A Large Deviation Analysis of Errors in Measurement Based Admission Control to Buffered and Bufferless Resources

N.G. Duffield

AT&T Labs,
Room B139, 180 Park Avenue,
Florham Park, NJ 07932, USA
duffield@research.att.com

Abstract

In measurement based admission control, measured traffic parameters are used to determine the maximum number of connections that can be admitted to a resource within a given quality constraint. It has been pointed out that the assumption that the measured parameters are the true ones can compromise admission control. This is because the measured parameters are themselves random quantities, and so contribute additional variability to the attained quality.

This paper analyzes the impact of measurement error within the framework of large deviation theory. For a class of admission controls, large deviation principles are established for the number of admitted connections, and for the attained overflow rates. The technique is applied initially to the case that many sources seek admission to a bufferless resource, but it is shown how to extend to buffered resources in both the many sources and large buffer asymptotic. The sampling properties of effective bandwidths are presented, together with a discussion the impact of the temporal extent of individual samples on estimator variability. Sample correlations are shown to increase estimator variance; procedures to make admission control robust with respect to these are described.

1 Introduction

Statistical multiplexing aims to provide a balance between the opposing goals of high utilization and high quality in integrated service networks. The function of connection admission control (CAC) is to maintain this balance.

Resource sharing is motivated by the desire to carry bursty traffic that has stringent Quality of Service (QoS) requirements, such as for delay and packet loss ratio. Experimental studies have revealed both source and network traffic to be bursty, manifesting variability at many time-scales; see e.g. [19, 40, 47]. This burstiness has the consequence that trying to achieve hard guaranteed QoS targets through allocation of dedicated resources will lead to low utilization. This observation applies to peak rate allocation and, it has been argued [23], also to allocation based on leaky bucket characterization [43]. Use of the latter has been

proposed as a means of allocating buffer and bandwidth through the network [38, 39], coupled with GPS scheduling [11].

On the other hand, resource sharing by aggregating flows can provide statistical guarantees at high utilization, even in the presence of high variability. This has been established by studies using trace-driven simulation [9], by model-based simulation [28], and supported by the analysis of the queueing properties of models of long-range dependent traffic sources and their aggregates [14]. But in order to provide statistical guarantees, it is necessary to know those characteristics of the traffic which determine the bandwidth required by the aggregate in order to obtain the desired QoS. Whereas this may be predicted from models, in practice the parameters of the model are not available for real traffic. This motivates characterizing the bandwidth requirements directly through measurements, and performing admission control based on these. A number of algorithms for Measurement Based Admission Control (MBAC) have been proposed in the recent literature; see [3, 8, 9, 21, 28, 32, 35].

The general form of the CACs is as follows. Measurements are made by drawing samples from connections, either those currently admitted, or those seeking admission, or both. Measurements may be performed on individual connections, or their aggregates, or both. The measurements may be supplemented by parameters declared by the connections. The CAC specifies a rule for calculating the QoS which would be received if all connections were admitted, and if this falls within an acceptable range, the connection or connections seeking admission are admitted. At short time-scales at least, it is assumed that the connections are stationary, so that their statistical properties after admission are the same as those measured, at least in some window after admission. In some of the CACs, longer term variations are guarded against by sampling strategies which give greater influence to the recent past than the distant past, for example by exponential smoothing, or windowing. Also, the longer term flux of connections arriving and terminating will provide some robustness against badly-characterized connections.

It has been pointed out by several authors [1, 21, 24] that the Certainty Equivalent approach, in which the measured traffic parameters are assumed to be the true ones, can lead to violations of QoS guarantees, even if the connections are assumed stationary. The measurements are themselves random quantities with a statistical distribution, and hence so is the number of connections admitted. This additional variability can degrade the QoS below that expected.

Grossglauser and Tse [24] were able to quantify this effect in a heavy traffic approximation. They considered (amongst other things) admission of statistically identical connections to a bufferless resource with some target blocking probability. They consider an infinite supply of potential connections. The i^{th} such connection offers a sample X_i of its bandwidth for CAC. For each N, form unbiased estimates

$$\widehat{a}_N = N^{-1}\sum_{i=1}^{N} X_i, \quad \text{and} \quad \widehat{v}_N = (N-1)^{-1}\sum_{i=1}^{N}(X_i - \widehat{a}_N)^2 \tag{1}$$

for the population mean and variance based on the first N samples. These can be used to parameterize a normal distribution with mean $N\widehat{a}_N$ and variance $N\widehat{v}_N$ as an approximation to the distribution of the aggregate

$S_N = \sum_{i=1}^{N} X_i$. Let Q_N denote the corresponding complementary CDF. Assume that the capacity of the channel to which the connections are admitted is na where $a = \mathsf{E}[X_i]$. In the Certainty Equivalent scheme, then for a target loss ratio of $e^{-\varepsilon}$ we can admit at most the random number $\widehat{N}_n$ of connections where

$$\widehat{N}_n = \inf\{N : Q_N(an) > e^{-\varepsilon}\} - 1. \tag{2}$$

Once this number of connections have been admitted, consider their *actual* aggregate bandwidth at some time sufficiently far in the future for the connections' bandwidth to be independent of the measurements. This is modeled by $\widehat{S}_n = \sum_{i=1}^{\widehat{N}_n} Y_i$ where the Y_i have the same distribution as the X_i but are independent of them. Denote by Q the complementary CDF of the standard normal distribution. In [23] it is shown that the attained blocking probability obeys

$$\lim_{n\to\infty} \mathsf{P}[\widehat{S}_n > na] = Q(Q^{-1}(e^{-\varepsilon})/\sqrt{2}) \tag{3}$$

(This is approximately $e^{-\varepsilon/2}$ to leading exponential order). We can interpret the factor $\sqrt{2}$ in (3) as saying that the effect of potential measurement error is to double the effective variance of the estimated bandwidth.

In order to be robust with respect to measurement errors and attain a blocking probability $e^{-\varepsilon}$, the admission control must "aim high" by (approximately) allocating resources as though the target probability were $e^{-2\varepsilon}$ rather than $e^{-\varepsilon}$. Robustness was also the focus from a Bayesian approach to analysis of admission of 2-state sources using Chernoff bounds by Bean [1] and Gibbens, Kelly and Key [21]; a Bayesian approach to predictive ruin probabilities using Dirichlet priors has been taken by Ganesh et. al. [16, 18]. Some effects of uncertainty between different source types was discussed by Courcoubetis and Weber [7].

In this paper the impact of measurement error is analyzed within the large deviation context. In this formulation, Gaussian approximations of the heavy traffic approximation in [24] are replaced by exponential approximations to the law of large numbers. The regime in which the latter approximations would be applied is that in which the target loss ratio is exponentially small in the scale of the typical number of admitted connections.

We separate out Large Deviation Principles for the number of sources admitted, and for the attained loss ratio. The analysis applies to a class of admission controls including those based on the measured large deviation properties of traffic. We investigate this class in some detail, and show how admission control can be made robust with respect to certain types of measurement error. We work initially in the many sources asymptotic for a bufferless resource, but the method is readily generalized to treat buffered resources, in both the large-buffer and many-sources asymptotic regimes. Moreover, we analyze the effects of correlations between measurements.

In more detail, our contributions are as follows.

(i) We recast the bufferless admission control problem within the framework of large deviations theory in the many-sources asymptotic regime, as exploited in [26]. In this framework, the asymptotic properties of the aggregates S_n for large n are described by a Large Deviation Principle (LDP); roughly

speaking this says that for large n

$$\mathsf{P}[S_n = nx] \approx e^{-nK(x)} \tag{4}$$

for some large deviation rate function K. $n^{-1}S_n$ converges almost surely to $a = \mathsf{E}[X_i]$; correspondingly $K(a) = 0$. (We outline the standard terms and tools that we use from Large Deviation theory in Section 8). For a target blocking probability of $e^{-n\varepsilon}$ and service rate nc, then knowing the actual distribution of the X_i, we would admit $N = \lfloor nm \rfloor$ connections, where m is the largest value such that

$$\mathsf{P}[S_N = nc] \approx e^{-nmK(c/m)} = e^{-n\varepsilon}. \tag{5}$$

Here we have used (4) to find the approximate form for $\mathsf{P}[S_N = nc]$.

We consider a class of CACs which predict the parameters of this asymptotic behavior from measurements. The number of connections $\widehat{N}_n$ to be admitted is now a random variable, being a function of the bandwidth samples X_i. We establish an LDP which says that for large n, the distribution of $\widehat{N}_n$ behaves for large n approximately as

$$\mathsf{P}[\widehat{N}_n = nx] \approx e^{-nJ(x)}, \tag{6}$$

for some large deviation rate function J whose form we derive. As $n \to \infty$, $n^{-1}\widehat{N}_n$ converges almost surely to m. In particular, $J(m) = 0$. The LDP holds because we can show that $n^{-1}\widehat{N}_n$ is, essentially, a continuous function of the empirical measure of the first n samples of X_i. As we explain, the LDP then follows by combining the Contraction Principle with Sanov's Theorem. In fact, this argument allows the conclusion of such an LDP for any admission control with such a continuity property.

(ii) We establish an LDP to describe the tail behavior of the aggregate of admitted connections. For large n,

$$\mathsf{P}[\widehat{S}_n = nc] \approx e^{-nI(c)}, \quad \text{where} \quad I(c) = \inf_y \{J(y) + yK(c/y)\}. \tag{7}$$

We can interpret this as saying that $e^{-nJ(y)}$ is roughly the probability to admit ny connections; conditional on this, $e^{-nyK(c/y)}$ is roughly the probability that $S_{ny} = nc$. Taking the infimum then picks out the y for which the product of these two probabilities is maximized. Inserting the value $y = m$ into the infimum (7), we see that $I(c) \leq mK(c/m) = \varepsilon$. This reflects the increase in the *attained* blocking probability as compared with that expected by Certainty Equivalence.

(iii) The exact form of the LDP depends on the detail of the CAC algorithm used. We give the forms of these LDPs for two CAC algorithms. One of these algorithms estimates the cumulant generating function of the bandwidth offered by connections through its empirical distribution. The other algorithm uses a quadratic approximation to the cumulant generating function which is parameterized by the empirical mean and variance. In both cases, a second order approximation agrees with the result of Grossglauser and Tse described above. The work described so far is presented in Section 2.

(iv) In Section 3, we consider the impact of Markovian correlations in the samples. At a bufferless resource it might be though that temporal correlations in the connections might not be of interest. However, if connections are sampled multiple times, the correlation can slow down convergence of measured estimates as the number of samples increases. We illustrate how the presence of Markovian correlations modifies the LDP derived in (i).

(v) Admission control can be made robust if the statistics of measurements errors are explicitly taken into account. We describe two approaches in Section 4. In the second order approximation one can adjust upwards the value of ε used to calculate the number of admitted sources; this method was available from [24]; here we extend it to cover Markovian samples too. More generally, the rate function in the LDP for the number of admitted sources can be used to estimate the required reduction in the number of admissions if the true sample distribution lie in some known set, but is otherwise unknown. This generalizes a previous scheme for two-level arrivals [1, 21]. Moreover, the approach can be further generalized to any models for which the sampling rate function is known; we apply it in the Markovian case.

(iv) The framework is sufficiently general to allow extensions to other regimes. In Section 5 we extend the theory to cover admission to buffered resources in the many-source asymptotic in the large deviation formalism that has been described by several authors [2, 7, 14, 42, 45]. The CAC algorithm is based upon estimation of the central objects used in these cited papers, namely the transient cumulant generating function of the arrival process. This approach has been implemented in [35]

(vi) Finally, in Section 6 we consider the problem of characterizing the effective bandwidth of a single connection from a number of samples. Here, the effective bandwidth [25, 29] is that appearing in the large-buffer asymptotic (see e.g. [4, 15, 20, 22, 31, 46]) rather than the many-sources asymptotic considered hitherto. The central object in this description is the (limiting) cumulant generating function. CAC based on measuring this directly has been proposed in [13] and implemented in [9]. We identify an interesting interaction between sampling by taking a large number of samples, and sampling in which the individual samples are extended over time. The latter is important for sampling the behavior of the arrival process at long timescales. Depending on the constraints under which sampling is performed, there may be a finite optimum length for the individual samples.

We summarize in Section 7; the longer proofs are given in Section 8.

2 LDP for Attained Loss Rates at a Bufferless Resource

2.1 LDP for the Number of Connections Admitted

The first part of our program is to demonstrate LDPs for the number of connections admitted. Each CAC rule potentially gives rise to a different LDP. The CAC rules that we consider in this section have the following

common framework. The connections seeking admission have bandwidth processes which are independent and identically distributed, with common marginal distribution ρ. As described in the introduction, we consider a potentially infinite supply of connections, the i^{th} connection providing a sample X_i of its bandwidth for admission control. Likewise Y_i represents the bandwidth of the connection in the distant future. All the X_i and Y_i are mutually independent with common distribution specified by the measure ρ on $\mathbb{R}_+$. Based on values of a finite subset $X^{(n)} = (X_1, X_2, \ldots, X_n)$, we will admit some number $\widehat{N}_n$ of connections. $\widehat{N}_n$ should be the largest number such that the probability that $\sum_{i=1}^{\widehat{N}_n} X_i$ exceeds a capacity nc is sufficiently small. Thus we work in a scaling in which the capacity scales proportionately with n, which, as we shall see, is also the scaling of the mean number of connections admitted.

Before considering specific admission control schemes, we describe a class of CAC for which an LDP exists for the number of admitted connections. An estimator of the distribution ρ is furnished from each sample set $X^{(n)}$ by the empirical measure $\widehat{\rho}_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i}$, where δ_x is the measure with unit mass at x. We regard an empirical measure $\widehat{\rho}_n$ as an element of $\mathcal{M}$ the space of probability measures on $\mathbb{R}_+$ equipped with the weak topology. Sanov's theorem [10] tells us that the $\widehat{\rho}_n$ satisfy an LDP with scale n and rate function $\nu \mapsto D(\nu, \rho)$, where for two probability measures ν, ρ, $D(\nu, \rho)$ is the entropy of ν relative to ρ:

$$D(\nu, \rho) = \int d\nu(x) \log \frac{d\nu}{d\rho}(x) \tag{8}$$

if ν is absolutely continuous w.r.t. ρ, and $+\infty$ otherwise. The basis of the LDP for $\widehat{N}_n$ is this: if $n^{-1}\widehat{N}_n$ is a continuous function $\widehat{\rho}_n$, the large deviation properties of $\widehat{N}_n$ as $n \to \infty$ follow from the Contraction Principle [10], formally at least.

We now give axioms for a class of admission control for which this argument can be made precise. We assume a control specified by a capacity function $C : \mathbb{R}_+ \times \mathbb{N} \times \mathcal{M} \to \mathbb{R}_+$, with $C(\varepsilon, N, \rho)$ being the minimum capacity required in order that an aggregate of N independent connection each with bandwidth distributed as ρ exceeds the capacity with frequency no more than $e^{-\varepsilon}$. We shall speak of ε as a "quality", higher values of ε corresponding to lower overflow probabilities. Let α_n be a sequence converging to some $\alpha \in (0, \infty)$, and β_n a sequence increasing to infinity. We assume that the target loss probability, the number of samples, and the capacity scale as $e^{-\beta_n \varepsilon}$, $\alpha_n n$ and nc respectively; the number of admitted connections is then

$$\widehat{N}_n = \inf\{N : C(\beta_n \varepsilon, N, \widehat{\rho}_{\alpha_n n}) \geq nc\} - 1; \tag{9}$$

Set $C_n(\varepsilon, m, \rho) = n^{-1} C(\beta_n \varepsilon, \lfloor nm \rfloor, \rho)$.

Theorem 1 *Assume that for each n, the (left) inverse $m_{n,\rho} = C_n(\varepsilon, \lfloor n \times \cdot \rfloor, \rho)^{-1}(c)$ is a weakly continuous function of ρ, converging uniformly as $\to \infty$ to a continuous limit, which we write as m_ρ. Then $n^{-1}\widehat{N}_n$ satisfies an LDP with scale n and good rate function*

$$J_\rho(x) = \inf_{\nu \in \mathcal{M}:\ m_\nu = x} \alpha D(\nu, \rho). \tag{10}$$

We will call J_ρ the *sampling rate function*. Note J_ρ depends on ε and c through the function $\rho \mapsto m_\rho$. Clearly $J_\rho(m)$ takes its minimum, 0, at $m = m_\rho$. Note that J_ρ is not necessarily convex.

