TECHNICAL RESEARCH REPORT

Modeling locality of reference via notions of positive dependence -- Some mixed news!

by Sarut Vanichpun and Armand M. Makowski

CSHCN TR 2005-7
(ISR TR 2005-95)

The Center for Satellite and Hybrid Communication Networks is a NASA-sponsored Commercial Space Center also supported by the Department of Defense (DOD), industry, the State of Maryland, the University of Maryland and the Institute for Systems Research. This document is a technical report in the CSHCN series originating at the University of Maryland.

Web site http://www.isr.umd.edu/CSHCN/

Modeling locality of reference via notions of positive dependence – Some mixed news!

Sarut Vanichpun
Qualcomm, Inc.
5775 Morehouse Dr.
San Diego, CA 92121
Email: sarut@qualcomm.com

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

***Abstract*— We introduce the notion of Temporal Correlations (TC) ordering as a way to compare strength of temporal correlations in streams of requests. This notion is based on the supermodular ordering, a concept of positive dependence used for comparing dependence structures in sequences of rvs. 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 also show how the comparison in the TC ordering is compatible with comparisons of some well-known locality of reference metrics, namely, the working set size and the inter-reference time. We establish a folk theorem to the effect that the stronger the temporal correlations, the smaller the miss rate for the PMM. Conjectures and simulations are offered regarding this folk theorem under the HOMM and under the LRUSM. The validity of this folk theorem is also discussed for general input streams under the Working Set algorithm.**

Keywords: Locality of reference in request streams, Temporal correlations, Positive dependence, Folk theorem for miss rates.

I. Introduction

The notion of *locality of reference* and its importance for caching were first recognized by Belady [8] in the context of computer memory. Subsequently, a number of studies have shown that request streams for Web objects exhibit strong locality of reference[1] [19, 20, 21]. Attempts at characterization were made early on by Denning through the working set model [15, 16]. Yet, like the notion of burstiness used in traffic modeling, locality of reference, while endowed with a clear intuitive content, admits no simple definition. Not surprisingly, in spite of numerous efforts, no consensus has been reached on how to formalize the notion, let alone *compare* streams of requests on the basis of their locality of reference. However, it is by now widely accepted that the two main components in locality of reference are *temporal correlations* in the streams of requests and the *popularity distribution* of requested objects.

[1]At least in the short timescales

To describe these two sources of locality, and to frame the subsequent discussion, we assume the following generic setup: We consider a universe of N cacheable items or documents, labeled $i = 1, \ldots, N$, and we write $\mathcal{N} = \{1, \ldots, N\}$. The successive requests arriving at the cache are modeled by a sequence $\boldsymbol{R} = \{R_t,\ t = 0, 1, \ldots\}$ of $\mathcal{N}$-valued rvs. For simplicity, we say that request R_t occurs at time $t = 0, 1, \ldots$.

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

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

whenever these limits exist (and they do in most models treated in the literature). Popularity represents a long-term expression of locality through the likelihood that a document will be requested in the future relative to other documents. Throughout we assume for the request stream $\boldsymbol{R}$ that the limits (1) exist and are constants. To avoid uninteresting situations, it is *always* the case that[2]

$$p(i) > 0, \quad i = 1, \ldots, N. \quad (2)$$

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

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

as a way to capture these temporal correlations as is traditionally done in other contexts. This is due to the *categorical nature* of the rvs $\{R_t,\ t = 0, 1, \ldots\}$ – they identify objects as values in a discrete set but their *actual* values are of no consequence. The focus should instead be on the *recurrence* patterns displayed by requests for particular documents over time.

The question naturally arises as to whether the popularity pmf and temporal correlations in streams of requests can be compared formally on the basis of some notions that simultaneously capture the intuitive content of locality of reference, and lead to useful implications for cache management. To

[2]A pmf $\boldsymbol{p}$ on $\{1, \ldots, N\}$ satisfying (2) is said to be *admissible*. Under this non-triviality condition (2), every document will eventually be requested by virtue of (1).

clarify this point, consider the following *folk theorem* which is widely expected to hold: For good caching policies, the stronger locality of reference, the smaller the miss rate. A natural step consists in relating locality of reference in a stream of requests to the skewness of its popularity pmf with the understanding that the more skewed the popularity pmf, the greater locality of reference. For instance, the notion of entropy [18] and the concept of majorization [22, 32, 33, 35, 36] have been used with some success precisely for that purpose. In [22, 33, 35] the authors then established a version of the folk theorem by showing (via majorization and Schur-concavity) that the more skewed the popularity pmf (thus, the stronger locality of reference), the smaller the miss rate of the cache. This was done for various cache replacement policies under the standard *Independent Reference Model (IRM)* according to which the requests $\{R_t,\ t = 0, 1, \ldots\}$ are i.i.d. rvs distributed according to the pmf $\boldsymbol{p}$.

When it comes to how temporal correlations contribute to locality of reference, the picture is far from complete: Several metrics have been proposed to capture the impact of temporal correlations, e.g., the inter-reference time [18, 19, 27], the working set size [15, 16] and the stack distance [1, 24]. However, none has been found appropriate for formalizing a folk theorem on miss rates. To make progress, we recall that the locality of reference present in a stream of requests is often coined as the property that "bursts of references are made in the near future to objects referenced in the recent past." Thus, if locality of reference is present in a stream of requests, it is not unreasonable to expect that it would manifest itself through positive temporal correlations of some form. Here, with this in mind, we turn to concepts of *positive dependence* as a way to model temporal correlations exhibited by Web request streams. These notions have been used previously in many contexts, e.g., traffic engineering [6, 7, 34] and reliability theory [4, 30]. The main contributions can be summarized as follows:

1. Temporal correlations and positive dependence – We make a connection between the concepts of positive dependence in sequence of rvs [Section II] and temporal correlations in the stream of requests [Section III]. Specifically, relying on the notion of supermodular ordering [Definition 2.3], we introduce the TC ordering [Definition 3.1] as a way of comparing two streams of requests on the basis of the strength of their temporal correlations.

2. Temporal correlations in Web request models – We make use of 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). For the HOMM [Section IV] and the LRUSM [Section VI], we demonstrate that both models exhibit temporal correlations in the sense that they have stronger strength of temporal correlations than the IRM with the same popularity pmf in the TC ordering. For the PMM [Section V], we show that its correlation parameter indeed captures the strength of temporal correlations, as expected.

3. Temporal correlations and some locality of reference metrics – We show in what sense the comparison of two request streams in the TC ordering is compatible with comparisons of some well-established locality of reference metrics, namely, the working set size [Section VII] and the inter-reference time [Section VIII].

4. Temporal correlations and miss rate – Regarding the aforementioned folk theorem for the miss rate [Section IX], we establish 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 the PMM [Section X-A]. Conjectures and simulations are offered as to when this folk theorem should hold under the HOMM [Section X-B] and under the LRUSM [Section X-C]. Lastly, we consider the miss rate of general input streams under the Working Set algorithm [Section XI]. The results indicate that the folk theorem does hold when the cache holds one document, but may fail to hold in some other situations where counterexamples are given.

We conclude in Section XII by explaining in what sense the news are indeed mixed! Many proofs have been omitted due to space limitations, but can be found in the thesis [33].

A word on the notation in use: Equivalence in law or in distribution between rvs (and stochastic processes) is denoted by $=_{st}$. Convergence in law or in distribution (as $t \to \infty$) is denoted by $\Longrightarrow_t$.

II. Modeling positive dependence

A. Conditionally increasing in sequence

Positive dependence in a collection of rvs can be captured in several ways. We begin with the following strong notion.

Definition 2.1: *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 $\boldsymbol{x} = (x_1, \ldots, x_k)$.*

This definition requires 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 $[X_{k+1}|(X_1, \ldots, X_k) = \boldsymbol{y}]$). In other words, we require

$$\begin{aligned} & \mathbf{E}\left[g(X_{k+1})|(X_1, \ldots, X_k) = \boldsymbol{x}\right] \\ \leq\ & \mathbf{E}\left[g(X_{k+1})|(X_1, \ldots, X_k) = \boldsymbol{y}\right] \end{aligned}$$

for all increasing function $g : \mathbb{R} \to \mathbb{R}$ provided the expectations exist.

The property in Definition 2.1 is sometimes called stochastic increasingness in sequence (SIS).

B. Supermodular ordering

The *supermodular* ordering has been found well suited for comparing the dependence structures of random vectors, e.g., see [6, 7, 30, 34] for recent applications in queueing and

reliability. The underlying class of functions associated with this ordering is first introduced.

