Towards an Erlang formula for multiclass networks

Matthieu Jonckheere, Jean Mairesse

Towards an Erlang formula for multiclass networks

Matthieu Jonckheere* Jean Mairesse†

Abstract

Consider a multiclass stochastic network with state dependent service rates and arrival rates describing bandwidth-sharing mechanisms as well as admission control and/or load balancing schemes. Given Poisson arrival and exponential service requirements, the number of customers in the network evolves as a multi-dimensional birth-and-death process on a finite subset of $\mathbb{N}^k$. We assume that the death (i.e. service) rates and the birth rates depending on the whole state of the system, satisfy a local balance condition. This makes the resulting network a so-called Whittle network and the stochastic process describing the state of the network is reversible with an explicit stationary distribution that is in fact insensitive to the service time distribution. Given routing constraints, we are interested in the optimal performance of such networks. We first construct bounds for generic performance criteria, that can be evaluated using recursive procedures, these bounds being attained in the case of a unique arrival process. We then study the case of several arrival processes, focusing in particular on the instance with admission control only. Building on convexity properties, we characterize the optimal policy, and give criteria on the service rates for which our bounds are again attained.

1 Introduction

The Erlang formula [6] has proved to be a central tool of performance evaluation for telephone networks. Its impressive and lasting success in an engineering context can be explained by both its simplicity and its robustness [1]. The Erlang formula is insensitive to the call duration distribution and depends on a unique parameter: the traffic intensity. The only assumptions which are required to apply the formula are Poisson arrivals. At the mathematical level, the key property is the reversibility of the one dimensional birth-and-death process. On the ground of this early work of Erlang has developed the whole theory of stochastic networks whose state evolves as a reversible stochastic process. Such networks became very popular with the seminal work of Kelly [8] further developed by Whittle [15]. The crucial insensitivity property of the Erlang model extends to circuit-switched networks (without admission control and dynamic load balancing). More recently, (quasi-)reversible networks have emerged as a powerful tool to capture the essential dynamics of complex and diverse real-life telecommunication networks, see for instance [2, 3]. Because the key performance indicators

*Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands, E-mail: m.t.s.jonckheere@tue.nl,

†LIAFA, CNRS et Université Paris 7, case 7014, 2 place Jussieu, 75251 Paris Cedex 05, France. E-mail: Jean.Mairesse@liafa.jussieu.fr.

of these models are independent of all traffic characteristics beyond the traffic intensity, they provide simple and robust engineering rules, just like the Erlang formula.

The situation gets more complicated in the presence of admission control and/or load balancing. The admission control consists in possibly rejecting customers even when the full capacity of the network is not utilized. Load balancing consists in routing an arriving customer to one of the queues in a subset of possible target queues. These techniques ensure an efficient utilization of resources by taking into account the system state to make decisions. They have become a key component of computer and communication systems.

The corresponding mathematical model can be described by a general multi-dimensional birth-and-death process on a finite subset of $\mathbb{N}^{\mathcal{I}}$, where $\mathcal{I}$ is the finite set of nodes (queues) of the network. The death (i.e. service) transition rates depend on the whole state of the system. Defining an admission control and/or a load balancing policy consists in tuning the birth (i.e. arrival) transition rates. More precisely, let us concentrate on one given class of customers arriving at rate ν. Admission control consists of defining, as a function of the state of the system x, the rate $\lambda(x) \leq \nu$ at which the customers are admitted in the network. Assume that for this class, the arriving customers can be served by any of the nodes in a subset $\mathcal{I}'$ of $\mathcal{I}$. Load balancing (aka routing) consists in choosing the rates $\lambda_i(x), i \in \mathcal{I}', \sum_i \lambda_i(x) = \lambda(x)$, at which the admitted customers are assigned to the different nodes.

The goal is to find the policy which optimizes some performance criterion, typically the blocking probability of an arriving customer. This leads in many cases to very difficult optimization problems. A discussion of some existing results appears in [2], see also the references therein. The book of Ross [13] is also devoted to a special instance of this question: the model considered, known as the stochastic knapsack model, is a model with admission control but without routing (each class of customer is assigned to a specific and different node). Partial optimality results in dimension 2 and 3 are obtained, for service rates depending only on the local number of customers at the node.

In the present paper, we restrict our attention to networks where the service rates satisfy the so-called "balance condition" and we consider the optimization problem within the restricted subclass of policies for arrivals which satisfy an analogous "balance condition", the so-called *insensitive policies*. Such policies are called insensitive because the resulting stochastic network is insensitive in the sense that its stationary distribution depends on the service time distribution only via its mean.

So the questions are three-fold: (1) Is it possible to find the optimal insensitive policy ? (2) Can we efficiently evaluate the optimal insensitive policy ? (3) Is the performance of the best insensitive policy close to the one of the best policy ? Here we address only the first two questions, having in mind the hope that the answer to the third question is yes for a broad range of parameters. In any case, the optimal performance of insensitive policies provides a bound for the optimal performance of general policies. This may bring enough motivation for studying the former.

Since these questions are difficult to answer in general, we first focus on policies with an admission region shaped as a rectangular hyper-parallelepiped and give efficient recursive formulas to evaluate their performance. This provides computable lower and upper bounds

for the performance of the optimal insensitive policy.

We then study the tightness of our bounds. When there is a unique arrival process, our bounds are always attained, i.e., the optimal insensitive policy can be described as a specific rectangular shaped policy.

Later on, we turn our attention to the case of several arrival processes and consider as its simplest instance, the case of networks with admission control only. This problem is of crucial importance when looking at the flow level modelling of fixed and wireless data networks [12]. We assume that the state space is coordinate-convex (see Figure 1). A *coordinate-convex* policy is a policy with full admission within a coordinate-convex subset of the state space, and full rejection out of it (see Figure 1). We prove that these policies are extremal within insensitive policies: any insensitive policy can be decomposed as a convex combination of coordinate-convex policies. Using this result, we prove that the minimal blocking probability for insensitive policies is reached by a coordinate-convex policy.

We define a *decentralized* policy to be a coordinate-convex policy in which the admission region is the intersection of the whole state space with a rectangular hyper-parallelepiped. On Figure 1, the left case corresponds to a non-decentralized coordinate-convex policy, while the other two cases correspond to decentralized policies.

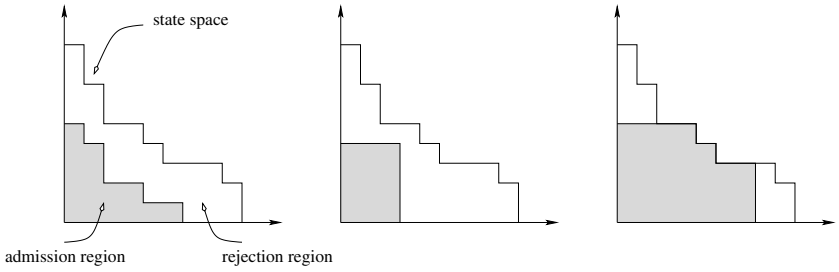

Figure 1: Coordinate-convex policy (left) and decentralized policies (middle and right)

Our results show that in general, decentralized policies are not optimal within insensitive policies. In other terms, there exist instances in which a coordinate-convex policy achieves a blocking probability which is strictly smaller than the one achieved by the best decentralized policy, which contrasts with the case of a unique arrival process. We provide a toy example of a network in which this phenomenon occurs. This emphasizes the intrinsic increased complexity of models where several classes of customers are competing. To give a complete picture, we also provide sufficient conditions on the model (monotonicity or light traffic) under which complete sharing policies (decentralized policies admitting all the traffic inside the state space) are optimal among insensitive policies.

The paper is organized as follows. Section 2 introduces the model, the notation and the objectives. In Section 3, we provide computable bounds for a broad set of performance criteria in the general case. Section 4 is a detailed analysis of networks with admission control only. In Section 5, we illustrate the concepts and the results on a toy example, which can be described as the simplest non trivial multiclass model.

2 General framework

Notation. Let e_i be the point of $\mathbb{N}^k$ defined by $(e_i)_i = 1,\ (e_i)_j = 0, j \neq i$. A *Ferrers set* is a finite subset E of $\mathbb{N}^k$ such that:

$$\big[x \in E, x_i > 0\big] \implies x - e_i \in E\,.$$

In other words, it is a coordinate-wise convex finite set. The set of all Ferrers sets is denoted by $\mathcal{F}(\mathbb{N}^k)$. The set of Ferrers sets included in $E \subset \mathbb{N}^k$ is denoted by $\mathcal{F}(E)$. The notation $x \leq y$ is used for the coordinate-wise ordering: $\forall i,\ x_i \leq y_i$. We further denote:

$$|x| = \sum_{i=1}^{k} x_i \quad \text{and} \quad \binom{|x|}{x} = \frac{|x|!}{x_1! \dots x_k!}.$$

We denote by $\mathbf{1}_S$ the indicator function of S, that is the map taking value 1 inside S and 0 outside. We denote respectively by $\mathbb{R}_+$ and $\mathbb{R}_+^*$ the set of non-negative and positive reals.

2.1 Model

Consider a network with a finite set of servers (nodes) $\mathcal{I}$. An arriving customer is served by one of the nodes and then leaves the network. More precisely, $\mathcal{I}$ is partitioned into finitely many non-empty subsets $\mathcal{I}_k, k \in \mathcal{K}$, each customer has a class which is an element of $\mathcal{K}$, and a customer of class k has to be served by one of the nodes in $\mathcal{I}_k$. The state of the system is described by a vector in $\mathbb{N}^{\mathcal{I}}$ corresponding to the number of customers in each node. The state space, denoted by $\mathcal{X}$, is assumed to be a Ferrers set:

$$\mathcal{X} \in \mathcal{F}(\mathbb{N}^{\mathcal{I}})\,.$$

Class-k customers arrive according to a Poisson process of intensity ν_k. The different arrival processes are mutually independent. An arriving customer of class k is either routed to a node in $\mathcal{I}_k$ or rejected (recall that the state space is finite). The routing/admission policy depends on the whole state of the system at the instant of arrival. The service requirements of the customers are independent and exponentially distributed with parameter 1. The service rate at a given node depends on the whole state of the system.

