Comparing strength of locality of reference – Popularity, majorization, and some folk theorems

Sarut Vanichpun   Armand M. Makowski
Department of Electrical and Computer Engineering
and the Institute for Systems Research
University of Maryland, College Park
College Park, Maryland 20742
Email: sarut@eng.umd.edu armand@isr.umd.edu

***Abstract*— The performance of demand-driven caching depends on the locality of reference exhibited by the stream of requests made to the cache. In particular, it is expected that the stronger the locality of reference, the smaller the miss rate of the cache. For the Independent Reference Model, this amounts to a smaller miss rate when the popularity distribution of requested objects in the stream is more skewed. In this paper, we formalize this "folk theorem" through the companion concepts of majorization and Schur-concavity. This folk theorem is established for caches operating under a Random On-demand Replacement Algorithm (RORA). However, the result fails to hold in general under the (popular) LRU and CLIMB policies, but can be established when the input has a Zipf-like popularity pmf with large skewness parameter. In addition, we explore how the majorization of popularity distributions translates into comparisons of three well-known locality of reference metrics, namely the inter-reference time, the working set size and the stack distance.**

Keywords: Locality of reference in request streams, Popularity, Majorization/Schur-concavity.

I. Introduction

Web caching aims to reduce network traffic, server load and user-perceived retrieval latency by replicating "popular" content on (proxy) caches that are strategically placed within the network. This approach is a natural outgrowth of caching techniques which were originally developed for computer memory and distributed file sharing systems, e.g., [1, 2, 3] (and references therein).

The performance of any form of caching is determined by a number of factors, chief amongst them the statistical properties of the streams of requests made to the cache. One important such property is the *locality of reference* present in a request stream whereby bursts of references are made in the near future to objects referenced in the recent past. The implications for cache management should be clear – Increased locality of reference should yield performance improvements for demand-driven caching that exploits recency of reference. In particular, under this form of cache management, we expect the following "folk theorem" to hold: The stronger the locality of reference in the stream of requests, the smaller the miss rate since the cache ends up being populated by Web objects with a higher likelihood of access in the near future.

The notion of locality and its importance for caching were first recognized by Belady [4] in the context of computer memory. Subsequently, a number of studies have shown that request streams for Web objects exhibit strong locality of reference[1] [5, 6, 7]. Attempts at characterization were made early on by Denning through the working set model [8, 9]. Yet, like the notion of burstiness used in traffic modeling, locality of reference, while endowed with a clear intuitive content, admits no simple definition. Not surprisingly, in spite of numerous efforts, no consensus has been reached on how to formalize the notion, let alone *compare* streams of requests on the basis of their locality of reference.[2] This has precluded a formal exploration of the folk theorem mentioned above, and it is one of the purposes of this paper to present a framework, albeit restricted, where such a discussion can take place.

Although several competing definitions are currently available, it is by now widely accepted that the two main contributors to locality of reference are *temporal correlations* in the streams of requests and the *popularity distribution* of requested objects. To describe these two sources of locality, and to frame the subsequent discussion, we assume the following generic setup: We consider a universe of N cacheable items or documents, labeled $i = 1, \ldots, N$, and we write $\mathcal{N} = \{1, \ldots, N\}$. The successive requests arriving at the cache are modeled by a sequence $\{R_t,\ t = 0, 1, \ldots\}$ of $\mathcal{N}$-valued rvs.

1. The *popularity* of the sequence of requests $\{R_t,\ t = 0, 1, \ldots\}$ is defined as the pmf $\boldsymbol{p} = (p(i), \ldots, p(N))$ on $\mathcal{N}$ given by

$$p(i) := \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbf{1}\left[R_\tau = i\right] \quad a.s., \quad i = 1, \ldots, N, \qquad (1)$$

whenever these limits exist (and they do in most models treated in the literature).

2. *Temporal correlations* are more delicate to define. Indeed, it is somewhat meaningless to use the covariance function

$$\gamma(s,t) := \mathrm{Cov}[R_s, R_t], \quad s, t = 0, 1, \ldots.$$

as a way to capture these temporal correlations as is traditionally done in other contexts. This is because the rvs $\{R_t,\ t = 0, 1, \ldots\}$ take values in a discrete set. We took $\{1, \ldots, N\}$

[1] At least in the short timescales

[2] An exception can be found in a recent paper by Fonseca et al. [10]; more on that later!

but we could have selected $\{1, \frac{1}{2}, \ldots, \frac{1}{N}\}$ instead; in fact *any* set of N distinct points in an arbitrary space would do the job. Thus, the *actual* values of the rvs $\{R_t,\ t = 0, 1, \ldots\}$ are of no consequence, and the focus should instead be on the *recurrence* patterns displayed by requests for particular documents over time. The literature contains several metrics to do this, e.g., the inter-reference time [3, 5, 10], the working set size [8, 9] and the stack distance [11, 12, 13].

To see how popularity indeed contributes to locality of reference, consider the situation where there is *no* temporal correlations in the stream of requests as would be the case under the standard *Independence Reference Model* (IRM). More precisely, under the IRM with popularity pmf $\boldsymbol{p}$, the successive requests $\{R_t,\ t = 0, 1, \ldots\}$ form a sequence of i.i.d. $\mathcal{N}$-valued rvs distributed according to the pmf $\boldsymbol{p}$.[3] Here, the *skewness* of $\boldsymbol{p}$ does act as an indicator of the strength of locality of reference present in the stream, under the intuition that the more "balanced" the pmf $\boldsymbol{p}$, the weaker the locality of reference. This is best appreciated by considering the limiting cases: If $\boldsymbol{p}$ is extremely unbalanced with $\boldsymbol{p} = (1-\delta, \varepsilon, \ldots, \varepsilon)$ (with $\delta = (N-1)\varepsilon$), a reference to document 1 is likely to be followed by a burst of additional references to document 1 provided $(N-1)\varepsilon \ll 1-\delta$. The exact opposite conclusion holds if the popularity pmf $\boldsymbol{p}$ were uniform, i.e., $p(1) = \ldots = p(N) = \frac{1}{N}$, for then the successive requests $\{R_t,\ t = 0, 1, \ldots\}$ form a truly random sequence.

Thus, even in the absence of temporal correlations, locality of reference is present, with its strength determined by the skewness of the underlying popularity distribution. In this paper, *as we restrict ourselves to the class of IRMs*,[4] the question naturally arises as to whether pmfs can be compared on the basis of their skewness so that the folk theorem discussed earlier can be established in some form. More formally, consider two IRMs with popularity pmfs $\boldsymbol{p}$ and $\boldsymbol{q}$ (on $\mathcal{N}$), and let $M(\boldsymbol{p})$ and $M(\boldsymbol{q})$ denote their miss rates under some cache replacement policy. We seek a way to compare the vectors $\boldsymbol{p}$ and $\boldsymbol{q}$, with the interpretation that if $\boldsymbol{p}$ is less skewed than $\boldsymbol{q}$, then the comparison

$$M(\boldsymbol{q}) \le M(\boldsymbol{p}) \tag{2}$$

holds. The main contributions along these lines are now summarized:

1. Majorization, Schur-concavity and entropy – In a recent paper, Fonseca et al. [10] introduced such a notion of comparison based on the entropy (6) of the popularity pmfs, i.e., the pmf $\boldsymbol{p}$ is considered to be less skewed (or more balanced) than the pmf $\boldsymbol{q}$ whenever the entropy of $\boldsymbol{p}$ is greater than the entropy of $\boldsymbol{q}$, i.e.,

$$H(\boldsymbol{q}) \le H(\boldsymbol{p}). \tag{3}$$

Unfortunately, this notion is not strong enough to allow for results of the form (2) to be established. Here, we turn instead to the stronger concept of *majorization* [16] as a way to characterize imbalance in the components of popularity pmfs. This notion is stronger than the concept of entropy-based comparison, and therefore holds the promise that comparison results such as (2) might indeed be obtainable under it. This will turn out to be the case as a result of the existence of a rich and structured class of monotone functions associated with majorization, the so-called Schur-convex/concave functions.

2. The folk theorem under RORA policies – The comparison (2) is shown to hold under the IRM for a number of policies, namely the optimal policy A_0, the random policy and the FIFO policy. These positive results are then extended to a very large class of replacement policies, the so-called Random On-demand Replacement Algorithms (RORA). To the best of the authors' knowledge, these results provide the first formal proof of folk theorems such as (2).

3. Counterexamples and asymptotics – However, the comparison (2) does *not* always hold under the LRU[5] and CLIMB replacement policies. We exhibit situations where under these policies, the IRM stream with pmf of higher entropy may have a smaller miss rate than the IRM stream with pmf of lower entropy. Yet, when the popularity pmfs are Zipf-like, simulations show that the comparison (2) *does* hold for the LRU and CLIMB policies. In fact this is formally established in the limiting regime where the skewness parameter of the Zipf-like pmf is large.

4. Popularity and other locality of reference metrics – In the spirit of the comparison (2), we investigate how the comparison by majorization of popularity pmfs is compatible with comparisons of three well-established locality of reference metrics, namely, the inter-reference time, the working set size and the stack distance.

Recently majorization has also been used for comparing the popularity pmf of the output of caches under the IRM for various policies [17]. Additional information on the material of this paper is available in [18].

The paper is organized as follows: Majorization and the companion notion of Schur-convexity are introduced in Section II. Zipf-like distributions are discussed in Section III. Some useful technical facts are summarized in Section IV. The basic model of cache management is given in Section V. The policy A_0 and the random policy are discussed in Sections VI and VII, respectively. The results on RORA cache policies can be found in Section VIII; Some preliminaries are briefly discussed in Appendix I. Results for the LRU and CLIMB policies are collected in Section IX with some proofs given in Appendix II. The effects of popularity on the inter-reference time, the working set size and the stack distance, are discussed in Section X, XI and XII, respectively. The paper closes with concluding remarks in Section XIII.

[3] Thus, $\mathbf{P}[R_t = i] = p(i)$ $(i = 1, \ldots, N)$ for all $t = 0, 1, \ldots$ and (1) holds with the given pmf $\boldsymbol{p}$ by the Strong Law of Large Numbers.

[4] This may not be too much of a limitation given that the IRM is the most basic request model; it is often used for checking various properties [14]. Moreover, recent results suggest some form of insensitivity to the statistics of streams of requests [15]. Of course, more work along these lines is needed.