Definition 2.2: *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 2.3: *For $\mathbb{R}^n$-valued rvs $\boldsymbol{X}$ and $\boldsymbol{Y}$, we say that $\boldsymbol{X}$ is smaller than $\boldsymbol{Y}$ in the supermodular ordering, written $\boldsymbol{X} \leq_{sm} \boldsymbol{Y}$, if $\mathbf{E}[\varphi(\boldsymbol{X})] \leq \mathbf{E}[\varphi(\boldsymbol{Y})]$ for all supermodular Borel measurable functions $\varphi : \mathbb{R}^n \to \mathbb{R}$ provided the expectations exist.*

It is a simple matter to check [6] that for any $\mathbb{R}^n$-valued rvs $\boldsymbol{X}$ and $\boldsymbol{Y}$, the comparison $\boldsymbol{X} \leq_{sm} \boldsymbol{Y}$ necessarily implies the distributional equalities

$$X_i =_{st} Y_i, \quad i = 1, \ldots, n, \tag{3}$$

as well as the covariance comparisons

$$\mathrm{Cov}[X_i, X_j] \leq \mathrm{Cov}[Y_i, Y_j], \quad i, j = 1, \ldots, n \tag{4}$$

whenever these quantities are well defined. Thus, the comparison $\boldsymbol{X} \leq_{sm} \boldsymbol{Y}$ represents a possible formalization of the statement that "$\boldsymbol{Y}$ is more correlated than $\boldsymbol{X}$" under the constraint that $\boldsymbol{X}$ and $\boldsymbol{Y}$ have the same marginals. Before stating a key comparison related to the supermodular ordering, we need the following definition.

Definition 2.4: *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 $\hat{X}_i =_{st} X_i$ for each $i = 1, \ldots, n$.*

Positive dependence between the components $X_1, \ldots, X_n$ of the $\mathbb{R}^n$-valued rv $\boldsymbol{X}$ can also be expressed by requiring that the rv $\boldsymbol{X}$ be larger in the supermodular ordering than its independent version $\hat{\boldsymbol{X}}$ [26].

Definition 2.5: *The $\mathbb{R}^n$-valued rv $\boldsymbol{X} = (X_1, \ldots, X_n)$ is said to be positive supermodular dependent (PSMD) if $\hat{\boldsymbol{X}} \leq_{sm} \boldsymbol{X}$ where $\hat{\boldsymbol{X}}$ is the independent version of $\boldsymbol{X}$.*

The next proposition is due to Meester and Shanthikumar [25, Thm. 3.8], and explores the relationships between the two notions of positive dependence introduced thus far.

Theorem 2.6: *If the $\mathbb{R}^n$-valued rv $\boldsymbol{X} = (X_1, \ldots, X_n)$ is CIS, then $\boldsymbol{X}$ is PSMD.*

The definitions above readily extend to infinite length sequences of rvs by requiring that each of the definitions holds for each finite section of the sequences.

III. Modeling temporal correlations in Web request streams

Given a stream of requests $\boldsymbol{R} = \{R_t,\ t = 0, 1, \ldots\}$, we set

$$V_t(i) = \mathbf{1}[R_t = i], \quad t = 0, 1, \ldots, \tag{5}$$

for each $i = 1, \ldots, N$, 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 locality of reference through some form of temporal correlations, 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 dependence in the sequence $\{V_t(i),\ t = 0, 1, \ldots\}$ and leads to the following notion of *Temporal Correlations (TC) ordering*.

Definition 3.1: *The request stream $\boldsymbol{R}^1 = \{R_t^1,\ t = 0, 1, \ldots\}$ is said to have weaker temporal correlations than the request stream $\boldsymbol{R}^2 = \{R_t^2,\ t = 0, 1, \ldots\}$, written $\boldsymbol{R}^1 \leq_{TC} \boldsymbol{R}^2$, 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 $\boldsymbol{R}^k$ via (5).

In this paper we use the comparison $\boldsymbol{R}^1 \leq_{TC} \boldsymbol{R}^2$ to formalize the fact that the stream $\boldsymbol{R}^1$ has less locality of reference than the stream $\boldsymbol{R}^2$. The difficulty associated with the "categorical" nature of streams of requests has been bypassed by focusing instead on the (numerical) indicator processes (5). The covariance comparison (4) might in principle have provided a natural way to compare the strength of positive dependencies between each pair of sequences $\{V_t^1(i),\ t = 0, 1, \ldots\}$ and $\{V_t^2(i),\ t = 0, 1, \ldots\}$, $i = 1, \ldots, N$. However, this second-order notion is too weak to establish the desired folk theorem for miss rates.

Now fix $i = 1, \ldots, N$. Whenever $\boldsymbol{R}^1 \leq_{TC} \boldsymbol{R}^2$, the equimarginal property (3) of the supermodular ordering yields $\mathbf{P}\left[V_t^1(i) = 1\right] = \mathbf{P}\left[V_t^2(i) = 1\right]$ for all $t = 0, 1, \ldots$, or equivalently,

$$\mathbf{P}\left[R_t^1 = i\right] = \mathbf{P}\left[R_t^2 = i\right], \quad t = 0, 1, \ldots. \tag{6}$$

Under the assumption that for each $k = 1, 2$, the limits (1) exist as constants for the request stream $\boldsymbol{R}^k$, we have

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

by the Bounded Convergence Theorem. Combining this last equation with (6) immediately leads to $\boldsymbol{p}^1 = \boldsymbol{p}^2$, i.e., the comparison $\boldsymbol{R}^1 \leq_{TC} \boldsymbol{R}^2$ requires that the request streams $\boldsymbol{R}^1$ and $\boldsymbol{R}^2$ have the same popularity profile. Thus, the TC ordering can capture only the contributions from temporal correlations to locality of reference.

Proposition 3.2: *If for each $i = 1, \ldots, N$, the indicator process $\{V_t(i),\ t = 0, 1, \ldots\}$ associated with a request stream $\boldsymbol{R}$ is PSMD, then $\hat{\boldsymbol{R}} \leq_{TC} \boldsymbol{R}$ where $\hat{\boldsymbol{R}}$ is the independent version of $\boldsymbol{R}$.*

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

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

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

for each $i=1,\ldots,N$ and the proof is completed. ■

In the next three sections, 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.

IV. Higher-order Markov chain model

Several higher-order Markov chain models have been proposed to characterize Web request streams (e.g., see [13, 17, 28] and references therein) due to their ability to capture some of the observed temporal correlations. In this section we present a model, recently proposed by Psounis et al. [28], which captures 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 $\mathbf{P}[Y_t=i]=p(i)$ for each $i=1,\ldots,N$. The pmf $\boldsymbol{p}=(p(1),\ldots,p(N))$ is assumed to be admissible (2). 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

$$\mathbf{P}[Z_t=k]=\alpha_k,\quad k=1,\ldots,h$$

and

$$\mathbf{P}[Z_t=0]=\beta=1-\sum_{k=1}^{h}\alpha_k>0$$

for all $t=0,1,\ldots$, 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

$$R_t=\mathbf{1}[Z_t=0]\,Y_t+\sum_{k=1}^{h}\mathbf{1}[Z_t=k]\,R_{t-k}. \tag{7}$$

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 is chosen independently of the past according to the popularity pmf $\boldsymbol{p}$ and $R_t=Y_t$.

The requests $\{R_t, t=0,1,\ldots\}$ form an h^{th}-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 (7) that for any $(i_0,\ldots,i_{t-1})$ in $\mathcal{N}^t$,

$$\begin{aligned}&\mathbf{P}[R_t=i|R_\tau=i_\tau,\tau=0,\ldots,t-1]\\ =\ &\beta p(i)+\sum_{k=1}^{h}\alpha_k\mathbf{1}[i_{t-k}=i] &(8)\\ =\ &\mathbf{P}[R_t=i|R_\tau=i_\tau,\tau=t-h,\ldots,t-1]. &(9)\end{aligned}$$

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 [28] that

$$\lim_{t\to\infty}\mathbf{P}[R_t=i]=\lim_{t\to\infty}\frac{1}{t}\sum_{s=1}^{t}\mathbf{1}[R_s=i]=p(i)\quad a.s. \tag{10}$$

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; its proof is available in Appendix I.

Theorem 4.1: *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 $\beta>0$. Then, it holds that $\hat{\boldsymbol{R}}\leq_{TC}\boldsymbol{R}$ where $\hat{\boldsymbol{R}}$ is the IRM with popularity pmf $\boldsymbol{p}$.*

V. The partial Markov chain model

The partial Markov chain model was introduced 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, for a given popularity pmf $\boldsymbol{p}$, it is natural to expect that the smaller the value of the correlation parameter β, the greater the 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 are 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 4.1 when $h=1$, formalizes these statements with the help of the TC ordering, thereby confirming the intuition that the parameter β of PMM($\beta,\boldsymbol{p}$) indeed constitutes a measure of the strength of temporal correlations.