Let X be the stochastic process valued in $\mathcal{X}$ describing the state of the system. The above assumptions result in X being a continuous-time jump Markov process, on the state space $\mathcal{X}$, with infinitesimal generator $Q = (q(x,y))_{x,y\in\mathcal{X}}$ given by: $\forall x \in \mathcal{X}$,

$$\begin{cases} q(x, x - e_i) \;=\; \phi_i(x) & \text{if } x - e_i \in \mathcal{X} \\ q(x, x + e_i) \;=\; \lambda_i(x) & \text{if } x + e_i \in \mathcal{X} \\ q(x,y) \;=\; 0 & \text{if } y \in \mathcal{X},\ y \neq x - e_i, x + e_i\,. \end{cases} \tag{1}$$

It is convenient to define ϕ_i, λ_i, for all x in $\mathbb{N}^{\mathcal{I}}$, so we set $\phi_i(x) = 0$ if $(x - e_i) \notin \mathcal{X}$, and $\lambda_i(x) = 0$ if $(x + e_i) \notin \mathcal{X}$. The scalars $\lambda_i(x)$ define the routing/admission policy. By definition of the model, they satisfy:

$$\text{Intensity constraints:} \quad \forall k \in \mathcal{K}, \quad \textstyle\sum_{i\in\mathcal{I}_k} \lambda_i(x) \leq \nu_k \tag{2}$$

By definition, the rejection rate of class-k customers in state x is equal to $\nu_k - \sum_{i\in\mathcal{I}_k} \lambda_i(x)$. The *intensity* $h : \mathcal{X} \longrightarrow \mathbb{R}_+^*$ of a routing is defined by

$$h(x) = \sum_{i\in\mathcal{I}} \lambda_i(x)\,. \tag{3}$$

The *maximum routing intensity* $\nu : \mathcal{X} \longrightarrow \mathbb{R}_+^*$ is defined by

$$\nu(x) = \sum_{k\in\mathcal{K}} \nu_k \mathbf{1}_{\{\exists i\in\mathcal{I}_k, x+e_i\in\mathcal{X}\}}\,. \tag{4}$$

Clearly, the intensity h of any routing satisfies: $\forall x \in \mathcal{X},\ h(x) \le \nu(x)$.

Concerning the service rates, we make the following assumption:

$$\forall i \in \mathcal{I}, \forall x \in \mathcal{X}, x_i > 0, \qquad \phi_i(x) > 0\,. \tag{5}$$

In the paper, the service rates $\phi_i(x)$ are assumed to be given and fixed. On the other hand, the routing rates $\lambda_i(x)$ will vary, and the actual state space of the process X will depend on this. For some choices of the routing rates, Q will not be irreducible. However, according to (5), the set of recurrent states of Q is always strongly connected, and belongs to $\mathcal{F}(\mathcal{X})$. In particular, there exists a unique stationary distribution for X that we denote by π.

Performance criterion. Below, the goal is to choose the routing rates in order to optimize a performance criterion. This criterion can be typically chosen as a given linear combination of the blocking probabilities of each class of customers. Consider $p = (p_k)_{k\in\mathcal{K}}, p_k > 0, \sum_k p_k = 1$. Set:

$$B_p = \sum_{x\in\mathcal{X}} \pi(x) \sum_{k\in\mathcal{K}} p_k \left(1 - \frac{\sum_{i\in\mathcal{I}_k} \lambda_i(x)}{\nu_k}\right). \tag{6}$$

Using the PASTA property, the blocking probability B of an arriving customer is a special case of this generic criterion: $B = B_{(\nu_k/\bar{\nu})_k}$, where $\bar{\nu} = \sum_k \nu_k$. When the routing policy is defined via a balance function Λ, we denote the associated blocking probability by $B_p(\Lambda)$.

We also give results which are valid for any criterion of the form $E[f(X)]$, for a given $f : \mathcal{X} \to \mathbb{R}$.

2.2 Balanced services and routing

The service rates are said to be *balanced* if there exists $\Phi : \mathcal{X} \to \mathbb{R}_+^*$ such that

$$\forall i,\ \forall x \in \mathcal{X}, x_i > 0, \quad \phi_i(x) = \frac{\Phi(x - e_i)}{\Phi(x)}\,. \tag{7}$$

Consider the following property:

$$\forall i, j, \quad \forall x \in \mathcal{X}, x_i > 0, x_j > 0, \qquad \phi_i(x)\phi_j(x - e_i) = \phi_j(x)\phi_i(x - e_j)\,. \tag{8}$$

Clearly (7) implies (8). Conversely, assume that (8) holds. Set $\mathbf{o} = (0, \ldots, 0)$. Consider a directed path from $\mathbf{o}$ to x in $\mathbb{N}^{\mathcal{I}}$, that is a sequence of points $(x_0 = \mathbf{o}, x_1, \ldots, x_n = x)$, such that $x_k - x_{k-1} = e_{i_k}$, $\forall 1 \leq k \leq n$. Define $\Phi : \mathcal{X} \to \mathbb{R}_+^*$ by the formula:

$$\Phi(x) = C\big[\phi_{i_1}(x_1)\phi_{i_2}(x_2)\cdots\phi_{i_n}(x_n)\big]^{-1} .$$

where C is some positive constant. Using (8), we can prove that the value of $\Phi(x)$ does not depend on the chosen directed path from $\mathbf{o}$ to x. Formula (7) follows readily.

Similarly, the routing rates are *balanced* if the following equivalent conditions are satisfied:

$$\begin{aligned} &\exists \Lambda : \mathcal{X} \to \mathbb{R}_+^*, \qquad \forall i,\ \forall x \in \mathcal{X}, x + e_i \in \mathcal{X}, \quad \lambda_i(x) = \Lambda(x + e_i)/\Lambda(x) . \\ &\forall i, j,\ \forall x \in \mathcal{X}, x + e_i \in \mathcal{X}, x + e_j \in X, \quad \lambda_i(x)\lambda_j(x + e_i) = \lambda_j(x)\lambda_i(x + e_j) . \end{aligned} \tag{9}$$

It is often convenient to define Λ or Φ on $\mathbb{N}^{\mathcal{I}}$ instead of $\mathcal{X}$. Of course, the actual rates λ_i and ϕ_i depend only on the restrictions of Λ and Φ to $\mathcal{X}$.

Assume that both the service and the routing rates are balanced. Using (7) and (9), we get, $\forall x \in \mathcal{X}, x + e_i \in \mathcal{X}$,

$$\Phi(x)\Lambda(x)\lambda_i(x) = \Phi(x + e_i)\Lambda(x + e_i)\phi_i(x + e_i) .$$

Since $\lambda_i(x) = q(x, x + e_i)$ and $\phi_i(x + e_i) = q(x + e_i, x)$, we conclude that the process X is reversible and that the stationary distribution is given by

$$\pi(x) = \frac{\Phi(x)\Lambda(x)}{\sum_{y\in\mathcal{X}} \Phi(y)\Lambda(y)} . \tag{10}$$

The balance conditions are obviously restrictive conditions. For instance, (9) is not satisfied for usual routing policies such as "join-the-shortest-queue".

The balance condition goes back to Kelly [8] (the slightly different concept of job-balance was developed by Hordijk et al [7]). Balance is satisfied for example if the rates at node i depend on x only via x_i or if the service capacities are fairly shared between classes i.e., $\phi_i(x) = c\frac{x_i}{|x|}$, or more generally in a bandwidth sharing network operated under the balanced fairness allocation [5].

It was recently proved that balance conditions are closely related to insensitivity. Assuming that the service rates are balanced, a necessary and sufficient condition for the insensitivity property to hold is that the routing rates be balanced, see [14, 4].

In the sequel, we always assume that the service rates **and** the routing rates are balanced which implies the insensitivity of the studied network. We shall then speak of *insensitive policies* to refer to those policies with balanced service and routing rates. The model defined in Section 2.1 corresponds, at the network level, to exponentially distributed service times. However, since we restrict our attention to insensitive policies, all the results remain true for i.i.d. generally distributed service times.

2.3 Admissible balance functions

From now on, assume that the service rates $(\phi(x))_{x\in\mathcal{X}}$ are balanced (by a balance function Φ) and fixed.

Let $\mathcal{A}$ be the set of normalized balance functions which satisfy the routing constraints, i.e.,

$$\mathcal{A} = \{\Lambda : \mathbb{N}^{\mathcal{I}} \to \mathbb{R}^+, \quad \forall x \notin \mathcal{X},\ \Lambda(x) = 0\ , \quad \sum_{x\in\mathcal{X}} \Lambda(x)\Phi(x) = 1\ , \tag{11}$$

$$\forall x, \forall i \in \mathcal{K},\ \sum_{j\in\mathcal{I}_i} \Lambda(x+e_j) \le \nu_i \Lambda(x)\}\ . \tag{12}$$

Such balance functions are called *admissible*. To each admissible balance function, we can associate the routing rates defined by (9). Below, we often identify an admissible balance function with the routing policy it defines.

In the following, it will often be convenient to relax the normalization condition (11), when considering a balance function. To differentiate between both cases, we shall use the notation $\tilde{\Lambda}$ instead of Λ when a balance function is not normalized.

We can characterize the structure of $\mathcal{A}$ as follows.

