Asymptotically Exact TTL-Approximations of the Cache Replacement Algorithms LRU(m) and h-LRU

Nicolas Gast, Benny Van Houdt

Asymptotically Exact TTL-Approximations of the Cache Replacement Algorithms LRU(m) and h-LRU

Nicolas Gast
Inria, France

Benny Van Houdt
University of Antwerp, Belgium

***Abstract*—Computer system and network performance can be significantly improved by caching frequently used information. When the cache size is limited, the cache replacement algorithm has an important impact on the effectiveness of caching. In this paper we introduce time-to-live (TTL) approximations to determine the cache hit probability of two classes of cache replacement algorithms: the recently introduced h-LRU and LRU(m). These approximations only require the requests to be generated according to a general Markovian arrival process (MAP). This includes phase-type renewal processes and the IRM model as special cases.**

We provide both numerical and theoretical support for the claim that the proposed TTL approximations are asymptotically exact. In particular, we show that the transient hit probability converges to the solution of a set of ODEs (under the IRM model), where the fixed point of the set of ODEs corresponds to the TTL approximation. We further show, by using synthetic and trace-based workloads, that h-LRU and LRU(m) perform alike, while the latter requires less work when a hit/miss occurs. We also show that, as opposed to LRU, h-LRU and LRU(m) are sensitive to the correlation between consecutive inter-request times.

I. Introduction

Caches form a key component of many computer networks and systems. A large variety of cache replacement algorithms has been introduced and analyzed over the last few decades. A lot of the initial work existed in deriving explicit expressions for the cache content distribution using a Markov chain analysis [1]. This approach, however, is not always feasible: Even if explicit expressions can be obtained, they are often only applicable to analyze small caches, because of the time it takes to evaluate them. This gave rise to various approximation algorithms to compute cache hit probabilities and most notably to time-to-live (TTL) approximations.

The first TTL approximation was introduced for the least recently used (LRU) policy under the Independent reference model (IRM) by Che *et al.* in [6]. The main idea behind this approximation is that an LRU cache behaves similar to a TTL cache. In a TTL cache, when an item enters the cache, it sets a deterministic timer with initial value T. When this timer expires the item is removed from the cache. If an item is requested before its timer expires, its timer is reset to T. When T is fixed, an item with popularity p_k is present in the cache at a random point in time with probability $1-e^{-p_kT}$ and $\sum_{k=1}^{N} 1-e^{-p_kT}$ is the average number of items in the cache. The Che approximation [6] consists in approximating an LRU cache of size m by a TTL cache with characteristic time $T(m)$, where $T(m)$ is the unique solution of the fixed point equation

$$m = \sum_{k=1}^{N}(1-e^{-p_kT}). \tag{1}$$

The above TTL approximation for LRU can easily be generalized to renewal requests as well as to other simple variations of LRU and RANDOM under both IRM and renewal requests, as well as to certain network setups [3], [7], [12], [13]. All of these TTL approximations have been shown to be (very) accurate by means of numerical examples, but except for LRU in [8], no theoretical support was provided thus far.

In this paper we introduce TTL approximations for two classes of cache replacement algorithms that are variants of LRU. The first class, called LRU($\mathbf{m}$), dates back to the 1980s [1], while the second, called h-LRU, was recently introduced in [12]. In fact, a TTL approximation for h-LRU was also introduced in [12], but this approximation relies on an additional approximation of independence between the different lists when $h > 2$. As we will show in the paper, this implies that the approximation error does not reduce to zero as the cache becomes large.

In this paper we make the following contributions:

- We present a TTL approximation for LRU($\mathbf{m}$) and h-LRU that is valid when the request process of an item is a Markovian arrival process (MAP). This includes any phase-type renewal process and the IRM model. In the special case of the IRM model, we derive simple closed-form expressions for the fixed point equations.
- Our TTL approximation for h-LRU can be computed in linear time in h and appears to be asymptotically exact as the cache size grows, in contrast to the TTL approximation in [12] for $h > 2$. Numerical results for the TTL approximation for LRU($\mathbf{m}$) also suggest that it is asymptotically exact.
- We prove that, under the IRM model, the transient behavior of both h-LRU and LRU($\mathbf{m}$) converges to the unique solution of a system of ODEs as the cache size goes to infinity. Our TTL approximations correspond to the unique fixed point of the associated systems of ODEs. This provides additional support for the claim that our TTL approximations are asymptotically exact and is the main technical contribution of the paper.

- We validate the accuracy of the TTL approximation. We show that h-LRU and LRU($\mathbf{m}$) perform alike in terms of the hit probability under both synthetic and trace-based workloads, while less work is required for LRU($\mathbf{m}$) when a hit/miss occurs.
- We indicate that both h-LRU and LRU($\mathbf{m}$) can exploit correlation in consecutive inter-request times of an item, while the hit probability of LRU is insensitive to this type of correlation.

The paper is structured as follows. We recall the definitions of LRU($\mathbf{m}$) and h-LRU in Section II. We show how to build and solve the TTL-approximation for LRU($\mathbf{m}$) in Section III-A, and for h-LRU in Section III-B. We demonstrate the accuracy of the TTL-approximation for any finite time period in Section IV. We compare LRU($\mathbf{m}$) and h-LRU in Section V, by using synthetic data and real traces. We conclude in Section VI.

II. Replacement Algorithms

We consider two families of cache replacement algorithms: h-LRU, introduced and called k-LRU in [12], and LRU($\mathbf{m}$), introduced in [1], [9]. Both operate on a cache that can store up to m items and both are variants of LRU, which replaces the least-recently-used item in the cache. One way to regard LRU is to think of the cache as an ordered list of m items, where the i-th position is occupied by the i-th most-recently-used item. When a miss occurs, the item in the last position of the list is removed and the requested item is inserted at the front of the list. If a hit occurs on the item in position i, item i moves to the front of the list, meaning the items in position 1 to $i-1$ move back one position.

The h-LRU replacement algorithm: h-LRU manages a cache of size m by making use of $h-1$ additional virtual lists of size m (called list 1 to list $h-1$) in which only meta-data is stored and one list of size m that correspond to the actual cache (called list h). Each list is ordered, and the item in the ith position of list ℓ is the ith most-recently-used item among the items in list ℓ. When item k is requested, two operations are performed:

- For each list ℓ in which item k appears (say in a position i), the item k moves to the first position of list ℓ and the items in positions 1 to $i-1$ move back one position.
- For each list ℓ in which item k does not appear *but appears in list* $\ell-1$, item k is inserted in the first position of list ℓ, all other items of list ℓ are moved back one position and the item that was in position m of list ℓ is discarded from list ℓ.

List 1 of h-LRU behaves exactly as LRU, except that only the meta-data of the items is stored. Also, an item can appear in any subset of the h lists at the same time. This implies that a request can lead to as many as h list updates. Note that there is no need for all of the h lists to have the same size m.

The LRU($\mathbf{m}$) replacement algorithm: LRU($\mathbf{m}$) makes use of h lists of sizes $m_1, \ldots, m_h$, where the first few lists may be virtual, i.e., contain meta-data only. If the first v lists are virtual we have $m_{v+1} + \cdots + m_h = m$ (that is, only the items in lists $v+1$ to h are stored in the cache). With LRU($\mathbf{m}$) each item appears in at most one of the h lists at any given time. Upon each request of an item:

- If this item is not in the cache, it moves to the first position of list 1 and all other items of list 1 move back one position. The item that was in position m_1 of list 1 is discarded.
- If this item is in position i of a list $\ell < h$, it is removed from list ℓ and inserted in the first position of list $\ell+1$. All other items of list $\ell+1$ move back one position and the item in the last position of list $\ell+1$ is removed from list $\ell+1$ and inserted in the first position of list ℓ. All previous items from position 1 to $i-1$ of list ℓ move back one position.
- If this item is in position i of list h, then this item moves to the first position of list h. All items that are in position 1 to $i-1$ of list h move back one position.

When using only one list, LRU($\mathbf{m}$) coincides with LRU, and therefore with 1-LRU.

III. TTL-approximations

A. TTL-approximation for LRU(m)

1) IRM setting: Under the IRM model the string of requested items is a set of i.i.d. random variables, where item k is requested with probability p_k. As far as the hit probability is concerned this corresponds to assuming that item k is requested according to a Poisson process with rate p_k.

The TTL-approximation for LRU($\mathbf{m}$) exists in assuming that, when an item is not requested, the time it spends in list ℓ is deterministic and independent of the item. We denote this characteristic time by T_ℓ. Let t_n be the n-th time that item k is either requested or moves from one list to another list (where we state that an item is part of list 0 when not in the cache). Using the above assumption, we define an $h+1$ states discrete-time Markov chain $(X_n)_{n\geq 0}$, where X_n is equal to the list id of the list containing item k at time t_n.

With probability $e^{-p_k T_\ell}$ the time between two requests for item k exceeds T_ℓ. Hence, the transition matrix of $(X_n)_n$ is

$$\mathbf{P}_k = \begin{bmatrix} 0 & 1 & & & \\ e^{-p_k T_1} & 0 & 1-e^{-p_k T_1} & & \\ & \ddots & \ddots & \ddots & \\ & & e^{-p_k T_{h-1}} & 0 & 1-e^{-p_k T_{h-1}} \\ & & & e^{-p_k T_h} & 1-e^{-p_k T_h} \end{bmatrix}.$$

The Markov chain X_n is a discrete-time birth-death process. Hence, its steady state vector $(\pi_{k,0}, \pi_{k,1}, \ldots, \pi_{k,h})$ obeys

$$\pi_{k,\ell} = \pi_{k,0} \frac{\prod_{s=1}^{\ell-1}(1-e^{-p_k T_s})}{\prod_{s=1}^{\ell} e^{-p_k T_s}} = \pi_{k,0} e^{p_k T_\ell} \prod_{s=1}^{\ell-1}(e^{p_k T_s} - 1), \tag{2}$$

for $\ell = 1, \ldots, h$.

Further for $\ell \in \{1, \ldots, h\}$, the average time spend in ℓ is

$$E[t_{n+1} - t_n | X_n = \ell] = \int_{t=0}^{T_\ell} e^{-p_k t} dt = \frac{1 - e^{-p_k T_\ell}}{p_k},$$

and $E[t_{n+1} - t_n | X_n = 0] = 1/p_k$. Combined with (2), this implies that when observing the system at a random point in time, that item k is in list $\ell \geq 1$ with probability

$$\frac{\pi_{k,\ell} E[t_{n+1} - t_n | X_n = \ell]}{\sum_{j=0}^{h} \pi_{k,j} E[t_{n+1} - t_n | X_n = j]} = \frac{(e^{p_k T_1} - 1) \ldots (e^{p_k T_\ell} - 1)}{1 + \sum_{j=1}^{h} (e^{p_k T_1} - 1) \ldots (e^{p_k T_j} - 1)}$$

The expected number of items part of list ℓ is the sum of the previous expression over all items k. As for the Che approximation, setting this sum equal to m_ℓ leads to the following set of fixed point equations for T_1 to T_h:

$$m_\ell = \sum_{k=1}^{n} \frac{(e^{p_k T_1} - 1) \ldots (e^{p_k T_\ell} - 1)}{1 + \sum_{j=1}^{h} (e^{p_k T_1} - 1) \ldots (e^{p_k T_j} - 1)}. \quad (3)$$

An iterative algorithm used to determine a solution of this set of fixed point equations is presented in Appendix A. In the next section we generalize this approximation to MAP arrivals.