Theorem 5.1: *Assume for each $k=1,2$ that the request stream $\boldsymbol{R}^{\beta_k}=\{R_t^{\beta_k},\ t=0,1,\ldots\}$ is modeled according to the stationary PMM($\beta_k,\boldsymbol{p}$) for some pmf $\boldsymbol{p}$ on $\mathcal{N}$. If $0<\beta_2<\beta_1$, then $\boldsymbol{R}^{\beta_1}\leq_{TC}\boldsymbol{R}^{\beta_2}$.*

The proof of this theorem relies on the following comparison of Markov chains under the supermodular ordering due to Bäuerle [6].

Theorem 5.2: *Let $\boldsymbol{X} = \{X_t,\ t = 0, 1, \ldots\}$ and $\boldsymbol{X}' = \{X'_t,\ t = 0, 1, \ldots\}$ be two stationary 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 $\boldsymbol{Q}(\gamma_0, \ldots, \gamma_n)$ by*

$$\begin{bmatrix} 1 - \sum_{j \neq 0} \gamma_j & \gamma_1 & \cdots & \gamma_n \\ \gamma_0 & 1 - \sum_{j \neq 1} \gamma_j & \cdots & \gamma_n \\ \vdots & \vdots & & \vdots \\ \gamma_0 & \gamma_1 & \cdots & 1 - \sum_{j \neq n} \gamma_j \end{bmatrix}. \quad (11)$$

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

A proof of Theorem 5.1. Fix $i = 1, \ldots, N$. Given a sequence $\boldsymbol{R}^\beta = \{R_t^\beta,\ t = 0, 1, \ldots\}$ modeled according to the stationary PMM($\beta, \boldsymbol{p}$), it follows from (47) (in Appendix) that the indicator sequence $\{V_t^\beta(i),\ t = 0, 1, \ldots\}$ associated with $\boldsymbol{R}^\beta$ is a Markov chain on $\{0, 1\}$ with

$$\begin{aligned} & \mathbf{P}\left[V_t^\beta(i) = 1 | V_0^\beta(i) = x_0, \ldots, V_{t-1}^\beta(i) = x_{t-1}\right] \\ = \quad & \beta p(i) + (1 - \beta) x_{t-1}, \quad t = 1, 2, \ldots \end{aligned}$$

for any $(x_0, \ldots, x_{t-1})$ in $\{0, 1\}^t$. 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 (11), by $\boldsymbol{P}^\beta(i) = \boldsymbol{Q}(\gamma_0, \gamma_1)$ 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 Markov chains $\{V_t^{\beta_1}(i),\ t = 0, 1, \ldots\}$ and $\{V_t^{\beta_2}(i),\ t = 0, 1, \ldots\}$ have transition matrices $\boldsymbol{P}^{\beta_1}(i) = \boldsymbol{Q}(\gamma_0, \gamma_1)$ and $\boldsymbol{P}^{\beta_2}(i) = \boldsymbol{Q}(c\gamma_0, c\gamma_1)$, 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 5.2, 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 $i = 1, \ldots, N$, whence $\boldsymbol{R}^{\beta_1} \leq_{TC} \boldsymbol{R}^{\beta_2}$. ■

VI. Least-Recently-Used stack model

The Least-Recently-Used stack model (LRUSM) has long been known to be a good model for generating sequences of requests whose statistical properties match those of observed reference streams [14, 31].

A. LRU stack and stack distance

With $\Lambda(\mathcal{N})$ denoting the set of all permutations of the N distinct documents $\{1, \ldots, N\}$, an element of $\Lambda(\mathcal{N})$ can be viewed as an ordered sequence of N distinct elements drawn from the set $\{1, \ldots, N\}$. It is convenient to picture such an element $\Omega = (\Omega(1), \ldots, \Omega(N))$ of $\Lambda(\mathcal{N})$ as a stack with $\Omega(1)$ in the top position, followed by $\Omega(2), \ldots, \Omega(N)$, in that order.

Given an initial stack Ω_0, 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 $t = 0, 1, \ldots$, the stack Ω_{t+1} is given by

$$\Omega_{t+1}(k) = \begin{cases} \Omega_t(D_t) & \text{if } k = 1 \\ \Omega_t(k-1) & \text{if } k = 2, \ldots, D_t \\ \Omega_t(k) & \text{if } k = D_t + 1, \ldots, N \end{cases} \quad (12)$$

where D_t denotes the position of the document R_t in the stack Ω_t, i.e., the rv D_t is the unique element of $\{1, \ldots, N\}$ such that

$$\Omega_t(D_t) = R_t. \quad (13)$$

In words, the stack Ω_{t+1} at time $t+1$ is obtained by moving the document $\Omega_t(D_t) = R_t$ up to the highest position (i.e., position 1) and shifting the documents $\Omega_t(1), \ldots, \Omega_t(D_t - 1)$ down by one position while the positions of 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 an initial stack Ω_0 in $\Lambda(\mathcal{N})$, with any sequence of $\{1, \ldots, N\}$-valued rvs $\{D_t,\ t = 0, 1, \ldots\}$, the stack operation (12) can be used to recursively generate a sequence of $\Lambda(\mathcal{N})$-valued rvs $\{\Omega_t,\ t = 0, 1, \ldots\}$. A request stream $\boldsymbol{R}$ is now readily extracted from this stack sequence via (13), i.e., we have

$$R_t = \Omega_t(D_t) = \Omega_{t+1}(1), \quad t = 0, 1, \ldots. \quad (14)$$

It is plain that the rvs $\{D_t,\ t = 0, 1, \ldots\}$ constitute the stack distance sequence associated with the request stream $\boldsymbol{R}$ defined at (14).

The stack and distance introduced above are often referred to as LRU stack and distance, respectively, in reference to the popular Least-Recently-Used (LRU) policy according to which the document to be evicted from the cache is the one which has been requested the least recently at the time of replacement. The dynamics of the LRU policy are best described through the notion of LRU stack and distance, with the resulting stack implementation of LRU being one of the factors behind its popularity.

B. The LRU stack model

The duality between streams of requests and stack distances embedded in (12)-(14) can be exploited to define correlated sequences of requests. 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.* $\{1, \ldots, N\}$-valued rvs distributed according to the pmf $\boldsymbol{a}$, i.e.,

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

given some arbitrary initial stack Ω_0 in $\Lambda(\mathcal{N})$.

Throughout we assume that the rv Ω_0 is independent of the stack distances $\{D_t, t = 1, 2, \ldots\}$, and uniformly distributed over $\Lambda(\mathcal{N})$. In that case, 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 request model is denoted by LRUSM($\boldsymbol{a}$).

The popularity pmf of the LRUSM is discussed in Proposition 6.1; a proof can be found in [37].

Proposition 6.1: *Assume the request stream* $\boldsymbol{R^a} = \{R_t^{\boldsymbol{a}},\ t = 0, 1, \ldots\}$ *to be modeled according to the stationary LRUSM(*$\boldsymbol{a}$*). If* $a_N > 0$*, then for each* $i = 1, \ldots, N$*, it holds that*

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

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 [3, 10].

C. Temporal correlations in LRUSM

The temporal correlations exhibited by the LRUSM are captured through the TC ordering as indicated by the next result.

Theorem 6.2: *Assume the request stream* $\boldsymbol{R^a} = \{R_t^{\boldsymbol{a}},\ t = 0, 1, \ldots\}$ *to be modeled according to the stationary LRUSM(*$\boldsymbol{a}$*) with stack distance pmf* $\boldsymbol{a}$ *satisfying*

$$a_1 \geq a_2 \geq \ldots \geq a_N > 0. \tag{15}$$

Then, it holds that $\hat{\boldsymbol{R}}^{\boldsymbol{a}} \leq_{TC} \boldsymbol{R^a}$ *where* $\hat{\boldsymbol{R}}^{\boldsymbol{a}}$ *is the independent version of* $\boldsymbol{R^a}$*.*

The proof of Theorem 6.2 is rather lengthy and is available in [37]. By virtue of Proposition 6.1, the independent version $\hat{\boldsymbol{R}}^{\boldsymbol{a}}$ of the stationary LRUSM($\boldsymbol{a}$) is simply the IRM with uniform popularity pmf $\boldsymbol{u} = (\frac{1}{N}, \ldots, \frac{1}{N})$. Moreover, it is not hard to see that the stationary LRUSM($\boldsymbol{u}$) indeed coincides with the IRM with uniform popularity pmf $\boldsymbol{u}$. Thus, under (15) we have

$$\boldsymbol{R^u} \leq_{TC} \boldsymbol{R^a}.$$

VII. Working set size

In the following two sections, we show how comparison in the TC ordering translates into comparisons of some well-established locality of reference metrics, namely, the working set size and the inter-reference time.

