JOURNAL OF OPTIMIZATION THEORY AND APPLICATIONS: Vol. 120, No. 3, pp. 533–557, March 2004 (©2004)

Path Capacity Dimensioning in a Multiprotocol Label Switched Network: Analysis of Optimal and Suboptimal Solutions[1]

C. Bruni[2], C. Scoglio[3], and S. Vergari[4]

Communicated by G. Di Pillo

Abstract. In recent years, the problem of capacity allocation for a label switched patch (LSP) in a multiprotocol label switched (MPLS) network has received great attention due to its relevance in the context of traffic control. In this paper, the problem of capacity allocation is formulated as an optimal control problem and its solution is obtained by assuming the knowledge of the bandwidth requests on the entire control interval. A suboptimal solution is also given which has the advantage of requiring limited information about future bandwidth requests. The analysis of the suboptimal solution is explored both analytically and numerically by using simulated and real data. This study demonstrates that the suboptimal solution, also with limited knowledge of the future, yields a good approximation of the optimal one and requires little additional cost.

Key Words. Internet traffic management, multiprotocol label switched networks, label switched paths, optimal and suboptimal capacity allocations.

1. Introduction

The introduction of new advanced multimedia applications in the Internet has yielded the need to provide quality of service (QoS) guarantees.

[1]The authors are indebted to Professor Ian F. Akyildiz for several useful suggestions.
[2]Professor, Dipartimento di Informatica e Sistemistica, University of Rome – La Sapienza, Rome, Italy.
[3]Researcher, Broadband and Wireless Networking Laboratory, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, Georgia.
[4]PhD Candidate, Dipartimento di Informatica e Sistemistica, University of Rome – La Sapienza, Rome, Italy.

In order to provide quantitative guarantees and an efficient use of network resources, the Internet Engineering Task Force (IETF) has proposed the differentiated services (DiffServ) architecture and several traffic engineering (TE) mechanisms (Ref. 1). The multiprotocol label switching (MPLS) technique is a novel method of TE and its combined use with DiffServ architecture can provide QoS guarantees and an efficient use of network resources (Refs. 2, 3).

MPLS is a switching technique that encapsulates packets at ingress nodes with short fixed labels and uses the label information to forward the packets along label switched paths (LSPs). Using labeling instead of long address matching, the complexity of the operations to be performed in the router is reduced and allows fast forwarding. The possibility of explicit routing of LSPs enables the choice of nonshortest paths achieving load balancing and, consequently, traffic engineering in MPLS networks (Refs. 4, 5).

The network framework consists in an Internet Protocol (IP) based network with an embedded MPLS domain. Then LSPs are characterized by the initial node, the end node, a path over the IP network with an instantaneous capacity. Eventually, an alternative path on the IP domain can be associated to the previously considered MPLS domain, at the same time, for the routing of packets which exceed the LSP capacity. It is possible to define a cost structure concerning the LSP connection and the forwarding over it. First, a cost is associated with the capacity of the LSP. A second cost must be associated as well with the variation of LSP capacity, since it requires a set of signaling operations to be performed and it can affect the dimension of other LSP in the MPLS network. A further cost term can be associated with the difference between the LSP capacity and the bandwidth request. When the bandwidth request is greater than the LSP capacity, this term takes into account the cost of exceeding packets routed on an alternative path of the IP network. When the bandwidth request is less than the LSP capacity, the same term takes into account the cost due to the wasted (reserved but not used) bandwidth. The LSP capacity allocation can be designed in order to minimize the above related costs. Obviously, the optimal solution will depend on the cost terms and their related weights. For example, if the capacity variation cost is negligible, the optimal solution is an LSP capacity that follows exactly the bandwidth requests; on the contrary, when the capacity variation cost is not negligible, the LSP optimal capacity will not follow closely the variation of the bandwidth request.

The optimal dynamic setup and dimensioning of LSP has been considered in Ref. 6 in the framework of Markov decision process theory. Under the assumption that bandwidth requests follows an M/M model, the optimal policy for LSP setup and dimensioning is derived as a control limit policy. However the above assumption is not appropriate for some Internet traffic.

In Ref. 7, we considered the problem of dimensioning a direct LSP as an optimization problem and solved it analytically by assuming the knowledge of the bandwidth requests over the whole control time interval. This is an unlikely assumption; in fact, the nature of IP traffic does not allow, in general, to predict exactly the traffic profile that will be offered to the network. However, by exploiting the very structure properties of the optimal solution, we were able to propose a suboptimal solution which requires bandwidth request knowledge reduced to a small sliding time window, centered on the current time. In fact, in the analytical expression of the optimal solution, the inverse of a modified-Toeplitz matrix appears (Refs. 8, 9): the proposed approximation is based on some properties of such matrices, proved in the Appendix (Section 6). The proposed suboptimal solution is completely independent of any stochastic assumption on the Internet traffic behavior and has the remarkable property of being almost on line. This suboptimal solution can be of interest in the context of some Internet services: in particular, there is a class of book-ahead guaranteed services which can be requested with enough notification time (Ref. 10). Examples of this class concern the access to supercomputer sites requiring high bandwidth or the booking of a pay-per-view show: the user has to request the service ahead of time by informing the service administrator one hour or more in advance.

The aim of this paper is to perform a detailed analysis of the suboptimal solution proposed in Ref. 7, considering both the approximation with respect to the optimal one and the required additional cost.

In Section 2, we restate the formulation of the optimal LSP capacity provisioning problem and provide the main results obtained in our previous paper (Ref. 7) related to optimal and suboptimal solutions. In Section 3, the main results of this paper are reported about the analysis of the suboptimal solution. The first result concerns the existence of a limited number of future bandwidth request samples which guarantees any desired level of approximation. The second result concerns the asymptotic behavior of the ratio between the number of future samples used by the suboptimal solution and the number of samples in the entire control time interval, which represent the dimension of the problem: it has been proved that this ratio goes to zero for increasing problem dimensions. Finally, the third result concerns the evaluation of the additional cost required by the suboptimal solution with respect to the optimal one. In Section 4, some concluding remarks are given both with respect to a simulated case and a real case study. The results confirmed the effectiveness of the suboptimal solution, which provides a very good approximation of the optimal one even with the knowledge of bandwidth requests on a small sliding window.

2. Optimal and Suboptimal Solutions for the LSP Capacity Provisioning Problem

Let us denote by [0, T] the time interval over which we consider the problem of the bandwidth allocation control for a generic LSP between two fixed routers in a MPLS network. We consider a uniform discretization of [0, T] and denote by $k = 1, \ldots, N$, the discrete time variable. Furthermore, let $b(k)$ and $x(k)$ denote the bandwidth request and the LSP capacity at time k, respectively. We assume that $b(k) \in [a, A]$, $k = 1, \ldots, N$, where $a > 0$ and A denotes the bandwidth availability on the LSP. We consider the LSP capacity like a simple linear discrete-time dynamical system,

$$x(k) = x(k-1) + \Delta(k), \qquad k = 1, 2, \ldots, N, \tag{1}$$

where the initial state $x(0) = x_0$ is assumed to be known and positive and the control variable $\Delta(k)$ represents the capacity variation of the LSP at time k.

The LSP capacity is constrained by the following inequalities:

$$x(k) \geq 0, \qquad k = 1, 2, \ldots, N, \tag{2}$$

which have an obvious physical meaning.