2) MAP arrivals: We now assume that the times that item k is requested are captured by a Markovian Arrival Process (MAP). MAPs have been developed with the aim of fitting a compact Markov model to workloads with statistical correlations and non-exponential distributions [5], [14]. A MAP is characterized by two $d \times d$ matrices $(D_0^{(k)}, D_1^{(k)})$, where the entry (j, j') of $D_1^{(k)}$ is the transition rate from state j to j' that is accompanied by an arrival and the entry (j, j') of $D_0^{(k)}$ is the transition rate from state j to j' (with $j \neq j'$) without arrival. Let $\phi^{(k)}$ such that $\phi^{(k)}(D_0^{(k)} + D_1^{(k)}) = \mathbf{0}$ and $\phi^{(k)}\mathbf{e} = 1$. Note, the request rate λ_k of item k can be expressed as $\lambda_k = \theta^{(k)} D_1^{(k)} \mathbf{e}$. Setting $D_0^{(k)} = -p_k$ and $D_1^{(k)} = p_k$ corresponds to the IRM case and letting $D_1^{(k)} = -D_0^{(k)} e v^{(k)}$ implies that item k is requested according to a phase-type renewal process characterized by $(v^{(k)}, D_0^{(k)})$.

Extending the previous section, we define a discrete-time Markov chain (X_n, S_n), where X_n is the list in which item k appears and S_n is the state of the MAP process at time t_n. This Markov chain has $d(h+1)$ states and its transition probability matrix $\mathbf{P}_k^{MAP}$ is given by

$$\begin{bmatrix} \mathbf{0} & (-D_0^{(k)})^{-1} D_1^{(k)} & & & \\ e^{D_0^{(k)} T_1} & 0 & A_{k,1} & & \\ & \ddots & \ddots & \ddots & \\ & & e^{D_0^{(k)} T_{h-1}} & 0 & A_{k,h-1} \\ & & & e^{D_0^{(k)} T_h} & A_{k,h} \end{bmatrix}$$

where

$$A_{k,\ell} = \int_{t=0}^{T_\ell} e^{D_0^{(k)} T_\ell} dt D_1^{(k)} = (I - e^{D_0^{(k)} T_\ell})(-D_0^{(k)})^{-1} D_1^{(k)}.$$

Due to the block structure of $\mathbf{P}_k^{MAP}$, its steady state vector $(\tilde{\pi}_{k,0}, \tilde{\pi}_{k,1}, \ldots, \tilde{\pi}_{k,h})$ obeys

$$\tilde{\pi}_{k,\ell} = \tilde{\pi}_{k,0} \prod_{s=1}^{\ell} R_{k,s}, \quad (4)$$

for $\ell = 1, \ldots, h$, where the matrices $R_{k,s}$ can be computed recursively as

$$R_{k,h} = A_{k,h-1}(I - A_{k,h})^{-1}, \quad (5)$$

$$R_{k,\ell} = A_{k,\ell-1} \left(I - R_{k,\ell+1} e^{D_0^{(k)} T_{\ell+1}} \right)^{-1}, \quad (6)$$

for $\ell = 1, \ldots, h-1$ and $h > 1$.

We also define the average time $(N_{k,\ell})_{j,j'}$ that item k spends in state j' in (t_n, t_{n+1}) given that $X_n = (\ell, j)$, for $j, j' \in \{1, \ldots, d\}$. Let $N_{k,\ell}$ be the matrix with entry (j, j') equal to $(N_{k,\ell})_{j,j'}$, then

$$N_{k,\ell} = \int_{t=0}^{T_\ell} e^{D_0^{(k)} t} dt = (I - e^{D_0^{(k)} T_\ell})(-D_0^{(k)})^{-1},$$

for $\ell \geq 1$ and $N_{k,0} = (-D_0^{(k)})^{-1}$. The fixed point equations for T_1 to T_h given in (3) generalize to

$$m_\ell = \sum_{k=1}^{n} \frac{\tilde{\pi}_{k,\ell} N_{k,\ell} \mathbf{e}}{\sum_{j=0}^{h} \tilde{\pi}_{k,j} N_{k,j} e}, \quad (7)$$

where $\mathbf{e}$ is a column vector of ones. The hit probability h_ℓ in list ℓ can subsequently be computed as

$$h_\ell = \frac{1}{\sum_{k=1}^{n} \lambda_k} \sum_{k=1}^{n} \frac{\tilde{\pi}_{k,\ell} N_{k,\ell} D_1^{(k)} \mathbf{e}}{\sum_{j=0}^{h} \tilde{\pi}_{k,j} N_{k,j} \mathbf{e}}, \quad (8)$$

for $\ell = 0, \ldots, h$.

B. TTL-approximation for h-LRU

1) IRM setting: As in [12], our approximation for h-LRU is obtained by assuming that an item that is not requested spends a deterministic time T_ℓ in list ℓ, independently of the identity of this item. For now we assume that $T_1 < T_2 < \ldots < T_h$. We will show that the fixed point solutions for T_1 to T_h always obey these inequalities.

We start by defining a discrete-time Markov chain $(Y_n)_{n \geq 0}$ by observing the system just prior to the time epochs that item k is requested. The state space of the Markov chain is given by $\{0, \ldots, h\}$. We say that $Y_n = 0$ if item k is not in any of the lists (just prior to the nth request). Otherwise, $Y_n = \ell$ if item k is in list ℓ, but is not in any of the lists $\ell + 1$ to h. In short, the state of the Markov chain is the largest id of the lists that contain item k.

If $Y_n = \ell$, then with probability $1 - e^{-p_k T_\ell}$, item k is requested before time T_ℓ in which case we have $Y_{n+1} = \ell + 1$. Otherwise, due to our assumption that $T_\ell \geq T_{\ell-1} \geq \ldots \geq T_1$ we have $Y_{n+1} = 0$ as in this case the item was discarded from

all lists. Therefore the transition probability matrix $\bar{P}_{h,k}$ of the $h+1$ state Markov chain $(Y_n)_{n\geq 0}$ is given by

$$\begin{bmatrix} e^{-p_k T_1} & 1-e^{-p_k T_1} & & & \\ e^{-p_k T_2} & & 1-e^{-p_k T_2} & & \\ \vdots & & & \ddots & \\ e^{-p_k T_h} & & & & 1-e^{-p_k T_h} \\ e^{-p_k T_h} & & & & 1-e^{-p_k T_h} \end{bmatrix}. \quad (9)$$

Let $\bar{\pi}^{(h,k)} = (\bar{\pi}_0^{(h,k)}, \ldots, \bar{\pi}_h^{(h,k)})$ be the stationary vector of $\bar{P}_{h,k}$, then the balance equations imply:

$$\bar{\pi}_\ell^{(h,k)} = \xi_\ell \bar{\pi}_0^{(h,k)} \prod_{s=1}^{\ell} (1 - e^{-p_k T_s}), \quad (10)$$

for $\ell = 1, \ldots, h$, where $\xi_\ell = 1$ for $\ell < h$ and $\xi_h = e^{p_k T_h}$. The probability $\bar{\pi}_h^{(h,k)}$ that item k is in the cache just before a request (which by the PASTA property is also the steady-state probability for the item to be in the cache) can therefore be expressed as

$$\frac{\prod_{s=1}^{h}(1-e^{-p_k T_s})}{\prod_{s=1}^{h}(1-e^{-p_k T_s}) + e^{-p_k T_h}\left(1+\sum_{\ell=1}^{h-1}\prod_{s=1}^{\ell}(1-e^{-p_k T_s})\right)}. \quad (11)$$

Due to the nature of h-LRU, T_1 can be found from analyzing LRU, T_2 from 2-LRU, etc. Thus, it suffices to define a fixed point equation for T_h. Under the IRM model this is simply $m = \sum_{k=1}^{n} \bar{\pi}_h^{(h,k)}$, due to the PASTA property. These fixed point equations can be generalized without much effort to renewal arrivals as explained in Appendix B.

The following property is proven in Appendix D, where we also show that $T_1 < T_2 < \ldots < T_h$ must hold to have a fixed point.

Proposition 1. *The fixed point equation* $m = \sum_{k=1}^{n} \pi_h^{(h,k)}$ *has a unique solution* T_h *which is such that* $T_h > T_{h-1}$.

When $h = 2$ Equation (11) simplifies to $(1 - e^{-p_k T_1})(1 - e^{-p_k T_2})/(1 - e^{-p_k T_1} + e^{-p_k T_2})$ which coincides with the hit probability of the so-called *refined* model for 2-LRU presented in [12, Eqn (9)]. For $h > 2$ only an approximation that relied on an additional approximation of independence between the h lists was presented in [12, see Eqn (10)]. In Figure 1 we plotted the ratio between our approximation and the one based on (10) of [12]. The results indicate that the difference grows with h. We show in Appendix C that it typically decreases as the popular items gain in popularity.

As (11) does not rely on the additional independence approximation, we expect that its approximation error is smaller and even tends to zero as m tends to infinity. This is confirmed by simulation and we list a small set of randomly chosen examples in Table I to illustrate.

2) MAP arrivals: For order d MAP arrivals, characterized by $(D_0^{(k)}, D_1^{(k)})$ for item k, we obtain a $(h+1)d$ state MC by additionally keeping track of the MAP state immediately after the requests. The transition probability matrix has the same form as $\bar{P}_{h,k}$, we only need to replace the probabilities

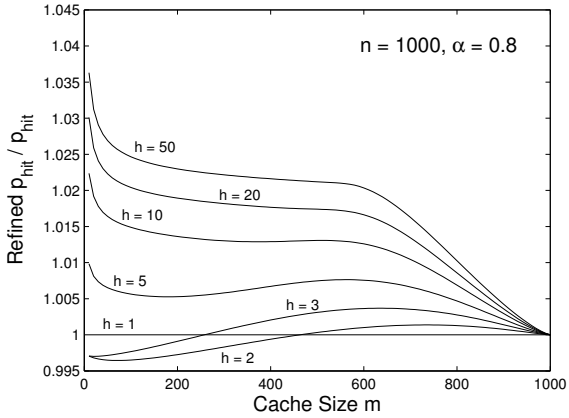

Fig. 1. Ratio of the approximation of the hit rate for h-LRU under the IRM model based on (11) and (10) of [12] as a function of the cache size for various values of h with $n = 1000$ items with a Zipf-like popularity distribution with $\alpha = 0.8$.

h	Simul.	Eq. (10) of [12] (err)	Eq. (11) (err)
$n = 1000$, $m = 10$			
2	0.19826	0.20139 (+1.576%)	0.20080 (+1.277%)
3	0.21139	0.21399 (+1.230%)	0.21336 (+0.932%)
5	0.21863	0.21780 (−0.381%)	0.21994 (+0.598%)
10	0.22357	0.21912 (−1.991%)	0.22402 (+0.201%)
$n = 1000$, $m = 100$			
2	0.47610	0.47808 (+0.415%)	0.47641 (+0.064%)
3	0.49535	0.49695 (+0.322%)	0.49579 (+0.089%)
5	0.50777	0.50521 (−0.504%)	0.50806 (+0.056%)
10	0.51506	0.50796 (−1.380%)	0.51552 (+0.088%)
$n = 10000$, $m = 100$			
2	0.27322	0.27404 (+0.302%)	0.27352 (+0.109%)
3	0.28453	0.28533 (+0.281%)	0.28477 (+0.085%)
5	0.29048	0.28873 (−0.602%)	0.29065 (+0.061%)
10	0.29427	0.28991 (−1.483%)	0.29430 (+0.011%)
$n = 10000$, $m = 1000$			
2	0.52589	0.52746 (+0.300%)	0.52596 (+0.013%)
3	0.54340	0.54453 (+0.207%)	0.54348 (+0.015%)
5	0.55452	0.55199 (−0.455%)	0.55457 (+0.009%)
10	0.56124	0.55447 (−1.206%)	0.56130 (+0.012%)

TABLE I
ACCURACY OF THE TWO APPROXIMATIONS FOR THE HIT PROBABILITY OF h-LRU UNDER THE IRM MODEL WITH A ZIPF-LIKE POPULARITY DISTRIBUTION WITH $\alpha = 0.8$. SIMULATION IS BASED ON 10 RUNS OF $10^3 n$ REQUESTS WITH A WARM-UP PERIOD OF 33%.

of the form $e^{-p_k T_\ell}$ by $e^{D_0^{(k)} T_\ell}(-D_0^{(k)})^{-1} D_1^{(k)}$ and $1 - e^{-p_k T_\ell}$ by $(I - e^{D_0^{(k)} T_\ell})(-D_0^{(k)})^{-1} D_1^{(k)}$. The fixed point equation for determining T_h is found as

$$m = \sum_{k=1}^{n} \frac{(\bar{\pi}_{h-1}^{(h,k)} + \bar{\pi}_h^{(h,k)})(I - e^{D_0^{(k)} T_h})(-D_0^{(k)})^{-1} e}{1/\lambda_k}, \quad (12)$$

where λ_k is the request rate of item k and

$$\bar{\pi}_\ell^{(h,k)} = \pi_0^{(h,k)} \left(\prod_{s=1}^{\ell} (I - e^{D_0^{(k)} T_s})(-D_0^{(k)})^{-1} D_1^{(k)} \right) \Xi_\ell,$$

for $\ell = 1, \ldots, h$, where $\Xi_\ell = I$ for $\ell < h$ and $\Xi_h = (I - (I - e^{D_0^{(k)} T_h})(-D_0^{(k)})^{-1} D_1^{(k)})^{-1}$. Finally, let $\nu^{(k)}$ be the stochastic invariant vector of $(-D_0^{(k)})^{-1} D_1^{(k)}$, that is, its d entries contain the probabilities to be in state 1 to d

immediately after an arrival. Hence, $\bar{\pi}_0^{(h,k)}$ can be computed by noting that $\sum_{\ell=0}^{h} \bar{\pi}_\ell^{(h,k)} = \nu^{(k)}$.

