k-server via multiscale entropic regularization

Sebastien Bubeck
Microsoft Research

Michael B. Cohen
MIT

James R. Lee
University of Washington

Yin Tat Lee
University of Washington

Aleksander Mądry
MIT

Abstract

We present an $O((\log k)^2)$-competitive randomized algorithm for the k-server problem on hierarchically separated trees (HSTs). This is the first $o(k)$-competitive randomized algorithm for which the competitive ratio is independent of the size of the underlying HST. Our algorithm is designed in the framework of online mirror descent where the mirror map is a multiscale entropy. When combined with Bartal's static HST embedding reduction, this leads to an $O((\log k)^2 \log n)$-competitive algorithm on any n-point metric space. We give a new dynamic HST embedding that yields an $O((\log k)^3 \log \Delta)$-competitive algorithm on any metric space where the ratio of the largest to smallest non-zero distance is at most Δ.

Contents

1 Introduction

Perhaps the most widely-studied problem in the field of online algorithms and competitive analysis is the *k-server problem,* introduced in [MMS90] to generalize and abstract a number of related problems arising in the study of paging and caching. The problem has been the object of intensive study since its inception, motivated largely by two long-standing conjectures about the competitive ratios that can be achieved by deterministic and randomized algorithms, respectively.

We recall the problem briefly; see Section 1.2 for a formal definition of the model. Fix a metric space (X, d) and $k \geqslant 1$, as well as an initial placement $\rho_0 \in X^k$ of k servers in X. An online k-server algorithm operates as follows. At each time step, a request $r_t \in X$ comes online, and the algorithm must respond to this request by moving one of the servers to r_t (unless there is already a server there). The cost of the algorithm is the total distance moved by all the servers over the course of the request sequence. An *offline* algorithm operates in the same manner, but is allowed access to the entire request sequence in advance. An online algorithm has competitive ratio α if, for every request sequence, its movement cost per unit time step is within an α factor of that achieved by the optimal offline algorithm.

Randomization. The authors of [MMS90] stated the *k-server conjecture:* On any metric space with at least $k+1$ points, the best competitive ratio achieved by *deterministic* online algorithms is precisely k. They showed that the ratio is always at least k. While the conjecture is still open in general, Koutsoupias and Papadimitriou resolved it within a factor of two: The work function algorithm obtains a competitive ratio of $2k-1$ on any metric space [KP95]. We refer the reader to the book [BE98] for further background on online algorithms and the k-server problem.

In the context of the *k-paging problem*, which is the special case of k-server on a metric space with all distances in the set $\{0, 1\}$, it was observed that randomness can help an online algorithm dramatically: It is known that the competitive ratio for k-paging is precisely the kth harmonic number H_k for every $k \geqslant 1$ [FKL+91, MS91].

Hierarchically separated trees. There is another class of metric spaces on which one can prove lower bounds on the competitive ratio even for randomized algorithms. Consider a finite, rooted tree $\mathcal{T} = (V, E)$ with vertex weights $w : V \to \mathbb{R}_+$ that are non-increasing along every root-leaf path. Let $\mathcal{L} \subseteq V$ denote the set of leaves and define the metric

$$d_w(\ell, \ell') := w_{\mathsf{lca}(\ell,\ell')} \qquad \forall \ell, \ell' \in \mathcal{L}\,.$$

In this case, d_w is an ultrametric (in fact, all finite ultrametrics are of this form). If $w_v \leqslant w_u/\tau$ whenever v is a child of u, then one says that $(\mathcal{T}, w)$ is a *τ-hierarchically separated tree (τ-HST)* and the metric space $(\mathcal{L}, d_w)$ is referred to as a *τ-HST metric.*

Lower bounds on the competitive ratio were established for any sufficiently large metric space [KRR94, BKRS00]. Following this framework, the authors of [BBM06] demonstrate a lower bound for every HST metric; showng that every large metric space contains a large subset that is close to an HST (and using the bounds from [BLMN05]) establishes that the competitive ratio for randomized algorithms is at least $\Omega\left(\frac{\log k}{\log\log k}\right)$ on every metric space with more than k points.

A lack of further lower bounds motivates the folklore "Randomized k-server conjecture" (see, e.g., [BBK99]) that the randomized competitive ratio is $(\log k)^{O(1)}$ on every metric space, or even $O(\log k)$, matching the lower bound for k-paging. One can also consult the survey [Kou09] (in particular, Conjecture 2 and the surrounding discussion).

Since the seminal work of Bartal on probabilistic embeddings of finite metric spaces into HSTs [Bar96, Bar98], it has been understood that obtaining *upper bounds* on the competitive ratio for HSTs is of central importance. Indeed, the competitive ratio for n-point metric spaces is at most an $O(\log n)$ factor larger than the competitive ratio for n-point HST metrics. This bound follows from Bartal's approach combined with the optimal distortion estimate of [FRT04].

The authors of [CMP08] give an $O(D)$-competitive randomized algorithm on binary 2-HSTs with combinatorial depth at most D, and in [BBMN15], a major breakthrough was achieved when the authors exhibited an $(\log n)^{O(1)}$-competitive algorithm for general n-vertex HSTs. In the present work, we obtain a competitive ratio independent of the size of the underlying HST, thereby verifying a long-held belief.

Theorem 1.1. *For every $k \geqslant 2$, there is an $O((\log k)^2)$-competitive randomized algorithm for the k-server problem on any HST.*

As mentioned previously, this yields an $O((\log k)^2 \log n)$-competitive randomized algorithm for every n-point metric space, via probabilistic embeddings of finite metric spaces into distributions over HSTs. The embedding underlying this reduction is *oblivious* to the request sequence. While this is a very convenient feature for the analysis, oblivious embeddings cannot avoid losing an $\Omega(\log n)$ factor in the competitive ratio. We show that in certain cases, this can be circumvented via the use of *dynamic* HST embeddings where the embedding is allowed to depend on the request sequence.

Theorem 1.2. *For every $k \geqslant 2$ and every finite metric space (X, d), there is an $O\left((\log k)^3 \log(1 + \mathcal{A}_X)\right)$-competitive randomized algorithm for the k-server problem on (X, d), where*

$$\mathcal{A}_X := \frac{\max_{x,y \in X} d(x, y)}{\min_{x \neq y \in X} d(x, y)}.$$

An amusing aspect of the dynamic embedding[1] is that it can be combined with known methods for trees (namely, the Double-Coverage algorithm [CKPV91, CL91]) to obtain a randomized algorithm for any finite metric space (X, d) that is poly(k)-competitive and which responds to a request in time poly(n, k), where $n = |X|$. It seems that this is the first such algorithm; e.g., the work function algorithm of [KP95] requires $n^{O(k)}$ time to respond to a request.

1.1 Mirror descent and entropic regularization

Our algorithm is most naturally stated in the framework of continuous-time *mirror descent*. This framework was originally introduced for convex optimization in [NY83] (see also [Bub15]), and recently it has played a key role in online decision making; see, e.g., [Haz16] for the online learning setting, and [ABBS10, BCN14] for applications to metrical task systems. Typically an entropy functional is used as a mirror map, and a key contribution of our work is to propose an appropriate multiscale entropy functional.

We establish some properties of a general setup in Section 2 and, as a warmup application, present in Section 2.2 an $O(\log k)$-competitive algorithm for the (fractional) weighted paging problem that is closely related to the algorithm of [BBN12]. This already exhibits a couple key ideas in a simplified setting, including the natural use of the Bregman divergence as a potential function, and the utility of using $k + \varepsilon$ servers for some $\varepsilon < 1$. In Section 3, we begin transferring these ideas to the setting of the k-server problem on trees. Notably, the polytope underlying our state representation is the one derived from the fractional allocation problem as employed in [CMP08] and [BBMN15]. In Section 3.4, we introduce the crucial idea of an auxilliary potential function that tracks the weighted depth of the underlying fractional server measure, and in Section 3.4.3 we show how using a time-varying weight can be leveraged to obtain an $O((\log k)^2)$ competitive ratio.

[1]This was pointed out by Yair Bartal.

1.2 Preliminaries

We use the notations $\mathbb{R}_+ := [0, \infty)$ and $\mathbb{Z}_+ = \mathbb{Z} \cap \mathbb{R}_+$. If X and Y are two metric spaces and $F : X \to Y$ is Lipschitz, we use $\|F\|_{\mathrm{lip}}$ to denote the Lipschitz constant of F. Consider a bounded, complete metric space (X, d) and two Borel probability measures μ and ν on X. We use $\mathbb{W}^1_X(\mu, \nu)$ to denote the L^1-transportation distance between μ and ν (sometimes called the Earthmover metric):

$$\mathbb{W}^1_X(\mu, \nu) := \inf \mathbb{E}\left[d(Y, Y')\right],$$

where the infimum is over all jointly distributed random variables (Y, Y') such that Y has marginal μ and Y' has marginal ν. The definition is extended in the natural way to any two Borel measures satisfying $\mu(X) = \nu(X)$.

Online algorithms and the competitive ratio. Let (X, d) be a metric space and fix $k \geqslant 1$. We now describe the k-server problem more formally. The input is a sequence $\langle \sigma_t \in X : t \geqslant 1\rangle$ of *requests*. At every time t, an online algorithm maintains a state $\rho_t \in X^k$ which can be thought of as the location of k *servers* in the space X. At time t, the algorithm is required to have a server at the requested site $r_t \in X$. In other words, a feasible state ρ_t is one that *services* r_t*:*

$$r_t \in \left\{(\rho_t)_1, \ldots, (\rho_t)_k\right\}.$$

Formally, an *online algorithm* is a sequence of mappings $\boldsymbol{\rho} = \langle \rho_1, \rho_2, \ldots, \rangle$ where, for every $t \geqslant 1$, $\rho_t : X^t \to X^k$ maps a request sequence $\langle r_1, \ldots, r_t\rangle$ to a k-server state that services r_t. In general, $\rho_0 \in X^k$ will denote some initial state of the algorithm.

The *cost of the algorithm* $\boldsymbol{\rho}$ *in servicing* $\boldsymbol{r} = \langle r_t : t \geqslant 1\rangle$ is defined as the sum of the movements of all the servers:

$$\mathsf{cost}_{\boldsymbol{\rho}}(\boldsymbol{r}; k) := \sum_{t \geqslant 1} d_{X^k}\left(\rho_t(r_1, \ldots, r_t), \rho_{t-1}(r_1, \ldots, r_{t-1})\right),$$

where

$$d_{X^k}\left((x_1, \ldots, x_k), (y_1, \ldots, y_k)\right) := \sum_{i=1}^{k} d_X(x_i, y_i) \qquad \forall x_1, \ldots, x_k, y_1, \ldots, y_k \in X,$$

and $\rho_0 \in X^k$ is some fixed initial configuration.

For a given request sequence $\boldsymbol{r} = \langle r_t : t \geqslant 1\rangle$, denote the cost of the *offline optimum* by

$$\mathsf{cost}^*(\boldsymbol{r}; k) := \inf_{\langle \rho_1, \rho_2, \ldots\rangle} \sum_{t \geqslant 1} d_{X^k}\left(\rho_t, \rho_{t-1}\right),$$

where the infimum is over all sequences $\langle \rho_1, \rho_2, \ldots\rangle$ such that ρ_t services r_t for each $t \geqslant 1$.

A *randomized online algorithm* $\boldsymbol{\rho}$ is a random online algorithm that is feasible with probability one. Such an algorithm is said to be α*-competitive* if for every $\rho_0 \in X^k$, there is a constant $c > 0$ such that for all $\boldsymbol{r}$:

$$\mathbb{E}\left[\mathsf{cost}_{\boldsymbol{\rho}}(\boldsymbol{r}; k)\right] \leqslant \alpha \cdot \mathsf{cost}^*(\boldsymbol{r}; k) + c.$$

2 Traversing a convex body online

Suppose that $\mathsf{K} \subseteq \mathbb{R}^n$ is a closed convex set and $f : \mathbb{R}_+ \times \mathsf{K} \to \mathbb{R}^n$ is a time-varying vector field defined on K. It is very natural to consider the *projected dynamical system:*

$$\begin{aligned} x'(t) &= \Pi_{\mathsf{K}}\left(x(t), f(t, x(t))\right) \\ x(0) &= x_0, \end{aligned} \tag{2.1}$$

where

$$\Pi_{\mathsf{K}}(x,v) := \lim_{\varepsilon\to 0} \frac{P_{\mathsf{K}}(x+\varepsilon v) - x}{\varepsilon},$$

and

$$P_{\mathsf{K}}(y) := \operatorname{argmin}\left\{\|y-z\|_2 : z \in \mathsf{K}\right\}.$$

One can interpret this as trying to "flow" along the vector field in direction $f(t, x(t))$ while being confined to remain in the convex body K. But since projection is a discontinuous operation (imagine hitting the boundary of a polytope, for instance), the classical theory of existence and uniqueness of ODEs no longer applies. Fortunately, there is now a well-established theory for projected dynamical systems.

