A Large Deviation Principle with Queueing Applications

Neil O'Connell
Basic Research Institute in the Mathematical Sciences
HP Laboratories Bristol

HPL-BRIMS-97-05
April, 1997

In this paper we present a large deviation principle, for partial sums processes indexed by the half line, which is particularly suited to queueing applications.

Internal Accession Date Only

1 Introduction

The main result in this paper provides a new tool for looking at large deviations for queueing systems in equilibrium. Equilibrium systems have generally been treated on a case-by-case basis, with much work and/or additional hypotheses necessary to prove large deviation principles (see, for example, Chang and Zajic [3], Ganesh and Anantharam [11], Ramanan and Dupuis [19]). We provide a simple sufficient condition for the usual sample path LDP (as in Mogulskii's theorem) to be strengthened to a topology for which the reflection mappings appearing in many queueing applications are continuous and the contraction principle can be applied. A step in this direction was made by Dobrushin and Pechersky [5], who introduce a finer topology (a guage topology) which allows one to treat the single server queue with constant service rate, and prove the LDP in this topology for a class of Markov jump processes. However, this does not easily extend to more complicated network configurations, or even to the single server queue with stochastic service rate. The main result in this paper can be (and has been) applied to some quite complicated multidimensional systems with interacting traffic [16, 17].

2 Background and motivation

The context in which the need for our main result arises is a general scheme which can be applied to an endless variety of network problems where the goal is to establish probability approximations for aspects of a system (such as queue lengths) under very general ergodicity and mixing assumptions about the network inputs.

Suppose that the inputs to a network can be represented by a sequence of random variables (X_k) in $\mathbb{R}^d$, and that the (sequence of) objects of interest, (O_n), can be expressed as a function of the partial sums process corresponding to X. To make this more precise, for $t \geq 0$ set

$$S_n(t) = \frac{1}{n} \sum_{k=1}^{[nt]} X_k, \tag{1}$$

and write $\tilde{S}_n$ for the polygonal approximation to S_n:

$$\tilde{S}_n(t) = S_n(t) + \left(t - \frac{[nt]}{n}\right)\left(S_n\left(\frac{[nt]+1}{n}\right) - S_n\left(\frac{[nt]}{n}\right)\right). \qquad (2)$$

Denote by $C(\mathbb{R}_+)$ the space of continuous functions on $\mathbb{R}_+$. Then $\tilde{S}_n \in C^d(\mathbb{R}_+)$ and our supposition is that there exists a function $f : C^d(\mathbb{R}_+) \to \mathcal{X}$, for some space $\mathcal{X}$, such that $O_n = f(\tilde{S}_n)$, for each n.

For example, suppose $d = 1$ and X_k is the difference between the amount of work arriving at time $-k$ at a single-server queue and the available service capacity at that time. Suppose also that the limit

$$\mu := \lim_{n\to\infty} \sum_{k=1}^{n} X_k/n$$

exists almost surely and is less than 0. Then the queue length at time zero is given by

$$Q_0 = \sup_{n\geq 0} \sum_{k=0}^{n} X_k, \qquad (3)$$

or, equivalently, $Q_0/n = f(\tilde{S}_n)$, where $f : \mathcal{A}_\mu(\mathbb{R}_+) \to \mathbb{R}_+$ is defined by

$$f(\phi) = \sup_{t>0} \phi(t). \qquad (4)$$

If the sequence X_k is stationary and ergodic, then Q_0 represents the equilibrium queue-length distribution. In this example, $O_n = Q_0/n$.

The idea is to deduce a large deviation principle (see below) for O_n from one which can generally be assumed for $\tilde{S}_n$. This can be done using the *contraction principle*, which we will now describe.