IV. Asymptotic Exactness of the approximations

In this section, we give evidences that the approximations presented in the previous section are asymptotically exact as the number of items tends to infinity. We first provide numerical evidence. We then show that the transient behavior of LRU($\mathbf{m}$) and h-LRU converges to a system of ODEs. By using a change of variable, these ODE can be transformed into PDEs whose fixed points are our TTL-approximations.

A. Numerical procedure and validation

For LRU($\mathbf{m}$), the fixed point of Equation (7) can computed by a iterative procedure that update the values T_ℓ in a round-robin fashion. This iterative procedure is described in Appendix A and works well for up to $h \approx 5$ lists but can be slow for a large number of lists. The computation for h-LRU is much faster and scales linearly with the number of lists: by construction, the first $h-1$ lists of a h-LRU cache behave like an $(h-1)$-LRU cache. Once T_{h-1} has been computed, the right-hand side of the fixed point equation (12) is increasing in T_h and can therefore be easily computed with a complexity that does not depend on h.

1) Accuracy for LRU(m): We show that our TTL-approximation for LRU($\mathbf{m}$) is accurate by comparing the approximation with a simulation. We assume that the inter-request times of item k follow a hyperexponential distribution with rate zp_k in state one and p_k/z in state two, while the popularity distribution is a Zipf-like distribution with parameter α, i.e., $p_k = (1/k^\alpha)/\sum_{i=1}^{n} 1/i^\alpha$. Correlation between consecutive inter-request times is introduced using the parameter $q \in (0,1]$. More precisely, let $(D_0^{(k)}, D_1^{(k)})$ equal

$$p_k \left(\begin{bmatrix} -z & 0 \\ 0 & -1/z \end{bmatrix}, q \begin{bmatrix} z \\ 1/z \end{bmatrix} [z \quad 1]/(z+1) - (1-q)D_0^{(k)} \right).$$

The squared coefficient of variation (SCV) of the inter-request times of item k is given by $2(z^2-z+1)/z-1$ and the lag-1 autocorrelation of inter-request times of item k is

$$\rho_1 = (1-q)\frac{(1-z)^2}{2(1-z)^2+z}.$$

In other words the lag-1 autocorrelation decreases linearly in q and setting $q=1$ implies that the arrival process is a renewal process with hyperexponential inter-request times.

Table II compares the accuracy of the model with time consuming simulations (based on 5 runs of $2 \cdot 10^6$ requests). We observe a good agreement between the TTL approximation and simulation that tends to improve with the size of the system (i.e., when n increases from 100 to 1000).

2) Accuracy for h-LRU: For the IRM model the TTL approximation was already validated by simulation in Table I. Using the same numerical examples as for LRU($\mathbf{m}$) we demonstrate the accuracy of the TTL-approximation under MAP arrivals in Table III. Simulation results are based on 5 runs containing $2 \cdot 10^6$ requests each and are in good agreement with the TTL-approximation.

n	q	z	method	h_0	h_1	h_2
100	1	2	model	0.26898	0.19304	0.53798
			simul.	0.27021	0.19340	0.53639
		10	model	0.03712	0.05889	0.90399
			simul.	0.03723	0.06106	0.90171
1000	1	2	model	0.22580	0.16262	0.61158
			simul.	0.22599	0.16256	0.61145
		10	model	0.03112	0.04963	0.91925
			simul.	0.03108	0.04969	0.91923
1000	0.1	2	model	0.21609	0.14510	0.63881
			simul.	0.21603	0.14526	0.63870
		10	model	0.03006	0.02044	0.94950
			simul.	0.02984	0.02032	0.94985

TABLE II
Accuracy of probability h_ℓ of finding a requested item in list ℓ for LRU($\mathbf{m}$). In this example $\alpha = 0.8$, $h = 2$ and $m_1 = m_2 = n/5$ (*i.e.,* 20 or 200).

n	q	z	method	$h=2$	$h=3$
100	1	2	model	0.53619	0.54292
			simul.	0.53449	0.54150
		10	model	0.88249	0.83718
			simul.	0.87936	0.83449
1000	1	2	model	0.61028	0.61605
			simul.	0.61016	0.61587
		10	model	0.90103	0.86300
			simul.	0.90071	0.86262
1000	0.1	2	model	0.64744	0.65841
			simul.	0.64807	0.65899
		10	model	0.94935	0.94646
			simul.	0.94924	0.94632

TABLE III
Accuracy of hit probability for h-LRU with MAP arrivals. In this example $\alpha = 0.8$ and $m = n/5$.

B. Asymptotic behavior and TTL-approximation

In this subsection, we construct two systems of ODEs that approximate the transient behavior of LRU($\mathbf{m}$) and h-LRU. These approximations become exact as the popularity of the most popular item decreases to zero:

Theorem 1. *Let $H_\ell(t)$ be the sum of the popularity of the items of list ℓ and $h_\ell(t)$ be the corresponding ODE approximation (Equation (18) for h-LRU and Equation (22) for LRU($\mathbf{m}$)). Then: for any time T, there exists a constant C such that*

$$\mathbf{E}\left[\sup_{t \le T/\sqrt{\max_k p_k}} |H_\ell(t) - h_\ell(t)| \right] \le C\sqrt{\max_k p_k},$$

where C does not depend on the probabilities $p_1 \dots p_n$, the cache size m or the number of items n.

Our proof of this results is to use an alternative representation of the state space that allows us to use techniques from stochastic approximation. We present the main ideas in this paper while the technical details are provided in Appendix E.

We associate to each item k a variable $\tau_k(t)$ that is called the *request time* of item k at time t and an additional variable that tracks if an item appears in a list. Our approximation is given by an ordinary differential equation (ODE) on $x_{k,b}(t)$ that is an approximation of the probability that $\tau_k(t)$ is greater than b while appearing in a list ℓ. A more natural representation

would be to consider the time since the last request. Our ODE approximation would then be replaced a partial differential equation (PDE) by replacing the ODE in $x_{k,\ell,b}$ by a PDE in $y_{k,\ell,s}$, where $y_{k,\ell,s}(t) = x_{k,\ell,t-s}(t)$. However, when working directly with the PDE, the proofs are much more complex. In each case, we show that the fixed point of the PDE corresponds to the TTL-approximation of LRU($\mathbf{m}$) and h-LRU presented in Sections III-A and III-B.

To ease the presentation, we present the convergence result when the arrivals follow an IRM model, where each item k has a probability p_k of being requested at each time step. This proof can be adapted to the case of MAP arrivals but at the price of more complex notations. Indeed, for IRM, our system of ODEs is given by the variables $x_{k,\ell,b}(t)$ which are essentially an approximation of the probability for item k to be in a list ℓ while having been requested between b and t. If the arrival process of an item is modeled by a MAP with d states, then our approximation would need to consider $x_{k,\ell,b,j}(t)$ which would approximate the probabilities for item k to be in state j, in list ℓ and having being requested between b and t. A detailed proof for the case of MAP arrivals is beyond the scope of this paper, both because of space constraints and for the sake of clarity of the exposition.

1) LRU: We first construct the ODE approximation for LRU. In this simpler case the proof of the validity of the Che-approximation could rely on a more direct argument that uses the close-form expression of the steady state distribution of LRU, as in [8]. Yet, the ideas presented in this section serve to illustrate the more complex cases of h-LRU and LRU($\mathbf{m}$).

The request time of an item k evolves as follows:

$$\tau_k(t+1) = \begin{cases} \tau_k(t) & \text{if } k \text{ is not requested} \\ t+1 & \text{if } k \text{ is requested.} \end{cases} \tag{13}$$

At time 0, $\tau_k(0) = -i$ if the item is in the ith position in the cache and $\tau_k(0) = -(m+1)$ if the item is not in the cache.

The cache contains m items. Let us denote[1] by $\Theta(t) = \sup\{b : \sum_{k=1}^n \mathbf{1}_{\{\tau_k(t)\geq b\}} \geq m\}$ the request time of the mth most recently requested item. When using LRU, an item that has a request time greater or equal to $\Theta(t)$ is in the cache. We denote by $H(t)$ the sum of the popularities of items in the cache:

$$H(t) = \sum_{k=1}^n p_k \mathbf{1}_{\{\tau_k(t)\geq\Theta(t)\}}.$$

Our approximation of the probability for item k to have a request time after b, is given by the following ODE (for $b < t$):

$$\dot{x}_{k,b}(t) = p_k(1 - x_{k,b}(t)). \tag{14}$$

with the initial conditions that for $t > 0$, $x_{k,t}(t) = 0$ and for $t = 0$, $x_{k,b}(0) = \mathbf{1}_{\{\tau_k(t)\geq b\}}$. Similarly to the stochastic system, we define $\theta(t) = \sup\{b : \sum_{k=1}^n x_{k,b}(t) \geq m\}$, which is the time for which the sum of the approximated probabilities of having items requested after b is equal to m. The approximation of the hit ratio for LRU is then given by

$$h(t) = \sum_{k=1}^n p_k x_{k,\theta(t)}(t).$$

Once these variables have been defined, the key ingredient of the proof is to use the same changes of variables as in the proof of Theorem 6 of [9], which is to consider $P_{\alpha,b}(t)$:

$$P_{\alpha,b}(t) = a^{1-\alpha} \sum_{k=1}^n (p_k)^\alpha \mathbf{1}_{\{\tau_k(t)\geq b\}},$$

where $a := \max_{k=1}^n p_k$. These variables are defined for $\alpha \in \{0, 1, \dots\}$ and $b \in \mathbf{Z}$. The collection of variables $\{P_{\alpha,b}\}_{\alpha,b}$ describes completely the state of the system at time t and live in an set of infinite dimension.

Similarly, we define a set of functions $\rho_{\alpha,b}$ by $\rho_{\alpha,b}(t) = a^{1-\alpha} \sum_{k=1}^n (p_k)^\alpha x_{k,b}(t)$. The functions $\rho_{\alpha,b}$ are solutions of the system of ODEs $d/dt\rho_{\alpha,b}(t) = f_{\alpha,b}(\rho)$, where:

$$f_{\alpha,b}(\rho) = a^{1-\alpha}(\sum_k (p_k)^{\alpha+1}) - a\rho_{\alpha+1,a}(t).$$

The proof of the theorem, detailed in Appendix E1, relies on classical results of stochastic approximation. It uses the fact that

- the function f is Lipchitz-continuous
- f is the $\mathbf{E}\left[P_{\alpha,b}(t+1) - P_{\alpha,b}(t) \mid P(t)\right] = f_{\alpha,b}(P(t))$
- The second moment of the variation of $P(t)$ is bounded: $\mathbf{E}\left[\|P(t+1) - P(t)\|_\infty^2 \mid \mathcal{P}\right] \leq a$.

