Worst-Case Rate-Based Flow Control with an ARMA Model of the Available Bandwidth

Eitan ALTMAN*, Tamer BAŞAR† and Naira HOVAKIMYAN ‡

Abstract

A central control problem that arises in High speed networks is the control of the rate of flow of information into the network. A rate that is too high may result in congestion and hence in the degradation of performance measures: high delays within the network, high loss probabilities of information packets and large delay variations; a rate that is too low may result in the under-utilization of the network and in low throughput. Most of the existing and future telecommunications networks therefore include dynamic flow control mechanisms, which have often been developed on growing available experience, using ad-hoc techniques that did not come as a result of a control-theoretical study. This is due to the high complexity of the controlled systems, that are typically decentralized, have non-linear dynamics, and may only use partial noisy delayed information. Some attempts have been made in recent years to use control theory to design flow controllers with, however, no explicit objective functions to be minimized; moreover, the class of control policies in existing theoretical work is quite restricted. In this paper we formulate explicitly some cost criteria to be minimized, related to performance measures mentioned above, including delays, throughputs and loss probabilities. We present an approximating linearized model with quadratic cost, and follow two approaches to model interfering traffic and other unknown data: the H^∞ approach, and the LQG one. We determine for both cases the optimal controllers, and illustrate the theoretical results using simulation.

1 Introduction

Congestion phenomena in high speed networks are quite frequent, both on the Internet as well as in ATM (Asynchronous Transfer Mode) both of which are widely used in today's high speed telecommunications networks. Congestion results in the degradation of many important performance measures; it causes losses of packets, it increases delays and decreases the transmission throughput. In order to protect the users from congestion, on one hand, and to allow for the efficient use of the network, on the other, flow control mechanisms have been introduced and adopted both in the Internet and in the ATM.

Several approaches to flow control in high speed networks, based on some feedback on the state of the network, have been proposed: window flow control (e.g. [13]), credit-based control (see [15, 16] and references therein) and rate-based flow control (see [9, 16] and references therein). The last one has been selected by the ATM forum in 1994 as the main approach for flow control in the Available Bit Rate (ABR) type of service, see [1]. This class of service is designed to support applications that can adjust their input rate to the congestion state of the network (e.g. file transfers), as opposed to some real-time applications (such as high quality interactive video or audio applications) that typically use other higher priority services and to which a fixed

*INRIA B.P. 93, 06902 Sophia Antipolis Cedex, France. Tel: (33) 4 92 38 77 86; Fax: (33) 4 92 38 77 65; Email: altman@sophia.inria.fr

†Coordinated Science Laboratory, University of Illinois, 1308 West Main Street, Urbana, IL 61801, USA. Tel: (1) 217 333-3607; Fax: (1) 217 244-1653; Email: tbasar@decision.csl.uiuc.edu. Research supported by Grants NSF ECS 92-20632 and NSF ECS 93-12807.

‡Institute of Mechanics, Armenian Academy of Sciences, Marshal Baghramian Street 24, Yerevan 375019, Armenia. Tel: 003-742-524890; E-mail: ercim@arminco.com

amount of bandwidth is allocated. In networks that integrate several types of traffic, the flow control is used to adjust the rate of transmission of the first type of traffic to the time-varying bandwidth left over by the higher priority real time applications. The ATM forum has not standardized the flow control mechanisms and does not wish to do so; its design and implementation are left to the switch designers.

We present in this paper two types of linear quadratic (LQ) models for designing optimal rate based control, so as to achieve optimal performance measures: LQG (Linear-Quadratic-Gaussian) and H^∞. In both cases we allow for noisy delayed information. The objectives in the design of the controllers are to (i) minimize the variation of the bottleneck queue size around some desired level; this objective measure is directly related to the minimization of losses and maximization of throughput; (ii) obtain good tracking between input and output rates; this objective is related to the efficient use of the network; and (iii) minimize the jitter. The relative importance (weight) of each one of these objectives is given as data in the design of the controller. The optimal controllers that are obtained are easy to implement: they turn to be linear in the (estimated) state.

We consider a discrete time, linearized model for the dynamics, as introduced in [2] (which extends the models in [8, 14]). A time unit corresponds to the time interval between two consecutive acquisitions of new information. According to [1], such information is available through RM (Resource Management) packets that travel from the source to the destination (forward packets) and then back from the destination to the source (backward packets). Information on the congestion (queue-length) and available bandwidth may be stamped onto these packets in the intermediate switches. We make the assumption that the RM packets are sent periodically. We model the interfering traffic, which may stem from higher priority real time applications, by a general ARMA process. This is a generalization of the AR model for interfering traffic that was introduced in [3, 4], which thus allows us to have a much larger class of models for the bandwidth available for the ABR traffic.

We describe briefly in the next section the linearized model for the state dynamics; detailed derivations of the non-linear model as well as its linear approximation can be found in [2]. Simulations further justify this linearization; see [2].

We assume that the performance measures (such as throughputs, delays and loss probabilities) are determined essentially by a bottleneck node. This assumption has both theoretical and experimental [7, 10] justification, and is often made in the literature in order to analyze or design controllers [2, 8, 12].

The control problem is solved in Sections 3 and 4 for the case when there is no delay in the measurements, and in Section 5 when the imperfect measurements are acquired with some delay, or alternatively when the control is updated more than once during a round trip. Simulation results are presented in Section 6, and some discussion on their interpretation is included in Section 7. The paper ends with the concluding remarks of Section 8, that summarize and highlight the advantages of the methodology and schemes introduced here.

2 The model

We begin by providing the discrete time model considered in this paper, where a time unit corresponds to a round trip delay. The more general model where a time unit corresponds to the period between consecutive instants of information acquisition, which may be smaller than a round trip delay, is formulated and studied later in Section 5. Let q_n denote the queue length at a bottleneck link, and μ_n denote the effective service rate available for traffic of the given source at that link at the beginning of the nth time slot. Let x_n denote the source rate during the nth time slot. We thus consider a rate-based flow control, where based on the current and previous noisy information on μ_n, q_n, and x_n (to be made precise below), the controller updates the transmission rate. Hence, the queue length evolves according to

$$q_{n+1} = q_n + x_n - \mu_n. \tag{1}$$

Since several sources with varying transmission rates may share the same bottleneck link, the service rate μ_n available to the controlled source may change over time in an unpredictable way. The other sources may be represented as some interference, which we model as a stable ARMA process:

$$\mu_n = \mu + \xi_n, \qquad \xi_n = \sum_{i=1}^{d} \ell_i \xi_{n-i} + \sum_{i=1}^{d} k_i \phi_{n-i}. \tag{2}$$

We assume that ℓ_i's, which will typically be obtained by some estimation/identification procedure (using, for example, the results and framework of [11]), are such that (2) describes a stable system. The variable ϕ_n in (2) stands for disturbance, the nature of which will be described shortly.

At the end of step n, the source obtains a noisy estimate of the value of μ_n, and a noisy estimate of the value of q_n. We use $\hat{\mu}_n$ to denote a measurement for the service rate at the bottleneck node, and $\hat{q}_n$ to denote a measurement for the bottleneck queue size. For our analysis, we assume that these measurements can be written as

$$\hat{\mu}_n = \mu_n + c v_n, \qquad \hat{q}_n = q_n + a w_n, \tag{3}$$

where v_n and w_n are disturbances whose nature will be described shortly, and a and c are some constants.

We introduce shifted versions of the variables $x_n, q_n, \hat{\mu}_n, \hat{q}_n$:

$$\tilde{x}_n := x_n - \mu, \qquad \tilde{q}_n := q_n - Q, \qquad \tilde{\hat{\mu}}_n := \hat{\mu}_n - \mu, \qquad \tilde{\hat{q}}_n := \hat{q}_n - Q. \tag{4}$$

We further define the (scaled) variation in the input rate by bu_n, where u_n will be called the control, and b is some constant parameter. In terms of the new variables, we now have the following dynamics:

$$\tilde{q}_{n+1} = \tilde{q}_n + \tilde{x}_n - \xi_n \tag{5}$$

$$\xi_n = \sum_{i=1}^{d} \ell_i \xi_{n-i} + \sum_{i=1}^{d} k_i \phi_{n-i} \tag{6}$$

$$\tilde{x}_{n+1} = \tilde{x}_n + b u_n \tag{7}$$

and the observations:

$$\tilde{\hat{\mu}}_n = \xi_n + c v_n, \qquad \tilde{\hat{q}}_n = \tilde{q}_n + a w_n. \tag{8}$$

For each pair of integers (k, K), with $K > k$, we introduce the following quadratic cost:

$$L_k^K = g\|\tilde{x} - \xi\|^2_{(k,K)} + \|\tilde{q}\|^2_{(k,K)} + \|u\|^2_{(k,K)} \qquad \text{where } g > 0, \quad \|u\|^2_{(k,K)} := \sum_{n=k}^{K} |u_n|^2, \ \text{ etc...} \tag{9}$$

This definition of the cost, over a period of length $K+1-k$ units, which is allowed to be infinite, allows us to quantify optimization criteria that have been used in previous control models [2]:

- The first term represents the quality with which the input rate tracks the available service rate, where g is a positive weighting term. A high rate of tracking is known to be desirable; see, e.g., [8].
- The second term is a penalty for deviating from a desirable queue length. Setting some desired level of queue reflects the fact that we do not wish to have a large queue, so as to avoid losses, and on the other hand, we do not wish to have a low level of queue, since if the queue is empty and the input rate is lower than the service rate then there is a waste of potential throughput. We have not included any additional weighting on this term, as any such positive weight can be absorbed into the other variables.
- The last term stands for a penalty for high jitter, i.e. for high variability of (input) transmission rate, which is known to be undesirable; see [8].[1] In many proposed telecommunication networks, the source will pay more for higher burstiness (variability) of its input rate, since highly variable input rate has typically a bad influence on other traffic. Again, as in the case of the second term, the weighting here has been normalized to 1, but any other weighting can be absorbed in the parameter b.

