does Assumption 2.1(ii). By Remark 2.2, the condition $C_i^* > C_i^{\#}$, $i \in \mathcal{I}$, holds whenever c_i are sufficiently large. In what follows, we give a more concrete sufficient condition for this.

Since $C^* = \sup_{z\geq 0}(cz - \ell(z))$ (where we again omit the dependence on i), and $\ell(z)$ is given in a form similar to (11), the maximizing z is the solution of $c = \ell'(z)$, namely $z = \ell'^{-1}(c)$. Hence

$$
C^* = c\ell'^{-1}(c) - \ell(\ell'^{-1}(c)) = \frac{1}{\theta}\big(e^{\frac{c}{\kappa}} - 1\big).
$$

We can write the condition $C^* > C^{\#}$ as

$$
\frac{1}{\theta_a}\big(e^{\frac{c}{\kappa_a}} - 1\big) > \frac{1}{\theta_s}\big(1 - e^{-\frac{c}{\kappa_s}}\big). \tag{13}
$$

Since the right hand side is bounded from above by $1/\theta_s$, the condition

$$
c > \kappa_a \log(1 + \theta_a/\theta_s)
$$

is sufficient.

As a special case, consider exponential interarrival and service times, with $\kappa_a = \kappa_s = 1$, $1/\theta_a = \lambda$, $1/\theta_s = \mu$. Then the condition (13) takes the form $\lambda\big(e^c - 1\big) > \mu\big(1 - e^{-c}\big)$, that can be written as $c > \log(\mu/\lambda)$.

4 Proof of main result

In the first part of the proof we provide an upper bound on the cost attained under the priority policy, showing

$$\limsup_n J^n(B^*) \le V. \tag{14}$$

Next we show that

$$\underline{V} \ge V. \tag{15}$$

Together, (14) and (15) imply

$$V \le \underline{V} \le \liminf_n J^n(B^*) \le \limsup_n J^n(B^*) \le V,$$

as well as

$$V \le \underline{V} \le \overline{V} \le \limsup_n J^n(B^*) \le V,$$

by which Theorem 2.1 follows.

4.1 An upper bound under the fixed priority policy

We provide an upper bound on the cost attained under the priority policy, by showing (14). To this end, fix $\eta > 0$ and define the mapping $\boldsymbol{X} : \mathbb{D}^2 \to \mathbb{R}$ by

$$\boldsymbol{X}(a,s) = \sup_{\alpha,\beta,\gamma,\delta} \big[a(\beta) - a(\beta-\alpha) - s(\delta) + s(\delta-\gamma) - \eta(\alpha-\gamma)^+\big], \qquad (a,s) \in \mathbb{D}^2,$$

where the supremum is performed over variables $\alpha, \beta, \gamma, \delta$ which satisfy

$$0 \le \alpha \le \beta \le T, \quad 0 \le \gamma \le \delta \le T. \tag{16}$$

The term involving η in the definition of $\boldsymbol{X}$ plays the role of a soft version of the hard constraint $\alpha \le \gamma$, when the parameter η is large. Defining $\boldsymbol{X}$ this way gives rise to a *continuous* map, as we argue at a later stage of the proof.

Let us argue that

$$X_1^n(T) \le \boldsymbol{X}(A_1^n, S_1^n) + n^{-1}. \tag{17}$$

Denote $r = r_n = \sup\{u \in [0,T] : X_1^n(u) = 0\}$. If $X_1^n(T) = 0$ then (17) holds by the nonnegativity of $\boldsymbol{X}(\cdot,\cdot)$. Otherwise, by right-continuity of the sample paths, and since the jumps of the arrival processes are all of size 1, we have $X_1^n(r) = n^{-1}$. The policy under consideration gives preemptive priority to class 1, by which $T_1(t) = \int_0^t 1_{\{X_1(u)>0\}} du$ hence $T_1^n(t) = \int_0^t 1_{\{X_1^n(u)>0\}} du$. Hence by (8),

$$X_1^n(T) = X_1^n(r) + A_1^n(T) - A_1^n(r) - S_1^n(T_1^n(T)) + S_1^n(T_1^n(r)).$$

