Adv. Appl. Prob. **44,** 995–1017 (2012)
Printed in Northern Ireland
© *Applied Probability Trust* 2012

A STRONG LAW FOR THE RATE OF GROWTH OF LONG LATENCY PERIODS IN A CLOUD COMPUTING SERVICE

SOUVIK GHOSH,* *Columbia University*
SOUMYADIP GHOSH,** *IBM T. J. Watson Research Centre*

Abstract

Cloud-computing shares a common pool of resources across customers at a scale that is orders of magnitude larger than traditional multiuser systems. Constituent physical compute servers are allocated multiple 'virtual machines' (VMs) to serve simultaneously. Each VM user should ideally be unaffected by others' demand. Naturally, this environment produces new challenges for the service providers in meeting customer expectations while extracting an efficient utilization from server resources. We study a new cloud service metric that measures prolonged latency or delay suffered by customers. We model the workload process of a cloud server and analyze the process as the customer population grows. The capacity required to ensure that the average workload does not exceed a threshold over long segments is characterized. This can be used by cloud operators to provide service guarantees on avoiding long durations of latency. As part of the analysis, we provide a uniform large deviation principle for collections of random variables that is of independent interest.

Keywords: Cloud computing; large deviations; long strange segment; latency period; moving average; nonstationary process

2010 Mathematics Subject Classification: Primary 60F10
Secondary 60F15; 60G99

1. Introduction

Cloud computing is a paradigm shift of multiple orders of magnitude in the pursuit of extracting greater utilization of server resources while serving the computing needs of a large collection of customers. This has been made possible primarily by the concept of workload *virtualization* wherein individual users operate on *virtual machines* (VMs), each with modest resource requirements, and multiple VMs are served by a single large computing server. Cloud service providers achieve greater utilization by over-provisioning VMs on computer nodes, acting on the assumption that rarely will multiple customers simultaneously require large quantities of resources.

The resources required over time by a user is a stochastic process, modeled here as a discrete-time moving average (MA) process. We allow for a heterogeneous population of customers, where they are partitioned only by their statistical/stochastic behavior, but are considered equal in terms of priority of service. Service guarantees currently provided by cloud computing providers (Amazon Web Services' EC2 , Google's Web Toolkit, Microsoft's

Received 21 October 2010; revision received 17 February 2012.

* Current address: LinkedIn Corporation, 2029 Stierlin Court, Mountain View, CA 94043, USA.

** Postal address: IBM T. J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, NY 10598, USA.

https://doi.org/10.1239/aap/1354716587 Published online by Cambridge University Press

Azure, etc.) are weak: service level agreements (SLAs) are available only for quick initial provisioning of a new VM from a user onto a compute node, but no guarantees are provided on the quality of service experienced by the customer over time. Large organizations with significant computing requirements, who are willing to pay for good service guarantees, are thus wary of using this architecture for any activity beyond their non-critical desktop usage; see Li *et al.* (2009) and Mendler (2010). This in particular impedes large-scale adoption of cloud computing for time-critical and resource-intensive workloads.

New techniques need to be developed to address the challenge of estimating performance from the user's perspective in this computing paradigm. A key performance indicator in multiuser systems measures the *latency* suffered by users. Latency occurs when access to computing resources is throttled because the total quantity of one or more resource required (CPU cycles, memory space, IO bandwidth, etc.) by all the VMs exceed the server's capacity. Then, under the most commonly used form of processor-sharing discipline, all customers on the server are provisioned proportionately lower resources than they had requested and, thus, are said to experience latency. Applications that are intolerant to latency are discouraged from being put on clouds in the absence of SLAs that penalize their incidence (see Li *et al.* (2009)). Therefore, for a company that wishes to guarantee its customers availability of the server's resources, it is important to understand how large and frequent such long time segments of continued latency can be. We provide a framework to construct such estimates. In particular, we use this framework to estimate the time till the first observation of continued latencies of a given large time length, and its dual, the largest period of latency experienced within a given time. Cloud service operators can utilize this technique to create SLA contracts. In addition, the relationship between the expected first observation time and the per-customer average capacity can help design system improvements to minimize SLA violations. An operator may also provide differentiated service to customers, where those willing to pay for better guarantees can be put on an isolated subcloud with capacity provisioning tailored to their growth, usage, and the agreed-upon SLA contract.

Our framework is built on analyzing *long strange segments* (see definitions (2) and (3)) of the underlying workload process of the cloud server; we refer the reader to Arratia *et al.* (1990) and Ghosh and Samorodnitsky (2010) for a review. Suppose that the server is allocated a capacity that maintains a steady per-customer average C_p above its expected value. Even if C_p is a large number, there will be long time segments during which the average workload of the server will exceed the total capacity. These are long strange segments. Li (2007) studied a similar but simpler metric that captures only the instantaneous latency suffered when instantaneous workload requests exceed capacity, while our goal here is sustained latency suffered over time intervals. A standard technique for analyzing the rate of growth of long strange segments for stationary processes involves an associated large deviation principle (see the discussion at the end of Section 2). While standard probabilistic models (for example, queues) operate on stationary processes, the cloud workload process is nonstationary (see the definition in Section 2). This is because the total number of VMs in the cloud environment increases over time. This is a consequence of the fact that VMs are software artifacts that are inexpensive to instantiate and operate, and so client organizations tend to encourage large-scale adoption and persistent usage of the VMs within their organization. In addition, a major new technological innovation allows fast migration of VMs between individual physical servers within the same cloud infrastructure. Thus, the cloud service environment is better modeled to consist of larger logical servers that each continually grow in capacity in order to serve a continually growing population of users, which yields a nonstationary workload process

https://doi.org/10.1239/aap/1354716587 Published online by Cambridge University Press

The standard large deviation tools that are vital to the analysis of long strange segments of stationary processes are thus not useful for our nonstationary workload process. This process however has a certain structure that can be gainfully exploited. To take advantage of this, we develop a tool for proving a uniform large deviation principle that in its most general form applies to collections of random variables that satisfy certain regulatory conditions (see Theorem 2 in Section 3). This tool, which is of independent interest, plays a crucial role in proving Theorem 1, the main result of this paper, which provides a strong law characterization of the rate of growth of the duration of latency periods as a function of the Fenchel–Legendre transform of the log-moment generation functions of the underlying process. The conditions imposed by the uniform large deviation principle (Theorem 3) are general enough to allow for a broad class of stochastic processes.

Li (2005) found that linear time series models provide accurate predictions of web-server workloads. Khan *et al.* (2012) presented a cloud workload model that augments a linear time series model with a hidden Markov model that changes the time series model parameters at a coarser timescale. Our choice of MA processes helps us provide a clear and simple exposition of the main points. While this does limit the results presented here to linear time series models of workloads, we expect to establish a similar analysis for state-space-based models in our future work.

To summarize, the main contributions of this paper are as follows.

(a) We provide a tool for proving the uniform large deviation principle for a collection of sequences of probability measures. Recall that the Gartner–Eliis theorem is a very helpful device for proving the large deviation principle for a single sequence of probability measures; see Gartner (1977), Ellis (1984), and Dembo and Zeitouni (1998, Theorem 2.3.6, p. 44). We view Theorem 2 as an analogue of the Gartner–Ellis theorem for proving the uniform large deviation principle for a collection of such sequences. The conditions imposed on the random variables restrict the set of admissible probability laws, but are sufficiently flexible to apply to a wide variety of situations.

(b) We provide strong laws characterizing the rate of growth of two performance measures of service under the cloud computing architecture, namely the minimum time taken to observe a continued latency period of a given length, and its dual, the maximum latency period that is observed within a given time.

(c) We show, using a motivating example, how these results can be used by a cloud service manager to (i) create SLA contracts representing a guarantee to the customer against chances of observing frequent long latencies, and (ii) design system improvements to minimize the frequency of long latencies, such as the rates at which new capacity should be procured/allocated to maintain or improve service.

In the following section we describe our model of the cloud environment and state the main result of this paper. We conclude the section with a discussion of a representative example. In Section 3 we state and prove the uniform large deviation principle for collections of random variables. This is used in Section 4 where the main result is proved.

2. Cloud model and main result

We model the workload of each user with respect to the instantaneous requirements for a single resource, e.g. CPU cycles required, over time. A total of K customer groups are served, where groups differ in their workload characterization. The cloud is managed in a manner

https://doi.org/10.1239/aap/1354716587 Published online by Cambridge University Press

that provisions $n_i(t)$ customers from the ith group at time t on each large logical server. The function $n_i(t)$ is assumed to be a power function, i.e. there exist a positive constant α and positive integers $c_1, \dots, c_K$ such that

$$n_i(t) = c_i \lfloor t^\alpha \rfloor \quad \text{for all } i = 1, \dots, K.$$

For any $x \in \mathbb{R}$, $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x and $\lceil x \rceil$ represents the smallest integer greater than or equal to x. The c_i are chosen to be positive integers rather than real numbers. This is solely because of convenience in handling the limit identities which appear below; we are certain that taking $n_i(t) = \lfloor c_i t^\alpha \rfloor$ for some positive real number c_i would not have any significant effect on the results. This form for $n_i(t)$ has two important implication. First, the relative mix of customers from each group, defined by the ratios of the parameters c_i, remains constant over time, and only the total population of users grows with time. Second, the number of customers remains a deterministic function of time. We believe this setting can be easily generalized to allow the number of customers to be a stochastic process, e.g. the case where $(n_1(t), \dots, n_K(t))$ are jointly regularly varying with index α and the number of customers in the ith group is a Poisson process with intensity $n_i(t)$, but we do not foresee this situation adding any extra insights to the studied problem.

The jth customer in the ith group has workload $W_{i,j}(t)$ at discrete time t, i.e.

$$W_{i,j}(t) = \mu_i + X_{i,j}(t) = \mu_i + \beta_i^\top Z(t) + \varepsilon_{i,j}(t) \quad \text{for all } 1 \le i \le K,\ 1 \le j \le n_i(t),\ t \ge 1,$$

where μ_i is a constant denoting the expected workload of customers in the ith group and $X_{i,j}(t)$ is the deviation from the mean workload of the jth customer in the ith group at time t. The zero-mean stochastic process $X_{i,j}(t)$ is further defined as the weighed sum of a K-dimensional MA process $Z(t)$ and additional pure-noise independent and identically distributed (i.i.d.) random variables $\varepsilon_{i,j}(t)$. The weights $\beta_i \in \mathbb{R}^K$ are group-specific constants. The noise process $(\varepsilon_{i,j}(t);\ 1 \le i \le K,\ t \ge 1,\ 1 \le j \le n_i(t))$ consists of i.i.d. random variables, independent of $(Z(t),\ t \in \mathbb{Z})$, with mean 0, satisfying

$$\Lambda_\varepsilon(\lambda) := \log \mathrm{E}[\exp\{\lambda \varepsilon_{i,j}(t)\}] < \infty \quad \text{in a neighborhood of } 0.$$

The process $Z(t)$ is a K-dimensional MA process with *coefficients* $(\phi_t,\ t \in \mathbb{Z})$ and *innovations* $(\xi_t,\ t \in \mathbb{Z})$ defined as

$$Z(t) = \sum_{k=-\infty}^{\infty} \phi_k \xi(t-k) \quad \text{for all } t \in \mathbb{Z},$$

where

(i) the coefficients satisfy $\sum_k |\phi_k| < \infty$ and $\phi := \sum_k \phi_k \neq 0$,

(ii) the innovations $(\xi(t),\ t \in \mathbb{Z})$ are K-dimensional i.i.d. random variables with mean 0, satisfying

$$\Lambda_\xi(\eta) := \log \mathrm{E}[\exp\{\eta \cdot \xi(t)\}] < \infty \quad \text{for all } \eta \in \mathbb{R}^K \tag{1}$$

(for any two vectors x and y, $x \cdot y$ denotes the scalar product).

