


A Decomposition Approach to the Weighted k -Server Problem

Nikhil Ayyadevara ✉ 

University of Michigan, Ann Arbor, MI, USA

Ashish Chiplunkar ✉ 

Indian Institute of Technology, New Delhi, India

Amatya Sharma ✉ 

University of Michigan, Ann Arbor, MI, USA

Abstract

A natural variant of the classical online k -server problem is the *weighted k -server problem*, where the cost of moving a server is its weight times the distance through which it moves. Despite its apparent simplicity, the weighted k -server problem is extremely poorly understood. Specifically, even on uniform metric spaces, finding the optimum competitive ratio of randomized algorithms remains an open problem – the best upper bound known is $2^{2^{k+O(1)}}$ due to a deterministic algorithm (Bansal et al., 2018), and the best lower bound known is $\Omega(2^k)$ (Ayyadevara and Chiplunkar, 2021).

With the aim of closing this exponential gap between the upper and lower bounds, we propose a decomposition approach for designing a randomized algorithm for weighted k -server on uniform metrics. Our first contribution includes two relaxed versions of the problem and a technique to obtain an algorithm for weighted k -server from algorithms for the two relaxed versions. Specifically, we prove that if there exists an α_1 -competitive algorithm for one version (which we call *Weighted k -Server – Service Pattern Construction*) and there exists an α_2 -competitive algorithm for the other version (which we call *Weighted k -server – Revealed Service Pattern*), then there exists an $(\alpha_1\alpha_2)$ -competitive algorithm for weighted k -server on uniform metric spaces. Our second contribution is a $2^{O(k^2)}$ -competitive randomized algorithm for Weighted k -server – Revealed Service Pattern. As a consequence, the task of designing a $2^{\text{poly}(k)}$ -competitive randomized algorithm for weighted k -server on uniform metrics reduces to designing a $2^{\text{poly}(k)}$ -competitive randomized algorithm for Weighted k -Server – Service Pattern Construction. Finally, we also prove that the $\Omega(2^k)$ lower bound for weighted k -server, in fact, holds for Weighted k -server – Revealed Service Pattern.

2012 ACM Subject Classification Theory of computation → Online algorithms; Theory of computation → Caching and paging algorithms

Keywords and phrases Online Algorithms, k -server, paging

Digital Object Identifier 10.4230/LIPIcs.FSTTCS.2024.6

1 Introduction

The k -server problem proposed by Manasse et al. [12] is a fundamental problem in online computation, and it has been actively studied for over three decades. In this problem, we are given a metric space M and k identical servers s_1, \dots, s_k located at points of M . In every round, a point of M is requested, and an online algorithm serves the request by moving (at least) one server to the requested point. The objective is to minimize the total distance traversed by all k servers.

Like several other online problems, the performance of algorithms for the k -server problem is measured using the framework of competitive analysis introduced by Sleator and Tarjan [15]. An online algorithm for a minimization problem is said to be α -competitive if, on every input, the ratio of the algorithm's (expected) cost to the cost of the optimal solution is at most α , possibly modulo an additive constant independent of the online input. In the deterministic setup, Manasse et al. [12] showed that no k -server algorithm can be better than



© Nikhil Ayyadevara, Ashish Chiplunkar, and Amatya Sharma;
licensed under Creative Commons License CC-BY 4.0

44th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2024).

Editors: Siddharth Barman and Sławomir Lasota; Article No. 6; pp. 6:1–6:17



Leibniz International Proceedings in Informatics

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

k -competitive on any metric space with more than k points. In their breakthrough result, Koutsoupias and Papadimitriou [11] gave the best known deterministic algorithm that is $(2k - 1)$ -competitive on every metric space, famously known as the Work Function Algorithm (WFA). In the setup of randomized algorithms, it is conjectured that the competitive ratio of k -server is $O(\text{poly}(\log k))$, and this remains unsolved. Very recently, refuting the so-called *randomized k -server conjecture*, Bubeck, Coester, and Rabani [6] exhibited a family of metric spaces on which the randomized competitive ratio of the k -server problem is $\Omega(\log^2 k)$.

The k -server problem is a generalization of the online paging problem. The paging problem concerns maintaining in a “fast” memory a subset of k pages out of the n pages in a “slow” memory. In each round, one of n ($\gg k$) pages is requested, and it must replace some page in the fast memory, unless it is already in the fast memory. The objective is to minimize the number of page replacements. The paging problem is exactly the k -server under the uniform metric on the set of pages. The paging problem has been well studied, and several deterministic algorithms like Least Recently Used (LRU), First In First Out (FIFO), etc. are known to be k -competitive [15]. The randomized algorithm by Achlioptas et al. [1] is known to be $H(k)$ -competitive, matching the lower bound by Fiat et al. [8]. Here $H(k) = 1 + 1/2 + \dots + 1/k = \Theta(\log k)$.

1.1 Weighted k -server

The weighted k -server problem, first defined by Newberg [13], is a natural generalization of the k -server problem. In the weighted k -server problem, the servers are distinguishable: the i 'th server has weight w_i , where $w_1 \leq \dots \leq w_k$. The cost incurred in moving a server is its weight times the distance it travels. The objective is to minimize the total weighted distance moved by all k servers. It is easy to see that an α -competitive algorithm for the (unweighted) k -server problem has a competitive ratio of at most $\alpha \cdot w_k/w_1$ for the weighted k -server problem. However, this bound can be arbitrarily bad as w_k/w_1 is unbounded. So, the challenge is to establish weight-independent bounds on the competitive ratio of the weighted k -server problem. Surprisingly, this simple introduction of weights makes this problem incredibly difficult, and a weight-independent upper bound on the competitive ratio for an arbitrary metric is only known for the case when $k \leq 2$ [14].

Owing to its difficulty on general metric spaces, it is natural to completely understand the weighted k -server problem on the simplest class of metric spaces, the uniform metric spaces first. Uniform metric spaces are the ones in which every pair of points is separated by a unit distance. The objective of this problem translates to minimizing the weighted sum of the number of movements of each server, and thus, this problem is equivalent to paging with the cost of a page replacement dependent on the cache slot it is stored in¹. In their seminal work, Fiat and Ricklin [9] gave a deterministic algorithm for the weighted k -server problem on uniform metrics with a competitive ratio doubly exponential in k , which was later improved by Bansal et al. [4] to $2^{2^{k+2}}$. This doubly exponential behavior of the competitive ratio was proven tight by Bansal et al. [3] when they showed that the deterministic competitive ratio is no less than $2^{2^{k-4}}$.

In the randomized setup, the only known algorithm which uses randomization in a non-trivial manner is the memoryless algorithm of Chiplunkar and Vishwanathan [7], which has a competitive ratio of 1.6^{2^k} . Chiplunkar and Vishwanathan also showed that this ratio is tight

¹ Note that this problem is different from weighted paging [16], where the weights are on the pages (points in the metric space) instead on the cache slots (servers). In fact, weighted paging is equivalent to unweighted k -server on star metrics.

for the class of randomized memoryless algorithms. Recently, Ayyadevara and Chiplunkar [2] showed that no randomized algorithm (memoryless or otherwise) can achieve a competitive ratio better than $\Omega(2^k)$. Closing the exponential gap between the 1.6^{2^k} upper bound and the $\Omega(2^k)$ lower bound on the randomized competitive ratio is still an open problem.

Very recently, Gupta et al. [10] studied the weighted k -server problem in the offline and resource augmentation settings, showing the first hardness of approximation result for polynomial-time algorithms.

1.2 Our Contributions

Throughout this paper, we focus on the weighted k -server problem on uniform metrics, and we avoid mentioning the metric space henceforward. Considering the fact that the competitive ratio of a server problem is typically exponentially better in the randomized setting than the deterministic setting, it is reasonable to conjecture that there exists a randomized $2^{\text{poly}(k)}$ -competitive randomized algorithm for weighted k -server. In this paper, we propose a way of designing such an algorithm using our key idea of decomposing the weighted k -server problem into the following two relaxed versions.

Weighted k -Server – Service Pattern Construction (WkS-SPC)

The input is the same as the weighted k -server. The difference is that, in response to each request, the algorithm must only commit to the movement of some subset of servers, without specifying where those servers move to. However, it is required that there exists some solution to the given instance that agrees with the algorithm's server movements. Note that the algorithm could potentially benefit from not being lazy, that is, by moving more than one server at the same time. The (expected) cost of the algorithm is, as defined earlier, the weighted sum of the number of movements of each server (recall that we are working on a uniform metric space so the distance between every pair of points is one unit). An algorithm is said to be α -competitive if the (expected) cost of its solution is at most α times the optimum cost.