Proposition 2.1. *The set $\mathcal{A}$ is convex and compact (for the topology of pointwise convergence). The blocking probability B_p is a linear function on $\mathcal{A}$:*

$$B_p(\Lambda) = 1 - \sum_{x\in\mathcal{X}} \Phi(x) \sum_{i\in\mathcal{K}} \frac{p_i}{\nu_i} \sum_{j\in\mathcal{I}_i} \Lambda(x+e_j)\ .$$

Proof. From the constraints (11) and (12), it is easily checked that if $\Lambda 1$ and $\Lambda 2$ belong to $\mathcal{A}$, so does $t\Lambda 1 + (1-t)\Lambda 2$ for $t \in [0,1]$. Since $\mathcal{A}$ is a set of bounded functions with finite support $\mathcal{X}$, it is compact. Consider now a balance function Λ. Expressing the routing rates as a function of Λ and using (10), we can rewrite equation (6) as:

$$B_p(\Lambda) \;=\; \sum_{x\in\mathcal{X}} \Lambda(x)\Phi(x) \sum_{k\in\mathcal{K}} p_k \Big(1 - \frac{1}{\nu_k} \sum_{i\in\mathcal{I}_k} \frac{\Lambda(x+e_i)}{\Lambda(x)}\Big) = 1 - \sum_{x\in\mathcal{X}} \Phi(x) \sum_{k\in\mathcal{K}} \frac{p_k}{\nu_k} \sum_{i\in\mathcal{I}_k} \Lambda(x+e_i)\ .$$

As a consequence, $B_p(t\Lambda 1 + (1-t)\Lambda 2) = tB_p(\Lambda 1) + (1-t)B_p(\Lambda 2)$. □

3 Rectangular balance functions and performance bounds

In this section, we focus on balance functions characterized by an admission region reduced to the intersection of $\mathcal{X}$ with a rectangular hyper-parallelepiped $y^{\downarrow} = \{x \le y\}$. We show how to use these balance functions to derive computable lower and upper bounds for the performance of the optimal insensitive policy.

3.1 Rectangular balance function

Definition 3.1. *Consider a point $y \in \mathcal{X}$ and a function $g : \mathcal{X} \to \mathbb{R}_+^*$ such that $g \leq \nu$, where ν is the maximum routing intensity defined in (4). The* rectangular *balance function $\tilde{\Lambda}^{y,g} : \mathbb{N}^{\mathcal{I}} \to \mathbb{R}_+$ associated with y and g is defined by:*

$$\tilde{\Lambda}^{y,g}(x) = \begin{cases} 1 & \text{if } x = y \\ g(x)^{-1} \sum_{i \in \mathcal{I}} \tilde{\Lambda}^{y,g}(x + e_i) & \text{if } x \leq y, x \neq y \\ 0 & \text{otherwise} \end{cases} .$$

Any admissible balance function defined on $\mathcal{X} \cap y^{\downarrow}$ can be represented as a rectangular balance function by choosing g appropriately. This can be shown recursively, starting with the extremal points of the state space, that is: $x \in \mathcal{X} \cap y^{\downarrow},\ \forall i,\ x + e_i \notin \mathcal{X} \cap y^{\downarrow}$.

On the other hand, a normalized rectangular balance function $\Lambda^{y,g}$ is not necessarily admissible. Consider the rates $\lambda_i(x)$ associated with $\Lambda^{y,g}$ and defined according to (9). We have $\sum_{i \in \mathcal{I}} \lambda_i(x) = g(x) \leq \nu(x)$. However, the rates may or may not satisfy the intensity constraints (2). This is an important point, so we illustrate it with an example.

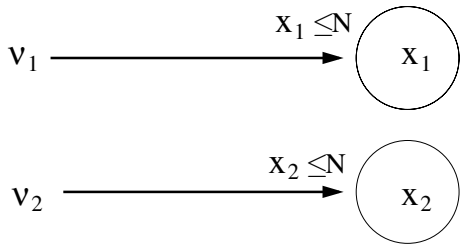

Figure 2: Network of example 1 : the admission constraints are $x_1 \leq N$, $x_2 \leq N$

Example 1. Let $\mathcal{I} = \mathcal{K} = \{1, 2\}$ with $\mathcal{I}_1 = \{1\}$ and $\mathcal{I}_2 = \{2\}$. Let the arrival rates be $\nu_1 = \nu_2 = 1/2$. The routing constraints are given by:

$$\lambda_1(x) \leq 1/2,$$

$$\lambda_2(x) \leq 1/2,$$

which for balance functions boil down to:

$$\Lambda(x) \geq 2\Lambda(x + e_1),$$

$$\Lambda(x) \geq 2\Lambda(x + e_2).$$

Consider the state space $\mathcal{X} = \{(x_1, x_2),\ x_1 \leq N,\ x_2 \leq N\}$. So the maximum routing intensity is

$$\nu(x) = \begin{cases} 1 & \text{for } x_1 < N, x_2 < N \\ 1/2 & \text{for } x_1 < N, x_2 = N \ \text{ or } \ x_1 = N, x_2 < N \\ 0 & \text{for } x_1 = N, x_2 = N \end{cases} .$$

Consider the point $y = (n, n)$ with $n < N$, and the function $g = \nu$. The corresponding rectangular balance function is:

$$\tilde{\Lambda}^{y,\nu}: \quad \begin{array}{cccccc} 1 & \cdots & 1 & 1 & 1 & 1 \\ \vdots & & 4 & 3 & 2 & 1 \\ \vdots & & 10 & 6 & 3 & 1 \\ \vdots & & 20 & 10 & 4 & 1 \\ \vdots & & & & & \vdots \\ \cdot & \cdots & \cdots & \cdots & \cdots & 1 \end{array} \tag{13}$$

The intensity constraints are not satisfied except on the diagonal $\{(i,i),\ 0 \le i \le n-1\}$. For instance, for the point $x = (n-1, n-2)$, we have

$$\lambda_1(x) = \tilde{\Lambda}^{y,\nu}(x+e_1)/\Lambda^{y,\nu}(x) = 1/3, \quad \lambda_2(x) = \tilde{\Lambda}^{y,\nu}(x+e_2)/\Lambda^{y,\nu}(x) = 2/3 >> \nu_2\,.$$

Assume now that $\nu_1 = x$ and $\nu_2 = 1-x$, for x irrational between 0 and 1. The maximum routing intensity is unchanged and equal to 1 on $y^{\downarrow}$. So the rectangular balance function associated with y and ν is still given by (13). But now the intensity constraints are violated at all the points of $y^{\downarrow}$.

We now come up with an explicit expression for rectangular balance functions. An *increasing path* from x to $y, y \ge x$, is a sequence of points $(x_0 = x, x_1, \ldots, x_k = y)$ such that $x_{j+1} - x_j = e_{i_j}$ for all j. Let $P(x,y)$ be the set of increasing paths from x to y.

Lemma 3.2. *We have*

$$\tilde{\Lambda}^{y,g}(x) = \begin{cases} 1 & \text{if } x = y \\ \sum_{p \in P(x,y)} \prod_{z \in p, z \neq y} g(z)^{-1} & \text{if } x \le y, x \neq y \\ 0 & \text{otherwise} \end{cases}. \tag{14}$$

Proof. Denote by $H(x,y)$ the right-hand side of (14). A path from x to y can be decomposed as a path from x to $x+e_i$ and a path from $x+e_i$ to y, for some i such that $x_i < y_i$. As a consequence, we have: $H(x,y) = \sum_{i: x_i < y_i} H(x+e_i, y) g(x)^{-1}$. We conclude that H and $\tilde{\Lambda}^{y,g}$ satisfy the same recursive equations and the same initial condition: $H(y,y) = 1 = \tilde{\Lambda}^{y,g}(y)$. □

3.2 Recursive evaluation of rectangular balance functions

We now show how to compute the performance of policies corresponding to rectangular balance functions.

We hence still assume that the service rates $(\phi(x))_{x \in \mathcal{X}}$ are balanced (by a balance function Φ) and fixed. We fix a rectangular balance function $\tilde{\Lambda}^{y,g}$ $(y \in \mathcal{X})$ and consider the routing rates $\lambda_i(x)$ associated with the balance function and defined according to (9).

As underlined by the Example 1, the intensity constraints may *not* be satisfied. However, let us consider the model operated under this (possibly non-admissible) policy. That is,

consider the stochastic process X with infinitesimal generator (1) for λ, ϕ. We first show how to compute the performance of the policy. Later on, we will use this to bound the performance of admissible policies.

Define:

$$C(y,g) = \sum_{x\in\mathcal{X}} \tilde{\Lambda}^{y,g}(x)\Phi(x), \qquad P^j(y,g) = \sum_{x\in\mathcal{X}} \tilde{\Lambda}^{y,g}(x+e_j)\Phi(x)\,.$$

Observe that $C(y,g)$ is the normalizing constant, that is, the stationary distribution $\pi_{y,g}$ satisfies:

$$\forall x\in\mathcal{X}, \qquad \pi_{y,g}(x) = C(y,g)^{-1}\tilde{\Lambda}^{y,g}(x)\Phi(x)\,.$$

Proposition 3.3. *Let $B_p = B_p(\Lambda^{y,g})$ be the blocking probability defined as in (6). We have*

$$B_p = 1 - \frac{\sum_{j\in\mathcal{K}} p_j\nu_j^{-1}\sum_{i\in\mathcal{I}_j} P^i(y,g)}{C(y,g)}\,. \tag{15}$$

The quantities P^j and C can be computed using the recursive schemes:

$$C(y,g) = \Phi(y) + \sum_{i\in\mathcal{I}:y_i>0} C(y-e_i,g)g(y-e_i)^{-1}, \tag{16}$$

$$P^j(y,g) = \Phi(y-e_j)\mathbf{1}_{\{y_j>0\}} + \sum_{i\in\mathcal{I}:y_i>0} P^j(y-e_i,g)g(y-e_i)^{-1}\,. \tag{17}$$

