Dynamic fluid-based scheduling in a multi-class abandonment queue

M. Larrañaga [b,e,*], U. Ayesta [b,c,d,e], I.M. Verloop [a,e]

[a] CNRS, IRIT, Toulouse, France
[b] CNRS, LAAS, 7 avenue du colonel Roche, F-31400 Toulouse, France
[c] IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Spain
[d] UPV/EHU, University of the Basque Country, 20018 Donostia, Spain
[e] Univ. de Toulouse, INP, LAAS, F-31400 Toulouse, France

A B S T R A C T

We investigate how to share a common resource among multiple classes of customers in the presence of abandonments. We consider two different models: (1) customers can abandon both while waiting in the queue and while being served, (2) only customers that are in the queue can abandon. Given the complexity of the stochastic optimization problem we propose a fluid model as a deterministic approximation. For the overload case we directly obtain that the $\tilde{c}\mu/\theta$ rule is optimal. For the underload case we use Pontryagin's Maximum Principle to obtain the optimal solution for two classes of customers; there exists a switching curve that splits the two-dimensional state-space into two regions such that when the number of customers in both classes is sufficiently small the optimal policy follows the $\tilde{c}\mu$-rule and when the number of customers is sufficiently large the optimal policy follows the $\tilde{c}\mu/\theta$-rule. The same structure is observed in the optimal policy of the stochastic model for an arbitrary number of classes. Based on this we develop a heuristic and by numerical experiments we evaluate its performance and compare it to several index policies. We observe that the suboptimality gap of our solution is small.

1. Introduction

Abandonment or reneging takes place when customers, unsatisfied of their long waiting time, decide to voluntarily leave the system. It has a huge impact in various real life applications such as the Internet or call centers, where customers may abandon while waiting in the queue, or even while being served. Abandonment is a very undesirable phenomena, both from the customers' and system's point of view, and it can have a big economical impact. It is thus not surprising that it has attracted considerable interest from the research community, with a surge in recent years.

An important line of research aims at characterizing the performance and impact of abandonments in systems, we refer to [1–6] for single-server models and [7–9] for papers dealing with the multi-class case. We also refer to [10] for a recent survey on abandonments in a many-server system. More related to our present work are the papers that deal with optimal scheduling or control aspects of multi-class queueing systems in the presence of abandonments, see for instance [11–18].

* Corresponding author at: CNRS, LAAS, 7 avenue du colonel Roche, F-31400 Toulouse, France.
E-mail addresses: maialen.larranaga@laas.fr (M. Larrañaga), urtzi@laas.fr (U. Ayesta), verloop@irit.fr (I.M. Verloop).

As the performance criteria the most common objectives are maximizing the service completion reward or minimizing a combination of the waiting time and abandonment penalty. In the case of two classes of customers and one server, the authors of [16] assume exponential distributed service requirements and impatience times and show that, under an additional condition on the ordering of the abandonment rates, an index policy is optimal. Since the abandonment rate is not bounded, it is not possible to uniformize the system. In order to prove the result, the authors use a continuous-time formulation of a Markov decision process (instead of the discrete-time equivalent) in addition to a truncation argument. In the case of no arrivals and non-preemptive service, the authors of [14] provide partial characterizations of the optimal policy and show that an optimal policy is typically state dependent. It is worthwhile to mention that [14] is inspired by a patient triage problem which illustrates that abandonments are as well an important issue in other areas than information technology. As far as the authors are aware, the above two settings are the only ones for which structural optimality results have been obtained. State-dependent heuristics for the multi-class queue are proposed in [14] for two classes and no arrivals and in [11] for an arbitrary number of classes including new arrivals.

In general, determining the exact optimal policy has so far proved analytically infeasible. Hence, researchers have focused on obtaining approximations of the optimal control. For example, in [18] the authors study a multi-class abandonment queue without arrivals and use the Lagrangian relaxation method [19] to construct an index policy, which is optimal for a relaxed optimization problem. Another approach is to study the system in the Halfin–Whitt heavy-traffic regime. That is, the total arrival rate and the number of servers both become large in such a way that the traffic intensity approaches one. For the abandonment queue this was first studied in [20]. This scaling gives rise to a diffusion control problem, for which the optimal controls are investigated in [12,13] and shown to be state dependent. In an overload setting the abandonment queue has been studied under a fluid scaling in [15,17], where the authors scale the number of servers and the arrival rate and show that the $\tilde{c}\mu/\theta$ rule (i.e., the policy where strict priority is given according to the indices $\tilde{c}\mu/\theta$) is asymptotically fluid optimal (here $\tilde{c}$ is the holding plus abandonment cost, θ is the abandonment rate and μ the service rate). The overload assumption is crucial in their analysis, since under this assumption the trajectories of the fluid model converge to a strictly positive state which completely characterizes the performance under the average performance criteria. The $\tilde{c}\mu/\theta$-rule emerges naturally as the policy that optimizes the performance associated to this absorbing state. Without abandonments, the $\tilde{c}\mu$-rule, i.e., strict priority is given according to the indices $\tilde{c}\mu$, is optimal in a multi-class single server queue for average reward and discounted cost criteria, in the preemptive and non-preemptive cases, see for example [21]. The $\tilde{c}\mu/\theta$ and the $\tilde{c}\mu$ index rules will play an important role throughout this paper.

In this paper, we investigate the fundamental question of how to share *one* common resource among multiple classes of customers in the presence of abandonments. We consider two different stochastic models: (1) customers can abandon when they are waiting in the queue and also while they are being served, (2) customers that are in the queue can abandon but customers in service cannot. Given the complexity of the problem it is not possible to solve the stochastic optimization problem. We thus propose a fluid model with non-linear dynamics, which can be interpreted as a deterministic approximation of the stochastic problem. We consider both the underload and the overload case, the former being considerably more difficult to solve. In an overload setting, we determine the optimal equilibrium point and show that under the $\tilde{c}\mu/\theta$ rule the dynamics converge to this point, which in fact is non-zero. In the underload case the fluid model will empty in finite time, hence we will seek for the optimal trajectory that minimizes the cost of draining the fluid. The latter makes the analysis considerably harder than in the overload case. Using Pontryagin's Maximum Principle [22] we solve completely the case with two classes of customers. The optimal solution has a remarkable structure, there exists a switching curve that splits the two-dimensional state-space into two regions such that: when the number of customers is sufficiently small the optimal policy follows the $\tilde{c}\mu$-rule and when the number of customers is sufficiently large the optimal policy follows the $\tilde{c}\mu/\theta$-rule. Solving the optimal *stochastic* control problem numerically we observe this same behavior where the shape of the switching curve is very well approximated by the one found in the fluid model. In fact, the combination of the $\tilde{c}\mu$ rule and the $\tilde{c}\mu/\theta$ rule is also observed numerically in the optimal stochastic control for more than two classes. We use this insight to propose a heuristic for the stochastic model (for an arbitrary number of classes). At last, by numerical experiments we evaluate the performance of the fluid-based heuristic and several index policies and observe that the suboptimality gap of our solution is small. We emphasize here that our heuristic works well across all loads, while the index policies $\tilde{c}\mu$ and $\tilde{c}\mu/\theta$, although being rather easy to implement, achieve only good performance in either the underload or the overload setting.

The approach of using the fluid model to find an approximation for the stochastic model finds its roots in the pioneering works by Avram et al. [23] and Weiss [24]. It is remarkable that in some cases the optimal control for the fluid model coincides with the optimal solution for the stochastic problem. See for example [23] where this is shown for the $c\mu$-rule in a multi-class single-server queue and [25] where this is shown for Klimov's rule in a multi-class queue with feedback. For other cases, researchers have aimed at establishing that the fluid control is asymptotically optimal, that is, the fluid-based control is optimal for the stochastic optimization problem after a suitable scaling, see for example [26–30]. We conclude by mentioning that the fluid approach owes its popularity to the groundbreaking result stating that if the fluid model drains in finite time, the stochastic process is stable, see [31,32].

The remainder of the paper is organized as follows. In Section 2 we present the stochastic model with abandonments and the optimization problem. In Section 3 we introduce the related fluid model and solve its fluid control problem for the underload case (for two classes of customers) and for the overload case. In Section 4 we develop a heuristic for the stochastic model for an arbitrary number of classes and in Section 5 we numerically compare the performance of the fluid-based heuristic with that of the optimal policy and several index policies proposed in the literature.

2. Model description

We consider a multi-class single-server queue with K classes of customers. Class-k customers arrive according to a Poisson process with rate λ_k and have an exponentially distributed service requirement with mean $1/\mu_k$. A class-k customer can abandon the system after an exponentially distributed amount of time with mean $1/\theta_k$. We define $\rho_k = \lambda_k/\mu_k$ as the traffic load of class k and $\rho = \sum_k \rho_k$ as the total load. We assume that the server has capacity 1 and can serve at most one customer at a time, where the service can be preemptive. At each moment in time, a policy π decides which class is served. Because of the Markov property we can focus on policies that base decisions on the current number of customers present in the various classes. For a given policy π, the control variable $(S_1^\pi(t), \ldots, S_K^\pi(t))$ denotes the class of the customer that is in service at time t, i.e., if at time t class k is in service, then $S_k^\pi(t) = 1$ and $S_l^\pi(t) = 0$ for $l \neq k$. Hence, it satisfies $S_k^\pi(t) \in \{0, 1\}$ and $\sum_{k=1}^K S_k^\pi(t) \leq 1$.

We are interested in two different models, depending on whether or not a customer in service becomes impatient and hence can abandon:

- Stochastic Model 1 (SM1): customers can abandon both while waiting in the queue and while being served, see Fig. 1(a).
- Stochastic Model 2 (SM2): customers can abandon only while waiting in the queue, see Fig. 1(b).

Both models have been studied in the literature, e.g., in [16] the model SM1 is studied, while the authors of [15,17,18] consider SM2.

For a given policy π, let $N_k^\pi(t)$ denote either the number of class-k customers in the system (SM1) or the number of class-k customers in the queue (SM2). Let c_k denote the holding cost per unit time for class-k customers. Let d_k denote the cost for each class-k customer that abandons. Our objective is to minimize the average cost, that is,

$$\min_\pi \limsup_{T\to\infty} \sum_{k=1}^K \frac{1}{T}\mathbb{E}\left(\int_0^T c_k N_k^\pi(t)\mathrm{d}t + d_k R_k^\pi(T)\right) = \min_\pi \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\int_0^T \sum_{k=1}^K \tilde{c}_k N_k^\pi(t)\mathrm{d}t\right),$$

where $R_k^\pi(T)$ denotes the number of class-k customers that abandoned in the interval $[0, T]$, $\tilde{c}_k := c_k + d_k\theta_k,\ k = 1, \ldots, K$, and we used that $\mathbb{E}(R_k^\pi(T)) = \theta_k\mathbb{E}(\int_0^T N_k^\pi(t)\mathrm{d}t)$. We note that for model SM2 only waiting customers can abandon and can hence contribute to the abandonment cost. In addition, implicitly we assumed that for SM2 only waiting customers contribute to the holding cost.[1] For model SM1, all customers will have a contribution to the abandonment cost (a customer in service can abandon) and we have assumed that all customers contribute to the holding cost.

The above described stochastic control problems have proved to be very difficult to solve. In [16] optimal dynamic scheduling is studied for the model SM1 for two classes of customers ($K = 2$) with $\mu_1 = \mu_2 = 1$. In case $\tilde{c}_1 \geq \tilde{c}_2$ and $\theta_1 \leq \theta_2$, the authors show that it is optimal to give strict priority to class 1, see [16, Theorem 3.5]. It is intuitively clear that giving priority to class 1 is the optimal thing to do, since serving class 1 myopically minimizes the (holding and abandonment) cost and in addition it is advantageous to keep the maximum number of class-2 customers in the system (without idling), since they have the highest abandonment rate. Outside this parameter setting, an optimal policy is expected to be state dependent, and as far as the authors are aware, no (structural) results exist for this stochastic optimal control problem. Therefore, in Section 3, in order to obtain further insight, we propose to solve a related fluid control model.