Mirror descent. For our applications, we will want to consider a projected dynamical system with respect to a non-Euclidean geometry on K. The resulting dynamics will correspond to continuous-time mirror descent.

Let $\Phi : \mathsf{K} \to \mathbb{R}$ denote a strongly convex function. For those familiar with mirror descent, it is natural to define an analogous Bregman projection operator

$$\tilde{P}^{\Phi}_{\mathsf{K}}(x) := \operatorname{argmin}\left\{D_\Phi(x;z) : z \in \mathsf{K}\right\},$$

where

$$D_\Phi(x;z) := \Phi(x) - \Phi(z) - \langle \nabla\Phi(z), x-z\rangle \tag{2.2}$$

is the Bregman divergence corresponding to Φ.

Problematically, this is not well-defined if $x \notin \mathsf{K}$. Instead, we should do projection in the local norm at x induced by $\nabla^2\Phi$:

$$P^{\Phi}_{\mathsf{K}}(y,x) := \operatorname{argmin}\left\{\|y-z\|^2_{\Phi,x} : z \in \mathsf{K}\right\},$$

where

$$\|w\|^2_{\Phi,x} := \left\langle w, \nabla^2\Phi(x) w\right\rangle \tag{2.3}$$

is the local norm at x (formally, a norm on the tangent space T_x). We can then sensibly define $\Pi^{\Phi}_{\mathsf{K}} : \mathsf{K} \times \mathbb{R}^n \to \mathbb{R}^n$ as before:

$$\Pi^{\Phi}_{\mathsf{K}}(x,v) := \lim_{\varepsilon\to 0} \frac{P^{\Phi}_{\mathsf{K}}(x+\varepsilon v, x) - x}{\varepsilon}.$$

This leads to a projected dynamical system corresponding to continuous-time mirror descent:

$$\begin{aligned} x'(t) &= \Pi^{\Phi}_{\mathsf{K}}\left(x(t), f(t,x(t))\right) \\ x(0) &= x_0 . \end{aligned}$$

It turns out that under mild regularity assumptions, one can establish existence and uniqueness of these dynamics.

Let us denote the *normal cone to* K *at* x by

$$N_{\mathsf{K}}(x) := \left\{p \in \mathbb{R}^n : \langle p, y-x\rangle \leqslant 0 \text{ for all } y \in \mathsf{K}\right\}.$$

If K is a polyhedron, then the normal cone to K at x is the cone spanned by the normals of the tight constraints at x.

Lemma 2.1. *Given any matrix $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, consider the polyhedron $\mathsf{K} := \{x \in \mathbb{R}^n : Ax \leqslant b\}$. For any $x \in \mathsf{K}$, it holds that*

$$N_{\mathsf{K}}(x) = \left\{A^T y \ : \ y \geqslant 0 \text{ and } y^T(b - Ax) = 0\right\}.$$

The following theorem is proved in Section 5.

Theorem 2.2. *Consider a compact convex set* $\mathsf{K} \subseteq \mathbb{R}^n$*, a strongly convex function* $\Phi : \mathsf{K} \to \mathbb{R}$*, and a continuous function* $f : [0, \infty) \times \mathsf{K} \to \mathbb{R}^n$*. Suppose furthermore that* $\nabla^2\Phi(x)^{-1}$ *is continuous. Then for any* $x_0 \in \mathsf{K}$*, there is an absolutely continuous solution* $x : [0, \infty) \to \mathsf{K}$ *satisfying*

$$\begin{aligned} x'(t) &\in \nabla^2\Phi(x(t))^{-1} \left(f(t, x(t)) - N_{\mathsf{K}}(x(t)) \right), \\ x(0) &= x_0. \end{aligned} \tag{2.4}$$

If we further assume that $\nabla^2\Phi(x)$ *is Lipschitz and* f *is locally Lipschitz, then the solution is unique.*

Note that the right-hand side of (2.4) is a set, and a solution is one that satisfies the inclusion. (This is known as a *differential inclusion* and is discussed in Section 5.) To understand the intuition behind (2.4), consider the following. If we considered the Euclidean projected dynamics (2.1), it would make sense that $x'(t) = f(t, x(t)) - \lambda(t)$, where $\lambda(t) \in N_{\mathsf{K}}(x(t))$: The dynamics are attempting to flow in the direction $f(t, x(t))$, and $\lambda(t)$ represents a set of forces in directions perpendicular to the tight constraints that are keeping $x(t)$ inside the body K. The inclusion (2.4) is analogous, except that "perpendicular" is a notion that now depends on the local geometry induced by $\nabla^2\Phi$ at $x(t)$ (cf. (2.3)).

2.1 Evolution of the Bregman divergence

Recall that the *Bregman divergence* associated to $\Phi : \mathsf{K} \to \mathbb{R}$ is given by (2.2). We will use D_Φ as a potential function to track the "discrepancy" between our algorithm and the optimal offline algorithm. In fact, it will be slightly easier to work with the function

$$\hat{D}_\Phi(y; x) := -\Phi(x) - \langle \nabla\Phi(x), y - x \rangle .$$

Suppose now that $x(t)$ is an absolutely continuous solution to the differential inclusion (2.4) and write

$$\nabla^2\Phi(x(t))\partial_t x(t) = f(t, x(t)) - \lambda(t)$$

with $\lambda(t) \in N_{\mathsf{K}}(x(t))$. One concludes immediately that for $y \in \mathsf{K}$:

$$\begin{aligned} \partial_t \hat{D}_\Phi(y; x(t)) &= \langle \nabla^2\Phi(x(t))\partial_t x(t), x(t) - y \rangle \\ &= \left\langle f(t, x(t)) - \lambda(t), x(t) - y \right\rangle \\ &\leqslant \left\langle f(t, x(t)), x(t) - y \right\rangle , \end{aligned} \tag{2.5}$$

where in the last inequality we have used that $\langle \lambda(t), y - x(t) \rangle \leqslant 0$ since $\lambda(t) \in N_{\mathsf{K}}(x(t))$.

2.2 Application: Fractional weighted paging

Fix $k \geqslant 1$. Consider the fractional weighted paging problem on pages in $[n]$ with a cache of size k and positive weights $\{w_i > 0 : i \in [n]\}$. For $z \in \mathbb{R}^n$, define the weighted ℓ_1 norm:

$$\|z\|_{\ell_1(w)} := \sum_{i=1}^{n} w_i \, |z_i| ,$$

and the dual norm:

$$\|z\|_{\ell_\infty(1/w)} := \max \left\{ \frac{|z_i|}{w_i} : i \in [n] \right\} .$$

Note that if $z(t) \in \mathbb{R}^n$ is an online algorithm for $t \in [0, T]$, then the movement cost is precisely

$$\int_0^T \|\partial_t z(t)\|_{\ell_1(w)}\, dt\,.$$

Moreover, up to a factor of two, we can charge our algorithm only for the cost of moving fractional mass *into* a node, i.e.,

$$\int_0^T \left\|(\partial_t z(t))_+\right\|_{\ell_1(w)}\, dt\,, \tag{2.6}$$

where for $z \in \mathbb{R}^n$, we denote $(z)_+ := (\max(0, z_1), \ldots, \max(0, z_n))$.

2.2.1 Entropy-regularized dynamics

Define the *fractional k-antipaging polytope*

$$\mathsf{P} := \left\{ x \in [0,1]^n : \sum_{i=1}^n x_i = n - k \right\}.$$

Here, we think of $1 - x_i$ as the fractional amount of page i that sits in the cache (hence x_i is the amount of fractional "antipage"). Define also the entropic regularizer

$$\Phi(x) := \sum_{i=1}^n w_i x_i \log x_i\,.$$

Suppose the current state of the cache is described by a point $x \in \mathsf{P}$ such that $x_i(0) > 0$ for all $i \in [n]$. When a request $r \in \{1, \ldots, n\}$ is received, we need to decrease x_r to 0. To this end, we use the (constant) control function:

$$f(t, x(t)) := -e_r \qquad \forall t \geqslant 0\,.$$

Now the intended trajectory is given by the differential inclusion:

$$\nabla^2\Phi(x(t))\partial_t x(t) \in -e_r - N_{\mathsf{P}}(x(t))\,.$$

Let us analyze the dynamics which are described by

$$\nabla^2\Phi(x(t))\partial_t x(t) = -e_r - \lambda(t)\,, \tag{2.7}$$

where $\lambda(t) \in N_{\mathsf{P}}(x(t))$. From Lemma 2.1, it is an exercise to compute that

$$\lambda(t) = \sum_{i=1}^n \lambda_i(t) e_i - \mu(t)\mathbb{1} \tag{2.8}$$

for some $\{\lambda_i(t) \geqslant 0\}$ such that $\lambda_i(t) > 0 \implies x_i(t) < 1$.

Here, the $\{\lambda_i\}$ functions are the Lagrangian multipliers for the constraints $\{x_i \leqslant 1\}$ of P, and μ is the multiplier for $\sum_{i=1}^n x_i = n - k$. The fact that $x_i(t) > 0$ for $t > 0$ is implicitly enforced by Φ (assuming some boundedness on the control f). Since $\|\nabla\Phi(x)\|_2 \to \infty$ as x approaches the boundary of the positive orthant, Φ acts as a barrier preventing the evolution from leaving $\mathbb{R}_+$. We do not stress this point formally at the moment since we will soon need to maintain a more restrictive condition.

Let $\hat{x} \in \{0,1\}^n \cap \mathsf{P}$ denote an integral antipaging point with $\hat{x}_r = 0$ (i.e., a state which has satisfied the request). Then (2.5) immediately yields

$$\partial_t \hat{D}_\Phi(\hat{x}; x(t)) \leqslant \langle e_r, \hat{x} - x(t) \rangle = -x_r(t)\,. \tag{2.9}$$

Moreover, from (2.7) and (2.8), one easily calculates:

$$\partial_t x_i(t) = \frac{x_i(t)}{w_i}\left(-e_r(i) - \lambda_i(t) + \mu(t)\right), \tag{2.10}$$

where

$$\lambda_i(t) = \begin{cases} 0 & x_i(t) < 1 \\ \mu(t) & \text{otherwise.} \end{cases}\,.$$

Using $\partial_t \sum_{i=1}^n x_i(t) = 0$ yields

$$\mu(t) = \frac{x_r(t)/w_r}{\sum_{i: x_i(t) < 1} x_i(t)/w_i}\,. \tag{2.11}$$

In particular, if $x_r(t) < 1$ then $\mu(t) \leqslant 1$, hence from (2.10), the instantaneous movement cost (recall (2.6)) is bounded by

$$w_r |\partial_t x_r(t)| \leqslant x_r(t)\,.$$

Thus the potential change (2.9) compensates for the movement cost.

Now we have to address convergence, and here we run into a problem: $\hat{D}_\Phi(\hat{x}; x(t))$ could be infinite! Therefore (2.9) does not show that $x_r(t) \to 0$ as $t \to \infty$.

2.2.2 Moving in the interior

Our solution to this problem will be to shift the variables away from the boundary of $\mathbb{R}_+$. For $\delta > 0$, define

$$\mathsf{P}_\delta := \mathsf{P} \cap [\delta, 1]^n = \left\{ x \in [\delta,1]^n : \sum_{i=1}^n x_i = n - k \right\}.$$

Clearly we cannot remain in this polytope and still service a request r by moving to a point with $x_r = 0$. Instead, we will allow our algorithm to satisfy the weaker constraint $x_r = \delta$, and then afterward show that any such algorithm can be transformed—in an online manner—to a valid fractional paging algorithm, as long as δ is chosen small enough. Furthermore, we can easily ensure that our dynamics remain inside P_δ by simplying stopping when $x_r(T) = \delta$ (if we can ensure that there is a time T at which this occurs).

Now note that

$$\sup_{x \in \mathsf{P}_\delta} \|\nabla \Phi(x)\|_{\ell_\infty(1/w)} \leqslant O\left(\log \frac{1}{\delta}\right). \tag{2.12}$$

Thus if we ensure that $x(t) \in \mathsf{P}_\delta$, then (2.9) implies that $x_r(T) = \delta$ occurs after some finite time T.

We are left to analyze how the potential changes when OPT moves. But from the definition and (2.12), we have

$$\partial_t \hat{D}_\Phi(y(t); x) = \langle \nabla \Phi(x), -\partial_t y(t) \rangle \leqslant O(\log \tfrac{1}{\delta}) \|\partial_t y(t)\|_{\ell_1(w)}\,.$$

Thus we obtain an $O(\log \frac{1}{\delta})$-competitive algorithm, where δ is the largest constant such that we can round (online) a fractional $\frac{k}{1-\delta}$-paging algorithm to a genuine fractional k-paging algorithm. As we will see now, this can be done when $\delta = \frac{1}{2k}$.

Transforming to a valid fractional paging algorithm. Consider a request sequence $\vec{r} = (r_1, r_2, \ldots, r_M)$, and a differentiable map $x : [0, T] \to \mathsf{P}_\delta$ that services $\vec{r}$ in the sense that there are times $t_1 < t_2 < \cdots < t_M$ such that $x_{r_i}(t_i) = \delta$.