Proof. Formula (15) is obtained directly from Prop. 2.1. Let us prove (16). If $x\le y, x\neq y$, we have

$$\tilde{\Lambda}^{y,g}(x) = \sum_{i:x_i<y_i} \tilde{\Lambda}^{y-e_i,g}(x)g(y-e_i)^{-1}\,.$$

So we get

$$\begin{aligned} C(y,g) &= \sum_{x\in\mathcal{X}} \tilde{\Lambda}^{y,g}(x)\Phi(x) \\ &= \tilde{\Lambda}^{y,g}(y)\Phi(y) + \sum_{x\le y, x\neq y} \Phi(x) \sum_{i:x_i<y_i} \tilde{\Lambda}^{y-e_i,g}(x)g(y-e_i)^{-1} \\ &= \Phi(y) + \sum_{i:y_i>0} C(y-e_i,g)g(y-e_i)^{-1}\,. \end{aligned}$$

The proof of (17) is analogous. □

An equivalent recursive formula implying the service rates instead of the balance function can be considered instead of (16)-(17). Define P'_j, $j\in\mathcal{I}$, and C' such that $P'_j(y,g)/C'(y,g) = P_j(y,g)/C(y,g)$ and

$$\begin{aligned} C'(y,g) &= 1 + \sum_{i:y_i>0} C'(y-e_i,g)\phi_i(y)g(y-e_i)^{-1}, \\ P'_j(y,g) &= \phi_j(y)\mathbf{1}_{\{y_j>0\}} + \sum_{i:y_i>0} P'_j(y-e_i,g)\phi_i(y)g(y-e_i)^{-1}\,. \end{aligned}$$

Other performance criteria. Consider a performance criterion of the form $\mathcal{P}_f = E[f(X)]$ for some function $f : \mathcal{X} \to \mathbb{R}_+$.

Proposition 3.4. *We have $\mathcal{P}_f(y) = D(y,g)/C(y,g)$, where $D(y,g)$ can be evaluated using the recursive scheme:*

$$D(y,g) \quad = \quad f(y)\Phi(y) + \sum_{i\in\mathcal{I}:y_i>0} D(y-e_i,g)g(y-e_i)^{-1}\,.$$

3.3 Decentralized policies

We introduce a subclass of admissible balanced functions called decentralized balance functions. Let k be a class, define $x^{(k)}$ to be the point such that:

$$x_j^{(k)} = \begin{cases} x_j & \text{for } j \in \mathcal{I}_k \\ 0 & \text{otherwise}\,. \end{cases}$$

Decentralized balance functions correspond to policies having the desirable property that the routing intensities concerning a given class of customers depend only on the number of customers of that class present in the network.

Definition 3.5. *Consider a point $y \in \mathbb{N}^{\mathcal{I}}$, not necessarily in $\mathcal{X}$. The* decentralized *balance function $\tilde{\Lambda}^y$ associated with y is defined by:*

$$\tilde{\Lambda}^y(x) = \begin{cases} \prod_{k\in\mathcal{K}} \binom{|y^{(k)}-x^{(k)}|}{y^{(k)}-x^{(k)}}\ \nu_k^{-|y^{(k)}-x^{(k)}|} & \textit{if } x \in y^{\downarrow}\cap\mathcal{X} \\ 0 & \textit{otherwise} \end{cases}\,. \tag{18}$$

The normalized version of $\tilde{\Lambda}^y$ is easily seen to be an admissible balance function. When $y \in \mathcal{X}$, the decentralized balance function is the rectangular balance function (see Definition 3.1) associated with y and ν.

Let us define the *decentralized* routing policy associated with the decentralized balance function. The intensity function (see (3)) of the decentralized routing policy is

$$\forall x \le y, \qquad h^y(x) = \sum_{k\in\mathcal{K}} \nu_k \sum_{i\in\mathcal{I}_k} \frac{y_i - x_i}{|y^{(k)} - x^{(k)}|}\mathbf{1}_{\{x+e_i\in y^{\downarrow}\cap\mathcal{X}\}}\,.$$

The decentralized policies work as follows. Do not accept customers outside the region $y^{\downarrow}$. Inside the region $y^{\downarrow}\cap\mathcal{X}$, all possible customers are accepted, *except* at points $x \in y^{\downarrow}\cap\mathcal{X}$ such that

$$\exists k \in \mathcal{K}, \exists i,j \in \mathcal{I}_k, \qquad x+e_i \in y^{\downarrow}\cap\mathcal{X},\ x+e_j \in y^{\downarrow}\cap\mathcal{X}^c\,. \tag{19}$$

Therefore, the decentralized policy becomes particularly simple when there exists no such point. This happens in the following two cases among others:

1. When $y \in \mathcal{X}$, implying that $y^{\downarrow}\cap\mathcal{X} = y^{\downarrow}$. When $|\mathcal{K}| = 1$, the decentralized policies with $y \in \mathcal{X}$ are extremal and boil down to the *simple policies* described in [2] (see Section 3.6).

2. When $|\mathcal{I}| = |\mathcal{K}|$ (each class is assigned to a specific node). In that case, such policies are sometimes called *threshold policies* in the literature, with the point y determining the threshold. We elaborate further on this case in Section 4.

Let us illustrate the above on a few examples.

Example 2. Consider the same model as in Example 1. Consider the point $y = (n, n)$ with $n < N$. The decentralized balance function associated with this point is:

$$\tilde{\Lambda}^y : \begin{array}{cccccc} 2^n & \cdots & 8 & 4 & 2 & 1 \\ \vdots & & 16 & 8 & 4 & 2 \\ \vdots & & 32 & 16 & 8 & 4 \\ \vdots & & 64 & 32 & 16 & 8 \\ \vdots & & & & & \vdots \\ 2^{2n} & \cdots & \cdots & \cdots & \cdots & 2^n \end{array} \tag{20}$$

Here we check that we have full admission of customers strictly inside the region $y^{\downarrow} \cap \mathcal{X} = y^{\downarrow}$.

Assume now that $n > N$. We obtain $\tilde{\Lambda}^y$ by restricting (20) to the state space $\mathcal{X}$. After renormalization, we obtain exactly the same function as before. So we still have full admission strictly inside $y^{\downarrow} \cap \mathcal{X}$. More generally, it is easily seen that full admission would hold for the decentralized policy associated with any point y. This is consistent with the fact that $|\mathcal{I}| = |\mathcal{K}|$.

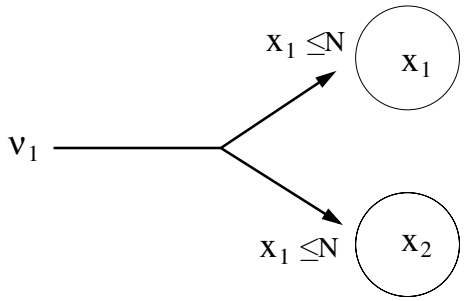

Figure 3: Network of example 3 : the admission constraints are $x_1 \leq N,\ x_2 \leq N$

Example 3. *Let $\mathcal{I} = \{1, 2\}$ and $\mathcal{K} = \{1\}$. Let the arrival rate be $\nu_1 = 1$. Consider the state space $\mathcal{X} = \{(x_1, x_2), x_1 \leq N, x_2 \leq N\}$. Choose $y = (y_1, y_2) \in \mathcal{X}$. Applying (18), we get*

$$\tilde{\Lambda}^y : \begin{array}{cccccc} 1 & \cdots & 1 & 1 & 1 & 1 \\ \vdots & & 4 & 3 & 2 & 1 \\ \vdots & & 10 & 6 & 3 & 1 \\ \vdots & & 20 & 10 & 4 & 1 \\ \vdots & & & & & \vdots \\ \cdot & \cdots & \cdots & \cdots & \cdots & 1 \end{array} \tag{21}$$

Observe that the balance function coincides with the balance function obtained in Example 1. This is natural since the balance function considered in Example 1 corresponds to the possibility of routing the total traffic indistinctively to both classes which boils down to a single-class system with two nodes.

For a point $x = (x_1, x_2)$, $x \in y^{\downarrow}$, $x \neq y$, *the routing is*

$$\lambda_1(x) = \frac{y_1 - x_1}{y_1 - x_1 + y_2 - x_2}, \quad \lambda_2(x) = \frac{y_2 - x_2}{y_1 - x_1 + y_2 - x_2}\,.$$

So we have $\lambda_1(x) + \lambda_2(x) = 1$, *and there is full admission at point* x *for* $x \in y^{\downarrow}, x \neq y$.

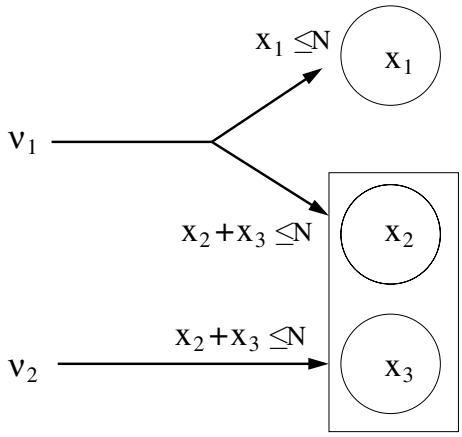

Figure 4: Network of example 4 : the admission constraints are $x_1 \leq N$, $x_2 + x_3 \leq N$

Example 4. *Let* $\mathcal{I} = \{1, 2, 3\}$ *and* $\mathcal{K} = \{1, 2\}$, $\mathcal{I}_1 = \{1, 2\}$, $\mathcal{I}_2 = \{3\}$. *Let the arrival rates be* ν_1 *and* ν_2. *Consider the state space* $\mathcal{X} = \{(x_1, x_2, x_3), x_1 \leq N, x_2 + x_3 \leq N\}$. *Choose* $y = (y_1, y_2, y_3)$. *Applying (18), we get*

$$\Lambda^y(x) = \binom{|y^{(1)} - x^{(1)}|}{y^{(1)} - x^{(1)}}\, \nu_1^{-|y^{(1)} - x^{(1)}|} \nu_2^{-y_3 + x_3}$$

For a point $x = (i, j, k) \in \mathcal{X}$, *the routing is*

$$\lambda_1(x) = \nu_1 \frac{y_1 - i}{y_1 - i + y_2 - j} \mathbf{1}_{\{i+1 \leq N\}}, \qquad \lambda_2(x) = \nu_1 \frac{y_2 - j}{y_1 - i + y_2 - j} \mathbf{1}_{\{j+k+1 \leq N\}}$$

and

$$\lambda_3(x) = \nu_2 \mathbf{1}_{\{k+1 \leq y_3\}} \mathbf{1}_{\{j+k+1 \leq N\}}\,.$$

Hence we have

$$\lambda_1(x) + \lambda_2(x) = \nu_1 \left(\frac{y_1 - i}{y_1 - i + y_2 - j} \mathbf{1}_{\{i+1 \leq N\}} + \frac{y_2 - j}{y_1 - i + y_2 - j} \mathbf{1}_{\{j+k+1 \leq N\}} \right)\,.$$

We do not have full admission at point x *when* $i + 1 \leq N$ *and* $j + k + 1 > N$.