3. Fluid control model

In this section the stochastic models (SM1, SM2) presented in Section 2 are approximated by the deterministic fluid model, where only the mean dynamics are taken into account. That is, let $n_k(t)$ be the amount of class-k fluid and $s_k(t)$ the control parameter. Then the fluid dynamics is described by the following set of differential equations:

$$\frac{\mathrm{d}n_k(t)}{\mathrm{d}t} = \lambda_k - \mu_k s_k(t) - \theta_k n_k(t), \quad \forall k \in \{1, \ldots, K\}, s_k(t) \in \mathcal{S},\ n_k(t) \geq 0,\ \forall k \in \{1, \ldots, K\},\ \forall t,$$

with

$$\mathcal{S} := \left\{ s = (s_1, \ldots, s_K) \text{ s.t. } \sum_{k=1}^K s_k \leq 1,\ s_k \geq 0\ \forall k \in \{1, \ldots, K\} \right\}.$$

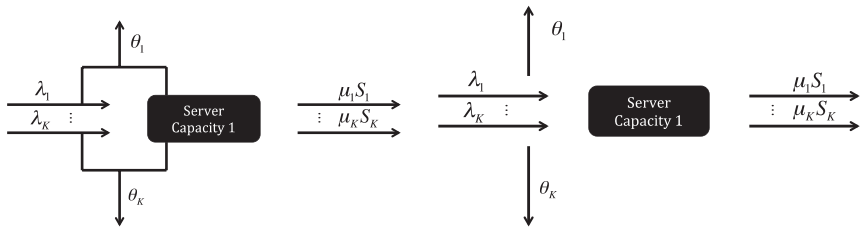
For the fluid analysis we will make a distinction between two different scenarios: (1) $\rho < 1$, which we refer to as the underload and (2) $\rho > 1$, which we refer to as the overload setting. Note that in the case $\rho < 1$, any non-idling control (i.e., $\sum_{k=1}^K s_k(t) = 1$ if $\sum_{k=1}^K n_k(t) > 0$) converges to the equilibrium point $(0, \ldots, 0)$.[2] Hence, when $\rho < 1$ we aim at minimizing the total cost until reaching the equilibrium point $(0, \ldots, 0)$. This can be written as

$$\min_{s(t)\in\mathcal{S}} \int_0^\infty \sum_{k=1}^K \tilde{c}_k n_k(t)\mathrm{d}t.$$

---

[1] For the model SM2, the latter was also assumed in [15,17], while [18] assumed customers in service contribute to the holding cost as well.

[2] Consider $w(t) := \sum_{k=1}^K n_k(t)/\mu_k$. Then $\frac{\mathrm{d}w(t)}{\mathrm{d}t} = \rho - \sum_{k=1}^K s_k(t) - \sum_{k=1}^K \frac{\theta_k}{\mu_k} n_k(t) < \rho - 1 < 0$, hence $w(t)$ converges to zero.



(a) Stochastic Model 1 (SM1). (b) Stochastic Model 2 (SM2).

Fig. 1. Multi-class single-server queue with abandonments.

When $\rho > 1$, an equilibrium point will necessarily be different than $(0, \ldots, 0)$. Hence, for $\rho > 1$ our objective is to minimize the average cost, i.e.,

$$\min_{s(t)\in\mathcal{S}} \lim_{T\to\infty} \frac{1}{T} \int_0^T \sum_{k=1}^{K} \tilde{c}_k n_k(t) \mathrm{d}t.$$

Throughout the study we refer to this optimal fluid control problem as Problem P.

3.1. Optimal policy in underload for two classes of customers

In this section we assume $\rho < 1$ and solve the fluid control model. We focus on the case of two classes of customers, whose solution is already rather involved. However, it gives us intuition on the structure of the optimal policy for an arbitrary number of classes.

We will see that an optimal policy can be of two possible shapes: either a switching curve emerges, *i.e.*, we prioritize one class above the switching curve and the other class below the switching curve, or one of the two classes is prioritized. This gives us four different type of strategies. As we show in the following proposition, the optimal strategy is fully characterized by the ordering of $\tilde{c}_1\mu_1$ and $\tilde{c}_2\mu_2$ and of $\tilde{c}_1\mu_1/\theta_1$ and $\tilde{c}_2\mu_2/\theta_2$.

Proposition 1. *Assume $K = 2$ and let λ_k, μ_k, θ_k, c_k and d_k be given for $k \in \{1, 2\}$. Assume $\tilde{c}_2\mu_2/\theta_2 \geq \tilde{c}_1\mu_1/\theta_1$. If $\rho < 1$, then, an optimal solution $s^*(\cdot)$ for Problem P under the total cost criteria is:*

- *if $\tilde{c}_2\mu_2 \leq \tilde{c}_1\mu_1$, then*
  - *$s^* = (0, 1)$ when $n_2 > h(n_1)$,*
  - *$s^* = (1, 0)$ when $n_2 \leq h(n_1)$ and $n_1 > 0$,*
  - *$s^* = (\rho_1, 1 - \rho_1)$ when $n_2 \leq h(0)$ and $n_1 = 0$,*

  where the switching curve $h(\cdot)$ is given by

$$h(n_1) := \frac{a_1 n_1 + a_2 + (a_3 n_1 - a_2)\left(\frac{\theta_1 n_1 + \mu_1 - \lambda_1}{\mu_1 - \lambda_1}\right)^{\frac{\theta_2}{\theta_1}}}{a_4 n_1} + \frac{\lambda_2}{\theta_2}, \tag{3.1}$$

  with

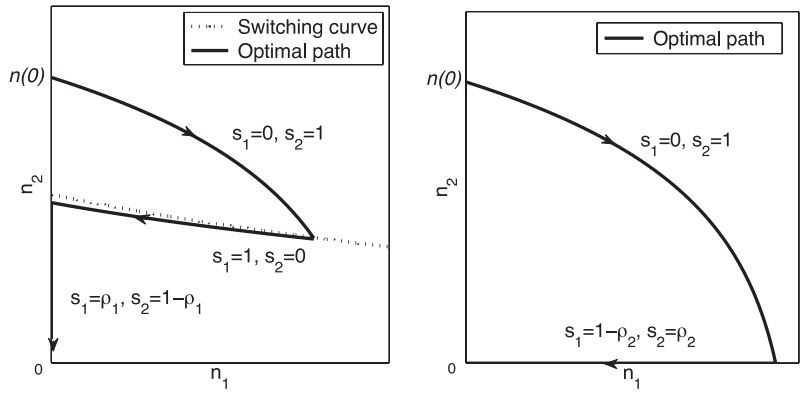
$$a_1 = \tilde{c}_2 \frac{\mu_2}{\theta_2}(1 - \rho); \qquad a_2 = a_1 \frac{\mu_1}{\theta_1}(1 - \rho_1);$$
$$a_3 = -\left(\tilde{c}_2 \frac{\mu_2}{\theta_2} - \tilde{c}_1 \frac{\mu_1}{\theta_1}\right)(1 - \rho_1), \quad \text{and} \quad a_4 = \left(\tilde{c}_2 \frac{\mu_2}{\theta_2} - \tilde{c}_1 \frac{\mu_1}{\theta_1}\right)\frac{\theta_2}{\mu_2}.$$

  That is, serve class 2 until the switching curve $h(\cdot)$ is reached, then serve class 1 until $n_1 = 0$. From that moment on, keep $n_1 = 0$ and give the rest of the service to class 2, see Fig. 2(a).
- *If $\tilde{c}_2\mu_2 \geq \tilde{c}_1\mu_1$, then*
  - *$s^* = (0, 1)$ when $n_2 > 0$,*
  - *$s^* = (1 - \rho_2, \rho_2)$ when $n_2 = 0$.*

  That is, serve class 2 until $n_2 = 0$. From that moment on, keep $n_2 = 0$ and give the rest of the service to class 1, see Fig. 2(b).

The solution in the case where $\tilde{c}_2\mu_2/\theta_2 \leq \tilde{c}_1\mu_1/\theta_1$ is equivalent with the indices swapped.

Remark 1 (*Arbitrary Number of Classes*)**.** Given the complexity to find an optimal solution for the fluid control model in underload when $K = 2$, we did not aim at obtaining an analytical solution for an arbitrary number of classes K. Instead, in Section 4 we develop a heuristic using the insights obtained for the case $K = 2$.



(a) Optimal path when $\tilde{c}_2\mu_2 \leq \tilde{c}_1\mu_1$. (b) Optimal path when $\tilde{c}_2\mu_2 \geq \tilde{c}_1\mu_1$.

Fig. 2. Optimal strategy assuming $\frac{\tilde{c}_2\mu_2}{\theta_2} \geq \frac{\tilde{c}_1\mu_1}{\theta_1}$, and the optimal path.

Remark 2. Observe that the switching curve $h(\cdot)$ defined in Proposition 1 is decreasing in n_1 and hence it will emerge in the first quadrant if and only if

$$h(0) = (1-\rho)\frac{\mu_2}{\theta_1\theta_2}\left(\frac{\tilde{c}_1\mu_1 - \tilde{c}_2\mu_2}{\frac{\tilde{c}_2\mu_2}{\theta_2} - \frac{\tilde{c}_1\mu_1}{\theta_1}}\right) \geq 0.$$

Since we are in the underload case this is equivalent to the condition

$$\left(\tilde{c}_1\mu_1 - \tilde{c}_2\mu_2\right) / \left(\frac{\tilde{c}_2\mu_2}{\theta_2} - \frac{\tilde{c}_1\mu_1}{\theta_1}\right) \geq 0.$$

Before proving the above proposition we first provide intuition for the structure of the optimal policy, which is characterized by a very simple rule based on the comparison of the indices $\tilde{c}\mu$ and $\tilde{c}\mu/\theta$. When the amount of fluid is small enough, the $\tilde{c}\mu$ rule is optimal. This can be explained as follows. Note that the derivative of the cost is given by $\sum_{k=1}^{K}\tilde{c}_k\frac{\mathrm{d}n_k(t)}{\mathrm{d}t} = \sum_{k=1}^{K}\tilde{c}_k\left(\lambda_k - \mu_k s_k(t) - \theta_k n_k(t)\right)$. The $\tilde{c}\mu$-rule myopically minimizes the derivative and is hence optimal in the short run. Close to the origin this is exactly what the optimal control prescribes. However, in the long term, one cannot neglect the effect of abandonments. For example, if $\tilde{c}_1\mu_1 > \tilde{c}_2\mu_2$, but $\theta_1 \gg \theta_2$, then the myopic rule would prioritize class 1. However, this minimizes $n_1(t)$, which has a negative impact on the derivative of the cost (cf. the term $\theta_1 n_1(t)$). Hence, in the long run it might be good to keep the amount of class-1 fluid high, since class 1 has a high abandonment rate. In Proposition 1 we showed that in a state far from the origin, the index that appropriately combines the above described effects is the $\tilde{c}\mu/\theta$ index. We will show in Proposition 2 that the $\tilde{c}\mu/\theta$ rule is in fact optimal when $\rho > 1$, i.e., in the overload setting.

The switching curve $h(\cdot)$, as defined in Proposition 1, describes the states in which it is optimal to switch from the $\tilde{c}_k\mu_k/\theta_k$ rule to the $\tilde{c}_k\mu$ rule. We can learn the following from the formula for $h(\cdot)$:

- as we can see from Remark 2, the ratio between $\tilde{c}_1\mu_1 - \tilde{c}_2\mu_2$ and $\frac{\tilde{c}_2\mu_2}{\theta_2} - \frac{\tilde{c}_1\mu_1}{\theta_1}$ determines $h(0)$, and hence the height of the switching curve. From this we observe that as the difference in the values for the $\tilde{c}\mu$ index grows large (small) relative to that of the $\tilde{c}\mu/\theta$ index, the height of the switching curve grows (goes to zero) and hence the optimal fluid control gets closer to the $\tilde{c}\mu$ rule ($\tilde{c}\mu/\theta$ rule).
- as the traffic load approaches one, i.e., $\rho \uparrow 1$, the switching curve $h(\cdot)$ converges to $\overline{h}(\cdot)$ with $\overline{h}(0) = 0$ and $\overline{h}(n_1) < 0$ for $n_1 > 0$. Hence, the $\tilde{c}\mu/\theta$ rule is optimal for the fluid model as $\rho \uparrow 1$. As we will see in Section 3.2, the $\tilde{c}\mu/\theta$ rule is optimal in the overload setting ($\rho > 1$) as well, showing continuity in the optimal solution.

Remark 3 (*Multi-Class Queue with Deadlines*)**.** In the case $c_k = 0,\ k = 1, \ldots, K$, the model becomes a multi-class queue with deadlines: customers need to be served before a deadline that is exponentially distributed with parameter θ_k and in the case they do not receive service before their deadline they abandon the queue giving a cost d_k. In this particular case the $\tilde{c}\mu$ rule reduces to $d\mu\theta$ rule and the $\tilde{c}\mu/\theta$ rule reduces to the $d\mu$ rule.

We now proceed to the proof of Proposition 1. The following two lemmas are necessary in order to prove Proposition 1. Their proofs are presented in the Appendix. The first lemma states that the index $\tilde{c}_k\mu_k$ determines the optimal action when the amount of fluid in both class 1 and class 2 is small.