Weighted k -Server – Revealed Service Pattern (WkS-RSP)

In this version, the adversary, in addition to giving requests, is obliged to help the algorithm by providing additional information as follows. The adversary must serve each request and reveal to the algorithm the subset of servers it moved. Note that the adversary does not reveal the destination to which it moved its servers – revealing destinations makes the problem trivial because the algorithm can simply copy the adversary's movements. Given a request and the additional information about the adversary's server movements, the algorithm is required to move its own servers to cover the request. In an ideal scenario where the adversary serves the requests optimally, we require the algorithm to produce a solution whose cost competes with the cost of the optimal solution. However, consider a malicious adversary which, in an attempt to be as unhelpful to the algorithm as possible, produces a far-from-optimum solution and shares its information with the algorithm. In this case, we do not require that the algorithm competes with the optimum solution – such an algorithm would already solve the weighted k -server problem without the adversary's help. Instead, we require the algorithm to compete with the adversary's revealed solution. Formally, an algorithm is said to be α -competitive if the (expected) cost of its output is at most α times the cost of the adversary's (possibly sub-optimal) solution.

Obviously, an algorithm for the weighted k -server problem gives an algorithm for each of the above problems. Interestingly, we prove that the converse is also true. Formally,

► **Theorem 1** (Composition Theorem). *If there exists an α_1 -competitive algorithm for WkS -SPC and there exists an α_2 -competitive algorithm for WkS -RSP, then there is an $(\alpha_1\alpha_2)$ -competitive algorithm for the weighted k -server on uniform metrics.*

We prove this theorem in Section 3. As a consequence of this theorem, it is enough to design $2^{\text{poly}(k)}$ -competitive algorithms for WkS -SPC and WkS -RSP to close the exponential gap between the upper and lower bounds on the randomized competitive ratio of weighted k -server. We already present such an algorithm for WkS -RSP in Section 4. We prove,

► **Theorem 2.** *There is a randomized algorithm for WkS -RSP with a competitive ratio of $2^{O(k^2)}$.*

This reduces the task of designing a $2^{\text{poly}(k)}$ -competitive algorithm for weighted k -server to designing such an algorithm for WkS -SPC, a potentially easier problem.

Can we improve the $2^{O(k^2)}$ upper bound for WkS -RSP to, for example, $\text{poly}(k)$? We answer this question in the negative. We show that, in fact, the lower bound construction by Ayyadevara and Chiplunkar [2] for weighted k -server applies to WkS -RSP too², giving the following result.

► **Theorem 3.** *The randomized competitive ratio of WkS -RSP is $\Omega(2^k)$.*

The proof of this result is deferred to Section A.

2 Preliminaries

In this section we define the problems WkS -SPC and WkS -RSP formally, but before that, we restate the definitions of some terms introduced by Bansal et al. [3] which will be needed in our problem definitions.

2.1 Service Patterns, Feasible Labelings, Extensions

Throughout this paper, we assume without loss of generality that all the servers of the algorithm and the adversary move in response to the first request. Given a solution to an instance of the weighted k -server problem with T requests, focus on the movements of the ℓ 'th server for an arbitrary ℓ . The time instants at which these movements take place partition the interval $[1, T + 1)$ into left-closed-right-open intervals so that the server stays put at some point during each of these intervals. Thus, ignoring the locations of the servers and focusing only on the time instants at which each server moves, we get a tuple of k partitions of $[1, T + 1)$, also known as a service pattern. Formally,

► **Definition 4** (Service Pattern and Levels [3]). *A k -tuple $\mathcal{I} = (\mathcal{I}^1, \dots, \mathcal{I}^k)$ is called a service pattern over an interval $[t_{begin}, t_{end})$ if each \mathcal{I}^ℓ is a partition of $[t_{begin}, t_{end})$ comprising of left-closed-right-open intervals with integer boundaries. We call \mathcal{I}^ℓ the ℓ 'th level of \mathcal{I} .*

² The construction by Bansal et al. [3] applies too, and for the same reason, implying a doubly exponential lower bound on the deterministic competitive ratio of WkS -RSP.

Observe that the cost of a solution is completely determined by its service pattern $\mathcal{I} = (\mathcal{I}^1, \dots, \mathcal{I}^k)$: the cost equals the sum over $\ell \in \{1, \dots, k\}$ of the number of intervals in \mathcal{I}^ℓ times the weight of the ℓ 'th server.

In order to completely specify a solution, in addition to a service pattern $\mathcal{I} = (\mathcal{I}^1, \dots, \mathcal{I}^k)$, we need to specify for each ℓ and each interval $I \in \mathcal{I}^\ell$ the location of the ℓ 'th server during the time interval I . We refer to this assignment as a labeling of the service pattern. Moreover, to serve the t 'th request σ_t , we need at least one server to occupy σ_t at time t , that is, we need that there exists a level of \mathcal{I} in which the (unique) interval containing t is labeled σ_t . Such a labeling is called a feasible labeling. Formally,

► **Definition 5** (Labeling and Feasibility [3]). *A labeling of a service pattern $\mathcal{I} = (\mathcal{I}^1, \dots, \mathcal{I}^k)$ is a function from the multi-set $\mathcal{I}^1 \uplus \dots \uplus \mathcal{I}^k$ to the set U of points in the metric space. We say that a labeling γ of \mathcal{I} is feasible with respect to a request sequence $\rho = (\sigma_1, \dots, \sigma_T)$, if for each time t , there exists an interval $I \in \mathcal{I}^1 \uplus \dots \uplus \mathcal{I}^k$ containing t such that $\gamma(I) = \sigma_t$. We say that a service pattern \mathcal{I} is feasible with respect to ρ if there exists a feasible labeling of \mathcal{I} with respect to ρ .*

Recall that in the definition of the weighted k -server problem in Section 1, we assumed that the servers are numbered in a non-decreasing order of their weights. Consider the more interesting case where the weights increase at least geometrically. If we enforce that every time a server moves, all servers lighter than it move too, we lose at most a constant factor in the competitive ratio. The advantage of this enforcement is that we have a more structured class of service patterns, called hierarchical service patterns.

► **Definition 6** (Hierarchical Service Pattern [3]). *A service pattern $\mathcal{I} = (\mathcal{I}^1, \dots, \mathcal{I}^k)$ is hierarchical if for every $\ell \in \{1, \dots, k-1\}$, the partition \mathcal{I}^ℓ refines the partition $\mathcal{I}^{\ell+1}$.*

Clearly, any service pattern can be made hierarchical in an online manner with at most a k factor loss in the cost. Since we aim to obtain $2^{\text{poly}(k)}$ -competitive algorithms, the k factor loss is affordable, and therefore, we only consider hierarchical service patterns throughout this paper. Henceforth, by service pattern we actually mean a hierarchical service pattern.

Next, consider some online algorithm for the weighted k -server problem and the solution it outputs on some request sequence. For each t , let \mathcal{I}_t denote the service pattern corresponding to the algorithm's solution until the t 'th request. Observe that \mathcal{I}_{t-1} and \mathcal{I}_t are closely related: if the algorithm moves the lightest ℓ servers to serve the t 'th request (possibly $\ell = 0$), then the interval $[t, t+1)$ gets added to the first ℓ levels of \mathcal{I}_{t-1} , whereas in each of the remaining levels, the last interval in the level merges with $[t, t+1)$. We call \mathcal{I}_t the ℓ -extension of \mathcal{I}_{t-1} . More formally,

► **Definition 7** (ℓ -extension). *Let $\mathcal{I}_{t-1} = (\mathcal{I}_{t-1}^1, \dots, \mathcal{I}_{t-1}^k)$ be a hierarchical service pattern over the interval $[1, t)$, and let L_{t-1}^i be the last interval in \mathcal{I}_{t-1}^i . For $\ell \in \{0, \dots, k\}$, we define the ℓ -extension of \mathcal{I}_{t-1} to be the service pattern $\mathcal{I}_t = (\mathcal{I}_t^1, \dots, \mathcal{I}_t^k)$ over the interval $[1, t+1)$ where:*

- $\forall i \leq \ell, \mathcal{I}_t^i = \mathcal{I}_{t-1}^i \cup \{[t, t+1)\}$.
- $\forall i > \ell, \mathcal{I}_t^i = (\mathcal{I}_{t-1}^i \setminus L_{t-1}^i) \cup \{L_{t-1}^i \cup [t, t+1)\}$.

Observe that the ℓ -extension of a hierarchical service pattern is a hierarchical service pattern.

2.2 Problem Definitions

Recall that our core idea to solve weighted k -server problem is to construct an algorithm using algorithms for its two relaxed versions. We defined them informally in Section 1. Their formal definitions are as follows.

