Analysis of rare events in queueing systems with model uncertainty

Rami Atar* Amarjit Budhiraja† Paul Dupuis‡ Ruoyu Wu‡

ABSTRACT

We describe a method to quantify robustness in the estimation of probabilities of rare events. Robustness to the underlying probabilistic model is expressed in terms of Rényi divergences. The main application area discussed is queueing networks.

1. INTRODUCTION

Queueing networks are used as models in various application areas such as cloud computing, telecommunications, manufacturing and human service centers including customer call centers and hospitals. Uncertainty in the underlying distributions is a central issue in applying these mathematical models to the real world applications. As far as theoretical asymptotic results are concerned, scaling according to the law of large numbers and central limit theorem (often referred to as fluid and diffusion approximations, respectively) are relatively tolerant to errors made in the distributional assumptions, due to the fact that limit results typically depend only on first and second moments. Model uncertainty is much harder to deal with at the large deviations (LD) scale, as probabilities of rare events are sensitive to the assumed tails of the primitives that drive the model. This paper describes a general method to quantify robustness of performance at the LD scale to the underlying probability distribution and outlines several directions in which it can be used for engineering systems, mostly queueing network models.

The method is based on recently developed perturbation bounds that, roughly stated, assert that risk-sensitive cost functionals corresponding to two given underlying distributions are comparable in terms of a certain information divergence called the *Rényi divergence*. The precise, general form of these bounds is given in §2, where it is also demonstrated that these bounds remain meaningful under LD scaling. Although in the discussion above we emphasized scaling limits, the perturbation bounds are also effective, as we shall demonstrate, at non-asymptotic settings.

In §3 we outline how the method applies to the *generalized Jackson network*. In such a network, service time distributions are general (rather than exponential) and as a result the model is non-Markovian. It is natural to formulate the perturbation bounds in such a way that they compare performance of a possibly difficult collection of models to one that is easier to analyze (analytically or numerically). We present the bounds so that they compare the performance of a collection of these networks to the *Jackson network* (i.e., the Markovian model). Whereas the discussion focuses on Jackson networks, it will be clear that these ideas go beyond the specifics of this model. As we show, in order to apply the proposed tools, calculations of the so-called *Rényi divergence rate* for renewal processes become significant. We provide new bounds on this rate, the proof of which will be reported elsewhere. §4 discusses several additional queueing models including another scaling (the many-server scaling), risk-sensitive cost and control systems, as well as the problem of small noise diffusions.

*Viterbi Faculty of Electrical Engineering, Technion
†Department of Statistics and Operations Research, University of North Carolina
‡Division of Applied Mathematics, Brown University

2. THE PROPOSED METHOD

2.1 Rényi divergence and logarithmic perturbation bounds

Fix a measurable space $(\mathcal{S}, \mathcal{F})$ and denote by $\mathcal{P}$ the set of probability measures on it. For $P, Q \in \mathcal{P}$, the *relative entropy* is given by

$$R(Q\|P) = \begin{cases} \int \log \frac{dQ}{dP}\, dQ & \text{if } Q \ll P \\ +\infty & \text{otherwise.} \end{cases}$$

Introduced in [17] (see [12] for a comprehensive treatment), the *Rényi divergence* of degree $\alpha > 1$, for $P, Q \in \mathcal{P}$, is defined by

$$R_\alpha(Q\|P) = \begin{cases} \frac{1}{\alpha(\alpha-1)} \log \int \left(\frac{dQ}{dP}\right)^\alpha dP & \text{if } Q \ll P \\ +\infty & \text{otherwise.} \end{cases}$$

For $\alpha = 1$, one sets $R_1(Q\|P) = R(Q\|P)$. Whereas two different formulas are used for the cases $\alpha = 1$ and $\alpha > 1$, it is a fact that $\alpha \mapsto R_\alpha(Q\|P)$ is continuous on $[1, \alpha^*]$ provided $R_{\alpha^*}(Q\|P) < \infty$ for some $\alpha^* > 1$. To mention few additional properties, one has that $\alpha \mapsto \alpha R_\alpha$ is nondecreasing on $[1, \infty)$, and given $\alpha \geq 1$, one always has $R_\alpha(Q\|P) \geq 0$, and $R_\alpha(Q\|P) = 0$ if and only if $Q = P$. A property that is of crucial importance in our use of Rényi divergence is its additivity for product measures, in the following sense:

$$R_\alpha(Q_1 \times Q_2\|P_1 \times P_2) = R_\alpha(Q_1\|P_1) + R_\alpha(Q_2\|P_2). \quad (1)$$