We initially allow the controller to be a function of all the past and present measurements and past actions:

$$u_n = f_n(\tilde{x}_m, \hat{\tilde{\mu}}_m, \hat{\tilde{q}}_m, u_{m-1}, m = k, \ldots, n)\,,\ n \geq k. \tag{10}$$

Later, in Section 5, we will consider the case where the dependence on $\hat{\tilde{\mu}}$ and $\hat{\tilde{q}}$ is with a delay of θ time units.

To obtain steady-state solutions (controllers), we take the number of periods (time slots) in the problem to be sufficiently large; in particular, we let $k=-\infty$ and $K=\infty$. Over this 'doubly-infinite' control horizon, we consider two types of models for the disturbances:

- **M1:** The Gaussian model Here v_n, w_n, ϕ_n are taken as independent identically distributed Gaussian random variables with zero mean and unit variance. It then follows that the measurements $\hat{\mu}_n$ and $\hat{q}_n$ are unbiased. The cost to be minimized in this case is the 'per time slot' expected value of the cost, written in precise mathematical terms as

$$J_G(f) = \limsup_{K\to\infty, k\to-\infty} \frac{1}{K+1-k} E\big[L_k^K(f;\phi,v,w)\big] \tag{11}$$

- **M2:** The H$^\infty$ approach In this case v_n, w_n and ϕ_n are taken as unknown (deterministic) disturbances. The cost to be minimized is the ℓ_2 gain from the disturbances to the cumulative cost $L_{-\infty}^{\infty}$, i.e.,

$$J_H(f) = \sup_{\{\phi_n, v_n, w_n\}_{n=-\infty}^{\infty}} \frac{L_{-\infty}^{\infty}(f;\phi,v,w)}{\|\phi,v,w\|^2} \tag{12}$$

where

$$\|\phi,v,w\|^2 := \|\phi\|^2 + \|v\|^2 + \|w\|^2 = \sum_{n=-\infty}^{\infty} \big(|\phi_n|^2 + |v_n|^2 + |w_n|^2\big)\,. \tag{13}$$

[1] Note that if flows from several sources are multiplexed in a FIFO queue, and if the input rate of one source is highly variable, then also the output rate for that source tends to be highly variable, since packets from the other sources come in between the packets of that source.

Here we also take the initial conditions for the state equations (5)-(7) in the infinite past to be *zero*, i.e.,

$$\lim_{n\to-\infty} \tilde{x}_n = 0\,, \qquad \lim_{n\to-\infty} \tilde{q}_n = 0\,, \qquad \lim_{n\to-\infty} \xi_n = 0\,. \tag{14}$$

Let us denote the square-root of the infimum of this cost function over all admissible controllers $\{f_n\}$ by γ^*:

$$(\gamma^*)^2 = \inf_{\{f_n\}_{n=-\infty}^{\infty}} J_H(f). \tag{15}$$

If there exists an actual minimizing control, say f^*, to the optimization problem above, then it can be shown [6] that, defining a soft-constrained cost function

$$L_\gamma(u;\phi,v,w) := L_{-\infty}^{\infty}(u;\phi,v,w) - \gamma^2 \left\|\phi,v,w\right\|^2,$$

to be viewed as the kernel of a two player game, to be minimized by Player 1 (controlling u) and maximized by Player 2 (controlling (ϕ,v,w)), f^* has the property: $\sup_{\phi,v,w} L_{\gamma^*}(f^*;\phi,v,w) = \inf_f \sup_{\phi,v,w} L_{\gamma^*}(f;\phi,v,w)$. The quantity above is the upper value of the zero-sum dynamic game with kernel L_{γ^*}, which is in fact equal to zero. It can actually be shown that for any $\gamma \geq \gamma^*$, the upper value of the game with parameterized kernel L_γ is zero, and for $\gamma < \gamma^*$, its upper value is infinite. Hence, γ^* is the smallest positive scalar γ for which the zero-sum game with kernel L_γ has a finite upper value.

Instead of obtaining f^* defined above, we will in fact solve a parameterized class of controllers, $\{f^\gamma, \gamma > \gamma^*\}$, where f^γ is obtained from $\sup_{\phi,v,w} L_\gamma(f^\gamma;\phi,v,w) = \inf_f \sup_{\phi,v,w} L_\gamma(f;\phi,v,w)$. The controller f^γ will clearly have the property that it ensures a performance level γ^2 for the index adopted for M2, i.e. the attenuation is bounded by

$$\mathcal{A}(f^\gamma;\phi,v,w) := \frac{\{L_{-\infty}^{\infty}(f^\gamma;\phi,v,w)\}^{1/2}}{\|\phi,v,w\|} \leq \gamma \quad \text{for all } \phi,v,w. \tag{16}$$

It will turn out that the limit $\lim_{\gamma\to\infty} f^\gamma =: f^\infty$ is a well-defined controller, and solves uniquely the control problem with the Gaussian model, M1.

3 Complete solution to the problem with Model M2 and with an additional noise perturbation in the measurements

The optimal control problem formulated above can be solved by a suitable modification of the theory of discrete-time H$^\infty$-control developed in [6]. Toward this end, we first write (5)-(7) in standard state variable form. Introduce the d-dimensional vector $\eta_n := (\xi_{n-d+1},\ldots,\xi_n)'$, that would be obtained from (6) by taking $k_1 = 1$, $k_i = 0$, $i \neq 1$, in terms of which (6) can be re-written as a first-order difference equation:

$$\eta_{n+1} = C\eta_n + \begin{pmatrix} 0 \\ . \\ . \\ . \\ 1 \end{pmatrix} \phi_n \qquad \text{where} \qquad C := \begin{pmatrix} 0 & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 0 & \ldots & 0 \\ . & . & . & . & \ldots & \\ \ell_d & \ell_{d-1} & \ell_{d-2} & & \ldots & \ell_1 \end{pmatrix} \tag{17}$$

Then the actual ξ_n generated by (6) is given by

$$\xi_n = (k_d, \cdots, k_1)\eta_n \tag{18}$$

Further introducing the $(d+2)$-dimensional vector $z_n := (\tilde{x}_n, \tilde{q}_n, \eta_n')'$, we can write (5)-(7) in the form:

$$z_{n+1} = Az_n + D\phi_{n+1} + Bu_n \tag{19}$$

where

$$A := \left(\begin{array}{cc|cccc} 1 & 0 & 0 & \dots & 0 & 0 \\ 1 & 1 & -k_d & \dots & -k_2 & -k_1 \\ \hline & & & & & \\ \multicolumn{2}{c|}{0} & \multicolumn{4}{c}{\mathrm{C}} \\ & & & & & \end{array} \right) \qquad D := \begin{pmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 1 \end{pmatrix} \qquad B := \begin{pmatrix} b \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{pmatrix} \tag{20}$$

Now, the stability of the ARMA process ξ_n generated by (2) is equivalent to the stability of the matrix C in (17), which is captured in the following assumption required to hold throughout the remaining analysis:

- **A1.** All eigenvalues of C are in the unit circle in the complex plane.

Under A1, we have the following fact[2], which will be useful in the statement of our main result.

- **F1.** Under A1, the pair (A,B) is stabilizable.

As far as the measurements available to the controller go, we have two disturbance corrupted state measurements, given by (8), and one perfect measurement of the state, $\tilde{x}_n$. This is a "mixed" structure, that leads to a *singular* measurement equation, which is not allowed in standard theory. To circumvent this difficulty, we will first perturb $\tilde{x}_n$ with a small independent disturbance, to arrive at the measurement equation

$$\hat{\tilde{x}}_n = \tilde{x}_n + \epsilon\tilde{\tilde{v}}_n, \qquad 0 < \epsilon << 1 \tag{21}$$

and assume that now $\hat{\tilde{x}}_n$ is available for control purposes, instead of $\tilde{x}_n$. In the above, $\tilde{\tilde{v}}_n$ is the auxiliary disturbance introduced, and ϵ is a small positive quantity. This modification is only for mathematical convenience, which will in the end be made to vanish asymptotically by letting $\epsilon \to 0$, so as to recover the original information structure.