The working set model was introduced by Denning [15] and some of its properties are discussed in [16]. 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 $\tau = 1, 2, \ldots$, the working set $W(t, \tau; \boldsymbol{R})$ of length τ at time t is the set of *distinct* documents which have occurred amongst the past τ consecutive requests $R_{(t-\tau+1)^+}, \ldots, R_t$.[3] The size of the working set $W(t, \tau; \boldsymbol{R})$ is denoted by $S(t, \tau; \boldsymbol{R})$.

A basic quantity of interest associated with the working set size is its long-run average defined by

$$\hat{S}(\tau; \boldsymbol{R}) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} S(t, \tau; \boldsymbol{R}) \quad a.s. \tag{16}$$

[3]For any x in $\mathbb{R}$, we set $x^+ = \max(0, x)$.

for each $\tau = 1, 2, \ldots$. The next lemma identifies conditions on the request stream $\boldsymbol{R}$ for the limits (16) to exist; its proof can be found in [33].

Lemma 7.1: *Assume that the request stream* $\boldsymbol{R} = \{R_t,\ t = 0, 1, \ldots\}$ *couples with a stationary sequence of* $\mathcal{N}$*-valued rvs* $\tilde{\boldsymbol{R}} = \{\tilde{R}_t,\ t = 0, 1, \ldots\}$*. Then, there exists an* $\{1, \ldots, \tau\}$*-valued rv* $S(\tau; \boldsymbol{R})$ *such that*

$$S(t, \tau; \boldsymbol{R}) \Longrightarrow_t S(\tau; \boldsymbol{R}), \quad \tau = 1, 2, \ldots. \tag{17}$$

and the a.s. limits (16) exist. If the stationary sequence $\tilde{\boldsymbol{R}}$ *is also ergodic, then*

$$\hat{S}(\tau; \boldsymbol{R}) = \mathbf{E}\left[S(\tau; \boldsymbol{R})\right], \quad \tau = 1, 2, \ldots. \tag{18}$$

The rv $S(\tau; \boldsymbol{R})$ at (17) can be viewed as the number of distinct documents in τ consecutive requests in the steady state. We expect that the stronger the strength of temporal correlations in the stream of requests, the smaller the working set size. The next result shows that such comparisons can indeed be formalized with the help of the TC ordering.

Theorem 7.2: *For request streams* $\boldsymbol{R}^1 = \{R_t^1,\ t = 0, 1, \ldots\}$ *and* $\boldsymbol{R}^2 = \{R_t^2,\ t = 0, 1, \ldots\}$ *such that* $\boldsymbol{R}^1 \leq_{TC} \boldsymbol{R}^2$*, it holds that*

$$\mathbf{E}\left[S(t, \tau; \boldsymbol{R}^2)\right] \leq \mathbf{E}\left[S(t, \tau; \boldsymbol{R}^1)\right], \quad t, \tau = 1, 2, \ldots \tag{19}$$

In addition, if for each $k = 1, 2$*, the request stream* $\boldsymbol{R}^k$ *couples with a stationary and ergodic sequence of* $\mathcal{N}$*-valued rvs* $\tilde{\boldsymbol{R}}^k = \{\tilde{R}_t^k,\ t = 0, 1, \ldots\}$*, then*

$$\hat{S}(\tau; \boldsymbol{R}^2) \leq \hat{S}(\tau; \boldsymbol{R}^1), \quad \tau = 1, 2, \ldots \tag{20}$$

where for each $k = 1, 2$*,* $\hat{S}(\tau; \boldsymbol{R}^k)$ *is the average working set size of the request stream* $\boldsymbol{R}^k$*.*

A proof of Theorem 7.2 is given in Appendix II.

VIII. Inter-reference time

The notion of inter-reference time in the stream of requests has recently received some attention as a way of characterizing locality of reference [18, 19, 27].

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

$$T(t; \boldsymbol{R}) := \inf\{\tau = 1, 2, \ldots, t : R_t = R_{t-\tau}\} \tag{21}$$

with the convention that $T(t; \boldsymbol{R}) = t+1$ if $R_{t-\tau} \neq R_t$ for all $\tau = 1, \ldots, t$.

Lemma 8.1: *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, there exists an* $\{1, 2, \ldots\}$*-valued rv* $T(\boldsymbol{R})$ *such that*

$$T(t; \boldsymbol{R}) \Longrightarrow_t T(\boldsymbol{R}). \tag{22}$$

The steady state inter-reference time $T(\boldsymbol{R})$ describes the time between two consecutive requests for the same document.

Our main comparison result for inter-reference times in the steady state is given in terms of the convex ordering[4] [29]:

Theorem 8.2: *Assume that for each $k = 1, 2$, the request stream $\boldsymbol{R}^k$ is asymptotically stationary, i.e., $\{R^k_{t+\ell},\ t = 0, 1, \ldots\} \Longrightarrow_\ell \{\tilde{R}^k_t,\ t = 0, 1, \ldots\}$ where $\tilde{\boldsymbol{R}}^k = \{\tilde{R}^k_t,\ t = 0, 1, \ldots\}$ is a stationary sequence of $\mathcal{N}$-valued rvs. If $\boldsymbol{R}^1 \leq_{TC} \boldsymbol{R}^2$, then it holds that*

$$T(\boldsymbol{R}^1) \leq_{cx} T(\boldsymbol{R}^2). \tag{23}$$

A proof of Lemma 8.1 is available in [33] while a proof of Theorem 8.2 is given in Appendix III. Theorem 8.2 states that the stronger the temporal correlations, the more variable the inter-reference time!

IX. The miss rate and its folk theorem

The *miss rate* of a caching policy is defined as the long-term frequency of the event that the requested document is not found in the cache; it provides a measure of the effectiveness of the caching policy. It is a commonly held belief that good caching takes advantage of locality of reference in that the stronger the strength of temporal correlations (i.e., the stronger locality of reference) in the stream of requests to the cache, the smaller the miss rate. We explore this "folk theorem" in the context of demand-driven caching which is briefly introduced in this section. Specific results and conjectures are provided in Section X under PMM, HOMM and LRUSM, and in Section XI under general Web request models exhibiting temporal correlations.

The system is composed of a server where a copy of each of the N cacheable documents is available, and of a cache of size M ($1 \leq M < N$). 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.

Let S_t denote the collection of documents in cache just before time t so that S_t is a subset of $\mathcal{N}$, and let U_t denote the decision to be performed according to the cache replacement policy π in force. Demand-driven caching is characterized by the dynamics

$$S_{t+1} = \begin{cases} S_t & \text{if } R_t \in S_t \\ S_t + R_t & \text{if } R_t \notin S_t, |S_t| < M \\ S_t - U_t + R_t & \text{if } R_t \notin S_t, |S_t| = M \end{cases} \tag{24}$$

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; and (iii) eviction is *mandatory* if the request R_t is not in cache S_t and the cache S_t is full.

The decisions $\{U_t,\ t = 0, 1, \ldots\}$ are determined through an eviction policy π. In most policies of interest, 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. 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 (24); 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

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

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

For example, under the random policy[5], we can take the cache state Ω_t to be the (unordered) set S_t of documents in the cache while under the LRU policy, the cache state Ω_t is is a permutation of the elements in S_t for all $t = 0, 1, \ldots$.

Under the cache replacement policy π, the miss rate $M_\pi(\boldsymbol{R})$ when the input to the cache is the request stream $\boldsymbol{R}$ is defined as the limiting constant

$$M_\pi(\boldsymbol{R}) = \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[R_\tau \notin S_\tau\right] \quad a.s. \tag{25}$$

whenever the limit exists. Almost sure convergence in (25) (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 $\boldsymbol{R}$ through the eviction policy π.

X. Folk theorems on various request models

A. PMM

The miss rates of demand-driven cache replacement policies under PMM have been previously considered in [2]. For particular caching policies such as LRU and FIFO, the miss rate under 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 (25) for the PMM($\beta, \boldsymbol{p}$) under the cache replacement policy π, we shall need the following definitions: For each $T = 1, 2, \ldots$, define

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

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

[5]Under the random policy, when the cache is full, the document to be evicted from the cache is selected randomly according to the uniform distribution.