3.4 Recursive evaluation of decentralized policies

When $y \in \mathcal{X}$, the decentralized policies are a subclass of the rectangular policies defined in Section 3.1. So the recursive formula (15) can be applied directly.

If $y \notin \mathcal{X}$, then the decentralized policy is not rectangular but is however the restriction of a rectangular policy defined on a larger state space containing both $\mathcal{X}$ and $y^{\downarrow}$. By following the exact same steps, we prove that the results of Proposition 3.3 hold in this generalized context. More precisely, we have the following.

Let $\mathcal{X}'$ be a finite Ferrers set containing both $\mathcal{X}$ and $y^{\downarrow}$. For $z \in \mathcal{X}'$, let $\tilde{\Lambda}^z$ be the decentralized balance function on $\mathcal{X}$ associated with z and defined according to (18). For $z \in \mathcal{X}'$, define:

$$C(z) = \sum_{x \in \mathcal{X}} \tilde{\Lambda}^z(x)\Phi(x), \qquad P^j(z) = \sum_{x \in \mathcal{X}} \tilde{\Lambda}^z(x + e_j)\Phi(x)\,.$$

Define $\nu' : \mathcal{X}' \longrightarrow \mathbb{R}_+^*$ by

$$\nu'(x) = \sum_{k \in \mathcal{K}} \nu_k \mathbf{1}_{\{\exists i \in \mathcal{I}_k, x + e_i \in \mathcal{X}'\}}\,.$$

Proposition 3.6. *Consider $y \notin \mathcal{X}$. The blocking probability of the decentralized policy on $\mathcal{X}$ associated with Λ^y satisfies*

$$B_p(\Lambda^y) = 1 - \frac{\sum_{j \in \mathcal{K}} p_j \nu_j^{-1} \sum_{i \in \mathcal{I}_j} P^i(y)}{C(y)}\,.$$

The quantities P^j and C can be computed using the recursive schemes:

$$\begin{aligned} C(z) &= \Phi(z)\mathbf{1}_{\{z \in \mathcal{X}\}} + \sum_{i \in \mathcal{I} : z_i > 0} C(z - e_i)\nu'(z - e_i)^{-1}, \\ P^j(z) &= \Phi(z - e_j)\mathbf{1}_{\{z - e_j \in \mathcal{X}\}} + \sum_{i \in \mathcal{I} : z_i > 0} P^j(z - e_i)\nu'(z - e_i)^{-1}\,. \end{aligned}$$

3.5 Bounds

We obtain bounds by applying the following simple principle. Consider a routing associated with the intensity function h. If $h \leq g$, then the balance function can be decomposed in terms of balance functions of rectangular policies with intensity g. We state more formally this result in the following lemma.

Theorem 3.7. *Consider an admissible balance function Λ with intensity h such that $h \leq g$, we have:*

$$\Lambda = \sum_{y \in \mathcal{X}} c_y \Lambda^{y,g}\,,$$

with $(c_y)_{y \in \mathcal{X}}$ defined by $c_y = \Lambda^{y,g}(y)^{-1}\big(\Lambda(y) - g(y)^{-1} \sum_{i \in \mathcal{I}} \Lambda(y + e_i)\big)$.

Proof. Introduce the frontier of the attainable states under Λ, $\mathcal{H} = \{x : \Lambda(x) \neq 0,\ \forall i \in \mathcal{I}, \Lambda(x+e_i) = 0\}$. Since Λ is admissible, for every reachable state not in $\mathcal{H}$, we have

$$h(x)^{-1} \sum_{i\in\mathcal{I}} \Lambda(x+e_i) = \Lambda(x) \implies g(x)^{-1} \sum_{i\in\mathcal{I}} \Lambda(x+e_i) \leq \Lambda(x)\,.$$

Set $c_y = \Lambda^{y,g}(y)^{-1}(\Lambda(y) - g(y)^{-1}\sum_{i\in\mathcal{I}}\Lambda(y+e_i))$. We have shown that c_y is non-negative. Now define $\Lambda' = \sum_{y\in\mathcal{X}} c_y \Lambda^{y,g}$. Using an induction on the state space (starting from the states in $\mathcal{H}$), we prove that $\Lambda = \Lambda'$. □

A direct consequence of the above Proposition is that:

$$B_p(\Lambda) \geq \min_{y\in\mathcal{X}} B_p(\Lambda^{y,g})\,. \tag{22}$$

Corollary 3.8. *Denote by B_p^* the minimum value of the blocking probability B_p over $\mathcal{A}$. We have the following bounds:*

$$\min_{y\in\mathcal{X}} B_p(\Lambda^{y,\nu}) \ \leq\ B_p^* \ \leq\ \min_{y\in\mathbb{N}^N} B_p(\Lambda^y) = \min_{y:\forall i,\ y^{(i)}\in\mathcal{X}} B_p(\Lambda^y)\,.$$

Proof. Since ν is an upper bound of the intensity for any admissible policy, we can always apply (22) with $g = \nu$. This provides the lower bound for B_p^*. The upper bound is clear: it follows from the fact that the decentralized balance functions Λ^y are admissible. □

Remark 3.1. *We could have stated the same result for any performance criterion $\mathcal{P}_f : \mathcal{A} \to \mathbb{R}_+$ which is convex, that is, which satisfies:*

$$\forall t \in [0,1],\ \forall \Lambda_1, \Lambda_2 \in \mathcal{A}, \qquad \mathcal{P}_f((1-t)\Lambda_1 + t\Lambda_2) \geq (1-t)\mathcal{P}_f(\Lambda_1) + t\mathcal{P}_f(\Lambda_2)\,. \tag{23}$$

3.6 One class of customers

Structural and optimality results for networks with only one arrival process have been given in [2]. We show that the results from [2] are a special case of the above results. Also the situation becomes much simpler, and for instance, the bounds of Corollary 3.8 are attained.

The set $\mathcal{I}$ is defined as before. The set $\mathcal{K}$ is reduced to a singleton and we simplify the notation accordingly. For instance, ν is the rate of the unique arrival process.

Consider a rectangular balance function associated with $y \in \mathcal{X}$ and ν. We have by definition $\sum_{i\in\mathcal{I}} \lambda_i(x) = \nu$. Since there is a single arrival process, the intensity constraint (2) is satisfied and $\tilde{\Lambda}^{y,g}$ is admissible. Also, we have (using (14) and (18))

$$\tilde{\Lambda}^{y,g}(x) = \tilde{\Lambda}^{y}(x) = \begin{cases} \binom{|y-x|}{y-x}\ \nu^{-|y-x|} & \text{if } x \in y^{\downarrow} \\ 0 & \text{otherwise} \end{cases}.$$

Therefore any rectangular balance function is decentralized. The decentralized balance functions (also called "simple" balance functions) are in that case 'extremal' and form a basis to decompose admissible balance functions. Next result is a corollary of Theorem 3.7 (see also [2, Theorem 1]).

Corollary 3.9. *Consider an admissible balance function* Λ. *There exists* $(c_x)_{x\in\mathcal{X}}, c_x \geq 0, \sum_x c_x = 1$, *such that:*

$$\Lambda = \sum_{y\in\mathcal{X}} c_y \Lambda^y \,. \tag{24}$$

This implies in particular, that for the blocking probability B defined in (6), **a simple balance function is optimal**. This holds more generally for any convex (see (23)) performance criterion.

Furthermore, the blocking probability can be evaluated very easily. By specializing the results in Section 3.2, we get that $B(\Lambda^y) = \pi_y(y) = \Phi(y)/C(y)$ which can be evaluated recursively using the formula (16) for $C(y)$:

$$(B(\Lambda^y))^{-1} = 1 + \sum_{i\in\mathcal{I}:y_i>0} \frac{\phi_i(y)}{\nu} (B(\Lambda^{y-e_i}))^{-1} \,, \tag{25}$$

with $B(\Lambda^{\mathbf{o}}) = 1$.

4 Networks with admission control

We now consider models with more than one arrival process. We restrict ourselves to the situation where each node is fed by only one arrival process: $\mathcal{K} = \mathcal{I}$. There is no "routing" but only admission control. The model includes for example bandwidth sharing processor sharing network operated under balance fairness [5] and subject to admission control schemes. Note that the service rates can be coupled in a very complex way.

The book of Ross [13] gives a good overview of existing results when the service rates are "uncoupled" i.e., depend on the local number of customers at each node only: $\phi_i(x) = \phi_i(x_i)$ (the so-called stochastic knapsack problem). The author studies several types of insensitive policies under more specific assumptions on the state space and/or the service rates of the network. The optimality results provided concern the case of locally dependent routing intensities, $\lambda_i(x) = \lambda_i(x_i)$, and mostly in dimension 2 or 3.