To introduce the modified measurement equation, let $y_n := (\hat{\tilde{x}}_n, \hat{\tilde{q}}_n, \hat{\tilde{\mu}}_n)'$. Then $y_n = Hz_n + E_\epsilon\tilde{v}_n$, where

$$H := \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix} \qquad E_\epsilon := \begin{pmatrix} \epsilon & 0 & 0 \\ 0 & a & 0 \\ 0 & 0 & c \end{pmatrix} \qquad \tilde{v}_n := \begin{pmatrix} \tilde{\tilde{v}}_n \\ w_n \\ v_n \end{pmatrix} \tag{22}$$

Since $E_\epsilon E_\epsilon' > 0$ for all $\epsilon > 0$, this is a nonsingular measurement equation. Compatible with this modification, we also modify the state equation (19), to write it as $z_{n+1} = Az_n + Bu_n + D_\epsilon\zeta_n$, where

$$D_\epsilon := \begin{pmatrix} 0 & 0 & \dots & 0 & 1 \\ \epsilon & 0 & \dots & 0 & 0 \end{pmatrix}' \qquad \zeta_n := (\phi_{n+1}, \tilde{\phi}_{n+1})' \tag{23}$$

[2]This follows from the Hautus test by observing that the unstable eigenvalue 1 (of multiplicity 2) of the matrix A is controllable, i.e. rank $\left(A - I \middle| B\right) = d+2$

with $\tilde{\phi}_{n+1}$ being the new disturbance introduced.

To complete the description, we have to write the cost function (9) in terms of these new variables

$$L = \sum_{n=-\infty}^{\infty} \left\{ |z_n|_N^2 + |u_n|^2 \right\} \tag{24}$$

where

$$N := \left(\begin{array}{cc|cccc} g & 0 & -gk_d & -gk_{d-1} & \dots & -gk_1 \\ 0 & 1 & 0 & 0 & \dots & 0 \\ \hline -gk_d & 0 & gk_d^2 & gk_dk_{d-1} & \dots & gk_dk_1 \\ -gk_{d-1} & 0 & gk_dk_{d-1} & gk_{d-1}^2 & \dots & gk_{d-1}k_1 \\ . & . & . & . & \dots & . \\ . & . & . & . & \dots & . \\ . & . & . & . & \dots & . \\ -gk_1 & 0 & gk_dk_1 & gk_{d-1}k_1 & \dots & gk_1^2 \end{array} \right) \tag{25}$$

and the H$^\infty$ criterion (to be minimized) now becomes

$$\sup_{\tilde{v},\zeta} \frac{L}{\|\tilde{v},\zeta\|^2}, \tag{26}$$

the square-root of whose infimum over all admissible controllers we now denote by $\gamma^*(\epsilon)$. It is continuous at $\epsilon = 0$, and the quantity $\gamma^*(0)$ is equal to the γ^* defined earlier (for the unmodified problem).

Three additional properties, related to the system matrices, are now worth noting, which we state below as facts:

- **F2.** The pair (A, H) is detectable.
- **F3.** The pair (A, N) is detectable.
- **F4.** For each $\epsilon > 0$, the pair (A, D_ϵ) is controllable.

We are now in a position to state the solution to the modified H$^\infty$ control problem formulated above, directly from Başar and Bernhard [6], Chap. 6.

Proposition 1 *Let $\epsilon > 0$ be fixed, and A1 hold. Then, $\gamma^*(\epsilon)$ defined above is finite, and for all $\gamma > \gamma^*(\epsilon)$:*
1) There exists a minimal nonnegative-definite solution, M_ϵ, to

$$M = N + A'M\Lambda_\epsilon^{-1}A \tag{27}$$

such that

$$\gamma^2 I - D_\epsilon' M_\epsilon D_\epsilon > 0, \tag{28}$$

where

$$\Lambda_\epsilon := I + (BB' - \gamma^{-2}D_\epsilon D_\epsilon')M. \tag{29}$$

2) There exists a minimal nonnegative-definite solution, Σ_ϵ, to

$$\Sigma = D_\epsilon D_\epsilon' + A\Sigma R_\epsilon^{-1}A' \tag{30}$$

such that

$$\gamma^2 I - N^{1/2}\Sigma_\epsilon N^{1/2} > 0, \tag{31}$$

where

$$R_\epsilon := I + \left(H'(E_\epsilon E_\epsilon')^{-1}H - \gamma^{-2}N\right)\Sigma. \tag{32}$$

3)

$$\Sigma_\epsilon^{1/2} M_\epsilon \Sigma_\epsilon^{1/2} < \gamma^2 I. \tag{33}$$

4) An H^∞-controller that ensures the bound γ^2 in the index (26) is

$$u_{n,\epsilon}^{\gamma} = f_\epsilon^{\gamma}(\hat{z}_{n|n}) = -B'M_\epsilon\Lambda_\epsilon^{-1}A\hat{z}_{n|n}, \tag{34}$$

where $\hat{z}_{n|n}$ is generated by

$$\hat{z}_{n|n} = \left(I + \Sigma_\epsilon H'(E_\epsilon E_\epsilon')^{-1}H - \gamma^{-2}\Sigma_\epsilon M_\epsilon\right)^{-1}\left(\breve{z}_n + \Sigma_\epsilon H'(E_\epsilon E_\epsilon')^{-1}y_n\right) \tag{35}$$

$$\breve{z}_{n+1} = A\breve{z}_n + Bu_n + A\Sigma_\epsilon R_\epsilon^{-1}\left[\gamma^{-2}N\breve{z}_n + H'(E_\epsilon E_\epsilon')^{-1}(y_n - H\breve{z}_n)\right] \tag{36}$$

5) The controller (34) leads to a bounded input-bounded state stable system, i.e. for all bounded disturbances $\tilde{v}, \zeta$, the system state z_n and the filter state $\hat{z}_{n|n}$ remain bounded. Equivalently, the two matrices

$$(I - BB'M_\epsilon\Lambda_\epsilon^{-1})A \quad \text{and} \quad \left(I - H'(E_\epsilon E_\epsilon')^{-1}H\Sigma_\epsilon R_\epsilon^{-1}\right)A'$$

have all their eigenvalues in the unit circle.

If any one of the three properties 1)-3) above does not hold, then $\gamma \leq \gamma^*(\epsilon)$. ◇

Remark 1 The results of Proposition 1 are valid not only for all finite γ's larger than $\gamma^*(\epsilon)$, but also as $\gamma \to \infty$. In this limiting case, which captures the solution to ϵ-perturbed problem under the Gaussian model (M1) of the disturbance, the constraints (28), (31) and (32) become irrelevant, and all other expressions admit "well-defined" forms, obtained by simply setting $\gamma^{-2} = 0$. ◇

4 The limiting solution as $\epsilon \downarrow 0$

We next study the limit of the solution given in Proposition 1 as $\epsilon \downarrow 0$. For the limit to be well defined, it will be sufficient to show that equations and relationships (27)-(33) are well-defined in the limit as $\epsilon \downarrow 0$. The analysis here is similar to that carried out in [4] for the case where interfering traffic was modeled by an AR process, with just A and N having different structures; it is reproduced here in full for the sake of completeness.

The first three of (27)-(33) depend continuously on $\epsilon \geq 0$, and hence in the limit (27) will be replaced by

$$M = N + A'M\Lambda_0^{-1}A \tag{37}$$

where

$$\Lambda_0 := I + \mathrm{diag}(b^2, 0, \ldots, 0, -\gamma^{-2})M \tag{38}$$

and (37) will have to be solved under the scalar condition (as counterpart of (28)):

$$(0,\ldots,0,1)M(0,\ldots,0,1)' < \gamma^2 \tag{39}$$

which guarantees invertibility of (38). Again, for each $\gamma > \gamma^*(0)$, (37) admits a unique minimal nonnegative-definite solution, which also satisfies (39).