► **Definition 8** (Weighted k -Server – Service Pattern Construction (WkS-SPC)). *For every online request $\sigma_t \in U$, an algorithm for WkS-SPC is required to output a service pattern \mathcal{I}_t , which is the ℓ_t -extension of \mathcal{I}_{t-1} for some $\ell_t \in \{0, \dots, k\}$, such that \mathcal{I}_t is feasible with respect to the request sequence $\sigma_1, \sigma_2, \dots, \sigma_t$. Equivalently, the algorithm outputs ℓ_t for each t . An algorithm for WkS-SPC is said to be α -competitive if the (expected) cost of the algorithm's service pattern is at most α times the optimal cost.*

► **Definition 9** (Weighted k -server – Revealed Service Pattern (WkS-RSP)). *For every online request $\sigma_t \in U$, the adversary reveals a service pattern \mathcal{I}_t , which is the ℓ_t -extension of \mathcal{I}_{t-1} for some $\ell_t \in \{0, \dots, k\}$, such that \mathcal{I}_t is feasible with respect to the request sequence $\sigma_1, \sigma_2, \dots, \sigma_t$. Equivalently, the algorithm's input is the pair (σ_t, ℓ_t) . An algorithm for WkS-RSP is required to serve the request σ_t , i.e., move servers to ensure that σ_t is covered by some server. An algorithm for WkS-RSP is said to be β -competitive if the (expected) cost of the algorithm's solution is at most β times the cost of the final service pattern revealed by the adversary.*

3 The Composition Theorem

In this section, we explain how we can construct a weighted k -server algorithm using algorithms for its two relaxations – WkS-SPC and WkS-RSP³.

► **Theorem 1** (Composition Theorem). *If there exists an α_1 -competitive algorithm for WkS-SPC and there exists an α_2 -competitive algorithm for WkS-RSP, then there is an $(\alpha_1\alpha_2)$ -competitive algorithm for the weighted k -server on uniform metrics.*

Proof. Let \mathcal{A}_1 be an α_1 -competitive algorithm for WkS-SPC, and \mathcal{A}_2 be an α_2 -competitive algorithm for WkS-RSP. Our algorithm \mathcal{A} for weighted k -server internally runs the two algorithms \mathcal{A}_1 and \mathcal{A}_2 . At all times, \mathcal{A} keeps each of its servers at the same point where the corresponding server of \mathcal{A}_2 is located. For every input request σ_t , \mathcal{A} performs the following sequence of steps.

1. \mathcal{A} passes σ_t to \mathcal{A}_1 .
2. In response, \mathcal{A}_1 outputs an ℓ_t such that the service pattern \mathcal{I}_t , which is the ℓ_t -extension to \mathcal{I}_{t-1} , is feasible for the request sequence $\sigma_1, \sigma_2, \dots, \sigma_t$.
3. \mathcal{A} passes (σ_t, ℓ_t) to \mathcal{A}_2 .
4. In response, \mathcal{A}_2 moves its servers to serve the request σ_t .
5. \mathcal{A} copies the movements of \mathcal{A}_2 's servers.

To analyze the competitiveness of \mathcal{A} , consider an arbitrary sequence ρ of requests, and let T denote its length. Let OPT denote the cost of an optimal solution for ρ . Denote the cost of a service pattern \mathcal{I} by $\text{cost}(\mathcal{I})$. Recall that \mathcal{I}_T is the final service pattern output by \mathcal{A}_1 . Let \mathcal{I}'_T denote the service pattern corresponding to \mathcal{A}_2 's output. Note that \mathcal{I}_T and \mathcal{I}'_T are random variables, and since \mathcal{A} 's output is same as \mathcal{A}_2 's output, the cost of \mathcal{A} 's output is $\text{cost}(\mathcal{I}'_T)$.

³ On a high level, our construction resembles the result by Ben-David et al. [5], which states that if there is an α_1 -competitive randomized algorithm against online adversary and an α_2 -competitive algorithm against any oblivious adversary, then there is an $(\alpha_1\alpha_2)$ competitive randomized algorithm for any adaptive offline adversary.

Since the output of \mathcal{A}_1 is a sequence of extensions such that the service pattern remains feasible with respect to the request sequence at all times, the sequence $(\sigma_t, \ell_t)_{t=1, \dots, T}$ is a valid instance of WkS-RSP (with probability one over the randomness of \mathcal{A}_1). Since \mathcal{A}_2 is α_2 -competitive, we have $\mathbb{E}[\text{cost}(\mathcal{I}'_T) \mid \mathcal{I}_T = \mathcal{I}] \leq \alpha_2 \cdot \text{cost}(\mathcal{I})$ for every service pattern \mathcal{I} feasible with respect to ρ . This implies $\mathbb{E}[\text{cost}(\mathcal{I}'_T)] \leq \alpha_2 \cdot \mathbb{E}[\text{cost}(\mathcal{I}_T)]$. Since \mathcal{A}_1 is α_1 -competitive, $\mathbb{E}[\text{cost}(\mathcal{I}_T)] \leq \alpha_1 \cdot \text{OPT}$. Thus, $\mathbb{E}[\text{cost}(\mathcal{I}'_T)] \leq \alpha_1 \alpha_2 \cdot \text{OPT}$. This implies that \mathcal{A} is $(\alpha_1 \alpha_2)$ -competitive. \blacktriangleleft

4 Competing with a Revealed Service Pattern

We organize this section as follows. In Section 4.1, we prove some structural results that are used in the definition and analysis of our algorithm. We define our algorithm formally in Section 4.2 and analyze its competitive ratio in Section 4.3. We use the following notation.

- U denotes the set of points in a uniform metric space.
- For t from 1 to T , The t 'th request is $\sigma_t \in U$, and $\rho_t = (\sigma_1, \dots, \sigma_t)$ denotes the sequence of requests received until time t .
- The service pattern revealed by the adversary with the t 'th request is denoted by $\mathcal{I}_t = (\mathcal{I}_t^1, \dots, \mathcal{I}_t^k)$. Recall that \mathcal{I}_t is the ℓ_t -extension of \mathcal{I}_{t-1} . Without loss of generality, we assume that the adversary moves all its servers with the first request, and therefore $\ell_1 = k$.
- L_t^ℓ denotes the last interval in \mathcal{I}_t^ℓ , that is, the unique interval in \mathcal{I}_t^ℓ that covers $[t, t+1)$.
- s_t^ℓ denotes the location of our algorithm's ℓ 'th server after processing the t 'th request. Since our algorithm is randomized, s_t^ℓ is a random variable. Note that for the t 'th request to be served, we must have $\sigma_t \in \{s_t^1, \dots, s_t^k\}$ with probability one.

Consider the adversary's service pattern $\mathcal{I}_t = (\mathcal{I}_t^1, \dots, \mathcal{I}_t^k)$. For an arbitrary ℓ , fix the labels of the last intervals $L_t^{\ell+1}, \dots, L_t^k$ in the top $k - \ell$ levels $\mathcal{I}_t^{\ell+1}, \dots, \mathcal{I}_t^k$ of \mathcal{I}_t , and consider all feasible labelings of \mathcal{I}_t with respect to ρ that agree with the fixed labels. The set of labels that these labelings assign to the last interval L_t^ℓ of \mathcal{I}_t^ℓ will be crucial for our algorithm. We now define this set formally.

► **Definition 10.** For any $t \in \{1, \dots, T\}$, $\ell \in \{1, \dots, k\}$, and $p^{\ell+1}, \dots, p^k \in U$, the set $Q_t^\ell(p^{\ell+1}, \dots, p^k)$ is defined to be the set of points p^ℓ for which there exists a feasible labeling γ of \mathcal{I}_t with respect to ρ_t such that $\gamma(L_t^i) = p^i$ for all $i \in \{\ell, \dots, k\}$.

4.1 Structural Results

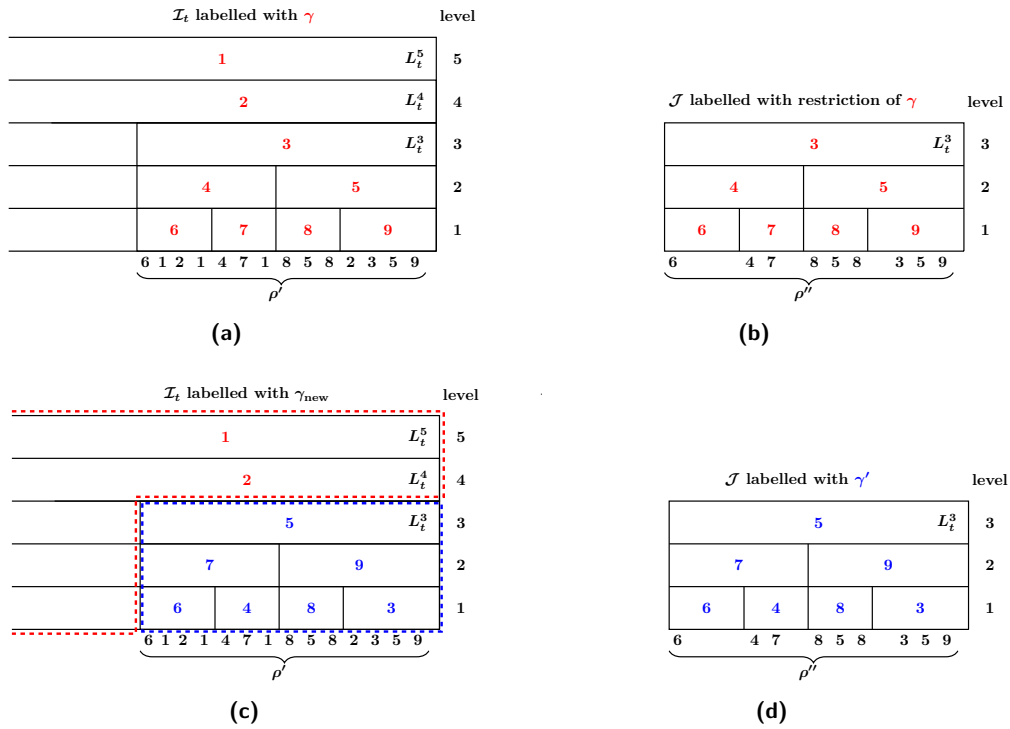
Bansal et al. [3] considered the following combinatorial question: given a service pattern \mathcal{I} and a request sequence ρ , how many labels can an interval in the k 'th level of \mathcal{I} get, over all possible feasible labelings of \mathcal{I} with respect to ρ ? They derived the following interesting property.