Lemma 1. *Let $K = 2$ and let $n(0) = (\varepsilon, \varepsilon)$ with $\varepsilon > 0$ small enough. If $\rho < 1$ and*

$$\tilde{c}_1\mu_1 \geq \tilde{c}_2\mu_2 \quad (\textit{resp. } \tilde{c}_1\mu_1 \leq \tilde{c}_2\mu_2),$$

then it is optimal to give priority to class 1 (resp. class 2) until the origin is reached.

Assuming that it is optimal to serve class 2 in the initial point, the next lemma describes the two possible strategies.

Lemma 2. *Let $K = 2$ and the initial conditions $n(0)$ be given. Assume that it is optimal to prioritize class 2 at time 0. Then the optimal solution of Problem P will be one out of the following two strategies:*

- *if $\left(\tilde{c}_1\mu_1 - \tilde{c}_2\mu_2\right) / \left(\frac{\tilde{c}_2\mu_2}{\theta_2} - \frac{\tilde{c}_1\mu_1}{\theta_1}\right) \geq 0$, then the following control is optimal:*
  - *$s^* = (0, 1)$ if $n_2 > h(n_1)$,*
  - *$s^* = (1, 0)$ if $n_2 \leq h(n_1)$ and $n_1 > 0$,*
  - *$s^* = (\rho_1, 1 - \rho_1)$ if $n_2 \leq h(0)$ and $n_1 = 0$,*

  with $h(\cdot)$ the switching curve as defined in Proposition 1.
- *if $\left(\tilde{c}_1\mu_1 - \tilde{c}_2\mu_2\right) / \left(\frac{\tilde{c}_2\mu_2}{\theta_2} - \frac{\tilde{c}_1\mu_1}{\theta_1}\right) \leq 0$, then the following control is optimal:*
  - *$s^* = (0, 1)$ if $n_2 > 0$,*
  - *$s^* = (1 - \rho_2, \rho_2)$ if $n_2 = 0$.*

The proof of Proposition 1 now follows easily.

Proof of Proposition 1. We first assume that $\tilde{c}_2\mu_2/\theta_2 \geq \tilde{c}_1\mu_1/\theta_1$ and $\tilde{c}_2\mu_2 \geq \tilde{c}_1\mu_1$. Now assume that there is a state n such that it is optimal to serve class 1. Since the values for the indices $\tilde{c}\mu/\theta$ and $\tilde{c}\mu$ are higher for class 2, Lemma 2 implies that class 1 should be given priority until it reaches 0. However, this is in contradiction with Lemma 1, which states that it is optimal to give priority to class 2 close to the origin. Hence, we have proved that it is optimal to give full priority to class 2.

Now assume that $\tilde{c}_2\mu_2/\theta_2 \geq \tilde{c}_1\mu_1/\theta_1$ and $\tilde{c}_2\mu_2 \leq \tilde{c}_1\mu_1$. Assume there exists a state such that $n_1 > h_2(n_2)$ where priority is given to class 1 (here h_2 denotes the switching curve of Lemma 2 if we had assumed class 1 was prioritized). Then Lemma 2 tells us that whenever the process reaches a point where $n_1 \leq h_2(n_2)$ the priority will be switched to class 2 until the equilibrium is reached. This, however, is in contradiction with Lemma 1, since it tells us that class 1 should be prioritized close to the origin. Hence, by contradiction we obtain that for states far enough from the origin, priority should be given to class 2. By Lemma 2, first item, we then obtain the result. □

3.2. Optimal policy in overload for an arbitrary number of classes

In this section we assume again an arbitrary number of classes, i.e., $K \geq 2$. To complete the analysis of Problem P we are left with the setting $\rho > 1$, in which case the objective is to minimize the average cost (the latter being strictly positive). The following proposition states an optimal control for the fluid model.

Proposition 2. *Let $\lambda_k, \mu_k, \theta_k, c_k$ and d_k be given for $k \in \{1, \ldots, K\}$, and assume the classes are ordered such that $\frac{\tilde{c}_1\mu_1}{\theta_1} \geq \frac{\tilde{c}_2\mu_2}{\theta_2} \geq \cdots \geq \frac{\tilde{c}_K\mu_K}{\theta_K}$. If $\rho > 1$, then an optimal solution $s^*(\cdot)$ for Problem P under the average cost criteria is:*

$$s^*(t) = \left(\rho_1, \ldots, \rho_l, 1 - \sum_{i=1}^{l(t)} \rho_i, 0, \ldots, 0\right),$$

with $l(t) := \min\{k : n_{k+1}(t) > 0\}$. That is, priority is given according to the index $\tilde{c}\mu/\theta$.

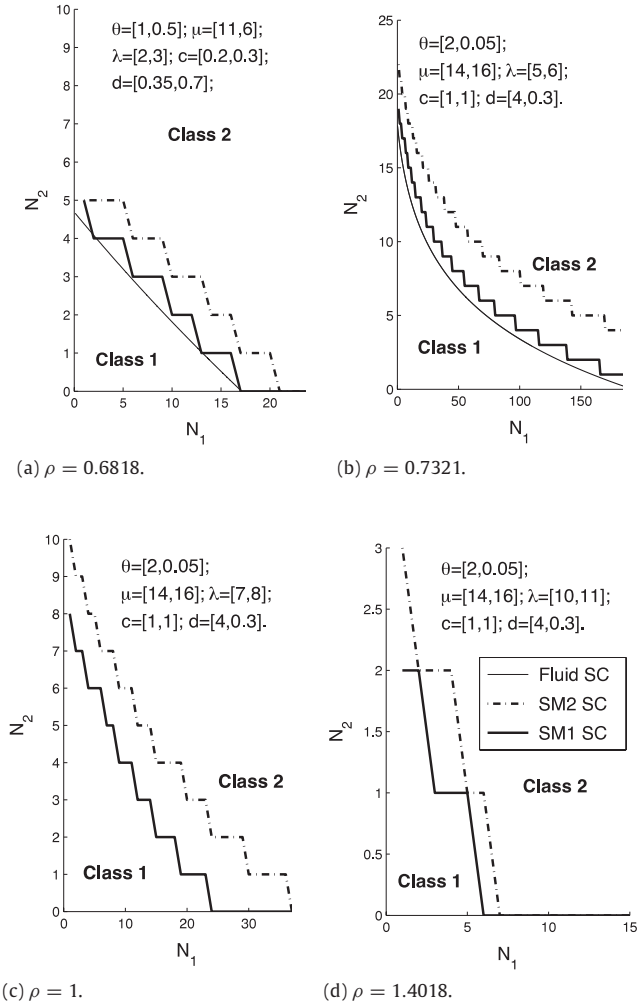
Proof. We first determine the optimal equilibrium point. An equilibrium point satisfies $0 = \lambda_k - \mu_k s_k - \theta_k n_k$, for all k. Hence, the optimal control (in equilibrium) that minimizes the equilibrium point is given by

$$\arg\min_{s \in \mathcal{S}} \sum_{k=1}^{K} \tilde{c}_k n_k = \arg\min_{s \in \mathcal{S}} \sum_{k=1}^{K} \tilde{c}_k \frac{\lambda_k - \mu_k s_k}{\theta_k} = \arg\min_{s \in \mathcal{S}} \sum_{k=1}^{K} -\frac{\tilde{c}_k \mu_k}{\theta_k} s_k.$$

This is minimized by giving the highest priority according to the $\tilde{c}\mu/\theta$ rule, that is the optimal equilibrium point is given by $n^* = (0, \ldots, 0, \frac{\lambda_{j+1} - \mu_{j+1}(1 - \sum_{i=1}^{j} \rho_i)}{\theta_{j+1}}, \frac{\lambda_{j+2}}{\theta_{j+2}}, \ldots, \frac{\lambda_K}{\theta_K})$ and $s^* = (\rho_1, \ldots, \rho_j, 1 - \sum_{i=1}^{j} \rho_i, 0, \ldots, 0)$, with j such that $\sum_{i=1}^{j} \rho_i < 1$ and $\sum_{i=1}^{j+1} \rho_i \geq 1$.

It remains to be checked that under the control $s^*(\cdot)$ as stated in the proposition, the fluid dynamics converge to the optimal equilibrium point. This can be seen as follows. Let $n^*(\cdot)$ denote the trajectory corresponding to the control $s^*(\cdot)$. Consider $w_j^*(t) := \sum_{k=1}^{j} n_k^*(t)/\mu_k$. By definition of $s^*(t)$ we have $\mathrm{d}w_j^*(t)/\mathrm{d}t = \sum_{k=1}^{j} \rho_k - 1 - \sum_{k=1}^{j} \theta_k n_k^*(t)/\mu_k < -(1 - \sum_{k=1}^{j} \rho_k)$ when $w_j^*(t) > 0$. Hence, in a finite time T the process hits zero, $w_j^*(T) = 0$, and stays there. From that moment on, class $j + 1$ is given capacity $1 - \sum_{k=1}^{j} \rho_k$ if present. Hence, it follows directly that this converges to the point n_{j+1}^*, which solves $0 = \lambda_{j+1} - \mu_{j+1}(1 - \sum_{k=1}^{j} \rho_k) - \theta_{j+1} n_{j+1}$. Since for $t > T$ we have $n_{j+1}^*(t) > 0$, classes $j+2, \ldots, K$ do not receive any service. Hence, their dynamics is described by $\mathrm{d}n_i^*(t)/\mathrm{d}t = \lambda_i - \theta_i n_i^*(t)$, and $n_i^*(t)$ converges to λ_i/θ_i, $i \in \{j + 2, \ldots, K\}$. □

We note that the $\tilde{c}\mu/\theta$ rule has previously been proposed by Atar et al. in [15,17], where optimal scheduling in the presence of abandonments was studied for the many-server setting. The rule was obtained by solving a fluid control model. The fluid model is similar to the one of Proposition 2, but has the additional condition $s_k \leq n_k$, which is due to the multi-server setting.



(a) $\rho = 0.6818$. (b) $\rho = 0.7321$.

(c) $\rho = 1$. (d) $\rho = 1.4018$.

Fig. 3. Switching curves for SM1, SM2 and the fluid control model.

3.3. Optimal control comparison of stochastic model with fluid model

In this section, we compare the switching curve that we obtained for the fluid model with the optimal solution for the stochastic models SM1 and SM2 obtained numerically by value iteration. In Fig. 3 we make this comparison for different sets of parameters. Note that the optimal stochastic switching curve of SM1 is always below the switching curve of SM2. This is due to the fact that allowing customers to abandon while being served, as in SM1, makes the effect of abandonments more significant.

We consider the underload case $\rho < 1$ in Fig. 3(a) and (b), the critical regime $\rho = 1$ in Fig. 3(c) and the overload setting $\rho > 1$ in Fig. 3(d). Moreover, we note that Fig. 3(b)–(d) correspond to the parameters of Example 1 in Section 5.

In Fig. 3(a)–(b) the parameters are such that $\tilde{c}_1\mu_1 \geq \tilde{c}_2\mu_2$ and $\tilde{c}_1\mu_1/\theta_1 \leq \tilde{c}_2\mu_2/\theta_2$, hence the optimal fluid solution is characterized by a switching curve and priority is given to class 2 above the curve and to class 1 below the curve. We observe that the fluid optimal switching curve approximates the stochastic optimal switching curve very well, except for a constant that apparently disappears after the fluid scaling.

In Fig. 3(c)–(d) the optimal stochastic policy is characterized by a switching curve where class 2 is served in states above the curve and class 1 in states below the curve. The optimal control in the fluid model is however to give strict priority to class 2 (in the case $\rho > 1$), since $\tilde{c}_1\mu_1/\theta_1 \leq \tilde{c}_2\mu_2/\theta_2$. For Fig. 3(c) the average number of customers in the system SM1 under the optimal policy is $(\overline{N}_1, \overline{N}_2) = (0.7796, 4.1194)$ which lies below the switching curve. Hence, the stochastic optimal policy will give most of the time priority to class 1. The optimal fluid control does not capture this property since the fluid switching curve $h(\cdot)$ vanishes for $\rho = 1$. In Example 1 of Section 5 we will see that the suboptimality gap when applying the optimal fluid control to the stochastic model is around 30%. In the case $\rho > 1$ is large enough, our index policy turns out to work well, see the numerical Section 5. This is explained by the fact that the process is living above the

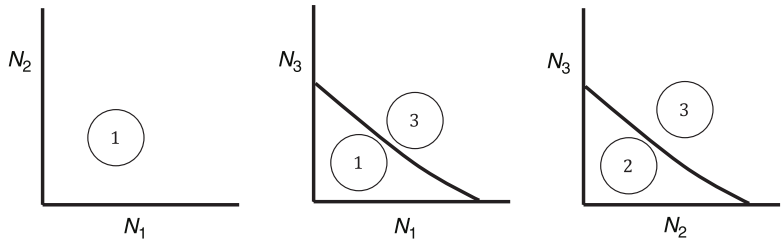


Fig. 4. An example of the heuristics for the case $K = 3$ when $\tilde{c}_1\mu_1 \geq \tilde{c}_2\mu_2 \geq \tilde{c}_3\mu_3$ and $\tilde{c}_3\mu_3/\theta_3 \geq \tilde{c}_1\mu_1/\theta_1 \geq \tilde{c}_2\mu_2/\theta_2$.

switching curve. For example, for the parameters as chosen in Fig. 3(d), the average number of customers in the system SM1 is $(\overline{N}_1, \overline{N}_2) = (3.0088, 3.4849)$ which lies above the switching curve. Hence, the stochastic optimal policy will give most of the time priority to class 2, which coincides with the optimal fluid control. In Example 1 of Section 5 we will see that the suboptimality gap when applying the optimal fluid control to the stochastic model is small.

