

# Stochastic Caching via Subset Entropy

Ravi Kumar   

Google Research, Mountain View, CA, USA

Roie Levin   

Rutgers University, Piscataway, NJ, USA

Joseph (Seffi) Naor   

Technion – Israel Institute of Technology, Haifa, Israel

Debmalya Panigrahi   

Duke University, Durham, NC, USA

---

## Abstract

A classic approach to beyond worst-case algorithm design is to impose stochastic assumptions on the input. However, a limiting feature of such stochastic analyses is that, by the min-max principle, performance on worst-case distributions mirrors that of randomized algorithms on worst-case inputs. In other words, the same shortcoming of worst-case analysis – its inability to distinguish between “easy” and “hard” instances – reappears as an inability to distinguish between “easy” and “hard” distributions. This raises a natural question: *Can we characterize “easy” input distributions with useful beyond worst-case bounds?*

A canonical example is the *stochastic caching* problem (Aho, Denning, and Ullman, 1971). When the page requests are drawn i.i.d. from the uniform distribution, the best achievable competitive ratio is  $O(\log k)$ , matching the performance of the best randomized algorithm on worst-case instances (Fiat, Karp, Luby, McGeoch, Sleator, and Young, 1991). However, when the input distribution has less entropy, intuition suggests that we should be able to do better by exploiting the information provided by the distribution. In this paper, we formalize this intuition by defining a new information-theoretic parameter of probability distributions called *subset entropy*. We then use this parameter to give a fine-grained characterization of the competitive ratio of stochastic caching, including a new analysis for the well-known LRU algorithm on stochastic inputs.

While our technical results are for the caching problem, we believe the broader principle – parameterizing algorithmic performance by an entropy measure of the input – is of independent interest and might apply to other online and stochastic optimization problems. Indeed, for many other fundamental problems such as (comparison-based) sorting, online matching, online load balancing, etc., the hardest stochastic instances involve high-entropy distributions. We hope our work is a step toward a broader theory of fine-grained algorithmic performance for this class of problems.

**2012 ACM Subject Classification** Theory of computation → Online algorithms

**Keywords and phrases** Online Algorithms, Beyond Worst-Case Analysis, Caching, Paging, Entropy

**Digital Object Identifier** 10.4230/LIPIcs.APPROX/RANDOM.2026.2

**Category** APPROX

**Related Version** *Arxiv version:* <https://arxiv.org/abs/2607.03947>

**Funding** *Joseph (Seffi) Naor:* Supported in part by ISF grant 3001/24 and United States – Israel BSF grant 2022418.

*Debmalya Panigrahi:* Supported in part by NSF grants CCF-2329230 and CCF-1955703.



© Ravi Kumar, Roie Levin, Joseph Naor, and Debmalaya Panigrahi;  
licensed under Creative Commons License CC-BY 4.0

Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2026).

Editors: Mohit Singh and Tom Gur; Article No. 2; pp. 2:1–2:24



Leibniz International Proceedings in Informatics

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

## 1 Introduction

Stochastic optimization is a well-known and extensively-studied framework for beyond worst-case analysis of algorithmic problems. The basic premise is to make an assumption about the input being drawn from a probability distribution, rather than being generated adversarially, thereby enabling algorithms that exploit structural properties of the distribution. However, a common shortfall of this strategy is that algorithmic performance on the worst-case input distribution often mirrors that of the best randomized algorithm for worst-case inputs. This equivalence, a consequence of the min-max principle (Yao [35]), is often useful for proving randomized lower bounds for algorithms but, conversely, also lower bounds the performance of the best stochastic algorithms. This is rather limiting because it brings back the main deficiency of worst-case analysis – its inability to separate “easy” from “hard” instances – as being unable to differentiate between “easy” and “hard” distributions. This raises the natural question: *can we obtain a finer-grained characterization of input distributions that yields beyond worst-case bounds for “easy” distributions?*

**The Caching Problem.** A case in point is the classical online *caching* (or *paging*) problem. In this problem, a sequence of page requests arrives online from a universe of  $n$  pages. The goal is to serve each request using a cache that can only hold  $k$  ( $\ll n$ ) pages at a time while minimizing the total number of page swaps. Caching is used in almost every extant computer system. It is also a focal point for modern developments in online algorithms such as the online primal-dual method, projections, and mirror descent [13, 12, 11]. The problem is considered well-understood in online adversarial models: in their seminal work, Sleator and Tarjan [33] showed that the *least-recently used* (LRU) policy is  $k$ -competitive, and that any deterministic online algorithm is at least  $k$ -competitive. When randomization is allowed, Fiat et al. [17] showed an elegant and improved  $O(\log k)$ -competitive algorithm. Unfortunately, this bound is asymptotically tight for randomized algorithms even when page requests are sampled i.i.d. from a known distribution, since no algorithm can do better on the uniform distribution over  $n = k + 1$  pages [17].

Yet, our understanding of the caching problem is wanting even in the simplest “beyond worst-case” setting, where each request is drawn independently from a fixed distribution. As noted, the uniform distribution *is* the worst-case distribution – but this is also the least compelling case for caching. If requests are uniform, there is no reason for a caching algorithm to prefer one page over another, and any caching algorithm is as good as any other. In contrast, if page requests are skewed toward a set of fewer than  $k$  pages, smart caching can be effective, as one can hope to identify the frequently requested pages and prioritize evicting others.

This leads to our main questions: *Which distributions admit better-than-worst-case competitive ratios? And which online caching algorithms attain tighter bounds for distributions that are caching-friendly?*

### 1.1 Our Results

We study the *stochastic caching* problem, where we assume the input sequence of requests is drawn independently from a fixed distribution. The *competitive ratio* of a stochastic caching algorithm is the ratio of the expected cost of the algorithm and the expected cost of an optimal offline solution, where both expectations are taken over the input distribution (and the algorithm, if randomized).

Our first contribution is a new measure called *subset entropy* to quantify the competitive ratio of a stochastic caching instance. A reasonable first guess for such a measure would be the *entropy* of the input distribution; indeed, this would align in the case of the uniform distribution over  $n = k + 1$  pages, where the entropy is  $\Theta(\log k)$ . However, if the distribution places  $1 - \varepsilon$  mass on one page, and  $\varepsilon/(n - 1)$  on each of the remaining pages, then the entropy is near zero, but this is essentially the same stochastic caching instance as before, but with a cache of size  $k - 1$ , and with the  $n - 1$  pages of low probability in the original instance.

This motivates our definition of subset entropy for a distribution  $\mathcal{D}$ : it is the maximum, taken over all subsets in the support of a distribution  $\mathcal{D}$ , of the entropy of  $\mathcal{D}$  conditioned to the subset. In the case of stochastic caching, we only need to consider subsets of size at most  $k$  in the support; we call this  $k$ -subset entropy and denote it  $\mathcal{H}(\mathcal{D}, k)$ , or  $\mathcal{H}$  for short when  $\mathcal{D}$  and  $k$  are clear from context. In Section 3, we show how to achieve competitive ratio  $O(\mathcal{H})$  for every input distribution  $\mathcal{D}$ .<sup>1</sup>

► **Theorem 1 (Informal).** *There is a caching algorithm that for a given distribution  $\mathcal{D}$  and a sequence of page requests drawn i.i.d. from  $\mathcal{D}$ , achieves a competitive ratio of  $O(\mathcal{H})$ .*

Note that  $\mathcal{H}$  is at most the worst case bound  $O(\log k)$ , but it is also much sharper for many distributions. In fact, we show in Section 2 that the only distributions for which  $\mathcal{H} = \Theta(\log k)$  are those that are nearly uniform, namely those for which some subset of  $\Omega(\text{poly}(k))$  pages have probabilities within a constant factor of each other. To our knowledge, the notion of subset entropy is novel and would be interesting to explore in other contexts as well. For example, does it encapsulate interesting information-theoretic properties of a distribution? We show one property in Section 2: informally, we show that it captures the cardinality of the largest (approximately) uniform distribution embedded in the overall distribution. A different question: is  $\mathcal{H}$  (exactly) computable efficiently? We answer this in the affirmative in Appendix E, given explicit access to the distribution. We hope this new parameter will be studied from other angles in the future.

The algorithm we use to establish Theorem 1 is natural: since  $\mathcal{D}$  is known, we keep the top  $k-1$  pages by probability in cache permanently and use the last cache slot to serve all other requests;<sup>2</sup> we call this algorithm TOP- $k$ . Indeed, in an important work from 1971, Aho, Delling, and Ullman [2] showed that this is the *optimal* online caching strategy in the i.i.d. setting.<sup>3</sup> Though the algorithm is unsurprising, the analysis is new. We show that TOP- $k$  is  $O(\mathcal{H})$ -competitive against the *offline* optimum. This quantifies the precise role of randomness in the input sequence on the competitive ratio, namely the difference between knowing the input *distribution* versus the actual *realization* of the input sequence.

The previous TOP- $k$  algorithm requires knowing the input distribution (we call this the *known* i.i.d. setting). Now, we consider the *unknown* i.i.d. setting where the caching algorithm no longer has access to the input distribution. Intuitively, we would still like to keep the top  $k$  pages by probability in the cache. Ordering pages by decreasing probability is equivalent to ordering them by increasing expected length of the interval between consecutive requests. Since we do not have access to the expected interval lengths, we use as a surrogate the length

<sup>1</sup> There do exist distributions for which  $\mathcal{H}$  is large and yet the distribution admits smaller competitive ratio. E.g., the uniform distribution with  $n = 10k$  pages has  $\mathcal{H} = O(\log k)$  but any algorithm is constant-competitive. To our knowledge this is true of most parameterized results, since it is rare to identify a parameter that guarantees *both* upper and lower bounds.

<sup>2</sup> For ease of analysis, we define the TOP- $k$  algorithm slightly differently from the optimal online strategy; see Section 3.

<sup>3</sup> In the stochastic setting, an optimal online solution is well-defined. It can be obtained from a Markov decision process via a dynamic program which, in general, can have an exponential-sized state space.

of the interval since the last request for every page – i.e., keep in cache the  $k$  most recently requested pages. This is none other than the well-known LRU algorithm, and our second result shows that it is  $O(\mathcal{H}^3)$ -competitive even in the unknown i.i.d. setting.

► **Theorem 2.** *When given a sequence of page requests drawn i.i.d. from an unknown distribution  $\mathcal{D}$ , the LRU caching algorithm achieves a competitive ratio of  $O(\mathcal{H}^3)$ .*

As intuition would predict, LRU approximately learns which pages in cache are most likely to be requested furthest in future and evicts these first. How well it does so depends on how predictable is this information, which in turn is captured by the subset entropy of  $\mathcal{D}$ . Since  $\mathcal{H} = O(\log k)$ , this also shows that LRU is  $O(\log^3 k)$ -competitive for unknown i.i.d. inputs. To our knowledge, this is the first analysis of LRU in a stochastic setting that yields an  $o(k)$ -competitive factor.

A more precise (although also not as lightweight) alternative to LRU is the *least-frequently used* (LFU) algorithm that always evicts the page with the least number of requests so far. Our third result shows that LFU is  $O(\mathcal{H})$ -competitive for i.i.d. inputs. This implies that up to constant factors, the known and unknown i.i.d. settings have the same competitive ratio.

► **Theorem 3.** *When given a sequence of page requests drawn i.i.d. from an unknown distribution  $\mathcal{D}$ , the LFU caching algorithm achieves a competitive ratio of  $O(\mathcal{H})$ .*

This theorem is not surprising in an asymptotic sense, since LFU eventually converges to the TOP- $k$  algorithm. But a priori, the rate of convergence could be arbitrarily slow depending on the difference between the probability values of pages requests. We show that the additive loss of the LFU algorithm vis-à-vis the TOP- $k$  algorithm is  $O(k\mathcal{H})$ , and is therefore absorbed in the multiplicative factor in Theorem 3.<sup>4</sup> In other words, LFU matches the performance of TOP- $k$  *even when it has not finished learning the input distribution*. We view this as an interesting insight into learning-augmented algorithms: it demonstrates the difference between statistically learning a distribution and learning it well enough to devise an effective algorithmic strategy.

## 1.2 Techniques and Overview

We start in Section 2 with our characterization of  $k$ -subset entropy in terms of the largest (up to size  $k$ ) approximately uniform distribution embedded in  $\mathcal{D}$ . Indeed, by definition, any conditional distribution in  $\mathcal{D}$  gives a lower bound on subset entropy. To show the converse, we use a factor-revealing LP to bound the maximum value of subset entropy when every approximately uniform distribution is size-constrained.

In Section 3, we show that the TOP- $k$  algorithm has a competitive ratio of  $O(\mathcal{H})$ , thereby establishing Theorem 1. The  $k$  pages with the largest values of  $p_i$  are permanently in the cache in the TOP- $k$  algorithm; we call them the *head* of the distribution and call the remaining pages the *tail*. The expected cost of the TOP- $k$  algorithm is immediate: it is the total probability in the tail, denoted  $p(\mathbb{T})$ . But how do we lower bound the expected cost of an offline optimal solution? The usual technique is to use dual fitting, i.e., write primal and dual LPs and then construct a feasible dual solution of sufficiently large value. Here, since the instance itself is random, so are the coefficients of the primal and dual LPs, and what we really have is a *distribution* over LPs. How do we argue about the expected value of the optimal solution of such a distribution? (Contrast this to randomized dual fitting [15] where the LPs are fixed but the primal and dual solutions are random.)