► **Fact 11** (Dichotomy Property [3]). *There exists a sequence n_1, n_2, \dots of integers with $n_k \leq 2^{2^{k+3 \log k}}$ such that the following holds: for every k , every sequence of requests $\rho = (\sigma_1, \dots, \sigma_T)$, every service pattern $\mathcal{I} = (\mathcal{I}^1, \dots, \mathcal{I}^k)$ over $[1, T+1)$, and every $I \in \mathcal{I}^k$, the set Q of labels of I over all feasible labelings of \mathcal{I} with respect to ρ is the entire U , or it has size at most n_k .*

The next lemma generalizes the above result to intervals in every level.

► **Lemma 12** (Generalized Dichotomy Property). *For every $t \in \{1, \dots, T\}$, $\ell \in \{1, \dots, k\}$, and $p^{\ell+1}, \dots, p^k \in U$, the set $Q_t^\ell(p^{\ell+1}, \dots, p^k)$ is the entire U , or it has size at most n_ℓ , where $n_\ell \leq 2^{2^{\ell+3} \log \ell}$ is the constant from Fact 11.*

Proof. The lemma holds trivially when $Q_t^\ell(p^{\ell+1}, \dots, p^k) = \emptyset$, so assume $Q_t^\ell(p^{\ell+1}, \dots, p^k) \neq \emptyset$. Let ρ' denote the request sequence during the interval L_t^ℓ and \mathcal{J} denote the restriction of the service pattern \mathcal{I}_t to the interval L_t^ℓ and levels $1, \dots, \ell$. Let ρ'' denote the subsequence of ρ' formed by removing all the requests to $p^{\ell+1}, \dots, p^k$. Let Q denote the set of labels of the interval L_t^ℓ over all the feasible labelings of the ℓ -level service pattern \mathcal{J} with respect to ρ'' . From Fact 11 we get that the set Q is either U or has size at most n_ℓ . We argue that $Q_t^\ell(p^{\ell+1}, \dots, p^k) = Q$, and this implies the claim. Refer to Figure 1 for a working illustration on an instance with $k = 5$ and $\ell = 3$.



■ **Figure 1** An illustration of Lemma 12 for $k = 5$ and $\ell = 3$. (a) Depicts \mathcal{I}_t with a labeling γ (colored in red) feasible with respect to ρ_t such that $\gamma(L_t^5) = 1$ and $\gamma(L_t^4) = 2$. (b) Depicts the 3-level service pattern \mathcal{J} labeled with the restriction of γ , along with ρ'' , the subsequence of ρ' formed by removing all the requests to points 1 and 2. (c) Depicts the labeling γ_{new} constructed by overwriting γ' onto γ for every interval in \mathcal{J} .

For any $p^\ell \in Q_t^\ell(p^{\ell+1}, \dots, p^k)$, from Definition 10 we get that there exists a feasible labeling γ of \mathcal{I}_t with respect to ρ_t such that $\gamma(L_t^i) = p^i$ for all $i \in \{\ell, \dots, k\}$. In the labeling γ , the servers $\ell + 1, \dots, k$ can only serve the requests to the points $p^{\ell+1}, \dots, p^k$ during the interval L_t^ℓ . This implies that all the requests in ρ'' must be served by servers $1, \dots, \ell$. Thus, the restriction γ' of γ to \mathcal{J} is feasible with respect to ρ'' . But $\gamma'(L_t^\ell) = \gamma(L_t^\ell) = p^\ell$. Therefore, $p^\ell \in Q$. Thus, $Q_t^\ell(p^{\ell+1}, \dots, p^k) \subseteq Q$.

Now consider any point $p^\ell \in Q$, and let γ' be a feasible labeling of \mathcal{J} with respect to ρ'' such that $\gamma'(L_t^\ell) = p^\ell$. Suppose γ is a feasible labeling of \mathcal{I}_t with respect to ρ_t such that $\gamma(L_t^i) = p^i$ for all $i \in \{\ell + 1, \dots, k\}$ (such a labeling exists because we assumed $Q_t^\ell(p^{\ell+1}, \dots, p^k) \neq \emptyset$). Overwrite the labeling γ' onto γ to get a new labeling γ_{new} . Formally, $\gamma_{\text{new}}(I) = \gamma'(I)$ if interval I is in \mathcal{J} , else $\gamma_{\text{new}}(I) = \gamma(I)$. We claim that γ_{new} is also a feasible labeling of \mathcal{I}_t with respect to ρ_t . This can be argued as follows.

The labeling γ_{new} serves all the requests in ρ'' because it agrees with γ' in the service pattern \mathcal{J} . γ_{new} serves all the requests during the interval L_t^ℓ other than those in ρ'' because all these requests are made at points in $\{p^{\ell+1}, \dots, p^k\}$, and $\gamma_{\text{new}}(L_t^i) = \gamma(L_t^i) = p^i$ for all $i \in \{\ell + 1, \dots, k\}$. Finally, γ_{new} serves all the requests before the interval L_t^ℓ because it agrees with γ before the interval L_t^ℓ .

Thus, γ_{new} is a feasible labeling of \mathcal{I}_t with respect to ρ_t such that $\gamma_{\text{new}}(L_t^i) = p^i$ for all $i \in \{\ell + 1, \dots, k\}$. From Definition 10, we get that $\gamma_{\text{new}}(L_t^\ell) = p^\ell \in Q_t^\ell(p^{\ell+1}, \dots, p^k)$. This implies $Q \subseteq Q_t^\ell(p^{\ell+1}, \dots, p^k)$. ◀

We now state a useful consequence of the simple fact that the restriction of a solution for the first t requests to the first $t - 1$ requests is a solution for the first $t - 1$ requests.

► **Lemma 13.** *For every $t \in \{2, \dots, T\}$, $\ell \in \{\ell_t + 1, \dots, k\}$, and $p^{\ell+1}, \dots, p^k \in U$, $Q_t^\ell(p^{\ell+1}, \dots, p^k) \subseteq Q_{t-1}^\ell(p^{\ell+1}, \dots, p^k)$.*

Proof. Recall that $\mathcal{I}_t = (\mathcal{I}_t^1, \dots, \mathcal{I}_t^k)$ is the ℓ_t -extension of $\mathcal{I}_{t-1} = (\mathcal{I}_{t-1}^1, \dots, \mathcal{I}_{t-1}^k)$, which means $L_{t-1}^\ell = L_t^\ell \setminus [t, t+1) \neq \emptyset$. By Definition 10, if some point p^ℓ is in $Q_t^\ell(p^{\ell+1}, \dots, p^k)$, then there exists a feasible labeling γ of \mathcal{I}_t with respect to ρ_t such that $\gamma(L_t^i) = p^i$ for all $i \in \{\ell, \dots, t\}$. Restrict γ to obtain a labeling γ' of \mathcal{I}_{t-1} in the obvious manner: $\gamma'(L_{t-1}^i) = \gamma(L_t^i) = p^i$ for all $i \in \{\ell, \dots, t\}$ and $\gamma'(I) = \gamma(I)$ for all other intervals I of \mathcal{I}_{t-1} (which are also intervals of \mathcal{I}_t). It is easy to check that γ' is a feasible labeling of \mathcal{I}_{t-1} with respect to ρ_{t-1} . Thus, from Definition 10 we get, $p^\ell \in Q_{t-1}^\ell(p^{\ell+1}, \dots, p^k)$. ◀

4.2 Algorithm

Before stating our WkS-RSP algorithm formally, we give some intuitive explanation. Recall that s_t^ℓ denotes the location of the algorithm's ℓ 'th server after serving the t 'th request. The critical invariant maintained by our algorithm is the following.

► **Invariant 14.** *For every t and ℓ , in response to the t 'th request, the algorithm keeps its ℓ 'th server at a uniformly random point in $Q_t^\ell(s_t^{\ell+1}, \dots, s_t^k)$.*

Lemma 16 states this claim formally. For now, let us just understand the scenarios in which the algorithm needs to move the ℓ 'th server at time t so that it occupies *some* point in $Q_t^\ell(s_t^{\ell+1}, \dots, s_t^k)$. These scenarios are as follows.