We will place the following additional restriction on the log-moment generating function $\Lambda_\xi(\cdot)$ to satisfy the conditions of the uniform large deviation principle (Theorem 3 below).

https://doi.org/10.1239/aap/1354716587 Published online by Cambridge University Press

Assumption 1. *It holds that* $|d\Lambda_\xi(\lambda\bar{\beta})/d\lambda| \to \infty$ *whenever* $|\lambda| \to \infty$*, where* $\bar{\beta} := C^{-1}(\sum_{i=1}^{K} c_i\beta_i)$ *with* $C := \sum_{i=1}^{K} c_i$*.*

This mild restriction on the parameters of the MA process is satisfied by realistic computing workloads. For example, it admits a Gaussian form for the innovations ξ (see Example 1 below). We also describe an example where the ξ follow a centered Poisson distribution (see Example 2 below).

The expected workload of the server at time t is given by $\sum_{i=1}^{K} n_i(t)\mu_i$. In our setup the number of customers in each group grows over time and so does the expected workload of the server. The capacity of the server must continually increase to handle this growth and prevent situations where the VMs are perpetually being throttled. Our imperative is to understand the deviations from the mean workload. Define $S(t)$ as the sum of all the deviations until time t, i.e.

$$S(t) := \sum_{k=1}^{t} \sum_{i=1}^{K} \sum_{j=1}^{n_i(k)} X_{i,j}(t) \quad \text{for all } t \geq 1,$$

and $N(t)$ as the associated normalizing term for time t, i.e.

$$N(t) = \sum_{k=1}^{t} \sum_{i=1}^{K} n_i(k) \quad \text{for all } t \geq 1.$$

By convention, we understand that $\sum_{l=i}^{j} x_l = 0$ if $j \leq i$. Furthermore, if i and j are not integers, $\sum_{l=i}^{j} x_l$ will denote $\sum_{l=\lceil i \rceil}^{\lfloor j \rfloor} x_l$.

We study the average deviation of the workload of the server from its mean over long segments of time. For any time segment (k, l), the average deviation is given by

$$\bar{X}(k, l) := \frac{S(l) - S(k)}{N(l) - N(k)}.$$

A simple argument using the law of large numbers tell us that $\bar{X}(k, l)$ should not be too far away from 0 if $l - k$ is large. If $\bar{X}(k, l)$ is not close to 0 then we refer to (k, l) as a *strange segment*. It is also easy to see that if we fix any number L and a threshold ϵ and wait sufficiently long, we will almost surely get a segment (k, l) such that $l - k \geq L$ and $\bar{X}(k, l) > \epsilon$. Our main result describes how the length of these strange segments grows over time.

For any measurable set A, we define the *long strange segments* as

$$R_t(A) := \sup\{m \colon \bar{X}(l - m, l) \in A \text{ for some } l = m, \dots, t\}, \tag{2}$$

and its dual characteristic as

$$T_r(A) := \inf\{l \colon \text{there exists } k,\ 0 \leq k \leq l - r, \text{ such that } \bar{X}(k, l) \in A\}. \tag{3}$$

The functional $R_n(A)$ is the maximum length of a segment from the first n observations, whose average is in set A, and $T_n(A)$ is the minimum number of observations required to have a segment of length at least n, whose average is in the set A. It is easy to see that $R_t(A)$ grows as $t \to \infty$ and $T_r(A)$ grows as $r \to \infty$. Theorem 1 below describes the rate of growth of these functionals. There is a duality relation between the rate of growth of these functionals which follows from the fact that $\{T_r(A) \leq m\} = \{R_m(A) \geq r\}$.

https://doi.org/10.1239/aap/1354716587 Published online by Cambridge University Press

For any convex function $f(\cdot)$, we will use $f^*(\cdot)$ to denote its Fenchel–Legendre transform:

$$f^*(x) := \sup_{\lambda\in\mathbb{R}}\{\lambda x - f(\lambda)\}.$$

For any set $A \subset \mathbb{R}$, A° and $\bar{A}$ will represent the interior and closure of A, respectively.

Theorem 1. *For any measurable set* A,

$$I_* \le \liminf_{r\to\infty} \frac{\log T_r(A)}{r} \le \limsup_{r\to\infty} \frac{\log T_r(A)}{r} \le I^* \quad \textit{almost surely (a.s.)} \tag{4}$$

and

$$\frac{1}{I^*} \le \liminf_{t\to\infty} \frac{R_t(A)}{\log t} \le \limsup_{t\to\infty} \frac{R_t(A)}{\log t} \le \frac{1}{I_*} \quad \textit{a.s.}, \tag{5}$$

where

$$I_* = \inf_{x\in\bar{A}} \Lambda^*(x), \qquad I^* = \inf_{x\in A^\circ} \Lambda^*(x),$$

and $\Lambda^*(x)$ *is the Fenchel–Legendre transform of* $\Lambda(\lambda) := \Lambda_\xi(\lambda\phi\bar{\beta})$.

Remark 1. Under our assumption that the customer group compositions remain constant, the customer groups are all jointly represented by their average $\bar{\beta} = (\sum_{i=1}^K c_i\beta_i)/\sum_{i=1}^K c_i$.

Remark 2. For the cloud server case, we are specifically interested in sets of the nature $A = (C_p, \infty)$. The parameter C_p represents the capacity of a cloud server, which is designed such that the per-customer capacity of the server is maintained at C_p units above the per-customer expected instantaneous workload $\bar{\mu} := C^{-1}(\sum_{i=1}^K c_i\mu_i)$ (recall that the deviations $X_{i,j}$ have zero mean). Our model assumes that the total number of customers grows over time, implying that the aggregate capacity of the server is also continually increased such that the average capacity per customer is always maintained at C_p units above the expected workload per customer at any time point. Thus, at time t, the server processes workload at a rate equal to $(\bar{\mu} + C_p)C\lfloor t^\alpha\rfloor$.

The workload imposed on the server by the VMs exceeds its capacity for precisely those time segments (k, l) over which the average of the excess workload process $\bar{X}(k, l)$ is in the set (C_p, ∞). This is true for any *work-conserving* service policy, i.e. any policy that requires the server to continually serve any buffered workload. Thus, the parameter C_p of the server encapsulates all the underlying queueing dynamics.

We assume that the parameter C_p is set a value greater than 0, i.e. the cloud servers are designed to be able to handle at least the mean workload. Then, the continuity and increasing nature of the Fenchel–Legendre transform over A ensures that the infimum over the sets $\bar{A}$ and A° are achieved at C_p. Thus, the upper and lower bounds in (4) and (5) collapse to give a limit result of the form:

$$\lim_{r\to\infty} \frac{\log T_r(A)}{r} = \lim_{t\to\infty} \frac{\log t}{R_t(A)} = \Lambda^*(C_p) \quad \text{a.s.}$$

Another interesting application is to check if there are long time periods when the server resources are being severely underutilized. With $A = (-\infty, -C_b)$, where $C_b > 0$, we can measure how long the server's per-customer capacity is underutilized by $C_p + C_b$ units.

https://doi.org/10.1239/aap/1354716587 Published online by Cambridge University Press

Remark 3. We discuss the case when the growth in the number of customers is a power law. Reid *et al.* (2011) forecast that cloud service providers will see their growth curve settle down in to a power-law-like growth phase; indeed, for some types of cloud service they posit that this will happen by 2012. Theorem 2 can easily be applied in other settings, for example, when the rate of growth is much slower, say logarithmic. However, if the rate of growth is exponential then the behavior of T_r is totally different. In this situation the averages $\bar{X}(k,l)$ converge weakly to a nondeterministic random variable and T_r behaves more like the maximum of a stationary sequence of random variables. We do not treat this situation in detail here.

Example 1. Suppose that the innovation vectors $\xi(t)$ are i.i.d. replicates of a K-dimensional joint-normal random vector with mean 0 and covariance matrix Σ. In that case $\Lambda(\lambda) = \lambda^2\phi^2\bar{\beta}^\top\Sigma\bar{\beta}/2$ and, hence,

$$\Lambda^*(x) = (2\phi^2\bar{\beta}^\top\Sigma\bar{\beta})^{-1}x^2 \quad \text{for all } x \in \mathbb{R}.$$

Therefore, if $A = (C_p, \infty)$ then

$$\lim_{r\to\infty} \frac{\log T_r(A)}{r} = \lim_{t\to\infty} \frac{\log t}{R_t(A)} = (2\phi^2\bar{\beta}^\top\Sigma\bar{\beta})^{-1}C_p^2 \quad \text{a.s.}$$

This yields the estimates $T_r \sim \exp\{rC_p^2/M\}$ and $R_t \sim M\log t/C_p^2$, where C_p represents the server's capacity and $M = 2\phi^2\bar{\beta}^\top\Sigma\bar{\beta}$ is a property of the customer classes. As expected, the duration T_r till one observes a latency period of length r grows with the server capacity C_p. On the other hand, higher variability of the innovation $\xi(t)$ results in a higher value of M and culminates in a faster growth of the long latency periods R_t observed in time t.

Example 2. If the innovations $\xi(t) = (\xi_1(t), \dots, \xi_K(t))$ are such that its components are independent 'centered' Poisson distributed with parameters $(\mu_1, \dots, \mu_K)$ then, for $\lambda = (\lambda_1, \dots, \lambda_K)$,

$$\Lambda(\lambda) = \sum_{i=1}^K \mu_i(\mathrm{e}^{\lambda_i}\phi\bar{\beta}_i - \lambda_i\phi\bar{\beta}_i - 1)$$

and, for $A = (C_p, \infty)$,

$$\begin{aligned}\lim_{r\to\infty} \frac{\log T_r(A)}{r} &= \lim_{t\to\infty} \frac{\log t}{R_t(A)} \\ &= \Lambda^*(C_p) \\ &= \sum_{i=1}^K \biggl(\frac{C_p}{\phi\bar{\beta}_i}\log\biggl(\frac{C_p}{\mu_i\phi\bar{\beta}_i}+1\biggr) - \mu_i\biggl(\frac{C_p}{\mu_i\phi\bar{\beta}_i} - \log\biggl(\frac{C_p}{\mu_i\phi\bar{\beta}_i}+1\biggr)\biggr)\biggr).\end{aligned}$$

Let $C_p^i = C_p/(\mu_i\phi\bar{\beta}_i)$. We obtain estimates $T_r \sim \exp\{r\sum_1^K \mu_i((1+C_p^i)\log(1+C_p^i) - C_p^i)\}$ and, similarly, $R_t \sim \log t/\sum_1^K \mu_i((1+C_p^i)\log(1+C_p^i) - C_p^i)$. In this case too the duration T_r till one observes a latency period of length r grows with the server capacity C_p. But, unlike the joint-normal innovations case above where $\log T_r$ grew as a quadratic of C_p, the growth of $\log T_r$ here, while superlinear, is slower.