[5] Least-Recently-Used

II. Majorization and Schur-concavity

Skewness in popularity distributions can be crisply formalized through the concept of *majorization* [16]. This notion formalizes statements concerning the relative size of components of two vectors, viz., the components $(x_1, \ldots, x_N)$ of the vector $\boldsymbol{x}$ are "more spread out" or "more balanced" than the components $(y_1, \ldots, y_N)$ of the vector $\boldsymbol{y}$: For vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^N$, we say that $\boldsymbol{x}$ is *majorized* by $\boldsymbol{y}$, and write $\boldsymbol{x} \prec \boldsymbol{y}$, whenever the conditions

$$\sum_{i=1}^{n} x_{[i]} \leq \sum_{i=1}^{n} y_{[i]}, \quad n = 1, 2, \ldots, N-1 \tag{4}$$

and

$$\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i \tag{5}$$

hold with $x_{[1]} \geq x_{[2]} \geq \ldots \geq x_{[N]}$ and $y_{[1]} \geq y_{[2]} \geq \ldots \geq y_{[N]}$ denoting the components of $\boldsymbol{x}$ and $\boldsymbol{y}$ arranged in decreasing order, respectively.

As elegantly demonstrated in the monograph of Marshall and Olkin [16], this notion has found widespread use in many diverse branches of mathematics and their applications, viz. in computer databases [19] and storage [20].

Key to the power of majorization is the companion notion of monotonicity associated with it: An $\mathbb{R}$-valued function φ defined on a set A of $\mathbb{R}^N$ is said to be *Schur-convex* (resp. *Schur-concave*) on A if

$$\varphi(\boldsymbol{x}) \leq \varphi(\boldsymbol{y}) \quad (\text{resp. } \varphi(\boldsymbol{x}) \geq \varphi(\boldsymbol{y}))$$

whenever $\boldsymbol{x}$ and $\boldsymbol{y}$ are elements in A satisfying $\boldsymbol{x} \prec \boldsymbol{y}$. If $A = \mathbb{R}^N$, then φ is simply said to be Schur-convex (resp. Schur-concave). In other words, Schur-convexity (resp. Schur-concavity) corresponds to monotone increasingness (resp. decreasingness) for majorization (viewed as a pre-order on subsets of $\mathbb{R}^N$).

With any permutation σ of $\{1, \ldots, N\}$, we associate the operator $\sigma : \mathbb{R}^N \to \mathbb{R}^N$ through the relation

$$\sigma(\boldsymbol{x}) := (x_{\sigma(1)}, \ldots, x_{\sigma(N)}), \quad \boldsymbol{x} \in \mathbb{R}^N.$$

Let $\{\sigma_i,\ i = 1, \ldots, N!\}$ be a given enumeration of all the $N!$ permutations of $\{1, \ldots, N\}$; this enumeration will be held fixed throughout the paper. A subset A of $\mathbb{R}^N$ is said to be *symmetric* if for any $\boldsymbol{x}$ in A, the element $\sigma_i(\boldsymbol{x})$ also belongs to A for *each* $i = 1, \ldots, N!$. Moreover, for any subset A of $\mathbb{R}^N$, a mapping $\varphi : A \to \mathbb{R}$ is said to be *symmetric* if A is symmetric and for any $\boldsymbol{x}$ in A, we have $\varphi(\sigma_i(\boldsymbol{x})) = \varphi(\boldsymbol{x})$ for *each* $i = 1, \ldots, N!$. If the mapping $\varphi : A \to \mathbb{R}$ is Schur-convex (resp. Schur-concave) with symmetric A, then φ is necessarily symmetric since $\sigma_i(\boldsymbol{x}) \prec \boldsymbol{x} \prec \sigma_i(\boldsymbol{x})$ implies $\varphi(\sigma_i(\boldsymbol{x})) = \varphi(\boldsymbol{x})$ for each $i = 1, \ldots, N!$.

Comparison results of the form (2) and (3) are essentially statements concerning the Schur-concavity of certain functionals. We provide an easy illustration of this idea to the entropy comparison (3). Recall that the entropy $H(\boldsymbol{p})$ of the pmf $\boldsymbol{p}$ on $\mathcal{N}$ is defined by

$$H(\boldsymbol{p}) := -\sum_{i=1}^{N} p(i) \log_2 p(i) \tag{6}$$

with the convention $t \log_2 t = 0$ for $t = 0$. By a classical result of Schur [16, C.1, p. 64] the mapping $\boldsymbol{x} \to -\sum_{i=1}^{N} x_i \log_2 x_i$ is a Schur-concave function on $\mathbb{R}_+^N$. This leads readily to the following well-known result [16, D.1, p. 71].

Proposition 1: *For pmfs $\boldsymbol{p}$ and $\boldsymbol{q}$ on $\mathcal{N}$, it holds that*

$$H(\boldsymbol{q}) \leq H(\boldsymbol{p})$$

whenever $\boldsymbol{p} \prec \boldsymbol{q}$.

Thus, majorization provides a stronger notion for comparing the imbalance in the components of pmfs than the entropy-based comparison (3) proposed by Fonseca et al. in [10].

III. Zipf-like pmfs

It has been observed in a number of studies that the popularity distribution of objects in request streams at Web caches is highly skewed. In [11] a good fit was provided by the *Zipf* distribution according to which the popularity of the i^{th} most popular object is inversely proportional to its rank, namely $1/i$. In more recent studies [14, 21], "Zipf-like" distributions[6] were found more appropriate; see [14] (and references therein) for an excellent summary. Such distributions form a one-parameter family. In our set-up, the popularity distribution $\boldsymbol{p}$ of the $\mathcal{N}$-valued rvs $\{R_t,\ t = 0, 1, \ldots\}$ is said to be Zipf-like with parameter $\alpha \geq 0$ if

$$p(i) = \frac{i^{-\alpha}}{C_\alpha(N)}, \quad i = 1, \ldots, N \tag{7}$$

with

$$C_\alpha(N) := \sum_{i=1}^{N} i^{-\alpha}. \tag{8}$$

The pmf (7) will be denoted by $\boldsymbol{p}_\alpha$. The case $\alpha = 1$ corresponds to the standard Zipf distribution. The value of α was found to be in the range $0.64 - 0.83$ [14].

Zipf-like pmfs are skewed towards the most popular objects. As $\alpha \to 0$, the Zipf-like pmf approaches the uniform distribution $\boldsymbol{u}$ while as $\alpha \to \infty$, it degenerates to the pmf $(1, 0, \ldots, 0)$. Extrapolating between these extreme cases, we expect the parameter α of Zipf-like pmfs (7)-(8) to measure the strength of skewness, with the larger α, the more skewed the pmf $\boldsymbol{p}_\alpha$. The next result can already be found in [16, B.2.b, p. 130] and shows that majorization indeed captures this fact.

Lemma 1: *For $0 \leq \alpha < \beta$, it holds that $\boldsymbol{p}_\alpha \prec \boldsymbol{p}_\beta$.*

In the spirit of the aforementioned folk theorem, we expect the miss rate of the cache replacement policy to decrease as α increases. This has been shown to be the case using simulations [22]. Zipf-like pmfs will be used in the discussion of the LRU and CLIMB policies given in Section IX.

[6] Such distributions are sometimes called generalized Zipf distributions.

IV. Some useful technical facts

We have collected in this section some useful technical results concerning Schur-convexity. We begin with some notation that will be used repeatedly: Let $\Lambda^\star(M;\mathcal{N})$ be the collection of all *unordered* subsets of size M of $\mathcal{N}=\{1,\ldots,N\}$, and let $\Lambda(M;\mathcal{N})$ be the collection of all *ordered* sequences of M *distinct* elements from $\mathcal{N}$. We write $\{i_1,\ldots,i_M\}$ (resp. $(i_1,\ldots,i_M)$) to denote an element in $\Lambda^\star(M;\mathcal{N})$ (resp. $\Lambda(M;\mathcal{N})$).

Next, as in [16, p. 78], for each $r=1,\ldots,N$, we define the *elementary symmetric* function $E_r:\mathbb{R}^N\to\mathbb{R}$ by

$$E_r(\boldsymbol{x}) := \sum_{\{i_1,\ldots,i_r\}\in\Lambda^\star(r;\mathcal{N})} x_{i_1}\cdots x_{i_r}, \quad \boldsymbol{x}\in\mathbb{R}^N. \tag{9}$$

By convention we write $E_0(\boldsymbol{x})=1$ ($\boldsymbol{x}\in\mathbb{R}^N$). It is well known [16, Prop. F.1., p. 78] that the function E_r is Schur-concave on $\mathbb{R}^N_+$.

We recall that any mapping $\varphi:A\to\mathbb{R}$ which is symmetric and convex (resp. concave) on some convex symmetric subset A of $\mathbb{R}^N$ is necessarily Schur-convex (resp. Schur-concave) [16, Prop. C.2, p. 67].

The following result is due to Schur [16, F.3, p. 80] and is key to a number of proofs.

Proposition 2: *For each $r=1,\ldots,N$, the mapping $\Phi_r:\mathbb{R}^N_+\to\mathbb{R}$ given*[7] *by*

$$\Phi_r(\boldsymbol{x}) := \frac{E_r(\boldsymbol{x})}{E_{r-1}(\boldsymbol{x})}, \quad \boldsymbol{x}\in\mathbb{R}^N_+$$

is increasing,[8] *symmetric and concave, hence increasing and Schur-concave.*

The next result is an easy byproduct of the definitions.

Proposition 3: *Let A be a convex symmetric subset of $\mathbb{R}^N$. Assume the mapping $\varphi:A\to\mathbb{R}$ to be concave and the mapping $h:\mathbb{R}^{N!}\to\mathbb{R}$ to be increasing, symmetric and concave. Then, the mapping $\varphi_h:A\to\mathbb{R}$ given by*

$$\varphi_h(\boldsymbol{x}) = h(\varphi(\sigma_1(\boldsymbol{x})),\ldots,\varphi(\sigma_{N!}(\boldsymbol{x}))), \quad \boldsymbol{x}\in A$$

is symmetric and concave, thus Schur-concave on A.

With vectors $\boldsymbol{t}$ and $\boldsymbol{x}$ in $\mathbb{R}^N$, we associate the element $\boldsymbol{t}\cdot\boldsymbol{x}$ of $\mathbb{R}^N$ with components

$$\boldsymbol{t}\cdot\boldsymbol{x} := (t_1x_1,\ldots,t_Nx_N).$$

With this notation we can state an important consequence of Proposition 3.

Proposition 4: *Assume the mapping $\psi:\mathbb{R}^N_+\to\mathbb{R}$ to be concave and the mapping $h:\mathbb{R}^{N!}\to\mathbb{R}$ to be increasing, symmetric and concave. For any non-zero vector $\boldsymbol{t}$ in $\mathbb{R}^N$, the mapping $\psi_{\boldsymbol{t}}:\mathbb{R}^N_+\to\mathbb{R}$ defined by*

$$\psi_{\boldsymbol{t}}(\boldsymbol{x}) = h(\psi(\boldsymbol{t}\cdot\sigma_1(\boldsymbol{x})),\ldots,\psi(\boldsymbol{t}\cdot\sigma_{N!}(\boldsymbol{x})))$$

for all $\boldsymbol{x}$ in $\mathbb{R}^N_+$, is symmetric and concave, thus Schur-concave.