Using $X_1^n(r) = n^{-1}$ and $T_1^n(T) - T_1^n(r) = T - r$, denoting $\hat{r} = T_1^n(r)$, we have

$$X_1^n(T) = n^{-1} + A_1^n(T) - A_1^n(r) - S_1^n(\hat{r} + T - r) + S_1^n(\hat{r}).$$

The bound (17) follows by taking $\alpha = \gamma = T - r$, $\beta = T$, $\delta = \hat{r} + T - r$.

For $i \geq 2$, use the bound $X_i^n(T) \leq A_i^n(T)$, which follows from (8). This and (17) give

$$\begin{aligned}\mathbb{E}\Big[e^{nc\cdot X^n(T)}\Big] &\leq \mathbb{E}\Big[e^{nc_1[\boldsymbol{X}(A_1^n,S_1^n)+n^{-1}]+\sum_{i\geq 2} nc_iA_i^n(T)}\Big]\\ &\leq \mathbb{E}\Big[e^{nc_1[\boldsymbol{X}(A_1^n,S_1^n)+n^{-1}]}\Big]\times\prod_{i\geq 2}\mathbb{E}\Big[e^{nc_iA_i^n(T)}\Big].\end{aligned}$$

Next we apply Varadhan's lemma (Theorem 4.3.1 of [4]) for each of the terms in the above expression. For the first term, note by Theorem 4.14 of [5], that the independence of the processes A_1^n and S_1^n for each n, and the fact that each of the corresponding sequences satisfies the LDP in $\mathbb{D}$ with a good rate function, imply that the sequence (A_1^n, S_1^n) satisfies the LDP on $\mathbb{D}^2$ in the product topology, with the good rate function formed by the sum. Below, we prove that $\boldsymbol{X}$ is continuous in the product topology. For the integrability condition of Varadhan's lemma, use the bound $\boldsymbol{X}(A_1^n, S_1^n) \leq A_1^n(T)$ and note that, in view of Assumption 2.1(i), there exists $\gamma_0 > 1$ such that

$$\limsup_n \frac{1}{n}\log\mathbb{E}[e^{\gamma_0 nc_iA_i^n(T)}] < \infty, \qquad i \in \mathcal{I}.$$

Thus, denoting $\mathbb{I}_1(a_1, s_1) = \mathbb{L}_1(a_1) + \mathbb{K}_1(s_1)$,

$$\begin{aligned}\limsup_n J^n(B^*) &= \limsup_n \frac{1}{n}\log\mathbb{E}\Big[e^{nc\cdot X^n(T)}\Big]\\ &\leq \limsup_n \frac{1}{n}\log\mathbb{E}\Big[e^{nc_1\boldsymbol{X}(A_1^n,S_1^n)}\Big] + \sum_{i\geq 2}\limsup_n\frac{1}{n}\log\mathbb{E}\Big[e^{nc_iA_i^n(T)}\Big]\\ &= \sup_{(a,s)\in\mathcal{AC}_0^2}[c_1\boldsymbol{X}(a,s) - \mathbb{I}_1(a,s)] + \sum_{i\geq 2}\sup_{a\in\mathcal{AC}_0}[c_ia(T) - \mathbb{L}_i(a)].\end{aligned}$$