<sup>4</sup>  $k$  is required for all algorithms as an initialization cost from an empty cache.

We start by writing the “expected” primal-dual LPs whose coefficients are the expected values of the (random) coefficients over the distributions of the primal and dual LPs. For this LP, we construct a feasible dual solution of value  $p(\mathbb{T})$  that satisfies these “expected constraints”. Interestingly, this dual solution to the “expected” LP turns out to be quite different from those used in standard primal-dual analysis of caching algorithms (e.g., in [5]). However, we want the dual solution to be feasible for the actual realized constraints, not the expected constraints. Constraints for which the realized dual coefficients are “close” to their expected values (in a specific sense that we define later) are already approximately satisfied by the expected dual. The main challenge is to handle the “tail constraints”, namely those that realize very differently from their expected form. These realizations can happen in many different ways and tracking their feasibility precisely is onerous. Instead, we carefully zero out dual variables to make sure that the tail constraints are also satisfied. This creates a distribution over dual solutions, since the set of tail constraints, and therefore, zeroed out variables, depends on the particular realization. Nevertheless, crucially, for each realized dual instance, the corresponding dual is *always* feasible. The key is to show that since each constraint is a tail constraint with small probability, we do not lose too much in the objective by zeroing out dual variables. In fact, we precisely lose a factor of  $O(\mathcal{H})$ . Hence this process gives a randomly defined dual solution that is deterministically feasible for the corresponding random instance, and the expected value of the dual objective over these random solutions is  $\Omega(p(\mathbb{T})/\mathcal{H})$ .

In Section 4, we study the LRU algorithm and establish Theorem 2. Recall that LRU is an empirical implementation of the TOP- $k$  algorithm which evicts pages based on the *expected* length time between subsequent requests. Thus it is tempting to imagine that the ideal scenario for LRU is when the empirical length of a request interval exactly matches its expected value for every page. Unfortunately, this intuition is misleading. Consider a uniform input distribution  $\mathcal{D}$  over  $n = k + 1$  pages. If an instantiation creates a round-robin request sequence, then the empirical length of a request interval is exactly  $k + 1$ , i.e., its expected value. However, LRU is  $k$ -competitive for this input sequence! Therefore, we need to use the symmetry-breaking properties of a random sequence in our analysis. Suppose a page  $i$  is requested at time  $t$ . If page  $i$  is not in the cache at time  $t$ , then by the properties of LRU, page  $i$  is not among the last  $k$  distinct pages requested before time  $t$ . What is the probability of this event? By reversing time, this is the same as the probability that page  $i$  is not going to be among the next  $k$  distinct pages to be requested after time  $t$ . Now, define a random variable for every page denoting the time till the first request of the page. Although these times are dependent, to first approximation, we can imagine that each page  $i$  is running an independent Poisson process with rate  $p_i$ , i.e., we can approximate the length of the request intervals by exponential variables. Then, we use the properties of Poisson processes to bound the probability that at least  $k$  other processes fired before process  $i$ , eventually proving Theorem 2.

In Section 5, we study the LFU algorithm and prove Theorem 3. We note that LFU might suffer more cache misses than TOP- $k$  because of inversions in the ranked order of pages by empirical frequency counts vis-à-vis their true probabilities. We bound the effect of these inversions in two parts. First, we show that inversions due to pairs that have probabilities within (say) a factor of 2 only incur an expected cost  $O(p(\mathbb{T}))$ . It is crucial that these inversions do not cost us in the additive term because such inversions might persist for an arbitrarily long time (e.g., if two pages differ in probability by  $\varepsilon$ ). It remains to bound the cost due to inversions where one page is twice as likely as the other. We show that after  $O(\mathcal{H})$  requests for a page, its count will forever be higher (with good probability) than the count of any page with less than half its probability. This incurs an additive loss of  $O(k\mathcal{H})$  on the pages in the head, which is ultimately absorbed in the multiplicative factor.

**The Bigger Picture.** Having discussed our technical results for stochastic caching, we now return to the broader perspective: obtaining finer-grained analyses for stochastic optimization problems. We argue that identifying a distributional quantity, such as entropy, and analyzing algorithms with respect to this quantity, offers a promising direction for beyond worst-case analysis in stochastic settings. Indeed, for many fundamental problems – such as comparison-based sorting, online matching, online load balancing, and related tasks – the hardest stochastic instances correspond to high-entropy (often uniform) distributions. Intuitively, this is consistent with the view that distributions with higher entropy are less informative than those with lower entropy. We view our work as a first step toward a theory, based in entropy and its variants, that provides a refined analysis of algorithmic performance across stochastic versions of this rich class of problems.

### 1.3 Related Work

In a classic paper, Belady [8] introduced the optimal “furthest in the future” rule for evicting pages from a cache in the offline setting. Soon after, stochastic versions of the online caching problem were considered and adaptations of Belady’s rule to online distributional settings were proposed and analyzed [2, 18]. Over the next few decades, a rich array of input models were introduced for online caching; some highlights include the work of Lund et al. [26] and Pabbaraju and Vakilian [29], who obtained approximations against the best online strategy even in partial information settings, the Markov paging model of Karlin et al. [23], where the input sequence is a Markov chain, and the access graph setting of Borodin et al. [10] (see also [9, Ch. 5]), where the request sequence is generated by a simple walk on an “access graph” defined on the pages. Somewhat related to our work is that of Pandurangan and Upfal [31], who studied the relationship between the performance of the best online caching algorithm and the entropy of a request sequence generated by a stochastic process. In particular, they showed that entropy is not sufficient to capture the performance of online caching algorithms.

We note that many further variants of online caching have been proposed, e.g., when pages have non-uniform weights [4, 21, 37], non-uniform sizes [3, 5, 22], as well as non-standard caching models such as elastic caches [19], caching with time windows [20], and caching with dynamic weights [16], interleaved caching [25], and caching with machine learning predictions [27, 6]. In contrast to the i.i.d. setting that we study in this paper, many models try to capture *locality of reference*, i.e., the phenomenon that pages requested recently tend to be re-requested [24, 36, 7, 30]. The validity of locality of reference varies by application domain, e.g., is more applicable for centralized settings like operating systems on computing devices compared to distributed and resource-shared settings like data centers where the requests come from a large number of independent processes. Nevertheless, we view the extensive literature on stochastic paging – including our work – and that on models capturing locality of reference as complementary, since they capture different aspects of a fundamental problem. Indeed, this complementarity can sometimes be leveraged to obtain richer models that try to incorporate both phenomenon, such as Markov paging [23].

In our work, the input model is the classic i.i.d. setting, but we are interested in characterizing input distributions that are amenable to caching and obtaining better-than-worst-case bounds for such inputs. In this sense, it aligns with the recent focus on “beyond worst-case” algorithm design, and in particular, with the emerging field of learning-augmented algorithms, e.g., [1]. The main difference is that while the typical goal in the latter context is to use additional information about the input and automatically adapt the performance bounds to the *quality* of this information, our goal is to adapt performance bounds to the quality of the input distribution itself without using additional information about the input. We point the reader to an excellent survey of “beyond worst-case” caching models in [32, Ch. 24] for a detailed discussion of related models.

## 2 Problem Definition and Background

### 2.1 The Stochastic Caching Problem

There is a universe of  $n$  pages indexed  $[n] = \{1, \dots, n\}$ , a cache of size  $k \leq n$ , a request sequence of length  $T$  indexed by  $[T]$ , and a distribution  $\mathcal{D}$  on  $[n]$ . At each time  $t \in [T]$ , the algorithm gets a page request  $r_t \in [n]$ , where  $r_t$  is drawn independently from  $\mathcal{D}$ . The algorithm maintains at most  $k$  pages in the cache at all times with the constraint that page  $r_t$  must be in the cache at time  $t$ . Swapping in/out a page in the cache incurs unit cost; the goal is to minimize the total cost, i.e., the number of page swaps. We call this the *stochastic caching* problem.

The distribution  $\mathcal{D}$  is defined by probabilities  $p_1, \dots, p_n$ , where  $p_i > 0$  is the probability that page  $i \in [n]$  is requested at any given time. Clearly,  $\sum_i p_i = 1$ . We also denote  $p(S) = \sum_{i \in S} p_i$  for a subset  $S \subseteq [n]$ . Wlog, we index pages in decreasing order of probabilities, i.e.,  $p_1 \geq p_2 \geq \dots \geq p_n$ , breaking ties arbitrarily. The top  $k$  pages are called the *head* of the distribution  $\mathcal{D}$  and denoted  $\mathbb{H} := [k]$ ; the remaining pages  $\mathbb{T} := [n] \setminus [k]$  comprise the *tail* of the distribution. The total probability mass in the tail is denoted  $p(\mathbb{T}) := \sum_{i=k+1}^n p_i = 1 - p(\mathbb{H})$ . In this paper, we consider two cases: when  $\mathcal{D}$  is *known* and when  $\mathcal{D}$  is *unknown* to the algorithm.

### 2.2 Subset Entropy

For a distribution  $\mathcal{D}$  with support  $U$  and for a subset  $S \subseteq U$ , let  $\mathcal{D}_S$  denote the distribution  $\mathcal{D}$  conditioned on  $S$ , i.e.,  $\mathcal{D}_S(u) = \frac{\mathcal{D}(u)}{\sum_{u \in S} \mathcal{D}(u)}$  for  $u \in S$ . Let  $H(\cdot)$  denote the *Shannon entropy* function:  $H(\mathcal{D}_S) := -\sum_{u \in S} \mathcal{D}_S(u) \cdot \log \mathcal{D}_S(u)$ . (All logarithms throughout the paper are base 2 unless otherwise stated.)

We define *k-subset entropy* of a distribution  $\mathcal{D}$  as follows:

► **Definition 4** (*k-Subset Entropy*). For a discrete distribution  $\mathcal{D} : U \rightarrow [0, 1]$  and any positive integer  $k \leq |U|$ , the *k-subset entropy* of  $\mathcal{D}$  is defined as the following:

$$\mathcal{H}(\mathcal{D}, k) = \max_{S \subseteq U: |S| \leq k} H(\mathcal{D}_S). \quad (2.1)$$

While *k-subset entropy* suffices in the context of caching, an interesting information-theoretic counterpart is the unrestricted version that maximizes entropy of conditional distributions over *all* subsets of  $U$ . We call this *subset entropy*, and it is the special case of *k-subset entropy* when  $k = n$ .

Next, we prove an important fact about *k-subset entropy*, which we rely on in the rest of the paper.

► **Lemma 5** (*k-Subset Entropy Lemma*). Suppose we have a distribution  $\mathcal{D} : U \rightarrow [0, 1]$ . Let  $S_\ell = \{u \in U : 1/2^{\ell+1} < \mathcal{D}(u) \leq 1/2^\ell\}$  for any  $\ell \geq 1$  and let  $n_{\max} := \max_\ell |S_\ell|$ . Then,

$$\mathcal{H}(\mathcal{D}, k) = \Theta(\log(\min\{k, n_{\max}\})).$$

One direction is immediate: clearly  $\mathcal{H}(\mathcal{D}, k) \geq \log(\min(k, n_{\max}))$  since the set  $S$  in the definition (2.1) could be chosen as (a subset of size  $\leq k$  of) the  $S_\ell$  for which  $n_{\max} = |S_\ell|$ , and by the definition of Shannon entropy,  $H(\mathcal{D}_{S_\ell}) = \Theta(\log |S_\ell|)$ . Surprisingly, the converse holds: if  $\mathcal{H}(\mathcal{D}, k) = h$ , then there is a set  $S \subseteq U$ ,  $|S| \geq 2^{\Theta(h)}$ , over which the conditional distribution  $\mathcal{D}_S$  is uniform up to constants. Intuitively, this means that the *k-subset entropy* of a distribution is dictated by the largest (up to  $k$ ) approximately uniform distribution

embedded in it. The proof (Appendix A) is via dual fitting of a factor revealing LP: we show that even if we ignore the original distribution  $\mathcal{D}$  and choose the probabilities of the pages in the maximizing set  $S$  completely freely, if we are constrained to choosing at most  $n_{\max}$  pages within a factor 2 of each other, we cannot give  $S$  a conditional entropy greater than  $O(\log n_{\max})$ .

In Appendix E, we also show that the maximizing set in the Definition (2.1) is always a subinterval of  $U$ , when arranged in sorted order of probabilities. This gives an efficient algorithm for computing  $k$ -subset entropy *exactly*. Nevertheless, the approximate characterization above is convenient to work with in our applications.

We use henceforth the shorthand  $\mathcal{H}$  to denote the  $k$ -subset entropy  $\mathcal{H}(\mathcal{D}, k)$ , since it is implicit that the distribution  $\mathcal{D}$  is the input distribution and the parameter  $k$  is the cache size.