Define

$$z(t) := \frac{1 - x(t)}{1 - \delta} \qquad \forall t \in [0, T] \,.$$

Then $z_{r_i}(t_i) = 1$, so z represents a trajectory on measures that services $\vec{r}$, but problematically we have $\|z(t)\|_1 = \frac{k}{1-\delta}$ for $t \in [0, T]$.

We fix this as follows: Let $\varepsilon := \frac{\delta k}{1-\delta}$ and define $\sigma : \mathbb{R}_+ \to \mathbb{R}_+$ so that $\sigma|_{[\ell, \ell+\varepsilon]} = \ell$ for every $\ell \in \mathbb{Z}_+$ and σ is extended affinely to the rest of $\mathbb{R}_+$. For $z \in \mathbb{R}^n_+$, define $\sigma(z) := (\sigma(z_1), \ldots, \sigma(z_n))$, and consider the trajectory $\sigma(z(t))$ for $t \in [0, T]$. Observe first that

$$\int_0^T \|\partial_t \sigma(z(t))\|_{\ell_1(w)} \, dt \leqslant \|\sigma\|_{\mathrm{lip}} \int_0^T \|\partial_t z(t)\|_{\ell_1(w)} \, dt \leqslant \frac{1}{1 - \varepsilon} \int_0^T \|\partial_t z(t)\|_{\ell_1(w)} \, dt \,.$$

Thus for $\delta = \frac{1}{2k}$, the movement cost has increased under σ by only an $O(1)$ factor.

Because σ is superadditive, it also holds that for every $t \in [0, T]$,

$$\sum_{i=1}^n \sigma(z_i(t)) \leqslant \sigma\left(\sum_{i=1}^n z_i(t)\right) = \sigma\,(k + \varepsilon) = k \,.$$

Therefore we use at most k fractional server mass at any point in time. We are left to show that, at no additional movement cost, this can be transformed into an algorithm that maintains fractional server mass exactly k.

To that end, we may assume that $\|\sigma(z(0))\|_1 = k$. It will be easiest to think of a weighted star metric on vertices $V = \{1, 2, \ldots, n\} \cup \{0\}$, where 0 is the center of the star and the edge $(0, i)$ has length w_i. When $\langle \partial_t \sigma(z(t)), e_i \rangle < 0$ for some $i \in [n]$, the instantaneous movement cost of $\sigma(z(t))$ in direction e_i is $-\langle \partial_t \sigma(z(t)), e_i \rangle$. Instead of deleting this mass, we can move it to 0 for the same cost. Similarly, when $\langle \partial_t \sigma(z(t)), e_i \rangle > 0$, the instantaneous movement cost in direction e_i is $\langle \partial_t \sigma(z(t)), e_i \rangle$. Instead of creating mass, we can move this mass from 0 to i for the same cost.

3 k-server on trees

Consider a rooted tree $\mathcal{T} = (V, E)$ with root $\mathbb{r} \in V$ and leaves $\mathcal{L} \subseteq V$. Let $\{w_v \geqslant 0 : v \in V\}$ be a collection of nonnegative weights on V with $w_{\mathbb{r}} = 0$. We will suppose that every leaf $\ell \in \mathcal{L}$ is at the same combinatorial distance from the root.

For $v \in V$, let $\mathcal{L}(v) \subseteq \mathcal{L}$ denote the set of leaves beneath v. For $u \in V \setminus \{\mathbb{r}\}$, let $p(u) \in V$ denote the parent of u, and write $\vec{E} = \{\vec{e} : e \in E\} = \{(p(u), u) : u \in V \setminus \{\mathbb{r}\}\}$ for the set of edges directed away from the root. For $(u, v) \in \vec{E}$, define $\mathrm{len}_w(u, v) := w_v$. Let $\mathrm{dist}_w(x, y)$ denote the weighted path distance between $x, y \in V$, where an edge $e \in E$ is given weight $\mathrm{len}_w(\vec{e})$. We say that the pair $(\mathcal{T}, w)$ is a *τ-adic HST* if for every $v \in V \setminus \{\mathbb{r}\}$, it holds that $w_v = \tau^j$ for some $j \in \mathbb{Z}$ and, moreover, if $(u, v) \in \vec{E}$ then $w_v = w_u / \tau$.

For $z \in \mathbb{R}^V$ and $w \in \mathbb{R}^V_+$, denote

$$\|z\|_{\ell_1(w)} := \sum_{v \in V} w_v \, |z_v| \,.$$

Leaf measures, internal measures, and supermeasures. A *leaf measure* is a point $z \in \mathbb{R}_+^{\mathcal{L}}$. The *mass* of a leaf measure is defined as the quantity $\sum_{\ell \in \mathcal{L}} z_\ell$. An *internal supermeasure* is a point $z \in \mathbb{R}_+^V$ such that

$$z_u \geqslant \sum_{v:(u,v)\in \vec{E}} z_v \qquad \forall u \in V \,. \tag{3.1}$$

The *mass* of an internal supermeasure is the quantity $z_{\mathbb{r}}$. We say that $z \in \mathbb{R}_+^V$ is an *internal measure* if (3.1) is satisfied with equality.

For a leaf measure $z \in \mathbb{R}_+^{\mathcal{L}}$, we define its *lifting to an internal measure* by

$$\hat{z}_v := \sum_{\ell \in \mathcal{L}(v)} z_\ell \qquad \forall v \in V \,.$$

Let $\widehat{M}$ denote the set of internal measures on V. It is straightforward to see that this is precisely the class of lifted leaf measures.

Lemma 3.1. *For leaf measures $y, z \in \mathbb{R}_+^{\mathcal{L}}$ with $\|y\|_1 = \|z\|_1$, it holds that*

$$\mathbb{W}_w^1(y,z) = \|\hat{y} - \hat{z}\|_{\ell_1(w)} \,.$$

A *fractional k-server algorithm for* $(\mathcal{L}, \mathrm{dist}_w)$ is an online sequence $\left\langle z^{(t)} \in \mathbb{R}_+^{\mathcal{L}} : t = 0,1,2,\ldots \right\rangle$ of leaf measures of mass k such that for every $t \geqslant 1$: $z^{(t)}_{\ell_t} \geqslant 1$ if ℓ_t is the requested leaf at time t. We also require that $z^{(0)}$ is integral. The *cost* of such an algorithm is defined by

$$\sum_{t\geqslant 0} \mathbb{W}_w^1\left(z^{(t)}, z^{(t+1)}\right) \,.$$

Lemma 3.2 ([BBMN15, §5.2])**.** *The following holds for all $\tau > 5$. If (X, d_X) is a τ-HST metric and there is a fractional online k-server algorithm for (X, d_X), then there is a randomized integral online k-server algorithm whose expected cost is at most $O(1)$ times larger.*

3.1 $k + \varepsilon$ fractional servers

For the remainder of the proof, we will work with continuous time trajectories $z : [0,T] \to \mathbb{R}_+^{\mathcal{L}}$ whose movement cost is measured by

$$\int_0^T \|\partial_t \hat{z}_t\|_{\ell_1(w)} \, dt \,,$$

in light of Lemma 3.1. Obviously such a trajectory can be mapped to a discrete-time algorithm by choosing times $T_1 \geqslant T_2 \geqslant \cdots$ that correspond to discrete times $t = 1,2,\ldots$.

Lemma 3.3. *If $y : [0,T] \to \mathbb{R}_+^V$ is a trajectory taking values in internal supermeasures of mass k, then there is an (adapted) trajectory $z : [0,T] \to \mathbb{R}_+^{\mathcal{L}}$ taking values in leaf measures of mass k such that:*

1. *For every leaf $\ell \in \mathcal{L}$: $z_t(\ell) \geqslant y_t(\ell)$, and*
2. *$\mathbb{W}_w^1(z(0), z(T)) \leqslant \int_0^T \|\partial_t y(t)\|_{\ell_1(w)} \, dt$.*

Proof. This lemma follows from a more general principle: If (X,d) is a metric space and $X' \subseteq X$, then an online (fractional) k-server algorithm on (X,d) servicing a sequence of requests in X' can be converted to an online (fractional) k-server algorithm on $(X', d|_{X'\times X'})$ without increasing the movement cost. This is a straightforward consequence of the triangle inequality.

Now observe that we can envision every trajectory on internal supermeasures $y(t) \in \mathbb{R}_+^V$ with $t \in [0,T]$ as a trajectory on genuine measures $\tilde{y}(t) \in \mathbb{R}_+^V$ defined by

$$\tilde{y}_u(t) := y_u(t) - \sum_{v:(u,v)\in\vec{E}} y_u(t) \qquad \forall u \in V\,.$$

And moreover,

$$\mathbb{W}_w^1(\tilde{y}(0), \tilde{y}(T)) \leqslant \int_0^T \|\partial_t y(t)\|_{\ell_1(w)}\,.$$

Since $(\mathcal{L}, \mathrm{dist}_w)$ is a subspace of (V, dist_w), this completes the proof by our earlier observation. □

Lemma 3.4. *For every $0 \leqslant \varepsilon < 1$, a fractional online $(k+\varepsilon)$-server algorithm on $(\mathcal{L}, \mathrm{dist}_w)$ can be converted to a fractional online k-server algorithm so that the movement cost increases by a factor of at most $\frac{1}{1-\varepsilon}$.*

Proof. The proof is similar to the case for fractional paging. Define $\sigma : \mathbb{R}_+ \to \mathbb{R}_+$ so that $\sigma|_{[\ell,\ell+\varepsilon]} = \ell$ for every $\ell \in \mathbb{Z}_+$ and σ is extended affinely to the rest of $\mathbb{R}_+$. For $y \in \mathbb{R}_+^V$, define $\sigma(y) := (\sigma(y_v))_{v\in V}$.

Consider a trajectory $z : [0,T] \to \mathbb{R}_+^{\mathcal{L}}$ taking values in leaf measures of mass $k+\varepsilon$. Then since σ is superadditive, it holds that $\sigma(\hat{z}(t))$ is an internal supermeasure for every $t \in [0,T]$. Moreover, $\sigma(\hat{z}_{\mathbb{r}}(t)) = \sigma(k+\varepsilon) = k$, so $\sigma(\hat{z}(t))$ is an internal supermeasure of mass k.

Finally, note that

$$\int_0^T \|\partial_t \sigma(\hat{z}(t))\|_{\ell_1(w)}\,dt \leqslant \|\sigma\|_{\mathrm{lip}} \int_0^T \|\partial_t \hat{z}(t)\|_{\ell_1(w)}\,dt \leqslant \frac{1}{1-\varepsilon}\int_0^T \|\partial_t \hat{z}(t)\|_{\ell_1(w)}\,dt\,.$$

Now applying Lemma 3.3 to the internal supermeasures $\sigma(\hat{z}(t))$ completes the proof. □

In light of the preceding lemma, it suffices to construct a competitive fractional $(k+\varepsilon)$-server algorithm with $\varepsilon < 1$ for any request sequence on $(\mathcal{L}, \mathrm{dist}_w)$.

3.2 The allocation polytope and multiscale entropy

For $u \in V$, write

$$\chi(u) := \left\{(v,j) : (u,v) \in \vec{E},\, j \in [N_u]\right\}$$

where N_u is the number of leaves in the subtree rooted at u. Denote

$$\Lambda := \left\{(\mathbb{r}, i) : i \in [N_{\mathbb{r}}]\right\} \cup \bigcup_{u\in V} \chi(u)\,.$$

With a slight abuse of notation, we sometimes write $\sum_{i\geqslant 1} f(x_{u,i})$ instead of $\sum_{i\in[N_u]} f(x_{u,i})$. The *allocation polytope on $\mathcal{T}$* is defined by

$$\mathsf{A} := \Bigg\{x \in [0,1]^\Lambda : \sum_{i\leqslant|S|} x_{u,i} \leqslant \sum_{(v,j)\in S} x_{v,j} \quad \forall u \in V, S \subseteq \chi(u)\,,$$
$$x_{\mathbb{r},i} = \mathbb{1}_{\{i>k\}} \quad \forall i \geqslant 1\Bigg\}\,.$$

Remark 3.5 (Interpretation of A)**.** In order to give an intuitive meaning to the variables of A, it makes sense to first consider their complements: $y_{u,i} = 1 - x_{u,i}$. The variable $y_{u,i}$ can be interpreted as "the probability there are at least i servers in the subtree rooted at u."

If $\{y_{u,i} : i \geqslant 1\}$ corresponds to a distribution over k-server configurations, then we should clearly have the constraints $y_{u,1} \geqslant y_{u,2} \geqslant \cdots$. We could add those constraints to A, but it turns out they would be superfluous (i.e., there would exist a dynamics for which they are never active). See Lemma 3.8 below.

Moreover, there is a natural set of constraints between $\{y_{u,i} : i \geqslant 1\}$ and the child variables $\{y_{v,j} : (v,j) \in \chi(u)\}$ which enforces that some joint distribution on k-server configurations is consistent with the variables. This is naturally phrased as a bipartite flow problem between the u-variables and the v-variables as v ranges over children of u.