In this Section, we do not restrict ourself to the assumption that $\phi_i(x) = \phi_i(x_i)$ but consider any balanced service rates (i.e. satisfying (7)). We first give a characterization of any insensitive policy in terms of coordinate-convex policies, i.e. policies such that customers are fully accepted in a Ferrers shaped subset of the state space. This allows us to conclude that an optimal insensitive policy is necessarily coordinate-convex. We then give conditions for which the optimal policy is the complete sharing policy, i.e. consists in always accepting customers when possible. In this last case, we will be able to compute efficiently the optimal performance of the network.

4.1 Extremal stationary measures

Let $\mathcal{X}$, $\mathcal{I} = \mathcal{K}$, and Φ be fixed.

The *complete sharing policy* is the policy in which customers are admitted whenever it is possible. It is characterized by the rates :

$$\forall x \in \mathcal{X}, \forall i \in \mathcal{I}, \qquad \lambda_i(x) = \begin{cases} \nu_i & \text{if } x + e_i \in \mathcal{X} \\ 0 & \text{otherwise .} \end{cases}$$

We denote by Λ_d the balance function corresponding to the complete sharing policy:

$$\forall x \in \mathcal{X}, \qquad \tilde{\Lambda}_d(x) = \prod_i \nu_i^{x_i} .$$

Observe that Λ_d coincides with the normalized decentralized balance function Λ^y associated with a point y such that $\mathcal{X} \subset y^{\downarrow}$.

Definition 4.1. *Consider a Ferrers set $\mathcal{C} \in \mathcal{F}(\mathcal{X})$. The* coordinate-convex *balance function associated with $\mathcal{C}$ is defined by,*

$$\tilde{\Lambda}^{\mathcal{C}}(x) = \tilde{\Lambda}_d(x)\mathbf{1}_{\{x \in \mathcal{C}\}} .$$

It corresponds to the following coordinate-convex *policy: if $x + e_i \in \mathcal{C}$, then $\lambda_i(x) = \nu_i$, if $x + e_i \notin \mathcal{C}$, then $\lambda_i(x) = 0$. In words, customers are accepted in a Ferrers shaped subset of the state space.*

We now state the main result of this section.

Theorem 4.2. *An admissible balance function Λ can be decomposed as:*

$$\Lambda(x) = \sum_{\mathcal{C} \in \mathcal{F}(\mathcal{X})} \beta(\mathcal{C}) \Lambda^{\mathcal{C}}(x),$$

with $\beta(\mathcal{C}) \geq 0$ for all $\mathcal{C}$ and $\sum_{\mathcal{C} \in \mathcal{F}(\mathcal{X})} \beta(\mathcal{C}) = 1$.

Observe the difference with Theorem 3.7: here all the balance functions $\Lambda^{\mathcal{C}}$ are **admissible**. Theorem 4.2 is hence similar to the decomposition obtained for the single-class systems in Corollary 3.9.

Proof. We use an induction on the cardinality of the state space. If the state space contains one state, the result is obviously true. Suppose it is satisfied for any $\mathcal{X} \in \mathcal{F}(\mathbb{N}^{\mathcal{I}})$ of cardinality less than or equal to $n-1$. Consider a state space $\mathcal{X} \in \mathcal{F}(\mathbb{N}^{\mathcal{I}})$ of cardinality n and an admissible balance function $\tilde{\Lambda}$. We are going to show that: $\tilde{\Lambda}(x) = \beta \tilde{\Lambda}^{\mathcal{X}}(x) + \tilde{G}(x)$, with $\tilde{G}$ being an admissible balance function defined on a strictly smaller state space $\mathcal{X}' \subset \mathcal{X}$.

Consider $H(x) = \tilde{\Lambda}(x) / \prod_i \nu_i^{x_i}$. Recall the intensity constraints (see (2) and (9)): $\forall i, \forall x,\ \tilde{\Lambda}(x + e_i)/\tilde{\Lambda}(x) \leq \nu_i$. These inequalities can be rewritten as: $\forall i, \forall x,\ H(x) \geq H(x + e_i)$.

Now define ω as the smallest value of H on $\mathcal{X}$ (which is attained on the frontier of $\mathcal{X}$ since H is coordinate-wise decreasing) and x_0 as a point of the frontier such that $H(x_0) = \omega$. Set $G(x) = H(x) - \omega \mathbf{1}_{\{x \in \mathcal{X}\}}$. We have, $\forall x \in \mathcal{X}$,

$$\begin{aligned} \tilde{\Lambda}(x) &= G(x) \prod_i \nu_i^{x_i} + \omega \prod_i \nu_i^{x_i} \\ \tilde{\Lambda}(x) &= G(x) \prod_i \nu_i^{x_i} + \omega \tilde{\Lambda}^{\mathcal{X}}(x) . \end{aligned}$$

For all $x \in \mathcal{X}, x + e_i \in \mathcal{X}$,

$$G(x + e_i) = H(x + e_i) - \omega \leq H(x) - \omega = G(x) \,. \tag{26}$$

Set $\tilde{G}(x) = G(x) \prod_i \nu_i^{x_i}$. Using (26), we obtain that $\tilde{G}(x + e_i)/\tilde{G}(x) \leq \nu_i$. Therefore $\tilde{G}$ is an admissible balance function. By construction $\tilde{G}(x_0) = 0$. Therefore, $\tilde{G}$ is non-zero on a set of cardinality less than or equal to $n - 1$. Also, since $\mathcal{X}$ is a Ferrers set and the point x_0 belongs to the frontier of $\mathcal{X}$, the set $\mathcal{X} \setminus \{x_0\}$ is still a Ferrers set. So, $\tilde{G}$ can be viewed as an admissible balance function on the Ferrers set $\mathcal{X} \setminus \{x_0\}$ of cardinality $n - 1$. This concludes the proof. □

Let the performance criterion be the blocking probability B_p defined in (6). We can deduce from Theorem 4.2 that the optimal insensitive policy is a coordinate-convex policy.

Corollary 4.3. *Let B_p^* be the infimum of B_p over all insensitive policies. We have:*

$$B_p^* = \min_{\mathcal{C} \in \mathcal{F}(\mathcal{X})} B_p(\Lambda^{\mathcal{C}}),$$

where $\Lambda^{\mathcal{C}}$ is introduced in Definition 4.1.

This optimality result extends to any convex criterion. In general, the minimum is not attained for a decentralized policy. This is illustrated in Section 5.

4.2 Recursive formulas and optimality of a complete sharing policy

We show how to compute the blocking probabilities recursively (Proposition 4.4). We also provide sufficient conditions to guarantee that the optimal policy is a complete sharing policy. As in Section 3, for a Ferrers set $\mathcal{C}$, define

$$C(\mathcal{C}) = \sum_{z \in \mathcal{C}} \tilde{\Lambda}^{\mathcal{C}}(z)\Phi(z), \qquad P^j(\mathcal{C}) = \sum_{z \in \mathcal{C}} \tilde{\Lambda}^{\mathcal{C}}(z + e_j)\Phi(z) \,.$$

Proposition 4.4. *Consider a Ferrers set $\mathcal{C} \in \mathcal{F}(\mathcal{X})$ and the corresponding coordinate-convex policy. The blocking probability $B_p(\Lambda^{\mathcal{C}})$ satisfies*

$$B_p(\Lambda^{\mathcal{C}}) = 1 - \frac{1}{C(\mathcal{C})} \sum_i \frac{p_i P^i(\mathcal{C})}{\nu_i} \,. \tag{27}$$

The quantities $P^i(\mathcal{C})$ and $C(\mathcal{C})$ can be evaluated recursively. For a point $x \notin \mathcal{C}$ such that $\mathcal{C} \cup \{x\}$ is also a Ferrers set, we have:

$$C(\mathcal{C} \cup \{x\}) = C(\mathcal{C}) + \tilde{\Lambda}_d(x)\Phi(x), \tag{28}$$

$$P^i(\mathcal{C} \cup \{x\}) = P^i(\mathcal{C}) + \tilde{\Lambda}_d(x)\Phi(x - e_i) \,. \tag{29}$$

Proof. Let us prove formula (27). Recall that the stationary distribution is given by $(\Phi(x)\Lambda^{\mathcal{C}}(x))_{x\in\mathcal{C}}$. Using $\lambda_i(x) = \Lambda^{\mathcal{C}}(x+e_i)/\Lambda^{\mathcal{C}}(x)$, the blocking probability $B_i(\Lambda^{\mathcal{C}})$ of customers of class i satisfies:

$$\begin{aligned} 1 - B_i(\Lambda^{\mathcal{C}}) &= \sum_{x\in\mathcal{C}} \Phi(x)\Lambda^{\mathcal{C}}(x)\frac{\Lambda^{\mathcal{C}}(x+e_i)}{\Lambda^{\mathcal{C}}(x)}\frac{1}{\nu_i} \\ &= \sum_{x\in\mathcal{C}} \frac{\Phi(x)\tilde{\Lambda}^{\mathcal{C}}(x)}{\sum_{y\in\mathcal{C}}\Phi(y)\tilde{\Lambda}^{\mathcal{C}}(y)}\frac{\tilde{\Lambda}^{\mathcal{C}}(x+e_i)}{\tilde{\Lambda}^{\mathcal{C}}(x)}\frac{1}{\nu_i} = \sum_{x\in\mathcal{C}} \frac{\Phi(x)\tilde{\Lambda}^{\mathcal{C}}(x+e_i)}{\sum_{y\in\mathcal{C}}\Phi(y)\tilde{\Lambda}^{\mathcal{C}}(y)}\frac{1}{\nu_i} = \frac{P^i(\mathcal{C})}{\nu_i C(\mathcal{C})}\,. \end{aligned}$$

Formula (27) follows readily.