In Section 4 we discuss how to translate the fluid optimal solution to the stochastic setting. In Section 5 we will numerically evaluate the performance of the heuristic when applied to the stochastic model. In fact, we will observe good performance. However, we do not have any result on the suboptimality gap. In the literature, asymptotic fluid optimality results have been obtained for various dynamic scheduling problems in queueing models, see for example [26–30]. More precisely, it is shown that when employing the optimal control resulting from the fluid model to the stochastic model, the fluid-scaled cost converges to the optimal cost of the fluid control model, the latter being in fact a provable lower bound on the stochastic cost. In this particular model the fluid model is presented as an approximation, there is no certainty that when applying the optimal fluid control in the stochastic model, this will be asymptotically optimal. We do believe though that when scaling λ_k's, μ_k's, the scaled queue length processes (when scaling space) behave according to the fluid dynamics. We note here that [15,17] show in fact that the $\tilde{c}\mu/\theta$ rule is asymptotically fluid optimal in a multi-server setting and assuming overload. Due to the multi-server setting, the authors of [15,17] need a different limiting regime: the arrival rates and the number of servers are scaled, while the service rate of each server is kept fixed to μ. We do expect though, in the case of overload, that a similar proof technique can be applied to our model.

4. Heuristics for an arbitrary number of classes

In this section, we will propose a heuristic for the stochastic optimization model with abandonments. This heuristic is based on the insights we obtained from the fluid control model.

We first consider the overload setting. In that case, the optimal fluid policy is to give priority according to the $\tilde{c}\mu/\theta$-rule. In Section 5 we will evaluate this policy when employed in the stochastic model (in overload).

We now consider the case of underload. Recall that in Proposition 1 we have seen that the optimal fluid control has a remarkable structure in the case of two classes: close to the origin the $\tilde{c}\mu$-rule is optimal, and when one of the fluids is sufficiently large the $\tilde{c}\mu/\theta$-rule is optimal. We observe the same structural property in the optimal solution for the stochastic control problem obtained numerically, see Section 3.3 and Fig. 5 (left). Our approach is thus to develop a heuristic that follows this insight, that is, close to the origin it will behave according to the $\tilde{c}\mu$-rule, and when the number of users in one of the classes is sufficiently large it will follow the $\tilde{c}\mu/\theta$-rule. It is not clear what should be the best choice for the threshold to decide whether the $\tilde{c}\mu$ rule or the $\tilde{c}\mu/\theta$ rule should be applied.

We propose the following heuristic, which is based on the two-class fluid analysis: for a general K-class queue we compare all classes pairwise and calculate the switching curves of the paired systems, see for example Fig. 4 where $\tilde{c}_1\mu_1 \geq \tilde{c}_2\mu_2 \geq \tilde{c}_3\mu_3$ and $\tilde{c}_3\mu_3/\theta_3 \geq \tilde{c}_1\mu_1/\theta_1 \geq \tilde{c}_2\mu_2/\theta_2$. Then, whenever the state $(N_1, N_2, \ldots, N_k)$ satisfies that all pairs (N_i, N_j) lie under their corresponding switching curves, we give priority to the class with the highest value for $\tilde{c}\mu$. However, if there is at least one state (N_i, N_j) that lies above its corresponding switching curve, we will give priority to the class with the highest value for $\tilde{c}\mu/\theta$. For example, for the parameters of Fig. 4 we will give priority to class 1 when both states (N_2, N_3) and (N_1, N_3) lie below their corresponding switching curves and otherwise priority is given to class 3. Whenever the queue of one class is empty we analyze the system in the same way but only take into account the $K - 1$ queues that are non-empty. For a better understanding we give a pseudo-code of the heuristic rule in Algorithm 1.

We propose an example with $K = 3$ to illustrate the heuristic we have just defined and to compare it to the optimal policy (obtained numerically by Value Iteration [33]). Let us consider the following set of parameters $\mu = [10, 10, 9]$; $\theta = [1, 0.5, 0.25]$; $c = [1.7, 1.7, 1.7]$; $d = [2, 2, 4]$; $\lambda = [2, 2, 1]$. Hence, $\tilde{c}_1\mu_1 \geq \tilde{c}_2\mu_2 \geq \tilde{c}_3\mu_3$ and $\tilde{c}_1\mu_1/\theta_1 \leq \tilde{c}_2\mu_2/\theta_2 \leq \tilde{c}_3\mu_3/\theta_3$. Under our heuristic, class 1 will be served when all three classes are close enough to the origin, according to the $\tilde{c}\mu$ rule, and class 3 will be served otherwise, according to the $\tilde{c}\mu/\theta$ rule. Class 2 will be served in the following two cases: (i) when class 1 is empty and (N_2, N_3) is sufficiently close to the origin (follows from the $\tilde{c}\mu$ rule) and (ii) when class 3 is empty and (N_1, N_2) is sufficiently far from the origin (follows from the $\tilde{c}\mu/\theta$ rule). In Fig. 5 we plot the actions under the optimal scheduling rule (calculated by Value Iteration) (left) and under our heuristic (right). We observe that the heuristic rule shows a qualitatively similar structure to the optimal solution. In Section 5.2 we will present a numerical comparison of its performance.

Algorithm 1 Algorithm to compute heuristic scheduling rule for an arbitrary K

Assume r queues are non empty.
Let N_i be the state of class $i \in \{1, \ldots, r\}$.
Compute the indices $\tilde{c}\mu$ and $\tilde{c}\mu/\theta$ for $i \in \{1, \ldots, r\}$.
Given a pair of classes i and j, such that $\tilde{c}_i\mu_i/\theta_i \geq \tilde{c}_j\mu_j/\theta_j$, compute the switching curve h_{ij} as given by Equation (3.1).
if for all i, j, $N_i \leq h_{ij}(N_j)$ **then**
    Give priority to the class with highest index $\tilde{c}\mu$
else
    Give priority to the class with highest index $\tilde{c}\mu/\theta$
end if

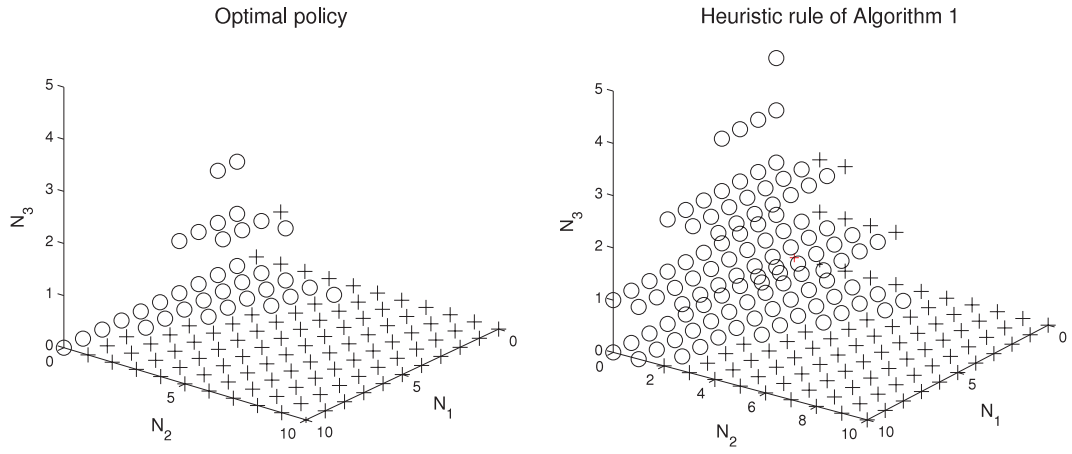


Fig. 5. Optimal policy and heuristic for a 3-class single server example for SM2. Circles indicate that class 1 is served, pluses that class 2 is served and absence of a sign corresponds to class 3 being served.

5. Numerical results

In this section we simulate the stochastic model and evaluate numerically the performance of the heuristic described in Algorithm 1. We compare the performance of our heuristic rule against the optimal policy. The latter is calculated using the Value Iteration algorithm [33]. We also simulate the following index policies available in the literature:

- the $\tilde{c}\mu/\theta$-rule. This rule was introduced in [15,17] where it was proved to be asymptotically fluid optimal for a multi-server system in overload. As shown in Proposition 2 this rule is also optimal for our fluid model in overload.
- the $\tilde{c}\mu/\theta - c$-rule. This rule was derived in [18] for the system SM2 (without arrivals) with the modification that the user in service also contributes to the cost.
- the $\tilde{c}\mu$-rule. This is the greedy or myopic rule that minimizes the instantaneous cost. This rule can be seen as a counterpart of the well-known $c\mu$-rule [21] for the system with abandonments.

Before we start describing in detail the results, we provide below our main conclusions:

- the qualitative performance in the SM1 and SM2 systems are very similar.
- the $\tilde{c}\mu/\theta$ and the $\tilde{c}\mu/\theta - c$-rules perform very well in overload.
- our heuristic (as proposed in Algorithm 1) performs very well across all loads.

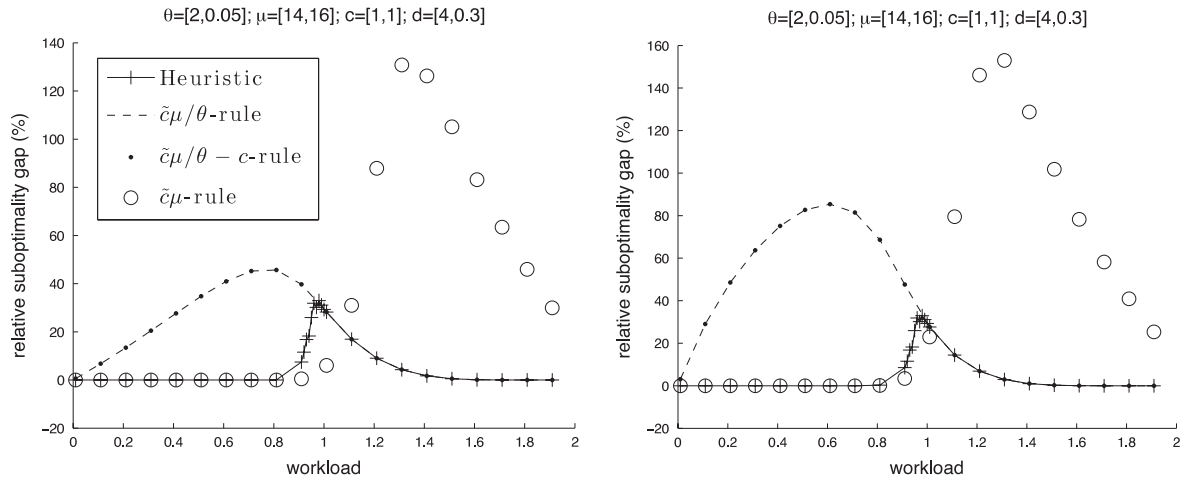
For the sake of fairness we can mention that even though the index rules $\tilde{c}\mu/\theta$ and $\tilde{c}\mu$ perform worse than the heuristic rule, they are simpler to implement since they are state independent.

We now present the scenarios we have simulated. In Section 5.1 we consider the case $K = 2$ and in Section 5.2 the case $K = 3$.

5.1. Performance analysis for $K = 2$

We consider the two models SM1 and SM2, and we calculate the relative suboptimality gap for the policies described above. In Example 1 and 2 we fix the parameters c, d, μ and θ and set $\rho_1 = \rho_2$ and vary the total workload ρ. In Example 3 we fix ρ and vary the value of θ_1.

- *Example* 1: in this first example we set $\theta = [2, 0.05]$; $\mu = [14, 16]$; $c = [1, 1]$; $d = [4, 0.3]$, such that $\tilde{c}_1\mu_1 \geq \tilde{c}_2\mu_2$ and $\tilde{c}_2\mu_2/\theta_2 \geq \tilde{c}_1\mu_1/\theta_1$. The results for this example are depicted in Fig. 6.