Note that Equation (14) can be transformed into a PDE by considering the change of variable $y_{k,s}(t) = x_{k,t-s}(t)$. The quantity $y_{k,s}(t)$ is an approximation of the probability for an item k to have been requested between $t-s$ and t. The set of ordinary differential Equations (14) can then be naturally transformed in the following PDE:

$$\frac{\partial}{\partial t} y_{k,s}(t) = p_k(1 - y_{k,s}(t)) - \frac{\partial}{\partial s} y_{k,s}(t). \tag{15}$$

The fixed point y of the PDE can be obtained by solving the equation $\frac{\partial}{\partial t} y = 0$. This fixed point satisfied $y_{k,s} = 1 - e^{-p_k s}$. For this fixed point, the quantity $T = t - \theta$ satisfies $m = \sum_{k=1}^n (1 - e^{-p_k T})$. This equation is the same as the Che-approximation, given by Equation (1).

2) h-LRU: The construction for LRU can be extended to the case of h-LRU by adding to each item h variables $L_{k,\ell}(t) \in \{\text{true}, \text{false}\}$. For item k and a list ℓ, $L_{k,\ell}(t)$ equals true if item k was present in list ℓ just after the last request[2] of item k and false otherwise. Similarly to the case of LRU, we define the quantity $\Theta_\ell(t)$ to be the request time of the least recently requested item that belongs to list ℓ at time t, that is,

$$\Theta_\ell(t) = \sup\{b : \sum_{k=1}^n \mathbf{1}_{\{\tau_k(t)\geq b \wedge L_{k,\ell}(t)\}} \geq m_\ell\}.$$

[1]Throughout the paper $\mathbf{1}_{\{A\}}$ is the indicator function of an event A. It is equal to 1 if A is true and 0 otherwise.

[2]Note that, after a request, an item is always inserted in list 1. This implies $L_{k,1}(t) = \text{true}$.

We then define a quantity $x_{k,\ell,b}(t)$ that is meant to be an approximation of the probability for item k to have $\tau_k(t) \geq b$ and $L_\ell(t) = \text{true}$.

As $L_1(t)$ is always equal to true, the ODE approximation for $x_{k,1,b}(t)$ is the same as (14). Moreover, this implies that $\Theta_1(t) \geq \Theta_\ell(t)$ for $\ell \geq 2$. For the list $\ell = 2$, the approximation is obtained by considering the evolution of $L_2(t)$. After a request, $L_2(t+1)$ is true if $\tau_k(t) \geq \Theta_1(t)$ or if ($\tau_k(t) \geq \Theta_2(t)$ and $L_2(t) = \text{true}$). Both these events occur if ($\tau_k(t) \geq \Theta_1(t)$ and $L_2(t) = \text{true}$) as $\Theta_1(t) \geq \Theta_2(t)$. This suggests that, if the item k is requested, then, in average $L_{k,2}(t+1)$ is approximately $x_{k,1,\theta_1(t)} + x_{k,2,\theta_2(t)} - x_{k,2,\theta_1(t)}$, which leads to the following ODE approximation for $x_{k,2,b}$:

$$\dot{x}_{k,2,b} = p_k(x_{k,2,\theta_2(t)} + x_{k,1,\theta_1(t)} - x_{k,2,\theta_1(t)} - x_{k,2,b}), \quad (16)$$

where $\theta_\ell(t) = \sup\{b : \sum_{k=1}^n x_{k,\ell,b}(t) \geq m_\ell\}$ for $\ell \in \{1, 2\}$.

The formulation for the third list and above is more complex. In Section III-B, we showed that the computation of the fixed point is simple because the quantities T_ℓ of the fixed point satisfy $T_1 \leq T_2 \cdots \leq T_h$. However, for the stochastic system, we do not necessarily have[3] $\Theta_\ell(t) \geq \Theta_{\ell+1}(t)$ when $\ell \geq 2$, which implies that the ODE approximation for h-LRU has 2^{h-1} terms.

Applying the reasoning of $L_{k,2}$ to compute $L_{k,\ell}$ ($\ell \geq 3$) involves computing the probability of ($\tau_k(t) \geq \Theta_{\ell-1}(t)$ and $L_{k,\ell-1}(t) = \text{true}$) or ($\tau_k(t) \geq \Theta_\ell(t)$ and $L_{k,\ell}(t) = \text{true}$). When $\Theta_\ell(t) \leq \Theta_{\ell-1}(t)$, both these events occur if ($\tau_k(t) \geq \Theta_{\ell-1}(t)$ and $L_{k,\ell}(t) = L_{k,\ell-1}(t) = \text{true}$). This suggest that the ODE for $x_{k,\ell,b}(t)$ has to involve a term $x_{k,\{\ell-1,\ell\},\theta_{\ell-1}(t)}(t)$, that is an approximation for the item k to have a request time after $\theta_{\ell-1}(t)$ and such that $L_{k,\ell-1}(t) = L_{k,\ell}(t) = \text{true}$. Note, for $\ell = 2$ we have $x_{k,\{\ell-1,\ell\},b}(t) = x_{k,\ell,b}(t)$ as $L_{k,1}(t)$ is always true, but this does not hold for $\ell > 2$. This leads to:

$$\begin{aligned}\dot{x}_{k,\ell,b} = p_k(&x_{k,\ell,\theta_\ell(t)} + x_{k,\ell-1,\theta_{\ell-1}(t)} \\ &- x_{k,\{\ell-1,\ell\},\max\{\theta_{\ell-1}(t),\theta_\ell(t)\}} - x_{k,\ell,b}), \quad (17)\end{aligned}$$

A similar reasoning can be applied to obtain an ODE for $x_{k,\{\ell-1,\ell\},b}(t)$ as a function of $x_{k,\{\ell-1,\ell\},b}(t)$, $x_{k,\{\ell-2,\ell-1,\ell\},b}(t)$ and $x_{k,\{\ell-2,\ell\},b}(t)$. For example, for $\ell = 3$ this approximation becomes

$$\dot{x}_{k,\{2,3\},b}(t) = x_{k,2,\theta_2(t)} + x_{k,3,\theta_1(t)} - x_{k,\{2,3\},\theta_1(t)} - x_{k,\{2,3\},b}$$

as $L_{k,1}(t)$ is always true.

The hit probability of list ℓ used in Theorem 1 is then

$$h_\ell(t) = \sum_{k=1}^n x_{k,\ell,\theta_\ell(t)}(t), \quad (18)$$

where the variables $x_{k,\ell,b}$ satisfy the above ODE.

[3]When $h = 3$ lists, the variables $\Theta_\ell(t)$ are not always ordered. For example, consider the case of four items $\{1, 2, 3, 4\}$ and $m_1 = m_2 = m_3 = 3$. If initially the three caches contain the three items $1, 2, 3$. Then, after a stream of requests: $4, 4, 3, 2, 1$, the cache 1 and 3 will contain the items $\{1, 2, 3\}$ while the cache 2 will contain $\{1, 2, 4\}$. This implies that $t - 3 = \Theta_2(t) < \Theta_3(t) = \Theta_1(t) = t - 2$.

The proof of Theorem 1 in the case of h-LRU is very similar as the one for LRU and uses the same stochastic approximation argument. Moreover, as for LRU, the ODE (16) can be transformed into a PDE by using the change of variables $y_{k,\ell,s}(t) = x_{k,\ell,t-s}(t)$ and $T_\ell(t) = t - \theta_\ell(t)$. This PDE has a unique fixed point that corresponds to Section III-B.

3) LRU(m): The construction of the approximation and the proof for the case of LRU($\mathbf{m}$) is more involved because of discontinuities in the dynamics. We replace the request time by a quantity that we call a *virtual request time* that is such that the m_h items that have the largest virtual request times are in list h. The next m_{h-1} are in list $h-1$, etc. The virtual request time of an item changes when this item is requested. If the item was in list h or $h-1$ prior to the request, its virtual request time becomes $t+1$. If the item was in a list $\ell \in \{0 \ldots h-2\}$, its virtual request time becomes the largest virtual request time of the items in list $\ell + 2$.

The approximation of the distribution of virtual request times is then given by an ODE on the quantities $x_{k,b}(t)$ that are meant to be an approximation of the probability that the item k has a virtual request time after b:

$$\dot{x}_{k,b}(t) = p_k(x_{k,\theta_{\zeta_b(t)}(t)-1}(t) - x_{k,b}(t)), \quad (19)$$

where $\theta_\ell(t)$ and $\zeta_b(t)$ are defined by:

$$\theta_\ell(t) = \sup\{b : \sum_{k=1}^n x_{k,b}(t) \geq m_h + \cdots + m_\ell\} \quad (20)$$

$$\zeta_b(t) = \max\{\ell : \theta_\ell(t) \leq b\} \quad (21)$$

The intuition behind Equation (19) is that the changes in $x_{k,b}$ are due to the items that had a virtual request time prior to b and that now have a virtual request time b or after. This only occurs if the item had a virtual request time between $\theta_{\zeta_b(t)-1}$ and b and was requested, in which case its new virtual request time is $\theta_{\zeta_b(t)+1} \geq b$. Otherwise, if an item had a virtual request time prior to $\theta_{\zeta_b(t)-1}$, then upon request it jumps to a list $\ell < \zeta_b(t) - 1$ and therefore its new virtual request time stays prior to b.

The hit ratio for LRU($\mathbf{m}$) used in Theorem 1 is given by

$$h_\ell(t) = \sum_{k=1}^n p_k(x_{k,\theta_{\ell+1}(t)} - x_{k,\theta_\ell(t)})(t) \quad (22)$$

The main difference between the proof for LRU($\mathbf{m}$) compared to the one of h-LRU is that the right-side of the differential equation (19) is not Lipschitz-continuous in ρ because the list in which an item that has an virtual request time b belongs to depends non-continuously on ρ (the list ζ_b is a discrete quantity). In Appendix E3, we explain how to prove the convergence of the stochastic approximation algorithm by using one-sided Lipschitz-continuous functions.

V. Comparison of LRU(m) and h-LRU

In this section we compare the performance of LRU($\mathbf{m}$) and h-LRU in terms of the achieved hit probability when subject to IRM, renewal, MAP requests and trace-based simulation. A good replacement algorithm should keep popular items in the

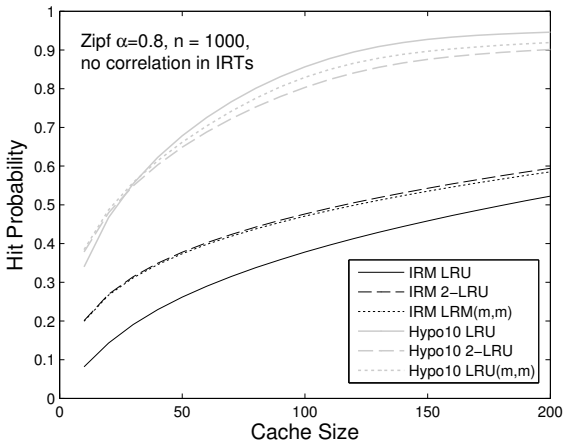

Fig. 2. Hit probability as a function of the cache size for LRU, LRU(m, m) and 2-LRU under the IRM model and with hyperexponential inter-request times (with $z = 10$).

