6th St.Petersburg Workshop on Simulation (2009) 1003-1007

# Comparative Study of Effective Bandwidth Estimators: Batch Means and Regenerative Cycles

Irina Dyudenko[1], Evsey Morozov[2], Michele Pagano[3], Werner Sandmann[4]

**Abstract**

We consider the estimation of effective bandwidths in single-server queueing networks with finite buffers and regenerative input process. Drawbacks of batch means estimators in simulation practice are discussed and a new regenerative estimator is suggested.

## 1. Introduction

In order to characterize the quality of service (QoS) offered by a communication network, one of the most relevant parameters is the packet loss *ratio*, which can be estimated as the buffer overflow probability where the buffer size $b$ of the bottleneck router along the path of the traffic is considered as threshold. The minimum capacity $C_\Gamma$ that guarantees a mean overflow probability of at most $\Gamma$ is called the effective bandwidth (EB) of the incoming traffic where in practice $\Gamma$ is given as a QoS constraint for the maximum acceptable packet loss rate [1, 2, 3].

Effective bandwidth estimation can be treated on the basis of large deviations theory (LDT).

If the queueing system is stable, the weak limit $W_n \Rightarrow W$ of the queue size (workload) process $(W_n)_{n\in\mathbb{N}}$ exists and the stationary workload $W$ *(under mild assumptions (see [10]))* satisfies a large deviations principle (LDP) such that the overflow probability has an asymptotically exponential form [10]. More precisely, let

$$\Lambda(\theta) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E} e^{\theta \sum_{i=1}^{n} X_i} \tag{1}$$

the logarithmic scaled cumulant generating function (LSCGF) of the arrival process where $X_i$ denotes the number of arrivals during the $i$th time unit. Define

$$\delta(C) := \sup\{\theta > 0 : \Lambda(\theta) \le C\theta\}. \tag{2}$$

The work of two first authors is supported by RFBR, project 07-07-00088.

[1]Institute of Applied Mathematical Research, Russian Academy of Sciences, E-mail: dyudenko@krc.karelia.ru

[2]Institute of Applied Mathematical Research, Russian Academy of Sciences, E-mail: emorozov@karelia.ru

[3]University of Pisa, Italy, E-mail: m.pagano@iet.unipi.it

[4]University of Bamberg, Germany, E-mail: werner.sandmann@uni-bamberg.de

Then, provided that the limit (1) exists,

$$\lim_{b\to\infty} \frac{1}{b} \log P(W > b) = -\delta(C), \tag{3}$$

which in turn implies the approximation $P(W > b) \approx e^{-\delta(C)b}$ for the overflow probability. The required effective bandwidth $C_\Gamma$ for a given maximum acceptable packet loss rate $\Gamma$ is given by $C_\Gamma = \min\{C : e^{-\delta(C)b} \le \Gamma\}$. As the equation $e^{-\delta(C)b} = \Gamma$ has a unique root $\theta^* := \delta(C_\Gamma) = -\log\Gamma/b$, it follows from (2) that $C_\Gamma$ can be expressed as

$$C_\Gamma = \frac{\Lambda(\theta^*)}{\theta^*}.$$

Thus, estimation $C_\Gamma$ is reduced to estimation of $\Lambda(\theta^*)$.

In this paper, we consider two different approaches to the estimation of $\Lambda(\theta^*)$, the batch means method and the regenerative method. The general framework addressed by both methods is the construction of confidence intervals for the steady-state mean of a covariance-stationary discrete-time stochastic process $(X_n)_{n\in\mathbb{N}}$. In the setting of queueing networks, typical covariance-stationary processes of interest include the arrival and the service process, the workload process, and the waiting time process, amongst others. The simplest approach to steady-state simulation is the replication-deletion approach where multiple independent realizations (replications) of the stochastic process under consideration are generated and for each realization an initial transient phase must be deleted, which causes an enormous overhead if a lot of replications are needed. In contrast, the batch means method and the regenerative method provide confidence intervals based on one single realization of the process. Before turning to effective bandwidth estimation we briefly outline the general underlying theory and some key properties of the methods.

