Multimodularity, Convexity and Optimization Properties

Eitan Altman, Bruno Gaujal, Arie Hordijk

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Multimodularity, Convexity and Optimization Properties

Eitan ALTMAN Bruno GAUJAL and Arie HORDIJK

N° 3181

Juin, 1997

THÈME 1

Rapport de recherche

ISSN 0249-6399

Multimodularity, Convexity and Optimization Properties

Eitan ALTMAN * Bruno GAUJAL ** and Arie HORDIJK ***

Thème 1 — Réseaux et systèmes
Projet Mistral, Sloop

Rapport de recherche n ° 3181 — Juin, 1997 — 25 pages

Abstract: We investigate in this paper the properties of multimodular functions. In doing so we give alternative proofs for properties already established by Hajek, and we extend his results. In particular, we show the relation between convexity and multimodularity, which allows us to restrict the study of multimodular functions to convex subsets of $\mathbb{Z}^m$. We then obtain general optimization results for average costs related to a sequence of multimodular functions. In particular, we establish lower bounds, and show that the expected average problem is optimized by using balanced sequences. We finally illustrate the usefulness of this theory in admission control into a D/D/1 queue with fixed batch arrivals, with no state information. We show that the balanced policy minimizes the average queue length for the case of an infinite queue, but not for the case of a finite queue. When further adding a constraint on the losses, it is shown that a balanced policy is also optimal for the case of finite queue.

Key-words: Multimodular functions, convexity, balanced sequences, admission control into a queue.

(Résumé : tsvp)

* INRIA, BP 93, 2004 Route des Lucioles, 06902 Sophia Antipolis Cedex, France. E-mail: altman@sophia.inria.fr. URL:http://www.inria.fr:80/mistral/personnel/Eitan.Altman/me.html

** INRIA, BP 93, 2004 Route des Lucioles, 06902 Sophia Antipolis Cedex, France. E-mail: gaujal@sophia.inria.fr. Bruno Gaujal is a member of a common project between CNRS, UNSA and INRIA.

*** Dept. of Mathematics and Computer Science, Leiden University, P.O.Box 9512, 2300RA Leiden, The Netherlands. E-mail: hordijk@wi.leidenuniv.nl. The research of Arie Hordijk was done while he was on sabbatical leave at INRIA, Sophia-Antipolis; it has been partially supported by the Ministère Français de l'Éducation Nationale et de l'Enseignement Supérieur et de la Recherche.

Multimodularité, convexité et propriétés d'optimisation

Résumé : Dans cet article, nous nous intéressons aux propriétés des fonctions multimodulaires. Ce faisant, nous montrons de nouvelles preuves des propriétés établies par Hajek et nous étendons certains de ses résultats. En particulier, nous montrons la relation qui existe entre multimodularité et convexité qui nous permet d'étudier la multimodularité sur des parties convexes de $\mathbb{Z}^m$. Ensuite, nous obtenons des résultats généraux d'optimisation pour le coût moyen d'une suite de fonctions multimodulaires. En particulier, on exhibe des bornes inférieures et on montre qu'elles sont atteintes pour les suites équilibrées. Finalement, on illustre l'utilité de cette théorie pour le contrôle d'admission dans une file D/D/1 avec des arrivées par paquets de taille fixe, sans information. On montre que la politique équilibrée minimise la longueur moyenne de la file pour une capacité infinie mais pas dans le cas d'une file à capacité finie. Dans ce cas, Si l'on ajoute une contrainte sur les pertes, on montre aussi qu'une politique équilibrée est optimale.

Mots-clé : Multimodularité, convexité, contrôle d'admission, suites équilibrées.

1 Introduction

The multimodularity property of functions was much investigated in the context of queueing systems. There are several cases in that field in which this property was exploited to solve stochastic control. Optimal admission control under no queue information was studied by Hajek [7]. The precise problem was to admit customers to a single queue, under the constraint that the long run fraction of customers admitted be at least p. The optimality of a policy based on a balanced sequence of admission actions was obtained in [7] for the number of customers in a one-server queue with exponential service and a renewal arrival process.

Another application of multimodular functions is in the control of queues with full state information. Weber and Stidham [8] (and later Glasserman and Yao [4]) obtained monotone properties of the optimal control policies as a function of the state, in a problem of control of service rates in a system of m queues in tandem. The methodology was strongly based on the multi-modularity properties of the immediate costs and the cost-to-go functions.

The purpose of this paper is to study the properties of multimodular functions, as a tool for further investigating the control of queueing systems. We provide alternative proofs for properties already established by Hajek. We further show the relation between convexity and multimodularity, which allows us to restrict the study of multimodular functions to convex subsets of $\mathbb{Z}^m$.

We then develop basic optimization tools for average costs, related to a sequence of multimodular functions. In particular, we establish lower bounds which are achieved by balanced sequences.

We finally illustrate the usefulness of this theory in admission control into a queue; we cite some results for the G/G/1 queue, and provide a detailed analysis of the D/D/1 queue with fixed batch arrivals, with no state information. We show, for the latter, that the policy which is defined through a balanced sequence minimizes the average queue length for the case of an infinite queue, but not for the case of a finite buffer. However, when further restricting to those policies for which no losses occur, we obtain again the optimality of balanced policies. To conclude that example, we study also the case where it is possible to admit a part of an arriving batch.

In follow-up papers, we shall make use of the theoretical results of this paper in order to study more general admission and service control problems in dynamic systems that can be described using the max-plus algebra, with general inter-arrival and service times.

2 Properties of multimodular functions

We present in this section a short overview and extension of Hajek's theory of multimodular functions. We begin by presenting the definition of multimodularity, and some general properties (Subsection 2.1). We then present in Subsection 2.2 the relation between multimodularity and convexity. The properties presented in Subsection 2.2 are those needed in the following sections on optimization and control.

Let $e_i \in \mathbb{N}^m$, $i = 1, ..., m$ denote the vector having all entries zero except for a 1 in its ith entry. Define $d_i = e_{i-1} - e_i$, $i = 1, ..., m$ (for an integer i taking values between 0 and m, we understand throughout $i - 1 = m$ for $i = 0$).

Let $\mathcal{F} = \{-e_1, d_2, ..., d_m, e_m\}$. Define $\mathcal{G} = \{e_i, -e_i, d_i, -d_i, \ i = 0, 1, ..., m\}$.

Definition 2.1 (Hajek). *A function f on $\mathbb{Z}^m$ is multimodular with respect to $\mathcal{F}$ if for all $x \in \mathbb{Z}^m$, $v, w \in \mathcal{F}$, $v \neq w$,*

$$f(x+v) + f(x+w) \geqslant f(x) + f(x+v+w). \tag{1}$$

Unless otherwise stated, we shall say that f is multimodular if it is multimodular with respect to $\mathcal{F}$.

2.1 General properties

For a function g defined on $\mathbb{Z}^m$, define

$$\Delta_i g(x) = \Delta_{e_i} g(x) = g(x+e_i) - g(x), \qquad \Delta_{d_i} = \Delta_{i-1} g - \Delta_i g.$$

We further define $\Delta_{-e_i} = -\Delta_{e_i}$, $\Delta_{-d_i} = -\Delta_{d_i}$. Note that $\Delta_{d_i} g(x) = g(x+e_i+d_i) - g(x+e_i)$.

It is easy to check that

Lemma 2.1. *Δ_v is a linear function for any $v \in \mathcal{G}$. For all $v, w \in \mathcal{G}$, $\Delta_v \Delta_w = \Delta_w \Delta_v$.*

Lemma 2.2. *f is multimodular if and only if*

$$\Delta_v \Delta_w f \leqslant 0 \tag{2}$$

for all $v, w \in \mathcal{F}$, $w \neq v$.

Proof. Consider first $w = d_i, v = d_j$ $(v \neq w)$. Then

$$\begin{aligned}
\Delta_v \Delta_w f(x) &= (\Delta_{i-1} - \Delta_i)(\Delta_{j-1} - \Delta_j) f(x) \\
&= (\Delta_{i-1} - \Delta_i)(f(x+e_{j-1}) - f(x+e_j)) \\
&= f(x+e_{j-1}+e_{i-1}) - f(x+e_{j-1}+e_i) - f(x+e_j+e_{i-1}) + f(x+e_j+e_i) \\
&= f(z+d_j+d_i) - f(z+d_j) - f(z+d_i) + f(z) \qquad (3)
\end{aligned}$$

where $z \stackrel{\triangle}{=} x + e_j + e_i$.

