Ph.D. Thesis

Comparing Strength of Locality of Reference: Popularity, Temporal Correlations, and Some Folk Theorems for the Miss Rates and Outputs of Caches

by Sarut Vanichpun
Advisor:

CSHCN PhD 2005-1
(ISR PhD 2005-1)

ABSTRACT

Title of dissertation: COMPARING STRENGTH OF LOCALITY
OF REFERENCE: POPULARITY, TEMPORAL
CORRELATIONS, AND SOME FOLK THEOREMS
FOR THE MISS RATES AND OUTPUTS OF CACHES

Sarut Vanichpun, Doctor of Philosophy, 2005

Dissertation directed by: Professor Armand M. Makowski
Department of Electrical and Computer Engineering and
Institute for Systems Research

The performance of demand-driven caching is known to depend on the locality of
reference exhibited by the stream of requests made to the cache. In spite of numerous
efforts, no consensus has been reached on how to formalize this notion, let alone on
how to compare streams of requests on the basis of their locality of reference. We take
on this issue with an eye towards validating operational expectations associated with the
notion of locality of reference. We focus on two "folk theorems," that is, (i) The stronger
the locality of reference, the smaller the miss rate of the cache; and (ii) Good caching
is expected to produce an output stream of requests exhibiting less locality of reference
than the input stream of requests. These two folk theorems are explored in the context
of demand-driven caching for the two main contributors of locality of reference, namely
popularity and temporal correlations.

We first focus exclusively on popularity by considering the situation where there
are no temporal correlations in the stream of requests, as would be the case under the

Independent Reference Model (IRM). As we propose to measure strength of locality of reference in a stream of requests through the skewness of its popularity distribution, we introduce the notion of majorization as a means for capturing this degree of skewness. We show that these folk theorems hold for caches operating under a large class of replacement policies, the so-called Random On-demand Replacement Algorithms (RORA), which includes the optimal policy A_0 and the random policy. However, counterexamples prove that this is not always the case under the (popular) Least-Recently-Used (LRU) and CLIMB policies. In such cases, conjectures are offered (and supported by simulations) as to when the folk theorems would hold under the LRU or CLIMB caching, given that the IRM input has a Zipf-like popularity pmf.

To compare the strength of temporal correlations in streams of requests, we define the notion of Temporal Correlations (TC) ordering based on the so-called supermodular ordering, a concept of positive dependence which has been successfully used for comparing dependence structures in sequences of random variables. We explore how the TC ordering captures the strength of temporal correlations in several Web request models, namely the higher-order Markov chain model (HOMM), the partial Markov chain model (PMM) and the Least-Recently-Used stack model (LRUSM). We establish the folk theorem to the effect that the stronger the strength of temporal correlations, the smaller the miss rate for the PMM under certain assumptions on the caching policy. Conjectures and simulations are offered as to when this folk theorem would hold under the HOMM and under the LRUSM. In addition, the validity of this folk theorem for general request streams under the Working Set algorithm is studied.

Lastly, we investigate how the majorization and TC orderings can be translated into comparisons of three well-known locality of reference metrics, namely the working set size, the inter-reference time and the stack distance.

COMPARING STRENGTH OF LOCALITY OF REFERENCE:
POPULARITY, TEMPORAL CORRELATIONS, AND SOME FOLK THEOREMS
FOR THE MISS RATES AND OUTPUTS OF CACHES

by

Sarut Vanichpun

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2005

Advisory Committee:

        Professor Armand M. Makowski, Chair
        Professor Manoj Franklin
        Professor Richard J. La
        Professor Adrian Papamarcou
        Professor A. Udaya Shankar

DEDICATION

To Dad and Mom

ACKNOWLEDGEMENTS

of writing this dissertation. My appreciation also extends to all my friends in College Park who have made this place home away from home for me.

Above all, I am deeply grateful to my parents and my sister. Without their love, support and belief in me, I would not have been able to accomplish this enormous task.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1   Web caching

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., [2, 24] (and references therein).

Since its inception, the World Wide Web has seen an exponential increase in the number of its users and in the volume of objects to be accessed. This trend, which is not likely to abate anytime soon, is challenging current cache architectures to meet the complementary mandates of *speed*, *scalability* and *reliability* which are central to delivering a satisfactory user experience.

Generally speaking, scalability requires some form of *hierarchical* organization. In the context of Web caching, this notion has led naturally to the deployment of *multi-layered* systems of *interconnected* caches which may be organized in a tree-like hierarchy or in more complicated meshes [12, 16, 29] (and references therein).

Even a cursory review of the literature [5, 54, 69] already reveals the large number

of difficult and challenging issues that need to be addressed in order to ensure proper operations of these distributed multi-level caching systems. Examples of these issues include (i) cache replacement strategies [15, 39, 54, 55]; (ii) prefetching algorithms [25] (and references therein); (iii) cache location [43, 44]; (iv) content placement [23, 57, 68]; and (v) cache cooperation techniques [16, 17, 30].

1.2   Locality of reference

Although these challenges have renewed interest in caching in general, some basic issues are still not well understood. Indeed, 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 stream of requests whereby "bursts of references are made in the near future to objects referenced in the recent past."

The notion of locality and its importance for caching were first recognized by Belady [10] in the context of computer memory, and attempts at characterization were made early on by Denning through the working set model [26, 27]. Subsequently, a number of studies have shown that request streams for Web objects exhibit strong locality of reference[1] [40, 41, 46] and various metrics have been proposed for characterizing the locality of reference in Web request streams [1, 34, 40].

Although several competing definitions for locality of reference are 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,

---

[1]At least in the short timescales.

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 $\boldsymbol{R} = \{R_t, \ t = 0, 1, \ldots\}$ 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

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

whenever these limits exist (and they do in most models treated in the literature). Popularity is usually viewed as a long-term expression of locality which captures the likelihood that a document will be requested in the future relative to other documents.

2. *Temporal correlations* are more delicate to define due to the "categorical" nature of the requests $\{R_t, \ t = 0, 1, \ldots\}$. Indeed, it is somewhat meaningless to use the covariance function

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

as a way to capture these temporal correlations as is traditionally done in other contexts. This is because of the *categorical nature* of the rvs $\{R_t, \ t = 0, 1, \ldots\}$ which take values in a discrete set – We took $\{1, \ldots, N\}$ but 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* exhibited by requests for particular documents over time. The literature contains several metrics for doing this, e.g., the inter-reference time [34, 40, 53], the working set size [26, 27] and the stack distance [1, 3, 50].

1.3　Folk theorems

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 on how to *compare* streams of requests on the basis of their locality of reference.[2] In addition, lacking in most of the work done thus far, is a clear recognition of the system-wide nature of Web caching, whereby local *transformative* actions shape the streams of requests as they pass through successive caches.[3] These problems have precluded a formal study of the following "folk theorems":

1. **Folk theorem on miss rates** – The stronger the locality of reference in the stream of requests, the smaller the miss rate, since the cache ends up being populated by objects with a higher likelihood of access in the near future. Such a property, if true, would confirm the central role played by locality of reference in shaping cache performance. In fact, the very presence of locality of reference in the stream of requests is what makes caching at all possible; and

2. **Folk theorem on output streams** – Good cache replacement strategies "absorb" locality of reference to a certain extent by producing a stream of misses from the cache – its so-called output – which exhibits *less* locality of reference than the input stream of requests. In the context of multi-level caching, this reduction property is often perceived as one of the main reasons for why caching looses its effectiveness after some level in a hierarchy of caches.

---

[2]Exceptions can be found in [34, 65].

[3]Recent works on this issue can be found in [17, 30, 32] for cache management and in [47, 70, 71] for Web traffic analysis.

Such folk theorems are expected to hold for demand-driven caching that exploits recency of reference. Interest in establishing them under *specific* definitions of locality of reference stems from a desire to validate their *operational* significance on caching systems. Counterexamples would cast some doubts as to whether a particular definition indeed captures the intuitive meaning of locality of reference and to whether a particular caching algorithm is indeed a well-behaved policy.

1.4   Contributions

In this dissertation, we identify notions of locality of reference which are capable of comparing the strength of locality of reference between streams of requests. Such notions allow a comparison statement of the form

$$\boldsymbol{R}^1 \leq_{LR} \boldsymbol{R}^2 \tag{1.1}$$

to the effect that "a request stream $\boldsymbol{R}^1$ has less locality of reference than a request stream $\boldsymbol{R}^2$" under some appropriate notion of locality of reference. With the comparison (1.1), we are able to formally investigate the folk theorems mentioned above, albeit in a simple framework under demand-driven cache replacement policies. Indeed, the folk theorem for miss rates can be formalized as

$$M_\pi(\boldsymbol{R}^2) \leq M_\pi(\boldsymbol{R}^1) \quad \text{whenever (1.1) holds} \tag{1.2}$$

where $M_\pi(\boldsymbol{R}^1)$ and $M_\pi(\boldsymbol{R}^2)$ denote the miss rates of the request streams $\boldsymbol{R}^1$ and $\boldsymbol{R}^2$ under the cache replacement policy π, respectively, while the folk theorem for output streams simply states that

$$\boldsymbol{R}_\pi^\star \leq_{LR} \boldsymbol{R} \tag{1.3}$$

where $\boldsymbol{R}_\pi^\star$ is the output stream of the cache operating under the policy π when the input stream is $\boldsymbol{R}$.

The tasks above have been carried out separately for the two main sources of locality of reference, namely popularity and temporal correlations. We now summarize the corresponding results in some details.

1.4.1 Majorization and popularity

We first focus exclusively on popularity as a way to formalize (1.1). To isolate its contributions, we consider the situation where there are *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 $\boldsymbol{p} = (p(1), \ldots, p(N))$, the requests $\{R_t, t = 0, 1, \ldots\}$ form a sequence of i.i.d. $\mathcal{N}$-valued rvs, each distributed according to the pmf $\boldsymbol{p}$. Even in the absence of temporal correlations, locality of reference is present, in that the *skewness* of $\boldsymbol{p}$ acts as an indicator of the strength of locality of reference under the intuition that the more "balanced" the pmf $\boldsymbol{p}$, the weaker the locality of reference.

In a recent paper, Fonseca et. al [34] introduced a notion of comparison based on the entropy 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 $\boldsymbol{q}$. Unfortunately, this notion is not strong enough to allow for results of the forms (1.2) and (1.3) to be established. Here, the degree of skewness in the popularity pmf is captured formally through the notion of *majorization (ordering)* [Chapter 2]. This concept has been used previously in the context of caching by van den Berg and Towsley [65]. With this notion, the comparison (1.2) can be recast as saying that the miss rate (as a function of popularity) belongs to the rich and structured class of monotone functions associated with majorization, the so-called Schur-convex/concave functions. Moreover, basic facts regarding majorization enable us to develop generic comparison results between the

popularity pmfs of the input and output streams [Chapter 6].

Equipped with the notion of majorization ordering, the folk theorems for the miss rates and output streams can be established for a number of policies, namely the optimal policy A_0, the random policy and the FIFO (First-In/First-Out) policy [Chapter 6]. These positive results are then extended to a very large class of replacement policies, the so-called Random On-demand Replacement Algorithms (RORA) [Chapter 7].

However, these folk theorems do *not* always hold under two self-organizing policies, namely the LRU (Least-Recently-Used) and CLIMB replacement policies [Chapter 8]. We first exhibit situations where under these policies, the IRM stream with more skewed popularity pmf may have a smaller miss rate than the IRM stream with less skewed popularity pmf. Yet, when the popularity pmfs are Zipf-like [Section 6.2], simulations show that the comparison (1.2) under these policies does hold. We formally establish this fact only in the limiting regime where the skewness parameter of the Zipf-like pmf is large, i.e., highly skewed.

It also happens that the LRU and CLIMB policies fail to reduce locality of reference in that under these policies, the input popularity pmf $\boldsymbol{p}$ (of $\boldsymbol{R}$) is not necessarily more skewed than the output popularity pmf $\boldsymbol{p}^\star$ (of $\boldsymbol{R}_\pi^\star$). We explore the issue through counterexamples which are developed within some classes of input popularity pmfs. In particular, when the input popularity pmf lies in the class of Zipf-like pmfs, we identify a condition involving the cache size and the number of cacheable documents under which reduction fails to occur at large enough values of the skewness parameter of the input Zipf-like pmf. Under this condition, which we expect to be satisfied in practice, we show that the output pmf $\boldsymbol{p}^\star$ may not exhibit less locality of reference than the input pmf $\boldsymbol{p}$ when the latter has too much of it to begin with. Additional simulations were carried out and suggest conjectures as to when LRU and CLIMB policies indeed reduce

locality of reference with Zipf-like input pmfs. All indications point to the possibility that for small enough cache sizes, the desired folk theorem will hold.

1.4.2 Positive dependence and temporal correlations

As mentioned earlier, the catagorical nature of the requests $\{R_t, t = 0, 1, \ldots\}$ makes it difficult to define appropriate notions of temporal correlations. Even though several metrics have been proposed, e.g., the inter-reference time, the working set size and the stack distance, none has been found appropriate for formalizing these folk theorems.

We take on this issue by applying the concepts of positive dependence [Chapter 3] to capture the strength of temporal correlations exhibited by streams of requests. Positive dependence has been used previously in a number of contexts, e.g., network traffic and queueing theory [8, 9, 66], and reliability theory [6, 60]. Specifically, relying on the notion of supermodular ordering [Definition 3.4] which has been used to compare dependence structures in sequences of rvs, we define the *Temporal Correlations (TC)* ordering [Definition 9.1] as a way to compare streams of requests on the basis of the strength of their temporal correlations. This new ordering is well suited for comparing the relative strength of temporal correlations as we note that request streams comparable in the TC ordering must have the same popularity profiles (under the assumption that they exist); in other words, the TC ordering cannot capture any contribution from popularity toward locality of reference.

We apply the TC ordering to capture the strength of temporal correlations present in several Web request models that are believed to exhibit such correlations, namely the higher-order Markov chain model (HOMM), the partial Markov chain model (PMM) and the Least-Recently-Used stack model (LRUSM). Indeed, we demonstrate that the HOMM exhibits temporal correlations in the sense that it has stronger strength of tempo-

ral correlations than the IRM with the same popularity pmf in the TC ordering [Section 9.2]. This property is shown to hold also for the LRUSM under a reasonable condition on its stack distance pmf [Section 9.4]. Lastly, for PMM, we show that the strength of temporal correlations is indeed captured by the correlation parameter as expected [Section 9.3].

With the TC ordering, we establish the folk theorem for miss rates when the input to the cache is modeled according to the PMM under certain assumptions on the cache replacement policies [Section 9.5.1]. Conjectures and simulations are offered as to when this folk theorem would hold under the HOMM [Section 9.5.2] and under the LRUSM [Section 9.5.2]. We also investigate this folk theorem with general input streams under the so-called Working Set (WS) algorithm [Section 10.4] which is a cache management policy associated with the working set model. The result indicates that (1.2) does hold when the cache holds only one document in which case the WS algorithm is identified with any demand-driven caching with unit cache size. However, the folk theorem may not hold in some other situations, as shown by counterexamples in the class of PMM request streams.

It is also desirable to establish the folk theorem for output streams via the TC ordering. However, there are only limited cases of interests as we recall that the output popularity pmf $p^\star$ is not necessarily the same as the input popularity pmf p and that the comparison in the TC ordering between the input stream and the output stream requires that both popularity pmfs be identical. This shortcoming calls for further study to develop orderings that can compare the strength of locality of reference contributed by both components, namely popularity and temporal correlations.

1.4.3 Locality of reference metrics

Lastly, we investigate whether the comparison in the majorization ordering of two IRM streams and the comparison in the TC ordering of two request streams translate into the expected comparisons for three well-established locality of reference metrics, namely, the working set size, the inter-reference time, and the stack distance.

For the working set size, the majorization ordering of two IRM streams implies the (strong) stochastic ordering between their working set sizes, while the TC ordering of two request streams only gives a comparison between their average working set sizes. In addition, both the majorization ordering and the TC ordering allow a comparison of the steady state inter-reference times in the convex ordering. However, implications of these orderings on the stack distances are not fully understood and require further investigation.

These locality of reference metrics are sometimes used for cache dimensioning and cache performance evaluation. Thus, the aforementioned relations naturally lead to various bounds on these performance metrics. For instance, because the IRM with uniform popularity pmf acts as a lower bound (in the sense of majorization ordering) for any IRM stream, its corresponding locality of reference metrics are bounds for those of other IRM streams. Furthermore, if the request stream R exhibits temporal correlations stronger than that of the IRM with similar popularity pmf in the sense of the TC ordering, then the performance metrics associated with this IRM, which are usually known or easier to be computed, can provide bounds for those of the request stream R.

1.5   Organization

The dissertation is organized as follows: The theory of majorization and its companion notion, Schur-convexity, are summarized in Chapter 2. Basic definitions and facts regarding positive dependence and stochastic orderings are collected in Chapter 3.

In Chapter 4, we introduce a simple framework of demand-driven caching and give the definitions of miss rate and output of a cache. We then use the concept of majorization ordering for comparing popularity pmfs of IRM request streams in Chapter 6. With the majorization ordering, we establish the folk theorems for miss rates and output streams under the random policy and the policy A_σ. These results are extended in Chapter 7 to a large class of demand-driven replacement policies, the so-called Random On-demand Replacement Algorithm (RORA). In Chapter 8, we show that the folk theorems do not hold in general for two well-known self-organizing policies, the LRU and CLIMB policies, where counterexamples are established. Asymptotics and conjectures under the class of IRM streams with Zipf-like popularity pmf are investigated.

In Chapter 9, we use the concepts of positive dependence and supermodular ordering to define the TC ordering as a means to compare strength of temporal correlations. This ordering is then used to capture the temporal correlations present in three request models, namely HOMM, PMM and LRUSM. The folk theorem for miss rates of the PMM is established under certain assumptions on the caching policy. Specific results and conjectures on this folk theorem under the HOMM and the LRUSM are provided.

The working set model is considered in Chapter 10 where we demonstrate how the majorization ordering between IRM streams and the TC ordering between request streams can be translated into comparisons of the working set sizes. Next, under the Working Set algorithm, we find that the folk theorems for miss rates and output streams do not always hold for IRM input streams. For general input models, the folk theorem

for miss rates holds when the cache holds only one document, but fails otherwise.

Lastly, in Chapter 11, we show that the majorization ordering and the TC ordering imply the comparison in the convex ordering of the steady state inter-reference times. We also investigate whether these orderings would lead to some appropriate comparisons of the stack distances.

Chapter 2

Majorization and Schur-convexity

2.1 Majorization – A primer

The concept of *majorization* [49] provides a powerful tool to formalize statements concerning the relative skewness in the 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

$$\sum_{i=1}^{n} x_{[i]} \leq \sum_{i=1}^{n} y_{[i]}, \quad n = 1, 2, \ldots, N - 1 \tag{2.1}$$

and

$$\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i \tag{2.2}$$

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 [49], this notion has found widespread use in many diverse branches of mathematics and their applications, viz. in computer databases [20] and storage [73].

We begin with a sufficient condition for majorization which is extracted from the discussion in [49, B.1, p. 129].

Proposition 2.1 *Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be distinct elements of $\mathbb{R}^N$ such that*

$$\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i. \tag{2.3}$$

Whenever, $x_1 \geq x_2 \geq \ldots \geq x_N$, if there exists some $k = 1, \ldots, N-1$ such that $x_i \leq y_i$, $i = 1, \ldots, k$ and $x_i \geq y_i$, $i = k+1, \ldots, N$, then the comparison $\boldsymbol{x} \prec \boldsymbol{y}$ holds.

The following sufficient condition for majorization will be useful in the sequel; it was already announced in [49, B.1.b, p. 129] without proof.

Theorem 2.2 *Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be distinct elements of $\mathbb{R}^N$ such that (2.3) holds. Whenever $x_1 \geq x_2 \geq \ldots \geq x_N > 0$, and the ratios $\frac{y_i}{x_i}$, $i = 1, \ldots, N$, are decreasing in i, we have the comparison $\boldsymbol{x} \prec \boldsymbol{y}$.*

Proof. Under the condition $x_i > 0$, $i = 1, \ldots, N$, we find that (2.3) can be rewritten as

$$\sum_{i=1}^{N} x_i \left(\frac{y_i}{x_i} - 1 \right) = 0. \tag{2.4}$$

If the ratios $\frac{y_i}{x_i}$, $i = 1, \ldots, N$, are decreasing in i, then by virtue of (2.4) there must exist some k with $1 \leq k < N$ such that

$$\frac{y_i}{x_i} - 1 \geq 0, \quad i = 1, \ldots, k$$

and

$$\frac{y_i}{x_i} - 1 \leq 0, \quad i = k+1, \ldots, N.$$

In other words, $x_i \leq y_i$ for $i = 1, \ldots, k$ and $y_i \leq x_i$ for $i = k + 1, \ldots, N$, and we readily obtain the comparison $x \prec y$ by applying Proposition 2.1. ∎

With any element of $\mathbb{R}^N$ such that $\sum_{i=1}^N x_i \neq 0$, we associate the *normalized* vector $\bar{x}$ as the element of $\mathbb{R}^N$ defined by

$$\bar{x} := \left(\sum_{i=1}^N x_i \right)^{-1} (x_1, \ldots, x_N). \tag{2.5}$$

With this notation, we can now present a useful corollary to Theorem 2.2.

Corollary 2.3 *Let x and y be distinct elements of $\mathbb{R}^N$ such that $\sum_{i=1}^N y_i > 0$. Whenever $x_1 \geq x_2 \geq \ldots \geq x_N > 0$, and the ratios $\frac{y_i}{x_i}$, $i = 1, \ldots, N$, are decreasing in i, we have the comparison $\bar{x} \prec \bar{y}$.*

Proof. Under the enforced assumptions, we note the inequalities $\sum_{i=1}^N x_i > 0$ and $\bar{x}_1 \geq \bar{x}_2 \geq \ldots \geq \bar{x}_N > 0$ with the ratios $\frac{\bar{y}_i}{\bar{x}_i}$, $i = 1, \ldots, N$, decreasing in i. Obviously, $\sum_{i=1}^N \bar{x}_i = \sum_{i=1}^N \bar{y}_i = 1$ and we get the desired result by applying Theorem 2.2 to $\bar{x}$ and $\bar{y}$. ∎

The following reformulation of Corollary 2.3 is used in the sequel.

Lemma 2.4 *Let x and y be distinct elements of $\mathbb{R}^N$ such that $x_i > 0$, $i = 1, \ldots, N$ and $\sum_{i=1}^N y_i > 0$. If*

$$\frac{y_i}{x_i} \geq \frac{y_j}{x_j} \tag{2.6}$$

whenever $x_i \geq x_j$ for distinct $i, j = 1, \ldots, N$, then the comparison $\bar{x} \prec \bar{y}$ holds.

Before giving a proof, we introduce the following notation: Let σ denote a permutation of $\{1, \ldots, N\}$. With any element x in $\mathbb{R}^N$, we associate the *permuted* vector $\sigma(x)$

in $\mathbb{R}^N$ through the relation

$$\sigma(\boldsymbol{x}) = \left(x_{\sigma(1)}, \ldots, x_{\sigma(N)}\right).$$

It is plain from the definition of majorization that for vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^N$, we have $\boldsymbol{x} \prec \boldsymbol{y}$ if and only if $\sigma(\boldsymbol{x}) \prec \boldsymbol{y}$ for any permutation σ of $\{1, \ldots, N\}$.

Proof. Let σ denote a permutation of $\{1, \ldots, N\}$ such that $x_{\sigma(1)} \geq x_{\sigma(2)} \geq \ldots \geq x_{\sigma(N)}$. The enforced monotonicity assumptions can be restated as

$$\frac{y_{\sigma(1)}}{x_{\sigma(1)}} \geq \frac{y_{\sigma(2)}}{x_{\sigma(2)}} \geq \ldots \geq \frac{y_{\sigma(N)}}{x_{\sigma(N)}},$$

and the desired result follows by an easy application of Corollary 2.3 to the elements $\sigma(\boldsymbol{x})$ and $\sigma(\boldsymbol{y})$. ∎

One such application of Lemma 2.4 is given in

Lemma 2.5 *For any $\varepsilon > 0$, define the N-dimensional vector $\boldsymbol{p}_\varepsilon$ by*

$$\boldsymbol{p}_\varepsilon = (1 - (N-1)\varepsilon, \varepsilon, \ldots, \varepsilon).$$

If ε and η satisfy the relation $0 < \eta \leq \varepsilon \leq \frac{1}{N}$, then it holds that $\boldsymbol{p}_\varepsilon \prec \boldsymbol{p}_\eta$.

Proof. As we have in mind to apply Lemma 2.4, we take $\bar{\boldsymbol{x}} = \boldsymbol{x} = \boldsymbol{p}_\varepsilon$ and $\bar{\boldsymbol{y}} = \boldsymbol{y} = \boldsymbol{p}_\eta$. It is plain that the requisite monotonicity assumptions of Lemma 2.4 hold when ε and η satisfy the relation $0 < \eta \leq \varepsilon \leq \frac{1}{N}$. ∎

2.2 Schur-convexity

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 $\mathbb{R}^N$).

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 this section. 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 $i = 1, \ldots, N!$.

In the following, we have collected some useful technical results concerning Schur-concave functions. As in [49, p. 78], for each $M = 1, \ldots, N$, the *elementary symmetric function* $E_{M,N} : \mathbb{R}^N \to \mathbb{R}$ is defined by

$$E_{M,N}(\boldsymbol{x}) := \sum_{\{i_1, \ldots, i_M\} \in \Lambda^\star(M;\mathcal{N})} x_{i_1} \cdots x_{i_M}, \quad \boldsymbol{x} \in \mathbb{R}^N \tag{2.7}$$

with $\Lambda^\star(M;\mathcal{N})$ denoting the collection of all *unordered* subsets of size M of $\mathcal{N} = \{1, \ldots, N\}$. By convention we write $E_{0,N}(\boldsymbol{x}) = 1$ for all $\boldsymbol{x}$ in $\mathbb{R}^N$. It is well known

[49, Prop. F.1., p. 78] that the function $E_{M,N}$ is Schur-concave on $\mathbb{R}_+^N$ for each $M = 0, 1, \ldots, N$.

We note from [49, Prop. C.2, p. 67] 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). The following result is due to Schur [49, F.3, p. 80] and will be key to a number of proofs.

Proposition 2.6 *For each $M = 1, \ldots, N$, the mapping $\Phi_{M,N} : \mathbb{R}_+^N \to \mathbb{R}$ given by[1]*

$$\Phi_{M,N}(\boldsymbol{x}) := \frac{E_{M,N}(\boldsymbol{x})}{E_{M-1,N}(\boldsymbol{x})}, \quad \boldsymbol{x} \in \mathbb{R}_+^N$$

is increasing,[2] symmetric and concave, hence increasing and Schur-concave.

Proposition 2.7 *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.

Proof. The mapping φ_h is symmetric by virtue of the symmetry of h. The concavity of φ_h can be shown as follows: First, for $i = 1, \ldots, N!$, we set $\varphi_i(\boldsymbol{x}) = \varphi(\sigma_i(\boldsymbol{x}))$ $(\boldsymbol{x} \in A)$; this definition is well posed since A is symmetric. The concavity of φ implies that of φ_i. For arbitrary $\boldsymbol{x}$ and $\boldsymbol{y}$ in A, and α in $[0, 1]$ (with $\bar{\alpha} = 1 - \alpha$), we see that $\alpha \boldsymbol{x} + \bar{\alpha} \boldsymbol{y}$ is

---

[1]For $\boldsymbol{x}$ in $\mathbb{R}_+^N$ such that $E_{M-1,N}(\boldsymbol{x}) = 0$, we have $E_{M,N}(\boldsymbol{x}) = 0$ and set $\Phi_{M,N}(\boldsymbol{x}) = 0$ by continuity.

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

also an element of A, and we obtain

$$
\begin{aligned}
\varphi_h(\alpha \boldsymbol{x} + \bar{\alpha}\boldsymbol{y}) \;&=\; h(\varphi_1(\alpha \boldsymbol{x} + \bar{\alpha}\boldsymbol{y}), \ldots, \varphi_{N!}(\alpha \boldsymbol{x} + \bar{\alpha}\boldsymbol{y})) \\[2mm]
&\geq\; h(\alpha\varphi_1(\boldsymbol{x}) + \bar{\alpha}\varphi_1(\boldsymbol{y}), \ldots, \alpha\varphi_{N!}(\boldsymbol{x}) + \bar{\alpha}\varphi_{N!}(\boldsymbol{y})) \\[2mm]
&\geq\; \alpha h(\varphi_1(\boldsymbol{x}), \ldots, \varphi_{N!}(\boldsymbol{x})) + \bar{\alpha}h(\varphi_1(\boldsymbol{y}), \ldots, \varphi_{N!}(\boldsymbol{y})) \\[2mm]
&=\; \alpha\varphi_h(\boldsymbol{x}) + \bar{\alpha}\varphi_h(\boldsymbol{y}).
\end{aligned}
$$

The first inequality follows from the concavity of each of the mappings φ_i, $i = 1, \ldots, N!$ and the increasingness of h, while the second inequality is implied by the concavity of h. ∎

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

$$
\boldsymbol{t} \cdot \boldsymbol{x} := (t_1 x_1, \ldots, t_N x_N).
$$

With this notation, we can state an important consequence of Proposition 2.7.

Proposition 2.8 *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*

$$
\psi_{\boldsymbol{t}}(\boldsymbol{x}) = h(\psi(\boldsymbol{t} \cdot \sigma_1(\boldsymbol{x})), \ldots, \psi(\boldsymbol{t} \cdot \sigma_{N!}(\boldsymbol{x}))), \quad \boldsymbol{x} \in \mathbb{R}_+^N
$$

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 2.7 with $A = \mathbb{R}_+^N$ and $\varphi = \tilde{\psi}_{\boldsymbol{t}}$. ∎

Chapter 3

Stochastic Orderings and Positive Dependence

3.1 Integral stochastic orderings

In this section, we summarize some important definitions and facts concerning the stochastic orderings of random vectors. Additional information can be found in the monographs by Müller and Stoyan [52] and by Shaked and Shanthikumar [59]. The basic definition of integral stochastic orderings can be stated as follows:

Definition 3.1 *Let $\mathcal{F}$ be a class of Borel measurable functions $\varphi : \mathbb{R}^n \to \mathbb{R}$. We say that the two $\mathbb{R}^n$-valued rvs $\boldsymbol{X}$ and $\boldsymbol{Y}$ satisfy the order relation $\boldsymbol{X} \leq_{\mathcal{F}} \boldsymbol{Y}$ if*

$$\mathbf{E}\left[\varphi(\boldsymbol{X})\right] \leq \mathbf{E}\left[\varphi(\boldsymbol{Y})\right] \tag{3.1}$$

for all functions φ in $\mathcal{F}$ whenever the expectations exist.

This generic definition has been specialized in the literature. Here are some important examples.

Definition 3.2 *For $\mathbb{R}^n$-valued rvs $\boldsymbol{X}$ and $\boldsymbol{Y}$, the rv $\boldsymbol{X}$ is said to be smaller than the rv $\boldsymbol{Y}$ according to*

- *the usual stochastic ordering, written $\boldsymbol{X} \leq_{st} \boldsymbol{Y}$, if (3.1) holds for all increasing functions $\varphi : \mathbb{R}^n \to \mathbb{R}$ whenever the expectations exist;*

- *the convex ordering, written $\boldsymbol{X} \leq_{cx} \boldsymbol{Y}$, if (3.1) holds for all convex functions $\varphi : \mathbb{R}^n \to \mathbb{R}$ whenever the expectations exist;*

- *the concave ordering, written $\boldsymbol{X} \leq_{cv} \boldsymbol{Y}$, if (3.1) holds for all concave functions $\varphi : \mathbb{R}^n \to \mathbb{R}$ whenever the expectations exist;*

- *the increasing convex ordering, written $\boldsymbol{X} \leq_{icx} \boldsymbol{Y}$, if (3.1) holds for all increasing convex functions $\varphi : \mathbb{R}^n \to \mathbb{R}$ whenever the expectations exist; and*

- *the increasing concave ordering, written $\boldsymbol{X} \leq_{icv} \boldsymbol{Y}$, if (3.1) holds for all increasing concave functions $\varphi : \mathbb{R}^n \to \mathbb{R}$ whenever the expectations exist.*

Let X and Y be $\mathbb{R}$-valued rvs. We note from [59, p. 3] that the comparison $X \leq_{st} Y$ is equivalent to

$$\mathbf{P}\left[X > t\right] \leq \mathbf{P}\left[Y > t\right], \quad t \in \mathbb{R}. \tag{3.2}$$

It is also known [59] that if $X \leq_{cx} Y$, we have $\mathbf{E}\left[X\right] = \mathbf{E}\left[Y\right]$ and $Var(X) \leq Var(Y)$. In other words, X has the same mean as Y but less variability than Y. When $X \leq_{icx} Y$, there exists an $\mathbb{R}$-valued rv Z such that $X \leq_{st} Z \leq_{cx} Y$ [48, Thm. 1], whence $\mathbf{E}\left[X\right] \leq \mathbf{E}\left[Y\right]$ and we can interpret Y as being greater than X in both "size and variability." Consequently, the orderings cx and icx are appropriate for comparing the variability of rvs. However, in the case of random vectors, it is also desirable to compare their degree of "dependence." In the next section, we describe a stochastic ordering which is well suited for comparing the dependence structures of random vectors and sequences.

A few words on the notation in use: Two $\mathbb{R}^n$-valued rvs $\boldsymbol{X}$ and $\boldsymbol{Y}$ are said to be *equal in law* if they have the same distribution, a fact we denote by $\boldsymbol{X} =_{st} \boldsymbol{Y}$. For two

sequences of rvs $\boldsymbol{X} = \{X_n, n = 1, 2, \ldots\}$ and $\boldsymbol{Y} = \{Y_n, n = 1, 2, \ldots\}$, the notation $\boldsymbol{X} =_{st} \boldsymbol{Y}$ indicates that $\boldsymbol{X}$ and $\boldsymbol{Y}$ have the same finite dimensional distributions, i.e., $(X_1, \ldots, X_n) =_{st} (Y_1, \ldots, Y_n)$ for all $n = 1, 2, \ldots$. Lastly, convergence in law or in distribution (with t going to infinity) is denoted by $\Longrightarrow_t$.

3.2   Supermodular ordering

Several stochastic orderings have been found well suited for comparing the dependence structures of random vectors. Here we rely on the *supermodular* ordering which has been used recently in several queueing and reliability applications [7, 8, 9, 60, 66]. We begin by introducing the class of functions associated with this ordering.

Definition 3.3 *A function $\varphi : \mathbb{R}^n \to \mathbb{R}$ is said to be supermodular (sm) if*

$$\varphi(\boldsymbol{x} \vee \boldsymbol{y}) + \varphi(\boldsymbol{x} \wedge \boldsymbol{y}) \geq \varphi(\boldsymbol{x}) + \varphi(\boldsymbol{y}), \quad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$$

where we set $\boldsymbol{x} \vee \boldsymbol{y} = (x_1 \vee y_1, \ldots, x_n \vee y_n)$ and $\boldsymbol{x} \wedge \boldsymbol{y} = (x_1 \wedge y_1, \ldots, x_n \wedge y_n)$.

The supermodular ordering is the integral ordering associated with the class of supermodular functions.

Definition 3.4 *For $\mathbb{R}^n$-valued rvs $\boldsymbol{X}$ and $\boldsymbol{Y}$, the rv $\boldsymbol{X}$ is said to be smaller than the rv $\boldsymbol{Y}$ according to the supermodular ordering, written $\boldsymbol{X} \leq_{sm} \boldsymbol{Y}$, if (3.1) holds for all supermodular Borel measurable functions $\varphi : \mathbb{R}^n \to \mathbb{R}$ whenever the expectations exist.*

It is a simple matter to check [8] that for any $\mathbb{R}^n$-valued rvs $\boldsymbol{X}$ and $\boldsymbol{Y}$, the comparison $\boldsymbol{X} \leq_{sm} \boldsymbol{Y}$ necessarily implies the stochastic equalities

$$X_i =_{st} Y_i, \quad i = 1, \ldots, n, \tag{3.3}$$

as well as the covariance comparisons

$$\mathrm{Cov}[X_i, X_j] \leq \mathrm{Cov}[Y_i, Y_j], \quad i, j = 1, 2, \ldots, n. \tag{3.4}$$

Thus, the comparison $\boldsymbol{X} \leq_{sm} \boldsymbol{Y}$ represents a possible formalization of the statement to the effect that "$\boldsymbol{Y}$ is more positively dependent than $\boldsymbol{X}$."

The definition of the supermodular ordering can be extended to sequences of rvs in a natural way.

Definition 3.5 *We say that the two $\mathbb{R}$-valued sequences $\boldsymbol{X} = \{X_n, n = 1, 2, \ldots\}$ and $\boldsymbol{Y} = \{Y_n, n = 1, 2, \ldots\}$ satisfy the relation $\boldsymbol{X} \leq_{sm} \boldsymbol{Y}$ if $(X_1, \ldots, X_n) \leq_{sm} (Y_1, \ldots, Y_n)$ for all $n = 1, 2, \ldots$.*

In what follows, we introduce several concepts of positive dependence.

3.3   Positive dependence

Positive dependence in a collection of rvs can be captured in several ways. The association of rvs is one of the most useful such characterizations; it was introduced by Esary, Proschan and Walkup [31] and has proved useful in various settings [6, 42] (and references therein).

Definition 3.6 *The $\mathbb{R}^n$-valued rv $\boldsymbol{X} = (X_1, \ldots, X_n)$ is said to be associated[1] if the inequality*

$$\mathbf{E}\left[f(\boldsymbol{X})g(\boldsymbol{X})\right] \geq \mathbf{E}\left[f(\boldsymbol{X})\right]\mathbf{E}\left[g(\boldsymbol{X})\right]$$

holds for all increasing functions $f, g : \mathbb{R}^n \to \mathbb{R}$ for which the expectations exist.

A stronger notion of positive dependence is given by

---

[1]Sometimes, we say that the $\mathbb{R}$-valued rvs $X_1, \ldots, X_n$ are associated.

Definition 3.7 *The $\mathbb{R}^n$-valued rv $\boldsymbol{X} = (X_1, \ldots, X_n)$ is said to be conditionally increasing in sequence (CIS) if for each $k = 1, 2, \ldots, n-1$, the family of conditional distributions $\{[X_{k+1}|X_1 = x_1, \ldots, X_k = x_k]\}$ is stochastically increasing in $\boldsymbol{x} = (x_1, \ldots, x_k)$.*

More precisely, this definition states that for each $k = 1, 2, \ldots, n-1$, for $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^k$ with $\boldsymbol{x} \leq \boldsymbol{y}$ componentwise, it holds that

$$[X_{k+1}|(X_1, \ldots, X_k) = \boldsymbol{x}] \leq_{st} [X_{k+1}|(X_1, \ldots, X_k) = \boldsymbol{y}]$$

where $[X_{k+1}|(X_1, \ldots, X_k) = \boldsymbol{x}]$ denotes any rv distributed according to the conditional distribution of X_{k+1} given $(X_1, \ldots, X_k) = \boldsymbol{x}$ (with a similar interpretation for $[X_{k+1}|(X_1, \ldots, X_k) = \boldsymbol{y}]$).

We next show how the supermodular ordering induces a notion of positive dependence but first, a definition:

Definition 3.8 *For $\mathbb{R}^n$-valued rvs $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$, we say that $\hat{\boldsymbol{X}} = (\hat{X}_1, \ldots, \hat{X}_n)$ is an independent version of $\boldsymbol{X} = (X_1, \ldots, X_n)$ if the rvs $\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_n$ are mutually independent with $\hat{X}_k =_{st} X_k$, for each $k = 1, \ldots, n$.*

From the concept of supermodular ordering, the positive dependence between the components $X_1, \ldots, X_n$ of the $\mathbb{R}^n$-valued rv $\boldsymbol{X}$ can be formalized by requiring that the rv $\boldsymbol{X}$ be larger in the supermodular ordering than its independent version $\hat{\boldsymbol{X}}$. This gives rise to the following notion of positive dependence [52]:

Definition 3.9 *The $\mathbb{R}^n$-valued rv $\boldsymbol{X} = (X_1, \ldots, X_n)$ is said to be positive supermodular dependent (PSMD) if*

$$\hat{\boldsymbol{X}} \leq_{sm} \boldsymbol{X} \tag{3.5}$$

where $\hat{\boldsymbol{X}}$ is the independent version of $\boldsymbol{X}$.

The next proposition explores the relationships between the various notions of positive dependence introduced thus far.

Theorem 3.10 *Consider an $\mathbb{R}^n$-valued rv $\boldsymbol{X} = (X_1, \ldots, X_n)$.*

(a) *If $\boldsymbol{X}$ is CIS, then $\boldsymbol{X}$ is associated; and*

(b) *If $\boldsymbol{X}$ is associated, then $\boldsymbol{X}$ is PSMD.*

Part (a) can be found in the monograph by Barlow and Proschan [6, Thm. 4.7, p. 146] while Part (b) has been established recently by Christofides and Vaggelatou [21, Thm. 1]. Earlier, Meester and Shanthikumar [51, Thm. 3.8] have shown that CIS implies PSMD.

Lastly, we naturally extend these definitions to sequences of rvs along the lines of Definition 3.5.

Definition 3.11 *For sequences of $\mathbb{R}$-valued rvs $\boldsymbol{X} = \{X_n, n = 1, 2, \ldots\}$ and $\hat{\boldsymbol{X}} = \{\hat{X}_n, n = 1, 2, \ldots\}$, we say that $\hat{\boldsymbol{X}}$ is an independent version of $\boldsymbol{X}$ if the rvs $\{\hat{X}_n, n = 1, 2, \ldots\}$ are mutually independent with $\hat{X}_n =_{st} X_n$ for all $n = 1, 2, \ldots$.*

Definition 3.12 *We say that the $\mathbb{R}$-valued sequence $\boldsymbol{X} = \{X_n, n = 1, 2, \ldots\}$ is associated (resp. CIS, PSMD) if for each $n = 1, 2, \ldots$, the $\mathbb{R}^n$-valued rv $(X_1, \ldots, X_n)$ is associated (resp. CIS, PSMD).*

Chapter 4

Demand-driven Caching

Consider a universe $\mathcal{N}$ of N cacheable documents, say $\mathcal{N} := \{1, \ldots, N\}$. The system is composed of a server where a copy of each of these N documents is available, and of a cache of size M ($1 \leq M < N$). 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. The document selected for eviction is determined through a *cache replacement* or *eviction* policy.[1]

We now develop below a mathematical framework to address some of the issues discussed in this dissertation. Additional details are available in the monographs by Aven, Coffman and Kogan [2] and by Coffman and Denning [24]. 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. $\Lambda(M; \mathcal{N})$). For each $i = 1, \ldots, N$, let $\Lambda_i^\star(M; \mathcal{N})$

---

[1] We use the terms interchangeably.

(resp. $\Lambda_i(M;\mathcal{N})$) denote the set of elements in $\Lambda^\star(M;\mathcal{N})$ (resp. $\Lambda(M;\mathcal{N})$) which do *not* contain i, i.e.,

$$\Lambda_i^\star(M;\mathcal{N}) := \{s = \{i_1, \ldots i_M\} \in \Lambda^\star(M;\mathcal{N}) : i \notin s\}$$

and

$$\Lambda_i(M;\mathcal{N}) := \{s = (i_1, \ldots i_M) \in \Lambda(M;\mathcal{N}) : i \notin s\}.$$

4.1   A simple framework

Consecutive user requests are modeled by a sequence of $\mathcal{N}$-valued rvs $\boldsymbol{R} = \{R_t, \; t = 0, 1, \ldots\}$. For simplicity we say that request R_t occurs at time $t = 0, 1, \ldots$. Let S_t denote the cache just before time t so that S_t is a subset of $\mathcal{N}$ with at most M elements. Also, the decision to be performed according to the eviction policy in force is the identity U_t of the document in S_t which needs to be evicted in order to make room for the request R_t (if the cache is already full).

Demand-driven caching considered here is characterized by the dynamics

$$S_{t+1} \;=\; \begin{cases} S_t & \text{if } R_t \in S_t \\ S_t + R_t & \text{if } R_t \notin S_t, |S_t| < M \\ S_t - U_t + R_t & \text{if } R_t \notin S_t, |S_t| = M \end{cases} \qquad (4.1)$$

for all $t = 0, 1, \ldots$, 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) Actions are taken only at the time requests are made, hence the terminology demand-driven caching; (ii) a requested document not in cache is *always* added to the cache if the cache is not full at the time of request; and (iii) eviction is *mandatory* if the request R_t is not in cache S_t and the cache S_t is full, i.e., $|S_t| = M$.

4.2 Web request models and reduced dynamics

Throughout we assume the following for the request stream $\boldsymbol{R} = \{R_t, \ t = 0, 1, \ldots\}$:
The popularity pmf $\boldsymbol{p} = (p(1), \ldots, p(N))$ of $\boldsymbol{R}$ exists and is defined as the *non-random* limits

$$p(i) = \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[R_\tau = i\right] \quad a.s., \quad i = 1, \ldots, N. \tag{4.2}$$

To avoid uninteresting situations, it is *always* the case that

$$p(i) > 0, \quad i = 1, \ldots, N. \tag{4.3}$$

A pmf $\boldsymbol{p}$ on $\{1, \ldots, N\}$ satisfying (4.3) is said to be *admissible*.[2]

Under this non-triviality condition (4.3), every document will eventually be requested as we note that

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

under the assumption (4.2). Thus, 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, i.e., for all $t = 0, 1, \ldots$, we have $|S_t| = M$ and (4.1) simplifies to

$$S_{t+1} = \begin{cases} S_t & \text{if } R_t \in S_t \\ S_t - U_t + R_t & \text{if } R_t \notin S_t. \end{cases} \tag{4.4}$$

A number of request models will be considered here, the best known one being the *Independent Reference Model (IRM)*. The IRM will serve as the first model for which we attempt to formalize the folk theorems introduced in this dissertation. It is a basic model which is often used for checking various properties of caching systems [13].

---

[2]Additional assumptions on the request streams, e.g., stationarity and ergodicity, will be required in some parts of the dissertation and will be stated when appropriate.

Moreover, recent results by Jelenkovic and Radovanovic [38] and by Sugimoto and Miyoshi [63] suggest some form of insensitivity of caching systems to the statistics of requests. However, the IRM does not possess any of the correlations which have been observed in Web reference streams, thus making it less suitable for modeling streams of requests with strong temporal correlations. Some examples of models displaying temporal correlations will be discussed later in Chapter 9.

4.3   Cache states and eviction policies

The decisions $\{U_t, \ t = 0, 1, \ldots\}$ are determined through an eviction policy; several examples will be presented shortly. For most eviction policies considered in the literature, as well as here, the dynamics of the cache can be characterized through the evolution of suitably defined variables $\{\Omega_t, \ t = 0, 1, \ldots\}$ where Ω_t is known as the *state of the cache* at time t.

Consider an eviction policy π. The cache state is specific to the eviction policy and is selected with the following in mind: (i) The set S_t of documents in the cache at time t can be recovered from Ω_t; (ii) the cache state Ω_{t+1} is fully determined through the knowledge of the triple (Ω_t, R_t, U_t) in a way that is compatible with the dynamics (4.4); and (iii) the eviction decision U_t at time t can be expressed as a function of the past $(\Omega_0, R_0, U_0, \ldots, \Omega_{t-1}, R_{t-1}, U_{t-1}, \Omega_t, R_t)$ (possibly through suitable randomization), i.e., for each $t = 0, 1, \ldots$, there exists a mapping π_t such that

$$U_t = \pi_t(\Omega_0, R_0, U_0, \ldots, \Omega_{t-1}, R_{t-1}, U_{t-1}, \Omega_t, R_t; \Xi_t) \tag{4.5}$$

where Ξ_t is a rv taken independent of the past $(\Omega_0, R_0, U_0, \ldots, \Omega_{t-1}, R_{t-1}, U_{t-1}, \Omega_t, R_t)$. Collectively, the mappings $\{\pi_t, \ t = 0, 1, \ldots\}$ define the eviction policy π.

We close this section with some examples of eviction policies which have been dis-

cussed in the literature (see e.g., [2, 24]):

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.

Any permutation σ of $\{1, \ldots, N\}$ induces an ordering of the documents by considering the documents $\sigma(1), \sigma(2), \ldots, \sigma(N)$ as "ordered" in decreasing order. This ranking of the documents allows us to define the eviction *policy A_σ* as follows: 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_σ prescribes the eviction of the document U_t given by

$$U_t = \arg\max \left(\sigma^{-1}(j) : \ j \in S_t \right). \tag{4.6}$$

The documents $\sigma(1), \ldots, \sigma(M-1)$, once loaded in the cache, will never be evicted, and in the steady state, the cache under the policy A_σ will contain the documents $\sigma(1), \ldots, \sigma(M-1)$.

The so-called *policy A_0* is associated with the underlying popularity pmf $\boldsymbol{p}$ of the request stream, and evicts the least popular document in the cache, i.e., when the replacement is required at time $t = 0, 1, \ldots$, select U_t to be

$$U_t = \arg\min \left(p(j) : \ j \in S_t \right). \tag{4.7}$$

This policy A_0 coincides with the policy $A_{\sigma^\star}$ associated with the permutation $\sigma^\star$ of $\{1, \ldots, N\}$ which orders the components of the underlying pmf $\boldsymbol{p}$ in decreasing order, namely $p(\sigma^\star(1)) \geq p(\sigma^\star(2)) \geq \ldots \geq p(\sigma^\star(N))$.

Under the random policy and the policies A_σ, we can take the cache state to be the (unordered) set of documents in the cache, i.e., the cache state is an element of $\Lambda^\star(M; \mathcal{N})$ and $\Omega_t = S_t$ for all $t = 0, 1, \ldots$.

The *First-in/First-out (FIFO)* policy replaces the document which has been in cache for the longest time, while the *Least-Recently-Used (LRU)* policy evicts the least recently requested document already in cache.

The *CLIMB* policy is a close relative of the LRU policy. It ranks documents in cache according to their recency of access: If the request 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 the cache at position 1.

The definition of the FIFO, LRU and CLIMB policies necessitates that the cache state be an element of $\Lambda(M; \mathcal{N})$ with Ω_t being a permutation of the elements in S_t for all $t = 0, 1, \ldots$.

4.4 Miss rate

A standard performance metric to evaluate and compare various caching policies is the *miss rate* of a cache. This quantity has the interpretation of being the long-term frequency of the event that the requested document is not in the cache, and therefore determines the effectiveness of a caching policy.

For a given request stream $\boldsymbol{R} = \{R_t, t = 0, 1 \ldots\}$, the miss rate $M_\pi(\boldsymbol{R})$ under a cache replacement policy π is defined as the a.s. limit

$$M_\pi(\boldsymbol{R}) = \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[R_\tau \notin S_\tau\right] \quad a.s. \tag{4.8}$$

(whenever the limit exists) where S_τ denotes the set of documents in cache operating under the replacement policy π at time τ when the input to the cache is the request stream $\boldsymbol{R}$. Almost sure convergence in (4.8) (and elsewhere) is taken under the probability measure on the sequence of rvs $\{\Omega_t, R_t, U_t, \ t = 0, 1, \ldots\}$ induced by the request stream $\{R_t, \ t = 0, 1, \ldots\}$ through the eviction policy π.