If rather than taking a number of samples proportion to the capacity, we instead take samples from connection as they attempt admission, (9) is replaced by

$$\widetilde{N}_n = \inf\{N : C(\beta_n\varepsilon, N, \widehat{\rho}_N) \geq nc\} - 1. \tag{11}$$

Formally, if we take approximately nm samples then the LDP for $n^{-1}\widetilde{N}_n$ should hold with $\alpha = m_\rho$ in (10). In fact, the corresponding asymptotic upper bound is then a simple consequence of $\widetilde{N}_n > nm$ implying $C_n(\varepsilon, m, \widehat{\rho}_{nm}) < c$ which we state without proof.

Theorem 2 $\limsup_{n\to\infty} n^{-1} \log \mathsf{P}[\widetilde{N}_n \geq nm] \leq -\inf_{x\geq m} \widetilde{J}_\rho(x)$ *where* $\widetilde{J}_\rho(x) = \inf_{\nu\in\mathcal{M}:\ m_\nu = x} xD(\nu, \rho)$.

In the sequel, except where stated, we shall establish large deviation limits based on Theorem 1 and set $\alpha = 1$. Generalization to arbitrary $\alpha \in (0, \infty)$ comprises multiplying the sampling rate function by α. Large deviation upper bounds for the framework of Theorem 2 can be obtained upon substitution of J_ρ by $\widetilde{J}_\rho$ in all cases. This leads to conservative estimators of loss probabilities and numbers of admitted connections.

2.2 CAC Using the Measured Cumulant Generating Function

Large deviation theory provides an asymptotic description of loss probability that can be used to formulate CACs of the type described in the previous section. The large deviation behavior is determined by the cumulant generating function (CGF) μ_ρ of the measure ρ namely

$$\log \mathsf{E}[e^{\theta X_1}] = \mu_\rho(\theta) := \log\langle\rho, g_\theta\rangle, \tag{12}$$

where $g_\theta(x) = e^{\theta x}$ and $\langle\rho, f\rangle = \int d\rho(x) f(x)$. A CAC scheme will associate with each empirical measure $\widehat{\rho}_n$ an estimator $\widehat{\mu}_{\widehat{\rho}_n}$ of the true CGF. One such estimator is the CGF $\mu_{\widehat{\rho}_n}(\theta)$ of the empirical measure, although we shall also consider another choice.

The CAC schemes that we consider will determine the number of connections to be admitted in the following manner, using $\widehat{\mu}_n := \widehat{\mu}_{\widehat{\rho}_n}$ in a large deviation approximation. Define the Legendre transform μ^* of an CGF μ by

$$\mu^*(x) = \sup_\theta\{x\theta - \mu(\theta)\}. \tag{13}$$

Let $S_n = \sum_{i=1}^n X_i$ denote the partial sums of the X_i. From Cramer's theorem [10] we know that for $c \geq a_\rho := \mathsf{E}[X_1]$,

$$\lim_{n\to\infty} n^{-1} \log \mathsf{P}[S_n \geq nc] = -\inf_{c'>c} \mu^*_\rho(c'). \tag{14}$$

Thus an estimated CGF $\widehat{\mu}_n$ furnishes the large deviation estimate

$$\mathsf{P}[S_n > nc] \approx e^{-(n \inf_{c'>c} \widehat{\mu}_n)^*(nc')}. \tag{15}$$

Thus, in the terminology of the previous section, for a given estimator $\widehat{\mu}$, we set $\beta_n = n$ and take for the capcity function

$$C(\varepsilon, N, \rho) = \inf\{c : (N\widehat{\mu}_\rho)^*(c) \le \varepsilon\}. \tag{16}$$

Note that $n^{-1}C(n\varepsilon, nm, \rho) = C(\varepsilon, m, \rho)$: the convergence assumption in Theorem 1 satisfied and consequently

$$m_\rho = \inf\{m : (m\widehat{\mu}_\rho)^*(c) \le \varepsilon\}. \tag{17}$$

Continuity of m_ρ will have to be checked for each $\widehat{\mu}$ considered.

Later on we will describe two CAC rules based on different schemes for estimating CGFs $\widehat{\mu}_n$ from the empirical measures $\widehat{\rho}_n$. In the *direct CGF* rule, the CGF is derived directly from the empirical distribution of the bandwidth samples: $\widehat{\mu}_n = \mu_{\widehat{\rho}_n}$. The *sample mean–variance* (SMV) rule, applied to Gaussian random variables, uses a quadratic approximation to the CGF parameterized by the bandwidth's mean and variance.

2.3 LDP for the Attained Loss Rate

$\widehat{S}_n = \sum_{i=1}^{\widehat{N}_n} Y_i$ is the admitted bandwidth, random through both the statistical properties of the Y_i and the number of connections admitted $\widehat{N}_n$. In this section we determine the asymptotic behavior of the attained loss rate $\mathsf{P}[\widehat{S}_n \ge nc]$ for large n.

Theorem 3 *Assume that $n^{-1}\widehat{N}_n$ satisfies an LDP with scale n and rate function αJ_ρ having the property that $\lim_{x \searrow 0} J_\rho(x) = \infty$. Then $n^{-1}\widehat{S}_n$ satisfies an LDP with scale n and rate function*

$$I(x) = \inf_{y>0} (\alpha J_\rho(y) + y\mu_\rho^*(x/y)). \tag{18}$$

When $x \ge m_\rho a_\rho$,

$$I(x) = \inf_{m_\rho \le y \le x/a_\rho} (\alpha J_\rho(y) + y\mu_\rho^*(x/y)). \tag{19}$$

The condition $x \ge m_\rho a_\rho$ in (19) says that the resources x are at least sufficient for the mean demand $m_\rho a_\rho$.

The *attained* loss rate is determined by the attained quality $I(c)$. Theorem 3 shows the reduction in quality incurred by measurement error since $I(c) \le \alpha J_\rho(m_\rho) + (m_\rho\mu_\rho)^*(c) = (m_\rho\mu_\rho)^*(c) \le \varepsilon$. However, the amount of reduction decreases as α, the proportionate number of measurements made, increases. As $\alpha \to \infty$, the location of the infimum in (18) approaches m_ρ, and the attained quality $I(c)$ approaches the target quality ε. In the following two sections we apply Theorem 3 to two examples of CAC rules.

2.4 CAC with the Direct CGF Rule

In the Direct CGF rule, the estimator $\widehat{\mu}_{\widehat{\rho}}$ is just $\mu_{\widehat{\rho}} = \log\langle\widehat{\rho}, g_\theta\rangle$. Set the estimated CGF $\widehat{\mu}_n = \mu_{\widehat{\rho}_n}$. Let p_ρ denote the essential supremum of the support of $\rho \in \mathcal{M}$. Define $\mathcal{M}_p = \{\rho \in \mathcal{M} : p_\rho \le p\}$, the measures in $\mathcal{M}$ with support in $[0, p]$. In order to apply Theorem 3, the main work is in establishing the weak continuity of $\rho \mapsto m_\rho$. In this case (17) becomes $m_\rho = \inf\{m : (m\mu_\rho)^*(c) \le \varepsilon\}$.

Theorem 4 *Assume X to be bounded, i.e. $\rho \in \mathcal{M}_p$ for some $p > 0$, and let $\varepsilon, c > 0$.*

(i) *$m \mapsto (m\mu_\rho)^*(c)$ is strictly decreasing for $m < c/a_\rho$, is convex, and maps $(c/p_\rho, c/a_\rho]$ to $[0, ((c/p_\rho)\mu_\rho)^*(c))$.*

(ii)

$$m_\rho = \begin{cases} \text{the unique solution in } (c/p_\rho, c/a_\rho] \text{ of } (m\mu_\rho)^*(c) = \varepsilon & \text{if } \varepsilon < ((c/p_\rho)\mu_\rho)^*(c) \\ c/p_\rho & \text{otherwise} \end{cases} . \quad (20)$$

(iii) *For each $p > 0$, the map $\rho \mapsto m_\rho$ is continuous on $\mathcal{M}_p$ when the latter is equipped with the weak topology inherited from $\mathcal{M}$.*

(iv) *$n^{-1}\widehat{N}_n$ and $n^{-1}\widehat{S}_n$ satisfy LDPs with scale n and rate function as described in Theorems 1 and 3 respectively.*

The expression for J_ρ in Theorem 1 is not convenient for calculations. In the next theorem we reduce the evaluation of J_ρ to a variational calculation in two dimensions. In Theorem 3 we were primarily interested in calculating $J_\rho(m)$ for $m \ge m_\rho$: this corresponds to more connections having been admitted than would be with perfect knowledge of ρ. We will find it convenient to define

$$\lambda_f(\phi) := \log\langle\rho, e^{\phi f}\rangle. \quad (21)$$

Note $\lambda_{\mathrm{id}} = \mu_\rho$ for id the identity function. λ_{g_θ} is the CGF of samples $e^{\theta X_i}$ used to form the estimates $\widehat{\mu}_n$. For simplicity we shall use the notation λ_θ in place of λ_{g_θ}.

Theorem 5 *(i) If $y \ge m_\rho$ then*

$$J_\rho(y) = \inf_\theta \lambda_\theta^*(e^{(c\theta-\varepsilon)/y}) \quad \text{where} \quad \lambda_\theta(\phi) = \log \mathsf{E}[\exp(\phi e^{\theta X_1})]. \quad (22)$$

(ii) *If $y \ge m_\rho$ then $\lambda_\theta^*(e^{(c\theta-\varepsilon)/y}) = \sup_{\phi \le 0}(\phi e^{(c\theta-\varepsilon)/y} - \lambda_\theta(\phi))$, i.e. the supremum can be restricted to $\phi \le 0$.*

(iii) *If $y \le c/a_\rho$ (resp. $y \ge c/a_\rho$), the infimum over θ in (22) can be restricted to $\theta \ge 0$ (resp. $\theta \le 0$).*

A consequence of Theorem 5 is that in order to evaluate $I(c)$, we can use (19) and (22) with the extrema for the latter restricted to $\phi \leq 0$ and $\theta \geq 0$. The proof of Theorem 5 is based upon the following proposition:

Proposition 1 *Assume λ_f to be essentially smooth (see Section 2.3. of [10]). Then*

$$\inf_{\nu:\langle \nu, f\rangle = k} D(\nu, \rho) = \lambda_f^*(k). \tag{23}$$

Example 1: Bernoulli Connections. We consider Bernoulli processes, taking the values 0 and 1 with probabilities $1-a$ and a respectively. (Consistent with the previous notation, a is then the mean arrival rate per connections). Then we have

$$\mu(\theta) = \log(ae^{\theta} + (1-a)) \tag{24}$$

$$\mu^*(x) = \begin{cases} x\log(x/a) + (1-x)\log(1-x)/(1-a), & x \in [0,1] \\ +\infty, & \text{otherwise} \end{cases} \tag{25}$$

$$\lambda_\theta(\phi) = \log(ae^{\phi e^{\theta}} + (1-a)e^{\phi}) \tag{26}$$

$$\lambda_\theta^*(x) = \begin{cases} -\log\left(a\left(\frac{(1-a)(x-1)}{a(e^{\theta}-x)}\right)^{\frac{e^{\theta}-x}{e^{\theta}-1}} + (1-a)\left(\frac{(1-a)(x-1)}{a(e^{\theta}-x)}\right)^{\frac{1-x}{e^{\theta}-1}}\right), \\ \qquad x \in [1, e^{\theta}] \text{ (or } [e^{\theta}, 1]) \text{ as } \theta > 0 \text{ (or } \theta < 0) \\ +\infty, \qquad \text{otherwise} \end{cases} \tag{27}$$

Here $\lambda_\theta^*(x)$ is defined by continuity at the endpoints of its effective domain, namely $x = 1$ and $x = e^{\theta}$. Numerical minimization can then be used to identify the rate functions J and I. We illustrate these graphically in Figure 1 for the parameters $a = c = 0.5$ and target quality $\varepsilon = 0.2$.

2.5 CAC with the Sample Mean–Variance Rule

The SMV rule is to parameterize an empirical measure $\widehat{\rho}$ of Gaussian bandwidths X_i by the sample mean and variance $\widehat{a}$ and $\widehat{v}$. (The same asymptotic results hold if we use the unbiased estimate of the population variance). The estimate of the CGF is $\widehat{\mu}_{\widehat{\rho}} = \mu_{\widehat{a},\widehat{v}}$ where

$$\mu_{a,v}(\theta) = a\theta + v\theta^2/2, \tag{28}$$

i.e. $\mu_{a,v}$ is the CGF of a Gaussian r.v. with mean a and variance v. $\mu_{a,v}$ has Legendre transform

$$\mu_{a,v}^*(x) = (x-a)^2/(2v). \tag{29}$$

Let a_ρ and v_ρ denote the mean and variance from a measure ρ. In this case (17) becomes

$$m_\rho = \inf\{m : (m\mu_{a_\rho, v_\rho})^*(c) \leq \varepsilon\}. \tag{30}$$

One shows from (29) and (30) that m_ρ takes the extended real value

$$m_\rho = \begin{cases} \left(\varepsilon v_\rho + a_\rho c - \sqrt{\varepsilon v_\rho(\varepsilon v_\rho + 2a_\rho c)}\right)/a_\rho^2 & a_\rho > 0 \\ +\infty & a_\rho \leq 0 \end{cases} \tag{31}$$

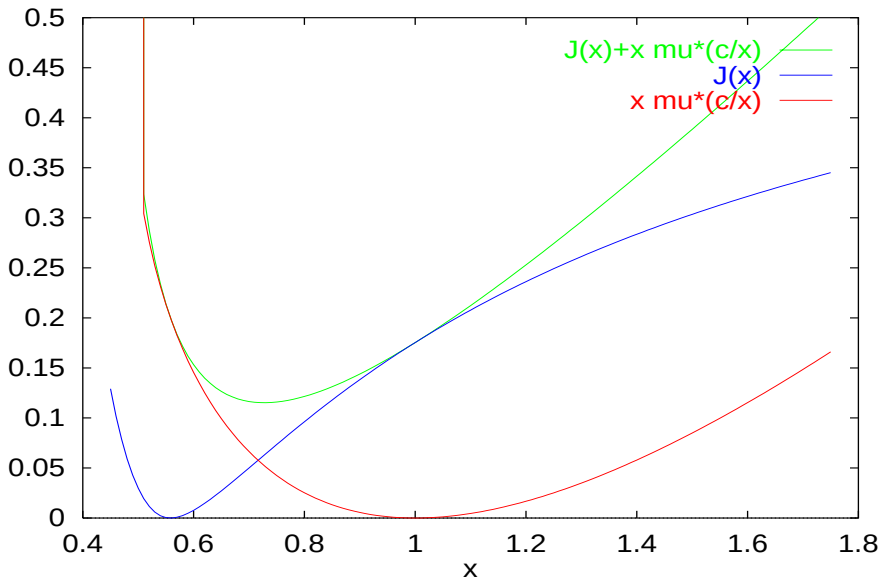

Figure 1: RATE FUNCTIONS FOR BERNOULLI CONNECTIONS At $x = m_\rho \approx 0.56$, $J(x) = 0$ and $x\mu^*(c/x) = \varepsilon$, corresponding to asymptotically most likely proportionate number of admitted connections for target quality ε. Asymptotic attained quality $\varepsilon' \approx 0.12' < \varepsilon$ occurs at the minimum $x = m' \approx 0.73 > m_\rho$ of function $J(x) + x\mu^*(c/x)$. The proportionately larger number of connections admitted due to measurement error has led to a lower attained quality.

m_ρ is an extended real-valued continuous function of the parameters (a_ρ, v_ρ). Thus if we can show that the parameters $(\widehat{a}_n = a_{\widehat{\rho}_n}, \widehat{v}_n = v_{\widehat{\rho}_n})$ satisfy an LDP, the LDP for $m_{\widehat{\rho}_n}$ will follow by the Contraction Principle.

Theorem 6 *Assume* $a_\rho, v_\rho > 0$.