### 3 Stochastic Caching for a Known Distribution

We begin with the known distribution setting. Recall our simple algorithm TOP- $k$ : keep the pages of the head  $\mathbb{H}$  in cache; when a page from the tail  $\mathbb{T}$  is requested, swap out an arbitrary page in cache for the requested page, then immediately swap the same page back. The main result of this section, which establishes Theorem 1, is the following:

► **Theorem 6.** *The competitive ratio of the algorithm TOP- $k$  for the stochastic caching problem for input distribution  $\mathcal{D}$  is  $O(\mathcal{H})$ .*

To establish Theorem 6, we first bound the expected cost of the TOP- $k$  algorithm.

► **Lemma 7.** *The expected cost of the TOP- $k$  algorithm for each request is  $2p(\mathbb{T})$ .*

**Proof.** For each request, the probability that the requested page is not in the cache is exactly  $p(\mathbb{T})$ , in which case the algorithm performs two evictions. ◀

The harder challenge is to lower bound the expected cost of an offline optimal solution. To this end, we fit a dual to the LP relaxation of the knapsack-cover (or KC) integer programming formulation for caching (see, e.g., [5])<sup>5</sup>. As noted in the introduction, while focusing on LP duality is a reasonable instinct, the main challenge is that the LP is defined via a distribution over coefficients, i.e., both the primal and dual LPs are random. We have to produce a feasible dual solution for *every* realization of the random instance, and then bound the expected objective value across the realizations.

Fortunately, since we are in an i.i.d. setting, the distribution over dual constraints that apply at any given time  $t$  is time-invariant, i.e., is independent of  $t$ . We view this distribution over dual constraints in two parts – the set of constraints that apply within a high probability interval, and the remaining constraints that are in the tail of the distribution. For the former, we set dual variables corresponding to the “expected” constraints and these are approximately feasible. For the latter, there are a large number of possibilities and it is onerous to track feasibility across all these realizations. Instead, we carefully zero out a subset of dual variables to ensure feasibility for the constraints in the tail of the distribution. Of course, the set of dual constraints that fall in the tail region is random, and therefore the set of zeroed out variables is also random. Nevertheless, we show that since a constraint

<sup>5</sup> Note that our choice of the KC inequality strengthened formulation is for convenience; it is not strictly necessary because the standard LP formulation is already integral.

is in the tail with low probability (by standard concentration inequalities), this last step of zeroing out variables to ensure feasibility has limited impact on the expected value of the dual objective.

► **Lemma 8.** *The expected cost of an offline optimal solution is  $\Omega\left(\frac{p(\mathbb{T})}{\mathcal{H}}\right)$ .*

**Proof Sketch.** Let  $x(i, r)$  be the indicator variable for the event that page  $i$  is evicted in the time interval  $I(i, r) := [t(i, r) + 1, t(i, r + 1) - 1]$ , where  $t(i, r)$  is the time of the  $r$ th request to page  $i$ . Let  $\rho(i, t)$  be the number of requests to page  $i$  up to and including time  $t$ . Consider the formulation:

|                                                                                                                                                                                                                                                                                                                                  |                                                                                                                                                                                                                                                                                                                                  |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| $\min \sum_{i=1}^n \sum_{r=1}^{\rho(i, T)} x(i, r) \quad \text{subject to}$ $\sum_{i \in S \setminus \{r_t\}} x(i, \rho(i, t)) \geq  S  - k \quad \begin{array}{l} \forall t \in [T] \\ \forall S \subseteq [n] \end{array}$ $x(i, r) \geq 0 \quad \begin{array}{l} \forall i \in [n] \\ \forall r \in [\rho(i, T)] \end{array}$ | $\max \sum_{t=1}^T \sum_{S \subseteq [n]} ( S  - k) \cdot y_S^t \quad \text{subject to}$ $\sum_{t \in I(i, r)} \sum_{S: i \in S} y_S^t \leq 1 \quad \begin{array}{l} \forall i \in [n] \\ \forall r \in [\rho(i, T)] \end{array}$ $y_S^t \geq 0 \quad \begin{array}{l} \forall t \in [T] \\ \forall S \subseteq [n] \end{array}$ |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

This formulation is standard, but our dual fitting is not (see [5] for the usual way the online dual is fit using multiplicative weights). We start by finding a stationary solution to the *expected* dual, i.e., find a vector  $y$  that is the same for each  $t \in [T]$ , such that the expected dual constraint is satisfied.

Intuitively, we should expect every time  $t$  to lie in an interval of length  $1/p_i$  between subsequent requests to page  $i$ , and hence these expected constraints would ask that there be less than 1 total  $y$  mass on page subsets containing  $i$  on these intervals, i.e., that there be less than  $p_i$  mass on these subsets per time step  $t$ . Define

$$y_S^t = \begin{cases} p_i - p_{i+1} & \text{if } S = [i] \text{ and } i \in \mathbb{T} \setminus \{n\}, \\ p_n & \text{if } S = [n], \\ 0 & \text{otherwise.} \end{cases}$$

Then, by design, for all  $i \in \mathbb{T}$  we have  $\sum_{S: i \in S} y_S^t = \sum_{j=i}^{n-1} (p_j - p_{j+1}) + p_n = p_i$ , as desired. For the remaining  $i \in \mathbb{H}$ , we have  $\sum_{S: i \in S} y_S^t = \sum_{j=k+1}^{n-1} (p_j - p_{j+1}) + p_n = p_{k+1} \leq p_i$ . Furthermore, the dual objective (per time step) is  $\sum_S (|S| - k) \cdot y_S^t = n \cdot p_n + \sum_{j=1}^n j \cdot (p_j - p_{j+1}) = \sum_{i \in \mathbb{T}} p_i$ .

However, this is too good to be true! If this were a legal dual solution, weak duality would imply that TOP- $k$  in fact achieves competitive ratio  $O(1)$ , which we know is impossible, as demonstrated by the uniform distribution case. Indeed, for the dual to witness a true lower bound on the primal program, it must be feasible with probability 1, and our stationary solution, as described, might violate dual constraints corresponding to intervals that happen to be longer than their expectation.

To construct our actual dual, we start with the stationary  $y$  described above. We then show how to carefully drop some coordinates to account for the random fluctuation in interval lengths. Since the dual is a packing problem, we can always make a solution feasible this way. The crux is to do this while approximately preserving the objective value. In Appendix B, we give the details of our modified dual construction and show that we only lose an  $O(\mathcal{H})$  factor in the objective. ◀

## 4 The LRU Algorithm

In this section, we analyze the popular *least recently used* (LRU) algorithm for stochastic caching. Note that LRU does not use distributional information, and hence can be used in the unknown distribution setting. Our main result is:

► **Theorem 9.** *The LRU algorithm has a competitive ratio  $O(\mathcal{H}^3)$  for the stochastic caching problem.*

We now provide the main ideas needed for proving this and defer the technical claims to Appendix C. We saw in Section 3 that the expected cost of an offline optimal solution is  $\Omega\left(\frac{p(\mathbb{T})}{\mathcal{H}}\right)$ . Hence, it suffices to bound the expected cost of LRU as  $O(\mathcal{H}^2 \cdot p(\mathbb{T}))$ . Note that LRU has a cache miss when the requested page is not among the last  $k$  distinct pages requested. Thus, for any page  $i$ , the probability that  $i$  incurs a cache miss (at any point of time) is the same as that of page  $i$  not being among the first  $k$  distinct pages requested in a sequence of pages drawn from  $\mathcal{D}$ . We denote this probability  $q_i$  for page  $i$ . With this notation, the expected cost of the LRU algorithm for any request is  $\sum_{i \in [n]} p_i q_i$ . Our goal from here on is to show

$$\sum_{i \in [n]} p_i q_i \leq O(\mathcal{H}^2 \cdot p(\mathbb{T})). \quad (4.2)$$

Before we begin, note that when  $p(\mathbb{T}) > 1/2$ , any algorithm, even one that incurs a cache miss at every time step, has cost  $O(p(\mathbb{T}))$ . Thus, we can assume without loss of generality that:

► **Assumption 10.** *The tail has total probability  $p(\mathbb{T}) \leq 1/2$ .*

Note that the contribution of the tail to the LHS of (4.2) is easy to handle, since the  $q_i$ 's are probabilities.

► **Observation 11** (Bounding  $\mathbb{T}$ ). *The contribution of  $\mathbb{T}$  is bounded as  $\sum_{i \in \mathbb{T}} p_i q_i \leq p(\mathbb{T})$ .*

This leaves us with the task of bounding the contribution of the head  $\mathbb{H}$  to the LHS of (4.2).

### 4.1 Probability Levels

We first group the pages in the head  $\mathbb{H}$  into doubling intervals by probability values  $p_i$ . In particular, for  $\ell \geq 0$ , we use  $L_\ell$  to denote the set of all pages  $i$  with  $p_i \in [1/2^{\ell+1}, 1/2^\ell]$ ; we say that the pages in  $L_\ell$  are at level  $\ell$  and denote  $|L_\ell| = n_\ell$ . Furthermore, we say that  $\ell$  is a *dominant level* if  $\sum_{i \in L_\ell} p_i > 1/2$ ; else,  $\ell$  is a *non-dominant level*. Note that there is either no dominant level or a single unique dominant level.

Let  $X_\ell$  be the random variable denoting the number of pages in  $L_\ell$  that do not appear among the first  $k$  distinct pages in any particular sequence of page requests. Let  $M_\ell$  be the multiset of page requests up to and including the first time all the pages in  $L_\ell$  are requested. We denote  $Y_\ell = |M_\ell|$  and  $Z_\ell$  to be the number of page requests from  $M_\ell$  to pages in  $\mathbb{T}$ .

We need certain structural properties of these random variables.

► **Lemma 12.** *The following three properties hold for  $X_\ell$ ,  $Y_\ell$ , and  $Z_\ell$ :*

(i) *For any level  $\ell$  with  $n_\ell > 0$ ,*

$$\sum_{i \in L_\ell} p_i q_i \leq \frac{1}{2^\ell} \cdot \sum_{i \in L_\ell} q_i = \frac{1}{2^\ell} \cdot \mathbb{E}[X_\ell]. \quad (4.3)$$

(ii) Let  $\ell$  be a level with  $n_\ell > 0$ , and let  $H_{n_\ell}$  denote the  $n_\ell$ th Harmonic number. Then,

$$\mathbb{E}[Y_\ell] \leq 2^{\ell+1} \cdot H_{n_\ell}. \quad (4.4)$$

(iii) Suppose  $\ell$  is a non-dominant level. Then, we have

$$\Pr[Z_\ell \geq t] \leq \begin{cases} 2 \cdot p(\mathbb{T}) \cdot 2^{\ell+1} \cdot \mathcal{H} & \text{for } t = 1, \\ \left(2e^2 \cdot p(\mathbb{T}) \cdot 2^{\ell+1} \cdot \mathcal{H}\right)^t & \text{for } t \geq 2. \end{cases} \quad (4.5)$$

$$(4.6)$$

Next, using these properties, we bound the total contribution of any single level  $L_\ell$  to  $\sum_i p_i q_i$  (i.e., to the LHS in (4.2)):

▷ Claim 13. For any level  $\ell$ , we have  $\sum_{i \in L_\ell} p_i q_i \leq 4\mathcal{H} \cdot p(\mathbb{T})$ .

Proof. By (4.3),  $\sum_{i \in L_\ell} p_i q_i \leq \frac{\mathbb{E}[X_\ell]}{2^\ell} \leq \frac{\mathbb{E}[Z_\ell]}{2^\ell}$  since  $X_\ell \leq Z_\ell$  for any level  $\ell$ .

So, it suffices to bound  $\mathbb{E}[Z_\ell]$ . Recall that  $M_\ell$  denotes the sequence of requests until all pages in  $L_\ell$  are requested. Color every request in  $M_\ell$  red, blue, or green depending on whether it is for a page in  $\mathbb{T}$ , for a page in  $\mathbb{H}$  but not in  $L_\ell$ , and for a page in  $L_\ell$ , respectively. Variable  $Z_\ell$  counts the number of red requests in  $M_\ell$ .

Let  $N$  denote the number of green requests in  $M_\ell$  and let  $\Gamma_i$ , for  $i \geq 1$ , be the number of red requests between the  $(i-1)$ st and  $i$ th green request. Note that the random variables  $\Gamma_i$  are independent and identically distributed; let  $\Gamma$  denote the common distribution. Furthermore,  $N$  is independent of the sequence  $\Gamma_1, \Gamma_2, \dots$ . Therefore, by Wald's identity we have

$$\mathbb{E}[Z_\ell] = \sum_i \mathbb{E}[\Gamma_i] = \mathbb{E}[\Gamma] \cdot \mathbb{E}[N]. \quad (4.7)$$

We bound the two terms in the RHS of (4.7) separately. For  $\mathbb{E}[N]$ , we use standard coupon collector bounds since the probabilities of pages in level  $\ell$  are uniform up to a factor of 2:  $\mathbb{E}[N] \leq 2n_\ell H_{n_\ell} \leq 2n_\ell \cdot \mathcal{H}$ .