It is well known that exponential integrals and relative entropy satisfy a convex duality relation, stated as follows. Let $Q \in \mathcal{P}$. Then for any bounded measurable $g : \mathcal{S} \to \mathbb{R}$,

$$\log \int e^g dQ = \sup_{P \in \mathcal{P}} \Big[\int g dP - R(P\|Q) \Big]. \tag{2}$$

Recently, an analogous relation has been shown for Rényi divergences ([2]; related calculations first appeared in [6]). Namely, fix $\alpha > 1$. Then

$$\frac{1}{\alpha} \log \int e^{\alpha g} dQ = \sup_{P \in \mathcal{P}} \Big[\frac{1}{\alpha - 1} \log \int e^{(\alpha-1)g} dP - R_\alpha(P\|Q) \Big]. \tag{3}$$

The identity (3) may indeed be viewed as an extension of (2), as the latter is recovered by taking the formal limit $\alpha \downarrow 1$ in the former.

Given P, Q and α, as well as an event $A \in \mathcal{F}$, it follows from (3) by taking $g(x) = 0$ [resp., $-M$] for $x \in A$ [resp., $x \in A^c$] and sending $M \to \infty$, that

$$\begin{aligned} &\frac{\alpha}{\alpha - 1} \log P(A) - \alpha R_\alpha(P\|Q) \\ &\quad \le \log Q(A) \le \frac{\alpha - 1}{\alpha} \log P(A) + (\alpha - 1) R_\alpha(Q\|P) \end{aligned} \tag{4}$$

(provided that $P(A) > 0$ and $Q(A) > 0$). In words: the logarithmic probability of an event under Q is estimated in terms of the same event under P and Rényi divergence. It is also a fact that both inequalities in (4) are tight, in the sense that given α, Q and A one can find P that makes them hold as equalities (with different P for each equality) [2].

Our point of view is to regard (4) as perturbation bounds. Given a nominal model P, (4) provides performance bounds on a true model Q in terms of performance under P and divergence terms.

2.2 LD scaling

A fact that makes the perturbation bounds particularly useful is that they remain meaningful under standard LD scaling. We first demonstrate this in the standard setting of independent and identically distributed (IID) random variables (RVs). Let $X_1, X_2, X_3 \ldots$ be a sequence of RVs, and let P and Q be two probability measures that make this sequence IID. Let P_n and Q_n denote the law of $X^n = (X_1, \ldots, X_n)$ under P and Q, resp. For each n, let A_n be an event that is measurable on $\sigma\{X^n\}$, the σ-algebra generated by X^n. We are interested in the exponential decay rate

$$E_n(P) = -\frac{1}{n} \log P(A_n).$$

By the IID assumption, we may appeal to (1), according to which $R_\alpha(Q_n\|P_n) = nR_\alpha(Q_1\|P_1)$. Thus by (4) we obtain the bounds

$$\begin{aligned} &\frac{\alpha - 1}{\alpha} E_n(P) - (\alpha - 1) R_\alpha(Q_1\|P_1) \\ &\quad \le E_n(Q) \le \frac{\alpha}{\alpha - 1} E_n(P) + \alpha R_\alpha(P_1\|Q_1). \end{aligned} \tag{5}$$

In these bounds, the divergence terms remain of order 1 under scaling, and so it is possible to compare the asymptotic behavior of $E_n(Q)$ to that of $E_n(P)$. Moreover, while standard problems in the theory of LD are concerned with limits of expressions such as $E_n(P)$ and $E_n(Q)$, we emphasize that the bounds (5) are valid *for all* n.

Regarding E_n as performance measures in this setting is indeed natural when one is interested in probabilities of rare events. Thus (5) allows one to quantify robustness in the following way. Let a collection $\mathcal{Q}$ of *true models* Q be specified in terms of their divergence from a certain *nominal model* P. That is, given a parameter r, let $\mathcal{Q} = \{Q : R_\alpha(P_1\|Q_1) \le r\}$. Then (5) gives the upper bound $E_n(Q) \le \alpha(\alpha - 1)^{-1} E_n(P) + \alpha r$, which holds for all $Q \in \mathcal{Q}$. A lower bound is obtained similarly. In applications, the selection of P should be based on considerations of tractability and design. Choosing for P a model that is tractable achieves guaranteed bounds on a set of possibly intractable true models Q. Engineering systems often operate under conditions that are distinct from those they are designed for. For such systems, the bounds provide guarantees on their true performance based on the designed performance.