We postpone the proof of Theorem 1 till Section 4, and develop the proper tools required for the proof in Section 3. We close this section with a discussion on why standard large deviation tools are inadequate for the proof of Theorem 1.

https://doi.org/10.1239/aap/1354716587 Published online by Cambridge University Press

The rate of growth of long strange segments have been studied in Mansfield *et al.* (2001) for MA processes with heavy-tailed innovations and then in Rachev and Samorodnitsky (2001) for long-range-dependent MA processes with heavy-tailed innovations. Recently, Ghosh and Samorodnitsky (2010) studied the effect of memory on the rate of growth of long strange segments for a MA process with light-tailed innovations. A strong law of the form (5) is often referred to as the Erdös–Rényi law of large numbers; Erdös and Rényi (1970) proved asymptotics for longest head runs in i.i.d. coin tosses.

It is instructive to take a heuristic look at the standard technique of proving the rate of growth of long strange segments for a stationary process, say (Y_t). A vital tool for analyzing this growth is a large deviation principle associated with the partial sums of (Y_t). Recall that a sequence of probability measures $(\mathrm{P}_t,\ t \geq 1)$ satisfies a *large deviation principle* (LDP) on $\mathbb{R}$ if there exists a nonnegative lower-semicontinuous function $I(\cdot)$ such that, for any measurable $A \subset \mathbb{R}$,

$$-\inf_{x\in A^\circ} I(x) \leq \liminf_{t\to\infty} \frac{1}{t}\log \mathrm{P}_t(A) \leq \limsup_{t\to\infty} \frac{1}{t}\log \mathrm{P}_t(A) \leq -\inf_{x\in\bar{A}} I(x). \tag{6}$$

The function $I(\cdot)$ is called the *rate function*. A rate function with compact level sets is called a *good* rate function.

Denote the average of the segment (k, l) by

$$\bar{Y}(k,l) = \frac{\sum_{i=k+1}^{l} Y_i}{l-k}.$$

It is often possible to show that the law of $\bar{Y}(0,t)$ satisfies an LDP under assumptions of mixing or other specific structure on (Y_t) and the existence of exponential moments of Y_t; see, for example, Bryc and Dembo (1996), Varadhan (1984), Dembo and Zeitouni (1998), and Deuschel and Stroock (1989). Then, for a 'nice' set A such that $\mathrm{E}(Y_0) \notin \bar{A}$, there exists $I > 0$ such that, for large t,

$$\log \mathrm{P}[\bar{Y}(0,t) \in A] \sim -It.$$

Using stationarity, this implies that $\log \mathrm{P}[\bar{Y}(l, l+t) \in A] \sim -It$ for every $l \geq 0$. Heuristically, this means that, for approximately e^{tI} segments of length t, we can expect to find one with an average in A. The segments $(0,t), (1,t+1), (2,t+2), \ldots$ are not independent, but that lack of independence is typically handled using mixing-type conditions borrowed from the process (Y_t) itself. Theorem 2.3 of Ghosh and Samorodnitsky (2010) is an example of this line of argument where the authors considered MA processes and used the LDP for partial sums proved in Ghosh and Samorodnitsky (2009) to obtain asymptotic results for the rate of growth of long strange segments.

In our application's setting, the distribution of $\bar{X}(l, l+t)$ differs from that of $\bar{X}(0,t)$ when $l > 0$. This is because the growing number of customers in the system implies that each $\bar{X}(l, l+t)$ represents an average over a different number of realizations $(N(t+l) - N(l)$ versus $N(t))$. So, in order to understand the rate of growth of the long strange segments, we need to estimate the probability $\mathrm{P}[\bar{X}(l, l+t) \in A]$ uniformly over $l \geq 0$. We address this problem by proving the uniform large deviation principle in Theorem 3. A collection of probability measures $(\mathrm{P}_{k,t},\ t \geq 1, k \in \Gamma)$ satisfies an LDP on $\mathbb{R}$ *uniformly over* $k \in \Gamma$ if there exist nonnegative lower-semicontinuous functions $(I_k(\cdot),\ k \in \Gamma)$ such that, for any measurable $A \subset \mathbb{R}$,

$$\liminf_{t\to\infty} \inf_{k\in\Gamma} \left\{ \frac{1}{t}\log \mathrm{P}_{k,t}(A) + \inf_{x\in A^\circ} I_k(x) \right\} \geq 0 \tag{7}$$

https://doi.org/10.1239/aap/1354716587 Published online by Cambridge University Press

and

$$\limsup_{t\to\infty}\sup_{k\in\Gamma}\left\{\frac{1}{t}\log \mathrm{P}_{k,t}(A)+\inf_{x\in\bar{A}} I_k(x)\right\}\leq 0. \tag{8}$$

Note that bounds (7) and (8) are generalizations of the left- and right-hand sides of the standard large deviation bounds in (6)

3. Uniform large deviation principle

The Gartner–Ellis theorem is an important tool for proving the LDP; cf. Gartner (1977), Ellis (1984), and Dembo and Zeitouni (1998, Theorem 2.3.6, p. 44). Theorem 2 is an analog of the Gartner–Ellis theorem for proving the uniform LDP. We use this theorem to prove the uniform LDP for the average of segments of the server workload process in Theorem 3, which is in fact the first step in proving Theorem 1.

Theorem 2. *Suppose that* $(Y_{k,t},\ t\geq 1,\ k\in\Gamma)$ *is a collection of random variables such that there exists* $(\Lambda^k(\cdot),\ k\in\Gamma)$ *which are differentiable and satisfy the following conditions: for all* $0<L<\infty$ *and* $\epsilon>0$, *there exists* $T>0$ *and* $\delta>0$ *such that*

$$\lim_{t\to\infty}\sup_{k\in\Gamma,\ |\lambda|\leq L}\left|\Lambda^k(\lambda)-\frac{1}{t}\log \mathrm{E}[\exp\{t\lambda Y_{k,t}\}]\right|=0, \tag{9}$$

$$\sup_{k\in\Gamma,\ t\geq T,\ |\lambda|\leq L}\left|\frac{1}{t}\log \mathrm{E}[\exp\{t\lambda Y_{k,t}]\right|<\infty, \tag{10}$$

$$\inf_{k\in\Gamma}|(\Lambda^k)'(\lambda)|\to\infty \quad \textit{whenever } |\lambda|\to\infty, \tag{11}$$

and

$$|(\Lambda^k)'(\lambda_1)-(\Lambda^k)'(\lambda_2)|<\epsilon \quad \textit{for all } |\lambda_1-\lambda_2|<\delta,\ \lambda_1,\lambda_2\in[-L,L],\ k\in\Gamma. \tag{12}$$

Then, for any closed set $F\subset\mathbb{R}$,

$$\limsup_{t\to\infty}\sup_{k\in\Gamma}\left\{\frac{1}{t}\log \mathrm{P}[Y_{k,t}\in F]+\inf_{x\in F}\Lambda^{k*}(x)\right\}\leq 0 \tag{13}$$

and, for any open set $G\subset\mathbb{R}$,

$$\liminf_{t\to\infty}\inf_{k\in\Gamma}\left\{\frac{1}{t}\log \mathrm{P}[Y_{k,t}\in G]+\inf_{x\in G}\Lambda^{k*}(x)\right\}\geq 0, \tag{14}$$

where the rate function $\Lambda^{k*}(\cdot)$ *is the Fenchel–Legendre transform of* $\Lambda^k(\cdot)$.

Remark 4. It can be observed from the proof below that conditions (9), (10), and (11) have been used to prove (13), whereas, all the conditions (9)–(12) are required to prove (14). Condition (9) requires that the normalized log-moment generating functions of $Y_{k,t}$ converge to $\Lambda^k(\lambda)$ uniformly over $k\in\Gamma$ and locally uniformly in $\lambda\in\mathbb{R}$. Condition (10) ensures uniform exponential tightness of the random variables $(Y_{k,t})$. Condition (11) is the equivalent of the steepness assumption imposed by the Gartner–Ellis theorem; cf. Dembo and Zeitouni (1998, Theorem 2.3.6, p. 44). Condition (12) requires that the functions $(\Lambda^k)'(\lambda)$ are continuous in λ, uniformly over $k\in\Gamma$ and λ in a compact subset of $\mathbb{R}$. This ensures that the Fenchel–Legendre transforms $\Lambda^{k*}(x)$ are continuous in x, uniformly over $k\in\Gamma$ and x in compact subsets of $\mathbb{R}$.

https://doi.org/10.1239/aap/1354716587 Published online by Cambridge University Press

Proof of Theorem 2. We will first prove (13). As (13) holds trivially when $F = \varnothing$, we can safely assume that F is nonempty. We begin by supposing that F is compact. Fix any $x \in F$ and $\delta > 0$. Since $\Lambda^k(\cdot)$ is convex, continuously differentiable, and satisfies (11), we can find $\lambda_x^k \in \mathbb{R}$ such that

$$(\Lambda^k)'(\lambda_x^k) = x.$$

This would imply that

$$\Lambda^{k*}(x) = \sup_{\lambda \in \mathbb{R}} \{\lambda x - \Lambda^k(\lambda)\} = \lambda_x^k x - \Lambda^k(\lambda_x^k).$$

From (11) we also know that $\{\lambda_x^k : k \in \Gamma\}$ is a bounded set. Hence, we can find an open neighborhood A_x of x such that

$$\inf_{y \in A_x} \lambda_x^k (y - x) \geq -\delta \qquad \text{for all } k \in \Gamma.$$

Then by Chebychev's inequality we obtain an upper bound for the probability

$$\mathrm{P}[Y_{k,t} \in A_x] \leq \mathrm{E}[\exp\{\lambda_x^k t (Y_{k,t} - x)\}] \exp\Big\{-t \inf_{y \in A_x} \lambda_x^k (y - x)\Big\},$$

which implies that

$$\frac{1}{t} \log \mathrm{P}[Y_{k,t} \in A_x] \leq \frac{1}{t} \log \mathrm{E}[\exp\{\lambda_x^k t Y_{k,t}\}] - \lambda_x^k x + \delta.$$

From (9) we can obtain $T \geq 1$ such that, for all $t \geq T$ and $k \in \Gamma$,

$$\frac{1}{t} \log \mathrm{E}[\exp\{\lambda_x^k t Y_{k,t}\}] \leq \Lambda^k(\lambda_x^k) + \delta,$$

and this means that, for $t \geq T$,

$$\frac{1}{t} \log \mathrm{P}[Y_{k,t} \in A_x] \leq \Lambda^k(\lambda_x^k) - \lambda_x^k x + 2\delta = -\Lambda^{k*}(x) + 2\delta.$$

Now, obviously, $\bigcup_{x \in F} A_x$ is an open cover of F and since F is compact, we can obtain $x_1, \dots, x_N \in F$ such that $F \subset \bigcup_{1 \leq i \leq N} A_{x_i}$. Then by a simple union-of-events bound we obtain, for $t \geq T$,

$$\frac{1}{t} \log \mathrm{P}[Y_{k,t} \in F] + \min_{1 \leq i \leq N} \Lambda^{k*}(x_i) \leq \frac{1}{t} \log N + 2\delta \qquad \text{for all } k \in \Gamma.$$