Let $v = e_j, w = -e_i$. Then

$$\begin{aligned}
\Delta_v \Delta_w f(x) &= -\Delta_j \Delta_i f(x) \\
&= -\Delta_j (f(x+e_i) - f(x)) \\
&= -f(x+e_i+e_j) + f(x+e_i) + f(x+e_j) - f(x) \\
&= -f(z+e_j) + f(z) + f(z+[-e_i]+e_j) - f(z+[-e_i]) \qquad (4)
\end{aligned}$$

where $z \stackrel{\triangle}{=} x + e_i$.

Let $v = e_i, w = d_j$ $(j \neq i+1)$.

$$\begin{aligned}
\Delta_v \Delta_w f(x) &= \Delta_i(\Delta_{j-1} - \Delta_j) f(x) \\
&= \Delta_i(f(x+e_{j-1}) - f(x+e_j)) \\
&= f(x+e_{j-1}+e_i) - f(x+e_{j-1}) - f(x+e_j+e_i) + f(x+e_j) \\
&= f(z+d_j+e_i) - f(z+d_j) - f(z+e_i) + f(z) \qquad (5)
\end{aligned}$$

where $z = x + e_j$.

Let $v = -e_i, w = d_j$ $(j \neq i)$.

$$\begin{aligned}
\Delta_v \Delta_w f(x) &= -\Delta_i(\Delta_{j-1} - \Delta_j) f(x) \\
&= -\Delta_i(f(x+e_{j-1}) - f(x+e_j)) \\
&= -f(x+e_{j-1}+e_i) + f(x+e_{j-1}) + f(x+e_j+e_i) - f(x+e_j) \\
&= -f(z+d_j) + f(z+d_j+[-e_i]) + f(z) + f(z+[-e_i]) \qquad (6)
\end{aligned}$$

where $z = x + e_j + e_i$.

The above equations easily imply the Lemma. ∎

Lemma 2.3. *(i) If f is multimodular then*
(i) For all i, j,

$$\Delta_i \Delta_j f \geqslant 0, \qquad (7)$$

and
(ii) For all i, j,

$$\Delta_j \Delta_j f \geqslant \Delta_i \Delta_j f. \qquad (8)$$

Proof. (i) Without loss of generality, assume that $i \leqslant j$. Then

$$\Delta_i \Delta_j f = \left(-\Delta_{-e_1} - \sum_{k=2}^{i} \Delta_{d_k} \right) \left(\sum_{l=j+1}^{m} \Delta_{d_l} + \Delta_{e_m} \right) f$$

The proof of (i) is established by applying Lemma 2.2.

For $i < j$ we have

$$\begin{aligned}
\Delta_j \Delta_j f &= (\Delta_j - \Delta_i)\Delta_j f + \Delta_i \Delta_j f \geqslant (\Delta_j - \Delta_i)\Delta_j f \\
&= \left(-\sum_{k=i+1}^{j} \Delta_{d_k} \right) \left(\sum_{l=j+1}^{m} \Delta_{d_l} + \Delta_{e_m} \right) f
\end{aligned}$$

and (ii) is established by applying Lemma 2.2. For $i > j$ we have

$$\Delta_j = \Delta_i - \sum_{k=i+1}^{m-1} \Delta_{d_k} - \Delta_m - \Delta_{-e_1} - \sum_{k=2}^{j} \Delta_k.$$

Hence

$$\begin{aligned}\Delta_j\Delta_j f &= \Delta_j(\Delta_j-\Delta_i)f+\Delta_j\Delta_i f \geqslant \Delta_j(\Delta_j-\Delta_i)f\\ &= \left(-\Delta_{-e_1}-\sum_{k=2}^{j}\Delta_{d_k}\right)\left(-\sum_{k=i+1}^{m}\Delta_{d_k}-\Delta_m-\Delta_{-e_1}-\sum_{k=2}^{j}\Delta_k\right)f.\end{aligned}$$

Again, (ii) is established by applying Lemma 2.2. ∎

Remark 2.1. The converse of the above Lemma holds for the 2-dimensional case: $\mathcal{F}=\{-e_1,d_2,e_2\}$. Indeed, assume that (7) and (8) hold. Then $\Delta_{-e_1}\Delta_{e_2}\leqslant 0$ due to (7); $\Delta_{-e_1}\Delta_{d_2}=\Delta_{-e_1}(\Delta_{e_1}-\Delta_{e_2})\leqslant 0$, and $\Delta_{e_2}\Delta_{d_2}=\Delta_{e_2}(\Delta_{e_1}-\Delta_{e_2})\leqslant 0$ due (8). Hence f is multimodular by Lemma 2.2.

∎

Lemma 2.4. *If f is multimodular then for all $i=1,...,m$,*
(i) $\Delta_{e_i}\Delta_{d_i}f\geqslant 0$,
(ii) $\Delta_{e_i}\Delta_{d_j}f\leqslant 0$, $j<i$ *and* $\Delta_{e_i}\Delta_{d_j}f\geqslant 0$, $j>i$,
(iii) $\Delta_{d_1}\Delta_{d_i}f\leqslant 0$, $i\neq 1$,
(iv) $\Delta_{d_i}\Delta_{d_i}f\geqslant 0$.
(v) $\Delta_{d_i}\Delta_{d_j}f\leqslant 0$ *for* $i\neq j$, $i\neq 1$, $j\neq 1$.

Proof. By taking $i=j-1$ in Lemma 2.3 (ii), we obtain (i).
(ii) For $j<i$,

$$\Delta_{e_i}\Delta_{d_j}f=\left(\sum_{k=i}^{m}\Delta_{d_k}+\Delta_{e_m}\right)\Delta_{d_j}f.$$

For $j>i$,

$$\Delta_{e_i}\Delta_{d_j}f=\left(-\Delta_{-e_1}-\sum_{k=2}^{i}\Delta_{d_k}\right)\Delta_{d_j}f.$$

For both cases, the proof is established by applying Lemma 2.2.
(iii)

$$\Delta_{d_1}\Delta_{d_i}f=(\Delta_{e_m}+\Delta_{-e_1})\Delta_{d_i}f.$$

The proof is established by applying Lemma 2.2.
(iv)

$$\Delta_{d_i}\Delta_{d_i}f=\Delta_{d_i}\left(-\sum_{j\neq i}\Delta_{d_j}\right)f$$

For $i\neq 1$, the proof is established by applying Lemma 2.2. For $i=1$ it follows from part (iii) of this Lemma. ∎

2.2 Multimodularity and convexity

In the space $\mathbb{R}^m$, $m+1$ extreme points in $\mathbb{Z}^m$ form a simplex.

A simplex consisting of the extreme points$\{x^0, ..., x^m\}$ of $\mathbb{Z}^m$ is called an *atom* (defined in [7] §3) if and only if for some ordering of the subset and for some permutation $\{i_1, \cdots, i_m\}$ of $(0, 1, ..., m)$,

$$\begin{array}{rcr} x^1 & = & x^0 + g_{i_1} \\ x^2 & = & x^1 + g_{i_2} \\ & . & \\ & . & \\ & . & \\ x^m & = & x^{m-1} + g_{i_m} \\ x^0 & = & x^m + g_{i_0} \end{array} \tag{9}$$

where $f_0, ..., f_m$ are the elements of $\mathcal{F}$.