With respect to the LSP capacity provisioning problem, in Ref. 7 we considered a cost function which takes into account the most relevant cost terms and allows us to get a manageable mathematical formulation. In particular, we considered the following cost terms:

(i) LSP Cost. It takes into account the cost due to the reserved capacity used to forward packets in MPLS mode. This cost, at time k, is assumed to be proportional to the LSP capacity,

$$J_l(k) = c_l \cdot x(k), \tag{3}$$

where $c_l > 0$ is the unitary cost for LSP capacity allocation.

(ii) Excess Cost. It takes into account mainly the cost due to packets switching performed in IP mode and their routing on an alternative path that occurs when the LSP capacity is less than the bandwidth request, $x(k) < b(k)$. Forwarding packets in MPLS mode is less expensive than the IP mode: to emphasize the advantage of MPLS techniques and to promote their utilization, we consider the above mentioned cost depending quadratically on the difference between the LSP capacity and the bandwidth request. On the other hand, it may happen that, at a generic time k, the LSP capacity is greater than the bandwidth request, $x(k) > b(k)$. In this case, a certain amount of bandwidth is reserved without utilization: we have assumed to penalize this event with a cost depending again quadratically on the amount of wasted bandwidth. For simplicity, we assume the same unitary cost coefficient $c_e > 0$ both for the bandwidth switched in IP mode and for the wasted bandwidth.

From the above consideration, at time k, we have the following excess cost term:

$$J_e(k) = c_e \cdot [b(k) - x(k)]^2. \tag{4}$$

(iii) Dimensioning Variation Cost. It takes into account the LSP dimensioning variation cost. Each change of LSP capacity is charged in order to avoid too much wide LSP capacity redimensioning, which in turn affects the dimensioning of the other LSPs in the MPLS network. The same term can also take into account the so-called signaling cost which occurs at each LSP capacity variation. The dimensioning variation cost at time k is assumed to depend quadratically on the size variation of the LSP capacity,

$$J_v(k) = c_v \cdot \Delta^2(k), \tag{5}$$

where $c_v > 0$ is the unitary dimensioning variation cost of the LSP.

The total cost function for the LSP allocation in the control interval is

$$\begin{aligned} J_t &= c_l \sum_{k=1}^{N} x(k) + c_e \sum_{k=1}^{N} [b(k) - x(k)]^2 + c_v \sum_{k=1}^{N} \Delta^2(k) \\ &= c_v \left[p \sum_{k=1}^{N} x(k) + q \sum_{k=1}^{N} [b(k) - x(k)]^2 + \sum_{k=1}^{N} \Delta^2(k) \right]. \end{aligned} \tag{6}$$

Taking (1) into account and disregarding in J_t either the terms independent on $x(k)$, $k = 1, \ldots, N$, or the constant multiplicative factor c_v, we can formulate the following quadratic programming problem.

Problem P. Find a global minimum for the cost function

$$J(x) = x^T H_N x + f^T x \tag{7}$$

in the admissible set

$$D = \{x \in \mathbb{R}^N : x \geq 0\}, \tag{8}$$

where x and f are the following N-vectors:

$$x = \begin{pmatrix} x(1) \\ \vdots \\ x(N) \end{pmatrix}, \qquad f = \begin{pmatrix} p - 2q \cdot b(1) - 2x_0 \\ p - 2q \cdot b(2) \\ \vdots \\ p - 2q \cdot b(N) \end{pmatrix}, \tag{9}$$

and H_N is the following $N \times N$ matrix:

$$H_N = \begin{pmatrix} 2+q & -1 & 0 & . & 0 & 0 \\ -1 & 2+q & -1 & . & 0 & 0 \\ 0 & -1 & . & . & . & . \\ . & . & . & . & -1 & 0 \\ 0 & 0 & . & -1 & 2+q & -1 \\ 0 & 0 & . & 0 & -1 & 1+q \end{pmatrix}. \tag{10}$$

Note that H_N is indeed a modified tridiagonal Toeplitz matrix (Refs. 8, 9).

The solution of Problem P is given in the following theorem.

Theorem 2.1. Assuming $p < 2qa$, the unique solution of Problem P is

$$x^o = -(1/2)P_N f, \tag{11}$$

where $P_N = H_N^{-1}$.

Proof. From the definite positivity of H_N proved in Theorem 6.1, J is strictly convex in $\mathbb{R}^N$ and therefore the unique global minimum of J in $\mathbb{R}^N$ is the solution of the equation

$$(dJ/dx)^T_{x^o} = 2H_N x^o + f = 0,$$

that is, (11). In order to prove that (11) is also the unique solution of Problem P, we have to verify that $x^o \in D$, that is,

$$x^o(k) \geq 0, \qquad k = 1, \ldots, N.$$

The generic component of x^o is

$$x^o(k) = -(1/2) \sum_{j=1}^{N} (P_N)_{kj} f(j), \qquad k = 1, 2, \ldots, N, \tag{12}$$

and the thesis follows immediately noting that

$$-f(j) > 0, \qquad j = 1, \ldots, N,$$

as a consequence of (9) and of the assumption, and that

$$(P_N)_{kj} > 0, \qquad k, j = 1, \ldots, N.$$

This latter statement is evident from (27) observing that $\det\{H_i\} > 0$, $\det\{K_i\} > 0$, $i = 1, 2, \ldots,$ as it results from Theorem 6.1 and Remark 6.1, respectively. □

In Ref. 7, the following definition was given.

Definition 2.1. For a fixed nonnegative integer $F \leq N-1$, a suboptimal solution for Problem P is given by

$$x^{so} = -(1/2)P_{NF}f, \tag{13}$$

where P_{NF} is the $(2F+1)$-diagonal matrix of dimension $N \times N$ with entries

$$(P_{NF})_{kj} = \begin{cases} (P_N)_{kj}, & |k-j| = 0, 1, \dots, F, \\ 0, & |k-j| = F+1, \dots, N-1. \end{cases} \tag{14}$$

Remark 2.1. Note that, while the optimal solution $x^o(k)$, for each k, depends on all the samples $b(j)$, $j = 1, \dots, N$, as it appears clear in (12), the generic component $x^{so}(k)$, $k = 1, \dots, N$, of the suboptimal solution depends on no more than F components $f(j)$ of f, with $j > k$, as it results from (13). The means that $x^{so}(k)$ can be seen as an almost on line suboptimal solution depending on the knowledge of the bandwidth request on a sliding window containing no more than $2F+1$ samples centered at time k.

Remark 2.2. As far as the dependence on x_0 is concerned, we note that the optimal solution exhibits an asymptotically stable behavior, although the free evolution of the dynamical system (1) is stable but not asymptotically. In fact, from (12), we have that the contribution of the initial state to $x^o(k)$ is

$$\hat{x}^o(k) = (P_N)_{k1}x_0.$$

From (27) and (30), we have

$$\hat{x}^o(k) = [\det\{H_{N-k}\} / \det\{H_N\}]x_0 < x_0/(1-q)^k.$$

Therefore,

$$\lim_{k \to \infty} \hat{x}^o(k) = 0.$$

It is also noteworthy that, while the weight p has not any influence on the free evolution, this latter goes to zero faster when the weight q increases.

As far as the dependence of x^{so} on x_0 is concerned, from (13) and (14) it appears clearly that $x^{so}(k)$ is independent of x_0 for $k > F+1$.

3. Analysis of the Suboptimal Solution

In this section, our aim is to analyze the suboptimal solution with reference to its capability of approximating the optimal one. In particular, we will prove two main results: the first one establishes that, for each fixed N, there exists a suitable value of $F \leq N-1$ which guarantees any desired approximation level; the second one concerns the asymptotic behavior of the suboptimal solution, when the problem dimension increases; in particular, we will prove that, when N increases, the ratio F/N converges to zero. Furthermore, an upper bound for the additional cost of the suboptimal solution is given in terms of F.