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, from the definition of the rvs $\{\gamma(k), k = 1, 2, \ldots\}$ it follows that

$$\begin{aligned} \sum_{t=1}^{T} \mathbf{1}\left[R^{\beta}_t \notin S_t\right] &= \sum_{k=1}^{\lambda(T)} \mathbf{1}\left[R^{\beta}_{\gamma(k)} \notin S_{\gamma(k)}\right] \\ &= \sum_{k=1}^{\lambda(T)} \mathbf{1}\left[Y_{\gamma(k)} \notin S_{\gamma(k)}\right] \end{aligned} \tag{26}$$

for all $T = 1, 2, \ldots$, and the miss rate under PMM($\beta, \boldsymbol{p}$) is given by

$$\begin{aligned} & M_{\pi}(\boldsymbol{R}^{\beta}) \\ = & \lim_{T\to\infty} \left(\frac{\lambda(T)}{T}\right) \left(\frac{1}{\lambda(T)} \sum_{k=1}^{\lambda(T)} \mathbf{1}\left[Y_{\gamma(k)} \notin S_{\gamma(k)}\right]\right). \end{aligned} \tag{27}$$

By the Strong Law of Large Numbers, we get

$$\lim_{T\to\infty} \frac{\lambda(T)}{T} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\left[Z_t = 0\right] = \beta \quad a.s. \tag{28}$$

The limit of the second factor in (27) 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 $t + 1$ are 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

$$\begin{aligned} & \lim_{T\to\infty} \frac{1}{\lambda(T)} \sum_{k=1}^{\lambda(T)} \mathbf{1}\left[Y_{\gamma(k)} \notin S_{\gamma(k)}\right] \\ = & \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}\left[Y_{\gamma(k)} \notin S_{\gamma(k)}\right] = \hat{M}_{\pi}(\boldsymbol{p}) \end{aligned} \tag{29}$$

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 (27), (28) and (29) yields the expression for the miss rate of PMM($\beta, \boldsymbol{p}$) as

$$M_{\pi}(\boldsymbol{R}^{\beta}) = \beta \cdot \hat{M}_{\pi}(\boldsymbol{p}). \tag{30}$$

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 [33]. Equipped with the expression (30), we can now conclude to the following monotonicity result.

Theorem 10.1: *Assume that the cache replacement policy π satisfies Condition ($\star$) and that for each $k = 1, 2$, the request stream $\boldsymbol{R}^{\beta_k} = \{R^{\beta_k}_t,\ t = 0, 1, \ldots\}$ is modeled according to the stationary PMM($\beta_k, \boldsymbol{p}$) for some pmf $\boldsymbol{p}$ on $\mathcal{N}$. Then, $M_{\pi}(\boldsymbol{R}^{\beta_2}) \le M_{\pi}(\boldsymbol{R}^{\beta_1})$ whenever $0 < \beta_2 < \beta_1$.*

In view of Theorem 5.1, we conclude that the folk theorem on the miss rate indeed holds for stationary PMMs under any cache replacement policy which satisfies Condition ($\star$).

B. HOMM

Let $\boldsymbol{R}$ be HOMM($h, \boldsymbol{\alpha}, \boldsymbol{p}$) for some pmf vectors $\boldsymbol{p}$ on $\mathcal{N}$ and $\boldsymbol{\alpha}$ on $\{0, \ldots, h\}$, respectively. 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 $\alpha^c_k = c\alpha_k$ for each $k = 1, \ldots, h$, and $\beta^c = 1 - c(1 - \beta) = \beta + (1 - c)(1 - \beta)$. Obviously, $\beta^c \ge \beta$ while $\alpha^c_k \le \alpha_k$ 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 4.1, 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., $\boldsymbol{R}^c \le_{TC} \boldsymbol{R}$. Taking our cue from Theorem 10.1, we would then expect the inequality $M_{\pi}(\boldsymbol{R}) \le M_{\pi}(\boldsymbol{R}^c)$ to hold for some good caching policies. We summarize these expectations as the following conjecture:

Conjecture 10.2: *Assume the request stream $\boldsymbol{R}$ to be modeled according to HOMM($h, \boldsymbol{\alpha}, \boldsymbol{p}$). For some $0 < c < 1$, if $\boldsymbol{R}^c$ 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 $\boldsymbol{R}^c \le_{TC} \boldsymbol{R}$ holds. Furthermore, under some appropriate cache replacement policy π, it holds that $M_{\pi}(\boldsymbol{R}) \le M_{\pi}(\boldsymbol{R}^c)$.*

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 $N = 1000$ and let the input popularity pmf be the Zipf-like distribution $\boldsymbol{p}_{\alpha}$ with parameter $\alpha = 0.8$, i.e., for each $i = 1, \ldots, N$,

$$p(i) = p_{\alpha}(i) = \frac{i^{-\alpha}}{C_{\alpha}(N)} \quad \text{with} \quad C_{\alpha}(N) := \sum_{i=1}^{N} i^{-\alpha}. \tag{31}$$

The Zipf-like distribution has been found appropriate for modeling the popularity distributions of observed reference streams in several data sets [12]. We consider six different classes of HOMM, each with different history window size $h = 1, 5, 10, 50, 100$ and 500. In each class, the input stream $\boldsymbol{R}^{\beta}$ (with $0 \le \beta \le 1$), 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 10.2 would require that the mapping $\beta \to M_{\mathrm{LRU}}(\boldsymbol{R}^{\beta})$ be increasing.

From Figure 1, the miss rate is indeed found to be increasing as the parameter β increases for all cases and for all cache sizes.[6] When $h = 1$, HOMM reduces to PMM and the

[6]Although parameters used in this simulation may not be representative of realistic situations, this simple example serves to establish the trend expected in Conjecture 10.2.

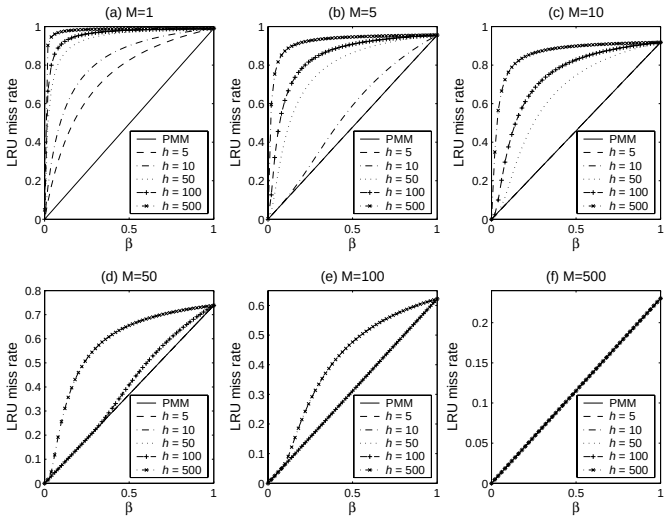

Fig. 1. LRU miss rates for various cache sizes when the input to the cache is the HOMM($h, \boldsymbol{\alpha}_h(\beta), \boldsymbol{p}_{0.8}$) with $\boldsymbol{\alpha}_h(\beta) = (\beta, \frac{1-\beta}{h}, \ldots, \frac{1-\beta}{h})$

results here confirm the validity of the expression (30) and of Theorem 10.1. 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 form of insensitivity of the LRU miss rate under the HOMM to the history window size h and to the pmf $\boldsymbol{\alpha}$.

C. LRUSM

According to Theorem 6.2, the stationary LRUSM($\boldsymbol{a}$) with stack distance pmf $\boldsymbol{a}$ satisfying condition (15) has stronger strength of temporal correlations than the stationary LRUSM($\boldsymbol{u}$). In the vein of Theorem 5.1, 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 6.2 suggests that this could happen when the pmf $\boldsymbol{a}$ is more skewed toward the smaller values of stack distance than the pmf $\boldsymbol{b}$, or equivalently, that the components of $\boldsymbol{b}$ are more balanced than the components of $\boldsymbol{a}$. The skewness in pmfs is naturally captured through the notion of *majorization* [23]: 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

$$\sum_{i=1}^{n} x_{[i]} \leq \sum_{i=1}^{n} y_{[i]},\ n = 1, \ldots, N-1,\ \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i \quad (32)$$

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. It is well known that $\boldsymbol{u} \prec \boldsymbol{a}$ for any pmf $\boldsymbol{a}$ on $\mathcal{N}$. With this notion, we can now state the following conjecture.