(i) The parameters $(\widehat{a}_n, \widehat{v}_n)$ *satisfies an LDP with scale* n *and rate function*

$$G(a,v) = \inf_{\nu: a_\nu = a, v_\nu = v} D(\nu,\rho) = \frac{(a-a_\rho)^2}{2v_\rho} + \frac{1}{2}\left(\left(\frac{v}{v_\rho} - 1\right) - \log\left(\frac{v}{v_\rho}\right)\right) \tag{32}$$

(ii) $n^{-1}\widehat{N}_n$ *satisfies an LDP with scale* n *and rate function*

$$J_\rho(x) = \inf_{a,v: m_{a,v} = x} G(a,v). \tag{33}$$

The infimum is achieved when

$$a = a(x) := c/x + \frac{a_\rho - c/x - \sqrt{(a_\rho - c/x)^2 + 4v_\rho(1 + x/(2\varepsilon))}}{2(1 + x/(2\varepsilon))}, \tag{34}$$

$$v = v(x) := x(a(x) - c/x)^2/(2\varepsilon). \tag{35}$$

(iii) $n^{-1}\widehat{S}_n$ *satisfies an LDP with scale* n *and rate function as described in Theorem 3.*

Example 2. Gaussian Connections A single numerical minimization applied to the rate functions of Theorem 6 yields the attained quality. We illustrate this in Figure 2 for $a_\rho = c = 0.5$ and $v_\rho = 0.125$.

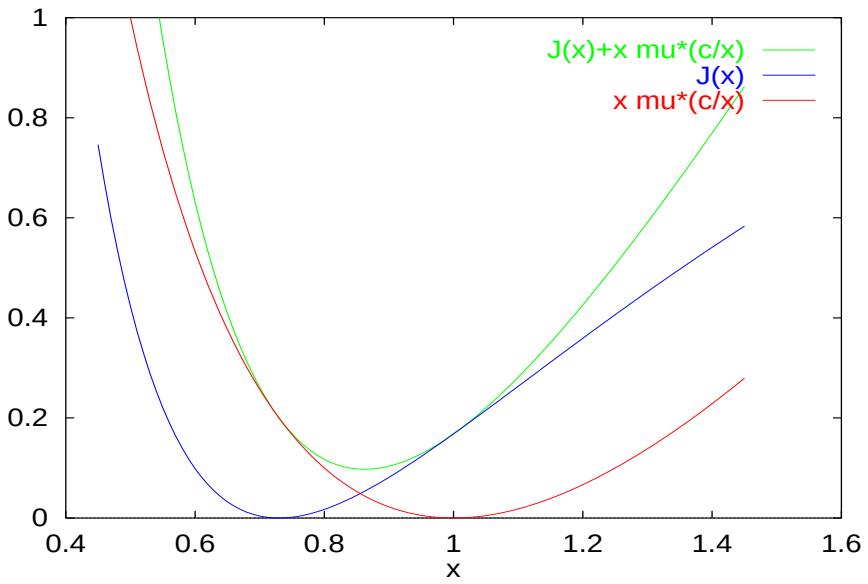

Figure 2: RATE FUNCTIONS FOR GAUSSIAN ARRIVALS Qualitatively similar to Figure 1 (see description there). $m_\rho \approx 0.73$. Reduced quality $\varepsilon' \approx 0.10 < \varepsilon$ at proportionate number of admissions $m' \approx 0.86 > m_\rho$.

2.6 Local Analysis

In this section we consider the heavy traffic limit regime $\varepsilon \to 0$. Until now we have suppressed the explicit dependence of I on the target quality ε; this enters through the dependence of m_ρ, and hence J_ρ, on ε. We indicate this dependence by writing $m_{\rho,\varepsilon}$ and the rate functions for $\widehat{S}_n$ and $\widehat{N}_n$ as I_ε and J_ε. Taking $\varepsilon \to 0$ is a heavy traffic limit in which the log-probability of overflow per connection becomes small. The following result establishes the relationship of the target quality ε, to the achieved quality $I_\varepsilon(c)$ in this limit for the Direct CGF rule; the identical result holds, with somewhat simpler detail, for the SMV rule. Set $H(y) = y\mu^*(c/y)$.

Theorem 7 *Assume H to be twice continuously differentiable on some interval $[m, c/a_\rho]$ and that each J_ε is twice continuously differentiable on $[m_{\rho,\varepsilon}, c/a_\rho]$. Then* $\lim_{\varepsilon\to 0} \varepsilon^{-1} I_\varepsilon(c) = (1 + c/(a\alpha))^{-1}$.

With $a = c$ we see that, asymptotically for $\varepsilon \to 0$, the attained blocking probability is the square root of the target blocking probability. To leading exponential order, this is the same reduction as observed in (3). We illustrate in Figure 3 by using numerical minimization to plot $\varepsilon^{-1} I_\varepsilon(c)$ near $\varepsilon = 0$ for Examples 1 and 2. In both cases $a = c$ and $\alpha = 1$, and the calculations are consistent with the theoretical limiting values $1/2$ as $\varepsilon \to 0$.

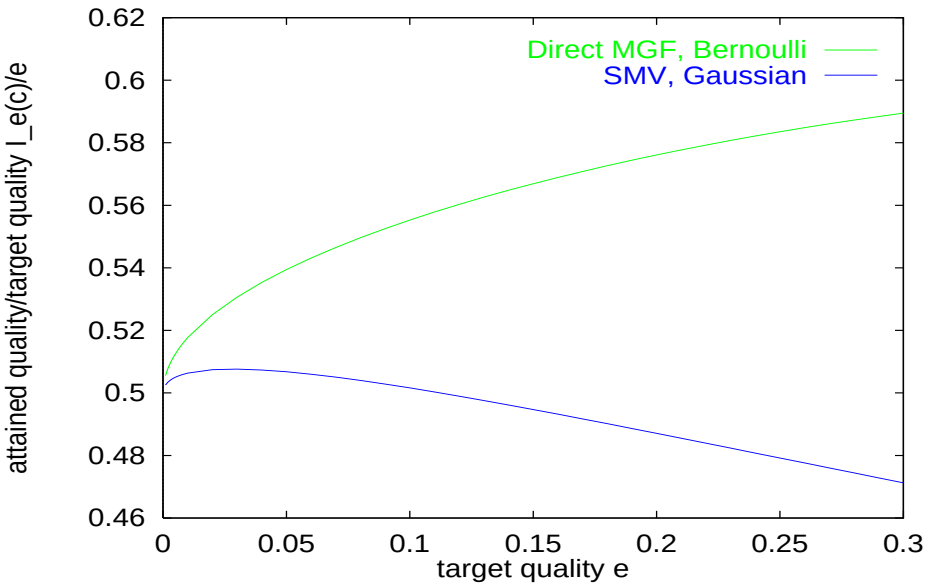

Figure 3: HEAVY TRAFFIC LIMIT: $\varepsilon^{-1} I_\varepsilon(c)$ approaches $1/2$ as $\varepsilon \to 0$, for Direct CGF applied to Bernoulli connections, and SMV applied to Gaussian connections.

3 Impact of Correlations Between Samples

So far, the samples X_i used for bandwidth estimation have been assumed independent; in this section we include the possibility of correlations between samples. We envisage that samples may be drawn from different connections, or from a single connection at different times, or some combination of these. Temporal correlations within individual connections have no impact when samples are drawn from independent instances of a process. But when samples are drawn sequentially from the same process, the presence of temporal correlations between the samples will effect the rate function which determines convergence of estimated connections characteristics to their true values. The presence of positive correlations causes sample means to converge more slowly to their true values as the sample size increases. Therefore we expect positively correlated samples will provide less accurate estimation of marginal quantities such as the empirical measure and all estimators based upon it.

3.1 LDPs for Empirical Measures of Markov Processes

In this section we assume that the X_i form a discrete-time stationary Markov process. The bandwidths of accepted connections are still a sequence of i.i.d. random variables Y_i, independent of X_i but with the same marginal distribution. (The Markov property of the samples X_i models an artifact of the sampling procedure). Recall that for independent samples we used Sanov's theorem to characterize the large deviation behavior of the empirical measures $\widehat{\rho}_n$. When the correlations are Markovian, we can quantify their effect by applying the large deviation theorem for empirical measures of a Markov process due to Donsker and Varadhan [12]. There are generalizations to k-step Markov processes or functions of Markov processes.

These could be used, for example, to formulate the appropriate LDP when the samples $(X_i(t))_{i\in\mathbb{N}}$ are constructed using a jumping window (of non-overlapping samples) or a sliding window (of overlapping samples).

For simplicty we assume that X_i form a discrete-time stationary Markov process in $\Omega \subset \mathbb{R}_+$ whose transtion measure $\pi(x, B) = \mathsf{P}[X_{i+1} \in B \mid X_i = x]$ satisfies the following mixing condition. For some probability measure β on Ω and $0 < a < b$, and some integer $j > 0$, then for all $x \in \mathbb{R}_+$

$$a\beta(B) < \pi^j(x, B) < b\beta(B). \tag{36}$$

where π^j denotes the j–step transition measure. If Ω is finite, irreducibility of (X_i) implies that (36) holds. Set $\mathcal{M}^{(2)} = \mathcal{M}(\Omega \times \Omega)$ and let $p_1\omega, p_2\omega$ denote the marginals of $\omega \in \mathcal{M}^{(2)}$, and set $\mathcal{M}(\rho, \sigma) = \{\omega \in \mathcal{M}^{(2)} \mid p_1\omega = \rho,\ p_2\omega = \sigma\}$, and $\mathcal{M}(\sigma) = \mathcal{M}(\sigma, \sigma)$. Define $\rho \otimes \pi \in \mathcal{M}^{(2)}$ by $(\rho \otimes \pi)(A \times B) = \int_A \rho(dx)\pi(x, B)$.

Proposition 2 *Under condition (36), the empirical measures $\widehat{\rho}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$ satisfy an LDP in $\mathcal{M}$ (with the weak topology) with scale n and rate function*

$$K(\nu) = K(\nu, \pi) = \inf_{\omega\in\mathcal{M}(\nu)} D(\omega, \nu \otimes \pi). \tag{37}$$

3.2 Direct CGF Estimation for Markov Chains

Equipped with Proposition 2 for the empirical measure of Markov chains, we can now obtain the LDP for attained overflow by a line of argument parallel that of Section 2.4. We preface this by recording the properties of the appropriate cumulant generating function.

Theorem 8 *Assume a transition measure π satisfying (36).*

(i) Let f be a bounded continuous function on Ω. The limit

$$\lambda^{(2)}_{f,\pi}(\phi) = \lim_{n\to\infty} n^{-1} \log \mathsf{E}[\exp(\phi \sum_{i=1}^n f(X_i))] \tag{38}$$

exists. Furthermore, $e^{\lambda^{(2)}_{f,\pi}(\phi)}$ is the unique maximal eigenvalue of the kernel $\pi_f(\phi; x, dy) = \pi(x, dy)e^{\phi f(y)}$; the (left) eigenmeasure $\eta(\phi)$ and (right) eigenfunction $\psi(\phi)$ of $\pi(\phi)$ are such that $(d\eta/d\beta)(\phi, \cdot)$ and $\psi(\phi, \cdot)$ are uniformly positive and bounded. $\lambda^{(2)}_{f,\pi}$ is essentially smooth.

(ii) Let f be a bounded continuous function on Ω.

$$\inf_{\nu:\langle\nu,f\rangle=k} K(\nu, \pi) = (\lambda^{(2)}_{f,\pi})^*(k) \tag{39}$$

(iii) Assume the marginal distribution of X has compact support. For service rate $c > 0$ and quality $\varepsilon > 0$, let $\widehat{N}_n$ be the number of connections admitted in the Direct CGF CAC on the basis of sampling n consecutive values of X. Then $n^{-1}\widehat{N}_n$ and $n^{-1}\widehat{S}_n$ satisfy an LDPs with scale n and respective rate functions

$$J^{(2)}(x) = \inf_{\theta}(\lambda^{(2)}_{g_\theta,\pi})^*(e^{(c\theta-\varepsilon)/y}) \quad \text{and} \quad I^{(2)}(x) = \inf_{y} J^{(2)}(y) + y\mu^*(x/y). \tag{40}$$

Example 3. Consider the effect of 1-step Markovian correlations on the direct CGF estimator for the bufferless model. We write the transition matrix on the state space $\{0,1\}$ for X_i as

$$\pi = \begin{pmatrix} 1-ab & (1-a)b \\ ab & 1-(1-a)b \end{pmatrix}, \tag{41}$$

with $a \in [0,1]$ and $b \in [0, 1/\max\{a, 1-a\}]$. In this parameterization, $\mathsf{E}[X_i] = a$, in agreement with the previous notation. b parameterizes the burstiness of X; successive samples are positively (or negatively) correlated when $b < 1$ (or $b > 1$) and X is Bernoulli when $b = 1$.

In Figure 4 we display the sampling rate function $J^{(2)}(x)$ above its zero at $x = m_\rho$ with burstiness parameter b from 0.5 to 1.75. Observe the rate function decreases to zero with b: the greater the correlations, the greater the probability of sampling error.

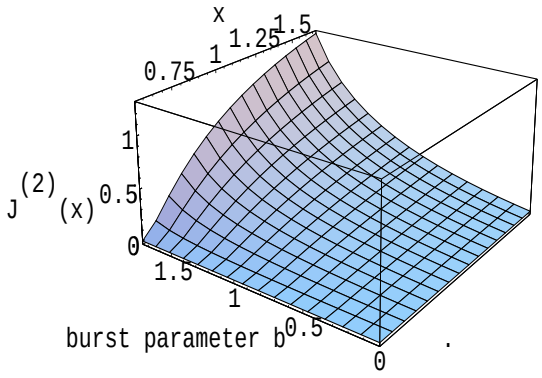

Figure 4: SAMPLING RATE FUNCTION WITH BURSTINESS: Impact of correlations in samples for admission to bufferless resource. As mean burst length grows to infinity ($b \to 0$) then sampling distribution widens about true value ($J^{(2)}(x) \to 0$).

3.3 Local Analysis

We can repeat the analysis of Section 2.6 using $\lambda^{(2)}_{g_\theta,\pi}$ in place of λ_θ. We state the following variant of Theorem 7 without proof. Here $J^{(2)}_\varepsilon, I^{(2)}_\varepsilon$ are the rate functions in the LDPs for $n^{-1}\widehat{N}_n$ and $n^{-1}\widehat{S}_n$ when the target quality is ε; $m^{(2)}_{\rho,\varepsilon}$ is the solution m to $J^{(2)}_\varepsilon(m) = 0$, and $H(y) = y\mu^*(c/y)$ as before. Again we insert a scale $\alpha > 0$, αn being the number of samples taken at service rate cn.

Theorem 9 *Assume H to be twice continuously differentiable on some interval $[m, c/a_\rho]$ and that each $J_\varepsilon^{(2)}$ is twice continuously differentiable on $[m_{\rho,\varepsilon}^{(2)}, c/a_\rho]$. Assume the limits*

$$v^{(2)} = \lim_{\theta\to 0} \theta^{-2}(\lambda^{(2)}_{g_\theta,\pi})''(0) \quad \text{and} \quad v = \lim_{\theta\to 0} \theta^{-2}\lambda''_\theta(0) \tag{42}$$

exist. Then

$$\lim_{\varepsilon\to 0} \varepsilon^{-1} I_\varepsilon^{(2)}(c) = \left(1 + \frac{v^{(2)}c}{va\alpha}\right)^{-1}. \tag{43}$$

Under appropriate regularity conditions, $\widehat{v}$ is the (rescaled) asymptotic variance of the partial sums $\lim_{n\to\infty} n^{-1}\operatorname{Var}(\sum_{i=1}^n X_i)$. Positive correlations amongst the X_i mean $v^{(2)} > v$; yielding lower quality than for the Bernoulli case.

Example 4. Recall the Markov process of Example 3. Routine calculations give

$$v^{(2)}/v = (b-2)/b \tag{44}$$

As b approaches the degenerate value 0 the mean length of bursts of X_i grows to infinity and the RHS of (43) approaches 0; in this limit, the variance of the samples is too large for them to provide any guarantee about the attained loss.

4 Robustness to Measurement Errors

In the previous sections we have quantified how measurement error increases the attained loss rate above the target loss rate under the assumption of Certainty Equivalence. In this section examine how to make admission control robust in the sense that the target loss rate is attained, at least in some asymptotic sense.