The existence of the limit (4.8) depends on the request stream $\boldsymbol{R}$ and on the cache

replacement policy π. Even in the case where the limit (4.8) exists, its expression is not known for general classes of request streams. However, when the request stream $\boldsymbol{R}$ is the IRM, the limit (4.8) exists under most cache replacement policies of interest. This special case will be treated in Chapter 5.

4.5 Output

Under the demand-driven caching operation (4.4), the output of the cache is the sequence of requests that incur a miss, i.e., when the incoming request cannot find the desired document in the cache. More precisely, a miss occurs at time t if R_t is *not* in S_t. Thus, we define recursively the time indices $\{\nu_k, k = 0, 1, \ldots\}$ by

$$\nu_0 = 0; \quad \nu_{k+1} := \nu_k + \eta_{k+1}, \quad k = 0, 1, \ldots$$

and

$$\eta_{k+1} := \inf \{\ell = 1, 2, \ldots : R_{\nu_k + \ell} \notin S_{\nu_k + \ell}\}$$

with the convention $\eta_{k+1} = \infty$ if either $\nu_k = \infty$ or if ν_k is finite but the set of indices entering the definition of η_{k+1} is empty. With δ denoting an element *not* in $\mathcal{N}$, we define the output process $\boldsymbol{R}^\star = \{R_k^\star, k = 1, 2, \ldots\}$ simply as

$$R_k^\star := \begin{cases} R_{\nu_k} & \text{if } \nu_k < \infty \\ \delta & \text{if } \nu_k = \infty \end{cases}$$

for each $k = 1, 2, \ldots$. The requests $\{R_k^\star, k = 1, 2, \ldots\}$ are those requests among $\{R_t, t = 0, 1, \ldots\}$ which incur a miss and which get forwarded to the server (or to the higher level cache in a hierarchical caching system).

The statistics of the output stream $\{R_k^\star, k = 1, 2, \ldots\}$ are determined by the statistics of the input stream $\{R_t, t = 0, 1, \ldots\}$ and by the cache replacement policy π in use. We

are interested in evaluating the popularity pmf $\boldsymbol{p}_\pi^\star = (p_\pi^\star(1), \ldots, p_\pi^\star(N))$ defined by

$$p_\pi^\star(i) := \lim_{K \to \infty} \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}\left[R_k^\star = i\right] \quad a.s. \tag{4.9}$$

for each $i = 1, \ldots, N$, whenever these limits exist.

As with the limit (4.8) of the miss rate, the existence and form of the limits (4.9) are not known for general classes of input models. However, as we shall see in the next chapter, when the input stream is modeled according to the IRM, the limits (4.9) exist and admit simple expressions for most cache replacement policies of interest.

Chapter 5

The Independent Reference Model (IRM)

The *Independent Reference Model (IRM)* is a basic model for Web reference streams; it is commonly used to evaluate various properties of caching policies [13]. We say that the request stream $\boldsymbol{R} = \{R_t, t = 0, 1, \ldots\}$ is an IRM with popularity pmf $\boldsymbol{p}$ if the rvs $\{R_t, t = 0, 1, \ldots\}$ are i.i.d. rvs distributed according to the pmf $\boldsymbol{p}$. In this chapter, we show that under the IRM with popularity pmf $\boldsymbol{p}$ and under a particular cache replacement policy π, the limit (4.8) for the miss rate and the limits (4.9) for the output popularity pmf $\boldsymbol{p}_\pi^\star$ exist and admit simple expressions whenever the a.s. limit

$$\mu_\pi^\star(s; \boldsymbol{p}) = \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[S_\tau = s\right] \quad a.s. \tag{5.1}$$

exists for each element s in $\Lambda^\star(M; \mathcal{N})$ with S_τ being the set of documents in cache at time τ. We now discuss these results for the miss rate and for the output popularity pmf, respectively.

5.1   Miss rate under the IRM

Before stating the main result, we note from the definition of the IRM that the requests $\{R_t, \ t = 0, 1, \ldots\}$ are characterized solely by the popularity pmf $\boldsymbol{p}$ and thus all IRM

streams with the same popularity pmf $\boldsymbol{p}$ must produce the same miss rate (4.8) under a given replacement policy π. Therefore, it is more appropriate to view the miss rate under the IRM as a function of the popularity pmf $\boldsymbol{p}$ and denote the limit (4.8) by $\hat{M}_\pi(\boldsymbol{p})$ to reflect this fact.

Theorem 5.1 *Consider an eviction policy π such that the limits (5.1) exist under the IRM with popularity pmf $\boldsymbol{p}$. Then, the limit (4.8) exists and is given by*

$$
\begin{aligned}
\hat{M}_\pi(\boldsymbol{p}) &= \sum_{i=1}^{N} p(i) \sum_{s \in \Lambda_i^\star(M;\mathcal{N})} \mu_\pi^\star(s;\boldsymbol{p}) && (5.2) \\
&= \sum_{s \in \Lambda^\star(M;\mathcal{N})} \mu_\pi^\star(s;\boldsymbol{p}) \sum_{i \notin s} p(i). && (5.3)
\end{aligned}
$$

Theorem 5.1 is established in the process of proving Theorem 5.2 in Section 5.3. The existence of the limits (5.1) is a mild assumption which is satisfied under all eviction policies of interest considered here (and in the literature). Indeed, under the IRM with popularity pmf $\boldsymbol{p}$, the sequence of cache states $\{\Omega_t,\ t = 0, 1, \ldots\}$ usually form a Markov chain over a finite state space, and standard ergodic results for finite state Markov chains readily yield the existence of the limits (5.1). This issue will be briefly discussed in each situation at the appropriate time. Note also that the limits (4.8) and (5.1) under the IRM are often constants which are independent of the initial cache state Ω_0. However this is not always the case as we shall see in the discussion of RORA policies [Chapter 7].

5.2 Output under the IRM

In this section, we establish the existence and form of the limits (4.9) when the input to the cache is the IRM with popularity pmf $\boldsymbol{p}$. We again do so under the assumption that the a.s. limit (5.1) exists for each s in $\Lambda^\star(M;\mathcal{N})$. The main result is contained in

Theorem 5.2 *Consider an eviction policy π such that the limits (5.1) exist under the IRM with popularity pmf $\boldsymbol{p}$. For each $i = 1, \ldots, N$, the limit (4.9) exists and is given by*

$$
\begin{aligned}
p_\pi^\star(i) &= \lim_{K \to \infty} \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}\left[R_k^\star = i\right] \\
&= \frac{p(i) m_\pi(i; \boldsymbol{p})}{\sum_{j=1}^{N} p(j) m_\pi(j; \boldsymbol{p})} \quad a.s.
\end{aligned}
\tag{5.4}
$$

where we have set

$$
m_\pi(i; \boldsymbol{p}) := \sum_{s \in \Lambda_i^\star(M;\mathcal{N})} \mu_\pi^\star(s; \boldsymbol{p}).
\tag{5.5}
$$

A proof of Theorem 5.2 is given in next section. Note that the existence of the limits (5.1) implies

$$
\begin{aligned}
m_\pi(i; \boldsymbol{p}) &= \sum_{s \in \Lambda_i^\star(M;\mathcal{N})} \left(\lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[S_\tau = s\right] \right) \\
&= \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \sum_{s \in \Lambda_i^\star(M;\mathcal{N})} \mathbf{1}\left[S_\tau = s\right] \\
&= \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[i \notin S_\tau\right] \quad a.s.
\end{aligned}
\tag{5.6}
$$

for each $i = 1, \ldots, N$, and $m_\pi(i; \boldsymbol{p})$ thus represents the fraction of times that document i will not be in the cache. This quantity is determined by the popularity pmf $\boldsymbol{p}$ of the IRM input and by the eviction policy π in use.

Inspection of (5.2) and (5.5) reveals that

$$
\sum_{i=1}^{N} p(i) m_\pi(i; \boldsymbol{p}) = \hat{M}_\pi(\boldsymbol{p}).
\tag{5.7}
$$

This leads via (5.4) to a simple connection between the miss rate of an eviction policy and the pmf of its output in the form

$$
p_\pi^\star(i) = \frac{p(i) m_\pi(i; \boldsymbol{p})}{\hat{M}_\pi(\boldsymbol{p})}, \quad i = 1, \ldots, N.
\tag{5.8}
$$

Thus, with the IRM input, we can view $p_\pi^\star(i)$ as the ratio of the miss rate of the cache when the requested document is i to the overall miss rate of the cache.

5.3 Proofs of Theorems 5.1 and 5.2

Key to the proofs of both Theorems 5.1 and 5.2 is the following observation: For each $t = 0, 1, \ldots$, the rvs Ω_t and R_t are independent. Hence, by independence of rvs $\{R_t, t = 0, 1, \ldots\}$, upon invoking Rajchman's version of the Strong Law of Large Numbers [22, Thm. 5.1.2., p. 103], we find

$$\lim_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[S_\tau = s\right] \left(\mathbf{1}\left[R_\tau = i\right] - p(i)\right) = 0 \quad a.s. \tag{5.9}$$

for *each* s in $\Lambda^\star(M; \mathcal{N})$ and $i = 1, \ldots, N$.

For each $t = 1, 2, \ldots$, let $K(t)$ denote the total number of misses up to time t. Obviously, we have

$$K(t) := \sum_{\tau=1}^{t} \mathbf{1}\left[R_\tau \notin S_\tau\right] = \sum_{i=1}^{N} \sum_{\tau=1}^{t} \mathbf{1}\left[i \notin S_\tau\right] \mathbf{1}\left[R_\tau = i\right]. \tag{5.10}$$

Fix $i = 1, \ldots, N$. We note that

$$\begin{aligned}
\sum_{k=1}^{K(t)} \mathbf{1}\left[R_k^\star = i\right] &= \sum_{\tau=1}^{t} \mathbf{1}\left[i \notin S_\tau\right] \mathbf{1}\left[R_\tau = i\right] \\
&= p(i) \sum_{\tau=1}^{t} \mathbf{1}\left[i \notin S_\tau\right] \\
&\quad + \sum_{\tau=1}^{t} \mathbf{1}\left[i \notin S_\tau\right] \left(\mathbf{1}\left[R_\tau = i\right] - p(i)\right).
\end{aligned} \tag{5.11}$$

It is now plain from (5.9) that

$$\begin{aligned}
&\lim_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[i \notin S_\tau\right] \left(\mathbf{1}\left[R_\tau = i\right] - p(i)\right) \\
&= \sum_{s \in \Lambda_i^\star(M; \mathcal{N})} \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[S_\tau = s\right] \left(\mathbf{1}\left[R_\tau = i\right] - p(i)\right) = 0 \quad a.s. \tag{5.12}
\end{aligned}$$

Next, combining (5.6) and (5.12), we get via (5.11) that

$$\lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[i \notin S_\tau\right] \mathbf{1}\left[R_\tau = i\right] \; = \; p(i) \sum_{s \in \Lambda_i^\star(M;\mathcal{N})} \mu_\pi^\star(s; \boldsymbol{p}) \quad a.s. \qquad (5.13)$$

Using the basic identity (5.10) for each $t = 1, 2, \ldots$, we conclude from (5.13) that

$$\lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[R_\tau \notin S_\tau\right] \; = \; \sum_{i=1}^{N} \left(\lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[i \notin S_\tau\right] \mathbf{1}\left[R_\tau = i\right] \right)$$

$$= \; \sum_{i=1}^{N} p(i) \sum_{s \in \Lambda_i^\star(M;\mathcal{N})} \mu_\pi^\star(s; \boldsymbol{p}) \quad a.s. \qquad (5.14)$$

This last limit yields the expression (5.2) for the miss rate (4.8).

To establish (5.3), we observe for each $t = 1, 2, \ldots$ that

$$\sum_{\tau=1}^{t} \mathbf{1}\left[R_\tau \notin S_\tau\right] \; = \; \sum_{\tau=1}^{t} \sum_{s \in \Lambda^\star(M;\mathcal{N})} \mathbf{1}\left[S_\tau = s\right] \left(\mathbf{1}\left[R_\tau \notin s\right] - \sum_{i \notin s} p(i) \right)$$

$$+ \sum_{\tau=1}^{t} \sum_{s \in \Lambda^\star(M;\mathcal{N})} \mathbf{1}\left[S_\tau = s\right] \cdot \left(\sum_{i \notin s} p(i) \right).$$

It then follows from (5.9) that

$$\lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \sum_{s \in \Lambda^\star(M;\mathcal{N})} \mathbf{1}\left[S_\tau = s\right] \left(\mathbf{1}\left[R_\tau \notin s\right] - \sum_{i \notin s} p(i) \right) = 0 \quad a.s.$$

so that

$$\lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[R_\tau \notin S_\tau\right] = \sum_{s \in \Lambda^\star(M;\mathcal{N})} \left(\lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[S_\tau = s\right] \right) \cdot \left(\sum_{i \notin s} p(i) \right) \quad a.s.$$

and the expression (5.3) is obtained under the existence of the limits (5.1). This completes the proof of Theorem 5.1.

It is now immediate that the following limit exists a.s., and is given by

$$\lim_{t \to \infty} \frac{1}{K(t)} \sum_{k=1}^{K(t)} \mathbf{1}\left[R_k^\star = i\right] \; = \; \frac{\lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[i \notin S_\tau\right] \mathbf{1}\left[R_\tau = i\right]}{\lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[R_\tau \notin S_\tau\right]}$$

$$= \; \frac{p(i) m_\pi(i; \boldsymbol{p})}{\sum_{j=1}^{N} p(j) m_\pi(j; \boldsymbol{p})} \quad a.s. \qquad (5.15)$$

38

as we note (5.13) and (5.14). The desired conclusion of Theorem 5.2 is readily obtained from (5.15) once we observe the convergence $\lim_{t\to\infty} K(t) = \infty$ a.s. monotonically so that the sequence $\{K(t),\ t = 1, 2, \ldots\}$ a.s. exhausts $\mathbb{N}$, and the a.s. existence of the limit in (5.15) implies the a.s. existence of the limit (4.9) with limiting value (5.4)-(5.5).

∎

Chapter 6

Comparing Popularity under the Independent Reference

Model

As we have in mind to study the strength of locality of reference present in streams of requests, we first focus on how *popularity* contributes to locality of reference by considering the situation where there are *no* temporal correlations in the stream of requests as would be the case under the IRM with popularity pmf $\boldsymbol{p}$. In this case, the *skewness* in the pmf $\boldsymbol{p}$ 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., $p(1) = \cdots = p(N) = \frac{1}{N}$, for then the successive requests $\{R_t, t = 0, 1, \ldots\}$ form a truly random sequence.

We capture the skewness in the popularity vector through the concept of *majorization* introduced in Chapter 2. From now on, the majorization comparison $\boldsymbol{p} \prec \boldsymbol{q}$ formalizes the notion that the IRM with popularity pmf $\boldsymbol{p}$ has less locality of reference than the

IRM with popularity pmf $\boldsymbol{q}$ as this comparison captures the fact that the pmf $\boldsymbol{q}$ is more skewed than the pmf $\boldsymbol{p}$. Under the IRM, the folk theorem for the miss rate associated with a particular eviction policy π can be restated as follows: If two IRM streams have popularity pmfs $\boldsymbol{p}$ and $\boldsymbol{q}$ satisfying $\boldsymbol{p} \prec \boldsymbol{q}$, then it holds that

$$\hat{M}_\pi(\boldsymbol{q}) \leq \hat{M}_\pi(\boldsymbol{p}), \tag{6.1}$$

i.e., "the more skewed the popularity pmf, the smaller the miss rate of a cache." Similarly, the folk theorem for the output of a cache under the IRM now reads as the comparison $\boldsymbol{p}_\pi^\star \prec \boldsymbol{p}$ in that the output popularity pmf $\boldsymbol{p}_\pi^\star$ is indeed more balanced than the popularity pmf $\boldsymbol{p}$ of the IRM input.

In this chapter, we first discuss some basic comparisons which are consequences of majorization comparison between pmf vectors. We then formally establish the folk theorems for the miss rate and for the output of a cache under the IRM with two well-known cache-replacement policies, namely, the random policy and the policy A_0. Results for more general policies are discussed in Chapter 7 for Random On-demand Replacement Algorithms, and in Chapter 8 for the LRU and CLIMB policies.

6.1 Entropy comparison

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

$$H(\boldsymbol{p}) := -\sum_{i=1}^{N} p(i) \log_2 p(i) \tag{6.2}$$

with the convention $t \log_2 t = 0$ for $t = 0$. It is known that the larger the entropy $H(\boldsymbol{p})$, the more balanced the pmf $\boldsymbol{p}$. This concept has been previously used by Fonseca et al.

[34] to capture the strength of locality of reference exhibited through the popularity pmf of the request stream.

By a classical result of Schur [49, C.1, p. 64] the mapping $x \to - \sum_{i=1}^{N} x_i \log_2 x_i$ is a Schur-concave function on $\mathbb{R}_+^N$. This leads readily to the following well-known result [49, D.1, p. 71].

Proposition 6.1 *For pmfs p and q on $\mathcal{N}$, it holds that*

$$H(\boldsymbol{q}) \leq H(\boldsymbol{p}) \tag{6.3}$$

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 (6.3) proposed by Fonseca et al. [34].

6.2   Zipf-like distributions

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 [1] a good fit was provided by the *Zipf* distribution according to which the popularity of the ith most popular object is inversely proportional to its rank, namely $1/i$.

In more recent studies [13, 39], "Zipf-like" distributions[1] were found more appropriate; see [13] (and references therein) for an excellent summary. Such distributions form a one-parameter family. In our set-up, for $\alpha \geq 0$, we say that the popularity distribution $\boldsymbol{p}$ of the $\mathcal{N}$-valued rvs $\{R_t, \ t = 0, 1, \ldots\}$ is Zipf-like with parameter α if

$$p(i) = \frac{i^{-\alpha}}{C_\alpha(N)}, \quad i = 1, \ldots, N \tag{6.4}$$

---

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

with

$$C_\alpha(N) := \sum_{i=1}^{N} i^{-\alpha}. \tag{6.5}$$

The pmf (6.4) will be denoted by $\boldsymbol{p}_\alpha$. It is always the case that

$$p_\alpha(1) \geq p_\alpha(2) \geq \ldots \geq p_\alpha(N). \tag{6.6}$$

The case $\alpha = 1$ corresponds to the standard Zipf distribution and the value of α was typically found to be in the range $0.64 - 0.83$ [13].

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 (6.4)-(6.5) to measure the strength of skewness, with the larger α, the more skewed the pmf $\boldsymbol{p}_\alpha$. The next result shows that majorization indeed captures this fact, and so it is warranted to call α the *skewness parameter* of the Zipf-like pmf.

Lemma 6.2 *For $0 \leq \alpha < \beta$, it holds that $\boldsymbol{p}_\alpha \prec \boldsymbol{p}_\beta$.*

Lemma 6.2 can already be found in [49, B.2.b, p. 130] and is an easy by-product of Lemma 2.4. Zipf-like distributions will be used in the discussion of the LRU and CLIMB policies in Chapter 8.

6.3 Comparing input and output

In the following two sections, we establish basic comparison results which provide the first step toward formalizing the folk theorem for the output of a cache. We begin with a comparison between the input popularity pmf and the output popularity pmf for a general caching policy.

43

Theorem 6.3 *Consider an eviction policy π such that the limits (5.1) exist under the IRM with popularity pmf $\boldsymbol{p}$.*

(i) If $m_\pi(i;\boldsymbol{p}) \leq m_\pi(j;\boldsymbol{p})$ whenever $p(i) \leq p(j)$ for distinct $i, j = 1, \ldots, N$, then it holds that $\boldsymbol{p} \prec \boldsymbol{p}_\pi^\star$;

(ii) If $m_\pi(i;\boldsymbol{p}) \geq m_\pi(j;\boldsymbol{p})$ whenever $p(i)m_\pi(i;\boldsymbol{p}) \leq p(j)m_\pi(j;\boldsymbol{p})$ for distinct $i, j = 1, \ldots, N$, then it holds that $\boldsymbol{p}_\pi^\star \prec \boldsymbol{p}$ provided $m_\pi(i;\boldsymbol{p}) > 0$ for each $i = 1, \ldots, N$.

Proof. Under the enforced assumptions, both claims are simple consequences of Lemma 2.4: For Claim (i), we use $\boldsymbol{x} = \boldsymbol{p}$ and $\boldsymbol{y}$ given by $y_i = p(i)m_\pi(i;\boldsymbol{p})$, $i = 1, \ldots, N$. Note that $\bar{\boldsymbol{x}} = \boldsymbol{p}$ while $\bar{\boldsymbol{y}} = \boldsymbol{p}_\pi^\star$, and that the monotonicity assumptions hold.

For Claim (ii), we take $\boldsymbol{y} = \boldsymbol{p}$ and $\boldsymbol{x}$ given by $x_i = p(i)m_\pi(i;\boldsymbol{p})$, $i = 1, \ldots, N$. This time, we have $\bar{\boldsymbol{x}} = \boldsymbol{p}_\pi^\star$ while $\bar{\boldsymbol{y}} = \boldsymbol{p}$, and the requisite monotonicity assumptions hold. ∎

Theorem 6.3 suggests the following definitions: We say that the caching algorithm π is *bad* if it has the property that the fraction of time that a document is not in cache increases as its popularity increases, i.e., for every admissible pmf $\boldsymbol{p}$, it holds that $m_\pi(i;\boldsymbol{p}) \leq m_\pi(j;\boldsymbol{p})$ whenever $p(i) \leq p(j)$ for distinct $i, j = 1, \ldots, N$. For a bad caching algorithm, Claim (i) states that the popularity pmf of the output is more skewed than the popularity pmf of the input, or equivalently that the output stream displays stronger locality of reference than the input stream.

The assumptions for Claim (ii) ensure that $m_\pi(i;\boldsymbol{p}) \leq m_\pi(j;\boldsymbol{p})$ and $p(j) \leq p(i)$ occur simultaneously for distinct $i, j = 1, \ldots, N$. This leads to defining a caching algorithm π as *good* if for every admissible pmf $\boldsymbol{p}$, we have $m_\pi(i;\boldsymbol{p}) \leq m_\pi(j;\boldsymbol{p})$ whenever

$p(j) \le p(i)$ for distinct $i, j = 1, \ldots, N$. Thus, a caching policy which satisfies the assumptions of Claim (ii) is necessarily a good policy. However, as we shall see in the case of the LRU and CLIMB policies [Chapter 8], this by itself is not sufficient to ensure that the output popularity pmf is more balanced than the input popularity pmf.

6.4 A useful comparison

Repeatedly we will encounter output pmfs which assume the generic form used in Theorem 6.4 below.

Theorem 6.4 *Let $\boldsymbol{p}$ be an admissible pmf on $\mathcal{N}$, and for each $i = 1, \ldots, N$, define the $(N-1)$-dimensional vector*

$$\boldsymbol{p}^{(i)} := (p(1), \ldots, p(i-1), p(i+1), \ldots, p(N)). \tag{6.7}$$

For each $M = 1, 2, \ldots, N-1$, the pmf $\boldsymbol{p}_M^\star$ on $\mathcal{N}$ defined by

$$p_M^\star(i) = \frac{p(i) E_{M,N-1}(\boldsymbol{p}^{(i)})}{\sum_{j=1}^N p(j) E_{M,N-1}(\boldsymbol{p}^{(j)})}, \quad i = 1, \ldots, N \tag{6.8}$$

satisfies the comparison $\boldsymbol{p}_M^\star \prec \boldsymbol{p}$ where the elementary symmetric function $E_{M,N-1} : \mathbb{R}^{N-1} \to \mathbb{R}$ is defined at (2.7).

Proof. Fix distinct $i, j = 1, \ldots, N$ and define the $(N-2)$-dimensional vector $\boldsymbol{p}^{(ij)}$ obtained from the pmf $\boldsymbol{p}$ by deleting the components associated with documents i and j. With this notation, we find

$$E_{M,N-1}(\boldsymbol{p}^{(i)}) - E_{M,N-1}(\boldsymbol{p}^{(j)})$$
$$= \sum_{s \in \Lambda_i^\star(M;\mathcal{N})} p(i_1) \cdots p(i_M) - \sum_{s \in \Lambda_j^\star(M;\mathcal{N})} p(i_1) \cdots p(i_M)$$

$$= \sum_{s \in \Lambda_i^\star(M;\mathcal{N}):\, j \in s} p(i_1) \cdots p(i_M) - \sum_{s \in \Lambda_j^\star(M;\mathcal{N}):\, i \in s} p(i_1) \cdots p(i_M)$$

$$= (p(j) - p(i))\, E_{M-1,N-2}(\boldsymbol{p}^{(ij)}). \tag{6.9}$$

On the other hand, we also have

$$p(i) E_{M,N-1}(\boldsymbol{p}^{(i)}) - p(j) E_{M,N-1}(\boldsymbol{p}^{(j)})$$

$$= p(i) \left(\sum_{s \in \Lambda_i^\star(M;\mathcal{N})} p(i_1) \cdots p(i_M) \right) - p(j) \left(\sum_{s \in \Lambda_j^\star(M;\mathcal{N})} p(i_1) \cdots p(i_M) \right)$$

$$= p(i) \left(\sum_{s \in \Lambda_i^\star(M;N):\, j \notin s} p(i_1) \cdots p(i_M) \right) - p(j) \left(\sum_{s \in \Lambda_j^\star(M;N):\, i \notin s} p(i_1) \cdots p(i_M) \right)$$

$$= (p(i) - p(j))\, E_{M,N-2}(\boldsymbol{p}^{(ij)}). \tag{6.10}$$

As we have in mind to apply Lemma 2.4, we take $\boldsymbol{y} = \boldsymbol{p}$ and $\boldsymbol{x}$ given by $x_i = p(i) E_{M,N-1}(\boldsymbol{p}^{(i)})$, $i = 1, \ldots, N$, whence $\bar{\boldsymbol{x}} = \boldsymbol{p}_M^\star$ and $\bar{\boldsymbol{y}} = \boldsymbol{p}$. For distinct $i, j = 1, \ldots, N$, we find from (6.9) and (6.10) that

$$\frac{x_i}{y_i} - \frac{x_j}{y_j} = (p(j) - p(i))\, E_{M-1,N-2}(\boldsymbol{p}^{(ij)}) \leq 0$$

whenever

$$x_i - x_j = (p(i) - p(j))\, E_{M,N-2}(\boldsymbol{p}^{(ij)}) \geq 0.$$

The assumptions of Lemma 2.4 are satisfied and the comparison $\boldsymbol{p}_M^\star \prec \boldsymbol{p}$ follows. ∎

6.5   The random policy

In the last two sections, we formalize the folk theorems under the IRM for the miss rate and the output of a cache under the random policy and the policy A_σ, respectively.

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. When the

input to the cache is the IRM with popularity pmf $\boldsymbol{p}$, the cache states $\{S_t,\ t = 0, 1, \ldots\}$ form a stationary ergodic Markov chain over the finite state space $\Lambda^\star(M; \mathcal{N})$ [2, Thm. 11, p. 132]. Its stationary distribution is given by

$$\mu_{\mathrm{Rand}}^\star(s; \boldsymbol{p}) = E_{M,N}(\boldsymbol{p})^{-1} p(i_1) \cdots p(i_M) \tag{6.11}$$

for every $s = \{i_1, \ldots, i_M\}$ in $\Lambda^\star(M; \mathcal{N})$ with normalizing constant $E_{M,N}(\boldsymbol{p})$ defined at (2.7).

6.5.1   The miss rate under the random policy

Under the IRM with popularity pmf $\boldsymbol{p}$, the corresponding miss rate is obtained from (5.3) and (6.11) (see also [2, Thm. 11, p. 132]) as

$$\hat{M}_{\mathrm{Rand}}(\boldsymbol{p}) = \frac{\sum_{\{i_1,\ldots,i_M\} \in \Lambda^\star(M;\mathcal{N})} p(i_1) \cdots p(i_M) \left(1 - \sum_{k=1}^M p(i_k)\right)}{\sum_{\{i_1,\ldots,i_M\} \in \Lambda^\star(M;\mathcal{N})} p(i_1) \cdots p(i_M)}. \tag{6.12}$$

That (6.1) indeed holds for the random policy is contained in

Theorem 6.5 *For admissible pmfs $\boldsymbol{p}$ and $\boldsymbol{q}$ on $\mathcal{N}$, it holds that*

$$\hat{M}_{\mathrm{Rand}}(\boldsymbol{q}) \leq \hat{M}_{\mathrm{Rand}}(\boldsymbol{p}) \tag{6.13}$$

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

Proof.   First, we note that

$$\sum_{\{i_1,\ldots,i_M\} \in \Lambda^\star(M;\mathcal{N})} p(i_1) \cdots p(i_M) = E_M(\boldsymbol{p}). \tag{6.14}$$

It is also a simple matter to see that

$$\sum_{\{i_1,\ldots,i_M\} \in \Lambda^\star(M;\mathcal{N})} p(i_1) \cdots p(i_M)(1 - \sum_{k=1}^M p(i_k))$$

$$
\begin{aligned}
&= \sum_{\{i_1,\ldots,i_M\}\in\Lambda^\star(M;\mathcal{N})} p(i_1)\cdots p(i_M) \cdot \sum_{i\notin\{i_1,\ldots,i_M\}} p(i) \\
&= (M+1) \sum_{\{i_1,\ldots,i_{M+1}\}\in\Lambda^\star(M+1;\mathcal{N})} p(i_1)\cdots p(i_{M+1}) \\
&= (M+1)E_{M+1}(\boldsymbol{p}).
\end{aligned}
\tag{6.15}
$$

Combining (6.14) and (6.15) through (6.12), we get

$$
\hat{M}_{\mathrm{Rand}}(\boldsymbol{p}) = (M+1)\frac{E_{M+1}(\boldsymbol{p})}{E_M(\boldsymbol{p})},
\tag{6.16}
$$

and the miss rate $\hat{M}_{\mathrm{Rand}}(\boldsymbol{p})$ is Schur-concave in $\boldsymbol{p}$ by Proposition 2.6 . ∎

Under the IRM, it is well known [2, p. 132] that the FIFO policy yields the same miss rate as the random policy, so that Theorem 6.5 holds for the FIFO policy as well.

In the special case $M = 1$, any demand-driven policy reduces to the policy that evicts the only document in cache if the requested document is not in cache. Specializing the results for the random policy, Theorem 6.5 immediately leads to

Corollary 6.6 *With $M = 1$, for admissible pmfs $\boldsymbol{p}$ and $\boldsymbol{q}$, it holds that*

$$
\hat{M}_\pi(\boldsymbol{q}) \le \hat{M}_\pi(\boldsymbol{p})
$$

whenever $\boldsymbol{p} \prec \boldsymbol{q}$ under any demand-driven replacement policy π.

6.5.2 The output under the random policy

As we report (6.11) into (5.5), we readily conclude that

$$
\begin{aligned}
m_{\mathrm{Rand}}(i;\boldsymbol{p}) &= E_{M,N}(\boldsymbol{p})^{-1} \sum_{s\in\Lambda_i^\star(M;\mathcal{N})} p(i_1)\cdots p(i_M) \\
&= \frac{E_{M,N-1}(\boldsymbol{p}^{(i)})}{E_{M,N}(\boldsymbol{p})}, \quad i = 1,\ldots,N
\end{aligned}
\tag{6.17}
$$

where $\boldsymbol{p}^{(i)}$ is the $(N-1)$-dimensional vector (6.7) obtained from the pmf $\boldsymbol{p}$ by deleting the component associated with document i. Consequently, (5.4) yields the output popularity distribution as

$$p_{\text{Rand}}^{\star}(i) = \frac{p(i)E_{M,N-1}(\boldsymbol{p}^{(i)})}{\sum_{j=1}^{N} p(j)E_{M,N-1}(\boldsymbol{p}^{(j)})}, \quad i = 1, \ldots, N \tag{6.18}$$

and Theorem 6.4 immediately implies

Theorem 6.7 *Under the random policy, it holds that $\boldsymbol{p}_{\text{Rand}}^{\star} \prec \boldsymbol{p}$.*

As in the case of miss rate, for the special case $M = 1$, by specializing the results for the random policy, the output pmf is given by

$$p^{\star}(i) = \frac{p(i)(1 - p(i))}{\sum_{j=1}^{N} p(j)(1 - p(j))}, \quad i = 1, \ldots, N \tag{6.19}$$

and Theorem 6.7 readily yields

Corollary 6.8 *With $M = 1$, under any demand-driven replacement policy π, the popularity pmf $\boldsymbol{p}_{\pi}^{\star}$ of the output is the pmf $\boldsymbol{p}^{\star}$ given at (6.19) with $\boldsymbol{p}^{\star} \prec \boldsymbol{p}$.*

6.6   The policy A_{σ}

Let σ denote a permutation of $\{1, \ldots, N\}$ which is held fixed throughout this section. Such a permutation can be used to induce an ordering of the documents by considering that the documents $\sigma(1), \sigma(2), \ldots, \sigma(N)$ are "ordered" in decreasing order. With this ranking of the documents, the policy A_{σ} can be defined as in Section 4.3 with the eviction rule (4.6).

6.6.1 Cache steady state under the policy A_σ

Under (4.3), every document is eventually requested with probability one, so that for sufficiently large time t, the cache S_t under the replacement policy A_σ is of the form

$$S_t := \Sigma + Y_t^\sigma \qquad (6.20)$$

with

$$\Sigma := \{\sigma(1), \sigma(2), \ldots, \sigma(M-1)\} \qquad (6.21)$$

and

$$Y_t^\sigma \in \Sigma^c = \{\sigma(M), \ldots, \sigma(N)\}. \qquad (6.22)$$

As explained earlier, there is then no loss of generality in assuming that the cache is indeed of the form (6.20)-(6.22), in which case the cache state S_t is determined completely by Y_t^σ. Under the IRM, the rvs $\{Y_t^\sigma, t = 0, 1, \ldots\}$ form a stationary ergodic Markov chain over the finite state space Σ^c with stationary distribution $\{\pi_\sigma(y), y \in \Sigma^c\}$ described in the following lemma.

Lemma 6.9 *The limits*

$$\lim_{t \to \infty} \mathbf{P}\left[Y_t^\sigma = y, R_t = x\right] = \pi_\sigma(y)p(x), \quad (x, y) \in \mathcal{N} \times \Sigma^c$$

exist with

$$\pi_\sigma(y) = \lim_{t \to \infty} \mathbf{P}\left[Y_t^\sigma = y\right] = \frac{p(y)}{\sum_{x \notin \Sigma} p(x)}, \quad y \notin \Sigma. \qquad (6.23)$$

The proof of Lemma 6.9 is omitted as it mimics the derivation of a similar result for the policy A_0 [24, Thm. 6.3, p. 268]. Note that (6.23) defines a pmf $\boldsymbol{\pi}_\sigma$ on Σ^c, which is simply the *conditional* pmf induced on Σ^c by the pmf $\boldsymbol{p}$.

6.6.2 The miss rate under the policy A_σ

Under the IRM with popularity pmf p, it follows from Lemma 6.9 and the expression (5.3) that the miss rate under the policy A_σ is given [24, Thm. 6.4, p. 269] by

$$\hat{M}_\sigma(\boldsymbol{p}) = \sum_{i=M}^{N} p(\sigma(i)) - \frac{\sum_{i=M}^{N} p(\sigma(i))^2}{\sum_{i=M}^{N} p(\sigma(i))}. \tag{6.24}$$

From the expression (6.24), it is not hard to see that the folk theorem (6.1) for miss rates under the policy A_σ does not hold in general. However, it does hold under a well-known instance of the policy A_σ, the policy A_0, defined earlier in Section 4.3. This policy A_0 is simply the policy $A_{\sigma^\star}$ where the permutation $\sigma^\star$ of $\{1, \dots, N\}$ orders the components of the underlying pmf p in decreasing order, i.e., $p(\sigma^\star(1)) \geq p(\sigma^\star(2)) \geq \dots \geq p(\sigma^\star(N))$. The analog of Theorem 6.5 for the policy A_0 is given in

Theorem 6.10 *For admissible pmfs $\boldsymbol{p}$ and $\boldsymbol{q}$ on $\mathcal{N}$, it holds that*

$$\hat{M}_{A_0}(\boldsymbol{q}) \leq \hat{M}_{A_0}(\boldsymbol{p}) \tag{6.25}$$

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

Proof. The policy A_0 is known [2, 24] to minimize the miss rate for the IRM amongst a large class of demand-driven policies, including the policies (4.6). In particular, we have

$$\hat{M}_{A_0}(\boldsymbol{p}) = \min_{i=1,\dots,N!} \hat{M}_{\sigma_i}(\boldsymbol{p}) \tag{6.26}$$

where $\{\sigma_i, i = 1, \dots, N!\}$ is a collection of all permutations of $\{1, \dots, N\}$. Furthermore, for any permutation σ of $\{1, \dots, N\}$, we can rewrite (6.24) as

$$\hat{M}_\sigma(\boldsymbol{p}) = \frac{\left(\sum_{i=M}^{N} p(\sigma(i))\right)^2 - \sum_{i=M}^{N} p(\sigma(i))^2}{\sum_{i=M}^{N} p(\sigma(i))}$$

$$\begin{aligned}
&= 2\frac{\sum_{i=M}^{N}\sum_{j=M}^{i-1} p(\sigma(i))p(\sigma(j))}{\sum_{i=M}^{N} p(\sigma(i))} \\
&= 2\frac{E_2(\boldsymbol{t}\cdot\sigma(\boldsymbol{p}))}{E_1(\boldsymbol{t}\cdot\sigma(\boldsymbol{p}))} \\
&= 2\Phi_2(\boldsymbol{t}\cdot\sigma(\boldsymbol{p})) \qquad\qquad\qquad\qquad (6.27)
\end{aligned}$$

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 2.6. Combining these facts with (6.26) and (6.27), we conclude by Proposition 2.8 that the miss rate functional under the policy A_0 is indeed Schur-concave in the pmf vector and the desired result follows. ∎

Without surprise, Corollary 6.6 also follows from Theorem 6.10 (with $M = 1$).

6.6.3 The output under the policy A_σ

From the expression of $\{\pi_\sigma(y), y \in \Sigma^c\}$ provided in Lemma 6.9, we obtain

$$m_\sigma(i; \boldsymbol{p}) = \begin{cases} 0 & \text{if } i \in \Sigma \\ 1 - \pi_\sigma(i) & \text{if } i \notin \Sigma \end{cases}$$

and Theorem 5.2 yields the output popularity distribution $\boldsymbol{p}_\sigma^\star$ as

$$p_\sigma^\star(i) = \begin{cases} 0 & \text{if } i \in \Sigma \\ \frac{p(i)(1-\pi_\sigma(i))}{\sum_{j\notin\Sigma} p(j)(1-\pi_\sigma(j))} & \text{if } i \notin \Sigma. \end{cases} \qquad (6.28)$$

Since $p_\sigma^\star(i) = 0$ whenever i belongs to Σ, it is more natural to seek a comparison between $\boldsymbol{p}_\sigma^\star$ (viewed as a pmf on Σ^c) and the conditional pmf $\boldsymbol{\pi}_\sigma$.

Theorem 6.11 *Under the policy A_σ, it holds that $\boldsymbol{p}_\sigma^\star \prec \boldsymbol{\pi}_\sigma$.*

Proof. We rewrite $p_\sigma^\star$ in (6.28) as a function of π_σ by dividing its numerator and denominator by $\sum_{j \notin \Sigma} p(j)$. This yields

$$p_\sigma^\star(i) = \frac{\pi_\sigma(i)(1 - \pi_\sigma(i))}{\sum_{j \notin \Sigma} \pi_\sigma(j)(1 - \pi_\sigma(j))}, \quad i \notin \Sigma.$$

With Lemma 2.4 in mind, we take x and y to be the elements of $\mathbb{R}^{N-M+1}$ given by $y = \pi_\sigma$ and $x_i = \pi_\sigma(i)(1 - \pi_\sigma(i))$, $i \notin \Sigma$, in which case

$$\frac{y_i}{x_i} = (1 - \pi_\sigma(i))^{-1}, \quad i \notin \Sigma. \tag{6.29}$$

Pick distinct i and j not in Σ. From (6.29), we see that $\frac{y_i}{x_i} \geq \frac{y_j}{x_j}$ if and only if $\pi_\sigma(i) \geq \pi_\sigma(j)$, and the assumptions of Lemma 2.4 will hold if we can show that $x_i \geq x_j$ whenever $\pi_\sigma(i) \geq \pi_\sigma(j)$. The analysis proceeds along two cases:

Case (a) – Assume $\pi_\sigma(i) \leq 1/2$. With $1/2 \geq \pi_\sigma(i) \geq \pi_\sigma(j)$, we find

$$x_i = \pi_\sigma(i)(1 - \pi_\sigma(i)) \geq \pi_\sigma(j)(1 - \pi_\sigma(j)) = x_j$$

by the increasing monotonicity of the mapping $p \to p(1 - p)$ on the interval $[0, \frac{1}{2}]$.

Case (b) – Assume $\pi_\sigma(i) > 1/2$, in which case $1/2 > 1 - \pi_\sigma(i) \geq \pi_\sigma(j)$ since $\sum_{k \notin \Sigma} \pi_\sigma(k) = 1$. We readily arrive at the conclusion $x_i \geq x_j$ by applying the argument in Case (a) to $1 - \pi_\sigma(i)$ and $\pi_\sigma(j)$.

The assumptions of Lemma 2.4 are satisfied and we get the desired result with $\bar{x} = p_\sigma^\star$ and $\bar{y} = \pi_\sigma$. ∎

Corollary 6.8 is also obtained from Theorem 6.11 (with $M = 1$) as expected.

Chapter 7

Random On-demand Replacement Algorithms (RORA)

We now introduce a large class of demand-driven eviction policies called *Random On-demand Replacement Algorithms* (RORA), and show that the folk theorems for the miss rate and the output of a cache hold under this class of policies when the input to the cache is the IRM. This class of policies generalizes many well-known caching policies, e.g., the random and FIFO policies, as well as the optimal policy A_0. Moreover, the Partially Preloaded Random Replacement Algorithms proposed by Gelenbe [35] form a subclass of RORAs.

7.1 Defining RORAs

A RORA policy follows the demand-driven caching rule (4.4) (under the customary assumption that the cache is initially full) and is characterized by an eviction/insertion pmf r on $\{1, \ldots, M\} \times \{1, \ldots, M\}$ which we organize as the $M \times M$ matrix $r = (r_{k\ell})$, 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 r is denoted RORA(r), and often referred to as the RORA(r) policy.

We select the cache state Ω_t at time t to be an element $(i_1, \ldots, i_M)$ of $\Lambda(M; \mathcal{N})$ with

the understanding that document i_k is in cache at position $k = 1, \ldots, M$, at time t. The RORA(r) policy implements the following eviction rule: 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 r, i.e., for each $t = 0, 1, \ldots$, we have

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

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

$$U_t = \mathbf{1}\left[R_t \notin S_t\right] i_{X_t}.$$

We have $U_t = 0$ whenever the requested document is in the cache (i.e., $R_t \in S_t$), in line with the convention that no replacement occurs and the cache state remains unchanged, i.e., $\Omega_{t+1} = \Omega_t$.

Next, if the requested document is not in the cache (i.e., $R_t \notin S_t$) and $(X_t, Y_t) = (k, \ell)$, then $U_t = i_k$, i.e., the 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 $k, k + 1 \ldots, \ell - 1$ (in that order) while if $k > \ell$, the documents $i_\ell, \ldots, i_{k-1}$ are shifted up to position $\ell + 1, \ldots, k$ (in that order). When $k = \ell$, the new document simply replaces the evicted document at position k.

Observe that the document initially at position i in the cache will *never* be replaced if

$$r_{k\ell} = 0 \quad \text{for} \quad \begin{cases} \text{all } k = 1, \ldots, i \text{ and } \ell = i, \ldots, M \\ \text{and} \\ \text{all } \ell = 1, \ldots, i \text{ and } k = i, \ldots, M. \end{cases} \quad (7.1)$$

If we use row i and column i to partition the matrix r into four blocks, then condition (7.1) expresses the fact that the entries in the northwest and southeast corners[1] *all* vanish

---

[1] With the understanding that the position of r_{11} is at the lower left corner of the matrix r.

55

(including row i and column i). Let $\Sigma_{\boldsymbol{r}}$ 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.,

$$\Sigma_{\boldsymbol{r}} := \{i = 1, \ldots, M : \text{ Eqn. (7.1) holds at } i\}. \tag{7.2}$$

Under the IRM with popularity pmf $\boldsymbol{p}$, the cache states $\{\Omega_t, t = 0, 1, \ldots\}$ form a Markov chain on the state space $\Lambda(M; \mathcal{N})$. The ergodic properties of this chain are determined by whether the set $\Sigma_{\boldsymbol{r}}$ is empty or not. This is done in Lemmas 7.1 and 7.2 in the next two sections. These basic results are established in Appendix A.

Throughout the discussion below we always assume that the cache size M and the number of cacheable documents N satisfy $M + 1 < N$. We do so in order to avoid technical cases of limited interest.[2] In addition, the input to the cache is assumed to be the IRM.

7.1.1 Case 1

The set $\Sigma_{\boldsymbol{r}}$ is *empty*, so that *every* document in cache is eventually replaced, i.e., for each $i = 1, \ldots, M$, there exists a pair k, ℓ (possibly depending on i) with either $1 \leq k \leq i \leq \ell \leq M$ or $1 \leq \ell \leq i \leq k \leq M$ such that

$$r_{k\ell} > 0.$$

Here are some well-known policies which fall in this case: The *random policy* corresponds to RORA($\boldsymbol{r}$) with $\boldsymbol{r}$ given by $r_{kk} = \frac{1}{M}$ for each $k = 1, \ldots, M$. The *FIFO policy* also belongs to RORA with two possibilities for $\boldsymbol{r}$, namely $r_{1M} = 1$ or $r_{M1} = 1$. The first (resp. second) choice corresponds to the cache state $(i_1, \ldots, i_M)$ being loaded from

---

[2]This is discussed in some details in Appendix A.

left to right with documents ordered from the oldest to the most recent (resp. from the most recent to the oldest).

In this case, the Markov chain $\{\Omega_t, t = 0, 1, \ldots\}$ is ergodic on the state space $\Lambda(M; \mathcal{N})$; its stationary distribution exists and is given in the following lemma.

Lemma 7.1 *Assume the input to be modeled according to the IRM with popularity pmf p. For any RORA(r) policy in Case 1 with Σ_r empty, the cache states $\{\Omega_t, t = 0, 1, \ldots\}$ form an ergodic Markov chain on the state space $\Lambda(M; \mathcal{N})$ with stationary pmf on $\Lambda(M; \mathcal{N})$ given by*

$$
\begin{aligned}
\mu_r(s; p) &= \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[\Omega_\tau = s\right] \quad a.s. \\
&= C(p)^{-1} p(i_1) p(i_2) \cdots p(i_M)
\end{aligned}
\tag{7.3}
$$

for every $s = (i_1, \ldots, i_M)$ in $\Lambda(M; \mathcal{N})$ with normalizing constant

$$
C(p) := \sum_{(i_1, \ldots, i_M) \in \Lambda(M; \mathcal{N})} p(i_1) p(i_2) \cdots p(i_M).
\tag{7.4}
$$

Note that the stationary pmf is the *same* for *all* RORAs in Case 1.

7.1.2 Case 2

The set Σ_r is *not* empty, and some documents, once put in cache, will never be replaced during the operation of the cache, i.e., if $\Omega_0 = (i_1, \ldots, i_M)$, then for all $t = 1, 2, \ldots$, with $\Omega_t = (j_1, \ldots, j_M)$, we have

$$
j_\ell = i_\ell, \quad \ell \in \Sigma_r.
\tag{7.5}
$$

Here are some examples of RORA policies in that category: For a permutation σ of $\{1, \ldots, N\}$, the policy A_σ evicts the "smallest" document in cache with documents

$\sigma(1), \sigma(2), \ldots, \sigma(N)$ "ordered" in decreasing order. The documents $\sigma(1), \ldots, \sigma(M -$
1), once loaded in the cache, will remain there, and in the steady state, the cache under
the policy A_σ will contain the documents $\sigma(1), \ldots, \sigma(M - 1)$.

This behavior can be recovered through the RORA(r) policy with matrix r of the
form $r_{kk} = 1$ for some $k = 1, \ldots, M$, in which case Σ_r has $M - 1$ elements, namely
$\{1, \ldots, k - 1, k + 1, \ldots, M\}$. If the documents $\sigma(1), \ldots, \sigma(M - 1)$ are initially put
in cache (i.e., preloaded) at the other positions $\ell \neq k$ in Σ_r, this RORA(r) policy will
behave like the policy A_σ in its *steady state* regime. The steady state behavior of the
cache under the policy A_0 is that of the RORA(r) policy above, this time, the preloaded
documents being the $M - 1$ *most popular* documents.

To describe the long-run behavior of the cache states $\{\Omega_t, t = 0, 1, \ldots\}$, we go back
to (7.5). First, with initial cache state $s_0 = (i_1, \ldots, i_M)$ in $\Lambda(M; \mathcal{N})$, we denote by
$\Sigma_r(s_0)$ the set of initial documents with positions in Σ_r, i.e.,

$$\Sigma_r(s_0) := \{i_\ell : \ell \in \Sigma_r\}. \tag{7.6}$$

Next, we introduce the component

$$\Lambda(r, s_0) := \{(j_1, \ldots, j_M) \in \Lambda(M; \mathcal{N}) : j_\ell = i_\ell, \ell \in \Sigma_r\}. \tag{7.7}$$

In view of (7.5), once the cache state is in $\Lambda(r, s_0)$, it remains there forever. In fact
all the states in the component $\Lambda(r, s_0)$ communicate with each other, and this set of
states is closed under the motion of the Markov chain $\{\Omega_t, t = 0, 1, \ldots\}$. Given that
$|\Sigma_r| = m$, there are $\binom{N-m}{M-m}(M - m)!$ elements in $\Lambda(r, s_0)$ and there are $\binom{N}{m} m!$ *distinct*
components which form a partition of $\Lambda(M; \mathcal{N})$.