Let $\mathcal{X}$ be a Hausdorff topological space with Borel σ-algebra $\mathcal{B}$, and let μ_n be a sequence of probability measures on $(\mathcal{X}, \mathcal{B})$. We say that μ_n satisfies the *large deviation principle* (LDP) with rate function I, if for all $B \in \mathcal{B}$,

$$-\inf_{x\in B^\circ} I(x) \leq \liminf_n \frac{1}{n} \log \mu_n(B) \leq \limsup_n \frac{1}{n} \log \mu_n(B) \leq -\inf_{x\in \bar{B}} I(x); \quad (5)$$

if, for each n, Z_n is a realisation of μ_n, it is sometimes convenient to say that the sequence Z_n satisfies the LDP. A rate function is *good* if its level sets are compact. The contraction principle states that if Z_n satisfies the

LDP in a Hausdorff topological space $\mathcal{X}$ with good rate function I, and f is a continuous mapping from $\mathcal{X}$ into another Hausdorff topological space $\mathcal{Y}$, then the sequence $f(Z_n)$ satisfies the LDP in $\mathcal{Y}$ with good rate function given by

$$J(y) = \inf\{I(x) : \ f(x) = y\}.$$

Now consider the partial sums process $\tilde{S}_n$. Denote by $\tilde{S}_n[0,1]$ the restriction of $\tilde{S}_n$ to the unit interval, by $C[0,1]$ the space of continuous functions on $[0,1]$, equipped with the uniform topology, and by $\mathcal{A}[0,1]$ the subspace of absolutely continuous functions on $[0,1]$ with $\phi(0) = 0$. Dembo and Zajic (1995) establish quite general conditions for which $\tilde{S}_n[0,1]$ satisfies the LDP in $\mathcal{A}[0,1]$ with good convex rate function given by

$$I(\phi) = \begin{cases} \int_0^1 \Lambda^*(\dot{\phi})ds & \phi \in \mathcal{A}[0,1] \\ \infty & \text{otherwise}, \end{cases} \tag{6}$$

where Λ^* is the Fenchel-Legendre transform of the scaled cumulant generating function

$$\Lambda(\lambda) = \lim_{n\to\infty} \frac{1}{n} \log Ee^{n\lambda\cdot S_n(1)}, \tag{7}$$

which is assumed to exist for each $\lambda \in \mathbb{R}^{d+1}$ as an extended real number. For such an LDP to hold in the *i.i.d.* case, it is sufficient that the moment generating function $Ee^{\lambda\cdot X_1}$ exists and is finite everywhere; this is a classical result, due to Varadhan (1966) and Mogulskii (1976). This is usually extended to the space $C(\mathbb{R}_+)$ (of continuous functions on $\mathbb{R}_+$), via the Dawson-Gärtner theorem for projective limits. However, the projective limit topology (the topology of uniform convergence on compact intervals) is not strong enough for many applications; in particular, the function f defined by (4) is not continuous in this topology on any supporting subspace, and so the contraction principle does not apply.

This was observed by Dobrushin and Pechersky [5], who introduce a finer topology (a guage topology) which allows one to treat the single server queue with constant service rate, and prove the LDP in this topology for a class of Markov jump processes. In this topology, the restriction of the mapping

$$\phi \mapsto \sup_{t\geq 0}[\phi(t) - t] \tag{8}$$

to a subspace of non-decreasing paths ϕ with limits

$$\lim_{t\to\infty} \phi(t)/t = \mu < 1,$$

is continuous. However, this does not easily extend to more complicated network configurations, or even to the single server queue with (stochastic) time-varying capacity.

We consider the set of paths

$$\mathcal{Y} = \bigcap_j \left\{ \phi \in C^d(\mathbb{R}_+) : \lim_{t\to\infty} \frac{\phi^j(t)}{1+t} \text{ exists} \right\},$$

and equip $\mathcal{Y}$ with the norm

$$||\phi||_u = \sup_j \sup_t \left| \frac{\phi^j(t)}{1+t} \right|.$$

Note that $\mathcal{Y}$ can be identified with the Polish space $C^d(\mathbb{R}_+^*)$ of continuous functions on the extended (and compactified) real line, equipped with the supremum norm, via the bijective mapping $\phi(t) \mapsto \phi(t)/(1+t)$. In particular, $\mathcal{Y}$ is a Polish space.

We will show that if the LDP holds in $C[0,1]$ and Λ is differentiable at the origin with $\nabla\Lambda(0) = \mu$, then the LDP holds in the subspace

$$\mathcal{Y}_\mu = \{\phi \in \mathcal{Y} : \lim_{t\to\infty} \frac{\phi^j(t)}{1+t} = \mu\}.$$

In our one-dimensional example, the function defined by (4) on $\mathcal{Y}_\mu$ is continuous provided $\mu < 0$ (see Section 4).