A preliminary remark may be done about the sign of the difference $x^o(k) - x^{so}(k)$, $k = 1, 2, \ldots, N$. In fact, from (12), (13), and (14), we have

$$\begin{aligned} x^o(k) - x^{so}(k) &= (1/2) \sum_{j=1}^{N} [(P_N)_{kj} - (P_{NF})_{kj}][-f(j)] \\ &= (1/2) \sum_{\substack{j=1 \\ |k-j|=F+1,\ldots,N-1}}^{N} (P_N)_{kj}[-f(j)], \qquad k = 1, 2, \ldots, N. \end{aligned}$$

From the positivity of $(P_N)_{kj}$ and of $-f(j)$, $k, j = 1, 2, \ldots, N$, already pointed out in the proof of Theorem 2.1, it follows immediately that

$$x^o(k) - x^{so}(k) > 0, \qquad k = 1, 2, \ldots, N.$$

This conclusion is obvious if one considers that indeed the approximation we are interested in concerns the possibility of disregarding those terms of the summation (12) related to the future samples far from the current instant k; on the other hand, each term in (12) is strictly positive.

In order to prove the above mentioned issues, we need some preliminary results which are given in the Appendix (Section 6). The main results of this paper follow.

Theorem 3.1. For each fixed $E > 0$ and for each fixed integer $N > [2(1+q)/Q]E$, where $Q = 2qA + 2x_0 - p > 0$, a nonnegative integer $F \leq N-1$ exists such that

$$\max_{k=1,\ldots,N} (x^o(k) - x^{so}(k)) < E. \tag{15}$$

Proof. Considering the infinite norm $\|\cdot\|_\infty$, we observe that the l.h.s. of (15) is $\|x^o - x^{so}\|_\infty$; then, recalling (11) and (13) and observing that $\|f\|_\infty \leq Q$, we have

$$\begin{aligned} &\max_{k=1,\ldots,N} (x^o(k) - x^{so}(k)) \\ &= \|x^o - x^{so}\|_\infty \\ &= (1/2)\|(P_N - P_{NF})f\|_\infty \\ &\leq (1/2)\|P_N - P_{NF}\|_\infty \|f\|_\infty \leq (Q/2)\|P_N - P_{NF}\|_\infty. \end{aligned} \tag{16}$$

From Theorem 6.3, we know that

$$\|(P_N - P_{NF})\|_\infty \leq (N-F-1)(P_N)_{N-F-1,N}.$$

So, recalling also (27), we have

$$\begin{aligned} &\max_{k=1,\ldots,N} (x^o(k) - x^{so}(k)) \\ &\leq (Q/2)(N-F-1)(P_N)_{N-F-1,N} \\ &= (Q/2)(N-F-1)\det\{K_{N-F-2}\}/\det\{H_N\}. \end{aligned} \tag{17}$$

From Lemmas 6.3 and 6.1, we have

$$\begin{aligned} &\max_{k=1,\ldots,N} (x^o(k) - x^{so}(k)) \\ &\leq (Q/2)(N-F-1) \sum_{h=0}^{N-F-2} \det\{H_h\}/\det\{H_N\} \\ &< (Q/2)(N-F-1) \sum_{h=0}^{N-F-2} [1/(1+q)^{N-h}] \\ &= (Q/2)(N-F-1) \sum_{l=F+2}^{N} [1/(1+q)^l] \\ &< (Q/2)[1/(1+q)^{F+2}](N-F-1)^2. \end{aligned}$$

Let us consider the function

$$\varphi(f) = \alpha[1/(1+q)^f](N-f-1)^2, \qquad f \in \mathbb{R}, \tag{18}$$

where

$$\alpha = Q/2(1+q)^2.$$

Since $1/(1+q) < 1$, we observe that φ is a monotonically decreasing function and that

$$\varphi(-1) = QN^2/2(1+q), \qquad \varphi(N-1) = 0.$$

Then, for each $E > 0$ and $N > [(2(1+q)/Q)E]^{1/2}$, the equation

$$\varphi(f) = E \tag{19}$$

admits a unique solution

$$\tilde{f} \in (-1, N-1).$$

From the decreasing behavior of φ, it follows that the nonnegative integer

$$F = \langle \tilde{f} \rangle + 1 \leq N - 1$$

satisfies the claim, where the notation $\langle \tilde{f} \rangle$ stands for the integer part of $\tilde{f}$.

$\square$

Theorem 3.2. For each fixed $E>0$ and for each fixed integer $N>[(2(1+q)/Q)E]^{1/2}$, where

$$Q=2qA+2x_0-p>0,$$

let F be the first integer which satisfies (15). Then,

$$\lim_{N\to+\infty} F/N=0.$$

Proof. From the proof of the previous theorem, we have that F is univocally defined as

$$F=\langle\tilde{f}\rangle+1,$$

where $\tilde{f}$ is the unique solution of equation (19) belonging to the interval $(-1, N-1)$. Denoting

$$z=\tilde{f}+1\in(0,N) \quad \text{and} \quad \beta=1+q>1,$$

from the definition (18), equation (19) becomes

$$N=z+((E/\alpha\beta)\beta^{z})^{1/2}$$

From the latter equation, we deduce easily that

$$N=\beta^{z/2}[(E/\alpha\beta)^{1/2}+\beta^{\log_\beta(z)-z/2}],$$

and consequently,

$$\begin{aligned} z/N &= 2\log_\beta(N)/N-(2/N)\log_\beta((E/\alpha\beta)^{1/2}+\beta^{\log_\beta(z)-z/2}) \\ &< 2\log_\beta(N)/N-(1/N)\log_\beta(E/\alpha\beta). \end{aligned}$$

Therefore, taking into account that $z/N>0$, we have

$$\begin{aligned} 0 &\le \lim_{N\to+\infty} z/N \\ &\le \lim_{N\to+\infty}[2\log_\beta(N)/N-(1/N)\log_\beta(E/\alpha\beta)]=0. \end{aligned}$$

From the definitions of z and F, the claim is proved. □

Theorem 3.1 is interesting because it assesses that a suboptimal solution exists allowing any desired level of approximation with respect to the optimal one. Moreover, Theorem 3.2 stresses that the advantages of the suboptimal solution, related to the minor information requested about the future, increase with the dimension of the problem; in fact, F is infinitesimal with respect to N as N increases without bound.

Finally note that, once N, E, Q are fixed, an a priori computation of F is possible by verifying the following inequality:

$$(Q/2)(N-F-1)(P_N)_{N-F-1,N} \leq E, \tag{20}$$

where $(P_N)_{N-F-1,\,N}$ may be computed by (27) for increasing values of F.

This possibility will be exploited in the next section, where some examples will be presented for both simulated and real data.

Remark 3.1. As far as the additional cost of the suboptimal solution is concerned, from (6) and (11), it follows that

$$\begin{aligned} 0 &\leq J(x^{so}) - J(x^o) \\ &= c_v (x^o - x^{so})^T H_N (x^o - x^{so}) \\ &\leq c_v \|H_N\|_\infty \|x^o - x^{so}\|_\infty^2. \end{aligned}$$

Recalling (10) and (17), we have

$$J(x^{so}) - J(x^o) \leq c_v (4+q)(Q^2/4)\|P_N - P_{NF}\|_\infty^2 = C. \tag{21}$$

From (21), it appears clear that the upper bound C for the additional cost depends on F and that, as it is obvious, it decreases when F approaches $N-1$. In particular, for $F = N-1$, we have $C = 0$, that is, $J(x^{so}) = J(x^o)$; in fact, in this case, the suboptimal solution coincides with the optimal one.