Proof. If the mapping ψ is concave, then the mapping $\tilde{\psi}_{\boldsymbol{t}}:\mathbb{R}^N_+\to\mathbb{R}$ given by

$$\tilde{\psi}_{\boldsymbol{t}}(\boldsymbol{x}) := \psi(\boldsymbol{t}\cdot\boldsymbol{x}), \quad \boldsymbol{x}\in\mathbb{R}^N_+$$

is also concave. We obtain the desired result by applying Proposition 3 with $A=\mathbb{R}^N_+$ and $\varphi=\tilde{\psi}_{\boldsymbol{t}}$. ■

V. Demand-driven caching

The system is composed of a server where a copy of each of the N cacheable documents is available, and of a cache of size M ($1\le M<N$). Documents are first requested at the cache: If the requested document has a copy already in cache (i.e., a hit), this copy is downloaded from the cache by the user. If the requested document is not in cache (i.e., a miss), a copy is requested instead from the server to be put in the cache. If the cache is already full, then a document already in cache is evicted to make place for the copy of the document just requested. A *demand-driven* cache replacement policy (to be specified shortly) is assumed to be in use.

Consecutive user requests are modeled by a sequence of $\mathcal{N}$-valued rvs $\{R_t,\ t=0,1,\ldots\}$. For simplicity we say that request R_t occurs at time $t=0,1,\ldots$. Let S_t denote the collection of documents in cache just before time t so that S_t is a subset of $\mathcal{N}$, and let U_t denote the decision to be performed according to the cache replacement policy in force. Demand-driven caching is characterized by the dynamics

$$S_{t+1} = \begin{cases} S_t & \text{if } R_t\in S_t \\ S_t+R_t & \text{if } R_t\notin S_t, |S_t|<M \\ S_t-U_t+R_t & \text{if } R_t\notin S_t, |S_t|=M \end{cases} \tag{10}$$

where $|S_t|$ denotes the cardinality of the set S_t, and $S_t-U_t+R_t$ denotes the subset of $\{1,\ldots,N\}$ obtained from S_t by removing U_t and then adding R_t to it, *in that order*.

These dynamics reflect the following operational assumptions: (i) a requested document not in cache is always added to the cache if the cache is not full; and (ii) eviction is mandatory if the request R_t is not in cache S_t and the cache S_t is full.

As mentioned earlier, the stream of requests $\{R_t,\ t=0,1,\ldots\}$ is modeled according to the standard IRM with popularity pmf $\boldsymbol{p}=(p(1),\ldots,p(N))$. To avoid uninteresting situations, it is *always* the case that

$$p(i)>0, \quad i=1,\ldots,N. \tag{11}$$

A pmf $\boldsymbol{p}$ on $\{1,\ldots,N\}$ satisfying (11) is said to be *admissible*. Under this non-triviality condition (11), every document is eventually requested as we note that (1) holds by the Strong Law of Large Numbers.

As we have in mind to study long term characteristics under demand-driven replacement policies, there is no loss of generality in assuming (as we do from now on) that the cache is full in that $|S_t|=M$ for all $t=0,1,\ldots$, and (10) simplifies to

$$S_{t+1} = \begin{cases} S_t & \text{if } R_t\in S_t \\ S_t-U_t+R_t & \text{if } R_t\notin S_t \end{cases} \quad t=0,1,\ldots. \tag{12}$$

[7] For $\boldsymbol{x}$ in $\mathbb{R}^N_+$ such that $E_{r-1}(\boldsymbol{x})=0$, we set $\Phi_r(\boldsymbol{x})=0$ by continuity.

[8] Here, increasing means increasing in each argument.

The miss rate is then defined as the limiting constant

$$M(\boldsymbol{p}) := \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[R_\tau \notin S_\tau\right] \quad a.s. \tag{13}$$

and depends on the replacement policy in use. This limiting constant exists under the demand-driven replacement policies of interest.

VI. The policy A_0

The requests are assumed described by the IRM with popularity pmf $\boldsymbol{p}$. When at time $t = 0, 1, \ldots$, the cache S_t is full and the requested document R_t is not in the cache, the policy A_0 prescribes the eviction of U_t given by

$$U_t = \arg\min\left(p(j) : \ j \in S_t\right). \tag{14}$$

This policy is an instance of the so-called policy A_σ associated with the permutation σ of $\{1, \ldots, N\}$, whereby

$$U_t = \arg\min\left(\sigma(j) : \ j \in S_t\right). \tag{15}$$

The policy A_0 is that policy (15) associated with the *inverse* of the permutation $\sigma^\star$ of $\{1, \ldots, N\}$ which orders the components of the underlying pmf $\boldsymbol{p}$ in increasing order, namely $p(\sigma^\star(1)) \leq p(\sigma^\star(2)) \leq \ldots \leq p(\sigma^\star(N))$.

Under the IRM with popularity pmf $\boldsymbol{p}$, we can easily modify classical arguments [2, Thm. 6.4, p. 269] in order to evaluate the miss rate under policy A_σ as

$$M_\sigma(\boldsymbol{p}) = \sum_{i=M}^{N} p(\sigma(i)) - \frac{\sum_{i=M}^{N} p(\sigma(i))^2}{\sum_{i=M}^{N} p(\sigma(i))}. \tag{16}$$

That (2) indeed holds for the policy A_0 is contained in

Theorem 1: *For admissible pmfs* $\mathbf{p}$ *and* $\mathbf{q}$ *on* $\mathcal{N}$*, it holds that*

$$M_{A_0}(\boldsymbol{q}) \leq M_{A_0}(\boldsymbol{p}) \tag{17}$$

whenever $\boldsymbol{p} \prec \boldsymbol{q}$.

Proof. The policy A_0 is known [1, 2] to minimize the miss rate amongst a large class of demand-driven policies, including the policies (15). In particular, we have

$$M_{A_0}(\boldsymbol{p}) = \min_{i=1,\ldots,N!} M_{\sigma_i}(\boldsymbol{p}). \tag{18}$$

Furthermore, for any permutation σ of $\{1, \ldots, N\}$, we can rewrite (16) as

$$\begin{aligned} M_\sigma(\boldsymbol{p}) &= \frac{\left(\sum_{i=M}^{N} p(\sigma(i))\right)^2 - \sum_{i=M}^{N} p(\sigma(i))^2}{\sum_{i=M}^{N} p(\sigma(i))} \\ &= 2\frac{\sum_{i=M}^{N}\sum_{j=M}^{i-1} p(\sigma(i))p(\sigma(j))}{\sum_{i=M}^{N} p(\sigma(i))} \\ &= 2\frac{E_2(\boldsymbol{t}\cdot\sigma(\boldsymbol{p}))}{E_1(\boldsymbol{t}\cdot\sigma(\boldsymbol{p}))} = 2\Phi_2(\boldsymbol{t}\cdot\sigma(\boldsymbol{p})) \end{aligned} \tag{19}$$

where the element $\boldsymbol{t}$ of $\mathbb{R}^N$ is specified by $t_1 = \ldots = t_{M-1} = 0$ and $t_M = \ldots = t_N = 1$.

The mapping $h : \mathbb{R}^{N!} \to \mathbb{R} : \boldsymbol{y} \to \min(y_1, \ldots, y_{N!})$ is clearly increasing, symmetric and concave, while the mapping Φ_2 is concave on $\mathbb{R}_+^N$ by Proposition 2. Combining these facts with (18) and (19), we conclude by Proposition 4 that the miss rate functional under the policy A_0 is indeed Schur-concave in the pmf vector and the desired result follows. ■

VII. The Random policy

According to the random policy, when the cache is full, the document to be evicted from the cache is selected randomly according to the uniform distribution. Under the IRM with popularity pmf $\boldsymbol{p}$, the corresponding miss rate is given by [1, Thm. 11, p. 132]

$$\begin{aligned} &M_{\mathrm{Rand}}(\boldsymbol{p}) \\ &= \frac{\sum_{\{i_1,\ldots,i_M\}} p(i_1)\cdots p(i_M)\left(1 - \sum_{k=1}^{M} p(i_k)\right)}{\sum_{\{i_1,\ldots,i_M\}} p(i_1)\cdots p(i_M)} \end{aligned} \tag{20}$$

where $\sum_{\{i_1,\ldots,i_M\}}$ denotes the summation over the collection $\Lambda^\star(M;\mathcal{N})$. The analog of Theorem 1 for the random policy is simply

Theorem 2: *For admissible pmfs* $\mathbf{p}$ *and* $\mathbf{q}$ *on* $\mathcal{N}$*, it holds that*

$$M_{\mathrm{Rand}}(\boldsymbol{q}) \leq M_{\mathrm{Rand}}(\boldsymbol{p}) \tag{21}$$

whenever $\boldsymbol{p} \prec \boldsymbol{q}$.

Proof. First, we note that

$$\sum_{\{i_1,\ldots,i_M\}\in\Lambda^\star(M;\mathcal{N})} p(i_1)\cdots p(i_M) = E_M(\boldsymbol{p}). \tag{22}$$

It is also a simple matter to see that

$$\begin{aligned} &\sum_{\{i_1,\ldots,i_M\}\in\Lambda^\star(M;\mathcal{N})} p(i_1)\cdots p(i_M)(1 - \sum_{k=1}^{M} p(i_k)) \\ &= \sum_{\{i_1,\ldots,i_M\}\in\Lambda^\star(M;\mathcal{N})} p(i_1)\cdots p(i_M)\cdot \sum_{i\notin\{i_1,\ldots,i_M\}} p(i) \\ &= (M+1) \sum_{\{i_1,\ldots,i_{M+1}\}\in\Lambda^\star(M+1;\mathcal{N})} p(i_1)\cdots p(i_{M+1}) \\ &= (M+1)E_{M+1}(\boldsymbol{p}). \end{aligned} \tag{23}$$

Combining (22) and (23) through (20) we get

$$M_{\mathrm{Rand}}(\boldsymbol{p}) = (M+1)\frac{E_{M+1}(\boldsymbol{p})}{E_M(\boldsymbol{p})}, \tag{24}$$

and by Proposition 2 the miss rate $M_{\mathrm{Rand}}(\boldsymbol{p})$ is Schur-concave in $\boldsymbol{p}$. ■

Under the IRM, it is well known [1, p. 132] that the FIFO policy yields the same miss rate as the random policy, so that Theorem 2 holds for the FIFO policy as well.

VIII. Random On-demand Replacement Algorithms

The results for the policy A_0 and for the random policy can be generalized to a large class of replacement policies called *Random On-demand Replacement Algorithms* (RORA): A RORA policy follows the demand-driven caching rule (10) (under the customary assumption that the cache is initially full). We represent the cache state as an element $(i_1, \ldots, i_M)$ in $\Lambda(M;\mathcal{N})$.