The limiting analysis of (30)-(32) is somewhat more complicated, since it involves the inverse of a matrix that becomes singular at $\epsilon = 0$ (which is $E_\epsilon E_\epsilon'$). However, this isolated singularity does not lead to any singularity for the solution of (30) as to be shown below. We will in fact show that in the limit as $\epsilon \downarrow 0$, the solution of (30) will be in the form

$$\Sigma_0 = \begin{pmatrix} 0 & . & . & . & 0 \\ . & & & & \\ . & & P & & \\ . & & & & \\ 0 & & & & \end{pmatrix} \tag{40}$$

where $P \geq 0$ is of dimensions $(d+1)\times(d+1)$. To see this, let us first rewrite ΣR_ϵ^{-1} as

$$\Sigma R_\epsilon^{-1} = \Sigma^{1/2}\left[I + \Sigma^{1/2}H'(E_\epsilon E_\epsilon')^{-1}H\Sigma^{1/2} - \gamma^{-2}\Sigma^{1/2}N\Sigma^{1/2}\right]^{-1}\Sigma^{1/2}$$

where $\Sigma^{1/2}$ denotes the unique nonnegative definite square root of Σ. Let the $(d+1)\times(d+1)$-dimensional lower block of N be denoted by $\tilde{N}$, which is

$$\tilde{N} := \left(\begin{array}{c|cccc} 1 & 0 & 0 & \ldots & 0 \\ \hline 0 & gk_d^2 & gk_dk_{d-1} & \ldots & gk_dk_1 \\ 0 & gk_dk_{d-1} & gk_{d-1}^2 & \ldots & gk_{d-1}k_1 \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ 0 & gk_dk_1 & gk_{d-1}k_1 & \ldots & gk_1^2 \end{array}\right) \tag{41}$$

and suppose that Σ is given by (40), where P is arbitrary. Then, straightforward manipulations lead to

$$\Sigma^{1/2}N\Sigma^{1/2} = \text{block-diag}\left(0,\ P^{1/2}\tilde{N}P^{1/2}\right)$$

$$\Sigma^{1/2}H'(E_\epsilon E_\epsilon')^{-1}H\Sigma^{1/2} = \text{block-diag}\left(0,\ P^{1/2}\tilde{H}'(\tilde{E}\tilde{E}')^{-1}\tilde{H}P^{1/2}\right)$$

where

$$\tilde{H} := \underbrace{\begin{pmatrix} 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & \ldots & 0 & 1 \end{pmatrix}}_{2\times(d+1)} \qquad \tilde{E} := \underbrace{\begin{pmatrix} a & 0 \\ 0 & c \end{pmatrix}}_{2\times 2} \tag{42}$$

Hence,

$$\Sigma R_\epsilon^{-1} \quad = \quad \Sigma^{1/2}\left[\text{block-diag}\left(1,\ I + P^{1/2}\left(\tilde{H}'(\tilde{E}\tilde{E}')^{-1}\tilde{H} - \gamma^{-2}\tilde{N}\right)P^{1/2}\right)\right]^{-1}\Sigma^{1/2}$$

$$
\begin{aligned}
&= \text{block-diag}\left(0,\ P^{1/2}\left[I + P^{1/2}\left(\tilde{H}'(\tilde{E}\tilde{E}')^{-1}\tilde{H} - \gamma^{-2}\tilde{N}\right)P^{1/2}\right]^{-1}P^{1/2}\right) \\
&= \text{block-diag}(0,\ \tilde{S}) \qquad (43)
\end{aligned}
$$

which is independent of ϵ. Pre- and post-multiplying this expression by A and A', we obtain

$$
A\Sigma R_\epsilon^{-1}A' = \text{block-diag}\left(0,\ \tilde{A}P^{1/2}\left[I + P^{1/2}\left(\tilde{H}'(\tilde{E}\tilde{E}')^{-1}\tilde{H} - \gamma^{-2}\tilde{N}\right)P^{1/2}\right]^{-1}P^{1/2}\tilde{A}'\right)
$$

where $\tilde{A}$ is the $(d+1)\times(d+1)$ dimensional lower block of A, i.e.

$$
\tilde{A} := \left(\begin{array}{c|c}
1 & -k_d \ \ldots - k_2 \ \ -k_1 \\
\hline
0 & \\
\cdot & \\
\cdot & \text{C} \\
\cdot & \\
0 &
\end{array}\right) \qquad (44)
$$

Furthermore, since $\lim_{\epsilon\to 0} D_\epsilon D_\epsilon' = \text{diag}(0,\ldots,0,1)$, it readily follows that the structure (40) is consistent with (30). Introducing a $(d+1)$-dimensional vector $\tilde{D}$,

$$
\tilde{D} := (0,\ldots,0,1)', \qquad (45)
$$

the equation to be satisfied by P can now be written as

$$
P = \tilde{D}\tilde{D}' + \tilde{A}P\left[I + \left(\tilde{H}'(\tilde{E}\tilde{E}')^{-1}\tilde{H} - \gamma^{-2}\tilde{N}\right)P\right]^{-1}\tilde{A}', \qquad (46)
$$

and (31) becomes equivalent to

$$
\gamma^2 I - \tilde{N}^{1/2}P\tilde{N}^{1/2} > 0. \qquad (47)
$$

Finally, (33) simplifies to

$$
P^{1/2}\tilde{M}P^{1/2} < \gamma^2 I\,, \qquad (48)
$$

where $\tilde{M}$ is the $(d+1)\times(d+1)$-dimensional lower block of M.

Hence, (27)-(29) are well-defined at $\epsilon = 0$, and (30)-(33) admit well-defined limits in the space of dimension $(d+1)$ instead of $(d+2)$.

To complete the limiting analysis, we still have to find the limiting form of the controller (34)-(36). Toward this end, write z, as $\epsilon \downarrow 0$, as

$$
z = \left(\alpha,\ \beta'\right)' \qquad (49)
$$

where α is of dimension 1, and β of dimension $d+1$. Compatibly, write (again as $\epsilon \downarrow 0$):

$$
\hat{z}_{n|n} = (\hat{\alpha}_{n|n},\ \hat{\beta}'_{n|n})'; \qquad \breve{z}_n = (\breve{\alpha}_n,\ \breve{\beta}'_n)'. \qquad (50)
$$

Then, using the earlier manipulations, and some straightforward extensions, it can be shown that $\hat{\alpha}_{n|n}$, $\hat{\beta}_{n|n}$, $\breve{\alpha}_n$, $\breve{\beta}_n$ are generated by (from (35)-(36), and using the notation introduced by (41)-(45)):

$$\begin{aligned}
\hat{\alpha}_{n|n} &= \breve{\alpha}_n && (51)\\
\breve{\alpha}_{n+1} &= \breve{\alpha}_n + bu_n && (52)\\
\hat{\beta}_{n|n} &= \left(I + P\tilde{H}'(\tilde{E}\tilde{E}')^{-1}\tilde{H} - \gamma^{-2}P\tilde{M}\right)^{-1}\left(\breve{\beta}_n + P\tilde{H}'(\tilde{E}\tilde{E}')^{-1}\tilde{y}_n\right) && (53)\\
\breve{\beta}_{n+1} &= \tilde{A}\breve{\beta}_n + \tilde{A}\tilde{S}\left[\gamma^{-2}\tilde{N}\breve{\beta}_n + \tilde{H}'(\tilde{E}\tilde{E}')^{-1}(\tilde{y}_n - \tilde{H}\breve{\beta}_n)\right] && (54)\\
\tilde{y}_n &:= (\hat{\tilde{q}}_n, \hat{\tilde{\mu}}_n)' && (55)
\end{aligned}$$

Clearly, (52) generates $\tilde{x}_n$ whose value is already available to the controller. Hence, the limiting controller is of dimension $d+1$, with the worst-case observer given by (53)-(54). The counterpart of (34) is then

$$u_n^{\gamma} = f^{\gamma}(x_n, \hat{\beta}_{n|n}) = -B'M_0\Lambda_0^{-1}A\begin{pmatrix} \tilde{x}_n \\ \hat{\beta}_{n|n} \end{pmatrix}, \tag{56}$$

and this achieves a level of disturbance attenuation no worse than γ, provided that $\gamma > \gamma^*$, where γ^* is the optimum level of disturbance attenuation for the unperturbed problem, which can be computed as $\gamma^* = \lim_{\epsilon\downarrow 0}\gamma^*(\epsilon)$.

To see this, first note that for each fixed controller, the attenuation is monotonically nonincreasing with decreasing $\epsilon > 0$, and hence its infsup value, $\gamma^*(\epsilon)$ also has the same property. Furthermore, since $\gamma^*(\epsilon) \geq 0$, the set $\{\gamma^*(\epsilon)\}_{\epsilon>0}$ has a limit as $\epsilon \downarrow 0$. It follows from the algebraic characterization of $\gamma^*(\epsilon)$, which was in terms of two Riccati equations and a spectral radius condition which are well defined at $\epsilon = 0$ (as discussed earlier), that its limit as $\epsilon \downarrow 0$ is indeed $\gamma^*(0)$. Furthermore, $\gamma^*(0) = \gamma^*$, because clearly $\gamma^* > \gamma^*(0)$ is not possible as this would imply $\gamma^* > \gamma^*(\epsilon)$ for some sufficiently small $\epsilon > 0$, contradicting the initial hypothesis that γ^* corresponds to a more refined information for the controller than $\gamma^*(\epsilon)$; $\gamma^* < \gamma^*(0)$ is not possible either, because if f^{γ} is a controller for the original problem, achieving a level of disturbance attenuation no worse than γ, with $\gamma^* < \gamma < \gamma^*(0)$, for the noise-perturbed problem with sufficiently small $\epsilon > 0$, still a performance level better than $\gamma^*(0)$ would be obtained using the same controller (not with noise perturbed measurement) by continuity of the controller with respect to the measurement, which would then violate the inequality $\gamma^*(0) \leq \gamma^*(\epsilon)$ for all $\epsilon > 0$. Now, finally under the controller (56), a level of disturbance attenuation no worse than $\gamma > \gamma^*$ would be achieved because the controller f_{ϵ}^{γ}, given by (34), is a continuous function of ϵ, which is also well defined at $\epsilon = 0$, and given any $\hat{\gamma} > \gamma^*$, there exists an $\epsilon_0 > 0$ such that for all $0 < \epsilon < \epsilon_0$, $\gamma^*(\epsilon) < \hat{\gamma}$, and

$$\sup_{\phi,v,w} L_{\hat{\gamma}}(f_{\epsilon}^{\hat{\gamma}}, \phi, v, w) = 0, \tag{57}$$

and by continuity at $\epsilon = 0$ (of also the concavity condition that assures finiteness of the supremum in (57)),

$$\sup_{\phi,v,w} L_{\hat{\gamma}}(f^{\hat{\gamma}}, \phi, v, w) = 0,$$

where $f^{\hat{\gamma}}$ given by (56) is also $f_{\epsilon}^{\hat{\gamma}}\big|_{\epsilon=0}$.