To bound  $\mathbb{E}[\Gamma]$ , note that  $\Gamma$  denotes the number of red requests before the first green request. Denote  $p(\ell) := \sum_{i \in L_\ell} p_i$  as the probability of a green request and recall that  $p(\mathbb{T})$  is the probability of a red request. Then,  $\mathbb{E}[\Gamma] = \left(\frac{p(\ell)}{p(\ell) + p(\mathbb{T})}\right)^{-1} - 1 = \frac{p(\mathbb{T})}{p(\ell)} \leq \frac{p(\mathbb{T})}{n_\ell \cdot 2^{-(\ell+1)}}$ .

Therefore, by (4.7),  $\mathbb{E}[Z_\ell] = \mathbb{E}[\Gamma] \cdot \mathbb{E}[N] < \frac{p(\mathbb{T})}{n_\ell \cdot 2^{-(\ell+1)}} \cdot 2n_\ell \cdot \mathcal{H} = 4\mathcal{H} \cdot p(\mathbb{T}) \cdot 2^\ell$ , thus yielding the claim. ◀

The above bound is sufficiently tight for levels  $\ell$  where the probabilities  $p_i$  are comparable to (or smaller than) the value of the tail probability  $p(\mathbb{T})$ . However, in higher levels with larger probabilities, the above bound is quite loose and we need a stronger guarantee.

Let  $r_1$  be the largest index  $i$  such that  $p_i \geq 4e^2 \cdot \mathcal{H} \cdot p(\mathbb{T})$  and let  $\ell_1$  be the level that contains page  $p_{r_1}$ . Roughly speaking, our tighter bound for levels  $\ell < \ell_1$  geometrically shrinks as we go to levels with larger probabilities; this will later allow us to add these bounds across levels in a lossless manner.

▷ Claim 14. For any non-dominant level  $\ell < \ell_1$ , we have  $\sum_{i \in L_\ell} p_i q_i \leq \frac{80\mathcal{H}^2 \cdot p(\mathbb{T})}{2^{\ell_1 - \ell}}$ .

Proof Sketch. Since  $X_\ell \leq Z_\ell$ , we get from Lemma 12(i) that

$$\sum_{i \in L_\ell} p_i q_i \leq \frac{1}{2^\ell} \cdot \mathbb{E}[X_\ell] \leq \Pr[X_\ell = 1 \mid Z_\ell = 1] \cdot \Pr[Z_\ell \geq 1] + \sum_{t \geq 2} \Pr[Z_\ell \geq t], \quad (4.8)$$

where we split between  $X_\ell = 1$  and  $X_\ell \geq 2$ , and replaced  $X_\ell$  with  $Z_\ell$ . We bound  $\sum_{t \geq 2} \Pr[Z_\ell \geq t]$  and  $\Pr[Z_\ell \geq 1]$  using Lemma 12 (iii). The interesting part is to bound  $\Pr[X_\ell = 1 \mid Z_\ell = 1]$ .

Note that since  $Z_\ell = 1$ , there is only one request from  $\mathbb{T}$  before all pages in level  $\ell$  are requested. Therefore, a necessary event for  $X_\ell = 1 \wedge Z_\ell = 1$  is that the last page requested in  $\mathbb{H}$  must be from level  $\ell$ ; in particular, page  $r_1$  (in level  $\ell_1 > \ell$ ) is requested before some page in  $L_\ell$ . Let  $\mathcal{E}$  denote this event; we bound

$$\Pr[X_\ell = 1 \mid Z_\ell = 1] \leq \Pr[\mathcal{E} \mid Z_\ell = 1] = \sum_{t \geq 1} \Pr[\mathcal{E} \mid Z_\ell = 1, Y_\ell = t] \cdot \Pr[Y_\ell = t]. \quad (4.9)$$

Consider any fixed number of requests, say  $t$ . The conditional probability of any page in the head  $\mathbb{H}$  increases when we decrease the number of requests for pages in the tail  $\mathbb{T}$  among these  $t$  requests. Therefore,

$$\Pr[\mathcal{E} \mid Z_\ell = 1, Y_\ell = t] \leq \Pr[\mathcal{E} \mid Z_\ell = 0, Y_\ell = t] = \frac{1/2^{\ell_1}}{1 - p(\mathbb{T})} \cdot t \leq \frac{t}{2^{\ell_1 - 1}}, \text{ since } p(\mathbb{T}) < 1/2.$$

Here, we used that given  $Z_\ell = 0$ , the probability that any request is for page  $r_1$  is  $\frac{1/2^{\ell_1}}{1 - p(\mathbb{T})}$ . The claim can now be established by putting together the bounds we have obtained in (4.8).  $\triangleleft$

## 4.2 Splitting the Head

We have established bounds on the contribution of individual levels to  $\sum_i p_i q_i$ . We are now ready to put these together to bound the total contribution of the head  $\mathbb{H}$ . To do this, we subdivide the head  $\mathbb{H} = [k]$  into three intervals  $\mathbb{H}_1 := [1, r_1]$ ,  $\mathbb{H}_2 = [r_1 + 1, r_2]$ , and  $\mathbb{H}_3 := [r_2 + 1, k]$  as follows: we define  $r_1$  to be the largest index  $i$  such that  $p_i \geq 4e^2 \mathcal{H} \cdot p(\mathbb{T})$  (this is the same definition of  $r_1$  that we used to define the level  $\ell_1$ ), and  $r_2$  to be the largest index  $i$  such that  $p_i \geq p(\mathbb{T})/2^{\mathcal{H}}$ . We will separately bound the contribution of each of these intervals to  $\sum_{i=1}^k p_i q_i$ .

► **Lemma 15** (Bounding  $\mathbb{H}_1$ ). *The contribution of  $\mathbb{H}_1$  is bounded as  $\sum_{i \in \mathbb{H}_1} p_i q_i \leq 88 \mathcal{H}^2 \cdot p(\mathbb{T})$ .*

**Proof Sketch.** This follows by using Claim 13 for a dominant level, if it exists, and level  $\ell_1$ , the last level in  $\mathbb{H}_1$ , and using Claim 14 for the remaining levels.  $\triangleleft$

► **Lemma 16** (Bounding  $\mathbb{H}_2$ ). *The contribution of  $\mathbb{H}_2$  is bounded as  $\sum_{i \in \mathbb{H}_2} p_i q_i \leq 12 \mathcal{H}^2 \cdot p(\mathbb{T})$ .*

**Proof.** By construction, the number of levels in  $\mathbb{H}_2$  is at most  $\log(4e^2 \mathcal{H} \cdot 2^{\mathcal{H}}) \leq 6 \mathcal{H}$ . The claim now follows from Claim 13.  $\triangleleft$

► **Lemma 17** (Bounding  $\mathbb{H}_3$ ). *The contribution of  $\mathbb{H}_3$  is bounded as  $\sum_{i \in \mathbb{H}_3} p_i q_i \leq 4p(\mathbb{T})$ .*

**Proof Sketch.** To bound the contribution of  $\mathbb{H}_3$ , we note that  $q_i \leq 1$  for all pages  $i$  and  $\sum_{i \in L_\ell} p_i$  for any level  $\ell$  is at most  $n_\ell \cdot \frac{1}{2^\ell} \leq 2^{\mathcal{H}} \cdot \frac{1}{2^\ell}$  by Lemma 5. The lemma now follows by adding over all levels in  $\mathbb{H}_3$  and using  $p_i \leq p(\mathbb{T})/2^{\mathcal{H}}$  for any page  $i \in \mathbb{H}_3$ .  $\triangleleft$

We conclude with the proof of the main theorem.

**Proof of Theorem 9.** By adding up the bounds from Observation 11 and Lemmas 15–17, we get

$$\sum_{i \in [n]} p_i q_i \leq (100 \mathcal{H}^2 + 5) \cdot p(\mathbb{T}) = O(\mathcal{H}^2 p(\mathbb{T})).$$

By Lemma 8, this implies that LRU is  $O(\mathcal{H}^3)$ -competitive for stochastic caching.  $\triangleleft$

## 5 The LFU Algorithm

In this section, we analyze the Least Frequently Used (LFU) algorithm for stochastic caching. Recall that in LFU, we maintain a count for every page of the number of times that page was requested, and we keep in cache the  $k$  pages with the highest count. When a page outside of cache is requested, we evict the lowest count page in order to serve the request, then revert back to the  $k$  pages with highest count if necessary. Clearly, in the limit, the empirical count distribution will converge to the true distribution, and LFU will eventually behave identically to TOP- $k$ . However, the “burn-in” period until convergence could be arbitrarily long as a function of  $n$ ; e.g., if  $p_k - p_{k+1} = \epsilon$ , then to even distinguish these two pages with probability  $\delta$  requires  $\Omega(\log(1/\delta)/\epsilon^2)$  samples (e.g., [14]).

Despite this, we show that LFU achieves an additive loss of  $O(\mathcal{H} \cdot k)$  in addition to the  $O(p(\mathbb{T}))$  cost paid by TOP- $k$ . Since  $k$  is incurred as an initialization cost for any algorithm, we show:

► **Theorem 18.** *The LFU algorithm has a competitive ratio  $O(\mathcal{H})$  for the stochastic caching problem.*

The general proof idea is to show that every page in  $\mathbb{H}$  (i.e., top  $k$ ) might be requested while it is outside of cache many times, but that after  $O(\mathcal{H})$  such cache misses, its count will forever be higher (from that point on, with good probability) than the count of any page with less than half its probability. If this is the case, we can charge any cache miss due to a request to  $\mathbb{T}$  to  $T \cdot p(\mathbb{T})$ . Each of the  $k$  pages in  $\mathbb{H}$  will only miss  $O(\mathcal{H})$  times before the only pages displacing it from cache are ones with very similar probability, and hence distinguishing between these does not matter; the total contribution from these pages in  $\mathbb{H}$  is  $O(k\mathcal{H} + T \cdot p(\mathbb{T}))$ . Below, we formalize this argument.

**Proof.** Define a parameter  $\gamma = 16 \log(\min(k, n_{\max})) = O(\mathcal{H})$ . For every  $i \in \mathbb{H}$  and  $j \in [n]$  such that  $p_j \leq p_i/2$ , define  $\mathcal{E}_{ij}$  to be the bad event that if the first time  $t$  that the empirical count of  $i$  is at least  $\gamma$ , there is at least one time  $t' \geq t$  such that the empirical count of  $j$  exceeds the count of  $i$ . We show that  $\mathcal{E}_{ij}$  is extremely unlikely for well separated  $p_i$  and  $p_j$ .

▷ **Claim 19.** For all  $i \in \mathbb{H}$ ,  $j \in [n]$  such that  $p_j \leq p_i/2^h$  for some  $h \geq 1$ , we have  $\Pr[\mathcal{E}_{ij}] \leq 2^{-h\gamma/8}$ .

The proof uses standard facts and we leave it to Appendix D. The last claim only gets easier if instead of considering a single page  $j$  with  $p_j \leq p_i/2^h$ , we consider a group of pages whose cumulative probability is at most  $p_i/2^h$ , and ask what are the odds that any of these pages pass  $i$  in count. (Since, even as we grouped these pages, their total count would still not exceed that of  $i$  with the same probability as before.)

► **Corollary 20.** *For all  $i \in \mathbb{H}$  and any subset  $S \subseteq [n]$  such that  $\sum_{j \in S} p_j \leq p_i/2^h$  for some  $h \geq 1$ , we have  $\Pr[\bigcup_{j \in S} \mathcal{E}_{ij}] \leq 2^{-h\gamma/8}$ .*

For every  $i \in \mathbb{H}$ , let  $R_i$  be the number of requests to page  $i$  until the count of  $i$  never goes below that of any page  $j$  with  $p_j \leq p_i/2$ . Using Claim 19 and Corollary 20, we can show that the expected number of requests until we achieve this count separation can be upper bounded.

▷ **Claim 21.** For every  $i \in \mathbb{H} \setminus D$  for some set  $D$  such that  $\sum_{i \in D} p_i = 2p(\mathbb{T})$ , we have  $\mathbb{E}[R_i] \leq 2\gamma$ .

Note that if we had set  $\gamma = \Theta(\log n)$ , we could guarantee by a union bound over all pairs of pages  $i$  and  $j$  that every page  $i \in \mathbb{H}$  sees at most  $\gamma$  requests before no page  $j$  with  $p_j \leq p_i/2$  ever has count higher than  $i$  (with high probability). However, we can get away with smaller  $\gamma$  using a more careful analysis, which we again defer to Appendix D. We now show how to conclude the proof.

Let  $S_t$  be the subset of pages  $i$  in  $\mathbb{H}$  that are both outside of cache at time  $t$  and that have been requested at least  $R_i$  times by time  $t$ . In other words, these are the pages  $i \in \mathbb{H}$  outside of cache at time  $t$  such that no page  $j$  with  $p_j \leq p_i/2$  will have a count exceeding that of page  $i$  after time  $t$ . Define an arbitrary perfect matching  $M$  between the pages of  $S_t$  and a subset of the pages from  $\mathbb{T}$  in cache at time  $t$ . If page  $i \in S_t$  is matched to page  $M(i)$ , then  $p_{M(i)} \geq p_i/2$ , since the count of  $M(i)$  must be higher than that of  $i$ , and  $i \in S_t$ .