As a result, when restricted to $\Lambda(r, s_0)$, this Markov chain is irreducible and aperi-
odic, and its ergodic behavior can be characterized as follows:

Lemma 7.2 *Assume the input to be modeled according to the IRM with popularity pmf*

*p. For any RORA(*r*) policy in Case 2 with $|\Sigma_{\boldsymbol{r}}| = m$ and initial cache state s_0, the cache states $\{\Omega_t, t = 0, 1, \ldots\}$ form an ergodic Markov chain on the component $\Lambda(\boldsymbol{r}, s_0)$. In particular the limit*

$$\mu_{\boldsymbol{r}, s_0}(s; \boldsymbol{p}) = \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}[\Omega_\tau = s] \quad a.s. \tag{7.8}$$

always exists for every $s = (i_1, \ldots, i_M)$ in $\Lambda(M; \mathcal{N})$ and is given by

$$\mu_{\boldsymbol{r}, s_0}(s; \boldsymbol{p}) = \begin{cases} C_{\boldsymbol{r}}(\boldsymbol{p}, s_0)^{-1} p(i_1) p(i_2) \cdots p(i_M) & , \quad s \in \Lambda(\boldsymbol{r}, s_0) \\ 0 & , \quad s \notin \Lambda(\boldsymbol{r}, s_0) \end{cases} \tag{7.9}$$

with normalizing constant

$$C_{\boldsymbol{r}}(\boldsymbol{p}, s_0) := \sum_{(i_1, \ldots, i_M) \in \Lambda(\boldsymbol{r}, s_0)} p(i_1) p(i_2) \cdots p(i_M). \tag{7.10}$$

From (7.7), we note the simplification

$$\mu_{\boldsymbol{r}, s_0}(s; \boldsymbol{p}) = C_{\boldsymbol{r}}'(\boldsymbol{p}, s_0)^{-1} \prod_{i_\ell \notin \Sigma_{\boldsymbol{r}}(s_0)} p(i_\ell) \tag{7.11}$$

for each $s = (i_1, \ldots, i_M)$ in $\Lambda(\boldsymbol{r}, s_0)$ with normalizing constant

$$C_{\boldsymbol{r}}'(\boldsymbol{p}, s_0) := \sum_{(i_1, \ldots, i_M) \in \Lambda(\boldsymbol{r}, s_0)} \prod_{i_\ell \notin \Sigma_{\boldsymbol{r}}(s_0)} p(i_\ell). \tag{7.12}$$

7.2 The miss rate under RORAs

7.2.1 Case 1

Fix $s = \{i_1, \ldots, i_M\}$ in $\Lambda^\star(M; \mathcal{N})$, and let $\Lambda(s|M; \mathcal{N})$ denote the subset of $\Lambda(M; \mathcal{N})$ defined by

$$\Lambda(s|M; \mathcal{N}) := \{(j_1, \ldots, j_M) \in \Lambda(M; \mathcal{N}) : \{j_1, \ldots, j_M\} = \{i_1, \ldots, i_M\}\}. \tag{7.13}$$

By Lemma 7.1, the limit (5.1) exists and is given by

$$
\begin{aligned}
\mu_{\boldsymbol{r}}^{\star}(s; \boldsymbol{p}) &= \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[S_{\tau} = s\right] \quad a.s. \\
&= \sum_{(j_1,\dots,j_M) \in \Lambda(s|M;\mathcal{N})} C(\boldsymbol{p})^{-1} p(j_1) p(j_2) \cdots p(j_M) \\
&= C(\boldsymbol{p})^{-1} M! \cdot p(i_1) p(i_2) \cdots p(i_M)
\end{aligned}
\tag{7.14}
$$

with normalizing constant $C(\boldsymbol{p})$ given by (7.4). The last equality at (7.14) follows from the fact that $|\Lambda(s|M;\mathcal{N})| = M!$.

Using (7.14) in conjunction with Theorem 5.1, we readily conclude that under the RORA($\boldsymbol{r}$) policy of Case 1 the miss rate (4.8) for the IRM exists as a constant which is independent of the initial cache state s_0. To acknowledge this fact, we simply denote this limiting constant by $\hat{M}_{\boldsymbol{r}}(\boldsymbol{p})$. Specializing (5.3) leads to

$$
\begin{aligned}
\hat{M}_{\boldsymbol{r}}(\boldsymbol{p}) &= C(\boldsymbol{p})^{-1} M! \sum_{\{i_1,\dots,i_M\} \in \Lambda^{\star}(M;\mathcal{N})} p(i_1) \cdots p(i_M) \sum_{i \notin \{i_1,\dots,i_M\}} p(i) \\
&= C(\boldsymbol{p})^{-1} (M+1)! \sum_{\{i_1,\dots,i_{M+1}\} \in \Lambda^{\star}(M+1;\mathcal{N})} p(i_1) \cdots p(i_{M+1}) \\
&= C(\boldsymbol{p})^{-1} (M+1)! \cdot E_{M+1,N}(\boldsymbol{p})
\end{aligned}
\tag{7.15}
$$

while the normalizing constant $C(\boldsymbol{p})$ given by (7.4) can be simplified as

$$
\begin{aligned}
C(\boldsymbol{p}) &= \sum_{(i_1,\dots,i_M) \in \Lambda(M;\mathcal{N})} p(i_1) \cdots p(i_M) \\
&= M! \sum_{\{i_1,\dots,i_M\} \in \Lambda^{\star}(M;\mathcal{N})} p(i_1) \cdots p(i_M) \\
&= M! \cdot E_{M,N}(\boldsymbol{p}).
\end{aligned}
\tag{7.16}
$$

Combining (7.15) and (7.16), we finally get

$$
\hat{M}_{\boldsymbol{r}}(\boldsymbol{p}) = (M+1) \cdot \frac{E_{M+1,N}(\boldsymbol{p})}{E_{M,N}(\boldsymbol{p})} = (M+1) \Phi_{M+1,N}(\boldsymbol{p})
\tag{7.17}
$$

and a straightforward application of Proposition 2.6 yields

Theorem 7.3 *Under any RORA($\boldsymbol{r}$) policy in Case 1, for admissible pmfs $\boldsymbol{p}$ and $\boldsymbol{q}$ on $\mathcal{N}$, it holds that*

$$\hat{M}_{\boldsymbol{r}}(\boldsymbol{q}) \leq \hat{M}_{\boldsymbol{r}}(\boldsymbol{p}) \tag{7.18}$$

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

7.2.2 Case 2

Consider now the RORA($\boldsymbol{r}$) policy under Case 2 when the set $\Sigma_{\boldsymbol{r}}$ is *not* empty, say with $|\Sigma_{\boldsymbol{r}}| = m$ for some $m = 1, \ldots, M - 1$, and let the cache be initially in state s_0 in $\Lambda(M; \mathcal{N})$. By Lemma 7.2, for each $s = \{i_1, \ldots, i_M\}$ in $\Lambda^\star(M; \mathcal{N})$ the limit (5.1) exists and is given by

$$
\begin{aligned}
\mu^\star_{\boldsymbol{r}, s_0}(s; \boldsymbol{p}) &= \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[S_\tau = s\right] \quad a.s. \\
&= \sum_{s' = (j_1, \ldots, j_M) \in \Lambda(s | \boldsymbol{r}, s_0)} \mu_{\boldsymbol{r}, s_0}(s'; \boldsymbol{p})
\end{aligned} \tag{7.19}
$$

where $\Lambda(s | \boldsymbol{r}, s_0)$ denotes the subset of $\Lambda(\boldsymbol{r}, s_0)$ defined by

$$\Lambda(s | \boldsymbol{r}, s_0) := \{(j_1, \ldots, j_M) \in \Lambda(\boldsymbol{r}, s_0) : \{j_1, \ldots, j_M\} = \{i_1, \ldots, i_M\}\} . \tag{7.20}$$

The set $\Lambda(s | \boldsymbol{r}, s_0)$ is *non*-empty if and only if

$$\Sigma_{\boldsymbol{r}}(s_0) \subseteq \{i_1, \ldots, i_M\} \tag{7.21}$$

and $\mu^\star_{\boldsymbol{r}, s_0}(s; \boldsymbol{p}) = 0$ whenever this inclusion (7.21) does not hold. With this in mind, we define

$$\Lambda^\star(\boldsymbol{r}, s_0) := \{s = \{i_1, \ldots, i_M\} \in \Lambda^\star(M; \mathcal{N}) : \text{Eqn. (7.21) holds at } s\}. \tag{7.22}$$

Going back to (7.11) and (7.12), we now conclude that for each $s = \{i_1, \ldots, i_M\}$ in $\Lambda^\star(\boldsymbol{r}, s_0)$, it holds

$$\mu^\star_{\boldsymbol{r}, s_0}(s; \boldsymbol{p}) = \sum_{(j_1, \ldots, j_M) \in \Lambda(s | \boldsymbol{r}, s_0)} C'_{\boldsymbol{r}}(\boldsymbol{p}, s_0)^{-1} \prod_{j_\ell \notin \Sigma_{\boldsymbol{r}}(s_0)} p(j_\ell)$$

$$= \quad C'_{\boldsymbol{r}}(\boldsymbol{p}, s_0)^{-1}(M - m)! \cdot \prod_{i_\ell \notin \Sigma_{\boldsymbol{r}}(s_0)} p(i_\ell) \qquad (7.23)$$

where in the last equality we combine the fact $\{j_1, \ldots, j_M\} = \{i_1, \ldots, i_M\}$ with (7.21),

and then made use of the identity $|\Lambda(s|\boldsymbol{r}, s_0)| = (M - m)!$.

Now, using (7.23) in conjunction with Theorem 5.1 we see that under the RORA($\boldsymbol{r}$)

policy of Case 2 the miss rate (4.8) for the IRM exists as a constant which *depends* on

the initial cache state s_0. We record this fact in the notation by denoting this limiting

constant by $\hat{M}_{\boldsymbol{r}}(\boldsymbol{p}; s_0)$. As in Case 1, specializing (5.3) leads to

$$
\begin{aligned}
\hat{M}_{\boldsymbol{r}}(\boldsymbol{p}; s_0) &= C'_{\boldsymbol{r}}(\boldsymbol{p}, s_0)^{-1}(M - m)! \sum_{\{i_1, \ldots, i_M\} \in \Lambda^\star(\boldsymbol{r}, s_0)} \prod_{i_\ell \notin \Sigma_{\boldsymbol{r}}(s_0)} p(i_\ell) \sum_{i \notin \{i_1, \ldots, i_M\}} p(i) \\
&= C'_{\boldsymbol{r}}(\boldsymbol{p}, s_0)^{-1}(M - m + 1)! \cdot E_{M-m+1, N}(\boldsymbol{t} \cdot \boldsymbol{p}) \qquad (7.24)
\end{aligned}
$$

where the element $\boldsymbol{t}$ in $\mathbb{R}^N_+$ is specified by $t_i = 0$ for i being a document in $\Sigma_{\boldsymbol{r}}(s_0)$ and

$t_i = 1$ otherwise. Moreover, by the same arguments as in Case 1, we can simplify the

normalizing constant $C'_{\boldsymbol{r}}(\boldsymbol{p}, s_0)$ as

$$
\begin{aligned}
C'_{\boldsymbol{r}}(\boldsymbol{p}, s_0) &= \sum_{(i_1, \ldots, i_M) \in \Lambda(\boldsymbol{r}, s_0)} \prod_{i_\ell \notin \Sigma_{\boldsymbol{r}}(s_0)} p(i_\ell) \\
&= (M - m)! \sum_{\{i_1, \ldots, i_M\} \in \Lambda^\star(\boldsymbol{r}, s_0)} \prod_{i_\ell \notin \Sigma_{\boldsymbol{r}}(s_0)} p(i_\ell) \\
&= (M - m)! \cdot E_{M-m, N}(\boldsymbol{t} \cdot \boldsymbol{p}) \qquad (7.25)
\end{aligned}
$$

with the element $\boldsymbol{t}$ given as above. It then follows from (7.24) and (7.25) that

$$
\begin{aligned}
\hat{M}_{\boldsymbol{r}}(\boldsymbol{p}; s_0) &= (M - m + 1) \cdot \frac{E_{M-m+1, N}(\boldsymbol{t} \cdot \boldsymbol{p})}{E_{M-m, N}(\boldsymbol{t} \cdot \boldsymbol{p})} \\
&= (M - m + 1)\Phi_{M-m+1, N}(\boldsymbol{t} \cdot \boldsymbol{p}). \qquad (7.26)
\end{aligned}
$$

Clearly, the documents in $\Sigma_{\boldsymbol{r}}(s_0)$ do not contribute to the miss rate since they never

generate a miss once loaded in cache – This is *regardless* of the order in which they

appear in the cache state s_0. This intuitively obvious fact is in agreement with the

expression (7.26) from which we see that for any two initial cache states s_0 and s'_0 in

$\Lambda(M; \mathcal{N})$ with $\Sigma_{\boldsymbol{r}}(s_0) = \Sigma_{\boldsymbol{r}}(s_0')$, we have the equality $\hat{M}_{\boldsymbol{r}}(\boldsymbol{p}; s_0) = \hat{M}_{\boldsymbol{r}}(\boldsymbol{p}; s_0')$. As a result, we shall find it appropriate to denote this common value by $\hat{M}_{\boldsymbol{r}, \Sigma_{\boldsymbol{r}}(s_0)}(\boldsymbol{p})$.

For any pmf $\boldsymbol{p}$ on $\mathcal{N}$, let $\Sigma^{\star}(\boldsymbol{p})$ denote the set of the m most popular documents according to the pmf $\boldsymbol{p}$. Equipped with the expression (7.26), we are now ready to establish the key result for RORA policies in Case 2.

Theorem 7.4 *Under any RORA(r) policy in Case 2 with $|\Sigma_{\boldsymbol{r}}| = m$ for some $m = 1, \ldots, M - 1$, for admissible pmfs $\boldsymbol{p}$ and $\boldsymbol{q}$ on $\mathcal{N}$, it holds that*

$$\hat{M}_{\boldsymbol{r}, \Sigma^{\star}(\boldsymbol{q})}(\boldsymbol{q}) \leq \hat{M}_{\boldsymbol{r}, \Sigma^{\star}(\boldsymbol{p})}(\boldsymbol{p}) \tag{7.27}$$

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

Proof. The desired result will be established if we can show that the miss rate function $\boldsymbol{p} \to \hat{M}_{\boldsymbol{r}, \Sigma_{\boldsymbol{r}}(s_0)}(\boldsymbol{p})$ as given in (7.26) is Schur-concave whenever s_0 is selected so that $\Sigma_{\boldsymbol{r}}(s_0) = \Sigma^{\star}(\boldsymbol{p})$.

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 (7.26) can be specified as $t_1 = \ldots = t_m = 0$ and $t_{m+1} = \ldots = t_N = 1$. By Proposition 2.6, the mapping $\Phi_{M-m+1,N}$ is increasing and Schur-concave on $\mathbb{R}_{+}^{N}$, and by virtue of the defining property of $\Sigma^{\star}(\boldsymbol{p})$, we have

$$\hat{M}_{\boldsymbol{r}, \Sigma^{\star}(\boldsymbol{p})}(\boldsymbol{p}) = \min_{i=1,\ldots,N!} (M - m + 1)\Phi_{M-m+1,N}(\boldsymbol{t} \cdot \sigma_i(\boldsymbol{p})) \tag{7.28}$$

where $\{\sigma_i, i = 1, \ldots, N!\}$ is a collection of all permutations of $\{1, \ldots, N\}$.

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 $\Phi_{M-m+1,N}$ is concave on $\mathbb{R}_{+}^{N}$ by Proposition 2.6. Combining these facts with the expression (7.28) for $\hat{M}_{\boldsymbol{r}, \Sigma^{\star}(\boldsymbol{p})}(\boldsymbol{p})$, we conclude by

Proposition 2.8 to the Schur-concavity (in the pmf vector) of the miss rate functional (7.26) under the RORA policy when $\Sigma_{\boldsymbol{r}}(s_0) = \Sigma^{\star}(\boldsymbol{p})$. ∎

7.3 The output under RORAs

We now discuss the popularity pmf of the output generated under the RORA policies still under the assumed IRM input stream.

7.3.1 Case 1

As we invoke Theorem 5.2, we can make use of the expressions (7.14) into the relation (5.5). For each $i = 1, \ldots, N$, this yields

$$
\begin{aligned}
m_{\boldsymbol{r}}(i; \boldsymbol{p}) &= \sum_{s \in \Lambda_i^{\star}(M; \mathcal{N})} C(\boldsymbol{p})^{-1} M! \cdot p(i_1) p(i_2) \cdots p(i_M) \\
&= \frac{E_{M,N-1}(\boldsymbol{p}^{(i)})}{E_{M,N}(\boldsymbol{p})}
\end{aligned}
\tag{7.29}
$$

where the last equality follows from (7.16) and by recalling the definition of $\boldsymbol{p}^{(i)}$ given at (6.7). Reporting (7.29) back into (5.4), we conclude that the popularity pmf $\boldsymbol{p}_{\boldsymbol{r}}^{\star}$ of the output produced by the RORA($\boldsymbol{r}$) policy in Case 1 is indeed of the form (6.8), and Theorem 6.4 gives us

Theorem 7.5 *Under any RORA(r) policy in Case 1, it holds that $\boldsymbol{p}_{\boldsymbol{r}}^{\star} \prec \boldsymbol{p}$.*

By going back to the proof of Theorem 6.4, the reader will readily check from (7.29) that the RORA($\boldsymbol{r}$) policy in Case 1 is indeed a good policy.

7.3.2 Case 2

Assume $|\Sigma_{\boldsymbol{r}}| = m$ for some $m = 1, \ldots, M - 1$, and let the cache be initially in state s_0 in $\Lambda(M; \mathcal{N})$. We define the pmf $\boldsymbol{\pi}$ on $\Sigma_{\boldsymbol{r}}(s_0)^c$ to be the *conditional* pmf induced on $\Sigma_{\boldsymbol{r}}(s_0)^c$ by $\boldsymbol{p}$; it is defined as

$$\pi(i) = \frac{p(i)}{\sum_{j \in \Sigma_{\boldsymbol{r}}(s_0)^c} p(j)}, \quad i \in \Sigma_{\boldsymbol{r}}(s_0)^c. \tag{7.30}$$

For all i in $\Sigma_{\boldsymbol{r}}(s_0)$, it is clear that $m_{\boldsymbol{r},s_0}(i; \boldsymbol{p}) = 0$ while for document i *not* in $\Sigma_{\boldsymbol{r}}(s_0)^c$, with the expression for $\mu^{\star}_{\boldsymbol{r},s_0}(s; \boldsymbol{p})$ given in (7.23), we find

$$
\begin{aligned}
m_{\boldsymbol{r},s_0}(i; \boldsymbol{p}) &= \sum_{s \in \Lambda^{\star}(\boldsymbol{r},s_0):\, i \notin s} C'_{\boldsymbol{r}}(\boldsymbol{p}, s_0)^{-1} (M - m)! \cdot \prod_{i_\ell \notin \Sigma_{\boldsymbol{r}}(s_0)} p(i_\ell) \\
&= \frac{E_{M-m,N}(\boldsymbol{t}^{(1)} \cdot \boldsymbol{p})}{E_{M-m,N}(\boldsymbol{t}^{(2)} \cdot \boldsymbol{p})} \\
&= \frac{E_{M-m,N-m-1}(\boldsymbol{\pi}^{(i)})}{E_{M-m,N-m}(\boldsymbol{\pi})}
\end{aligned}
\tag{7.31}
$$

where the element $\boldsymbol{t}^{(1)}$ and $\boldsymbol{t}^{(2)}$ of $\mathbb{R}^N_+$ are specified by $t^{(1)}_j = t^{(2)}_j = 0$ for j being a document in $\Sigma_{\boldsymbol{r}}(s_0)$, $t^{(1)}_i = 0$, $t^{(2)}_i = 1$ and $t^{(1)}_j = t^{(2)}_j = 1$ for all $j \neq i$ being a document in $\Sigma_{\boldsymbol{r}}(s_0)^c$. In the second equality we made use of the expression (7.25).

On revisiting the proof of Theorem 6.4, we note that for distinct i, j in $\Sigma_{\boldsymbol{r}}(s_0)^c$, we have $m_{\boldsymbol{r},s_0}(i; \boldsymbol{p}) \leq m_{\boldsymbol{r},s_0}(j; \boldsymbol{p})$ whenever $p(j) \leq p(i)$. Consequently, since $m_{\boldsymbol{r},s_0}(i; \boldsymbol{p}) = 0$ for all i in $\Sigma_{\boldsymbol{r}}(s_0)$, we conclude that the RORA policy in Case 2 is a good policy if the documents in $\Sigma_{\boldsymbol{r}}(s_0)$ are the m most popular documents, i.e., $\Sigma_{\boldsymbol{r}}(s_0) = \Sigma^{\star}(\boldsymbol{p})$.

Combining (7.31) with (5.4), we immediately get

$$
p^{\star}_{\boldsymbol{r},s_0}(i) =
\begin{cases}
0 & \text{if } i \in \Sigma(s_0) \\
\frac{\pi(i) E_{M-m,N-m-1}(\boldsymbol{\pi}^{(i)})}{\sum_{j \in \Sigma(s_0)^c} \pi(j) E_{M-m,N-m-1}(\boldsymbol{\pi}^{(j)})} & \text{if } i \notin \Sigma(s_0).
\end{cases}
\tag{7.32}
$$

Since $p^{\star}_{\boldsymbol{r},s_0}(i) = 0$ whenever i belongs to $\Sigma_{\boldsymbol{r}}(s_0)$, it is more natural to seek a comparison between $\boldsymbol{p}^{\star}_{\boldsymbol{r},s_0}$ and the conditional pmf $\boldsymbol{\pi}$.

Theorem 7.6 *Under any RORA(r) policy in Case 2, it holds that $p_{r,s_0}^{\star} \prec \pi$.*

Proof. The arguments are essentially those given in the proof of Theorem 6.4. We immediately obtain the desired result upon identifying π and $\Sigma_r(s_0)^c$ with p and $\mathcal{N}$ in Theorem 6.4, respectively. ∎

Chapter 8

Self-organizing Policies

In this chapter, we investigate the folk theorems under the IRM for the miss rate and the output of a cache operated by well-known self-organizing policies, namely, the LRU and CLIMB policies. The LRU and CLIMB policies are described in Section 4.3. From the positive results achieved under the RORA policies, one might expect that the folk theorems would hold under these two self-organizing policies. However, both folk theorems for the miss rate and the output under the LRU and CLIMB policies fail to hold in general. Nonetheless, as we restrict ourself to the class of IRM inputs with Zipf-like popularity pmf (6.4)-(6.5), simulation results and asymptotics suggest that the folk theorems might hold under the IRM with this class of popularity pmfs.

We now discuss the results for the LRU and CLIMB policies, respectively.

8.1   The miss rate under the LRU policy

Under the IRM with admissible popularity pmf $\boldsymbol{p}$, it is known [2, Thm. 9, p. 130] [24, Thm. 6.5, p. 272] that the LRU cache states $\{\Omega_t, t = 0, 1, \ldots\}$ form a stationary ergodic Markov chain over the finite state space $\Lambda(M; \mathcal{N})$ with stationary distribution given by

$$\mu_{\mathrm{LRU}}(s; \boldsymbol{p}) \;=\; \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[\Omega_\tau = s\right] \quad a.s.$$

$$= \frac{p(i_1) \cdots p(i_M)}{\prod_{k=1}^{M-1} (1 - \sum_{j=1}^{k} p(i_j))} \tag{8.1}$$

for every $s = (i_1, \ldots, i_M)$ in $\Lambda(M; \mathcal{N})$. Consequently, the limit (5.1) exists for each $s = \{i_1, \ldots, i_M\}$ in $\Lambda^\star(M; \mathcal{N})$ as

$$\mu_{\mathrm{LRU}}^\star(s; \boldsymbol{p}) = \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[S_\tau = s\right] \quad a.s.$$

$$= \sum_{(j_1, \ldots, j_M) \in \Lambda(s|M; \mathcal{N})} \frac{p(j_1) \cdots p(j_M)}{\prod_{k=1}^{M-1} (1 - \sum_{\ell=1}^{k} p(j_\ell))} \tag{8.2}$$

where $\Lambda(s|M; \mathcal{N})$ is defined at (7.13).

The miss rate of the LRU policy under IRM can then be evaluated from (5.3) (see also [2, Chap. 4]) as

$$\hat{M}_{\mathrm{LRU}}(\boldsymbol{p}) = \sum_{(i_1, \ldots, i_M) \in \Lambda(M; \mathcal{N})} \frac{p(i_1) \cdots p(i_M) \left(1 - \sum_{j=1}^{M} p(i_j)\right)}{\prod_{k=1}^{M-1} (1 - \sum_{j=1}^{k} p(i_j))}. \tag{8.3}$$

If instead we use (5.2), as we note that

$$\sum_{s \in \Lambda_i^\star(M; \mathcal{N})} \left(\sum_{(j_1, \ldots, j_M) \in \Lambda(s|M; \mathcal{N})} \cdots \right) = \sum_{s \in \Lambda_i(M; \mathcal{N})} \cdots,$$

it is now plain that

$$\hat{M}_{\mathrm{LRU}}(\boldsymbol{p}) = \sum_{i=1}^{N} p(i) \sum_{s \in \Lambda_i(M; \mathcal{N})} \frac{p(i_1) \cdots p(i_M)}{\prod_{k=1}^{M-1} (1 - \sum_{\ell=1}^{k} p(i_\ell))}. \tag{8.4}$$

8.1.1 A counterexample

Contrary to what transpired with RORA policies, the miss rate under the LRU policy is *not* Schur-concave in general, and consequently the folk theorem (6.1) does not hold. This is demonstrated through the following example developed for $M = 3$ and $N = 4$:

In this case, simple algebraic manipulations transform (8.3) into the simpler expression

$$\hat{M}_{\mathrm{LRU}}(\boldsymbol{p}) = \sum_{(i_1, i_2) \in \Lambda(2; \mathcal{N})} \frac{2p(1)p(2)p(3)p(4)}{\prod_{k=1}^{2} (1 - \sum_{j=1}^{k} p(i_j))}. \tag{8.5}$$
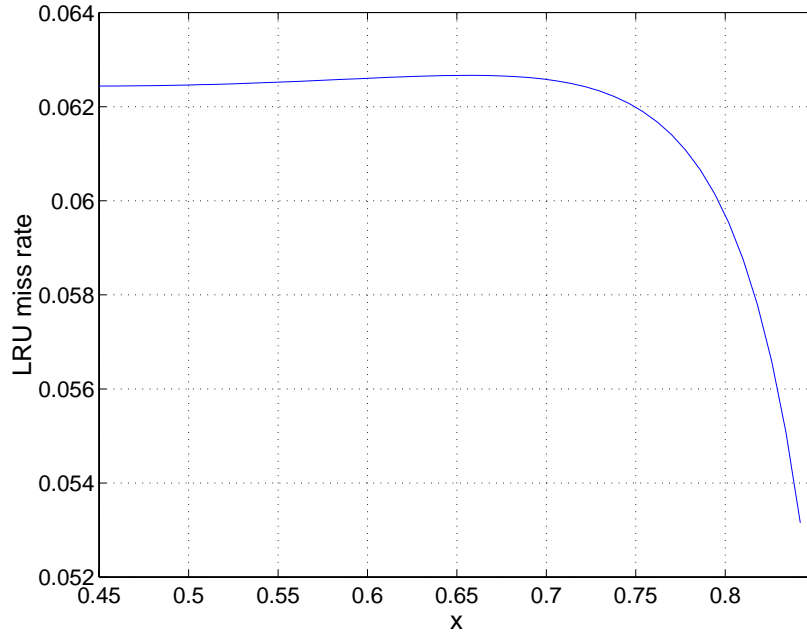
Figure 8.1: LRU miss rate when $M = 3$, $N = 4$, $y = p(3) = p(4) = 0.05$, $p(1) = x$ and $p(2) = 0.9 - p(1)$

We evaluated the expressions (8.5) for the family of pmfs

$$\boldsymbol{p}(x, y) = (x, 1 - 2y - x, y, y), \quad 0 < y < \frac{1}{4} \tag{8.6}$$

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 rate under the LRU policy were indeed a Schur-concave function in the popularity pmf, the functions $x \rightarrow \hat{M}_{\mathrm{LRU}}(\boldsymbol{p}(x, y))$ should be monotone decreasing in x on the interval $[\frac{1}{2} - y, 1 - 3y]$.

Figures 8.1 and 8.2 display the numerical values of $\hat{M}_{\mathrm{LRU}}(\boldsymbol{p}(x, y))$ as a function of x with $y = 0.05$ and $y = 0.01$, respectively. In both cases, the miss rate of the LRU policy is *not* monotone decreasing in x on the range $[\frac{1}{2} - y, 1 - 3y]$, with the trend becoming more pronounced with decreasing y. In short, the miss rate is not Schur-concave under the LRU policy.

69
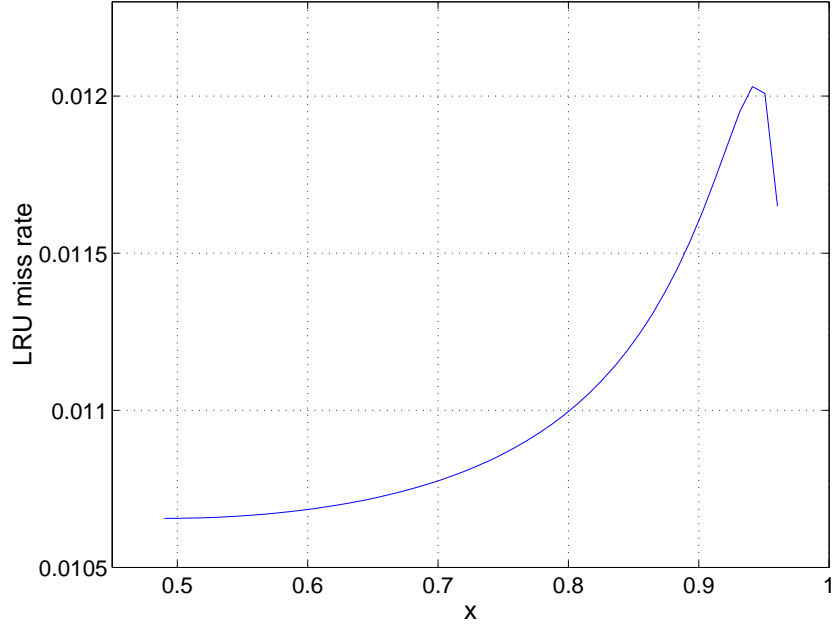
Figure 8.2: LRU miss rate when $M = 3$, $N = 4$, $y = p(3) = p(4) = 0.01$, $p(1) = x$ and $p(2) = 0.98 - p(1)$

8.1.2 LRU miss rate and IRM with Zipf-like popularity pmfs

While the miss rate is *not* Schur-concave under the LRU policy, the desired monotonicity (6.1) is nevertheless true in an asymptotic sense when the popularity pmf is restricted to the class of Zipf-like pmfs.

Theorem 8.1 *Assume the IRM input to have a Zipf-like popularity pmf $\boldsymbol{p}_\alpha$ for some $\alpha \geq 0$. Then, there exists $\alpha^\star = \alpha^\star(M, N) > 0$ and $\Delta > 0$ such that $\hat{M}_{\mathrm{LRU}}(\boldsymbol{p}_\beta) < \hat{M}_{\mathrm{LRU}}(\boldsymbol{p}_\alpha)$ whenever $\alpha^\star < \alpha$ and $\alpha + \Delta < \beta$.*

This result is a byproduct of the asymptotic equivalence

$$\lim_{\alpha \to \infty} \frac{\hat{M}_{\mathrm{LRU}}(\boldsymbol{p}_\alpha)}{(M+1)^{-\alpha}} = 2 \tag{8.7}$$

70

established in Appendix B.1. Indeed, for every ε in the interval $(0, 1)$, there exists $\alpha^\star(M, N) > 0$ such that for $\alpha > \alpha^\star$,

$$1 - \varepsilon \leq \frac{\hat{M}_{\text{LRU}}(\boldsymbol{p}_\alpha)}{2(M+1)^{-\alpha}} \leq 1 + \varepsilon. \tag{8.8}$$

Thus, for $\alpha^\star < \alpha < \beta$, we conclude that

$$\frac{1 - \varepsilon}{1 + \varepsilon} \cdot (M+1)^{\beta - \alpha} \leq \frac{\hat{M}_{\text{LRU}}(\boldsymbol{p}_\alpha)}{\hat{M}_{\text{LRU}}(\boldsymbol{p}_\beta)} \leq \frac{1 + \varepsilon}{1 - \varepsilon} \cdot (M+1)^{\beta - \alpha} \tag{8.9}$$

and the desired result follows whenever $\beta - \alpha > \Delta$ with $\Delta > 0$ selected such that

$$\frac{1 + \varepsilon}{1 - \varepsilon} = (M+1)^\Delta.$$

Of course such a selection is always possible.

We have also carried out simulations of a cache operating under the LRU policy when the IRM input has a Zipf-like popularity pmf $\boldsymbol{p}_\alpha$.[1] The number of documents is set at $N = 1,000$ while the cache size is $M = 100$. The miss rate of the LRU policy is displayed in Figure 8.3 and 8.4 for small α ($0 \leq \alpha \leq 1$) and large α ($\alpha > 1$), 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 for miss rates probably holds under the LRU policy when the comparison is made within the class of Zipf-like popularity pmfs, hence the following

Conjecture 8.2 *For arbitrary cache size M and number of documents N, the function $\alpha \to \hat{M}_{\text{LRU}}(\boldsymbol{p}_\alpha)$ is strictly decreasing on $[0, \infty)$.*

---

[1]We choose simulations over numerical evaluation of (8.3) because this expression is not suitable for numerical evaluation due to a combinatorial explosion, as pointed out in [33].
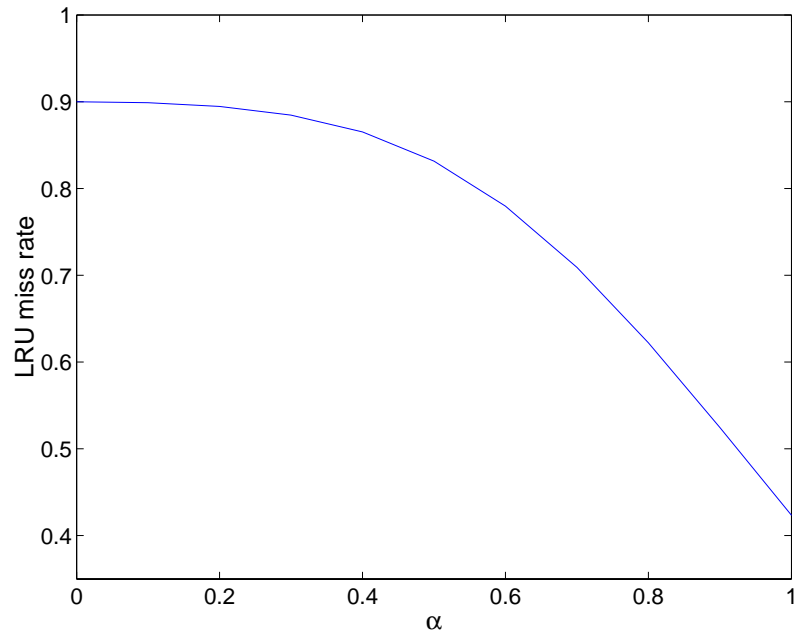
Figure 8.3: LRU miss rate when the IRM input has a Zipf-like popularity pmf p_α for α small ($0 \leq \alpha \leq 1$)
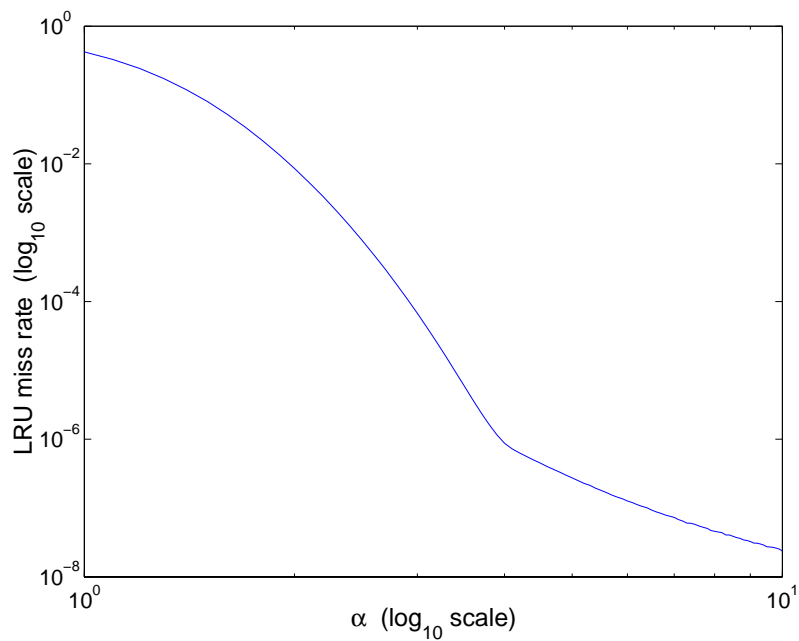
Figure 8.4: LRU miss rate when the IRM input has a Zipf-like popularity pmf p_α for α large ($\alpha > 1$)

8.2 The output under the LRU policy

With the expressions (8.1) for the LRU cache stationary distribution under the IRM, it is a simple matter to check for each $i = 1, \ldots, N$, that

$$
\begin{aligned}
m_{\mathrm{LRU}}(i; \boldsymbol{p}) &= \sum_{s \in \Lambda_i(M; \mathcal{N})} \mu_{\mathrm{LRU}}(s; \boldsymbol{p}) \\
&= \sum_{s \in \Lambda_i(M; \mathcal{N})} \frac{p(i_1) \cdots p(i_M)}{\prod_{k=1}^{M-1} (1 - \sum_{j=1}^{k} p(i_j))}.
\end{aligned}
\tag{8.10}
$$

Theorem 5.2 then gives the output popularity pmf in the form

$$
p_{\mathrm{LRU}}^{\star}(i) = \frac{p(i)}{\hat{M}_{\mathrm{LRU}}(\boldsymbol{p})} \sum_{s \in \Lambda_i(M; \mathcal{N})} \frac{p(i_1) \cdots p(i_M)}{\prod_{k=1}^{M-1} (1 - \sum_{j=1}^{k} p(i_j))}
\tag{8.11}
$$

for each $i = 1, \ldots, N$, as we make use of (5.8).

8.2.1 LRU is a good policy

We begin with a positive result.

Lemma 8.3 *The LRU policy is a good policy.*

Proof. Pick distinct $i, j = 1, \ldots, N$ with $p(j) \leq p(i)$. We need to show that

$$
m_{\mathrm{LRU}}(i; \boldsymbol{p}) \leq m_{\mathrm{LRU}}(j; \boldsymbol{p}).
\tag{8.12}
$$

We begin by writing $m_{\mathrm{LRU}}(i; \boldsymbol{p})$ as

$$
m_{\mathrm{LRU}}(i; \boldsymbol{p}) = \sum_{s \in \Lambda_i(M; \mathcal{N}): \, j \in s} \mu_{\mathrm{LRU}}(s; \boldsymbol{p}) + \sum_{s \in \Lambda_i(M; \mathcal{N}): \, j \notin s} \mu_{\mathrm{LRU}}(s; \boldsymbol{p})
\tag{8.13}
$$

with a similar expression for $m_{\mathrm{LRU}}(j; \boldsymbol{p})$. The fact that the sets $\{s \in \Lambda_i(M; \mathcal{N}) : \, j \notin s\}$ and $\{s \in \Lambda_j(M; \mathcal{N}) : \, i \notin s\}$ coincide leads to

$$
\begin{aligned}
m_{\mathrm{LRU}}(i; \boldsymbol{p}) - m_{\mathrm{LRU}}(j; \boldsymbol{p}) &= \sum_{s \in \Lambda_i(M; \mathcal{N}): \, j \in s} \mu_{\mathrm{LRU}}(s; \boldsymbol{p}) \\
&\quad - \sum_{s \in \Lambda_j(M; \mathcal{N}): \, i \in s} \mu_{\mathrm{LRU}}(s; \boldsymbol{p}).
\end{aligned}
\tag{8.14}
$$

The sets $\{s \in \Lambda_i(M;\mathcal{N}) : j \in s\}$ and $\{s \in \Lambda_j(M;\mathcal{N}) : i \in s\}$ can be put into *one-to-one* correspondence with each other as follows: Each element s in the former set does not contain i but contains j in exactly one position, say position k for some $k = 1, \ldots, M$, with all other positions occupied by neither i nor j. Thus, with such an element s we can associate an element $T(s)$ in $\Lambda_j(M;\mathcal{N})$ by substituting i for j at position k and letting all other positions unchanged. This element $T(s)$ now contains i but not j anymore, and is therefore an element of the latter set. Moreover, for such an element $T(s)$ it holds that

$$\mu_{\mathrm{LRU}}(s; \boldsymbol{p}) \leq \mu_{\mathrm{LRU}}(T(s); \boldsymbol{p}) \tag{8.15}$$

as a consequence of the assumption $p(j) \leq p(i)$ and of the expression (8.1). With these observations in mind, we find that

$$
\begin{aligned}
\sum_{s \in \Lambda_j(M;\mathcal{N}):\ i \in s} \mu_{\mathrm{LRU}}(s; \boldsymbol{p}) \ &= \sum_{s \in \Lambda_i(M;\mathcal{N}):\ j \in s} \mu_{\mathrm{LRU}}(T(s); \boldsymbol{p}) \\
&\geq \sum_{s \in \Lambda_i(M;\mathcal{N}):\ j \in s} \mu_{\mathrm{LRU}}(s; \boldsymbol{p})
\end{aligned}
$$

and the conclusion (8.12) is now immediate via (8.14). ∎

8.2.2 Counterexamples

In view of Lemma 8.3, it is tempting to expect that the majorization comparison $\boldsymbol{p}_{\mathrm{LRU}}^{\star} \prec \boldsymbol{p}$ also holds under the LRU policy. This is not true in general as the following counterexamples show: Fix $N = 2, 3, \ldots$. Assume that the input to the cache is the IRM with popularity pmf $\boldsymbol{p}_\varepsilon$ where we set

$$\boldsymbol{p}_\varepsilon = (1 - (N-1)\varepsilon, \varepsilon, \ldots, \varepsilon) \tag{8.16}$$

for some $0 < \varepsilon \leq \frac{1}{N}$. Note that $p_\varepsilon(1) \geq p_\varepsilon(2) = \cdots = p_\varepsilon(N)$, and as $\varepsilon \to \frac{1}{N}$, the pmf p_ε approaches the uniform distribution $\boldsymbol{u}$ while as $\varepsilon \to 0$, it degenerates to $(1, 0, \ldots, 0)$. Indeed, from Lemma 2.5, we find that $\boldsymbol{p}_{\varepsilon_1} \prec \boldsymbol{p}_{\varepsilon_2}$ whenever $\varepsilon_2 \leq \varepsilon_1$.

Under the LRU policy, it is plain from (8.10)-(8.11) that the output popularity pmf $\boldsymbol{p}^\star_{\mathrm{LRU},\varepsilon}$ is of the form

$$\boldsymbol{p}^\star_\varepsilon = (1 - (N-1)\delta(\varepsilon), \delta(\varepsilon), \ldots, \delta(\varepsilon)). \tag{8.17}$$

for some mapping $\delta : (0, \frac{1}{N}] \to (0, \frac{1}{N-1})$. Because of their special structures, (8.16) and (8.17), the comparison between $\boldsymbol{p}_\varepsilon$ and $\boldsymbol{p}^\star_{\mathrm{LRU},\varepsilon}$ depends only on the value of $\delta(\varepsilon)$; this fact is stated in

Proposition 8.4 *For each $0 < \varepsilon \leq \frac{1}{N}$, let $\boldsymbol{p}_\varepsilon$ and $\boldsymbol{p}^\star_\varepsilon$ be the pmfs of the form (8.16) and (8.17), respectively.*

   (i) If $0 < \delta(\varepsilon) \leq \varepsilon$, then the comparison $\boldsymbol{p}_\varepsilon \prec \boldsymbol{p}^\star_\varepsilon$ holds;

   (ii) If $\varepsilon \leq \delta(\varepsilon) \leq \frac{1-\varepsilon}{N-1}$, then the comparison $\boldsymbol{p}^\star_\varepsilon \prec \boldsymbol{p}_\varepsilon$ holds;

   (iii) If $\frac{1-\varepsilon}{N-1} < \delta(\varepsilon) < \min(1 - (N-1)\varepsilon, \frac{1}{N-1})$, then neither the comparison $\boldsymbol{p}^\star_\varepsilon \prec \boldsymbol{p}_\varepsilon$ nor the comparison $\boldsymbol{p}_\varepsilon \prec \boldsymbol{p}^\star_\varepsilon$ holds; and

   (iv) If $\min(1 - (N-1)\varepsilon, \frac{1}{N-1}) \leq \delta(\varepsilon) < \frac{1}{N-1}$, then the comparison $\boldsymbol{p}_\varepsilon \prec \boldsymbol{p}^\star_\varepsilon$ holds.

Proof. Fix $0 < \varepsilon \leq \frac{1}{N}$. The discussion is separated into 2 cases, namely (a) $0 < \delta(\varepsilon) \leq \frac{1}{N}$ and (b) $\frac{1}{N} < \delta(\varepsilon) < \frac{1}{N-1}$.

Case (a) – With $0 < \delta(\varepsilon) \leq \frac{1}{N}$, we note that $p^\star_\varepsilon(1) \geq p^\star_\varepsilon(2) = \cdots = p^\star_\varepsilon(N)$. By Lemma 2.5, the comparison $\boldsymbol{p}^\star_\varepsilon \prec \boldsymbol{p}_\varepsilon$ (resp. $\boldsymbol{p}_\varepsilon \prec \boldsymbol{p}^\star_\varepsilon$) holds whenever

$$\delta(\varepsilon) \geq (\leq) \ \varepsilon, \tag{8.18}$$

and Claim (i) is obtained.

Case (b) – When $\frac{1}{N} < \delta(\varepsilon) < \frac{1}{N-1}$, we have $p_\varepsilon^\star(1) < p_\varepsilon^\star(2) = \cdots = p_\varepsilon^\star(N)$. In this case, the conditions (2.1) for the majorization comparison $\boldsymbol{p}_\varepsilon^\star \prec \boldsymbol{p}_\varepsilon$ (resp. $\boldsymbol{p}_\varepsilon \prec \boldsymbol{p}_\varepsilon^\star$) are simply

$$k\delta(\varepsilon) + (N-k)\varepsilon \ \leq (\geq) \ 1, \quad k = 1, \ldots, N-1. \tag{8.19}$$

Because $\delta(\varepsilon) > \varepsilon$ in this case, the left-hand side of (8.19) is monotone increasing in k. From this observation and (8.19), the comparison $\boldsymbol{p}_\varepsilon^\star \prec \boldsymbol{p}_\varepsilon$ will hold if

$$\delta(\varepsilon) \leq \frac{1-\varepsilon}{N-1}, \tag{8.20}$$

while the comparison $\boldsymbol{p}_\varepsilon \prec \boldsymbol{p}_\varepsilon^\star$ will hold if

$$\delta(\varepsilon) \geq 1 - (N-1)\varepsilon. \tag{8.21}$$

However, neither the comparison $\boldsymbol{p}_\varepsilon \prec \boldsymbol{p}_\varepsilon^\star$ nor the comparison $\boldsymbol{p}_\varepsilon^\star \prec \boldsymbol{p}_\varepsilon$ holds if

$$\frac{1-\varepsilon}{N-1} < \delta(\varepsilon) < 1 - (N-1)\varepsilon. \tag{8.22}$$

Combining (8.18) and (8.20) yields Claim (ii). Upon recalling that $\delta(\varepsilon) < \frac{1}{N-1}$, we obtain Claim (iii) and (iv) from (8.22) and (8.21), respectively. ■

Using Proposition 8.4, we show under the LRU policy that it is possible to find some $0 < \varepsilon < \frac{1}{N}$ such that $\delta(\varepsilon) > \frac{1-\varepsilon}{N-1}$, and thus the desired comparison $\boldsymbol{p}_{\mathrm{LRU},\varepsilon}^\star \prec \boldsymbol{p}_\varepsilon$ does not hold. This result is given in the following theorem: its proof is available in Appendix C.1.

Theorem 8.5 *Assume the IRM input to have the popularity pmf $\boldsymbol{p}_\varepsilon$ for some $0 < \varepsilon \leq \frac{1}{N}$. Under the LRU policy, whenever*

$$0 < \varepsilon < \frac{\left(\sum_{\ell=1}^{M-1} \frac{1}{N-\ell}\right) - 1}{\left(\sum_{\ell=1}^{M-1} \frac{\ell}{N-\ell}\right)}, \tag{8.23}$$

the comparison $\boldsymbol{p}_{\text{LRU},\varepsilon}^{\star} \prec \boldsymbol{p}_{\varepsilon}$ does not hold provided that the number of documents N and the cache size M satisfy the condition $\sum_{\ell=1}^{M-1} \frac{1}{N-\ell} > 1$.

For example, if we take $\boldsymbol{p}_{\varepsilon}$ with parameters $N = 10$ and $\varepsilon = 0.05$ and set the cache size $M = 8$, a simple calculation yields $\delta(\varepsilon) = 0.1111$ and the assumptions of Theorem 8.5 are satisfied. Thus, the comparison $\boldsymbol{p}_{\text{LRU},\varepsilon}^{\star} \prec \boldsymbol{p}_{\varepsilon}$ does not hold. However, the entropy of $\boldsymbol{p}_{\varepsilon}$ is smaller than the entropy of $\boldsymbol{p}_{\text{LRU},\varepsilon}^{\star}$, i.e.,

$$0.7283 = H(\boldsymbol{p}_{\varepsilon}) \leq H(\boldsymbol{p}_{\text{LRU},\varepsilon}^{\star}) = 0.9554.$$

This suggests that $\boldsymbol{p}_{\text{LRU},\varepsilon}^{\star}$ is more balanced than $\boldsymbol{p}_{\varepsilon}$ in the sense of entropy comparison. Hence, even though the comparison in the majorization ordering does not hold, the entropy comparison might still be valid. This should not come as a surprise since the majorization comparison is a stronger notion than the entropy comparison.

As for the case of the LRU miss rate, we would expect that the comparison $\boldsymbol{p}_{\text{LRU}}^{\star} \prec \boldsymbol{p}$ under the LRU policy would hold within the class of IRM inputs with Zipf-like popularity pmf $\boldsymbol{p}_{\alpha}$. However, this is not the case as the following example demonstrates: With $M = 3$ and $N = 4$ under the Zipf-like popularity pmf (6.4)-(6.5) with $\alpha = 3$, we have computed the output popularity pmf under the LRU policy using (8.11). The numerical values of both input and output popularity pmfs are given in Table 8.1.

Table 8.1: $\boldsymbol{p}_{\alpha}$ and $\boldsymbol{p}_{\text{LRU},\alpha}^{\star}$ under the LRU policy when the IRM input has a Zipf-like popularity pmf $\boldsymbol{p}_{\alpha}$ with parameter $\alpha = 3$

i	1	2	3	4
$\boldsymbol{p}_{\alpha}$	0.8491	0.1061	0.0314	0.0133
$\boldsymbol{p}_{\text{LRU},\alpha}^{\star}$	0.0118	0.2031	0.3853	0.3998

By the definition of majorization (2.1)-(2.2), the comparison $\boldsymbol{p}_{\text{LRU},\alpha}^{\star} \prec \boldsymbol{p}_{\alpha}$ requires

$$\min_{i=1,\ldots,N} p_{\alpha}(i) \leq \min_{i=1,\ldots,N} p_{\text{LRU},\alpha}^{\star}(i), \tag{8.24}$$

in clear contradiction with Table 8.1, and therefore does not hold. On the other hand, the comparison $\boldsymbol{p}_\alpha \prec \boldsymbol{p}^\star_{\mathrm{LRU},\alpha}$ is not valid either since it calls for the unmet requirement

$$\max_{i=1,\ldots,N} p_\alpha(i) \leq \max_{i=1,\ldots,N} p^\star_{\mathrm{LRU},\alpha}(i). \tag{8.25}$$

In short, $\boldsymbol{p}_\alpha$ and $\boldsymbol{p}^\star_{\mathrm{LRU},\alpha}$ are not comparable in the majorization ordering. This situation does not represent an isolated incident as the next theorem shows; its proof is available in Appendix B.2.

Theorem 8.6 *Assume the IRM input to have a Zipf-like popularity pmf $\boldsymbol{p}_\alpha$ for some $\alpha \geq 0$. If the number of documents N and the cache size M satisfy the condition*

$$N < M!, \tag{8.26}$$

then under the LRU policy, there exists $\alpha^\star = \alpha^\star(M, N)$ such that $\boldsymbol{p}^\star_{\mathrm{LRU},\alpha} \prec \boldsymbol{p}_\alpha$ does not hold whenever $\alpha > \alpha^\star$.