It is now easy to see that, for $t \geq T$,

$$\sup_{k \in \Gamma} \bigg\{ \frac{1}{t} \log \mathrm{P}[Y_{k,t} \in F] + \inf_{x \in F} \Lambda^{k*}(x) \bigg\} \leq \frac{1}{t} \log N + 2\delta,$$

and since $0 < \delta < 1$ is arbitrary,

$$\limsup_{t \to \infty} \sup_{k \in \Gamma} \bigg\{ \frac{1}{t} \log \mathrm{P}[Y_{k,t} \in F] + \inf_{x \in F} \Lambda^{k*}(x) \bigg\} \leq 0. \tag{15}$$

This proves (13) when F is compact.

https://doi.org/10.1239/aap/1354716587 Published online by Cambridge University Press

Next we extend the above result to any nonempty closed set F. First we note a few facts. Using (9) and (10), we obtain, for any $\delta > 0$,

$$c := \sup_{k\in\Gamma,\, |\lambda|<\delta} |\Lambda^k(\lambda)| < \infty.$$

Since $\{\lambda_x^k : k \in \Gamma\}$ is bounded and $\Lambda^{k*}(x) = \lambda_x^k x - \Lambda^k(\lambda_x^k)$, we obtain $\sup_{k\in\Gamma} \Lambda^{k*}(x) < \infty$. Furthermore, for all $k \in \Gamma$,

$$\Lambda^{k*}(x) = \sup_{\lambda\in\mathbb{R}}\{\lambda x - \Lambda^k(\lambda)\} \geq \sup_{|\lambda|<\delta}\{\lambda x - \Lambda^k(\lambda)\} \geq \delta|x| - c.$$

Hence, for any closed set F, there exists $M_1 > 0$ such that

$$\inf_{x\in F} \Lambda^{k*}(x) = \inf_{x\in F\cap[-M_1,M_1]} \Lambda^{k*}(x) \quad \text{for all } k \in \Gamma. \tag{16}$$

Also, note that, for any $k \in \Gamma$ and $t \geq 1$,

$$\frac{1}{t}\log \mathrm{P}[|Y_{k,t}| > \theta] \leq -\theta + \sup_{k\in\Gamma,\, t\geq 1} \frac{1}{t}\log \mathrm{E}[\mathrm{e}^{tY_{k,t}}] + \sup_{k\in\Gamma,\, t\geq 1} \frac{1}{t}\log \mathrm{E}[\mathrm{e}^{-tY_{k,t}}]$$

and, therefore,

$$\lim_{\theta\to\infty} \limsup_{t\to\infty} \sup_{k\in\Gamma} \frac{1}{t}\log \mathrm{P}[|Y_{k,t}| > \theta] = -\infty. \tag{17}$$

Now set

$$c' = \sup_{k\in\Gamma} \inf_{x\in F} \Lambda^{k*}(x).$$

Since, for any x, $\sup_{k\in\Gamma} \Lambda^{k*}(x) < \infty$, we obtain $c' < \infty$. Note that if $c' = 0$ then the proof is immediate. So we examine the case when $c' > 0$. Using (17), we can obtain $M_2 > 0$ such that

$$\mathrm{P}[|Y_{k,t}| > M_2] \leq \mathrm{e}^{-2c't} \quad \text{for all } k \in \Gamma,\ t \geq 1.$$

Let $M = \max\{M_1, M_2\}$. Note that, from (15) and (16),

$$\begin{aligned}
&\limsup_{t\to\infty} \sup_{k\in\Gamma}\left\{\frac{1}{t}\log \mathrm{P}[Y_{k,t} \in F\cap[-M,M]] + \inf_{x\in F}\Lambda^{k*}(x)\right\}\\
&\quad = \limsup_{t\to\infty} \sup_{k\in\Gamma}\left\{\frac{1}{t}\log \mathrm{P}[Y_{k,t} \in F\cap[-M,M]] + \inf_{x\in F\cap[-M,M]}\Lambda^{k*}(x)\right\}\\
&\quad \leq 0.
\end{aligned}$$

This means that, for any given $\delta > 0$, we can find $T \geq 1$ such that

$$\frac{1}{t}\log \mathrm{P}[Y_{k,t} \in F\cap[-M,M]] + \inf_{x\in F}\Lambda^{k*}(x) \leq \delta \quad \text{for all } k \in \Gamma,\ t \geq T.$$

Now if $\mathrm{P}[Y_{k,t} \in F\cap[-M,M]] \leq \mathrm{P}[|Y_{k,t}| > M]$ then

$$\frac{1}{t}\log \mathrm{P}[Y_{k,t} \in F] \leq \frac{1}{t}\log 2 - 2c'.$$

https://doi.org/10.1239/aap/1354716587 Published online by Cambridge University Press

Otherwise,

$$\frac{1}{t}\log \mathrm{P}[Y_{k,t} \in F] \le \frac{1}{t}\log 2 + \frac{1}{t}\log \mathrm{P}[Y_{k,t} \in F \cap [-M, M]].$$

Therefore, in both cases,

$$\frac{1}{t}\log \mathrm{P}[Y_{k,t} \in F] + \inf_{x\in F} \Lambda^{k*}(x) \le \frac{1}{t}\log 2 + \delta \qquad \text{for all } k \in \Gamma,\ t \ge T,$$

and, hence,

$$\limsup_{t\to\infty}\sup_{k\in\Gamma}\biggl\{\frac{1}{t}\log \mathrm{P}[Y_{k,t} \in F] + \inf_{x\in F} \Lambda^{k*}(x)\biggr\} \le 0.$$

This completes the proof of (13).

We will now prove (14). Note that we can find $M > 0$ such that

$$\inf_{x\in G} \Lambda^{k*}(x) = \inf_{x\in G\cap[-M,M]} \Lambda^{k*}(x) \qquad \text{for all } k \in \Gamma.$$

Fix any $\epsilon > 0$, and obtain $x^k \in G \cap [-M, M]$ such that

$$\Lambda^{k*}(x^k) < \inf_{x\in G} \Lambda^{k*}(x) + \tfrac{1}{2}\epsilon.$$

Another observation that we need to make is that we can find $\delta > 0$ such that

$$|\Lambda^{k*}(x) - \Lambda^{k*}(y)| < \tfrac{1}{2}\epsilon \qquad \text{for all } |x - y| < \delta,\ x, y \in [-M, M],\ k \in \Gamma.$$

This follows easily from (12). Now, obviously, $\bigcup_{x\in G\cap[-M,M]} B_{x,\delta}$ is an open cover of $G \cap [-M, M]$, where $B_{x,\delta} = (x - \delta, x + \delta)$. Since $G \cap [-M, M]$ is precompact, we can find $x_1, \dots, x_n \in G \cap [-M, M]$ such that, for all x^k, there exists $1 \le i_k \le n$ for which $|x^k - x_{i_k}| < \delta$. This implies that

$$\inf_{1\le i\le n} \Lambda^{k*}(x_i) < \inf_{x\in G} \Lambda^{k*}(x) + \epsilon \qquad \text{for all } k \in \Gamma.$$

For notational simplicity, we define $X = \{x_1, \dots, x_n\}$. Let $\delta' > 0$ be such that $B_{x,\delta'} \subset G$ for all $x \in X$. Now fix any $x \in X$. Define the random variables $\tilde{Y}_{k,t}$ by an exponential change of measure such that

$$\mathrm{P}[\tilde{Y}_{k,t} \in B] = \frac{\mathrm{E}[\exp\{t\lambda^k_x Y_{k,t}\}\,\mathbf{1}_{[Y_{k,t}\in B]}]}{\mathrm{E}[\exp\{t\lambda^k_x Y_{k,t}\}]}.$$

Then

$$\mathrm{P}[Y_{k,t} \in B_{x,\delta'}] = \mathrm{E}[\exp\{t\lambda^k_x Y_{k,t}\}]\,\mathrm{E}[\exp\{-t\lambda^k_x \tilde{Y}_{k,t}\}\,\mathbf{1}_{[\tilde{Y}_{k,t}\in B_{x,\delta'}]}]$$

and

$$\begin{aligned}\frac{1}{t}\log \mathrm{P}[Y_{k,t} \in B_{x,\delta'}] &= \frac{1}{t}\log \mathrm{E}[\exp\{t\lambda^k_x Y_{k,t}\}] + \frac{1}{t}\log \mathrm{E}[\exp\{-t\lambda^k_x \tilde{Y}_{k,t}\}\,\mathbf{1}_{[\tilde{Y}_{k,t}\in B_{x,\delta'}]}]\\ &\ge \frac{1}{t}\log \mathrm{E}[\exp\{t\lambda^k_x Y_{k,t}\}] - \lambda^k_x x - |\lambda^k_x|\delta' + \frac{1}{t}\log \mathrm{P}[\tilde{Y}_{k,t} \in B_{x,\delta'}].\end{aligned}$$

We claim that

$$\lim_{t\to\infty}\inf_{k\in\Gamma,x\in X}\frac{1}{t}\log \mathrm{P}[\tilde{Y}_{k,t} \in B_{x,\delta'}] = 0. \qquad (18)$$

https://doi.org/10.1239/aap/1354716587 Published online by Cambridge University Press

We complete the proof of (14) assuming (18), which we prove at the end. Let $M' > 0$ be such that $|(\Lambda^k)'(\lambda)| > M$ for all $|\lambda| > M'$ and $k \in \Gamma$. From assumption (11) we know that $M' < \infty$. We can also obtain $T \geq 1$ such that, for all $t \geq T$ and $x \in X$,

$$\inf_{k\in\Gamma} \frac{1}{t} \log \mathrm{P}[\tilde{Y}_{k,t} \in B_{x,\delta'}] \geq -\epsilon$$

and

$$\sup_{k\in\Gamma} \left| \Lambda^k(\lambda_x^k) - \frac{1}{t} \log \mathrm{E}[\exp\{t\lambda_x^k Y_{t,k}\}] \right| < \epsilon.$$

This implies that, for all $t \geq T$, $x \in X$, and $k \in \Gamma$,

$$\begin{aligned} \frac{1}{t} \log \mathrm{P}[Y_{k,t} \in G] &\geq \frac{1}{t} \log \mathrm{P}[Y_{k,t} \in B_{x,\delta'}] \\ &\geq \Lambda^k(\lambda_x^k) - \lambda_x^k x - M'\delta' - 2\epsilon \\ &= -\Lambda^{k*}(x) - M'\delta' - 2\epsilon. \end{aligned}$$

Since $x \in X$ is arbitrary and M', δ', and ϵ are independent of the choice of x, we obtain, for all $t \geq T$ and $k \in \Gamma$,

$$\frac{1}{t} \log \mathrm{P}[Y_{k,t} \in G] \geq -\inf_{x\in X} \Lambda^{k*}(x) - M'\delta' - 2\epsilon \geq -\inf_{x\in G} \Lambda^{k*}(x) - M'\delta' - 3\epsilon.$$

Hence, we obtain

$$\liminf_{t\to\infty} \sup_{k\in\Gamma} \left\{ \frac{1}{t} \log \mathrm{P}[Y_{k,t} \in G] + \inf_{x\in G} \Lambda^{k*}(x) \right\} \geq -M'\delta' - 3\epsilon.$$

This completes the proof of (14) since δ' and ϵ can be chosen arbitrarily close to 0.