Conjecture 10.3: *Consider request streams $\boldsymbol{R}^{\boldsymbol{a}}$ and $\boldsymbol{R}^{\boldsymbol{b}}$ which are modeled according to the stationary LRUSM($\boldsymbol{a}$) and LRUSM($\boldsymbol{b}$), respectively. If both pmfs $\boldsymbol{a}$ and $\boldsymbol{b}$ satisfy (15) with $\boldsymbol{b} \prec \boldsymbol{a}$, then the comparison $\boldsymbol{R}^{\boldsymbol{b}} \leq_{TC} \boldsymbol{R}^{\boldsymbol{a}}$ holds.*

When both pmfs $\boldsymbol{a}$ and $\boldsymbol{b}$ satisfy (15), the conditions (32) for the majorization comparison $\boldsymbol{b} \prec \boldsymbol{a}$ to hold reduce to

$$\sum_{i=1}^{n} b_i \leq \sum_{i=1}^{n} a_i, \quad n = 1, \ldots, N-1. \quad (33)$$

This condition is a formalization of the statement that the pmf $\boldsymbol{a}$ is more skewed toward the smaller values of stack distance than the pmf $\boldsymbol{b}$.[7]

To glean evidence in favor of Conjecture 10.3, consider the LRU policy. The miss rate of the LRU policy under LRUSM($\boldsymbol{a}$) is given [14, p. 277] by

$$M_{\mathrm{LRU}}(\boldsymbol{R}^{\boldsymbol{a}}) = \mathbf{P}\left[D_t > M\right] = \sum_{k=M+1}^{N} a_k \quad (34)$$

when the cache size is M. Combining (33) and (34), we conclude that $M_{\mathrm{LRU}}(\boldsymbol{R}^{\boldsymbol{a}}) \leq M_{\mathrm{LRU}}(\boldsymbol{R}^{\boldsymbol{b}})$ for two LRUSM request streams $\boldsymbol{R}^{\boldsymbol{a}}$ and $\boldsymbol{R}^{\boldsymbol{b}}$ satisfying the conditions of Conjecture 10.3. This is of course the desired inequality expressing the folk theorem for miss rates under the LRU policy which would be expected if Conjecture 10.3 were to hold.

XI. The working set (WS) 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 $W(-1, \tau; \boldsymbol{R}) = \phi$. The number of documents in the cache at time t under the WS algorithm is the number of distinct documents in $W(t-1, \tau; \boldsymbol{R})$ which is the working set size $S(t-1, \tau; \boldsymbol{R})$. 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 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

$$S_t = \{i = 1, \ldots, N : i \in \Omega_t\} = W(t-1, \tau; \boldsymbol{R})$$

for $t = 0, 1, \ldots$. 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.

A. The miss rate under the WS algorithm

The miss rate of the WS algorithm with length τ is defined as in the case of demand-driven caching. For the input stream $\boldsymbol{R} = \{R_t,\ t = 0, 1, \ldots\}$, it is given by the a.s. limit

$$M_{\mathrm{WS}}(\boldsymbol{R}; \tau) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\left[R_t \notin S_t\right]$$

[7] The condition (33) is equivalent to the usual stochastic ordering $\boldsymbol{a} \leq_{st} \boldsymbol{b}$ between the pmfs $\boldsymbol{a}$ and $\boldsymbol{b}$ [29].

whenever this limit exists. Observe that a miss occurs at time t when the document R_t is not in the working set $W(t-1,\tau;\boldsymbol{R})$. Therefore, with $\{V_t(i),\ t=0,1,\ldots\}$, $i=1,\ldots,N$, denoting the indicator sequences (5) associated with $\boldsymbol{R}$, whenever $\tau \le t$, we find

$$\begin{aligned}\mathbf{1}\left[R_t \notin S_t\right] &= \mathbf{1}\left[R_t \notin W(t-1,\tau;\boldsymbol{R})\right] \\ &= \mathbf{1}\left[R_t \notin \{R_{t-\tau},\ldots,R_{t-1}\}\right] \\ &= \sum_{i=1}^{N} \mathbf{1}\left[R_t=i\right] \prod_{\ell=1}^{\tau} \mathbf{1}\left[R_{t-\ell}\neq i\right] \\ &= \sum_{i=1}^{N} V_t(i) \prod_{\ell=1}^{\tau} (1-V_{t-\ell}(i)) \\ &= \sum_{i=1}^{N} g(V_{t-\tau}(i),\ldots,V_t(i)) \qquad (35)\end{aligned}$$

where for $(x_0,\ldots,x_\tau)\in\mathbb{R}^{\tau+1}$, we have set

$$g(x_0,\ldots,x_\tau) = x_\tau \prod_{\ell=0}^{\tau-1}(1-x_\ell). \qquad (36)$$

Consequently,

$$\begin{aligned}&M_{\mathrm{WS}}(\boldsymbol{R};\tau) \qquad (37)\\ =\ &\lim_{T\to\infty} \frac{1}{T-\tau+1} \sum_{t=\tau}^{T}\sum_{i=1}^{N} g(V_{t-\tau}(i),\ldots,V_t(i)) \quad a.s.\end{aligned}$$

provided the limit exists. The next lemma gives conditions for the existence of the limit (37); a proof is available in [33].

Lemma 11.1: *Fix $\tau=1,2,\ldots$. If the request stream $\boldsymbol{R}=\{R_t,\ t=0,1,\ldots\}$ couples with a stationary and ergodic sequence of $\mathcal{N}$-valued rvs $\tilde{\boldsymbol{R}}=\{\tilde{R}_t,\ t=0,1,\ldots\}$, then the a.s. limit (37) exists and is given by*

$$M_{\mathrm{WS}}(\boldsymbol{R};\tau) = \lim_{t\to\infty} \sum_{i=1}^{N} \mathbf{E}\left[g(V_{t-\tau}(i),\ldots,V_t(i))\right] \quad a.s. \qquad (38)$$

B. On the folk theorem under the WS algorithm

The folk theorem to the effect that the stronger the temporal correlations, the smaller the miss rate, holds if we can show that

$$M_{\mathrm{WS}}(\boldsymbol{R}^2;\tau) \le M_{\mathrm{WS}}(\boldsymbol{R}^1;\tau) \quad \text{if} \quad \boldsymbol{R}^1 \le_{TC} \boldsymbol{R}^2. \qquad (39)$$

From the definitions of the TC and sm orderings, we see from (38) that establishing (39) can be achieved by showing that the mapping g given in (36) is submodular.[8] We discuss these issues by first showing a positive result when $\tau=1$, and then providing counterexamples when $\tau>1$.

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 11.2: *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 $\boldsymbol{R}^1 \le_{TC} \boldsymbol{R}^2$, then it holds that*

$$\mathbf{P}\left[R_t^2 \notin S_t^2\right] \le \mathbf{P}\left[R_t^1 \notin S_t^1\right], \quad t=1,2,\ldots. \qquad (40)$$

Proof. Fix $k=1,2$. For each $t=1,2,\ldots$, we have from (35)-(36) that

$$\mathbf{1}\left[R_t^k \notin S_t^k\right] = \mathbf{1}\left[R_t^k \neq R_{t-1}^k\right] = \sum_{i=1}^{N} g(V_{t-1}^k(i), V_t^k(i)) \qquad (41)$$

with the mapping $g:\mathbb{R}^2\to\mathbb{R}$ being given by (36). Because the mapping $(x_0,x_1)\to x_0x_1$ is supermodular, the mapping $(x_0,x_1)\to -x_0x_1$ 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.

The assumption $\boldsymbol{R}^1 \le_{TC} \boldsymbol{R}^2$ implies the comparisons $\{V_t^1(i),\ t=0,1,\ldots\} \le_{sm} \{V_t^2(i),\ t=0,1,\ldots\}$ for each $i=1,\ldots,N$. The submodularity of g readily yields

$$\sum_{i=1}^{N} \mathbf{E}\left[g(V_{t-1}^2(i),V_t^2(i))\right] \le \sum_{i=1}^{N} \mathbf{E}\left[g(V_{t-1}^1(i),V_t^1(i))\right] \qquad (42)$$

for each $t=1,2,\ldots$, and the comparisons (40) follow from (41) and (42). ■

Thus, combining Lemma 11.1 and Theorem 11.2 we find that the folk theorem (39) indeed holds for $\tau=1$ whenever the request streams $\boldsymbol{R}^1$ and $\boldsymbol{R}^2$ couple with stationary and ergodic sequences.

When $\tau>1$, the folk theorem (39) does not necessarily hold. To construct a counterexample, we consider the situation where the PMM is taken to be the input to the cache. Then, the miss rate of the WS algorithm with length τ for PMM($\beta,\boldsymbol{p}$) [2] is given by

$$M_{\mathrm{WS}}(\beta,\boldsymbol{p};\tau) = \beta\sum_{i=1}^{N} p(i)(1-p(i))(1-\beta p(i))^{\tau-1}. \qquad (43)$$

From Section V, we expect that as the strength of temporal correlations increases, i.e., the value of the parameter β decreases, the miss rate $M_{\mathrm{WS}}(\beta,\boldsymbol{p};\tau)$ should decrease. To put it differently, the mapping $\beta\to M_{\mathrm{WS}}(\beta,\boldsymbol{p};\tau)$ should be increasing when the popularity pmf $\boldsymbol{p}$ is held fixed. That this may not always be so becomes clear when considering the uniform popularity pmf $\boldsymbol{u}=(\frac{1}{N},\ldots,\frac{1}{N})$.