(a) Relative suboptimality gap for SM1.

(b) Relative suboptimality gap for SM2.

Fig. 6. Performance comparison of policies for Example 1.

In underload the heuristic and the $\tilde{c}\mu$-rule behave optimally, while the $\tilde{c}\mu/\theta$ and $\tilde{c}\mu/\theta - c$-rules behave very poorly. In Fig. 3(b) we plotted the switching curves corresponding to the load $\rho = 0.73$. In fact, the average number of users (in the SM1 system with $\rho = 0.73$) is given by $(\overline{N}_1, \overline{N}_2) = (0.4047, 1.5092)$ which is a state far below both the SM1 switching curve and the fluid switching curve. This shows why both our heuristic and the $\tilde{c}\mu$-rule (this is the control below the switching curve) behave close to optimal.

In the overload case though, the $\tilde{c}\mu$-rule incurs a high relative suboptimality gap while our heuristic and the $\tilde{c}\mu/\theta$ and $\tilde{c}\mu/\theta - c$-rules are close to optimal. The latter conforms with what we expected as described in Section 3.3.

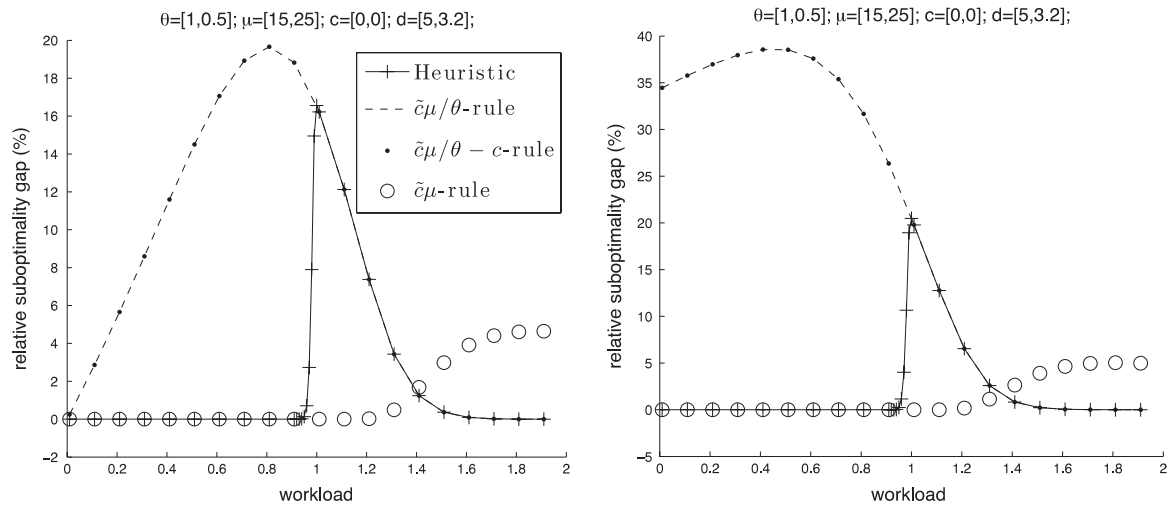
We observe that when the load is close to the critical regime $\rho = 1$ the suboptimality gap is around 30%. Our heuristic will give priority to class 2 in the case $\rho > 1$ and has a switching curve very close to the origin in the case $\rho = 1 - \epsilon$. In Fig. 3(c), which corresponds to the current example, we see that the optimal policy for the stochastic optimization problem is described by a switching curve for $\rho = 1$. Hence, when we are in a state below the switching curve, class 1 will be given priority. The process when $\rho = 1$ lives on average close to the stochastic switching curve, therefore, our policy can be far from optimal, as discussed in Section 3.3.

- *Example* 2: in this second example we set $\theta = [1, 0.5]$; $\mu = [15, 25]$; $c = [0, 0]$; $d = [5, 3.2]$, so that $\tilde{c}_1\mu_1 \geq \tilde{c}_2\mu_2$ and $\tilde{c}_2\mu_2/\theta_2 \geq \tilde{c}_1\mu_1/\theta_1$. As explained in Remark 3, setting $c_1 = c_2 = 0$ gives a different interpretation of the model: customers will abandon the system when a certain deadline is met before they have attained full service. In this case the $\tilde{c}\mu/\theta$, $\tilde{c}\mu/\theta - c$ and the $\tilde{c}\mu$ rules reduce to the $d\mu$, $d\mu$, and $d\theta\mu$ rules, respectively. We observe that the index $d\mu$ does not perform well in underload but is close to optimal in overload. The opposite holds for the $d\theta\mu$ rule. Our policy is optimal in underload and as good as the $d\mu$ index in overload, see Fig. 7.

  In Fig. 8 we plotted the optimal switching curves for the stochastic models SM1 and SM2 (obtained by value iteration), as well as the optimal fluid switching curve $h(\cdot)$. Fig. 8(a) corresponds to load $\rho = 0.8867$. In that case, the average number of customers is given by $(\overline{N}_1, \overline{N}_2) = (0.6859, 2.6963)$, which is a state far below all switching curves. Hence, this shows why our heuristic and the $\tilde{c}\mu$-rule perform close to optimal. On the other hand, when $\rho = 1$, Fig. 8(b), the optimal control in the fluid model is to serve class 2, so there is no switching curve. Under the optimal policy for the stochastic model, the average number of customers is given by $(\overline{N}_1, \overline{N}_2) = (0.763, 4.2703)$. This is a state far below the switching curves of the stochastic model. Hence, most of the time priority is given to class 1 under the optimal policy. This explains why our heuristic gives a positive optimality gap of 16%. However, as the load of the system increases ($\rho > 1$) the process will live more above the optimal switching curve for the stochastic model. See for example Fig. 8(c) for load $\rho = 1.52$ for which the average number of customers under the optimal policy is given by $(\overline{N}_1, \overline{N}_2) = (6.8054, 3.7244)$. This explains why our heuristic, which gives priority to class 2, has a suboptimality gap very close to 0%.

Remark 4 (*Peak When Workload Close to 1*)**.** We observe in Examples 1 and 2 that when the workload is close to 1 a peak appears in the suboptimality gap for the heuristic rule. This can be explained by the following. In the proof of Lemma 2, see Appendix, we observe a switching curve whenever

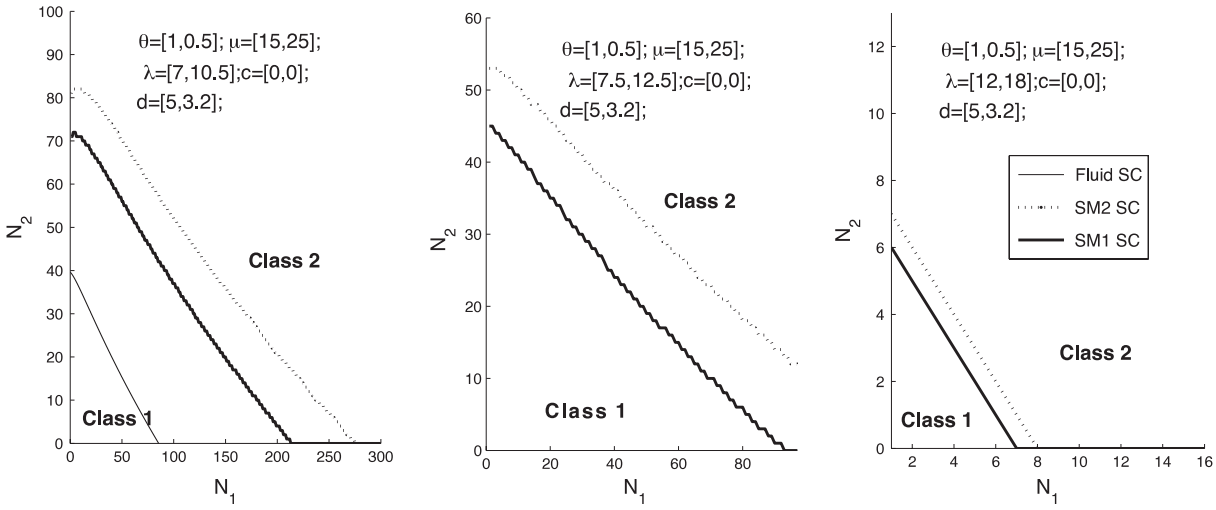
$$h(0) = (1 - \rho_1 - \rho_2)\frac{\mu_2}{\theta_1\theta_2}\left(\frac{\tilde{c}_1\mu_1 - \tilde{c}_2\mu_2}{\frac{\tilde{c}_2\mu_2}{\theta_2} - \frac{\tilde{c}_1\mu_1}{\theta_1}}\right) > 0.$$



(a) Relative suboptimality gap for SM1.

(b) Relative suboptimality gap for SM2.

Fig. 7. Performance comparison of policies for Example 2.



(a) $\rho = 0.8867$.

(b) $\rho = 1$.

(c) $\rho = 1.52$.

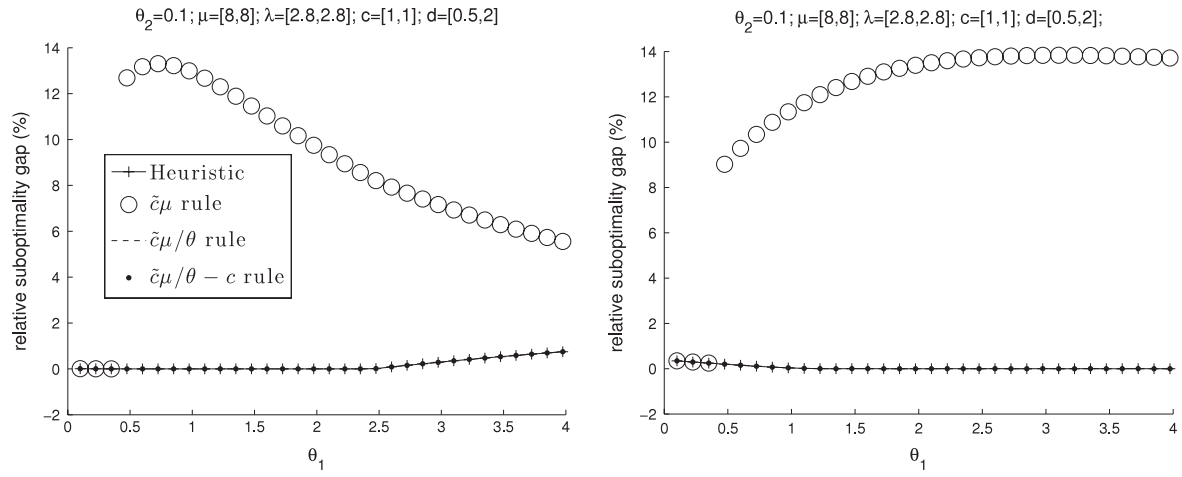
Fig. 8. Comparison of switching curves for Example 2.

Therefore, as $1 - \rho_1 - \rho_2 \to 0$ the switching curve vanishes, that is, the heuristic becomes equivalent to the $\tilde{c}\mu/\theta$-rule. However, around $\rho = 1$ the optimal stochastic control still follows the $\tilde{c}\mu$-rule in a non-negligible part of the state space, see for instance Fig. 8(b).

- *Example* 3: we consider the following parameters: $\theta_2 = 0.1$; $\mu = [8, 8]$; $\lambda = [2.8, 2.8]$; $c = [1, 1]$; $d = [0.5, 2]$, and we let θ_1 vary. Hence, $\rho = 0.7$, i.e., we are in underload. The results are plotted in Fig. 9.
  When $\theta_1 \in [0, 0.4]$, we have $\tilde{c}_2\mu_2/\theta_2 \geq \tilde{c}_1\mu_1/\theta_1$ and $\tilde{c}_2\mu_2 \geq \tilde{c}_1\mu_1$, in which case the heuristic gives priority to class 2, as do all the index policies. On the other hand, when $\theta_1 \in (0.4, 4]$, then $\tilde{c}_2\mu_2/\theta_2 \geq \tilde{c}_1\mu_1/\theta_1$ and $\tilde{c}_1\mu_1 \geq \tilde{c}_2\mu_2$ and a switching curve does appear in the heuristic. For the cases where no switching curve appears ($\theta_1 \in [0, 0.4]$), all the index rules are optimal, but as soon as a switching curve emerges in the heuristic the $\tilde{c}\mu$ rule gives a positive suboptimality gap. The reason why the $\tilde{c}\mu$-rule performs bad in this particular case is that as soon as the ratio $\frac{\tilde{c}_1\mu_1 - \tilde{c}_2\mu_2}{\tilde{c}_2\mu_2/\theta_2 - \tilde{c}_1\mu_1/\theta_1}$ becomes small, the switching curve gets close to zero and hence the $\tilde{c}\mu/\theta$ rule becomes optimal.