4.1 Robustness and Local Analysis

A simple approach to robustness is to use the results of Theorems 7 and 9 to calculate an $\varepsilon' > \varepsilon$ to be used in CAC in order that the attained loss rate is ε. In the case of Markovian samples, for example, this would entail using $\varepsilon' = \varepsilon(1 + \widehat{v}^{(2)}c/(\widehat{v}\widehat{a}\alpha))$, where notation such as $\widehat{v}^{(2)}$ denotes the estimator of $v^{(2)}$ obtained, e.g., by differentiation of the measured CGF.

4.2 Asymptotic Analysis of Robustness for Independent Samples

We now use the large deviation formulation of Section 2 to develop robust admission control outside the asymptotic region considered in Section 4.1. Let the true but unknown distribution of independent samples X_i be ρ^0. Consider nm sources admitted to capacity nc after constructing the empirical distribution $\widehat{\rho}_{\alpha_n n}$

from $\alpha_n n$ samples (with $\lim_{n\to\infty} \alpha_n = \alpha \in (0,\infty)$), where for the moment we chose m fixed, independent of $\widehat{\rho}_n$. Then the pair of independent random variables $(\widehat{\rho}_{\alpha_n n}, S_{nm}/n)$ satisfies a large deviation principle with scale n and rate function $(\rho, s) \mapsto \alpha D(\rho,\rho^0) + (m\mu_{\rho^0})^*(s)$. Suppose we want to make admission control robust w.r.t. unknown distributions $\rho^0 \in \mathcal{M}^0 \subset \mathcal{M}$, in the sense that a target loss probability of $\approx e^{-n\varepsilon}$ is attained independently of ρ^0. Then the LDP for $(\widehat{\rho}_{\alpha_n n}, S_{nm}/n)$ suggests we should admit $\widetilde{N}_n(\widehat{\rho}_n)$ connections where $\widetilde{N}_n(\rho) = \lfloor n\widetilde{m}_\rho \rfloor - 1$. and

$$\widetilde{m}_\rho = \inf\{m : \inf_{c'>c} \inf_{\rho^0 \in \mathcal{M}} \left(\alpha D(\rho,\rho^0) + (m\mu_{\rho^0})^*(c')\right) < \varepsilon\}. \tag{45}$$

This gives us the desired robustness, asymptotically as $n \to \infty$, at least when $\mathcal{M}^0 \subseteq \mathcal{M}_p$ for some $p > 0$.

Theorem 10 *Let $\mathcal{M}^0 \subseteq \mathcal{M}_p$ for some $p > 0$. Then for any true distribution $\rho^0 \in \mathcal{M}^0$ of the X_i,*

$$\limsup_{n\to\infty} n^{-1} \log \mathsf{P}[S_{\widetilde{N}_n(\widehat{\rho}_n)} \geq nc] \leq -\varepsilon. \tag{46}$$

It is not difficult to see that the asymptotic upper bound (46) also holds in the regime where one sample is taken per admitted source, provided we substitute α by m in the definition (45) of $\widetilde{m}_\rho$, the proportionate number of admitted sources; see the discussion at the end of Section 2.1.

We can identify the minimizing ρ^0 in (45) as follows. Using the contraction principle to re-express μ^* from Sanov's theorem, let us rewrite the rate function for the LDP for $(\widehat{\rho}_{\alpha_n n}, S_{nm}/n)$ as

$$(\rho, s) \mapsto \inf_{\nu : s = m\int \nu(dx)x} \inf_{\rho^0 \in \mathcal{M}^0} \left(\alpha D(\rho,\rho^0) + mD(\nu,\rho^0)\right). \tag{47}$$

The infimum over ρ^0 is achieved when $\rho^0 = (\alpha\rho + m\nu)/(\alpha + m)$. In the case that the samples are taken one per admitted, this reduces to $\rho^0 = (\rho+\nu)/2$ through the substitution of α by m in (45). Thus, admission control is done as if the true distribution ρ^0 were the mean of the measured distribution ρ and the distribution ν that saturates the capacity c. The same applies to any linear functional of the distributions, for example, their means. This generalizes an observation made for the two-level model in [21], where the same property was observed for the means of the corresponding distributions. For the two-level model, stating this property in terms of means or the distribution themselves is equivalent, since there is a affine bijection between mean and distribution in this case: the distribution $(1-p, p)$ on $\{0, 1\}$ has mean p.

It is worth noting that under stronger assumptions (the finiteness of Ω) the same target bound as (46) is obtained for predictive probabilities of loss using Bayesian inference. Briefly, this involves determining the posterior distribution π_n for the X_i from a given prior and the empirical measures $\widehat{\rho}_n$. One then admits $\widetilde{N}_n(\widehat{\rho}_n)$ sources, the role of $\mathcal{M}^0$ being played by the support of the prior distribution. The bound then follows by combining an LDP for the sequence $\{\pi_n\}$ of posterior distributions (recently established in [17]) with Varadhan's theorem.

4.3 Asymptotic Analysis of Robustness for Markov Samples

To provide robustness against measurement errors due to Markov sampling, it is necessary to incorporate the Markovian sampling properties into (45); this was the approach in Section 4.1. For samples $(X_i)_{i=1,\ldots,n}$ for the pairwise empirical measure $\widehat{\rho}_n^{(2)} \in \mathcal{M}^{(2)}$:

$$\widehat{\rho}_n^{(2)} = n^{-1} \sum_{i=1}^{n} \delta_{(X_i, X_{i+1})} \tag{48}$$

with indices taken modulo n, and where $\delta_{(x,y)}$ is the unit mass at $(x\,, y) \in \mathbb{R}_+^2$. Under the previous conditions (36) on the transition measure π, it is known [10] that $\widehat{\rho}_n^{(2)}$ satisfies an LDP with rate function

$$K^{(2)}(\omega) = K^{(2)}(\omega, \pi) = \begin{cases} D(\omega, p_1\omega \otimes \pi) & \text{if} \quad \omega(dx\,, dy) = \omega(dy, dx) \\ +\infty & \text{otherwise} \end{cases} . \tag{49}$$

Let $\mathcal{P}$ be a set of transition measures satisfying (36). Then (49) suggests that we employ the following adaptation of (45) in order to make admission control robust w.r.t. measurements from Markovian samples (X_i) with arbitrary transition measure $\pi \in \mathcal{P}$. Admit $\widetilde{N}_n^{(2)} = \lfloor n \widetilde{m}^{(2)}_{\widehat{\rho}_n^{(2)}} \rfloor$ connections, where for $\omega \in \mathcal{M}^{(2)}$

$$\widetilde{m}_\omega^{(2)} = \inf\{m : \inf_{\pi \in \mathcal{P}} \inf_{c' > c} \left(\alpha K^{(2)}(\omega, \pi) + (m\mu_{\rho_\pi})^*(c') \right) < \varepsilon\}, \tag{50}$$

where ρ_π is the stationary measure for π. One can prove and analog of Theorem 10, namely that when (X_i) has any transition measure in $\mathcal{P}$,

$$\limsup_{n \to \infty} n^{-1} \log \mathsf{P}[S_{\widetilde{N}_n^{(2)}} \geq nc] \leq -\varepsilon. \tag{51}$$

Example 5. We illustrate the impact of Markovian sample correlations on the admission controls discussed above. If we only use the Bernoulli-robust control (45) when the samples are actually Markovian with a transition measure π, the corresponding loss exponent is $\inf_\rho \left(K(\rho, \pi) + (\widetilde{m}_\rho \mu_{\rho_\pi})^*(c) \right)$ We illustrate the corresponding loss probabilities in a system with 100 sources and target overflow probability 10^{-4} in Figure 5. The transition matrix takes the form (41), with $a = .25$ (left figure), $a = .5$ (middle) and $a = .75$ (right), with b from $.1$ to 1 (the Bernoulli case). In each figure we show the log overflow probability for Certainty Equivalent, Bernoulli-robust and Markov-robust admission. As one might expect, except for Markov-robust admission, the attained loss rates increase as sampling correlation increases (i.e. as b decreases). However, there is little variation with the activity a of the sources. Below each graph is shown the admitted load (asymptotically as $n \to \infty$) in the three schemes, respectively $a m_{\rho_\pi}$, $a \widetilde{m}_{\rho_\pi}$ and $a \widetilde{m}^{(2)}_{\rho_\pi \otimes \pi}$ for Certainty Equivalent, Bernoulli-robust and Markov-robust admission. Observe as a increases, the relative difference between the admitted load of the three schemes narrows.

A final generalization. Suppose that instead of an being an artifact of sampling, the Markov property is shared by the admitted connections Y_i. Then for robust admission control, the second term in (50) should be replaced by the appropriate rate function governing the LDP for sums of Markov processes, namely

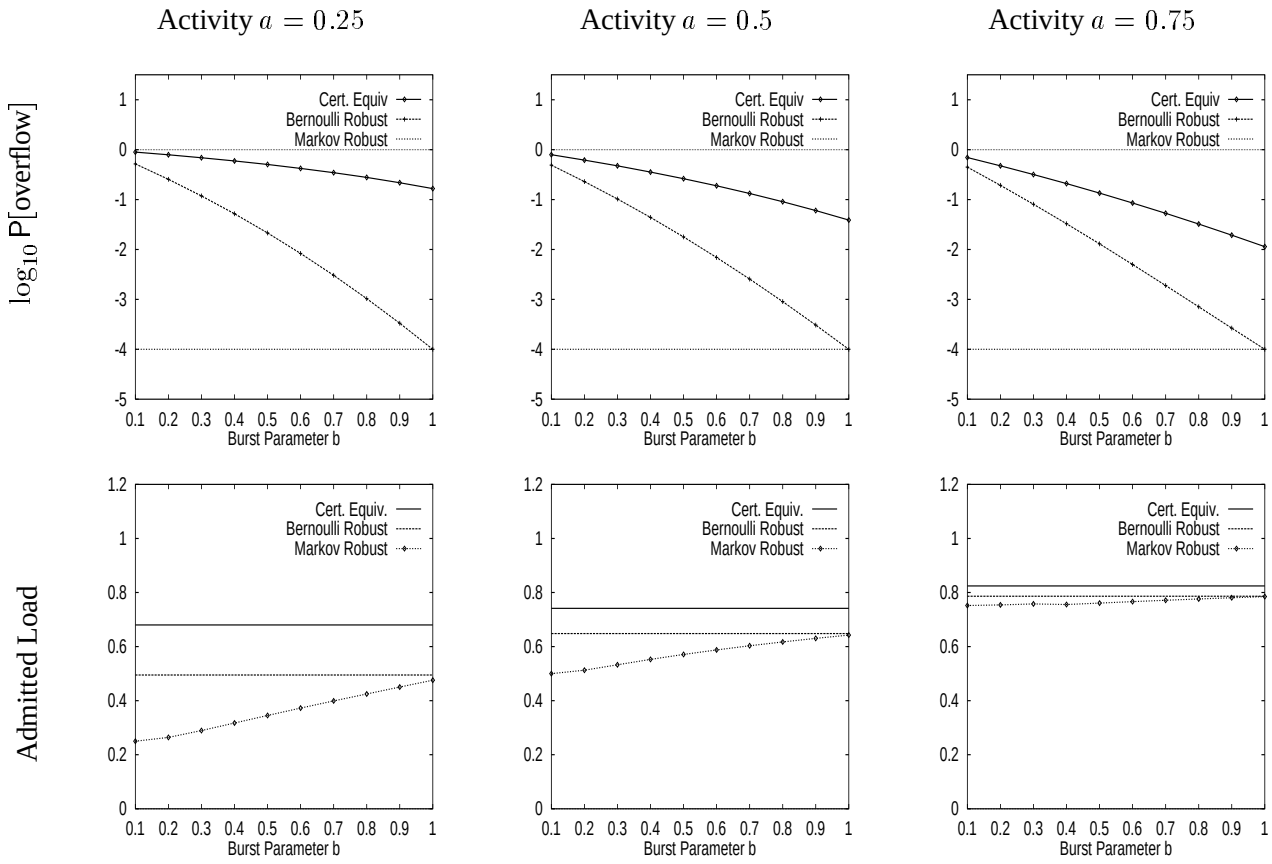

Figure 5: ROBUST ADMISSION AND MARKOVIAN SAMPLES: Certainty Equivalent, Bernoulli-robust and Markov-robust admission. (upper row) attained (logarithmic) overflow probability; (lower row) admitted load.

$(m\mu_\pi)^*(c)$ where $\mu_\pi = \lambda^{(2)}_{\mathrm{id},\pi}$. Observe that

$$\mu^*_\pi(c) = (\lambda^{(2)}_{\mathrm{id},\pi})^*(c) = \inf_{\nu:\int d\nu(x)x=c} K(\nu,\pi) = \inf_{\nu:\int d\nu(x)x=c} \inf_{\omega\in\mathcal{M}(\nu)} K^{(2)}(\omega,\pi). \tag{52}$$

Using arguments similar to those following (47) one can show that the extremizing π in (50) is given by

$$(\alpha+m)\pi(x\,,dy) = \alpha\frac{d\omega(x\,,y)}{dp_1\omega(x)} + m\frac{d\nu(x\,,y)}{dp_1\nu(x)} \tag{53}$$

where ν is the minimizer in (52). We see again that admission control is done as if the true transition measure were a mixture of the measured transition measure and that which causes capacity to be saturated.

5 Admission of Many Connections to a Buffered System

We now extend the work of the Section 2 to deal with buffered resources. Instead of single samples X_i, each connection presents values from a cumulative arrival processes $(X_i(t))_{t\in T}$ where $T = \mathbb{Z}_+$ or $\mathbb{R}_+$. $X_i(t)$ is the amount of work arriving in an interval of duration t. Similarly, the admitted connections will be described by processes $Y_i(\cdot)$. The $X_i(\cdot)$ and $Y_i(\cdot)$ are assumed to be independent copies of a process $X(\cdot)$ with stationary increments.

We assume that the trajectories $X(\cdot)$ lie in a complete separable metric space Ω equipped with a topology in which the evaluation maps $f_t : \Omega \to \mathbb{R}; X(\cdot) \mapsto X(t)$ are continuous. An example is motivated by the expectation that $X(\cdot)$ will be an increasing process with bounded derivative. Specifically, when $T = \mathbb{R}_+$ an example is when Ω is the subset of increasing homeomorphisms of $\mathbb{R}_+$ with $X(0) = 0$ and finite Lipschitz distance $\|X\| = \sup_{0\le x<y<\infty}(X(y) - X(x))/(y-x)$. It is simple to verify that Ω is a complete separable metric space when topologized with the norm $\|\cdot\|$. Another example for $T = \mathbb{N}$ is when we equip $\Omega = \mathbb{R}^{\times\mathbb{N}}$ with the product topology.

In all cases, the trajectories $X(\cdot)$ are assumed distributed according to some measure Ξ on Ω. Each $X(t)$ is distributed according to the image measure $\rho_t = \Xi f_t^{-1}$. $\boldsymbol{\rho} = (\rho_t)_{t\in\mathbb{R}}$ will denote the measure-valued path of the one-dimensional distributions of X (to be distinguished from the measure Ξ on the space of paths Ω).

Define $S_N(t) = \sum_{i=1}^N Y_i(t)$, $Q_{N,n} = \sup_{t\ge 0}(S_N(t) - nct)$. $Q_{N,n}$ has the same distribution as the virtual queue length due to N arrival streams served at total rate nc. Under very general conditions on the process $X(\cdot)$ (see [14]) the probability of the queue exceeding a buffer level nb has the asymptotic form

$$\lim_{n\to\infty} n^{-1}\log \mathsf{P}[Q_{nm,n} \ge nb] = -mL_{\boldsymbol{\rho}}(c/m, b/m), \tag{54}$$

where

$$L_{\boldsymbol{\rho}}(c,b) = \inf_{t>0}\mu^*_{\rho_t}(b+ct). \tag{55}$$

Assume a target loss ratio $e^{-n\varepsilon}$ with service rate nc and buffer size nb. Given an estimate $\widehat{\boldsymbol{\rho}}$ of the true distributions $\boldsymbol{\rho}$ a CAC rule associates the family of CGF estimates as $(\widehat{\mu}_{\widehat{\rho}_t})_{t\in\mathbb{R}}$. Based on these, the Certainty

Equivalent number of connections to be admitted would be

$$\widehat{N}_n = \inf_{t>0} \widehat{N}_{n,t} \quad \text{where} \quad \widehat{N}_{n,t} = \inf\{N : N\widehat{\mu}^*_{\widehat{\rho}_t}(nc/N, bn/N) \leq n\varepsilon\} - 1. \tag{56}$$

In practice we expect that CAC to be based on observations a set of n arrival processes $X^{(n)}(t)$ with t being in some finite subset $T_{\rm obs}$ of T. In this case we would restrict the infimum in (56) to the times of observations, i.e.

$$\widehat{N}_n(T_{\rm obs}) = \inf_{t \in T_{\rm obs}} \widehat{N}_{n,t} \tag{57}$$

Define $m_{\nu,t} = \inf\{x : (x\widehat{\mu}_\nu)^*(b + ct) \leq \varepsilon\}$.