The above instance of the bounds addresses models driven by an IID sequence. It has various significant generalizations. First, one can treat models that do not follow an IID structure. An important class of such models is queueing systems, which are discussed in the next section. If the driving noise does not consist of independent RVs, one cannot appeal to (1), and the Rényi divergence terms must be computed or estimated. Second, the formula (3) allows for general functionals g that need not express probabilities. In particular, one may bound *risk-sensitive costs*. Third, one may handle settings that accommodate control.

As a final general remark, given a particular event or a sequence of events that are of interest, one can optimize over the parameter α for the tightest upper and lower bounds.

3. QUEUEING APPLICATIONS

This section describes how the approach might be applied to queueing models. The proposed approach is highly relevant to these models because much is known regarding their LD behavior in the Markovian setting (e.g., [3, 5, 7, 18]), but considerably less on models with general service time distributions (e.g., [13]). Since in practice service times are often non-exponential, there is interest in the study of non-Markovian models. It is natural to seek comparison to the Markovian models. Moreover, the authors are not aware of work on the robustness of these models at the LD scale.

We outline directions of study for the Generalized Jackson Network (GJN), which we regard a prototype for a far greater class of models.

3.1 The generalized Jackson network

We are interested in developing performance guarantees for GJN based on known sample path LD properties of the JN. The latter appear in work of Dupuis-Ellis [5] and work that has emanated from it.

The GJN consists of a fixed, finite number, K, of service stations, where at each station jobs queue up to get service in a FIFO manner, and upon departure they are routed probabilistically to one of the stations or leave the system. Exogenous arrivals follow renewal processes, and service times are IID (for each station). A Jackson Network (JN) corresponds to the special case in which exogenous arrivals are Poisson and service times are exponential. Thus the queueing process in a JN forms a Markov process on $\mathbb{Z}_+^K$.

A key observation regarding the applicability of our approach to these models is based on viewing them as dynamical systems driven by renewal processes. To make this point, we shall keep the discussion as simple as possible as far as the initial configuration is concerned by assum-

ing that there is no stochasticity associated to it (as can be achieved, for example, by assuming that the initial number of jobs at each station is deterministic, all servers at stations that are initially nonempty start a service cycle at time zero, and the arrival clock also starts afresh at time zero). Let S_k denote the potential service process for server k. Namely, the number of jobs this server processes by the time that it has been busy for t time units is given by $S_k(t)$. If $T_k(t)$ denotes the busyness time of that server by time t, $D_k(t) = S_k(T_k(t))$ jobs have been processed by that time. Denote by A_k the exogenous arrival processes; these are renewal processes by assumption. Denote by $\{R_k(i)\}_{i\in\mathbb{N}}$ the routing variables; for each k, these are IID RVs taking values in $\{1,\dots,K+1\}$. The *primitive processes* $\{A_k, S_k, R_k\}_{k\le K}$ fully determine the behavior of the system. Moreover, denote

$$\Delta(t) = \{A_k|_{[0,t]}, S_k|_{[0,t]}, R_k \circ S_k|_{[0,t]}\}_{k\le K}.$$

Then owing to the fact that $T_k(t) \le t$, the sample paths $\{X_k|_{[0,t]}\}_{k\le K}$ of the queue length processes X_k by time t are determined by $\Delta(t)$. However complicated the dependence of the queueing process on the data might be, this observation allows us to use the approach in a manner similar to that outlined above for IID driven systems, where the process Δ plays the role of the driving process.