Let  $C^t$  be the  $k$  pages with highest counts at time  $t$ . Then we can write the total cost of the LFU algorithm as

$$\begin{aligned} \sum_t \sum_{i=1}^n p_i \Pr[i \notin C^t] &\leq \sum_t \sum_{i \notin S_t} p_i \Pr[i \notin C^t] + \sum_t \sum_{i \in S_t} p_i \Pr[i \notin C^t] + \sum_t p(\mathbb{T}) \\ &\leq 2T \cdot p(\mathbb{T}) + \sum_{i \in \mathbb{H}} \mathbb{E}[R_i] + \sum_t \sum_{i \in S_t} p_i \Pr[i \notin C^t] + T \cdot p(\mathbb{T}) \end{aligned} \quad (5.10)$$

$$\leq 2\gamma \cdot k + 2 \cdot \sum_t \sum_{i \in S_t} p_{M(i)} + 3T \cdot p(\mathbb{T}) \quad (5.11)$$

$$\leq 2\gamma \cdot k + 5T \cdot p(\mathbb{T}). \quad (5.12)$$

Step (5.10) follows because the expected number of misses for each page  $i \in \mathbb{H}$  until it belongs to  $S_t$  is  $\mathbb{E}[R_i]$ , step (5.11) follows by Claim 21 and since  $p_i \leq 2p_{M(i)}$ , and step (5.12) follows since the matched pages in cache form a subset of  $\mathbb{T}$ , and hence their total probability is less than  $p(\mathbb{T})$ .  $\blacktriangleleft$

## 6 Closing Remarks

In this work, we study the stochastic caching problem as a way to move beyond worst-case competitive ratio analysis for online algorithms. We give new evidence for the effectiveness of the popular LRU and LFU caching policies, by showing that for many distributions, they incur approximation factors significantly better than classical bounds predict. En route, we identify subset entropy as an important parameter of the distribution governing the achievable competitive ratio of stochastic caching; this information-theoretic quantity might be of independent interest.

There are several avenues for future work. The first obvious direction is to either improve the competitive ratio of LRU from  $O(\mathcal{H}^3)$  to  $O(\mathcal{H})$ , or show that this is not possible. In Appendix F, we show that there are distributions for which the cost of LRU is  $\Omega(\mathcal{H} \cdot p(\mathbb{T}))$ , meaning that our techniques can at best prove an  $O(\mathcal{H}^2)$  bound.

Another direction is to generalize the result to more complex stochastic models, e.g., the Markov paging model [23] (this would entail developing an analogous quantity to subset entropy). Going beyond caching, we often prove lower bounds for randomized online algorithms by exhibiting pathological input distributions. Parameterizing what makes these distributions difficult and, more generally, understanding and quantifying the performance of an online algorithm through a fine-grained measure such as subset entropy are promising directions of future research.

## References

- 1 Algorithms with predictions (ALPS). <https://algorithms-with-predictions.github.io/>.
- 2 Alfred V. Aho, Peter J. Denning, and Jeffrey D. Ullman. Principles of optimal page replacement. *J. ACM*, 18(1):80–93, 1971. doi:10.1145/321623.321632.
- 3 Susanne Albers, Sanjeev Arora, and Sanjeev Khanna. Page replacement for general caching problems. In *SODA*, pages 31–40, 1999. URL: <http://dl.acm.org/citation.cfm?id=314500.314528>.
- 4 Nikhil Bansal, Niv Buchbinder, and Joseph Naor. A primal-dual randomized algorithm for weighted paging. *J. ACM*, 59(4):19, 2012.
- 5 Nikhil Bansal, Niv Buchbinder, and Joseph Naor. Randomized competitive algorithms for generalized caching. *SICOMP*, 41(2):391–414, 2012. doi:10.1137/090779000.
- 6 Nikhil Bansal, Christian Coester, Ravi Kumar, Manish Purohit, and Erik Vee. Learning-augmented weighted paging. In *SODA*, pages 67–89, 2022. doi:10.1137/1.9781611977073.4.
- 7 Luca Becchetti. Modeling locality: A probabilistic analysis of LRU and FWF. In *ESA*, pages 98–109, 2004. doi:10.1007/978-3-540-30140-0\_11.
- 8 Laszlo A. Belady. A study of replacement algorithms for a virtual-storage computer. *IBM Systems Journal*, 5(2):78–101, 1966. doi:10.1147/SJ.52.0078.
- 9 Allan Borodin and Ran El-Yaniv. *Online Computation and Competitive Analysis*. Cambridge U. Press, 1998.
- 10 Allan Borodin, Sandy Irani, Prabhakar Raghavan, and Baruch Schieber. Competitive paging with locality of reference. *JCSS*, 50(2):244–258, 1995. doi:10.1006/JCSS.1995.1021.
- 11 Sébastien Bubeck, Michael B Cohen, Yin Tat Lee, James R Lee, and Aleksander Madry. K-server via multiscale entropic regularization. In *STOC*, pages 3–16, 2018. doi:10.1145/3188745.3188798.
- 12 Niv Buchbinder, Shahar Chen, and Joseph Naor. Competitive analysis via regularization. In *SODA*, pages 436–444, 2014. doi:10.1137/1.9781611973402.32.
- 13 Niv Buchbinder, Joseph Seffi Naor, et al. The design of competitive online algorithms via a primal–dual approach. *Foundations and Trends® in Theoretical Computer Science*, 3(2–3):93–263, 2009. doi:10.1561/04000000024.
- 14 Clément L Canonne. A short note on an inequality between KL and TV. *arXiv*, 2202.07198, 2022. arXiv:2202.07198.
- 15 Nikhil R. Devanur, Kamal Jain, and Robert D. Kleinberg. Randomized primal-dual analysis of RANKING for online bipartite matching. In *SODA*, pages 101–107, 2013. doi:10.1137/1.9781611973105.7.
- 16 Guy Even, Moti Medina, and Dror Rawitz. Online generalized caching with varying weights and costs. In *SPAA*, pages 205–212, 2018. doi:10.1145/3210377.3210404.
- 17 Amos Fiat, Richard M. Karp, Michael Luby, Lyle A. McGeoch, Daniel Dominic Sleator, and Neal E. Young. Competitive paging algorithms. *J. ALG*, 12(4):685–699, 1991. doi:10.1016/0196-6774(91)90041-V.
- 18 Peter A. Franaszek and Terry J. Wagner. Some distribution-free aspects of paging algorithm performance. *J. ACM*, 21(1):31–39, 1974. doi:10.1145/321796.321800.
- 19 Anupam Gupta, Ravishankar Krishnaswamy, Amit Kumar, and Debmalya Panigrahi. Elastic caching. In *SODA*, pages 143–156, 2019. doi:10.1137/1.9781611975482.10.
- 20 Anupam Gupta, Amit Kumar, and Debmalya Panigrahi. Caching with time windows and delays. *SICOMP*, 51(4):975–1017, 2022. doi:10.1137/20M1346286.
- 21 S. Irani. Randomized weighted caching with two page weights. *Algorithmica*, 32(4):624–640, 2002. doi:10.1007/S00453-001-0095-6.
- 22 Sandy Irani. Page replacement with multi-size pages and applications to web caching. In *STOC*, pages 701–710, 1997. doi:10.1145/258533.258666.
- 23 Anna R. Karlin, Steven J. Phillips, and Prabhakar Raghavan. Markov paging. *SICOMP*, 30(3):906–922, 2000. doi:10.1137/S0097539794268042.

- 24 Elias Koutsoupias and Christos H. Papadimitriou. Beyond competitive analysis. *SICOMP*, 30(1):300–317, 2000. doi:10.1137/S0097539796299540.
- 25 Ravi Kumar, Manish Purohit, Zoya Svitkina, and Erik Vee. Interleaved caching with access graphs. In *SODA*, pages 1846–1858, 2020. doi:10.1137/1.9781611975994.113.
- 26 Carsten Lund, Steven J. Phillips, and Nick Reingold. Paging against a distribution and IP networking. *JCSS*, 58(1):222–232, 1999. doi:10.1006/JCSS.1997.1498.
- 27 Thodoris Lykouris and Sergei Vassilvitskii. Competitive caching with machine learned advice. In *ICML*, pages 3302–3311, 2018. URL: <http://proceedings.mlr.press/v80/lykouris18a.html>.
- 28 Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms*. Cambridge U. Press, 1995. doi:10.1017/CB09780511814075.
- 29 Chirag Pabbaraju and Ali Vakilian. New and improved bounds for Markov paging. In *ICALP*, pages 123:1–123:20, 2025. doi:10.4230/LIPIcs.ICALP.2025.123.
- 30 Konstantinos Panagiotou and Alexander Souza. On adequate performance measures for paging. In *STOC*, pages 487–496, 2006. doi:10.1145/1132516.1132587.
- 31 Gopal Pandurangan and Eli Upfal. Entropy-based bounds for online algorithms. *TALG*, 3(1):7:1–7:19, 2007. doi:10.1145/1219944.1219953.
- 32 Tim Roughgarden, editor. *Beyond the Worst-Case Analysis of Algorithms*. Cambridge U. Press, 2020.
- 33 Daniel D. Sleator and Robert E. Tarjan. Amortized efficiency of list update and paging rules. *C. ACM*, 28(2):202–208, 1985. doi:10.1145/2786.2793.
- 34 J Michael Steele. *Stochastic Calculus and Financial Applications. Vol I*. Springer, 2001.
- 35 Andrew Chi-Chih Yao. Probabilistic computations: Toward a unified measure of complexity (extended abstract). In *FOCS*, pages 222–227, 1977. doi:10.1109/SFCS.1977.24.
- 36 Neal E. Young. On-line paging against adversarially biased random inputs. *J. ALG*, 37(1):218–235, 2000. doi:10.1006/JAGM.2000.1099.
- 37 Neal E. Young. Online paging and caching. In *Encyclopedia of Algorithms*, pages 1457–1461. Springer, 2016. doi:10.1007/978-1-4939-2864-4\_267.

## A Missing Proofs from Section 2

**Proof of Lemma 5.** Note that for every  $\ell$ , the distribution  $\mathcal{D}_{S_\ell}$  is uniform up to a factor of 2 (i.e., the maximum probability is at most twice the minimum probability). As noted, the Shannon entropy of such a distribution is  $\Theta(\log n_\ell)$ , where  $n_\ell$  is the size of its support. We choose  $S$  as follows: if  $n_{\max} \leq k$ , then  $S = \arg \max_\ell |S_\ell|$  and hence  $|S| = n_{\max}$ ; else,  $S$  is any  $k$ -sized subset of  $\arg \max_\ell |S_\ell|$  and hence  $|S| = k$ . Then, we get

$$\mathcal{H}(\mathcal{D}, k) \geq H(\mathcal{D}_S) = \Omega(\log |S|) = \Omega(\log(\min\{k, n_{\max}\})).$$

We now show the other direction, by proving that for any distribution  $\mathcal{D}'$  with  $\max_\ell |\{u \in U : 1/2^{\ell+1} < \mathcal{D}'(u) \leq 1/2^\ell\}| \leq N$ , we have that  $H(\mathcal{D}') \leq O(\log N)$ . Note that  $N \leq \min\{2n_{\max}, k\}$  for  $\mathcal{D}_S$ , where  $S \subseteq [n]$  is any subset of cardinality at most  $k$ , since pages in  $S$  whose probabilities in  $\mathcal{D}_S$  differ by at most a factor of two can come from at most two adjacent levels of the original distribution  $\mathcal{D}$ .

Letting  $S'_\ell := \{u : 1/2^{\ell+1} < \mathcal{D}'(u) \leq 1/2^\ell\}$ , we have,

$$H(\mathcal{D}') = - \sum_u \mathcal{D}'(u) \cdot \log \mathcal{D}'(u) = - \sum_\ell \sum_{u \in S'_\ell} \mathcal{D}'(u) \cdot \log \mathcal{D}'(u).$$

Now, for any  $u \in S'_\ell$ , we have  $\mathcal{D}'(u) \leq 1/2^\ell$  and  $-\log \mathcal{D}'(u) = \log 1/\mathcal{D}'(u) < \log 2^{\ell+1} = \ell + 1$ .

Therefore,  $H(\mathcal{D}') \leq \sum_\ell n_\ell \cdot \frac{\ell+1}{2^\ell}$ , where  $n_\ell := |S'_\ell|$ .