Writing $a(T)$ as the integral of its derivative and using the integral expression (6) for $\mathbb{L}_i$ shows that the second term above is given by $T\sum_{i\geq 2} C_i^*$. As for the first term, write

$$\begin{aligned}\sup_{a,s}[c_1\boldsymbol{X}(a,s) - \mathbb{I}_1(a,s)] = \sup_{a,s}\sup_{\alpha,\beta,\gamma,\delta}\Big[&\int_{\beta-\alpha}^{\beta} c_1\dot{a}(t)dt - \int_{\delta-\gamma}^{\delta} c_1\dot{s}(t)dt\\ &- \int_0^T(\ell_1(\dot{a}(t)) + k_1(\dot{s}(t)))dt - \eta(\alpha-\gamma)^+\Big],\end{aligned}$$

where the supremum over $\alpha, \beta, \gamma, \delta$ is as in (16). Interchanging the order of the suprema, fix $\alpha, \beta, \gamma, \delta$, and note that the expression is maximized by selecting $\dot{a}(t) = \lambda_1$ for t outside the interval $[\beta - \alpha, \beta]$ (by which $\ell_1(\dot{a}(t)) = 0$), and $\dot{s}(t) = \mu_1$ outside $[\delta - \gamma, \delta]$ (so $k_1(\dot{s}(t)) = 0$). Moreover, the maximum over $\dot{a}(t)$ of $c_1\dot{a}(t) - \ell_1(\dot{a}(t))$ is given by C_1^*, whereas that of $-c_1\dot{s}(t) - k_1(\dot{s}(t))$ by $-C_1^{\#}$. Hence

$$\sup_{a,s}[c_1\boldsymbol{X}(a,s) - \mathbb{I}_1(a,s)] = \sup_{\alpha,\gamma\in[0,T]}\{\alpha C_1^* - \gamma C_1^{\#} - \eta(\alpha-\gamma)^+\}.$$

Now, $\alpha C_1^* - \gamma C_1^{\#} = \alpha(C_1^* - C_1^{\#}) + (\alpha - \gamma)C_1^{\#}$, and by assumption, $C_1^* > C_1^{\#}$. Hence

$$\sup_{a,s}[c_1\boldsymbol{X}(a,s) - \mathbb{I}_1(a,s)] \leq T(C_1^* - C_1^{\#}) + \sup_{\alpha,\gamma\in[0,T]}\{(\alpha-\gamma)C_1^{\#} - \eta(\alpha-\gamma)^+\}. \tag{18}$$

Assume, without loss of generality, that $\eta > C_1^{\#}$. Then the last term above is equal to zero. We obtain

$$\limsup_n J^n(B^*) \le T(C_1^* - C_1^{\#}) + T \sum_{i\ge 2} C_i^*,$$

which proves (14).

It remains to prove the continuity of $\boldsymbol{X}$ in $\mathbb{D}^2$ with the product topology. Let d denote the metric

$$d(x,y) = \inf_{\lambda\in\Lambda}(\| \lambda \|^\circ + \|x - y\circ\lambda\|), \qquad x, y \in \mathbb{D},$$

where Λ denotes the class of strictly increasing, continuous mappings of $[0,T]$ onto itself,

$$\| \lambda \|^\circ = \sup_{s\ne t}\Big| \log \frac{\lambda(t)-\lambda(s)}{t-s}\Big|,$$

and $\|x\| = \sup_t |x(t)|$. The required continuity will be established once we show that, for any $\varepsilon > 0$ there exists $\rho > 0$ such that

$$d(a_1,a_2)\vee d(s_1,s_2) \le \rho$$

implies

$$|\boldsymbol{X}(a_1,s_1) - \boldsymbol{X}(a_2,s_2)| \le \varepsilon. \tag{19}$$

To this end, let $\varepsilon > 0$ be given. Let $\rho > 0$ be so small that $8\rho + 8\eta T(e^{2\rho} - 1) \le \varepsilon$. Let a_1, a_2, s_1, s_2 be such that $d(a_1,a_2)\vee d(s_1,s_2) \le \rho$. Fix λ_a such that $\|\lambda_a\|^\circ + \|a_1 - a_2\circ\lambda_a\| < 2\rho$, and λ_s such that a similar statement holds for s_1, s_2. We have