4. Application to Simulated and Real Data

In this section, we test the capabilities of the optimal and suboptimal LSP capacity allocation procedure by exploiting both simulated and real data. For this, we consider a case study obtained by suitably simulating a sequence of bandwidth requests. Moreover, a second case study, concerning real data traffic, is presented also in order to demonstrate the effectiveness of the proposed suboptimal solution.

To generate the simulated bandwidth request profile, we used three stochastic processes:

(a) the first generates the requests arrival times and is simulated as a Poisson process with parameter $\lambda = 1/2$;
(b) the second concerns the time duration of each request and is characterized by an exponential distribution with parameter $\mu = 1/10$;
(c) the last is related to the amount of bandwidth of each request and follows a uniform distribution on the integers of the interval [1, 10].

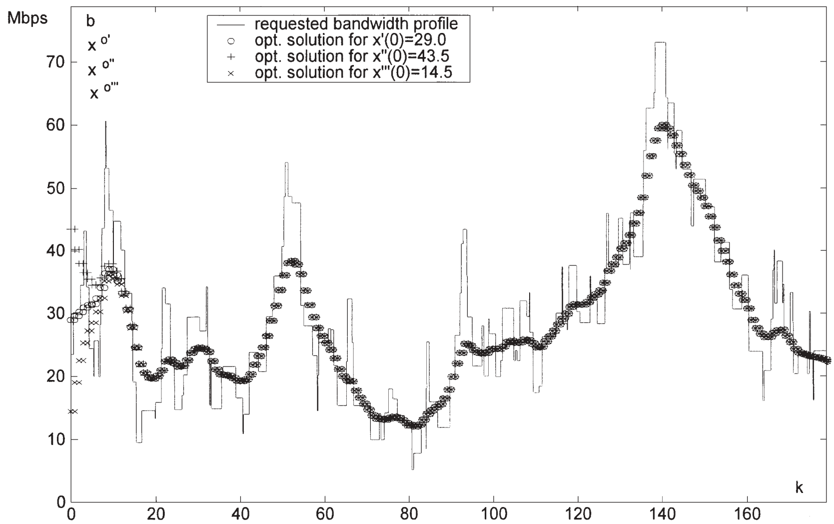

Fig. 1. Optimal solutions for three different values of x_0 (simulated data, $p = q = 0.1$).

On the generated bandwidth profile, we selected a time window containing $N = 180$ samples. We observe that, for the generated bandwidth profile, we can assume $a = 5$, so that the assumption of Theorem 2.1 becomes $p > 10q$.

With reference to the above simulated data, the analysis of the optimal and suboptimal solution will deal with the following three points:

(A) the sensitivity of the optimal solution with request to the initial condition choice;
(B) the influence of weights p, q on the optimal and suboptimal solutions;
(C) the approximation level given by the suboptimal solution, as a function of F, with respect to the optimal solution in terms of both the maximum error and additional cost.

The analysis of the influence of the initial condition on the optimal solution is performed on the basis of a fixed pair of parameters $p = q = 0.1$. In Fig. 1, the behavior of the optimal solution is shown with respect to three different values of x_0, namely:

$x_0' = 29.0$, which is the mean value of the bandwidth request over the entire control time interval;
$x_0'' = 43.5$, which corresponds to a perturbation of +50% with respect to x_0';

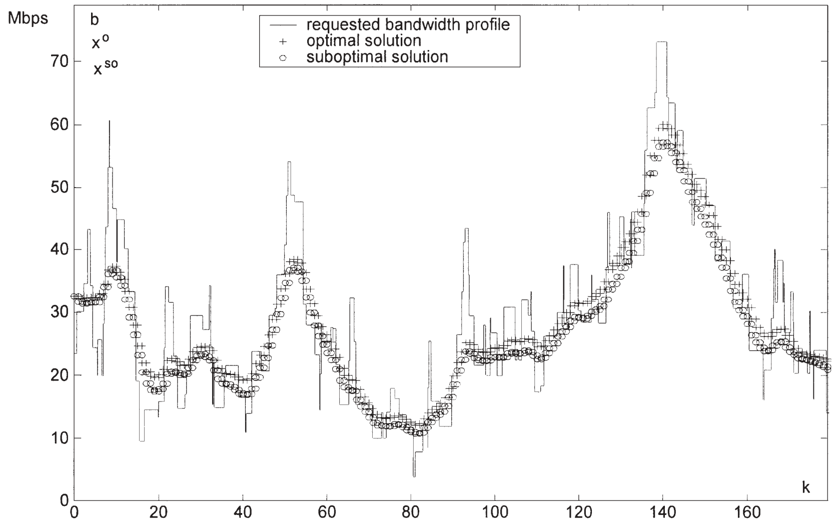

Fig. 2. Optimal and suboptimal solutions (simulated data, $p = q = 0.1$, $F = 8$, $x_0 = 32.5$).

$x_0''' = 14.5$, which corresponds to a perturbation of –50% with respect to x_0'.

As expected from Remark 2.2, the effect of x_0 vanishes for increasing k; from Fig. 1, it is evident that the choice of x_0 influences the behavior of the optimal solution only on a small initial subinterval and its effect runs out very quickly.

As far as the second issue is concerned, we have considered the values $p = q = 0.1$ as a reference; for the suboptimal solution, we fixed $F = 8$; both the optimal and suboptimal solution have been computed assuming $x_0 = 32.5$, which is the mean value of the first eight samples of the requested bandwidth profile. In Fig. 2, the optimal and the suboptimal solutions are reported for the reference case. We have considered three comparison cases obtained by increasing the coefficients p, q one at the time or by decreasing both, in any case satisfying the inequality $p < 10q$. The results thus obtained are shown in Figs. 3, 4, 5, respectively. The following remarks follow:

(i) Comparing the results of Figs. 2 and 3, we observe that an increase of p gives rise to a worse fitting of x^o with respect to the requested bandwidth profile; that occurs because of the increase of the unitary cost of the LSP capacity allocation. The same influence occurs on the suboptimal solution. Moreover, the approximation of x^{so} with respect to x^o is virtually unchanged because p only influences x^o and x^{so} through the vector f. On the other hand,

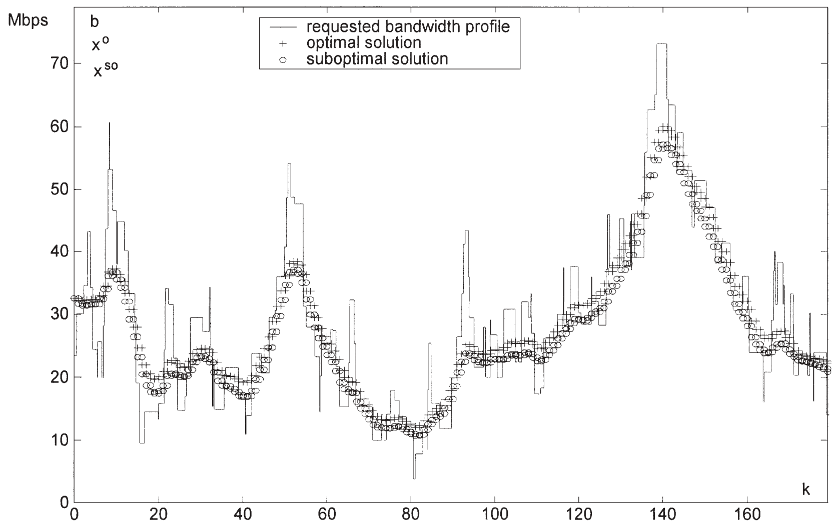

Fig. 3. Optimal and suboptimal solutions (simulated data, $p = 0.9$, $q = 0.1$, $F = 8$, $x_0 = 32.5$).