Next we present a characterization of an atom [7], which is essential for the optimization result that we obtain in the following sections. Denote by $\lfloor x \rfloor$ the largest integer smaller than or equal to x. Then the following trivially holds

$$\int_0^1 \lfloor x + \theta \rfloor d\theta = x. \tag{10}$$

Given $z \in \mathbb{R}^m$, $\theta \in \mathbb{R}$, define the vector $u^z(\theta)$ in $\mathbb{Z}^m$:

$$u_i^z(\theta) = \lfloor \theta + z_1 + ... + z_i \rfloor - \lfloor \theta + z_1 + ... + z_{i-1} \rfloor,$$

$i = 1, ..., m$. Then by (10),

$$\int_0^1 u^z(\theta) d\theta = z. \tag{11}$$

Since

$$\int_0^1 \lfloor \theta + z_1 + ... + z_i \rfloor d\theta = z_1 + ... + z_i,$$

$$\int_0^1 \lfloor \theta + z_1 + ... + z_{i-1} \rfloor d\theta = z_1 + ... + z_{i-1}.$$

$u^z(\theta)$ is periodic in θ with period 1, and piecewize constant with at most $m+1$ jumps per period. Thus, the set $\{u^z(\theta) : 0 \leqslant \theta \leqslant 1\}$ contains at most $m+1$ vectors, all integer valued. The next Lemma follows from [7]:

Lemma 2.5. *A point z is contained in an atom called $S(z)$ if and only if the extreme points of $S(z)$ contain $\{u^z(\theta) : 0 \leqslant \theta \leqslant 1\}$. A point z is in the interior of an atom $S(z)$ if and only if the extreme points of $S(z)$ equal $\{u^z(\theta) : 0 \leqslant \theta \leqslant 1\}$.*

Each point $z \in \mathbb{R}^m$ is contained is some atom $S(z)$. It can thus be expressed as a convex combination of the extreme points of $S(z)$.

For any function f on $\mathbb{Z}^m$, we define the corresponding function $\widetilde{f}$ on $\mathbb{R}^m$ as follows. It agrees with f on $\mathbb{Z}_m$, and its value on an arbitrary point in $z \in \mathbb{R}^m$ is obtained as the corresponding linear interpolation of the values of f on the extreme points of the atom $S(z)$.

Theorem 2.1. *f is multimodular if and only if $\widetilde{f}$ is convex.*

Proof. "only if":

We check convexity at a point z; it is established by showing that at point z, at any direction d, the right derivative is greater that or equal to the left derivative. It obviously suffices to check at points that are on the boundary of an atom, since, by definition, $\widetilde{f}$ is linear in the interior of atoms. Hence, we first assume that the point z is on the interior of a face (of dimension $m-1$) which is common between two adjacent atoms. Without loss of generality, assume that the atoms (defined below by their extreme points) are

$$A = A(x_0, x_1, ..., x_m) \qquad \text{and} \qquad \overline{A} = A(x_0, x_1^*, ..., x_m).$$

where x_i satisfy (9) and

$$x_1^* = x_0 + g_{i_2}, \qquad x_2 = x_0^* + g_{i_1}.$$

Case 1: $g_{i_1} = -e_1 = (-1, 0, ..., 0)$, $g_{i_2} = e_m = (0, 0, ..., 1)$.

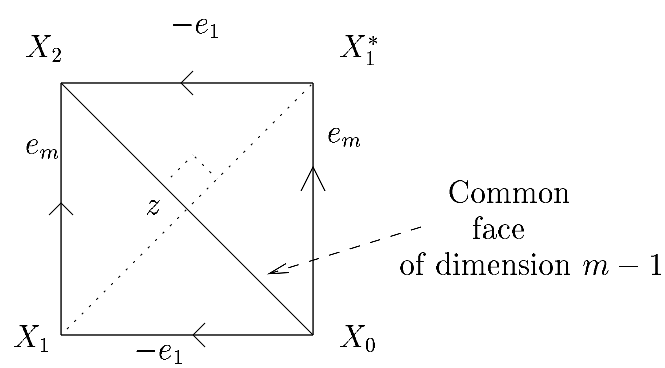

Figure 1: Checking convexity at a point z, Case 1.

Decompose direction d in its projection $d^{||}$ over the common face and in the direction $d^{\perp}$ perpendicular to that face. In the direction $d^{||}$, the left and right derivatives are equal. Note that the hyper-plane through x_1, x_0, x_2, x_1^* is perpendicular to the common face. In the direction $d^{\perp}$, the right derivative is a constant c times $\widetilde{f}(x_1^*) - \widetilde{f}(z)$. The left derivative is $c(\widetilde{f}(z)-\widetilde{f}(x_1))$. Omitting the constant c, we get for the difference

$$\begin{aligned}
&(\widetilde{f}(x_1^*)-\widetilde{f}(z))-(\widetilde{f}(z)-\widetilde{f}(x_1))\\
&= \frac{1}{2}\big[(f(x_1^*)-f(x_0))-(f(x_2)-f(x_1))\big] \qquad (12)\\
&\quad+\frac{1}{2}\big[(f(x_1)-f(x_0))-(f(x_2)-f(x_1^*))\big] \qquad (13)
\end{aligned}$$

The fact that both (12) as well as (13) are nonnegative follows by applying (1) with $x=x_0$. Then,

$$(12)\geqslant 0 \quad \text{since} \quad \begin{cases} x_1^* = x_0+e_m\\ x_2 = x_0-e_1+e_m\\ x_1 = x_0-e_1 \end{cases}$$

$$(13)\geqslant 0 \quad \text{since} \quad \begin{cases} x_1 = x_0-e_1\\ x_2 = x_0-e_1+e_m\\ x_1^* = x_0+e_m \end{cases}$$

Case 2: $g_{i_1}=e_m$ and $g_{i_2}=-e_1$. It is handled as Case 1.
Case 3: $g_{i_1}=d_2=(1,-1,0,...,0)$, $g_{i_2}=-e_1$. We thus set $x=x_0$, and $x_1^*=x_0-e_1, x_2=x_0-e_1+d_2, x_1=x_0+d_2$. In this case, x_2-x_1 is perpendicular to the common face. We decompose again d to the projection on the common face $d^{||}$, and to the direction $d^{\perp}$ perpendicular to the face. As in Case 1, it suffices to consider the direction $d^{\perp}$. The right derivative in this direction is $f(x_1^*)-f(x_0)$, and the left derivative is $f(x_2)-f(x_1)$ (both up to a multiplicative constant).

The difference between the right and left derivatives is indeed nonnegative: $f(x_1^*)-f(x_0)-(f(x_2)-f(x_1))\geqslant 0$. This is obtained again by applying (1).

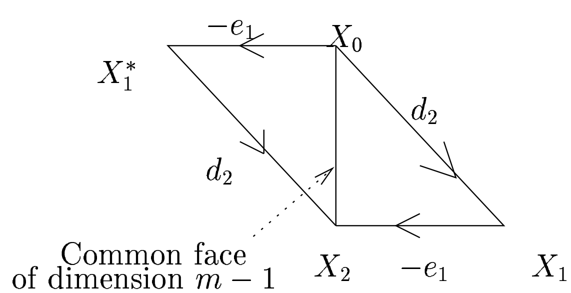

Figure 2: Checking convexity at a point on the common face, Case 3

Case 4: $g_{i_1}=d_2=(1,-1,0,...,0)$, $g_{i_2}=d_3$. In this case, $x_1^*-x_1$ is perpendicular to the common face, and the analysis is as for Case 1.

All other cases, in which z is in the interior of a face (of dimension $m-1$), common to two adjacent atoms, are similar to one of those considered above. It now remains to

consider the case where z lies on a subspace of $A \cap \overline{A}$ of dimension at most $m-2$. We note that the point z can be expressed as the limit of a sequence z_n of points that are on the boundary between two atoms which have a face (of dimension $m-1$) in common. Since f is continuous (this follows from the linear interpolation), and since the limit of convex functions is also convex, the convexity at the point z is established as well.
"if": Consider an arbitrary point x_0 and any two distinct elements g_i, g_j in $\mathcal{F}$. We have to show that

$$f(x_0) + f(x_2) - f(x_1) - f(x_1^*) \leqslant 0, \tag{14}$$

where $x_1 \triangleq x_0 + g_i$, $x_1^* \triangleq x_0 + g_j$, $x_2 \triangleq x_1 + g_j = x_1^* + g_i$.