This result is now summarized in the following Theorem:

Theorem 1 *Consider the original control problem formulated in Section 2, under the H^∞ model* (M2) *of the disturbance. Let Assumption A1 hold. Then γ^* defined by (15) is finite, and for all $\gamma > \gamma^*$:*
1) There exists a minimal nonnegative-definite solution, M_0, to (37) such that (39) holds.
2) There exists a minimal nonnegative-definite solution, P, to (46), such that (47) holds.
3) Condition (48) holds.
4) An H^∞-controller that ensures the bound γ^2 in the index (12) is given by (56).
5) The controller (56) leads to a bounded input-bounded state stable system (see Proposition 1 part 5)).
For a given $\gamma > 0$, if any one of the properties 1)-3) above does not hold, then $\gamma \leq \gamma^*$. $\diamond$

We immediately have the following corollary to Theorem 1, obtained by letting $\gamma \to \infty$, which solves the original problem under the Gaussian interpretation (M1) of the disturbances (see Remark 1).

Corollary 1 *Under A1, the optimal control problem with the disturbance model M1 admits the solution:*

$$u_n^\infty = f^\infty(x_n, \hat{\beta}_{n|n}) = -B'\overline{M}(I + BB'\overline{M})^{-1} A \begin{pmatrix} \tilde{x}_n \\ \hat{\beta}_{n|n} \end{pmatrix} =: \mathcal{K} \begin{pmatrix} \tilde{x}_n \\ \hat{\beta}_{n|n} \end{pmatrix}, \tag{58}$$

where $\overline{M}$ is the unique nonnegative-definite solution of

$$\overline{M} = N + A'\overline{M}(I + BB'\overline{M})^{-1} A \tag{59}$$

in the class of nonnegative-definite matrices, and

$$\hat{\beta}_{n|n} = \left(I + \overline{P}\tilde{H}'(\tilde{E}\tilde{E}')^{-1}\tilde{H}\right)^{-1} \left(\breve{\beta}_n + \overline{P}\tilde{H}'(\tilde{E}\tilde{E}')^{-1}\tilde{y}_n\right) \tag{60}$$

$$\breve{\beta}_{n+1} = \tilde{A}\breve{\beta}_n + \tilde{A}\overline{S}\tilde{H}'(\tilde{E}\tilde{E}')^{-1}(\tilde{y}_n - \tilde{H}\breve{\beta}_n) \tag{61}$$

where $\overline{S} := \overline{P}\left[I + \tilde{H}'(\tilde{E}\tilde{E}')^{-1}\tilde{H}\overline{P}\right]^{-1}$, and $\overline{P}$ is the unique nonnegative-definite solution of $\overline{P} = \tilde{D}\tilde{D}' + \tilde{A}\overline{S}\tilde{A}'$ in the class of nonnegative definite matrices.
Furthermore, the controller (58) leads to a stable system, and the optimal average (incremental) values of the three different additive terms in the cost (11) are given by (for $-\infty \leq n \leq \infty$)

$$E\big[|\tilde{x}_n - \xi_n|^2\big] = \zeta_1'\Xi\zeta_1\,, \qquad E\big[|\tilde{q}_n|^2\big] = \zeta_2'\Xi\zeta_2\,, \qquad E\big[|u_n|^2\big] = \mathcal{K}\big(\Xi - \text{ blockdiag } (0\,,\overline{S})\big)\mathcal{K}'\,, \tag{62}$$

where Ξ is the unique nonnegative definite solution of the linear matrix equation:

$$\Xi = (A - B\mathcal{K})\Xi(A - B\mathcal{K})' + B\mathcal{K}\big[\text{ blockdiag } (0\,,\overline{S})\big]\mathcal{K}'B' + DD'\,, \tag{63}$$

and ζ_1 and ζ_2 are $d+2$-dimensional vectors, defined by: $\zeta_1 := \big(1, 0, \ldots, 0, -1\big)'$ and $\zeta_2 := \big(0, 1, 0, \ldots, 0\big)'$. $\diamond$

Remark 2 The expressions (62) can be related to other performance measures of interest, and in particular, to blocking probabilities. Let Φ denote the probability distribution of a standard Gaussian random variable.

Then (62) enables us to use the following standard heuristic in order to approximate the blocking probabilities in steady state for a finite buffer of size $K >> Q$:

$$P(\text{blocking}) \sim P(q_n > K) = P(\tilde{q}_n > K - Q). \tag{64}$$

Since $\tilde{q}_n$ is Gaussian with zero mean and second moment given by $\zeta_2' \Xi \zeta_2$ (see (62)), this leads to

$$P(\text{blocking}) \sim 1 - \Phi\left(\frac{K-Q}{\sqrt{\zeta_2' \Xi \zeta_2}}\right). \tag{65}$$

5 Delayed measurements

In our basic model presented in Section 2, a time unit corresponded to a round trip delay. In applications where the end-to-end delays are quite large, however, it might be highly inefficient to wait for a whole round trip delay until the control is updated. In such cases, one may choose one of the following two options:
(i) Consider some intermediate nodes as virtual sources and destinations, and use a hop-by-hop flow control. In that case, one may use at each hop the optimal control schemes proposed in the previous sections.
(ii) Allow for updating of the control several times (say $\theta \geq 1$ times) during a round trip. In that case, a basic time unit corresponds to the round trip time divided by $(1+\theta)$. In an ATM-type architecture where ABR (Available Bit Rate) traffic is controlled, this is realized by sending RM (Resource Management) packets that travel from the source to the destination (forward packets) and then back from the destination to the source (backward packets). Information on queue-length and available bandwidth are then stamped onto these packets at the intermediate switches.

One may also use a combination of the two possibilities above, and allow for a hop-by-hop control where, at each hop, control is updated at times which are shorter than the round trip hop time.

We shall focus in this section on the second possibility above, and design the corresponding optimal controllers. The system can still be modeled by the dynamics described in Section 2. However, since the basic time unit is now shorter (by a factor of $1+\theta$) than the time unit considered in the previous sections, the basic parameters of the model have to be rescaled accordingly. For example μ appearing in (2) will now be one-$(\theta+1)th$ of the earlier one, and so will μ_n, x_n and u_n. In terms of the new time unit, the problem formulation will then be exactly as the one in Section 2 – with one major difference. Now the measurements $\hat{\tilde{\mu}}$ and $\hat{\tilde{q}}$ are acquired (for control purposes) with a delay of $\theta \geq 1$ time units. Hence, the controller (10) will be of the form

$$u_n = f_n(\tilde{x}_m, \hat{\tilde{\mu}}_{m-\theta}, \hat{\tilde{q}}_{m-\theta}, u_{m-1}, m = k, \ldots, n), \tag{66}$$

where k is actually taken to be $-\infty$. Here we could also have taken the dependence on $\tilde{x}_m$ to be with a delay of θ units, but since $\tilde{x}_m$ is generated by noise-free dynamics, this would not make any difference in the end result; that is, any performance that is achievable using $\tilde{x}_m$ could also be achieved using $\tilde{x}_{m-\theta}$.

We note that even though this structure was arrived at by increasing the frequency of control updates and by rescaling the parameters of the model (as discussed above), it can also be interpreted for a fixed time scale as the control having access to measurements with a delay of θ time units. In the framework of this interpretation, we denote the minimum H^∞ performance level $\gamma^*(\epsilon)$ by $\gamma_\theta^*(\epsilon)$, to show explicitly its dependence on the delay factor θ. When we vary θ (as we will do in the sequel) this will be done precisely in this framework, with the time-scale kept fixed. This now brings us to the point where we present the counterpart of Proposition 1 for the delayed measurements case. The small-noise perturbed problem is precisely the one formulated in Section 3, with the only difference being that now the control has access to $y_{m-\theta}, m \leq n$, at time n.