The eviction rule of RORA is characterized by a pmf $\boldsymbol{r}$ which we organize as the $M \times M$ matrix $\boldsymbol{r} = (r_{k\ell})$, i.e., for each $k, \ell = 1, \ldots, M$, we have $r_{k\ell} \geq 0$ and $\sum_{k=1}^{M} \sum_{\ell=1}^{M} r_{k\ell} = 1$. The RORA associated with the pmf matrix $\boldsymbol{r}$ is denoted RORA($\boldsymbol{r}$). Suppose that the current cache is in state $S_t = (i_1, \ldots, i_M)$ (in $\Lambda(M;\mathcal{N})$). If the requested document R_t is not in cache, then with probability $r_{k\ell}$, the document i_k (document at position k) is evicted and the new document is inserted in the cache at position ℓ. If $k < \ell$, the documents $i_{k+1}, \ldots, i_\ell$ are shifted down to position $k, k+1 \ldots, \ell - 1$ while if $k > \ell$, the documents $i_\ell, \ldots, i_{k-1}$ are shifted up to position $\ell + 1, \ldots, k$. When $k = \ell$, the new document simply replaces the evicted document at position k.

RORAs constitute a large class of replacement algorithms which contains many known policies: The random policy corresponds to RORA($\boldsymbol{r}$) with $\boldsymbol{r}$ given by $r_{kk} = \frac{1}{M}$ for each $k = 1, \ldots, M$, while the FIFO algorithm is associated with two possibilities for $\boldsymbol{r}$, either $r_{1M} = 1$ or $r_{M1} = 1$. Lastly, the Partially Preloaded Random Replacement Algorithms proposed by Gelenbe [23] also form a subclass of RORA.

RORAs fall into one of two classes. To define them, we observe that the document initially at position i will *never* be replaced if and only if

$$r_{k\ell} = 0 \quad \text{for } k \leq i \leq \ell \quad \text{and} \quad \ell \leq i \leq k. \tag{25}$$

If we use row i and column i to partition the matrix $\boldsymbol{r}$ into four blocks, then condition (25) expresses the fact that the entries in the northwest and southeast corners *all* vanish (including row i and column i). Let Σ denote the set of positions in the cache with the property that any document initially put there will never be evicted during the operation of the cache, i.e.,

$$\Sigma := \{i = 1, \ldots, M : \text{ Eqn. (25) holds at } i\}. \tag{26}$$

Case 1 – The set Σ *empty*, so that every document in cache can be replaced. This will be the case for the random and FIFO policies. In Appendix I, we show that the miss rate can be written as

$$M_{\boldsymbol{r}}(\boldsymbol{p}) = (M+1)\frac{E_{M+1}(\boldsymbol{p})}{E_M(\boldsymbol{p})}. \tag{27}$$

Because this expression is identical with that for M_{Rand} in (24), we readily obtain

Theorem 3: *Under Case 1, for admissible pmfs* $\boldsymbol{p}$ *and* $\boldsymbol{q}$ *on* $\mathcal{N}$*, it holds that*

$$M_{\boldsymbol{r}}(\boldsymbol{q}) \leq M_{\boldsymbol{r}}(\boldsymbol{p}) \tag{28}$$

whenever $\boldsymbol{p} \prec \boldsymbol{q}$.

Case 2 – The set Σ is *not* empty, and some documents, once put in cache, is never replaced during the operation of the cache. Consider for instance the matrix $\boldsymbol{r}$ of the form $r_{kk} = 1$ for some $k = 1, \ldots, M$, in which case Σ contains $M - 1$ elements, namely $\{1, \ldots, k-1, k+1, \ldots, M\}$. For any permutation σ of $\{1, \ldots, N\}$, if the documents $\sigma(1), \ldots, \sigma(M-1)$ are initially put in cache (i.e., preloaded) at the other positions $\ell \neq k$, this RORA($\boldsymbol{r}$) policy will behave like the policy A_σ in steady state. With initial cache state s_0 in $\Lambda(M;\mathcal{N})$, we denote by $\Sigma(s_0)$ the set of initial documents with positions in Σ. The documents in $\Sigma(s_0)$ is never replaced during the operation of the cache.

If the set Σ is non-empty with $|\Sigma| = m$ for some $m = 1, \ldots, M-1$, then the miss rate is shown in Appendix I to be given by

$$M_{\boldsymbol{r}}(\boldsymbol{p}; s_0) = (M - m + 1)\frac{E_{M-m+1}(\boldsymbol{t} \cdot \boldsymbol{p})}{E_{M-m}(\boldsymbol{t} \cdot \boldsymbol{p})} \tag{29}$$

where the element $\boldsymbol{t}$ in $\mathbb{R}^N$ is specified by $t_i = 0$ for i being a document in $\Sigma(s_0)$ and $t_i = 1$ otherwise. The documents in $\Sigma(s_0)$ do not contribute to the miss rate since they never generate a miss once loaded in cache. It is easy to see that for any two initial cache states s_0 and s_0' in $\Lambda(M;\mathcal{N})$ with $\Sigma(s_0) = \Sigma(s_0')$, we have $M_{\boldsymbol{r}}(\boldsymbol{p}; s_0) = M_{\boldsymbol{r}}(\boldsymbol{p}; s_0')$. Hence, we shall find it appropriate to denote this common value by $M_{\boldsymbol{r},\Sigma(s_0)}(\boldsymbol{p})$.

Let $\Sigma^\star(\boldsymbol{p})$ denote the set of the m most popular documents for the pmf $\boldsymbol{p}$. Equipped with the expression (29), we are now ready to establish

Theorem 4: *Under Case 2 with* $|\Sigma| = m$*, for admissible pmfs* $\boldsymbol{p}$ *and* $\boldsymbol{q}$ *on* $\mathcal{N}$*, it holds that*

$$M_{\boldsymbol{r},\Sigma^\star(\boldsymbol{q})}(\boldsymbol{q}) \leq M_{\boldsymbol{r},\Sigma^\star(\boldsymbol{p})}(\boldsymbol{p}) \tag{30}$$

whenever $\boldsymbol{p} \prec \boldsymbol{q}$.

Proof. Consider a RORA($\boldsymbol{r}$) policy with $|\Sigma| = m$ for some $m = 1, \ldots, M-1$, We need to show that the miss rate function $M_{\boldsymbol{r},\Sigma(s_0)}(\boldsymbol{p})$ in (29) is Schur-concave whenever s_0 is selected so that $\Sigma(s_0) = \Sigma^\star(\boldsymbol{p})$. As we can always relabel the documents, there is no loss of generality in assuming $p(1) \geq p(2) \geq \ldots \geq p(N)$, whence $\Sigma^\star(\boldsymbol{p}) = \{1, \ldots, m\}$ and the element $\boldsymbol{t}$ in (29) can be specified as $t_1 = \ldots = t_m = 0$ and $t_{m+1} = \ldots = t_N = 1$.

By Proposition 2, the mapping $\frac{E_{M-m+1}}{E_{M-m}}$ is increasing and Schur-concave on $\mathbb{R}_+^N$, and by virtue of the defining property of $\Sigma^\star(\boldsymbol{p})$, we have

$$\begin{aligned} & M_{\boldsymbol{r},\Sigma^\star(\boldsymbol{p})}(\boldsymbol{p}) \\ &= \min_{i=1,\ldots,N!} (M - m + 1)\frac{E_{M-m+1}(\boldsymbol{t} \cdot \sigma_i(\boldsymbol{p}))}{E_{M-m}(\boldsymbol{t} \cdot \sigma_i(\boldsymbol{p}))} \end{aligned} \tag{31}$$

with the element $\boldsymbol{t}$ defined above. The expression (31) is similar to (18) and (19) given in the proof of Theorem 1 with 2 replaced by $M - m + 1$ and the desired result readily follows by similar arguments. ■

IX. The LRU and CLIMB policies

The LRU policy evicts the document which was requested the least recently at the time the replacement is required. The CLIMB policy ranks documents in cache according to their recency of access: If the requested document is not in the cache, the document at the last position (position M) is evicted and replaced by the new document. If the requested document is in the cache at position i, $i = 2, \ldots, M$, it exchanges position with the document at position $i-1$. The cache remains unchanged if the requested document is in position 1.

The miss rates of the LRU and CLIMB policies have been evaluated under the IRM with popularity pmf $\boldsymbol{p}$ [1, Chap. 4]. We have the expressions

$$M_{\rm LRU}(\boldsymbol{p}) = \sum_{(i_1,\ldots,i_M)\in\Lambda(M;\mathcal{N})} \frac{\prod_{\ell=1}^{M} p(i_\ell)\left(1-\sum_{j=1}^{M} p(i_j)\right)}{\prod_{k=1}^{M-1}(1-\sum_{j=1}^{k} p(i_j))} \tag{32}$$

and

$$M_{\rm CL}(\boldsymbol{p}) = \frac{\sum_{(i_1,\ldots,i_M)} \prod_{\ell=1}^{M} p(i_\ell)^{M-\ell+1}\left(1-\sum_{j=1}^{M} p(i_j)\right)}{\sum_{(i_1,\ldots,i_M)} \prod_{\ell=1}^{M} p(i_\ell)^{M-\ell+1}}, \tag{33}$$

where the summation $\sum_{(i_1,\ldots,i_M)}$ is taken over the set $\Lambda(M;\mathcal{N})$.

Contrary to what transpired with the policy A_0 and the random policy, the miss rate for either the LRU or CLIMB policies is *not* Schur-concave in general, and consequently the folk theorem (2) may fail to hold. This is demonstrated through the following example developed for $M = 3$ and $N = 4$:

In this case, simple algebraic manipulations transform the expressions (32) and (33) into the simpler expressions

$$M_{\rm LRU}(\boldsymbol{p}) = \sum_{(i_1,i_2)\in\Lambda(2;\mathcal{N})} \frac{2\prod_{i=1}^{4} p(i)}{\prod_{k=1}^{2}(1-\sum_{j=1}^{k} p(i_j))} \tag{34}$$

and

$$M_{\rm CL}(\boldsymbol{p}) = \frac{2\prod_{j=1}^{4} p(j)\left(\sum_{i=1}^{4} p(i)^2(1-p(i))\right)}{\sum_{(i_1,i_2,i_3)\in\Lambda(3;\mathcal{N})} p(i_1)^3 p(i_2)^2 p(i_3)}, \tag{35}$$

respectively.