Theorem 11 *Assume $X(\cdot)$ is such that (54) holds; that $\rho \mapsto m_{\rho,t}$ is weakly continuous for each t; and let $T_{\rm obs}$ be a finite set of observation times.*

(i) $n^{-1}\widehat{N}_n(T_{\rm obs})$ satisfies an LDP with scale n and rate function

$$J_\Xi(x) \quad = \quad \inf_{\Upsilon \in \mathcal{M}(\Omega, T_{\rm obs}, x)} D(\Upsilon, \Xi) \tag{58}$$

where

$$\mathcal{M}(\Omega, T_{\rm obs}, x) \quad = \quad \{\Upsilon \in \mathcal{M}(\Omega) : x = \inf_{t \in T_{\rm obs}} m_{\Upsilon f_t^{-1}, t}\} \tag{59}$$

(ii) Together with the assumption of (i) assume also that $\lim_{x \searrow 0} J_\Xi(x) = \infty$. Then

$$\lim_{n\to\infty} n^{-1} \log \mathsf{P}[Q_{\widehat{N}_n, n} \geq nb] = -\inf_{m>0} \left(J_\Xi(m) + mL(c/m, b/m)\right). \tag{60}$$

Application: The conclusions of Theorem 11 hold when $\widehat{\mu}$ is specified by the direct CGF rule when $X(\cdot)$ has bounded (and stationary) increments: i.e. $p_{\rho t} = p_{\rho_1} t$. To see this we only have to verify the assumption that $J_\Xi(0^+) = \infty$. To see this we can easily adapt the arguments from Theorem 4 to conclude that $m_{\rho_t, t} \geq (b + ct)/(p_{\rho_1} t) > c/p_{\rho_1}$, and so $J_\Xi(x) = \infty$ for $x < c/p_{\rho_1}$.

6 Admission Control in the Large Buffer Asymptotic

6.1 LDP for Effective Bandwidths

An axiomatic scheme for establishing large deviation queue tail asymptotics for short-range dependent traffic has been established by a number of authors; see [4, 22, 31]. The central object for characterizing the cumulative arrival process $X(\cdot)$ of a connection is the effective bandwidth. Following [30] we define this first as a function c on $\mathbb{R}_+ \times \mathbb{R}$ by

$$c(t, \theta) = (\theta t)^{-1} \log \mathsf{E}[e^{\theta X(t)}] \tag{61}$$

The theory relates the asymptotic properties of $c(t,\theta)$ for large t to those of the tail of the distribution of the virtual length

$$Q(c) = \sup_{t>0}\{X(t) - ct\} \tag{62}$$

of queue in an infinite buffer fed by X and served at rate c. We summarize the relevant results of [4, 22, 31] here. Set $\xi(t,\theta) = t^{-1}\log \mathsf{E}[e^{\theta X(t)}]$ (= $\theta c(t,\theta)$ when $\theta \neq 0$).

Proposition 3 *Let $c > 0$.*

(i) Assume the limit $\xi(\theta) = \lim_{t\to\infty}\xi(t,\theta)$ exists in a neighborhood of some $\Delta(c)$ such that $\xi(\Delta(c)) = c$.

(ii) Assume $\xi(\theta)$ is differentiable at $\Delta(c)$ with $\xi'(\Delta(c)) \geq 0$.

(iii) Assume $\xi(t,\Delta(c))$ finite for all $t > 0$.

Then

$$\lim_{b\to\infty} b^{-1}\log \mathsf{P}[Q(c) > b] = -\Delta(c). \tag{63}$$

$c(\delta) = \delta^{-1}\xi(\delta)$ is the bandwidth to be allocated the connection in order to achieve and asymptotic decay rate of δ for the log tail probability. We remark the well-known fact that approximations based on (63) can be quite bad, even to leading exponential order, when sources are highly aggregated; see e.g. [6]. Indeed, this was the motivation for using (54) for queues of aggregate traffic.

For MBAC the effective bandwidth is to be *measured*. We expect that, in practice, the estimate of $c(\delta)$ is not constructed as a limit, but rather through a direct estimate of the CGF of $X(t)$ for some large t. The sample duration t should be chosen greater than the correlation time of the increment process of X in order to avoid contaminating estimates of the effective bandwidth by finite size effects. (However, we shall see that making t large can also adversely effect sampling properties). This motivates finding the LDP for estimates of the general effective bandwidth function $c(t,\theta)$. Let $(X_i(t))_{i\in\mathbb{N}}$ be a sequence of samples of $X(t)$. An estimate of $c(t,\theta)$ based on the first n samples is

$$\widehat{c}_n(t,\theta) = (t\theta)^{-1}\log\frac{1}{n}\sum_{i=1}^{n} e^{\theta X_i(t)} \tag{64}$$

We can rewrite (64) in terms of the empirical measures $\widehat{\rho}_{n,t} = n^{-1}\sum_{i=1}^{n}\delta_{X_i(t)}$ as $\widehat{c}_n(t,\theta) = c_{\widehat{\rho}_{n,t},t}(\theta)$ where $c_{\nu,t}(\theta) := (t\theta)^{-1}\log\langle\nu, g_\theta\rangle$. When the samples $X_i(t)$ are i.i.d. then we will obtain an LDP for $\widehat{c}_n(t,\theta)$ with scale n by combining the contraction principle with a LDP for $\widehat{\rho}_{n,t}$ due to Sanov's theorem.

To examine the effect of measurement errors on the attained queue tail, we consider a queue whose input is an independent copy Y of the process X, served at the constant rate $\widehat{c}_n(t,\delta)$, and scale the queue level of interest in proportion to the number of samples of $X(t)$.

Theorem 12 *(i) Assume $p_{\rho t} < \infty$. For each $\theta > 0$, $\widehat{c}_n(t,\theta)$ satisfies an LDP with scale n and rate function*

$$J(c,t,\theta) = \inf_{\nu;\, \langle \nu, g_\theta \rangle = e^{c\theta t}} D(\nu, \rho_t) = \lambda^*_{t,\theta}(e^{c\theta t}), \tag{65}$$

where $\lambda_{t,\theta}(\phi) = \log \mathsf{E}[\exp(\phi e^{\theta X(t)})]$.

(ii) Assume the LDP from (i), the assumptions of Proposition 3, and that the convergence in (63) is uniform in c. Then

$$\lim_{n\to\infty} \frac{1}{n} \log \mathsf{P}[Q(\widehat{c}_n(t,\delta)) > nb] = -I(b,t,\delta) := -\inf_c \left(J(c,t,\delta) + b\Delta(c)\right) \tag{66}$$

Example 6: Bernoulli Connections The queueing properties of Bernoulli processes form an uninteresting example for effective bandwidth theory; the effective bandwidth is independent of the sample duration. We use them here rather to investigate the sampling properties of the effective bandwidth and their dependent on the sample duration. With the notation of Example 1 we find that

$$\lambda_{t,\theta}(\phi) = \sum_{s=0}^{t} \binom{t}{s} a^s (1-a)^{t-s} \exp(\phi e^{s\theta}). \tag{67}$$

In Figure 6(i) we plot the rate functions occurring in (66) for $t = 10$, $\delta = 1$ and $b = 1$. Observe that for these parameters, the infimum in (66) is obtained at $c = a = 0.5$. The interpretation is that violation of the actual quality $\varepsilon' = J(a, 10, 1)$ is most likely to occur entirely through estimating the effective bandwidth by the mean a. But generally the minimum will not be located at $c = a$. For example, if we increase the proportionate number of sample taken by a factor $\alpha > 1$, as in Section 2.3, the distribution of the effective bandwidth narrows around $c(t,\delta)$. We illustrate by displaying the corresponding rate functions $\alpha J(c,t,\delta) + b\Delta(c)$ in Figure 6(ii). More generally, this effect is controlled by the ratio α/b.

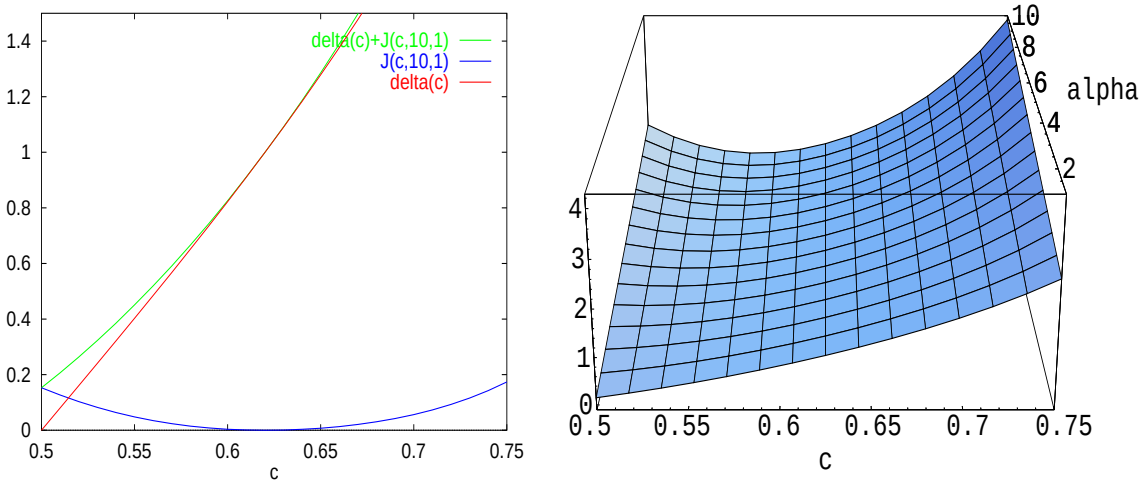

Figure 6: RATE FUNCTIONS FOR LARGE BUFFER ASYMPTOTIC: BERNOULLI ARRIVALS (i) Left: $t = 10$, minimum achieved at $c = a = 0.5$. (ii) Right: $\alpha J(c,t,1) + \Delta(c)$, where α scales the number of measurements. With $\alpha \geq 5$ the minimum detaches from $c = a$.

6.2 Sampling Properties and the Sample Duration

We now examine the dependence of the rate function $I(c,\delta,t)$ on the sample duration t. We mentioned before that it is desirable to have t at least as large as the correlation time of the increment process of X. However, the sampling properties of $\widehat{c}_n(t,\delta)$ can actually be quite bad for large t. One indication of this is the *non-commutativity* of the limits $n\to\infty$ and $t\to\infty$ for $\widehat{c}_n(t,\delta)$. From Theorem 12(ii) we know that for fixed t, $\widehat{c}_n(t,\delta)$ converges in probability to $c(t,\delta)$. However, for a wide class of processes X with stationary increments, for any finite n,

$$\lim_{t\to\infty}\widehat{c}_n(t,\delta)=a=\mathsf{E}[X(1)],\quad \text{a.e.} \tag{68}$$

Consider the zero mean process $Y(t)=X(t)-at$. Then (68) occurs whenever for any $\varepsilon\neq 0$, $Y_t+\varepsilon t$ converges to $\infty\cdot\mathrm{sgn}(\varepsilon)$ almost surely. For $T=\mathbb{N}$ this happens when X is a random walk with jumps which satisfy certain mixing conditions; see [36]. What is happening is that $X(t)$ self-averages as $t\to\infty$, so that both $X(t)/t$ and $\widehat{c}_n(t,\delta)$, for fixed n, converge to the mean a as $t\to\infty$.

This non-commutativity is also manifest through the behavior, at least in examples, of the sampling rate function $J(x,t,\delta)$ for large t. We illustrate this in Figure 7 for the Bernoulli arrivals of Example 1; we set $\delta=1$ and vary x and t. For each finite t we know that $x\mapsto J(x,t,\delta)$ takes its minimum 0 at $c(\delta)$. As t increases to $\tau=5$, $x\mapsto J(x,t,\delta)$ initially becomes steeper, but for $t>\tau$ flattens down towards 0; this makes the estimation of the bandwidth by the mean a progressively more likely, as we might expect from (68). This indicates that when the sample duration t can be chosen independently of n, there is some value of t ($\tau=5$ in the example) at which the sampling proper ties of $\widehat{c}(\tau,\delta)$ are optimal in the sense that $J(c,t,\delta)$ is steepest about $c=c(\delta)$ for $t=\tau$: the effective bandwidth estimates are most likely to be near $c(\delta)$.

When sampling a single connection sequentially, we may expect that increasing t will decrease the number of samples proportionately. The sampling rate function to be considered now is $t^{-1}J(x,t,\delta)$. We display this for the Bernoulli example in Figure 8. In this case, $t^{-1}J(x,t,\delta)$ is decreasing in t; the sampling properties are best for small t. But if we move beyond Bernoulli arrivals to the correlated one, we will have to balance the increased accuracy due to smaller (and hence more numerous) samples, against the bias they introduce in the estimation of the effective bandwidth; see e.g. the discussion in [13].

6.3 Local Analysis of the Sampling Rate Function

Some more analysis can be done to substantiate the above observations in a more general setting. We expand $J(\cdot,t,\delta)$ to leading order about its zero at $c(\delta)$. By applying arguments similar to those in the proof of Theorem 7 with Theorem 12(i) we have

$$J''(c(t,\delta),t,\delta) = \frac{d^2}{dc^2}\lambda^*_{t,\delta}(e^{c\delta t})\Big|_{c=c(t,\delta)}=\lambda^{*,\prime\prime}_{t,\delta}(e^{c(\delta)\delta t})(e^{c(t,\delta)\delta t}\delta t)^2 \tag{69}$$

$$= \frac{(e^{c(t,\delta)\delta t}\delta t)^2}{\lambda''_{t,\delta}(0)}=\frac{(\delta t)^2}{\exp(t\delta c(t,2\delta)-2t\delta c(t,\delta))-1} \tag{70}$$

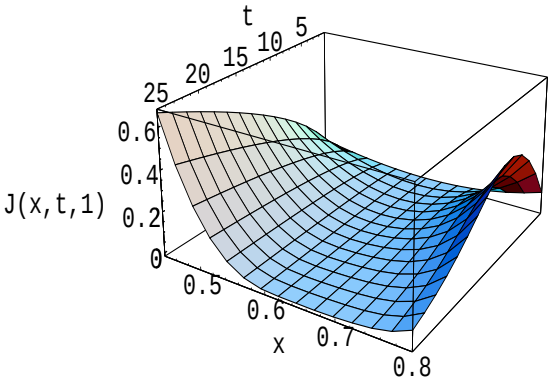

Figure 7: SAMPLING RATE FUNCTION FOR EFFECTIVE BANDWIDTH: TIME DEPENDENCE. Bernoulli Arrivals. Plot shows $J(x,t,1)$. Observe flattening of rate function about its minimum at $c(\delta)$ as t increases beyond ≈ 10.

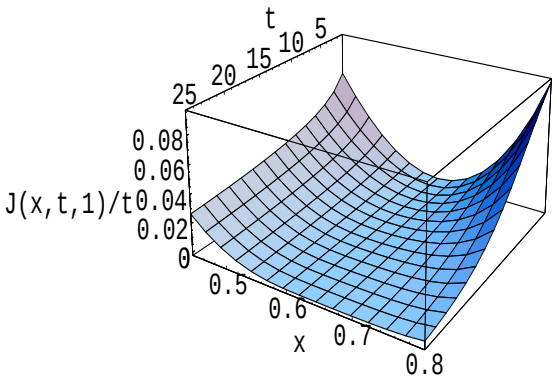

Figure 8: SAMPLING RATE FUNCTIONS FOR EFFECTIVE BANDWIDTH: Bernoulli Arrivals under time constraint. Plot shows $t^{-1}J(x,t,1)$. Sampling rate function is decreasing in t.