8.2.3 A conjecture

Theorems 7.5 and 7.6 were valid for *all* values of M and N, and for *arbitrary* admissible pmfs. While the counterexamples discussed earlier dash our hope to get an analogous result for the LRU policy, the possibility remains, fueled by Corollary 6.8, that the positive result is nevertheless valid in some appropriate range of the parameters M and N. We now explore this issue still with Zipf-like popularity pmfs (6.4)-(6.5).

Conjecture 8.7 *Assume the IRM input to have a Zipf-like popularity pmf $\boldsymbol{p}_\alpha$ for some $\alpha \geq 0$. For each $N = 1, 2, \ldots$, under the LRU policy, there exists an integer $M^\star = M^\star(\alpha; N)$ with $1 \leq M^\star < N$ such that $\boldsymbol{p}^\star_{\mathrm{LRU},\alpha} \prec \boldsymbol{p}_\alpha$ whenever $M = 1, \ldots, M^\star$.*

In support of this conjecture, we have carried out simulations of the cache operating under the LRU policy when the IRM input has Zipf-like popularity pmf with parameter

$\alpha = 0.8, 1$ and 2 and $N = 1,000$. We find the output popularity pmfs for different values of cache size, namely $M = 10, 50, 100$ and 500. The resulting output popularity pmfs in the original order of documents are shown in Figure 8.5, while the results after rearranging documents in the decreasing order of their output probabilities are displayed in Figure 8.6.

From Figure 8.6 (a), when $\alpha = 0.8$, the comparison $\boldsymbol{p}_{\mathrm{LRU},\alpha}^{\star} \prec \boldsymbol{p}_{\alpha}$ holds for $M = 10, 50$. This follows from the sufficient condition for majorization comparison provided in Proposition 2.1. Indeed, from their respective plots, we observe that the pmfs $\boldsymbol{p}_{\alpha}$ and $\boldsymbol{p}_{\mathrm{LRU},\alpha}^{\star}$ when arranged in decreasing order intersect only once, namely $p_{\mathrm{LRU},\alpha}^{\star}([i]) \leq p_{\alpha}(i), i = 1, \ldots, k$, and $p_{\mathrm{LRU},\alpha}^{\star}([i]) \geq p_{\alpha}(i), i = k+1, \ldots, N$, for some $k = 1, \ldots, N-1$, where $p_{\mathrm{LRU},\alpha}^{\star}([1]) \geq p_{\mathrm{LRU},\alpha}^{\star}([2]) \geq \ldots \geq p_{\mathrm{LRU},\alpha}^{\star}([N])$ are the components of $\boldsymbol{p}_{\mathrm{LRU},\alpha}^{\star}$ arranged in decreasing order.

However, for $\alpha = 0.8$ and $M = 100, 500$, despite the fact that in Figure 8.6 (a), $\boldsymbol{p}_{\alpha}^{\star}$ of both cases look uniform in the range where document rank is smaller than M, the comparison $\boldsymbol{p}_{\mathrm{LRU},\alpha}^{\star} \prec \boldsymbol{p}_{\alpha}$ is invalid since the necessary condition (8.24) does not hold. This violation, $\min_{i=1,\ldots,N} p_{\mathrm{LRU},\alpha}^{\star}(i) < p_{\alpha}(N)$, can be easily seen from Figure 8.5 (a) or from the subfigure inside Figure 8.6 (a).

For $\alpha = 1$ and $\alpha = 2$, by the same arguments, we conclude from Figures 8.5 (b)-(c) and 8.6 (b)-(c) that the comparison $\boldsymbol{p}_{\mathrm{LRU},\alpha}^{\star} \prec \boldsymbol{p}_{\alpha}$ holds for $M = 10$ but does not hold for other cache sizes $M = 50, 100, 500$. Therefore, these experimental findings agree with Conjecture 8.7 and suggest that the value of $M^{\star}(\alpha; N)$ in Conjecture 8.7 decreases as α increases. This last observation is supported by the observation that for $\alpha = 0$, both $\boldsymbol{p}_0$ and $\boldsymbol{p}_{\mathrm{LRU},0}^{\star}$ are the uniform pmf $\boldsymbol{u}$ on $\mathcal{N}$, thus the comparison $\boldsymbol{p}_{\mathrm{LRU},0}^{\star} \prec \boldsymbol{p}_0$ holds for all $M = 1, \ldots, N-1$, whence $M^{\star}(0; N) = N - 1$.
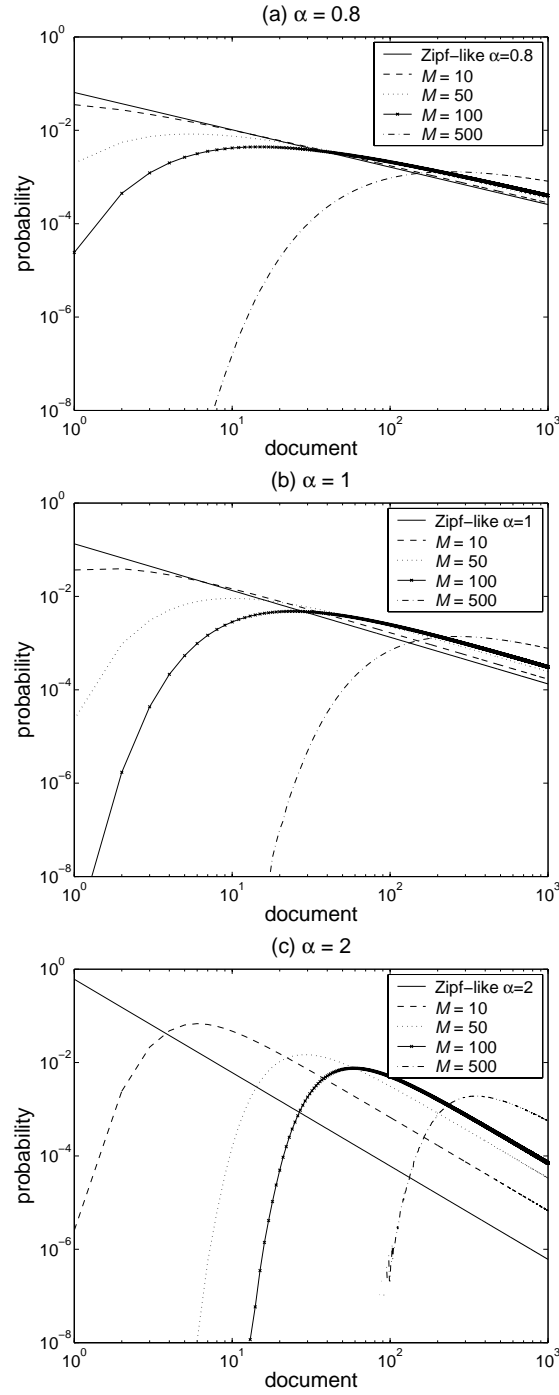
Figure 8.5: LRU output popularity pmf with different cache sizes M when the IRM input has a Zipf-like popularity pmf $\boldsymbol{p}_\alpha$ with (a) $\alpha = 0.8$, (b) $\alpha = 1$ and (c) $\alpha = 2$. Documents are arranged in the original order of the input pmf $\boldsymbol{p}_\alpha$.
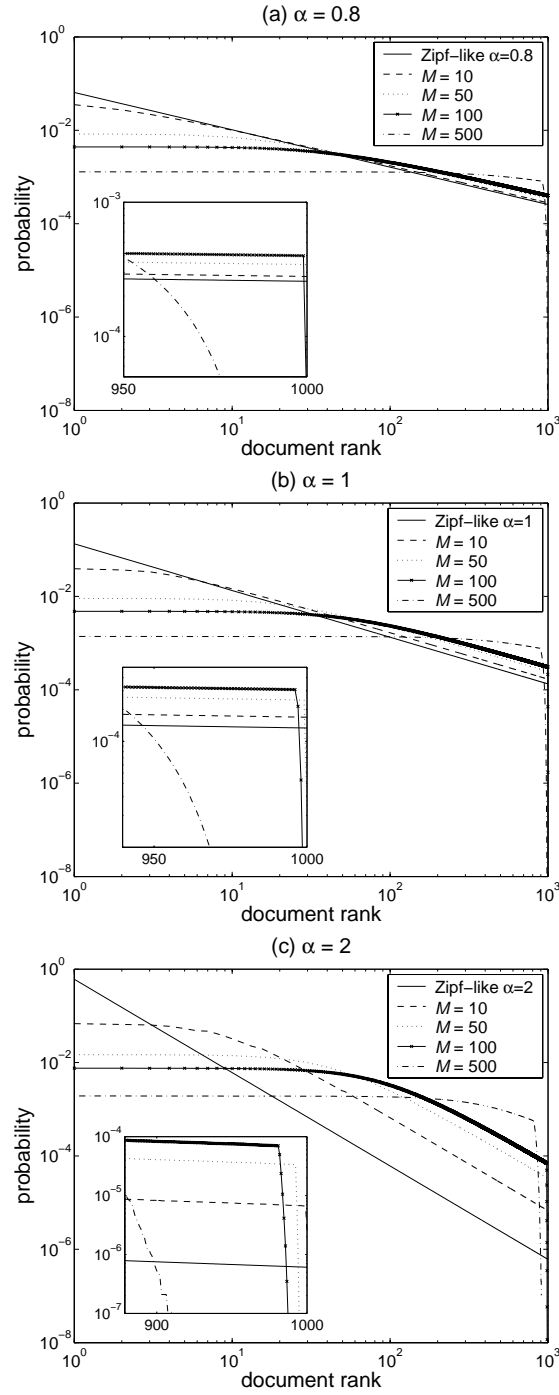
Figure 8.6: LRU output popularity pmf with different cache sizes M when the IRM input has a Zipf-like popularity pmf $\boldsymbol{p}_\alpha$ with (a) $\alpha = 0.8$, (b) $\alpha = 1$ and (c) $\alpha = 2$. Documents are ranked according to their probabilities.

8.3 The miss rate under the CLIMB policy

Under the IRM assumption on the input, the CLIMB cache states $\{\Omega_t, t = 0, 1, \ldots\}$ form a stationary ergodic Markov chain on the finite state space $\Lambda(M; \mathcal{N})$ with stationary distribution [2, p. 133] given by

$$
\begin{aligned}
\mu_{\mathrm{CL}}(s; \boldsymbol{p}) &= \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[\Omega_\tau = s\right] \quad a.s. \\
&= \frac{1}{K_{\mathrm{CL}}} \prod_{\ell=1}^{M} p(i_\ell)^{M-\ell+1}
\end{aligned}
\tag{8.27}
$$

for each $s = (i_1, \ldots, i_M)$ in $\Lambda(M; \mathcal{N})$, where the normalizing constant is simply

$$
K_{\mathrm{CL}} := \sum_{(i_1, \ldots, i_M) \in \Lambda(M; \mathcal{N})} \prod_{\ell=1}^{M} p(i_\ell)^{M-\ell+1}.
$$

The limit (5.1) then exists for each $s = \{i_1, \ldots, i_M\}$ in $\Lambda^\star(M; \mathcal{N})$ as

$$
\begin{aligned}
\mu_{\mathrm{CL}}^\star(s; \boldsymbol{p}) &= \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[S_\tau = s\right] \quad a.s. \\
&= \frac{1}{K_{\mathrm{CL}}} \sum_{(j_1, \ldots, j_M) \in \Lambda(s|M; \mathcal{N})} \prod_{\ell=1}^{M} p(j_\ell)^{M-\ell+1}.
\end{aligned}
\tag{8.28}
$$

The miss rate of the CLIMB policy under IRM can now be obtained [2, Chap. 4] from (5.3) as

$$
\hat{M}_{\mathrm{CL}}(\boldsymbol{p}) = \frac{1}{K_{\mathrm{CL}}} \sum_{(i_1, \ldots, i_M) \in \Lambda(M; \mathcal{N})} \prod_{\ell=1}^{M} p(i_\ell)^{M-\ell+1} \left(1 - \sum_{j=1}^{M} p(i_j)\right)
\tag{8.29}
$$

or from (5.2) as

$$
\hat{M}_{\mathrm{CL}}(\boldsymbol{p}) = \sum_{i=1}^{N} p(i) \sum_{s \in \Lambda_i(M; \mathcal{N})} \frac{1}{K_{\mathrm{CL}}} \prod_{\ell=1}^{M} p(i_\ell)^{M-\ell+1}.
\tag{8.30}
$$

8.3.1 A counterexample

As in the case of the LRU miss rate, the miss rate for the CLIMB policy is in general *not* a Schur-concave function, and thus the folk theorem (6.1) does not hold. We
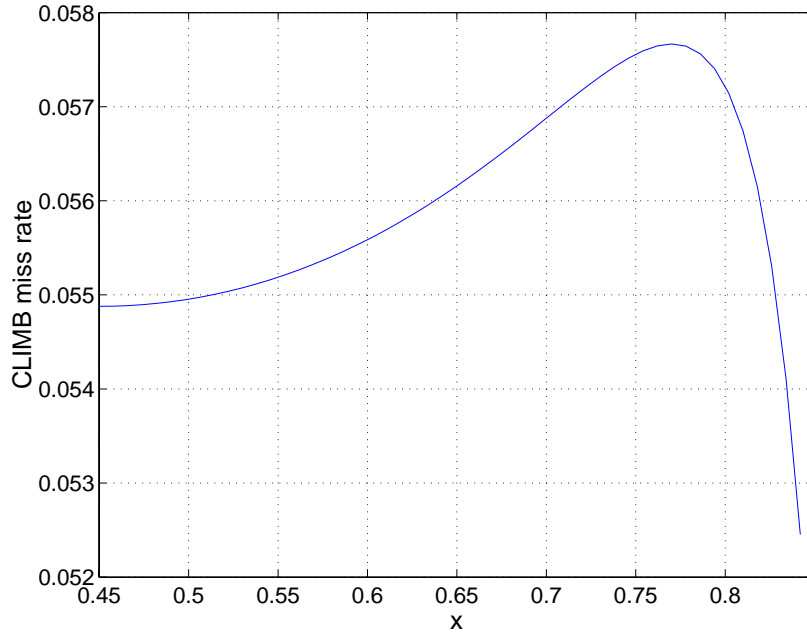
Figure 8.7: CLIMB miss rate when $M = 3$, $N = 4$, $y = p(3) = p(4) = 0.05$, $p(1) = x$ and $p(2) = 0.9 - p(1)$

demonstrate this fact through the same counterexample developed for the LRU policy in Section 8.1.1.

In that case, we set $M = 3$ and $N = 4$ and the expression (8.29) can be simplified as

$$\hat{M}_{\mathrm{CL}}(\boldsymbol{p}) = \frac{2 \prod_{j=1}^{4} p(j) \left(\sum_{i=1}^{4} p(i)^2 (1 - p(i)) \right)}{\sum_{(i_1, i_2, i_3) \in \Lambda(3; \mathcal{N})} p(i_1)^3 p(i_2)^2 p(i_3)}. \tag{8.31}$$

The numerical values of the expression (8.31) are evaluated for the family of pmfs (8.6) with x in the interval $[\frac{1}{2} - y, 1 - 3y]$. Under these constraints, it holds that $\boldsymbol{p}(x, y) \prec \boldsymbol{p}(x', y)$ whenever $x < x'$ in the interval $[\frac{1}{2} - y, 1 - 3y]$ and for the CLIMB miss rate to be Schur-concave, the function $x \to \hat{M}_{\mathrm{CL}}(\boldsymbol{p}(x, y))$ must be monotone *decreasing* on the interval $[\frac{1}{2} - y, 1 - 3y]$.

Figures 8.7 and 8.8 display the numerical values of $\hat{M}_{\mathrm{CL}}(\boldsymbol{p}(x, y))$ as a function of x with $y = 0.05$ and $y = 0.01$, respectively. In both cases, the miss rate of the CLIMB

83
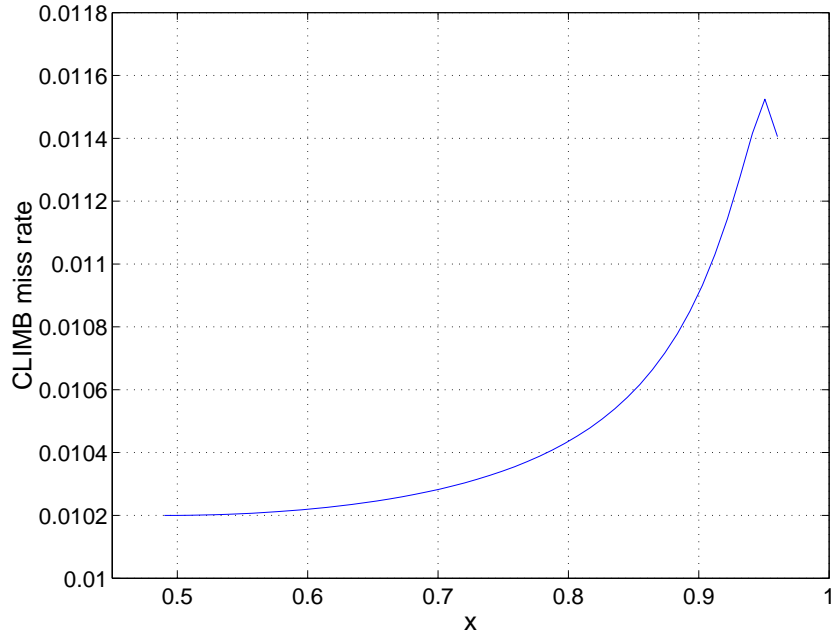
Figure 8.8: CLIMB miss rate when $M = 3$, $N = 4$, $y = p(3) = p(4) = 0.01$, $p(1) = x$ and $p(2) = 0.98 - p(1)$

policy is *not* monotone decreasing in x on the entire range and thus the miss rate is not always Schur-concave under the CLIMB policy.

8.3.2 CLIMB miss rate and IRM with Zipf-like popularity pmfs

Although the CLIMB miss rate is *not* Schur-concave in general, the desired monotonicity (6.1) holds asymptotically when the popularity pmf of the IRM input lies in the class of Zipf-like pmfs.

Theorem 8.8 *Assume the IRM input to have a Zipf-like popularity pmf $\boldsymbol{p}_\alpha$ for some $\alpha \geq 0$. Then, there exists $\alpha^\star = \alpha^\star(M, N) > 0$ and $\Delta > 0$ such that $\hat{M}_{\mathrm{CL}}(\boldsymbol{p}_\beta) < \hat{M}_{\mathrm{CL}}(\boldsymbol{p}_\alpha)$ whenever $\alpha^\star < \alpha$ and $\alpha + \Delta < \beta$.*

Similarly to Theorem 8.1, this theorem is a by-product of the asymptotics

$$\lim_{\alpha \to \infty} \frac{\hat{M}_{\mathrm{CL}}(\boldsymbol{p}_\alpha)}{(M+1)^{-\alpha}} = 2 \tag{8.32}$$

obtained in the Appendix B.3.

In addition, we carry out simulations of a cache operating under the CLIMB policy when the IRM input has a Zipf-like popularity pmf $\boldsymbol{p}_\alpha$. We set the number of documents $N = 1,000$ and cache size $M = 100$. Figure 8.9 and 8.10 show the miss rate of the CLIMB policy when α is small ($0 \le \alpha \le 1$) and large ($\alpha > 1$), respectively. As for the LRU miss rate, the CLIMB miss rate appears to be decreasing as the skewness parameter α increases across the entire range of α, thereby suggesting the following

Conjecture 8.9 *For arbitrary cache size M and number of documents N, the function $\alpha \to \hat{M}_{\mathrm{CL}}(\boldsymbol{p}_\alpha)$ is strictly decreasing on $[0, \infty)$.*

8.4 The output under the CLIMB policy

8.4.1 CLIMB is a good policy

From the expression (8.27), for each $i = 1, \ldots, N$, we have

$$
\begin{aligned}
m_{\mathrm{CL}}(i; \boldsymbol{p}) &= \sum_{s \in \Lambda_i(M;\mathcal{N})} \mu_{\mathrm{CL}}(s; \boldsymbol{p}) \\
&= \frac{1}{K_{\mathrm{CL}}} \sum_{s \in \Lambda_i(M;\mathcal{N})} \prod_{\ell=1}^{M} p(i_\ell)^{M-\ell+1}
\end{aligned} \tag{8.33}
$$

and by Theorem 5.2,

$$p_{\mathrm{CL}}^\star(i) = \frac{p(i)}{\hat{M}_{\mathrm{CL}}(\boldsymbol{p}) K_{\mathrm{CL}}} \sum_{s \in \Lambda_i(M;\mathcal{N})} \prod_{\ell=1}^{M} p(i_\ell)^{M-\ell+1} \tag{8.34}$$

for each $i = 1, \ldots, N$, where we have used the expression (5.8).
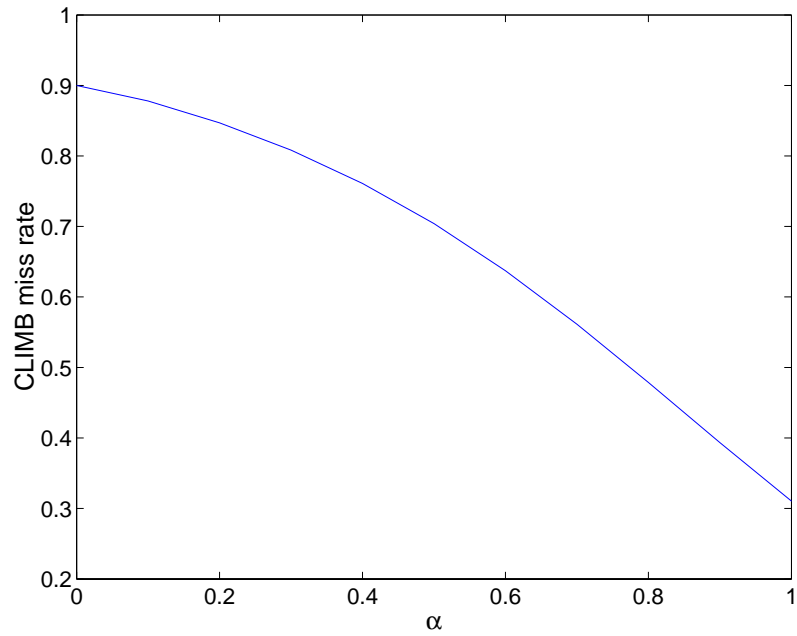
Figure 8.9: CLIMB miss rate when the IRM input has a Zipf-like popularity pmf p_α for α small ($0 \leq \alpha \leq 1$)
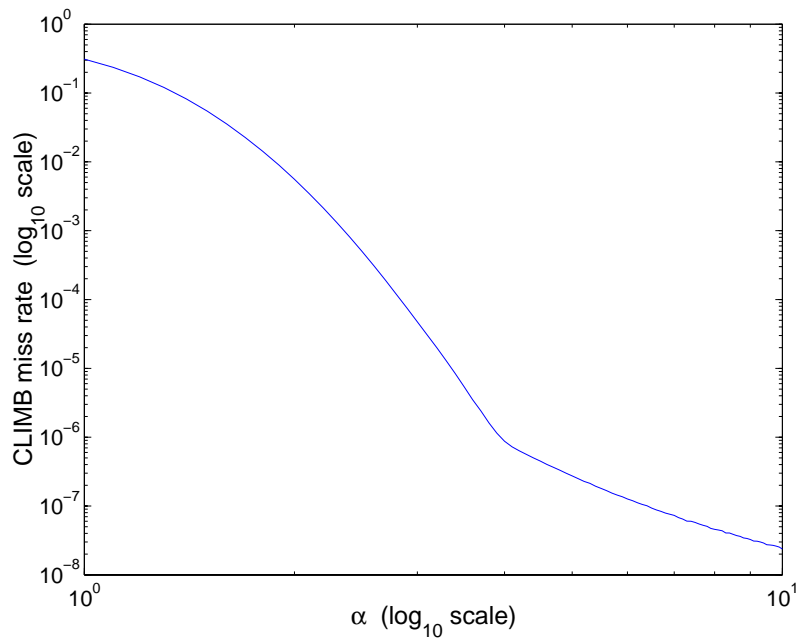
Figure 8.10: CLIMB miss rate when the IRM input has a Zipf-like popularity pmf p_α for α large ($\alpha > 1$)

Lemma 8.10 *The CLIMB policy is a good policy.*

Proof. The proof is essentially that for the analogous result for the LRU policy given in Lemma 8.3. Here the validity of (8.15) follows from the expressions (8.27). ∎

8.4.2 Counterexamples

Again, Corollary 6.8 and Lemma 8.10 might have created the expectation that the majorization comparison $p^{\star}_{\text{CL}} \prec p$ also holds under the CLIMB policy for arbitrary input pmf p. This is not the case as we show by counterexamples when the IRM input has the popularity pmf p_{ε} defined at (8.16). Under this IRM input, it is a simple matter to see from (8.33) and (8.34) that the output popularity pmf $p^{\star}_{\text{CL},\varepsilon}$ is of the form (8.17). Therefore, by Proposition 8.4, the comparison $p^{\star}_{\text{CL},\varepsilon} \prec p_{\varepsilon}$ will not hold if $\delta(\varepsilon) > \frac{1-\varepsilon}{N-1}$. This is indeed the case when ε is small enough; this result is demonstrated in the next theorem whose proof can be found in Appendix C.2.

Theorem 8.11 *Assume the IRM input to have the popularity pmf p_{ε} for some $0 < \varepsilon \leq \frac{1}{N}$. Under the CLIMB policy, whenever*

$$0 < \varepsilon < \frac{1}{2N - 1} \tag{8.35}$$

the comparison $p^{\star}_{\text{CL},\varepsilon} \prec p_{\varepsilon}$ does not hold provided that the number of documents N and the cache size M satisfy the condition $N > M > 2$.

For instance, consider p_{ε} with parameters $N = 10$ and $\varepsilon = 0.05$ and set the cache size $M = 4$. With these parameters, $\delta(\varepsilon) = 0.1110$ and the assumptions of Theorem 8.11 are satisfied. Thus, the comparison $p^{\star}_{\text{CL},\varepsilon} \prec p_{\varepsilon}$ does not hold. However, as was

found in the case of the LRU policy, the entropy comparison is valid in that the entropy of $\boldsymbol{p}_\varepsilon$ is smaller than the entropy of $\boldsymbol{p}^\star_{\mathrm{CL},\varepsilon}$, i.e.,

$$0.7283 = H(\boldsymbol{p}_\varepsilon) \leq H(\boldsymbol{p}^\star_{\mathrm{CL},\varepsilon}) = 0.9560,$$

suggesting that $\boldsymbol{p}^\star_{\mathrm{CL},\varepsilon}$ is more balanced than $\boldsymbol{p}_\varepsilon$ in the sense of entropy comparison.

We next give counterexamples when the IRM input has Zipf-like popularity pmf (6.4)-(6.5). Assume $M = 3$, $N = 4$ and the IRM input has Zipf-like popularity pmf (6.4)-(6.5) with $\alpha = 3$. With these parameters, we have computed the output popularity pmf under the CLIMB policy using (8.34). The numerical values of both input and output popularity pmfs are presented in Table 8.2.

Table 8.2: $\boldsymbol{p}_\alpha$ and $\boldsymbol{p}^\star_{\mathrm{CL},\alpha}$ under the CLIMB policy when the IRM input has a Zipf-like popularity pmf $\boldsymbol{p}_\alpha$ with parameter $\alpha = 3$

i	1	2	3	4
$\boldsymbol{p}_\alpha$	0.8491	0.1061	0.0314	0.0133
$\boldsymbol{p}^\star_{\mathrm{CL},\alpha}$	0.0027	0.1386	0.4000	0.4587

As in the case of the LRU policy, the pmfs $\boldsymbol{p}_\alpha$ and $\boldsymbol{p}^\star_{\mathrm{CL},\alpha}$ are not comparable in the majorization ordering. The arguments are similar to the one given for the LRU policy, and are therefore omitted. Moreover, a result analogous to Theorem 8.6 holds for the CLIMB policy. It is given next, with a proof available in Appendix B.4.

Theorem 8.12 *Assume the IRM input to have a Zipf-like popularity pmf $\boldsymbol{p}_\alpha$ for some $\alpha \geq 0$. If the number of documents N and the cache size M satisfy the condition (8.26), then under the CLIMB policy, there exists $\alpha^\star = \alpha^\star(M, N)$ such that $\boldsymbol{p}^\star_{\mathrm{CL},\alpha} \prec \boldsymbol{p}_\alpha$ does not hold whenever $\alpha > \alpha^\star$.*

8.4.3 A conjecture

Here as well, we venture that a conjecture similar to Conjecture 8.7 is also valid for the CLIMB policy when the IRM input popularity pmf is a Zipf-like distribution (6.4)-(6.5).

Conjecture 8.13 *Assume the IRM input to have a Zipf-like popularity pmf $\boldsymbol{p}_\alpha$ for some $\alpha \geq 0$. For each $N = 1, 2, \ldots$, under the CLIMB policy, there exists an integer $M^\star = M^\star(\alpha; N)$ with $1 \leq M^\star < N$ such that $\boldsymbol{p}^\star_{\mathrm{CL},\alpha} \prec \boldsymbol{p}_\alpha$ whenever $M = 1, \ldots, M^\star$.*

A number of simulation experiments have been carried out under the CLIMB policy, as was done for the LRU policy, to support Conjecture 8.13. The discussion of the experimental results shown in Figure 8.11 and 8.12 is similar to that given in Section 8.2.3 for the LRU policy and shall be omitted.
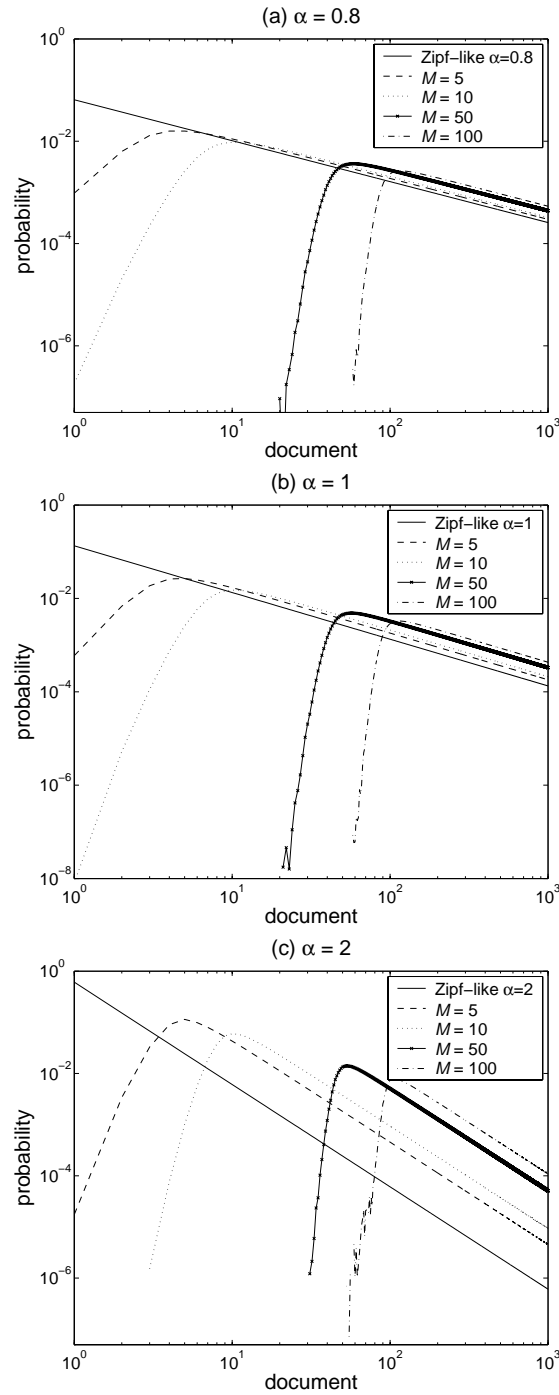
Figure 8.11: CLIMB output popularity pmf with different cache sizes M when the IRM input has a Zipf-like popularity pmf $\boldsymbol{p}_\alpha$ with (a) $\alpha = 0.8$, (b) $\alpha = 1$ and (c) $\alpha = 2$. Documents are arranged in the original order of the input pmf $\boldsymbol{p}_\alpha$.
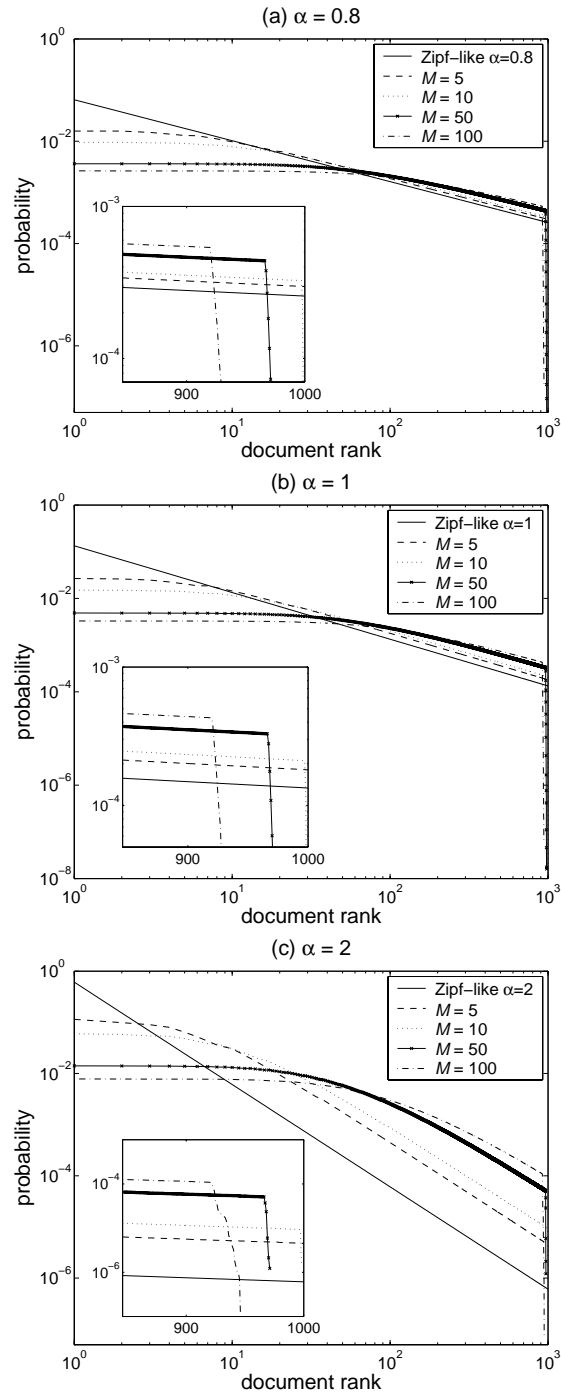
Figure 8.12: CLIMB output popularity pmf with different cache sizes M when the IRM input has a Zipf-like popularity pmf $\boldsymbol{p}_\alpha$ with (a) $\alpha = 0.8$, (b) $\alpha = 1$ and (c) $\alpha = 2$. Documents are ranked according to their probabilities.

Chapter 9

Comparing Temporal Correlations

As was done for popularity, it is natural to seek an appropriate notion which can capture the strength of temporal correlations in streams of requests. Loosely speaking, temporal correlations are understood as the likelihood that a document will be requested in the near future, given that it has been requested in the recent past. Indeed, it is observed in [56] that Web traces usually exhibit short-term temporal correlations in the sense that the probability of requesting a particular document given that the document was recently requested is higher than what it would be if the document has not been recently requested.

In this chapter, we develop a notion that can capture the strength of temporal correlations in Web request streams using the concepts of positive dependence introduced in Chapter 3. Specifically, relying on the notion of supermodular ordering [Definition 3.4], we define the TC ordering [Definition 9.1] for comparing two streams of requests on the basis of the strength of their temporal correlations.

We then apply the TC ordering to investigate the existence of temporal correlations in several Web request models that are believed to exhibit such correlations, namely, the higher-order Markov chain model (HOMM), the partial Markov chain model (PMM) and the Least-Recently-Used stack model (LRUSM). Lastly, with the help of the TC

ordering, we establish a version of the statement to the effect that "the stronger the strength of temporal correlations, the smaller the miss rate" when the input to the cache is modeled by the PMM. Specific results and conjectures on this folk theorem when the input streams are modeled by the HOMM and by the LRUSM are provided.

9.1 Temporal correlations via positive dependence

Given a stream of requests $\boldsymbol{R} = \{R_t, \ t = 0, 1, \ldots\}$, we define for each $i = 1, \ldots, N$, the rvs

$$V_t(i) = \mathbf{1}\left[R_t = i\right], \quad t = 0, 1, \ldots, \tag{9.1}$$

i.e., the rv $V_t(i)$ is the indicator function of the event that the request at time t is made to document i. If the sequence of requests $\{R_t, \ t = 0, 1, \ldots\}$ were to exhibit some form of temporal correlations, then a request to document i would likely be followed by a burst of references to document i in the near future. This corresponds to the presence of positive dependencies in the sequence $\{V_t(i), \ t = 0, 1, \ldots\}$ and leads naturally to the following definition of *Temporal Correlations ordering* (*TC ordering*, for short):

Definition 9.1 *The request stream* $\boldsymbol{R}^1 = \{R_t^1, \ t = 0, 1, \ldots\}$ *is said to have weaker temporal correlations than the request stream* $\boldsymbol{R}^2 = \{R_t^2, \ t = 0, 1, \ldots\}$, *a situation denoted*

$$\boldsymbol{R}^1 \leq_{TC} \boldsymbol{R}^2, \tag{9.2}$$

if for each $i = 1, \ldots, N$, *the comparison*

$$\{V_t^1(i), \ t = 0, 1, \ldots\} \leq_{sm} \{V_t^2(i), \ t = 0, 1, \ldots\}$$

holds where for each $k = 1, 2$, *the rvs* $\{V_t^k(i), \ t = 0, 1, \ldots\}$ *denote the indicator process associated with* $\boldsymbol{R}^k$ *through (9.1).*

Under this definition, whenever $R^1 \leq_{TC} R^2$, it follows from the equi-marginal property (3.3) of the sm ordering that

$$\mathbf{P}\left[V_t^1(i) = 1\right] = \mathbf{P}\left[V_t^2(i) = 1\right], \quad i = 1, \ldots, N,$$

or equivalently that

$$\mathbf{P}\left[R_t^1 = i\right] = \mathbf{P}\left[R_t^2 = i\right], \quad i = 1, \ldots, N, \tag{9.3}$$

for all $t = 0, 1, \ldots$. Therefore, under the assumption that for each $k = 1, 2$, the limits (4.2) exist as constants for the request stream R^k, we have

$$\begin{aligned}
p^k(i) &= \mathbf{E}\left[\lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[R_\tau^k = i\right]\right] \\
&= \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{P}\left[R_\tau^k = i\right], \quad i = 1, \ldots, N,
\end{aligned}$$

by the Bounded Convergence Theorem. Combining this last equation and (9.3) immediately leads to $p^1 = p^2$, i.e., the comparison $R^1 \leq_{TC} R^2$ requires that the request streams R^1 and R^2 must have the same popularity profile. In other words, the TC ordering captures only the contribution from temporal correlations to locality of reference.

Proposition 9.2 *For a request stream R, if each of the indicator processes $\{V_t(i), \; t = 0, 1, \ldots\}$, $i = 1, \ldots, N$, associated with R is PSMD, then it holds that*

$$\hat{R} \leq_{TC} R$$

where $\hat{R}$ is the independent version of R.

When the request stream R is a stationary sequence, the independent version $\hat{R}$ of R is simply the IRM whose popularity pmf is the common marginal of the request stream R.

Proof. Fix $i = 1, \ldots, N$. Under the enforced assumptions, the sequence $\{V_t(i), \; t =$

$0, 1, \ldots\}$ associated with $\boldsymbol{R}$ is PSMD. This amounts to

$$\{\hat{V}_t(i), \ t = 0, 1, \ldots\} \leq_{sm} \{V_t(i), \ t = 0, 1, \ldots\}$$

where the sequence $\{\hat{V}_t(i), \ t = 0, 1, \ldots\}$ is the independent version of the indicator sequence $\{V_t(i), \ t = 0, 1, \ldots\}$. With $\hat{\boldsymbol{R}} = \{\hat{R}_t, \ t = 0, 1, \ldots\}$ being the independent version of the request stream $\boldsymbol{R}$, it is plain that

$$\{\hat{V}_t(i), \ t = 0, 1, \ldots\} =_{st} \{\mathbf{1}\left[\hat{R}_t = i\right], \ t = 0, 1, \ldots\}, \quad i = 1, \ldots, N,$$

and the proof is completed. ∎

In what follows, we investigate whether various request models of interest display temporal correlations in the sense of the TC ordering. These models include the higher-order Markov chain model, the partial Markov chain model and the Least-Recently-Used stack model.

9.2 Higher-order Markov chain models (HOMM)

Several higher-order Markov chain models have been used to characterize Web request streams (e.g., see [19, 28, 56] and references therein) due to their ability to capture some of the observed temporal correlations. Here we rely on a model, recently proposed by Psounis et al. [56], which is capable of capturing both the long-term popularity and short-term temporal correlations of Web request streams.

The model can be described as follows: Let $\mathcal{N}$-valued rvs $\{R_0, \ldots, R_{h-1}\}$ be the initial requests and let $\{Y_t, t = 0, 1, \ldots\}$ be a sequence of i.i.d. $\mathcal{N}$-valued rvs with $\mathbf{P}\left[Y_t = i\right] = p(i)$ for each $i = 1, \ldots, N$. The pmf $\boldsymbol{p} = (p(1), \ldots, p(N))$ is assumed to be admissible (4.3) and as we shall see shortly, it will turn out to be the popularity pmf

of this model. Next, with $0 \leq \alpha_1, \ldots, \alpha_h < 1$ and $\sum_{k=1}^{h} \alpha_k < 1$, let $\{Z_t, t = 0, 1, \ldots\}$ be another sequence of i.i.d. $\{0, 1, \ldots, h\}$-valued rvs with

$$\mathbf{P}\left[Z_t = k\right] = \alpha_k, \quad k = 1, \ldots, h \quad \text{and} \quad \mathbf{P}\left[Z_t = 0\right] = \beta = 1 - \sum_{k=1}^{h} \alpha_k > 0,$$

i.e., the rv Z_t is distributed according to the pmf $\boldsymbol{\alpha} = (\beta, \alpha_1, \ldots, \alpha_h)$. The collections of rvs $\{R_0, \ldots, R_{h-1}\}$, $\{Y_t, t = 0, 1, \ldots\}$ and $\{Z_t, t = 0, 1, \ldots\}$ are mutually independent. For each $t = h, h+1, \ldots$, the request R_t is described by the evolution

$$R_t = \mathbf{1}\left[Z_t = 0\right] Y_t + \sum_{k=1}^{h} \mathbf{1}\left[Z_t = k\right] R_{t-k}. \tag{9.4}$$

In words, the request R_t is made to the same document requested at time $t - k$, namely R_{t-k}, with probability α_k, for some $k = 1, \ldots, h$; otherwise $R_t = Y_t$, i.e., it is chosen independently of the past according to the popularity pmf $\boldsymbol{p}$.

The requests $\{R_t, t = 0, 1, \ldots\}$ form an hth-order Markov chain since the value of R_t depends only on the rvs $R_{t-1}, \ldots, R_{t-h}$. In fact, for $t = h, h+1, \ldots$, we have from (9.4) that for any $(i_0, \ldots, i_{t-1})$ in $\mathcal{N}^t$,

$$\begin{aligned}
\mathbf{P}\left[R_t = i | R_\tau = i_\tau, \tau = 0, \ldots, t-1\right] &= \beta p(i) + \sum_{k=1}^{h} \alpha_k \mathbf{1}\left[i_{t-k} = i\right] \\
&= \mathbf{P}\left[R_t = i | R_\tau = i_\tau, \tau = t-h, \ldots, t-1\right].
\end{aligned} \tag{9.5}$$

With $\beta > 0$, this h^{th}-order Markov chain is irreducible and aperiodic on its finite state space; its stationary distribution exists and is unique. It can be shown [56] that

$$\lim_{t \to \infty} \mathbf{P}\left[R_t = i\right] = \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[R_\tau = i\right] = p(i) \quad a.s.$$

for each $i = 1, \ldots, N$, and it is therefore warranted to call the pmf $\boldsymbol{p}$ the long-term popularity pmf of this request model. Moreover, there exists a unique stationary version, still denoted thereafter by $\{R_t, t = 0, 1, \ldots\}$. The parameters of the model are the history window size h, the pmf $\boldsymbol{\alpha}$ and the popularity pmf $\boldsymbol{p}$, and we shall refer to this model by HOMM($h, \boldsymbol{\alpha}, \boldsymbol{p}$).

That the HOMM($h, \boldsymbol{\alpha}, \boldsymbol{p}$) exhibits temporal correlations is formalized in the next result.

Theorem 9.3 *Assume the request stream $\boldsymbol{R} = \{R_t, t = 0, 1, \ldots\}$ to be modeled according to the stationary HOMM($h, \boldsymbol{\alpha}, \boldsymbol{p}$). Then, for each $i = 1, \ldots, N$, the indicator sequence $\{V_t(i), t = 0, 1, \ldots\}$ associated with the request stream $\boldsymbol{R}$ is PSMD, whence*

$$\hat{\boldsymbol{R}} \leq_{TC} \boldsymbol{R} \tag{9.6}$$

where $\hat{\boldsymbol{R}}$ is the IRM with popularity pmf $\boldsymbol{p}$.

Proof. In order to show that the sequences $\{V_t(i), t = 0, 1, \ldots\}$, $i = 1, \ldots, N$ are PSMD, we shall make use of another sequence of $\mathcal{N}$-valued rvs $\tilde{\boldsymbol{R}} = \{\tilde{R}_t, t = 0, 1, \ldots\}$ constructed as follows: The rvs $\{\tilde{R}_0, \ldots, \tilde{R}_{h-1}\}$ are i.i.d. rvs distributed according to the pmf $\boldsymbol{p}$ and the rvs $\{\tilde{R}_t, t = h, h+1, \ldots\}$ are generated through the evolution (9.4) with the help of mutually independent sequences of i.i.d. rvs $\{\tilde{Y}_t, t = 0, 1, \ldots\}$ and $\{\tilde{Z}_t, t = 0, 1, \ldots\}$ distributed according to the pmfs $\boldsymbol{p}$ and $\boldsymbol{\alpha}$, respectively. The collections of rvs $\{\tilde{Y}_t, t = 0, 1, \ldots\}$ and $\{\tilde{Z}_t, t = 0, 1, \ldots\}$ are taken to be independent of the rvs $\{\tilde{R}_0, \ldots, \tilde{R}_{h-1}\}$. From this construction, the process $\tilde{\boldsymbol{R}} = \{\tilde{R}_t, t = 0, 1, \ldots\}$ is an hth-order Markov chain and with $\beta > 0$, we get

$$\{\tilde{R}_{t+\tau}, t = 0, 1, \ldots\} \Longrightarrow_\tau \{R_t, t = 0, 1, \ldots\}. \tag{9.7}$$

Fix $i = 1, \ldots, N$. Let $\{\tilde{V}_t(i) = \mathbf{1}\left[\tilde{R}_t = i\right], t = 0, 1, \ldots\}$ be the indicator sequence associated with the sequence $\tilde{\boldsymbol{R}}$ defined earlier. We will show that this sequence $\{\tilde{V}_t(i), t = 0, 1, \ldots\}$ is CIS. To do so, for each $t = 0, 1, \ldots$, set $\tilde{\boldsymbol{V}}^t(i) = (\tilde{V}_0(i), \ldots, \tilde{V}_t(i))$. Because the sequence $\{\tilde{V}_t(i), t = 0, 1, \ldots\}$ is a sequence of $\{0, 1\}$-valued rvs, it is CIS [59, 67] if for each $t = 0, 1, \ldots$, the inequality

$$\mathbf{P}\left[\tilde{V}_{t+1}(i) = 1|\tilde{\boldsymbol{V}}^t(i) = \boldsymbol{x}^t\right] \leq \mathbf{P}\left[\tilde{V}_{t+1}(i) = 1|\tilde{\boldsymbol{V}}^t(i) = \boldsymbol{y}^t\right] \tag{9.8}$$

holds for all vectors $\boldsymbol{x}^t = (x_0, \dots, x_t)$ and $\boldsymbol{y}^t = (y_0, \dots, y_t)$ in $\{0,1\}^{t+1}$ with $\boldsymbol{x}^t \le \boldsymbol{y}^t$ componentwise.

For $t = 0, 1, \dots, h - 2$, it holds for all $\boldsymbol{x}^t = (x_0, \dots, x_t)$ in $\{0,1\}^{t+1}$ that

$$\mathbf{P}\left[\tilde{V}_{t+1}(i) = 1 | \tilde{\boldsymbol{V}}^t(i) = \boldsymbol{x}^t\right] = \mathbf{P}\left[\tilde{V}_{t+1}(i) = 1\right] = \mathbf{P}\left[\tilde{R}_{t+1} = i\right] = p(i) \qquad (9.9)$$

by independence of the rvs $\tilde{R}_0, \dots, \tilde{R}_{h-1}$, and the inequality (9.8) is obtained for each $t = 0, 1, \dots, h - 2$. Next, for $t = h - 1, h, \dots$, and $\boldsymbol{x}^t = (x_0, \dots, x_t)$ in $\{0,1\}^{t+1}$, let $(i_0, \dots, i_t)$ be an element in $\mathcal{N}^{t+1}$ with the property that for each $k = 0, \dots, t$, $i_k = i$ if $x_k = 1$ and $i_k \ne i$ if $x_k = 0$. With such an element, we obtain from (9.5) that

$$\mathbf{P}\left[\tilde{V}_{t+1}(i) = 1 | (\tilde{R}_0, \dots, \tilde{R}_t) = (i_0, \dots, i_t)\right]$$
$$= \mathbf{P}\left[\tilde{R}_{t+1} = i | (\tilde{R}_0, \dots, \tilde{R}_t) = (i_0, \dots, i_t)\right]$$
$$= \beta p(i) + \sum_{k=1}^{h} \alpha_k \mathbf{1}\left[i_{t+1-k} = i\right]$$
$$= \beta p(i) + \sum_{k=1}^{h} \alpha_k x_{t+1-k}. \qquad (9.10)$$

Since (9.10) holds for any $(i_0, \dots, i_t)$ in $\mathcal{N}^{t+1}$ satisfying the property above, a standard preconditioning argument readily yields

$$\mathbf{P}\left[\tilde{V}_{t+1}(i) = 1 | \tilde{\boldsymbol{V}}^t(i) = \boldsymbol{x}^t\right] = \beta p(i) + \sum_{k=1}^{h} \alpha_k x_{t+1-k}. \qquad (9.11)$$

This last expression being monotone increasing in $\boldsymbol{x}^t = (x_0, \dots, x_t)$, we obtain the inequality (9.8) for each $t = h - 1, h, \dots$.