It now remains to prove (18). Since X is a finite set, it suffices to show that, for any $x \in X$,

$$\lim_{t\to\infty} \inf_{k\in\Gamma} \frac{1}{t} \log \mathrm{P}[\tilde{Y}_{k,t} \in B_{x,\delta'}] = 0.$$

We will use the upper large deviation bound (13) for this purpose. Note that

$$\begin{aligned} \frac{1}{t} \log \mathrm{E}[\exp\{t\lambda\tilde{Y}_{k,t}\}] &= \frac{1}{t} \log \mathrm{E}[\exp\{t(\lambda + \lambda_x^k)Y_{k,t}\}] - \frac{1}{t} \log \mathrm{E}[\exp\{t\lambda_x^k Y_{k,t}\}] \\ &\to \tilde{\Lambda}^k(\lambda) \\ &:= \Lambda^k(\lambda + \lambda_x^k) - \Lambda^k(\lambda_x^k). \end{aligned}$$

It is easy to check that $\tilde{\Lambda}^k(\cdot)$ inherits properties (9), (10), (11), and (12) from $\Lambda^k(\cdot)$. Therefore, since $B_{x,\delta'}^c := \{x \in \mathbb{R} \colon x \notin B_{x,\delta'}\}$ is a closed set, by (13),

$$\limsup_{t\to\infty} \sup_{k\in\Gamma} \left\{ \frac{1}{t} \log \mathrm{P}[\tilde{Y}_{k,t} \in B_{x,\delta'}^{\mathrm{c}}] + \inf_{y\in B_{x,\delta'}^{\mathrm{c}}} \tilde{\Lambda}^{k*}(y) \right\} \geq 0. \tag{19}$$

Note that $(\tilde{\Lambda}^k)'(0) = x$ for all $k \in \Gamma$, which implies that $\tilde{\Lambda}^{k*}(x) = 0$ for all $k \in \Gamma$. Since $\tilde{\Lambda}^{k*}(\cdot)$ is nonnegative and convex, $\inf_{y\in B_{x,\delta'}^{\mathrm{c}}} \tilde{\Lambda}^{k*}(y) \geq \min\{\tilde{\Lambda}^{k*}(x-\delta'), \tilde{\Lambda}^{k*}(x+\delta')\}$. Now obtain a compact set K' such that $|(\tilde{\Lambda}^k)'(\lambda)| > |x| + \delta'$ and then find $\eta > 0$ such that

$$|(\tilde{\Lambda}^k)'(\lambda') - (\tilde{\Lambda}^k)'(\lambda'')| < \tfrac{1}{2}\delta' \quad \text{for all } |\lambda' - \lambda''| < \eta,\ \lambda', \lambda'' \in K',\ k \in \Gamma. \tag{20}$$

https://doi.org/10.1239/aap/1354716587 Published online by Cambridge University Press

Then obtain $\tilde{\lambda}^k_{x+}$ and $\tilde{\lambda}^k_{x-}$ such that $(\tilde{\Lambda}^k)'(\tilde{\lambda}^k_{x+}) = x + \delta'$ and $(\tilde{\Lambda}^k)'(\tilde{\lambda}^k_{x-}) = x - \delta'$. From (20) we know that $\tilde{\lambda}^k_{x+} > \eta$ and $\tilde{\lambda}^k_{x-} < -\eta$ for all $k \in \Gamma$. Therefore, for all $k \in \Gamma$,

$$\begin{aligned}
\tilde{\Lambda}^{k*}(x + \delta') &= \tilde{\lambda}^k_{x+}(x + \delta') - \tilde{\Lambda}^k(\tilde{\lambda}^k_{x+}) \\
&= \tilde{\lambda}^k_{x+}(x + \delta') - \int_0^{\tilde{\lambda}^k_{x+}} (\tilde{\Lambda}^k)'(z)\,\mathrm{d}z \\
&\geq \tilde{\lambda}^k_{x+}(x + \delta') - \big(x + \tfrac{1}{2}\delta'\big)\eta - (\tilde{\lambda}^k_{x+} - \eta)(x + \delta') \\
&= \tfrac{1}{2}\eta\delta'
\end{aligned}$$

and

$$\begin{aligned}
\tilde{\Lambda}^{k*}(x - \delta') &= \tilde{\lambda}^k_{x-}(x - \delta') - \tilde{\Lambda}^k(\tilde{\lambda}^k_{x-}) \\
&= \tilde{\lambda}^k_{x+}(x + \delta') + \int_{\tilde{\lambda}^k_{x-}}^0 (\tilde{\Lambda}^k)'(z)\,\mathrm{d}z \\
&\geq \tilde{\lambda}^k_{x-}(x - \delta') + \big(x - \tfrac{1}{2}\delta'\big)\eta + (\tilde{\lambda}^k_{x+} - \eta)(x - \delta') \\
&= \tfrac{1}{2}\eta\delta'.
\end{aligned}$$

This implies that $\min\{\tilde{\Lambda}^{k*}(x - \delta'), \tilde{\Lambda}^{k*}(x + \delta')\} \geq \eta\delta'/2$ for all $k \in \Gamma$ and, hence, using (19), we obtain

$$\limsup_{t\to\infty} \sup_{k\in\Gamma} \frac{1}{t} \log \mathrm{P}[\tilde{Y}_{k,t} \in B^{\mathrm{c}}_{x,\delta'}] \leq -\frac{\eta\delta'}{2}.$$

This also means that

$$\lim_{t\to\infty} \inf_{k\in\Gamma} \mathrm{P}[\tilde{Y}_{k,t} \in B_{x,\delta'}] = 1.$$

This proves (18) and, hence, completes the proof of the theorem.

Theorem 3 below allows us to approximate the probability of deviation from 0 of the average $\bar{X}(k, l)$ for different segments (k, l) when $l - k$ is large. This is a vital component in the proof of Theorem 1.

Theorem 3. *If Assumption 1 holds then, for any measurable set $A \subset \mathbb{R}$,*

$$\limsup_{t\to\infty} \sup_{k\geq 0} \bigg\{ \frac{1}{t} \log \mathrm{P}[\bar{X}(kt, (k+1)t) \in A] + \inf_{x\in\bar{A}} \Lambda^{k*}(x) \bigg\} \leq 0$$

and

$$\liminf_{t\to\infty} \inf_{k\geq 0} \bigg\{ \frac{1}{t} \log \mathrm{P}[\bar{X}(kt, (k+1)t) \in A] + \inf_{x\in A^\circ} \Lambda^{k*}(x) \bigg\} \geq 0,$$

where the rate function $\Lambda^{k*}(\cdot)$ is the Fenchel–Legendre transform of

$$\Lambda^k(\lambda) := \int_k^{k+1} \Lambda_\xi\bigg(\frac{(\alpha+1)\lambda\phi y^\alpha}{(k+1)^{\alpha+1} - k^{\alpha+1}} \bar{\beta} \bigg)\,\mathrm{d}y$$

and $\Lambda_\xi(\cdot)$ is as defined in (1).

Proof. The result will follow once we check that the conditions of Theorem 2 hold by setting

$$Y_{k,t} := \bar{X}(kt, (k+1)t) = \frac{S((k+1)t) - S(kt)}{N((k+1)t) - N(kt)} \qquad \text{for all } t \in \mathbb{N},\ k \in \mathbb{R}_+.$$

https://doi.org/10.1239/aap/1354716587 Published online by Cambridge University Press

The most complicated part is to check the uniform convergence condition (9), that is, for any $0 < \Delta < \infty$,

$$\lim_{t\to\infty} \sup_{k\geq 0,\, |\lambda|\leq\Delta} \left|\Lambda^k(\lambda) - \frac{1}{t}\log \mathrm{E}[\exp\{t\lambda \bar{X}(kt,\,(k+1)t)\}]\right| = 0. \tag{21}$$

We begin by observing that, for any $u \in \mathbb{R}$,

$$\begin{aligned}
&\log \mathrm{E}[\exp\{u(S((k+1)t) - S(kt))\}] \\
&\quad = \log \mathrm{E}\biggl[\exp\biggl\{u \sum_{l=kt+1}^{(k+1)t} \sum_{i=1}^{K} \sum_{j=1}^{n_i(l)} X_{i,j}(l)\biggr\}\biggr] \\
&\quad = \log \mathrm{E}\biggl[\exp\biggl\{u \sum_{l=kt+1}^{(k+1)t} \sum_{i=1}^{K} n_i(l)\beta_i^{\top} Z(l) + u \sum_{l=kt+1}^{(k+1)t} \sum_{i=1}^{K} \sum_{j=1}^{n_i(l)} \varepsilon_{i,j}(l)\biggr\}\biggr] \\
&\quad = \log \mathrm{E}\biggl[\exp\biggl\{u \sum_{l=kt+1}^{(k+1)t} \sum_{i=1}^{K} n_i(l)\beta_i^{\top} Z(l)\biggr\}\biggr] + \log \mathrm{E}\biggl[\exp\biggl\{u \sum_{l=kt+1}^{(k+1)t} \sum_{i=1}^{K} \sum_{j=1}^{n_i(l)} \varepsilon_{i,j}(l)\biggr\}\biggr], \qquad (22)
\end{aligned}$$

where the last equality follows from the independence of the εs and the Zs. To understand the first component of (22), define $\beta = \sum_{i=1}^{K} c_i\beta_i$ and note that

$$\begin{aligned}
&\log \mathrm{E}\biggl[\exp\biggl\{u \sum_{l=kt+1}^{(k+1)t} \sum_{i=1}^{K} n_i(l)\beta_i^{\top} Z(l)\biggr\}\biggr] \\
&\quad = \log \mathrm{E}\biggl[\exp\biggl\{u \sum_{i=1}^{K} \beta_i^{\top} \biggl(\sum_{l=kt+1}^{(k+1)t} n_i(l) \sum_{j=-\infty}^{\infty} \phi_k \xi(l-j)\biggr)\biggr\}\biggr] \\
&\quad = \log \mathrm{E}\biggl[\exp\biggl\{u \biggl(\sum_{i=1}^{K} \beta_i c_i\biggr) \cdot \biggl(\sum_{l=kt+1}^{(k+1)t} \lfloor l^{\alpha} \rfloor \sum_{j=-\infty}^{\infty} \phi_j \xi(l-j)\biggr)\biggr\}\biggr] \\
&\quad = \log \mathrm{E}\biggl[\exp\biggl\{u\beta \cdot \biggl(\sum_{j=-\infty}^{\infty} \xi(j) \sum_{l=kt+1}^{(k+1)t} \lfloor l^{\alpha} \rfloor \phi_{l-j}\biggr)\biggr\}\biggr] \\
&\quad = \sum_{j=-\infty}^{\infty} \Lambda_{\xi}\biggl(u\beta \sum_{l=kt+1}^{(k+1)t} \lfloor l^{\alpha} \rfloor \phi_{l-j}\biggr).
\end{aligned}$$