Fix t, and consider a sequence of events A_n, where for each n, A_n is measurable on $\sigma\{X|_{[0,nt]}\}$. Then by the discussion above, it is also measurable on $\sigma\{\Delta(nt)\}$. Let Q and P correspond to the GJN and JN, respectively. Thus under Q, A_k and S_k are renewal processes, and under P these are Poisson processes. (One has a degree of freedom in choosing the parameters of the Poissons.) Denote by P_n and Q_n the respective laws of $\Delta(nt)$ under the two measures. Then analogously to the derivation of (5) we obtain

$$\begin{aligned}\frac{\alpha-1}{\alpha}E_n(P) - (\alpha-1)\frac{R_\alpha(Q_n\|P_n)}{n} \\ \le E_n(Q) \le \frac{\alpha}{\alpha-1}E_n(P) + \alpha\frac{R_\alpha(P_n\|Q_n)}{n}.\end{aligned} \tag{6}$$

The two Rényi divergence terms above are basically concerned with the *Rényi divergence rate* (RDR) between a general renewal process and a Poisson process. We report on some progress on this direction in §3.2.

3.2 RDR estimates for renewal processes

Calculations and bounds of entropy rate and Rényi entropy rate have been studied for several families of stochastic processes, including Markov chains and hidden Markov models [8], [14]. However, the precise question that arises from the above discussion is concerned with the RDR of a renewal process with respect to a Poisson process, which seems not to have been addressed before. In this section we present some results in this direction.

Let N_t be a simple counting process. Assume that under Q it is a renewal with inter-event distribution π and under P it is a Poisson(1) process. Let $P_t = P \circ N|^{-1}_{[0,t]}$ and $Q_t = Q \circ N^{-1}_{[0,t]}$, and let r_α denote the RDR defined by

$$r_\alpha = \limsup_{t\to\infty} t^{-1} \log R_\alpha(Q_t\|P_t).$$

Assume that π has a density, denoted by g, and let h denote the hazard rate, $h(x) = g(x)/\pi[x,\infty)$. Let also

$$H(x) = \int_0^x (1-h(s))ds + \log h(x).$$

Let Z be a RV distributed according to π. Let β denote the logarithmic moment generating function of $(Z, H(Z))$, and let β^* be its Legendre transform:

$$\beta(\lambda) = \log Ee^{\lambda_1 Z + \lambda_2 H(Z)}, \quad \lambda\in\mathbb{R}^2,$$

$$\beta^*(x) = \sup_\lambda\{\langle\lambda, x\rangle - \beta(\lambda)\}, \quad x\in\mathbb{R}^2.$$

Let

$$G_\alpha(\theta) = \theta \sup_{x\in\mathbb{R}^2: x_1\in[0,\theta^{-1}]} [\alpha x_2 - \beta^*(x_1,x_2)], \qquad \theta\in(0,\infty)$$

Then we have the following general bound on r_α.

THEOREM 3.1. *If β is finite in a neighborhood of the origin then*

$$r_\alpha \le \frac{1}{\alpha(\alpha-1)} \sup_\theta G_\alpha(\theta)^+.$$

The proof uses the explicit expression

$$\Lambda_t = e^{-\int_0^t (1-h(V_s))ds + \int_0^t \log h(V_{s-})dN_s}$$

of the Radon-Nikodym derivative, where V_t denotes the time since the last jump preceding t (or zero). The RDR is bounded in terms of exponential integrals involving the pair of partial sums $(\sum_{i=1}^n Z_i, \sum_{i=1}^n H(Z_i))$, where Z_i are IID RVs distributed according to π. LD estimates based on Laplace's principle and Cramer's theorem then provide the bound stated in the theorem.

In several families of processes we have more concrete bounds. These include the compound Poisson process and counting processes with Gamma-distributed inter-events.

Consider a compound Poisson with intensity $\lambda(\cdot)$ under Q. Set $k_\alpha(x) = x^\alpha - 1 - \alpha x + \alpha$. If $0 < a \le \lambda(t) \le b < \infty$ a.s., then we have the bound

$$R_\alpha(Q_t\|P_t) \le \frac{k_\alpha(a)\vee k_\alpha(b)}{\alpha(\alpha-1)}t,$$

for all t. (In particular, the RDR is bounded by the constant in front of t above).

Next, the Gamma distribution has been proposed as a model for service times in applications (eg. in [1], [15]). For renewal processes with Gamma inter-event distribution we have the following. Assume $\pi = \Gamma(k,\rho)$ with $k \ge 1$, $\rho > 1$, namely $g(x) = \frac{\rho^k}{\Gamma(k)}x^{k-1}e^{-\rho x}$. Then

$$\begin{aligned}&\sup_\theta G_\alpha(\theta) \\ &\quad\le \Big(\frac{\Gamma(1+\alpha(k-1))}{\Gamma(k)^\alpha}\rho^{\alpha k}\Big)^{\frac{1}{1+\alpha(k-1)}} - \alpha(\rho-1) - 1.\end{aligned} \tag{7}$$

Along with Theorem 3.1, this gives a bound on the RDR r_α in this case.