3 The main result

We consider the set of paths

$$\mathcal{Y} = \bigcap_j \left\{ \phi \in C^d(\mathbb{R}_+) : \lim_{t\to\infty} \frac{\phi^j(t)}{1+t} \text{ exists} \right\},$$

and equip $\mathcal{Y}$ with the norm

$$||\phi||_u = \sup_j \sup_t \left| \frac{\phi^j(t)}{1+t} \right|.$$

Note that $\mathcal{Y}$ can be identified with the Polish space $C^d(\mathbb{R}_+^*)$ of continuous functions on thc extended (and compactified) real line, equipped with the supremum norm, via the bijective mapping $\phi(t) \mapsto \phi(t)/(1+t)$. In particular, $\mathcal{Y}$ is a Polish space.

Although this topology is quite different from the gauge topology introduced by Dobrushin and Pechersky [5], conceptually it is quite similar: the idea is to get some kind of uniform control over the sample average. We have also used some ideas from their paper in the proof of Theorem 1 below, in order to construct compact sets that support most of the measure. We remark also that Deuschel and Stroock [8] prove a version of Schilder's theorem in the space $\mathcal{Y}$, using essentially Gaussian techniques.

Theorem 1 *Suppose that for each $\theta \in \mathbb{R}^d$, the limit*

$$\Lambda(\theta) = \lim_{n\to\infty} \frac{1}{n} \log E e^{n\theta \cdot S_n(1)}, \tag{9}$$

exists as an extended real number, and the sequence $\tilde{S}_n[0,1]$ satisfies the LDP in $C^d[0,1]$ with good rate function given by

$$I_1(\phi) = \begin{cases} \int_0^1 \Lambda^*(\dot{\phi})ds & \phi \in \mathcal{A}^d[0,1], \\ \infty & \mathit{otherwise} \end{cases}$$

where Λ^* is the convex dual of Λ. If Λ is differentiable at the origin, then $\tilde{S}_n$ satisfies the LDP in $\mathcal{Y}$ with good rate function

$$I_\infty(\phi) = \begin{cases} \int_0^\infty \Lambda^*(\dot{\phi})ds & \phi \in \mathcal{A}^d(\mathbb{R}_+) \cap \mathcal{Y}, \\ \infty & \mathit{otherwise} \end{cases}$$

Proof. By considering $\tilde{S}_n(t) - t\nabla\Lambda(0)$ we can, without loss of generality, assume that $\nabla\Lambda(0) = 0$. We first show that $\mathcal{D}_I \subset \mathcal{Y}$ and $P(\tilde{S}_n \in \mathcal{Y}) = 1$, for all n. By the convexity of Λ^*, and Jensen's inequality,

$$t\Lambda^*(\phi(t)/t) \leq I(\phi).$$

Since this holds for all t, we must have $\phi(t)/t \to 0$ as $t \to \infty$ ($\nabla\Lambda(0) = 0$ implies that Λ^* has a unique zero at the origin). By hypothesis we can choose, for each j, $\theta > 0$ such that $\Lambda_j(\theta)$ and $\Lambda_j(-\theta)$ are finite; if we let

$$\epsilon_+(n) = \left|\frac{1}{n}\log Ee^{n\theta S_n^j(1)} - \Lambda_j(\theta)\right|$$

and

$$\epsilon_-(n) = \left|\frac{1}{n}\log Ee^{-n\theta S_n^j(1)} - \Lambda_j(-\theta)\right|,$$

then $\epsilon_+(n) \vee \epsilon_-(n) \to 0$, as $n \to \infty$. Thus, for each $\delta > 0$,

$$\begin{aligned} &P\left\{\left|\frac{\tilde{S}_n^j(t)}{1+t}\right| > \delta, \text{ for some } t > t_0\right\} \\ &\leq \sum_{k>nt_0} e^{-\delta k\theta+\Lambda(\theta)+\epsilon_+(k)} + \sum_{k>nt_0} e^{-\delta k\theta+\Lambda(-\theta)+\epsilon_-(k)}; \end{aligned}$$