Thus, the inequalities (9.8) hold for *all* $t = 0, 1, \dots$. This implies that the sequence $\{\tilde{V}_t(i),\ t = 0, 1, \dots\}$ is CIS, whence indeed PSMD by Theorem 3.10, i.e.,

$$\{\hat{\tilde{V}}_t(i),\ t = 0, 1, \dots\} \le_{sm} \{\tilde{V}_t(i),\ t = 0, 1, \dots\} \qquad (9.12)$$

where $\{\hat{\tilde{V}}_t(i),\ t = 0, 1, \dots\}$ is the independent version of $\{\tilde{V}_t(i), t = 0, 1, \dots\}$. Now, recalling (9.7), it is plain that

$$\{\hat{\tilde{V}}_{t+\tau}(i),\ t = 0, 1, \dots\} \Longrightarrow_\tau \{\hat{V}_t(i),\ t = 0, 1, \dots\} \qquad (9.13)$$

where $\{\hat{V}_t(i), t = 0, 1, \ldots\}$ is a sequence of i.i.d. $\{0, 1\}$-valued rvs with $\mathbf{P}\left[\hat{V}_0(i) = 1\right] = p(i)$ and is exactly the independent version of $\{V_t(i), t = 0, 1, \ldots\}$. By invoking the fact that the sm ordering is closed under weak convergence [52, Thm. 3.9.8, p. 116], we conclude from (9.7), (9.12) and (9.13) that

$$\{\hat{V}_t(i),\ t = 0, 1, \ldots\} \leq_{sm} \{V_t(i),\ t = 0, 1, \ldots\}.$$

Therefore, the sequence $\{V_t(i), t = 0, 1, \ldots\}$ is PSMD for each $i = 1, \ldots, N$, and by Proposition 9.2, the comparison $\hat{\boldsymbol{R}} \leq_{TC} \boldsymbol{R}$ holds with $\hat{\boldsymbol{R}}$ being the independent version of $\boldsymbol{R}$. ∎

9.3 Partial Markov chain models (PMM)

The partial Markov chain model was introduced early on in the literature as a reference model for computer memory paging [2]. It is a subclass of higher-order Markov chain models and corresponds to HOMM$(h, \boldsymbol{\alpha}, \boldsymbol{p})$ with parameter $h = 1$. In that case, we have $\boldsymbol{\alpha} = (\beta, \alpha_1)$ where $\alpha_1 = 1 - \beta$ and we refer to this model as PMM$(\beta, \boldsymbol{p})$.

Under this model, with probability $1 - \beta$, $R_t = R_{t-1}$, otherwise with probability β, $R_t = Y_t$, i.e., R_t is drawn independently of the past according to the popularity pmf $\boldsymbol{p}$. Therefore, it is natural to expect that when the popularity pmf $\boldsymbol{p}$ is held fixed, the smaller the value of correlation parameter β, the greater temporal correlations exhibited by the PMM$(\beta, \boldsymbol{p})$. In the extreme cases, as $\beta \uparrow 1$, the PMM$(\beta, \boldsymbol{p})$ becomes the IRM with popularity pmf $\boldsymbol{p}$ and there is no temporal correlations. On the other hand, as $\beta \downarrow 0$, all the requests are made to the same document, hence displaying the strongest possible form of temporal correlations. The following result, which contains Theorem 9.3 when $h = 1$, formalizes these statements with the help of the TC ordering, thereby confirming

the intuition that the parameter β of PMM($\beta, \boldsymbol{p}$) is indeed a measure of the strength of temporal correlations.

Theorem 9.4 *Assume that for each* $k = 1, 2$, *the request stream* $\boldsymbol{R}^{\beta_k} = \{R_t^{\beta_k}, t = 0, 1, \ldots\}$ *is modeled according to the stationary PMM($\beta_k, \boldsymbol{p}$). If* $0 < \beta_2 \leq \beta_1$, *then*

$$\boldsymbol{R}^{\beta_1} \leq_{TC} \boldsymbol{R}^{\beta_2}. \tag{9.14}$$

The proof of this theorem relies on the following comparison of Markov chains under the supermodular ordering due to Bäuerle [8].

Theorem 9.5 *Let* $\boldsymbol{X} = \{X_t, t = 0, 1, \ldots\}$ *and* $\boldsymbol{X}' = \{X_t', t = 0, 1, \ldots\}$ *be two station-ary Markov chains on* $\{0, 1, \ldots, n\}$ *with transition matrices* $\boldsymbol{P}$ *and* $\boldsymbol{P}'$, *respectively. For* $\gamma_0, \ldots, \gamma_n \geq 0$ *with* $0 < \sum_{j=0}^n \gamma_j \leq 1$, *define the* $(n+1) \times (n+1)$ *matrix*

$$\boldsymbol{Q}(\gamma_0, \ldots, \gamma_n) = \begin{bmatrix} 1 - \sum_{j \neq 0} \gamma_j & \gamma_1 & \cdots & \gamma_n \\ \gamma_0 & 1 - \sum_{j \neq 1} \gamma_j & \cdots & \gamma_n \\ \vdots & \vdots & & \vdots \\ \gamma_0 & \gamma_1 & \cdots & 1 - \sum_{j \neq n} \gamma_j \end{bmatrix}. \tag{9.15}$$

With $\boldsymbol{P} = \boldsymbol{Q}(\gamma_0, \ldots, \gamma_n)$ *and* $\boldsymbol{P}' = \boldsymbol{Q}(c\gamma_0, \ldots, c\gamma_n)$ *for some* $0 \leq c \leq 1$, *it holds that*

$$\boldsymbol{X} \leq_{sm} \boldsymbol{X}'.$$

Proof of Theorem 9.4. Fix $i = 1, \ldots, N$. Given a sequence $\boldsymbol{R}^\beta = \{R_t^\beta, t = 0, 1, \ldots\}$ modeled according to the PMM($\beta, \boldsymbol{p}$), it follows from (9.11) that the sequence $\{V_t^\beta(i), t = 0, 1, \ldots\}$ associated with $\boldsymbol{R}^\beta$ is a Markov chain on $\{0, 1\}$ with

$$\mathbf{P}\left[V_{t+1}^\beta(i) = 1 | V_t^\beta(i) = x_t, \ldots, V_0^\beta(i) = x_0\right] = \beta p(i) + (1 - \beta)x_t, \quad t = 0, 1, \ldots,$$

for any $(x_0, \ldots, x_t)$ in $\{0, 1\}^{t+1}$. Its transition matrix $\boldsymbol{P}^\beta(i)$ is simply given by

$$\boldsymbol{P}^\beta(i) = \begin{bmatrix} 1 - \beta p(i) & \beta p(i) \\ \beta(1 - p(i)) & 1 - \beta(1 - p(i)) \end{bmatrix},$$

or equivalently, in the notation (9.15), $\boldsymbol{P}^\beta(i) = \boldsymbol{Q}(\gamma_0, \gamma_1)$ where $\gamma_0 = \beta(1 - p(i))$ and $\gamma_1 = \beta p(i)$ with $0 < \gamma_0 + \gamma_1 = \beta \leq 1$.

For two stationary PMM request streams $\boldsymbol{R}^{\beta_1}$ and $\boldsymbol{R}^{\beta_2}$ with $0 < \beta_2 \leq \beta_1$, we can always write $\beta_2 = c\beta_1$ with $0 < c = \frac{\beta_2}{\beta_1} \leq 1$. Thus, the sequences $\{V_t^{\beta_1}(i), t = 0, 1, \ldots\}$ and $\{V_t^{\beta_2}(i), t = 0, 1, \ldots\}$ have transition matrices

$$\boldsymbol{P}^{\beta_1}(i) = \boldsymbol{Q}(\gamma_0, \gamma_1) \quad \text{and} \quad \boldsymbol{P}^{\beta_2}(i) = \boldsymbol{Q}(c\gamma_0, c\gamma_1),$$

respectively, with $\gamma_0 = \beta_1(1 - p(i))$, $\gamma_1 = \beta_1 p(i)$ and $c = \frac{\beta_2}{\beta_1}$. By applying Theorem 9.5, we obtain the comparison

$$\{V_t^{\beta_1}(i), t = 0, 1, \ldots\} \leq_{sm} \{V_t^{\beta_2}(i), t = 0, 1, \ldots\}$$

for each $i = 1, \ldots, N$, and the conclusion (9.14) follows upon recalling Definition 9.1 of the TC ordering. ∎

9.4 Least-Recently-Used stack models (LRUSM)

The Least-Recently-Used stack model (LRUSM) has long been known to be a good model for generating the sequence of requests whose statistical properties match those of observed reference streams [24, 61]. We first state the definition and basic properties of the LRUSM, and then show that under some appropriate assumptions on the model, the LRUSM exhibits stronger strength of temporal correlations than its independent version in the TC ordering.

9.4.1 LRU stack and stack distance

We begin with the notion of *LRU stack* and *stack distance*. For each $t = 0, 1, \ldots$, the stack $\Omega_t = (\Omega_t(1), \ldots, \Omega_t(N))$ is defined as an element in $\Lambda(N; \mathcal{N})$, i.e., Ω_t is an *ordered* sequence of the documents $\{1, \ldots, N\}$. It is customary to assume that $\Omega(1)$ is in the top position of the stack, followed by $\Omega_t(2), \ldots, \Omega_t(N)$, in that order.

Given an initial stack Ω_0 in $\Lambda(N; \mathcal{N})$, with any stream of requests $\boldsymbol{R} = \{R_t,\ t = 0, 1, \ldots\}$, we can associate a stack sequence $\{\Omega_t, t = 0, 1, \ldots\}$ through the following recursive mechanism: For each $t = 0, 1, \ldots$, let D_t denotes the position of the document R_{t+1} in the stack Ω_t, i.e., the rv D_t is the unique element of $\{1, \ldots, N\}$ such that

$$\Omega_t(D_t) = R_{t+1}.$$

The stack Ω_{t+1} is then given by

$$\Omega_{t+1}(k) = \begin{cases} \Omega_t(D_t) & \text{if } k = 1 \\ \Omega_t(k-1) & \text{if } k = 2, \ldots, D_t \\ \Omega_t(k) & \text{if } k = D_t + 1, \ldots, N. \end{cases} \qquad (9.16)$$

In words, the document $\Omega_t(D_t) = R_{t+1}$ is moved up to the highest position (i.e., position 1) in the stack Ω_{t+1} at time $t + 1$ and the documents $\Omega_t(1), \ldots, \Omega_t(D_t - 1)$ are shifted down by one position while the documents $\Omega_t(D_t + 1), \ldots, \Omega_t(N)$ remain unchanged. We refer to the rvs $\{D_t, t = 0, 1, \ldots\}$ so defined as the stack distance sequence associated with the request stream $\boldsymbol{R}$.

Conversely, given the initial stack Ω_0 in $\Lambda(N; \mathcal{N})$, with any sequence of $\{1, \ldots, N\}$-valued rvs $\{D_t, t = 0, 1, \ldots\}$, we can use the stack operation (9.16) to generate a sequence of $\Lambda(N; \mathcal{N})$-valued rvs $\{\Omega_t, t = 0, 1, \ldots\}$. A request stream $\boldsymbol{R}$ is readily generated from this stack sequence by reading off the top of the stack, i.e., with $R_0 = \Omega_0(1)$, we have

$$R_{t+1} = \Omega_t(D_t) = \Omega_{t+1}(1), \quad t = 0, 1, \ldots. \qquad (9.17)$$

Note that the rvs $\{D_t, t = 0, 1, \ldots\}$ constitute the stack distance sequence associated with the request stream $\boldsymbol{R}$ defined at (9.17).

The stack and stack distance introduced above are often referred to as LRU stack and stack distance, respectively, in reference to the popular LRU policy. The dynamics of the LRU policy are best described through the notion of LRU stack and stack distance as we now briefly explain: Returning to (9.16), we see that the stack Ω_t at time t ranks the documents according to their recency of reference with the most recently requested document remaining at the highest stack position. For each $k = 1, \ldots, N$, the document $\Omega_t(k)$ at position k in the stack Ω_t is the kth most recently referenced document at time t, hence the name, LRU stack. Consequently, the documents $\Omega_t(1), \ldots, \Omega_t(M)$ in the first M positions of the stack Ω_t simply yield the documents in cache under the LRU policy with cache size M when the requests $R_0, \ldots, R_t$ have already been served, i.e., $S_{t+1} = \{\Omega_t(1), \ldots, \Omega_t(M)\}$ where S_{t+1} is the LRU cache at time $t + 1$. With this observation in mind, a miss of the LRU cache of size M will occur at time $t + 1$ if $D_t > M$ and thus the miss rate (4.8) under the LRU policy can alternatively be given by the limit

$$M_{\mathrm{LRU}}(\boldsymbol{R}) = \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbf{1}\left[D_\tau > M\right] \quad a.s. \tag{9.18}$$

whenever the limit exists.

9.4.2 The LRU stack model

The duality between streams of requests and stack distances embedded in (9.16) can be used to advantage in defining sequences of requests with temporal correlations. We present one of the simplest ways to do just that: The *Least-Recently-Used stack model* (LRUSM) with pmf $\boldsymbol{a}$ on $\mathcal{N}$ is defined as the request stream $\boldsymbol{R}^{\boldsymbol{a}} = \{R_t^{\boldsymbol{a}}, t = 0, 1, \ldots\}$ whose stack distance sequence $\{D_t, t = 0, 1, \ldots\}$ is a collection of *i.i.d.* rvs distributed

according to the pmf $\boldsymbol{a}$, i.e.,

$$\mathbf{P}\left[D_t = k\right] = a_k, \quad k = 1, \ldots, N; \ t = 0, 1, \ldots,$$

given some arbitrary initial stack Ω_0 in $\Lambda(N; \mathcal{N})$. Throughout we assume that the rv Ω_0 is independent of the stack distances $\{D_t, t = 0, 1, \ldots\}$. In fact, provided $a_N > 0$, when the initial stack rv Ω_0 is uniformly distributed over $\Lambda(N; \mathcal{N})$, the stack rvs $\{\Omega_t, \ t = 0, 1, \ldots\}$ form a stationary sequence, and so do the request rvs $\{R_t^{\boldsymbol{a}}, t = 0, 1, \ldots\}$. This fact is established in the process of proving Proposition 9.6 in Appendix D.1. We shall denote this request model by LRUSM($\boldsymbol{a}$).

From (9.18), the miss rate of the LRUSM($\boldsymbol{a}$) under the LRU policy with cache size M is simply

$$M_{\mathrm{LRU}}(\boldsymbol{R^a}) = \mathbf{P}\left[D_t > M\right] = \sum_{k=M+1}^{N} a_k \tag{9.19}$$

by the Strong Law of Large Number. The LRU policy is known to be an optimal policy for the LRUSM($\boldsymbol{a}$) in the sense that the LRU policy minimizes the miss rate of the request stream $\boldsymbol{R^a}$ over the class of replacement policies (4.5) if the stack distance pmf $\boldsymbol{a}$ satisfies the LRU optimality condition [58]

$$(N - k)a_k \geq \sum_{j=k+1}^{N} a_j, \quad k = 1, \ldots, N. \tag{9.20}$$

The popularity pmf of the LRUSM is discussed first in Proposition 9.6; its proof can be found in Appendix D.1.

Proposition 9.6 *Assume the request stream $\boldsymbol{R^a} = \{R_t^{\boldsymbol{a}}, t = 0, 1, \ldots\}$ to be modeled according to the LRUSM($\boldsymbol{a}$). If $a_N > 0$, then for each $i = 1, \ldots, N$, it holds that*

$$p_{\boldsymbol{a}}(i) = \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[R_\tau^{\boldsymbol{a}} = i\right] = \frac{1}{N} \quad a.s. \tag{9.21}$$

Thus, under LRUSM, as every document is equally popular, locality of reference is expressed solely through temporal correlations with no contribution from the popularity of documents. This was found to be a drawback of the LRUSM for characterizing Web request streams and several variants of this model have been proposed to accommodate this shortcoming [4, 14, 18].

9.4.3 Temporal correlations in LRUSM

As was done with the HOMM, we show that the TC ordering also captures the strength of temporal correlations exhibited by the LRUSM. Recall the sequence of indicator functions $\{V_t^{\boldsymbol{a}}(i) = \mathbf{1}\left[R_t^{\boldsymbol{a}} = i\right], t = 0, 1, \ldots\}$, $i = 1, \ldots, N$, associated with the LRUSM request stream $\{R_t^{\boldsymbol{a}}, t = 0, 1, \ldots\}$. The main result is contained in

Theorem 9.7 *Assume the request stream $\boldsymbol{R^a} = \{R_t^{\boldsymbol{a}}, t = 0, 1, \ldots\}$ to be modeled according to the LRUSM($\boldsymbol{a}$) with stack distance pmf $\boldsymbol{a}$ satisfying*

$$a_1 \geq a_2 \geq \ldots \geq a_N > 0. \tag{9.22}$$

Then, for each $i = 1, \ldots, N$, the indicator sequence $\{V_t^{\boldsymbol{a}}(i), t = 0, 1, \ldots\}$ associated with the request stream $\boldsymbol{R^a}$ is CIS, whence

$$\hat{\boldsymbol{R}}^a \leq_{TC} \boldsymbol{R}^a \tag{9.23}$$

where $\hat{\boldsymbol{R}}^a$ is the independent version of $\boldsymbol{R}^a$.

A proof of Theorem 9.7 can be found in Appendix D.2. In view of Proposition 9.6, when the LRUSM request stream $\boldsymbol{R}^a$ is stationary, its independent version $\hat{\boldsymbol{R}}^a$ is simply the IRM with uniform popularity pmf $\boldsymbol{u} = (\frac{1}{N}, \ldots, \frac{1}{N})$. In fact, it is not hard to see that the stationary LRUSM($\boldsymbol{u}$) indeed coincides with the IRM with uniform popularity pmf $\boldsymbol{u}$. Notice that the condition (9.22) for the LRUSM($\boldsymbol{a}$) to exhibit temporal correlations

in the sense of the TC ordering (9.23) does imply the LRU optimality condition (9.20). This confirms the intuition that the LRU policy is designed to work best with the stream that exhibits temporal correlations amongst its requests.

9.5 Folk theorem on miss rates

With the help of the TC ordering, we can now use the results of Theorems 9.3, 9.4 and 9.7 to explore the folk theorem to the effect that the stronger the strength of temporal correlations, the smaller the miss rate under the PMM, the HOMM and the LRUSM, respectively. Specific results and conjectures are provided next for the PMM, the HOMM and the LRUSM, respectively.

9.5.1 PMM

The miss rates of PMM under demand-driven cache replacement policies have been previously considered in [2]. For particular caching policies such as LRU and FIFO, the miss rate under $\mathrm{PMM}(\beta, \boldsymbol{p})$ is shown to be proportional to the miss rate of the IRM with the same popularity pmf $\boldsymbol{p}$. We first demonstrate this fact in some generality and then use it to compare the miss rates of two PMM streams with different strength of temporal correlations.

As we seek to evaluate the limit (4.8) for the $\mathrm{PMM}(\beta, \boldsymbol{p})$ under the cache replacement policy π, we shall need the following definitions: For each $T = 1, 2, \ldots$, define

$$\lambda(T) = \sum_{t=1}^{T} \mathbf{1} \left[Z_t = 0 \right]$$

as the number of times from time 1 up to time T that the requests are chosen independently of the past according to the popularity pmf $\boldsymbol{p}$. Also, for each $k = 1, 2, \ldots$, let

$\gamma(k) = \inf\{t = 1, 2, \ldots : \lambda(t) = k\}$. Under demand-driven caching with the PMM input, a miss can only occur at the time epochs $\gamma(k)$ $(k = 1, 2, \ldots)$ at which point we have $R^{\beta}_{\gamma(k)} = Y_{\gamma(k)}$. Therefore, it follows from the definition of the rvs $\{\gamma(k), k = 1, 2, \ldots\}$ that

$$
\begin{aligned}
\sum_{t=1}^{T} \mathbf{1}\left[R^{\beta}_t \notin S_t\right] &= \sum_{k=1}^{\lambda(T)} \mathbf{1}\left[R^{\beta}_{\gamma(k)} \notin S_{\gamma(k)}\right] \\
&= \sum_{k=1}^{\lambda(T)} \mathbf{1}\left[Y_{\gamma(k)} \notin S_{\gamma(k)}\right], \quad T = 1, 2, \ldots,
\end{aligned}
$$

and the miss rate under PMM$(\beta, \boldsymbol{p})$ is given by

$$
\begin{aligned}
M_{\pi}(\boldsymbol{R}^{\beta}) &= \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\left[R^{\beta}_t \notin S_t\right] \\
&= \lim_{T \to \infty} \left(\frac{\lambda(T)}{T}\right)\left(\frac{1}{\lambda(T)} \sum_{k=1}^{\lambda(T)} \mathbf{1}\left[Y_{\gamma(k)} \notin S_{\gamma(k)}\right]\right). \quad (9.24)
\end{aligned}
$$

By the Strong Law of Large Numbers, we see that the limit of the first term in (9.24) is simply

$$
\lim_{T \to \infty} \frac{\lambda(T)}{T} = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\left[Z_t = 0\right] = \beta \quad a.s. \quad (9.25)
$$

The limit of the second term in (9.24) in general does not necessarily have a closed-form expression. However, It does admit a simple expression in the special case when the cache replacement policy π satisfies the following condition:

($\star$) For all $t = 1, 2, \ldots$, if $R_t = R_{t-1}$, then the cache state and eviction rule at time $t + 1$ is the same as those at time t, i.e., $\Omega_{t+1} = \Omega_t$ and $U_{t+1} = U_t$.

Under this condition, we can write the second limit as

$$
\begin{aligned}
\lim_{T \to \infty} \frac{1}{\lambda(T)} \sum_{k=1}^{\lambda(T)} \mathbf{1}\left[Y_{\gamma(k)} \notin S_{\gamma(k)}\right] &= \lim_{K \to \infty} \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}\left[Y_{\gamma(k)} \notin S_{\gamma(k)}\right] \\
&= \hat{M}_{\pi}(\boldsymbol{p}) \quad (9.26)
\end{aligned}
$$

where $\hat{M}_{\pi}(\boldsymbol{p})$ is the miss rate of the IRM with popularity pmf $\boldsymbol{p}$ under the policy π. The last equality follows from the fact that the rvs $\{Y_{\gamma(k)}, k = 1, 2, \ldots\}$ form an IRM

with popularity pmf $\boldsymbol{p}$ and that by Condition ($\star$), the cache sets $\{S_{\gamma(k)}, k = 1, 2, \ldots\}$ are similar to the cache sets under the policy π when the input is the IRM sequence $\{Y_{\gamma(k)}, k = 1, 2, \ldots\}$. Combining (9.24), (9.25) and (9.26) yields the expression for the miss rate of PMM($\beta, \boldsymbol{p}$) as

$$M_\pi(\boldsymbol{R}^\beta) = \beta \cdot \hat{M}_\pi(\boldsymbol{p}). \tag{9.27}$$

Condition ($\star$) is satisfied by many cache replacement policies of interest, e.g., the policy A_0, the LRU, FIFO and random policies but not by the CLIMB policy. Equipped with the expression (9.27), we can now conclude to the following monotonicity result.

Theorem 9.8 *Assume that the cache replacement policy π satisfies Condition ($\star$) and that for each $k = 1, 2$, the request stream $\boldsymbol{R}^{\beta_k}$ is modeled according to PMM($\beta_k, \boldsymbol{p}^k$). If $\boldsymbol{p}^1 = \boldsymbol{p}^2$ and $0 < \beta_2 \leq \beta_1$, then it holds that*

$$M_\pi(\boldsymbol{R}^{\beta_2}) \leq M_\pi(\boldsymbol{R}^{\beta_1}). \tag{9.28}$$

Moreover, if the mapping $\boldsymbol{p} \to \hat{M}_\pi(\boldsymbol{p})$ is Schur-concave, then whenever $\boldsymbol{p}^1 \prec \boldsymbol{p}^2$ and $0 < \beta_2 \leq \beta_1$, the comparison (9.28) also holds.

In view of Theorem 9.4, we conclude that the folk theorem on the miss rate indeed holds for the PMM under any cache replacement policy which satisfies Condition ($\star$).

9.5.2 HOMM

Consider the following situation: Let $\boldsymbol{R}$ be HOMM($h, \boldsymbol{\alpha}, \boldsymbol{p}$) for some pmf vectors $\boldsymbol{p}$ on $\mathcal{N}$ and $\boldsymbol{\alpha}$ on $\{0, \ldots, h\}$. For some $0 < c < 1$, let $\boldsymbol{R}^c$ denote HOMM($h, \boldsymbol{\alpha}^c, \boldsymbol{p}$) where $\boldsymbol{\alpha}^c$ is obtained from $\boldsymbol{\alpha}$ by taking $\alpha_k^c = c\alpha_k$ for each $k = 1, \ldots, h$, and $\beta^c = 1 - c(1 - \beta) = \beta + (1 - c)(1 - \beta)$. Obviously, $\beta^c \geq \beta$ while $\alpha_k^c \leq \alpha_k$ for each $k = 1, \ldots, h$. In other words, under HOMM($h, \boldsymbol{\alpha}, \boldsymbol{p}$), there is a smaller probability to generate a new request

independently of past requests than under HOMM($h, \boldsymbol{\alpha}^c, \boldsymbol{p}$). Therefore, in an attempt to generalize Theorem 9.3, it is reasonable to think that HOMM($h, \boldsymbol{\alpha}^c, \boldsymbol{p}$) has less temporal correlations than HOMM($h, \boldsymbol{\alpha}, \boldsymbol{p}$) according to the TC ordering, i.e., $\boldsymbol{R}^c \leq_{TC} \boldsymbol{R}$. Taking our cue from Theorem 9.8, we would then expect the inequality $M_\pi(\boldsymbol{R}) \leq M_\pi(\boldsymbol{R}^c)$ to hold for some good caching policies. We summarize these expectations as the following conjecture:

Conjecture 9.9 *Assume the request stream $\boldsymbol{R} = \{R_t, t = 0, 1, \ldots\}$ to be modeled according to HOMM($h, \boldsymbol{\alpha}, \boldsymbol{p}$). For some $0 < c < 1$, if the request stream $\boldsymbol{R}^c = \{R_t^c, t = 0, 1, \ldots\}$ is modeled according to HOMM($h, \boldsymbol{\alpha}^c, \boldsymbol{p}$) with $\boldsymbol{\alpha}^c = (1 - c(1 - \beta), c\alpha_1, \ldots, c\alpha_h)$, then the comparison $\boldsymbol{R}^c \leq_{TC} \boldsymbol{R}$ holds. Furthermore, under some appropriate cache replacement policy π, it holds that $M_\pi(\boldsymbol{R}) \leq M_\pi(\boldsymbol{R}^c)$.*

Establishing this conjecture appears to be much more difficult than for the PMM, and requires further investigation. However, in support of this conjecture, we have carried out several experiments under the LRU policy when the input to the cache is modeled according to the HOMM. Throughout, we fix $N = 100$ and let the input popularity pmf $\boldsymbol{p}$ be the Zipf-like distribution $\boldsymbol{p}_\alpha$ (6.4)-(6.5) with parameter $\alpha = 0.8$. We consider five different classes of HOMM, each with different history window size $h = 1, \ldots, 5$. In each class, the input stream $\boldsymbol{R}^\beta$ (with $0 \leq \beta \leq 1$), is generated according to HOMM($h, \boldsymbol{\alpha}_h(\beta), \boldsymbol{p}_\alpha$) with $\boldsymbol{\alpha}_h(\beta) = (\beta, \frac{1-\beta}{h}, \ldots, \frac{1-\beta}{h})$. The validity of Conjecture 9.9 would require that the mapping $\beta \to M_{\mathrm{LRU}}(\boldsymbol{R}^\beta)$ be increasing.

From Figure 9.1, the miss rate is indeed found to be increasing as the parameter β increases for all cases and for all cache sizes. When $h = 1$, HOMM reduces to PMM and the results here confirm the validity of the expression (9.27) and of Theorem 9.8. It is interesting to note that for a given cache size M, the miss rates of all HOMM input streams with $h \leq M$ are the same as the miss rate of the PMM. This suggests some
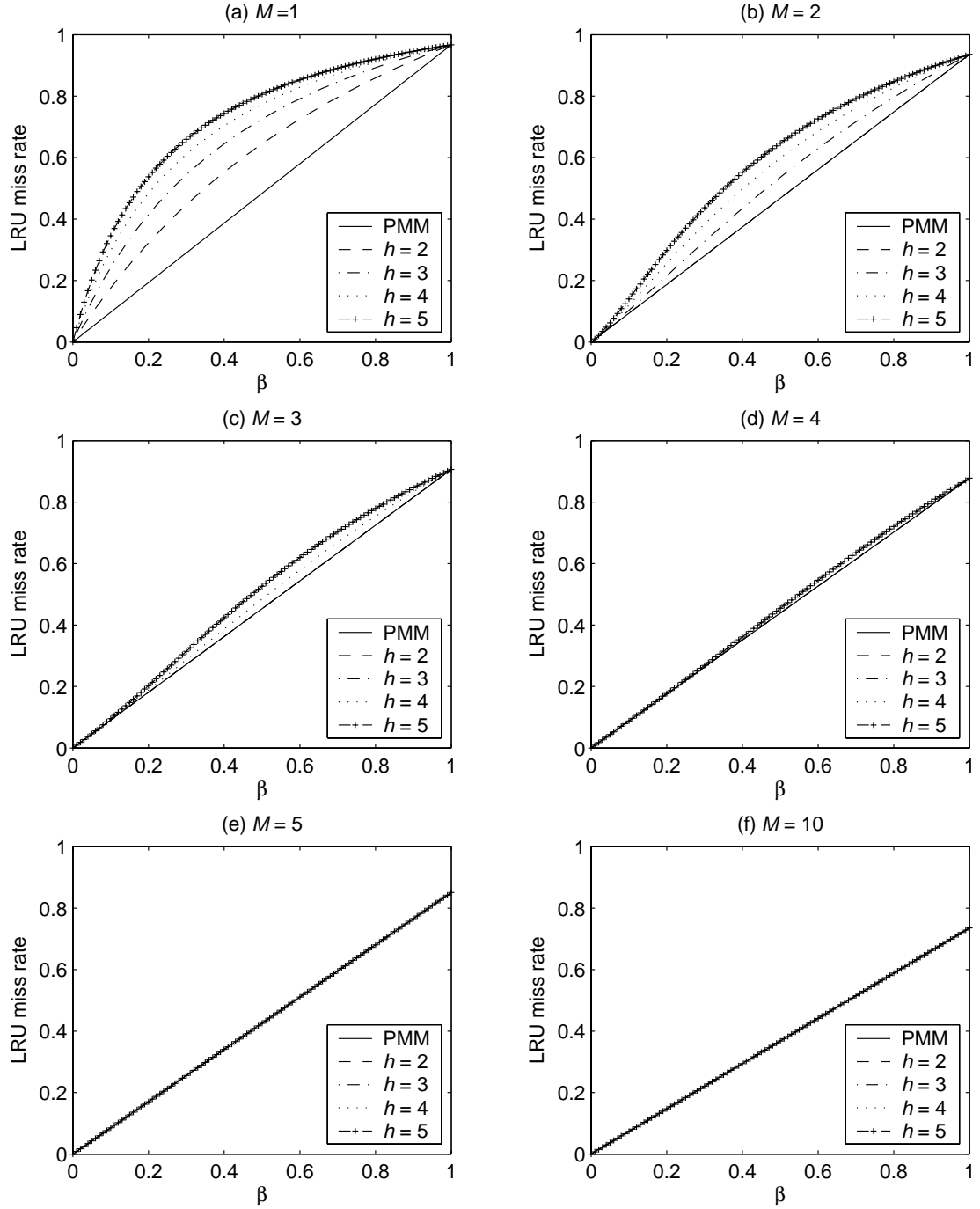
Figure 9.1: LRU miss rates for various cache sizes M when the input to the cache is the HOMM$(h, \boldsymbol{\alpha}_h(\beta), \boldsymbol{p}_{0.8})$ with $\boldsymbol{\alpha}_h(\beta) = (\beta, \frac{1-\beta}{h}, \ldots, \frac{1-\beta}{h})$

form of insensitivity of the LRU miss rate under the HOMM to the history window size h and to the pmf $\boldsymbol{\alpha}$. Lastly, for all cases and for all cache sizes, the miss rate always goes to 0 as β goes to 0. This is due to the fact that $\lim_{t \to \infty} \mathbf{P}\left[R_t^0 = R_{t-1}^0 \right] = 1$ where $\boldsymbol{R}^0$ denotes the HOMM$(h, \boldsymbol{\alpha}_h(0), \boldsymbol{p}_\alpha)$.[1]

9.5.3 LRUSM

According to Theorem 9.7, the stationary LRUSM($\boldsymbol{a}$) with stack distance pmf $\boldsymbol{a}$ satisfying condition (9.22) has stronger strength of temporal correlations than the stationary LRUSM($\boldsymbol{u}$). In the vein of Theorem 9.4, it is then natural to wonder when does the LRUSM($\boldsymbol{b}$) have weaker temporal correlations than the LRUSM($\boldsymbol{a}$) for pmf $\boldsymbol{b}$ not necessarily uniform. Theorem 9.7 suggests that this could happen when the pmf $\boldsymbol{a}$ is more skewed toward the smaller values of stack distance than the pmf $\boldsymbol{b}$. To capture the skewness in the pmf vectors, we recall the notion of majorization introduced in Chapter 2 and note that for any pmf $\boldsymbol{a}$ on $\mathcal{N}$, it holds that $\boldsymbol{u} \prec \boldsymbol{a}$. With majorization, we can now state the following conjecture.

Conjecture 9.10 *Consider request streams $\boldsymbol{R}^a$ and $\boldsymbol{R}^b$ which are modeled according to the stationary LRUSM($\boldsymbol{a}$) and LRUSM($\boldsymbol{b}$), respectively. If both pmfs $\boldsymbol{a}$ and $\boldsymbol{b}$ satisfy (9.22) with $\boldsymbol{b} \prec \boldsymbol{a}$, then the comparison $\boldsymbol{R}^b \leq_{TC} \boldsymbol{R}^a$ holds.*

When both pmfs $\boldsymbol{a}$ and $\boldsymbol{b}$ satisfy (9.22), the conditions (2.1)-(2.2) for the majorization comparison $\boldsymbol{b} \prec \boldsymbol{a}$ to hold reduce to

$$\sum_{i=1}^{n} b_i \leq \sum_{i=1}^{n} a_i, \quad n = 1, \ldots, N - 1. \tag{9.29}$$

---

[1]Indeed, if $\boldsymbol{R}$ is modeled according to the HOMM$(h, \boldsymbol{\alpha}, \boldsymbol{p}_\alpha)$ with $\beta = 0$, then it can be shown that $\lim_{t \to \infty} \mathbf{P}\left[R_t = R_{t-1} \right] = 1$ provided that the hth-order Markov chain $\{R_t, t = 0, 1, \ldots\}$ is aperiodic.

This condition is a possible formalization of the statement that the pmf a is more skewed toward the smaller values of stack distance than the pmf b.[2]

To glean evidence in favor of Conjecture 9.10, we consider the LRU policy and recall that the miss rate under the LRU policy with cache size M for the LRUSM(a) is given by (9.19). Combining (9.19) and (9.29), we conclude that for two LRUSM request streams R^a and R^b satisfying the conditions of Conjecture 9.10, it holds that $M_{\mathrm{LRU}}(R^a) \leq M_{\mathrm{LRU}}(R^b)$. This is of course the desired inequality expressing the folk theorem for miss rates under the LRU policy which would be expected if Conjecture 9.10 were to hold.

---

[2]The condition (9.29) is equivalent to the usual stochastic ordering between the stack distance rvs D_t^a and D_t^b associated with the request streams R^a and R^b, respectively, where $D_t^a \leq_{st} D_t^b$.

Chapter 10

The Working Set Model

In the last two chapters, we show how comparisons in the majorization ordering of popularity and in the TC ordering of temporal correlations can be translated into comparisons of some well-known metrics, namely, the working set size, the inter-reference time and the stack distance. In this chapter, we discuss results for the working set model and some folk theorems under its companion memory management policy, the so-called Working Set algorithm.

10.1 Definition

The working set model was introduced by Denning [26] and some of its properties are discussed in [27]. It can be defined as follows: Consider a request stream $\boldsymbol{R} = \{R_t, t = 0, 1, \ldots\}$. Fix $t = 0, 1, \ldots$. For each $\tau = 1, 2, \ldots$, we define the working set $W(t, \tau; \boldsymbol{R})$ of length τ at time t to be the set of *distinct* documents occurring amongst the past τ consecutive requests $R_{(t-\tau+1)^+}, \ldots, R_t$.[1] The size of the working set $W(t, \tau; \boldsymbol{R})$ is denoted by $S(t, \tau; \boldsymbol{R})$. Under some appropriate conditions on the request stream $\boldsymbol{R}$, it holds that $S(t, \tau; \boldsymbol{R}) \Longrightarrow_t S(\tau; \boldsymbol{R})$ where $S(\tau; \boldsymbol{R})$ is the steady state working set size

---

[1]For any $x \in \mathbb{R}$, we set $(x)^+ = \max(0, x)$.

of length τ. The rv $S(\tau; \boldsymbol{R})$ can be viewed as the number of distinct documents in τ consecutive requests in the steady state.

A basic quantity of interest associated with the working set size is its long-run average defined by

$$\hat{S}(\tau; \boldsymbol{R}) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} S(t, \tau; \boldsymbol{R}) \quad a.s. \tag{10.1}$$

for each $\tau = 1, 2, \ldots$. In the next lemma, we identify conditions on the request stream $\boldsymbol{R}$ for the existence of these limits (10.1), in the process making a connection between the limits (10.1) and the steady state working set sizes.

Lemma 10.1 *Assume the request stream $\boldsymbol{R} = \{R_t, t = 0, 1, \ldots\}$ to couple with a stationary sequence of $\mathcal{N}$-valued rvs $\tilde{\boldsymbol{R}} = \{\tilde{R}_t, t = 0, 1, \ldots\}$. Then, the a.s. limits (10.1) exist and it holds that*[2]

$$S(t, \tau; \boldsymbol{R}) \Longrightarrow_t S(\tau; \boldsymbol{R}), \quad \tau = 1, 2, \ldots. \tag{10.2}$$

If, in addition, the sequence $\tilde{\boldsymbol{R}}$ is ergodic, then

$$\hat{S}(\tau; \boldsymbol{R}) = \mathbf{E}\left[S(\tau; \boldsymbol{R})\right], \quad \tau = 1, 2, \ldots. \tag{10.3}$$

A proof of Lemma 10.1 can be found in Appendix E.1. A special case of Lemma 10.1 occurs when the request stream $\boldsymbol{R}$ itself is stationary. In that case, the distribution of $S(t, \tau; \boldsymbol{R})$ does not depend on t when $t \geq \tau - 1$, i.e., for each $\tau = 1, 2, \ldots$, we have

$$S(t, \tau; \boldsymbol{R}) =_{st} S(\tau - 1, \tau; \boldsymbol{R}), \quad t = \tau, \tau + 1, \ldots. \tag{10.4}$$

Therefore, (10.2) automatically holds. Furthermore, if the request stream $\boldsymbol{R}$ is stationary and ergodic, then (10.3) is also obtained.

---

[2]In fact, (10.2) holds under the weaker assumption that the request stream $\boldsymbol{R} = \{R_t, t = 0, 1, \ldots\}$ is asymptotically stationary in that $\{R_{t+\ell}, t = 0, 1, \ldots\} \Longrightarrow_\ell \{\tilde{R}_t, t = 0, 1, \ldots\}$ with $\tilde{\boldsymbol{R}} = \{\tilde{R}_t, t = 0, 1, \ldots\}$ being a stationary sequence of $\mathcal{N}$-valued rvs.

10.2 The effect of popularity

Assume the request stream $\mathbf{R} = \{R_t, t = 0, 1, \ldots\}$ to be the IRM with popularity pmf $\mathbf{p}$. Under these enforced i.i.d. assumption, the request stream $\mathbf{R}$ is stationary and ergodic, and from (10.4), we obtain

$$S(\tau; \mathbf{R}) =_{st} S(\tau - 1, \tau; \mathbf{R}) = |\{R_0, \ldots, R_{\tau-1}\}|. \tag{10.5}$$

Since the IRM request stream $\mathbf{R}$ is characterized solely by its popularity pmf $\mathbf{p}$, the pmf of $S(\tau; \mathbf{R})$ clearly depends only on the pmf $\mathbf{p}$ and we shall recognize this fact by denoting the working set size of length τ of the IRM by $S(\tau; \mathbf{p})$. Similarly, we let $\hat{S}(\tau; \mathbf{p})$ denote the average working set size (10.1) of length τ of the IRM request stream.

For positive integer $n = 1, 2, \ldots$ and pmf $\boldsymbol{\theta} = (\theta(1), \ldots, \theta(N))$ on $\{1, \ldots, N\}$, 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

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

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

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

115

this rv records the number of *distinct* outcomes that occur amongst the n trials. The following result was established by Wong and Yue [72] and deals with the Schur-concavity of the tails probabilities

$$\pi_\ell(n, \boldsymbol{\theta}) := \mathbf{P}\left[K(n, \boldsymbol{\theta}) > \ell\right], \quad \ell = 1, 2, \ldots, \min(N, n).$$

Theorem 10.2 *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.*

From (10.5), 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, by combining Theorem 10.2 with the basic fact (3.2) on the usual stochastic ordering, we get the following corollary.

Corollary 10.3 *For admissible pmfs* $\boldsymbol{p}$ *and* $\boldsymbol{q}$ *on* $\mathcal{N}$, *it holds that*

$$S(\tau; \boldsymbol{q}) \leq_{st} S(\tau; \boldsymbol{p}), \quad \tau = 1, 2, \ldots, \tag{10.6}$$

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!

A simple consequence of Corollary 10.3 is the comparisons of the average working set sizes, namely

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

provided $\boldsymbol{p} \prec \boldsymbol{q}$. This is due to the facts that the comparisons (10.6) imply

$$\mathbf{E}\left[S(\tau; \boldsymbol{q})\right] \leq \mathbf{E}\left[S(\tau; \boldsymbol{p})\right], \quad \tau = 1, 2, \ldots,$$

and that under the IRM, Lemma 10.1 yields $\hat{S}(\tau; \boldsymbol{p}) = \mathbf{E}\left[S(\tau; \boldsymbol{p})\right]$ for all $\tau = 1, 2, \ldots$.

10.3 The effect of temporal correlations

As for popularity, it is expected that the stronger the strength of temporal correlations in the stream of requests, the smaller the working set size. We wish to formalize this statement as was done for popularity in Corollary 10.3. However, with the help of the TC ordering, we obtain only the comparison of the expectations of the working set sizes.

Theorem 10.4 *For two request streams $\boldsymbol{R}^1 = \{R_t^1, t = 0, 1, \ldots\}$ and $\boldsymbol{R}^2 = \{R_t^2, t = 0, 1, \ldots\}$, if $\boldsymbol{R}^1 \leq_{TC} \boldsymbol{R}^2$, then for each $t = 0, 1, \ldots$, it holds that*

$$\mathbf{E}\left[S(t, \tau; \boldsymbol{R}^2)\right] \leq \mathbf{E}\left[S(t, \tau; \boldsymbol{R}^1)\right], \quad \tau = 1, 2, \ldots. \tag{10.7}$$

A proof of this theorem relies on the fact that the rv $S(t, \tau; \boldsymbol{R})$ can be expressed as a combination of supermodular functions of the indicator sequences $\{V_t(i), t = 0, 1, \ldots\}$, $i = 1, \ldots, N$, associated with the request stream $\boldsymbol{R}$. Before giving a proof, we note the following lemma [7, Lemma 2.1].

Lemma 10.5 *If the mapping $\psi : \mathbb{R}^\tau \to \mathbb{R}$ is given by*

$$\psi(\boldsymbol{x}) = \prod_{i=1}^{\tau} \psi^\star(x_i), \quad \boldsymbol{x} = (x_1, \ldots, x_\tau) \in \mathbb{R}^\tau \tag{10.8}$$

for some monotone mapping $\psi^\star : \mathbb{R} \to \mathbb{R}$, then ψ is supermodular.

Proof of Theorem 10.4. Fix $t = 0, 1, \ldots$ and $\tau = 1, \ldots, t + 1$. The working set size $S(t, \tau; \boldsymbol{R})$ of length τ at time t for the request stream $\boldsymbol{R}$ can be expressed in terms of the corresponding indicator sequences $\{V_t(i), t = 0, 1, \ldots\}$, $i = 1, \ldots, N$, as follows: From the definition of $S(t, \tau; \boldsymbol{R})$, we can write

$$S(t, \tau; \boldsymbol{R}) = \sum_{i=1}^{N} \mathbf{1}\left[i \in \{R_{(t-\tau+1)^+}, \ldots, R_t\}\right]$$

117

$$
\begin{aligned}
&= \sum_{i=1}^{N} \mathbf{1}\left[i \in \{R_{t-\tau+1}, \ldots, R_t\}\right] \\
&= \sum_{i=1}^{N}(1 - \mathbf{1}\left[i \notin \{R_{t-\tau+1}, \ldots, R_t\}\right]) \\
&= \sum_{i=1}^{N}(1 - \prod_{\ell=0}^{\tau-1} \mathbf{1}\left[R_{t-\ell} \neq i\right]) \\
&= \sum_{i=1}^{N}(1 - \prod_{\ell=0}^{\tau-1}(1 - \mathbf{1}\left[R_{t-\ell} = i\right])) \\
&= \sum_{i=1}^{N}(1 - \prod_{\ell=0}^{\tau-1}(1 - V_{t-\ell}(i))). \\
&= \sum_{i=1}^{N}(1 - \psi(V_{t-\tau+1}(i), \ldots, V_t(i))) \quad\quad (10.9)
\end{aligned}
$$

where the mapping $\psi : \mathbb{R}^\tau \to \mathbb{R}$ is of the form (10.8) with mapping $\psi^\star : \mathbb{R} \to \mathbb{R}$ given by

$$
\psi^\star(x) = 1 - x, \quad x \in \mathbb{R}. \quad\quad (10.10)
$$

By Lemma 10.5, the mapping ψ is supermodular since $\psi^\star$ defined at (10.10) is monotone.

Equipped with the expressions (10.8)-(10.10), we are now ready to prove Theorem 10.4. Recall that for any two request streams $\boldsymbol{R}^1$ and $\boldsymbol{R}^2$ such that $\boldsymbol{R}^1 \leq_{TC} \boldsymbol{R}^2$, we have the comparison $\{V_t^1(i), t = 0, 1, \ldots\} \leq_{sm} \{V_t^2(i), t = 0, 1, \ldots\}$ for each $i = 1, \ldots, N$. From the supermodularity of ψ and the definition of the sm ordering, it then follows that

$$
\mathbf{E}\left[\psi(V_{t-\tau+1}^1(i), \ldots, V_t^1(i))\right] \leq \mathbf{E}\left[\psi(V_{t-\tau+1}^2(i), \ldots, V_t^2(i))\right] \quad\quad (10.11)
$$

for all $i = 1, \ldots, N$. Combining inequalities (10.11) with (10.9) yields the comparison (10.7) for each $\tau = 1, \ldots, t+1$. Upon noting that for all $\tau > t+1$,

$$
S(t, \tau; \boldsymbol{R}^k) = S(t, t+1; \boldsymbol{R}^k), \quad k = 1, 2,
$$

we get the desired comparisons (10.7) for all $\tau = 1, 2, \ldots$. ∎

Corollary 10.6 *Assume that for each $k = 1, 2$, the request stream $\boldsymbol{R}^k = \{R_t^k, t = 0, 1, \ldots\}$ couples with a stationary sequence of $\mathcal{N}$-valued rvs $\tilde{\boldsymbol{R}}^k = \{\tilde{R}_t^k, t = 0, 1, \ldots\}$. If $\boldsymbol{R}^1 \leq_{TC} \boldsymbol{R}^2$, then it holds that*

$$\mathbf{E}\left[S(\tau; \boldsymbol{R}^2)\right] \leq \mathbf{E}\left[S(\tau; \boldsymbol{R}^1)\right], \quad \tau = 1, 2, \ldots, \tag{10.12}$$

where for each $k = 1, 2$, $S(\tau; \boldsymbol{R}^k)$ is the steady state working set size of the request stream $\boldsymbol{R}^k$. In addition, if $\tilde{\boldsymbol{R}}^1$ and $\tilde{\boldsymbol{R}}^2$ are stationary and ergodic, then it holds that

$$\hat{S}(\tau; \boldsymbol{R}^2) \leq \hat{S}(\tau; \boldsymbol{R}^1), \quad \tau = 1, 2, \ldots, \tag{10.13}$$

where for each $k = 1, 2$, $\hat{S}(\tau; \boldsymbol{R}^k)$ is the average working set size of the request stream $\boldsymbol{R}^k$.

Proof. Fix $\tau = 1, 2, \ldots$ and $k = 1, 2$. Under the assumptions above, Lemma 10.1 already yields the convergence

$$S(t, \tau; \boldsymbol{R}^k) \Longrightarrow_t S(\tau; \boldsymbol{R}^k). \tag{10.14}$$

Next, because $S(t, \tau; \boldsymbol{R}^k) \leq N$ for every $t = 0, 1, \ldots$, the sequence $\{S(t, \tau; \boldsymbol{R}^k), t = 0, 1, \ldots\}$ is uniformly integrable. Combining this fact with (10.14), it follows from [11, Thm. 5.4, p. 32] that

$$\lim_{t \to \infty} \mathbf{E}\left[S(t, \tau; \boldsymbol{R}^k)\right] = \mathbf{E}\left[S(\tau; \boldsymbol{R}^k)\right]. \tag{10.15}$$

Invoking (10.7) and (10.15), we obtain the steady state comparisons (10.12). The comparisons (10.13) for the average working set sizes follow from (10.12) under the additional ergodicity assumption of the coupling processes associated with $\boldsymbol{R}^1$ and $\boldsymbol{R}^2$.

∎

Corollary 10.6 demonstrates that for a request stream $\boldsymbol{R}$ exhibiting temporal correlations, the independent version $\hat{\boldsymbol{R}}$ of $\boldsymbol{R}$ can be used to provide various performance bounds, which in turn can be used for cache dimensioning associated with the request stream $\boldsymbol{R}$. We illustrate this argument with three request models, namely the HOMM, PMM and LRUSM request streams, with the help of Theorems 9.3, 9.4 and 9.7, respectively. Upon noting that the stationary HOMM and PMM are ergodic Markov chains, we obtain

Corollary 10.7 *Assume the request stream* $\boldsymbol{R} = \{R_t, t = 0, 1, \ldots\}$ *to be modeled according to the stationary HOMM($h, \boldsymbol{\alpha}, \boldsymbol{p}$) with admissible popularity pmf $\boldsymbol{p}$. Then, it holds that*

$$\hat{S}(\tau; \boldsymbol{R}) \leq \hat{S}(\tau; \hat{\boldsymbol{R}}), \quad \tau = 1, 2, \ldots,$$

where $\hat{\boldsymbol{R}}$ is the IRM with popularity pmf $\boldsymbol{p}$.

Corollary 10.8 *Assume that for each $k = 1, 2$, the request stream $\boldsymbol{R}^{\beta_k} = \{R_t^{\beta_k}, t = 0, 1, \ldots\}$ is modeled according to the stationary PMM($\beta_k, \boldsymbol{p}$) with admissible popularity pmf $\boldsymbol{p}$. If $0 < \beta_2 \leq \beta_1$, then it holds that*

$$\hat{S}(\tau; \boldsymbol{R}^{\beta_2}) \leq \hat{S}(\tau; \boldsymbol{R}^{\beta_1}), \quad \tau = 1, 2, \ldots.$$

Lastly, we note the comparison of the working set sizes under the LRUSM.