Proposition 2 *Consider the framework of Section 3, but with the controller having the structural form discussed above, with $\theta \geq 1$. Let $\epsilon > 0$ be fixed, and assumption A1 hold. Then:*
1) $\gamma_\theta^*(\epsilon)$ is finite, and for each fixed $\epsilon > 0$, the ordering $\theta_2 > \theta_1 \geq 1$ implies that $\gamma_{\theta_2}^*(\epsilon) \geq \gamma_{\theta_1}^*(\epsilon) \geq \gamma^*(\epsilon)$, where $\gamma^*(\epsilon)$ is as defined in Proposition 1.
2) An H^∞-controller that ensures a bound $\gamma^2 > (\gamma_\theta^*(\epsilon))^2$ in the index (26) is

$$u_{n,\epsilon,\theta}^{\gamma} = f_{\epsilon,\theta}^{\gamma}(\hat{z}_{n|n-\theta}) = -B' M_\epsilon \Lambda_\epsilon^{-1} A \left(I - \gamma^{-2} \tilde{\Sigma}_{\epsilon,\theta} M_\epsilon \right)^{-1} \hat{z}_{n|n-\theta}, \tag{67}$$

where $\hat{z}_{n|n-\theta} = \xi_n^{(n)}$, with $\xi_n^{(n)}$ being the last step of the iteration $\{\xi_k^{(n)}\}_{k=n-\theta+1}^{n}$ generated by

$$\xi_{k+1}^{(n)} = A \left(I - \gamma^{-2} \tilde{\Sigma}_{\epsilon,\theta} N \right)^{-1} \xi_k^{(n)} + B u_k \, ; \quad \xi_{n-\theta+1}^{(n)} = \check{z}_{n-\theta+1} \, , \tag{68}$$

and $\{\check{z}_n\}_{n=0}^{\infty}$ is as generated by (36). The matrix $\tilde{\Sigma}_{\epsilon,\theta}$ is the θ-th term generated by the recursion

$$\tilde{\Sigma}_{\epsilon,k+1} = A \tilde{\Sigma}_{\epsilon,k} \left(I - \gamma^{-2} N \tilde{\Sigma}_{\epsilon,k} \right)^{-1} A' + D_\epsilon D_\epsilon' \, ; \quad \tilde{\Sigma}_{\epsilon,1} = \Sigma_\epsilon \, , \tag{69}$$

satisfying the two spectral radius conditions

$$N^{1/2} \tilde{\Sigma}_{\epsilon,k} N^{1/2} < \gamma^2 I \, , \quad k = 1, \ldots, \theta \tag{70}$$

$$M_\epsilon^{1/2} \tilde{\Sigma}_{\epsilon,k} M_\epsilon^{1/2} < \gamma^2 I \, , \quad k = 1, \ldots, \theta + 1. \tag{71}$$

The initializing matrix Σ_ϵ in (69) is a minimal nonnegative-definite solution of (30), satisfying (31), and all other terms, such as M_ϵ, Λ_ϵ, D_ϵ, are as defined in Proposition 1.

If any one of the conditions of Proposition 1, or (70)-(71) do not hold, then $\gamma < \gamma_\theta^*(\epsilon)$.

Proof. The first statement above follows from the fact that in H^∞-optimal control, less on-line information on the uncertainties cannot lead to improved optimum performance, which in this case is quantified by $\gamma_\theta^*(\epsilon)$. The other statements of the Proposition follow directly from Theorem 6.6 (p. 276) of the 2nd edition of [6]. Even though this Theorem states the solution of only the finite-horizon case, its extension to the infinite-horizon case (which is dealt with here) follows by mimicking the arguments used in the non-delay case; see Theorem 5.5, p. 226 of the 2nd edition of [6]. $\diamond$

Remark 3 The ordering of the γ^*'s in part 1 of Proposition 2 is valid, as discussed prior to the statement of the Proposition, under the interpretation that the round-trip delay is fixed, and θ reflects the delay in the acquisition of the measurements. The problem of relevance, however, is the one where the number of control updates is increased (from 1 to $\theta+1$) during each round trip. Under this interpretation, the time scaling will not be fixed, and we would then expect the H$^\infty$ performance level to be a nonincreasing function of θ (instead of being nondecreasing, as in the case of Proposition 2), because the control is applied more frequently on a given time interval. $\diamond$

To present the counterpart of Theorem 1 for the $\theta \geq 1$ case, we have to study the limiting process as $\epsilon \downarrow 0$. The analysis of Section 4 applies here intact, with the exception of the control law; we also have to determine the limiting expression for $\tilde{\Sigma}_{\epsilon,k}$.

Compatible with that of $\Sigma_0 = \lim_{\epsilon\downarrow 0} \Sigma_\epsilon$, it is not difficult to show that $\tilde{\Sigma}_{0,k} =$ block-diag $\left(0,\ \tilde{P}_k\right)$, where the $(d+1)$-dimensional square matrix $\tilde{P}_k$ is generated by the recursion

$$\tilde{P}_{k+1} = \tilde{A}\tilde{P}_k\left(I - \gamma^{-2}\tilde{N}\tilde{P}_k\right)^{-1}\tilde{A}' + \tilde{D}\tilde{D}'\,; \quad \tilde{P}_1 = P\,, \tag{72}$$

where P is a minimal nonnegative-definite solution of (46), satisfying (47), and $\tilde{A}$, $\tilde{D}$ are as defined by (44) and (45), respectively. In view of this simplification, the two spectral radius conditions (70) and (71) become, respectively

$$\tilde{N}^{1/2}\tilde{P}_k\tilde{N}^{1/2} < \gamma^2 I\,, \quad k = 1,\ldots,\theta, \qquad \text{and} \qquad \tilde{M}^{1/2}\tilde{P}_k\tilde{M}^{1/2} < \gamma^2 I\,, \quad k = 1,\ldots,\theta+1\,, \tag{73}$$

where $\tilde{N}$ was defined by (41) and $\tilde{M}$ in conjunction with (48).

To obtain the limiting form of the controller, we write (compatibly with (49) and (50)):

$$\hat{z}_{n|n-\theta} = \left(\hat{\alpha}_{n|n-\theta},\ \hat{\beta}'_{n|n-\theta}\right)'\,; \qquad \xi_k^{(n)} = \left(\tilde{\alpha}_k^{(n)},\ \tilde{\beta}_k^{(n)'}\right)' \tag{74}$$

Then, some manipulations yield:

$$\begin{array}{l} \tilde{\alpha}_{k+1}^{(n)} = \tilde{\alpha}_k^{(n)} + bu_k\,; \quad \tilde{\alpha}_{n-\theta+1}^{(n)} = \left\{ \begin{array}{cc} \tilde{x}_{n-\theta+1} & n \geq \theta \\ 0 & \text{else} \end{array} \right. \\ \Rightarrow \quad \tilde{\alpha}_k^{(n)} = \tilde{x}_k\,, \quad k = 0,1,\ldots\ \forall\, n \geq 0 \qquad \Rightarrow \qquad \hat{\alpha}_{n|n-\theta} = \tilde{x}_n \end{array} \tag{75}$$

and

$$\begin{array}{c} \hat{\beta}_{n|n-\theta} = \tilde{\beta}_n^{(n)} \\ \tilde{\beta}_{k+1}^{(n)} = \underbrace{\begin{pmatrix} 0 & 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 \\ \cdot & \cdot & \cdot & \ldots & \cdot & \cdot \\ 0 & 0 & 0 & \ldots & 1 & 0 \\ 0 & 0 & 0 & \ldots & 0 & 1 \end{pmatrix}}_{(d+1)\times(d+2)} A\left(I - \gamma^{-2}\tilde{\Sigma}_{0,\theta}N\right)^{-1}\begin{pmatrix} \tilde{x}_k \\ \tilde{\beta}_k^{(n)} \end{pmatrix}\,; \\ \tilde{\beta}_{n-\theta+1}^{(n)} = \left\{ \begin{array}{cc} \breve{\beta}_{n-\theta+1} & n \geq \theta \\ 0 & \text{else} \end{array} \right. \end{array} \tag{76}$$

where $\breve{\beta}_n$, $n \geq 1$ is generated by (54). Note that in the delayed measurement case we do not have a separation between $\hat{\alpha}$ and $\hat{\beta}$ as clean as in the case with $\theta = 0$. The dynamics for $\tilde{\beta}^{(n)}$ above, and thereby those of $\hat{\beta}_{n|n-\theta}$ use also $\tilde{x}$ as an input, whereas (53) and (54) did not.

In view of these limiting expressions, the limiting expression for the controller (67), as $\epsilon \downarrow 0$, is

$$u^{\gamma}_{n,\theta} = f^{\gamma}_{\theta}(x_n, \hat{\beta}_{n|n-\theta}) = -B' M_0 \Lambda_0^{-1} A \left(I - \gamma^{-2}\tilde{\Sigma}_{0,\theta} M_0\right)^{-1} \begin{pmatrix} \tilde{x}_n \\ \hat{\beta}_{n|n-\theta} \end{pmatrix}, \tag{77}$$

This now leads to the following Theorem:

Theorem 2 *Consider the original control problem formulated in Section 2, under the H^∞ model* (M2) *of the disturbance, and with delayed measurements, with the optimum H^∞ performance level denoted by γ^*_θ. Let Assumption A1 hold. Then γ^*_θ is finite, and for all $\gamma > \gamma^*_\theta$:*
1) Statements 1)-3) of Theorem 1 hold.
2) Conditions (73) hold, and the matrix sequence generated by (72) is well defined and nonnegative definite.
3) An H^∞-controller that ensures the bound γ^2 in (12) under θ-delayed measurements is given by (77).
For a given $\gamma > 0$, if any one of the conditions above does not hold, then $\gamma \leq \gamma^*_\theta$. ◇

By letting $\gamma \to \infty$ in Theorem 2, we arrive at the following counterpart of Corollary 1 under delayed measurements.