letting $t_0 \to \infty$ we see that $\tilde{S}_n^j(t)/(1+t) \to 0$, almost surely, as $t \to \infty$. We have thus shown that $\mathcal{D}_I \subset \mathcal{Y}$ and $P(\tilde{S}_n \in \mathcal{Y}) = 1$. Now by (H1), the Dawson-Gärtner theorem for projective limits, and [7, Lemma 4.1.5], we have that $\tilde{S}_n$ satisfies the LDP in $\mathcal{Y}$ when equipped with the topology of uniform convergence on compact intervals. To strengthen this to the topology induced by the norm $\|\cdot\|_u$, we appeal again to the inverse contraction principle, by which it suffices to prove exponential tightness in the space $(\mathcal{Y}, \|\cdot\|_u)$.

For each t, denote by $\mathcal{C}^d[0,t]$ the projection of $\mathcal{C}^d(\mathbb{R}_+)$ onto the interval $[0,t]$, equipped with the uniform topology, and by $\phi[0,t]$, for $\phi \in \mathcal{C}^d(\mathbb{R}_+)$, the restriction of ϕ to the interval $[0,t]$. Goodness of the rate function in (H1) implies that the sequence $\tilde{S}_n[0,1]$ is exponentially tight in the uniform topology on $\mathcal{A}^d[0,1]$. In other words, for each $\alpha > 0$, there exists a compact set K_α in $\mathcal{A}^d[0,1]$ such that

$$\limsup_n \frac{1}{n}\log P(\tilde{S}_n[0,1] \notin K_\alpha) \leq -\alpha.$$

It follows that for each $t > 0$,

$$K_\alpha(t) := \{\phi \in \mathcal{C}^d[0,t] : \{s \mapsto \phi(st)\} \in K_\alpha\}$$

is compact in $\mathcal{C}^d[0,t]$, and for each $0 < \epsilon < \alpha$,

$$\begin{aligned}\limsup_n \frac{1}{n}\log P\bigcup_{t>1}\{\tilde{S}_n[0,t] \notin K_\alpha(t)\} &\leq \limsup_n \frac{1}{n}\log P\bigcup_{t>1}\{\tilde{S}_{nt}[0,1] \notin K_\alpha(1)\} \\ &\leq \limsup_n \frac{1}{n}\log \sum_{k>n} e^{-(\alpha-\epsilon)k} \\ &\leq -\alpha+\epsilon.\end{aligned}$$

Since ϵ is arbitrary we have, for each $\alpha > 0$,

$$\limsup_n \frac{1}{n}\log P\bigcup_{t>1}\{\tilde{S}_n[0,t] \notin K_\alpha(t)\} \leq -\alpha. \tag{10}$$

For $\alpha, t > 0$, set

$$d_\alpha(t) = \begin{cases} \alpha^2 & t \leq \alpha^2 \\ t^{-1/2} & t > \alpha^2 \end{cases}$$

and consider the sets

$$D_\alpha = \bigcap_j \left\{\phi \in \mathcal{Y} : \left|\frac{\phi^j(t)}{1+t}\right| \leq d_\alpha(t), \text{ for all } t, \ \phi[0,t] \in K_\alpha(t) \text{ for all } t > 1\right\}.$$

The exponential tightness of $\tilde{S}_n$ in $(\mathcal{Y}, \|\cdot\|_u)$ will be established by the following two lemmas.

Lemma 1 *For each $\alpha > 0$, D_α is compact in $(\mathcal{Y}, \|\cdot\|_u)$.*

Proof. Let ϕ_n be a sequence in D_α. By Tychonoff's theorem, the set $\cap_{t>1}K_\alpha(t)$ is compact in $\mathcal{Y}$ when equipped with the topology of uniform convergence on compact intervals, so there exists a subsequence $n(k)$ such that ϕ converges to some $\phi \in \cap_{t>1}K_\alpha(t)$ in this topology. It follows that, for each $T > 0$, and for each j,

$$\lim_{k\to\infty}\sup_{t\leq T}\left|\frac{\phi^j_{n(k)}(t)}{1+t} - \frac{\phi^j(t)}{1+t}\right| = 0.$$