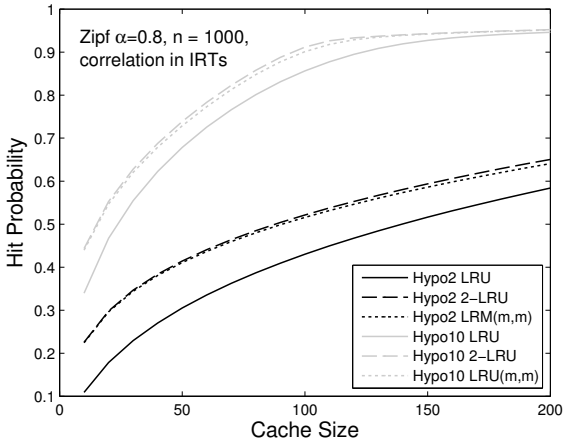

Fig. 3. Hit probability as a function of the cache size for LRU, LRU(m, m) and 2-LRU when subject to MAP arrivals (with $z = 2, 10$ and $q = 1/20$).

cache, but needs to be sufficiently responsive to changes in the popularity. As LRU($\mathbf{m}$) and h-LRU are clearly better suited to keep popular items in the cache than LRU, they perform better under static workloads (IRM). We demonstrate that they often also outperform LRU when the workload is dynamic.

A. Synthetic (static) workloads

For the synthetic workloads we restrict ourselves to LRU, 2-LRU and LRU(m, m). The latter two algorithms both use a cache of size m and additionally keep track of meta-data only for the m items in list 1.

Figure 2 depicts the hit probability as a function of the cache size when $n = 1000$, items follow a Zipf-like popularity distribution with parameter 0.8 under IRM and renewal requests (with $z = 10$, see Section IV-A1). Figure 3 shows the impact of having correlation between consecutive inter-request times (that is, $q = 1/20$ instead of $q = 1$ for $z = 2, 10$).

One of the main observations is that LRU(m, m) performs very similar to 2-LRU under IRM, renewal and MAP requests.

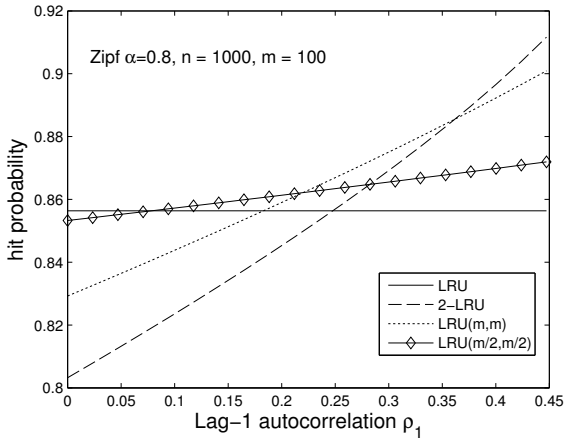

Fig. 4. Hit probability as a function of the lag-1 autocorrelation ρ_1 for LRU, LRU(m, m), LRU$(m/2, m/2)$ and 2-LRU when subject to MAP arrivals (with $z = 10$).

In fact, 2-LRU performs slightly better, unless the workload is very dynamic ($z = 10$ and $q = 1$ case). Another important observation that can be drawn from comparing Figures 2 and 3 is that the hit rate of both 2-LRU and LRU(m, m) significantly improves in the presence of correlation between consecutive inter-request times (that is, when $q < 1$), while LRU does not. Recall that LRU($\mathbf{m}$) needs to update at most one list per hit, as opposed to h-LRU. Thus, whenever both algorithms perform alike, LRU($\mathbf{m}$) may be more attractive to use.

Figure 4 shows that the hit rate of 2-LRU and LRU(m, m) both increase with increasing lag-1 autocorrelation and more importantly that the hit probability of LRU is completely *insensitive* to any correlation between consecutive inter-request times. Figure 4 further indicates that the hit probability also increases with ρ_1 when splitting the cache in two lists of equal size (although the gain is less pronounced). As indicated by the following theorem, whose proof is given in Appendix F, the insensitivity of LRU is a general result.

Theorem 2. *Assume that the items' request processes are stationary, independent of each other and that the expected number of requests per unit time is positive and finite. Then, the hit probability of LRU only depends on the inter-arrival time distribution. In particular, it does not depend on the correlation between inter-arrival time.*

This theorem complements the results of Jelenkovic and Radovanovic who showed in [11], [10] that for dependent request processes, the hit probability is asymptotically, for large cache sizes, the same as in the corresponding LRU system with i.i.d. requests. Our insensitivity result is valid not just asymptotically but requires the request processes of the various items to be independent.

B. Trace-based simulation

To perform the trace-based simulations we rely on the same 4 IR cache traces as in [4, Section 4]. In this section, we only report the result for the trace collected on Monday 18th Feb

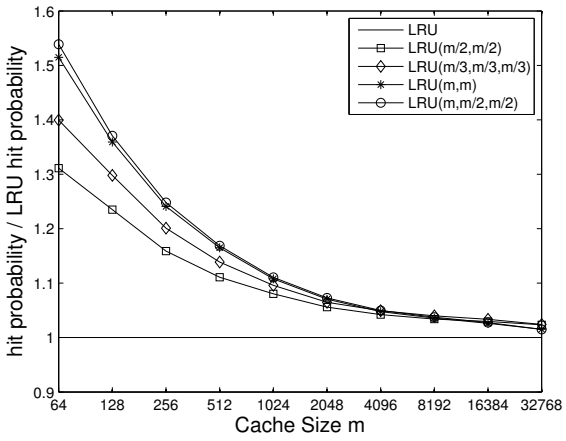

Fig. 5. Hit probability as a function of the cache size for LRU($\mathbf{m}$) compared to LRU using trace-based simulation.

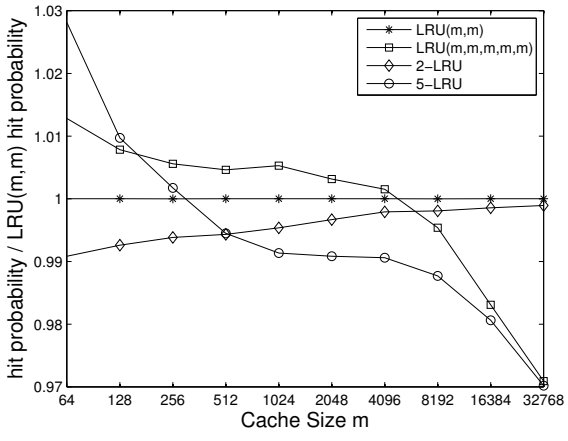

Fig. 6. Hit probability as a function of the cache size for LRU(m, m, m, m, m) and h-LRU compared to LRU(m, m) using trace-based simulation.

2013. We also simulated the other traces and obtained very similar results.

The hit probability of LRU($\mathbf{m}$) with a split cache and/or virtual lists normalized by the LRU hit probability is depicted in Figure 5 as a function of the cache size m. It indicates that LRU($\mathbf{m}$) is more effective than LRU, especially when the cache is small. For small caches using a virtual list is better than splitting the cache and using both a virtual list and split cache offers only a slight additional gain. While not depicted here, we should note that using more virtual lists or splitting the cache in more parts sometimes result in a hit probability below the LRU hit probability for larger caches.

Figure 6 compares h-LRU with LRU($\mathbf{m}$) using virtual lists, where the hit probability is now normalized by the hit probability of LRU(m, m) to better highlight the differences. We observe that 2-LRU differs by less than 1% from LRU(m, m), while 5-LRU and LRU(m, m, m, m, m) differ by less than 2%. Given that h-LRU may require an update of up to h lists, while LRU($\mathbf{m}$) requires only one update in case of a hit, LRU($\mathbf{m}$) seems preferential in this particular case.

VI. Conclusion

In this paper, we developed algorithms to approximate the hit probability of the cache replacement policies LRU($\mathbf{m}$) and h-LRU. These algorithms rely on an equivalence between LRU-based and TTL-based cache replacement algorithms. We showed numerically that the TLL-approximations are very accurate for moderate cache sizes and appear asymptotically exact as the cache size grows. We also provide theoretical support for this claim, by establishing a bound between the transient dynamics of both policies and a set of ODEs whose fixed-point coincides with the proposed TTL-approximation.

A possible extension of our results would be to study networks of caches in which LRU, LRU($\mathbf{m}$) or h-LRU is used in each node. Further, our TTL-approximation with MAP arrivals can be readily adapted to other policies such as FIFO($\mathbf{m}$) and RAND($\mathbf{m}$) introduced in [9]. In fact, a generalization to a network of caches would be fairly straightforward for the class of RAND($\mathbf{m}$) policies.

References

[1] O. I. Aven, E. G. Coffman, Jr., and Y. A. Kogan. *Stochastic Analysis of Computer Storage*. Kluwer Academic Publishers, Norwell, MA, USA, 1987.

[2] F. Baccelli and P. Brémaud. *Elements of queueing theory: Palm Martingale calculus and stochastic recurrences*, volume 26. Springer Science & Business Media, 2013.

[3] D. S. Berger, P. Gland, S. Singla, and F. Ciucu. Exact analysis of TTL cache networks. *Performance Evaluation*, 79:2–23, 2014.

[4] G. Bianchi, A. Detti, A. Caponi, and N. Blefari-Melazzi. Check before storing: What is the performance price of content integrity verification in LRU caching? *SIGCOMM Comput. Commun. Rev.*, 43(3):59–67, July 2013.

[5] G. Casale. Building accurate workload models using markovian arrival processes. In *ACM SIGMETRICS*, SIGMETRICS '11, pages 357–358, New York, NY, USA, 2011. ACM.

[6] H. Che, Y. Tung, and Z. Wang. Hierarchical web caching systems: modeling, design and experimental results. *IEEE J.Sel. A. Commun.*, 20(7):1305–1314, 2002.

[7] N. C. Fofack, P. Nain, G. Neglia, and D. Towsley. Analysis of TTL-based cache networks. In *Valuetools 2012*, pages 1–10. IEEE, 2012.

[8] C. Fricker, P. Robert, and J. Roberts. A versatile and accurate approximation for LRU cache performance. In *Proceedings of the 24th International Teletraffic Congress*, ITC '12, pages 8:1–8, 2012.

[9] N. Gast and B. Van Houdt. Transient and steady-state regime of a family of list-based cache replacement algorithms. In *Proceedings of ACM SIGMETRICS*. ACM, 2015.

[10] P. Jelenkovic and A. Radovanovic. Asymptotic insensitivity of least-recently-used caching to statistical dependency. In *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies*, volume 1, pages 438–447. IEEE, 2003.

[11] P. R. Jelenković and A. Radovanović. Least-recently-used caching with dependent requests. *Theoretical computer science*, 326(1):293–327, 2004.

[12] V. Martina, M. Garetto, and E. Leonardi. A unified approach to the performance analysis of caching systems. In *INFOCOM 2014*, pages 2040–2048, 2014.

[13] E. J. Rosensweig, J. Kurose, and D. Towsley. Approximate models for general cache networks. In *INFOCOM'10*, pages 1100–1108, Piscataway, NJ, USA, 2010. IEEE Press.

[14] M. Telek and G. Horváth. A minimal representation of markov arrival processes and a moments matching method. *Perform. Eval.*, 64(9-12):1153–1168, Oct. 2007.

APPENDIX

A. Fixed point computations

This subsection contains some details on the computation of the fixed point for LRU($\mathbf{m}$) and h-LRU when subject to MAP arrivals. For LRU($\mathbf{m}$) the fixed point equations in case of MAP arrivals are given by (7) and an iterative algorithm to compute the fixed point is presented in Figure 1. For h-LRU we first determine T_1 via (12) with $h = 1$, then we determine T_2 by considering the TTL approximation for 2-LRU with T_1 fixed, etc. In other words:

- for h-LRU computing $T_1, \ldots, T_h$ corresponds to solving h one dimensional problems
- for LRU($\mathbf{m}$), computing $T_1 \ldots T_h$ corresponds to solving a single h-dimensional one.

In practice, the computation for h-LRU is much faster.