We evaluate numerically the expressions (34) and (35) for the family of pmfs

$$\boldsymbol{p}(x,y) = (x, 1-2y-x, y, y), \quad 0 < y < \frac{1}{4} \tag{36}$$

with x in the interval $[\frac{1}{2}-y, 1-3y]$. Under these constraints, the components of the pmf $\boldsymbol{p}(x,y)$ are listed in decreasing order and for any given y it holds that $\boldsymbol{p}(x,y) \prec \boldsymbol{p}(x',y)$ whenever $x < x'$ in the interval $[\frac{1}{2}-y, 1-3y]$. Therefore, if the miss rates under LRU and CLIMB *were* indeed Schur-concave functions in the popularity pmf, we would expect the functions $x \to M_{\rm LRU}(\boldsymbol{p}(x,y))$ and $x \to M_{\rm CL}(\boldsymbol{p}(x,y))$ to be monotone *decreasing* on the interval $[\frac{1}{2}-y, 1-3y]$.

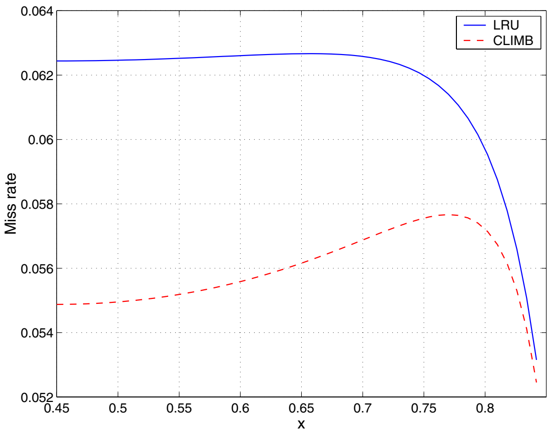

Fig. 1. LRU and CLIMB miss rates when $M = 3$, $N = 4$, $y = p(3) = p(4) = 0.05$, $p(1) = x$ and $p(2) = 0.9 - p(1)$

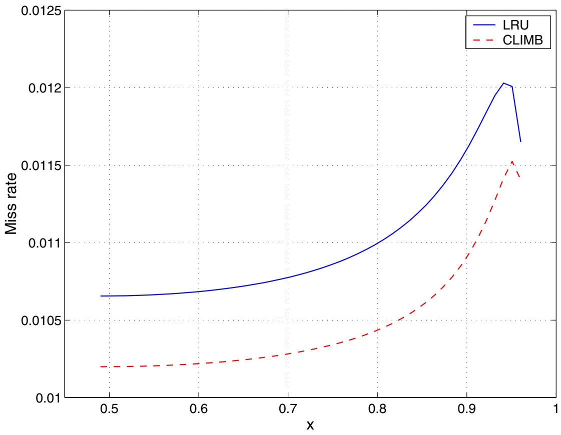

Fig. 2. LRU and CLIMB miss rates when $M = 3$, $N = 4$, $y = p(3) = p(4) = 0.01$, $p(1) = x$ and $p(2) = 0.98 - p(1)$

Figures 1 and 2 display the numerical values of $M_{\rm LRU}(\boldsymbol{p}(x,y))$ and $M_{\rm CL}(\boldsymbol{p}(x,y))$ as a function of x with $y = 0.05$ and $y = 0.01$, respectively. In both cases, the miss rates of the LRU and CLIMB policies are *not* monotone decreasing in x on the entire range $[\frac{1}{2}-y, 1-3y]$, with the trend becoming more pronounced with decreasing y.

While the miss rate is not always Schur-concave under the LRU and CLIMB policies, in the case of Zipf-like popularity distributions, the desired monotonicity (2) is nevertheless true in the following asymptotic sense.

Theorem 5: *Assume the input to have a Zipf-like popularity pmf $\boldsymbol{p}_\alpha$ for some $\alpha \geq 0$. Then, there exists $\alpha^\star = \alpha^\star(M,N)$ such that for $\alpha > \beta > \alpha^\star$, we have $M_{\rm LRU}(\boldsymbol{p}_\alpha) < M_{\rm LRU}(\boldsymbol{p}_\beta)$ and $M_{\rm CL}(\boldsymbol{p}_\alpha) < M_{\rm CL}(\boldsymbol{p}_\beta)$.*

A proof of Theorem 5 is available in Appendix II. We have also carried out simulations of a cache operating under the LRU and CLIMB policies when the input has a Zipf-like popularity pmf $\boldsymbol{p}_\alpha$. The number of documents is set at $N = 1,000$ while the cache size is $M = 100$. The miss rates of both policies are displayed in Figure 3 and 4 for α small

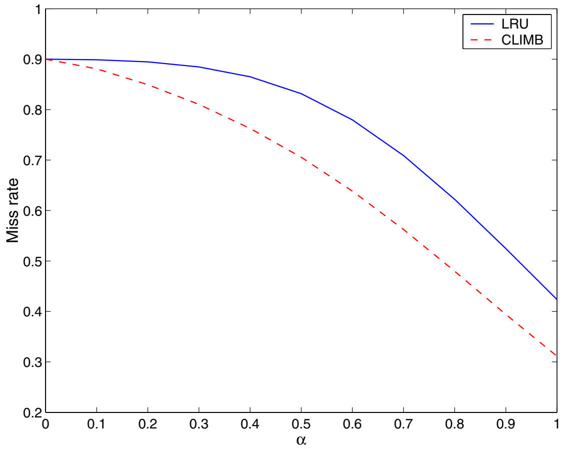

Fig. 3. LRU and CLIMB miss rates when the input has a Zipf-like popularity pmf $\boldsymbol{p}_\alpha$ for α small ($0 \leq \alpha \leq 1$)

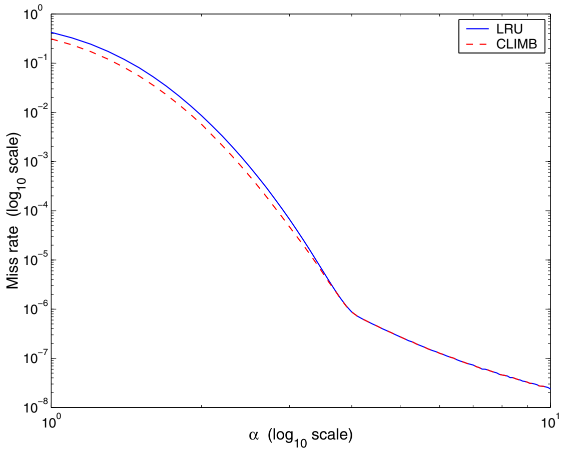

Fig. 4. LRU and CLIMB miss rates when the input has a Zipf-like popularity pmf $\boldsymbol{p}_\alpha$ for α large ($\alpha > 1$)

($0 \leq \alpha \leq 1$) and α large ($\alpha > 1$), respectively. It appears that the miss rate is indeed decreasing as the skewness parameter α increases across the entire range of α. This suggests that the folk theorem should hold for the LRU and CLIMB policies when the comparison is made within the class of Zipf-like popularity pmfs. Work is in progress on this issue.

X. The Inter-reference time

In the next three sections, we turn to the discussion of how majorization of popularity pmfs translates into comparisons of three well-established metrics for locality of reference, namely, the inter-reference time, the working set size and the stack distance, in that order. We begin in this section with the notion of inter-reference time in the stream of requests, a notion which has recently received some attention as a way of characterizing temporal correlations [3, 5, 10].

First a definition. Given an IRM with popularity pmf $\boldsymbol{p}$, we define the inter-reference time $T(\boldsymbol{p})$ as the rv given by

$$T(\boldsymbol{p}) := \inf\{t = 1, 2, \ldots : R_t = R_0\}. \tag{37}$$

Our main comparison result for inter-reference times is given in terms of the convex ordering[9] [24]:

Theorem 6: *For admissible pmfs $\boldsymbol{p}$ and $\boldsymbol{q}$ on $\mathcal{N}$, it holds that*

$$T(\boldsymbol{p}) \leq_{cx} T(\boldsymbol{q}) \tag{38}$$

whenever $\boldsymbol{p} \prec \boldsymbol{q}$.

Thus, the more skewed the popularity pmf, the stronger the locality of reference in the IRM, and the more variable the inter-reference time!

Proof. It is well known [24, Thm. 2.A.1, p. 57] that the comparison (38) between the $\{1, 2, \ldots\}$-valued rvs $T(\boldsymbol{p})$ and $T(\boldsymbol{q})$ is equivalent to

$$\sum_{k=n}^{\infty} \mathbf{P}\left[T(\boldsymbol{p}) \geq k\right] \leq \sum_{k=n}^{\infty} \mathbf{P}\left[T(\boldsymbol{q}) \geq k\right] \tag{39}$$

for *all* $n = 1, 2, \ldots$, with

$$\mathbf{E}\left[T(\boldsymbol{p})\right] = \mathbf{E}\left[T(\boldsymbol{q})\right]. \tag{40}$$

Consider a given pmf $\boldsymbol{p}$ on $\mathcal{N}$ and fix $i = 1, \ldots, N$. For each $t = 1, 2, \ldots$, we note that

$$\mathbf{P}\left[T(\boldsymbol{p}) = t | R_0 = i\right] = (1 - p(i))^{t-1} p(i),$$

i.e., conditional on $R_0 = i$, the inter-reference time $T(\boldsymbol{p})$ is geometrically distributed with parameter $p(i)$. Consequently, for each $n = 1, 2, \ldots$, we find

$$\begin{aligned} \mathbf{P}\left[T(\boldsymbol{p}) \geq n | R_0 = i\right] &= \sum_{t=n}^{\infty} \mathbf{P}\left[T(\boldsymbol{p}) = t | R_0 = i\right] \\ &= (1 - p(i))^{n-1}, \end{aligned}$$

whence

$$\mathbf{P}\left[T(\boldsymbol{p}) \geq n\right] = \sum_{i=1}^{N} p(i)(1 - p(i))^{n-1}.$$

Next, we obtain

$$\psi_n(\boldsymbol{p}) := \sum_{k=n}^{\infty} \mathbf{P}\left[T(\boldsymbol{p}) \geq k\right] = \sum_{i=1}^{N} (1 - p(i))^{n-1}.$$

In particular, with $n = 1$, this last calculation yields

$$\mathbf{E}\left[T(\boldsymbol{p})\right] = \sum_{k=1}^{\infty} \mathbf{P}\left[T(\boldsymbol{p}) \geq k\right] = N,$$

and this independently of $\boldsymbol{p}$! In other words, (40) holds.