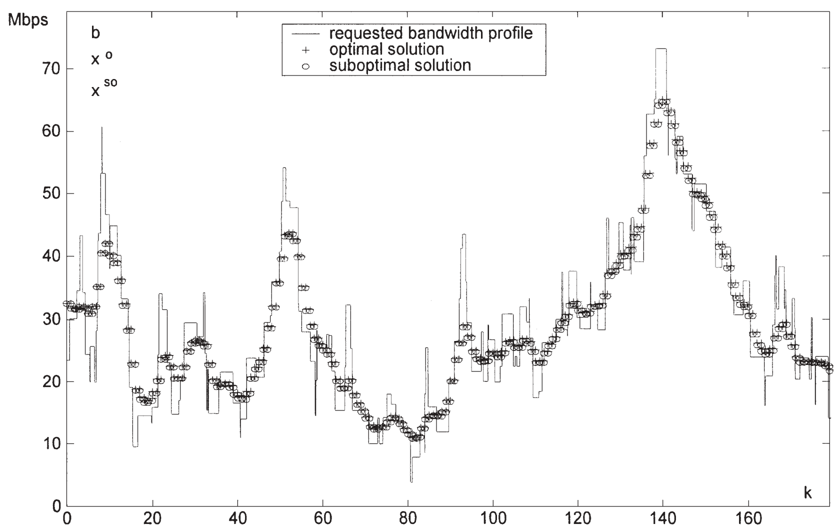

Fig. 4. Optimal and suboptimal solutions (simulated data, $p = 0.1$, $q = 0.3$, $F = 8$, $x_0 = 32.5$).

the approximation error $x^o - x^{so}$ contains components of f which, for each value of p, are multiplied by very small weights provided that $\|P_N - P_{NF}\|_\infty$ is small too.

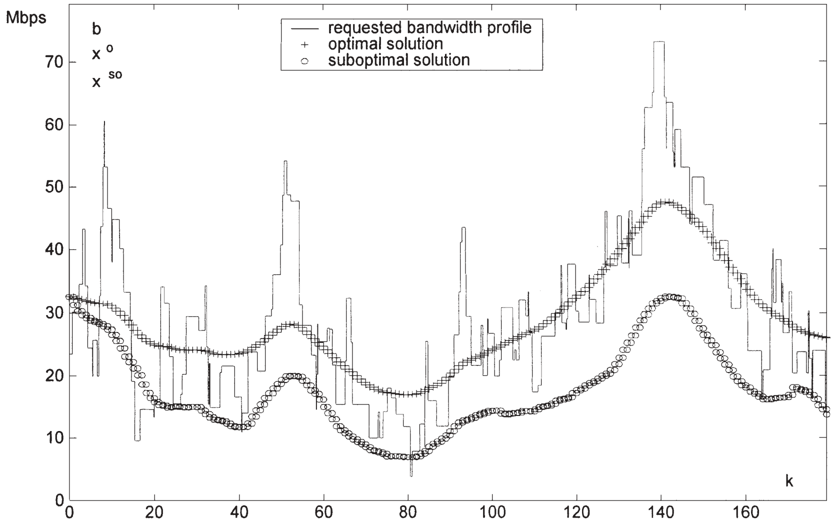

Fig. 5. Optimal and suboptimal solutions (simulated data, $p = q = 0.01$, $F = 8$, $x_0 = 32.5$).

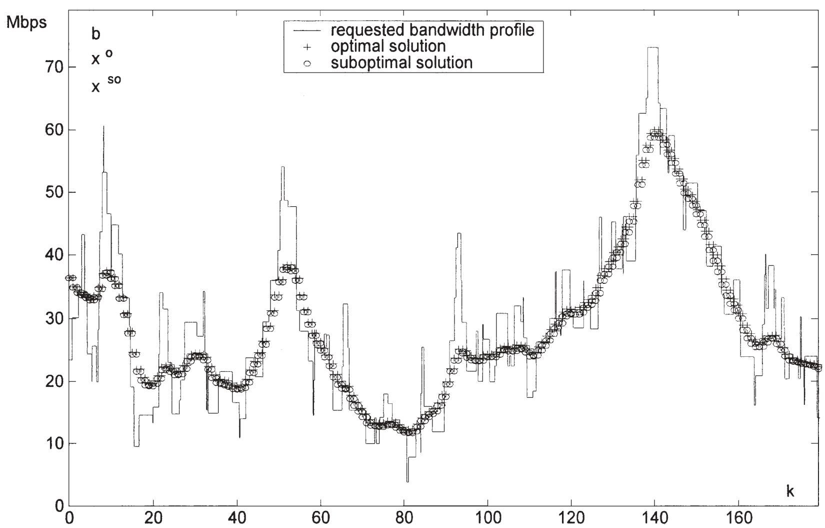

Fig. 6. Optimal and suboptimal solutions (simulated data, $p = q = 0.1$, $F = 12$).

(ii) Comparing the results of Figs. 2 and 4, we observe that an increase of q corresponds obviously to a better fitting of x^o with respect to the requested bandwidth profile due to the increase of the unitary cost of the

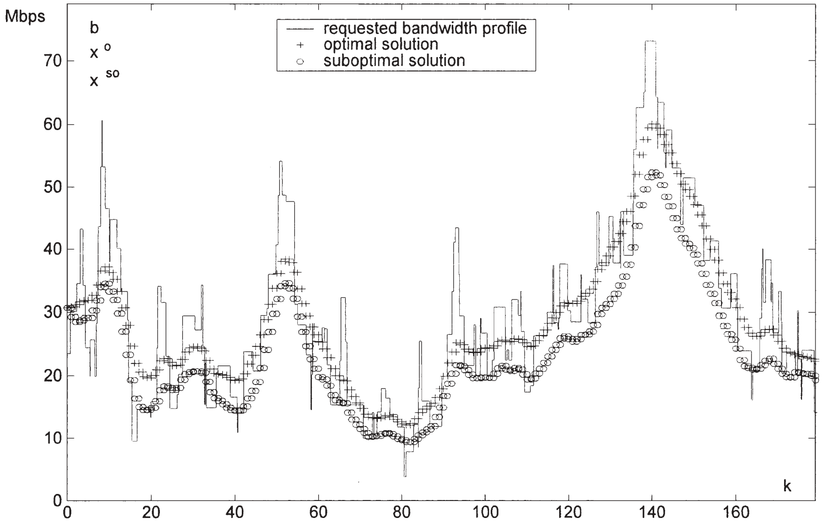

Fig. 7. Optimal and suboptimal solutions (simulated data, $p = q = 0.1$, $F = 5$).

term J_e. Furthermore, the approximation of x^{so} with respect to x^o improves because q influences P_N in such way that P_N approaches the identity matrix when q increases.

(iii) Comparing Figs. 2 and 5, we observe that a consistent decrease of both p and q corresponds to a smoothing effect on x^o and x^{so}; this happens because the dimensioning variation cost becomes dominant on the other cost terms. Moreover, the approximation of x^{so} with respect to x^o gets worse because $\|P_N - P_{NF}\|_\infty$ increases when q decreases.

As far as the third point is concerned, with respect to the reference case in Fig. 2, we have considered two comparison cases, the first one increasing F (that is the amount of future knowledge) and the second one decreasing it; both for the optimal and suboptimal solution, the initial condition x_0 is changed according to the F variations. The results obtained are shown in Figs. 6 and 7, respectively.