Define $z \triangleq 0.5(x_1 + x_1^*) = 0.5(x_0 + x_2)$ and consider the line segment $x_1 \to z \to x_1^*$. The left derivative (l.d.) and right derivative (r.d.) in z are given by

$$l.d. = \widetilde{f}(\frac{1}{2}(x_1 + x_1^*)) - \widetilde{f}(x_1) = \frac{1}{2}f(x_0) + \frac{1}{2}f(x_2) - f(x_1),$$

$$r.d. = \widetilde{f}(x_1^*) - \widetilde{f}(\frac{1}{2}(x_0 + x_2)) = f(x_1^*) - \frac{1}{2}f(x_0) - \frac{1}{2}f(x_2).$$

Since $\widetilde{f}$ is convex, $r.d. - l.d.$ is non-positive, and hence (14) holds. ∎

Remark 2.2. *(Restriction of Multimodularity to a convex set)*
It is clear from the proof that we can restrict the domain of f in Theorem 2.1. Indeed, let A be a convex set which is a union of a set of atoms. The equivalence of the multimodularity of f and the convexity of $\widetilde{f}$ still holds if we restrict the function f to A, and restrict the definition of multimodularity to directions that lead to points in A. In other words, f is multimodular in A if the following holds. If $x_0, x_0 + g_i, x_0 + g_j, x_0 + g_i + g_j$ are all elements of A then

$$f(x_0) + f(x_0 + g_i + g_j) - f(x_0 + g_i) - f(x_0 + g_j) \leqslant 0.$$

Next we consider the integer convexity properties of a function f. A function f is said to be integer convex if the following holds. For vectors x and d in $\mathbb{Z}^m$, we have

$$f(x + d) - f(x) \geqslant f(x) - f(x - d).$$

Theorem 2.2. *Let f be multimodular. Then it is integer convex.*

Proof. Define $\delta_d^+(x)$:= the right derivative of $\widetilde{f}$ at x in the direction d and $\delta_d^-(x)$:= the left derivative of $\widetilde{f}$ at x in the direction d. Since $\widetilde{f}$ is convex (Theorem 2.1) then

$$\delta_d^+(x) \geqslant \delta_d^-(x). \tag{15}$$

Since $\widetilde{f}(y) = f(y)$ at the integer points, and since $\widetilde{f}$ is convex, we have

$$\delta_d^+(x) \leqslant \frac{f(x+d) - f(x)}{|d|}, \qquad \delta_d^-(x) \geqslant \frac{f(x) - f(x-d)}{|d|}.$$

This, together with (15) imply the integer convexity of f. ∎

The converse of the above theorem is not true:

Counter-example 2.1. Consider the convex function $f : \mathbb{N}^m \to \mathbb{R}$ given by $f(x) = \max_{i=1,\ldots,m} x_i$. It is integer convex since it is the maximum of convex (linear) functions. However, it is not multimodular. Indeed, consider $m = 2$, $x = (i+1, i)$ for some integer i. Then

$$2i + 2 = f(x - e_1 + e_m) + f(x) > f(x - e_1) + f(x + e_2) = 2i + 1.$$

Hence f is not multimodular.

Theorem 2.3. *Let f be multimodular, and let $h : \mathbb{R} \to \mathbb{R}$ be convex nondecreasing. Assume that $f(x+d_i) \leqslant f(x), i = 2, ..., m$, and $f(x-e_m) \leqslant f(x)$. Then $h(f)$ is multimodular.*

Proof. Let $g \in \mathcal{F}$, $g \neq d_i$. Since f is multimodular, we have:

$$f(x + d_i) - f(x) \geqslant f(x + d_i + g) - f(x + g) =: \alpha. \tag{16}$$

Hence,

$$\begin{aligned} h(f(x + d_i + g)) - h(f(x + g)) &= h(f(x + d_i) + \alpha) - h(f(x + d_i)) \leqslant h(f(x) + \alpha) - h(f(x)) \\ &= h(f(x) + f(x + d_i + g) - f(x + g)) - h(f(x)) \\ &\leqslant h(f(x + d_i)) - h(f(x)). \end{aligned}$$

The first inequality follows from the convexity of h, and since $f(x + d_i) \leqslant f(x)$. The second inequality follows from the fact that h it is nondecreasing, and the fact that by (16), $f(x + d_i) \geqslant f(x) + f(x + d_i + g) - f(x + g)$.

The same argument holds for e_m replacing d_i, which establishes the proof. ∎

3 The optimality of regular policies for a single criterion

In this section, we will use multimodularity to optimize a cost function based on a sequence of functions which will represent a quantity of interest in a given model, such as workload in a queue for example (see § 6 for a more precise instance of the problem).

Consider a sequence of nonnegative functions $f_k : \mathbb{N}^k \to \mathbb{R}$ that satisfy the following assumptions:

- <1> f_k is multimodular.

- <2> $f_k(a_1, ..., a_k) \geqslant f_{k-1}(a_2, ..., a_k), \ \forall K > 1;$

For a given sequence $\{a_k\}$, we define the cost $g(a)$ as

$$g(a) = \overline{\lim_{N\to\infty}} \frac{1}{N} \sum_{n=1}^{N} f_n(a_1, ..., a_n).$$

Let p and θ be two positive reals. We define the *balanced sequence* $\{a_k^p(\theta)\}$ with rate p and initial phase θ as,

$$a_k^p(\theta) = \lfloor kp + \theta \rfloor - \lfloor (k-1)p + \theta \rfloor, \tag{17}$$

where $\lfloor x \rfloor$ is the largest integer smaller than or equal to x. Note that the set $\{a_k^p(\theta), 0 \leqslant \theta \leqslant 1\}$ are extreme points of an atom containing the point $(p, p, ..., p)$.

The aim of this section will be to prove that this sequence minimizes the function g, provided that some conditions (including <1> and <2>) above hold. This sequence was used by Hajek in [7], and we use several properties of the balanced sequence established therein. To establish the main optimization results, we need the following technical Lemma.

Lemma 3.1. *If f_k satisfies assumption <2>, then the function $\widetilde{f}_k$ satisfies assumption <2> for positive real numbers.*

Proof. Let $z = (z_1, \cdots, z_k) \in \mathbb{R}_+^k$. This point belongs to an atom $S(z)$ made by the extreme points $x^0, x^1, \cdots, x^k$. The numbering of the extreme points of the atom is chosen such that according to the base $\mathcal{F}^k = (-e_1^k, d_1^k, \cdots, d_{k-1}^k, e_k^k)$, $x^1 = x^0 - e_1^k$. The other indices are arbitrary. This implies that $x_j^0 = x_j^1$ for all $j > 2$. If we call P the projection of $\mathbb{R}_+^k$ onto $\mathbb{R}_+^{k-1}$ along the first coordinate,

$$\begin{aligned}
P(d_j^k) &= P(d_j^{k-1}) \text{ if } 1 < j < k \\
P(d_1^k) &= -e_1^{k-1} \\
P(e_1^k) &= 0 \\
P(e_k^k) &= e_k^{k-1} \\
P(x^i) &= (x_2^i, \cdots, x_k^i)
\end{aligned}$$

Theses equalities imply that $P(x^0) = P(x^1)$ and $P(x^1), \cdots, P(x^k)$ form an atom in $\mathbb{R}_+^{k-1}$, using the definition of an atom. Also, P(z) belongs to this atom, and if

$$(z_1, z_2, \cdots, z_k) = \left(1 - \sum_{i=1}^{k} \alpha_i\right) x^0 + \alpha_1 x^1 + \cdots + \alpha_k x^k,$$

then

$$\begin{aligned}
(z_2, \cdots, z_k) &= P(z_1, z_2, \cdots, z_k) \\
&= \left(1 - \sum_{i=1}^{k} \alpha_i\right) P(x^0) + \alpha_1 P(x^1) + \cdots + \alpha_k P(x^k) \\
&= \left(1 - \sum_{i=2}^{k} \alpha_i\right) P(x^1) + \cdots + \alpha_k P(x^k).
\end{aligned}$$

Now,

$$\begin{aligned}
\widetilde{f}_k(z_1, z_2, \cdots, z_k) &= \left(1 - \sum_{i=1}^{k} \alpha_i\right) f_k(x^0) + \alpha_1 f_k(x^1) + \cdots + \alpha_k f_k(x^k) \\
&\geqslant \left(1 - \sum_{i=1}^{k} \alpha_i\right) f_{k-1}(P(x^0)) + \alpha_1 f_{k-1}(P(x^1)) + \cdots + \alpha_k f_{k-1}(P(x^k)) \\
&= \left(1 - \sum_{i=2}^{k} \alpha_i\right) f_{k-1}(P(x^1)) + \cdots + \alpha_k f_{k-1}(P(x^k)) \\
&= \widetilde{f}_{k-1}(z_2, \cdots, z_k).
\end{aligned}$$

■

Theorem 3.1. *Under assumptions <1> and <2>, let* Θ *be a random variable, uniformly distributed in* $[0,1]$*, and denote the expectation w.r.t.* Θ *by* E_Θ*. Then*

$$\lim_{N\to\infty} E_\Theta f_N(a_1^p(\Theta), ..., a_n^p(\Theta)) = \lim_{N\to\infty} \widetilde{f}_N(p, p, ..., p). \tag{18}$$

Proof We have for all n,

$$E_\Theta f_n(a_1^p(\Theta), ..., a_n^p(\Theta)) = \widetilde{f}_n(p, ..., p). \tag{19}$$

(This follows (11), from Lemma 2.5, and fact that $\widetilde{f}_n$ is affine on each atom, and agrees with f_n for the extreme points of the atom.) Since $\widetilde{f}_N(p, p, ..., p)$ is increasing in N by Lemma 3.1, the limit in N exists (it is possibly infinite). ■

We call the sequence $\{a^p(\Theta)\}$ the randomized balanced policy.