It is a simple matter to see that for each $n = 1, 2, \ldots$, the mapping $t \to (1-t)^{n-1}$ is convex on $\mathbb{R}_+$. By a classical result of Schur [16, C.1, p. 64] the mapping $\boldsymbol{x} \to \sum_{i=1}^{N} (1 - x_i)^{n-1}$ is a Schur-convex function on $\mathbb{R}_+^N$. To put it differently, the mapping $\boldsymbol{p} \to \psi_n(\boldsymbol{p})$ is Schur-convex, and (39) indeed holds when $\boldsymbol{p} \prec \boldsymbol{q}$. ■

[9] Recall that for $\mathbf{R}$-valued rvs X and Y, Y is greater than X in the convex ordering, written $X \leq_{cx} Y$ if $\mathbf{E}\left[\varphi(X)\right] \leq \mathbf{E}\left[\varphi(Y)\right]$ for any convex mapping $\varphi : \mathbf{R} \to \mathbf{R}$ for which the expectations are well defined.

XI. The working set size

Consider an IRM request stream $\{R_t, t = 0, 1, \ldots\}$ with popularity pmf $\boldsymbol{p}$. Fix $t = 0, 1, \ldots$. For each $\tau = 1, 2, \ldots$, we define the working set $W(t;\tau)$ of length τ (starting at time t) to be the set of *distinct* documents occurring amongst the next τ consecutive requests $\{R_t, \ldots, R_{t+\tau-1}\}$. The size $|W(t;\tau)|$ of the working set $W(t;\tau)$ is denoted by $S(t;\tau)$.[10] Under the enforced i.i.d. assumption on the request stream, the pmf of the rv $S(t;\tau)$ does not depend on t. To recognize this fact, we write $S(\tau;\boldsymbol{p})$ to represent the number of distinct requested documents in τ timeslots under the IRM with popularity pmf $\boldsymbol{p}$.

For positive integer $n = 1, 2, \ldots$ and pmf $\boldsymbol{\theta} = (\theta(1), \ldots, \theta(N))$ on $\{1, \ldots, N\}$, it is customary to imagine the following experimental setup: An experiment has N distinct outcomes, outcome i occurring with probability $\theta(i)$ ($i = 1, \ldots, N$). We carry out this experiment n times under independent and statistically identical conditions. Let $X_i(n,\boldsymbol{\theta})$ denote the number of times that outcome i occurs amongst these n trials ($i = 1, \ldots, N$). These N rvs are organized into an $\mathbb{N}^N$-valued rv $\boldsymbol{X}(n,\boldsymbol{\theta})$ known as the *multinomial* rv with parameters n and $\boldsymbol{\theta}$. Its distribution is given by

$$\mathbf{P}\left[\boldsymbol{X}(n,\boldsymbol{\theta}) = \boldsymbol{x}\right] = \left(\begin{array}{c} n \\ x_1, \ldots, x_N \end{array} \right) \cdot \prod_{i=1}^{N} \theta(i)^{x_i}$$

whenever the integer components $(x_1, \ldots, x_N)$ of $\boldsymbol{x}$ satisfy $x_i \geq 0$ ($i = 1, \ldots, N$) and $\sum_{i=1}^N x_i = n$.

With $\boldsymbol{X}(n,\boldsymbol{\theta})$, we can associate the rv $K(n,\boldsymbol{\theta})$ given by

$$K(n,\boldsymbol{\theta}) := \sum_{i=1}^{N} \mathbf{1}\left[X_i(n,\boldsymbol{\theta}) > 0\right];$$

this rv records the number of *distinct* outcomes that occur amongst the n trials. The following result was established by Wong and Yue [25] and deals with the Schur-concavity of the tails probabilities

$$\pi_\ell(n,\boldsymbol{\theta}) := \mathbf{P}\left[K(n,\boldsymbol{\theta}) > \ell\right], \quad \ell = 0, 1, \ldots, \min(N, n).$$

Theorem 7: *For each* $n = 1, 2, \ldots$ *and each* $\ell = 1, 2, \ldots, \min(N, n)$, *the mapping* $\boldsymbol{\theta} \to \pi_\ell(n,\boldsymbol{\theta})$ *is Schur-concave.*

The working set size $S(\tau;\boldsymbol{p})$ of the IRM request stream with popularity pmf $\boldsymbol{p}$ is simply the number of distinct outcomes $K(\tau,\boldsymbol{p})$ for the multinomial rv with parameters τ and $\boldsymbol{p}$. Thus, as a direct implication of Theorem 7, we obtain the following corollary using [24, p. 3].[11]

Corollary 1: *For admissible pmfs* $\boldsymbol{p}$ *and* $\boldsymbol{q}$ *on* $\mathcal{N}$, *it holds that*

$$S(\tau;\boldsymbol{q}) \leq_{st} S(\tau;\boldsymbol{p}), \quad \tau = 1, 2, \ldots,$$

whenever $\boldsymbol{p} \prec \boldsymbol{q}$.

In words, the more skewed the popularity pmf, the stronger the locality of reference in the IRM, and the smaller (in the strong stochastic sense) the working set size, in line with one's intuition!

XII. The stack distance

The notion of stack distance has been widely used as a metric for temporal correlations [11, 12, 13]: The stack distance of the IRM request stream with popularity pmf $\boldsymbol{p}$ is the rv $D(\boldsymbol{p})$ defined by

$$D(\boldsymbol{p}) = |\{R_0, , \ldots, R_{T(\boldsymbol{p})}\}| \tag{41}$$

where $T(\boldsymbol{p})$ is the inter-reference time (37); this rv records the number of *distinct* documents requested from time $t = 0$ until that time when the initial request R_0 is made again for the first time.[12]

It is not difficult to see that the stochastic equivalence

$$D(\boldsymbol{p}) =_{st} S(\tau;\boldsymbol{p})_{\tau=T(\boldsymbol{p})}$$

holds. Hence, in view of the results obtained in Corollary 1, one might expect the comparison

$$D(\boldsymbol{q}) \leq_{st} D(\boldsymbol{p}) \tag{42}$$

to hold whenever the pmfs $\boldsymbol{p}$ and $\boldsymbol{q}$ on $\mathcal{N}$ satisfy $\boldsymbol{p} \prec \boldsymbol{q}$. However, the comparison (42) can not be established as we explain below: Indeed, it is known [26, 27] that the stack distance is related to the miss rate of the LRU replacement policy. Specifically, given an IRM request stream with popularity pmf $\boldsymbol{p}$, the miss rate $M_{\mathrm{LRU}}(\boldsymbol{p})$ of LRU with cache size M can be expressed in terms of the tail distribution of $D(\boldsymbol{p})$ through

$$M_{\mathrm{LRU}}(\boldsymbol{p}) = \mathbf{P}\left[D(\boldsymbol{p}) > M\right].$$

But, in Section IX, we have seen that it is possible to find pmfs $\boldsymbol{p}$ and $\boldsymbol{q}$ on $\mathcal{N}$ such that $\boldsymbol{p} \prec \boldsymbol{q}$ and yet $M_{\mathrm{LRU}}(\boldsymbol{p}) < M_{\mathrm{LRU}}(\boldsymbol{q})$, or equivalently, $\mathbf{P}\left[D(\boldsymbol{p}) > M\right] < \mathbf{P}\left[D(\boldsymbol{q}) > M\right]$. In short, the comparison (42) does not hold in general [24, p. 3].

Although somewhat annoying from the point of view of intuition, this state of affairs is perhaps not too surprising given the opposite direction of the comparison of inter-reference times in Theorem 6. It is possible that some comparison other than (42) might hold, say in the increasing concave ordering[13] [24], i.e., for pmfs $\boldsymbol{p}$ and $\boldsymbol{q}$ on $\mathcal{N}$ such that $\boldsymbol{p} \prec \boldsymbol{q}$,

$$D(\boldsymbol{q}) \leq_{icv} D(\boldsymbol{p}). \tag{43}$$

This comparison is compatible with the *weaker* result of Yue and Wong [20] that $\mathbf{E}\left[D(\boldsymbol{q})\right] < \mathbf{E}\left[D(\boldsymbol{p})\right]$ whenever $\boldsymbol{p} \prec \boldsymbol{q}$.

[10] A slightly different definition of the working set is usually adopted in the literature [8, 9], namely with $\tau \leq t$, the working set $W(t;\tau)$ of length τ (ending at time t) is defined as the set of *distinct* documents occurring amongst the last τ requests $\{R_{t-\tau+1}, \ldots, R_t\}$ up to time t. Under the IRM assumption, this "backward in time" definition is stochastically equivalent to the "forward in time" definition given here since the pmf of the rv $S(t;\tau)$ does not depend on t whenever $t \geq \tau$.

[11] For $\mathbf{R}$-valued rvs X and Y, Y is greater than X in the usual stochastic ordering, written $X \leq_{st} Y$, if $\mathbf{E}\left[\varphi(X)\right] \leq \mathbf{E}\left[\varphi(Y)\right]$ for any monotone non-decreasing mapping $\varphi : \mathbf{R} \to \mathbf{R}$ for which the expectations are well defined.

[12] As for the notions of working set and its size, the stack distance is usually given through a "backward in time" definition. These two definitions coincide under the i.i.d. assumption enforced on the request stream.

[13] For $\mathbf{R}$-valued rvs X and Y, Y is greater than X in the increasing concave ordering, written $X \leq_{icv} Y$, if $\mathbf{E}\left[\varphi(X)\right] \leq \mathbf{E}\left[\varphi(Y)\right]$ for any increasing and concave mapping $\varphi : \mathbf{R} \to \mathbf{R}$ for which the expectations are well defined.

XIII. Concluding remarks

Under the assumption that the request stream is modeled by IRM, we have used the concepts of majorization and Schur-concavity to formalize the "folk theorem" that the stronger the locality of reference, the smaller the miss rate of the cache. This folk theorem was shown to hold for a large class of replacement policies, including the random and FIFO policies, as well as the optimal policy A_0. However, it fails to hold in general for the (popular) LRU and CLIMB policies. This suggests that popularity alone may not be strong enough to capture the *operational* meaning of locality of reference.

The results obtained here are based on the basic IRM which exhibits neither temporal nor spatial correlations. It would be desirable to explore these issues for models with correlations in order to further understand the operational meaning of locality of reference to demand-driven caching.

References

[1] O.I. Aven, E.G. Coffman and Y.A. Kogan, *Stochastic Analysis of Computer Storage*, D. Reidel Publishing Company, Dordrecht (Holland), 1987.

[2] E. Coffman and P. Denning, *Operating Systems Theory*, Prentice-Hall, NJ, 1973.

[3] V. Phalke and B. Gopinath, "An interference gap model for temporal locality in program behavior," in Proceedings of ACM SIGMETRICS'1995, May 1995, pp. 291–300.

[4] L.A. Belady, "A study of replacement algorithms for a virtual-storage computer," *IBM Systems Journal* **5** (1966), pp. 78–101.

[5] S. Jin and A. Bestavros, "Sources and characteristics of Web temporal locality," in Proceedings of MASCOTS'2000, San Francisco (CA), August 2000.