Input: D_0, D_1, $m_1, \ldots, m_h$, ϵ
Output: fixed point solution $\hat{T}_1, \ldots, \hat{T}_m$

for $\ell = 1$ *to* h **do**
  $\hat{T}_\ell = n$;
end
$\hat{T}_{h+1} = \infty$, $x = 1$;
while $x > \epsilon$ **do**
  for $\ell = 1$ *to* h **do**
    Find $x \in (-\hat{T}_\ell, \hat{T}_{\ell+1})$ such that $(T_1, \ldots, T_h)$ equal to $(\hat{T}_1, \ldots, \hat{T}_\ell + x, \hat{T}_{\ell+1} - x, \ldots, \hat{T}_h)$ minimizes $|m_\ell -$ rhs of (7)|;
    $\hat{T}_\ell = \hat{T}_\ell + x$; $\hat{T}_{\ell+1} = \hat{T}_{\ell+1} - x$;
  end
end

Algorithm 1: Iterative algorithm used to solve fixed point equations in (7)**.**

B. h-LRU with renewal arrivals

The same approach as for the IRM model can be used to obtain a TTL approximation when the requests for item k follow a renewal process, characterized by a distribution with cumulative distribution function $F_k(x)$. Let $\bar{F}_k(x) = 1 - F_k(x)$. In this case we get $(\bar{P}_{h,k})_{j,0} = \bar{F}_k(T_{\min(h,j+1)})$ and $(\bar{P}_{h,k})_{j,\min(h,j+1)} = F_k(T_{\min(h,j+1)})$. The hit probability for item k can therefore be expressed as

$$\bar{\pi}_h^{(h,k)} = \frac{\prod_{s=1}^{h} F_k(T_s)}{\prod_{s=1}^{h} F_k(T_s) + \bar{F}_k(T_h)\left(1 + \sum_{j=1}^{h-1}\prod_{s=1}^{j} F_k(T_s)\right)},$$

while for $j = 1, \ldots, h-1$ we have

$$\bar{\pi}_j^{(h,k)} = \frac{\bar{F}_k(T_h)\prod_{s=1}^{j} F_k(T_s)}{\prod_{s=1}^{h} F_k(T_s) + \bar{F}_k(T_h)\left(1 + \sum_{j=1}^{h-1}\prod_{s=1}^{j} F_k(T_s)\right)}.$$

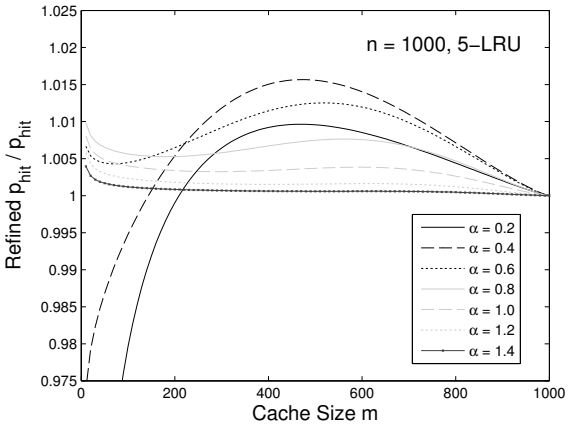

Fig. 7. Ratio of the approximation of the hit rate for 5-LRU under the IRM model based on (11) and (10) of [12] as a function of the cache size for various values of α with $n = 1000$ items with a Zipf-like popularity distribution with α.

The fixed point equation for determining T_h is found as

$$\begin{aligned} m &= \sum_{k=1}^{n} \frac{(\bar{\pi}_{h-1}^{(h,k)} + \bar{\pi}_h^{(h,k)}) \int_{x=0}^{T_h} x dF_k(x)}{\int_{x=0}^{\infty} \bar{F}_k(x)dx} \\ &= \sum_{k=1}^{n} \frac{(\bar{\pi}_{h-1}^{(h,k)} + \bar{\pi}_h^{(h,k)}) \left(T_h - \int_{x=0}^{T_h} F_k(x)dx\right)}{\int_{x=0}^{\infty} \bar{F}_k(x)dx}, \end{aligned}$$

as $(\bar{\pi}_{h-1}^{(h,k)} + \bar{\pi}_h^{(h,k)}) \int_{x=0}^{T_h} x dF_k(x)$ is the mean time that item k spends in the cache between two requests for item k and $\int_{x=0}^{\infty} \bar{F}_k(x)dx$ is simply the mean time between two requests.

C. Comparison of TTL-approximations for h-LRU

In Figure 1 we depicted the difference between our TTL-approximation for h-LRU under the IRM model based on (11) and (10) of [12], where the popularity of the items followed a Zipf-like distribution with $\alpha = 0.8$. Figure 7 depicts the impact of the parameter α of the Zipf-like distribution when $h = 5$. The difference between both approximations appears to grow as α decreases. In other words, the difference decreases as the popular items gain in popularity.

D. Some proofs for h-LRU under IRM

1) Proof of Proposition 1: The fixed point equation for $h \geq 2$ can be written as $m = f_h(T_h)$, where

$$f_h(x) = \sum_{k=1}^{n} \frac{(1 - e^{-p_k x})\prod_{s=1}^{h-1} e_{k,s}}{\prod_{s=1}^{h-1} e_{k,s} + e^{-p_k x}\left(1 + \sum_{j=1}^{h-2}\prod_{s=1}^{j} e_{k,s}\right)},$$

with $e_{k,s} = (1 - e^{-p_k T_s})$. The function $f_h(x)$ is clearly an increasing function in x and therefore $m = f(x)$ has a unique

solution T_h. Further,

$$\begin{aligned} f_h(T_{h-1}) &= \\ &\sum_{k=1}^{n} \frac{(1-e^{-p_k T_{h-1}}) \prod_{s=1}^{h-1} e_{k,s}}{\prod_{s=1}^{h-1} e_{k,s} + e^{-p_k T_{h-1}} \left(1 + \sum_{j=1}^{h-2} \prod_{s=1}^{j} e_{k,s}\right)} \\ &< \sum_{k=1}^{n} \frac{\prod_{s=1}^{h-1} e_{k,s}}{\prod_{s=1}^{h-1} e_{k,s} + e^{-p_k T_{h-1}} \left(1 + \sum_{j=1}^{h-2} \prod_{s=1}^{j} e_{k,s}\right)} \\ &= \sum_{k=1}^{n} \bar{\pi}_{h-1}^{(h-1,k)} = m, \end{aligned}$$

meaning $T_h > T_{h-1}$.

2) Structure of the fixed point: Using induction we prove that the fixed point solutions obey $T_1 < \ldots < T_h$. We assume that $T_1 < \ldots < T_{h-1}$ (which trivially holds for $h = 2$) and show that the fixed point equation for T_h does not have a solution for $T_h \in (0, T_{h-1})$. The key thing to note is that when $T_h \leq T_{h-1}$ item k is part of list $h-1$ whenever it is part of list h. As such we still obtain a Markov chain by observing the system just prior to the request times, but now the last two rows of the transition probability matrix are both equal to

$$(e^{-p_k T_{h-1}}, 0, \ldots, 0, e^{-p_k T_h} - e^{-p_k T_{h-1}}, 1 - e^{-p_k T_h}).$$

Let $(\hat{\pi}_0^{(h,k)}, \ldots, \hat{\pi}_h^{(h,k)})$ be the invariant vector of this modified Markov chain, then it is easy to see that

$$\hat{\pi}_{h-1}^{(h,k)} + \hat{\pi}_h^{(h,k)} = \pi_{h-1}^{(h-1,k)},$$

as lumping the last two states into a single state results in the matrix $\bar{P}_{h-1,k}$. Hence the fixed point equation $\sum_{k=1}^{n} \hat{\pi}_h^{(h,k)} = m$ cannot have a solution as

$$\sum_{k=1}^{n} \hat{\pi}_h^{(h,k)} < \sum_{k=1}^{n} (\hat{\pi}_{h-1}^{(h,k)} + \hat{\pi}_h^{(h,k)}) = \sum_{k=1}^{n} \bar{\pi}_{h-1}^{(h-1,k)} = m.$$

E. Proof of Theorem 1

1) LRU: Let us denote $a = \max_{k=1}^{n} p_k$. For $\alpha \in \{0, 1, \ldots\}$ and $b \in \mathbf{Z}$, we define $P_{\alpha,b}(t)$:

$$P_{\alpha,b}(t) = a^{1-\alpha} \sum_{k=1}^{n} (p_k)^{\alpha} \mathbf{1}_{\{\tau_k(t) \geq b\}}.$$

$P_{\alpha,b}(t)$ is the sum of the popularity to the power α of all items that have a request time greater or equal to b at time t.

The collection of variables $\{P_{\alpha,b}\}_{\alpha,b}$ describe completely the state of the system at time t. They live in a set $\mathcal{P}$, that is the set of infinite vectors such that there exists a vector $x_{k,b}$ that is bounded by 1, non-increasing in b and such that $P_{\alpha,b} = a^{1-\alpha} \sum_{k=1}^{n} (p_k)^{\alpha} x_{k,b}$.

$$\mathcal{P} = \Big\{ (P_{\alpha,b})_{\alpha,b} : \exists (x_{k,b}) \text{ non-increasing in } b, \text{ bounded by } 1 \text{ such that for all } \alpha, b\text{: } P_{\alpha,b} = a^{1-\alpha} \sum_{k=1}^{n} (p_k)^{\alpha} x_{k,b} \Big\}.$$

We equip $\mathcal{P}$ with the L_∞ norm and denote $\|\rho\|_\infty = \sup_{\alpha,b} |\rho_{\alpha,b}|$ the norm of a vector $\rho \in \mathcal{P}$.

Similarly, we define a set of functions $\rho_{\alpha,b}$ by $\rho_{\alpha,b}(t) = a^{1-\alpha} \sum_{k=1}^{n} (p_k)^{\alpha} x_{k,b}(t)$. The functions $\rho_{\alpha,b}$ are solutions of the system of ODEs $d/dt \rho_{\alpha,b}(t) = f_{\alpha,b}(\rho)$, where:

$$f_{\alpha,b}(\rho) = a^{1-\alpha} (\sum_{k} (p_k)^{\alpha+1}) - a \rho_{\alpha+1,a}(t).$$

To prove the result, we use the following two lemmas, whose proofs are given below. The first one is a classical result from stochastic approximation that uses the fact that $X(t)$ is noisy Euler discretization of a Lipchitz-continuous differential equation. The second lemma states that the popularity in the cache is a continuous function of ρ.

Lemma 1. *Let $f : \mathcal{P} \to \mathrm{span}(\mathcal{P})$ be a Lipchitz continuous function with constant aL such that $\sup_{x \in \mathcal{P}} \|f(x)\|_\infty \leq a \leq 1$ and $f(x) - x \in \mathcal{P}$. Let X be a $\mathcal{P}$-valued stochastic process adapted to a filtration $\mathcal{F}$ such that $\mathbf{E}[X(t+1) - X(t) \mid \mathcal{F}_t] = f(X(t))$ and $\mathbf{E}\left[\|X(t+1) - X(t)\|_\infty^2\right] \leq a$. Then, the ODE $\dot{x} = f(x)$ has a unique solution $x_{X(0)}$ that starts in $X(0)$ and for any $T > 0$,*

$$\mathbf{E}\left[\sup_{t \leq T/a} \left\|X(t) - x_{X(0)}(t)\right\|_\infty^2\right] \leq T(2L+1)\exp(2TL)a.$$

Lemma 2. *Let $g_m : \mathcal{P} \to [0,1]$ be the function defined by $g_m(\rho) = \rho_{1,\theta}$, where $\theta = \sup\{b : \rho_{0,b} \geq m\}$. The function $g_m(\rho)$ is Lipschitz-continuous on $\mathcal{P}$ with the constant 2.*

Let us show that f satisfies the assumption of Lemma 1. It should be clear that f is Lipschitz-continuous with constant a. Moreover, $P_{\alpha,b}(t)$ changes if the requested item has a request time prior to b. If this item is k, then $P_{\alpha,b}(t+1) = P_{\alpha,b}(t) + a^{1-\alpha}(p_k)^{\alpha}$. This shows that

$$\begin{aligned} &\mathbf{E}[P_{\alpha,b}(t+1) - P_{\alpha,b}(t) \mid \mathcal{F}_t] \\ &\quad = \sum_{k=1}^{n} p_k a^{1-\alpha} (p_k)^{\alpha} \mathbf{1}_{\{\tau_k(t) < b\}} = f_{\alpha,b}(P(t)) \end{aligned}$$

Last, the second moment of the variation of $P(t)$ is bounded:

$$\begin{aligned} \mathbf{E}\left[\|P(t+1) - P(t)\|_\infty^2 \,|\mathcal{P}\right] &= \mathbf{E}\left[\sup_{\alpha,b} |P_{\alpha,b}(t+1) - P_{\alpha,b}(t)|^2 \,|\mathcal{F}_t\right] \\ &= \mathbf{E}\left[|P_{0,t}(t+1) - P_{0,t}(t)|^2 \,|\mathcal{F}_t\right] \\ &= \sum_{k=1}^{n} a p_k = a. \end{aligned}$$

This implies that for each $T > 0$, there exists a constant C such that $\mathbf{E}\left[\sup_{t \leq T/a} \|P(t) - \rho(t)\|_\infty^2\right] \leq Ca/4$. Lemma 2 concludes the proof for LRU.