## 2. Batch means method

With the batch means method the data from one single simulation run of length $n$ is grouped into $k$ batches of size $m$ such that $n = km$. Hence, for $i = 1, \ldots, k$ the $i$-th batch mean and the sample variance from $k$ batches are given by

$$Y_i = \frac{1}{m} \sum_{j=1}^{m} X_{(i-1)m+j}, \qquad S_k^2 = \frac{1}{k-1} \sum_{i=1}^{k} \left(Y_i - \bar{Y}_k\right)^2, \tag{4}$$

where the batch means sample mean equals the overall sample mean of $X_1, \ldots, X_n$ :

$$\bar{Y}_k = \frac{1}{k} \sum_{i=1}^{k} Y_i = \frac{1}{km} \sum_{i=1}^{k} \sum_{j=1}^{m} X_{(i-1)m+j} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}_n. \tag{5}$$

Based on the batch means, a $100(1-\alpha)\%$ confidence interval is constructed by

$$I = \left[\bar{Y}_k - t_{k-1,1-\alpha/2} \frac{S_k}{\sqrt{k}} \;,\; \bar{Y}_k + t_{k-1,1-\alpha/2} \frac{S_k}{\sqrt{k}}\right], \tag{6}$$

where $t_{k-1,1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of the Student $t$-distribution with $k-1$ degrees of freedom.

The justification of constructing confidence intervals based on batch means relies on central limit theorems for covariance-stationary processes, according to which the batch means are asymptotically normal as the batch size approaches infinity. Furthermore, it is shown in [7] that if

$$0 < \sum_{i=-\infty}^{\infty} |Cov(X_j, X_{j+i})| < \infty, \tag{7}$$

then all lag-autocorrelations in the batch means process vanish as the batch size approaches infinity. These results seem to indicate that everything becomes fine when the batch size is chosen sufficiently large. However, asymptotic results do not strictly apply in practice and can be misleading for finite simulation run lengths.

The crucial point is to find a reasonable balance between the batch size and the number of batches. On the one hand, we have to assure by a sufficiently large batch size that the batch means are at least approximately i.i.d. normal in order to achieve a coverage probability close to the nominal value given by the confidence level. On the other hand, we need sufficiently many batches in order to construct reliable confidence intervals that are reasonably narrow and stable.

The study that probably most influenced the choice of the number of batches in practical applications of the batch means method is [8], which is often summarized overly simplified by just citing the recommendation of $10 \leq k \leq 30$ batches. We believe that it is important to know the framework and some of the details that led to this recommendation. In [8], the existence of a maximum number of batches $k^* \geq 2$ and a corresponding minimum batch size $m^* = n/k^*$ is assumed such that for all $k \leq k^*$, the dependency and the nonnormality of the batch means are "negligible" (in an intuitive sense, where a formalization remains open), and only the effects of batch sizes $m \geq m^*$ are studied. In this setting more batches imply a smaller expected confidence interval width but also a smaller coverage probability. According to [8], for all confidence levels the expected width of the confidence interval monotonically increases but the decrease rate quickly decays with increasing number of batches. The standard deviation and the coefficient of variation are much more sensitive to choice of $k$. Consequently, more than 30 batches can be reasonable if the confidence interval stability is important.

Another important point to note is that in practice we usually do not know suitable $k^*$ or $m^*$ as assumed in [8]. In most simulations some nonnormality or dependencies are actually present and cause biased estimators. Moreover, guidelines that are useful in a classical queueing setting may break down when considering realistic Internet traffic models. In particular, condition (7) hardly holds in the presence of long range dependencies. Despite a great deal work on modifications has been carried out, no generally satisfactory choices of $k$ and $m$ are available.

## 3. Regenerative method

The process $(X_n)_{n\in\mathbb{N}}$ is called (zero-delayed) classically regenerative if an infinite sequence $0 = T_0 < T_1 < \ldots$ of *regeneration instants* exists such that the distribution of $X_{n+T_k}$ is the same for each $k \geq 1$ and independent of the pre-history $X_n$, $n < T_k$, $n \geq 1$. The i.i.d. regeneration cycles are defined as $G_n = (X_k,\, T_{n-1} \leq k < T_n)$, and the cycle periods $\tau_n = T_n - T_{n-1}$ are also i.i.d., $n \geq 1$.

$$T_0 = 0,\ T_{n+1} = \min(k > T_n : X_k = 0),\ n \geq 0. \tag{8}$$

We assume the regenerative process to be positive recurrent, that is $\mathbb{E}\tau < \infty$. (Throughout the paper we suppress an index to denote a generic element.)