1. The algorithm moves the ℓ' 'th server for some $\ell' > \ell$. This means that, potentially, $s_t^{\ell'} \neq s_{t-1}^{\ell'}$, so $Q_t^\ell(s_t^{\ell+1}, \dots, s_t^k)$ could be different from $Q_{t-1}^\ell(s_{t-1}^{\ell+1}, \dots, s_{t-1}^k)$.
2. $\ell_t \geq \ell$. This means that $L_t^\ell = [t, t+1)$ is disjoint from L_{t-1}^ℓ , and again, potentially, $Q_t^\ell(s_t^{\ell+1}, \dots, s_t^k)$ could be different from $Q_{t-1}^\ell(s_{t-1}^{\ell+1}, \dots, s_{t-1}^k)$.
3. None of the above happens, so by Lemma 13, $Q_t^\ell(s_t^{\ell+1}, \dots, s_t^k) \subseteq Q_{t-1}^\ell(s_{t-1}^{\ell+1}, \dots, s_{t-1}^k)$. However, $s_{t-1}^\ell \notin Q_t^\ell(s_t^{\ell+1}, \dots, s_t^k)$ (that is, $s_{t-1}^\ell \in Q_{t-1}^\ell(s_{t-1}^{\ell+1}, \dots, s_{t-1}^k) \setminus Q_t^\ell(s_t^{\ell+1}, \dots, s_t^k)$), so the ℓ 'th server can no longer remain at the same place s_{t-1}^ℓ as before.

6:10 A Decomposition Approach to the Weighted k -Server Problem

We call the movement in the first two scenarios a *forced movement* (because the movement of some other server forced this movement), and the movement in the third scenario an *unforced movement*. If none of the above scenarios arises, then the ℓ 'th server stays put. Algorithm 1 is the formal description of our algorithm for WkS-RSP.

■ **Algorithm 1** WkS-RSP.

```

1: for  $t = 1$  to  $T$  do
2:   Input: request  $\sigma_t \in U$ , and  $\ell_t \in \{0, \dots, k\}$ .
3:   {Recall:  $\mathcal{I}_t$  is the  $\ell_t$ -extension of  $\mathcal{I}_{t-1}$ .}
4:   flag  $\leftarrow$  FALSE
5:   for  $\ell = k$  to  $1$  do
6:     {Decide movements of servers in decreasing order of weight.}
7:     {flag = TRUE indicates that an unforced movement of some server heavier than the
       $\ell$ 'th has happened.}
8:     Compute  $Q_t^\ell(s_t^{\ell+1}, \dots, s_t^k)$  (by brute force).
9:     if flag OR  $\ell \leq \ell_t$  then
10:       $s_t^\ell \leftarrow$  a uniformly random point in  $Q_t^\ell(s_t^{\ell+1}, \dots, s_t^k)$ . {forced movement}
11:      else if  $s_{t-1}^\ell \notin Q_t^\ell(s_t^{\ell+1}, \dots, s_t^k)$  then
12:         $s_t^\ell \leftarrow$  a uniformly random point in  $Q_t^\ell(s_t^{\ell+1}, \dots, s_t^k)$ . {unforced movement}
13:        flag  $\leftarrow$  TRUE.
14:      else
15:         $s_t^\ell \leftarrow s_{t-1}^\ell$ . {no movement}
16:      end if
17:    end for
18:  end for

```

Note that it is unclear so far why Algorithm 1 is well-defined – why the set $Q_t^\ell(s_t^{\ell+1}, \dots, s_t^k)$ is nonempty when we attempt to send the ℓ 'th server to a uniformly random point in it in steps 10 and 12 – and why every request gets served. We provide an answer now.

► **Lemma 15.** *For every $t \in \{1, \dots, T\}$ the following statements hold with probability one.*

1. *For every $\ell \in \{0, \dots, k\}$, there exists a feasible labeling γ of \mathcal{I}_t with respect to ρ_t such that $\gamma(L_t^i) = s_t^i$ for all $i \in \{\ell + 1, \dots, k\}$.*
2. *For every $\ell \in \{1, \dots, k\}$, the set $Q_t^\ell(s_t^{\ell+1}, \dots, s_t^k)$ is non-empty.*
3. *Algorithm 1 serves the t 'th request.*

Proof. For an arbitrary $t \in \{1, \dots, T\}$, we prove the lemma by reverse induction on $\ell \in \{0, \dots, k\}$ in an interleaved manner. More precisely, as the base case, we prove the first claim for $\ell = k$. Assuming that the first claim holds for an arbitrary $\ell > 0$, we prove that the second claim holds for the same ℓ . Assuming that the second claim holds for an arbitrary $\ell > 0$, we prove that the first claim holds for $\ell - 1$. Finally, assuming that the first claim holds for $\ell = 0$, we prove that the third claim holds.

As the base case, we need to prove the first claim for $\ell = k$. We know that the service pattern \mathcal{I}_t that the adversary provides is feasible. This implies that there exists a feasible labeling γ of \mathcal{I}_t with respect to ρ_t . The condition $\gamma(L_t^i) = s_t^i$ for all $i \in \{\ell + 1, \dots, k\}$ is vacuously true.

For the inductive step, assume that the first claim holds for some $0 < \ell \leq k$. Hence, there exists a feasible labeling γ of \mathcal{I}_t such that $\gamma(L_t^i) = s_t^i$, for all $i \in \{\ell + 1, \dots, k\}$. By Definition 10, the point $\gamma(L_t^\ell)$ lies in $Q_t^\ell(s_t^{\ell+1}, \dots, s_t^k)$. Thus, $Q_t^\ell(s_t^{\ell+1}, \dots, s_t^k) \neq \emptyset$.

We designed the algorithm so that s_t^ℓ is guaranteed to be in $Q_t^\ell(s_t^{\ell+1}, \dots, s_t^k)$. By Definition 10, there exists a feasible labeling γ' of \mathcal{I}_t such that $\gamma'(L_t^i) = s_t^i$ for all $i \in \{\ell, \dots, k\}$. Hence, the first claim holds for $\ell - 1$ as well. This proves the first two claims.

Finally, since the first claim holds for $\ell = 0$, there exists a feasible labeling γ of \mathcal{I}_t with respect to ρ_t such that $\gamma(L_t^i) = s_t^i$ for all $i \in \{1, \dots, k\}$. By definition of feasibility of a labeling (Definition 5), γ must assign the label σ_t to some interval in \mathcal{I}_t that contains t . Since the only intervals in \mathcal{I}_t that contains t are the L_t^i 's, we must have $\sigma_t = s_t^i$ for some i . Since s_t^i 's are the positions of the algorithm's servers after processing the t 'th request, the request gets served. \blacktriangleleft

4.3 Competitive Analysis

We begin by proving that the algorithm indeed maintains Invariant 14 after serving every request.

► **Lemma 16.** *For every $t \in \{1, \dots, T\}$, every $\ell \in \{1, \dots, k\}$, and every $p^{\ell+1}, \dots, p^k \in U$, conditioned on $s_t^i = p^i$ for all $i \in \{\ell + 1, \dots, k\}$ and $Q_t^\ell(p^{\ell+1}, \dots, p^k) \neq \emptyset$, s_t^ℓ is a uniformly random point in $Q_t^\ell(p^{\ell+1}, \dots, p^k)$.*

Proof. We prove this lemma by induction of time t . The base case of $t = 1$ is true because we are assuming $\ell_1 = k$, which makes the condition in step 9 of the algorithm true.

Consider the inductive case, where $t > 1$. Note that the algorithm executes exactly one step out of 10, 12, and 15. Conditioned on the algorithm executing step 10 or 12, s_t^ℓ is located at a uniformly random point in $Q_t^\ell(p^{\ell+1}, \dots, p^k)$ by design. On the other hand, suppose the algorithm executes step 15, that is, the checks in steps 9 and 11 fail. Then the fact that the check of step 9 failed implies that the algorithm did not move any server heavier than the ℓ 'th. Thus $(s_{t-1}^{\ell+1}, \dots, s_{t-1}^k) = (s_t^{\ell+1}, \dots, s_t^k) = (p^{\ell+1}, \dots, p^k)$. Therefore, by induction hypothesis, s_{t-1}^ℓ is a uniformly random point in $Q_{t-1}^\ell(p^{\ell+1}, \dots, p^k)$. Additionally, conditioned on the failure of the check in step 11, $s_{t-1}^\ell \in Q_t^\ell(s_t^{\ell+1}, \dots, s_t^k)$, so $s_t^\ell = s_{t-1}^\ell$ is a uniformly random point in $Q_{t-1}^\ell(p^{\ell+1}, \dots, p^k) \cap Q_t^\ell(p^{\ell+1}, \dots, p^k)$. But by Lemma 13, $Q_{t-1}^\ell(p^{\ell+1}, \dots, p^k) \cap Q_t^\ell(p^{\ell+1}, \dots, p^k) = Q_t^\ell(p^{\ell+1}, \dots, p^k)$, so s_t^ℓ is a uniformly random point in $Q_t^\ell(p^{\ell+1}, \dots, p^k)$. Thus, irrespective of which one of steps 10, 12, and 15 is executed, s_t^ℓ is a uniformly random point in $Q_t^\ell(p^{\ell+1}, \dots, p^k)$, and this implies the claim. \blacktriangleleft

Having established that our algorithm is well-defined and that it indeed serves every request, we now focus on bounding the cost of algorithm's solution. For each $\ell \in \{1, \dots, k\}$, define the random variables X^ℓ and Y^ℓ to be the number of forced movements and unforced movements respectively, of the algorithm's ℓ 'th server. First, we bound X^ℓ for all ℓ as follows.