$$\begin{aligned}\boldsymbol{X}(a_1,s_1) - \boldsymbol{X}(a_2,s_2) = {}&\sup_{\alpha,\beta,\gamma,\delta}\big[a_1(\beta) - a_1(\beta-\alpha) - s_1(\delta) + s_1(\delta-\gamma) - \eta(\alpha-\gamma)^+\big]\\ &- \sup_{\bar\alpha,\bar\beta,\bar\gamma,\bar\delta}\big[a_2(\bar\beta) - a_2(\bar\beta-\bar\alpha) - s_2(\bar\delta) + s_2(\bar\delta-\bar\gamma) - \eta(\bar\alpha-\bar\gamma)^+\big],\end{aligned}$$

where the supremum is over $(\alpha,\beta,\gamma,\delta)$ and $(\bar\alpha,\bar\beta,\bar\gamma,\bar\delta)$ satisfying (16). Given $(\alpha,\beta,\gamma,\delta)$, select $(\bar\alpha,\bar\beta,\bar\gamma,\bar\delta)$ such that

$$\bar\beta = \lambda_a(\beta), \quad \bar\beta - \bar\alpha = \lambda_a(\beta-\alpha), \quad \bar\delta = \lambda_s(\delta), \quad \bar\delta - \bar\gamma = \lambda_s(\delta-\gamma). \tag{20}$$

Note that $0 < \bar\alpha \le \bar\beta \le T$ and $0 < \bar\gamma \le \bar\delta \le T$. Therefore

$$\boldsymbol{X}(a_1,s_1) - \boldsymbol{X}(a_2,s_2) \le \sup_{\alpha,\beta,\gamma,\delta} Y(\alpha,\beta,\gamma,\delta), \tag{21}$$

where

$$\begin{aligned}Y(\alpha,\beta,\gamma,\delta) = {}& a_1(\beta) - a_2(\lambda_a(\beta)) + a_2(\lambda_a(\beta-\alpha)) - a_1(\beta-\alpha)\\ &+ s_2(\lambda_s(\delta)) - s_1(\delta) + s_1(\delta-\gamma) - s_2(\lambda_s(\delta-\gamma))\\ &- \eta(\alpha-\gamma)^+ + \eta(\bar\alpha-\bar\gamma)^+.\end{aligned}$$

Note that

$$\begin{aligned}Y(\alpha,\beta,\gamma,\delta) &\le 8\rho - \eta(\alpha-\gamma)^+ + \eta(\bar\alpha-\bar\gamma)^+\\ &\le 8\rho + \eta|\alpha-\bar\alpha| + \eta|\gamma-\bar\gamma|.\end{aligned}$$

Now, the bound $\|\lambda_a\| \le 2\rho$ implies that for every $t \in [0,T]$, $|t-\lambda_a(t)| \le \hat\rho := T(e^{2\rho}-1)$. A similar statement holds for λ_s. Hence by (20), $|\alpha-\bar\alpha| \le 4\hat\rho$ and $|\gamma-\bar\gamma| \le 4\hat\rho$. Thus

$$\boldsymbol{X}(a_1,s_1)-\boldsymbol{X}(a_2,s_2) \le 8\rho + 8\eta\hat\rho \le \varepsilon.$$

Interchanging the roles of (a_1,s_1) and (a_2,s_2) gives a similar bound, and (19) follows.

4.2 A lower bound on the risk-sensitive value

We now show (15), by considering a sequence of general admissible controls, providing a lower bound on their performance. Recall the model equations

$$X_i^n(t) = A_i^n(t) - S_i^n(T_i^n(t)), \qquad T_i^n(t) = \int_0^t B_n^i(s)ds, \tag{22}$$

where B^n takes values in $\mathbb{U}$, and, for every t,

$$X_i^n(t) \ge 0, \qquad X_i^n(t) = 0 \text{ implies } B_i^n(t) = 0. \tag{23}$$

We introduce a model that is similar, but does not adhere to constraints of the form (23). Namely, we consider

$$Y_i^n(t) = A_i^n(t) - S_i^n(P_i^n(t)), \qquad P_i^n(t) = \int_0^t Q_i^n(s)ds, \tag{24}$$