[6] S. Jin and A. Bestavros, "Temporal locality in Web request streams: Sources, characteristics, and caching implications" (Extended Abstract), in Proceedings of ACM SIGMETRICS'2000, Santa Clara (CA), June 2000.

[7] A. Mahanti, C. Williamson and D. Eager, "Temporal locality and its impact on Web proxy cache performance," *Performance Evaluation* **42** (2000), Special Issue on Internet Performance Modelling, pp. 187–203.

[8] P.J. Denning, "The working set model for program behavior," *Communications of the ACM* **11** (1968), pp. 323–333.

[9] P.J. Denning and S.S. Schwartz, "Properties of the working set model," *Communications of the ACM* **15** (1972), pp. 191–198.

[10] R. Fonseca, V. Almeida, M. Crovella and B. Abrahao, "On the intrinsic locality of Web reference streams," in Proceedings of IEEE INFOCOM 2003, San Francisco (CA), April 2003.

[11] V. Almeida, A. Bestavros, M. Crovella and A. de Oliveira, "Characterizing reference locality in the Web," in Proceedings of PDIS'96, December 1996, Miami (FL), pp. 92–107.

[12] A. Balamash and M. Krunz, "Application of multifractals in the characterization of WWW traffic," in Proceedings of ICC 2002, April 2002.

[13] R.L. Mattson, J. Gecsei, D.R. Slutz and L. Traiger, "Evaluation techniques for storage hierarchies," *IBM Systems Journal* **9** (1970), pp. 78–117.

[14] L. Breslau, P. Cao, L. Fan, G. Phillips and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in Proceedings of IEEE INFOCOM 1999, New York (NY), March 1999.

[15] P. Jelenkovic and A. Radovanovic, "Asymptotic insensitivity of Least-Recently-Used caching to statistical dependency," in Proceedings of IEEE INFOCOM 2003, San Francisco (CA), April 2003.

[16] A.W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York (NY), 1979.

[17] S. Vanichpun and A.M. Makowski, "The output of a cache under the Independent Reference Model – Where did the locality of reference go?," submitted for inclusion in the program of Performance 2004, New York (NY), June 2004.

[18] A.M. Makowski and S. Vanichpun, "Comparing strength of locality of reference – Popularity, majorization, and some folk theorems for miss rates and the output of cache," in *Performance Evaluation and Planning Methods for the Next Generation Internet*, A. Girard, B. Sansó and F. J. Vázquez-Abad, Editors, Kluwer Academic Press.

[19] S. Christodoulakis, "Implications of certain assumptions in database performance evaluation," *ACM Transactions on Database Systems* **9** (1984), pp. 163–186.

[20] P.C. Yue and C.K. Wong, "On the optimality of the probability ranking scheme in storage applications," *Journal of the ACM* **20** (1973), pp. 624–633.

[21] S. Jin and A. Bestavros, "GreedyDual* Web caching algorithm: Exploiting the two sources of temporal locality in Web request streams," in Proceedings of the 5th International Web Caching and Content Delivery Workshop, Lisbon, Portugal, May 2000.

[22] S. Gadde, J.S. Chase and M. Rabinovich, "Web caching and content distribution: A view from the interior," *Computer Communications* **24** (2001), pp. 222–231.

[23] E. Gelenbe, "A unified approach to the evaluation of a class of replacement algorithms," *IEEE Transactions on Computers* **22** (1973), pp. 611–618.

[24] M. Shaked and J.G. Shanthikumar, *Stochastic Orders and Their Applications*, Academic Press, San Diego (CA), 1994.

[25] C.K. Wong and P.C. Yue, "A majorization theorem for the number of distinct outcomes in N independent trials," *Discrete Mathematics* **6** (1973), pp. 391–398.

[26] P. Flajolet, D. Gardy and L. Thimonier, "Birthday paradox, coupon collector, caching algorithms and self-organizing search," *Discrete Applied Mathematics* **39** (1992), pp. 207–229.

[27] P.R. Jelenkovic, "Asymptotic approximation of the Move-To-Front search cost distribution and Least-Recently-Used caching fault probabilities," *Annals of Applied Probability* **9** (1999), pp. 420–469.

Appendix I
Some expressions for RORA policies

Considera RORA with pmf matrix $\boldsymbol{r}$ and let $\{\Omega_t,\ t = 0, 1, \ldots\}$ denote the sequence of cache states under RORA (with the cache initially full as explained earlier).[14] Introduce a sequence of i.i.d. rvs $\{(X_t, Y_t),\ t = 0, 1, \ldots\}$ taking values in $\{1, \ldots, M\} \times \{1 \ldots, M\}$ with common pmf $\boldsymbol{r}$, i.e., for each $t = 0, 1, \ldots$,

$$\mathbf{P}\left[(X_t, Y_t) = (k, \ell)\right] = r_{k\ell}, \quad k, \ell = 1, \ldots, M.$$

The sequences of rvs $\{(X_t, Y_t),\ t = 0, 1, \ldots\}$ and $\{R_t,\ t = 0, 1, \ldots\}$ are assumed mutually independent. Under RORA, the document U_t to be evicted at time t is given by

$$U_t = \mathbf{1}\left[R_t \notin S_t\right] \Omega_{t, X_t}$$

where $\Omega_{t,k}$ denotes the document in cache at position $k = 1, \ldots, M$. If $U_t \neq 0$, the new document is inserted at position Y_t while if $U_t = 0$, no replacement occurs. Under the IRM, the cache states $\{\Omega_t, t = 0, 1, \ldots\}$ is easily seen to form a Markov Chain on the state space $\Lambda(M; \mathcal{N})$.

The irreducibility properties of this Markov chain are determined by the eviction/insertion matrix $\boldsymbol{r}$. With this in mind, let $\mathcal{S}(\boldsymbol{r}, s_0)$ denote the irreducible component that is reachable from the initial cache state s_0 in $\Lambda(M; \mathcal{N})$. On this component $\mathcal{S}$[15], the Markov chain $\{\Omega_t,\ t = 0, 1, \ldots\}$ is ergodic; its stationary distribution exists and is given by

$$\pi(s) = C^{-1} p(i_1) p(i_2) \cdots p(i_M) \tag{44}$$

[14] Observe that the set S_t of documents in cache at time t is recoverable from the cache state Ω_t.

[15] We have suppressed the dependence on $\boldsymbol{r}$ and s_0 for notational simplicity.

for each cache state $s = (i_1, \ldots, i_M)$ (in $\mathcal{S}$) with normalizing constant given by

$$C = \sum_{s\in\mathcal{S}} p(i_1)p(i_2)\cdots p(i_M). \tag{45}$$

This is readily verified using the Global Balance Equation [18]. Moreover, if we denote by $M_{\boldsymbol{r}}(\boldsymbol{p}; s_0)$ the miss rate achieved under RORA($\boldsymbol{r}$) when starting in cache state $\Omega_0 = s_0$, we find

$$\begin{aligned} M_{\boldsymbol{r}}(\boldsymbol{p}; s_0) &= \lim_{t\to\infty}\frac{1}{t}\sum_{\tau=1}^{t} \mathbf{1}\left[R_\tau \notin S_\tau\right] \quad a.s. \\ &= \sum_{s\in\mathcal{S}}\pi(s)\sum_{i\notin s} p(i) \end{aligned} \tag{46}$$

where $i \notin s$ denotes the set of elements in $\mathcal{N}$ which are not in s [18].

The exact form of the stationary distribution depends on $\mathcal{S}$, thus on $\boldsymbol{r}$ and on any initial condition s_0 that gives rise to $\mathcal{S}$. We recall the definition (25) of the set Σ of non-replaceable positions, and we refer the reader to the discussion in Section VIII before embarking on the derivation of the stationary distribution and miss rate under the two basic cases. Additional details are available elsewhere [18].

Case 1 – The set Σ being empty, the Markov chain has exactly one irreducible component $\mathcal{S} = \Lambda(M;\mathcal{N})$, and the stationary distribution is given by (44) and (45). By reporting (44) and (45) with $\mathcal{S} = \Lambda(M;\mathcal{N})$ into (46), we find

$$\begin{aligned} & M_{\boldsymbol{r}}(\boldsymbol{p}; s_0) \\ &= C^{-1} \sum_{(i_1,\ldots,i_M)\in\Lambda(M;\mathcal{N})} p(i_1)\cdots p(i_M) \sum_{i\notin\{i_1,\ldots,i_M\}} p(i) \\ &= C^{-1} \sum_{(i_1,\ldots,i_{M+1})\in\Lambda(M+1;\mathcal{N})} p(i_1)\cdots p(i_{M+1}). \end{aligned} \tag{47}$$

As in the proof of Theorem 2, we note the relations

$$\sum_{(i_1,\ldots,i_K)\in\Lambda(K;\mathcal{N})} p(i_1)\cdots p(i_K) = K!E_K(\boldsymbol{p}) \tag{48}$$

for all $K = 1, \ldots, N$, and the expression (27) is now immediate from (45), (47) and (48).

A special case occurs when $N = M + 1$ under the FIFO policy with either $r_{1M} = 1$ or $r_{M1} = 1$: If $s_0 = (i_1, \ldots, i_M)$, then only $M + 1$ states can be reached from s_0, i.e., $\mathcal{S}$ contains $(i_1, \ldots, i_M)$, $(i_2, \ldots, i_M, i_{M+1})$, $(i_3, \ldots, i_{M+1}, i_1)$, $\ldots, (i_{M+1}, i_1, \ldots, i_{M-1})$. Thus, the state space reduces to $\Lambda^\star(M;\mathcal{N})$ and the stationary distribution simplifies to

$$\pi(s) = \frac{p(i_1)\cdots p(i_M)}{\sum_{\{i_1,\ldots,i_M\}\in\Lambda^\star(M;\mathcal{N})} p(i_1)\cdots p(i_M)} \tag{49}$$

with $s = \{i_1, \ldots, i_M\}$ arbitrary in $\Lambda^\star(M;\mathcal{N})$. Upon utilizing (49) and following the steps above, we conclude that the miss rate $M_{\boldsymbol{r}}(\boldsymbol{p}; s_0)$ is also given by (27).

In view of this discussion, in Case 1 it is appropriate to drop the dependence of the miss rate on s_0 as was done in Section VIII.