Consider a Ferrers set $\mathcal{C}$ and a point $x \notin \mathcal{C}$ such that $\mathcal{C}\cup\{x\}$ is also a Ferrers set. We have:

$$\begin{aligned} C(\mathcal{C}\cup\{x\}) &= \sum_{z\in\mathcal{C}\cup\{x\}} \tilde{\Lambda}_d(z)\mathbf{1}_{\{z\in\mathcal{C}\cup\{x\}\}}\Phi(z) = C(\mathcal{C}) + \tilde{\Lambda}_d(x)\Phi(x). \\ P_j(\mathcal{C}\cup\{x\}) &= \sum_{z\in\mathcal{C}\cup\{x\}} \mathbf{1}_{\{z+e_j\in\mathcal{C}\cup\{x\}\}}\tilde{\Lambda}_d(z+e_j)\Phi(z), \end{aligned}$$

Using that $\mathcal{C}\cup\{x\}$ is a Ferrers set, $z+e_j \in \mathcal{C}\cup\{x\}$ and $z \in \mathcal{C}\cup\{x\}$ implies $z\in\mathcal{C}$ and $z+e_j\in\mathcal{C}$ or $z = x-e_j$. Hence:

$$\begin{aligned} P_j(\mathcal{C}\cup\{x\}) &= \sum_{z\in\mathcal{C}} \mathbf{1}_{\{z+e_j\in\mathcal{C}\}}\tilde{\Lambda}_d(z+e_j)\Phi(z) + \tilde{\Lambda}_d(x)\Phi(x-e_j), \\ P_j(\mathcal{C}\cup\{x\}) &= P_j(\mathcal{C}) + \tilde{\Lambda}_d(x)\Phi(x-e_j). \end{aligned}$$

□

Proposition 4.5. *Let X be a r.v. distributed according to the stationary number of customers. We have:*

$$B_p(\Lambda^{\mathcal{C}}) = 1 - \sum_{i\in\mathcal{I}} \frac{p_i}{\nu_i} E[\phi_i(X)]\,. \tag{30}$$

Let x be a point such that $\mathcal{C}\cup\{x\} \in \mathcal{F}(\mathcal{X})$. The blocking probabilities satisfy:

$$[B_p(\Lambda^{\mathcal{C}\cup\{x\}}) \le B_p(\Lambda^{\mathcal{C}})] \iff [\sum_{i\in\mathcal{I}} \frac{p_i}{\nu_i} E[\phi_i(X)] \le \sum_{i\in\mathcal{I}} \frac{p_i}{\nu_i}\phi_i(x)]\,. \tag{31}$$

Proof. Denote by $B_i(\Lambda^{\mathcal{C}})$ the blocking probability of class i customers. We first prove the rate conservation law:

$$\nu_i(1 - B_i(\Lambda^{\mathcal{C}})) = E[\phi_i(X)]\,.$$

Let $(X(t))_t$ be a Markov process describing the state of the network in stationary behavior. The processes $M_t = X_i(t) - \int_0^t (\lambda_i(X(s)) - \phi_i(X(s))ds$ are square integrable martingales, for all $i\in\mathcal{I}$, see for instance [11]. By Doob's martingale convergence theorem, M_t converges a.s. to a finite limit as t goes to infinity, so M_t/t converges a.s. to 0. Since the state space is finite, we also have that $X_i(t)/t$ converges a.s. to 0. It implies that:

$$\frac{1}{t}\int_0^t (\nu_i \mathbf{1}_{\{X(s)+e_i\in\mathcal{C}\}} - \phi_i(X(s)))ds \xrightarrow{t} \nu_i(1 - B_i(\Lambda^{\mathcal{C}})) - E[\phi_i(X)] = 0\,.$$

By summing over i, we get (30).

Now let us prove (31). Write $P_p(\mathcal{C}) = \sum_i (p_i/\nu_i) P_i(\mathcal{C})$. We have

$$[B_p(\Lambda^{\mathcal{C}\cup\{x\}}) \leq B_p(\Lambda^{\mathcal{C}})] \iff [\frac{P_p(\mathcal{C}\cup\{x\})}{C(\mathcal{C}\cup\{x\})} \geq \frac{P_p(\mathcal{C})}{C(\mathcal{C})}] \,.$$

Using Proposition 4.4, simple computations, and (30), this is equivalent to:

$$\sum_{i\in\mathcal{I}} \frac{p_i}{\nu_i}\phi_i(x) \geq \frac{P_p(\mathcal{C})}{C(\mathcal{C})} = 1 - B_p(\Lambda^{\mathcal{C}}) = \sum_{i\in\mathcal{I}} \frac{p_i}{\nu_i} E[\phi_i(X)] \,.$$

□

The simple comparison rule (31) allows to conclude that a complete sharing policy is optimal when the load of the network is small enough, or when the network is work conserving. We thus have the following results as direct consequences of Proposition 4.5.

Corollary 4.6 (Light traffic regime)**.** *Suppose that:*

$$\min_{x\in\mathcal{X}-\mathbf{o}} \sum_{i\in\mathcal{I}} \frac{p_i}{\nu_i}\phi_i(x) \geq 1 \,, \tag{32}$$

then a complete sharing policy is optimal for the blocking probability B_p.

Corollary 4.7 (Work-conserving network)**.** *Suppose that:*

$$\sum_i \phi_i(x) = c\mathbf{1}_{\{x\neq\mathbf{o}\}}, \tag{33}$$

for some constant $c \in \mathbb{R}_+^*$. Then a complete sharing policy is optimal for the blocking probability of an arrival customer, that is B_p with $p_i = \nu_i/(\sum_j \nu_j)$.

Finding the optimal routing can be done by using Proposition 4.4. However, the complexity of such an optimization program might be too big to be considered as a practical scheme. In any case, the results of Section 3 still applies: by focusing on policies having rectangular hyper-parallelepiped state spaces, we get easily computable upper and lower bounds of the optimal performance.

5 A four state example

Consider a model with two classes of customers and two nodes ($|\mathcal{K}| = |\mathcal{I}| = 2$) and a state space $\mathcal{X} = \{(0,0),(1,0),(0,1),(1,1)\}$. The arrival rates of the classes are denoted by λ_1 and λ_2 respectively. The service rates $\phi_i(x)$ are supposed to be balanced which means that: $\phi_1(1,1)\phi_2(0,1) = \phi_2(1,1)\phi_1(1,0)$. We use the notation

$$\phi_1(1,0) = a, \qquad \phi_2(0,1) = b, \qquad \phi_1(1,0)\phi_2(1,1) = c \,.$$

In Figure 5, we have represented all the different coordinate-convex insensitive policies as well as various sensitive policies (for which the traffic of a given class in a given state is either

fully accepted or fully rejected). The vertices represent the reachable states for each policy while the edges correspond to the transitions between states, of intensity λ_1 for an horizontal edge and λ_2 for a vertical one. We denote by P_a, P_b, P_{abc} the three decentralized policies, and by P_{ab} the unique non-decentralized coordinate-convex policy. The blocking probability of an arriving customer for the four insensitive policies are given by:

$$B(P_a) = \frac{\lambda_1}{(\lambda_1+\lambda_2^2)} + \frac{\lambda_2}{a(\lambda_1+\lambda_2)(1+\lambda_2/a)}, \quad B(P_b) = \frac{\lambda_2}{(\lambda_1+\lambda_2)} + \frac{\lambda_1^2}{b(\lambda_1+\lambda_2)(1+\lambda_1/b)},$$

$$B(P_{ab}) = \frac{\lambda_2/a + \lambda_1/b}{1+\lambda_1/b+\lambda_2/a},$$

$$B(P_{abc}) = \frac{\lambda_1(\lambda_1/b+\lambda_1\lambda_2/c)+\lambda_2(\lambda_2/a+\lambda_1\lambda_2/c)}{(\lambda_1+\lambda_2)(1+\lambda_1/b+\lambda_2/a+\lambda_1\lambda_2/c)}\,.$$

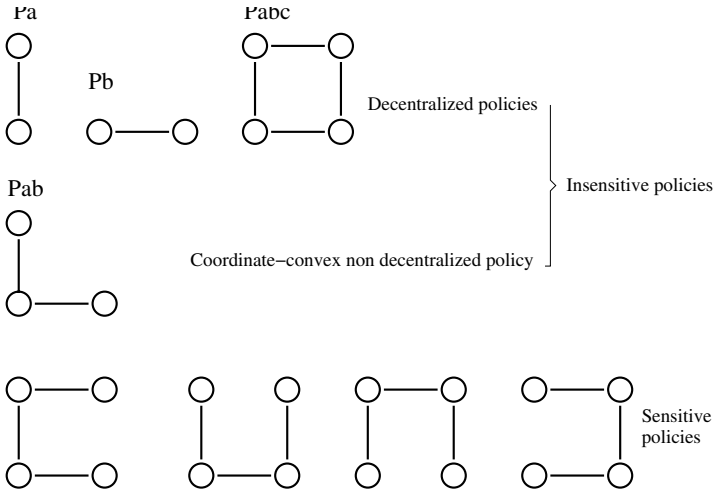

Figure 5: Insensitive policies and sensitive policies

Consider now the rectangular (not necessarily admissible) policies of Sections 3.1 and 3.2. They are all represented in Figure 6, with the transitions intensities being represented on the edges. The two policies on the left of Figure 6 are non-admissible (let us call them P_1 and P_2), while the one on the right is P_{abc}. The blocking probabilities of an arriving customer for P_1 and P_2 are:

$$B(P_1) = \frac{\lambda_1}{\lambda} + \frac{\lambda_2}{a(1+\lambda/a)}, \qquad B(P_2) = \frac{\lambda_2}{\lambda} + \frac{\lambda_2}{b(1+\lambda/b)}\,.$$

In particular, we check that it is possible to have

$$\min(B(P_1), B(P_2)) < \min(B(P_a), B(P_b), B(P_{ab}), B(P_{abc}))\,,$$

in which case the lower bound computed in Corollary 3.8 is not attained.