► **Lemma 17.** *The following inequalities hold with probability one.*

1. $X^\ell \leq X^{\ell+1} + Y^{\ell+1} + |\mathcal{I}_T^\ell|$ for all $\ell \in \{1, \dots, k - 1\}$.
2. $X^k \leq |\mathcal{I}_T^k|$.

Proof. For $\ell < k$, every forced movement of the ℓ 'th server happens at time t only if **flag** is true or $\ell \leq \ell_t$. Observe that in the former case, the $(\ell + 1)$ 'th server must have moved at time t too, so we charge the movement of the ℓ 'th server to the movement of the $(\ell + 1)$ 'th server, which could be either forced or unforced. In the latter case, since \mathcal{I}_T is a hierarchical service pattern, a new interval starts at time t in the ℓ 'th level \mathcal{I}_T^ℓ of \mathcal{I}_T , so we charge the movement of the ℓ 'th server to that interval. The argument for the second claim is the same as above except that **flag** is never true; a forced movement of the k 'th server happens at time t only if $\ell_t = k$. \blacktriangleleft

6:12 A Decomposition Approach to the Weighted k -Server Problem

Next, we bound Y^ℓ by first bounding the number of unforced movements of the ℓ 'th server in an arbitrary interval in which no forced movement of the ℓ 'th server happens.

► **Lemma 18.** *For every $\ell \in \{1, \dots, k\}$, every $p^{\ell+1}, \dots, p^k \in U$, and every $t_{\text{begin}}, t_{\text{end}}$ such that $1 \leq t_{\text{begin}} < t_{\text{end}} \leq T + 1$, the following holds. Conditioned on the event that $s_t^j = p^j$ for all $j \in \{\ell + 1, \dots, k\}$, and all $t \in (t_{\text{begin}}, t_{\text{end}})$, the expected number of unforced movements of the algorithm's ℓ 'th server is at most $H(n_\ell)$, where H denotes the harmonic function defined as $H(n) = 1 + 1/2 + \dots + 1/n$.*

Proof. Conditioned on the event that $s_t^j = p^j$ for all $j \in \{\ell + 1, \dots, k\}$ and all $t \in (t_{\text{begin}}, t_{\text{end}})$, we use Lemma 13 to claim that $Q_{t-1}^\ell(p^{\ell+1}, \dots, p^k) \supseteq Q_t^\ell(p^{\ell+1}, \dots, p^k)$ for every $t \in (t_{\text{begin}}, t_{\text{end}})$. For brevity we write Q_t for $Q_t^\ell(p^{\ell+1}, \dots, p^k)$. Let Z_t be the indicator random variable of the event that an unforced movement happens at time t . From the algorithm, we know that this event happens if and only if $s_{t-1}^\ell \in Q_{t-1} \setminus Q_t$. From Lemma 16, we know that s_{t-1}^ℓ is a uniformly random point in Q_{t-1} . Thus,

$$\mathbb{E}[Z_t] = \frac{|Q_{t-1}| - |Q_t|}{|Q_{t-1}|} \leq \frac{1}{|Q_{t-1}|} + \frac{1}{|Q_{t-1}| - 1} + \dots + \frac{1}{|Q_t| + 1} = H(|Q_{t-1}|) - H(|Q_t|)$$

Let t_1 denote the earliest time t for which $Q_t \subsetneq Q_{t-1}$. Then the expected number of unforced movements is bounded as

$$\sum_{t=t_1}^{t_{\text{end}}-1} \mathbb{E}[Z_t] \leq 1 + \sum_{t=t_1+1}^{t_{\text{end}}-1} H(|Q_{t-1}|) - H(|Q_t|) = 1 + H(|Q_{t_1}|) - H(|Q_{t_{\text{end}}-1}|).$$

Since $Q_{t_1} \subsetneq Q_{t_1-1} \subseteq U$, by Lemma 12, we have $|Q_{t_1}| \leq n_\ell$. By the second claim of Lemma 15, $|Q_{t_{\text{end}}-1}| \geq 1$. Thus, the expected number of unforced movements is at most $H(n_\ell)$. ◀

► **Lemma 19.** *For every $\ell \in \{1, \dots, k\}$, we have $\mathbb{E}[Y^\ell] \leq H(n_\ell) \cdot \mathbb{E}[X^\ell]$.*

Proof. Let us condition on the sequence of timestamps at which a forced movement of the ℓ 'th server takes place. If t_1, t_2 are two consecutive timestamps in this sequence, it is easy to notice that the servers $\ell + 1, \dots, k$ remain at the same position throughout this interval. Then Lemma 18 applied to the interval (t_1, t_2) implies that the expected number of unforced movements of the ℓ 'th server in this interval is at most $H(n_\ell)$. Summing up over all pairs t_1, t_2 of consecutive timestamps, we get $\mathbb{E}[Y^\ell | X^\ell = x] \leq H(n_\ell) \cdot x$, and therefore, $\mathbb{E}[Y^\ell] \leq H(n_\ell) \cdot \mathbb{E}[X^\ell]$. ◀

► **Theorem 2.** *There is a randomized algorithm for WkS -RSP with a competitive ratio of $2^{O(k^2)}$.*

Proof. From Lemma 15, we already know that Algorithm 1 serves every request in ρ_T . Towards proving competitiveness of the algorithm, we first define the constants c_k, \dots, c_1 inductively as follows: $c_k = H(n_k) + 1$ and $c_\ell = (H(n_\ell) + 1) \cdot (c_{\ell+1} + 1)$, for every $\ell \in \{1, \dots, k-1\}$. We claim that for every $\ell \in \{1, \dots, k\}$, the expected number of movements of the algorithm's ℓ 'th server, which equals $\mathbb{E}[X^\ell] + \mathbb{E}[Y^\ell]$, is at most c_ℓ times the number of movements of the adversary's ℓ 'th server, which equals $|\mathcal{I}_T^\ell|$. We prove this claim using reverse induction on ℓ from k to 1.

For the base case, i.e. $\ell = k$, from Lemma 19 we have that $\mathbb{E}[Y^k] \leq H(n_k) \cdot \mathbb{E}[X^k]$. From Lemma 17, we know that $\mathbb{E}[X^k] \leq |\mathcal{I}_T^k|$. Thus, the expected number of algorithm's k 'th server movements is,

$$\mathbb{E}[X^k] + \mathbb{E}[Y^k] \leq (H(n_k) + 1) \cdot \mathbb{E}[X^k] = c_k \cdot |\mathcal{I}_T^k|.$$

For the inductive case, assume that $\mathbb{E}[X^{\ell+1}] + \mathbb{E}[Y^{\ell+1}] \leq c_{\ell+1} \cdot |\mathcal{I}_T^{\ell+1}|$, for an arbitrary $\ell \in \{1, \dots, k-1\}$. From Lemma 19, we have that $\mathbb{E}[Y^\ell] \leq H(n_\ell) \cdot \mathbb{E}[X^\ell]$, and from Lemma 17, we have that $\mathbb{E}[X^\ell] \leq \mathbb{E}[X^{\ell+1}] + \mathbb{E}[Y^{\ell+1}] + |\mathcal{I}_T^\ell|$. Thus, we have,

$$\mathbb{E}[X^\ell] + \mathbb{E}[Y^\ell] \leq (H(n_\ell) + 1) \cdot \mathbb{E}[X^\ell] \leq (H(n_\ell) + 1) \cdot (\mathbb{E}[X^{\ell+1}] + \mathbb{E}[Y^{\ell+1}] + |\mathcal{I}_T^\ell|).$$

Recall that $\mathcal{I}_T = (\mathcal{I}_T^1, \dots, \mathcal{I}_T^k)$ is a hierarchical service pattern, and therefore, $|\mathcal{I}_T^{\ell+1}| \leq |\mathcal{I}_T^\ell|$. Using this fact, the induction hypothesis, and the definition of c_ℓ , we get,

$$\mathbb{E}[X^\ell] + \mathbb{E}[Y^\ell] \leq (H(n_\ell) + 1) \cdot (c_{\ell+1} \cdot |\mathcal{I}_T^{\ell+1}| + |\mathcal{I}_T^\ell|) \leq (H(n_\ell) + 1) \cdot (c_{\ell+1} + 1) \cdot |\mathcal{I}_T^\ell| = c_\ell \cdot |\mathcal{I}_T^\ell|,$$

as required.