Comparing Figs. 2, 6, 7, it appears that an increase of F obviously corresponds to an improvement for the approximation of x^{so} with respect to x^o. In order to quantify the above remark, by exploiting inequality (20) for our particular numerical parameters choice ($p = q = 0.1$, $Q = 100$), we have computed the minimum values of F which guarantees a desired upper bound E for the approximation error, for $N = 200$:

$$E = 0.01,\quad 0.05,\quad 0.1,\quad 0.5,\quad 1,\quad 3,\quad 5,\quad 8,\quad 11,\quad 15, \tag{22a}$$

$$F = 46,\quad 41,\quad 39,\quad 34,\quad 31,\quad 28,\quad 26,\quad 25,\quad 24,\quad 23. \tag{22b}$$

Comparing Figs. 2, 6, 7 with (22), the twofold use of inequality (20) is confirmed: for a given N, one can fix the admissible upper bound E for the approximation error and consequently foresee the needed value of F or inversely.

A second issue of interest is related to the evaluation of the additional cost of the suboptimal solution with respect to the optimal one. With reference to the cases of Figs. 2, 6, 7 and assuming $c_v = 1$, we have computed the values of the additional cost $\Delta J = J(x^{so})–J(x^o)$ and its percent values:

$$F = \quad 5, \qquad\qquad 8, \qquad\qquad 12, \tag{23a}$$

$$\Delta J = 488.6\,(36.9\%), \quad 71.8\,(5.4\%), \quad 5.4\,(0.4\%). \tag{23b}$$

We observe from (23) that the numerical values of ΔJ quickly decrease for increasing F; finally, we stress that, also for small values of F, the percent additional cost keeps to reasonable values.

In order to have a case study based on real data, we have considered a sample of the traffic on the New York–Cleveland link of the Abilene network from the viewpoint of the Cleveland router in the time interval from 16th January 2002 at 6.00AM to 20th January 2002 at 9.30AM. The discrete time requested bandwidth was obtained by averaging the incoming bits per second at the Cleveland router on 30 minutes subintervals, thus obtaining a string of 200 samples (the data have been obtained from the web site http://hydra.uits.iu.edu/~abilene/traffic).

Although the above considered traffic data are only the recording of actual traffic without any forecasting, it is possible to use the same data a posteriori in the context of book-ahead service, thus simulating the knowledge of the future bandwidth request on a limited subinterval. From this point of view, we assumed the traffic request with $T_F = 2.5$ hours and $T_F = 4$ hours in advance, corresponding to $F = 5$ and $F = 8$, respectively, over a control time interval of 100 hours, that is, $N = 200$. Optimal and suboptimal solutions have been computed in both cases and they are reported in Figs. 8 and 9, respectively. Once again it appears that the approximation improves when F increases.

Finally, we have tested the robustness of the proposed suboptimal solution by considering several different traffic traces: in all these cases, we observed that the maximum errors were very little and substantially comparable; the same was true for the additional costs.

5. Concluding Remarks

In this paper, we have considered the problem of the capacity dimensioning for a LSP in a MPLS network, assuming to satisfy the bandwidth request by forwarding packets in IP or MPLS mode.

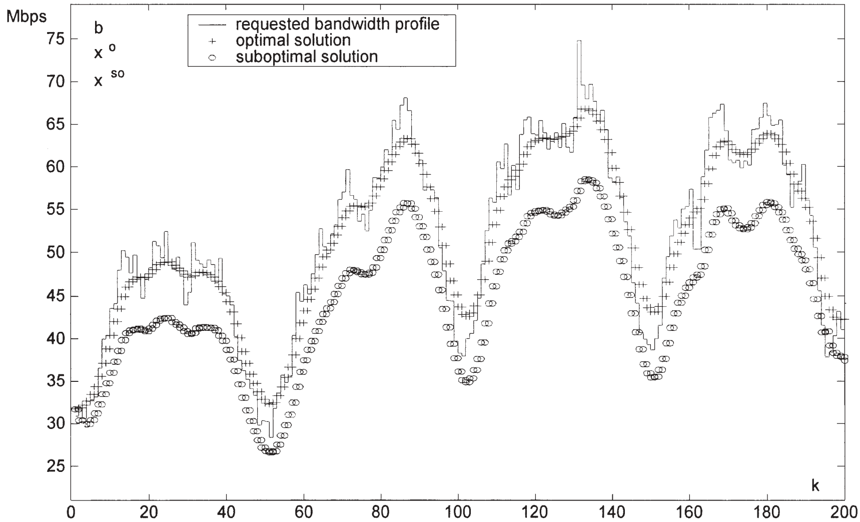

Fig. 8. Optimal and suboptimal solutions (real data, $p = q = 0.1$, $F = 5$, $T_F = 2.5$ h).

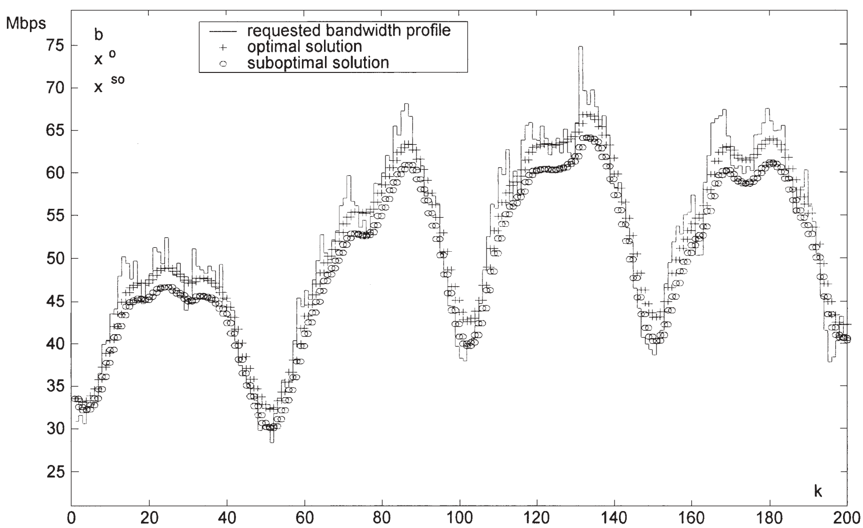

Fig. 9. Optimal and suboptimal solutions (real data, $p = q = 0.1$, $F = 8$, $T_F = 4$ h).

An optimal control problem is formulated and solved by assuming complete information about the future bandwidth request over the control time interval. In order to put this work in a more realistic framework, and in particular in the context of band-booking with a fixed advance (e.g., pay per view service), we have proposed a suboptimal solution which deserves the advantage of requiring the knowledge of a narrow sliding window on the bandwidth request profile, centered at the current time.

The main objective of this paper consists in an in-depth analysis of the suboptimal solution with respect to the optimal one, with reference to both the approximation capability and the additional cost. The analysis has been performed analytically and the results achieved have been validated by numerical case studies related to simulated and real traffic data.

This study allows to state that a limited knowledge of the future bandwidth request with respect to the entire control time interval is enough to yield a suboptimal solution characterized by negligeable approximation errors, as well as very limited additional cost with respect to the optimal one.

6. Appendix

In this appendix, we consider a particular tridiagonal symmetric Toeplitz matrix of dimension i,

$$K_i = \begin{pmatrix} 2+q & -1 & 0 & . & 0 \\ -1 & 2+q & . & . & . \\ 0 & . & . & -1 & 0 \\ . & . & -1 & 2+q & -1 \\ 0 & . & 0 & -1 & 2+q \end{pmatrix}, \quad q>0, \qquad i=1,2,\ldots,$$

and a tridiagonal symmetric modified Toeplitz matrix of dimension i,

$$H_i = \begin{pmatrix} 2+q & -1 & 0 & . & 0 \\ -1 & 2+q & . & . & . \\ 0 & . & . & -1 & 0 \\ . & . & -1 & 2+q & -1 \\ 0 & . & 0 & -1 & 1+q \end{pmatrix}, \quad q>0, \qquad i=2,3,\ldots.$$

Results will be given related to their definite positivity, the expression of their determinants, and their inverses.