To estimate a stationary characteristic $\gamma = \mathbb{E}f(X)$ of the process for a measurable function $f$, assuming the weak limit $f(X_n) \Rightarrow f(X)$ exists, we define the i.i.d. variables

$$Y_i = \sum_{k=T_{i-1}}^{T_i-1} f(X_k),\ i \geq 1. \tag{9}$$

If $\mathbb{E}|Y| < \infty$ and the cycle period $\tau$ is aperiodic, then with probability 1,

$$\gamma_n = \frac{1}{n}\sum_{k=1}^{n} f(X_k) \to \frac{\mathbb{E}Y}{\mathbb{E}\tau} = \mathbb{E}f(X),\ n \to \infty. \tag{10}$$

Moreover, assuming that $\sigma^2 \equiv Var(Y - \tau\gamma) \in (0, \infty)$, a (regenerative) central limit theorem states that the $100(1-\alpha)\%$ confidence interval for $\gamma$ is

$$\left[\gamma_n - \frac{z_\alpha\sqrt{v_n}}{\sqrt{n}},\, \gamma_n + \frac{z_\alpha\sqrt{v_n}}{\sqrt{n}}\right], \tag{11}$$

where $z_\alpha$ satisfies $P(N(0,1) \leq z_\alpha) = 1 - \alpha/2$, and the empirical variance

$$v_n = \frac{1}{n}\frac{\sum_{i=1}^{n}(Y_i - \gamma_n\tau_i)^2}{\overline{\tau}_n^2} \Rightarrow \sigma^2. \tag{12}$$

(Here $\overline{\tau}_n$ stands for the sample mean cycle period.) If the number of regenerative cycles $k \leq 30$ then the $1-\alpha/2$ quantile of the Student $t$-distribution with $k-1$ degrees of freedom is more appropriate to use instead of $z_\alpha$. A minimal sufficient condition for $v_n$ to be weakly consistent is $\mathbb{E}(Y - \gamma\tau)^2 < \infty$, while under stronger assumptions, $\mathbb{E}Y^2 < \infty$, $\mathbb{E}\tau^2 < \infty$, the estimate is strongly consistent [11].

## 4. Application to effective bandwidth estimation

With regard to effective bandwidth estimation based on the batch means method, assuming that the batch means $Y_i$ according to (4) are i.i.d. as a random variable $Y$, we obtain

$$\log \mathbb{E}e^{\theta^* \sum\limits_{i=1}^{n} X_i} = \log \mathbb{E}e^{\theta^* \sum\limits_{i=1}^{k} mY_i} = \log \mathbb{E}e^{\theta^* mkY} = \log \mathbb{E}e^{\theta^* nY} = n \log \mathbb{E}e^{\theta^* Y}, \tag{13}$$

which suggests the estimator

$$\hat{\Lambda}(\theta^*) = \frac{1}{n} n \log \frac{1}{k} \sum_{i=1}^{k} e^{\theta^* Y_i} = \log \frac{1}{k} \sum_{i=1}^{k} e^{\theta^* Y_i}. \tag{14}$$

Alternatively, if we consider the batch sums $\hat{X}_i = mY_i$ and assume that they are i.i.d. as a random variable $\hat{X}$, we obtain

$$\log \mathbb{E} e^{\theta^* \sum_{i=1}^{n} X_i} = \log \mathbb{E} e^{\theta^* \sum_{i=1}^{k} \hat{X}_i} = \log \mathbb{E} e^{\theta^* k \hat{X}} = k \log \mathbb{E} e^{\theta^* \hat{X}}, \tag{15}$$

which suggests the estimator [13]

$$\hat{\Lambda}(\theta^*) = \frac{1}{n} k \log \frac{1}{k} \sum_{i=1}^{k} e^{\theta^* \hat{X}_i} = \frac{1}{m} \log \frac{1}{k} \sum_{i=1}^{k} e^{\theta^* \hat{X}_i} \tag{16}$$