Assume δ small and t larger than the correlation time of X. Then $t\delta c(t,2\delta) - 2t\delta c(t,\delta) \approx vt\delta$, where we assume the existence of $v = \lim_{t\to\infty} t^{-1}\mathrm{Var}X(t)$. For the specified ranges of t and δ we have the approximation

$$J''(c,t,\delta) \approx \frac{(\delta t)^2}{e^{vt\delta}-1}. \tag{71}$$

Consider the quadratic approximation for $J(c,t,\delta)$ about $c = c(\delta)$. When the choice of t is unconstrained, we see that the sampling rate function $J(c,t,\delta)$ will be maximized for c in a neighborhood on $c(\delta)$ by taking $t = z/(v\delta)$ where z is the unique maximizer in $(0,\infty)$ of the function $z^2/(e^z-1)$.

When the choice of t is constrained, then the relevant sampling rate function is $t^{-1}J(c,t,\delta)$ with second derivative $\approx \delta^2 t/(e^{vt\delta}-1)$ at $c(\delta)$ for δ small and t greater than the correlation time of X. In distinction with the unconstrained case, this is decreasing in t.

7 Discussion and Further Work

In this paper we provided a large deviation framework in which to describe errors in measurement-based admission control. We have extensively investigated these for many sources seeking admission at a bufferless resource; determined the effect of sample correlations, and showing how to make admission control robust against a class of measurement errors. We have also used the framework to describe measurement errors in buffered resources, both in the many-source and large-buffer asymptotic.

The logarithmic asymptotics used in this paper can be improved upon. Multiplicative corrections to e.g. (15) are provided by Bahadur-Rao (see e.g. [10]); this approach can also be used for buffered systems; see recent work in [34, 37]. It should be possible to determine the finer asymptotics of attained loss, and also to take account of such corrections in the admission controls themselves.

We intend to publish elsewhere results on the qualitative features of attained loss for admission to buffered systems. One matter of interest here is the relation between the so-called critical timescale for loss (the optimizing t in (55), the corresponding measured timescale (the optimizing t in (56)), and any optimal sampling timescale of the type discussed in Section 6. The discussion there indicates that the sampling rate function may become quite flat–leading to higher likelihood of measurement error–if the critical queueing timescale is large which the number of samples remains fixed.

Acknowledgment

David Tse contributed to the formulation of the problem considered during many useful conversations. Thanks are due to the referees for their constructive suggestions.

8 Definitions and Proofs

Large Deviations Terminology. We collect together (from [10]) the terms and tools from Large Deviation theory which we will use. A *rate function* I on a space S is a lower semicontinuous function $S \to [0, \infty]$ with closed level sets. I is *good* if its level sets are compact also. A family of random variables $(X_n)_{n\in\mathbb{N}}$ satisfies a *Large Deviation Principle* (LDP) with *scale* n and rate function I if for each subset $B \subset S$:

$$-\inf_{x\in B^0} I(x) \le \liminf_{n\to\infty} \frac{1}{n}\log \mathsf{P}[X_n \in B] \le \limsup_{n\to\infty} \frac{1}{n}\log \mathsf{P}[X_n \in B] \le -\inf_{x\in \bar{B}} I(x), \tag{72}$$

where B^o is the interior and $\bar{B}$ the closure of B. Define $\mu_n(\theta) = n^{-1}\log \mathsf{E}[e^{\theta X_n}]$. The Gärtner-Ellis theorem says sufficient conditions for $(X_n)_{n\in\mathbb{N}}$ to satisfy an LDP with scale n and good rate function I are that the limit $\mu(\theta) = \lim_{n\to\infty} \mu_n(\theta)$ exists as an extended real function, and that it is essentially smooth. In this case the rate function is $I = \mu^*$ where $\mu^*(x) = \sup_{\theta\in\mathbb{R}}\{x\theta - \mu(\theta)\}$ is the *Legendre Transform* of μ. The *Contraction Principle* says that if this LDP holds and $f : S \to T$ is a continuous function with S and T Hausdorff, then $(f(X_n))_{n\in\mathbb{N}}$ satisfies an LDP with scale n and rate function $J(y) = \inf_{y:f(x)=y} I(x)$.

Proof of Theorem 1: First set $\widehat{m}_n = m_{\widehat{\alpha}_n \rho_n}$. It follows from the Contraction Principle and Sanov's Theorem that $\widehat{m}_n$ satisfies an LDP with scale n and good rate function αJ_ρ. By the uniform convergence of the C_n^{-1}, the sequences $n^{-1}\widehat{N}_n$ and $\widehat{m}_n$ are exponentially equivalent. Thus the same LDP holds for $n^{-1}\widehat{N}_n$; see Theorem 4.2.16 in [10]. ■

Proof of Theorem 3: We perform the proof for $\alpha = 1$. When ρ has finite support, then the proof of the Theorem follows quickly by combining Example 6.6.24 in [10] with the well known Contraction Principle resulting from expressing the empirical mean as a function of the empirical measure; see Section 2.1.2 in [10]. More generally, the empirical mean is not a weakly continuous function of the empirical measure, so the same argument cannot be used directly. (Another possible approach, indicated in Remark (c) following Theorem 4.2.1 of [10] seems hard to apply in this case). More generally, we use the following argument.

For clarity we denote μ_ρ by μ. For any Borel subset B of $\mathbb{R}_+$,

$$\mathsf{P}[\widehat{S}_n \in nB] = \sum_{M \in \mathbb{Z}_+} \mathsf{P}[\widehat{N}_n = k]\mathsf{P}[N_k \in nB] = \int dQ_n(m) e^{nf_n(m,B)}, \tag{73}$$

where

$$Q_n[B] = \sum_{k \in nB} \mathsf{P}[\widehat{N}_n = k] \quad \text{and} \quad f_n(m,B) = n^{-1} \log \mathsf{P}[S_{\lfloor nm \rfloor} \in nB]. \tag{74}$$

From Cramer's theorem on the LDP for sums of i.i.d. random variables we expect $f_n(m,B) \approx -\inf_{x \in B} m\mu^*(x/m)$ for large n. Coupling this with the LDP for $\widehat{N}_n$, then the stated result follows, formally at least, by applying Varadhan's theorem; see Section 4.3 of [10].

More precisely, for any $\widetilde{m} > 0$ observe

$$\int_{m \geq \widetilde{m}} dQ_n(m) e^{nf_n(m,B)} \leq \int dQ_n(m) e^{nf_n(m,B)} \leq \int_{m \geq \widetilde{m}} dQ_n(m) e^{nf_n(m,B)} + \mathsf{P}[\widehat{N}_n < n\widetilde{m}]. \tag{75}$$

Set $f^{(\varepsilon)}(m,B) = -\inf_{x \in B} m(\mu^*(x/m) + \varepsilon)$ and $f = f^{(0)}$. By Cramer's theorem, for each $m > 0$,

$$f(m, B^\circ) \leq \liminf_{n \to \infty} f_n(m,B) \leq \limsup_{n \to \infty} f_n(m,B) \leq f(m, \bar{B}). \tag{76}$$

The inferior and superior limits of $m^{-1} f_n(m,B)$ as $n \to \infty$ exist uniformly for m bounded away from 0. Hence, for all $\varepsilon > 0$ we can find n_ε such that for all $n > n_\varepsilon$,

$$f^{(-\varepsilon)}(m, B^\circ) \leq f_n(m,B) \leq f^{(\varepsilon)}(m, \bar{B}) \tag{77}$$

for all $m > \widetilde{m}$. Furthermore, $f^{(\varepsilon)}(\cdot, B)$ is continuous, since μ^*, being convex, is continuous in the interior of its effective domain; see Theorem 10.1 of [41]. Using Varadhan's Theorem we conclude that

$$\begin{aligned} -\inf_{m \geq \widetilde{m}} (J_\rho(m) - f^{(-\varepsilon)}(m, B^\circ)) &\leq \liminf_{n \to \infty} n^{-1} \log \int_{m \geq \widetilde{m}} dQ_n(m) e^{nf_n(m,B)} && (78) \\ &\leq \limsup_{n \to \infty} n^{-1} \log \int_{m \geq \widetilde{m}} dQ_n(m) e^{nf_n(m,B)} && (79) \\ &\leq -\inf_{m \geq \widetilde{m}} (J_\rho(m) - f^{(\varepsilon)}(m, \bar{B})). && (80) \end{aligned}$$

The result then follows by taking $\varepsilon \to 0$. In full detail: $\lim_{\varepsilon \to 0} -f^{(\varepsilon)}(m,B) = \lim_{\varepsilon \to 0} \inf_{x \in B} m(\mu^*(x/m) + \varepsilon) = \inf_{\varepsilon > 0} \inf_{x \in B} m(\mu^*(x/m) + \varepsilon) = \inf_{x \in B} \inf_{\varepsilon > 0} m(\mu^*(x/m) + \varepsilon) = \inf_{x \in B} m\mu^*(x/m) = -f(m,B)$, while $\lim_{\varepsilon \to 0} -f^{(-\varepsilon)}(m,B) = \sup_{\varepsilon > 0} \inf_{x \in B} m(\mu^*(x/m) - \varepsilon) \leq \inf_{x \in B} \sup_{\varepsilon > 0} m(\mu^*(x/m) - \varepsilon) = \inf_{x \in B} m\mu^*(x/m) = -f(m,B)$. The limits as $\varepsilon \to 0$ can be pulled though the infima over m in (78) and (80) in a similar way.

So far $\widetilde{m} > 0$ in (75) was arbitrary. Now observe that from the assumed LDP for $\widehat{N}_n$,

$$-\inf_{m < \widetilde{m}} J_\rho(m) \leq \liminf_{n \to \infty} n^{-1} \log \mathsf{P}[\widehat{N}_n < n\widetilde{m}] \leq \limsup_{n \to \infty} n^{-1} \log \mathsf{P}[\widehat{N}_n < n\widetilde{m}] \leq -\inf_{m \leq \widetilde{m}} J_\rho(m) \tag{81}$$

With the assumption that $J_\rho(0^+) = \infty$ we then obtain (18) by taking $\widetilde{m} \to 0$.

The restriction of the supremum in (19) arise as since J_ρ takes its minimum value, 0, at m_ρ and is non-increasing on $[0, m_\rho)$ and non decreasing on (m_ρ, ∞), while $y \mapsto y\mu_\rho^*(x/y) = (y\mu_\rho)^*(x)$ is convex and takes its minimum value 0 at $y = x/a_\rho \geq m_\rho$. ■

Proof of Theorem 4: (i) *Strict decreasing property.* Since X is bounded, μ_ρ is differentiable. The supremum in the Legendre transform $(m\mu_\rho)^*(c) = \sup_\theta(c\theta - m\mu_\rho(\theta))$ can be restricted to $\theta > 0$ since it occurs when $\mu_\rho'(\theta) = c/m > a_\rho = \mu_\rho'(0)$, and μ_ρ' is increasing since (by Hölder's Inequality) μ_ρ is convex. Furthermore, since X is non-negative and not identically zero, so is $\mu_\rho(\theta)$ for $\theta > 0$. Hence $(m\mu_\rho)^*(c)$ is decreasing in m for $m < c/a_\rho$.

Convexity and Continuity. Clearly $\limsup_{\theta\to\infty} \mu_\rho'(\theta) = p_\rho$, and so $(m\mu_\rho)^*(c)$ is finite for $m \in (c/p_\rho, c/a_\rho)$. It is also convex in m since for $q \in [0, 1]$,

$$\begin{aligned}((qm_1 + (1-q)m_2)\mu_\rho)^*(c) &= \sup_\theta \left(q(c\theta - m_1\mu_\rho(\theta)) + (1-q)(c\theta - m_2\mu_\rho(\theta))\right) &(82)\\ &\leq q(m_1\mu_\rho)^*(c) + (1-q)(m_2\mu_\rho)^*(c). &(83)\end{aligned}$$

Thus, by Theorem 10.1 of [41], $(m\mu_\rho)^*(c)$ is continuous in m on $(c/p_\rho, c/a_\rho)$.

Range. We establish the range by continuity. First the lower boundary at $m = c/p_\rho$. As a Legendre transform, μ_ρ^* is lower semicontinuous, and hence so is $m \mapsto (m\mu_\rho)^*(c) = m\mu_\rho^*(c/m)$. Thus, $((c/p_\rho)\mu_\rho)^*(c) \geq \lim_{m\searrow c/p_\rho}(m\mu_\rho)^*(c) = \liminf_{m\searrow c/p_\rho}(m\mu_\rho)^*(c) \geq ((c/p_\rho)\mu_\rho)^*(c)$, where the first inequality and the equality follow from the decreasing property, and the second inequality by lower semicontinuity. Finally, $((c/a_\rho)\mu_\rho)^*(c) = 0$ and the continuity at the upper boundary at $m = c/a_\rho$ follows from the non-degeneracy of ρ.

(ii) Suppose first that ρ is an atom, i.e., it has support at a single point, so that $a_\rho = p_\rho$. Then $\mu^*(x) = 0$ if $x = p_\rho$ and ∞ if $x \neq p_\rho$. Thus $m_\rho = c/p_\rho$ for all ε. Otherwise, if ρ is not an atom, the existence and uniqueness of the solution follow from the properties established in (i). If $((c/p_\rho)\mu_\rho)^*(c) = \infty$ then this solution exists for all $\varepsilon > 0$, as it does when $\varepsilon < ((c/p_\rho)\mu_\rho)^*(c) < \infty$. If $(c/p_\rho)\mu_\rho)^*(c) < \infty$, then $(m\mu_\rho)^*(c) = \infty$ for all $m < c/p_\rho$. Hence, if $\varepsilon > ((c/p_\rho)\mu_\rho)^*(c)$, then $m_\rho = c/p_\rho$.

(iii) Suppose first that ρ is an atom. If ρ_n converges weakly to ρ in some $\mathcal{M}_p$, then $a_{\rho_n} = \langle\rho_n, h\rangle$ converges to $a_\rho = \langle\rho_n, h\rangle$ since $h(x) = x$ is a bounded continuous function. Since $m_{\rho_n} \leq c/a_{\rho_n}$ then $\limsup_{n\to\infty} m_{\rho_n} \leq \lim_{n\to\infty} c/a_{\rho_n} = c/a_\rho = c/p_\rho = m_\rho$. It remains to prove the complementary lower bound. For any $k > a_\rho$, $\lim_{n\to\infty}\rho_n[k, \infty) \to 0$ and hence $\limsup_{n\to\infty}\mu_{\rho_n}(\theta) \leq k\theta$ for $\theta > 0$, thus $\liminf_{n\to\infty}\mu_{\rho_n}^*(x) = \infty$ for $x > a_\rho$ and so finally $\liminf_{n\to\infty} m_{\rho_n} \geq c/a_\rho = m_\rho$.

In the remainder of the proof we suppose that ρ is not an atom (this means that a_ρ and p_ρ are distinct). For each θ, $e^{\theta x}$ is a bounded continuous function of x in $[0, p]$, and hence $\rho \mapsto \mu_\rho(\theta)$ is weakly continuous on $\mathcal{M}_p$. So by Lemma 1 of [14] we have the following continuity property: if $\rho_n \to \rho$ weakly as $n \to \infty$, then $\mu_{\rho_n}^*(x) \to \mu_\rho^*(x)$ for all x in the interior of the effective domain of μ_ρ^*, including (a_ρ, p_ρ). We use this to show that the map $\rho \mapsto m_\rho$ is weakly sequentially continuous; since $\mathcal{M}$ is metrizable in its weak topology,

the maps is then also weakly continuous; see Section 12 of [5].