The constraints in the definition of A correspond to one side of those constraints: That there is enough flow from the u-variables to the v-variables. But they allow for the possibility that there is an excess of flow (i.e., there are possibly more fractional servers at u than the sum of servers at the children). The other side of the constraints (which ensure there is no excess) is enforced implicitly by the dynamics. See Lemma 3.10.

Even if all the constraints were present, a point of A does *not* represent a distribution over k-server configurations in $\mathcal{T}$. For instance, if u, u' are siblings in $\mathcal{T}$, then the "covariances" between the children of u and u' are completely unspecified.

Example 3.6. An example helps in understanding the utility of the allocation variables. We will use the complement variables $\{y_{u,i}\}$ from Remark 3.5. Consider a node u with three children a, b, c, and suppose that $y_a = y_b = y_c = (\frac{1}{3}, 0, 0, \ldots, 0)$. Let (k_a, k_b, k_c) denote the number of servers at a, b, c. We now examine three situations and how they encode the joint distribution of (k_a, k_b, k_c).

1. $y_u = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \ldots, 0)$

   This corresponds to $(k_a, k_b, k_c) = (1,1,1)$ with probability 1/3, and $(k_a, k_b, k_c) = (0,0,0)$ otherwise.

2. $y_u = (1, 0, \ldots, 0)$

   This corresponds to (k_a, k_b, k_c) being each of $(1,0,0), (0,1,0), (0,0,1)$ with equal probability.

3. $y_u = \big(1-(1-p)^3, \frac{3}{2}p^2(1-p)+p^3, p^3, 0, \ldots, 0\big)$ where $p = \frac{1}{3}$.

   This corresponds to k_a, k_b, k_c being i.i.d. random variables, each equal to one with probability 1/3.

The relevance to k-server is as follows: Situation (2) is a strong signal that there should always be at least one server in T_u, the subtree rooted at u. If there is only one server there, we will be reluctant to use that server to service requests elsewhere. Situation (1), on the other hand, suggests that three servers are needed in T_u, but only 1/3 of the time. Thus we will be more likely to "borrow" server mass sitting there, knowing that we can send it back when it is needed again.

Fix some $\delta > 0$ and define the shifted multiscale entropy by

$$\Phi(x) := \sum_{u \in V} w_u \sum_{i \geqslant 1} (x_{u,i} + \delta) \log(x_{u,i} + \delta) \,.$$

For $\ell \in \mathcal{L}$, denote $x_\ell := x_{\ell,1}$. Finally, for $u \in V$ and $x \in \mathsf{A}$, define the *associated server measure* $z \in \mathbb{R}_+^V$ by

$$z_u := \frac{1}{1-\delta} \sum_{i \geqslant 1} (1 - x_{u,i}) \,.$$

Suppose we receive a request at $\ell \in \mathcal{L}$. Let $x : [0, \infty) \to \mathsf{A}$ be an absolutely continuous trajectory satisfying

$$\partial_t x(t) = \nabla^2 \Phi(x(t))^{-1} (-e_{\ell,1} - \lambda(t)) \tag{3.2}$$

with $\lambda(t) \in N_{\mathsf{A}}(x(t))$ for almost every $t \geqslant 0$. Denote

$$T = \inf\{t \geqslant 0 : x_\ell(t) < \delta\} \,.$$

By construction, as long as $T < \infty$ (see Lemma 3.14 below), it holds that $\{z_u(t) : t \in [0, T]\}$ is a fractional $\frac{k}{1-\delta}$-server trajectory that services the request at ℓ. Now set $\delta := \frac{1}{2k}$ so that $\frac{k}{1-\delta} < k+1$ for $k \geqslant 2$.

3.3 Dynamics

We now describe in detail the dynamics of $x(t)$ on $[0, T]$. We allow (momentarily) for the possibility that $T = +\infty$. Our main goal is to establish that $\partial_t z(t)$ is a flow from the non-requested leaves to the requested leaf and that $T < \infty$.

Lemma 3.7. *Suppose that $x_{\ell,1}(0) > \delta$. The continuous trajectory $x(t)$ defined in (3.2) exists uniquely for $t \in [0, T]$ and satisfies $x_{\ell,1}(t) \geqslant \delta$ for all $t \in [0, T]$. Furthermore, $x(t)$ is absolutely continuous and its derivative is given by*

$$\partial_t x_{u,i}(t) = \frac{x_{u,i}(t) + \delta}{w_u} \left(-\mathbb{1}_{\{(u,i)=(\ell,1)\}} - \lambda_{u,i}(t) \right) , \tag{3.3}$$

for all $u \in V \setminus \{\mathbb{r}\}$ and all $i \geqslant 1$, and

$$\lambda_{u,i}(t) = \sum_{\substack{S \subseteq \chi(u): \\ i \leqslant |S|}} \lambda_S(t) - \sum_{\substack{S \subseteq \chi(p(u)): \\ (u,i) \in S}} \lambda_S(t) , \tag{3.4}$$

where $\lambda_S(t) \geqslant 0$ are the Lagrangian multipliers for the constraints $\{\sum_{i \leqslant |S|} x_{u,i}(t) \leqslant \sum_{(v,j) \in S} x_{v,j}(t)\}$. Also, we have that

$$\lambda_S(t) > 0 \implies \sum_{i \leqslant |S|} x_{u,i}(t) = \sum_{(v,j) \in S} x_{v,j}(t) . \tag{3.5}$$

Remark. We will establish that $T < \infty$ in Lemma 3.14.

Proof. Since the allocation polytope A is compact and convex, Φ is strongly convex and smooth, the existence and the uniqueness of the path $x(t)$ defined in (3.2) follows from Theorem 5.7 with $f(t, x) = -e_{\ell,1}$. In particular, using the formula for Φ, we have that

$$\partial_t x(t) = \frac{x_{u,i}(t) + \delta}{w_u} (-e_{\ell,1} - \lambda(t)) \tag{3.6}$$

with $\lambda(t) \in N_{\mathsf{A}}(x(t))$.

To calculate $N_{\mathsf{A}}(x(t))$, we note that the constraints $\{x_{u,i}(t) \geqslant 0\}$ are redundant, as they can be expressed by the constraints $\{\sum_{i \leqslant |S|} x_{u,i}(t) \leqslant \sum_{(v,j) \in S} x_{v,j}(t)\}$ and $x_{\mathbb{r},1}(t) = 0$ using the sequence of singleton sets $S = (u, i), (p(u), 1), (p(p(u)), 1), \cdots$. Hence, we can ignore the constraints $\{x_{u,i}(t) \geqslant 0\}$ from the polytope.

Using the definition of the allocation polytope A, Lemma 2.1 asserts that

$$\begin{aligned} & N_{\mathsf{A}}(x(t)) \\ & = \Bigg\{ \sum_{u \in V, S \subseteq \chi(u)} \lambda_S(t) \left(\sum_{i \leqslant |S|} e_{u,i} - \sum_{(v,j) \in S} e_{v,j} \right) + \sum_{i \geqslant 1} \mu_i(t) e_{\mathbb{r},i} + \sum_{u \in V, i \geqslant 1} \eta_{u,i}(t) e_{u,i} \\ & \qquad \text{where } \lambda_S(t), \eta_{u,i}(t) \geqslant 0,\ \mu_i(t) \in \mathbb{R}, \eta_{u,i}(t) \cdot (1 - x_{u,i}(t)) = 0, \\ & \qquad \text{and } \lambda_S(t) \cdot \left(\sum_{i \leqslant |S|} x_{u,i}(t) - \sum_{(v,j) \in S} x_{v,j}(t) \right) = 0 \Bigg\} . \end{aligned}$$

Rearranging the terms in (3.6) and $N_{\mathsf{A}}(x(t))$, and ignoring the terms for the root $\mathbb{r}$, we see that it only remains to show that one can take $\eta_{u,i}(t)$ (the Lagrange multiplier for $\{x_{u,i}(t) \leqslant 1\}$) to be 0.

Let $\widetilde{\mathsf{A}}$ be the polytope just as A, except with $[0,1]$ replaced by $[0,2]$. Assume now that $x(t)$ is defined with A replaced by $\widetilde{\mathsf{A}}$. We will show that one has $x_{u,i}(t) \leqslant 1$. This implies that the Lagrange multipliers for the path defined on $\widetilde{\mathsf{A}}$ are valid Lagrange multipliers for the path on A, and in particular one can take $\eta_{u,i}(t) = 0$.

Toward deriving a contradiction, let us assume that there exists a time $t > 0$, $u \in V_h$, and $i \geqslant 1$ such that $x_{u,i}(t) > 1$ and $\partial_t x_{u,i}(t) > 0$. We prove by induction on h that this impossible.

For $h = 0$ this follows from the equality constraints at the root. Now consider $h \geqslant 1$ and observe that by (3.3) and (3.4), one must have $\lambda_S(t) \neq 0$ for some $S \subseteq \chi(p(u))$, which means that $\sum_{i \leqslant |S|} x_{p(u),i}(t) = \sum_{(v,j)\in S} x_{v,j}(t)$. However, the induction hypothesis implies that for any $j \geqslant 1$, $x_{p(u),j}(t) \leqslant 1$, and thus the constraint corresponding to $S \setminus \{(u,i)\}$ is violated for $x(t)$, yielding a a contradiction. $\square$

We now prove several lemmas giving a more refined understanding of the dynamics (3.3). The reader is encouraged to skip these arguments upon a first reading. The main technical property we need to establish is that $z(t) \in \widehat{M}$ for all times $t \in [0,T]$, i.e., the mass per level remains constant. This is proved in Lemma 3.10.

For $h \geqslant 0$, let V_h denote the set of vertices with a simple path to the root using h edges. Define $C(t) \supseteq \{S : \lambda_S(t) \neq 0\}$ to be the set of *active constraints:*

$$C(t) := \left\{ S \subseteq \chi(u) : u \in V \text{ and } \sum_{i \leqslant |S|} x_{u,i}(t) = \sum_{(v,j)\in S} x_{v,j}(t) \right\}.$$

Lemma 3.8. *For any $t \geqslant 0$, $u \in V$, and $i \geqslant j \geqslant 1$, it holds that $x_{u,i}(t) \geqslant x_{u,j}(t)$.*

Proof. We will show that $x_{u,i}(t) > x_{u,i+1}(t)$ implies $\partial_t x_{u,i}(t) \leqslant \partial_t x_{u,i+1}(t)$. Recalling (3.3) and (3.4), it is enough to show that

$$\sum_{\substack{S \subseteq \chi(p(u)):\\ (u,i+1)\in S}} \lambda_S(t) \geqslant \sum_{\substack{S \subseteq \chi(p(u)):\\ (u,i)\in S}} \lambda_S(t)\,.$$

Let us show that if $(u,i) \in S$ and $(u,i+1) \notin S$ then $\lambda_S(t) = 0$, yielding the desired conclusion.

Using the constraint for $S \cup \{(u,i+1)\} \setminus \{(u,i)\}$ gives

$$\sum_{(v,j)\in S} x_{v,j}(t) > \sum_{(v,j)\in S\cup\{(u,i+1)\}\setminus\{(u,i)\}} x_{v,j}(t) \geqslant \sum_{i \leqslant |S|} x_{p(u),i}(t)\,,$$

implying that $\lambda_S(t) = 0$. $\square$

Lemma 3.9. *Consider $u \in V$ and $S, S' \subseteq \chi(u)$ such that $S, S' \in C(t)$. Then $S \cup S' \in C(t)$ as well.*

Proof. First we claim that $M_S := \max_{(v,j)\in S} x_{v,j}(t)$ and $M_{S'} := \max_{(v,j)\in S'} x_{v,j}(t)$ are equal. Let (v_*, j_*) denote some pair for which $x_{v,j}(t)$ is maximal among $(v,j) \in S$. If $M_S > M_{S'}$, then for any $(v', j') \in S'$, the constraint corresponding to $S \cup \{(v',j')\} \setminus \{(v_*, j_*)\}$ is violated because $S, S' \in C(t)$. Let us denote $M := M_S = M_{S'}$.