Proof of Lemma 1. The solution of the ODE $\dot{x} = f(x)$ that starts in $X(0)$ satisfies $x(t) = X(0) + \int_{s=0}^{t} f(x(s))ds$. Let

$E(t)$ be such that

$$X(t) = X(0) + \sum_{s=0}^{t-1} f(X(s)) + E(t)$$
$$= X(0) + \int_0^{t-1} f(X(\lfloor s \rfloor))ds + E(t).$$

We have:

$$\|X(t) - x(t)\|_\infty \le \int_{s=0}^{t} \|f(X(\lfloor s \rfloor)) - f(x(s))\|_\infty + \|E(t)\|_\infty$$
$$\le aL \int_{s=0}^{t} \|X(\lfloor s \rfloor) - x(s)\|_\infty + \|E(t)\|_\infty ,$$

where we used that f is Lipschitz-continuous of constant La.

Let $\bar{X}(t)$ be the the piecewise-linear interpolation of X such that $\bar{X}(t) = X(t)$ when $t \in \mathbf{Z}^+$. We have:

$$\|X(\lfloor s \rfloor) - x(s)\|_\infty \le \left\|X(\lfloor s \rfloor) - \bar{X}(s)\right\|_\infty + \left\|\bar{X}(s) - x(s)\right\|_\infty$$
$$\le a + \left\|\bar{X}(s) - x(s)\right\|_\infty ,$$

where we used that $\|f(x)\|_\infty \le a$.

This shows that for any $t \le T/a$ (with $t \in \mathbf{Z}^+$):

$$\left\|\bar{X}(t) - x(t)\right\|_\infty \le aL \int_{s=0}^{t} \left\|\bar{X}(s) - x(s)\right\|_\infty + a^2 Lt + \|E(t)\|_\infty$$
$$\le \exp(aLt)(a^2 Lt + \sup_{s \le t} \|E(s)\|_\infty)$$
$$\le \exp(LT)(aLT + \sup_{t \le T/a} \|E(t)\|_\infty),$$

using Gronwall's inequality.
By assumption,

$$\mathbf{E}\left[\|E(t+1) - E(t)\|_\infty^2 \mid \mathcal{F}_t\right]$$
$$= \mathrm{var}\left[\|X(t+1) - X(t) \mid \mathcal{F}_t\|_\infty\right]$$
$$\le \mathbf{E}\left[\|X(t+1) - X(t)\|_2^2 \mid \mathcal{F}_t\right]$$
$$\le a^2.$$

As $E(t)$ is a martingale, this implies that

$$\mathbf{E}\left[\sup_{t \le T/a} \|E(t)\|_\infty^2\right] \le \mathbf{E}\left[\|E(T)\|_\infty^2\right] \le aT.$$

□

Proof of Lemma 2. Let $\rho, \rho' \in \mathcal{P}$. By definition of $\mathcal{P}$, there exist x and x' such that $\rho_{\alpha,b} = a^{1-\alpha}\sum_{k=1}^n (p_k)^\alpha x_{k,b}$ and $\rho'_{\alpha,b} = a^{1-\alpha}\sum_{k=1}^n (p_k)^\alpha x'_{k,b}$. Let θ, θ' be such that $\rho_{0,\theta} = \rho'_{0,\theta'} = m$ and assume without loss of generality that $\theta' \le \theta$. As $x_{k,b}$ is non-increasing in b, this implies that $x_{k,\theta} \ge x_{k,\theta'}$. Hence, we have:

$$|\rho_{1,\theta} - \rho_{1,\theta'}| = \sum_{k=1}^n p_k(x_{k,\theta} - x_{k,\theta'}) \le \sum_{k=1}^n a(x_{k,\theta} - x_{k,\theta'})$$
$$= |\rho_{0,\theta} - \rho_{0,\theta'}| \le \left|\rho_{0,\theta} - \rho'_{0,\theta'}\right| + \left|\rho'_{0,\theta'} - \rho_{0,\theta'}\right|$$
$$= |\rho_{0,\theta'} - \rho_{0,\theta'}| \le \|\rho - \rho'\|_\infty . \quad (23)$$

Therefore:

$$|g_m(\rho) - g_m(\rho')| = \left|\rho_{1,\theta} - \rho'_{1,\theta'}\right|$$
$$\le |\rho_{1,\theta} - \rho_{1,\theta'}| + \left|\rho_{1,\theta'} - \rho'_{1,\theta'}\right|$$
$$\le 2\|\rho - \rho'\|_\infty ,$$

where the last inequality comes from (23). □

2) Generalization to h-LRU: The proof for h-LRU is almost identical to the proof for LRU. For simplicity, we focus on the case of 2-LRU. The proof is similar for $h \ge 3$.

We define the quantities $\rho_{\alpha,\ell,b}(t)$ and $P_{\alpha,\ell,b}(t)$ by

$$\rho_{\alpha,\ell,b}(t) = a^{1-\alpha}\sum_{k=1}^n (p_k)^\alpha x_{k,\ell,b}(t);$$
$$P_{\alpha,\ell,b}(t) = a^{1-\alpha}\sum_{k=1}^n (p_k)^\alpha \mathbf{1}_{\{\tau_k(t) \ge b \wedge L_{k,\ell}(t)\}}$$

and Equation (16) implies that

$$\dot{\rho}_{\alpha,2,b} = a(\rho_{\alpha+1,2,\theta_2(t)} + \rho_{\alpha+1,1,\theta_1(t)} - \rho_{\alpha+1,2,\theta_1(t)} - \rho_{\alpha+1,2,b}). \quad (24)$$

Lemma 2 implies that the quantity $g_{m,\ell}(\rho) = \rho_{1,\ell,\theta}$, where θ is such that $\rho_{0,\ell,\theta} = m_\ell$, is a Lipschitz function of ρ with constant 2. It follows that the right-side of the ODE Equation (24) is Lipschitz-continuous with constant $4a$. As for LRU, the right side of Equation (24) is the average variation of $P_{\alpha,2,b}$ and that the second moment of the variation is bounded by a. Lemma 1 concludes the proof for 2-LRU.

As for LRU, we can transform (16) into a PDE by using the change of variables $y_{k,\ell,s}(t) = x_{k,\ell,t-s}(t)$ and $T_\ell(t) = t - \theta_\ell(t)$. For example, for $\ell = 2$, the fixed point y of this PDE satisfies

$$0 = p_k(y_{k,2,T_2} + y_{k,1,T_1} - y_{k,2,T_1} - y_{k,2,s}) - \frac{\partial}{\partial s} y_{k,2,s}.$$

The solution of this ODE in s is given by

$$y_{k,2,s} = (y_{k,2,T_2} - y_{k,2,T_1} + y_{k,1,T_1})(1 - e^{-p_k s}) \quad (25)$$
$$= \frac{y_{k,1,T_1}}{1 + e^{-p_k T_2} - e^{p_k T_1}}(1 - e^{-p_k s}), \quad (26)$$

where we use (25) for $s = T_1$ and $s = T_2$ to obtain (26).

In Section IV-B1, we have shown that $y_{k,1,T_1} = 1 - e^{-p_k T_1}$ where T_1 is such that $\sum_{k=1}^n y_{k,1,T_1} = m$. One can verify that replacing $y_{k,1,T_1}$ by $1 - e^{-p_k T_1}$ in Equation (26) with $s = T_2$ leads to Equation (11).

3) LRU($\mathbf{m}$): We now highlight the main differences with the case of h-LRU. They are mainly due to the non-continuity of the right-side of the differential equation (19).

As before, let $P_{\alpha,b}(t) = a^{1-\alpha}\sum_{k=1}^n (p_k)^\alpha \mathbf{1}_{\{\sigma_k(t) \ge b\}}$, where $a = \max_{k=1}^n p_k$. We also define $f : \mathcal{P} \to \mathrm{span}(\mathcal{P})$ by $f_{\alpha,b}(\rho) = a(\rho_{\alpha+1,\theta_{\zeta_b}-1} - \rho_{\alpha+1,b})$, where θ_ℓ and ζ_b are two functions of ρ that are defined by

$$\rho_{0,\theta_\ell} = m_\ell + \cdots + m_h \text{ and } \theta_{\zeta_b} \le b < \theta_{\zeta_b+1}.$$

As for the the cases of LRU and h-LRU, one can verify that $f_{\alpha,b}$ is the average variation of $P_{\alpha,b}(t)$ during one time

step and that the second moment of the average variation is bounded by a^2. Moreover, if x is a solution of the differential equation (19), then $\rho_{\alpha,b}(t) = \sum_{k=1}^{n} x_{k,b}(t)$ is a solution of the differential equation $\dot{\rho} = f(\rho)$.

The next lemma states some key properties of the function f. In particular, *(i)* quantifies what we mean by partially one-sided Lipschitz. Its proof is given below.

Lemma 3. *For any $\rho, \rho' \in \mathcal{P}$ and $\alpha \geq 1$, we have:*
(i) $(\rho_{0,b} - \rho'_{0,b})(f_{0,b}(\rho) - f_{0,b}(\rho')) \leq 2a \left\|\rho - \rho'\right\|_\infty^2$;
(ii) $\left\|f(\rho)\right\|_\infty \leq a$;
(iii) $|f_{\alpha,b}(\rho) - f_{\alpha,b}(\rho')| \leq |f_{0,b}(\rho) - f_{0,b}(\rho')| + 3\left\|\rho - \rho'\right\|_\infty$.