First we show that m_{ρ_n} is the unique solution of $(m_{\rho_n}\mu_{\rho_n})^*(c) = \varepsilon$ for n sufficiently large. Let $\rho_n \to \rho$ weakly as $n \to \infty$. Suppose first that $\varepsilon < ((c/p_\rho)\mu_\rho)^*(c)$. By (i), there exists for any sufficiently small $\varepsilon' > 0$ a $\bar{m} \in (c/p_\rho, m_\rho) \subset (c/p_\rho, c/a_\rho)$ such that $\varepsilon + \varepsilon' < (\bar{m}\mu_\rho)^*(c) < \infty$. So by the weak continuity described in the preceding paragraph,

$$\varepsilon < (\bar{m}\mu_{\rho_n})^*(c) < \infty \tag{84}$$

for all n sufficiently large. Clearly $\liminf_n p_{\rho_n} \geq p_\rho$ and so $c/p_{\rho_n} < \bar{m}$ for all n sufficiently large. Combining this with (84) and the decreasing property (i), we have $\varepsilon < ((c/p_{\rho_n})\mu_{\rho_n})^*(c)$, and hence by (ii), m_{ρ_n} is the unique solution of $(m_{\rho_n}\mu_{\rho_n})^*(c) = \varepsilon$.

Suppose now that $m_{\rho_n} \not\to m_\rho$. Then for $\delta > 0$ sufficiently small, and along some subsequence, either (a) $m_{\rho_n} > m_\rho + \delta$, or (b) $m_{\rho_n} < m_\rho - \delta$. Suppose (a). Taking the limit along the subsequence, and using the decreasing property (i),

$$\varepsilon = (m_{\rho_n}\mu_{\rho_n})^*(c) < ((m_\rho + \delta)\mu_{\rho_n})^*(c) \stackrel{n\to\infty}{\longrightarrow} ((m_\rho + \delta)\mu_\rho)^*(c) < (m_\rho\mu_\rho)^*(c) = \varepsilon, \tag{85}$$

a contradiction. Case (b) yields a similar contradiction.

We now treat the remaining case that $((c/p_\rho)\mu_\rho)^*(c) \leq \varepsilon < \infty$; from (ii) this means $m_\rho = c/p_\rho$. First observe $\liminf_{n\to\infty} m_{\rho_n} \geq m_\rho = c/p_\rho$; for otherwise, there exists $\bar{m}$ such that $m_{\rho_n} < \bar{m} < m_\rho$ infinitely often. Passing to a subsequence using the decreasing property (i),

$$\liminf_{n\to\infty}(m_{\rho_n}\mu_{\rho_n})^*(c) \geq \liminf_{n\to\infty}(\bar{m}\mu_{\rho_n})^*(c) \geq (\bar{m}\limsup_{n\to\infty}\mu_{\rho_n})(c) = (\bar{m}\mu_\rho)(c) = 0. \tag{86}$$

The last equality is because $c/\bar{m} > p_\rho$, the previous one because $\mu_\rho(\theta)$ is weakly continuous in $\rho \in \mathcal{M}_p$ for each θ. This means $(m_{\rho_n}\mu_{\rho_n})^*(c) > \varepsilon$ for some n. But the second part of the definition (20) then requires $m_{\rho_n} = c/p_{\rho_n}$ and hence $(c/p_{\rho_n}\mu_{\rho_n})^*(c) > \varepsilon$ from which the first part requires $(m_{\rho_n}\mu_{\rho_n})^*(c) = \varepsilon$, a contradiction.

To complete the proof it remains to show that $\limsup_{n\to\infty} m_{\rho_n} \leq m_\rho$. If not, then $\lim_{n\to\infty} m_{\rho_n} > m_\rho = c/p_\rho$ along some subsequence. Together with $\liminf_{n\to\infty} p_{\rho_n} \geq p_\rho$, this implies for sufficiently large n in the subsequence that $m_{\rho_n} > c/p_{\rho_n}$. Hence (20) only allows $(m_{\rho_n}\mu_{\rho_n})^*(c) = \varepsilon$ for these n. The conclusion then follows by establishing a contradiction as in (85).

(iv) The conclusions follow as a corollary of (iii) once we observe that, by assumption, the support of Q_n in (74) is contained in $[c/p_\rho, \infty)$ with p_ρ finite, so $J_\rho(m) = \infty$ for $m < c/p_\rho$. ■

Proof of Theorem 5: (i) Observe that $\{m_\nu \geq x\} = \{(x\mu_\nu)^*(c) > \varepsilon\}$. Hence if $x \geq m_\rho$, then

$$J_\rho(x) = \inf_{y\geq x} J_\rho(y) = \inf_{\nu: m_\nu \geq x} D(\nu,\rho) \tag{87}$$

$$= \inf_{\nu:(x\mu_\nu)^*(c)>\varepsilon} D(\nu,\rho) \tag{88}$$

$$= \inf_{\theta} \inf_{\nu: c\theta - x\mu_\nu(\theta) > \varepsilon} D(\nu, \rho) \tag{89}$$

$$= \inf_{\varepsilon' > \varepsilon} \inf_{\theta} \inf_{\nu: c\theta - x\mu_\nu(\theta) = \varepsilon'} D(\nu, \rho) \tag{90}$$

$$= \inf_{\varepsilon' > \varepsilon} \inf_{\theta} \inf_{\nu: \langle \nu, g_\theta \rangle = e^{(c\theta - \varepsilon')/x}} D(\nu, \rho) \tag{91}$$

$$= \inf_{\varepsilon' > \varepsilon} \inf_{\theta} \lambda_\theta^*(e^{(c\theta - \varepsilon')/x}). \tag{92}$$

The last step follows by Proposition 1. The result then follows if we can show that λ_θ^* is non-increasing on $(0, e^{(c\theta - \varepsilon)/x})$. Now note that $\lambda_\theta'(0) = e^{\mu_\rho(\theta)}$, so that λ_θ^* takes its minimum value at $e^{\mu_\rho(\theta)}$. But now $x \geq m_\rho$ means $(x\mu_\rho)^*(c) \leq \varepsilon$ and hence $(c\theta - \varepsilon)/x \leq \mu_\rho(\theta)$ for all θ. Since λ_θ^* is convex and non-negative, λ_θ^* is non-increasing on $(0, e^{\mu_\rho(\theta)}) \supseteq (0, e^{(c\theta - \varepsilon)/x})$.

(ii) Since $e^{(c\theta - \varepsilon)/x} \leq e^{\mu_\rho(\theta)} = \lambda_\theta'(0)$ and λ_θ is convex, the supremum in $\lambda_\theta^*(e^{(c\theta - \varepsilon)/x})$ occurs at some $\phi_{e^{(c\theta - \varepsilon)/x}} \leq 0$.

(iii) $y \leq c/a_\rho$ can be rewritten $c \geq y\mu_\rho'(0)$, in which case the supremum over θ in $(y\mu_\rho)^*(c)$ can be restricted to $\theta \geq 0$. For the reversed in equality, the argument is similar. ■

Proof of Proposition 1: By the assumptions on λ_f, there exists a unique ϕ_k such that $k = \lambda_f'(\phi_k)$, and $\lambda_f^*(k) = k\phi_k - \lambda_f(\phi_k)$. Let ν_k denote the probability measure absolutely continuous with respect to ρ with Radon-Nikodym derivative $h(x) = \frac{d\nu_k}{d\rho}(x) = e^{\phi_k f(x) - \lambda_f(\phi_k)}$. We see that $\langle \nu_k, f \rangle = \langle \rho, f e^{\phi_k f} \rangle / \langle \rho, e^{\phi_k f} \rangle = \lambda_f'(\phi_k) = k$. From the definition (8) of D we find $D(\nu_k, \rho) = k\phi_k - \lambda_f(\phi_k) = \lambda_f^*(k)$.

The stated result follows if we can show that $D(\nu, \rho) \geq D(\nu_k, \rho)$ for any $\nu \in \mathcal{M}$ with $\langle \nu, f \rangle = k$. Any such ν absolutely continuous w.r.t. ρ has Radon-Nikodym derivative $g \in L^1(\mathbb{R}_+, d\rho)$ with the following properties:

$$\int d\rho(x) g(x) = 1 \quad \text{and} \quad \int d\rho(x) g(x) f(x) = k. \tag{93}$$

From convexity of the function $y \mapsto y \log y$

$$D(\nu, \rho) - D(\nu_k, \rho) = \int d\rho(x) \, (g(x) \log g(x) - h(x) \log h(x)) \tag{94}$$

$$\geq \int d\rho(x) \, (g(x) - h(x)) \, (1 + \log h(x)) \tag{95}$$

$$= \int d\rho(x) \, (g(x) - h(x)) \, (1 + \phi_k f(x) - \lambda_f(\phi_k)) = 0, \tag{96}$$

where the last equality follows from (93), the latter holding also in the particular case $g = h$. ■

Proof of Theorem 6: (i) The form of the putative rate function given by the second equality in (32) can be obtained by an argument similar to that used in Proposition 1. But we cannot immediately use the Contraction Principle to conclude the LDP for $(\widehat{a}_n, \widehat{v}_n)$ because $\rho \mapsto a_\rho$ and $\rho \mapsto v_\rho$ are not weakly continuous functions on $\mathcal{M}$. Although we expect that the result may be established by the exponential approximation methods given in Section 4.2.2 of [10], we give here instead a more elementary argument.

First we find the joint CGF of $\widehat{a}_n = n^{-1}\sum_{i=1}^n X_i$ and $\widehat{v}_n = n^{-1}\sum_{i=1}^n (X_i - \widehat{a}_n)^2$. Some notation: let $\mathrm{id_n}$ be the n-dimensional identity matrix, B_n the n-dimensional matrix whose entries are all $1/n$, and $\mathbf{1}$ the n-tuple with unit entries.

$$\begin{aligned}
\mathsf{E}[e^{n(\theta\widehat{a}_n+\phi\widehat{v}_n)}] &= \mathsf{E}[e^{\theta\sum_{i=1}^n X_i+\phi\sum_{i,j=1}^n X_i(\mathrm{id}_n-B_n)_{ij}X_j}] &(97)\\
&= \frac{1}{(2\pi v_\rho)^{n/2}}\int dx_1\ldots dx_n e^{\left(\sum_{i=i}^n\left(\theta x_i-\frac{(x_i-a_\rho)^2}{2v_\rho}\right)+\phi\sum_{i,j=1}^n x_i(\mathrm{id}_n-B_n)_{ij}x_j\right)} &(98)\\
&= \frac{1}{|R_n|^{1/2}}\exp\left(\left((\theta v_\rho+a_\rho)^2\mathbf{1}\cdot R_n^{-1}\cdot\mathbf{1}-na_\rho^2\right)/(2v_\rho)\right), &(99)
\end{aligned}$$

where $R_n = (1-2\phi v_\rho)\mathrm{id}_n + 2\phi v_\rho B_n$. This requires that R_n be positive definite, which we now verify. Since $B_n^k = B_n$ for all $k\in\mathbb{N}$, one verifies that R_n has inverse $R_n^{-1} = (\mathrm{id}_n - 2\phi v_\rho B_n)/(1-2\phi v_\rho)$ when $2\phi v_\rho \neq 1$. A simple induction shows that R_n has determinant $(1-2\phi v_\rho)^{n-1}$. This establishes that when $2\phi v_\rho < 1$, the principal minors of R_n are positive, and so R_n is positive definite as required. $\mathbf{1}\cdot R_n^{-1}\cdot\mathbf{1} = n$ and so we obtain for the CGF

$$\lim_{n\to\infty} n^{-1}\log\mathsf{E}[e^{n(\theta\widehat{a}_n+\phi\widehat{v}_n)}] = \nu(\theta,\phi) := a_\rho\theta + v\theta^2/2 - \frac{1}{2}\log(1-2\phi v_\rho), \tag{100}$$

when $2\phi v_\rho < 1$, and ∞ otherwise. ν satisfies the conditions of the Gärtner-Ellis theorem, and so $(\widehat{a}_n,\widehat{v}_n)$ satisfies an LDP with scale n and rate function ν^*. A short calculation shows that $\nu^* = G$ from (32).

(ii) The LDP for $n^{-1}\widehat{N}_n$ now follows from (i) and the Contraction Principle since $m_{a,v}$ depends continuously on a and v. The form of the minimizing a and v can be found by using Lagrange multipliers to solve the variation problem under the constraint $m_{a,v} = x$.

(iii) We see from (31) that in order to make $m_{a,v} = x \to 0$ in (33) we require $a\to\infty$. From (32) this means that $K(a,v)\to\infty$ and hence $J_\rho(x)\to\infty$. Hence the conditions of Theorem 3 are satisfied. ■

Proof of Theorem 7: We will see that $J_\varepsilon'(m_{\rho,\varepsilon}) = 0$ and $H'(c/a_\rho) = 0$ under the assumptions stated. Then using Taylor's theorem, expanding J_ε near $m_{\rho,\varepsilon}$ and H near c/a_ρ, we can write for all ε sufficiently small, and all y in $[m_{\rho,\varepsilon}, c/a_\rho]$

$$\alpha J_\varepsilon(y) + H(y) = \alpha(y-m_{\rho,\varepsilon})^2 J_\varepsilon''(m_{\rho,\varepsilon})/2 + (y-c/a_\rho)^2 H''(c/a_\rho)/2 + O\left((c/a_\rho - m_{\rho,\varepsilon})^3\right). \tag{101}$$

The variational formula (19) then gives

$$I_\varepsilon(c) = ((c/a_\rho - m_{\rho,\varepsilon})^2\frac{\alpha J_\varepsilon''(m_{\rho,\varepsilon})H''(c/a_\rho)}{2(\alpha J_\varepsilon''(m_{\rho,\varepsilon}) + H''(c/a_\rho))} + O\left((c/a_\rho - m_{\rho,\varepsilon})^3\right). \tag{102}$$

We now find the asymptotics as $\varepsilon\to 0$ of the various terms in (102). By explicit differentiation $H'(c/a_\rho) = \mu_\rho^*(a_\rho) - a_\rho\mu_\rho^{*\prime}(a_\rho) = \mu_\rho(\mu_\rho^{*\prime}(a_\rho)) = \mu_\rho(0) = 0$ and $H''(c/a_\rho) = (a^3/c)\mu_\rho^{*\prime\prime}(a_\rho) = a_\rho^3/(c\mu_\rho''(0)) = a_\rho^3/(cv)$. Let $\varepsilon = (m_\rho\mu_\rho)^*(c) = \sup_\theta(c\theta - m_{\rho,\varepsilon}\mu_\rho(\theta))$ be attained at θ_ε. Since ρ is non-degenerate, μ_ρ is strictly convex in at least some neighborhood of 0 and so clearly $m_{\rho,\varepsilon}\to c/a_\rho$ and $\theta_\varepsilon\to 0$ as $\varepsilon\to 0$. Let $\mu_\rho(\theta) = a_\rho\theta + v\theta^2/2 + \ldots$ be the first two terms in the Taylor expansion of μ_ρ about 0. Then

$$c - m_{\rho,\varepsilon}a_\rho = m_{\rho,\varepsilon}\theta_\varepsilon v + O({\theta_\varepsilon}^2) \quad\text{and}\quad \varepsilon = (c-m_{\rho,\varepsilon}a_\rho)\theta_\varepsilon + v{\theta_\varepsilon}^2/2 + O({\theta_\varepsilon}^3), \tag{103}$$

for ε is a neighborhood of 0, and hence

$$\theta_\varepsilon{}^2 \sim \varepsilon \frac{2a_\rho}{vc} \quad \text{and} \quad c/a_\rho - m_{\rho,\varepsilon} \sim \theta_\varepsilon \frac{cv}{a^2}, \quad \text{as } \varepsilon \to 0. \tag{104}$$

The extremum for $J_\varepsilon(m_{\rho,\varepsilon})$ in (22) is achieved at $\theta = \theta_\varepsilon$ and $\phi = \phi_0 := 0$: for $\theta = \theta_\varepsilon$, the supremum over ϕ is at $\phi = 0$ since $\lambda'_{\theta_\varepsilon}(0) = \langle \rho, g_{\theta_\varepsilon} \rangle = e^{\mu_\rho(\theta_\varepsilon)} = e^{(c\theta_\varepsilon - \varepsilon)/m_{\rho,\varepsilon}}$; and $\lambda^*_{\theta_\varepsilon}(m_{\rho,\varepsilon}) = 0$ since $\lambda_{\theta_\varepsilon}(\phi_0) = 0$.