Corollary 10.9 *Assume the request stream $\boldsymbol{R}^{\boldsymbol{a}} = \{R_t^{\boldsymbol{a}}, t = 0, 1, \ldots\}$ to be modeled according to the stationary LRUSM($\boldsymbol{a}$) with stack distance pmf $\boldsymbol{a}$ satisfying (9.22). Then, it holds that*

$$\mathbf{E}\left[S(\tau; \boldsymbol{R}^{\boldsymbol{a}})\right] \leq \mathbf{E}\left[S(\tau; \hat{\boldsymbol{R}}^{\boldsymbol{a}})\right], \quad \tau = 1, 2, \ldots,$$

where $\hat{\boldsymbol{R}}^{\boldsymbol{a}}$ is the IRM with uniform popularity pmf $\boldsymbol{u}$.

10.4 The Working Set algorithm

Fix $\tau = 1, 2, \ldots$. The Working Set (WS) algorithm with length τ is the algorithm that maintains the previous τ consecutive requested documents $R_{(t-\tau)^+}, \ldots, R_{t-1}$ in the cache S_t at time t. In other words, the cache S_t is simply the working set $W(t-1, \tau; \boldsymbol{R})$ with the convention $W(-1, \tau; \boldsymbol{R}) = \phi$. This algorithm differs from other demand-driven caching policies in that the number of documents in the cache may change over time while demand-driven caching policies have a fixed cache size M (as soon as each document has been called at least once). The number of documents in the cache at time t under the WS algorithm is basically the number of distinct documents in $W(t-1, \tau; \boldsymbol{R})$ which is the working set size $S(t-1, \tau; \boldsymbol{R})$.

The operation of the WS algorithm can be described as follows: For each $t = 0, 1, \ldots$, let Ω_t be the state of the cache at time t defined by

$$\Omega_t = (R_{(t-\tau)^+}, \ldots, R_{t-1}).$$

It is easy to see from this definition that the cache state Ω_{t+1} is completely determined by the previous cache state Ω_t and the current request R_t. Furthermore, the cache set S_t can be recovered from Ω_t by taking

$$S_t = \{i = 1, \ldots, N : i \in \Omega_t\} = W(t-1, \tau; \boldsymbol{R}), \quad t = 0, 1, \ldots.$$

For $t \geq \tau$, regardless of a cache miss, the WS algorithm will evict the document $R_{t-\tau}$ if $R_{t-\tau} \notin W(t, \tau; \boldsymbol{R})$ and does not evict any document, otherwise.

The miss rate of the WS algorithm with length τ can be defined in the same way as in the case of demand-driven caching; it is given by the a.s. limit

$$
\begin{aligned}
M_{\mathrm{WS}}(\boldsymbol{R}) &= \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\left[R_t \notin S_t\right] \quad a.s. \\
&= \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\left[R_t \notin W(t-1, \tau; \boldsymbol{R})\right] \quad a.s. \quad (10.16)
\end{aligned}
$$

We next explore the folk theorems for miss rates and for output streams under the WS algorithm. We do so for both the IRM input stream and general input stream exhibiting temporal correlations, respectively.

10.4.1   Under the IRM

We first assume the input to the cache to be modeled according the IRM with popularity pmf p. Under this assumption, we show that the folk theorems for the miss rate and the output of a cache under the WS algorithm do not hold in general. This comes as no surprise since the WS algorithm is a close cousin of the LRU policy in that the LRU policy of cache size M can be obtained from the WS algorithm that keeps the M most recent distinct documents in the cache by varying its length τ.

Miss rate of WS algorithm

It is known [2, 27] that the miss rate $\hat{M}_{\mathrm{WS}}(p)$ of the WS algorithm with length τ under the IRM with popularity pmf p is given by

$$\hat{M}_{\mathrm{WS}}(p) = \sum_{i=1}^{N} p(i)(1 - p(i))^{\tau}. \tag{10.17}$$

Unfortunately, the miss rate function $\hat{M}_{\mathrm{WS}}(p)$ is not Schur-concave in p for $\tau = 2, 3, \ldots$. However, it is Schur-concave only when $\tau = 1$ in which case the WS algorithm coincides with any demand-driven caching policy of cache size $M = 1$. These results are contained in

Theorem 10.10 *Assume the input to be modeled according to the IRM with popularity pmf p. The miss rate function $\hat{M}_{\mathrm{WS}}(p)$ under the WS algorithm with length τ is Schur-concave in the pmf p when $\tau = 1$ and is not Schur-concave in the pmf p when $\tau = 2, 3, \ldots$.*

Proof. For each $\tau = 1, 2, \ldots$, the miss rate function $\hat{M}_{\text{WS}}(\boldsymbol{p})$ in (10.17) is of the form

$$\hat{M}_{\text{WS}}(\boldsymbol{p}) = \sum_{i=1}^{N} g_\tau(p(i))$$

where the mapping $g_\tau : [0,1] \to [0, 0.25]$ is given by $x \to x(1-x)^\tau$. As we note from [49, 3.C.1, p. 64 and 3.C.1.c, p. 67], the function $\hat{M}_{\text{WS}}(\boldsymbol{p})$ is Schur-concave if and only if the mapping g_τ is concave. It is now a simple matter to check that the mapping g_τ is concave only when $\tau = 1$ and *not* concave when $\tau = 2, 3, \ldots$, whence the desired result. ■

Output of WS algorithm

By restricting the input streams to be in the class of IRM, the output of the WS algorithm with length τ can be analyzed along the same lines as Theorem 5.2 for demand-driven caching policies. Indeed, for the IRM with popularity pmf $\boldsymbol{p}$, the output popularity pmf $\boldsymbol{p}_{\text{WS}}^\star$ under the WS algorithm with length τ is given by

$$p_{\text{WS}}^\star(i) = \frac{p(i)(1-p(i))^\tau}{\sum_{j=1}^{N} p(j)(1-p(j))^\tau}, \quad i = 1, \ldots, N. \tag{10.18}$$

As for the case of miss rate, the folk theorem for the output that $\boldsymbol{p}_{\text{WS}}^\star \prec \boldsymbol{p}$ does not hold when $\tau = 2, 3, \ldots$, but does hold only for $\tau = 1$ in which case the WS algorithm reduces to any demand-driven caching policy with cache size $M = 1$. The counterexamples when $\tau = 2, 3, \ldots$, are given below where the IRM input has a Zipf-like popularity pmf with large α.

Theorem 10.11 *Assume the input to be modeled according to the IRM with Zipf-like popularity pmf $\boldsymbol{p}_\alpha$ for some $\alpha \geq 0$. If the number of documents N and the length τ of*

the WS algorithm satisfy the condition

$$N < 2^{\tau-1} \quad with \quad \tau > 1, \tag{10.19}$$

then under the WS algorithm, there exists $\alpha^\star = \alpha^\star(\tau, N)$ *such that* $\boldsymbol{p}^\star_{\mathrm{WS},\alpha} \prec \boldsymbol{p}_\alpha$ *does not hold for* $\alpha > \alpha^\star$.

A proof of this theorem is given in Appendix B.5.

10.4.2 Miss rate under input with temporal correlations

Given an input stream $\boldsymbol{R} = \{R_t, t = 0, 1, \ldots\}$, let $\{V_t(i), t = 0, 1, \ldots\}$, $i = 1, \ldots, N$, be the indicator sequences (9.1) associated with it. Recall from (10.16) that a miss occurs at time t when the document R_t is not in the working set $W(t - 1, \tau; \boldsymbol{R})$. Thus, the indicator function for the miss event at time $t \geq \tau$ can be written as

$$
\begin{aligned}
\mathbf{1}\left[R_t \notin W(t - 1, \tau; \boldsymbol{R})\right] &= \mathbf{1}\left[R_t \notin \{R_{t-\tau}, \ldots, R_{t-1}\}\right] \\
&= \sum_{i=1}^{N} \mathbf{1}\left[R_t = i\right] \mathbf{1}\left[i \notin \{R_{t-\tau}, \ldots, R_{t-1}\}\right] \quad (10.20) \\
&= \sum_{i=1}^{N} \mathbf{1}\left[R_t = i\right] \prod_{\ell=1}^{\tau} \mathbf{1}\left[R_{t-\ell} \neq i\right] \\
&= \sum_{i=1}^{N} V_t(i) \prod_{\ell=1}^{\tau} (1 - V_{t-\ell}(i)) \\
&= \sum_{i=1}^{N} g(V_{t-\tau}(i), \ldots, V_t(i)) \quad (10.21)
\end{aligned}
$$

where we have set

$$g(x_0, \ldots, x_\tau) = x_\tau \prod_{\ell=0}^{\tau-1} (1 - x_\ell), \quad (x_0, \ldots, x_\tau) \in \mathbb{R}^{\tau+1}. \tag{10.22}$$

Combining (10.16), (10.21) and (10.22) yields the miss rate under the WS algorithm as the limit

$$
M_{\mathrm{WS}}(\boldsymbol{R}) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{\tau-1} \mathbf{1}\left[R_t \notin W(t - 1, \tau; \boldsymbol{R})\right]
$$

$$+ \lim_{T \to \infty} \left(\frac{T - \tau + 1}{T} \right) \frac{1}{T - \tau + 1} \sum_{t=\tau}^{T} \sum_{i=1}^{N} g(V_{t-\tau}(i), \ldots, V_t(i))$$

$$= \lim_{T \to \infty} \frac{1}{T} \sum_{t=\tau}^{T+\tau-1} \sum_{i=1}^{N} g(V_{t-\tau}(i), \ldots, V_t(i)) \quad a.s. \tag{10.23}$$

and if the request stream $\boldsymbol{R}$ admits some form of ergodicity, then the limit (10.23) exists. A condition for the existence of the limit (10.23) is given in the next lemma whose proof is available in Appendix E.2.

Lemma 10.12 *Fix $\tau = 1, 2, \ldots$. Assume the request stream $\boldsymbol{R} = \{R_t, t = 0, 1, \ldots\}$ to couple with a stationary and ergodic sequence of $\mathcal{N}$-valued rvs $\tilde{\boldsymbol{R}} = \{\tilde{R}_t, t = 0, 1, \ldots\}$. Then, the a.s. limit (10.23) exists and is given by*

$$M_{\mathrm{WS}}(\boldsymbol{R}) = \lim_{t \to \infty} \sum_{i=1}^{N} \mathbf{E}\left[g(V_{t-\tau}(i), \ldots, V_t(i))\right] \quad a.s. \tag{10.24}$$

In particular, if $\boldsymbol{R}$ is stationary and ergodic, then

$$M_{\mathrm{WS}}(\boldsymbol{R}) = \sum_{i=1}^{N} \mathbf{P}\left[R_\tau = i, R_\ell \neq i, \ell = 0, \ldots, \tau - 1\right]. \tag{10.25}$$

To establish the folk theorem to the effect that the stronger the temporal correlations, the smaller the miss rate, we need to show that

$$M_{\mathrm{WS}}(\boldsymbol{R}^2) \leq M_{\mathrm{WS}}(\boldsymbol{R}^1) \quad \text{whenever} \quad \boldsymbol{R}^1 \leq_{TC} \boldsymbol{R}^2. \tag{10.26}$$

Therefore, upon recalling the definitions of the TC and sm orderings, we see that establishing (10.26) amounts to showing that the mapping g given in (10.22) is submodular.[3] Unfortunately, the mapping g is *not* submodular in general; only in the special case $\tau = 1$ is g a submodular function. We shall discuss these issues by first showing the

---

[3]A function $\varphi : \mathbb{R}^n \to \mathbb{R}$ is said to be submodular if $-\varphi$ is supermodular.

positive result when $\tau = 1$ and then providing counterexamples using the PMM when $\tau > 1$.

$[\tau = 1]$ – When $\tau = 1$, we note that $S(t - 1, \tau; \boldsymbol{R}) = 1$ for all $t = 1, 2, \ldots$, and the WS algorithm coincides with *any* demand-driven caching policy having cache size $M = 1$. In that case, the only document in the cache at time t is the document R_{t-1} and a miss occurs when $R_t \neq R_{t-1}$. The folk theorem holds in this special case for all demand-driven caching policies.

Theorem 10.13 *Consider an arbitrary demand-driven replacement policy π with $M = 1$. If the request streams $\boldsymbol{R}^1$ and $\boldsymbol{R}^2$ satisfy the relation $\boldsymbol{R}^1 \leq_{TC} \boldsymbol{R}^2$, then it holds that*

$$\mathbf{P}\left[R_t^2 \notin S_t^2\right] \leq \mathbf{P}\left[R_t^1 \notin S_t^1\right], \quad t = 1, 2, \ldots. \tag{10.27}$$

Proof. For each $t = 1, 2, \ldots$, we have from (10.21)-(10.22) that

$$
\begin{aligned}
\mathbf{1}\left[R_t \notin S_t\right] &= \mathbf{1}\left[R_t \neq R_{t-1}\right] \\
&= \sum_{i=1}^{N} g(V_{t-1}(i), V_t(i))
\end{aligned}
$$

with the mapping $g : \mathbb{R}^2 \to \mathbb{R}$ being given by

$$g(x_0, x_1) = x_1 - x_0 x_1, \quad (x_0, x_1) \in \mathbb{R}^2.$$

Because the mapping $(x_0, x_1) \to x_0 x_1$ is supermodular, the mapping $(x_0, x_1) \to -x_0 x_1$ is submodular. The mapping $(x_0, x_1) \to x_1$ being submodular, the mapping g is therefore submodular since the sum of two submodular functions is still a submodular function.

Given two request streams $\boldsymbol{R}^1$ and $\boldsymbol{R}^2$ such that $\boldsymbol{R}^1 \leq_{TC} \boldsymbol{R}^2$, we recall the comparisons $\{V_t^1(i), t = 0, 1, \ldots\} \leq_{sm} \{V_t^2(i), t = 0, 1, \ldots\}$ for each $i = 1, \ldots, N$. Thus by

126

the definition of the sm ordering, we obtain for each $t = 1, 2, \ldots$,

$$
\begin{aligned}
\mathbf{P}\left[R_t^2 \notin S_t^2\right] &= \sum_{i=1}^{N} \mathbf{E}\left[g(V_{t-1}^2(i), V_t^2(i))\right] \\
&\leq \sum_{i=1}^{N} \mathbf{E}\left[g(V_{t-1}^1(i), V_t^1(i))\right] \\
&= \mathbf{P}\left[R_t^1 \notin S_t^1\right].
\end{aligned}
$$

∎

Corollary 10.14 *Consider an arbitrary demand-driven replacement policy π with $M = 1$. If the request streams $\boldsymbol{R}^1$ and $\boldsymbol{R}^2$ couple with stationary and ergodic sequences of $\mathcal{N}$-valued rvs $\tilde{\boldsymbol{R}}^1$ and $\tilde{\boldsymbol{R}}^2$, respectively, and satisfy the relation $\boldsymbol{R}^1 \leq_{TC} \boldsymbol{R}^2$, then it holds that*

$$
M_{\mathrm{WS}}(\boldsymbol{R}^2) \leq M_{\mathrm{WS}}(\boldsymbol{R}^1).
$$

Proof. Under the assumptions above, the miss rate of the request stream $\boldsymbol{R}^k$ for each $k = 1, 2$, can be obtained using Lemma 10.12 and is given by

$$
M_{\mathrm{WS}}(\boldsymbol{R}^k) = \lim_{t \to \infty} \mathbf{P}\left[R_t^k \notin S_t^k\right] \quad a.s.
$$

The desired result is now immediate from (10.27). ∎

$[\tau > 1]$ – The folk theorem (10.26) does not necessarily hold when $\tau > 1$ as we now demonstrate via counterexamples when the PMM is taken to be the input to the cache.

The miss rate of the WS algorithm with length τ for PMM$(\beta, \boldsymbol{p})$ [2] is given by

$$
M_{\mathrm{WS}}(\beta, \boldsymbol{p}) = \beta \sum_{i=1}^{N} p(i)(1 - p(i))(1 - \beta p(i))^{\tau-1}. \tag{10.28}
$$

From Section 9.3, we would expect that as the strength of temporal correlations increases, i.e., the value of the parameter β decreases, the miss rate $M_{\mathrm{WS}}(\beta, \boldsymbol{p})$ should be decreasing. To put it differently, the mapping $\beta \to M_{\mathrm{WS}}(\beta, \boldsymbol{p})$ should be increasing when the popularity pmf $\boldsymbol{p}$ is held fixed.

However, this is not always the case as we show in the counterexamples where the PMM stream is assumed to have the uniform popularity pmf $\boldsymbol{u} = (\frac{1}{N}, \ldots, \frac{1}{N})$.

Theorem 10.15 *Fix $\tau = 2, 3, \ldots$, and assume the input to be modeled according to PMM($\beta, \boldsymbol{u}$). Under the WS algorithm with length τ, the miss rate function $M_{\mathrm{WS}}(\beta, \boldsymbol{u})$ given in (10.28) is increasing in β when $\beta \leq \frac{N}{\tau}$ and decreasing in β when $\beta > \frac{N}{\tau}$.*

Thus, the folk theorem always holds when the length τ of the WS algorithm is smaller than the number of documents N but may fail to hold otherwise.

Proof. When the PMM has the uniform popularity pmf $\boldsymbol{u}$, the expression (10.28) for the miss rate under the WS algorithm becomes

$$M_{\mathrm{WS}}(\beta, \boldsymbol{u}) = \beta \left(1 - \frac{1}{N}\right) \left(1 - \frac{\beta}{N}\right)^{\tau - 1}.$$

Differentiating this expression with respect to β yields

$$\frac{d}{d\beta} M_{\mathrm{WS}}(\beta, \boldsymbol{u}) = \left(1 - \frac{1}{N}\right) \left(1 - \frac{\beta}{N}\right)^{\tau - 2} \left(1 - \frac{\tau \beta}{N}\right).$$

Thus, the miss rate function $M_{\mathrm{WS}}(\beta, \boldsymbol{u})$ is increasing when $1 - \frac{\tau\beta}{N} \geq 0$, or equivalently, $\beta \leq \frac{N}{\tau}$, and is decreasing when $1 - \frac{\tau\beta}{N} < 0$, or equivalently, $\beta > \frac{N}{\tau}$. ∎

Chapter 11

Inter-reference Time and Stack Distance

In this chapter, we continue the program announced in Chapter 10 as we seek the appropriate comparisons for the inter-reference times and the stack distances when the request streams are comparable in either the majorization or the TC orderings.

11.1   Inter-reference time

The notion of inter-reference time in the stream of requests has recently received some attention as a way of characterizing temporal correlations [34, 40, 53].

First a definition. Given a request stream $\boldsymbol{R} = \{R_t, t = 0, 1, \ldots\}$, for each $t = 0, 1, \ldots$, we define the inter-reference time $T(t; \boldsymbol{R})$ as the rv given by

$$T(t; \boldsymbol{R}) := \inf\{\tau = 1, 2, \ldots, t : R_t = R_{t-\tau}\} \tag{11.1}$$

with the convention that $T(t; \boldsymbol{R}) = t + 1$ if $R_{t-\tau} \neq R_t$ for all $\tau = 1, \ldots, t$. As for the working set size, under some appropriate conditions on the request stream $\boldsymbol{R}$, $T(t; \boldsymbol{R}) \Longrightarrow_t T(\boldsymbol{R})$ where the steady state inter-reference time $T(\boldsymbol{R})$ describes the time between two consecutive requests for the same document. One such condition is given in

Lemma 11.1 *Assume the request stream* $\boldsymbol{R} = \{R_t, t = 0, 1, \ldots\}$ *to be asymptoti-cally stationary, i.e.,* $\{R_{t+\ell}, t = 0, 1, \ldots\} \Longrightarrow_\ell \{\tilde{R}_t, t = 0, 1, \ldots\}$ *with* $\tilde{\boldsymbol{R}} = \{\tilde{R}_t, t = 0, 1, \ldots\}$ *being a stationary sequence of* $\mathcal{N}$*-valued rvs. Then, it holds that*

$$T(t; \boldsymbol{R}) \Longrightarrow_t T(\boldsymbol{R}). \tag{11.2}$$

A proof of Lemma 11.1 is given in Appendix E.3. Lastly, we note that if the request stream $\boldsymbol{R}$ is stationary and ergodic, then the pmf of the steady state inter-reference time $T(\boldsymbol{R})$ is given by the limits

$$\mathbf{P}\left[T(\boldsymbol{R}) = k\right] = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\left[T(t; \boldsymbol{R}) = k\right] \quad a.s., \quad k = 1, 2, \ldots.$$

11.1.1 The effect of popularity

We first study the effect of popularity on the inter-reference time by assuming the request stream $\boldsymbol{R}$ to be the IRM with popularity pmf $\boldsymbol{p}$. Under the IRM, the request stream $\boldsymbol{R}$ is stationary and ergodic in which case (11.2) holds. In fact, $T(\boldsymbol{R})$ can be represented by

$$T(\boldsymbol{R}) =_{st} \inf\{t = 1, 2, \ldots : R_t = R_0\} \tag{11.3}$$

since the i.i.d. process $\{R_t, t = 0, 1, \ldots\}$ is reversible. The main comparison for the steady state inter-reference times is given in terms of the convex ordering.

Theorem 11.2 *Assume that request streams* $\boldsymbol{R}^1$ *and* $\boldsymbol{R}^2$ *are modeled according to the IRM with admissible popularity pmfs* $\boldsymbol{p}^1$ *and* $\boldsymbol{p}^2$, *respectively. Then, it holds that*

$$T(\boldsymbol{R}^1) \leq_{cx} T(\boldsymbol{R}^2) \tag{11.4}$$

whenever $\boldsymbol{p}^1 \prec \boldsymbol{p}^2$.

Thus, the more skewed the popularity pmf, the stronger the locality of reference in the IRM, and the more variable the inter-reference time in that (11.4) implies $\mathbf{E}\left[T(\boldsymbol{R}^1)\right] = \mathbf{E}\left[T(\boldsymbol{R}^2)\right]$ and $Var(T(\boldsymbol{R}^1)) \le Var(T(\boldsymbol{R}^2))$. This can be explained by observing that a document with high probability of request is likely to be requested again in the near future, leading to smaller values for $T(\boldsymbol{R})$ and correspondingly larger deviation from its mean.

Proof. It is well known [59, Thm. 2.A.1, p. 57] that the comparison (11.4) between the $\{1, 2, \ldots\}$-valued rvs $T(\boldsymbol{R}^1)$ and $T(\boldsymbol{R}^2)$ is equivalent to

$$\sum_{\tau=n}^{\infty} \mathbf{P}\left[T(\boldsymbol{R}^1) > \tau\right] \le \sum_{\tau=n}^{\infty} \mathbf{P}\left[T(\boldsymbol{R}^2) > \tau\right] \tag{11.5}$$

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

$$\mathbf{E}\left[T(\boldsymbol{R}^1)\right] = \mathbf{E}\left[T(\boldsymbol{R}^2)\right]. \tag{11.6}$$

Consider an IRM request stream $\boldsymbol{R}$ with popularity pmf $\boldsymbol{p}$ and fix $i = 1, \ldots, N$. By using the representation (11.3), we note that

$$\mathbf{P}\left[T(\boldsymbol{R}) = \tau | R_0 = i\right] = p(i)(1 - p(i))^{\tau-1}, \quad \tau = 1, 2, \ldots,$$

i.e., conditional on $R_0 = i$, the inter-reference time $T(\boldsymbol{R})$ is geometrically distributed with parameter $p(i)$. Consequently, for each $n = 0, 1, \ldots$, we find

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

whence

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

Next, we obtain

$$\psi_n(\boldsymbol{p}) := \sum_{\tau=n}^{\infty} \mathbf{P}\left[T(\boldsymbol{p}) > \tau\right] = \sum_{i=1}^{N} (1 - p(i))^n, \quad n = 0, 1, \ldots.$$

131

In particular, with $n = 0$, this last calculation yields

$$\mathbf{E}\left[T(\boldsymbol{R})\right] = \sum_{\tau=0}^{\infty} \mathbf{P}\left[T(\boldsymbol{R}) > \tau\right] = N,$$

and this independently of $\boldsymbol{p}$! In other words, (11.6) holds.

It is a simple matter to see that for each $n = 1, 2, \ldots$, the mapping $t \rightarrow (1 - t)^n$ is convex on $\mathbb{R}_+$. By a classical result of Schur [49, C.1, p. 64], the mapping $\boldsymbol{x} \rightarrow \sum_{i=1}^{N}(1 - x_i)^n$ is a Schur-convex function on $\mathbb{R}_+^N$. To put it differently, the mapping $\boldsymbol{p} \rightarrow \psi_n(\boldsymbol{p})$ is Schur-convex, and (11.5) indeed holds when $\boldsymbol{p}^1 \prec \boldsymbol{p}^2$. ∎

11.1.2 The effect of temporal correlations

We now turn to the comparison (11.4) for the steady state inter-reference times when the request streams $\boldsymbol{R}^1$ and $\boldsymbol{R}^2$ are comparable in the TC ordering.

Theorem 11.3 *Assume that for each $k = 1, 2$, the request stream $\boldsymbol{R}^k$ is asymptotically stationary, i.e., $\{R_{t+\ell}^k, t = 0, 1, \ldots\} \Longrightarrow_\ell \{\tilde{R}_t^k, t = 0, 1, \ldots\}$ where $\tilde{\boldsymbol{R}}^k = \{\tilde{R}_t^k, t = 0, 1, \ldots\}$ is a stationary sequence of $\mathcal{N}$-valued rvs, and has admissible popularity pmf $\boldsymbol{p}^k$. If $\boldsymbol{R}^1 \leq_{TC} \boldsymbol{R}^2$, then the comparison (11.4) holds.*

Theorem 11.3 states that the stronger the temporal correlations, the more variable the inter-reference time! To establish Theorem 11.3, we shall rely on the following lemma whose proof is available in Appendix E.4.

Lemma 11.4 *Assume that the request stream $\boldsymbol{R} = \{R_t, t = 0, 1, \ldots\}$ is asymptotically stationary, i.e., $\{R_{t+\ell}, t = 0, 1, \ldots\} \Longrightarrow_\ell \{\tilde{R}_t, t = 0, 1, \ldots\}$ where $\tilde{\boldsymbol{R}} = \{\tilde{R}_t, t = 0, 1, \ldots\}$ is a stationary sequence of $\mathcal{N}$-valued rvs, and has admissible popularity pmf*

p. Then, it holds that

$$\sum_{\tau=n}^{\infty} \mathbf{P}\left[T(\boldsymbol{R}) > \tau\right] = \sum_{i=1}^{N} \mathbf{P}\left[\tilde{R}_{\ell} \neq i, \ell = 0, \ldots, n-1\right], \quad n = 1, 2, \ldots, \tag{11.7}$$

and

$$\mathbf{E}\left[T(\boldsymbol{R})\right] = \sum_{\tau=0}^{\infty} \mathbf{P}\left[T(\boldsymbol{R}) > \tau\right] = N. \tag{11.8}$$

Proof of Theorem 11.3. The proof of this theorem proceeds along lines similar to ones found in the proof of Theorem 11.2. The comparison (11.4) is established by showing that (11.5) and (11.6) hold whenever $\boldsymbol{R}^1 \leq_{TC} \boldsymbol{R}^2$.

Fix $k = 1, 2$. For each $i = 1, \ldots, N$, let $\{V_t^k(i), t = 0, 1, \ldots\}$ and $\{\tilde{V}_t^k(i), t = 0, 1, \ldots\}$ be the indicator sequences (9.1) associated with $\boldsymbol{R}^k$ and $\tilde{\boldsymbol{R}}^k$, respectively. From Lemma 11.4, the expression (11.7) for each $n = 1, 2, \ldots$, can be rewritten as

$$\begin{aligned}
\sum_{\tau=n}^{\infty} \mathbf{P}\left[T(\boldsymbol{R}^k) > \tau\right] &= \sum_{i=1}^{N} \mathbf{E}\left[\mathbf{1}\left[\tilde{R}_{\ell}^k \neq i, \ell = 0, \ldots, n-1\right]\right] \\
&= \sum_{i=1}^{N} \mathbf{E}\left[\prod_{\ell=0}^{n-1} (1 - \tilde{V}_{\ell}^k(i))\right] \\
&= \sum_{i=1}^{N} \mathbf{E}\left[\psi(\tilde{V}_0^k(i), \ldots, \tilde{V}_{n-1}^k(i))\right] \tag{11.9}
\end{aligned}$$

where the mapping $\psi : \mathbb{R}^n \to \mathbb{R}$ is of the form (10.8) and (10.10). By Lemma 10.5, the mapping ψ is supermodular.

For each $k = 1, 2$, the assumption $\{R_{t+\ell}^k, t = 0, 1, \ldots\} \Longrightarrow_{\ell} \{\tilde{R}_t^k, t = 0, 1, \ldots\}$ yields

$$\{V_{t+\ell}^k(i), t = 0, 1, \ldots\} \Longrightarrow_{\ell} \{\tilde{V}_t^k(i), t = 0, 1, \ldots\}, \quad i = 1, \ldots, N. \tag{11.10}$$

But $\boldsymbol{R}^1 \leq_{TC} \boldsymbol{R}^2$ implies the comparison $\{V_t^1(i), t = 0, 1, \ldots\} \leq_{sm} \{V_t^2(i), t = 0, 1, \ldots\}$ for each $i = 1, \ldots, N$, and the sm comparison being closed under weak con-

vergence [52, Thm. 3.9.8, p. 116], it is now plain from (11.10) that

$$\{\tilde{V}_t^1(i), t = 0, 1, \ldots\} \leq_{sm} \{\tilde{V}_t^2(i), t = 0, 1, \ldots\}, \quad i = 1, \ldots, N. \tag{11.11}$$

In short, $\tilde{\boldsymbol{R}}^1 \leq_{TC} \tilde{\boldsymbol{R}}^2$ and the required condition (11.5) follows upon combining (11.11) with (11.9).

Lastly, under the assumptions of the theorem, we recall from Lemma 11.4 that $\mathbf{E}\left[T(\boldsymbol{R}^1)\right] = \mathbf{E}\left[T(\boldsymbol{R}^2)\right] = N$, and (11.6) holds. ∎

The following results are obtained upon combining Theorem 11.3 with Theorems 9.3, 9.4 and 9.7, respectively.

Corollary 11.5 *Assume the request stream $\boldsymbol{R} = \{R_t, t = 0, 1, \ldots\}$ to be modeled according to the stationary HOMM($h, \boldsymbol{\alpha}, \boldsymbol{p}$) with admissible popularity pmf $\boldsymbol{p}$. Then, it holds that*

$$T(\hat{\boldsymbol{R}}) \leq_{cx} T(\boldsymbol{R})$$

where $\hat{\boldsymbol{R}}$ is the IRM with popularity pmf $\boldsymbol{p}$.

Corollary 11.6 *Assume that for each $k = 1, 2$, the request stream $\boldsymbol{R}^{\beta_k} = \{R_t^{\beta_k}, t = 0, 1, \ldots\}$ is modeled according to the stationary PMM($\beta_k, \boldsymbol{p}$) with admissible popularity pmf $\boldsymbol{p}$. If $0 < \beta_2 \leq \beta_1$, then it holds that*

$$T(\boldsymbol{R}^{\beta_1}) \leq_{cx} T(\boldsymbol{R}^{\beta_2}).$$

Corollary 11.7 *Assume the request stream $\boldsymbol{R}^a = \{R_t^a, t = 0, 1, \ldots\}$ to be modeled according to the stationary LRUSM($\boldsymbol{a}$) with stack distance pmf $\boldsymbol{a}$ satisfying (9.22). Then, it holds that*

$$T(\hat{\boldsymbol{R}}^a) \leq_{cx} T(\boldsymbol{R}^a)$$

where $\hat{\boldsymbol{R}}^a$ is the IRM with uniform popularity pmf $\boldsymbol{u}$.

11.2 Stack distance

The notion of stack distance has been widely used as a metric for temporal correlations [1, 3, 50]: For each $t = 1, 2, \ldots$, the stack distance of the request stream $\boldsymbol{R} = \{R_t, t = 0, 1, \ldots\}$ at time t is the rv $D(t; \boldsymbol{R})$ defined by

$$D(t; \boldsymbol{R}) = |\{R_{t-T(t;\boldsymbol{R})+1,}, \ldots, R_t\}| \tag{11.12}$$

where $T(t; \boldsymbol{R})$ is the inter-reference time (11.1). It is not hard to see that the relation

$$D(t; \boldsymbol{R}) = S(t, T(t; \boldsymbol{R}); \boldsymbol{R}) \tag{11.13}$$

holds. In words, $D(t; \boldsymbol{R})$ can be interpreted as the working set size where the length of the working set is taken to be the inter-reference time $T(t; \boldsymbol{R})$. Hence, $D(t; \boldsymbol{R})$ records the number of *distinct* documents requested from the time the document R_t was last requested before time t.

Under some appropriate conditions on the request stream $\{R_t, t = 0, 1, \ldots\}$, the weak convergence $D(t; \boldsymbol{R}) \Longrightarrow_t D(\boldsymbol{R})$ holds with the steady state stack distance $D(\boldsymbol{R})$ being the rv representing the number of distinct documents requested between two consecutive requests for the same document. This fact is given in the next lemma whose proof can be found in Appendix E.5.

Lemma 11.8 *Assume the request stream* $\boldsymbol{R} = \{R_t, t = 0, 1, \ldots\}$ *to be asymptotically stationary, i.e.,* $\{R_{t+\ell}, t = 0, 1, \ldots\} \Longrightarrow_\ell \{\tilde{R}_t, t = 0, 1, \ldots\}$ *with* $\tilde{\boldsymbol{R}} = \{\tilde{R}_t, t = 0, 1, \ldots\}$ *being a stationary sequence of* $\mathcal{N}$*-valued rvs. Then, it holds that*

$$D(t; \boldsymbol{R}) \Longrightarrow_t D(\boldsymbol{R}). \tag{11.14}$$

It is known [33, 37] that the stack distance is related to the miss rate of the LRU replacement policy. Specifically, given a request stream $\boldsymbol{R}$ such that the steady state stack distance $D(\boldsymbol{R})$ exists, the miss rate $M_{\mathrm{LRU}}(\boldsymbol{R})$ of LRU with cache size M can be expressed in terms of the tail distribution of $D(\boldsymbol{R})$ through

$$M_{\mathrm{LRU}}(\boldsymbol{R}) = \mathbf{P}\left[D(\boldsymbol{R}) > M\right]. \tag{11.15}$$

11.2.1 The effect of popularity

To see the effect of popularity, we restrict the request streams to be in the class of IRMs, in which case the steady state stack distances exist by Lemma 11.8. From (11.13), in view of the results obtained in Corollary 10.3, we might expect that for two IRM request streams $\boldsymbol{R}^1$ and $\boldsymbol{R}^2$ with popularity pmfs $\boldsymbol{p}^1$ and $\boldsymbol{p}^2$, respectively, the comparison

$$D(\boldsymbol{R}^2) \leq_{st} D(\boldsymbol{R}^1) \tag{11.16}$$

should hold if $\boldsymbol{p}^1 \prec \boldsymbol{p}^2$. However, the comparison (11.16) can not be established as we explain below: Recall the relation (11.15) between the miss rate of the LRU policy and the tail distribution of the stack distance. In Section 8.1, we have seen that it is possible to find pmfs $\boldsymbol{p}^1$ and $\boldsymbol{p}^2$ on $\mathcal{N}$ such that $\boldsymbol{p}^1 \prec \boldsymbol{p}^2$ and yet $\hat{M}_{\mathrm{LRU}}(\boldsymbol{p}^1) < \hat{M}_{\mathrm{LRU}}(\boldsymbol{p}^2)$, or equivalently, $\mathbf{P}\left[D(\boldsymbol{R}^1) > M\right] < \mathbf{P}\left[D(\boldsymbol{R}^2) > M\right]$. As we recall (3.2), we conclude that the comparison (11.16) does not hold in general.

Although somewhat annoying from the point of view of intuition, this state of affairs is perhaps not too surprising (in view of (11.13)) given the opposite direction of the comparison of inter-reference times in Theorem 11.2. It is possible that some comparison other than (11.16) might hold, say in the increasing concave ordering, i.e., for two IRM request streams $\boldsymbol{R}^1$ and $\boldsymbol{R}^2$ with popularity pmfs $\boldsymbol{p}^1$ and $\boldsymbol{p}^2$, respectively, it holds

$$D(\boldsymbol{R}^2) \leq_{icv} D(\boldsymbol{R}^1) \tag{11.17}$$

136

whenever $\boldsymbol{p}^1 \prec \boldsymbol{p}^2$. This comparison is compatible with the *weaker* result of Yue and Wong [73] that the comparison $\mathbf{E}\left[D(\boldsymbol{R}^2)\right] < \mathbf{E}\left[D(\boldsymbol{R}^1)\right]$ holds whenever $\boldsymbol{p}^1 \prec \boldsymbol{p}^2$.

11.2.2 The effect of temporal correlations

Inspired by the results obtained for the working set size in Corollary 10.6, we would expect that the stronger the strength of temporal correlations, the smaller the stack distance. Unfortunately, we have not yet been able to formalize this statement and will pose this problem in the following conjecture.

Conjecture 11.9 *Assume that for each* $k = 1, 2$, *the request stream* $\boldsymbol{R}^k$ *is asymptotically stationary, i.e.,* $\{R_{t+\ell}^k, t = 0, 1, \ldots\} \Longrightarrow_\ell \{\tilde{R}_t^k, t = 0, 1, \ldots\}$ *where* $\tilde{\boldsymbol{R}}^k = \{\tilde{R}_t^k, t = 0, 1, \ldots\}$ *is a stationary sequence of* $\mathcal{N}$-*valued rvs. If* $\boldsymbol{R}^1 \leq_{TC} \boldsymbol{R}^2$, *then it holds that*

$$\mathbf{E}\left[D(\boldsymbol{R}^2)\right] \leq \mathbf{E}\left[D(\boldsymbol{R}^1)\right].$$

A support for this conjecture is given under the class of PMM request streams: For this class of request streams, we have from Theorem 9.8 that if $\boldsymbol{R}^{\beta_1}$ and $\boldsymbol{R}^{\beta_2}$ are modeled according to the PMM$(\beta_1, \boldsymbol{p})$ and PMM$(\beta_2, \boldsymbol{p})$, respectively, with $0 < \beta_2 \leq \beta_1$ (i.e., $\boldsymbol{R}^{\beta_1} \leq_{TC} \boldsymbol{R}^{\beta_2}$), then $M_{\mathrm{LRU}}(\boldsymbol{R}^{\beta_2}) \leq M_{\mathrm{LRU}}(\boldsymbol{R}^{\beta_1})$ for all cache sizes $M = 1, \ldots, N - 1$. It then follows from the relation (11.15) that $\mathbf{P}\left[D(\boldsymbol{R}^{\beta_2}) > M\right] \leq \mathbf{P}\left[D(\boldsymbol{R}^{\beta_1}) > M\right]$ for each $M = 1, 2, \ldots, N - 1$, or equivalently, that

$$D(\boldsymbol{R}^{\beta_2}) \leq_{st} D(\boldsymbol{R}^{\beta_1}) \tag{11.18}$$

by the property (3.2) of the usual stochastic ordering. Conjecture 11.9 holds under the class of PMM request streams since (11.18) implies $\mathbf{E}\left[D(\boldsymbol{R}^{\beta_2})\right] \leq \mathbf{E}\left[D(\boldsymbol{R}^{\beta_1})\right]$.

Appendix A

A Discussion of Lemmas 7.1 and 7.2

Consider the RORA(r) policy for some eviction/insertion pmf r. As pointed out in Sections 7.1.1 and 7.1.2, under the IRM input, the cache states $\{\Omega_t, t = 0, 1, \ldots, \}$ form a Markov chain with state space $\Lambda(M; \mathcal{N})$ whose ergodic properties are determined through the set Σ_r.

Fix the cache state $s = (i_1, \ldots, i_M)$ in $\Lambda(M; \mathcal{N})$, and for each $k, \ell = 1, \ldots, M$, define the set $\Gamma_{k,\ell}(s)$ as the collection of states which can reach s in one step when the eviction and insertion are occurring at positions k and ℓ, respectively. Thus,

$$
\Gamma_{k,\ell}(s) = \begin{cases} \{s' = (i_1, \ldots, i_{k-1}, i, i_k, \ldots, i_{\ell-1}, i_{\ell+1}, \ldots, i_M) : i \notin s\} & \text{if } k < \ell \\ \{s' = (i_1, \ldots, i_{\ell-1}, i_{\ell+1}, \ldots, i_k, i, i_{k+1}, \ldots, i_M) : i \notin s\} & \text{if } k > \ell \\ \{s' = (i_1, \ldots, i_{\ell-1}, i, i_{\ell+1}, \ldots, i_M) : i \notin s\} & \text{if } k = \ell. \end{cases}
$$

Lemma A.1 *Fix* $t = 0, 1, \ldots$. *For each cache state* $s = (i_1, \ldots, i_M)$ *in* $\Lambda(M; \mathcal{N})$, *we have*

$$
\begin{aligned}
\mathbf{P}\left[\Omega_{t+1} = s\right] &= \left(\sum_{i \in s} p(i)\right) \mathbf{P}\left[\Omega_t = s\right] \\
&\quad + \sum_{i \notin s} p(i) \sum_{k=1}^{M} \sum_{\ell=1}^{M} r_{k\ell} \left(\sum_{s' \in \Gamma_{k\ell}(s)} \mathbf{P}\left[\Omega_t = s'\right]\right). \quad \text{(A.1)}
\end{aligned}
$$

Proof. Fix $t = 0, 1, \ldots$. Obviously, we have

$$
\begin{aligned}
\mathbf{P}\left[\Omega_{t+1} = s\right] &= \mathbf{P}\left[\Omega_{t+1} = s, R_t \in S_t\right] + \mathbf{P}\left[\Omega_{t+1} = s, R_t \notin S_t\right] \\
&= \mathbf{P}\left[\Omega_t = s, R_t \in S_t\right] + \mathbf{P}\left[\Omega_{t+1} = s, R_t \notin S_t\right] \quad (\text{A.2})
\end{aligned}
$$

because the cache state remains unchanged if the requested document is in cache.

Next, by independence,

$$
\begin{aligned}
\mathbf{P}\left[\Omega_t = s, R_t \in S_t\right] &= \sum_{i=1}^{N} \mathbf{P}\left[\Omega_t = s, R_t = i, i \in S_t\right] \\
&= \left(\sum_{i \in s} p(i)\right) \mathbf{P}\left[\Omega_t = s\right] \quad (\text{A.3})
\end{aligned}
$$

since S_t is determined by Ω_t. Similarly,

$$
\begin{aligned}
\mathbf{P}\left[\Omega_{t+1} = s, R_t \notin S_t\right] &= \sum_{i=1}^{N} \mathbf{P}\left[\Omega_{t+1} = s, R_t = i, i \notin S_t\right] \\
&= \sum_{i \notin s} \mathbf{P}\left[\Omega_{t+1} = s, R_t = i\right] \\
&= \sum_{i \notin s} \sum_{k=1}^{M} \sum_{\ell=1}^{M} \sum_{s' \in \Gamma_{k\ell}(s)} \mathbf{P}\left[\Omega_t = s', \Omega_{t+1} = s, R_t = i\right] \\
&= \sum_{i \notin s} \sum_{k=1}^{M} \sum_{\ell=1}^{M} \sum_{s' \in \Gamma_{k\ell}(s)} p(i) r_{k\ell} \mathbf{P}\left[\Omega_t = s'\right] \\
&= \sum_{i \notin s} p(i) \sum_{k=1}^{M} \sum_{\ell=1}^{M} r_{k\ell} \left(\sum_{s' \in \Gamma_{k\ell}(s)} \mathbf{P}\left[\Omega_t = s'\right]\right). \quad (\text{A.4})
\end{aligned}
$$

We obtain (A.1) by collecting (A.3) and (A.4) into (A.2). ∎

Case 1 – The set Σ_r being empty, the Markov chain has exactly one irreducible component, namely $\Lambda(r, s_0) = \Lambda(M; \mathcal{N})$ regardless of the initial condition s_0, with

$$
\mu_r(s; p) = \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[\Omega_\tau = s\right] = \lim_{t \to \infty} \mathbf{P}\left[\Omega_t = s\right] \quad a.s.
$$

for each s in $\Lambda(M;\mathcal{N})$. Letting t go to infinity in (A.1), we conclude by the standard theory of Markov chains that $\{\mu_{\boldsymbol{r}}(s;\boldsymbol{p}),\ s \in \Lambda(M;\mathcal{N})\}$ given in (7.3)-(7.4) of Lemma 7.1 is indeed the stationary pmf of this Markov chain since it satisfies the Global Balance Equations

$$\mu_{\boldsymbol{r}}(s;\boldsymbol{p}) \;=\; \left(\sum_{i\in s} p(i)\right) \mu_{\boldsymbol{r}}(s;\boldsymbol{p}) \;+\; \sum_{i\notin s} p(i) \sum_{k=1}^{M}\sum_{\ell=1}^{M} r_{k\ell} \left(\sum_{s'\in\Gamma_{k\ell}(s)} \mu_{\boldsymbol{r}}(s';\boldsymbol{p})\right). \quad \text{(A.5)}$$

We now discuss the technical issues which arise when $N = M + 1$. In this case, the analysis that we have done so far holds for all RORA($\boldsymbol{r}$) policies in Case 1 but the FIFO policy with either $r_{1M} = 1$ or $r_{M1} = 1$. Under this particular case, if $s_0 = (i_1,\ldots,i_M)$, then only $M + 1$ states can be reached from s_0, i.e., $\Lambda(\boldsymbol{r},s_0)$ contains the elements $(i_1,\ldots,i_M), (i_2,\ldots,i_M,i_{M+1}), (i_3,\ldots,i_{M+1},i_1), \ldots, (i_{M+1},i_1,\ldots,i_{M-1})$. This state space $\Lambda(\boldsymbol{r},s_0)$ is equivalent to the set $\Lambda^\star(M;\mathcal{N})$ and it can be verified using the Global Balance Equations (A.5) that the stationary pmf is given by

$$\mu_{\boldsymbol{r}}(s;\boldsymbol{p}) = \frac{p(i_1)\cdots p(i_M)}{\sum_{\{j_1,\ldots,j_M\}\in\Lambda^\star(M;\mathcal{N})} p(j_1)\cdots p(j_M)} \quad \text{(A.6)}$$

with $s = (i_1,\ldots,i_M)$ arbitrary in $\Lambda(\boldsymbol{r},s_0)$. Finally, with the stationary pmf (A.6) and $N = M + 1$, it is plain that the miss rate $\hat{M}_{\boldsymbol{r}}(\boldsymbol{p})$ and the output popularity pmf $\boldsymbol{p}_{\boldsymbol{r}}^\star$ in this case are still given by (7.17) and (6.8), respectively, independently of the initial cache state s_0.

Case 2 – The set $\Sigma_{\boldsymbol{r}}$ is *non*-empty with $|\Sigma_{\boldsymbol{r}}| = m$ for some $m = 1,\ldots,M-1$. As discussed in Section 7.1.2, if the Markov chain starts in the initial state s_0 in $\Lambda(M;\mathcal{N})$, it will always stay within the component $\Lambda(\boldsymbol{r},s_0)$ defined at (7.7). On this component $\Lambda(\boldsymbol{r},s_0)$, the Markov chain is irreducible and aperiodic; its stationary pmf exists for each s in $\Lambda(\boldsymbol{r},s_0)$. It is a simple matter to check that the pmf $\{\mu_{\boldsymbol{r},s_0}(s), s \in \Lambda(\boldsymbol{r},s_0)\}$ given in (7.9)-(7.10) of Lemma 7.2 satisfies the Global Balance Equations (A.5) and hence it is a stationary pmf for this Markov chain.

In the case when $N = M + 1$, the analysis still holds for all RORA(r) policies in Case 2 with the exception of FIFO-like policies, i.e., the RORA(r) policy with $r_{k\ell} = 1$ for some $k, \ell = 1, \ldots, M$ and $|\Sigma_r| = m$, for some $m = 1, \ldots, M - 1$. For this special case, under the same reasons as in Case 1, the state space $\Lambda(r, s_0)$ has only $M - m + 1$ elements and coincides with the set $\Lambda^\star(r, s_0)$ defined at (7.22). We again use the Global Balance Equations (A.5) to show that the stationary pmf is given by

$$\mu_{r,s_0}(s; p) = \frac{\prod_{i_\ell \notin \Sigma_r(s_0)} p(i_\ell)}{\sum_{\{j_1,\ldots,j_M\} \in \Lambda^\star(r,s_0)} \prod_{j_\ell \notin \Sigma_r(s_0)} p(j_\ell)} \tag{A.7}$$

where $s = (i_1, \ldots, i_M)$ arbitrary in $\Lambda(r, s_0)$. It is easy to check in this case that with the stationary pmf given in (A.7), the miss rate $\hat{M}_r(p; s_0)$ and the output popularity pmf $p_{r,s_0}^\star$ also admit the expressions (7.26) and (7.32), respectively.

Appendix B

Proofs of Theorems 8.1, 8.6, 8.8, 8.12 and 10.11

Throughout, the notion of 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 $\lim_{\alpha \to \infty} \frac{f(\alpha)}{g(\alpha)} = 1$. We shall have repeated use for the next two elementary lemmas.

Lemma B.1 *Consider a finite family $a_1, \ldots, a_K$ of positive scalars. We have*

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

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

Lemma B.2 *Consider $2K$ mappings $f_1, g_1, \ldots, f_K, g_K : \mathbb{R}_+ \to \mathbb{R}_+$ such that for each $k = 1, \ldots, K$, we have $f_k(\alpha) \sim g_k(\alpha)$ as $\alpha \to \infty$. Then, it holds that*

$$\sum_{k=1}^{K} f_k(\alpha) \sim \sum_{k=1}^{K} g_k(\alpha) \quad (\alpha \to \infty).$$

From now on, without further mention, all asymptotics are understood in the regime where α is large, and the qualifier $\alpha \to \infty$ is dropped from the notation. In particular, by recalling the normalizing constant $C_\alpha(N)$ of Zipf-like distributions defined at (6.5), we note that

$$C_\alpha(N) \sim 1. \tag{B.1}$$

B.1 A proof of Theorem 8.1

Fix $\alpha \geq 0$. Upon substituting (6.4)-(6.5) into the expression (8.4), we find

$$\hat{M}_{\text{LRU}}(\boldsymbol{p}_\alpha) = \frac{1}{C_\alpha(N)^2} \sum_{i=1}^{N} i^{-\alpha} \nu_\alpha(i) \tag{B.2}$$

with

$$\nu_\alpha(i) = \sum_{s \in \Lambda_i(M;\mathcal{N})} \frac{\prod_{\ell=1}^{M} i_\ell^{-\alpha}}{\prod_{k=1}^{M-1} \left(\sum_{j \notin \{i_1,\ldots,i_k\}} j^{-\alpha} \right)}, \quad i = 1,\ldots,N, \tag{B.3}$$

where for each element $s = (i_1,\ldots,i_M)$ of $\Lambda_i(M;\mathcal{N})$, we have denoted by $j \notin \{i_1,\ldots,i_k\}$ the set of elements j in $\mathcal{N}$ which are not in the set $\{i_1,\ldots,i_k\}$.