Using the triangle inequality, we obtain the obvious bound

$$\begin{aligned}
&\lim_{t\to\infty} \sup_{k\geq 0,\, |\lambda|\leq\Delta} \left|\Lambda^k(\lambda) - \frac{1}{t}\log \mathrm{E}[\exp\{t\lambda \bar{X}(kt,\,(k+1)t)\}]\right| \\
&\quad \leq \lim_{t\to\infty} \sup_{k\geq 0,\, |\lambda|\leq\Delta} \left|\Lambda^k(\lambda) - \frac{1}{t} \sum_{j=kt+1}^{(k+1)t} \Lambda_{\xi}\biggl(\frac{t\lambda}{N((k+1)t) - N(kt)} \beta \sum_{l=kt+1}^{(k+1)t} \lfloor l^{\alpha} \rfloor \phi_{l-j}\biggr)\right| \\
&\qquad + \lim_{L\to\infty} \lim_{t\to\infty} \sup_{k\geq 0,\, |\lambda|\leq\Delta} \left|\frac{1}{t} \sum_{j=-\infty}^{kt-L} \Lambda_{\xi}\biggl(\frac{t\lambda}{N((k+1)t) - N(kt)} \beta \sum_{l=kt+1}^{(k+1)t} \lfloor l^{\alpha} \rfloor \phi_{l-j}\biggr)\right|
\end{aligned}$$

$$+\lim_{L\to\infty}\lim_{t\to\infty}\sup_{k\geq 0,\,|\lambda|\leq\Delta}\left|\frac{1}{t}\sum_{j=(k+1)t+L}^{\infty}\Lambda_\xi\left(\frac{t\lambda}{N((k+1)t)-N(kt)}\beta\sum_{l=kt+1}^{(k+1)t}\lfloor l^\alpha\rfloor\phi_{l-j}\right)\right|$$

$$+\lim_{t\to\infty}\sup_{k\geq 0,\,|\lambda|\leq\Delta}\left|\frac{1}{t}\sum_{\substack{kt-L<j\leq kt \text{ or}\\(k+1)t<j\leq(k+1)t+L}}\Lambda_\xi\left(\frac{t\lambda}{N((k+1)t)-N(kt)}\beta\sum_{l=kt+1}^{(k+1)t}\lfloor l^\alpha\rfloor\phi_{l-j}\right)\right|$$

$$+\lim_{t\to\infty}\sup_{k\geq 0,\,|\lambda|\leq\Delta}\left|\frac{1}{t}\log\mathrm{E}\left[\exp\left\{\frac{t\lambda}{N((k+1)t)-N(kt)}\sum_{l=1}^{t}\sum_{i=1}^{K}\sum_{j=1}^{n_i(l)}\varepsilon_{i,j}(l)\right\}\right]\right|. \qquad (23)$$

We will prove (21) by showing that each of the terms in the above expression is equal to 0. For that purpose, we make use of the following facts.

(i) There exists $M' > 0$ such that

$$\frac{t((k+1)t)^\alpha}{(kt+1)^\alpha+\cdots+((k+1)t)^\alpha}\leq M' \quad \text{for all } t\geq 1,\ k\geq 0.$$

(ii) Given any $0<\epsilon<\frac{1}{2}$, there exists $\kappa_1>0$ such that

$$|\Lambda_\xi(u)-\Lambda_\xi(v)|\leq\kappa_1\|u-v\| \quad \text{whenever } \|u\|\leq M,\ \|v\|\leq M, \text{ and } \|u-v\|\leq\epsilon,$$

where $\|\cdot\|$ denotes the supnorm on $\mathbb{R}^K$ and

$$M=M'\Delta\|\bar\beta\|\sum_{k=-\infty}^{\infty}|\phi_k|.$$

(iii) There exists $L\geq 1$ such that $\sum_{|k|>L}|\phi_k|<\epsilon/(M'\Delta\|\bar\beta\|)$.

We obtain (ii) since $\Lambda_\xi(\cdot)$ is convex and differentiable (cf. Lemma 2.2.5 of Dembo and Zeitouni (1998)) and (iii) follows from the summability of the coefficients (ϕ_k).

Define the function $f_{t,k}\colon (k,k+1)\to\mathbb{R}$ by

$$f_{t,k}(y):=\Lambda_\xi\left(\frac{t\lambda}{N((k+1)t)-N(kt)}\left(\sum_{l=kt+1}^{(k+1)t}\lfloor l^\alpha\rfloor\phi_{l-\lceil ty\rceil}\right)\beta\right),$$

and note that

$$\frac{1}{t}\sum_{j=kt+1}^{(k+1)t}\Lambda_\xi\left(\frac{t\lambda}{N((k+1)t)-N(kt)}\beta\sum_{l=kt+1}^{(k+1)t}\lfloor l^\alpha\rfloor\phi_{l-j}\right)=\int_k^{k+1}f_{t,k}(y)\,\mathrm{d}y.$$

Choose t large enough such that $kt+1\leq\lceil ty\rceil-L$, $\lceil ty\rceil+L\leq(k+1)t$, and

$$\left|\frac{t\lfloor l^\alpha\rfloor}{N((k+1)t)-N(kt)}-\frac{(\alpha+1)y^\alpha}{C((k+1)^{\alpha+1}-k^{\alpha+1})}\right|\leq\frac{\epsilon}{\Delta\|\beta\|}\left(\sum_{k=-\infty}^{\infty}|\phi_k|\right)^{-1}$$

for all $k\geq 0$, $k+\epsilon<y<k+1-\epsilon$, and $\lceil ty\rceil-L\leq l\leq\lceil ty\rceil+L$. It is easy to check that, for y in this range and $|\lambda|\leq\Delta$,

$$\left\|\frac{t\lambda\beta}{N((k+1)t)-N(kt)}\sum_{l=kt+1}^{(k+1)t}\lfloor l^\alpha\rfloor\phi_{l-\lceil ty\rceil}-\frac{t\lambda\beta}{N((k+1)t)-N(kt)}\sum_{l=\lceil ty\rceil-L}^{\lceil ty\rceil+L}\lfloor l^\alpha\rfloor\phi_{l-\lceil ty\rceil}\right\|\leq\epsilon,$$

$$\left\| \frac{t\lambda\beta}{N((k+1)t)-N(kt)} \sum_{l=\lceil ty\rceil-L}^{\lceil ty\rceil+L} \lfloor l^{\alpha}\rfloor \phi_{l-\lceil ty\rceil} - \frac{(\alpha+1)\lambda y^{\alpha}\bar{\beta}}{(k+1)^{\alpha+1}-k^{\alpha+1}} \sum_{l=\lceil ty\rceil-L}^{\lceil ty\rceil+L} \phi_{l-\lceil ty\rceil} \right\| \le \epsilon,$$

and

$$\left\| \frac{(\alpha+1)\lambda y^{\alpha}\bar{\beta}}{(k+1)^{\alpha+1}-k^{\alpha+1}} \sum_{l=\lceil ty\rceil-L}^{\lceil ty\rceil+L} \phi_{l-\lceil ty\rceil} - \frac{(\alpha+1)\lambda y^{\alpha}\bar{\beta}}{(k+1)^{\alpha+1}-k^{\alpha+1}} \sum_{l=-\infty}^{\infty} \phi_{l} \right\| \le \epsilon.$$

This implies that, for all $k \ge 0$, $k+\epsilon < y < k+1-\epsilon$, and $|\lambda| \le \Delta$,

$$\left| \Lambda_{\xi}\left(\frac{(\alpha+1)\lambda\phi y^{\alpha}}{(k+1)^{\alpha+1}-k^{\alpha+1}} \bar{\beta} \right) - f_{t,k}(y) \right| \le 3\kappa_1\epsilon,$$

and, hence, we obtain

$$\begin{aligned} \lim_{t\to\infty} \sup_{k\ge 0,\, |\lambda|\le\Delta} & \left| \Lambda^{k}(\lambda) - \frac{1}{t} \sum_{j=kt+1}^{(k+1)t} \Lambda_{\xi}\left(\frac{t\lambda}{N((k+1)t)-N(kt)} \beta \sum_{l=kt+1}^{(k+1)t} \lfloor l^{\alpha}\rfloor \phi_{l-j} \right) \right| \\ & \le 3\kappa_1\epsilon + 4M_1\epsilon, \end{aligned} \tag{24}$$

where

$$M_1 = \max\left\{ \Lambda_{\xi}\left(M'\Delta\|\beta\| \sum_{k=-\infty}^{\infty} |\phi_k| \right), \Lambda_{\xi}\left(-M'\Delta\|\beta\| \sum_{k=-\infty}^{\infty} |\phi_k| \right) \right\}.$$

Obviously, since ϵ is arbitrary, the limit in (24) is 0.

The other parts in (23) are handled much more easily. Note that, for any $k \ge 0$,

$$\begin{aligned} \left| \sum_{j=-\infty}^{kt-L} \Lambda_{\xi}\left(\frac{t\lambda}{N((k+1)t)-N(kt)} \beta \sum_{l=kt+1}^{(k+1)t} \lfloor l^{\alpha}\rfloor \phi_{l-j} \right) \right| & \le \kappa_1 M'\Delta\|\bar{\beta}\| \sum_{j=-\infty}^{kt-L} \sum_{l=kt+1}^{(k+1)t} |\phi_{l-j}| \\ & \le t\kappa_1\epsilon, \end{aligned}$$

and, hence,

$$\lim_{L\to\infty} \lim_{t\to\infty} \sup_{k\ge 0,\, |\lambda|\le\Delta} \left| \frac{1}{t} \sum_{j=-\infty}^{kt-L} \Lambda_{\xi}\left(\frac{t\lambda}{N((k+1)t)-N(kt)} \beta \sum_{l=kt+1}^{(k+1)t} \lfloor l^{\alpha}\rfloor \phi_{l-j} \right) \right| = 0.$$

Using a similar argument, we also obtain

$$\lim_{L\to\infty} \lim_{t\to\infty} \sup_{k\ge 0,\, |\lambda|\le\Delta} \left| \frac{1}{t} \sum_{j=(k+1)t+L}^{\infty} \Lambda_{\xi}\left(\frac{t\lambda}{N((k+1)t)-N(kt)} \beta \sum_{l=kt+1}^{(k+1)t} \lfloor l^{\alpha}\rfloor \phi_{l-j} \right) \right| = 0.$$

Furthermore, it is also easy to check that, for every $L \ge 1$,

$$\begin{aligned} \lim_{t\to\infty} \sup_{k\ge 0,\, |\lambda|\le\Delta} & \left| \frac{1}{t} \sum_{\substack{kt-L<j\le kt \text{ or} \\ (k+1)t<j\le(k+1)t+L}} \Lambda_{\xi}\left(\frac{t\lambda}{N((k+1)t)-N(kt)} \beta \sum_{l=kt+1}^{(k+1)t} \lfloor l^{\alpha}\rfloor \phi_{l-j} \right) \right| \\ & \le \lim_{t\to\infty} \frac{1}{t} 2(L+1)M_1 \\ & = 0. \end{aligned}$$

https://doi.org/10.1239/aap/1354716587 Published online by Cambridge University Press

For the final part of the proof of (21), we note the following facts about $\Lambda_\varepsilon(\cdot)$: $\Lambda_\varepsilon(0) = 0$, $\Lambda'_\varepsilon(0) = 0$ because $\mathrm{E}[\varepsilon_{i,j}(t)] = 0$, and $\Lambda_\varepsilon(\cdot)$ is nonnegative and twice continuously differentiable in a neighborhood of 0. The last fact can be easily derived following Lemma 2.2.5 of Dembo and Zeitouni (1998). This implies that there exist positive constants κ and η such that

$$|\Lambda_\varepsilon(u)| \le \kappa u^2 \quad \text{for all } |u| \le \eta.$$

Choose t large enough such that $t\Delta/N(t) < \eta$. This also means that $|t\lambda/(N((k+1)t) - N(kt))| < \eta$ for all $k \ge 0$ and $|\lambda| \le \Delta$. Hence, we have

$$\begin{aligned}
&\left|\log \mathrm{E}\left[\exp\left\{\frac{t\lambda}{N((k+1)t) - N(kt)} \sum_{l=1}^{t}\sum_{i=1}^{K}\sum_{j=1}^{n_i(l)} \varepsilon_{i,j}(l)\right\}\right]\right| \\
&\quad = \sum_{l=kt+1}^{(k+1)t}\sum_{i=1}^{K}\sum_{j=1}^{n_i(l)} \Lambda_\varepsilon\left(\frac{t\lambda}{N((k+1)t) - N(kt)}\right) \\
&\quad \le (N((k+1)t) - N(kt))\kappa \frac{t^2\lambda^2}{(N((k+1)t) - N(kt))^2}.
\end{aligned}$$

This immediately gives

$$\begin{aligned}
&\lim_{t\to\infty} \sup_{k\ge 0,\, |\lambda|\le\Delta} \left|\frac{1}{t}\log \mathrm{E}\left[\exp\left\{\frac{\lambda}{N((k+1)t) - N(kt)} \sum_{l=1}^{t}\sum_{i=1}^{K}\sum_{j=1}^{n_i(l)} \varepsilon_{i,j}(l)\right\}\right]\right| \\
&\le \lim_{t\to\infty} \kappa \sup_{k\ge 0,\, |\lambda|\le\Delta} \frac{t\lambda}{(N((k+1)t) - N(kt))} \\
&= 0,
\end{aligned}$$

completing the proof of (21).