Theorem 11.3: *Assume the input stream to be modeled according to PMM($\beta,\boldsymbol{u}$). Under the WS algorithm with length τ, the miss rate function $M_{\mathrm{WS}}(\beta,\boldsymbol{u};\tau)$ given in (43) is increasing in β when $\beta\le\frac{N}{\tau}$ and decreasing in β when $\beta>\frac{N}{\tau}$.*

[8] A function $\varphi:\mathbb{R}^n\to\mathbb{R}$ is said to be submodular if $-\varphi$ is supermodular.

Proof. When the PMM has the uniform popularity pmf $\boldsymbol{u}$, the expression (43) becomes

$$M_{\mathrm{WS}}(\beta, \boldsymbol{u}; \tau) = \beta \left(1 - \frac{1}{N}\right) \left(1 - \frac{\beta}{N}\right)^{\tau - 1}.$$

Differentiating this expression with respect to β yields

$$\frac{d}{d\beta} M_{\mathrm{WS}}(\beta, \boldsymbol{u}; \tau) = \left(1 - \frac{1}{N}\right) \left(1 - \frac{\beta}{N}\right)^{\tau - 2} \left(1 - \frac{\tau\beta}{N}\right).$$

Thus, the miss rate function $M_{\mathrm{WS}}(\beta, \boldsymbol{u}; \tau)$ 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, $\beta > \frac{N}{\tau}$. ■

Thus, under the PMM, 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.

XII. Concluding remarks

Here, we have attempted to model the (positive) temporal correlations present in streams of requests with the help of the TC ordering, an approach based on the concept of positive dependence called supermodular ordering. On the positive side, we show that (i) the comparison under the TC ordering is compatible with comparisons of some well-known metrics of locality of reference, namely, the working set size and the inter-reference time; (ii) this TC ordering captures to a certain extent the strength of temporal correlations present in Web request models which are expected to exhibit temporal correlations, e.g., the HOMM, PMM and LRUSM; and (iii) the folk theorem on miss rates holds under PMM input to the cache for a large class of replacement policies. These preliminary results suggest that the TC ordering might indeed provide a useful way to compare streams of requests in terms of their locality of reference, especially when correlations are present.

However, the folk theorem fails to hold as evidenced by the counterexample found for the Working Set algorithm. This state of affairs is certainly disappointing and provides yet another confirmation that locality of reference, while an extensively studied (and allegedly understood) notion, still remains elusive in some of its characteristics. That locality of reference is about positive correlations, there is little doubt about it! The TC ordering captures only some aspects of the notion but undoubtedly there is more to it!

References

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

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

[3] P. Barford and M. Crovella, "Generating representative Web workloads for network and server performance evaluation," in *Proceedings of the 1998 ACM SIGMETRICS Conference*, June 1998, Madison (WS).

[4] 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.

[5] N. Bäuerle, "Inequalities for stochastic models via supermodular orderings," *Communication in Statistics – Stochastic Models* **13** (1997), pp. 181–201.

[6] N. Bäuerle, "Monotonicity results for $MR|GI|1$ queues," *Journal of Applied Probability* **34** (1997), pp. 514–524.

[7] N. Bäuerle and T. Rolski, "A monotonicity result for the work-load in Markov-modulated queues," *Journal of Applied Probability* **35** (1998), pp. 741–747.

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

[9] P. Billingsley, *Convergence of Probability Measures* John Wiley & Sons, New York (NY), 1968.

[10] M. Busari and C. Williamson, "Prowgen: a synthetic workload generation tool for simulation evaluation of Web proxy caches," *Computer Networks* **38** (2002), pp. 779–794.

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

[12] 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.

[13] 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** (2004), pp. 557–574.

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

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

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

[17] M. Deshpande and G. Karypis, "Selective Markov models for predicting Web-page accesses," in *Proceedings of SIAM Data Mining Conference 2001*, Chicago (IL), April 2001.

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

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

[20] 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.

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

[22] A.M. Makowski and S. Vanichpun, "Comparing strength of locality of reference – Popularity, majorization, and some folk theorems for miss rates and the output of cache," in *Performance Evaluation and Planning Methods for the Next Generation Internet*, A. Girard, B. Sansó and F. J. Vázquez-Abad, Editors, Kluwer Academic Press, 2005.

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

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

[25] L.E. Meester and J.G. Shanthikumar, "Regularity of stochastic processes: A theory of directional convexity," *Probability in the Engineering and Informational Sciences* **7** (1993), pp. 343–360.

[26] A. Müller and D. Stoyan, *Comparison Methods for Stochastic Models and Risks*, John Wiley & Sons, Chichester, 2002.

[27] V. Phalke and B. Gopinath, "An inter-reference gap model for temporal locality in program behavior," in *Proceedings of the 1995 ACM SIGMETRICS Conference*, May 1995, pp. 291–300.

[28] K. Psounis, A. Zhu, B. Prabhakar and R. Motwani, "Modeling correlations in Web-traces and implications for designing replacement policies," *Computer Networks* **45** (2004), pp. 379–398.

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

[30] M. Shaked and J.G. Shanthikumar, "Supermodular stochastic orders and positive dependence of random vectors," *Journal of Multivarate Analysis* **61** (1997), pp. 86–101.

[31] G. Shedler and C. Tung, "Locality in page reference strings," *SIAM Journal of Computing* **1** (1972), pp. 218–241.

[32] 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** (1993), pp. 508–512.
[33] S. Vanichpun, *Comparing Strength of Locality of Reference: Popularity, Temporal Correlations, and Some Folk Theorems for the Miss Rates and Outputs of Caches*, Ph.D. Dissertation, Department of Electrical and Computer Engineering, University of Maryland, College Park (MD), May 2005. Available as ISR Technical Report No. PhD 2005-1.
[34] 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.
[35] S. Vanichpun and A.M. Makowski, "Comparing strength of locality of reference – Popularity, majorization, and some folk theorems," in *Proceedings of IEEE INFOCOM 2004*, Hong Kong (PRC), April 2004.
[36] S. Vanichpun and A.M. Makowski, "The output of a cache under the Independent Reference Model – Where did the locality of reference go?," in *Proceedings of the 2004 ACM SIGMETRICS-PERFORMANCE Conference*, New York (NY), June 2004.
[37] S. Vanichpun and A.M. Makowski, "Positive dependence in the Least-Recently-Used stack model," in preparation (2005).

APPENDIX I
A PROOF OF THEOREM 4.1

By Proposition 3.2, it suffices to show for each $i = 1, \ldots, N$, that the indicator sequence $\{V_t(i),\ t = 0, 1, \ldots\}$ associated with the request stream $\boldsymbol{R}$ is PSMD. To do so, we construct another sequence of $\mathcal{N}$-valued rvs $\tilde{\boldsymbol{R}} = \{\tilde{R}_t,\ t = 0, 1, \ldots\}$ 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 (7) 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 $\{\tilde{R}_0, \ldots, \tilde{R}_{h-1}\}$. By construction, the process $\tilde{\boldsymbol{R}} = \{\tilde{R}_t,\ t = 0, 1, \ldots\}$ is an h^{th}-order Markov chain and with $\beta > 0$, we get

$$\{\tilde{R}_{t+\tau},\ t = 0, 1, \ldots\} \Longrightarrow_\tau \{R_t,\ t = 0, 1, \ldots\}. \tag{44}$$

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. 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 [29] if for each $t = 0, 1, \ldots$, the inequality

$$\begin{aligned} & \mathbf{P}\left[\tilde{V}_{t+1}(i) = 1 | \tilde{\boldsymbol{V}}^t(i) = \boldsymbol{x}^t\right] \\ \leq\ & \mathbf{P}\left[\tilde{V}_{t+1}(i) = 1 | \tilde{\boldsymbol{V}}^t(i) = \boldsymbol{y}^t\right] \end{aligned} \tag{45}$$

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

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

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

by independence of the rvs $\tilde{R}_0, \ldots, \tilde{R}_{h-1}$, and the inequality (45) is obtained for each $t = 0, 1, \ldots, h-2$. Next, for $t = h-1, h, \ldots$, and $\boldsymbol{x}^t = (x_0, \ldots, x_t)$ in $\{0, 1\}^{t+1}$, let $(i_0, \ldots, i_t)$ be an element in $\mathcal{N}^{t+1}$ with the property that for each $k = 0, \ldots, t$, $i_k = i$ if $x_k = 1$ and $i_k \neq i$ if $x_k = 0$. With such an element, we obtain from (8) that

$$\begin{aligned} & \mathbf{P}\left[\tilde{V}_{t+1}(i) = 1 | (\tilde{R}_0, \ldots, \tilde{R}_t) = (i_0, \ldots, i_t)\right] \\ =\ & \mathbf{P}\left[\tilde{R}_{t+1} = i | (\tilde{R}_0, \ldots, \tilde{R}_t) = (i_0, \ldots, i_t)\right] \\ =\ & \beta p(i) + \sum_{k=1}^{h} \alpha_k \mathbf{1}\left[i_{t+1-k} = i\right] \\ =\ & \beta p(i) + \sum_{k=1}^{h} \alpha_k x_{t+1-k}. \end{aligned} \tag{46}$$

Since (46) holds for any $(i_0, \ldots, i_t)$ in $\mathcal{N}^{t+1}$ satisfying the property above, a standard preconditioning argument readily yields

$$\mathbf{P}\left[\tilde{V}_{t+1}(i) = 1 | \tilde{\boldsymbol{V}}^t(i) = \boldsymbol{x}^t\right] = \beta p(i) + \sum_{k=1}^{h} \alpha_k x_{t+1-k}. \tag{47}$$

This last expression being monotone increasing in $\boldsymbol{x}^t = (x_0, \ldots, x_t)$, we obtain the inequality (45) for each $t = h, h+1, \ldots$.