As a consequence of the above inductive claim, the total cost of the algorithm is at most $\max\{c_1, \dots, c_k\} = c_1$ times the cost of the adversary's service pattern. Moreover, from the recurrence relation defining c_k, \dots, c_1 and the upper bound on n_ℓ from Fact 11, it is clear that c_1 is $2^{O(k^2)}$. Thus, the competitive ratio of our algorithm is $2^{O(k^2)}$. \blacktriangleleft

5 Concluding Remarks and Open Problems

The main open question of finding the randomized competitive ratio of weighted k -server on uniform metrics still remains unresolved. Our decomposition approach and the randomized algorithm for WkS -RSP imply that the task of designing a $2^{\text{poly}(k)}$ -competitive randomized algorithm for weighted k -server on uniform metrics is equivalent to designing a $2^{\text{poly}(k)}$ -competitive algorithm for WkS -SPC. We do not know any non-trivial bounds on the competitive ratio of WkS -SPC and it is not even clear whether it is easier or harder than WkS -RSP in terms of competitiveness. While the known lower bound constructions for weighted k -server also apply to WkS -RSP, these constructions fail to get a lower bound on the competitive ratio of WkS -SPC. We therefore propose the open problem of finding bounds on the competitive ratio of WkS -SPC in the deterministic as well as randomized setting.

In the deterministic setting, the competitive ratio of WkS -SPC is bounded from below by the competitive ratio of weighted k -server divided by the competitive ratio of WkS -RSP, again due to Theorem 1. However, for this to give a non-trivial lower bound on the competitive ratio of WkS -SPC, we require an upper bound on the competitive ratio of WkS -RSP that is less than the known lower bound on the competitive ratio of weighted k -server. Unfortunately, no such upper bound is known. Thus, showing a separation between weighted k -server and WkS -RSP is an interesting open problem.

Additionally, we also believe that closing the quadratic gap between the exponents in the upper and lower bounds on the randomized competitive ratio of WkS -RSP is an interesting open problem, because it will result in a better understanding of the weighted k -server problem.

Finally, for the weighted k -server problem with $k > 2$, no weight-independent and metric-independent upper bounds on the competitive ratio are known on any well-structured class of metrics larger than the class of uniform metrics. Proving such bounds seems rather ambitious, given our limited understanding of weighted k -server on uniform metrics.

References

- 1 Dimitris Achlioptas, Marek Chrobak, and John Noga. Competitive analysis of randomized paging algorithms. *Theor. Comput. Sci.*, 234(1-2):203–218, 2000. doi:10.1016/S0304-3975(98)00116-9.

- 2 Nikhil Ayyadevara and Ashish Chiplunkar. The randomized competitive ratio of weighted k -server is at least exponential. In *ESA*, volume 204 of *LIPICs*, pages 9:1–9:11. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021. doi:10.4230/LIPICs.ESA.2021.9.
- 3 Nikhil Bansal, Marek Eliás, and Grigorios Koumoutsos. Weighted k -server bounds via combinatorial dichotomies. In *FOCS*, pages 493–504, 2017. doi:10.1109/FOCS.2017.52.
- 4 Nikhil Bansal, Marek Eliás, Grigorios Koumoutsos, and Jesper Nederlof. Competitive algorithms for generalized k -server in uniform metrics. In *SODA*, pages 992–1001, 2018. doi:10.1137/1.9781611975031.64.
- 5 Shai Ben-David, Allan Borodin, Richard M. Karp, Gábor Tardos, and Avi Wigderson. On the power of randomization in on-line algorithms. *Algorithmica*, 11(1):2–14, 1994. doi:10.1007/BF01294260.
- 6 Sébastien Bubeck, Christian Coester, and Yuval Rabani. The randomized k -server conjecture is false! In *STOC*, pages 581–594. ACM, 2023. doi:10.1145/3564246.3585132.
- 7 Ashish Chiplunkar and Sundar Vishwanathan. Randomized memoryless algorithms for the weighted and the generalized k -server problems. *ACM Trans. Algorithms*, 16(1):14:1–14:28, 2020. doi:10.1145/3365002.
- 8 Amos Fiat, Richard M. Karp, Michael Luby, Lyle A. McGeoch, Daniel Dominic Sleator, and Neal E. Young. Competitive paging algorithms. *J. Algorithms*, 12(4):685–699, 1991. doi:10.1016/0196-6774(91)90041-V.
- 9 Amos Fiat and Moty Ricklin. Competitive algorithms for the weighted server problem. *Theoretical Computer Science*, 130(1):85–99, 1994. doi:10.1016/0304-3975(94)90154-6.
- 10 Anupam Gupta, Amit Kumar, and Debmalya Panigrahi. Efficient algorithms and hardness results for the weighted k -server problem. In *APPROX/RANDOM*, volume 275 of *LIPICs*, pages 12:1–12:19. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023. doi:10.4230/LIPICs.APPROX/RANDOM.2023.12.
- 11 Elias Koutsoupias and Christos H. Papadimitriou. On the k -server conjecture. *J. ACM*, 42(5):971–983, 1995. doi:10.1145/210118.210128.
- 12 Mark S. Manasse, Lyle A. McGeoch, and Daniel Dominic Sleator. Competitive algorithms for on-line problems. In *STOC*, pages 322–333, 1988. doi:10.1145/62212.62243.
- 13 Lee A. Newberg. The k -server problem with distinguishable servers. Master’s thesis, University of California, Berkeley, 1991.
- 14 René Sitters. The generalized work function algorithm is competitive for the generalized 2-server problem. *SIAM J. Comput.*, 43(1):96–125, 2014. doi:10.1137/120885309.
- 15 Daniel Dominic Sleator and Robert Endre Tarjan. Amortized efficiency of list update and paging rules. *Commun. ACM*, 28(2):202–208, 1985. doi:10.1145/2786.2793.
- 16 Neal E. Young. On-line file caching. *Algorithmica*, 33(3):371–383, 2002. doi:10.1007/s00453-001-0124-5.

A Lower Bound for WkS -RSP

In this section, we show that the lower-bound construction for weighted k -server by Ayyadevara and Chiplunkar [2] applies to WkS -RSP and gives the same lower bound. In [2] the generation of an adversarial request sequence involves repeatedly calling a randomized recursive procedure named **strategy**. Their analysis bounds the adversary’s cost and the expected cost of an arbitrary deterministic algorithm, both amortized per **strategy** call. Then the exponential lower bound is established by an application of Yao’s principle. We show that essentially the same adversarial construction of input distribution works for WkS -RSP. The only change needed in the construction is that now the adversary must provide its service pattern along with the requested point in an online manner.

The weights of the servers are $1, \beta, \dots, \beta^{k-1}$ where β is a large integer. The sequence n_0, n_1, \dots is defined as $n_0 = 1$ and for $\ell > 0$,

$$n_\ell = \left(\left\lceil \frac{n_{\ell-1}}{2} \right\rceil + 1 \right) \cdot \left(\left\lfloor \frac{n_{\ell-1}}{2} \right\rfloor + 1 \right).$$

The adversarial strategy in [2] uses the following combinatorial result from [3].

► **Fact 20** ([3]). *Let $\ell \in \mathbb{N}$ and let P be a set of n_ℓ points. There exists a set-system $\mathcal{Q}_\ell \subseteq 2^P$ satisfying the following properties.*

1. \mathcal{Q}_ℓ contains $\lceil n_{\ell-1}/2 \rceil + 1$ sets, each of size $n_{\ell-1}$.
2. For every $p \in P$, there exists a set in \mathcal{Q}_ℓ not containing p .
3. For every $p \in P$, there exists a $q \in P$ such that every set in \mathcal{Q}_ℓ contains at least one of p and q .

We modify the adversarial strategy from [2] for weighted k -server to get the following strategy for WkS-RSP.

■ **Procedure 2** adversary.

Mark all points in S ;
repeat *infinitely many times*
 Pick a point p uniformly at random from S (with replacement);
 Mark p ;
 if *All points in S are marked* **then**
 $\ell_{ext} \leftarrow k$;
 Unmark all points $q \in S$ other than p ;
 else
 $\ell_{ext} \leftarrow k - 1$;
 Call **strategy**($k - 1, S \setminus \{p\}, \ell_{ext}$);

■ **Procedure 3** **strategy**(ℓ, P, ℓ_{ext}) (Promise: $|P| = n_\ell$ and $\ell_{ext} \geq \ell$).

if $\ell = 0$ (and therefore, $|P| = n_0 = 1$) **then**
 Output (p, ℓ_{ext}) , where p is the unique point in P ;
else
 Construct the set-system $\mathcal{Q}_\ell \subseteq 2^P$ using Fact 20;
 repeat $(\beta - 1) \cdot (\lceil n_{\ell-1}/2 \rceil + 1)$ **times**
 Pick a set P' uniformly at random from \mathcal{Q}_ℓ (with replacement);
 Call **strategy**($\ell - 1, P', \ell_{ext}$);
 $\ell_{ext} \leftarrow \ell - 1$;

The following claim bounds the expected cost of an arbitrary online algorithm for every **strategy** call made by **adversary**. Its proof is identical to the proof of Corollary 6 in [2]. It is noteworthy that all the arguments involved in that proof go through even in the revealed service pattern setting.