Suppose that $|S| \geqslant |S'|$. The same argument shows that for any $(v,j) \in S \setminus S'$, one has $x_{v,j}(t) = M$. Using the constraint for $S \setminus \{(v_*, j_*)\}$ and the fact that $S \in C(t)$ shows that $x_{u,|S|}(t) \geqslant M$, and thus by Lemma 3.8, one has $x_{u,|S|+m}(t) \geqslant M$ for any $m \geqslant 0$. This implies:

$$\sum_{i \leqslant |S\cup S'|} x_{u,i}(t) \geqslant \sum_{i \leqslant |S|} x_{u,i}(t) + M \cdot |S \setminus S'| = \sum_{(v,j)\in S} x_{v,j}(t) + M \cdot |S \setminus S'| = \sum_{(v,j)\in S\cup S'} x_{v,j}(t)\,.$$

Furthermore, since $x(t) \in \mathsf{A}$ it also holds that $\sum_{i \leqslant |S \cup S'|} x_{u,i}(t) \leqslant \sum_{(v,j) \in S \cup S'} x_{v,j}(t)$, demonstrating that $S \cup S' \in \mathcal{C}(t)$. □

Lemma 3.10. *Suppose that $z(0) \in \widehat{M}$. Then $z(t) \in \widehat{M}$ for $t \geqslant 0$.*

Proof. For $u \in V$, let $S_u(t)$ be the maximum (w.r.t. inclusion) active set at time t in $\chi(u)$ (cf. Lemma 3.9). Since $\partial_t x(t) \in N_{\mathsf{A}}(x(t))^{\perp}$ (Lemma 5.8), one has

$$\sum_{i \leqslant |S_u(t)|} \partial_t x_{u,i}(t) = \sum_{(v,j) \in S_u(t)} \partial_t x_{v,j}(t) \,,$$

which in turns gives, for any $h \geqslant 1$,

$$\sum_{u \in V_h, i \leqslant |S_u(t)|} \partial_t x_{u,i}(t) = \sum_{(v,j) \in \cup_{u \in V_h} S_u(t)} \partial_t x_{v,j}(t) \,. \tag{3.7}$$

Thus to compare the derivatives of the mass at two adjacent levels, it remains to establish that

$$\sum_{u \in V_h, i > |S_u(t)|} \partial_t x_{u,i}(t) \geqslant \sum_{(v,j) \notin \cup_{u \in V_h} S_u(t) : v \in V_{h+1}} \partial_t x_{v,j}(t) \,.$$

We show that every term in the first sum is nonnegative and every term in the second sum is nonpositive. In particular, since $\partial_t x_{\mathbb{r},i}(t) = 0$ for all $i \geqslant 1$, this will imply by induction that

$$\sum_{v \in V_{h+1}, j \geqslant 1} \partial_t x_{v,j}(t) \leqslant \sum_{u \in V_h, i \geqslant 1} \partial_t x_{u,i}(t) \leqslant 0 \,,$$

yielding $\sum_{u \in V_h} \partial_t z_u(t) \geqslant 0$. Then the proof is concluded using $z(0) \in \widehat{M}$ and $z(t) \in \mathsf{A}$ for all $t \geqslant 0$.

Thus it remains to show that $\partial_t x_{u,i}(t) \geqslant 0$ for all $i > |S_u|$ and $u \notin \mathcal{L}$, and $\partial_t x_{v,j}(t) \leqslant 0$ for all $(v, j) \notin \cup_{u \in V_h} S_u(t)$. For $u = \mathbb{r}$, one has $\partial_t x_{u,i}(t) = 0$, and for $u \neq r$ we have thanks to (3.3) and Lemma 3.7:

$$\partial_t x_{u,i}(t) = -\frac{x_{u,i}(t) + \delta}{w_u} \lambda_{u,i}(t) \,.$$

Since $i > |S_u(t)|$ and $S_u(t)$ is the maximum active set in $\chi(u)$, it holds that $\lambda_S(t) = 0$ for any $S \subseteq \chi(u)$ with $|S| \geqslant i$. Thus from (3.4), we see that $\lambda_{u,i}(t) \leqslant 0$, and in turn $\partial_t x_{u,i}(t) \geqslant 0$. On the other hand for $(v, j) \in \chi(u)$ one has

$$\partial_t x_{v,j}(t) \leqslant -\frac{x_{v,j}(t) + \delta}{w_v} \lambda_{v,j}(t) \,.$$

Assume $(v, j) \notin S_u(t)$. Then since $S_u(t)$ is the maximum active set in $\chi(u)$, it holds that $\lambda_S(t) = 0$ for any S with $(v, j) \in S$. Using (3.4), we see that $\lambda_{v,j}(t) \geqslant 0$, concluding the proof □

We have established that $z(t)$ is an internal measure for every $t \geqslant 0$, and thus $\partial_t z(t)$ is a flow. We now we show that $\partial_t z(t)$ is a flow directed toward the request ℓ.

Lemma 3.11. *It holds that $\partial_t z_u(t) \geqslant 0$ if u is an ancestor of the request ℓ, and $\partial_t z_u(t) \leqslant 0$ otherwise.*

Proof. Since $z(t) \in \widehat{M}$ (Lemma 3.10), it suffices to show that for any leaf $\ell' \neq \ell$, one has $\partial_t z_{\ell'}(t) \leqslant 0$. Indeed by preservation of mass at every node (i.e., $\partial_t z_u(t) = \sum_{v:(u,v) \in \vec{E}} \partial_t z_v(t)$), this implies that $\partial_t z_u(t) \leqslant 0$ for any u which is not an ancestor of ℓ. Furthermore, by preservation of mass per level (i.e., $\sum_{u \in V_h} \partial_t z_u(t) = 0$), and the fact that there is a single ancestor of ℓ per level, this also gives $\partial_t z_u(t) \geqslant 0$ for any ancestor u of ℓ.

Notice that $\partial_t z_{\ell'}(t) = -\frac{1}{1-\delta} \sum_{i \geqslant 1} \partial_t x_{\ell',i}(t)$, and thus it suffices to show that for any $i \geqslant 1$, $\partial_t x_{\ell',i}(t) \geqslant 0$. The latter inequality is straightforward from (3.3) and (3.4) since $\chi(\ell') = \emptyset$. □

The next lemma follows immediately from Lemma 3.11 since $\partial_t x_{\ell'}(t) \geqslant 0$ for all $\ell' \neq \ell$.

Lemma 3.12. *If $x_{\ell'}(0) \geqslant \delta$ for all $\ell' \in \mathcal{L}$, then $x_{\ell'}(T) \geqslant \delta$ for all $\ell' \in \mathcal{L}$.*

Let us extend the definition of $\|\cdot\|_{\ell_1(w)}$ to $x \in \mathbb{R}^\Lambda$ by

$$\|x\|_{\ell_1(w)} := \sum_{u \in V} w_u \sum_{i \geqslant 1} |x_{u,i}| \,.$$

Observe that

$$\sup_{x \in \mathsf{A}} \|\nabla \Phi(x)\|_{\ell_\infty(1/w)} \leqslant O(\log \tfrac{1}{\delta}) \,.$$

This yields the following.

Lemma 3.13. *It holds that for every $x \in \mathsf{A}$ and $\{y(t)\} \subseteq \mathsf{A}$ differentiable:*

$$\partial_t \hat{D}_\Phi(y(t); x) \leqslant \|\partial_t y(t)\|_{\ell_1(w)} O(\log \tfrac{1}{\delta}) \,.$$

The next lemma is an immediate consequence of (2.5).

Lemma 3.14. *If $y \in \mathsf{A}$ satisfies $y_\ell = 0$, then*

$$\partial_t \hat{D}_\Phi(y; x(t)) \leqslant -x_{\ell,1}(t) \,.$$

In particular, we have that $T < \infty$ and hence $x_{\ell,1}(T) = \delta$.

Proof. The first conclusion follows from (2.5) using $y_\ell = 0$ and $f(t, x(t)) = -e_{\ell,1}$. Since the divergence is nonnegative and it is decreasing with rate at least δ whenever $x_{\ell,1}(t) \geqslant \delta$, the trajectory ends in finite time. □

3.4 The weighted depth potential

We now define an auxiliary potential function Ψ_t. We relate it to the dynamics, and then present applications in Section 3.4.1–Section 3.4.3, culminating in the assertion that our algorithm is $O((\log k)^2)$-competitive.

Consider a differentiable function $\Psi(t)$ such that

$$\partial_t \Psi(t) = \sum_{u \in V \setminus \{\mathbb{r}\}} w_u \left(\Delta_u(t) + \Delta_{p(u)}(t)\right) \sum_{i \geqslant 1} \partial_t x_{u,i}(t) \tag{3.8}$$

for some functions $\{\Delta_u(t) \geqslant 0 : u \in V\}$ satisfying $\Delta_u(t) \leqslant \Delta_v(t)$ for all $(u, v) \in \vec{E}$ as well as $\Delta_{\mathbb{r}}(t) \equiv 0$. Note that an important special case is simply when

$$\Psi(t) = \sum_{u \in V \setminus \{\mathbb{r}\}} w_u (\Delta_u + \Delta_{p(u)}) \sum_{i \geqslant 1} x_{u,i}(t)$$

where $\{\Delta_u : u \in V\}$ are independent of t. For an edge $(u, v) \in \vec{E}$, define

$$q_t(v) := \Delta_v(t) - \Delta_u(t) \,. \tag{3.9}$$

Lemma 3.15. *The following holds for almost every $t \in [0, T]$:*

$$\|\partial_t x(t)\|_{\ell_1(q_t w)} \leqslant 3\Delta_\ell(t)(x_\ell(t) + \delta) + \partial_t \Psi(t) \,. \tag{3.10}$$

Before proving the lemma, let us explain the intuition. We are bounding, not the true movement cost, but the movement cost with weights given by $q_t w$. We know that the requested leaf ℓ is the unique sink for server flow, and by (3.3), the instantaneous w-weighted cost of the flow entering ℓ is at most $x_\ell(t) + \delta$. Therefore the $q_t w$-weighted cost of server mass "descending to ℓ" is bounded by $O(\Delta_\ell(t)(x_\ell(t) + \delta))$, and this accounts for the first term in (3.10). The rest of the movement is due to mass moving "up" the tree (toward the root), and this is accounted for by $\partial_t \Psi(t)$, since Ψ can be thought of as the average weighted depth of the server mass (plus some additive constant). The proof is now a technical verification of this description.

Proof of Lemma 3.15. Note that from Lemma 3.7, for every $u \in V \setminus \{\mathbb{r}\}$, we have:

$$\partial_t x_{u,i}(t) = \frac{x_{u,i}(t) + \delta}{w_u} \left(-\mathbb{1}_{\{(u,i)=(\ell,1)\}} - \lambda_{u,i}(t)\right), \tag{3.11}$$

where

$$\lambda_{u,i}(t) = \sum_{\substack{S \subseteq \chi(u):\\ i \leqslant |S|}} \lambda_S(t) - \sum_{\substack{S \subseteq \chi(p(u)):\\ (u,i) \in S}} \lambda_S(t). \tag{3.12}$$

Recalling that $\Delta_{\mathbb{r}}(t) = 0$ for all $t \in [0, T]$, we calculate:

$$\begin{aligned}
&\partial_t \Psi(t) + 2\Delta_\ell(t)(x_\ell(t) + \delta) \\
&\quad \geqslant \partial_t \Psi(t) + \left(\Delta_\ell(t) + \Delta_{p(\ell)}(t)\right)(x_\ell(t) + \delta) \\
&\quad = \left(\Delta_\ell(t) + \Delta_{p(\ell)}(t)\right)(x_\ell(t) + \delta) + \sum_{u \in V \setminus \{\mathbb{r}\}} w_u \left(\Delta_u(t) + \Delta_{p(u)}(t)\right) \sum_{i \geqslant 1} \partial_t x_{u,i}(t) \\
&\quad = \sum_{u \in V \setminus \{\mathbb{r}\}} \sum_{S \subseteq \chi(u)} \lambda_S(t) \left(\sum_{(v,j) \in S} \left(\Delta_v(t) + \Delta_{p(v)}(t)\right)(x_{v,j}(t) + \delta) - \left(\Delta_u(t) + \Delta_{p(u)}(t)\right) \sum_{i \leqslant |S|} (x_{u,i}(t) + \delta) \right) \\
&\quad\qquad + \sum_{S \subseteq \chi(\mathbb{r})} \lambda_S(t) \sum_{(v,j) \in S} \Delta_v(t)(x_{v,j}(t) + \delta) \\
&\quad = \sum_{u \in V \setminus \{\mathbb{r}\}} \sum_{S \subseteq \chi(u)} \lambda_S(t) \left(\sum_{(v,j) \in S} \left[(\Delta_v(t) - \Delta_u(t)) + \left(\Delta_{p(v)}(t) - \Delta_{p(u)}(t)\right)\right](x_{v,j}(t) + \delta) \right) \\
&\quad\qquad + \sum_{S \subseteq \chi(\mathbb{r})} \lambda_S(t) \sum_{(v,j) \in S} \Delta_v(t)(x_{v,j}(t) + \delta) \\
&\quad = \sum_{u \in V \setminus \{\mathbb{r}\}} \sum_{S \subseteq \chi(u)} \lambda_S(t) \left(\sum_{(v,j) \in S} \left(q_t(v) + q_t(u)\right)(x_{v,j}(t) + \delta) \right) \\
&\quad\qquad + \sum_{S \subseteq \chi(\mathbb{r})} \lambda_S(t) \sum_{(v,j) \in S} \Delta_v(t)(x_{v,j}(t) + \delta),
\end{aligned}$$

where in the penultimate equality we have used that for $S \subseteq \chi(u)$,

$$\lambda_S(t) > 0 \implies \sum_{i \leqslant |S|} x_{u,i}(t) = \sum_{(v,j) \in S} x_{v,j}(t). \tag{3.13}$$