To bound the RHS of this inequality, we use a factor-revealing LP. We note that  $H(\mathcal{D}')$  is at most the optimal primal objective in the following LP:

$$\text{maximize } \sum_{\ell \geq 1} n_\ell \cdot \frac{\ell+1}{2^\ell} \text{ subject to } \sum_{\ell} \frac{n_\ell}{2^\ell} \leq 1 \text{ and } \{n_\ell \leq N\}_{\forall \ell}.$$

The dual LP is

$$\text{minimize } \alpha + N \sum_{\ell} \beta_\ell, \text{ subject to } \left\{ \frac{\alpha}{2^\ell} + \beta_\ell \geq \frac{\ell+1}{2^\ell} \right\}_{\forall \ell}.$$

Now, a feasible dual solution is given by:

$$\alpha = \log N \quad \beta_\ell = \begin{cases} \frac{\ell+1}{2^\ell} & \text{if } \ell > \log N - 1 \\ 0 & \text{if } \ell \leq \log N - 1. \end{cases}$$

The dual objective for this solution is  $O(\log N) = O(\log(\min\{k, n_{\max}\}))$ . Therefore,  $\mathcal{H}(\mathcal{D}, k) = H(\mathcal{D}_S) \leq O(\log(\min\{k, n_{\max}\})) = O(\log(\min\{k, \max_{\ell} |S_\ell|\}))$ .  $\blacktriangleleft$

## B Missing Proofs from Section 3

We first define our true dual vector  $y$ . For any time  $t$ , let  $G(t) \subseteq [n]$  be the set of pages  $i \notin [k]$ , for which the interval  $I(i, \rho(i, t))$  has  $|I(i, \rho(i, t))| \leq 3/p_i$ . Let  $\sigma_1^t, \dots, \sigma_{m^t}^t$  be the pages of  $G(t)$  in decreasing order of  $p_i$ . Then define  $S_j^t = [k] \cup \{\sigma_1^t, \dots, \sigma_j^t\}$ , and

$$y_S^t = \begin{cases} p_{\sigma_j^t} - p_{\sigma_{j+1}^t} & \text{if } S = S_j^t \text{ and } j < m^t. \\ p_{\sigma_{m^t}^t} & \text{if } S = S_j^t \text{ and } j = m^t. \\ 0 & \text{otherwise.} \end{cases}$$

In words, we define dual variables only for prefixes of the pages in decreasing order of probability (after removing pages whose intervals exceed their expected length by a constant), and include the  $k$  pages in  $\mathbb{H}$  in all these prefixes.

We argue that this dual is approximately feasible with constant probability, and that its objective value is sufficiently large. We start with the first claim. Let  $n_{\max} := \max_{\ell} |S_\ell|$  in Lemma 5 and let  $\eta := \min\{n_{\max}, k\}$ . Recall that by Lemma 5, we have that  $\mathcal{H} = \Theta(\log \eta)$ .

► **Lemma 22.** *Fix a time step  $t$ . With probability at least  $3/4$ , for all  $i \in [n]$ , the solution  $y$  satisfies the dual constraints associated with  $(i, \rho(i, t))$  up to a factor of  $3 \ln(2\eta)$ .*

**Proof.** Since  $t$  is fixed, for convenience we define  $I(i) := I(i, \rho(i, t))$ .

For pages  $i \in [n] \setminus ([k] \cup G(t))$ , for all  $t' \in I(i)$  and  $S$  such that  $i \in S$ , note that  $y_S^{t'} = 0$ . So, the corresponding dual constraint is satisfied vacuously.

Next, consider constraints associated with pages  $i \in G(t)$ . For all  $t' \in I(i)$ , if  $j^*(t')$  is the index of  $i$  in  $\sigma_1^{t'}, \dots, \sigma_{m^{t'}}^{t'}$ , then by design

$$\sum_{S: i \in S} y_S^{t'} = \sum_{j=j^*(t')}^{m^{t'}-1} (p_{\sigma_j^{t'}} - p_{\sigma_{j+1}^{t'}}) + p_{\sigma_{m^{t'}}^{t'}} = p_{\sigma_{j^*(t')}^{t'}} = p_i,$$

which means

$$\sum_{t' \in I(i)} \sum_{S: i \in S} y_S^{t'} = p_i \cdot |I(i)| \leq p_i \cdot \frac{3}{p_i} = 3.$$

It remains to handle constraints associated with  $i \in [k]$ . This time, for all  $t' \in I(i)$ , we have

$$\sum_{S: i \in S} y_S^{t'} = \sum_{j=1}^{m^{t'}-1} (p_{\sigma_j^{t'}} - p_{\sigma_{j+1}^{t'}}) + p_{\sigma_{m^{t'}}^{t'}} = p_{\sigma_1^{t'}} \leq p_k,$$

and hence

$$\Pr \left[ \sum_{t' \in I(i)} \sum_{S: i \in S} y_S^{t'} \geq 3 \ln(2\eta) \right] \leq \Pr [|I(i)| \cdot p_k \geq 3 \ln(2\eta)] \leq \Pr \left[ |I(i)| \geq E[|I(i)|] \cdot 3 \ln(2\eta) \cdot \frac{p_i}{p_k} \right].$$

Since  $|I(i, j)|$  is a geometric random variable with parameter  $p_i$ , we get that

$$\Pr \left[ \sum_{t' \in I(i)} \sum_{S: i \in S} y_S^{t'} \geq 3 \ln(2\eta) \right] \leq (1 - p_i)^{E[|I(i)|] \cdot 3 \ln(2\eta) \cdot \frac{p_i}{p_k}} \leq e^{-3 \ln(2\eta) \cdot \frac{p_i}{p_k}} \leq (2\eta)^{-3 \cdot \frac{p_i}{p_k}}.$$

Let  $L_\ell$  be the pages  $i$  in  $[k]$  such that  $p_k \cdot 2^{\ell+1} > p_i \geq p_k \cdot 2^\ell$ . Note that  $|L_\ell| \leq 2n_{\max}$  since it contains pages from at most two sets  $S_{\ell'}, S_{\ell'+1}$  for some  $\ell'$ . Moreover,  $|L_\ell| \leq k$ . Hence,  $|L_\ell| \leq 2\eta$ .

Now, let  $\mathcal{E}$  be the event that there exists an  $i \in [k]$  for which  $\sum_{t' \in I(i)} \sum_{S: i \in S} y_S^{t'} \geq 3 \ln(2\eta)$ , i.e., the dual constraint for page  $i$  is violated beyond a factor of  $3 \ln(2\eta)$ . By a union bound,

$$\Pr[\mathcal{E}] \leq \sum_{\ell \geq 0} \sum_{i \in [k]: i \in L_\ell} (2\eta)^{-3 \cdot \frac{p_i}{p_k}} \leq \sum_{\ell \geq 0} \sum_{i \in [k]: i \in L_\ell} (2\eta)^{-3 \cdot 2^\ell} \leq \sum_{\ell \geq 0} 2\eta \cdot (2\eta)^{-3 \cdot 2^\ell} \leq (2\eta)^{-2} \leq \frac{1}{4}. \quad \blacktriangleleft$$

► **Lemma 23.** *The expected objective of the dual  $y$  at time  $t$  is at least  $2p(\mathbb{T})/3$ .*

**Proof.** By construction, the dual objective at time  $t$  is

$$\sum_S (|S| - k) \cdot y_S^t = m^t \cdot p_{\sigma_{m^t}} + \sum_{j=1}^{m^t-1} j \cdot (p_{\sigma_j^t} - p_{\sigma_{j+1}^t}) = \sum_{i \in G(t)} p_i.$$

By a Markov bound, every page  $i \in [n] \setminus [k]$  appears in  $G(t)$  with probability at least  $2/3$ , hence the claim.  $\blacktriangleleft$

Using these, we can conclude the proof of Lemma 8, which in turn implies Theorem 6 by weak duality.

**Proof of Lemma 8.** From Lemma 22 and Lemma 23, there exists a dual solution  $y$  such that:

- (a) At every time  $t$ , with probability  $\geq 3/4$ , solution  $y$  simultaneously satisfies all constraints  $(i, j)$  such that  $t \in I(i, j)$  up to a multiplicative factor of  $3 \ln(2\eta)$ .
- (b) The expected dual objective value of this solution  $y$  is  $2p(\mathbb{T})/3$ .

Define a modified dual solution  $\hat{y}$  as follows. For every time  $t$ , if there exists a dual constraint  $(i, j)$  such that  $t \in I(i, j)$  that is violated by a multiplicative factor greater than  $3 \ln(2\eta)$  (call this event  $\mathcal{E}(t)$ ), then set  $\hat{y}_S^t = 0$  for all  $S \subseteq [n]$ . Otherwise, set  $\hat{y}_S^t = y_S^t / (3 \ln(2\eta))$ .

The new dual is always feasible. Furthermore, when  $y_S^t > 0$ , we have  $\hat{y}_S^t = y_S^t / (3 \ln(2\eta))$  with probability at least  $3/4$ . Hence, by the linearity of expectation, the dual objective of  $\hat{y}$  is

$$\begin{aligned} \sum_S (|S| - k) \cdot \mathbb{E}[\hat{y}_S^t] &\geq \frac{1}{3 \ln(2\eta)} \sum_S (|S| - k) \cdot (\mathbb{E}[y_S^t] - \mathbb{E}[y_S^t \mid \mathcal{E}(t)] \cdot \Pr[\mathcal{E}(t)]) \\ &\geq \frac{2p(\mathbb{T})/3 - p(\mathbb{T})/4}{3 \ln(2\eta)} \\ &\geq \frac{5p(\mathbb{T})/12}{3 \ln(2 \cdot \min\{k, n_{\max}\})}, \end{aligned} \tag{B.13}$$

where in (B.13) we used  $\sum_S (|S| - k) \cdot y_S^t \leq p(\mathbb{T})$  with probability 1, together with Lemmas 22 and 23.

The characterization of  $\mathcal{H} = \Theta(\log \min\{k, n_{\max}\})$  from Lemma 5 concludes the proof.  $\blacktriangleleft$

## C Missing Proofs from Section 4

### C.1 Proof of Lemma 12

**Proof of Lemma 12.** We prove the three parts separately.

- (i) It follows from the fact that  $0 \leq X_\ell \leq n_\ell$ , and the fact that  $\mathbb{E}[X_\ell] = \sum_{i \in L_\ell} q_i$ .
- (ii) This fact follows from standard coupon collector bounds, since each request is for a page in  $L_\ell$  with probability  $\geq 1/2^{\ell+1}$  and there are  $n_\ell$  distinct pages in  $L_\ell$ .
- (iii) Since  $\ell$  is a non-dominant level, even after conditioning on the page requests being from level  $\ell$ , the probability that any request is for a page in  $\mathbb{T}$  is at most  $2p(\mathbb{T})$ . We use this fact in the rest of this proof.

For  $t = 1$ , we have

$$\begin{aligned} \Pr[Z_\ell \geq 1] &\leq \mathbb{E}[Z_\ell] = \sum_{L \geq 1} \mathbb{E}[Z_\ell \mid Y_\ell = L] \cdot \Pr[Y_\ell = L] \leq 2p(\mathbb{T}) \sum_{L \geq 1} L \cdot \Pr[Y_\ell = L] \\ &= 2p(\mathbb{T}) \cdot \mathbb{E}[L] \leq p(\mathbb{T}) \cdot 2^{\ell+1} \cdot \mathcal{H}. \end{aligned}$$

For  $t \geq 2$ , let  $\mu := \mathbb{E}[Y_\ell] \leq 2^{\ell+1} \cdot \mathcal{H}$ . Now, the value of  $Y_\ell$  can be partitioned into intervals  $(e(\gamma-1)\mu, e\gamma\mu]$  for integers  $\gamma \geq 1$ . Using this partitioning, we get

$$\begin{aligned} \Pr[Z_\ell \geq t] &= \sum_{\gamma \geq 1} \Pr[Z_\ell \geq t \mid e(\gamma-1)\mu < Y_\ell \leq e\gamma\mu] \cdot \Pr[e(\gamma-1)\mu < Y_\ell \leq e\gamma\mu] \\ &= \sum_{\gamma \geq 1} \Pr[Z_\ell \geq t \mid Y_\ell \leq e\gamma\mu] \cdot \Pr[Y_\ell > e(\gamma-1)\mu]. \end{aligned}$$

Using the Binomial distribution, we get  $\Pr[Z_\ell \geq t \mid Y_\ell \leq e\gamma\mu] \leq \binom{e\gamma\mu}{t} (2p(\mathbb{T}))^t \leq \left(\frac{e^2\gamma\mu}{t}\right)^t (2p(\mathbb{T}))^t$ , where the last step used Stirling's approximation.