Note that this implies, for each t and j,

$$\left|\frac{\phi^j(t)}{1+t}\right| \leq d_\alpha(t),$$

and so $\phi \in D_\alpha$. Now for each $\epsilon > 0$ (sufficiently small), we have for k sufficiently large,

$$\begin{aligned}\|\phi_{n(k)} - \phi\|_u &\leq \sup_{t\leq 1/\epsilon^2}\left|\frac{\phi^j_{n(k)}(t)}{1+t} - \frac{\phi^j(t)}{1+t}\right| + \sup_{t>1/\epsilon^2}\left|\frac{\phi^j_{n(k)}(t)}{1+t} - \frac{\phi^j(t)}{1+t}\right| \\ &\leq \epsilon + 2d_\alpha(1/\epsilon^2) = 3\epsilon.\end{aligned}$$

The set D_α is therefore sequentially compact, and hence compact, in $(\mathcal{Y}, \|\cdot\|_u)$. □

Lemma 2 *If (H1) is satisfied, then*

$$\lim_{\alpha\to\infty}\limsup_n \frac{1}{n}\log P(\tilde{S}_n \notin D_\alpha) = -\infty.$$

Proof. First we have, by the contraction principle,

$$\limsup_n P\bigcup_{t\leq\alpha^2}\{|\tilde{S}_n(t)| > \alpha^2(1+t)\} \leq -\inf_{0<\tau<\alpha^2} \tau R_j(\alpha^2/\tau) \leq -\alpha^2 R_j(1). \tag{11}$$

Here we have used that fact that $\tau\Lambda^*(\alpha^2/\tau)$ and $\tau\Lambda^*(-\alpha^2/\tau)$ are both non-increasing functions of τ, which can be checked using Jensen's inequality. We also have, for each j and some $\theta > 0$,

$$\begin{aligned}P\bigcup_{t>\alpha^2}\{|\tilde{S}^j_n(t)| > (1+t)d_\alpha(t)\} &\leq P\bigcup_{i=0}^{n-1}\bigcup_{k=[\alpha^2]}^{\infty}\{|\tilde{S}^j_n(k+i/n)| > (1+k)d_\alpha(k)\} \\ &\leq n\sum_{k=[\alpha^2]}^{\infty} C(\theta)e^{-\theta nkd_\alpha(k)} \\ &\leq nC(\theta)De^{-\theta n\sqrt{\alpha^2-1}/2}.\end{aligned}$$

Here we have used (H1), Chebyshev's inequality, and the inequality

$$\sum_{k\geq k_0} e^{-\rho\sqrt{k}} \leq De^{-\rho\sqrt{k_0-1}/2}.$$

It follows that

$$\limsup_n \frac{1}{n}\log P\bigcup_{t>\alpha^2}\{|\tilde{S}^j_n(t)| > (1+t)d_\alpha(t)\} \leq -\theta\sqrt{\alpha^2-1}/2. \tag{12}$$

The statement can now be obtained from (10), (11) and (12), via the principle of the largest term. □

This concludes the proof of the theorem. □

4 How to apply the main result

Theorem 1 provides a new tool for looking at large deviations for queueing systems in equilibrium. We will now illustrate how it is used by working through the single-server queue example: suppose $d = 1$ and consider the function f defined by (4). Recall that $f(\tilde{S}_n)$ is equal in distribution to the normalised queue length at a single-server queue. If $\Lambda'(0) = \mu$, say, then a corollary of Theorem 1 is that the LDP holds in the subspace

$$\mathcal{Y}^\mu = \left\{ \phi \in \mathcal{Y} : \lim_{t \to \infty} \frac{\phi(t)}{1+t} = \mu \right\}.$$