Fix $i = 1,2,\ldots,N$. For each element $s = (i_1,\ldots,i_M)$ in $\Lambda_i(M;\mathcal{N})$, we invoke Lemma B.1 to claim that

$$\sum_{j \notin \{i_1,\ldots,i_k\}} j^{-\alpha} \sim \left(\min_{j \notin \{i_1,\ldots,i_k\}} j \right)^{-\alpha}, \quad k = 1,\ldots,M-1,$$

whence

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

where we have set

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

Lemmas B.1 and B.2 together yield

$$\nu_\alpha(i) \sim \sum_{s \in \Lambda_i(M;\mathcal{N})} \left(\frac{\prod_{\ell=1}^{M} i_\ell}{\rho(s)} \right)^{-\alpha} \sim c(i) \cdot \nu(i)^{-\alpha} \tag{B.4}$$

where

$$\nu(i) := \min_{s \in \Lambda_i(M;\mathcal{N})} \left(\frac{\prod_{\ell=1}^{M} i_\ell}{\rho(s)} \right) \tag{B.5}$$

and $c(i)$ is the number of elements s in $\Lambda_i(M;\mathcal{N})$ which achieve the minimum in (B.5).

To proceed we note the obvious inequality

$$\nu(i) \geq \frac{\min_{s \in \Lambda_i(M;\mathcal{N})} \left(\prod_{\ell=1}^{M} i_\ell \right)}{\max_{s \in \Lambda_i(M;\mathcal{N})} \rho(s)}. \tag{B.6}$$

We shall show the existence of element(s) s in $\Lambda_i(M;\mathcal{N})$ which *simultaneously* achieve the minimum in

$$\min_{s\in\Lambda_i(M;\mathcal{N})}\left(\prod_{\ell=1}^{M}i_\ell\right) \tag{B.7}$$

and the maximum in

$$\max_{s\in\Lambda_i(M;\mathcal{N})}\rho(s). \tag{B.8}$$

This will imply that (B.6) holds as an equality, and in the process both the minimal value of $\nu(i)$ and the integer $c(i)$ will be determined.

For $i = M+1,\ldots,N$, it is plain that $s = (1,\ldots,M)$ is the only element in $\Lambda_i(M;\mathcal{N})$ achieving both the minimum (B.7) with minimal value $M!$ and the maximum (B.8) with maximal value $M!$. This last claim can be established by easy interchange arguments. Thus, $c(i) = 1$ and

$$\nu(i) = \frac{M!}{M!} = 1. \tag{B.9}$$

Similarly, when $i = 2,\ldots,M$, the element $s = (1,\ldots,i-1,i+1,\ldots,M,M+1)$ of $\Lambda_i(M;\mathcal{N})$ yields the minimum (B.7) with minimal value $\prod_{\ell=1}^{i-1}\ell \cdot \prod_{\ell=i+1}^{M+1}\ell$ and the maximum (B.8) with maximal value $\prod_{\ell=2}^{i-1}\ell \cdot i^{M-i+1}$, whence $c(i) = 1$ and

$$\nu(i) = \frac{\prod_{\ell=1}^{i-1}\ell \cdot \prod_{\ell=i+1}^{M+1}\ell}{\prod_{\ell=2}^{i-1}\ell \cdot i^{M-i+1}} = \frac{(M+1)!}{i!i^{M-i+1}}. \tag{B.10}$$

For $i = 1$, $\rho(s) = 1$ for *any* element s in $\Lambda_1(M;\mathcal{N})$ so that the maximum (B.8) has value 1. On the other hand, the minimum (B.7) is achieved by *any* of the $M!$ permutations of $(2,3,\ldots,M,M+1)$, yielding the minimal value $(M+1)!$. Hence, $c(1) = M!$ and

$$\nu(1) = (M+1)! \tag{B.11}$$

which is simply (B.10) at $i = 1$.

Invoking Lemmas B.1 and B.2 again, we find

$$\sum_{i=1}^{N}i^{-\alpha}\nu_\alpha(i) \sim c \cdot \left(\min_{i=1,\ldots,N}i\nu(i)\right)^{-\alpha} \tag{B.12}$$

144

for some integer c to be determined. It follows from (B.9) that

$$\min_{i=M+1,\ldots,N} i\nu(i) = \min_{i=M+1,\ldots,N} i = M+1 \tag{B.13}$$

and (B.10) allows us to write

$$\min_{i=1,\ldots,M} i\nu(i) = (M+1) \min_{i=1,\ldots,M} \varphi(i). \tag{B.14}$$

with

$$\varphi(i) := \frac{M!}{i!\,i^{M-i}}, \quad i = 1,\ldots,M. \tag{B.15}$$

It is a simple matter to check that

$$M! = \varphi(1) > \varphi(2) > \ldots > \varphi(M) = 1 \tag{B.16}$$

so that the minimum in (B.14) is achieved at $i = M$ with minimal value $M+1$. It then follows from this fact and (B.13) that

$$\min_{i=1,\ldots,N} i\nu(i) = M+1 \tag{B.17}$$

and $c = 2$. Finally, combining (B.1) (B.2), (B.12) and (B.17) readily leads to

$$\hat{M}_{\mathrm{LRU}}(\boldsymbol{p}_\alpha) \sim 2(M+1)^{-\alpha} \tag{B.18}$$

and the desired conclusion (8.7) is obtained. ∎

B.2   A proof of Theorem 8.6

First, in order to lighten up the notation, let $\boldsymbol{p}_\alpha^\star$ denote $\boldsymbol{p}_{\mathrm{LRU},\alpha}^\star$. The proof of Theorem 8.6 relies on the following observation: By the definition of majorization (2.1)-(2.2), the comparison $\boldsymbol{p}_\alpha^\star \prec \boldsymbol{p}_\alpha$ requires the condition

$$\min_{i=1,\ldots,N} p_\alpha(i) \leq \min_{i=1,\ldots,N} p_\alpha^\star(i) \tag{B.19}$$

to hold. Thus, as we recall (6.6), this comparison will not hold if we can show that

$$C_\alpha(N)N^\alpha \cdot \min_{i=1,\ldots,N} p_\alpha^\star(i) < 1. \tag{B.20}$$

We show under the appropriate conditions on M and N that (B.20) indeed holds for large enough values of α.

Fix $\alpha \geq 0$ and substitute (6.4)-(6.5) into the expression (8.11) for the pmf $\boldsymbol{p}_\alpha^\star$. For each $i = 1, \ldots, N$, we find

$$p_\alpha^\star(i) = \frac{i^{-\alpha}\nu_\alpha(i)}{\sum_{j=1}^N j^{-\alpha}\nu_\alpha(j)} \tag{B.21}$$

with $\nu_\alpha(i)$, $i = 1, \ldots, N$, given at (B.3). By virtue of (B.4), (B.12), (B.17) and (B.21), we can now write

$$p_\alpha^\star(i) \sim \frac{c(i)}{2}\left(\frac{M+1}{i\nu(i)}\right)^\alpha, \quad i = 1, \ldots, N.$$

Consequently,

$$\min_{i=1,\ldots,N} p_\alpha^\star(i) \sim \frac{1}{2}\min_{i=1,\ldots,N}\left(c(i)\left(\frac{M+1}{i\nu(i)}\right)^\alpha\right). \tag{B.22}$$

By recalling (B.9), we get

$$\min_{i=M+1,\ldots,N}\left(c(i)\left(\frac{M+1}{i\nu(i)}\right)^\alpha\right) = \left(\frac{M+1}{N}\right)^\alpha \tag{B.23}$$

where the minimum is achieved at $i = N$. Next, by using (B.10), we get with the help of (B.15) and (B.16) that

$$\min_{i=2,\ldots,N}\left(c(i)\left(\frac{M+1}{i\nu(i)}\right)^\alpha\right) = \left(\frac{2^{M-1}}{M!}\right)^\alpha \tag{B.24}$$

where the minimum is achieved at $i = 2$. Finally, $\nu(1) = (M+1)!$ and $c(1) = M!$ yield

$$c(1)\left(\frac{M+1}{\nu(1)}\right)^\alpha = M!\frac{1}{(M!)^\alpha}. \tag{B.25}$$

Combining (B.1), (B.23), (B.24) and (B.25), we conclude from (B.22) that

$$C_\alpha(N)N^\alpha \cdot \min_{i=1,\dots,N} p_\alpha^\star(i) \sim \frac{1}{2} \min\left(M!\left(\frac{N}{M!}\right)^\alpha, \left(\frac{2^{M-1}N}{M!}\right)^\alpha, (M+1)^\alpha\right).$$

Under (8.26), as α grows large, the first term in the minimum above will have the smallest value, so

$$C_\alpha(N)N^\alpha \cdot \min_{i=1,\dots,N} p_\alpha^\star(i) \sim \frac{M!}{2}\left(\frac{N}{M!}\right)^\alpha,$$

and the condition (B.20) indeed holds for large enough values of α. ∎

B.3   A proof of Theorem 8.8

Fix $\alpha \geq 0$. By substituting (6.4)-(6.5) into the expression (8.30), we find

$$\hat{M}_{\mathrm{CL}}(\boldsymbol{p}_\alpha) = \frac{1}{C_\alpha(N)K_{\mathrm{CL},\alpha}} \sum_{i=1}^{N} i^{-\alpha}\eta_\alpha(i) \tag{B.26}$$

with

$$\eta_\alpha(i) = \sum_{s\in\Lambda_i(M;\mathcal{N})} \prod_{\ell=1}^{M} i_\ell^{-\alpha(M-\ell+1)}, \quad i=1,\dots,N, \tag{B.27}$$

and

$$K_{\mathrm{CL},\alpha} = \sum_{s\in\Lambda(M;\mathcal{N})} \prod_{\ell=1}^{M} i_\ell^{-\alpha(M-\ell+1)}. \tag{B.28}$$

Fix $i=1,\dots,N$. By Lemma B.1 we immediately get

$$\eta_\alpha(i) \sim c'(i)\eta(i)^{-\alpha} \tag{B.29}$$

with

$$\eta(i) := \min_{s\in\Lambda_i(M;\mathcal{N})} \left(\prod_{\ell=1}^{M} i_\ell^{M-\ell+1}\right) \tag{B.30}$$

and $c'(i)$ is the number of elements s in $\Lambda_i(M;\mathcal{N})$ that achieve the minimum in (B.30).

Elementary interchange arguments show that the minimal value in (B.30) is achieved at

some unique element $s = (i_1, \ldots, i_M)$ of $\Lambda_i(M; \mathcal{N})$ with the property $i_1 < i_2 < \ldots < i_M$, so that $c'(i) = 1$.

Using this observation, we first conclude that

$$\eta(M+1) = \ldots = \eta(N) = \prod_{\ell=1}^{M} \ell^{M-\ell+1}. \tag{B.31}$$

On the other hand, whenever $i = 1, \ldots, M$, direct inspection shows that

$$
\begin{aligned}
\eta(i) &= (M+1) \prod_{1 \le \ell < i} \ell^{M-\ell+1} \cdot \prod_{i < \ell \le M} \ell^{M-\ell+2} \\
&= \frac{\prod_{i < \ell \le M} \ell}{i^{M-i+1}} \cdot (M+1)\eta(M+1) \\
&= (M+1)\eta(M+1)\frac{\varphi(i)}{i} \tag{B.32}
\end{aligned}
$$

where the quantities $\varphi(i), i = 1, \ldots, M$, are defined at (B.15).

Next, upon making use of Lemmas B.1 and B.2, we see that

$$\sum_{i=1}^{N} i^{-\alpha} \eta_\alpha(i) \sim c' \cdot \left(\min_{i=1,\ldots,N} i\eta(i) \right)^{-\alpha} \tag{B.33}$$

with c' denoting the number of indices achieving the minimum in $\min_{i=1,\ldots,N} i\eta(i)$. Obviously, by virtue of (B.31), we find

$$\min_{i=M+1,\ldots,N} i\eta(i) = (M+1)\eta(M+1) \tag{B.34}$$

where the minimum is achieved at $i = M + 1$. On the other hand, as we rely on (B.32), we get

$$\min_{i=1,\ldots,M} i\eta(i) = (M+1)\eta(M+1) \min_{i=1,\ldots,M} \varphi(i) \tag{B.35}$$

and by (B.16), the minimum in (B.35) is achieved at $i = M$ with minimal value $(M+1)\eta(M+1)$. Combining this fact with (B.34), we obtain $c' = 2$ and

$$\min_{i=1,\ldots,N} i\eta(i) = (M+1)\eta(M+1). \tag{B.36}$$

Lastly, invoking Lemma B.1 with (B.28) leads to

$$
\begin{aligned}
K_{\mathrm{CL},\alpha} \;\sim\; & \left(\min_{s \in \Lambda(M;\mathcal{N})} \prod_{\ell=1}^{M} i_\ell^{M-\ell+1} \right)^{-\alpha} \\
=\; & \left(\prod_{\ell=1}^{M} \ell^{M-\ell+1} \right)^{-\alpha} \;=\; \eta(M+1)^{-\alpha}.
\end{aligned}
\tag{B.37}
$$

It is now plain to see from (B.1), (B.26), (B.33), (B.36) and (B.37) that

$$
\hat{M}_{\mathrm{CL}}(\boldsymbol{p}_\alpha) \sim 2(M+1)^{-\alpha}
\tag{B.38}
$$

and the conclusion (8.32) follows. ∎

B.4   A proof of Theorem 8.12

To simplify the notation, we shall write $\boldsymbol{p}_\alpha^\star$ to denote $\boldsymbol{p}_{\mathrm{CL},\alpha}^\star$. The proof of this theorem proceeds along the same line as the proof of Theorem 8.6. We need to show under the appropriate conditions on M and N that (B.20) holds for large enough values of α.

Fix $\alpha \geq 0$. Substitute (6.4)-(6.5) into the expression (8.34) yields

$$
p_\alpha^\star(i) = \frac{i^{-\alpha}\eta_\alpha(i)}{\sum_{j=1}^{N} j^{-\alpha}\eta_\alpha(j)}, \quad i = 1, \ldots, N,
\tag{B.39}
$$

with $\eta_\alpha(i)$, $i = 1, \ldots, N$, given at (B.27). With the help of (B.29), (B.33), (B.36) and (B.39), we can now write

$$
p_\alpha^\star(i) \;\sim\; \frac{1}{2} \left(\frac{(M+1)\eta(M+1)}{i\eta(i)} \right)^\alpha, \quad i = 1, \ldots, N.
\tag{B.40}
$$

Therefore, we obtain

$$
\min_{i=1,\ldots,N} p_\alpha^\star(i) \;\sim\; \frac{1}{2} \left(\frac{(M+1)\eta(M+1)}{\max_{i=1,\ldots,N} i\eta(i)} \right)^\alpha.
\tag{B.41}
$$

Upon noting (B.31), it is a simple matter to check that

$$\max_{i=M+1,\ldots,N} i\eta(i) = N \cdot \eta(M+1) \tag{B.42}$$

and from (B.32), it follows from the fact (B.16) that

$$\max_{i=1,\ldots,M} i\eta(i) = (M+1)! \cdot \eta(M+1). \tag{B.43}$$

As a result of (B.42) and (B.43), we find

$$\max_{i=1,\ldots,N} i\eta(i) = \max\left((M+1)!, N\right) \cdot \eta(M+1). \tag{B.44}$$

To conclude the proof, we note from (B.1), (B.41) and (B.44) that

$$C_\alpha(N)N^\alpha \cdot \min_{i=1,\ldots,N} p_\alpha^\star(i) \sim \frac{1}{2}\left(\frac{(M+1)N}{\max\left((M+1)!, N\right)}\right)^\alpha$$

with $\max\left((M+1)!, N\right) = (M+1)!$ under (8.26). Consequently, the last asymptotics takes the simplified form

$$C_\alpha(N)N^\alpha \cdot \min_{i=1,\ldots,N} p_\alpha^\star(i) \sim \frac{1}{2}\left(\frac{N}{M!}\right)^\alpha$$

and the validity of (B.20) for large enough values of α follows. ∎

B.5   A proof of Theorem 10.11

To simplify the notation, the output pmf $\boldsymbol{p}_{\mathrm{WS},\alpha}^\star$ will be denoted by $\boldsymbol{p}_\alpha^\star$. As in the proof of Theorem 8.6, we try to establish (B.20) under the appropriate condition on τ and N for large enough value of α.

Fix $\alpha \geq 0$ and $\tau > 1$. By substituting (6.4)-(6.5) into the expression (10.18), we have

$$p_\alpha^\star(i) = \frac{i^{-\alpha}(\sum_{j \neq i} j^{-\alpha})^\tau}{\sum_{k=1}^N k^{-\alpha}(\sum_{\ell \neq k} \ell^{-\alpha})^\tau}, \quad i = 1, \ldots, N, \tag{B.45}$$

where we have denoted by $j \neq i$ the set of elements j in $\mathcal{N}$ which are different from i.

As a direct application of Lemma B.1, it follows that

$$i^{-\alpha}(\sum_{j \neq i} j^{-\alpha})^\tau \sim i^{-\alpha}(\min_{j \neq i} j)^{-\alpha\tau} = \begin{cases} 2^{-\alpha\tau}, & i = 1 \\ i^{-\alpha}, & i = 2, \dots, N \end{cases} \tag{B.46}$$

and therefore by Lemma B.2, we find

$$\begin{aligned} \sum_{i=1}^{N} i^{-\alpha}(\sum_{j \neq i} j^{-\alpha})^\tau &\sim 2^{-\alpha\tau} + \sum_{i=2}^{N} i^{-\alpha} \\ &\sim 2^{-\alpha}. \end{aligned} \tag{B.47}$$

Combining (B.45), (B.46) and (B.47) yields

$$p_\alpha^\star(i) \sim \begin{cases} 2^{-\alpha(\tau-1)}, & i = 1 \\ \left(\frac{i}{2}\right)^{-\alpha}, & i = 2, \dots, N. \end{cases} \tag{B.48}$$

From the expressions (B.48), it is a simple matter to check that

$$\begin{aligned} \min_{i=1,\dots,N} p_\alpha^\star(i) &\sim \min(2^{-\alpha(\tau-1)}, \min_{i=2,\dots,N} \left(\frac{i}{2}\right)^{-\alpha}) \\ &= \min\left(2^{-\alpha(\tau-1)}, \left(\frac{N}{2}\right)^{-\alpha}\right). \end{aligned} \tag{B.49}$$

Finally, we note from (B.1) and (B.49) that

$$C_\alpha(N)N^\alpha \cdot \min_{i=1,\dots,N} p_\alpha^\star(i) \sim \min\left(\left(\frac{N}{2^{\tau-1}}\right)^\alpha, 2^\alpha\right)$$

and by the enforced condition (10.19), this asymptotics reduces to

$$C_\alpha(N)N^\alpha \cdot \min_{i=1,\dots,N} p_\alpha^\star(i) \sim \left(\frac{N}{2^{\tau-1}}\right)^\alpha.$$

Hence, the condition (B.20) is satisfied for large enough values of α. $\blacksquare$

Appendix C

Proofs of Theorems 8.5 and 8.11

C.1 A proof of Theorem 8.5

To lighten up the notation, we shall write $\boldsymbol{p}_\varepsilon^\star$ to denote $\boldsymbol{p}_{\mathrm{LRU},\varepsilon}^\star$. From Proposition 8.4, the comparison $\boldsymbol{p}_\varepsilon^\star \prec \boldsymbol{p}_\varepsilon$ does not hold whenever $\delta(\varepsilon) > \frac{1-\varepsilon}{N-1}$, or equivalently, whenever

$$p_\varepsilon^\star(1) < \epsilon. \tag{C.1}$$

Under the pmf (8.16), we find from (8.10) that

$$p_\varepsilon(i) m(i; \boldsymbol{p}_\varepsilon) = \frac{(N-2)!}{(N-M-1)!} \frac{(1-(N-1)\varepsilon)\varepsilon^M}{\prod_{k=1}^{M-1}(1-k\varepsilon)} \cdot a(i) \tag{C.2}$$

with

$$a(i) = \begin{cases} N-1 & \text{if } i = 1 \\ 1 + \frac{(N-M-1)\varepsilon}{(1-(N-1)\varepsilon)} + \sum_{\ell=1}^{M-1} \prod_{k=\ell}^{M-1} \frac{(1-k\varepsilon)}{(N-k)\varepsilon} & \text{if } i = 2, \ldots, N. \end{cases} \tag{C.3}$$

Reporting (C.2)-(C.3) into (5.4), we get

$$
\begin{aligned}
p_\varepsilon^\star(1) &= \left[2 + \frac{(N-M-1)\varepsilon}{(1-(N-1)\varepsilon)} + \sum_{\ell=1}^{M-1} \prod_{k=\ell}^{M-1} \frac{(1-k\varepsilon)}{(N-k)\varepsilon} \right]^{-1} \\
&\le \left[\sum_{\ell=1}^{M-1} \prod_{k=\ell}^{M-1} \frac{(1-k\varepsilon)}{(N-k)\varepsilon} \right]^{-1} \\
&\le \left[\sum_{\ell=1}^{M-1} \frac{(1-\ell\varepsilon)}{(N-\ell)\varepsilon} \right]^{-1} \tag{C.4}
\end{aligned}
$$

where the last inequality follows from the fact that for each $k = 1, \ldots, M-1$, $\frac{(1-k\varepsilon)}{(N-k)\varepsilon} \geq 1$ since $\varepsilon \leq \frac{1}{N}$.

Consequently, the condition (C.1) will hold if

$$1 < \sum_{\ell=1}^{M-1} \frac{(1 - \ell\varepsilon)}{(N - \ell)}$$

or equivalently, if

$$\varepsilon < \frac{\left(\sum_{\ell=1}^{M-1} \frac{1}{N-\ell}\right) - 1}{\left(\sum_{\ell=1}^{M-1} \frac{\ell}{N-\ell}\right)}.$$

Hence, provided that N and M satisfy the condition $\sum_{\ell=1}^{M-1} \frac{1}{N-\ell} > 1$, there exists ε in the range (8.23) for which the comparison $\boldsymbol{p}_\varepsilon^\star \prec \boldsymbol{p}_\varepsilon$ does not hold. ∎

C.2   A proof of Theorem 8.11

First, to simplify the notation, the output popularity pmf $\boldsymbol{p}_{\mathrm{CL},\varepsilon}^\star$ will be denoted by $\boldsymbol{p}_\varepsilon^\star$. The proof of this theorem proceeds along the same lines as in the proof of Theorem 8.5. We seek ε such that the condition (C.1) holds.

For the input pmf (8.16), we have from (8.33) that

$$p_\varepsilon(i)m(i;\boldsymbol{p}_\varepsilon) = \frac{(N-2)!}{(N-M-1)!} \frac{(1-(N-1)\varepsilon)\varepsilon^{\frac{M(M+1)}{2}}}{K_{\mathrm{CL}}} \cdot b(i) \qquad \text{(C.5)}$$

with

$$b(i) = \begin{cases} N-1 & \text{if } i = 1 \\ \frac{N-M-1}{1-(N-1)\varepsilon} + \sum_{\ell=1}^{M} \left(\frac{1-(N-1)\varepsilon}{\varepsilon}\right)^{\ell-1} & \text{if } i = 2, \ldots, N. \end{cases} \qquad \text{(C.6)}$$

Combining (C.5)-(C.6) with (5.4), we find

$$p_\varepsilon^\star(1) = \left[1 + \frac{N-M-1}{1-(N-1)\varepsilon} + \sum_{\ell=1}^{M} \left(\frac{1-(N-1)\varepsilon}{\varepsilon}\right)^{\ell-1}\right]^{-1}$$

$$\leq \left[\sum_{\ell=1}^{M}\left(\frac{1-(N-1)\varepsilon}{\varepsilon}\right)^{\ell-1}\right]^{-1}$$

$$= \left[1+\sum_{\ell=1}^{M-1}\left(\frac{1-(N-1)\varepsilon}{\varepsilon}\right)^{\ell}\right]^{-1}$$

$$\leq \left[\sum_{\ell=1}^{M-1}\left(\frac{1-(N-1)\varepsilon}{\varepsilon}\right)^{\ell}\right]^{-1}. \qquad \text{(C.7)}$$

Provided $M > 2$, we obtain

$$p_\varepsilon^\star(1) \leq \left[\frac{\varepsilon}{1-(N-1)\varepsilon}\right]^2. \qquad \text{(C.8)}$$

Thus, the condition (C.1) holds if

$$\left[\frac{\varepsilon}{1-(N-1)\varepsilon}\right]^2 < \varepsilon,$$

or equivalently, if

$$\varepsilon < (1-(N-1)\varepsilon)^2.$$

This last inequality indeed holds when ε is in the range (8.35) and the desired result follows.

Appendix D

Proofs of Proposition 9.6 and Theorem 9.7

D.1 A proof of Proposition 9.6

To facilitate the proof, we shall need the following notion of stack position: Fix $i = 1, \ldots, N$. For each $t = 0, 1, \ldots$, let the rv $X_t^a(i)$ denote the position of document i in the LRU stack Ω_t at time t associated with the request stream $\{R_t^a, \ t = 0, 1, \ldots\}$. From the stack operation (9.16), the sequence $\{X_t^a(i), \ t = 0, 1, \ldots\}$ is seen to evolve according to the recursion

$$
X_{t+1}^a(i) = \begin{cases} 1 & \text{if } D_t = X_t^a(i) \\ X_t^a(i) & \text{if } D_t < X_t^a(i) \\ X_t^a(i) + 1 & \text{if } D_t > X_t^a(i) \end{cases} \tag{D.1}
$$

for all $t = 0, 1, \ldots$ with the initial position $X_0^a(i)$ given and assumed independent of the i.i.d. stack distances $\{D_t, t = 0, 1, \ldots\}$.

By independence of the rvs $\{D_t, t = 0, 1, \ldots\}$, it follows from (D.1) that the sequence $\{X_t^a(i), t = 0, 1, \ldots\}$ is a Markov chain on the state space $\{1, \ldots, N\}$ with one-step transition probability matrix $\boldsymbol{P^a} = (P_{kj}^a, \ j, k = 1, \ldots, N)$ given by

$$
P_{kj}^a = \mathbf{P}\left[X_{t+1}^a(i) = j | X_t^a(i) = k\right]
$$

$$
\begin{aligned}
&= \quad \delta(j,1)\mathbf{P}\left[D_t = k\right] + \delta(j,k)\mathbf{P}\left[D_t < k\right] + \delta(j,k+1)\mathbf{P}\left[D_t > k\right] \\
&= \quad \delta(j,1)a_k + \delta(j,k) \cdot \left(\sum_{\ell=1}^{k-1} a_\ell\right) + \delta(j,k+1) \cdot \left(\sum_{\ell=k+1}^{N} a_\ell\right)
\end{aligned}
$$

for $j, k = 1, \ldots, N$, where we set $\delta(x,y) = \mathbf{1}\left[x = y\right]$ for any $x, y \in \mathbb{R}$. This transition matrix $\boldsymbol{P^a}$ is a doubly stochastic matrix, i.e., $\sum_{j=1}^{N} P_{kj}^{\boldsymbol{a}} = \sum_{k=1}^{N} P_{kj}^{\boldsymbol{a}} = 1$ for all $j, k = 1, \ldots, N$. An invariant distribution for $\boldsymbol{P^a}$ then exists, is unique and is given by the uniform pmf $\boldsymbol{u}$ on $\{1, \ldots, N\}$.

The condition $a_N > 0$ is necessary and sufficient for the Markov chain $\{X_t^{\boldsymbol{a}}(i), t = 0, 1, \ldots\}$ to be irreducible on its finite state space $\{1, \ldots, N\}$, hence to be positive recurrent. For $0 < a_N < 1$, the Markov chain $\{X_t^{\boldsymbol{a}}(i), t = 0, 1, \ldots\}$ is aperiodic while for $a_N = 1$, it is periodic with period N. Regardless of its periodicity [36, Thm. 6.4.3, p. 227], when $a_N > 0$, the fraction of time that $\{X_t^{\boldsymbol{a}}(i), t = 0, 1, \ldots\}$ spends in a given state k will a.s. converge to the corresponding entry of invariant distribution. The latter being the uniform pmf on $\{1, \ldots, N\}$, we conclude that

$$
\lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[X_\tau^{\boldsymbol{a}}(i) = k\right] = \frac{1}{N} \quad a.s., \quad k = 1, \ldots, N. \tag{D.2}
$$

Moreover, in the stationary regime, when $a_N > 0$, we have

$$
\mathbf{P}\left[X_\tau^{\boldsymbol{a}}(i) = k\right] = \frac{1}{N}, \quad k = 1, \ldots, N,
$$

for all $i = 1, \ldots, N$. This implies that in stationarity, the stack rvs $\{\Omega_t, t = 0, 1, \ldots\}$ are uniformly distributed over $\Lambda(N; \mathcal{N})$.

With the fact (D.2), we are now ready to prove Proposition 9.6: Fix $i = 1, \ldots, N$. Recall that $R_t^{\boldsymbol{a}} = i$ if and only if $X_t^{\boldsymbol{a}}(i) = 1$ since this corresponds to document i being in position 1 of the LRU stack Ω_t associated with the request stream $\boldsymbol{R^a}$. Under the assumption $a_N > 0$, we can combine this observation with the convergence (D.2) to get

$$
p_{\boldsymbol{a}}(i) \quad = \quad \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[R_\tau^{\boldsymbol{a}} = i\right]
$$

$$= \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1} \left[X_\tau^a(i) = 1 \right] \quad a.s.$$

and the desired result is obtained. ∎

D.2   A proof of Theorem 9.7

Throughout, for each $i = 1, \ldots, N$, we set

$$V_t^{\boldsymbol{a}}(i) = \mathbf{1} \left[R_t^{\boldsymbol{a}} = i \right], \quad t = 0, 1, \ldots, \tag{D.3}$$

and for each $t = 0, 1, \ldots$, write $\boldsymbol{V}^{\boldsymbol{a},t}(i) = (V_0^{\boldsymbol{a}}(i), \ldots, V_t^{\boldsymbol{a}}(i))$.

Fix $i = 1, \ldots, N$. In order to establish the CIS property of the sequence $\{V_t^{\boldsymbol{a}}(i), t = 0, 1, \ldots\}$, it suffices to show that for each $t = 0, 1, \ldots$, the inequality

$$\mathbf{P} \left[V_{t+1}^{\boldsymbol{a}}(i) = 1 | \boldsymbol{V}^{\boldsymbol{a},t}(i) = \boldsymbol{x}^t \right] \leq \mathbf{P} \left[V_{t+1}^{\boldsymbol{a}}(i) = 1 | \boldsymbol{V}^{\boldsymbol{a},t}(i) = \boldsymbol{y}^t \right] \tag{D.4}$$

holds for *any* pair of vectors $\boldsymbol{x}^t = (x_0, \ldots, x_t)$ and $\boldsymbol{y}^t = (y_0, \ldots, y_t)$ in $\{0,1\}^{t+1}$ satisfying $\boldsymbol{x}^t \leq \boldsymbol{y}^t$ componentwise.

Our first task is to provide a simpler expression for the probabilities of interest. To that end, for $\xi = 1, \ldots, N$, we introduce the quantities $\{P_t(\xi), \ t = 0, 1, \ldots\}$ given by

$$P_t(\xi) := \mathbf{P} \left[X_{t+1}^{\boldsymbol{a}}(i) = 1 | X_0^{\boldsymbol{a}}(i) = \xi, X_1^{\boldsymbol{a}}(i) \neq 1, \ldots, X_t^{\boldsymbol{a}}(i) \neq 1 \right] \tag{D.5}$$

for all $t = 1, 2, \ldots$ with

$$P_0(\xi) := \mathbf{P} \left[X_1^{\boldsymbol{a}}(i) = 1 | X_0^{\boldsymbol{a}}(i) = \xi \right].$$

Moreover, for each $t = 0, 1, \ldots$, and any non-zero element $\boldsymbol{x}^t$ in $\{0,1\}^{t+1}$, we set

$$\tau(\boldsymbol{x}^t) := \max \left(s = 0, \ldots, t : \ x_s = 1 \right).$$

Proposition D.1 *For each $t = 0, 1, \ldots$, and any non-zero vector $\boldsymbol{x}^t$ in $\{0, 1\}^{t+1}$, it holds that*

$$\mathbf{P}\left[V_{t+1}^a(i) = 1 | \boldsymbol{V}^{a,t}(i) = \boldsymbol{x}^t\right] = P_{t-\tau(\boldsymbol{x}^t)}(1). \qquad \text{(D.6)}$$

Proof. Fix $t = 0, 1, \ldots$ and consider a non-zero vector $\boldsymbol{x}^t = (x_0, \ldots, x_t)$ in $\{0, 1\}^{t+1}$. Writing $\tau = \tau(\boldsymbol{x}^t)$ to simplify the notation, we see from the definitions that

$$
\begin{aligned}
& [\boldsymbol{V}^{a,t}(i) = \boldsymbol{x}^t] \\
&= \left[\boldsymbol{V}^{a,\tau-1}(i) = \boldsymbol{x}^{\tau-1}, V_\tau^a(i) = 1, V_{\tau+1}^a(i) = 0, \ldots, V_t^a(i) = 0\right] \\
&= \left[\boldsymbol{V}^{a,\tau-1}(i) = \boldsymbol{x}^{\tau-1}, R_\tau^a = i, R_{\tau+1}^a \neq i, \ldots, R_t^a \neq i\right] \\
&= \left[\boldsymbol{V}^{a,\tau-1}(i) = \boldsymbol{x}^{\tau-1}, X_\tau^a(i) = 1, X_{\tau+1}^a(i) \neq 1, \ldots, X_t^a(i) \neq 1\right] \qquad \text{(D.7)}
\end{aligned}
$$

where we have set $\boldsymbol{x}^{\tau-1} = (x_0, \ldots, x_{\tau-1})$ and that

$$[V_{t+1}^a(i) = 1] = [X_{t+1}^a(i) = 1]. \qquad \text{(D.8)}$$

Assume first that $\tau < t$. Now observe that the event $[\boldsymbol{V}^{a,\tau-1}(i) = \boldsymbol{x}^{\tau-1}, X_\tau^a(i) = 1]$ is determined by the rvs $X_0^a(i), \ldots, X_\tau^a(i)$. Thus, by preconditioning with respect to these rvs, we readily conclude from (D.7) that

$$
\begin{aligned}
& \mathbf{P}\left[\boldsymbol{V}^{a,t}(i) = \boldsymbol{x}^t\right] \\
&= \mathbf{P}\left[\boldsymbol{V}^{a,\tau-1}(i) = \boldsymbol{x}^{\tau-1}, X_\tau^a(i) = 1, X_{\tau+1}^a(i) \neq 1, \ldots, X_t^a(i) \neq 1\right] \\
&= \mathbf{P}\left[\boldsymbol{V}^{a,\tau-1}(i) = \boldsymbol{x}^{\tau-1}, X_\tau^a(i) = 1\right] \\
& \qquad \cdot \mathbf{P}\left[X_{\tau+1}^a(i) \neq 1, \ldots, X_t^a(i) \neq 1 | X_\tau^a(i) = 1\right] \qquad \text{(D.9)}
\end{aligned}
$$

where in the last step we used the fact that the stack position sequence $\{X_t^a(i), t = 0, 1, \ldots\}$ is a Markov chain. Similarly, this time making use of (D.7) and (D.8), we get

$$\mathbf{P}\left[\boldsymbol{V}^{a,t}(i) = \boldsymbol{x}^t, V_{t+1}^a(i) = 1\right]$$

$$= \mathbf{P}\left[\boldsymbol{V}^{\boldsymbol{a},\tau-1}(i) = \boldsymbol{x}^{\tau-1}, X_\tau^{\boldsymbol{a}}(i) = 1, X_{\tau+1}^{\boldsymbol{a}}(i) \neq 1, \ldots, X_t^{\boldsymbol{a}}(i) \neq 1, X_{t+1}^{\boldsymbol{a}}(i) = 1\right]$$

$$= \mathbf{P}\left[\boldsymbol{V}^{\boldsymbol{a},\tau-1}(i) = \boldsymbol{x}^{\tau-1}, X_\tau^{\boldsymbol{a}}(i) = 1\right]$$

$$\cdot \mathbf{P}\left[X_{\tau+1}^{\boldsymbol{a}}(i) \neq 1, \ldots, X_t^{\boldsymbol{a}}(i) \neq 1, X_{t+1}^{\boldsymbol{a}}(i) = 1 | X_\tau^{\boldsymbol{a}}(i) = 1\right]. \tag{D.10}$$

It is now plain that

$$\mathbf{P}\left[V_{t+1}^{\boldsymbol{a}}(i) = 1 | \boldsymbol{V}^{\boldsymbol{a},t}(i) = \boldsymbol{x}^t\right]$$

$$= \frac{\mathbf{P}\left[\boldsymbol{V}^{\boldsymbol{a},t}(i) = \boldsymbol{x}^t, V_{t+1}^{\boldsymbol{a}}(i) = 1\right]}{\mathbf{P}\left[\boldsymbol{V}^{\boldsymbol{a},t}(i) = \boldsymbol{x}^t\right]}$$

$$= \frac{\mathbf{P}\left[X_{\tau+1}^{\boldsymbol{a}}(i) \neq 1, \ldots, X_t^{\boldsymbol{a}}(i) \neq 1, X_{t+1}^{\boldsymbol{a}}(i) = 1 | X_\tau^{\boldsymbol{a}}(i) = 1\right]}{\mathbf{P}\left[X_{\tau+1}^{\boldsymbol{a}}(i) \neq 1, \ldots, X_t^{\boldsymbol{a}}(i) \neq 1 | X_\tau^{\boldsymbol{a}}(i) = 1\right]}$$

$$= \frac{\mathbf{P}\left[X_\tau^{\boldsymbol{a}}(i) = 1, X_{\tau+1}^{\boldsymbol{a}}(i) \neq 1, \ldots, X_t^{\boldsymbol{a}}(i) \neq 1, X_{t+1}^{\boldsymbol{a}}(i) = 1\right]}{\mathbf{P}\left[X_\tau^{\boldsymbol{a}}(i) = 1, X_{\tau+1}^{\boldsymbol{a}}(i) \neq 1, \ldots, X_t^{\boldsymbol{a}}(i) \neq 1\right]}$$

$$= \mathbf{P}\left[X_{t+1}^{\boldsymbol{a}}(i) = 1 | X_\tau^{\boldsymbol{a}}(i) = 1, X_{\tau+1}^{\boldsymbol{a}}(i) \neq 1, \ldots, X_t^{\boldsymbol{a}}(i) \neq 1\right]$$

and the desired conclusion follows by the homogeneity of the Markov chain $\{X_t^{\boldsymbol{a}}(i), t = 0, 1, \ldots\}$.

The case $\tau = t$ is straightforward. ∎

D.2.1 Some preliminary calculations

Since the expressions for the probabilities of interest involve the stack position sequences $\{X_t^{\boldsymbol{a}}(i), t = 0, 1, \ldots\}$, $i = 1, \ldots, N$, associated with the LRUSM request stream $\boldsymbol{R}^{\boldsymbol{a}}$, we shall need some basic facts concerning them in order to show the desired CIS property. Throughout the discussion of the results in this and the next sections, we fix the index $i = 1, \ldots, N$ and the pmf $\boldsymbol{a}$, and lighten up the notation by writing X_t to denote the stack position $X_t^{\boldsymbol{a}}(i)$ of the document i at time t. For each $t = 0, 1, \ldots$, let $\mathcal{A}_t$ denote the event $[X_t \neq D_t, \ldots, X_0 \neq D_0]$.

Recall that the stack distance rvs $\{D_t, t = 0, 1, \ldots\}$ associated with $\{R_t^{\boldsymbol{a}}, t = 0, 1, \ldots\}$ are i.i.d. rvs distributed according to the generic rv D with pmf $\boldsymbol{a}$. We set

$$\alpha(y) = \mathbf{P}[D < y] \quad \text{and} \quad \beta(y) = \mathbf{P}[D > y], \quad y = 0, 1, \ldots, N.$$

and define the quantities

$$Q_t(y; \xi) := \mathbf{P}[X_t = y, \mathcal{A}_{t-1}, X_0 = \xi], \quad y, \xi = 1, \ldots, N,$$

for each $t = 1, 2, \ldots$.

Proposition D.2 *For each* $t = 1, 2, \ldots$ *and* $\xi = 1, \ldots, N$, *it holds that*

$$Q_{t+1}(y; \xi) = \alpha(y)Q_t(y; \xi) + \beta(y - 1)Q_t(y - 1; \xi) \tag{D.11}$$

for all $y = 1, \ldots, N$.

Proof. Fix $t = 1, 2, \ldots$ and $\xi = 1, \ldots, N$. The case $y = 1$ requires a separate analysis: The evolution (D.1) precludes $X_{t+1} = 1$ under the condition $X_t \neq D_t$. Therefore, we must have $\mathbf{P}[X_{t+1} = 1, \mathcal{A}_t, X_0 = \xi] = 0$ and the expression (D.11) holds as we observe that $\alpha(1) = 0$ and $\mathbf{P}[X_t = 0, \mathcal{A}_{t-1}, X_0 = \xi] = 0$.

Next we turn to the case $y = 2, \ldots, N$. The evolution (D.1) implies the relation $X_{t+1} = X_t$ if $D_t < X_t$ and $X_{t+1} = X_t + 1$ if $X_t < D_t$. Thus, the event $[X_{t+1} = y, X_t \neq D_t]$ is the union of the two disjoint events $[X_t = y - 1, X_t < D_t]$ and $[X_t = y, D_t < X_t]$. This leads naturally to

$$
\begin{aligned}
\mathbf{P}[X_{t+1} = y, \mathcal{A}_t, X_0 = \xi] &= \mathbf{P}[X_{t+1} = y, X_t \neq D_t, \mathcal{A}_{t-1}, X_0 = \xi] \\
&= \mathbf{P}[X_t = y - 1, X_t < D_t, \mathcal{A}_{t-1}, X_0 = \xi] \\
&\quad + \mathbf{P}[X_t = y, D_t < X_t, \mathcal{A}_{t-1}, X_0 = \xi]
\end{aligned}
$$

$$\begin{aligned}
&= \quad \mathbf{P}\left[X_t = y-1, y-1 < D_t, \mathcal{A}_{t-1}, X_0 = \xi\right] \\
&\quad + \mathbf{P}\left[X_t = y, D_t < y, \mathcal{A}_{t-1}, X_0 = \xi\right] \\
&= \quad \mathbf{P}\left[y-1 < D_t\right]\mathbf{P}\left[X_t = y-1, \mathcal{A}_{t-1}, X_0 = \xi\right] \\
&\quad + \mathbf{P}\left[D_t < y\right]\mathbf{P}\left[X_t = y, \mathcal{A}_{t-1}, X_0 = \xi\right]
\end{aligned}$$

as we make use of the fact that the rv D_t is independent of the rvs $\{X_s, D_s, \ s = 0, 1, \ldots, t-1, X_t\}$. ∎

The case $t = 0$ in (D.11) is somewhat different but by essentially the same arguments, we get that

$$Q_1(y; \xi) = (\delta(y, \xi)\alpha(\xi) + \delta(y, \xi+1)\beta(\xi)) \cdot \mathbf{P}\left[X_0 = \xi\right] \qquad \text{(D.12)}$$

for arbitrary $y, \xi = 1, \ldots, N$. This follows from the fact that constraints exist between the stack positions X_0 and X_1 on the event $\mathcal{A}_0$.

D.2.2 Monotonicity under the likelihood ratio ordering

We also make use of the so-called *likelihood ratio* ordering, which is now defined.

Definition D.3 *For $\mathbb{N}$-valued rvs X and Y, we say that X is smaller than Y according to the likelihood ratio (lr) ordering, written $X \leq_{lr} Y$, if*

$$\mathbf{P}\left[X = y\right]\mathbf{P}\left[Y = x\right] \leq \mathbf{P}\left[X = x\right]\mathbf{P}\left[Y = y\right] \qquad \text{(D.13)}$$

for all x and y in $\mathbb{N}$ with $x < y$.

The likelihood ratio ordering is stronger than the usual stochastic ordering [59, Thm. 1.C.2, p. 29], i.e., if the $\mathbb{N}$-valued rvs X and Y satisfy $X \leq_{lr} Y$, then $X \leq_{st} Y$.

In what follows, we shall find it convenient to use the following notation: If X is an $\mathbb{N}$-valued rv and $\mathcal{A}$ is an event, then $[X|\mathcal{A}]$ denotes any rv whose distribution is the conditional distribution of X given $\mathcal{A}$. The comparison

$$[X|\mathcal{A}] \leq_{lr} [X|\mathcal{B}]$$

for some other event $\mathcal{B}$ then amounts to

$$\mathbf{P}\left[X = y|\mathcal{A}\right]\mathbf{P}\left[X = x|\mathcal{B}\right] \leq \mathbf{P}\left[X = x|\mathcal{A}\right]\mathbf{P}\left[X = y|\mathcal{B}\right] \tag{D.14}$$

whenever $x < y$ in $\mathbb{N}$, or equivalently

$$\mathbf{P}\left[X = y, \mathcal{A}\right]\mathbf{P}\left[X = x, \mathcal{B}\right] \leq \mathbf{P}\left[X = x, \mathcal{A}\right]\mathbf{P}\left[X = y, \mathcal{B}\right] \tag{D.15}$$

provided $\mathbf{P}\left[\mathcal{A}\right] > 0$ and $\mathbf{P}\left[\mathcal{B}\right] > 0$. With the likelihood ratio ordering, we can now state the following

Theorem D.4 *For $\xi, \zeta = 1, \ldots, N$ with $\xi \leq \zeta$, it holds that*

$$[X_t|\mathcal{A}_{t-1}, X_0 = \xi] \leq_{lr} [X_t|\mathcal{A}_{t-1}, X_0 = \zeta], \quad t = 1, 2, \ldots. \tag{D.16}$$

Before giving a proof we observe that the comparison (D.16) holds for some $t = 1, 2, \ldots$ if

$$\mathbf{P}\left[X_t = y, \mathcal{A}_{t-1}, X_0 = \xi\right]\mathbf{P}\left[X_t = x, \mathcal{A}_{t-1}, X_0 = \zeta\right]$$
$$\leq \quad \mathbf{P}\left[X_t = x, \mathcal{A}_{t-1}, X_0 = \xi\right]\mathbf{P}\left[X_t = y, \mathcal{A}_{t-1}, X_0 = \zeta\right] \tag{D.17}$$

for $x, y = 1, \ldots, N$ with $x < y$.

Proof. The proof proceeds by induction on $t = 1, 2, \ldots$. Throughout we fix arbitrary $\xi, \zeta = 1, \ldots, N$ such that $\xi \leq \zeta$.

162

The basis step: For $t = 1$ the comparison (D.16) (when interpreted through (D.17)) requires that

$$Q_1(y; \xi)Q_1(x; \zeta) \leq Q_1(x; \xi)Q_1(y; \zeta) \tag{D.18}$$

for all $x, y = 1, \ldots, N$ with $x < y$.

In view of (D.12), the inequality (D.18) is certainly implied by

$$\left(\delta(y, \xi)\alpha(\xi) + \delta(y, \xi + 1)\beta(\xi)\right)\left(\delta(x, \zeta)\alpha(\zeta) + \delta(x, \zeta + 1)\beta(\zeta)\right)$$

$$\leq \left(\delta(x, \xi)\alpha(\xi) + \delta(x, \xi + 1)\beta(\xi)\right)\left(\delta(y, \zeta)\alpha(\zeta) + \delta(y, \zeta + 1)\beta(\zeta)\right),$$

an inequality we can rewrite as

$$\delta(y, \xi)\delta(x, \zeta)\alpha(\xi)\alpha(\zeta) + \delta(y, \xi)\delta(x, \zeta + 1)\alpha(\xi)\beta(\zeta)$$

$$+ \delta(y, \xi + 1)\delta(x, \zeta)\beta(\xi)\alpha(\zeta) + \delta(y, \xi + 1)\delta(x, \zeta + 1)\beta(\xi)\beta(\zeta)$$

$$\leq \delta(x, \xi)\delta(y, \zeta)\alpha(\xi)\alpha(\zeta) + \delta(x, \xi)\delta(y, \zeta + 1)\alpha(\xi)\beta(\zeta)$$

$$+ \delta(x, \xi + 1)\delta(y, \zeta)\beta(\xi)\alpha(\zeta) + \delta(x, \xi + 1)\delta(y, \zeta + 1)\beta(\xi)\beta(\zeta). \tag{D.19}$$

Comparing like terms in (D.19), we see that (D.18) will hold since the four inequalities

$$\delta(y, \xi)\delta(x, \zeta) \leq \delta(x, \xi)\delta(y, \zeta),$$

$$\delta(y, \xi)\delta(x, \zeta + 1) \leq \delta(x, \xi)\delta(y, \zeta + 1),$$

$$\delta(y, \xi + 1)\delta(x, \zeta) \leq \delta(x, \xi + 1)\delta(y, \zeta)$$

and

$$\delta(y, \xi + 1)\delta(x, \zeta + 1) \leq \delta(x, \xi + 1)\delta(y, \zeta + 1)$$

all hold under the constraints $x < y$ and $\xi \leq \zeta$.

The induction step: Now assuming that (D.16) holds for some $t = 1, 2, \ldots$, namely

$$[X_t | \mathcal{A}_{t-1}, X_0 = \xi] \leq_{lr} [X_t | \mathcal{A}_{t-1}, X_0 = \zeta], \tag{D.20}$$

163

we seek to show that

$$[X_{t+1}|\mathcal{A}_t, X_0 = \xi] \leq_{lr} [X_{t+1}|\mathcal{A}_t, X_0 = \zeta]. \tag{D.21}$$

As discussed earlier, the comparison (D.20) is equivalent to

$$Q_t(y'; \xi)Q_t(x'; \zeta) \leq Q_t(x'; \xi)Q_t(y'; \zeta) \tag{D.22}$$

for all $x', y' = 1, \ldots, N$ with $x' < y'$, while the desired comparison (D.21) is equivalent to

$$Q_{t+1}(y; \xi)Q_{t+1}(x; \zeta) \leq Q_{t+1}(x; \xi)Q_{t+1}(y; \zeta) \tag{D.23}$$

for all $x, y = 1, \ldots, N$ with $x < y$.

To establish (D.23), we fix $x, y = 1, \ldots, N$ with $x < y$. From Proposition D.2, we have the expressions

$$
\begin{aligned}
Q_{t+1}(y; \xi)Q_{t+1}(x; \zeta) \;=\;& \alpha(y)\alpha(x)Q_t(y; \xi)Q_t(x; \zeta) & \text{(D.24)} \\
& + \alpha(y)\beta(x-1)Q_t(y; \xi)Q_t(x-1; \zeta) & \text{(D.25)} \\
& + \beta(y-1)\alpha(x)Q_t(y-1; \xi)Q_t(x; \zeta) & \text{(D.26)} \\
& + \beta(y-1)\beta(x-1)Q_t(y-1; \xi)Q_t(x-1; \zeta) & \text{(D.27)}
\end{aligned}
$$

and

$$
\begin{aligned}
Q_{t+1}(x; \xi)Q_{t+1}(y; \zeta) \;=\;& \alpha(x)\alpha(y)Q_t(x; \xi)Q_t(y; \zeta) & \text{(D.28)} \\
& + \alpha(x)\beta(y-1)Q_t(x; \xi)Q_t(y-1; \zeta) & \text{(D.29)} \\
& + \beta(x-1)\alpha(y)Q_t(x-1; \xi)Q_t(y; \zeta) & \text{(D.30)} \\
& + \beta(x-1)\beta(y-1)Q_t(x-1; \xi)Q_t(y-1; \zeta). & \text{(D.31)}
\end{aligned}
$$

Comparing the last two expressions term by term, namely (D.24) with (D.28), (D.25) with (D.30), (D.26) with (D.29), and (D.27) with (D.31), we conclude from (D.22) that

(D.23) holds. This completes the proof of the induction step. ∎

Before we can state the main results of this section, we pause for an easy technical lemma.