On the other hand, the $q_t \cdot w$-movement cost is equal to

$$\sum_{u \in V \setminus \{\mathbb{r}\}} \sum_{i \geqslant 1} q_t(u) w_u \left|\partial_t x_{u,i}(t)\right| = \sum_{u \in V \setminus \{\mathbb{r}\}} \sum_{i \geqslant 1} q_t(u) w_u \left(\partial_t x_{u,i}(t)\right)_+ - \sum_{u \in V \setminus \{\mathbb{r}\}} q_t(u) w_u \left(\partial_t x_{u,i}(t)\right)_- .$$

Using (3.11) and (3.12) gives

$$\sum_{u \in V \setminus \{\mathbb{r}\}} \sum_{i \geqslant 1} w_u q_t(u) \left(\partial_t x_{u,i}(t)\right)_+ \leqslant \sum_{u \in V} \sum_{S \subseteq \chi(u)} \lambda_S(t) \sum_{(v,j) \in S} q_t(v)(x_{v,j}(t) + \delta)$$

and

$$-\sum_{u \in V \setminus \{\mathbb{r}\}} \sum_{i \geqslant 1} w_u q_t(u) \left(\partial_t x_{u,i}(t)\right)_- \leqslant q_t(\ell)(x_\ell(t) + \delta) + \sum_{u \in V \setminus \{\mathbb{r}\}} \sum_{S \subseteq \chi(u)} \lambda_S(t) q_t(u) \sum_{i \leqslant |S|} (x_{u,i}(t) + \delta)$$

Using $q_t(\ell) \leqslant \Delta_t(\ell)$ and (3.13) again (as well as $q_t(v) = \Delta_v(t)$ for $(v, j) \in S \subseteq \chi(\mathbb{r})$), we have

$$\sum_{u \in V \setminus \{\mathbb{r}\}} \sum_{i \geqslant 1} q_t(u) w_u \left|\partial_t x_{u,i}(t)\right| \leqslant \Delta_t(\ell)(x_\ell(t) + \delta) + \sum_{u \in V \setminus \{\mathbb{r}\}} \sum_{S \subseteq \chi(u)} \lambda_S(t) \sum_{(v,j) \in S} \big(q_t(u) + q_t(v)\big)(x_{v,j}(t) + \delta)$$
$$+ \sum_{S \subseteq \chi(\mathbb{r})} \lambda_S(t) \sum_{(v,j) \in S} \Delta_v(t)(x_{v,j}(t) + \delta),$$

yielding the desired result. ◻

Note that $\|\partial_t x(t)\|_{\ell_1(q_t w)} = (1 - \delta)\|\partial_t z(t)\|_{\ell_1(q_t w)}$. Thus combining Lemma 3.15 with Lemma 3.14 and using $x_\ell(t) \geqslant \delta$ for $t \in [0, T]$ yields the following.

Corollary 3.16. *For almost every $t \in [0, T]$, if $y \in \mathsf{A}$ satisfies $y_\ell = 0$, then*

$$(1 - \delta) \|\partial_t z(t)\|_{\ell_1(q_t w)} \leqslant 3\Delta_\ell(t)(x_\ell(t) + \delta) + \partial_t \Psi(t) \leqslant \partial_t \Psi(t) - 6\Delta_\ell(t)\partial_t \hat{D}_\Phi(y; x(t)).$$

For a function $f : V \to \mathbb{R}_+$, define $\widehat{f} : V \setminus \mathcal{L} \to \mathbb{R}_+$ by

$$\widehat{f}(u) := \min\left\{f(v) + f(v') : v \neq v', (u, v), (u, v') \in \vec{E}\right\}.$$

For concreteness, let us define $\widehat{f}(u) := 0$ if u has only one child.

Lemma 3.17. *For almost every $t \in [0, T]$, the following holds. Suppose that $(\mathcal{T}, w)$ is a τ-adic HST for $\tau \geqslant 2$, and there is some $c > 0$ such that*

$$\widehat{q_t}(v) \geqslant c \qquad \forall v \in V \setminus \mathcal{L}.$$

If $y \in \mathsf{A}$ satisfies $y_\ell = 0$, then

$$\frac{c(1 - \delta)}{4} \|\partial_t z(t)\|_{\ell_1(w)} \leqslant \partial_t \Psi(t) - 6\Delta_\ell(t)\partial_t \hat{D}_\Phi(y; x(t)).$$

Proof. From Lemma 3.10, it holds that $z(t)$ is an internal measure for all $t \in [0, T]$, and moreover $\partial_t z(t)$ is a flow towards ℓ (cf. Lemma 3.11). Therefore we can decompose

$$\partial_t z(t) = \sum_{\ell' \in \mathcal{L}} y^{(\ell')}(t),$$

where $y^{(\ell')}(t)$ is a flow on the unique ℓ'-ℓ path in $\mathcal{T}$.

Let us use $|y^{(\ell')}(t)|$ to denote the magnitude of the corresponding flow. Since $\partial_t z(t)$ is a flow towards ℓ, we have

$$\begin{aligned}\|\partial_t z(t)\|_{\ell_1(q_t w)} &\geqslant \frac{1}{\tau} \sum_{\ell' \in \mathcal{L} \setminus \{\ell\}} \widehat{q_t}(\mathsf{lca}(\ell, \ell')) w_{\mathsf{lca}(\ell,\ell')} \left|y^{(\ell')}(t)\right| \\ &\geqslant \frac{c}{\tau} \sum_{\ell' \in \mathcal{L} \setminus \{\ell\}} w_{\mathsf{lca}(\ell,\ell')} \left|y^{(\ell')}(t)\right| \geqslant \frac{c}{4} \|\partial_t z(t)\|_{\ell_1(w)},\end{aligned} \tag{3.14}$$

where the first and last inequality use the fact that w is τ-adic. ◻

3.4.1 Combinatorial depth for general trees

Let $\mathrm{dist}_{\mathcal{T}}$ denote the unweighted shortest-path metric on $\mathcal{T}$. Define $\Delta_u(t) := \Delta_u = \mathrm{dist}_{\mathcal{T}}(\mathbb{r}, u)$, and

$$\Psi(t) := \sum_{u \in V \setminus \{\mathbb{r}\}} w_u \left(\Delta_u + \Delta_{p(u)}\right) \sum_{i \geqslant 1} x_{u,i}(t)\,,$$

Note that q_t (cf. (3.9)) satisfies $q_t \equiv 1$ for all $t \in [0, T]$. Therefore applying Corollary 3.16 yields

$$(1-\delta) \|\partial_t z(t)\|_{\ell_1(w)} \leqslant \partial_t \Psi(t) - 6\, \mathrm{dist}_{\mathcal{T}}(\mathbb{r}, \ell) \cdot \partial_t \hat{D}_\Phi(y; x(t))\,. \tag{3.15}$$

Combined with Lemma 3.13, this gives the following result.

Corollary 3.18. *For any tree metric with combinatorial depth D, there is an $O(D \log k)$-competitive* fractional *k-server algorithm.*

Proof. Consider any trajectory $\{y(t) \in \mathbb{R}_+^V : t \in [0, T]\}$ where $y(t)$ services a request sequence $\boldsymbol{\sigma}$, and such that $y(t)$ is almost surely an internal server measure for all $t \in [0, T]$. Let $z(t)$ denote our algorithm for servicing $\boldsymbol{\sigma}$. Combining Lemma 3.13 and (3.15), we see that

$$\begin{aligned}
&\hat{D}_\Phi(y(T); z(T)) - \hat{D}_\Phi(y(0); z(0)) \\
&\qquad = \int_0^T \partial_t \hat{D}_\Phi(y(t); z(t))\, dt \\
&\qquad \leqslant O(\log k) \int_0^T \|\partial_t y(t)\|_{\ell_1(w)}\, dt + \frac{1}{6D} \left([\Psi(T) - \Psi(0)] - (1-\delta) \int_0^T \|\partial_t z(t)\|_{\ell_1(w)}\, dt \right).
\end{aligned}$$

Rearranging yields

$$\begin{aligned}
\int_0^T \|\partial_t z(t)\|_{\ell_1(w)}\, dt \leqslant O(D \log k) &\int_0^T \|\partial_t y(t)\|_{\ell_1(w)}\, dt \\
&+ O(1) [\Psi(T) - \Psi(0)] + O(D) \left[\hat{D}_\Phi(y(T); z(T)) - \hat{D}_\Phi(y(0); z(0))\right].
\end{aligned}$$

To conclude, observe that $|\hat{D}_\Phi(\cdot\,; \cdot)|$ is uniformly bounded by $\mathrm{diam}(\mathcal{T}) O(k \log k)$, and $|\Psi(T) - \Psi(0)| \leqslant \mathrm{diam}(\mathcal{T}) O(k)$. $\square$

Note that, as opposed to the situation for HSTs, it is not known how to round online a fractional k-server algorithm on a tree to a random integral algorithm while losing only an $O(1)$ factor in the competitive ratio.

3.4.2 Cardinality for an HST

Assume now that $(\mathcal{T}, w)$ is a τ-adic HST for some $\tau \geqslant 2$. Recall that N_u is number of leaves in the subtree rooted at u and define $\Delta_u(t) := \Delta_u = \log\left(\frac{n}{N_u}\right)$, and

$$\Psi(t) := \sum_{u \in V \setminus \{\mathbb{r}\}} w_u \left(\Delta_u + \Delta_{p(u)}\right) \sum_{i \geqslant 1} x_{u,i}(t)\,.$$

Define $q = q_t$ as in (3.9). Then for every $u \in V$:

$$q(u) = \log \frac{N_{p(u)}}{N_u}\,.$$

In particular, for any two children v, v' of u:

$$\widehat{q}(u) \geqslant \log \frac{N_u}{N_v} + \log \frac{N_u}{N_{v'}} \geqslant \log 2\,.$$

Applying Lemma 3.17 with $c := \log 2$ yields

$$\frac{(1-\delta)\log 2}{4} \|\partial_t z(t)\|_{\ell_1(w)} \leqslant \partial_t \Psi(t) - 6\log(n)\cdot \partial_t \hat{D}_\Phi(y; x(t))\,.$$

Combined with Lemma 3.13, this gives the following consequence, as in the proof of Corollary 3.18.

Corollary 3.19. *If $(\mathcal{T}, w)$ is a τ-adic HST for some $\tau \geqslant 2$, then there is an $O(\log(k)\log(n))$-competitive online fractional k-server algorithm on $(\mathcal{L}, \mathrm{dist}_w)$.*

The preceding construction motivates our approach to obtaining an $O((\log k)^2)$ competitive ratio: Try to replace N_u by the fractional server mass in the subtree beneath u.

3.4.3 Fractional server-weighted depth

Finally, let us establish the $O((\log k)^2)$ bound. Suppose now that $(\mathcal{T}, w)$ is a τ-adic HST. Define:

$$\Psi(t) := \sum_{u \in V\setminus\{\mathbb{r}\}} w_u \left[\left(z_u(t) + \left(1+\tau^{-1}\mathbb{1}_{\{u\notin\mathcal{L}\}}\right)\varepsilon\right)\log(z_u(t)+\varepsilon) + z_u(t)\log(z_{p(u)}(t)+\varepsilon)\right]\,.$$

Consider some node $u \in V \setminus \{\mathbb{r}\}$, and the terms in $\partial_t \Psi(t)$ corresponding to $\partial_t z_u(t)$:

$$\begin{aligned} & w_u \partial_t z_u(t) \left(1 + \log\left(z_u(t)+\varepsilon\right) + \log\left(z_{p(u)}(t)+\varepsilon\right) + \frac{\tau^{-1}\varepsilon \mathbb{1}_{\{u\notin\mathcal{L}\}}}{z_u(t)+\varepsilon} + \tau^{-1}\mathbb{1}_{\{u\notin\mathcal{L}\}} \sum_{v:(u,v)\in\vec{E}} \frac{z_v(t)}{z_u(t)+\varepsilon}\right) \\ &\quad = w_u \partial_t z_u(t) \left(1 + \log\left(z_u(t)+\varepsilon\right) + \log\left(z_{p(u)}(t)+\varepsilon\right) + \mathbb{1}_{\{u\notin\mathcal{L}\}}\tau^{-1}\right)\,, \end{aligned}$$

where in the last equality we used that $z(t) \in \widehat{M}$.