If $\mu < 1$, then the restriction of f to $\mathcal{Y}^\mu$ is finite and continuous. To see this, observe that if $\|\phi - \phi'\|_u < \epsilon$, then there exists a t_0, *independent of* ϕ, such that $|f(\phi) - f(\phi')| < 2t_0\epsilon$. We can therefore apply the contraction principle and Jensen's inequality to get that the sequence $Q/n = f(\tilde{S}_n)$ satisfies the LDP in $\mathbb{R}_+$ with rate function given by

$$\begin{aligned} J(q) &= \inf\left\{ \int_0^\infty \Lambda^*(\dot{\phi})ds : \sup_{t>0}[\phi(t) - ct] = q \right\} \\ &= \inf_{\tau>0} \inf\left\{ \int_0^\tau \Lambda^*(\dot{\phi})ds : \phi(\tau) - c\tau = q \right\} \\ &= \inf_{\tau>0} \tau\Lambda^*(c + q/\tau). \end{aligned}$$

This fact has previously been demonstrated by several authors [2, 9, 10, 12], under similar conditions. The *i.i.d.* case is due to Cramér [4] and Borovkov [1]. The advantage of the above approach is that the existence of an LDP is established by continuity which, using the above topology, is quite accessible, and the rate function is easier to compute.

The main result in this paper, combined with the general approach we have discussed, is quite widely applicable. It is ideally suited to problems where reflection mappings exist. It has been used, for example, to obtain comprehensive equilibrium large deviations results for a multiclass FIFO queue [16] and can also be applied to systems with dedicated buffers [15, 17] (the latter corresponds to the random walk in a quadrant, and is the subject of many recent papers).

References

[1] A.A. Borovkov. *Random Processes in Queueing Theory.* Springer-Verlag, 1976.

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

[3] C.-S. Chang and T. Zajic. Effective bandwidths of departure processes from queues with time varying capacities. INFOCOM, 1995.

[4] H. Cramér. On some questions connected with mathematical risk. Univ. Calif. Publications in Statistics, vol. 2, 99–125, 1954.

[5] R.L. Dobrushin and E.A. Pechersky. Large deviations for random processes with independent increments on infinite intervals. Preprint.

[6] Amir Dembo and Tim Zajic. Large deviations: from empirical mean and measure to partial sums process. *Stoch. Proc. Appl.* 57:191–224, 1995.

[7] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications.* Jones and Bartlett, 1993.

[8] Jean-Dominique Deuschel and Daniel W. Stroock. *Large deviations.* Academic Press, 1989.

[9] G. de Veciana, C. Courcoubetis and J. Walrand. Decoupling bandwidths for networks: a decomposition approach to resource management. Memorandum No. UCB/ERL M93/50, University of California, 1993.

[10] N.G. Duffield and Neil O'Connell. Large deviations and overflow probabilities for the general single server queue, with applications. *Proc. Camb. Phil. Soc.* 118(1), 1995.

[11] A. Ganesh and V. Anantharam. Stationary tail probabilities in exponential server tandems with renewal arrivals. To appear in *Queueing Systems.*

[12] Peter W. Glynn and Ward Whitt. Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *J. Appl. Prob.*, to appear.

[13] J. A. Johnson. Banach spaces of Lipschitz functions and vestor valued Lipschitz functional. *Trans. Amer. Math. Soc.* 148: 147–169, 1970.

[14] A.A. Mogulskii. Large deviations for trajectories of multi dimensional random walks. *Th. Prob. Appl.* 21:300–315, 1976.

[15] Neil O'Connell. Queue lengths and departures at single-server resources. To appear in the *Proceedings of the RSS Workshop on Stochastic Networks*, Edinburgh, 1995.

[16] Neil O'Connell. Large deviations for departures from a shared buffer. Revision submitted to *J. Appl. Prob.*

[17] Neil O'Connell. Large deviations for queue lengths at a multi-buffered resource. Revision submitted to *J. Appl. Prob.*

[18] I.F. Pinelis. A problem on large deviations in the space of trajectories. *Th. Prob. Appl.* 26:69–84, 1981.

[19] Kavita Ramanan and Paul Dupuis. Large deviation properties of data streams that share a buffer. Technical Report LCDS 95-8, Division of Applied Mathematics, Brown University.

[20] R.T. Rockafeller. *Convex Analysis.* Princeton University Press, Princeton, New Jersey, 1970.

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

BRIMS, Hewlett-Packard Labs

Filton Road, Stoke Gifford

Bristol BS12 6QZ, UK.