Let us denote by $V(t) \in \mathcal{P}$ the vector defined by $V_{\alpha,b}(t) = P_{\alpha,b}(t+1) - P_{\alpha,b}(t) - f_{\alpha,b}(P(t))$. We have $\mathbf{E}\left[V(t) \mid \mathcal{F}_t\right] = 0$ and $\mathbf{E}\left[\left\|V(t)\right\|_\infty^2 \mid \mathcal{F}_t\right] \leq a^2$. Moreover, the definition of $\rho(t+1) = \rho(t) + \int_0^1 f(\rho(t+s))ds$ implies that $(P_{0,b}(t+1) - \rho_{0,b}(t+1))^2$ equals

$$\begin{aligned}
&(P_{0,b}(t) - \rho_{0,b}(t) + V_{0,b}(t) + f_{0,b}(P(t)) + \int_0^1 f_{0,b}(\rho(t+s))ds)^2 \\
&= (P_{0,b}(t) - \rho_{0,b}(t))^2 + \Big[V_{0,b}(t) + f_{0,b}(P(t)) \\
&+ \int_0^1 f_{0,b}(\rho(t+s))ds\Big]^2 + 2\big(P_{0,b}(t) - \rho_{0,b}(t)\big)\big(V_{0,b}(t)\big) \\
&+ 2\Big(P_{0,b}(t) - \rho_{0,b}(t)\Big)\Big(f_{0,b}(P(t)) + \int_0^1 f_{0,b}(\rho(t+s))ds\Big) \quad (27)
\end{aligned}$$

In expectation, the second term is smaller than $9a^2$, the third is 0. For $\alpha = 0$, the last term equals

$$\begin{aligned}
&2\int_0^1 \Big(P_{0,b}(t) - \rho_{0,b}(t)\Big)\Big(f_{0,b}(P(t)) + f_{0,b}(\rho(t+s))\Big)ds \\
&= 2\int_0^1 \Big(P_{0,b}(t) - \rho_{0,b}(t+s)\Big)\Big(f_{0,b}(P(t)) + f_{0,b}(\rho(t+s))\Big)ds \\
&+ 2\int_0^1 \Big(\rho_{0,b}(t+s) - \rho_{0,b}(t)\Big)\Big(f_{0,b}(P(t)) + f_{0,b}(\rho(t+s))\Big)ds \\
&\leq 4a\int_0^1 \left\|P(t) - \rho(t+s)\right\|_\infty^2 ds + 2a^2, \quad (28)
\end{aligned}$$

where we use Lemma 3(i) to bound the first term of the equality and Lemma 3(ii) for the second.

As $P_{\alpha,b}(b) = \rho_{\alpha,b}(b)$, Lemma 3(iii) implies that

$$\begin{aligned}
|P_{\alpha,b}(t) - \rho_{\alpha,b}(t)| &\leq \int_{s=b}^{t} |f_{0,b}(P(\lfloor s \rfloor)) - f_{0,b}(\rho(s))|\, ds \\
&+ 3\int_{s=b}^{t} \left\|P(\lfloor s \rfloor) - \rho(s)\right\|_\infty ds + \sum_{s=b}^{t-1} |V_{\alpha,b}(s)|\,. \quad (29)
\end{aligned}$$

As $\zeta_b(t)$ is a decreasing function of time that can take at most h values, it can be shown that

$$\begin{aligned}
\int_{s=b}^{t} |f_{0,b}(P(\lfloor s \rfloor)) - f_{0,b}(\rho(s))| &\leq h\left|\int_{s=b}^{t} f_{0,b}(P(\lfloor s \rfloor)) - f_{0,b}(\rho(s))\right| \\
&+ a\int_{s=b}^{t} \left\|P(\lfloor s \rfloor) - \rho(s)\right\|_\infty \quad (30)
\end{aligned}$$

Combining Equation (27), (28), (29) and (30) shows that

$$\mathbf{E}\left[\left\|P(t) - \rho(t)\right\|_\infty^2\right] \leq 9ah\int_0^t \mathbf{E}\left[\left\|P(\lfloor s \rfloor) - \rho(s)\right\|_\infty^2\right] ds + 14ha^2t.$$

By Gronwall's inequality, this is less than $\exp(9aht)14ha^2t$, which, when t is less than T/a this is less than Ca for $C = 14hT\exp(9Th)$. Lemma 2 concludes the result.

Proof of Lemma 3. The function $f : \mathcal{P} \to \mathrm{span}(\mathcal{P})$ is given by

$$f_{\alpha,b}(\rho) = a(\rho_{\alpha+1,\theta_{\zeta_b}-1} - \rho_{\alpha+1,b}), \quad (31)$$

where θ_ℓ and ζ_b are two functions of ρ that are defined by

$$\rho_{0,\theta_\ell} = m_\ell + \cdots + m_h \text{ and } \theta_{\zeta_b} \leq b < \theta_{\zeta_b+1}. \quad (32)$$

We begin by the proof of *(i)* which states that $(\rho_{0,b} - \rho'_{0,b})(f_{0,b}(\rho) - f_{0,b}(\rho')) \leq 2a\left\|\rho - \rho'\right\|_\infty^2$. Let $\rho, \rho' \in \mathcal{P}$ and let ζ_b and ζ'_b be defined as in Equation (32). We have

$$\begin{aligned}
&(\rho_{0,b} - \rho'_{0,b})(f_{0,b}(\rho) - f_{0,b}(\rho')) \\
&= a(\rho_{0,b} - \rho'_{0,b})(\rho'_{\alpha+1,b} - \rho_{\alpha+1,b} + \rho_{\alpha+1,\theta_{\zeta_b-1}} - \rho'_{\alpha+1,\theta'_{\zeta'_b-1}}) \\
&\leq a\left\|\rho - \rho'\right\|_\infty^2 + a(\rho_{0,b} - \rho'_{0,b})(\rho_{\alpha+1,\theta_{\zeta_b-1}} - \rho'_{\alpha+1,\theta'_{\zeta'_b-1}}).
\end{aligned}$$

We then distinguish three cases:

- If $\zeta_b = \zeta'_b$, then we can use Lemma 2 to show that we have $\left|\rho_{\alpha+1,\theta_{\zeta_b-1}} - \rho'_{\alpha+1,\theta'_{\zeta'_b-1}}\right| \leq \left\|\rho - \rho'\right\|_\infty$, which implies that $(\rho_{0,b} - \rho'_{0,b})(\rho_{\alpha+1,\theta_{\zeta_b-1}} - \rho'_{\alpha+1,\theta'_{\zeta'_b-1}}) \leq \left\|\rho - \rho'\right\|_\infty^2$.
- If $\zeta_b > \zeta'_b$, then Equation (32) implies that $\rho_{0,b} \geq m_{\zeta_b+1} + \cdots + m_h > \rho'_{0,b}$. $\zeta_b > \zeta'_b$ also implies that $\rho'_{\alpha+1,\theta'_{\zeta'_b-1}} > \rho'_{\alpha+1,\theta'_{\zeta_b-1}}$. Hence,

$$\begin{aligned}
&(\rho_{0,b} - \rho'_{0,b})(\rho_{\alpha+1,\theta_{\zeta_b-1}} - \rho'_{\alpha+1,\theta'_{\zeta'_b-1}}) \\
&\leq (\rho_{0,b} - \rho'_{0,b})(\rho_{\alpha+1,\theta_{\zeta_b-1}} - \rho'_{\alpha+1,\theta'_{\zeta_b-1}}) \leq \left\|\rho - \rho'\right\|_\infty^2,
\end{aligned}$$

  where the last inequality comes from Lemma 2.
- The case $\zeta_b < \zeta'_b$ is symmetric.

This concludes the proof of *(i)*. Point *(ii)* follows directly from Equation (31).

For point *(iii)*, we can mimic the proof of Equation (23). By definition of $\mathcal{P}$, there exists non-decreasing functions x and

x'. Assume without loss of generality that $\theta_{\zeta_b} \leq \theta'_{\zeta'_b}$ which implies that $x_{k,\theta_{\zeta_b}} \geq x_{k,\theta'_{\zeta'_b}}$ for all $k \in \{1 \dots n\}$. Thus:

$$\begin{aligned}
\left|\rho_{\alpha,\theta_{\zeta_b-1}} - \rho'_{\alpha,\theta'_{\zeta_b-1}}\right| &= \left|a^{1-\alpha}\sum_{k=1}^{n}(p_k)^\alpha(x_{k,\theta_{\zeta_b}} - x'_{k,\theta'_{\zeta'_b}})\right| \\
&\leq \left|a^{1-\alpha}\sum_{k=1}^{n}(p_k)^\alpha(x_{k,\theta_{\zeta_b}} - x_{k,\theta'_{\zeta'_b}})\right| + \|\rho-\rho'\|_\infty \\
&= a^{1-\alpha}\sum_{k=1}^{n}(p_k)^\alpha(x_{k,\theta_{\zeta_b}} - x_{k,\theta'_{\zeta'_b}}) + \|\rho-\rho'\|_\infty \\
&\leq \sum_{k=1}^{n} p_k(x_{k,\theta_{\zeta_b}} - x_{k,\theta'_{\zeta'_b}}) + \|\rho-\rho'\|_\infty \\
&\leq \left|\rho_{0,\theta_{\zeta_b-1}} - \rho'_{0,\theta'_{\zeta_b-1}}\right| + 2\|\rho-\rho'\|_\infty
\end{aligned}$$

□

F. Proof of the Insensitivity of LRU

For each k, the requests of k are generated according to a stationary point process R_k. For $t < s$, $R_k[t,s)$ is the number of requested of item k during a time interval $[t,s]$. Let $\vartheta_k(t)$ be the time elapsed since the last request of item k. Without loss of generality, in the rest of the proof, we assume that the request process is simple (*i.e.* that with probability 1, the time between two consecutive requests of an item is never 0). If it is not the case, one can suppress any of the two request and obtain the same behavior of the LRU cache. Hence, the process (R_k, ϑ_k) is a stationary marked point process that satisfies the Hypothesis 1.1.1 of [2].

As R is stationary, the probability that the item k is requested during a time interval $[t, t+x]$ does not depend on t. Let $\tilde{F}_k(x)$ denote this quantity. We have:

$$\tilde{F}_k(x) = \mathbf{P}\left[R_k[t,t+x] \geq 1\right] = \mathbf{P}\left[R_k[0,x] \geq 1\right].$$

We also define $F_k(x)$ that is the probability that the time between two arrivals is smaller than x. As (R_k, ϑ_k) is a stationary marked point process, this quantity is well defined and can be expressed as

$$\begin{aligned}
F_k(x) &= \mathbf{P}\left[R_k[t,t+x] \geq 1 \mid \text{ a request occurred a time } t\right] \\
&= \mathbf{P}\left[R_k[0,x] \geq 1 \mid \text{ a request occurred a time } 0\right]
\end{aligned}$$

Note that the definition of $F_k(x)$ only requires the process R_k to be stationary. When the process is a renewal process, $F_k(x)$ is the cumulative distribution function of the inter-request time.

By the inversion formula [2, Section 1.2.4], $\tilde{F}_k$ can be expressed as a function of F_k:

$$\tilde{F}_k(x) = \lambda_k \int_0^x (1 - F_k(t))dt, \tag{33}$$

where $\lambda_k = 1/\int_0^\infty (1-F_k(t))dt$ is the request rate of item k. This quantity only depends on F_k and not on the correlation between two arrivals.

To conclude the proof, we remark that the probability that an item k is in the cache when it is requested can be expressed in terms of the functions F_k and $\tilde{F}_\ell$ for $\ell \neq k$. Indeed, Let $\mathfrak{S}_{n,-k}$ be the set of permutation of $\{1 \dots k-1\} \bigcup \{k+1 \dots n\}$ (*i.e.* all integers between 1 and n except k). An item is in the cache at time t if it is among the m items that were last requested. Hence, the probability for item k to be in the cache at time t is

$$\sum_{\sigma \in \mathfrak{S}_{n,-k}} \mathbf{P}\left[\vartheta_k(t) \leq \vartheta_{\sigma(m)}(t), \vartheta_{\sigma(1)}(t) \leq \dots \leq \vartheta_{\sigma(n-1)}(t)\right].$$

This event conditioned on the fact that item k is requested at time t is the probability that item k is in the cache when it is requested. Hence, the hit rate is:

$$\sum_k \lambda_k \sum_{\sigma \in \mathfrak{S}_{n,-k}} \mathbf{P}\left(\begin{matrix}\vartheta_k(t) \leq \vartheta_{\sigma(m)}(t), \\ \vartheta_{\sigma(1)}(t) \leq \dots \leq \vartheta_{\sigma(n-1)}(t)\end{matrix}\middle| \begin{matrix}\text{item } k \text{ is} \\ \text{requested at } t\end{matrix}\right).$$

This quantity can clearly be expressed as a function of the F_k and $\tilde{F}_k$ which by Equation (33) can be expressed solely as a function of the F_k.