Since $z(t) \in \widehat{M}$ and $(\mathcal{T}, w)$ is a τ-adic HST, it holds that for every $j \in \mathbb{Z}$:

$$\sum_{\substack{u\in V:\\ w_u=\tau^j}} \partial_t z_u(t) = 0\,,$$

and therefore we conclude that

$$\begin{aligned} \partial_t \Psi(t) &= \sum_{u\in V\setminus\{\mathbb{r}\}} w_u \partial_t z_u(t) \left[\log\left(z_u(t)+\varepsilon\right) + \log\left(z_{p(u)}(t)+\varepsilon\right)\right] \\ &= -\sum_{u\in V\setminus\{\mathbb{r}\}} w_u \partial_t z_u(t) \left[\log\frac{k+2\varepsilon}{z_u(t)+\varepsilon} + \log\frac{k+2\varepsilon}{z_{p(u)}(t)+\varepsilon}\right] \\ &= \frac{1}{1-\delta} \sum_{u\in V\setminus\{\mathbb{r}\}} w_u \left[\log\frac{k+2\varepsilon}{z_u(t)+\varepsilon} + \log\frac{k+2\varepsilon}{z_{p(u)}(t)+\varepsilon}\right] \sum_{i\geqslant 1} \partial_t x_{u,i}(t)\,. \end{aligned}$$

Therefore (3.8) holds with

$$\Delta_u(t) = \frac{1}{1-\delta}\log\frac{k+2\varepsilon}{z_u(t)+\varepsilon}\,,$$

and in this case:

$$q_t(u) = \frac{1}{1-\delta}\log\frac{z_{p(u)}(t)+\varepsilon}{z_u(t)+\varepsilon}\,.$$

Now we can prove the main technical theorem of this section.

Theorem 3.20. *The trajectory $\sigma(z(t))$ for $t \in [0,T]$ is an internal supermeasure of mass k that services the request at $\ell \in \mathcal{L}$. Moreover, if $y \in \mathsf{A}$ satisfies $y_\ell = 0$, then for almost every $t \in [0,T]$:*

$$\frac{1-\varepsilon}{4} \log \tfrac{4}{3} \|\partial_t \sigma(z(t))\|_{\ell_1(w)} \leqslant (1-\delta)\partial_t \Psi(t) - 6\log(2+k/\varepsilon) \cdot \partial_t \hat{D}_\Phi(y; x(t)) .$$

Proof. First, note that if $u \in V \setminus \mathcal{L}$ and $z_u(t) \geqslant \varepsilon$, then for any children v, v' of u:

$$(1-\delta)\widehat{q}_t(u) \geqslant \log \frac{z_u(t)+\varepsilon}{z_v(t)+\varepsilon} + \log \frac{z_u(t)+\varepsilon}{z_{v'}(t)+\varepsilon} \geqslant \log \frac{4}{3} ,$$

and therefore

$$\widehat{q}_t(u) \geqslant \frac{\log \frac{4}{3}}{1-\delta} \mathbb{1}_{\{z_u(t) \geqslant \varepsilon\}} . \tag{3.16}$$

Let $y^{(\ell')}(t)$ be as in the proof of Lemma 3.17. Partition $\mathcal{L} \setminus \{\ell\} = \bigcup_{v \in V \setminus \mathcal{L}} \mathcal{L}_v$, where $\mathcal{L}_v$ is the set of leaves ℓ' with $v = \mathsf{lca}(\ell, \ell')$, and define

$$y^{(v)}(t) := \sum_{\ell' \in \mathcal{L}_v} y^{(\ell')}(t) .$$

Note that by Lemma 3.11, $\partial_t z(t)$ is a flow towards ℓ, and thus there are no cancellations in the preceding sum.

Now use inequality (3.14) and (3.16) to write

$$\begin{aligned} \|\partial_t z(t)\|_{\ell_1(q_t w)} &\geqslant \frac{\log \frac{4}{3}}{(1-\delta)\tau} \sum_{\ell' \in \mathcal{L} \setminus \{\ell\}} w_{\mathsf{lca}(\ell,\ell')} \mathbb{1}_{\{z_{\mathsf{lca}(\ell,\ell')}(t) \geqslant \varepsilon\}} \left| y^{(\ell')}(t) \right| \\ &= \frac{\log \frac{4}{3}}{(1-\delta)\tau} \sum_{v \in V \setminus \mathcal{L}} w_v \mathbb{1}_{\{z_v(t) \geqslant \varepsilon\}} \left| y^{(v)}(t) \right| \\ &\geqslant \frac{1-\varepsilon}{1-\delta} \log \tfrac{4}{3} \|\partial_t \sigma(z(t))\|_{\ell_1(w)} , \end{aligned}$$

where in the final line we have used the fact that σ is $\frac{1}{1-\varepsilon}$-Lipschitz and $\sigma(z_v(t)) = 0$ when $z_v(t) < \varepsilon$. Combined with Corollary 3.16, this yields the desired result, noting that for any leaf $\ell \in \mathcal{L}$:

$$\Delta_\ell(t) \leqslant \frac{1}{1-\delta} \log \frac{k+2\varepsilon}{\varepsilon} . \qquad \square$$

Using Lemma 3.13 and Lemma 3.3 (as in the proof of Corollary 3.18), this yields an $O((\log k)^2)$-competitive online fractional k-server algorithm for any HST metric. (It is not difficult to see that every HST metric embeds with $O(1)$ distortion into the metric of a 2-adic HST.)

4 Dynamic HST embeddings

Consider a discrete metric space (X, d). Denote the *aspect ratio of* (X,d) by

$$\mathcal{A}_X := \frac{\max_{x,y \in X} d(x,y)}{\min_{x \neq y \in X} d(x,y)} .$$

Theorem 4.1. *If there is an α-competitive algorithm for k-server on HSTs, then there is an $O(\alpha \log(\mathcal{A}_X) \log k)$-competitive algorithm on any metric space (X,d).*

We use $\mathcal{R}(X) := X^{\mathbb{N}}$ to denote the space of request sequences. For $\boldsymbol{\sigma} \in \mathcal{R}(X)$ and $s \leqslant t$, denote $\boldsymbol{\sigma}_{[s,t]} := \langle \sigma_s, \sigma_{s+1}, \ldots, \sigma_t \rangle$. We will consider sequences of random variables that are implicitly functions of $\boldsymbol{\sigma} \in \mathcal{R}(X)$. Say that such a sequence $\boldsymbol{Z} = \langle Z_t : t \geqslant 0 \rangle$ is *adapted* if Z_t is a function of $\boldsymbol{\sigma}_{[1,t]}$ for every $t \geqslant 1$.

This allows one to encode state that depends on the underlying request sequence $\boldsymbol{\sigma}$ in a time-dependent way. For instance, to count the number of requests that fall into a subset $X' \subseteq X$, one would set $\mathbf{Z} = 0$. Then given that σ_t arrives at time t, we would update $\mathbf{Z} := \mathbf{Z} + \mathbb{1}_{X'}(\sigma_t)$. The meaning is: $Z_0 = 0$ and $Z_t := Z_{t-1} + \mathbb{1}_{X'}(\sigma_t)$ for all $t \geqslant 1$.

An *online algorithm for k-server on* X is an adapted sequence $\boldsymbol{A} = \langle A_1, A_2, \ldots \rangle$ where for every $t \geqslant 1$: $A_t \in X^k$ and $\sigma_t \in \{(A_t)_1, \ldots, (A_t)_k\}$. For a function f with domain X, write $f^{\otimes k}$ for the function with domain X^k given by $f^{\otimes k}(x_1, \ldots, x_k) := (f(x_1), \ldots, f(x_k))$.

For an algorithm $\boldsymbol{A}$ and a request sequence $\boldsymbol{\sigma}$, we write $\mathsf{cost}_X(\boldsymbol{A}(\boldsymbol{\sigma}))$ for the total movement cost incurred in servicing $\boldsymbol{\sigma}$. We denote by $\mathrm{opt}^X : \mathcal{R}(X) \to (X^k)^{\mathbb{N}}$ an optimal *offline* algorithm and $\mathsf{cost}^*_X(\boldsymbol{\sigma}) := \mathsf{cost}_X(\mathrm{opt}^X(\boldsymbol{\sigma}))$ the optimal offline movement cost.

4.1 Hierarchical partitions and canonical HSTs

Let us suppose that $\mathrm{diam}(X, d) = 1$ and $\mathcal{A}_X < \infty$. Let $\tau := 4$ be a scale parameter, and let $M \in \mathbb{N}$ denote the smallest number for which $\tau^{-M} < d(x, y)$ for all $x \neq y \in X$.

A sequence of subsets $\xi = (\xi_0, \xi_1, \ldots, \xi_\ell)$ of X for $0 \leqslant \ell \leqslant M$ is a *chain* if $\xi_0 = X$ and

$$\xi_0 \supseteq \xi_1 \supseteq \cdots \supseteq \xi_\ell .$$

Define the *length* of such a chain by $\mathrm{len}(\xi) := \ell$. Denote $\min(\xi) := \xi_{\mathrm{len}(\xi)}$. A chain is *complete* if it has length M and $|\xi_M| = 1$. Let $\mathcal{C}_X$ denote the set of chains in X and let $\bar{\mathcal{C}}_X$ denote the set of complete chains.

Define a rooted tree structure on $\mathcal{C}_X$ as follows. The root of $\mathcal{C}_X$ is X. For two chains $\xi, \xi' \in \mathcal{C}_X$: ξ' is a child of ξ if and only if ξ is a prefix of ξ' and $\mathrm{len}(\xi') = \mathrm{len}(\xi) + 1$. For $\xi, \xi' \in \mathcal{C}_X$, let $\mathsf{lca}(\xi, \xi') \in \mathcal{C}_X$ denote their least common ancestor. Define a τ-HST metric on $\bar{\mathcal{C}}_X$ by

$$\mathbb{d}_\tau(\xi, \xi') := \tau^{-\mathrm{len}(\mathsf{lca}(\xi,\xi'))} \qquad \xi \neq \xi' .$$

Embedding into complete chains. For a partition P of X we write $P(x)$ for the unique set $S \in P$ containing x. A τ*-stack* $\mathcal{P}$ *of* X is a sequence $\mathcal{P} = (P^0, P^1, \ldots, P^M)$ of partitions of X such that: $P^0 = \{X\}$ and for all $j = 1, 2, \ldots, M$, it holds that

$$S \in P^j \implies \mathrm{diam}(S) \leqslant \tau^{-j} .$$

Note that $P^M(x) = \{x\}$ because $\mathrm{diam}(S) \leqslant \tau^{-M}$ implies $|S| \leqslant 1$. Every τ-stack $\mathcal{P}$ induces a canonical mapping $F_{\mathcal{P}} : X \to \bar{\mathcal{C}}_X$ into the set of complete chains on X as follows. First define the *forced refinement* $\hat{\mathcal{P}} = \left(\hat{P}^0, \hat{P}^1, \ldots, \hat{P}^M\right)$ inductively by $\hat{P}^0 := P^0$ and $\hat{P}^j := \left\{S \cap S' : S \in P^j, S' \in P^{j-1}\right\}$ for $j = 1, 2, \ldots, M$. Next define $F_{\mathcal{P}}$ by

$$F_{\mathcal{P}}(x) := \left(\hat{P}^0(x), \hat{P}^1(x), \ldots, \hat{P}^M(x)\right) .$$

The following two lemmas will help to estimate the distortion of the embedding $F_{\mathcal{P}}$.

Lemma 4.2. *For any τ-stack $\mathcal{P}$, the map $F_{\mathcal{P}} : X \to (\bar{\mathcal{C}}_X, \mathbb{d}_\tau)$ is non-contracting.*

Proof. Consider $x, y \in X$. If $\mathbb{d}_\tau(F_{\mathcal{P}}(x), F_{\mathcal{P}}(y)) = \tau^{-\ell}$, then $F_{\mathcal{P}}(x)$ and $F_{\mathcal{P}}(y)$ share a common prefix of length ℓ, and therefore $\mathcal{P}^\ell(x) = \mathcal{P}^\ell(y)$. Now property (ii) of a τ-stack guarantees that $d(x, y) \leqslant \mathrm{diam}(\mathcal{P}^\ell(x)) \leqslant \tau^{-\ell}$. $\square$

Lemma 4.3. *For any τ-stack $\mathcal{P} = (P^0, P^1, \ldots, P^M)$, it holds that*

$$\mathbb{d}_\tau\left(F_{\mathcal{P}}(x), F_{\mathcal{P}}(y)\right) \leqslant \tau \sum_{j \geqslant 1} \tau^{-j} \mathbb{1}_{P_j(x) \neq P_j(y)} \qquad \forall x, y \in X\,.$$

Proof. Consider $x, y \in X$ and suppose that $\mathbb{d}_\tau\left(F_{\mathcal{P}}(x), F_{\mathcal{P}}(y)\right) = \tau^{-\ell}$ for some $\ell < M$. It is straightforward to check that $\ell + 1 = \min\{j : P^j(x) \neq P^j(y)\}$. □

Next we observe that there is a universal inverse to those embeddings. Define the mapping $F_{\mathsf{in}} : \bar{C}_X \to X$ by $F_{\mathsf{in}}(\xi) := \min(\xi)$. One has *for any* τ-stack $\mathcal{P}$:

$$F_{\mathsf{in}} \circ F_{\mathcal{P}} = \mathrm{id}_X\,. \tag{4.1}$$

4.2 The HST reduction

Let A^C denote an α-competitive k-server algorithm for the metric space $(\bar{C}_X, \mathbb{d}_\tau)$ over some probability space Ω_C. Suppose that $\boldsymbol{\sigma} = \langle \sigma_1, \sigma_2, \ldots\rangle$ is a request sequence for X, and we have an adapted sequence $\boldsymbol{\mathcal{P}} = \langle \mathcal{P}_1, \mathcal{P}_2, \ldots\rangle$ of τ-stacks of X over an independent probability space Ω_X.