Lemma D.5 *Let X and Y be $\{1, \ldots, N\}$-valued rvs with $X \leq_{st} Y$, and let D be another $\{1, \ldots, N\}$-valued rv independent of X and Y with pmf $\boldsymbol{a} = (a_1, \ldots, a_N)$, i.e., $\mathbf{P}[D = k] = a_k$, $k = 1, \ldots, N$. If the pmf $\boldsymbol{a}$ satisfies the condition (9.22), then it holds that*

$$\mathbf{P}[Y = D] \leq \mathbf{P}[X = D]. \tag{D.32}$$

Proof. Set $b_\ell = a_\ell - a_{\ell+1}$ for $\ell = 1, \ldots, N-1$ and $b_N = a_N$, so that $a_k = \sum_{\ell=k}^{N} b_\ell$ for each $k = 1, \ldots, N$. The independence of the rvs X and D leads to

$$
\begin{aligned}
\mathbf{P}[X = D] &= \sum_{j=1}^{N} \mathbf{P}[X = j]\, \mathbf{P}[D = j] \\
&= \sum_{j=1}^{N} \mathbf{P}[X = j]\, a_j \\
&= \sum_{j=1}^{N} \left(\sum_{\ell=j}^{N} b_\ell \right) \mathbf{P}[X = j] \\
&= \sum_{\ell=1}^{N} b_\ell \sum_{j=1}^{\ell} \mathbf{P}[X = j] \\
&= \sum_{\ell=1}^{N} b_\ell \mathbf{P}[X \leq \ell] \tag{D.33}
\end{aligned}
$$

and we similarly find

$$\mathbf{P}[Y = D] = \sum_{\ell=1}^{N} b_\ell \mathbf{P}[Y \leq \ell]. \tag{D.34}$$

Under the assumption $X \leq_{st} Y$, we have from (3.2) that $\mathbf{P}[Y \leq \ell] \leq \mathbf{P}[X \leq \ell]$ for all $\ell = 1, \ldots, N$. It is plain from (D.33) and (D.34) that (D.32) holds once it is noted that

$b_\ell \geq 0$ for each $\ell = 1, \ldots, N$, under the monotonicity condition (9.22).   ∎

Proposition D.6 *Assume the stack distance pmf $\mathbf{a}$ to satisfy the condition (9.22). Then, for $\xi, \zeta = 1, \ldots, N$ with $\xi \leq \zeta$, it holds that*

$$P_t(\zeta) \leq P_t(\xi), \quad t = 0, 1, \ldots. \tag{D.35}$$

Proof. First, consider the case $t = 0$. For any $\xi = 1, \ldots, N$, we find

$$P_0(\xi) = \mathbf{P}\left[X_1 = 1 | X_0 = \xi\right] = a_\xi.$$

Hence, for any $\xi, \zeta = 1, \ldots, N$ with $\xi \leq \zeta$, it holds that

$$P_0(\zeta) \leq P_0(\xi)$$

under the condition (9.22).

Fix $t = 1, 2, \ldots$. Recall from (D.1) that

$$[X_1 \neq 1, \ldots, X_t \neq 1] = [X_0 \neq D_0, \ldots, X_{t-1} \neq D_{t-1}] \tag{D.36}$$

and that

$$[X_{t+1} = 1] = [X_t = D_t]. \tag{D.37}$$

Using (D.36) and (D.37), for any $\xi = 1, \ldots, N$, we can rewrite (D.5) as

$$
\begin{aligned}
P_t(\xi) &= \mathbf{P}\left[X_t = D_t | X_0 = \xi, X_0 \neq D_0, \ldots, X_{t-1} \neq D_{t-1}\right] \\
&= \mathbf{P}\left[X_t = D_t | \mathcal{A}_{t-1}, X_0 = \xi\right].
\end{aligned}
\tag{D.38}
$$

Now, fix $\xi, \zeta = 1, \ldots, N$ with $\xi \leq \zeta$. Because the lr ordering implies the st ordering, Theorem D.4 readily yields

$$[X_t | \mathcal{A}_{t-1}, X_0 = \xi] \leq_{st} [X_t | \mathcal{A}_{t-1}, X_0 = \zeta]. \qquad \text{(D.39)}$$

Under the monotonicity condition (9.22), combining (D.39) with Lemma D.5 leads to

$$\mathbf{P}\left[X_t = D_t | \mathcal{A}_{t-1}, X_0 = \zeta\right] \leq \mathbf{P}\left[X_t = D_t | \mathcal{A}_{t-1}, X_0 = \xi\right],$$

and the desired conclusion (D.35) is obtained upon noting (D.38). ∎

Proposition D.7 *Assume the stack distance pmf $\boldsymbol{a}$ to satisfy the condition (9.22). Then, it holds that*

$$P_{t+1}(1) \leq P_t(1), \quad t = 0, 1, \ldots. \qquad \text{(D.40)}$$

Proof. The inequalities (D.40) are simple consequences of Proposition D.6. Fix $t = 1, 2, \ldots$. Under the observation that $[X_0 = 1, X_0 \neq D_0] = [X_1 = 2]$, we find via (D.38) that

$$
\begin{aligned}
P_{t+1}(1) &= \mathbf{P}\left[X_{t+1} = D_{t+1} | \mathcal{A}_t, X_0 = 1\right] \\
&= \mathbf{P}\left[X_{t+1} = D_{t+1} | X_0 = 1, X_0 \neq D_0, \ldots, X_t \neq D_t\right] \\
&= \mathbf{P}\left[X_{t+1} = D_{t+1} | X_1 = 2, X_1 \neq D_1 \ldots, X_t \neq D_t\right] \\
&= \mathbf{P}\left[X_t = D_t | \mathcal{A}_{t-1}, X_0 = 2\right] \\
&= P_t(2) \qquad \text{(D.41)}
\end{aligned}
$$

where the forth equality follows from the homogeneity of the Markov chain $\{X_t, t = 0, 1, \ldots\}$ and by the independence of the rvs $\{D_t, t = 0, 1, \ldots\}$. Invoking Proposition D.6 with (D.41), we get the inequality (D.40).

The case $t = 0$ uses essentially the same argument. We write

$$
\begin{aligned}
P_1(1) &= \mathbf{P}\left[X_1 = D_1 | X_0 = 1, X_0 \neq D_0\right] \\
&= \mathbf{P}\left[X_0 = D_0 | X_0 = 2\right] \\
&= P_0(2)
\end{aligned}
\tag{D.42}
$$

and the inequality $P_1(1) \leq P_0(1)$ simply follows from Proposition D.6 and (D.42). ∎

D.2.3 Main proof

We now return to proving Theorem 9.7 by showing that the sequences $\{V_t^{\boldsymbol{a}}(i), t = 0, 1, \ldots\}$, $i = 1, \ldots, N$, are CIS: Fix $i = 1, \ldots, N$. Given $t = 0, 1, \ldots$, we need to show that (D.4) holds for *any* pair of vectors $\boldsymbol{x}^t = (x_0, \ldots, x_t)$ and $\boldsymbol{y}^t = (y_0, \ldots, y_t)$ in $\{0, 1\}^{t+1}$ satisfying $\boldsymbol{x}^t \leq \boldsymbol{y}^t$ componentwise.

The case $t = 0$ is rather straightforward as (D.4) then reduces to establishing

$$
\mathbf{P}\left[V_1^{\boldsymbol{a}}(i) = 1 | V_0^{\boldsymbol{a}}(i) = 0\right] \leq \mathbf{P}\left[V_1^{\boldsymbol{a}}(i) = 1 | V_0^{\boldsymbol{a}}(i) = 1\right]
$$

or equivalently,

$$
\mathbf{P}\left[X_1^{\boldsymbol{a}}(i) = 1 | X_0^{\boldsymbol{a}}(i) \neq 1\right] \leq \mathbf{P}\left[X_1^{\boldsymbol{a}}(i) = 1 | X_0^{\boldsymbol{a}}(i) = 1\right].
\tag{D.43}
$$

Conditioning on $X_0^{\boldsymbol{a}}(i)$, the condition (D.43) becomes

$$
\sum_{\xi=2}^{N} P_0(\xi) \mathbf{P}\left[X_0^{\boldsymbol{a}}(i) = \xi | X_0^{\boldsymbol{a}}(i) \neq 1\right] \leq P_0(1)
$$

which indeed holds by Proposition D.6.

From now on, as we assume $t = 1, 2, \ldots$, two basic cases need to be considered:

Case 1: Assume $\boldsymbol{x}^t$ to be a non-zero element in $\{0, 1\}^{t+1}$, in which case $\boldsymbol{y}^t$ is also a non-zero element in $\{0, 1\}^{t+1}$. By Proposition D.1, we get that (D.4) holds provided

$$
P_{t-\tau(\boldsymbol{x}^t)}(1) \leq P_{t-\tau(\boldsymbol{y}^t)}(1),
\tag{D.44}
$$

an inequality which is automatically satisfied by virtue of Proposition D.7 given that $\tau(\boldsymbol{x}^t) \leq \tau(\boldsymbol{y}^t)$ whenever $\boldsymbol{x}^t \leq \boldsymbol{y}^t$.

Case 2: Assume that $\boldsymbol{x}^t$ is the zero element $\boldsymbol{0}^t = (0, \ldots, 0)$ in $\{0, 1\}^{t+1}$ and note that

$$\mathbf{P}\left[V_{t+1}^{\boldsymbol{a}}(i) = 1 | \boldsymbol{V}^{\boldsymbol{a},t}(i) = \boldsymbol{0}^t\right] = \mathbf{P}\left[X_{t+1}^{\boldsymbol{a}}(i) = 1 | X_0^{\boldsymbol{a}}(i) \neq 1, \ldots, X_t^{\boldsymbol{a}}(i) \neq 1\right].$$

Invoking again Proposition D.1 for any non-zero element $\boldsymbol{y}^t$ in $\{0, 1\}^{t+1}$, we see that the desired inequality (D.4) reduces to

$$\mathbf{P}\left[X_{t+1}^{\boldsymbol{a}}(i) = 1 | X_0^{\boldsymbol{a}}(i) \neq 1, \ldots, X_t^{\boldsymbol{a}}(i) \neq 1\right] \leq P_{t-\tau(\boldsymbol{y}^t)}(1), \qquad \text{(D.45)}$$

and by Proposition D.7, it then clearly suffices to establish the inequality

$$\mathbf{P}\left[X_{t+1}^{\boldsymbol{a}}(i) = 1 | X_0^{\boldsymbol{a}}(i) \neq 1, \ldots, X_t^{\boldsymbol{a}}(i) \neq 1\right] \leq P_t(1). \qquad \text{(D.46)}$$

Conditioning on $X_0^{\boldsymbol{a}}(i)$, we find

$$\begin{aligned}
&\mathbf{P}\left[X_{t+1}^{\boldsymbol{a}}(i) = 1 | X_0^{\boldsymbol{a}}(i) \neq 1, \ldots, X_t^{\boldsymbol{a}}(i) \neq 1\right] \\
&= \sum_{\xi=2}^{N} P_t(\xi) \mathbf{P}\left[X_0^{\boldsymbol{a}}(i) = \xi | X_0^{\boldsymbol{a}}(i) \neq 1, X_1^{\boldsymbol{a}}(i) \neq 1, \ldots, X_t^{\boldsymbol{a}}(i) \neq 1\right] \\
&\leq P_t(1) \sum_{\xi=2}^{N} \mathbf{P}\left[X_0^{\boldsymbol{a}}(i) = \xi | X_0^{\boldsymbol{a}}(i) \neq 1, X_1^{\boldsymbol{a}}(i) \neq 1, \ldots, X_t^{\boldsymbol{a}}(i) \neq 1\right] \\
&= P_t(1)
\end{aligned}$$

where the inequality follows from Proposition D.6. Thus, the required condition (D.46) holds. This completes the proof of the CIS property of the sequence $\{V_t^{\boldsymbol{a}}(i), t = 0, 1, \ldots\}$.

Finally, since the sequence $\{V_t^{\boldsymbol{a}}(i), t = 0, 1, \ldots\}$ is CIS for each $i = 1, \ldots, N$ and CIS implies PSMD, the desired comparison between $\boldsymbol{R}^{\boldsymbol{a}}$ and its independent version $\hat{\boldsymbol{R}}^{\boldsymbol{a}}$ follows from Proposition 9.2. ∎

Appendix E

Proofs of Lemmas 10.1, 10.12, 11.1, 11.4 and 11.8

E.1   A proof of Lemma 10.1

First, consider the case when the request stream $\boldsymbol{R} = \{R_t, t = 0, 1, \ldots\}$ is stationary. In this case, we have for each $\tau = 1, 2, \ldots$ and for all $t \geq \tau - 1$ that

$$
\begin{aligned}
S(t, \tau; \boldsymbol{R}) &= |\{R_{(t-\tau+1)^+}, \ldots, R_t\}| \\
&= |\{R_{t-\tau+1}, \ldots, R_t\}| \\
&=_{st} |\{R_0, \ldots, R_{\tau-1}\}| \\
&= S(\tau - 1, \tau; \boldsymbol{R}).
\end{aligned}
$$

By letting t go to infinity, we obtain (10.2) with $S(\tau; \boldsymbol{R}) =_{st} S(\tau - 1, \tau; \boldsymbol{R})$.

Next, we show that the limit (10.1) exists for each $\tau = 1, 2, \ldots$. From the definition of the working set size, for $t \geq \tau - 1$, we can write

$$
S(t, \tau; \boldsymbol{R}) = \sum_{i=1}^{N} (1 - \mathbf{1} [R_{t-\ell} \neq i, \ell = 0, \ldots, \tau - 1]). \tag{E.1}
$$

Consequently, the limit (10.1) can be rewritten as

$$
\hat{S}(\tau; \boldsymbol{R}) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{\tau-2} S(t, \tau; \boldsymbol{R})
$$

$$+ \lim_{T \to \infty} \left(\frac{T - \tau + 1}{T} \right) \frac{1}{T - \tau + 1} \sum_{t=\tau-1}^{T-1} S(t, \tau; \boldsymbol{R})$$

$$= \lim_{T \to \infty} \frac{1}{T} \sum_{t=\tau-1}^{\tau+T-2} S(t, \tau; \boldsymbol{R})$$

$$= \sum_{i=1}^{N} \left(1 - \lim_{T \to \infty} \frac{1}{T} \sum_{t=\tau-1}^{\tau+T-2} \mathbf{1} \left[R_{t-\ell} \neq i, \ell = 0, \dots, \tau - 1 \right] \right). \quad \text{(E.2)}$$

Because the limits on the right-hand side of (E.2) are guaranteed to exist a.s. by the stationarity assumption of the request stream $\boldsymbol{R}$ [62, Chap. 5], the limit (10.1) exists a.s. for each $\tau = 1, 2, \dots$.

In addition, if the request stream $\{R_t, t = 0, 1, \dots\}$ is stationary and ergodic, then [62, Chap. 5] for each $i = 1, \dots, N$,

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=\tau-1}^{\tau+T-2} \mathbf{1} \left[R_{t-\ell} \neq i, \ell = 0, \dots, \tau - 1 \right] = \mathbf{P} \left[R_\ell \neq i, \ell = 0, \dots, \tau - 1 \right] \quad a.s.,$$

and it follows from (E.1) and (E.2) that

$$\begin{aligned} \hat{S}(\tau; \boldsymbol{R}) &= \sum_{i=1}^{N} \left(1 - \mathbf{P} \left[R_\ell \neq i, \ell = 0, \dots, \tau - 1 \right] \right) \\ &= \mathbf{E} \left[S(\tau - 1, \tau; \boldsymbol{R}) \right] \\ &= \mathbf{E} \left[S(\tau; \boldsymbol{R}) \right], \quad \tau = 1, 2, \dots. \end{aligned}$$

We now assume that the request stream $\boldsymbol{R} = \{R_t, t = 0, 1, \dots\}$ couples with a stationary sequence of $\mathcal{N}$-valued rvs $\tilde{\boldsymbol{R}} = \{\tilde{R}_t, t = 0, 1, \dots\}$. By coupling, we mean that there exists a coupling time $T^\star$ such that $R_t = \tilde{R}_t$ for all $t \geq T^\star$, with the $\{0, 1, \dots\}$-valued rv $T^\star$ being finite a.s. (see e.g., [45, 64]). Under this assumption, it holds for each $\tau = 1, 2, \dots$ that

$$S(t, \tau; \boldsymbol{R}) = S(t, \tau; \tilde{\boldsymbol{R}}), \quad t \geq T^\star + \tau - 1, \quad \text{(E.3)}$$

or equivalently, the sequence $\{S(t, \tau; \boldsymbol{R}), t = 0, 1, \dots\}$ couples with the sequence $\{S(t, \tau; \tilde{\boldsymbol{R}}), t = 0, 1, \dots\}$ where the coupling time is given by $T^\star + \tau - 1$. By the

first part of the proof, $S(t, \tau; \tilde{\boldsymbol{R}}) \implies_t S(\tau; \tilde{\boldsymbol{R}})$ for each $\tau = 1, 2, \ldots$, and from (E.3), we get $S(t, \tau; \boldsymbol{R}) \implies_t S(\tau; \boldsymbol{R})$ with $S(\tau; \boldsymbol{R}) = S(\tau; \tilde{\boldsymbol{R}})$.

By a similar argument, we find

$$
\lim_{T \to \infty} \frac{1}{T} \sum_{t=\tau-1}^{\tau+T-2} \mathbf{1}\left[R_{t-\ell} \neq i, \ell = 0, \ldots, \tau - 1\right]
$$

$$
= \lim_{T \to \infty} \frac{1}{T} \sum_{t=\tau-1}^{T^\star+\tau-2} \mathbf{1}\left[R_{t-\ell} \neq i, \ell = 0, \ldots, \tau - 1\right]
$$

$$
+ \lim_{T \to \infty} \left(\frac{T - T^\star}{T}\right) \frac{1}{T - T^\star} \sum_{t=T^\star+\tau-1}^{\tau+T-2} \mathbf{1}\left[\tilde{R}_{t-\ell} \neq i, \ell = 0, \ldots, \tau - 1\right]
$$

$$
= \lim_{T \to \infty} \frac{1}{T} \sum_{t=\tau-1}^{\tau+T-2} \mathbf{1}\left[\tilde{R}_{t-\ell} \neq i, \ell = 0, \ldots, \tau - 1\right].
$$

By virtue of (E.2), the limit (10.1) exists for each $\tau = 1, 2, \ldots$, and coincides with $\hat{S}(\tau; \tilde{\boldsymbol{R}})$. Lastly, if the sequence $\tilde{\boldsymbol{R}}$ is stationary and ergodic, the argument above yields

$$
\hat{S}(\tau; \boldsymbol{R}) = \hat{S}(\tau; \tilde{\boldsymbol{R}}) = \mathbf{E}\left[S(\tau; \tilde{\boldsymbol{R}})\right] = \mathbf{E}\left[S(\tau; \boldsymbol{R})\right]
$$

for each $\tau = 1, 2, \ldots$. ∎

E.2 A proof of Lemma 10.12

Fix $\tau = 1, 2, \ldots$. We first consider the case when the request stream $\boldsymbol{R} = \{R_t, t = 0, 1, \ldots\}$ is stationary and ergodic. Fix $i = 1, \ldots, N$. Recalling from (10.20) and (10.21) that

$$
g(V_{t-\tau}(i), \ldots, V_t(i)) = \mathbf{1}\left[R_t = i, R_{t-\ell} \neq i, \ell = 1, \ldots, \tau\right], \tag{E.4}
$$

we can write

$$
\lim_{T \to \infty} \frac{1}{T} \sum_{t=\tau}^{T+\tau-1} g(V_{t-\tau}(i), \ldots, V_t(i))
$$

$$= \lim_{T \to \infty} \frac{1}{T} \sum_{t=\tau}^{T+\tau-1} \mathbf{1}\left[R_t = i, R_{t-\ell} \neq i, \ell = 1, \ldots, \tau\right]$$

$$= \mathbf{P}\left[R_\tau = i, R_\ell \neq i, \ell = 0, \ldots, \tau - 1\right] \quad a.s. \tag{E.5}$$

where the last equality is due to stationarity and ergodicity of the request stream $\boldsymbol{R}$ [62, Chap. 5]. Consequently, the limit (10.23) exists and is given by

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=\tau}^{T+\tau-1} \sum_{i=1}^{N} g(V_{t-\tau}(i), \ldots, V_t(i)) = \sum_{i=1}^{N} \mathbf{P}\left[R_\tau = i, R_\ell \neq i, \ell = 0, \ldots, \tau - 1\right],$$

whence the conclusion (10.25).

Next, we assume that the request stream $\boldsymbol{R} = \{R_t, t = 0, 1, \ldots\}$ couples with a stationary and ergodic sequence of $\mathcal{N}$-valued rvs $\tilde{\boldsymbol{R}} = \{\tilde{R}_t, t = 0, 1, \ldots\}$. Let $\{0, 1, \ldots\}$-valued rv $T^\star$ be the coupling time where $T^\star$ is finite a.s. and $R_t = \tilde{R}_t$ for all $t \geq T^\star$. Fix $i = 1, \ldots, N$ and let $\{\tilde{V}_t(i), t = 0, 1, \ldots\}$ be the indicator sequence associated with $\tilde{\boldsymbol{R}}$ through (9.1). Under this assumption, it is plain from (E.4) that

$$g(V_{t-\tau}(i), \ldots, V_t(i)) = g(\tilde{V}_{t-\tau}(i), \ldots, \tilde{V}_t(i)), \quad t \geq T^\star + \tau, \tag{E.6}$$

hence,

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=\tau}^{T+\tau-1} g(V_{t-\tau}(i), \ldots, V_t(i))$$

$$= \lim_{T \to \infty} \frac{1}{T} \sum_{t=\tau}^{T^\star+\tau-1} g(V_{t-\tau}(i), \ldots, V_t(i))$$

$$+ \lim_{T \to \infty} \left(\frac{T - T^\star}{T}\right) \frac{1}{T - T^\star} \sum_{t=T^\star+\tau}^{T+\tau-1} g(\tilde{V}_{t-\tau}(i), \ldots, \tilde{V}_t(i))$$

$$= \lim_{T \to \infty} \frac{1}{T} \sum_{t=\tau}^{T+\tau-1} g(\tilde{V}_{t-\tau}(i), \ldots, \tilde{V}_t(i))$$

$$= \mathbf{P}\left[\tilde{R}_\tau = i, \tilde{R}_\ell \neq i, \ell = 0, \ldots, \tau - 1\right] \quad a.s. \tag{E.7}$$

where the last equality follows from (E.5).

As a result, the limit (10.23) exists and is given by

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=\tau}^{T+\tau-1} \sum_{i=1}^{N} g(V_{t-\tau}(i), \ldots, V_t(i))$$

$$= \sum_{i=1}^{N} \mathbf{P}\left[\tilde{R}_\tau = i, \tilde{R}_\ell \neq i, \ell = 0, \ldots, \tau - 1\right]. \tag{E.8}$$

Upon noting that

$$\lim_{t \to \infty} \sum_{i=1}^{N} \mathbf{E}\left[g(V_{t-\tau}(i), \ldots, V_t(i))\right] = \lim_{t \to \infty} \sum_{i=1}^{N} \mathbf{P}\left[R_t = i, R_{t-\ell} \neq i, \ell = 1, \ldots, \tau\right]$$

$$= \sum_{i=1}^{N} \mathbf{P}\left[\tilde{R}_\tau = i, \tilde{R}_\ell \neq i, \ell = 0, \ldots, \tau - 1\right],$$

the desired result (10.24) is immediate from (E.8). ∎

E.3   A proof of Lemma 11.1

As in the proof of Lemma 10.1, we first assume that the request stream $\boldsymbol{R} = \{R_t, t = 0, 1, \ldots\}$ is stationary. From the definition of the inter-reference time, we have for each $\tau = 1, 2, \ldots$ and $t = \tau, \tau + 1, \ldots$, that

$$\mathbf{P}\left[T(t; \boldsymbol{R}) > \tau\right] = \mathbf{P}\left[R_{t-\ell} \neq R_t, \ell = 1, \ldots, \tau\right]$$

$$= \sum_{i=1}^{N} \mathbf{P}\left[R_t = i, R_{t-\ell} \neq i, \ell = 1, \ldots, \tau\right] \tag{E.9}$$

$$= \sum_{i=1}^{N} \mathbf{P}\left[R_\tau = i, R_\ell \neq i, \ell = 0, \ldots, \tau - 1\right]$$

$$= \mathbf{P}\left[T(\tau; \boldsymbol{R}) > \tau\right], \tag{E.10}$$

where the third equality follows from the stationarity of the request stream $\boldsymbol{R}$. By letting t go to infinity in (E.10), we obtain $T(t; \boldsymbol{R}) \Longrightarrow_t T(\boldsymbol{R})$ with $\mathbf{P}\left[T(\boldsymbol{R}) > \tau\right] = \mathbf{P}\left[T(\tau; \boldsymbol{R}) > \tau\right]$ for each $\tau = 1, 2, \ldots$.

Next, assume that the request stream $\boldsymbol{R}$ is asymptotically stationary, i.e., $\{R_{t+\ell}, t = 0, 1, \ldots\} \Longrightarrow_\ell \{\tilde{R}_t, t = 0, 1, \ldots\}$ where $\tilde{\boldsymbol{R}} = \{\tilde{R}_t, t = 0, 1, \ldots\}$ is a stationary sequence of $\mathcal{N}$-valued rvs. Under this assumption, we note for each $i = 1, \ldots, N$ that

$$\lim_{t \to \infty} \mathbf{P}\left[R_t = i, R_{t-\ell} \neq i, \ell = 1, \ldots, \tau\right] = \mathbf{P}\left[\tilde{R}_\tau = i, \tilde{R}_\ell \neq i, \ell = 0, \ldots, \tau - 1\right]$$

and invoking (E.9), thus yields

$$\lim_{t \to \infty} \mathbf{P}\left[T(t; \boldsymbol{R}) > \tau\right] = \mathbf{P}\left[T(\tilde{\boldsymbol{R}}) > \tau\right], \quad \tau = 1, 2, \ldots.$$

As a result, the weak convergence $T(t; \boldsymbol{R}) \Longrightarrow_t T(\boldsymbol{R})$ holds with $T(\boldsymbol{R}) =_{st} T(\tilde{\boldsymbol{R}})$, i.e., $T(\boldsymbol{R})$ is characterized by setting $\mathbf{P}\left[T(\boldsymbol{R}) > \tau\right] = \mathbf{P}\left[T(\tilde{\boldsymbol{R}}) > \tau\right]$ for each $\tau = 1, 2, \ldots$.

∎

E.4   A proof of Lemma 11.4

Under the assumptions of the lemma, we note from Appendix E.3 that

$$
\begin{aligned}
\mathbf{P}\left[T(\boldsymbol{R}) > \tau\right] &= \mathbf{P}\left[T(\tilde{\boldsymbol{R}}) > \tau\right] \\
&= \sum_{i=1}^{N} \mathbf{P}\left[\tilde{R}_\tau = i, \tilde{R}_\ell \neq i, \ell = 0, \ldots, \tau - 1\right].
\end{aligned}
$$

Consequently, for each $n = 0, 1, \ldots$, we find

$$\sum_{\tau=n}^{\infty} \mathbf{P}\left[T(\boldsymbol{R}) > \tau\right] = \sum_{i=1}^{N} \sum_{\tau=n}^{\infty} \mathbf{P}\left[\tilde{R}_\tau = i, \tilde{R}_\ell \neq i, \ell = 0, \ldots, \tau - 1\right]. \qquad \text{(E.11)}$$

First, we consider the expression (E.11) for $n = 0$ in which case $\mathbf{E}\left[T(\boldsymbol{R})\right] = \sum_{\tau=0}^{\infty} \mathbf{P}\left[T(\boldsymbol{R}) > \tau\right]$. For each $k = 0, 1, \ldots$, we observe that

$$
\begin{aligned}
&\sum_{\tau=0}^{k} \mathbf{P}\left[\tilde{R}_\tau = i, \tilde{R}_\ell \neq i, \ell = 0, \ldots, \tau - 1\right] \\
&= 1 - \mathbf{P}\left[\tilde{R}_0 \neq i\right] + \sum_{\tau=1}^{k} \mathbf{P}\left[\tilde{R}_\tau = i, \tilde{R}_\ell \neq i, \ell = 0, \ldots, \tau - 1\right] \\
&= 1 - \mathbf{P}\left[\tilde{R}_0 \neq i, \tilde{R}_1 \neq i\right] + \sum_{\tau=2}^{k} \mathbf{P}\left[\tilde{R}_\tau = i, \tilde{R}_\ell \neq i, \ell = 0, \ldots, \tau - 1\right] \\
&\vdots \\
&= 1 - \mathbf{P}\left[\tilde{R}_\ell \neq i, \ell = 0, \ldots, k\right]. \qquad\qquad\qquad\qquad \text{(E.12)}
\end{aligned}
$$

175

By letting k go to infinity, we obtain

$$\sum_{\tau=0}^{\infty} \mathbf{P}\left[\tilde{R}_{\tau} = i, \tilde{R}_{\ell} \neq i, \ell = 0, \ldots, \tau - 1\right] = 1 - \lim_{k \to \infty} \mathbf{P}\left[\tilde{R}_{\ell} \neq i, \ell = 0, \ldots, k\right]$$
$$= 1 \qquad (E.13)$$

under the assumptions (4.2) and (4.3) that the popularity pmf p of R (which coincides with that of $\tilde{R}$) exists and is admissible. It is now immediate from (E.11) and (E.13) that

$$\mathbf{E}\left[T(\boldsymbol{R})\right] = \sum_{i=1}^{N} \sum_{\tau=0}^{\infty} \mathbf{P}\left[\tilde{R}_{\tau} = i, \tilde{R}_{\ell} \neq i, \ell = 0, \ldots, \tau - 1\right] = N.$$

From (E.12) and (E.13), it is plain that the expression (E.11) for the case $n = 1, 2, \ldots$, can be rewritten as

$$\sum_{\tau=n}^{\infty} \mathbf{P}\left[T(\boldsymbol{R}) > \tau\right] = \sum_{i=1}^{N}\left(1 - \sum_{\tau=0}^{n-1} \mathbf{P}\left[\tilde{R}_{\tau} = i, \tilde{R}_{\ell} \neq i, \ell = 0, \ldots, \tau - 1\right]\right)$$
$$= \sum_{i=1}^{N} \mathbf{P}\left[\tilde{R}_{\ell} \neq i, \ell = 0, \ldots, n - 1\right],$$

whence the desired result.

E.5 A proof of Lemma 11.8

To establish Lemma 11.8, we shall make use of the following

Lemma E.1 *For a request stream* $\boldsymbol{R} = \{R_t, t = 0, 1, \ldots\}$ *with admissible popularity pmf* p*, it holds for each* $i = 1, \ldots, N$ *and for each* $k = 1, \ldots, N$ *that*

$$\lim_{t \to \infty} \mathbf{P}\left[R_t = i, R_{\ell} \neq i, \ell = 0, \ldots, t - 1, |\{R_0, \ldots, R_t\}| = k\right] = 0. \qquad (E.14)$$

Proof. For each $i = 1, \ldots, N$ and $k = 1, \ldots, N$, it holds that

$$\mathbf{P}\left[R_t = i, R_\ell \neq i, \ell = 0, \ldots, t - 1, |\{R_0, \ldots, R_t\}| = k\right]$$

$$\leq \quad \mathbf{P}\left[R_t = i, R_\ell \neq i, \ell = 0, \ldots, t - 1\right], \quad t = 1, 2, \ldots, \tag{E.15}$$

and that

$$\lim_{t \to \infty} \mathbf{P}\left[R_t = i, R_\ell \neq i, \ell = 0, \ldots, t - 1\right] = 0 \tag{E.16}$$

under the assumptions (4.2) and (4.3) that the popularity pmf $\boldsymbol{p}$ of $\boldsymbol{R}$ exists and is admissible. Combining (E.15) and (E.16) simply yields (E.14). ∎

Proof of Lemma 11.8. First, we assume that the request stream $\boldsymbol{R} = \{R_t, t = 0, 1, \ldots\}$ is stationary. Fix $k = 1, \ldots, N$. For each $t = 0, 1, \ldots$, the definition of the stack distance gives

$$\mathbf{P}\left[D(t; \boldsymbol{R}) = k\right]$$

$$= \quad \mathbf{P}\left[|\{R_{t-T(t; \boldsymbol{R})+1}, \ldots, R_t\}| = k\right]$$

$$= \quad \sum_{\tau=1}^{t+1} \mathbf{P}\left[T(t; \boldsymbol{R}) = \tau, |\{R_{t-\tau+1}, \ldots, R_t\}| = k\right]$$

$$= \quad \sum_{\tau=1}^{t} \sum_{i=1}^{N} \mathbf{P}\left[R_t = R_{t-\tau} = i, R_{t-\ell} \neq i, \ell = 1, \ldots, \tau - 1, |\{R_{t-\tau+1}, \ldots, R_t\}| = k\right]$$

$$+ \sum_{i=1}^{N} \mathbf{P}\left[R_t = i, R_\ell \neq i, \ell = 0, \ldots, t - 1, |\{R_0, \ldots, R_t\}| = k\right] \tag{E.17}$$

$$= \quad \sum_{i=1}^{N} \sum_{\tau=1}^{t} \mathbf{P}\left[R_\tau = R_0 = i, R_\ell \neq i, \ell = 1, \ldots, \tau - 1, |\{R_1, \ldots, R_\tau\}| = k\right]$$

$$+ \sum_{i=1}^{N} \mathbf{P}\left[R_t = i, R_\ell \neq i, \ell = 0, \ldots, t - 1, |\{R_0, \ldots, R_t\}| = k\right] \tag{E.18}$$

where the last equality follows from the stationarity of the request stream $\boldsymbol{R}$.

We now verify the existence of the limit of (E.18) as t goes to infinity. For each $i = 1, \ldots, N$ and $t = 1, 2, \ldots$, we have

$$
\begin{aligned}
\psi_{k,t}(i) &:= \sum_{\tau=1}^{t} \mathbf{P}\left[R_\tau = R_0 = i, R_\ell \neq i, \ell = 1, \ldots, \tau - 1, |\{R_1, \ldots, R_\tau\}| = k\right] \\
&\leq \sum_{\tau=1}^{t} \mathbf{P}\left[R_\tau = R_0 = i, R_\ell \neq i, \ell = 1, \ldots, \tau - 1\right] \\
&\leq \sum_{\tau=1}^{\infty} \mathbf{P}\left[R_\tau = R_0 = i, R_\ell \neq i, \ell = 1, \ldots, \tau - 1\right] \\
&= \mathbf{P}\left[R_0 = i\right].
\end{aligned}
$$

Consequently, for each $i = 1, \ldots, N$, the monotone sequence $\{\psi_{k,t}(i), t = 1, 2, \ldots\}$ is bounded above by $\mathbf{P}\left[R_0 = i\right]$, thus its limit exists, is finite and is given by

$$
\begin{aligned}
\psi_k(i) &:= \lim_{t \to \infty} \psi_{k,t}(i) \\
&= \sum_{\tau=1}^{\infty} \mathbf{P}\left[R_\tau = R_0 = i, R_\ell \neq i, \ell = 1, \ldots, \tau - 1, |\{R_1, \ldots, R_\tau\}| = k\right].
\end{aligned}
$$

Combining this fact with (E.18) and Lemma E.1 yields

$$
\lim_{t \to \infty} \mathbf{P}\left[D(t; \boldsymbol{R}) = k\right] = \sum_{i=1}^{N} \psi_k(i), \quad k = 1, \ldots, N,
$$

whence $D(t; \boldsymbol{R}) \Longrightarrow_t D(\boldsymbol{R})$ with $D(\boldsymbol{R})$ characterized by setting $\mathbf{P}\left[D(\boldsymbol{R}) = k\right] = \sum_{i=1}^{N} \psi_k(i)$ for each $k = 1, \ldots, N$.

Now, assume that the request stream $\boldsymbol{R}$ is asymptotically stationary, i.e., $\{R_{t+\ell}, t = 0, 1, \ldots\} \Longrightarrow_\ell \{\tilde{R}_t, t = 0, 1, \ldots\}$ where $\tilde{\boldsymbol{R}} = \{\tilde{R}_t, t = 0, 1, \ldots\}$ is a stationary sequence of $\mathcal{N}$-valued rvs. Fix $k = 1, \ldots, N$. Under this assumption, we note that

$$
\begin{aligned}
&\lim_{t \to \infty} \mathbf{P}\left[R_t = R_{t-\tau} = i, R_{t-\ell} \neq i, \ell = 1, \ldots, \tau - 1, |\{R_{t-\tau+1}, \ldots, R_t\}| = k\right] \\
&= \mathbf{P}\left[\tilde{R}_\tau = \tilde{R}_0 = i, \tilde{R}_\ell \neq i, \ell = 1, \ldots, \tau - 1, |\{\tilde{R}_1, \ldots, \tilde{R}_\tau\}| = k\right]. \quad \text{(E.19)}
\end{aligned}
$$

for each $i = 1, \ldots, N$ and $\tau = 1, 2, \ldots$

We shall establish the existence of the limit of $\mathbf{P}\left[D(t; \boldsymbol{R}) = k\right]$ as t goes to infinity by using the expression (E.17). As in the first part of the proof, for each $i = 1, \ldots, N$, it is plain that

$$
\begin{aligned}
&\tilde{\psi}_{k,t}(i) \\
:=\ &\sum_{\tau=1}^{t} \mathbf{P}\left[R_t = R_{t-\tau} = i, R_{t-\ell} \neq i, \ell = 1, \ldots, \tau-1, |\{R_{t-\tau+1}, \ldots, R_t\}| = k\right] \\
\leq\ &\sum_{\tau=1}^{t} \mathbf{P}\left[R_t = R_{t-\tau} = i, R_{t-\ell} \neq i, \ell = 1, \ldots, \tau-1\right] \\
\leq\ &\mathbf{P}\left[R_t = i\right], \quad t = 1, 2, \ldots,
\end{aligned}
$$

and the monotone sequence $\{\tilde{\psi}_{k,t}(i), t = 1, 2, \ldots\}$ is bounded above by 1. Consequently, for each $i = 1, \ldots, N$, $\lim_{t\to\infty} \tilde{\psi}_{k,t}(i)$ exists, is finite and is given by

$$
\begin{aligned}
&\lim_{t\to\infty} \tilde{\psi}_{k,t}(i) \\
=\ &\lim_{t\to\infty} \sum_{\tau=1}^{t} \mathbf{P}\left[R_t = R_{t-\tau} = i, R_{t-\ell} \neq i, \ell = 1, \ldots, \tau-1, |\{R_{t-\tau+1}, \ldots, R_t\}| = k\right] \\
=\ &\sum_{\tau=1}^{\infty} \mathbf{P}\left[\tilde{R}_\tau = \tilde{R}_0 = i, \tilde{R}_\ell \neq i, \ell = 1, \ldots, \tau-1, |\{\tilde{R}_1, \ldots, \tilde{R}_\tau\}| = k\right] \quad (\text{E.20})
\end{aligned}
$$

as we make use of (E.19).

By virtue of Lemma E.1 and (E.20), it now follows from (E.17) that

$$
\begin{aligned}
&\lim_{t\to\infty} \mathbf{P}\left[D(t; \boldsymbol{R}) = k\right] \\
=\ &\sum_{i=1}^{N} \sum_{\tau=1}^{\infty} \mathbf{P}\left[\tilde{R}_\tau = \tilde{R}_0 = i, \tilde{R}_\ell \neq i, \ell = 1, \ldots, \tau-1, |\{\tilde{R}_1, \ldots, \tilde{R}_\tau\}| = k\right] \\
=\ &\mathbf{P}\left[D(\tilde{\boldsymbol{R}}) = k\right], \quad k = 1, \ldots, N,
\end{aligned}
$$

and $D(t; \boldsymbol{R}) \implies D(\boldsymbol{R})$ with $D(\boldsymbol{R}) =_{st} D(\tilde{\boldsymbol{R}})$, i.e., $\mathbf{P}\left[D(\boldsymbol{R}) = k\right] = \mathbf{P}\left[D(\tilde{\boldsymbol{R}}) = k\right]$ for each $k = 1, \ldots, N$. ∎

BIBLIOGRAPHY

[1] V. Almeida, A. Bestavros, M. Crovella, and A. de Oliveira. Characterizing reference locality in the WWW. In *Proceedings of the Fourth International Conference on Parallel and Distributed Information Systems 1996*, pages 92–107, Miami (FL), December 1996.

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

[3] A. Balamash and M. Krunz. Application of multifractals in the characterization of WWW traffic. In *Proceedings of IEEE ICC 2002*, New York (NY), April 2002.

[4] P. Barford and M. Crovella. Generating representative Web workloads for network and server performance evaluation. In *Proceedings of the 1998 ACM SIGMETRICS Conference*, Madison (WS), June 1998.

[5] G. Barish and K. Obraczka. World Wide Web caching: Trends and techniques. *IEEE Communications Magazine, Internet Technology Series*, pages 178–185, May 2000.

[6] R.E. Barlow and F. Proschan. *Statistical Theory of Reliability and Life Testing*. International Series in Decision Processes. Holt, Rinehart and Winston, New York (NY), 1975.

[7] N. Bäuerle. Inequalities for stochastic models via supermodular orderings. *Communication in Statistics – Stochastic Models*, 13(1):181–201, 1997.

[8] N. Bäuerle. Monotonicity results for $MR|GI|1$ queues. *Journal of Applied Probability*, 34(2):514–524, 1997.

[9] N. Bäuerle and T. Rolski. A monotonicity result for the workload in Markov-modulated queues. *Journal of Applied Probability*, 35(3):741–747, 1998.

[10] L.A. Belady. A study of replacement algorithms for a virtual-storage computer. *IBM Systems Journal*, 5(2):78–101, 1966.

[11] P. Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, New York (NY), 1968.

[12] C.M. Bowman, P.B. Danzig, D.R. Hardy, U. Manber, and M.F. Schwartz. The Harvest information discovery and access system. *Computer Networks and ISDN Systems*, 28(1-2):119–125, 1995.

[13] 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.

[14] M. Busari and C. Williamson. Prowgen: a synthetic workload generation tool for simulation evaluation of Web proxy caches. *Computer Networks*, 38(6):779–794, 2002.

[15] P. Cao and S. Irani. Cost-aware WWW proxy caching algorithms. In *Proceedings of the USENIX Symposium on Internet Technologies and Systems 1997*, Monterey (CA), December 1997.

[16] A. Chankhunthod, P.B. Danzig, C. Neerdaels, M.F. Schwartz, and K.J. Worrell. A hierarchical Internet object cache. In *Proceedings of the USENIX 1996 Technical Conference*, San Diego (CA), 1996.

[17] H. Che, Z. Wang, and Y. Tung. Analysis and design of hierarchical Web caching systems. In *Proceedings of IEEE INFOCOM 2001*, Anchorage (AL), April 2001.

[18] L. Cherkasova and G. Ciardo. Characterizing temporal locality and its impact on Web server performance. In *Proceedings of IEEE ICCCN 2000*, Las Vegas (NV), October 2000.

[19] W.K. Ching, E.S. Fung, and M.K. Ng. Higher-order Markov chain models for categorical data sequences. *International Journal of Naval Research Logistics*, 51(4):557–574, 2004.

[20] S. Christodoulakis. Implications of certain assumptions in database performance evaluation. *ACM Transactions on Database Systems*, 9(2):163–186, 1984.

[21] T. Christofides and E. Vaggelatou. A connection between supermodular ordering and positive/negative association. *Journal of Multivariate Analysis*, 88(1):138–151, 2004.

[22] K.L. Chung. *A Course in Probability Theory*. Academic Press, New York (NY), second edition, 1974.

[23] I. Cidon, S. Kutten, and R. Soffer. Optimal allocation of electronic content. In *Proceedings of IEEE INFOCOM 2001*, Anchorage (AL), April 2001.

[24] E.G. Coffman and P. Denning. *Operating Systems Theory*. Prentice-Hall, Englewood Cliffs (NJ), 1973.

[25] B.D. Davison. *The Design and Evaluation of Web Prefetching and Caching Techniques*. PhD thesis, Department of Computer Science, Rutgers University, New Brunswick (NJ), October 2002.

[26] P.J. Denning. The working set model for program behavior. *Communications of the ACM*, 11(5):323–333, 1968.

[27] P.J. Denning and S.C. Schwartz. Properties of the working-set model. *Communications of the ACM*, 15(3):191–198, 1972.

[28] M. Deshpande and G. Karypis. Selective Markov models for predicting Web page accesses. *ACM Transactions on Internet Technology*, 4(2):163–184, 2004.

[29] S.G. Dykes, C.L. Jeffrey, and S. Das. Taxonomy and design analysis for distributed Web caching. In *Proceedings of the 32th Hawaii International Conference on System Sciences*, Maui (HI), January 1999.

[30] S.G. Dykes and K.A. Robbins. A viability analysis of cooperative proxy caching. In *Proceedings of IEEE INFOCOM 2001*, Anchorage (AL), April 2001.

[31] J.D. Esary, F. Proschan, and D.W. Walkup. Association of random variables, with applications. *Annals of Mathematical Statistics*, 38(5):1466–1474, 1967.

[32] J. Escorcia, D. Ghosal, and D. Sarkar. A novel cache distribution heuristic algorithm for a mesh of caches and its performance evaluation. *Computer Communications*, 25(3):329–340, 2002.

[33] P. Flajolet, D. Gardy, and L. Thimonier. Birthday paradox, coupon collector, caching algorithms and self-organizing search. *Discrete Applied Mathematics*, 39(3):207–229, 1992.

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

[35] E. Gelenbe. A unified approach to the evaluation of a class of replacement algorithms. *IEEE Transactions on Computers*, 22:611–618, 1973.

[36] G. Grimmett and D. Stirzaker. *Probability and Random Processes*. Oxford University Press, New York (NY), third edition, 2001.

[37] 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(2):430–469, 1999.

[38] P.R. 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.

[39] 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.

[40] S. Jin and A. Bestavros. Sources and characteristics of Web temporal locality. In *Proceedings of MASCOTS 2000: The IEEE/ACM International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*, San Francisco (CA), August 2000.

[41] S. Jin and A. Bestavros. Temporal locality in Web request streams: Sources, characteristics, and caching implications (extended abstract). In *Proceedings of the 2000 ACM SIGMETRICS Conference*, Santa Clara (CA), June 2000.

[42] K. Joag-Dev, M.D. Perlman, and L.D. Pitt. Association of normal random variables and Slepian's inequality. *Annals of Probability*, 11(2):451–455, 1983.

[43] P. Krishnan, D. Raz, and Y. Shavitt. The cache location problem. *IEEE/ACM Transactions on Networking*, 8(5):568–582, 2000.

[44] B. Li, M.J. Golin, G.F. Italiano, and X. Deng. On the optimal placement of Web proxies in the Internet. In *Proceedings of IEEE INFOCOM 1999*, New York (NY), March 1999.

[45] T. Lindvall. *Lectures on the Coupling Method*. John Wiley & Sons, New York (NY), 1992.

[46] A. Mahanti, D. Eager, and C. Williamson. Temporal locality and its impact on Web proxy cache performance. *Performance Evaluation, Special Issue on Performance Modelling*, 42(2-3):187–203, 2000.

[47] A. Mahanti, C. Williamson, and D. Eager. Traffic analysis of a Web proxy caching hierarchy. *IEEE Network*, 14(3):16–23, May-June 2000.

[48] A.M. Makowski. On an elementary characterization of the increasing convex ordering, with an application. *Journal of Applied Probability*, 31:834–840, 1994.

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

[50] R.L. Mattson, J. Gecsei, D.R. Slutz, and L. Traiger. Evaluation techniques for storage hierarchies. *IBM Systems Journal*, 9(2):78–117, 1970.

[51] L.E. Meester and J.G. Shanthikumar. Regularity of stochastic processes: A theory of directional convexity. *Probability in the Engineering and Informational Sciences*, 7:343–360, 1993.

[52] A. Müller and D. Stoyan. *Comparison Methods for Stochastic Models and Risks*. John Wiley & Sons, Chichester, UK, 2002.

[53] V. Phalke and B. Gopinath. An inter-reference gap model for temporal locality in program behavior. In *Proceedings of the 1995 ACM SIGMETRICS Conference*, pages 291–300, Ottawa, Canada, May 1995.

[54] S. Podlipnig and L. Boszormenyi. A survey of Web cache replacement strategies. *ACM Computing Surveys*, 35(4):331–373, December 2003.

[55] K. Psounis and B. Prabhakar. Efficient randomized Web-cache replacement schemes using samples from past eviction times. *IEEE/ACM Transactions on Networking*, 10(4):441–454, 2002.

[56] K. Psounis, A. Zhu, B. Prabhakar, and R. Motwani. Modeling correlations in Web-traces and implications for designing replacement policies. *Computer Networks*, 45(4):379–398, 2004.

[57] L. Qiu, V.N. Padmanabhan, and G.M. Voelker. On the placement of Web server replicas. In *Proceedings of IEEE INFOCOM 2001*, Anchorage (AL), April 2001.

[58] B. Ramakrishna Rau. Properties and applications of the Least-Recently-Used stack model. Technical Report CSL-TR-77-139, Stanford University, 1977.

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

[60] M. Shaked and J.G. Shanthikumar. Supermodular stochastic orders and positive dependence of random vectors. *Journal of Multivariate Analysis*, 61(1):86–101, 1997.

[61] G. Shedler and C. Tung. Locality in page reference strings. *SIAM Journal of Computing*, 1(3):218–241, 1972.

[62] A.N. Shiryaev. *Probability*. Springer-Verlag, New York (NY), second edition, 1995.

[63] T. Sugimoto and N. Miyoshi. On the asymptotics of fault probability in Least-Recently-Used caching with Zipf-type request distribution. Technical Report B-407, Department of Mathematical and Computer Sciences, Tokyo Institute of Technology, July 2004.

[64] H. Thorisson. *Coupling, Stationarity, and Regeneration*. Springer-Verlag, New York (NY), 2000.

[65] J. van den Berg and D. Towsley. Properties of the miss ratio for a 2-level storage model with LRU or FIFO replacement strategy and independent references. *IEEE Transactions on Computers*, 42(4):508–512, 1993.

[66] S. Vanichpun and A.M. Makowski. The effects of positive correlations on buffer occupancy: Lower bounds via supermodular ordering. In *Proceedings of IEEE INFOCOM 2002*, New York (NY), June 2002.

[67] S. Vanichpun and A.M. Makowski. When are on-off sources SIS?: Conditions and applications. *Probability in the Engineering and Informational Sciences*, 18(4):423–443, 2004.

[68] D.C. Verma. *Content Distribution Networks: An Engineering Approach*. John Wiley & Sons, New York (NY), January 2002.

[69] J. Wang. A survey of Web caching schemes for the Internet. *ACM Computer Communication Review*, 25(9):36–46, 1999.

[70] C. Williamson. On filter effects in caching hierarchies. *ACM Transactions on Internet Technology*, 2(1):47–77, 2002.

[71] A. Wolman, G.M. Voelker, N. Sharma, N. Cardwell, A. Karlin, and H.M. Levy. On the scale and performance of cooperative Web proxy caching. *Operating Systems Review*, 33(5):16–31, 1999.

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

[73] P.C. Yue and C.K. Wong. On the optimality of the probability ranking scheme in storage applications. *Journal of the ACM*, 20(4):624–633, 1973.