Theorem 3.2. *Under assumptions <1> and <2>, for every* $\theta \in [0,1]$*,*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f_n(a_1^p(\theta), ..., a_n^p(\theta)) \leqslant \lim_{N\to\infty} \widetilde{f}_N(p, p, ..., p). \tag{20}$$

Proof. Define

$$f_m(\theta, p) \triangleq f_m(a_1^p(\theta), ..., a_m^p(\theta)).$$

f_m is periodic (in θ) with period 1. Define

$$f'_m(\theta, p) \triangleq f_m(a_{-m+1}^p(\theta), ..., a_0^p(\theta)).$$

Then we have

$$f'_m(\theta', p) = f_m(\theta, p) \quad \text{where} \quad \theta' = \theta - mp, \tag{21}$$

Indeed,

$$\begin{aligned} f'_m(\theta',p) &= f_m(a^p_{-m+1}(\theta'),...,a^p_0(\theta')) \\ &= f_m(a^p_{-m+1}(\theta+mp),...,a^p_0(\theta+mp)) = f_m(\theta,p), \end{aligned}$$

where the last equality follows from the fact that $a^p_{-m+k}(\theta+mp) = a^p_k(\theta)$, $k=1,...,m$. f'_m is again periodic w.r.t. θ, with period 1, and is increasing in m so that the following limit exists (possibly infinite):

$$f'_\infty(\theta,p) \stackrel{\triangle}{=} \lim_{m\to\infty} f'_m(\theta,p).$$

Moreover, we have that $E_\Theta f'_m(\Theta,p) = \widetilde{f}_m(p,...,p)$, where Θ be a random variable, uniformly distributed in $[0,1]$ (this follows (11), from Lemma 2.5, and fact that $\widetilde{f}_n$ is affine on each atom, and agrees with f_n for the extreme points of the atom). Hence,

$$E_\Theta f'_\infty(\Theta,p) = \lim_{N\to\infty} \widetilde{f}_N(p,p,...,p). \tag{22}$$

Consider now the balanced sequence for fixed θ. Then

$$\begin{aligned} &\frac{1}{N}\sum_{m=1}^{N} f_m(a^p_1(\theta),...,a^p_m(\theta)) \\ &\leqslant \frac{1}{N}\sum_{m=1}^{N} f_N(a^p_{-(N-m)}(\theta),...,a^p_0(\theta),...,a^p_m(\theta)) \\ &\leqslant \frac{1}{N}\sum_{m=1}^{N} f'_\infty(-mp+\theta,p). \end{aligned}$$

The last inequality follows from assumption <2> for the functions f_k, as well as an argument similar to the one used in (21).

If p is irrational, applying the ergodic theorem of Weyl and Von Neumann ([5]), we have

$$\lim_{N\to\infty} \frac{1}{N}\sum_{m=1}^{N} f'_\infty(-mp+\theta,p) = E_\Theta f'_\infty(\Theta,p).$$

From Equation (22), we have $E_\Theta f'_\infty(\Theta,p) = \lim_{N\to\infty} \widetilde{f}_N(p,p,...,p)$. This implies that if p is irrational,

$$\lim_{N\to\infty} \frac{1}{N}\sum_{m=1}^{N} f'_\infty(-mp+\theta,p) = \lim_{N\to\infty} \widetilde{f}_N(p,p,...,p) \tag{23}$$

If p is rational, then $p=q/d$ were q and d are relatively prime and $d \geqslant 1$. This implies that the sequence $(a^p_{-(N-m)}(\theta),...,a^p_0(\theta),...,a^p_m(\theta))$ is constant if θ mod $1 \in [j/d,(j+1)/d]$, for all j. Therefore, $f'_m(\theta, p)$ is also constant on these intervals and by passage to the limit, $f'_\infty(\theta, p)$ is constant on these intervals. Now, note that $\mathrm{Frac}(\theta - mp) \in [j/d, (j+1)/d]$ for exactly one value of m out of d consecutive values of m because q are d are relatively prime.

Now, we have

$$\lim_{N\to\infty} \frac{1}{N} \sum_{m=1}^{N} f'_\infty(-mp+\theta, p) = \frac{1}{d} \sum_{m=1}^{d-1} f'_\infty(m/d, p) = E_\Theta f'_\infty(\Theta, p).$$

Equation (22) concludes this case as well. ∎

3.1 Lower Bounds

In this subsection, we establish lower bounds for the discounted cost for all sequences $\{a_k\}$. This then serves for obtaining a lower bound on the average cost. Here, we use the following assumption for the functions f_k.

- <3> For any sequence $\{a_k\}$ $\exists$ a sequence $\{b_k\}$ such that $\forall k, m, \quad k > m,\ f_k(b_1, \cdots, b_{k-m}, a_1, ..., a_m) = f_m(a_1, ..., a_m)$.

We use the notions defined in the previous sections.

Let us fix the sequence $\{a_k\}$, as well as some arbitrary integer, N. We define $p_\alpha \triangleq (1-\alpha)\sum_{k=1}^{\infty} \alpha^{k-1} a_k$.

Now, using assumptions <1> through <3>, we have

$$\begin{aligned}
&\sum_{n=1}^{\infty}(1-\alpha)\alpha^{n-1} f_n(a_1, a_2 ..., a_n) \\
&\geqslant \sum_{n=1}^{N}(1-\alpha)\alpha^{n-1} f_N(b_1, \cdots b_{N-n}, a_1, a_2 ..., a_n) + \sum_{n=N+1}^{\infty}(1-\alpha)\alpha^{n-1} f_N(a_{n-N+1}, a_2 ..., a_n) \\
&\geqslant \tilde{f}_N\left(b_1 \sum_{n=1}^{N-1}(1-\alpha)\alpha^{n-1} + \alpha^N p_\alpha, b_2 \sum_{n=1}^{N-2}(1-\alpha)\alpha^{n-1} + \alpha^{N-1} p_\alpha, \cdots, p_\alpha\right) \\
&= B(N, \alpha, p_\alpha),
\end{aligned} \tag{24}$$

where

$$B(N, \alpha, p) \triangleq \tilde{f}_N\left(b_1 \sum_{n=1}^{N-1}(1-\alpha)\alpha^{n-1} + \alpha^N p, b_2 \sum_{n=1}^{N-2}(1-\alpha)\alpha^{n-1} + \alpha^{N-1} p, \cdots, p\right) \tag{25}$$

Note that B is defined for a fixed sequence $\{a_k\}$. Also note that $B(N, \alpha, p)$ is lower semi-continuous in α for $0 \leqslant \alpha < 1$ and in p.

Using Lemma 7.1 in the Appendix, we find,

$$\begin{aligned}\overline{\lim_{m\to\infty}} \frac{1}{m}\sum_{n=1}^{m} f_n(a_1,\cdots,a_n) &\geqslant \overline{\lim_{\alpha\uparrow 1}}(1-\alpha)\sum_{n-1}^{\infty}\alpha^{n-1} f_n(a_1,\cdots,a_n)\\ &\geqslant \overline{\lim_{\alpha\uparrow 1}} B(N,\alpha,p_\alpha)\\ &\geqslant \inf_{q\in\mathcal{L}} B(N,1,q), \qquad (26)\end{aligned}$$

where $\mathcal{L}$ is the set of all limit points of p_α as $\alpha \uparrow 1$.