It is simpler to check the other conditions of Theorem 2. Note that we can find M such that

$$\frac{y^\alpha}{(k+1)^{\alpha+1} - k^{\alpha+1}} \le M \quad \text{for all } k \ge 0,\ k \le y \le k+1. \tag{25}$$

This implies that, for any $\Delta > 0$,

$$\sup_{k\ge 0,\, |\lambda|\le\Delta} |\Lambda^k(\lambda)| < \infty,$$

and this combined with (21) shows that condition (10) holds.

Next we check that $\Lambda^k(\cdot)$ is differentiable. Since $\Lambda_\xi(\cdot)$ is finite everywhere, by Lemma 2.2.5 of Dembo and Zeitouni (1998), $\Lambda_\xi(\cdot)$ is differentiable and

$$\Lambda'_\xi(\eta) = \frac{\mathrm{E}[\xi(0)\mathrm{e}^{\eta\cdot\xi(0)}]}{\mathrm{E}[\mathrm{e}^{\eta\cdot\xi(0)}]}.$$

For any δ satisfying $0 < \|\delta\| < 1$,

$$z\mathrm{e}^{(\eta+\delta)\cdot z} - z\mathrm{e}^{\eta\cdot z} \to 0 \quad \text{and} \quad \|z\mathrm{e}^{(\eta+\delta)\cdot z} - z\mathrm{e}^{\eta\cdot z}\| \le h(z) := \|z\|\mathrm{e}^{\eta\cdot z}(\mathrm{e}^{\|z\|} + 1).$$

https://doi.org/10.1239/aap/1354716587 Published online by Cambridge University Press

Since $\mathrm{E}[h(\xi(0))] < \infty$ using the dominated convergence theorem, $\mathrm{E}[\xi(0)\mathrm{e}^{\lambda\xi(0)}]$ is continuous. This implies that $\Lambda'_\xi(\cdot)$ is continuous. Now we can use the Leibniz integral rule (cf. Theorem 7.40 of Apostol (1974)) to show that $\Lambda^k(\cdot)$ is differentiable and

$$(\Lambda^k)'(\lambda) := \int_k^{k+1} \frac{(\alpha+1)\phi y^\alpha}{(k+1)^{\alpha+1} - k^{\alpha+1}} \bar{\beta} \cdot \Lambda'_\xi\bigg(\frac{(\alpha+1)\lambda\phi y^\alpha}{(k+1)^{\alpha+1} - k^{\alpha+1}} \bar{\beta}\bigg)\, \mathrm{d}y.$$

It is easy to see that $\|\Lambda'_\xi(\eta)\| \to \infty$ whenever $\|\eta\| \to \infty$. This combined with (25) shows that (11) holds. Finally, (12) follows from the fact that $\Lambda'_\xi(\cdot)$ is continuous on compact sets and (25). This completes the proof of the theorem.

4. Proof of Theorem 1 and required lemmas

Proof of Theorem 1. We will first prove the lower inequality in (4). The inequality is obvious when $I_* = 0$. Also, if A is nonempty then $I_* < \infty$ from Assumption 1. So it suffices to consider $0 < I_* < \infty$. We will use the simple inclusion bound: for all $m \geq 1$ and $r \geq 1$,

$$\{T_r(A) \leq m\} \subset \bigcup_{l=r}^{\infty} \bigcup_{j=0}^{m-1} \{\bar{X}(j, j+l) \in A\}.$$

Thus, we obtain

$$\mathrm{P}[T_r(A) \leq m] \leq \sum_{l=r}^{\infty} \sum_{j=0}^{m-1} \mathrm{P}[\bar{X}(j, j+l) \in A].$$

Lemma 1 below shows that the $\Lambda^{k*}(x)$ are increasing functions of k for fixed x. Lemma 2 below, which builds on this, gives the existence of a K_0 such that

$$\inf_{x\in\bar{A}} \Lambda^{k*}(x) \geq \inf_{x\in\bar{A}} \Lambda^*(x) - \tfrac{1}{3}\epsilon = I_* - \tfrac{1}{3}\epsilon \qquad \text{for all } k \geq K_0.$$

We can also find, from Lemma 3 below, a constant $I > 0$ such that $I \leq \inf_{x\in\bar{A}} \Lambda^{k*}(x)$ for all $k \geq 0$. Now, for any $0 < \epsilon < I$, by Theorem 3 we can obtain $T \geq 1$ such that, for all $l \geq T$ and all $k \geq 0$,

$$\mathrm{P}[\bar{X}(kl, (k+1)l) \in A] \leq \exp\Big\{-l\Big(\inf_{x\in\bar{A}} \Lambda^{k*}(x) - \tfrac{1}{3}\epsilon\Big)\Big\}.$$

This gives, for $r \geq T$,

$$\begin{aligned}
\mathrm{P}[T_r(A) \leq m] &\leq \sum_{l=r}^{\infty} \sum_{j=0}^{m-1} \mathrm{P}[\bar{X}(j, j+l) \in A] \\
&\leq \sum_{l=r}^{\infty} \sum_{j=0}^{K_0 l} \mathrm{P}[\bar{X}(j, j+l) \in A] + \sum_{l=r}^{\infty} \sum_{j=K_0 l}^{m-1} \mathrm{P}[\bar{X}(j, j+l) \in A] \\
&\leq \sum_{l=r}^{\infty} K_0 l \mathrm{e}^{-l(I-\epsilon/3)} + m \sum_{l=r}^{\infty} \mathrm{e}^{-l(I_*-2\epsilon/3)}.
\end{aligned}$$

Now set $m = \lfloor \mathrm{e}^{r(I_* - \epsilon)} \rfloor$, and note that

$$\sum_{r=1}^{\infty} \mathrm{P}[T_r(A) \le \lfloor \mathrm{e}^{r(I_*-\epsilon)} \rfloor] \le T + \sum_{r=T}^{\infty} \sum_{l=r}^{\infty} K_0 l \mathrm{e}^{-l(I-\epsilon/3)} + \sum_{r=T}^{\infty} \mathrm{e}^{r(I_*-\epsilon)} \sum_{l=r}^{\infty} \mathrm{e}^{-l(I_*-2\epsilon/3)}$$
$$< \infty.$$

Hence, using the Borel–Cantelli lemma, we obtain

$$\liminf_{r\to\infty} \frac{\log T_r(A)}{r} \ge I_* - \epsilon \quad \text{a.s.}$$

The lower inequality in (4) is thus proved by letting $\epsilon \to 0$.

Also, observe that, using the relation $\{T_r(A) \le m\} = \{R_m(A) \ge r\}$, we obtain

$$\limsup_{t\to\infty} \frac{R_t}{\log t} \le \frac{1}{I_*} \quad \text{a.s.}$$

In order to prove the upper bound in (4), it suffices to consider the case $I^* < \infty$. In this case the set A has a nonempty interior. Define two new random variables by

$$Y'_{k,t} := \beta \frac{\sum_{j=kt+1}^{(k+1)t} \sum_{l=kt+1}^{(k+1)t} \lfloor l^{\alpha} \rfloor \phi_{l-j} \xi(j)}{N((k+1)t) - N(kt)} \quad \text{and} \quad Y''_{k,t} := \bar{X}(kt, (k+1)t) - Y'_{k,t},$$

where, as before, $\beta = \sum_{i=1}^{K} c_i \beta_i$. For a set A and $\eta > 0$, define

$$A(\eta) := \{x : d(x, A^{\mathrm{c}}) > \eta\},$$

where $d(x, A^{\mathrm{c}})$ is the distance from the point x to the complement A^{c}. Now observe that, for any positive integers r and q with $q > r$,

$$\begin{aligned} \mathrm{P}[T_r(A) > q] &\le \mathrm{P}\bigg[\bar{X}(kr, (k+1)r) \notin A,\ k = 0, \ldots, \bigg\lfloor \frac{q}{r} \bigg\rfloor\bigg] \\ &\le \mathrm{P}\bigg[Y'_{k,r} \notin A(\eta),\ k = 0, \ldots, \bigg\lfloor \frac{q}{r} \bigg\rfloor\bigg] + \sum_{l=1}^{\lfloor q/r \rfloor} \mathrm{P}[|Y''_{k,r}| > \eta]. \end{aligned} \tag{26}$$

Since the $Y'_{k,r}$, $k = 0, 1, \ldots, \lfloor q/r \rfloor$, are independent, the right-hand side of (26) equals

$$\begin{aligned} &\prod_{k=0}^{\lfloor q/r \rfloor} (1 - \mathrm{P}[Y'_{k,r} \in A(\eta)]) + \sum_{l=1}^{\lfloor q/r \rfloor} \mathrm{P}[|Y''_{k,r}| > \eta] \\ &\qquad \le \exp\bigg(-\sum_{k=0}^{\lfloor q/r \rfloor} \mathrm{P}[Y'_{k,r} \in A(\eta)]\bigg) + \sum_{l=1}^{\lfloor q/r \rfloor} \mathrm{P}[|Y''_{k,r}| > \eta]. \end{aligned}$$

From the arguments following (24), it is easy to check that the law of $Y'_{k,t}$ satisfies the LDP uniformly over $k \ge 0$ with rate function $\Lambda^{k*}(\cdot)$. We can therefore obtain $T \ge 1$ such that

$$\frac{1}{t} \log \mathrm{P}[Y'_{k,t} \in A(\eta)] \ge -\inf_{x\in A(\eta)} \Lambda^{k*}(x) - \frac{1}{4}\epsilon \quad \text{for all } t \ge T,\ k \ge 0.$$

https://doi.org/10.1239/aap/1354716587 Published online by Cambridge University Press

Lemma 1(ii) below then implies that

$$\frac{1}{t}\log \mathrm{P}[Y'_{k,t} \in A(\eta)] \ge -\inf_{x\in A(\eta)} \Lambda^*(x) - \frac{1}{4}\epsilon \quad \text{for all } t \ge T,\ k \ge 0.$$