where Q^n is $\mathbb{U}$-valued. Note, in particular, that Y_i^n is not constrained to remain nonnegative. Denote the collection of all processes taking values in $\mathbb{U}$ by $\mathcal{Q}$. Since for a given n and $B^n \in \mathcal{B}$ there exists a $\mathbb{U}$-valued process Q^n (specifically, $Q^n = B^n$) such that $Y^n = X^n$, clearly

$$V^n = \inf_{B^n\in\mathcal{B}} J^n(B^n) = \inf_{B^n\in\mathcal{B}} \frac{1}{n}\log\mathbb{E}\Big[e^{nc\cdot X^n(T)}\Big] \ge \inf_{Q^n\in\mathcal{Q}} \frac{1}{n}\log\mathbb{E}\Big[e^{nc\cdot Y^n(T)}\Big].$$

Noting that $Y_i^n(T) = A_i^n(T) - S_i^n(P_i^n(T))$, we can write

$$V^n \ge V_*^n := \inf_{u\in\mathcal{U}} \frac{1}{n}\log\mathbb{E}\Big[e^{n\sum_i c_i(A_i^n(T)-S_i^n(u_iT))}\Big], \tag{25}$$

where $u=(u_i)_{i\in\mathcal{I}}$ and $\mathcal{U}$ denotes the set of all $\mathbb{U}$-valued random variables. We proceed by deriving a lower bound on the right hand side of (25).

For each n, let u^n be a $\mathbb{U}$-valued random variable for which

$$V_*^n + \frac{1}{n} \ge \frac{1}{n}\log R^n \qquad \text{where} \qquad R^n = \mathbb{E}\Big[e^{n\sum_i c_i(A_i^n(T)-S_i^n(u_i^nT))}\Big].$$

Then

$$\underline{V} \ge \liminf_n \frac{1}{n}\log R^n.$$

Fix $\varepsilon > 0$. Denote by $B(v,r) \subset \mathbb{R}^I$ the open ball of radius $r>0$ around $v\in\mathbb{R}^I$. Fix a finite collection of balls $B_k := B(v_k,\varepsilon)$, $k \in \mathcal{K}_\varepsilon := \{1,\dots,K_\varepsilon\}$, with $v_k = (v_{k,i})_{i\in\mathcal{I}} \in \mathbb{U}$, such that $\cup_k B_k \supset \mathbb{U}$. Then for every n and k we have

$$R^n \ge \mathbb{E}\Big[1_{\{u^n\in B_k\}}e^{n\sum_i c_i(A_i^n(T)-S_i^n(t_{k,i}))}\Big], \tag{26}$$

where $t_{k,i} = ((v_{k,i} + \varepsilon)T) \wedge T$. Moreover, if k_n denotes a variable k which maximizes the right hand side of (26) then

$$R^n \geq \frac{1}{K_\varepsilon} \mathbb{E}\Big[e^{n \sum_i c_i (A_i^n(T) - S_i^n(t_{k_n,i}))} \Big].$$

As a result,

$$R^n \geq \min_{k \in \mathcal{K}_\varepsilon} \frac{1}{K_\varepsilon} \mathbb{E}\Big[e^{n \sum_i c_i (A_i^n(T) - S_i^n(t_{k,i}))} \Big].$$

Hence

$$\underline{V} \geq \min_{k \in \mathcal{K}_\varepsilon} \liminf_n \frac{1}{n} \log \mathbb{E}\Big[e^{n \sum_i c_i (A_i^n(T) - S_i^n(t_{k,i}))} \Big].$$

Using the independence of the $2I$ processes A_i^n, S_i^n, and Varadhan's lemma,

$$\begin{aligned} \underline{V} &\geq \min_{k \in \mathcal{K}_\varepsilon} \sup_{(a,s) \in \mathcal{AC}_0^{2I}} \sum_i \{ c_i [a_i(T) - s_i(t_{k,i})] - \mathbb{L}_i(a_i) - \mathbb{K}_i(s_i) \} \\ &\geq \inf_{u \in \mathbb{U}} \sup_{(a,s) \in \mathcal{AC}_0^{2I}} \sum_i \{ c_i [a_i(T) - s_i((u_i + \varepsilon)T \wedge T)] - \mathbb{L}_i(a_i) - \mathbb{K}_i(s_i) \}. \end{aligned} \tag{27}$$