3.2 Optimality of the Balanced Sequences

Theorem 3.3. *Under assumptions <1>,<2> and <3>, and given some $p \in [0,1]$, and any $\theta \in [0,1]$, if the functions $f_k(a_1,\cdots,a_k)$ are increasing in all a_i, then the balanced sequence $a^p(\theta)$ minimizes the average cost $g(a)$ over all sequences that satisfy the constraint:*

$$\underline{\lim_{N\to\infty}} \frac{1}{N}\sum_{n=1}^{N} a_n \geqslant p.$$

Proof.
We denote by

$$\underline{p} \stackrel{\triangle}{=} \underline{\lim_{N\to\infty}} \frac{1}{N}\sum_{n=1}^{N} a_n$$

By using Lemma 7.1 in the Appendix,

$$p \leqslant \underline{p} \leqslant \underline{\lim_{\alpha\uparrow 1}} p_\alpha = \inf\{q, q\in\mathcal{L}\}.$$

If the functions $\{f_k\}$ are increasing, then B is increasing in p, therefore,

$$g(a) \geqslant \inf_{q\in\mathcal{L}} B(N,1,q) \geqslant B(N,1,p), \qquad (27)$$

by Equation (26). Note that for any given p, by definition of B, $B(N,1,p) = \widetilde{f}_N(p,p,\cdots,p)$. If we let N go to infinity, we get

$$g(a) = \overline{\lim_{m\to\infty}} \frac{1}{m}\sum_{n-1}^{m} f_n(a_1,\cdots,a_n) \geqslant \overline{\lim_{N\to\infty}} \widetilde{f}_N(p,p,p). \qquad (28)$$

Theorem 3.2 shows that $\overline{\lim}_{N\to\infty} \widetilde{f}_N(p,p,p) \geqslant g(a^p(\theta))$. Thus $g(a) \geqslant g(a^p(\theta))$. ∎

When the functions f_k are decreasing, we have the analogous result.

Theorem 3.4. *Under assumptions <1>,<2> and <3>, and given some $p \in [0,1]$, and any $\theta \in [0,1]$, if the functions $f_k(a_1, \cdots, a_k)$ are decreasing in all a_i, then the balanced sequence $a^p(\theta)$ minimizes the average cost $g(a)$ over all sequences that satisfy the constraint:*

$$\overline{\lim_{N\to\infty}} \frac{1}{N} \sum_{n=1}^{N} a_n \leqslant p.$$

Proof. The proof is similar to the previous one, using the fact that if

$$\overline{p} \stackrel{\triangle}{=} \overline{\lim_{N\to\infty}} \frac{1}{N} \sum_{n=1}^{N} a_n,$$

then

$$p \geqslant \overline{p} \geqslant \overline{\lim_{\alpha\uparrow 1}} p_\alpha = \sup\{q, q \in \mathcal{L}\}.$$

■

4 The optimality of balanced policies for multiple criteria

In this section, we establish general conditions under which the balanced policy is optimal when the cost function depends on multiple criteria. This has applications for routing control in several queues rather than admission control in one queue as in the previous section.

Now, we study the following general optimization problem. Consider K sequences of functions $f_n^{(i)}, i = 1, ..., K$. Each set of functions $f_n^{(i)}$ will satisfy assumptions <1>,<2> and <3>, as in Section 3.

A policy is a sequence $a = (a_1, a_2, ...)$, where a_n is a vector taking values in $\{0,1\}^K$. We consider the additional constraint that for every integer j, only one of the components of a_j may be different than 0. A policy satisfying this constraint is called feasible.

Let h be a convex increasing function from $\mathbb{R}^K$ to $\mathbb{R}$. Define

$$g(a) \stackrel{\triangle}{=} \overline{\lim_{N\to\infty}} \frac{1}{N} \sum_{n=1}^{N} h(f_n^1(a), ..., f_n^K(a)),$$

Following notations introduced in Section 3, we get a bound called $B_i(N, \alpha, p_i)$ for sequence i. Here, we denote by

$$B_i(\alpha, p_i) \stackrel{\triangle}{=} \sup_N B_i(N, \alpha, p_i),$$

and

$$B_i(p_i) \stackrel{\triangle}{=} \sup_{\alpha \leqslant 1} B_i(\alpha, p_i).$$

Note that by convexity of $f_n^{(i)}$, $B_i(p_i)$ is continuous from below.

Our objective is to minimize $g(a)$ (with no constraints on the asymptotic fractions).

Theorem 4.1. *Assume that for all i, the functions $f_n^{(i)}$ satisfy assumptions <1>,<2> and <3>. The following lower bound holds for all policies:*

$$g(a) \geqslant \inf_{p_1+\ldots+p_K=1} h(B_1(p_1), ..., B_K(p_K)).$$

Proof. Due to Lemma 7.1 in the Appendix, Jensen's inequality and Equation 24, we have

$$\begin{aligned}
&\overline{\lim_{N\to\infty}} \frac{1}{N} \sum_{n=1}^{N} h(f_n^1, ..., f_n^K) \\
\geqslant\; &\overline{\lim_{\alpha\to1}} (1-\alpha) \sum_{n=1}^{\infty} \alpha^{n-1} h(f_n^1, ..., f_n^K) \\
\geqslant\; &\overline{\lim_{\alpha\to1}}\, h\left((1-\alpha)\sum_{n=1}^{\infty}\alpha^{n-1} f_n^1, ..., (1-\alpha)\sum_{n=1}^{\infty}\alpha^{n-1} f_n^K\right) \\
\geqslant\; &\overline{\lim_{\alpha\to1}}\, h\left(B_1(\alpha, p_1^a(\alpha)), ..., B_K(\alpha, p_K^a(\alpha))\right) \qquad (29)
\end{aligned}$$

where

$$p_i^a(\alpha) \stackrel{\triangle}{=} (1-\alpha)\sum_{k=1}^{\infty} \alpha^{k-1} a_k^i. \qquad (30)$$

We note that $\sum_{i=1}^K p_i^a(\alpha) = 1$. Hence, one may choose a sequence $\alpha_n \uparrow 1$ such that the following limits exist:

$$\lim_{n\to\infty} p_i^a(\alpha_n) = p_i, \qquad i = 1, ..., K \qquad (31)$$

and $\sum_{i=1}^K p_i = 1$. From the continuity of $B_i(\alpha, p_i)$ in p and α we get from (29)

$$\begin{aligned}
g(a) &\geqslant h(B_1(1,p_1), ..., , B_K(1,p_K)) \qquad (32) \\
&\geqslant \inf_{p_1+\ldots+p_K=1} h(B_1(1,p_1), ..., B_K(1,p_K)).
\end{aligned}$$

■

Note that there exists some p^* that achieves the infimum

$$\inf_{p_1+\ldots+p_K=1} h(B_1(1,p_1), ..., B_K(1,p_K)),$$

since $h(B_1(1,p_1),...,B_K(1,p_K))$ is continuous in $p=(p_1,\cdots,p_K)$.

Consider the policy $a^{p^*}(\theta)$ given by

$$a^{p^*}_{k,i}(\theta) = \lfloor kp^*_i + \theta_i \rfloor - \lfloor (k-1)p^*_i + \theta_i \rfloor. \tag{33}$$

There are some p^* for which the condition of feasibility of the policy $a^{p^*}(\theta)$ is satisfied, that is, there exists some $\theta = (\theta_1,...,\theta_K)$, such that the policy $a^{p^*}(\theta)$ given in (33) is feasible. These p^* are called *balanced* and are more exhaustively studied in [3] and references therein.

Theorem 4.2. *Assume that h is linear nondecreasing and that p^* is balanced. Then $a^{p^*}(\theta)$ is optimal for the average cost, i.e. it minimizes $g(a)$ over all feasible policies.*

Proof. The proof follows directly from Theorem 3.2 together with Theorem 4.1. ∎

The balance condition on p^* is still not completely characterized, however, we can mention two simple cases for which p^* is balanced. i.e. for which there exist some $\theta_1,...,\theta_K$, such that $a^{p^*}(\theta)$ is feasible.