Corollary 2 *Under assumption A1, the optimal control problem with the disturbance model M1, and θ-delayed measurements, admits the unique solution:*

$$u^{\infty}_{n,\theta} = f^{\infty}_{\theta}(x_n, \hat{\beta}_{n|n-\theta}) = -B'\overline{M}(I + BB'\overline{M})^{-1} A \begin{pmatrix} \tilde{x}_n \\ \hat{\beta}_{n|n-\theta} \end{pmatrix}, \tag{78}$$

where $\overline{M}$ is the unique nonnegative-definite solution of (59) in the class of nonnegative-definite matrices, and

$$\hat{\beta}_{n|n-\theta} = \tilde{\beta}^{(n)}_n; \qquad \tilde{\beta}^{(n)}_{k+1} = \tilde{A}\tilde{\beta}^{(n)}_k, \; \tilde{\beta}^{(n)}_{n-\theta+1} = \begin{cases} \breve{\beta}_{n-\theta+1} & n \geq \theta \\ 0 & else \end{cases} \tag{79}$$

where $\breve{\beta}_n$, $n \geq 1$, is generated by (61), and $\tilde{A}$ is as defined by (44). ◇

6 Simulation Results

First paralleling the simulations presented in [4], but for the genuine ARMA model of interfering traffic, we present in this section some simulation results using both the H$^\infty$ and Gaussian models, and with no delay in the measurements (or equivalently with control updated only once during each round trip); hence, this first set of simulations will constitute an illustration of the results of Section 4. Subsequently, we present simulations for the case of delayed information, thus illustrating the results of Section 5. For simulation purposes, the observation noises v_n and w_n are chosen as standard i.i.d. Gaussian sequences. In order to illustrate the effectiveness of the H$^\infty$ (worst-case) controller in the context of telecommunications, we present results obtained under different choices of perturbations (interfering traffic) ϕ_n (in (2)), such as:
(i) highly correlated perturbations: $\phi_n = \sin(0.2n)$,
(ii) i.i.d. Gaussian perturbations,
(iii) i.i.d. uniformly distributed perturbations (over $[-1, 1]$).

We illustrate below how the design objectives influence the behavior of the system when driven by the optimal controller. More precisely, we shall focus on the influence of the parameters g (see (9)) and b (see (7)) on the system.

The parameter g will influence the tracking between input and output rates, as can be seen from equations (5) and (9). The larger g is, the more costly the deviations between the input and output rates will be. Indirectly, this may result in reducing the variations of the queue size (see (5)).

The parameter b is expected to influence the jitter, i.e. the variability of the transmission rate $\tilde{x}$, which is given by (see (7)) $\hat{u}_n := bu_n = \tilde{x}_{n+1} - \tilde{x}_n$. It is easily seen from (7) and (9) that choosing a smaller value of b results in a higher weight on deviations of u from zero, and thus a higher weight on the variability of $\tilde{x}$. Thus, we would expect that smaller values for b will be useful when designing controllers for video or voice traffic, where the transmission rate is typically required to be regular (i.e. low jitter). (Note that if one considers $\hat{u}_n$ as the control, instead of u_n, then the cost L in (9) becomes $L = g\left\|\tilde{x} - \xi\right\|^2 + \left\|\tilde{q}\right\|^2 + b^{-2}\left\|\hat{u}\right\|^2$. This shows that indeed lower values of b result in higher weight on the jitter.)

Finally, if g is small and b is large, then the second term in (9) becomes important, that is we pay more for deviations of the queue length from its target value Q. In practice this part will be responsible for decreasing loss probabilities (that occur when queue sizes are large) and increasing throughputs (since we may lose potential throughput when the queue is empty, if, at that time, the input rate is lower than the output rate).

We shall examine three situations in the simulations below:

Case 1: The three different criteria are weighted equally in the cost L to be minimized (see (7) and (9)): $g = b = 1$.

Case 2: $b = g = 0.1$. We thus place relatively less emphasis on the tracking between input and output rates and thus on the variability of the queue length; the objective that is most weighted is that of minimizing the jitter $\hat{u}$.

Case 3: $b = 10, g = 0.1$. The objective that is most weighted is that of minimizing the deviations of the queue size from Q.

To describe the interfering traffic, we have chosen an ARMA model of order two with $\ell_1 = 0.7, \ell_2 = -0.3$, $k_1 = 1, k_2 = 2$. The parameters a and c (appearing in (3)) were chosen to be $\sqrt{2}$. We chose $Q = 30$ and $\mu = 40$ units. (A unit corresponds to the number of packets transmitted during a round trip delay.) The duration of each simulation was 140 time units. The queues were initially empty. A steady state was reached in each case after 20 time units. The design of the controller is for a steady-state behavior. We present in Table 1 below the performance (attenuation) for a transient period (first 100 units) as well as the steady-state period (last 100 units). In all cases, the performance is indeed considerably better at the steady state.

For each of the cases above, we considered both of the scenarios below:

- an almost optimal H^∞ controller, with γ very close to (slightly larger than) the optimal (smallest) value γ^* (see (15)), so as to operate under stable conditions.
- very high values of γ ($\gamma = 100$), which resulted in a controller that is (almost) optimal for the Gaussian model.

In addition to the values obtained for the attenuation according to the simulations, which are presented

	Parameters					*Simulation Results*	
Sim.	γ	b	g	*pert.*	obs. noise	s.s. attenuation	tr. attenuation
1	100	0.1	0.1	sin	sin	10.49	11.54
2	γ^*=12.03, 12.05	0.1	0.1	sin	sin	5.75	10.75
3	100	1	1	sin	sin	4.93	6.29
4	γ^*=5.17, 5.2	1	1	sin	sin	2.61	6.45
5	100	0.1	0.1	sin	Gauss.	7.90	8.87
6	γ^*=12.03, 12.05	0.1	0.1	sin	Gauss.	7.44	9.55
7	100	1	1	sin	Gauss.	3.33	4.67
8	γ^*=5.17, 5.2	1	1	sin	Gauss.	4.06	5.37
9	100	10	0.1	sin	Gauss.	2.27	3.57
10	γ^*=3.89, 4	10	0.1	sin	Gauss.	3.11	5.41
11	100	10	0.1	sin	sin	3.54	4.9
12	γ^*=3.89, 4	10	0.1	sin	sin	2.06	5.27
13	100	1	1	Gauss.	Gauss.	2.36	3.60
14	γ^*=5.17, 5.2	1	1	Gauss.	Gauss.	3.41	5.28
15	100	10	0.1	Gauss.	Gauss.	1.81	3.08
16	γ^*=3.89, 4	10	0.1	Gauss.	Gauss.	3.14	4.53
17	100	0.1	0.1	Gauss.	Gauss.	3.2	5.07
18	γ^*=12.03, 12.05	0.1	0.1	Gauss.	Gauss.	8.01	9.43
19	100	1	1	Unif.	Unif.	2.05	5.41
20	γ^*=5.17, 5.2	1	1	Unif.	Unif.	3.77	7.66
21	100	0.1	0.1	Unif.	Unif.	3.19	7.34
22	γ^*=12.03, 12.05	0.1	0.1	Unif.	Unif.	6.56	13
23	100	10	0.1	Unif.	Unif.	1.49	4.51
24	γ^*=3.89, 4	10	0.1	Unif.	Unif.	3.17	6.18

Table 1: Steady-state and transient attenuation levels in the absence of delay in measurements

in Table 1, we also present some figures (Fig. 1-4) with each one depicting a set of 4 curves: the evolution of the queue size, the controlled input versus available (output) rate over the entire simulation interval as well as over a typical shorter period (a zoomed-in curve), and the evolution of the control. (Multiplying the control evolution by the parameter b provides us with the jitter.)

In Table 2 we illustrate the effect of delay on steady-state attenuation. We chose for all simulations $b = 10$ and $g = 1$.

7 Conclusions drawn from the simulations

As expected, the curves below indicate that smaller values of b and g result in low jitter. Indeed, the jitter (bu_n) is the smallest in Figure 1 (corresponding to simulation 2) for which $g = b = 0.1$; its values are between -2 and +2. In Figure 2 (corresponding to simulation 4), in which $g = b = 1$, the jitter has much higher oscillations, of amplitude 8. In Figures 3 (corresponding to simulation 6) in which $g = 0.1, b = 10$, it has oscillations of amplitude 10. (Note that for a fair comparison of all these figures, u has to be multiplied by b).

Large b and relatively small g result in the lowest queue lengths. Indeed, in Figure 3, for which $g = 0.1, b = 10$, the amplitude of the oscillations of the queue length is 5, whereas in Figure 1 (g=b=0.1) the oscillations are of amplitude 9.