In the exponential case, $k = 1$, the expression in (7) above reduces to

$$\rho^\alpha - \alpha(\rho-1) - 1,$$

which, as we can verify, gives a tight upper bound on the RDR.

4. DISCUSSION AND OPEN PROBLEMS

4.1 Exact expressions for the RDR

While the RDR estimate of Theorem 3.1 can be useful to obtain perturbation bounds along the lines described above, it is desirable to obtain exact expressions. Moreover, the lower perturbation bound requires information on the RDR of a Poisson process with respect to the renewal process. This motivates the following.

PROBLEM 4.1. *Compute the RDR of a renewal process with respect to a Poisson, and for a Poisson with respect to a renewal process, for broad families of renewal processes. More generally, compute (or provide useful bounds) for the RDR of one renewal process with respect to another.*

Carrying out the proposed program for a given queueing model depends on the bounds one has on the Rényi divergence term as well as on the particular choice of events of interest. Quality of service considerations in heavily loaded cloud computing applications may be approached by considering large delays building up in a GJN. In other applications one may be concerned with buffer overflows.

4.2 The $G/G/n$ model

There has been much interest in recent years in the $G/G/n$ model in a setting where the number of servers n plays also the role of the scaling parameter, and the arrival process is scaled proportionally to n. This specific scaling has been referred to as a *many-server* scaling. This model and scaling were studied for their LLN and CLT asymptotics in [10], [11], [16]. The method outlined above seems useful in studying robust LD estimates for the $G/G/n$ model based on the much easier $M/M/n$ model.

For $G/G/n$ models that accommodate abandonment [9], one may be interested in the event of a large abandonment count over a given time interval.

4.3 Risk-sensitive control

We focus on one out of various RS control problems that are of interest in the many-server $G/G/n$ setting. Consider a parallel server model consisting of a many-server pool with multiple customers classes and customer abandonment. The term 'parallel server' refers to the fact that each arrival requires a single service that can be attained at any one of the servers. A recurring theme in the literature on this model is how to choose service policy to minimize abandonment count over a given time interval. The motivation comes from large call centers. The need to cover general service time distribution as well as patience time distribution, has been recognized many times in earlier work on this model. The question of RS cost has not been addressed before. It is reasonable to expect that one might solve the problem for the Markovian setting (which here means exponential service and patience times). The perturbation bounds then can be used to yield performance guarantees for the non-Markovian setting.

4.4 Small noise diffusion

In this work we have focused on queuing systems where the driving 'noise' is given by renewal processes. Many engineering systems are described through noise processes that are Gaussian or appropriate perturbations of Gaussian processes. Bounds as in (6) can be used for deriving robust large deviation bounds for such systems as well. A basic example of such bounds is as follows. Suppose that ν is the standard Wiener measure on $\mathcal{C} = C([0,1] : \mathbb{R}^d)$ (the space of $\mathbb{R}^d$-valued continuous functions on $[0,1]$ equipped with the uniform topology). Denote the canonical coordinate process on $\mathcal{C}$ by W and the canonical filtration by $\{\mathcal{F}_t\}$. For $\varepsilon \in (0,1)$ and $K \in (0,\infty)$ let $\mathcal{U}^\varepsilon$ be the collection of $\mathbb{R}^d$-valued progressively measurable processes that satisfy $\int_{[0,1]} \|u(s)\|^2 ds \le K/\varepsilon$. For $u \in \mathcal{U}^\varepsilon$ let θ_u denote the probability law of $W(\cdot) + \int_0^\cdot u(s)ds$, and let $\mathcal{M}^\varepsilon$ be the collection of all such measures. Then we have the bound

$$\limsup_{\varepsilon\to 0} \sup_{\theta\in\mathcal{M}^\varepsilon} \varepsilon \log P_\theta(\sqrt{\varepsilon}W(\cdot) \in A) \le -([\sqrt{I(A)} - \sqrt{K/2}]^+)^2,$$

where I is the classical rate function in Schilder's theorem, namely for a Borel set A in $\mathcal{C}$

$$I(A) = \inf\left\{ \frac{1}{2}\int_{[0,1]} \|\dot\varphi(s)\|^2 ds : \varphi \in A \right\}.$$

This bound is obtained by appealing to (4), evaluating the Rényi divergence and optimizing over the parameter α. A research program that aims to develop such robust large deviation bounds for various quantities of interest, such as path probabilities for diffusion processes, probabilities associated with exit times and locations from bounded domains, invariant measure probabilities, etc., is still in very early stages.