- **P1:** $K=2$.
- **P2:** K criteria with symmetric costs ($h(x)=\sum_i x_i$ and all f^i are equal).

Corollary 4.1. *(i) Consider problem P1. There exists some p such that the balanced policy is optimal for any initial phase θ.*
(ii) Consider problem P2. By symmetry, the balanced policy with $p=1/K$ is optimal for any initial phase θ.

Next, we restrict again to the case of a single objective ($K=1$), and show that the results of the previous section can be extended. More precisely, we show that balanced policy is optimal in a stronger sense.

Corollary 4.2. *Given some $p\in[0,1]$, and any $\theta\in[0,1]$, balanced policy $a^p(\theta)$ minimizes the average cost $g(a)$ over all policies that satisfy the constraint:*

$$\overline{\lim_{\alpha\to 1}}\, p_1(\alpha) \geqslant p_1. \tag{34}$$

where $p_1^a(\alpha)$ is defined in (30).

Note that the constraint $\underline{\lim}_{N\to\infty}\frac{1}{N}\sum_{n=1}^N a_n \geqslant p_1$ (in Theorem 3.3) implies (34), due to Lemma 7.1 in the Appendix. Therefore the minimization in Theorem 3.3 is over a subclass of the set of policies on which minimization is performed in Corollary 4.2. Thus, Corollary 4.2 implies that a policy a that satisfies (34) does not perform better than Hajek's policy (with $p=p_1$) even if $\underline{\lim}_{N\to\infty}\frac{1}{N}\sum_{n=1}^N a_n < p_1$.

Proof of Corollary 4.2: Choose an arbitrary policy a that satisfies (34). Choose a subsequence $\alpha_n \uparrow 1$ such that $\overline{\lim}_{n\to\infty} p_1^a(\alpha_n) = p_1$. The proof now follows by combining Theorem 3.2 with (32). ∎

5 Admission control in a queue

In this section we give a typical application of the multimodularity theory by looking at a G/G/1 queue with batch arrivals and an admission control a that must accept a given proportion p of the arriving customers. A more detailed analysis of this system can be found in [2]. The model is illustrated in Figure 3.

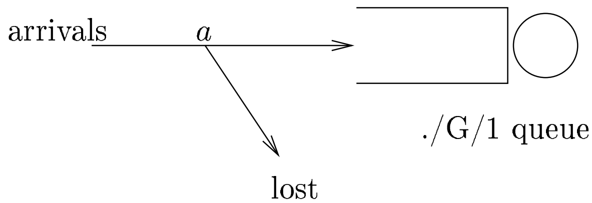

Figure 3: Admission control in a ./G/1 queue.

Let $\{T_i\}_{i\in\mathbb{N}}$ be the sequence of arrival times, with the convention that $T_1 = 0$, the queue being empty at time 0. The admission control is defined through an *arrival sequence.* The arrival sequence is a sequence of integer numbers, $a = (a_1, a_2, \cdots, a_N, \cdots)$, where a_i gives the number of customers admitted to the queue at time T_i. We also denote by $\{\sigma_k\}$ the sequence of service times in the server.

We denote by $W_k(a_1, \cdots, a_k)$ the workload in the queue at time T_k under the admission control a. Here, W_k will play the role of the functions f_k.

The following result is proved in [2]. (The special case of D/D/1 queue is analyzed in details in the next section).

Theorem 5.1. *The function $\mathbf{E}_{\sigma,T} W_k(a_1, \cdots, a_k)$ (where $\mathbf{E}_{\sigma,T}$ denotes the expectation w.r.t. the service times and the inter-arrival times) has the following properties:*

- $\mathbf{E}_{\sigma,T} W_k(a_1, \cdots, a_k)$ *is multimodular.*
- $\mathbf{E}_{\sigma,T} W_k(a_1, \cdots, a_k) \geqslant \mathbf{E}_{\sigma,T} W_m(a_{k-m+1}, \cdots, a_k) \geqslant$*, for* $k > m$*.*
- $\mathbf{E}_{\sigma,T} W_k(a_1, \cdots, a_k) = \mathbf{E}_{\sigma,T} W_m(0, \cdots, 0, a_1, \cdots, a_k)$*, for* $k < m$*.*
- $\mathbf{E}_{\sigma,T} W_k(a_1, \cdots, a_k)$ *is increasing in* a_i*.*

The expectation of (any nondecreasing convex function h of) the workload satisfies conditions <1>,<2> and <3>. The general theorem 3.3 applies and the balanced admission policy $a^p(\theta)$ with rate p minimizes the Cezaro limit

$$g(a) = \overline{\lim_{N\to\infty}} \frac{1}{N} \sum_{n=1}^{N} \mathbf{E}_{\sigma,T} W_n(a_1, ..., a_n),$$

among all policies with rate at least p.

This example can be generalized. In [2], the traveling time of a customer in an arbitrary network of queues which forms an event graph is shown to be multimodular with respect to the admission sequence. The optimality of the balanced sequence in this case is proved in [2]. More general applications of these results can also be found in a forthcoming paper [3].

6 Applications in high-speed telecommunication systems

In this section we present another illustration of the theorems that are given in sections 3 and 4 that we fully develop. Here, we consider a simple model composed of a controlled D/D/1 queue with service times $\sigma_n = \sigma$ and inter-arrival times $\tau_n = \tau$ all deterministic. Assume that the available actions are 0 (corresponding to rejecting an arriving customer) and 1 (corresponding to acceptance of an arriving customer).

The type of problem we consider is typical in high speed telecommunications networks, and in particular, to the ATM (Asynchronous Transfer Mode). The latter has been chosen by the standardization committee ITU-T [1] as the main standard for integration of services in broadband networks. In order to handle efficiently a large variety of applications, such as voice, data, video and file transfer, cells of fixed size are used, giving rise to our model that uses fixed service times. Fixed inter-arrival times are typical for isochronous applications (voice, video) and also for large file transfer.

Two important measures of quality of services in ATM networks are loss probabilities (CLR - Cell Loss Ratios) and delays. According to the ATM standard [1], when a CBR (Constant Bit Rate) session is established, the network should provide a guarantee that these two measures are bounded by given constants. Since the available sources are limited and, moreover, might be shared with other applications, a typical objective of the network is to minimize the delay of the CBR session while meeting the constraint on the loss probabilities. Losses might be due either to overflow, or to deliberate packet discarding by the network (e.g. to allow the resources to be available for other applications). The problem can be formulated in our framework as one of discarding cells so as to minimize the average queue size (i.e. the workload in the system) which is known to be proportional to the average sojourn time (due to Little's law), subject to a lower bound p on the average cell discarding rate.

We now describe the state evolution of the system. If x_n denotes the amount of workload in the system immediately after the nth arrival that occurs after time 0, and the system is initially empty (at time 0), then

$$x_{n+1} = \max(x_n - \tau, 0) + a_n\sigma.$$

The solution of this recursion is given by the expansion of the Lindley equation:

$$x_{n+1} = f_n(a_1, ..., a_n) = \max\left\{0, \sum_{k=j}^{n-1}(a_j\sigma - \tau), j = 1, ..., n-1\right\} + a_n\sigma. \tag{35}$$

We show by a simple inductive argument that f_m is indeed multimodular for all integers m over $\mathbb{Z}^m$. The function $x_0(a) \stackrel{\triangle}{=} 0$, $a \in \{0,1\}^N$ is clearly multimodular, as well as the function $x_1(a) = a_1\sigma$. Assume that $x_n = f_n(a)$ is multimodular. $x_{n+1}(a)$ is a convex increasing function of f_n, and is therefore multimodular by Theorem 2.3. Note that the assumptions $f(x+d_i) \leqslant f(x), i = 2, ..., N$, and $f(x - e_N) \leqslant f(x)$ indeed hold since $a \geqslant 0$. In particular, the assumption $f(x+d_i) \leqslant f(x), i = 2, ..., N$, follows from the fact that

$$x_{i+2} = \max(0, a_i\sigma - \tau, x_i - 2\tau + a_i\sigma) + a_{i+1}\sigma.$$

Our goal is to obtain a policy a^* that minimizes an expected average cost related to the amount of work in the system at arrival epochs. The cost to be minimized is thus

$$g(a) \stackrel{\triangle}{=} \overline{\lim_{N\to\infty}} \frac{1}{N} \sum_{n=1}^{N} f_n(a_1, ..., a_n),$$

subject to the constraint:

$$\underline{\lim_{N\to\infty}} \frac{1}{N} \sum_{n=1}^{N} a_n \geqslant p^*.$$

Consider first the case of a queue with infinite capacity. Then, it follows from Theorem 3.3 that a balanced policy (with arbitrary θ) is optimal. The assumptions of the Theorem indeed hold:

- f_n (in (35)) is indeed monotone increasing in a_i;
- Property <3> (in Subsection 3.1) holds by choosing $b_k = 0$, since

$$f_k(a_1, ..., a_k) = f_m(\underbrace{0, ..., 0}_{m-k}, a_1, ..., a_k), \; k < m; \tag{36}$$

- By combining (36) with the first monotonicity property, we get

$$f_{k-1}(a_2, ..., a_k) = f_k(0, a_2, ..., a_k) \leqslant f_k(a_1, ..., a_k),$$

  which establishes Property <2> (in the beginning of Section 3).