► **Lemma 21.** *For $\ell_{ext} = k$ or $k-1$, the expected cost of the algorithm per **strategy**($k-1, P, \ell_{ext}$) call made by **adversary** is $(\beta - 1)^{k-1}/(n_{k-1} + 1)$.*

We now show that the service pattern $\mathcal{I} = (\mathcal{I}^1, \dots, \mathcal{I}^k)$ created by the procedure **adversary** is feasible, that is, it can be labeled in a way that all requests get served. We start by noting the following.

6:16 A Decomposition Approach to the Weighted k -Server Problem

► **Observation 22.** Let $(p_{t_1}, \ell_{t_1}), \dots, (p_{t_2}, \ell_{t_2})$ be the input generated by a $\text{strategy}(\ell, P, \ell_{ext})$ call which starts at time t_1 and ends at time t_2 . Then $\ell_{t_1} = \ell_{ext} \geq \ell$ and $\ell_{t_1+1}, \dots, \ell_{t_2}$ are all less than ℓ . As a consequence, the following statements about the adversary's service pattern $\mathcal{I} = (\mathcal{I}^1, \dots, \mathcal{I}^k)$ hold.

1. For all $\ell' \geq \ell$, a single interval in $\mathcal{I}^{\ell'}$ covers the interval $[t_1, t_2 + 1)$.
2. For all $\ell' \leq \ell$, the interval in $\mathcal{I}^{\ell'}$ covering t_1 starts at t_1 , and the interval in $\mathcal{I}^{\ell'}$ covering t_2 ends at $t_2 + 1$.

In particular, $[t_1, t_2 + 1)$ is an interval in \mathcal{I}^ℓ .

► **Lemma 23.** Consider an arbitrary $\text{strategy}(\ell, P, \ell_{ext})$ call which starts at time t_1 , ends at time t_2 , and generates the input $(p_{t_1}, \ell_{t_1}), \dots, (p_{t_2}, \ell_{t_2})$. Suppose for all $\ell' > \ell$, the single interval in $\mathcal{I}^{\ell'}$ covering $[t_1, t_2 + 1)$ is labeled with some point $p_{\ell'}$, such that $P \cap \{p_{\ell+1}, \dots, p_k\} \neq \emptyset$. Then the intervals in $\mathcal{I}^1, \dots, \mathcal{I}^\ell$ that intersect $[t_1, t_2 + 1)$ (and therefore, are subsets of $[t_1, t_2 + 1)$) can be labeled in such a way that for all $t \in \{t_1, \dots, t_2\}$ there exists $i \in \{1, \dots, k\}$ such that the unique interval in \mathcal{I}^i covering t is labeled with p_t .

Proof. We prove by induction on ℓ . For $\ell = 0$, the set P contains a single point, which gets requested. Since $P \cap \{p_1, \dots, p_k\} \neq \emptyset$, the claim holds.

For $\ell > 0$, consider a point $p \in P \cap \{p_{\ell+1}, \dots, p_k\}$. From the third property in Fact 20, there exists a point $q \in P$ such that every set in the set system \mathcal{Q}_ℓ contains at least one of the points p or q . Label the interval $[t_1, t_2 + 1)$ in \mathcal{I}^ℓ by such a point q . Consider an arbitrary recursive call $\text{strategy}(\ell - 1, P', \ell_{ext})$ which starts at time $t'_1 \geq t_1$ and ends at time $t'_2 \leq t_2$. We have $P' \cap \{q, p_{\ell+1}, \dots, p_k\} \neq \emptyset$. By induction hypothesis, the intervals in $\mathcal{I}^1, \dots, \mathcal{I}^{\ell-1}$ that intersect $[t'_1, t'_2 + 1)$ can be labeled in such a way that for all $t \in \{t'_1, \dots, t'_2\}$ there exists $i \in \{1, \dots, k\}$ such that the unique interval in \mathcal{I}^i covering t is labeled with p_t . ◀

► **Lemma 24.** The service pattern $\mathcal{I} = (\mathcal{I}^1, \dots, \mathcal{I}^k)$ created by the procedure adversary is feasible.

Proof. Every interval in \mathcal{I}^k starts exactly when all points in S are found to be marked. For every interval I^k in \mathcal{I}^k do the following. Label it by the point q whose marking results in the beginning of the next interval in \mathcal{I}^k . In other words, q is the last point to get marked after the unmarking step in the beginning of I^k . Thus, q is never sampled by adversary during the interval I^k , and therefore q belongs to the set $S \setminus \{p\}$ passed to every strategy call made by adversary during the interval I^k . Thus, by Lemma 23 for $\ell = k - 1$, the intervals in $\mathcal{I}^1, \dots, \mathcal{I}^{k-1}$ that are subsets of I^k can be labeled in such a way that all requests given during the interval I^k are served. Thus, the service pattern \mathcal{I} is feasible. ◀

Now, we bound the cost of the service pattern $\mathcal{I} = (\mathcal{I}^1, \dots, \mathcal{I}^k)$ created by the procedure adversary. We start by bounding the total cost of intervals created during a $\text{strategy}(\ell, P, \ell_{ext})$ call.

► **Lemma 25.** Define the sequence c_0, c_1, \dots inductively as follows: $c_0 = 0$ and for $\ell > 0$

$$c_\ell = \beta^{\ell-1} + \beta \cdot (\lceil n_{\ell-1}/2 \rceil + 1) \cdot c_{\ell-1}$$

For an arbitrary $\ell \in \{0, \dots, k-1\}$ and $\ell_{ext} \geq \ell$, consider a call of $\text{strategy}(\ell, P, \ell_{ext})$ which starts at time t_1 and ends at time t_2 . The total cost of the intervals in layers $\mathcal{I}_1, \dots, \mathcal{I}_\ell$ that intersect the interval $[t_1, t_2 + 1)$ (equivalently, are subsets of $[t_1, t_2 + 1)$, by Observation 22) is at most c_ℓ .

Proof. We prove this by induction on ℓ . The claim is trivially true for $\ell = 0$. For $\ell > 0$, the $\text{strategy}(\ell, P, \ell_{ext})$ makes $(\beta - 1) \cdot (\lceil n_{\ell-1}/2 \rceil + 1)$ recursive calls of $\text{strategy}(\ell - 1, P', \ell_{ext})$. For each of these calls, the following holds by induction hypothesis. If the call starts at time $t'_1 \geq t_1$ and ends at time $t'_2 \leq t_2$, then the total cost of intervals in layers $\mathcal{I}_1, \dots, \mathcal{I}_{\ell-1}$ that intersect the interval $[t'_1, t'_2 + 1)$ is at most $c_{\ell-1}$. Since there are $(\beta - 1) \cdot (\lceil n_{\ell-1}/2 \rceil + 1)$ such recursive calls the total cost of intervals in layers $\mathcal{I}_1, \dots, \mathcal{I}_{\ell-1}$ that intersect the interval $[t_1, t_2 + 1)$ is at most $(\beta - 1) \cdot (\lceil n_{\ell-1}/2 \rceil + 1) \cdot c_{\ell-1}$. Adding to this the cost $\beta^{\ell-1}$ of the interval $[t_1, t_2 + 1) \in \mathcal{I}_\ell$ gives the required bound. \blacktriangleleft

► **Theorem 3.** *The randomized competitive ratio of WkS-RSP is $\Omega(2^k)$.*

Proof. Consider the service pattern $\mathcal{I} = (\mathcal{I}^1, \dots, \mathcal{I}^k)$ of the adversary. Every interval in \mathcal{I}^k starts with one marked point and ends just before all the points in the set S are marked in the procedure adversary. Using the standard coupon collector argument, we get that the expected number of $\text{strategy}(k-1, S \setminus \{p\}, \ell_{ext})$ made during an interval in \mathcal{I}^k is $(n_{k-1} + 1)H(n_{k-1})$. Thus, the amortized cost of intervals in \mathcal{I}^k per $\text{strategy}(k-1, S \setminus \{p\}, \ell_{ext})$ call is $\beta^{k-1}/((n_{k-1} + 1)H(n_{k-1}))$. From Lemma 25, we get that the total cost of intervals in $\mathcal{I}^1, \dots, \mathcal{I}^{k-1}$ per $\text{strategy}(k-1, S \setminus \{p\}, \ell_{ext})$ call is at most c_{k-1} . The cost of the revealed service pattern per $\text{strategy}(k-1, S \setminus \{p\}, \ell_{ext})$ call is at most $\beta^{k-1}/((n_{k-1} + 1)H(n_{k-1})) + c_{k-1}$. The rest of the proof is identical to the proof of Theorem 2 in [2]. Essentially, for a large β , the dominant term in the adversary's cost is $\beta^{k-1}/((n_{k-1} + 1)H(n_{k-1}))$, while the algorithm's cost is $\beta^{k-1}/(n_{k-1} + 1)$ modulo a lower order term due to Lemma 21, thus implying a lower bound of $H(n_{k-1}) = \Omega(2^k)$ on the competitive ratio. \blacktriangleleft