Thus, the inequalities (45) hold for *all* $t = 0, 1, \ldots$. This implies that the sequence $\{\tilde{V}_t(i),\ t = 0, 1, \ldots\}$ is CIS, whence indeed PSMD by Theorem 2.6, i.e.,

$$\{\hat{\tilde{V}}_t(i),\ t = 0, 1, \ldots\} \leq_{sm} \{\tilde{V}_t(i),\ t = 0, 1, \ldots\} \tag{48}$$

where $\{\hat{\tilde{V}}_t(i),\ t = 0, 1, \ldots\}$ is the independent version of $\{\tilde{V}_t(i),\ t = 0, 1, \ldots\}$. Now, recalling (44), it is plain that

$$\{\hat{\tilde{V}}_{t+\tau}(i),\ t = 0, 1, \ldots\} \Longrightarrow_\tau \{\hat{V}_t(i),\ t = 0, 1, \ldots\} \tag{49}$$

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 [26, Thm. 3.9.8, p. 116], we conclude from (44), (48) and (49) 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 the proof is completed. ■

APPENDIX II
A PROOF OF THEOREM 7.2

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

$$\begin{aligned} S(t,\tau;\boldsymbol{R}) &= \sum_{i=1}^{N} \mathbf{1}\left[i \in \{R_{t-\tau+1},\ldots,R_t\}\right] \\ &= \sum_{i=1}^{N} (1 - \mathbf{1}\left[i \notin \{R_{t-\tau+1},\ldots,R_t\}\right]) \\ &= \sum_{i=1}^{N} (1 - \prod_{\ell=0}^{\tau-1} \mathbf{1}\left[R_{t-\ell} \neq i\right]) \\ &= \sum_{i=1}^{N} (1 - \prod_{\ell=0}^{\tau-1} (1 - V_{t-\ell}(i))). \\ &= \sum_{i=1}^{N} (1 - \psi(V_{t-\tau+1}(i),\ldots,V_t(i))) \quad (50) \end{aligned}$$

where the mapping $\psi : \mathbb{R}^\tau \to \mathbb{R}$ given by

$$\psi(\boldsymbol{x}) = \prod_{i=1}^{\tau} (1 - x_i), \quad \boldsymbol{x} = (x_1,\ldots,x_\tau) \in \mathbb{R}^\tau, \qquad (51)$$

is a supermodular function [5, Lemma 2.1].

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, the inequality

$$\begin{aligned} & \mathbf{E}\left[\psi(V_{t-\tau+1}^1(i),\ldots,V_t^1(i))\right] \\ \leq\ & \mathbf{E}\left[\psi(V_{t-\tau+1}^2(i),\ldots,V_t^2(i))\right] \qquad (52) \end{aligned}$$

follows for all $i = 1,\ldots,N$. Combining inequalities (52) with (50) yields the comparison (19) 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 comparisons (19) for all $\tau = 1,2,\ldots$.

Next, fix $\tau = 1,2,\ldots$ and $k = 1,2$. Under the assumptions that the request stream $\boldsymbol{R}^k$ couples with a stationary and ergodic sequence of $\mathcal{N}$-valued rvs $\tilde{\boldsymbol{R}}^k$, Lemma 7.1 already yields the convergence

$$S(t,\tau;\boldsymbol{R}^k) \Longrightarrow_t S(\tau;\boldsymbol{R}^k). \qquad (53)$$

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 (53), it follows from [9, Thm. 5.4, p. 32] that

$$\lim_{t\to\infty} \mathbf{E}\left[S(t,\tau;\boldsymbol{R}^k)\right] = \mathbf{E}\left[S(\tau;\boldsymbol{R}^k)\right] = \hat{S}(\tau;\boldsymbol{R}^k) \qquad (54)$$

where the last equality is due to (18). Invoking (19) and (54), we obtain the comparisons (20) for each $\tau = 1,2,\ldots$. ■

APPENDIX III
A PROOF OF THEOREM 8.2

To establish Theorem 8.2, we shall rely on the following lemma whose proof is available in [33].

Lemma C.1 *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 $\boldsymbol{p}$. Then, it holds that*

$$\sum_{\tau=n}^{\infty} \mathbf{P}\left[T(\boldsymbol{R}) > \tau\right] = \sum_{i=1}^{N} \mathbf{P}\left[\tilde{R}_\ell \neq i, \ell = 0,\ldots,n-1\right] \qquad (55)$$

for $n = 1,2,\ldots$ and $\mathbf{E}\left[T(\boldsymbol{R})\right] = \sum_{\tau=0}^{\infty} \mathbf{P}\left[T(\boldsymbol{R}) > \tau\right] = N$.

Proof of Theorem 8.2. It is well known [29, Thm. 2.A.1, p. 57] that the comparison (23) between the $\{1,2,\ldots\}$-valued rvs $T(\boldsymbol{R}^1)$ and $T(\boldsymbol{R}^2)$ is equivalent to

$$\sum_{\tau=n}^{\infty} \mathbf{P}\left[T(\boldsymbol{R}^1) > \tau\right] \leq \sum_{\tau=n}^{\infty} \mathbf{P}\left[T(\boldsymbol{R}^2) > \tau\right] \qquad (56)$$

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

$$\mathbf{E}\left[T(\boldsymbol{R}^1)\right] = \mathbf{E}\left[T(\boldsymbol{R}^2)\right]. \qquad (57)$$

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 (5) associated with $\boldsymbol{R}^k$ and $\tilde{\boldsymbol{R}}^k$, respectively. From Lemma C.1, the expression (55) for $n = 1,2,\ldots$, can be rewritten as

$$\begin{aligned} & \sum_{\tau=n}^{\infty} \mathbf{P}\left[T(\boldsymbol{R}^k) > \tau\right] \\ =\ & \sum_{i=1}^{N} \mathbf{E}\left[\mathbf{1}\left[\tilde{R}_\ell^k \neq i, \ell = 0,\ldots,n-1\right]\right] \\ =\ & \sum_{i=1}^{N} \mathbf{E}\left[\prod_{\ell=0}^{n-1} (1 - \tilde{V}_\ell^k(i))\right] \\ =\ & \sum_{i=1}^{N} \mathbf{E}\left[\psi(\tilde{V}_0^k(i),\ldots,\tilde{V}_{n-1}^k(i))\right] \qquad (58) \end{aligned}$$

where the mapping $\psi : \mathbb{R}^n \to \mathbb{R}$ of the form (51) is a supermodular function.

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

$$\{V_{t+\ell}^k(i),\ t = 0,1,\ldots\} \Longrightarrow_\ell \{\tilde{V}_t^k(i),\ t = 0,1,\ldots\}, \qquad (59)$$

for each $i = 1,\ldots,N$. 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 convergence [26, Thm. 3.9.8, p. 116], it is now plain from (59) that

$$\{\tilde{V}_t^1(i),\ t = 0,1,\ldots\} \leq_{sm} \{\tilde{V}_t^2(i),\ t = 0,1,\ldots\}, \qquad (60)$$

for each $i = 1,\ldots,N$. In short, $\tilde{\boldsymbol{R}}^1 \leq_{TC} \tilde{\boldsymbol{R}}^2$ and the required condition (56) follows upon combining (60) with (58). Lastly, under the assumptions of the theorem, we recall from Lemma C.1 that $\mathbf{E}\left[T(\boldsymbol{R}^1)\right] = \mathbf{E}\left[T(\boldsymbol{R}^2)\right] = N$, and (57) holds. ■