Consider now a queue with a finite storage capacity for the workload, i.e. the workload at the queue at each time instant is bounded by C. When the queue is full, the overflow workload is lost. Hajek's policy need not be optimal anymore, as the following example shows.

Counter-example 6.1. *(Non optimality of a balanced policy)*
Let $\tau = 1$, $\sigma = 100$, $C = 100$, $p^* = 0.01$. Assume that the cost to be minimized is the average queue length. Hajek's policy achieves an average queue length of 50.5 for any θ.

Consider now the periodic policy of period 200 that accepts 2 consecutive customers and rejects all following ones. After the second acceptance, the amount of work in the system is 100 due to the limit on the queue capacity, and there is loss of workload (of 99 units). The average queue length is 25.75. Thus the new policy achieves half the queue length as the previous one.

■

Although the balanced policy in the above counter example results in a larger queue, it has the advantage over the other policy of not creating losses. As we now show, a balanced policy is optimal if we restrict to policies with the further constraint that no losses are allowed. Thus, consider the class of policies that satisfy the constraint:

$$x_n(a_1, ..., a_n) \leqslant C$$

where x_n is given by (35).

Since x_n is multimodular for all n, the set

$$\{(a_1, ..., a_n) : \widetilde{f}_n(a_1, ..., a_n) \leqslant C\}$$

is convex for all n. Using now the remark 2.2 and Theorem 3.3, we conclude that a balanced policy is again optimal.

In the above admission control we considered only the possibility of accepting or rejecting the whole arriving batch (of 100). In practice, arriving batches may correspond to cells originating from different sources, and it is often possible to reject only a part of the batch.

Assume, thus, that the available actions are $a \in \{0, 1, ..., \overline{N}\}$, where $a = i$ means accepting $i(100/\overline{N})$ units of workload. Assume that the batch size of 100 is an integer multiple of $\overline{N}$. We can thus split an arrival batch and accept only a fraction of it; more precisely, we can either reject it, or accept $1/\overline{N}$th of the batch, or $2/\overline{N}$th, etc.... The smallest unit of batch which we can accept (i.e. $\overline{N}^{-1}$ is called a mini batch.

Consider now the balanced policy $a^*[\overline{N}]$ that is given in (17) corresponding to $p = p^*\overline{N}$. In other words, instead of considering a target fraction p of the whole batch to be accepted, which is smaller than (or equal to) 1, the new target corresponds to the average number of mini batches to be accepted, and can be any real number between 0 and $\overline{N}$ (in particular, $p = \overline{N}$ will correspond to accepting $\overline{N}$ mini-batches, i.e. the whole original batch).

We may repeat the above calculation and show that this policy is optimal for the cases of (i) the infinite queue and (ii) the bounded queue, restricting to policies that do not generate losses. Moreover, for both cases, this policy is better than the one that consists of accepting or rejecting the whole batch according to the policy a^* defined above, since a^* is a feasible policy in our new problem, for which $a^*[\overline{N}]$ is optimal (Theorem 3.3).

In order to illustrate the last point, consider $\overline{N} = 10$. A balanced policy corresponds to acceptance of a mini-batch of 10 units, once every 10 time slots. The average queue length obtained by that policy is 5.5, i.e. about ten times less than the one obtained when the whole batch was to be accepted or rejected.

7 Appendix

The following Lemma is often used in applications of optimal control (or games) with an average cost criteria (see e.g. [6]) yet it is not easy to find its proof in the literature in the format in which it has been applied.

Lemma 7.1. *Consider a sequence a_n of real numbers all having the same sign. Then*

$$\overline{\lim_{n\to\infty}} \frac{1}{N}\sum_{n=1}^{N} a_n \geqslant \overline{\lim_{\alpha\to 1}}(1-\alpha)\sum_{k=1}^{\infty}\alpha^{k-1}a_k \tag{37}$$

$$\geqslant \underline{\lim}_{\alpha\to 1}(1-\alpha)\sum_{k=1}^{\infty}\alpha^{k-1}a_k \geqslant \underline{\lim}_{n\to\infty}\frac{1}{N}\sum_{n=1}^{N} a_n \triangleq p \tag{38}$$

Proof. Note that

$$\frac{1}{1-\alpha}\sum_{k=1}^{\infty}\alpha^{k-1}a_k = \sum_{k=1}^{\infty}\left(\sum_{l=1}^{k} a_l\right)\alpha^{k-1} \tag{39}$$

$$\frac{1}{(1-\alpha)^2} = \sum_{k=1}^{\infty} k\alpha^{k-1}. \tag{40}$$

Hence

$$(1-\alpha)\sum_{k=1}^{\infty}\alpha^{k-1}a_k - p = (1-\alpha)^2\sum_{k=1}^{\infty}\left(\frac{1}{k}\sum_{l=1}^{k}a_l - p\right)k\alpha^{k-1}. \tag{41}$$

For any $\epsilon > 0$, choose N_ϵ such that

$$\frac{1}{N}\sum_{n=1}^{N} a_n \geqslant p - \epsilon$$

for all $N \geqslant N_\epsilon$. Then the right-hand side of (41) is bounded below by

$$(1-\alpha)^2\left(\sum_{k=1}^{N_\epsilon -1}\left(\frac{1}{k}\sum_{l=1}^{k}a_l - p\right)k\alpha^{k-1} - \epsilon\sum_{k=N_\epsilon}^{\infty}k\alpha^{k-1}\right)$$

$$\geqslant (1-\alpha)^2\left[\left(N_\epsilon \max_{1\leqslant k\leqslant N_\epsilon}\left|\sum_{l=1}^{k}a_l - kp\right|\right) - \epsilon(1-\epsilon)^{-2}\right] \geqslant -2\epsilon$$

for α sufficiently close to 1. This establishes (38). (37) is obtained similarly. ∎

References

[1] ITU-T Recommendation I371, "Traffic control and congestion control in B-ISDN", Perth, Nov. 1995

[2] E. Altman, B. Gaujal and A. Hordijk, "Admission Control in Stochastic Event Graph", *unpublished*, 1997.

[3] E. Altman, B. Gaujal and A. Hordijk, "Balanced Sequences and Optimal Routing", *unpublished*, 1997.

[4] P. Glasserman and D.D. Yao, "Monotone Optimal Control of Permutable GSMPs", *Math. of Oper. Res.* **19** pp. 449–476, 1994.

[5] Y.G. Sinai, "Introduction to Ergodic Theory" *Princeton University Press* Princeton, N.J., 1976.

[6] R. Sznajder and J. A. Filar, "Some comments on a theorem of Hardy and Littlewood", *J. of Optimization Theory and Appl.* **75**, No. 1, 1992, pp. 201 - 208, 1982.

[7] B. Hajek, "Extremal splitting of point processes", *Mathematics of Operations Research*, **10**, pp. 543-556, 1985.

[8] R. R. Weber and Stidham, Jr., S., "Optimal Control of Service Rates in Networks of Queues", *Advances of Applied Probability* **19**, pp. 202–218, 1987.

Unité de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unité de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unité de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unité de recherche INRIA Sophia Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
ISSN 0249-6399