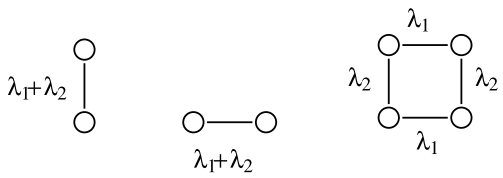

Figure 6: Rectangular (not necessarily admissible) policies

We represent in Figure 7 the value of the blocking for each coordinate-convex policy when $c = 0.1$ (left), $c = 2$ (right) for a and b varying from 0 to 4. The color code is as follows: the green curve corresponds to P_a, the blue curve corresponds to P_b, the yellow one to P_{ab}, and the red one to P_{abc}. Hence, the optimality of the non-decentralized policy P_{ab} can be observed for $c = 0.1$, while for $c = 2$ the complete sharing policy P_{abc} is always optimal. In accordance with Corollary 4.6, the complete sharing policy P_{abc} is optimal in light traffic in both cases.

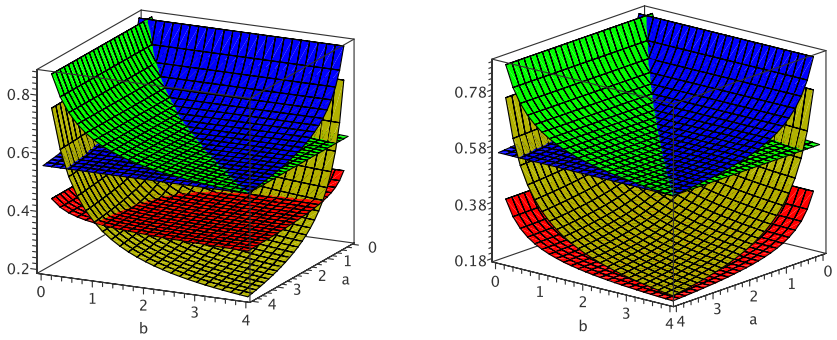

Figure 7: Blocking probabilities of the coordinate-convex policies

We now compare, for different values of the parameters, the minimum blocking probabilities of the three sets of policies represented in Figure 5: decentralized, coordinate-convex and sensitive policies, together with the lower bound of Corollary 3.8. The parameters of the four scenarios are gathered in the following table.

Scenario	a	b	c/a	λ_2
1	2	2	2	1
2	0.3	2	2	1
3	2	6	2	1
4	2	6	1	1

Table 1: Parameters of the 4 scenarios

Different observations can be made:

- Scenario 1 falls into a well studied case of a loss network with two circuits and symmetric service requirements since $\phi_1(1,0) = \phi_2(0,1) = \phi_1(1,1) = \phi_2(1,1)$. The optimal policy is known to be one of the policies of Figure 5, see [13]. We are in the domain of application of Proposition 4.6, so the complete sharing policy is optimal among insensitive policies. Since all the service rates are equal, it is clear that this policy is actually optimal among sensitive policies as well. It explains why all the curves correspond in Figure 8 (left). The non-monotonicity of the blocking probability with respect to the load (due to the trade-off between the influences of the two classes) can be observed.
- In the scenarios 2,3, and 4, the asymmetry of the service rates makes the insensitive policies perform worse than the sensitive ones for some loads. The lower bound of Corollary 3.8 becomes very loose in light traffic, while it is attained in heavy traffic.
- In scenarios 2 and 3, decentralized policies are optimal among insensitive policies while a non-decentralized policy is the best insensitive policy in scenario 4 for moderate loads.

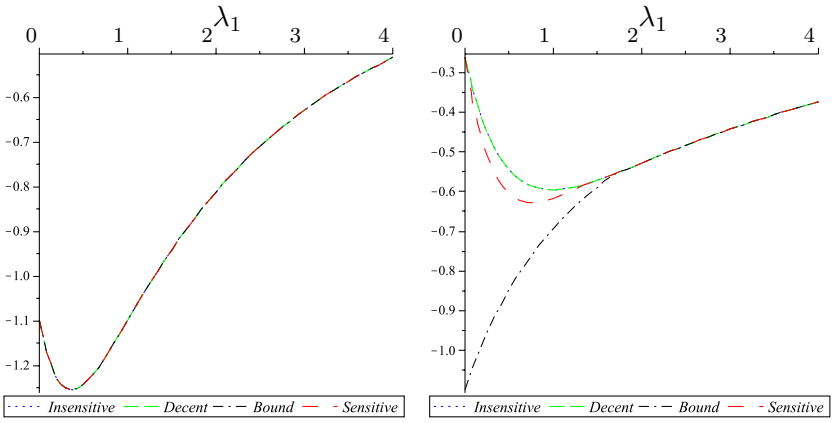

Figure 8: Blocking probabilities (log scale) for scenarios 1 (left) and 2 (right)

Other numerical studies. Of course, realistic examples have a much larger number of states. Several recent papers numerically compare insensitive and sensitive policies with the help of Markov decision process techniques [9, 10]. The example of Section 5 in [9] is enlightening. It has 4 nodes, $\mathcal{I} = \{a, b, c, d\}$, and 3 classes of customers, $\mathcal{K} = \{1, 2, 3\}$, with $\mathcal{I}_1 = \{a\}, \mathcal{I}_2 = \{b, c\}, \mathcal{I}_3 = \{d\}$. The service rates are balanced and given by a fair sharing between classes (i.e. of the type $\phi_i(x) = x_i/(\sum_j x_j)$). The authors show numerically that decentralized policies are actually optimal in the whole class of insensitive policies, for the whole range of load parameters, when classes 1 and 3 have the same mean service requirement. (Note that our results of Section 5 allow to efficiently compute the performance of these decentralized policies.) This example gives hope that the optimality of decentralized policies holds more generally under certain, still unknown, conditions.

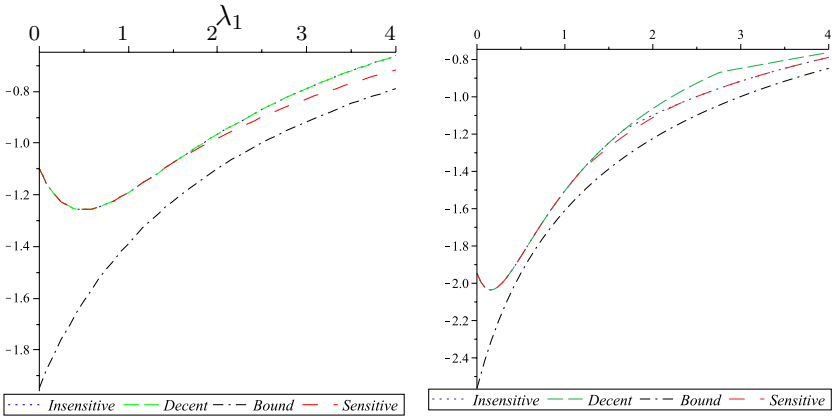

Figure 9: Blocking probabilities (log scale) for scenarios 3 (left) and 4 (right)

6 Conclusion

We give efficient recursive formulas to evaluate rectangular policies in the general case. This enables us to obtain computable bounds for the performance of the optimal insensitive policy. We then give a precise characterization of the optimal insensitive policies for networks with admission control. To find conditions ensuring the optimality of decentralized policies is still a challenging open question for general network topologies. Another important remaining issue is to determine whether the performance of the best insensitive policy is close to the one of the best policy.

References

[1] T. Bonald. Insensitive traffic models for communication networks. *Disc. Event Dyn. Syst.*, 17(3):405–421, 2007.

[2] T. Bonald, M. Jonckheere, and A. Proutière. Insensitive load balancing. In *Perf. Eval. Review: Proc. ACM Sigmetrics/ Performance 2004*, pages 367–377, 2004.

[3] T. Bonald, L. Massoulié, A. Proutière, and J. Virtamo. A queueing analysis of max-min fairness, proportional fairness and balanced fairness. *Queueing Syst.*, 53(1-2):65–84, 2006.

[4] T. Bonald and A. Proutière. Insensitivity in processor-sharing networks. *Perf. Eval.*, 49:193–209, 2002.

[5] T. Bonald and A. Proutière. Insensitive bandwidth sharing in data networks. *Queueing Syst.*, 44(1):69–100, 2003.

[6] A.K. Erlang. Solutions of some problems in the theory of probabilities of significance

in automatic telephone exchange. In E. Brockmeyer, H.L. Halstrom, and A. Jensen, editors, *The life and works of A.K. Erlang*. 1948 (1917 in danish).

[7] A. Hordijk and N.M. van Dijk. Networks of queues, part 1: Job-local-balance and the adjoint process; part 2: General routing and service characteristics. *Lecture Notes in Control and Information Sciences*, 60:158–205, 1983.

[8] F. Kelly. *Reversibility and Stochastic Networks*. Wiley, 1979.

[9] J. Leino and J. Virtamo. Insensitive load balancing in data networks. *Computer Networks*, 50(8):1059–1068, 2006.

[10] V. Pla, J. Virtamo, and J. Martï¿½nez-Bauset. Optimal robust policies for bandwidth allocation and admission control in wireless networks. *Computer Networks*, 52:3258–3272, 2008.

[11] P. Robert. *Stochastic Networks and Queues*. Springer, 2003.

[12] J.W. Roberts. A survey on statistical bandwidth sharing. *Comp. Netw.*, 45:319–332, 2004.

[13] Keith W. Ross. *Multiservice Loss Models for Broadband Telecommunication Networks*. Springer, 1995.

[14] R. Serfozo. *Introduction to Stochastic Networks*. Springer, 1999.

[15] P. Whittle. *Systems in stochastic equilibrium*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Ltd., 1986.