Theorem 6.1. The $i \times i$ matrix

$$H_i = \begin{pmatrix} 2+q & -1 & 0 & . & 0 & 0 \\ -1 & 2+q & -1 & . & 0 & 0 \\ 0 & -1 & . & . & 0 & 0 \\ . & . & . & . & . & 0 \\ 0 & 0 & . & . & 2+q & -1 \\ 0 & 0 & . & 0 & -1 & 1+q \end{pmatrix}, \qquad i = 2, 3, \ldots, \tag{24}$$

is definite positive for $q > 0$.

Proof. Let us set

$$\alpha = 2 + q, \qquad \beta = 1 + q,$$

and for $x \in \mathbb{R}^i$, consider the quadratic form

$$x^T \cdot H_i \cdot x = (x_1 \vdots x_2 \ \ldots \ x_i) \begin{pmatrix} \alpha & \vdots & -1 & 0 & . & 0 \\ \cdots & . & \cdots & \cdots & \cdots & \cdots \\ -1 & \vdots & & & & \\ 0 & \vdots & & H_{i-1} & & \\ . & \vdots & & & & \\ 0 & \vdots & & & & \end{pmatrix} \begin{pmatrix} x_1 \\ \cdots \\ x_2 \\ \vdots \\ x_i \end{pmatrix}$$

$$= \alpha x_1^2 - 2x_1x_2 + (x_2 \quad x_3 \ \ldots \ x_i) H_{i-1} (x_2 \quad x_3 \ \ldots \ x_i)^T. \tag{25}$$

By iterating the previous argument, we have

$$x^T \cdot H_i \cdot x = \alpha(x_1^2 + x_2^2 + \cdots + x_{i-2}^2) - 2(x_1x_2 + x_2x_3 + \cdots + x_{i-2}x_{i-1})$$

$$+ (x_{i-1} x_i) \begin{pmatrix} \alpha & -1 \\ -1 & \beta \end{pmatrix} \begin{pmatrix} x_{i-1} \\ x_i \end{pmatrix}$$

$$= \alpha(x_1^2 + x_2^2 + \cdots + x_{i-2}^2 + x_{i-1}^2) + \beta x_i^2 - 2(x_1x_2 + x_2x_3 + \cdots + x_{i-1}x_i)$$

$$= q(x_1^2 + x_2^2 + \cdots + x_i^2) + x_1^2 + (x_1 - x_2)^2 + (x_2 - x_3)^2 + \cdots + (x_{i-1} - x_i)^2.$$

It follows that

$$x^T H_i x > 0, \quad \forall x \in \mathbb{R}^i, \qquad x \neq 0. \qquad \square$$

Remark 6.1. Note also that the $i \times i$ matrix

$$K_i = \begin{pmatrix} 2+q & -1 & 0 & . & 0 \\ -1 & 2+q & . & . & . \\ 0 & . & . & -1 & 0 \\ . & . & -1 & 2+q & -1 \\ 0 & . & 0 & -1 & 2+q \end{pmatrix}, \qquad i = 1, 2, \ldots, \tag{26}$$

for $q > 0$ is positive definite. The proof of this statement is similar to the proof of previous theorem.

The expression of the inverse of H_i is known; letting $P_i = H_i^{-1}$, we have (Ref. 9)

$$(P_i)_{kj} = [1/\det\{H_i\}]\det\{K_{\min\{k,j\}-1}\} \cdot \det\{H_{i-\max\{k,j\}}\}. \tag{27}$$

For the computation of the determinant of K_i, an iterative formula is given in Ref. 9,

$$\det\{K_i\} = (2+q)\det\{K_{i-1}\} - \det\{K_{i-2}\}, \qquad i = 1, 2, \ldots, \tag{28}$$

where

$$\det\{K_0\} = 1 \quad \text{and} \quad \det\{K_{-1}\} = 0.$$

We now provide a similar formula for $\det\{H_i\}$.

Theorem 6.2. Defining $\det\{H_0\} = 1$ and $\det\{H_1\} = 1 + q$, it results that

$$\det\{H_i\} = (2+q)\det\{H_{i-1}\} - \det\{H_{i-2}\}, \qquad i = 2, 3, \ldots. \tag{29}$$

Proof. Taking into account the partition of H_i, given in (25), we have

$$\begin{aligned}
\det\{H_i\} &= \det\{H_{i-1}\}\det\{\alpha - (-1\ 0\ \ldots\ 0)P_{i-1}(-1\ 0\ \ldots\ 0)^T\} \\
&= \det\{H_{i-1}\}\det\{\alpha - (P_{i-1})_{1,1}\} \\
&= \det\{H_{i-1}\}\det\{\alpha - \det\{H_{i-2}\}/\det\{H_{i-1}\}\} \\
&= (2+q)\det\{H_{i-1}\} - \det\{H_{i-2}\}, \qquad i = 2, 3, \ldots. \qquad \square
\end{aligned}$$

Lemma 6.1. The following useful inequality holds:

$$\det\{H_{i-h}\}/\det\{H_i\} < 1/(1+q)^h, \qquad i = 2, 3, \ldots, \; h = 1, 2, \ldots, i. \tag{30}$$

Proof. First of all, by exploiting (29), we have

$$\begin{aligned} \det\{H_i\} - \det\{H_{i-1}\} &= (1+q)\det\{H_{i-1}\} - \det\{H_{i-2}\} \\ &> \det\{H_{i-1}\} - \det\{H_{i-2}\} \\ &> \ldots \\ &> \det\{H_1\} - \det\{H_0\} > q > 0, \qquad i = 2, 3, \ldots . \end{aligned} \tag{31}$$

From (31), it follows that

$$\det\{H_i\} > \det\{H_{i-1}\};$$

therefore, taking (29) into account, it results that

$$\begin{aligned} \det\{H_i\} &> (2+q)\det\{H_{i-1}\} - \det\{H_{i-1}\} = (1+q)\det\{H_{i-1}\} \\ &> \ldots \\ &> (1+q)^h \det\{H_{i-h}\}, \qquad i = 2, 3, \ldots, \; h = 1, 2, \ldots, i. \end{aligned}$$

□

Lemma 6.2. Letting

$$R_i^H = \det\{H_i\}/\det\{H_{i-1}\}, \qquad i = 1, 2, \ldots, \tag{32}$$

$$R_i^K = \det\{K_i\}/\det\{K_{i-1}\}, \qquad i = 1, 2, \ldots, \tag{33}$$

the following recursive formulas hold:

$$R_{i+1}^H = (2+q) - 1/R_i^H, \qquad i = 1, 2, \ldots, \tag{34}$$

$$R_{i+1}^K = (2+q) - 1/R_i^K, \qquad i = 1, 2, \ldots . \tag{35}$$

Moreover, the sequence $\{R_i^H\}_{i=1,2,\ldots}$ is monotone increasing, while the sequence $\{R_i^K\}_{i=1,2,\ldots}$ is monotone decreasing.