5.2. Performance analysis for $K > 2$

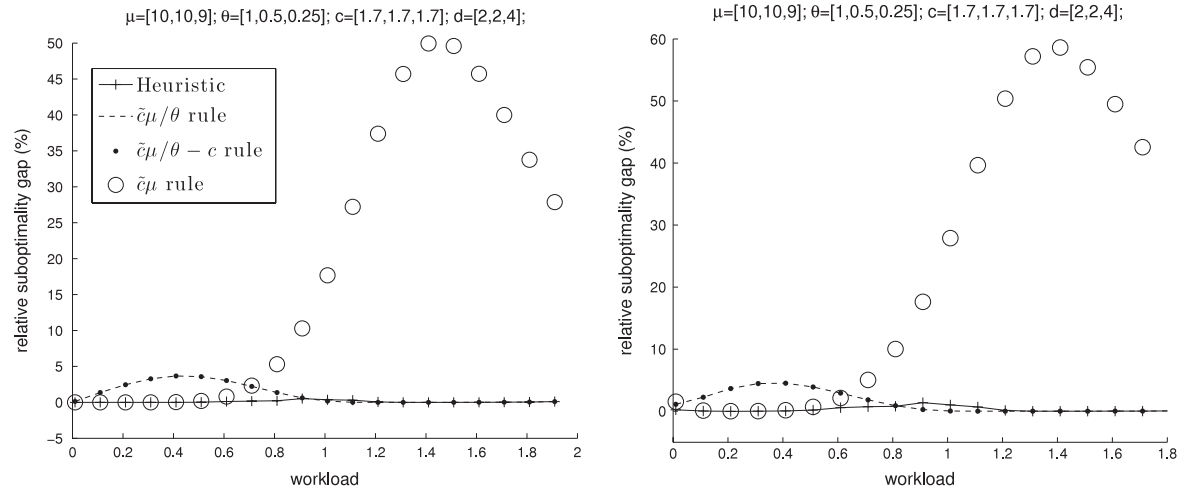
We analyze the relative performance of the heuristic as explained in Section 4. Here we take the same example that was introduced in Section 4 with parameters $\mu = [10, 10, 9]$; $\theta = [1, 0.5, 0.25]$; $c = [1.7, 1.7, 1.7]$; $d = [2, 2, 4]$. Let



(a) Relative suboptimality gap for SM1.

(b) Relative suboptimality gap for SM2.

Fig. 9. Performance comparison of policies for Example 3, with load $\rho = 0.7$.



(a) Relative suboptimality gap for SM1.

(b) Relative suboptimality gap for SM2.

Fig. 10. Performance comparison of policies for $K = 3$.

$\lambda_i = \lambda\beta_i,\ i = 1, 2, 3$, denote the arrival rate of class-i jobs, where λ denotes the total arrival rate and β_i is the fraction of class-i customers. We choose $\beta_i,\ i = 1, 2, 3$, in such a way that $\rho_1 = \rho_2 = \rho_3$. We vary the value of λ to change the total load in the system and we compute the relative suboptimality gap of all the policies, including the heuristic. The results are depicted in Fig. 10. We observe that our policy is optimal together with the $\tilde{c}\mu$ rule for very low loads but at some point, as the workload ρ increases, the $\tilde{c}\mu$-rule starts performing increasingly worse. On the other hand, the $\tilde{c}\mu/\theta$ and the $\tilde{c}\mu/\theta - c$ rules become optimal when the load becomes larger than 1. Our heuristic keeps a low suboptimality gap throughout.

6. Conclusions

In this paper, we have proposed a fluid model to obtain an approximative optimal control for a multi-class single server queue with abandonments. Using Pontryagin's Maximum Principle we have completely characterized the fluid optimal control in the case of two classes, which provides very valuable insights into the solution of the stochastic problem for an arbitrary number of classes: close to the origin the optimal control behaves according to the $\tilde{c}\mu$-rule and far from the origin according to the $\tilde{c}\mu/\theta$-rule. We then proposed a heuristic for the stochastic model for an arbitrary number of classes that follows this structure, and with numerical experiments we have shown that it performs very well.

There are several interesting research avenues that are worthwhile pursuing. An interesting problem is to establish the existence of a switch curve in the optimal policy of the stochastic problem. Another interesting open problem would be to show that the fluid-based heuristic is asymptotically optimal for the stochastic problem.

Acknowledgment

The Ph.D. fellowship of Maialen Larrañaga is funded by a research grant of the EADS Foundation (http://fondation.eads.com/).

Appendix

In this section we present the proofs of Lemmas 1 and 2, which were used in order to prove Proposition 1. In addition, we give the statement of Pontryagin's Maximum Principle.

Proof of Lemma 1. Lemma 1 states which class is optimal to serve close to the origin. We calculate the cost function when starting in a point very close to the origin $(n_1(0), n_2(0)) = (\varepsilon, \varepsilon)$ when priority is given to class 1. We do the same for the case when priority is given instead to class 2. When comparing both cost functions, we get the condition under which prioritizing class 1 gives lower cost than prioritizing class 2. We note that, it is sufficient to compare the above described two policies: since the control appears linearly, we can assume that under the optimal policy full priority will be given to one class as long as we start close enough to the origin, see [34].

We first consider the control that gives full priority to class 1. When class 1 hits zero, ρ_1 is given to class 1 and $1-\rho_1$ to class 2, until the equilibrium (0, 0) is reached. The cost under this policy, starting in state $(n_1(0), n_2(0)) = (\varepsilon, \varepsilon)$, is

$$C_1(t, n) := \int_0^T \tilde{c}_1 n_1(t) + \tilde{c}_2 n_2(t) \mathrm{d}t.$$

In order to compute the trajectories $n_1(t)$ and $n_2(t)$, we will split up the time into two time intervals, $[0, t_1]$ and $[t_1, t_2]$, where t_1 is the moment when class 1 hits zero and t_2 when class 2 hits zero. After some algebra, we obtain that for the interval $[0, t_1]$ the trajectories are as follows:

$$\begin{cases} n_1(t) = \left(\varepsilon + \dfrac{\mu_1 - \lambda_1}{\theta_1}\right) \mathrm{e}^{-\theta_1 t} + \dfrac{\lambda_1 - \mu_1}{\theta_1} = -\theta_1 t \left(\varepsilon + \dfrac{\mu_1 - \lambda_1}{\theta_1}\right) + \varepsilon + o(\varepsilon) & t \in [0, t_1], \\ n_2(t) = \left(\varepsilon - \dfrac{\lambda_2}{\theta_2}\right) \mathrm{e}^{-\theta_2 t} + \dfrac{\lambda_2}{\theta_2} = \theta_2 t \left(\dfrac{\lambda_2}{\theta_2} - \varepsilon\right) + \varepsilon + o(\varepsilon) & t \in [0, t_1]. \end{cases}$$

We used here that $t_2 \leq \frac{n_1(0)/\mu_1 + n_2(0)/\mu_2}{1-\rho} = O(\varepsilon)$,[3] hence $\mathrm{e}^{-\theta_1 t} = -\theta_1 t + 1 + o(\varepsilon)$, for $t \leq t_2$. (Here $o(\varepsilon) = g(\varepsilon)$ for $g(\cdot)$ a function that satisfies $\lim_{\varepsilon \to 0} g(\varepsilon)/\varepsilon = 0$.) We note that since ε is chosen small enough, $n_2(t) > 0$ for all $t < t_1$. Time t_1 being the moment at which class 1 empties, we obtain

$$t_1 = \frac{\varepsilon}{\theta_1 \left(\varepsilon + \frac{\mu_1 - \lambda_1}{\theta_1}\right)} = \frac{\varepsilon}{\mu_1 - \lambda_1} + o(\varepsilon),$$

therefore

$$n_2(t_1) = \frac{\lambda_2}{\theta_2} \frac{\varepsilon \theta_2}{\mu_1 - \lambda_1} + \varepsilon + o(\varepsilon). \tag{A.1}$$

Recall that t_2 is the time at which class 2 is emptied. In the interval $[t_1, t_2]$ class 1 receives service ρ_1 and class 2 service $1-\rho_1$. Hence, after some algebra we obtain that

$$\begin{cases} n_1(t) = 0, & t \in [t_1, t_2], \\ n_2(t) = A_2' \mathrm{e}^{-\theta_2 t} + \dfrac{\lambda_2 - \mu_2(1-\rho_1)}{\theta_2} = A_2' \left(-\theta_2 t + 1\right) + \dfrac{\lambda_2 - \mu_2(1-\rho_1)}{\theta_2} + o(\varepsilon), & t \in [t_1, t_2], \end{cases}$$

where A_2' is the constant of integration. Here we used that $t = O(\varepsilon)$, hence $\mathrm{e}^{-\theta_2 t} = -\theta_2 t + 1 + o(\varepsilon)$. Moreover, from (A.1) we obtain

$$A_2' = \frac{-\frac{\lambda_2}{\theta_2}\left(\frac{-\varepsilon\theta_2}{\mu_1 - \lambda_1}\right) + \varepsilon + \frac{\mu_2(1-\rho_1) - \lambda_2}{\theta_2}}{1 - \frac{\theta_2 \varepsilon}{\mu_1 - \lambda_1}} + o(\varepsilon),$$

[3] This follows from the fact that the workload $w(t) := n_1(t)/\mu_1 + n_2(t)/\mu_2$ has a negative drift smaller than or equal to $\rho - 1$, see the footnote in Section 3.

hence, we have

$$\begin{aligned} n_2(t) &= \frac{(\lambda_2\theta_2+\theta_2(\mu_1-\lambda_1))\varepsilon+(\mu_2(1-\rho_1)-\lambda_2)(\mu_1-\lambda_1)}{-\theta_2^2\varepsilon+(\mu_1-\lambda_1)\theta_2}(-\theta_2 t+1)+\frac{\lambda_2-\mu_2(1-\rho_1)}{\theta_2}+o(\varepsilon),\\ &= \frac{(\lambda_2\theta_2+\theta_2(\mu_1-\lambda_1))\varepsilon+(\mu_2(1-\rho_1)-\lambda_2)(\mu_1-\lambda_1)}{-\theta_2^2\varepsilon+(\mu_1-\lambda_1)\theta_2}(-\theta_2 t)\\ &\quad+\frac{(\mu_1-\lambda_1+\mu_2(1-\rho_1))\varepsilon}{-\theta_2\varepsilon+\mu_1-\lambda_1}+o(\varepsilon),\quad t\in[t_1,t_2], \end{aligned}$$

and from $n_2(t_2)=0$ we obtain

$$\begin{aligned} t_2 &= \frac{(\mu_1-\lambda_1+\mu_2(1-\rho_1))\varepsilon}{-\theta_2^2\varepsilon^2+(\lambda_2\theta_2+\theta_2(\mu_1-\lambda_1))\varepsilon+(\mu_2(1-\rho_1)-\lambda_2)(\mu_1-\lambda_1)}+o(\varepsilon)\\ &= \frac{(\mu_1-\lambda_1+\mu_2(1-\rho_1))\varepsilon}{(\mu_2(1-\rho_1)-\lambda_2)(\mu_1-\lambda_1)}+o(\varepsilon). \end{aligned}$$