Case 2 – The set Σ is not empty with $|\Sigma| = m$ for some $m = 1, \ldots, M - 1$. Given an initial cache state s_0 in $\Lambda(M;\mathcal{N})$, we let $\Sigma(s_0)$ be the set of initial documents with their positions in Σ and these documents will never be replaced during the operation of the cache. With $\Sigma(s_0)^c$ denoting the documents not in $\Sigma(s_0)$, the state space of Case 2 reduces to $\mathcal{S} = \{\Sigma(s_0) \cup s' : s' \in \Lambda(M - m; \Sigma(s_0)^c)\}$ and the stationary distribution is given by

$$\pi(s) = \frac{p(i_1)\cdots p(i_{M-m})}{\sum_{(i_1,\ldots,i_{M-m})\in\Lambda(M-m;\Sigma(s_0)^c)} p(i_1)\cdots p(i_{M-m})}$$

where we have set $s = \Sigma(s_0) \cup s'$ with $s' = (i_1, \ldots, i_{M-m})$ arbitrary in $\Lambda(M - m; \Sigma(s_0)^c)$. By injecting this last expression into (46) and following the same steps as in Case 1, we get

$$M_{\boldsymbol{r}}(\boldsymbol{p}; s_0) = \frac{\sum_{(i_1,\ldots,i_{M-m+1})} p(i_1)\cdots p(i_{M-m+1})}{\sum_{(i_1,\ldots,i_{M-m})} p(i_1)\cdots p(i_{M-m})}$$

where $\sum_{(i_1,\ldots,i_{M-m+1})}$ and $\sum_{(i_1,\ldots,i_{M-m})}$ denote the summations taken over the sets $\Lambda(M - m + 1; \Sigma(s_0)^c)$ and $\Lambda(M - m; \Sigma(s_0)^c)$, respectively. Define the element $\boldsymbol{t}$ in $\mathbb{R}^N$ by $t_i = 0$ for i in $\Sigma(s_0)$ and $t_i = 1$ otherwise. With this element $\boldsymbol{t}$, it is plain from (48) that the expression for the miss rate above becomes (29).

Again, a special case occurs when $N = M+1$ under FIFO-like policies, i.e., for some $k, \ell = 1, \ldots, M$, $r_{k\ell} = 1$ with $|\Sigma| = m$. We now have $\mathcal{S} = \{\Sigma(s_0) \cup s' : s' \in \Lambda^\star(M - m; \Sigma(s_0)^c)\}$ and the stationary distribution is given by

$$\pi(s) = \frac{p(i_1)\cdots p(i_{M-m})}{\sum_{\{i_1,\ldots,i_{M-m}\}\in\Lambda^\star(M-m;\Sigma(s_0)^c)} p(i_1)\cdots p(i_{M-m})}$$

where $s = \Sigma(s_0) \cup s'$ with $s' = \{i_1, \ldots, i_{M-m}\}$ arbitrary in $\Lambda^\star(M - m; \Sigma(s_0)^c)$. It is a simple matter to check that the miss rate $M_{\boldsymbol{r}}(\boldsymbol{p}; s_0)$ also has the expression (29).

APPENDIX II
A PROOF OF THEOREM 5

We shall have repeated use for the next elementary lemma where asymptotic equivalence is defined as follows: For mappings $f, g : \mathbb{R}_+ \to \mathbb{R}$, we write $f(\alpha) \sim g(\alpha)$ $(\alpha \to \infty)$ if $\lim_{\alpha\to\infty} \frac{f(\alpha)}{g(\alpha)} = 1$.

Lemma 2: *Consider a finite family $a_1, \ldots, a_K$ of positive scalars. We have*

$$\sum_{k=1}^{K} a_k^{-\alpha} \sim c \cdot \left(\min_{k=1,\ldots,K} a_k\right)^{-\alpha} \qquad (\alpha \to \infty).$$

where c denotes the number of indices ℓ for which it holds $a_\ell = \min_{k=1,\ldots,K} a_k$.

In what follows, without further mention, all asymptotics are understood in the regime where α is large, and the qualifier $\alpha \to \infty$ is now dropped from the notation. In particular, we have $C_\alpha(N) \sim 1$.

The LRU policy – Fix $\alpha \geq 0$. Substituting (7)-(8) into the expression (32) for the miss rate under the LRU policy readily leads to

$$\begin{aligned} & C_\alpha(N)^2 M_{\mathrm{LRU}}(\boldsymbol{p}_\alpha) \qquad (50) \\ = & \sum_{s\in\Lambda(M;\mathcal{N})} \frac{\left(\prod_{\ell=1}^{M} i_\ell^{-\alpha}\right)\left(\sum_{j\notin\{i_1,\ldots,i_M\}} j^{-\alpha}\right)}{\prod_{k=1}^{M-1}\left(\sum_{j\notin\{i_1,\ldots,i_k\}} j^{-\alpha}\right)} \end{aligned}$$

where we set $s = (i_1, \ldots, i_M)$ be an element in $\Lambda(M;\mathcal{N})$ and for each $k = 1, \ldots, M$, $j \notin \{i_1, \ldots, i_k\}$ denotes the set of elements j in $\mathcal{N}$ which are not in the set $\{i_1, \ldots, i_k\}$.

For any element $s = (i_1, \ldots, i_M)$ in $\Lambda(M;\mathcal{N})$, Lemma 2 yields

$$\sum_{j\notin\{i_1,\ldots,i_k\}} j^{-\alpha} \sim \left(\min_{j\notin\{i_1,\ldots,i_k\}} j\right)^{-\alpha} \tag{51}$$

for each $k = 1, \ldots, M$, whence

$$\prod_{k=1}^{M-1}\left(\sum_{j\notin\{i_1,\ldots,i_k\}} j^{-\alpha}\right) \sim \rho(s)^{-\alpha} \tag{52}$$

where we have set

$$\rho(s) := \prod_{k=1}^{M-1}\left(\min_{j\notin\{i_1,\ldots,i_k\}} j\right).$$

By combining (50) and (51) with (52), we readily have

$$M_{\mathrm{LRU}}(\boldsymbol{p}_\alpha) \sim \sum_{s\in\Lambda(M;\mathcal{N})} \nu(s)^{-\alpha} \tag{53}$$

where we have set

$$\nu(s) := \frac{\left(\prod_{\ell=1}^{M} i_\ell\right)\left(\min_{j\notin s} j\right)}{\rho(s)}$$

for any element $s = (i_1, \ldots, i_M)$ in $\Lambda(M;\mathcal{N})$. Utilizing Lemma 2 on (53), we obtain

$$M_{\mathrm{LRU}}(\boldsymbol{p}_\alpha) \sim c \cdot \left(\min_{s\in\Lambda(M;\mathcal{N})} \nu(s)\right)^{-\alpha} \tag{54}$$

where c is the number of elements s in $\Lambda(M;\mathcal{N})$ that achieve the minimum in (54).

It is clear that

$$\begin{aligned} & \min_{s\in\Lambda(M;\mathcal{N})} \nu(s) \qquad (55) \\ \geq & \frac{\min_{s\in\Lambda(M;\mathcal{N})}\left(\left(\prod_{\ell=1}^{M} i_\ell\right)\cdot\left(\min_{j\notin s} j\right)\right)}{\max_{s\in\Lambda(M;\mathcal{N})} \rho(s)}. \end{aligned}$$

It is not too difficult to check that $s = (1, \ldots, M)$ and $s = (1, \ldots, M-1, M+1)$ are the only two elements in $\Lambda(M;\mathcal{N})$ that *simultaneously* achieve the minimum $(M+1)!$ of the quantity $\left(\prod_{\ell=1}^{M} i_\ell\right)\cdot\left(\min_{j\notin s} j\right)$ and the maximum $M!$ of $\rho(s)$. Hence, (55) does hold as an equality with

$$\min_{s\in\Lambda(M;\mathcal{N})} \nu(s) = \frac{(M+1)!}{M!} = M+1$$

and $c = 2$. It then follows from (54) that

$$M_{\mathrm{LRU}}(\boldsymbol{p}) \sim 2(M+1)^{-\alpha} \tag{56}$$

and the desired conclusion readily follows. ■

The CLIMB policy – Fix $\alpha \geq 0$. Upon substituting (7)-(8) into the expression (33) for the miss rate under the CLIMB policy we find

$$\begin{aligned} & C_\alpha(N) M_{\mathrm{CL}}(\boldsymbol{p}_\alpha) \qquad (57) \\ = & \frac{\sum_{s\in\Lambda(M;\mathcal{N})}\left(\prod_{\ell=1}^{M} i_\ell^{-\alpha(M-\ell+1)}\right)\left(\sum_{j\notin s} j^{-\alpha}\right)}{\sum_{s\in\Lambda(M;\mathcal{N})}\prod_{\ell=1}^{M} i_\ell^{-\alpha(M-\ell+1)}} \end{aligned}$$

where $j \notin s$ denotes the set of elements j in $\mathcal{N}$ which are not in s.

Invoking Lemma 2, we immediately get

$$\begin{aligned} \sum_{s\in\Lambda(M;\mathcal{N})}\prod_{\ell=1}^{M} i_\ell^{-\alpha(M-\ell+1)} & \sim \left(\min_{s\in\Lambda(M;\mathcal{N})}\prod_{\ell=1}^{M} i_\ell^{M-\ell+1}\right)^{-\alpha} \\ & = \left(\prod_{\ell=1}^{M} \ell^{M-\ell+1}\right)^{-\alpha} \qquad (58) \end{aligned}$$

where the minimum is readily seen to be achieved by $s = (1, \ldots, M)$. Next, by using of Lemma 2 again, we see that

$$\sum_{j\notin s} j^{-\alpha} \sim \left(\min_{j\notin s} j\right)^{-\alpha} \tag{59}$$

for any element s in $\Lambda(M;\mathcal{N})$.

By combining (57)-(59) and making use of Lemma 2, we readily obtain

$$M_{\mathrm{CL}}(\boldsymbol{p}_\alpha) \sim c \cdot \left(\frac{\min_{s\in\Lambda(M;\mathcal{N})} \mu(s)}{\prod_{\ell=1}^{M} \ell^{M-\ell+1}}\right)^{-\alpha} \tag{60}$$

where we have set

$$\mu(s) = \prod_{\ell=1}^{M} i_\ell^{M-\ell+1} \cdot \left(\min_{j\notin s} j\right)$$

for $s = (i_1, \ldots, i_M)$ in $\Lambda(M;\mathcal{N})$ and c denotes the number of indices achieving the minimum in (60). It is a simple matter to check that $s = (1, \ldots, M)$ and $s = (1, \ldots, M-1, M+1)$ are the only two elements in $\Lambda(M;\mathcal{N})$ achieving the minimum $(M+1)\prod_{\ell=1}^{M} \ell^{M-\ell+1}$ of $\mu(s)$. Thus, with $c = 2$, (60) yields

$$M_{\mathrm{CL}}(\boldsymbol{p}_\alpha) \sim 2(M+1)^{-\alpha} \tag{61}$$

and the desired conclusion is now immediate. ■

From (56) and (61), it is plain that

$$M_{\mathrm{LRU}}(\boldsymbol{p}_\alpha) \sim M_{\mathrm{CL}}(\boldsymbol{p}_\alpha) \sim 2(M+1)^{-\alpha}$$

and this is consistent with plots (in log-log scale) displayed in Figure 4 when α is large.