4.5 Partly uncertain models

In some systems there may be parts of the model that are well modelled (e.g., exponential interarrival time distributions), and one would like bounds which can distinguish these parts of the model from the parts one wishes to consider as uncertain. An approach to this issue in the setting of ordinary performance measures and using relative entropy, spelled out in [4], could possibly be adapted to the situation where performance is determined by risk-sensitive costs.

5. REFERENCES

[1] J. Abate, G. L. Choudhury, and W. Whitt. Waiting-time tail probabilities in queues with long-tail service-time distributions. *Queueing Systems*, 16(3): 311–338, Sep 1994.

[2] R. Atar, K. Chowdhary, and P. Dupuis. Robust bounds on risk-sensitive functionals via Rényi divergence. *SIAM/ASA J. Uncertain. Quantif.*, 3(1): 18–33, 2015.

[3] R. Atar and P. Dupuis. Large deviations and queueing networks: methods for rate function identification. *Stoch. Proc. and Their Appl.*, 84:255–296, 1999.

[4] K. Chowdhary and P. Dupuis. Distinguishing and integrating aleatoric and epistemic variation in uncertainty quantification. *ESAIM: Mathematical Modelling and Numerical Analysis*, 47:635–662, 2013.

[5] P. Dupuis and R. S. Ellis. The large deviation principle for a general class of queueing systems, I. *Trans. Amer. Math. Soc.*, 347:2689–2751, 1996.

[6] K. Dvijotham and E. Todorov. A unified theory of linearly solvable optimal control. *Artificial Intelligence (UAI)*, page 1, 2011.

[7] I. Ignatiouk-Robert. Large deviations of jackson networks. *Ann. Appl. Probab.*, 10(3):962–1001, 08 2000.

[8] P. Jacquet, G. Seroussi, and W. Szpankowski. On the entropy of a hidden Markov process. *Theoretical computer science*, 395(2-3):203–219, 2008.

[9] W. Kang and K. Ramanan. Fluid limits of many-server queues with reneging. *Ann. Appl. Probab.*, 20(6):2204–2260, 2010.

[10] H. Kaspi and K. Ramanan. Law of large numbers limits for many-server queues. *Annals of Applied Probability*, 21(1):33–114, 2011.

[11] H. Kaspi and K. Ramanan. SPDE limits of many-server queues. *Ann. Appl. Probab.*, 23(1): 145–229, 2013.

[12] F. Liese and I. Vajda. *Convex Statistical Distances*. Teubner-Texte zur Mathematik. Teubner, 1987. ISBN 9783322004284.

[13] K. Majewski. Large deviation bounds for single class queueing networks and their calculation. *Queueing Systems*, 48(1):103–134, Sep 2004.

[14] E. Ordentlich and T. Weissman. New bounds on the entropy rate of hidden Markov processes. In *Information Theory Workshop, 2004. IEEE*, pages 117–122. IEEE, 2004.

[15] M. Rajaratnam and F. Takawita. Hand-off traffic characterisation in cellular networks under non-classical arrivals and gamma service time distributions. In *Personal, Indoor and Mobile Radio Communications, 2000. PIMRC 2000. The 11th IEEE International Symposium on*, volume 2, pages 1535–1539. IEEE, 2000.

[16] J. Reed. The $G/GI/N$ queue in the Halfin-Whitt regime. *Ann. Appl. Probab.*, 19(6):2211–2269, 2009.

[17] A. Rényi. On measures of entropy and information. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I*, pages 547–561, Berkeley, Calif., 1961. Univ. California Press.

[18] A. Shwartz and A. Weiss. *Large Deviations for Performance Analysis: Queues, Communication and Computing*. Chapman and Hall, New York, 1995.