We can now compute the cost function:

$$\begin{aligned} C_1(t,(\varepsilon,\varepsilon)) &= \int_0^{t_1}\tilde{c}_1\left(\left(\varepsilon+\frac{\mu_1-\lambda_1}{\theta_1}\right)(-\theta_1 t)+\varepsilon\right)+\tilde{c}_2\left(\left(\varepsilon-\frac{\lambda_2}{\theta_2}\right)(-\theta_2 t)+\varepsilon\right)\mathrm{d}t\\ &\quad+\int_{t_1}^{t_2}\tilde{c}_2\left(\frac{(\lambda_2\theta_2+\theta_2(\mu_1-\lambda_1))\varepsilon+(\mu_2(1-\rho_1)-\lambda_2)(\mu_1-\lambda_1)}{-\theta_2^2\varepsilon+(\mu_1-\lambda_1)\theta_2}(-\theta_2 t)\right)\mathrm{d}t\\ &\quad+\int_{t_1}^{t_2}\tilde{c}_2\left(\frac{(\mu_1-\lambda_1+\mu_2(1-\rho_1))\varepsilon}{-\theta_2\varepsilon+\mu_1-\lambda_1}\right)\mathrm{d}t+o(\varepsilon^2)\\ &= \varepsilon^2\left(\frac{\tilde{c}_1}{2(\mu_1-\lambda_1)}+\tilde{c}_2\frac{2(\mu_1-\lambda_1)+\lambda_2}{2(\mu_1-\lambda_1)^2}\right)\\ &\quad-\tilde{c}_2\frac{(\lambda_2\theta_2+\theta_2(\mu_1-\lambda_1))\varepsilon+(\mu_2(1-\rho_1)-\lambda_2)(\mu_1-\lambda_1)}{2(-\theta_2\varepsilon+(\mu_1-\lambda_1))}\left((t_2)^2-(t_1)^2\right)\\ &\quad+\tilde{c}_2\left(\frac{(\mu_1-\lambda_1+\mu_2(1-\rho_1))\varepsilon}{-\theta_2\varepsilon+\mu_1-\lambda_1}\right)(t_2-t_1)+o(\varepsilon^2), \end{aligned}$$

where

$$\begin{aligned} (t_2)^2-(t_1)^2 &= \frac{(b_1\varepsilon^2+b_2\varepsilon)^2}{(-b_1^2\varepsilon^2+b_3\varepsilon+b_4)^2}-\varepsilon^2 b_5^2+o(\varepsilon^2)=\frac{b_2^2\varepsilon^2}{b_4^2}-\varepsilon^2 b_5^2+o(\varepsilon^2),\\ t_2-t_1 &= \frac{b_1\varepsilon^2+b_2\varepsilon}{b_4}-b_5\varepsilon+o(\varepsilon)=\left(\frac{b_2}{b_4}-b_5\right)\varepsilon+o(\varepsilon^2)+o(\varepsilon), \end{aligned}$$

with

$$\begin{aligned} &b_1=-\theta_2,\qquad b_2=\mu_1-\lambda_1+\mu_2(1-\rho_1),\\ &b_3=\lambda_2\theta_2+\theta_2(\mu_1-\lambda_1),\qquad b_4=(\mu_2(1-\rho_1)-\lambda_2)(\mu_1-\lambda_1),\qquad b_5=\frac{1}{\mu_1-\lambda_1}. \end{aligned}$$

After some calculations, we then obtain

$$C_1(t,(\varepsilon,\varepsilon))=\tilde{c}_1\varepsilon^2\left(\frac{1}{2(\mu_1-\lambda_1)}\right)+\tilde{c}_2\varepsilon^2\left(\frac{2(\mu_1-\lambda_1)+\lambda_2}{2(\mu_1-\lambda_1)^2}+\frac{(\mu_1-\lambda_1+\lambda_2)^2}{2(\mu_2(1-\rho_1)-\lambda_2)(\mu_1-\lambda_1)^2}\right)+o(\varepsilon^2).$$

By symmetry, the cost when instead class 2 is given priority is given by

$$C_2(t,(\varepsilon,\varepsilon))=\tilde{c}_2\varepsilon^2\left(\frac{1}{2(\mu_2-\lambda_2)}\right)+\tilde{c}_1\varepsilon^2\left(\frac{2(\mu_2-\lambda_2)+\lambda_1}{2(\mu_2-\lambda_2)^2}+\frac{(\mu_2-\lambda_2+\lambda_1)^2}{2(\mu_1(1-\rho_2)-\lambda_1)(\mu_2-\lambda_2)^2}\right)+o(\varepsilon^2).$$

It can now be checked that $C_1(t,(\varepsilon,\varepsilon))\le C_2(t,(\varepsilon,\varepsilon))$ if and only if $\tilde{c}_1\mu_1\ge\tilde{c}_2\mu_2$, (given that we are in underload ($\rho<1$)), which proves the result. □

We now present the proof of Lemma 2. We will make use of Pontryagin's Maximum Principle and we will therefore first present the statement of the theorem adapted to our problem formulation.

Theorem 1. *Necessary conditions for an optimal control of problem P (for $\rho \leq 1$) are given by the* Pontryagin's Maximum Principle [22, *Theorem* 3.26]*: let $s^*(\cdot)$ be an optimal control, piecewise continuous, and let $n^*(\cdot)$ be the associated optimal trajectory. Let T be the optimal final time subject to optimization, i.e., T is such that $n^*(T) = 0$. Then, there exists a continuous function $\gamma^*(t) = (\gamma_1^*(t), \ldots, \gamma_K^*(t)) \neq (0, \ldots, 0)$ with piecewise continuous derivatives that for all $t \in [0, T]$ satisfies,*

1.
$$\dot{\gamma}_k^* = -\frac{\partial \mathcal{H}(n^*, s^*, \gamma^*, t)}{\partial n_k} \quad \forall k \in \{1, \ldots, K\}, \tag{A.2}$$
in all the continuity points of $s^*(t)$ where $\mathcal{H}$ is the Hamiltonian of the system given by,
$$\mathcal{H}(n(t), s(t), \gamma(t), t) = \sum_{k=1}^{K} \tilde{c}_k n_k(t) + \gamma^{\mathsf{T}}(t) \begin{bmatrix} \lambda_1 - \mu_1 s_1(t) - \theta_1 n_1(t) \\ \vdots \\ \lambda_K - \mu_K s_K(t) - \theta_K n_K(t) \end{bmatrix}, \tag{A.3}$$
2.
$$s^*(t) = \arg\min_{s \in \mathcal{S}} \mathcal{H}(n^*(t), s, \gamma^*(t), t), \tag{A.4}$$
3.
$$\dot{n}_k^*(t) = \lambda_k - \mu_k s_k^*(t) - \theta_k n_k^*(t) \tag{A.5}$$
in all the continuity points of $s^*(t)$, with $n^*(0) = n_0$, $n^*(T) = 0$,
4. *and the transversality condition*
$$\mathcal{H}(n^*(t), s^*(t), \gamma^*(t), t) = 0, \quad \forall t \in [0, T]. \tag{A.6}$$

Proof of Lemma 2. We assumed that $\rho_1 + \rho_2 < 1$ and as explained in Section 3 under any non-idling control the optimal final time is finite. We first solve for Eq. (A.2), which gives us $\gamma_k^*(t) = C_k' e^{\theta_k t} + \frac{\tilde{c}_k}{\theta_k}$ where C_k' are constants of integration. Hence, Eq. (A.4) is equivalent to solving
$$\arg\min_{\mathbf{s} \in \mathcal{S}} \sum_{k=1}^{2} -\mu_k \left(C_k' e^{\theta_k t} + \frac{\tilde{c}_k}{\theta_k} \right) s_k. \tag{A.7}$$
Hence, under the optimal control the class with higher value for $\mu_k \left(C_k' e^{\theta_k t} + \frac{\tilde{c}_k}{\theta_k} \right)$ value will be prioritized. Without loss of generality, we have assumed that for the given initial conditions $n(0)$ priority is given to class 2 when $t = 0$. To determine whether there is a switch in priorities we study the following switching function:
$$\phi(t) := \mu_1 \left(C_1' e^{\theta_1 t} + \frac{\tilde{c}_1}{\theta_1} \right) - \mu_2 \left(C_2' e^{\theta_2 t} + \frac{\tilde{c}_2}{\theta_2} \right).$$
Hence, at time t, if $\phi(t) < 0$, then it is optimal to prioritize class 2 and if $\phi(t) > 0$, then it is optimal to prioritize class 1. Recall that we assumed it to be optimal to serve class 2 at time 0, so that $\phi(0) < 0$. Note that, this function can at most have one zero. Hence, two things can happen:

1. $\exists\, t \in [0, T]$, s.t $\phi(t) = 0$. We denote this time by t_1, i.e., $\phi(t_1) = 0$. Hence, the optimal action in the interval $[0, t_1]$ is to prioritize class 2, and the optimal action in the interval $[t_1, T]$ is to prioritize class 1.
2. $\nexists\, t \in [0, T]$ s.t $\phi(t) = 0$. Hence, the optimal action is to prioritize class 2 for all $t \in [0, T]$.

Hence, it remains to be derived under which conditions a switch can occur, and in which state (n_1, n_2) this happens. For the ease of notation we assume $t_1 = 0$, so the state $(n_1(0), n_2(0)) = (n_{10}, n_{20})$ is a point on the switching curve. We further define t_2 as the moment at which the amount of fluid in class 1 empties and t_3 as the time at which the equilibrium $(0, 0)$ is reached. Hence, $s^*(t) = (0, 1)$ for $t = 0$, $s^*(t) = (1, 0)$ for $t \in (0, t_2]$, $s^*(t) = (\rho_1, 1 - \rho_1)$ for $t \in (t_2, t_3]$.

We will now study the switching function, which fully characterizes which class is given priority. Note that the constants C_1', C_2', which appear in the switching function $\phi(t)$, are still to be determined. In order to obtain C_1', C_2', we will apply the transversality conditions of the Pontryagin's Maximum Principle in Eq. (A.6).

If $t = t_1 = 0$, then $s_2^*(t) = 1$ and $s_1^*(t) = 0$. Using that $\gamma_k^*(t) = C_k' e^{\theta_k t} + \tilde{c}_k/\theta_k$, we obtain that the Hamiltonian for $t = t_1 = 0$ is given by
$$\mathcal{H}(n^*(t), s^*(t), \gamma^*(t), t) = \tilde{c}_1 \frac{\lambda_1}{\theta_1} - C_1' \theta_1 \left(n_{10} - \frac{\lambda_1}{\theta_1} \right) + \tilde{c}_2 \left(\frac{\lambda_2 - \mu_2}{\theta_2} \right) - \theta_2 C_2' \left(n_{20} + \frac{\mu_2 - \lambda_2}{\theta_2} \right). \tag{A.8}$$

If $t \in (0, t_2]$, then $s_1^*(t) = 1$ and $s_2^*(t) = 0$. Hence,

$$n_1^*(t) = \left(n_{10} + \frac{\mu_1 - \lambda_1}{\theta_1}\right) \mathrm{e}^{-\theta_1 t} + \frac{\lambda_1 - \mu_1}{\theta_1}, \qquad n_2^*(t) = \left(n_{20} - \frac{\lambda_2}{\theta_2}\right) \mathrm{e}^{-\theta_2 t} + \frac{\lambda_2}{\theta_2},$$

so that for $t \in (0, t_2]$ we have

$$\begin{aligned}
\mathcal{H}(n^*(t), s^*(t), \gamma^*(t), t) &= \tilde{c}_1 \left(\left(n_{10} + \frac{\mu_1 - \lambda_1}{\theta_1}\right) \mathrm{e}^{-\theta_1 t} + \frac{\lambda_1 - \mu_1}{\theta_1}\right) + \tilde{c}_2 \left(\left(n_{20} - \frac{\lambda_2}{\theta_2}\right) \mathrm{e}^{-\theta_2 t} + \frac{\lambda_2}{\theta_2}\right) \\
&\quad + \left(C_1' \mathrm{e}^{\theta_1 t} + \frac{\tilde{c}_1}{\theta_1}\right) \left(\lambda_1 - \mu_1 - \theta_1 \left(\left(n_{10} + \frac{\mu_1 - \lambda_1}{\theta_1}\right) \mathrm{e}^{-\theta_1 t} + \frac{\lambda_1 - \mu_1}{\theta_1}\right)\right) \\
&\quad + \left(C_2' \mathrm{e}^{\theta_2 t} + \frac{\tilde{c}_2}{\theta_2}\right) \left(\lambda_2 - \theta_2 \left(\left(n_{20} - \frac{\lambda_2}{\theta_2}\right) \mathrm{e}^{-\theta_2 t} + \frac{\lambda_2}{\theta_2}\right)\right) \\
&= \tilde{c}_1 \left(\frac{\lambda_1 - \mu_1}{\theta_1}\right) + \tilde{c}_2 \frac{\lambda_2}{\theta_2} - \theta_1 C_1' \left(n_{10} + \frac{\mu_1 - \lambda_1}{\theta_1}\right) - \theta_2 C_2' \left(n_{20} - \frac{\lambda_2}{\theta_2}\right). \qquad \text{(A.9)}
\end{aligned}$$

Setting (A.8) and (A.9) equal to 0, we obtain the following expressions:

$$\begin{aligned}
C_1' &= \frac{\tilde{c}_1 \left(\frac{\lambda_1 - \mu_1}{\theta_1}\right) + \tilde{c}_2 \frac{\lambda_2}{\theta_2} - \theta_2 C_2' \left(n_{20} - \frac{\lambda_2}{\theta_2}\right)}{\theta_1 \left(n_{10} + \frac{\mu_1 - \lambda_1}{\theta_1}\right)}, \\
C_2' &= \frac{\left(n_{10} - \frac{\lambda_1}{\theta_1}\right) \left(\tilde{c}_1 \frac{\mu_1}{\theta_1} - \tilde{c}_2 \frac{\mu_2}{\theta_2}\right) + \left(\tilde{c}_1 \frac{\lambda_1}{\theta_1} + \tilde{c}_2 \frac{(\lambda_2 - \mu_2)}{\theta_2}\right) \frac{\mu_1}{\theta_1}}{\theta_2 \left(n_{20} + \frac{\mu_2 - \lambda_2}{\theta_2}\right) \frac{\mu_1}{\theta_1} + \mu_2 \left(n_{10} - \frac{\lambda_1}{\theta_1}\right)}. \qquad \text{(A.10)}
\end{aligned}$$