For both versions, an estimator of the effective bandwidth is given by

$$\hat{C}_\Gamma = \frac{\hat{\Lambda}(\theta^*)}{\theta^*}. \tag{17}$$

However, the problems with appropriately choosing the batch size $m$ and the number $k$ of batches as outlined in Section 2 for the general framework carries over to effective bandwidth estimation and the $Y_i$ or $\hat{X}_i$, respectively, are only approximately i.i.d. even with a good choice. Therefore, we supposed in [13] not to group into batches of fixed size but to consider regenerative cycles. Indeed, if the arrival process has regeneration instants $T_k$, then the variables

$$\hat{X}_k = \sum_{i=T_k}^{T_{k+1}-1} X_i, \; k \geq 1 \tag{18}$$

are really i.i.d., not only approximately as with the batch means method. Grouping in such a way we form an alternative estimator of the LSCGF function:

$$\hat{\Lambda}(\theta^*) = \frac{k}{T_k} \log \frac{1}{k} \sum_{i=1}^{k} e^{\theta^* \hat{X}_i}, \tag{19}$$

where $k$ is the number of regeneration cycles. Assuming $\mathbb{E} e^{\theta^* \hat{X}} < \infty$, we obtain with probability 1

$$\lim_{n \to \infty} \hat{\Lambda}_k(\theta^*) = \frac{1}{\mathbb{E}\tau} \log \mathbb{E} e^{\theta^* \hat{X}}. \tag{20}$$

Preliminary analysis shows that the desired equality $\Lambda(\theta) = \log \mathbb{E} e^{\theta^* \hat{X}} / \mathbb{E}\tau$ seems plausible [12, 13].

## 5. Summary of simulation results

Due to lack of space, we do not present excessive simulation results but rather briefly describe our simulation setup and summarize our findings. To compare the properties of the two estimators, we consider a two-station tandem network with Poisson input to the first station, a (desired) constant service rate $C_2$ at the second one, and with the finite buffers $b_1$, $b_2$, respectively. In such a system the arrival process to the second station regenerates when an arriving customer sees an empty first station. This allows to construct the regenerative estimator by an evident way. Our goal is to find (by estimation) the required constant rate $C_2$ which guarantees given loss probability $\Gamma$.

The simulation has revealed an advantage of the regenerative estimator of $\Lambda(\theta)$ over the batch mean one (in the terms of variance reduction) when exponential or constant service time at the first station is used [12, 13]. On the other hand, simulation shows that the batch mean estimator, being optimistic, has an advantage when regenerative period has a large variance. Also we consider the state-dependent service rate at the first station: it is $C_1^1$ until queue size exceeds a threshold $L$, and becomes $C_1^2(> C_1^1)$ when buffer exceeds level $L$.

## References


[1] Chang C.-S., Thomas J.A. (1995). Effective Bandwidths on High-Speed Digital Networks. *IEEE Journal on Selected Areas in Communications*, 13(6), 1091–1100.

[2] Gibbens R.J. (1996). Traffic Characterisation and Effective Bandwidths for Broadband Network Traces. In *Stochastic Networks: Theory and Applications*. Ed. by F.P.Kelly, S.Zachary and I. B. Ziedins. Oxford University Press, 169–181.

[3] Kelly F.P. (1996). Notes on Effective Bandwidths. In *Stochastic Networks: Theory and Applications*. Ed. by F.P.Kelly, S.Zachary and I. B. Ziedins. Oxford University Press, 141–168.

[4] Wischik D. (1999). The Output of a Switch, or, Effective Bandwidths for Networks. *Queueing Systems*, 32(4), 383–396.

[5] Asmussen S. (2003). *Applied Probability and Queues*, 2nd, Springer, NY.

[6] Bischak D.P., Kelton W.D. and Pollock S.M. (1993). Weighted Batch Means for Confidence Intervals in Steady-state Simulations. *Management Science*, 39(8), 1002–1019.

[7] Law A.M. and Carson J.S. (1979). A Sequential Procedure for Determining the Length of a Steady-State Simulation. *Operations Research*, 27(5), 1011–1025.

[8] Schmeiser B. (1982). Batch Size Effects in the Analysis of Simulation Output. *Operations Research*, 30(3), 556–568.