We now bound the second term,  $\Pr[Y_\ell > e(\gamma-1)\mu]$ . Observe that for  $Y_\ell > e(\gamma-1)\mu$  to hold, each of the first  $\gamma-1$  blocks of  $e\mu$  requests must fail to request all the pages from level  $\ell$ . By Markov's inequality, since  $\mu = \mathbb{E}[Y_\ell]$ , each block of  $e\mu$  requests fails to request all pages from level  $\ell$  with probability at most  $1/e$ , and since these failure events are independent,  $\Pr[Y_\ell > e(\gamma-1)\mu] \leq e^{-(\gamma-1)}$ . Therefore, for  $\gamma \geq 1$ , we have  $\Pr[Y_\ell > e(\gamma-1)\mu] \leq e^{-(\gamma-1)}$ . Putting these together, we get

$$\begin{aligned} \Pr[Z_\ell \geq t] &\leq \sum_{\gamma \geq 1} \left(\frac{e^2\gamma\mu}{t}\right)^t (2p(\mathbb{T}))^t \cdot e^{-(\gamma-1)} \\ &\leq \left(\frac{e^2\mu}{t}\right)^t \cdot (2p(\mathbb{T}))^t \cdot \sum_{\gamma=1}^{\infty} \frac{\gamma^t}{e^{\gamma-1}} \leq \left(\frac{e^2\mu}{t}\right)^t \cdot (2p(\mathbb{T}))^t \cdot \left(\sum_{\gamma=1}^{2t \ln(t)-1} \frac{\gamma^t}{e^{\gamma-1}} + \sum_{\gamma=2t \ln t}^{\infty} \frac{\gamma^t}{e^{\gamma-1}}\right) \\ &\leq \left(\frac{e^2\mu}{t}\right)^t \cdot (2p(\mathbb{T}))^t \cdot \left((2t \ln(t)-1) \cdot \frac{t^t}{e^{t-1}} + 1\right) \tag{C.14} \\ &\leq (2e^2\mu)^t \cdot p(\mathbb{T})^t. \tag{since  $t \geq 2$ } \end{aligned}$$

Step (C.14) follows since the expression  $\gamma^t/e^{\gamma-1}$  is maximized at  $\gamma = t$ , so the first partial sum can be bounded by its length times the maximum  $t^t/e^{t-1}$ . The suffix sum converges to less than one.  $\blacktriangleleft$

### C.2 Proof of Claim 14

**Proof of Claim 14.** Recall that by (4.3), we have that for any level  $\ell$ ,  $\sum_{i \in L_\ell} p_i q_i \leq \frac{1}{2^\ell} \cdot \mathbb{E}[X_\ell]$ , where  $X_\ell$  denotes the number of pages in  $L_\ell$  that do not appear among the first  $k$  distinct pages. We proceed to bound  $\mathbb{E}[X_\ell]$ . By definition,  $X_\ell \leq Z_\ell$ . Therefore, we can write

$$\begin{aligned}
\mathbb{E}[X_\ell] &= \sum_{t \geq 1} t \cdot \Pr[X_\ell = t] = \Pr[X_\ell = 1] + \sum_{t \geq 2} t \cdot \Pr[X_\ell = t] = \Pr[X_\ell = 1] + \sum_{t \geq 2} \Pr[X_\ell \geq t] \\
&\leq \Pr[X_\ell = 1 \mid Z_\ell = 1] \cdot \Pr[Z_\ell = 1] + \sum_{t \geq 2} \Pr[Z_\ell \geq t] \\
&\leq \Pr[X_\ell = 1 \mid Z_\ell = 1] \cdot \Pr[Z_\ell \geq 1] + \sum_{t \geq 2} \Pr[Z_\ell \geq t]. \tag{C.15}
\end{aligned}$$

We can bound factors of (C.15) of the form  $\Pr[Z_\ell \geq t]$  using (4.6). We get

$$\sum_{t \geq 2} \Pr[Z_\ell \geq t] \leq \sum_{t \geq 2} \left( 2e^2 \cdot p(\mathbb{T}) \cdot 2^{\ell+1} \cdot \mathcal{H} \right)^t \leq 2 \left( 2e^2 \cdot p(\mathbb{T}) \cdot 2^{\ell+1} \cdot \mathcal{H} \right)^2,$$

where in the last step, we used

$$2e^2 \cdot p(\mathbb{T}) \cdot 2^{\ell+1} \cdot \mathcal{H} \leq 2e^2 \cdot p(\mathbb{T}) \cdot 2^{\ell_1} \cdot \mathcal{H} \leq \frac{2e^2 \cdot p(\mathbb{T}) \cdot \mathcal{H}}{4e^2 \mathcal{H} \cdot p(\mathbb{T})} = 1/2.$$

Thus,  $\sum_{t \geq 2} \Pr[Z_\ell \geq t] \leq 2 \left( 2e^2 \cdot p(\mathbb{T}) \cdot 2^{\ell+1} \cdot \mathcal{H} \right)^2 \leq \frac{\mathcal{H} \cdot p(\mathbb{T})}{2^{\ell_1 - 2\ell - 6}}$ , since  $p(\mathbb{T}) \leq 1/(4e^2 \mathcal{H} \cdot 2^{\ell_1})$ .

It remains to bound the first part of (C.15). From (4.5), we have that  $\Pr[Z_\ell \geq 1] \leq 2p(\mathbb{T}) \cdot 2^{\ell+1} \cdot \mathcal{H}$ . We are left to bound  $\Pr[X_\ell = 1 \mid Z_\ell = 1]$ . Since there is only one request from  $\mathbb{T}$  before all pages in level  $\ell$  are requested, a necessary event for  $X_\ell$  to be 1 is that the last page requested in  $\mathbb{H}$  must be from level  $\ell$ . In particular, consider page  $r_1$  which is in level  $\ell_1$  and therefore not in level  $\ell < \ell_1$ . Let  $\mathcal{E}$  denote the event that page  $r_1$  is requested before all pages in level  $\ell$  are requested. Then,

$$\Pr[X_\ell = 1 \mid Z_\ell = 1] \leq \Pr[\mathcal{E} \mid Z_\ell = 1] = \sum_{t \geq 1} \Pr[\mathcal{E} \mid Z_\ell = 1, Y_\ell = t] \cdot \Pr[Y_\ell = t]. \tag{C.16}$$

Consider any fixed number of requests, say  $t$ . The conditional probability of any page in the head  $\mathbb{H}$  increases when we decrease the number of requests for pages in the tail  $\mathbb{T}$  among these  $t$  requests. Therefore,  $\Pr[\mathcal{E} \mid Z_\ell = 1, Y_\ell = t] \leq \Pr[\mathcal{E} \mid Z_\ell = 0, Y_\ell = t]$ .

Now, let  $I_t(\mathcal{E})$  denote the number of requests for page  $r_1$  among the first  $t$  requests. Given that  $Z_\ell = 0$ , the probability that any request is for page  $r_1$  is  $2^{-\ell_1}/(1 - p(\mathbb{T}))$ . Therefore,

$$\Pr[\mathcal{E} \mid Z_\ell = 1, Y_\ell = t] \leq \mathbb{E}[I_t(\mathcal{E}) \mid Z_\ell = 0] = \frac{1/2^{\ell_1}}{1 - p(\mathbb{T})} \cdot t \leq \frac{t}{2^{\ell_1 - 1}},$$

since  $p(\mathbb{T}) < 1/2$ . Substituting in (C.16), we get

$$\begin{aligned}
\Pr[X_\ell = 1 \mid Z_\ell = 1] &= \sum_{t \geq 1} \Pr[\mathcal{E} \mid Z_\ell = 1, Y_\ell = t] \cdot \Pr[Y_\ell = t] \leq \sum_{t \geq 1} \frac{t}{2^{\ell_1 - 1}} \cdot \Pr[Y_\ell = t] \\
&\leq \frac{\mathbb{E}[Y_\ell]}{2^{\ell_1 - 1}} \leq \frac{\mathcal{H} \cdot 2^{\ell+1}}{2^{\ell_1 - 1}} = \frac{4\mathcal{H}}{2^{\ell_1 - \ell}},
\end{aligned}$$

where we again used (4.4) in the last inequality. Therefore,

$$\Pr[X_\ell = 1 \mid Z_\ell = 1] \cdot \Pr[Z_\ell \geq 1] \leq \frac{4\mathcal{H}}{2^{\ell_1 - \ell}} \cdot 2p(\mathbb{T}) \cdot 2^{\ell+1} \cdot \mathcal{H} = \frac{\mathcal{H}^2 \cdot p(\mathbb{T})}{2^{\ell_1 - 2\ell - 4}}.$$

This bounds the first expression in (C.15).

Substituting our bounds for both expressions in (C.15), we get

$$\mathbb{E}[X_\ell] \leq \frac{\mathcal{H}^2 \cdot p(\mathbb{T})}{2^{\ell_1 - 2\ell - 4}} + \frac{\mathcal{H} \cdot p(\mathbb{T})}{2^{\ell_1 - 2\ell - 6}} \leq \frac{80\mathcal{H}^2 \cdot p(\mathbb{T})}{2^{\ell_1 - 2\ell}}.$$

Therefore,

$$\sum_{i \in L_\ell} p_i q_i \leq \frac{1}{2^\ell} \cdot \mathbb{E}[X_\ell] \leq \frac{1}{2^\ell} \cdot \frac{80\mathcal{H}^2 \cdot p(\mathbb{T})}{2^{\ell_1 - 2\ell}} = \frac{80\mathcal{H}^2 \cdot p(\mathbb{T})}{2^{\ell_1 - \ell}}. \quad \triangleleft$$

### C.3 Proof of Lemma 15

**Proof of Lemma 15.** We use Claim 13 to bound both a dominant level, if it exists, and the last level in  $\mathbb{H}_1$ , i.e.,  $\ell_1$ . For both these levels  $\ell = \ell^*$  and  $\ell = \ell_1$ , we get  $\sum_{i \in L_\ell} p_i q_i \leq 4\mathcal{H} \cdot p(\mathbb{T})$ , by Claim 13.

For non-dominant levels  $\ell < \ell_1$ , we use Claim 14. Thus,  $\sum_{\ell < \ell_1} \sum_{i \in L_\ell} p_i q_i \leq \sum_{\ell_1 - \ell \geq 1} \frac{80\mathcal{H}^2 \cdot p(\mathbb{T})}{2^{\ell_1 - \ell}} \leq 80\mathcal{H}^2 \cdot p(\mathbb{T})$ . Summing the bounds, we get  $\sum_{i \in \mathbb{H}_1} p_i q_i \leq 4\mathcal{H} \cdot p(\mathbb{T}) + 80\mathcal{H}^2 \cdot p(\mathbb{T}) \leq 88\mathcal{H}^2 \cdot p(\mathbb{T})$ .  $\blacktriangleleft$

### C.4 Proof of Lemma 17

**Proof of Lemma 17.** Let  $\ell_3$  be the first level  $\ell$  that contains pages from  $\mathbb{H}_3$ , i.e., all pages in  $\mathbb{H}_3$  are in levels  $\ell_3$  and above. Then,  $\sum_{i \in \mathbb{H}_3} p_i q_i \leq \sum_{\ell \geq \ell_3} \frac{1}{2^\ell} \sum_{i \in L_\ell} q_i \leq \sum_{\ell \geq \ell_3} \frac{1}{2^\ell} \cdot n_\ell$ , since  $q_i \leq 1$  by virtue of being a probability.

By Lemma 5, for any level  $\ell$ , the number pages  $n_\ell$  is at most  $2^{\mathcal{H}}$ . Therefore,

$$\sum_{i \in \mathbb{H}_3} p_i q_i \leq \sum_{\ell \geq \ell_3} \frac{1}{2^\ell} \cdot n_\ell \leq 2^{\mathcal{H}} \cdot \sum_{\ell \geq \ell_3} \frac{1}{2^\ell} \leq 2^{\mathcal{H}} \cdot \frac{1}{2^{\ell_3 - 1}}.$$

Note that  $p_i \leq p(\mathbb{T})/2^{\mathcal{H}}$  for any page  $i \in \mathbb{H}_3$ . If page  $i$  is in level  $\ell$ , then  $p_i \geq 1/2^{\ell+1}$ . Combining the two, we get  $\frac{1}{2^{\ell+1}} \leq p_i \leq \frac{p(\mathbb{T})}{2^{\mathcal{H}}}$ , and, in particular,  $\frac{1}{2^{\ell_3-1}} \leq \frac{4p(\mathbb{T})}{2^{\mathcal{H}}}$ . Putting these together, we get  $\sum_{i \in \mathbb{H}_3} p_i q_i \leq 2^{\mathcal{H}} \cdot \frac{1}{2^{\ell_3-1}} \leq 4p(\mathbb{T})$ .  $\blacktriangleleft$

## D Missing Proofs from Section 5

### D.1 Proof of Claim 19

**Proof of Claim 19.** Fix  $i \in \mathbb{H}$  and  $j$  such that  $p_j \leq p_i/2$ . Consider the following random walk on the line that starts at 0, moves +1 with probability  $p_i/(p_i + p_j) \geq 1/(2^h + 1)$ , and moves -1 with probability  $p_j/(p_i + p_j) \leq 1/(2^h + 1)$ . After  $\gamma$  time steps, the position is at least  $\gamma/2$  with probability at least  $1 - 2^{-h\gamma/4}$ . Indeed, observe that the probability the walk is at position  $< \gamma/2$  after  $\gamma$  time steps is the probability that  $\gamma$  i.i.d. Bernoullis  $X_1, \dots, X_\gamma$  with probability  $q = 1/(2^h + 1)$  sum to  $\gamma/4$ . The sum  $X = \sum_i X_i$  has mean  $\mu = q\gamma$  and if we set  $\delta = \gamma/(4\mu) - 1$ , by a Chernoff bound,

$$\Pr \left[ \sum_i X_i \geq \frac{\gamma}{4} \right] = \Pr \left[ \sum_i X_i \geq (1 + \delta)\mu \right] \leq \frac{e^{(1+\delta)\mu}}{(1 + \delta)^{(1+\delta)\mu}} \leq e^{\gamma/4} (4q)^{-\gamma/4} \leq 2^{-h\gamma/4}.$$

A random walk with bias  $1 - 1/(2^h + 1)$  at position  $z$  returns to 0 with probability exactly  $2^{-hz}$  (see, e.g., [34, page 6]).