Proof. From the recursive formulas (29) and (28), and from (32), (33), the recursive formulas (34) and (35) follow respectively. The sequence $\{R_i^H\}_{i=1,2,\ldots}$ is monotone increasing; in fact, by direct computation we have $R_2^H > R_1^H$. By exploiting (34), it follows that

$$R_2^H > R_1^H \quad \Rightarrow \quad R_3^H > R_2^H \quad \Rightarrow \cdots \Rightarrow \quad R_{i+1}^H > R_i^H, \qquad i = 1, 2, \ldots .$$

The proof of the decreasing behavior of $\{R_i^K\}_{i=1,2,\ldots}$ can be given along the same line. □

Theorem 6.3. For each integer $i>1$ and for each nonnegative integer $l \leq i-1$, we have

$$\|P_i - P_{il}\|_\infty = \max_{k=1,\dots,i} \left\{ \sum_{j=1}^{i} |P_i - P_{il}|_{k,j} \right\}$$
$$\leq (i-l-1)(P_i)_{i-l-1,\, i} = (i-l-1)(P_i)_{i,\, i-l-1}, \tag{36}$$

where P_{il} is the $(2l+1)$-diagonal matrix of dimension $i \times i$ defined according to (14).

Proof. The proof is in two steps. The first one shows that

$$(P_i)_{k,j} > (P_i)_{k,j+1}, \qquad k=1,\dots,i-2 \text{ and } j=k,\dots,i-1. \tag{37}$$

In fact, from (31), it is immediate to verify that

$$[1/\det\{H_i\}]\det\{K_{k-1}\}\det\{H_{i-j}\} > [1/\det\{H_i\}]\det\{K_{k-1}\}\det\{H_{i-(j+1)}\},$$

and recalling (27), we see that (37) is proved.

The second step is to show that

$$(P_i)_{k,j} < (P_i)_{k+1,j+1}, \qquad k=1,\dots,i-1 \text{ and } j=1,\dots,i-1. \tag{38}$$

Recalling (32) and (33), by direct computation we have that $R_1^K > R_1^H$. From the recursive formulas (34) and (35), it is easy to show that $R_i^K > R_i^H$, $i=2,3,\dots$. Recalling now the monotone properties of the sequences $\{R_i^H\}_{i=1,2,\dots}$ and $\{R_i^K\}_{i=1,2,\dots}$, proved in Lemma 6.2, it follows that

$$R_k^K > R_h^H, \qquad k=1,\dots,i \text{ and } h=1,\dots,i,$$

and even more it is true that

$$R_k^K > R_h^H, \qquad k=1,\dots,i-1 \text{ and } h=1,\dots,i-k.$$

Defining $h=i-j$, the previous inequality is equivalent to

$$R_k^K > R_{i-j}^H, \qquad k=1,\dots,i-1 \text{ and } j=k,\dots,i-1,$$

which, recalling (32), (33), (27), and taking into account the symmetry property of the inverse matrix P_i, proves (38) by easy computation.

Finally, from (37) and (38), the thesis (36) follows immediately. □

Lemma 6.3. For the matrices H_i and K_i, introduced by (24) and (26), respectively, the following relation holds:

$$\det\{K_i\} = \sum_{h=0}^{i} \det\{H_h\}. \tag{39}$$

Proof. We will prove the thesis by induction over three steps. From (29) and (28), the thesis follows for $i = 1$; in fact,

$$\begin{aligned}\det\{K_1\} &= 2 + q = 1 + (1 + q)\\ &= 1 + \det\{H_1\} = \sum_{h=0}^{1} \det\{H_h\}.\end{aligned}$$

Again from (29) and (28), the thesis is also true for $i = 2$,

$$\begin{aligned}\det\{K_2\} &= (2 + q)\det\{K_1\} - \det\{K_0\}\\ &= (2 + q)(1 + \det\{H_1\}) - \det\{H_0\}\\ &= \det\{H_2\} + (1 + q) + 1\\ &= \sum_{h=0}^{2} \det\{H_h\}.\end{aligned}$$

Similarly, for $i = 3$, we have

$$\begin{aligned}\det\{K_3\} &= (2 + q)\det\{K_2\} - \det\{K_1\}\\ &= (2 + q)(1 + \det\{H_1\} + \det\{H_2\}) - (1 + \det\{H_1\})\\ &= \det\{H_3\} + \det\{H_0\} + \det\{H_1\} + \det\{H_2\}\\ &= \sum_{h=0}^{3} \det\{H_h\}.\end{aligned}$$

Assuming that (39) is true for $i - 1$, $i - 2$, and $i - 3$, we now verify that it is also true for each $i > 3$:

$$\begin{aligned}\det\{K_i\} &= (2 + q)\det\{K_{i-1}\} - \det\{K_{i-2}\}\\ &= (2 + q)\sum_{h=0}^{i-1} \det\{H_h\} - \sum_{h=0}^{i-2} \det\{H_h\}\\ &= (2 + q)\sum_{h=0}^{i-2} \det\{H_h\} + (2 + q)\det\{H_{i-1}\} - \sum_{h=0}^{i-3} \det\{H_h\} - \det\{H_{i-2}\}\\ &= (2 + q)\det\{K_{i-2}\} - \det\{K_{i-3}\} + \det\{H_i\}\\ &= \det\{K_{i-1}\} + \det\{H_i\}\\ &= \sum_{h=0}^{i-1} \det\{H_h\} + \det\{H_i\}\\ &= \sum_{h=0}^{i} \det\{H_h\}.\end{aligned}$$

□

References

1. Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and Weiss, W., *An Architecture for Differentiated Services*, Internet Engineering Task Force, RFC 2475, 1998.
2. Trimintzios, P., Griffin, D., Georgatsos, P., Goderis, Egan, R., L., Jacquenet, C., and Georgiadis, D., *A Management and Control Architecture for Providing IP Differentiated Services in MPLS-Based Networks*, IEEE Communications Magazine, Vol. 39, pp. 80–88, 2001.
3. Awduche, D. O., Malcolm, J., Agogbua, J., O'Dell, M., and McManus, J., *Requirement for Traffic Engineering over MPLS*, Internet Engineering Task Force, RFC 2702, 1999.
4. Rosen, E., Viswanathan, A., and Callon, R., *Multiprotocol Label Switching Architecture*, Internet Engineering Task Force, RFC 3031, 2001.
5. Gonzales, F., Chang, C., Chen, L., and Lin, C., *Using MultiProtocol Label Switching (MPLS) to Improve IP Network Traffic Engineering*, Proceedings of Interdisciplinary Telecommunications Program, 2000.
6. Anjali, T., Scoglio, C., De Oliveira, J. C., Akyildiz, I. F., and Uhl, G., *Optimal Policy for Label Switched Path Setup in MPLS Networks*, Computer Networks Journal, Vol. 39, pp. 165–193, 2002.
7. Bruni, C., Scoglio, C., and Vergari, S., *Optimal Capacity Provisioning for Label Switched Paths in MPLS Networks*, Proceedings of Networking 2002, Pisa, Italy; Lecture Notes in Computer Science, Springer Verlag, Berlin, Germany, Vol. 2345, 2002.
8. Bruni, C., and Vergari, S., *Controlled Approximation for the Inverses of Special Toeplitz and Modified Toeplitz Matrices*, Calcolo (to appear).
9. Fischer, C. F., and Usmani, R. A., *Properties of Some Tridiagonal Matrices and Their Application to Boundary-Value Problems*, SIAM Journal on Numerical Analysis, Vol. 6, pp. 127–142, 1969.
10. Thirimalasetty, S. R., and Medhi, D., *MPLS Traffic Engineering for Survivable Book-Ahead Guaranteed Services*, Computer Science Telecommunications Technical Report, University of Missoury, Kansas City, 2001.