	Parameters				*Simulation Results*
Sim.	θ	γ	*pert.*	obs. noise	s.s. attenuation
a	2	100	Gauss.	Gauss.	2.45
b	2	100	Unif.	Unif.	2.76
c	2	100	sin.	Gauss.	2.34
d	2	100	sin.	sin.	3.09
e	2	γ^*=9.5, 10.5	Gauss.	Gauss.	4.51
f	2	γ^*=9.5, 10.5	Unif.	Unif.	7.20
g	2	γ^*=9.5, 10.5	sin.	Gauss.	2.52
h	2	γ^*=9.5, 10.5	sin.	sin.	2.91
i	5	100	Gauss.	Gauss.	9.84
j	5	100	Unif.	Unif.	9.77
k	5	100	sin.	Gauss.	3.57
l	8	100	Gauss.	Gauss.	16.60
m	8	100	Unif.	Unif.	14.01
n	8	100	sin.	Gauss.	7.62

Table 2: Effect of the delay θ on attenuation

Finally, relatively large values of g indeed lead to much better tracking of the input with respect to the output rate, as can be seen from Figures 2 and 4, where $g = b = 1$. We observe from the figures that relatively large values of g also lead to smaller variation in the queue length around its target level.

An interesting question that can be raised here is whether in specific applications one should choose an optimal H$^\infty$ controller (designed for the worst case behavior of noise and perturbations) or a Gaussian one. When both the noise and the perturbations are Gaussian, we see from Table 1 (Simulations 13-18) that, as expected, the Gaussian controller performs better (leads to smaller attenuation). In the presence of non-Gaussian noise, however, the H$^\infty$ controller performs sometimes better and sometimes worse (note that we did not conduct any simulations under the worst-case noise, which is typically linear in the state). Indeed, in simulations 1-4, we see that for the case of sinusoidal measurement perturbations and noise, the attenuation is better under an H$^\infty$ controller, whereas in the case of uniformly distributed perturbations and noise (simulations 19-24) the situation is reversed.

When the measurements are delayed, we see in Table 2, as to be expected, that the attenuation increases with delay. As in the nondelayed case, we see that for the case of sinusoidal measurement perturbations and noise, the attenuation is better under an H$^\infty$ controller, whereas in the other cases the situation is reversed.

8 Concluding Remarks

We have presented in this paper a control-theoretic approach to designing optimal rate-based flow controllers. Using the H$^\infty$-optimization based controller design theory that was developed in the last decade, we were able to handle very general measurement noise and interfering traffic: these may be bursty, non-stationary, non-ergodic, highly correlated and even periodic. In spite of this generality in the description of the noise and interfering traffic, the resulting (optimal) controller is easy to implement, being linear in the (augmented) state of the system.

It turned out that the underlying control problem did not fall precisely into the standard H$^\infty$ setting with imperfect information, and to bring it to the standard setting, the perfect state measurements had to be perturbed by small noise. The limit of the solution thus obtained, as the perturbation noise vanishes, provided the optimal control for the original problem.

We further studied the problem of delayed information; due to the *certainty-equivalence* property of both H^∞ and LQG optimal controllers (see [6]), it turned out that the problem with delayed information did not require an increase in the dimension of the state space (as is generally the case in optimal control problems with delayed information, see e.g. [5]). The optimal controller thus remains simple, and easy to implement also in the presence of delay, and is thus appropriate for applications in high speed networks.

The linear-quadratic control theoretic framework adopted in this paper offers an additional important advantage besides leading to a design that is *optimal* with respect to a performance index. It enables us to obtain *analytical expressions for the performance measures*, as in (62), as well as expressions for expected blocking probabilities, as in Remark 2. This is especially important in Call Admission Control (CAC) of ABR-type traffic, since, when accepting a new call, the network has to guarantee some Quality of Service which has been pre-negotiated (and includes in particular an upper bound on the cell loss rate, a lower bound on the available transmission rate, etc). To the best of our knowledge, there do not exist analytical expressions for performance measures for the control schemes proposed in [1], and this limits the possibility of the network to guarantee performance measures when using them.

The H^∞ approach allows for the measurement noises to be state-dependent. This makes it possible, in particular, to consider quantized measurements. For instance, we may consider queue length information that is conveyed in a single congestion bit, that is turned on if and only if the queue length is beyond some threshold. We are currently investigating the computation of achievable performances for such scenarios.

Effectiveness and versatility of the control schemes developed here hinge on the modeling assumption that it is the source that computes the input rates, according to the information it acquires. A major limitation of this approach in the context of the control of ABR traffic in ATM is that, according to [1], the source receives through the RM packets only two bits of information on the status of the congestion (one is used to tell the source not to increase its rate, and the other one is used to tell the source to decrease its rate; if both bits are zero, then the source may increase its sending rate). An alternative possibility would be to let the switches compute the desired input rate from the source. The advantage is that the switch indeed possesses full (non-delayed) information about its own queue length and the available bandwidth. Derivation of the optimal control in this case will be similar to those carried out in this paper. A drawback of this approach is that, in the presence of many sources and destinations, it might require high computational effort in the switches.

References

[1] The ATM Forum Technical Committee, *Traffic Management Specification*, Version 4.0, 95-0013R8, October 1995.

[2] E. Altman, F. Baccelli and J. C. Bolot, "Discrete-time analysis of adaptive rate control mechanisms", *High Speed Networks and their performance*, H.G. Perros and Y. Viniotis Eds., North Holland, 121-140, 1994.

[3] E. Altman and T. Başar, "Optimal Rate Control for High Speed Telecommunication Networks", *Proc. 34th IEEE Conference on Decision and Control*, pp. 1389-1394, December 13-15, 1995, New Orleans, Louisiana.

[4] E. Altman and T. Başar, "Optimal Rate Control for High Speed Telecommunication Networks: the case of delayed information", First Workshop on ATM Traffic Management, WATM'95, IFIP, WG.6.2 Broadband Communication, pp. 115-122, Paris, France, December 6-8 1995.

[5] E. Altman and G. Koole, "Control of a random walk with noisy delayed information", in *Systems and Control Letters*, **24**, pp. 207-213, 1995.

[6] T. Başar and P. Bernhard, *H^∞-Optimal Control and Relaxed Minimax Design Problems: A Dynamic Game Approach*, Birkhäuser, Boston, MA, 1991 (2nd edition, 1995).

[7] J.-C. Bolot, "End-to-end delay and loss behavior in the Internet," *Proc. ACM Sigcomm '93*, pp. 289-298, San Francisco, CA, Sept. 1993.

[8] J-C. Bolot and A. U. Shankar, "Analysis of a fluid approximation to flow control dynamics," *Proc. IEEE INFOCOM '92*, Florence, Italy, May 1992, pp. 2398-2407.

[9] F. Bonomi and K. W. Fendick, "The rate-base flow control framework for the Available Bit Rate ATM service", *IEEE Network*, pp. 25-39, March/April 1995.

[10] O. J. Boxma, "Sojourn times in cyclic queues – the influence of the slowest server", *Computer Performance and Reliability*, G. Iazeolla, P. J. Courtois and O. J. Boxma (Editors), Elsevier Science Publishers B.V. (North-Holland), pp. 13-24, 1988.

[11] G. Didinsky, Z. Pan and T. Başar, "Parameter identification for uncertain plants using H^∞ methods," *Automatica*, vol. 31, no. 9, pp. 1227-1250, September 1995.

[12] K. Fendick, M. A. Rodrigues and A. Weiss, "Analysis of a rate-based control strategy with delayed feedback," *Proc. ACM SIGCOMM '92*, Baltimore, MD, August 1992, pp. 136-148.

[13] V. Jacobson, "Congestion avoidance and control," *Proc. ACM SIGCOMM '88*, Stanford, CA, 314-329, August 1988.

[14] S. Keshav, "A control-theoretic approach to flow control," *Proc. ACM SIGCOMM '91*, Zurich, Switzerland, September 1991.

[15] H. T. Kung and R. Morris, "Credit-based flow control for ATM networks", *IEEE Network*, 40-48, March/April 1995.

[16] K. K. Ramakrishnan and P. Newman, "Integration of rate and credit schemes for ATM flow control", *IEEE Network*, pp. 49-56, March/April 1995.

Figure 1: Simulation 6; $\gamma = 12.03$, unequally weighted costs $g = b = 0.1$, sinusoidal perturbations and Gaussian noise, $k_1 = k_2 = 1, l_1 = 0.7, l_2 = -0.3$.

Figure 2: Simulation 8; $\gamma = 5.17$, equally weighted costs $g = b = 1$, sinusoidal perturbations and Gaussian noise, $k_1 = k_2 = 1, l_1 = 0.7, l_2 = -0.3$.

Figure 3: Simulation 10; $\gamma = 3.89$, unequally weighted costs $g = 0.1, b = 10$, sinusoidal perturbations and Gaussian noise, $k_1 = k_2 = 1, l_1 = 0.7, l_2 = -0.3$.

Figure 4: Simulation 13; $\gamma = 100$, equally weighted costs $g = b = 1$, Gaussian perturbations and noise, $k_1 = k_2 = 1, l_1 = 0.7, l_2 = -0.3$.