Hence, for small enough $\eta > 0$,

$$\frac{1}{t}\log \mathrm{P}[Y'_{k,t} \in A(\eta)] \ge -I^* - \frac{1}{2}\epsilon \quad \text{for all } t \ge T,\ k \ge 0.$$

Therefore, by setting $q_r = \lceil \mathrm{e}^{r(I^*+\epsilon)} \rceil$ and using the above inequality, we obtain

$$\begin{aligned}\sum_{r=1}^{\infty} \exp\biggl\{-\sum_{k=0}^{\lfloor q_r/r \rfloor} \mathrm{P}[Y'_{k,r} \in A(\eta)]\biggr\} &\le T + \sum_{r=T}^{\infty} \exp\biggl\{-\frac{\mathrm{e}^{r(I^*+\epsilon)}}{r}\mathrm{e}^{-r(I^*+\epsilon/2)}\biggr\} \\ &\le T + \sum_{r=T}^{\infty} \exp\biggl\{-\frac{\mathrm{e}^{r\epsilon/2}}{r}\biggr\} \\ &< \infty. \end{aligned} \tag{27}$$

Furthermore, note that, for $\epsilon > 0$ and $\eta > 0$ such that the above holds,

$$\limsup_{t\to\infty}\sup_{k\ge 0}\frac{1}{t}\log \mathrm{P}[|Y''_{k,t}| > \eta] \le -\lambda\eta + \limsup_{t\to\infty}\sup_{k\ge 0}\frac{1}{t}\log \mathrm{E}[\lambda t \mid Y''_{k,t}] = -\lambda\eta.$$

The last equality follows from the steps used in the proof of Theorem 3. Now by choosing $\lambda > (I^* + \epsilon)/\eta$ we obtain

$$\sum_{r=1}^{\infty}\sum_{l=1}^{\lfloor q_r/r \rfloor} \mathrm{P}[|Y''_{k,r}| > \eta] \le \sum_{r=1}^{\infty}\biggl\lfloor \frac{q_r}{r} \biggr\rfloor \sup_{k\ge 0} \mathrm{P}[|Y''_{k,r}| > \eta] < \infty. \tag{28}$$

Combining (27) and (28) we obtain

$$\sum_{r=1}^{\infty} \mathrm{P}[T_r(A) > q] < \infty.$$

Finally, by applying the first Borel–Cantelli lemma and then letting $\epsilon \to 0$, we complete the proof of the upper bound of (4). The lower bound in (5) is again proved using the same identity $\{T_r(A) \le m\} = \{R_m(A) \ge r\}$. Hence, the proof is complete.

Lemma 1. (i) *For any* $\lambda \in \mathbb{R}$, $\Lambda^k(\lambda)$ *is a decreasing function of* k.

(ii) *For any* $x \in \mathbb{R}$, $\Lambda^{k*}(x)$ *is an increasing function of* k.

Proof. Suppose that F_k is the distribution function of the random variables

$$U_k := \frac{(\alpha+1)(k+U)^{\alpha}}{(k+1)^{\alpha+1} - k^{\alpha+1}} \quad \text{where } U \sim \mathrm{Uniform}(0,1),\ k \ge 0.$$

Observe that $\mathrm{E}(U_k) = 1$ for all $k \ge 0$. Also, for any nonnegative random variable X with mean 1 and distribution F_X, define the Lorenz function

$$L_X(p) := \int_0^p F_X^{-1}(u)\,\mathrm{d}u \quad \text{for all } 0 \le p \le 1.$$

https://doi.org/10.1239/aap/1354716587 Published online by Cambridge University Press

Note that the Lorenz function of U_k is given by

$$L_{U_k}(p) = \frac{(k+p)^{\alpha+1} - k^{\alpha+1}}{(k+1)^{\alpha+1} - k^{\alpha+1}} \quad \text{for all } 0 \le p \le 1$$

and

$$\frac{\partial}{\partial k} L_{U_k}(p) = \frac{(\alpha+1)[(k+p)^{\alpha}((k+1)^{\alpha}(1-p) + k^{\alpha}p) - k^{\alpha}(k+1)^{\alpha}]}{((k+1)^{\alpha+1} - k^{\alpha+1})^2} > 0$$

for all $k \ge 0$ and $0 \le p \le 1$. This implies that

$$L_{U_{k'}}(p) \ge L_{U_{k''}}(p) \quad \text{for all } 0 \le p \le 1,\ k' \ge k'',$$

which means that U_k is decreasing in Lorenz order as k increases. Hence, by Arnold (1980, Theorem 3.2, p. 37) and using the fact that $\Lambda_{\xi}(\cdot)$ is convex and continuous, we find that

$$\Lambda^k(\lambda) = \mathrm{E}[\Lambda_{\xi}(\lambda\phi U_k \bar{\beta})]$$

is decreasing in k.

Part (ii) of the lemma follows easily from part (i) using the definition of the Fenchel–Legendre transform.

Lemma 2. *For any measurable set $A \subset \mathbb{R}$ and $\epsilon > 0$, there exists K_0 such that*

$$\inf_{x\in A} \Lambda^{k*}(x) \ge \inf_{x\in A} \Lambda^*(x) - \epsilon \quad \textit{for all } k \ge K_0,$$

where $\Lambda^{k*}(\cdot)$ and $\Lambda^*(\cdot)$ are as described in Theorem 3 and Theorem 1, respectively.

Proof. Fix any $\epsilon > 0$. From the arguments leading to (16) we can find $M_1 > 0$ such that

$$\inf_{x\in A} \Lambda^*(x) = \inf_{x\in A\cap[-M_1,M_1]} \Lambda^*(x).$$

Lemma 1(ii) then gives

$$\inf_{x\in A} \Lambda^{k*}(x) = \inf_{x\in A\cap[-M_1,M_1]} \Lambda^{k*}(x) \quad \text{for all } k \ge 0.$$

Using Assumption 1, we obtain $M_2 > 0$ such that $|\lambda| > M_2$ implies that $|(\Lambda^0)'(\lambda)| > 2M_1$. Since $\Lambda^k(\cdot)$ converges locally uniformly to $\Lambda(\cdot)$, we know that there exists K_0 such that

$$\sup_{\lambda\in[-M_2,M_2]} |\Lambda^k(\lambda) - \Lambda(\lambda)| < \tfrac{1}{4}\epsilon \quad \text{for all } k \ge K_0.$$

Now, for any $x \in [-M_1, M_1]$, we can obtain $\lambda_x \in [-M_2, M_2]$ such that $\lambda_x x - \Lambda(\lambda_x) > \Lambda^*(x) - \epsilon/4$ and, therefore, for all $k \ge K_0$,

$$\Lambda^{k*}(x) \ge \lambda_x x - \Lambda^k(\lambda_x) \ge \lambda_x x - \Lambda(\lambda_x) - \tfrac{1}{4}\epsilon \ge \Lambda^*(x) - \tfrac{1}{2}\epsilon.$$

This implies that, for all $k \ge K_0$,

$$\inf_{x\in A\cap[-M_1,M_1]} \Lambda^{k*}(x) \ge \inf_{x\in A\cap[-M_1,M_1]} \Lambda^*(x) - \epsilon,$$

completing the proof.

https://doi.org/10.1239/aap/1354716587 Published online by Cambridge University Press

Lemma 3. *For any measurable set $A \subset \mathbb{R}$,*

$$\inf_{x \in A} \Lambda^*(x) > 0 \quad \Longrightarrow \quad \inf_{k \geq 0} \inf_{x \in A} \Lambda^{k*}(x) > 0. \tag{29}$$

Proof. Using Lemma 1(ii), it suffices to show that (29) implies that $\inf_{x \in A} \Lambda^{0*}(x) > 0$. Fix any $x \neq 0$. Since $\Lambda^0(\lambda)$ is strictly convex and finite everywhere and $(\Lambda^0)'(0) = 0$, if $(\Lambda^0)'(\lambda_x^0) = x$ then $\lambda_x^0 \neq 0$. Then $\Lambda^{0*}(x) = \lambda_x^0 x - \Lambda^0(\lambda_x^0) \neq 0$. For any measurable $A \subset \mathbb{R}$,

$$\inf_{x \in A} \Lambda^{0*}(x) = 0 \quad \Longrightarrow \quad 0 \in \bar{A}.$$

This would imply that $\inf_{x \in A} \Lambda^*(x) = 0$. This proves the lemma.

References

Apostol, T. M. (1974). *Mathematical Analysis.* Addison-Wesley, Reading, MA.

Arnold, B. C. (1980). *Majorization and the Lorenz Order: A Brief Introduction* (Lecture Notes Statist. **43**). Springer.

Arratia, R., Gordon, L. and Waterman, M. S. (1990). The Erdös-Rényi law in distribution, for coin tossing and sequence matching. *Ann. Statist.* **18,** 539–570.

Bryc, W. and Dembo, A. (1996). Large deviations and strong mixing. *Ann. Inst. H. Poincaré Prob. Statist.* **32,** 549–569.

Dembo, A. and Zeitouni, O. (1998). *Large Deviations Techniques and Applications,* 2nd edn. Springer, New York.

Deuschel, J.-D. and Stroock, D. W. (1989). *Large Deviations.* Academic Press, Boston, MA.

Ellis, R. S. (1984). Large deviations for a general class of random vectors. *Ann. Prob.* **12,** 1–12.

Erdös, P. and Rényi, A. (1970). On a new law of large numbers. *J. Analyse Math.* **23,** 103–111.

Gartner, J. (1977). On large deviations from the invariant measure. *Theory Prob. Appl.* **22,** 24–39.

Ghosh, S. and Samorodnitsky, G. (2009). The effect of memory on functional large deviations of infinite moving average processes. *Stoch. Process. Appl.* **119,** 534–561.

Ghosh, S. and Samorodnitsky, G. (2010). Long strange segments, ruin probabilities and the effect of memory on moving average processes. *Stoch. Process. Appl.* **120,** 2302–2330.

Khan, A., Yan, X., Tao, S. and Anerousis, N. (2012). Workload characterization and prediction in the cloud: a multiple time series approach. Submitted.

Li, H. *et al.* (2009). Developing an enterprise cloud computing strategy. White Paper, Intel Corporation.

Li, T.-H. (2005). A hierarchical framework for modeling and forecasting web server workload. *J. Amer. Statist. Assoc.* **100,** 748–763.

Li, T.-H. (2007). A statistical framework of optimal workload consolidation with application to capacity planning for on-demand computing. *J. Amer. Statist. Assoc.* **102,** 841–855.

Mansfield, P., Rachev, S. T. and Samorodnitsky, G. (2001). Long strange segments of a stochastic process. *Ann. Appl. Prob.* **11,** 878–921.

Mendler, C. (2010). Cloud 99.99: the small print exposed. Analyst Report, Yankee Group.

Rachev, S. T. and Samorodnitsky, G. (2001). Long strange segments in a long-range-dependent moving average. *Stoch. Process. Appl.* **93,** 119–148.

Reid, S. *et al.* (2011). Sizing the cloud: understanding and quantifying the future of cloud computing. Tech. Rep., Forrester Research. Available at http://forrester.com/rb/Research/sizing_cloud/q/id/58161/t/2.

Varadhan, S. R. S. (1984). *Large Deviations and Applications.* SIAM, Philadelphia, PA.

https://doi.org/10.1239/aap/1354716587 Published online by Cambridge University Press