Fix $u \in \mathbb{U}$ and i. The problem of maximizing $c_i a_i(T) - \int_0^T \ell_i(\dot{a}_i(t))dt$ over $a_i \in \mathcal{AC}_0$ is solved by writing this expression as $\int_0^T (c_i \dot{a}_i(t) - \ell_i(\dot{a}_i(t)))dt$ and maximizing the integrand. A calculation shows that the maximum is given by TC_i^*. Maximizing $-c_i s_i((u_i + \varepsilon)T \wedge T) - \int_0^T k_i(\dot{s}_i(t))dt$ over $s_i \in \mathcal{AC}_0$ is attained by letting $\dot{s}_i(t) = \mu_i$ for $t \in [(u_i + \varepsilon)T \wedge T, T]$, and a calculation shows that the maximum is then given by $-\{(u_i + \varepsilon)T \wedge T\} C_i^{\#}$. Thus by (27),

$$\begin{aligned} \underline{V} &\geq \inf_{u \in \mathbb{U}} \sum_i [TC_i^* - \{(u_i + \varepsilon)T \wedge T\} C_i^{\#}] \\ &\geq \inf_{u \in \mathbb{U}} \sum_i [TC_i^* - u_i T C_i^{\#}] - c_0 \varepsilon, \end{aligned} \tag{28}$$

where $c_0 = T \sum_i C_i^{\#}$. Taking $\varepsilon \to 0$, and using $C_1^{\#} \geq C_i^{\#}$ for all i, by which the infimum is attained with $u_1 = 1$, gives (15).

Remark 4.3. *The assumption $C_i^* > C_i^{\#}$, $i \in \mathcal{I}$ is used in two places in the proof of the result. First, in Section 4.1, it is used in the argument leading to* (18)*. Then, in Section 4.2, it is used to argue that the minimization of the expression in* (28) *is attained by setting $u_1 = 1$. The fact that $u_1 = 1$ is selected indicates that a policy that asymptotically achieves the lower bound must act to provide (nearly) all effort to class 1. The priority policy studied in the upper bound also gives highest priority to class 1. Thus we can interpret the role that the aforementioned assumption plays as follows. It dictates that the contribution of the cost associated with one dominating class to the overall cost is large compared to the other classes, to the degree that an AO policy must devote all effort to this class. An attempt to go beyond this case must deal with more general target distribution of effort. However, the techniques we have demonstrated in this paper break down, as the resulting upper and lower bounds that they give rise to no longer match each other.*

References

[1] R. Atar, A. Goswami, and A. Shwartz. Risk-sensitive control for the parallel server model. *SIAM J. Control Optim.*, 51(6):4363–4386, 2013.

[2] R. Atar, A. Goswami, and A. Shwartz. Asymptotic optimality of a fixed priority rule for a queueing problem in the large deviation regime. *Electron. Commun. Probab.*, 19(11):1–13, 2014.

[3] A. Biswas. Private communication.

[4] A. Dembo and O. Zeitouni. *Large deviations techniques and applications*, volume 38 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 1998. ISBN 0-387-98406-2. xvi+396 pp.

[5] A. Ganesh, N. O'Connell, and D. Wischik. *Big Queues*, volume 1838 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2004. ISBN 3-540-20912-3. xii+254 pp.

[6] A. A. Puhalskii and W. Whitt. Functional large deviation principles for first-passage-time processes. *Ann. Appl. Probab.*, 7(2):362–381, 1997.

[7] A. L. Stolyar and K. Ramanan. Largest weighted delay first scheduling: large deviations and optimality. *Ann. Appl. Probab.*, 11(1):1–48, 2001.