The first and second derivatives of J_ε at $m_{\rho,\varepsilon}$ are

$$J'_\varepsilon(m_{\rho,\varepsilon}) = -\lambda^*_{\theta_\varepsilon}{}'(e^{(c\theta_\varepsilon - \varepsilon)/m_{\rho,\varepsilon}}) e^{(c\theta_\varepsilon - \varepsilon)/m_{\rho,\varepsilon}} (c\theta_\varepsilon - \varepsilon)/m_\rho \propto \phi_0 = 0, \tag{105}$$

and, discarding some terms proportional to $\lambda^*_{\theta_\varepsilon}{}'(e^{(c\theta_\varepsilon - \varepsilon)/m_{\rho,\varepsilon}}) = 0$:

$$\begin{aligned} J''_\varepsilon(m_{\rho,\varepsilon}) &= \lambda^*_{\theta_\varepsilon}{}''(e^{(c\theta_\varepsilon - \varepsilon)/m_{\rho,\varepsilon}}) \left(e^{(c\theta_\varepsilon - \varepsilon)/m_{\rho,\varepsilon}} (c\theta_\varepsilon - \varepsilon)/m_\rho \right)^2 & (106) \\ &= \frac{1}{\lambda''_{\theta_\varepsilon}(\phi_0)} \left(\frac{\mu_\rho(\theta_\varepsilon) e^{\mu_\rho(\theta_\varepsilon)}}{m_{\rho,\varepsilon}} \right)^2 & (107) \\ &= \left(e^{\mu_\rho(2\theta_\varepsilon) - 2\mu_\rho(\theta_\varepsilon)} - 1 \right)^{-1} \left(\frac{\mu_\rho(\theta_\varepsilon)}{m_{\rho,\varepsilon}} \right)^2. & (108) \end{aligned}$$

The result follows by combining (104), (108) and the expression for H''. ■

Proof of Theorem 8: (i) under (36), these properties are given in Section 3 of [27]

(ii) The proof parallels that of Proposition 1. Since $\lambda^{(2)}_{f,\pi}$ is essentially smooth, there exists a unique ϕ_k for which $(\lambda^{(2)}_{f,\pi})'(\phi_k) = k$. Set $\omega(dx, dy) = \eta(\phi_k, dx)\pi(x, dy)\psi(\phi_k, y)e^{\phi_k f(y) - \lambda_{f,\pi}({}^{(2)}\phi_k)}$. Using the eigen-properties from part (i) one sees that $p_1\omega(dx) = p_2\omega(dx) = \eta(\phi_k, dx)\psi(\phi_k, x)$. From Lemma 4.1 on [27] we have $\int \eta(\phi_k, dx)\psi(\phi_k, x)f(x) = (\lambda^{(2)}_{f,\pi})'(\phi_k) = k$, i.e. $\langle p_1\omega, f \rangle = k$. Then since

$$\frac{d\omega}{d(p_1\omega \otimes \pi)}(x, y) = \frac{\psi(\phi_k, y)}{\psi(\phi_k, x)} e^{\phi_k f(y) - \lambda^{(2)}_{f,\pi}(\phi_k)} \tag{109}$$

we have

$$\begin{aligned} D(\omega, p_1\omega \otimes \pi) &= \int \omega(dx, dy) \left(\phi_k f(y) - \lambda^{(2)}_{f,\pi}(\phi_k) + \log((\psi(\phi_k, x)/\psi(\phi_k, y)) \right) & (110) \\ &= \phi_k(\lambda^{(2)}_{f,\pi})'(\phi_k) - \lambda^{(2)}_{f,\pi}(\phi_k) = (\lambda^{(2)}_{f,\pi})^*(k) & (111) \end{aligned}$$

Now let $\alpha \in \mathcal{M}^{(2)}$ be absolutely continuous w.r.t. ω and such that

$$p_1\alpha = p_2\alpha, \quad \text{and} \quad \langle p_1\omega, f \rangle = k. \tag{112}$$

Then

$$D(\alpha, p_1\alpha \otimes \pi) - D(\omega, p_1\omega \otimes \pi) = D(\alpha, \omega) - D(p_1\alpha, p_1\omega) - \int (d\alpha - d\omega) \log \left(\frac{d\omega}{d(p_1\omega \otimes \pi)} \right) \tag{113}$$

Since p_1 is the adjoint of a unit-preserving positive map, we have $D(\alpha,\omega) - D(p_1\alpha, p_1\omega) \geq 0$, while the last term in (113) is zero, using (109) and (112). Thus $D(\alpha, p_1\alpha \otimes \pi) \geq D(\omega, p_1\omega \otimes \pi)$

(iii) now follows by arguments parallel to those of Theorems 4(iii) and 5. ■

Proof of Theorem 10: Let $Q_{n,\rho}$ denote the distribution of $\widehat{\rho}_n$ when the X_i are distributed as ρ. Then

$$\begin{aligned} \mathsf{P}[S_{\widetilde{N}_n(\widehat{\rho}_n)} \geq nc] &\leq \int dQ_{n,\rho}(\nu) \exp(-n \inf_{c'>c} (n^{-1}\widetilde{N}_n\mu_\rho)^*(c')) &(114)\\ &\leq e^{-n\varepsilon} \int dQ_{n,\rho}(\nu) e^{nD(\nu,\rho)} = e^{-n\varepsilon} \int d\widetilde{Q}_{n,\rho}(x) e^{n\mu_\rho^*(x)}, &(115)\end{aligned}$$

where $\widetilde{Q}_{n,\rho}$ is the distribution of the empirical mean $\int \widehat{\rho}_n(dx)x$. The result then follows from Varadhan's Theorem on observing that because the support of ρ is bounded, then μ_ρ is bounded and continuous on its effective domain. ■

Proof of Theorem 11: (i) By Sanov's Theorem, the family empirical measures of trajectories associated with the trajectories $X^{(n)}(\cdot)$, namely $\widehat{\Xi}_n = n^{-1}\sum_{i=1}^n \delta_{X_i(\cdot)}$, satisfies an LDP with rate function $\Upsilon \mapsto D(\Upsilon, \Xi)$. The LDP then follows from the Contraction Principle if we can show that $\Upsilon \mapsto \inf_{t\in T_{\rm obs}} m_{\Upsilon f_t^{-1},t}$ is weakly continuous. Set $\widehat{m}_{n,t} = m_{\widehat{\Xi}_n f_t^{-1},t}$. Then $\widehat{N}_n(T_{\rm obs}) = \lfloor \widehat{N}'_n(T_{\rm obs})\rfloor$ where $\widehat{N}'_n(T_{\rm obs}) = n\inf_{t\in T_{\rm obs}} \widehat{m}_{n,t}$. Now $\Upsilon \mapsto \Upsilon f_t^{-1}$ is a continuous map from $\mathcal{M}(\Omega)$ to $\mathcal{M}$, both spaces equipped with their weak topologies. For observe with any bounded continuous function g that $\langle \Upsilon f_t^{-1}, g\rangle = \langle \Upsilon, g\circ f_t\rangle$, and $g\circ f_t$ is, by assumption of continuity of f_t, bounded and continuous. We have assumed that $\rho \to m_{\rho,t}$ is weakly continuous. The LDP for $n^{-1}\widehat{N}'_n(T_{\rm obs})$ then follows since taking the infimum over a finite set is continuous. The LDP for $n^{-1}\widehat{N}_n(T_{\rm obs})$ follows by exponential equivalence as before.

(ii) The proof of this parallels that of Theorem 3 and will be omitted. ■

Proof of Theorem 12 (i) By Sanov's Theorem, $\widehat{\rho}_{n,t}$ satisfies an LDP with scale n and rate function $\nu \mapsto D(\nu,\rho_t)$. $\nu \mapsto c_{\nu,t}(\theta)$ is continuous on each $\mathcal{M}_p$, and so assuming $p_{\rho_t} < \infty$, the LDP for $\widehat{c}_n(t,\theta)$ follows by the contraction principle. The alternate form of the rate function follows by application of Proposition 1.

(ii) Write $\mathsf{P}[Q(\widehat{c}_n(t,\delta)) > nb] = \int dW_{n,t}(c) e^{nbf_{nb}(c)}$ where $Q_{n,t}$ is the distribution of $\widehat{c}_n(t,\delta)$ and $f_n(c) = n^{-1}\log \mathsf{P}[Q(c) > n]$. The result then follows by applying Varadhan's theorem. ■

References

[1] N.G. Bean, Robust connection acceptance control for ATM networks with incomplete source information, *Annals of Operations Research,* 48:357–379, 1994

[2] D.D. Botvich and N.G. Duffield, Large deviations, the shape of the loss curve, and economies of scale in large multiplexers, Queueing Systems Theory Appl., 20:293–320, 1995.

[3] C. Casetti, J. Kurose, D. Towsley, An Adaptive Algorithm for Measurement-based Admission Control in Integrated Services Packet Networks, Int. Workshop on Protocols for High Speed Networks, (Sophia Antipolis, Oct. 1996).

[4] C.-S. Chang, Stability, queue length and delay of deterministic and stochastic queueing networks. *IEEE Trans. on Automatic Control*, 39:913–931, 1994.

[5] G. Choquet. *Lectures on Analysis*, Vol. 1. Benjamin, New York, 1969.

[6] G.L. Choudhury, D.M. Lucantoni and W. Whitt, Squeezing the most out of ATM. *IEEE Transactions on Communications*, 44:203–217, 1993.

[7] C. Courcoubetis and R. Weber. Buffer overflow asymptotics for a switch handling many traffic sources. *J. Appl. Prob.*, 33:886–903, 1996.

[8] C. Courcoubetis, G. Kesidis, A. Ridder, J. Walrand and R. Weber, Call Acceptance and Routing using Inferences from Measured Buffer Occupancy, IEEE Trans. Comm., 43:1778–1784, 1995.

[9] S. Crosby, I. Leslie, J.T. Lewis, R. Russell, F. Toomey and B. McGurk, Practical Connection Admission Control for ATM Networks Based on On-line Measurements", Proceedings IEEE ATM '97, June 1997, Lisbon.

[10] A. Dembo and O. Zeitouni, *Large Deviation Techniques and Applications.* Jones and Bartlett, Boston-London, 1993.

[11] A. Demers, S. Keshav, S. Shenker, Analysis and Simulation of a Fair Queueing Algorithm, Internetworking: Research and Experience, 1:3–26, 1990.

[12] M.D. Donsker and S.R.S. Varadhan, Asymptotic evaluation of certain Markov process expectations for large time. III. *Commun. Pure Appl. Math.*, 29:389–461, 1976.

[13] N.G. Duffield, J.T. Lewis, N. O'Connell, R. Russell and F. Toomey, Entropy of ATM traffic streams: a tool for estimating QoS parameters. *IEEE J. Selected Areas in Commun.* **13** 981–990: 1995.

[14] N. G. Duffield, Economies of scale in queues with sources having power-law large deviation scalings, *J. Appl. Prob.*, 33:840–857, 1996.

[15] A.I. Elwalid, D. Mitra and T.E. Stern, Statistical multiplexing of Markov modulated sources: theory and computational algorithms. In: *Teletraffic and Datatraffic in a period of change, ITC-13*, A. Jensen & V.B. Iversen (Eds.) Elsevier Science Publishers B.V. (North-Holland), 1991.

[16] A. Ganesh, P. Green, N. O'Connell, S. Pitts, Bayesian network management. Queueing Systems Theory Appl. 28:267–282, 1998.

[17] A. Ganesh, N. O'Connell, An inverse of Sanov's theorem. Statist. Probab. Letters, to appear, 1998.

[18] A. Ganesh, N. O'Connell, A large deviations principle for Dirichlet posteriors, preprint, 1998.

[19] M.W. Garrett and W. Willinger, Analysis, modeling and generation of self-similar VBR traffic. In *Proceedings ACM SIGCOMM'94*, London, UK, August 1994, pp.269-280.

[20] R.J. Gibbens and P.J. Hunt, Effective Bandwidths for the multi-type UAS channel *Queueing Systems Theory Appl.*, 9:17–28, 1991.

[21] R.J. Gibbens, F.P. Kelly & P.B. Key, A decision-theoretic approach to call admission control in ATM networks IEEE Journal on Selected Areas of Communications, 13:1101–1114, 1995.

[22] P.W. Glynn and W. Whitt, Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. In: *Studies in Applied Probability* Eds. J. Galambos and J. Gani, *Journal of Applied Probability, Special Volume* **31A** 131–159, 1994.

[23] M. Grossglauser, S. Keshav and D.N.C. Tse, RCBR: A simple and efficient service for multiple time-scale traffic, in *Proc. ACM SIGCOMM'95*, 219–230.

[24] M. Grossglauser and D.N.C. Tse, A Framework for robust measurement-based admission control, In Proc. ACM SIGCOMM 97, Cannes, France, September 1997.

[25] R. Guerin, H. Ahmadi and M. Naghshineh, Equivalent capacity and its application to bandwidth allocation in high-speed networks, *IEEE Journal on Selected Areas in Communications*, 9:968-981, 1991.

[26] J.Y. Hui, Resource allocation for broadband networks. *IEEE J. Selected Areas in Commun.* 6:1598–1608, 1988.

[27] I. Iscoe, P. Ney and E. Nummelin, Large deviations of uniformly recurrent Markov additive processes. *Adv. in Appl. Math.* 6:373–412, 1985.

[28] S. Jamin, P.B. Danzig, S. Shenker, and L. Zhang, A measurement-based admission control algorithm for integrated services packet networks, *Proc. ACM SIGCOMM'95* Cambridge, MA, Sept. 1995.

[29] F.P. Kelly. Effective bandwidths at multi-type queues. *Queueing Systems Theory Appl.* 9:5–16, 1991.

[30] F.P. Kelly. Notes on effective bandwidths. in: *Stochastic Networks, Theory and Applications*, Eds. F.P Kelly, S. Zachary and I. Ziedens, Royal Statistical Society Lecture Notes Series, vol. 4, pp,141–168, 1996.

[31] G. Kesidis, J. Walrand and C.S. Chang, Effective bandwidths for multiclass Markov fluids and other ATM Sources. *IEEE/ACM Trans. Networking*, 1:424-428, 1993.

[32] E. Knightly, Second Moment Resource Allocation in Multi-Service Networks, in: Proceedings of ACM SIGMETRICS '97, Seattle, WA, June 1997.

[33] S. Kullback, *Information Theory and Statistics*, Wiley, New York, 1959.

[34] M. Likhanov and R.R. Mazumdar, Cell loss asymptotics in buffers fed with a large number of independent stationary sources, Proc. IEEE INFOCOM'98.

[35] B. McGurk and C. Walsh. Investigations of the performance of a measurement-based Connection Admission Control Algorithm. Proceedings 5th IFIP Workshop on Performance Modelling and Evaluation of ATM Networks, Ilkley, UK, July 1997

[36] S.P. Meyn and R.L. Tweedie, Markov chains and stochastic stability, Springer, New York, 1993.

[37] M. Montgomery and G. de Veciana, On the relevance of time scales in performance oriented traffic characterizations, Proc. IEEE INFOCOM'96.

[38] A.K. Parekh and R.G. Gallagher, A generalized processor sharing approach to flow control in Integrated Services networks: the single node case, *IEEE/ACM Transactions on Networking*, 1:344–357, 1993.

[39] A.K. Parekh and R.G. Gallagher, A generalized processor sharing approach to flow control in Integrated Services networks: the multiple node case *IEEE/ACM Transactions on Networking*, 2:137–150, 1994.

[40] V. Paxson and S. Floyd, Wide-area traffic: the failure of Poisson modeling, *IEEE/ACM Trans. Networking*, 3:226–244, 1995.

[41] R.T. Rockafellar, *Convex Analysis.* Princeton University Press, Princeton, 1970.

[42] A. Simonian and J. Guibert, Large deviations approximation for fluid queues fed by a large number of on-off sources. *Proceedings of ITC 14, Antibes, 1994* pp. 1013–1022.

[43] J.S. Turner, Managing Bandwidth in ATM networks with burst traffic, *IEEE Network Magazine*, September 1992.

[44] S.R.S. Varadhan, Asymptotic probabilities and differential equations, *Commun. Pure Appl. Math.*, 19:261–286, 1966.

[45] A. Weiss, A new technique for analysing large traffic systems. *J. Appl. Prob.* 18:506–532, 1986.

[46] W. Whitt, Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues. *Telecommunications Systems.* 2:71-107, 1993.

[47] W. Willinger, M.S. Taqqu, R. Sherman, D.V. Wilson, Self-Similarity Through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level, *Proceedings of ACM SIGCOMM 1995*.