Now, it is easy to see that the event  $\mathcal{E}_{ij}$  has the same probability as that of this random walk returning to 0 after  $\gamma$  time steps. Therefore, if  $\mathcal{E}_{ij}$  holds either (1) the random walk was at position  $\leq \gamma/4$  after  $\gamma$  steps, or (2) starting at position  $\gamma/4$ , the walk returned to 0 after a finite amount of time. By a union bound,  $\Pr[\mathcal{E}_{ij}] \leq 2^{-h\gamma/8}$ .  $\blacktriangleleft$

### D.2 Proof of Claim 21

**Proof of Claim 21.** Recall that  $\gamma = 16\mathcal{H} = 16 \log \min(k, n_{\max})$ . We do a case analysis.

**Case 1.** Suppose that  $k \leq n_{\max}$ , so  $\gamma = 16 \log k$ . Recalling that the pages of  $\mathbb{T}$  are ordered by decreasing probability, define  $G_\ell := \{i \in \mathbb{T} \mid i \geq 2k+1, i \equiv \ell \pmod k\}$ . Define  $q := 2 \cdot \sum_{i \in G'_0} p_i$ , and let  $I = \{i \in \mathbb{H} \mid p_i \geq q\}$ .

Note that the total probability mass of  $\mathbb{H} \setminus I$  is at most  $k \cdot q = O(p(\mathbb{T}))$ . Also, for all  $i \in I$  and all  $\ell$ , we have  $p_i \geq 2 \sum_{j \in G_\ell} p_j$ . By Corollary 20, after  $\gamma$  requests to  $i$ , the probability any page in  $G_\ell$  ever has count exceeding that of  $i$  is  $2^{-\gamma} = 1/\text{poly}(k)$ . By the same argument, the probability any page in  $\{k+1, \dots, 2k\}$  exceeds the count of  $i$  after  $\gamma$  requests to  $i$  is also  $1/\text{poly}(k)$ . By a union bound, no page in the tail exceeds the count of  $i$  after  $\gamma$  requests to  $i$  with probability  $1 - \text{poly}(k)$ , and hence the expected number of requests to  $i$  before this condition holds is at most  $2\gamma$ .

**Case 2.** Suppose now that  $n_{\max} \leq k$ , so  $\gamma = 16 \log n_{\max}$ . Then let  $L_\ell = \{j \in \mathbb{T} \mid p_j \in [p_k 2^\ell, p_k 2^{\ell+1})\}$ . By definition  $|L_\ell| \leq n_{\max}$ . For every  $i \in \mathbb{H}$  and every  $\ell \geq 1$ , by Corollary 20 after  $\gamma$  requests to  $i$ , a page  $j$  in  $L_\ell$  will have count exceeding  $i$  with probability no more than  $2^{-h\gamma/8} = (n_{\max})^{-2\ell}$ . By a union bound, no page in  $L_\ell$  has count exceeding  $i$  with probability  $(n_{\max})^{-2\ell} \leq 2^{-\ell-1}$ , and hence by another union bound, no page in  $\mathbb{T}$  with probability more than  $p_i/2$  will have count exceeding  $i$  with probability at least  $1/2$ . Once again, this means the expected number of requests to  $i$  before this condition holds is  $2\gamma$ .  $\triangleleft$

## E Computing Subset Entropy

We recall the definition of  $k$ -subset entropy (2.1). We show the following lemma, which immediately implies an efficient algorithm for computing  $\mathcal{H}(\mathcal{D}, k)$  (we call this the interval property of  $\mathcal{H}(\mathcal{D}, k)$ ):

► **Lemma 24** (Interval Property of  $\mathcal{H}(\mathcal{D}, k)$ ). *Consider a discrete distribution  $\mathcal{D} : [n] \rightarrow [0, 1]$  such that  $\mathcal{D}(i) \geq \mathcal{D}(j)$  for any  $i \geq j$ . Let  $k \leq n$  be any positive integer. Then, there is always a subset  $S \subseteq [n]$ ,  $|S| \leq k$  that maximizes  $H(\mathcal{D}_S)$  among subsets of  $[n]$  of cardinality at most  $k$  which is of the form  $S = \{r, r+1, \dots, t\}$ .*

Next, we prove this lemma. An important parameter in our proof will be the weighted geometric mean defined as  $\tau = \prod_{i \in S} p_i^{q_i}$ .

► **Claim 25.** The weighted geometric mean  $\tau = \prod_{i \in S} p_i^{q_i}$  satisfies the following condition:

$$\frac{\partial H(S)}{\partial p_j} \begin{cases} < 0 & \text{if } p_j > \tau \\ = 0 & \text{if } p_j = \tau \\ > 0 & \text{if } p_j < \tau. \end{cases}$$

*Proof.* Let us use the notation  $p_i := \mathcal{D}(i)$ , i.e.,  $p_1 \geq p_2 \geq \dots \geq p_n$ . We have:

$$H(S) = - \frac{\sum_{i \in S} p_i \log \frac{p_i}{\sum_{i \in S} p_i}}{\sum_{i \in S} p_i} = - \frac{\sum_{i \in S} p_i \log p_i}{\sum_{i \in S} p_i} + \log \left( \sum_{i \in S} p_i \right).$$

Note that the entropy of the conditional distribution on  $S$  given by  $H(\mathcal{D}_S)$  is a function of the values  $p_i, i \in S$ . We consider the partial derivative of this function with respect to any specific  $p_j, j \in S$ :

$$\begin{aligned} \frac{\partial H(S)}{\partial p_j} &= - \frac{(\sum_{i \in S} p_i) (1 + \log p_j) - \sum_{i \in S} p_i \log p_i}{(\sum_{i \in S} p_i)^2} + \frac{1}{\sum_{i \in S} p_i} \\ &= - \frac{\log p_j}{\sum_{i \in S} p_i} + \frac{\sum_{i \in S} p_i \log p_i}{(\sum_{i \in S} p_i)^2} = \frac{\frac{\sum_{i \in S} p_i \cdot \log p_i}{\sum_{i \in S} p_i} - \log p_j}{\sum_{i \in S} p_i} = \frac{\sum_{i \in S} q_i \log p_i - \log p_j}{\sum_{i \in S} p_i}, \end{aligned}$$

where  $q_i := \frac{p_i}{\sum_{i \in S} p_i}$ . Recall that  $\tau = \prod_{i \in S} p_i^{q_i}$ , i.e.,  $\log \tau = \sum_{i \in S} q_i \log p_i$ . The claim follows.  $\triangleleft$

Next, we state some simple properties of the weighted geometric mean  $\tau$ .

▷ **Claim 26.** The weighted geometric mean  $\tau$  over a set  $S$  does not change if  $\tau$  is added to  $S$ .

*Proof.* Recall that  $\tau = \prod_{i \in S} p_i^{q_i}$  where  $q_i = p_i/p$ , where  $p = \sum_{i \in S} p_i$ . Let  $q'_i$  be the weighted geometric mean if we include  $\tau$ , i.e.,

$$q'_i = \begin{cases} p_i/p+\tau & \text{if } i \in S \\ \tau/p+\tau & \text{if } p_i = \tau. \end{cases}$$

The new weighted geometric mean  $\tau'$  is given by

$$\tau' = \left( \prod_{i \in S} p_i^{q'_i} \right) \cdot \tau^{\tau/p+\tau} = \left( \prod_{i \in S} p_i^{p_i/p+\tau} \right) \cdot \tau^{\tau/p+\tau} = \left( \prod_{i \in S} p_i^{p_i/p} \right)^{p/p+\tau} \cdot \tau^{\tau/p+\tau} = \tau^{p/p+\tau} \cdot \tau^{\tau/p+\tau} = \tau. \triangleleft$$

▷ **Claim 27.** Consider two sets  $S_1 = S \cup \{p_{\max}\}$ ,  $S_2 = S \cup \{p_{\min}\}$  such that for any  $i \in S$ , we have  $p_{\min} \leq p_i \leq p_{\max}$ . Let  $\tau_1, \tau_2$  be the weighted geometric means of  $S_1, S_2$  respectively. Then,  $\tau_1 \geq \tau_2$ .

*Proof.* Let  $p = \sum_{i \in S} p_i$ . We have

$$\begin{aligned} \frac{\tau_1}{\tau_2} &= \frac{\prod_{i \in S} p_i^{p_i/p+p_{\max}}}{\prod_{i \in S} p_i^{p_i/p+p_{\min}}} \cdot \frac{p_{\max}^{p_{\max}/p+p_{\max}}}{p_{\min}^{p_{\min}/p+p_{\min}}} \geq \prod_{i \in S} p_i^{\frac{p_i(p_{\min}-p_{\max})}{(p+p_{\max})(p+p_{\min})}} \cdot \frac{p^{p(p_{\max}-p_{\min})}}{p_{\max}^{(p+p_{\max})(p+p_{\min})}} \\ &= \prod_{i \in S} \left( \frac{p_{\max}}{p_i} \right)^{\frac{p_i(p_{\max}-p_{\min})}{(p+p_{\max})(p+p_{\min})}} \geq 1. \quad \triangleleft \end{aligned}$$

Now, we are ready to prove Lemma 24.

**Proof of Lemma 24.** Let  $S_{\max}$  denote a subset that maximizes  $H(\mathcal{D}_S)$ . Let  $p_{\min}, p_{\max}$  respectively denote the minimum and maximum values in  $S_{\max}$ . Assume for contradiction that there is some  $q \notin S_{\max}$  such that  $p_{\max} < q < p_{\min}$ . We will show that  $\max(H(S_{\max} \cup \{q\} \setminus \{p_{\max}\}), H(S_{\max} \cup \{q\} \setminus \{p_{\min}\})) \geq H(S_{\max})$ . Clearly, this would establish the lemma.

We consider two continuous processes. The first is to decrease the value  $p_{\max}$ . As we do so, it might not remaining the maximum value in set  $S_{\max}$  but we still denote the dynamic value by  $p_{\max}$ . Note that the value of the weighted geometric mean of  $S_{\max}$ , denoted  $\tau$ , also changes; the termination condition we define below is with respect to the current value of the geometric mean and not its original value. This process ends when either of the following two events happens:

- either,  $p_{\max}$  decreases to the current value of the weighted geometric mean  $\tau$ .
- or,  $p_{\max}$  decreases all the way to  $q$  while always staying above the (dynamic) weighted geometric mean  $\tau$ . Then, by Claim 25, the value of  $H(S_{\max})$  is non-decreasing throughout this entire process. Therefore, we have replaced  $p_{\max}$  by  $q$  and not decreased  $H(S_{\max})$ , which is a contradiction.

Next, we consider the second process. This is similar to the first process, except that we now increase  $p_{\min}$  instead of decreasing  $p_{\max}$ . As earlier, as we increase the value of  $p_{\min}$ , it might not remaining the minimum value in set  $S_{\max}$  but we still denote the dynamic value by  $p_{\min}$ . Also, as in the previous process, the value of the weighted geometric mean  $\tau$  changes during this process, and the termination condition we define below is with respect to the current value of the geometric mean and not its original value. This process ends when either of the following two events happens:

- either,  $p_{\min}$  increases to the current value of the weighted geometric mean  $\tau$ .
- or,  $p_{\min}$  increases all the way to  $q$  while always staying below the (dynamic) weighted geometric mean  $\tau$ . Then, by Claim 25, the value of  $H(S_{\max})$  is non-decreasing throughout this entire process. Therefore, we have replaced  $p_{\min}$  by  $q$  and not decreased  $H(S_{\max})$ , which is a contradiction.

We are left with the case that in both processes, the first event happened. Let  $\tau_1, \tau_2$  be the final values of  $\tau$  in these two processes respectively. For the first event to happen in the first process, we must have  $q < \tau_1$ ; similarly, for the first event to happen in the second process, we must have  $q > \tau_2$ . Thus

$$\tau_2 < q < \tau_1. \quad (\text{E.17})$$

By Claim 26, the value of  $\tau_1$  is the weighted geometric mean of the set  $S_{\max} \setminus \{p_{\max}\}$ ; similarly, the value of  $\tau_2$  is the weighted geometric mean of the set  $S_{\max} \setminus \{p_{\min}\}$ . Therefore, by Claim 27,  $\tau_1 < \tau_2$ . This contradicts (E.17). ◀

## F Lower Bound on LRU

► **Theorem 28.** *There are distributions  $\mathcal{D}$  such that the cost of LRU is  $\Omega(p(\mathbb{T})) \cdot \mathcal{H}$ .*

**Proof.** Let  $\mathcal{D}$  be a distribution that has one page with probability  $\varepsilon = 4/(k \log k)$ , and  $k$  pages with probability  $(1 - \varepsilon)/k$ . By a Markov bound, the  $\varepsilon$  probability page appears by time  $2/\varepsilon = k \log k/2$  with probability at least  $1/2$ . Furthermore, by lower bounds on the coupon collector time (see, e.g., [28, pages 60-63]) the last page of the top  $k$  appears before time  $9/10 \cdot k \log k$  with probability less than  $1/4$ . Hence, with probability at least  $1/4$ , the algorithm pays  $(1 - \varepsilon)/k = \Omega(p(\mathbb{T}) \cdot \mathcal{H})$ . ◀

This does not rule out that LRU is  $\mathcal{H}$ -competitive, since this bound is witnessed by a distribution for which  $p(\mathbb{T}) \approx \text{opt}$ , but it means a different proof technique than ours is required.