If $t \in (t_2, t_3]$, then $s_1^*(t) = \rho_1$ and $s_2^*(t) = 1 - \rho_1$. Hence,

$$\begin{aligned}
n_1^*(t) &= 0, \\
n_2^*(t) &= \left(n_{20} - \frac{\lambda_2}{\theta_2} + \frac{\mu_2 \mu_1 - \lambda_1 \mu_2}{\mu_1 \theta_2} \left(\frac{\mu_1 - \lambda_1}{n_{10}\theta_1 - \lambda_1 + \mu_1}\right)^{-\frac{\theta_2}{\theta_1}}\right) \mathrm{e}^{-\theta_2 t} - \left(\frac{\mu_2 \mu_1 - \lambda_1 \mu_2 - \lambda_2 \mu_1}{\mu_1 \theta_2}\right),
\end{aligned}$$

so that for $t \in (t_2, t_3]$ we have

$$\begin{aligned}
\mathcal{H}(n^*(t), s^*(t), \gamma^*(t), t) &= \tilde{c}_2 n_2^*(t) + \left(C_2' \mathrm{e}^{\theta_2 t} + \frac{\tilde{c}_2}{\theta_2}\right) \left(\lambda_2 - \mu_2 \left(1 - \frac{\lambda_1}{\mu_1}\right) - \theta_2 n_2^*(t)\right) \\
&= -\frac{\tilde{c}_2 \mu_2}{\theta_2} (1 - \rho_1 - \rho_2) - \theta_2 C_2' \\
&\quad \times \left(n_{20} - \frac{\lambda_2}{\theta_2} + \frac{\mu_2 \mu_1 - \lambda_1 \mu_2}{\mu_1 \theta_2} \left(\frac{\mu_1 - \lambda_1}{n_{10} - \lambda_1 + \mu_1}\right)^{-\frac{\theta_2}{\theta_1}}\right). \qquad \text{(A.11)}
\end{aligned}$$

Setting Eq. (A.11) equal to 0, and using Eq. (A.10), we obtain that a state on the switching curve satisfies the following relation:

$$n_{20} = \frac{a_1 n_{10} + a_2 + (a_3 n_{10} - a_2) \left(\frac{\theta_1 n_{10} + \mu_1 - \lambda_1}{\mu_1 - \lambda_1}\right)^{\frac{\theta_2}{\theta_1}}}{a_4 n_{10}} + \frac{\lambda_2}{\theta_2}, \qquad \text{(A.12)}$$

where

$$\begin{aligned}
a_1 &= \tilde{c}_2 \frac{\mu_2}{\theta_2} (1 - \rho_1 - \rho_2); \qquad a_2 = a_1 \frac{\mu_1}{\theta_1} (1 - \rho_1); \\
a_3 &= \left(\tilde{c}_1 \frac{\mu_1}{\theta_1} - \tilde{c}_2 \frac{\mu_2}{\theta_2}\right) (1 - \rho_1); \qquad a_4 = -\left(\tilde{c}_1 \frac{\mu_1}{\theta_1} - \tilde{c}_2 \frac{\mu_2}{\theta_2}\right) \frac{\theta_2}{\mu_2}.
\end{aligned}$$

A switch will only appear if (n_{10}, n_{20}) is positive. Since Eq. (A.12) is decreasing in n_{10}, this is equivalent to Eq. (A.12) to be positive in the point $n_{10} = 0$. Using l'Hopital we obtain

$$n_{20} \xrightarrow{n_{10} \to 0} \frac{a_1}{a_4} + \frac{\lambda_2}{\theta_2} + \frac{a_3}{a_4} - \frac{a_2 \theta_2}{a_4 \mu_1 (1 - \rho_1)} = (1 - \rho_1 - \rho_2) \frac{\mu_2}{\theta_1 \theta_2} \left(\frac{\tilde{c}_1 \mu_1 - \tilde{c}_2 \mu_2}{\frac{\tilde{c}_2 \mu_2}{\theta_2} - \frac{\tilde{c}_1 \mu_1}{\theta_1}}\right).$$

Since we assumed that the system is in under-load ($\rho_1 + \rho_2 < 1$), we have

$$(1-\rho_1-\rho_2)\frac{\mu_2}{\theta_1\theta_2}\left(\frac{\tilde{c}_1\mu_1-\tilde{c}_2\mu_2}{\frac{\tilde{c}_2\mu_2}{\theta_2}-\frac{\tilde{c}_1\mu_1}{\theta_1}}\right)\geq 0 \Longleftrightarrow \begin{cases}\tilde{c}_2\mu_2/\theta_2 > \tilde{c}_1\mu_1/\theta_1 & \text{and} \quad \tilde{c}_1\mu_1 \geq \tilde{c}_2\mu_2,\\ \tilde{c}_1\mu_1/\theta_1 > \tilde{c}_2\mu_2/\theta_2 & \text{and} \quad \tilde{c}_2\mu_2 \geq \tilde{c}_1\mu_1.\end{cases}$$

Therefore, the condition so that a switch of priority occurs is that the $\tilde{c}\mu$ and the $\tilde{c}\mu/\theta$ have the opposite ordering. This proves the result. □

References

[1] P. Brill, M. Posner, Level crossings in point processes applied to queues: single-server case, Operations Research 25 (1977) 662–674.
[2] F. Baccelli, P. Boyer, G. Hebuterne, Single-server queues with impatient customers, Advances in Applied Prabability 16 (1984) 887–905.
[3] A. Brandt, M. Brandt, On the two-class M/M/1 system under preemptive resume and impatience of the prioritized customers, Queueing Systems 47 (2004) 147–168.
[4] C. Gromoll, P. Robert, B. Zwart, Fluid limits for processor sharing queues with impatience, Mathematics of Operations Research 33 (2008) 375–402.
[5] F. Iravani, B. Balcioğlu, On priority queues with impatient customers, Queueing Systems 58 (2008) 239–260.
[6] B. Ata, M. Tongarlak, On scheduling a multiclass queue with abandonments under general delay costs, Queueing Systems (2012) 1–40.
[7] O. Boxma, P. de Waal, Multiserver queues with impatient customers, in: Proceedings of ITC-14, 1994, pp. 743–756.
[8] N. Boots, H. Tijms, A multiserver queueing systems with impatient customers, Management Science 45 (1999) 444–448.
[9] W. Whitt, Efficiency-driven heavy-traffic approximations for many-server queues with abandonments, Management Science 50 (2004) 1449–1461.
[10] J. Dai, S. He, Many-server queues with customer abandonment: a survey of diffusion and fluid approximations, Journal of Systems Science and Systems Engineering 21 (2012) 1–36.
[11] K. Glazebrook, P. Ansell, R. Dunn, R. Lumley, On the optimal allocation of service to impatient tasks, Journal of Applied Probability 41 (2004) 51–72.
[12] R. Atar, A. Mandelbaum, M. Reiman, Scheduling a multi-class queue with many exponential servers: asymptotic optimality in heavy-traffic, The Annals of Applied Probability 14 (2004) 1084–1134.
[13] J. Harrison, A. Zeevi, Dynamic scheduling of a multiclass queue in the Halfin–Whitt heavy traffic regime, Operations Research 52 (2004) 243–257.
[14] N. Argon, S. Ziya, R. Righter, Scheduling impatient jobs in a clearing system with insights on patient triage in mass-casualty incidents, Probability in the Engineering and Informational Sciences 22 (2010) 301–332.
[15] R. Atar, C. Giat, N. Shimkin, The $c\mu/\theta$ rule for many-server queues with abandonment, Operations Research 58 (2010) 1427–1439.
[16] D. Down, G. Koole, M. Lewis, Dynamic control of a single server system with abandonments, Queueing Systems 67 (2011) 63–90.
[17] R. Atar, C. Giat, N. Shimkin, On the asymptotic optimality of the $c\mu/\theta$ rule under ergodic cost, Queueing Systems 67 (2011) 127–144.
[18] U. Ayesta, P. Jacko, V. Novak, A nearly-optimal index rule for scheduling of users with abandonment, in: IEEE Infocom 2011, 2011.
[19] P. Whittle, Restless bandits: activity allocation in a changing world, Journal of Applied Probability 25 (1988) 287–298.
[20] O. Garnett, A. Mandelbaum, M. Reiman, Designing a call center with impatient customers, Manufacturing and Service Operations Management 4 (2002) 208–227.
[21] C. Buyukkoc, P. Varaya, J. Walrand, The $c\mu$ rule revisited, Advances in Applied Probability 17 (1985) 237–238.
[22] D. Bertsekas, Dynamic Programming and Optimal Control, Vol. II, Athena Scientific, 2007.
[23] F. Avram, D. Bertsimas, M. Richard, Stochastic networks, in: Proceedings of the IMA, 1994, pp. 199–234.
[24] G. Weiss, On optimal draining of reentrant fluid lines, in: F.P. Kelly, R.J. Williams (Eds.), Stochastic Networks, 1995, pp. 91–103.
[25] N. Bäuerle, U. Rieder, Optimal control of single-server fluid networks, Queueing Systems 35 (2000) 185–200.
[26] S. Meyn, Stability and optimization of queueing networks and their fluid models, in: Mathematics of Stochastic Manufacturing Systems, in: Lectures in Applied Mathematics, vol. 33, 1997, pp. 175–199.
[27] N. Bäuerle, Asymptotic optimality of tracking policies in stochastic networks, The Annals of Applied Probability 10 (2000) 1065–1083.
[28] A. Gajrat, A. Hordijk, Fluid approximation of a controlled multiclass tandem network, Queueing Systems 35 (2000) 349–380.
[29] C. Maglaras, Discrete-review policies for scheduling stochastic networks: trajectory tracking and fluid-scale asymptotic optimality, The Annals of Applied Probability 10 (2000) 897–929.
[30] I.M. Verloop, R. Núñez-Queija, Asymptotically optimal parallel resource assignment with interference, Queueing Systems 65 (2010) 43–92.
[31] J. Dai, On positive harris recurrence of multiclass queueing networks: a unified approach via fluid limit models, The Annals of Applied Probability 5 (1995) 49–77.
[32] P. Robert, Stochastic Networks and Queues, Springer-Verlag, 2003.
[33] M.L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, John Wiley & Sons, 2005.
[34] B.C. Chachuat, Nonlinear and dynamic optimization: from theory to practice, IC-32: Winter Semester, 2006–2007.

Maialen Larrañaga is a Ph.D. student in CNRS-LAAS and INP-ENSEEIHT in Toulouse since October 2012 under the supervision of Urtzi Ayesta and Ina Maria Verloop. She received her Master degree in Mathematics in September 2012 from the University of the Basque Country (UPV/EHU). The research for her Master thesis was carried out at the Basque Center for Applied Mathematics (BCAM), Derio, Spain. Her main research interests are: optimal control theory, dynamic programming, stochastic processes, scaling methods, with applications into communication networks.

Urtzi Ayesta received a Ph.D. degree from Université de Nice-Sophia Antipolis (France), a M.S. degree in Electrical Engineering from Columbia University (US) and a B.S./M.S. degree in Telecommunication Engineering from Nafarroako Unibertsitate Publikoa-Universidad Publica de Navarra (Spain). His Ph.D. research work was carried out at the research laboratories of INRIA Sophia-Antipolis and France Telecom R&D. He is currently a CNRS researcher working at LAAS, Toulouse, and an IKERBASQUE Research Professor in the Computer Science Faculty at the University of the Basque Country.

I.M. Verloop received the M.Sc. degree in Mathematics from Utrecht University, The Netherlands, in 2005, and the Ph.D. degree from the Mathematics and Computer Science Department of Eindhoven University of Technology, in 2009. Her Ph.D. research was carried at the Probability, Networks and Algorithms Department of the Center for Mathematics and Computer Science (CWI) in Amsterdam, The Netherlands. From 2009 to 2011 she was a postdoc at the Basque Center for Applied Mathematics, Derio, Spain. Since October 2011 she has been a CNRS researcher at IRIT, Toulouse, France. Her research interests are in the performance analysis of communication networks, scheduling and queueing theory.