This yields a mapping $F_{\boldsymbol{\mathcal{P}}} : \mathcal{R}(X) \to \mathcal{R}(C_X)$ given by

$$F_{\boldsymbol{\mathcal{P}}}(\sigma) := \left\langle F_{\mathcal{P}_1}(\sigma_1), F_{\mathcal{P}_2}(\sigma_2), \ldots \right\rangle\,.$$

From this one derives a k-server algorithm for X:

$$\boldsymbol{A}^X := F_{\mathsf{in}}^{\otimes k} \circ \boldsymbol{A}^C \circ F_{\boldsymbol{\mathcal{P}}}\,,$$

Note that $\boldsymbol{A}^X$ is a valid k-server algorithm precisely because of (4.1).

Moreover, because of Lemma 4.2, the inverse map F_{in} is non-expanding, and thus:

$$\mathsf{cost}_X(\boldsymbol{A}^X(\boldsymbol{\sigma})) \leqslant \mathop{\mathbb{E}}_{\Omega_C}\left[\mathsf{cost}_{\mathbb{d}_\tau}\big((\boldsymbol{A}^C \circ F_{\boldsymbol{\mathcal{P}}})(\boldsymbol{\sigma})\big)\right] \leqslant O_{X,k}(1) + \alpha\,\mathsf{cost}^*_{\mathbb{d}_\tau}(F_{\boldsymbol{\mathcal{P}}}(\boldsymbol{\sigma}))\,. \tag{4.2}$$

Thus our goal becomes clear: We would like to choose $\boldsymbol{\mathcal{P}}$ so that

$$\mathop{\mathbb{E}}_{\Omega_X}\left[\mathsf{cost}_{\mathbb{d}_\tau}\left((F_{\boldsymbol{\mathcal{P}}}^{\otimes k} \circ \mathrm{opt}^X)(\boldsymbol{\sigma})\right)\right] \leqslant O_{X,k}(1) + \beta\,\mathsf{cost}^*_X(\boldsymbol{\sigma}) \qquad \forall \sigma \in \mathcal{R}(X)\,. \tag{4.3}$$

Indeed since $F_{\boldsymbol{\mathcal{P}}}^{\otimes k} \circ \mathrm{opt}^X$ services the request sequence $F_{\boldsymbol{\mathcal{P}}}(\boldsymbol{\sigma})$, one has $\mathsf{cost}^*_{\mathbb{d}_\tau}(F_{\boldsymbol{\mathcal{P}}}(\boldsymbol{\sigma})) \leqslant \mathsf{cost}_{\mathbb{d}_\tau}\left((F_{\boldsymbol{\mathcal{P}}}^{\otimes k} \circ \mathrm{opt}^X)(\boldsymbol{\sigma})\right)$, and thus (4.3) in conjunction with (4.2) show that $\boldsymbol{A}^X$ is an $\alpha\beta$-competitive algorithm for X.

In essence, (4.3) asks that the embedding $F_{\boldsymbol{\mathcal{P}}}$ has β-distortion *on the subset of X that currently matters*. Focusing on such a subset is the reason why one could hope to the usual $\Omega(\log n)$ lower bound by something depending on k and $\mathcal{A}_X$.

4.3 A dynamic embedding

The algorithm will produce an adapted sequence $\boldsymbol{\mathcal{P}} = \langle \mathcal{P}_1, \mathcal{P}_2, \ldots\rangle$ of τ-stacks verifying (4.3) with $\beta \leqslant O(M \log k)$. In fact it is slightly easier for the algorithm's description to allow P_t^j to be a partial partition, i.e., simply a collection of pairwise disjoint subsets of X. In the case it is understood that the embeddings $F_{\mathcal{P}}$ use the *completion* of any such partial partition P, that is all the elements from $X \setminus [P]$ (we denote $[P] := \cup_{S \in P} S$) are added as singletons to form a complete partition.

The embedding algorithm. For $j = 1, \ldots, M$, we will maintain an adapted set $\boldsymbol{N}^j \subseteq X$ of *level-j centers* as well as an adapted mapping $\boldsymbol{R}^j : \boldsymbol{N}^j \to \mathbb{R}_+$ of radii. These will be used to construct our adapted τ-stack $\boldsymbol{\mathcal{P}} = \left(\boldsymbol{P}^0, \boldsymbol{P}^1, \ldots, \boldsymbol{P}^M\right)$.

For every $j \in \mathbb{Z}$, consider the probability distribution μ_j with density:

$$d\mu_j(r) := \frac{k\tau^j \log k}{k-1} \exp\left(-r\tau^j \log k\right) \mathbb{1}_{[0,\tau^{-j}]}(r)\,.$$

This is simply a truncated exponential distribution, as employed by Bartal [Bar96].

Initially, $\boldsymbol{N}_0^j = \boldsymbol{P}_0^j = \emptyset$ for $j = 1, \ldots, M$. Upon receiving request $\sigma_t \in X$, we proceed as follows:

For $j = 1, 2, \ldots, M$:

1. If $|\boldsymbol{N}^j| \geqslant 2k$, then
   [level-j reset]:
   For $i = j, j+1, \ldots, M$, set $\boldsymbol{N}^j := \emptyset$ and $\boldsymbol{P}^j := \emptyset$.
2. If $d(\sigma_t, \boldsymbol{N}^j) > \tau^{-j-1}$, then
   [level-j insertion]:
   (a) $\boldsymbol{N}^j := \boldsymbol{N}^j \cup \{\sigma_t\}$
   (b) $\boldsymbol{R}^j(\sigma_t) := \tau^{-j-1} + \hat{R}_t^j$, where $\hat{R}_t^j$ is sampled independently according to the distribution μ_{j+1}.
   (c) $\boldsymbol{P}^j := \boldsymbol{P}^j \cup \left\{B(\sigma_t, \boldsymbol{R}^j(\sigma_t)) \setminus [\boldsymbol{P}^j]\right\}$.

Analysis. Let us now prove that for the stack $\boldsymbol{\mathcal{P}}$ generated in this way, (4.3) holds for $\beta \leqslant O(M \log k)$. We need a few preliminary results.

Lemma 4.4. *$\mathcal{P}_t$ is a τ-stack for every $t \geqslant 1$.*

Proof. This follows from the fact that the distribution μ_{j+1} is supported on the interval $[0, \tau^{-j-1}]$ and thus every set in P_t^j is contained in a ball of radius $2\tau^{-j-1}$, which is a set of diameter at most $4\tau^{-j-1} \leqslant \tau^{-j}$ for $\tau \geqslant 4$. □

We defer the proof of the next lemma to the end of this section.

Lemma 4.5. *For all $x \in X$ and $t \geqslant 1$, it holds that*

$$\mathop{\mathbb{E}}_{\Omega_X}\left[\mathbb{d}_\tau(F_{\boldsymbol{\mathcal{P}}_t}(x), F_{\boldsymbol{\mathcal{P}}_t}(\sigma_t))\right] \leqslant O(M \log k)\, d(x, \sigma_t)\,.$$

Lemma 4.6. *For all $t \geqslant 1$ and $j = 1, 2, \ldots, M$: If $K_{j,t}$ denotes the number of level-j resets up to time t, then*

$$\mathsf{cost}_X^*(\sigma_{[1,t]}) \geqslant k\tau^{-j-1} \cdot K_{j,t}\,.$$

Proof. Suppose that between time $t_1 + 1$ and t_2 there are requests made at points $x_1, x_2, \ldots, x_{k+k'} \in X$ that satisfy $d(x_i, x_j) > D$ for all $i \neq j$. Then clearly:

$$\mathsf{cost}_X^*(\sigma_{[1,t_2]}) \geqslant \mathsf{cost}_X^*(\sigma_{[1,t_1]}) + k'D\,. \qquad \square$$

Theorem 4.7. *For every time $t \geqslant 1$:*

$$\mathop{\mathbb{E}}_{\Omega_X}\left[\mathsf{cost}_{\mathbb{d}_\tau}\left((F_{\boldsymbol{\mathcal{P}}}^{\otimes k} \circ \mathrm{opt}^X)(\sigma_{[1,t]})\right)\right] \leqslant O(M \log k)\, \mathsf{cost}_X^*(\sigma_{[1,t]}) + O_{X,k}(1)\,.$$

Proof. We can split the movement of $F_{\mathcal{P}}^{\otimes k} \circ \mathrm{opt}^X$ into three parts: First the stack $\mathcal{P}$ is possibly updated by the algorithm, either with a reset or an insertion (each induce movement), and then we mirror the move of opt^X.

Let us first consider the case of a level-j reset at time t. The key observation is that the mapping $F_{\mathcal{P}_t}$ remains identical to $F_{\mathcal{P}_t}$ on the $(j-1)$-prefix, that is $(P_t^0(x), \ldots, P_t^{j-1}(x)) = (P_{t-1}^0(x), \ldots, P_{t-1}^{j-1}(x))$, $\forall x \in X$. Thus the movement cost induced by a level-j reset is at most $k\tau^{1-j}$. In particular the total cost of resets up to time t is upper bounded by

$$\sum_{j=1}^{M} K_{j,t} \cdot k\tau^{1-j} \leqslant \tau^2 M \cdot \mathsf{cost}_X^*(\sigma_{[1,t]}) ,$$

where the last inequality is Lemma 4.6.

For the cost resulting from a level-j insertion at time t, we use the following argument. Assume that j is the smallest index in $[M]$ with a level-j insertion. Note that, by construction of the radii, one has $P_t^h(\sigma_t) \subseteq P_t^j(\sigma_t)$ for $h \geqslant j$. In particular, the movement comes from the set of servers $I_t = \{i \in [k] : (\mathrm{opt}_{t-1}^X)_i \in P_t^j(\sigma_t)\}$, for which the mapping $F_{\mathcal{P}_t}$ will change the j-suffix (i.e., $(P_t^j, \ldots, P_t^M)$) compared to $F_{\mathcal{P}_{t-1}}$. However, importantly the $(j-1)$-prefix remains identical (this is because there is no insertion at level $j-1$ and thus these servers remain part of the same non-singleton cluster at level $j-1$).

In other words, the total movement is $O(\tau^{-j}|I_t|)$. Now we argue that we can match this movement with either reset movement or movement coming from opt^X as follows. First we ignore the set of servers in $J_t \subseteq I_t$ such that their $(j-1)$-prefix remain the same forever, indeed one has $\sum_{t \geqslant 1} |J_t| = O_{X,k}(1)$. Now for a server $i \in I_t \setminus J_t$ consider the first time $s \geqslant t$ such that the $(j-1)$-prefix of $F_{\mathcal{P}_{t-1}}((\mathrm{opt}_s^X)_i)$ is different from $F_{\mathcal{P}_t}((\mathrm{opt}_{t-1}^X)_i)$. The corresponding movement at time s comes either from a reset or from a movement of opt^X, and moreover its cost is larger than τ^{-j}. Thus we just showed that at the expense of an additive term of $O_{X,k}(1)$ and a multiplicative factor 2 for movement cost induced by resets and the movement of opt^X, one can ignore the movement cost induced by insertions.

Finally, we deal with the cost coming from movement of opt^X. We may assume that opt^X is conservative: If $\sigma_t \in \mathrm{opt}_{t-1}^X$, then $\mathrm{opt}_t^X = \mathrm{opt}_{t-1}^X$ and otherwise $\mathrm{opt}_{t-1}^X \setminus \mathrm{opt}_t^X = \{x_t\}$ for some $x_t \in X$, and opt^X pays $d(\sigma_t, x_t)$. From Lemma 4.5, the expected cost of mirroring this move in $\mathbb{d}_\tau$ is at most $O(M \log k)\, d(\sigma_t, x_t)$. ◻

Thus we are left only to analyze the stretch. Note that since $d(\sigma_t, N_t^j) \leqslant \tau^{-j-1}$ by construction, the next lemma yields Lemma 4.5 when combined with Lemma 4.3.

Lemma 4.8. *For every $y \in X$ satisfying $d(y, N_t^j) \leqslant \tau^{-j-1}$ and every $x \in X$:*

$$\mathbb{P}\left[P_t^j(x) \neq P_t^j(y)\right] \leqslant (2+4e)d(x,y)\tau^{j+1} \log k .$$

Proof. Note that $|N_t^j| \leqslant 2k$ by construction. Let us arrange the centers in the order which they were added: $N_t^j = \{x_1, x_2, \ldots, x_N\}$. Let $\hat{R}_1, \ldots, \hat{R}_N$ denote the corresponding random radii $\hat{R}_i := \hat{R}_t^j(x_i)$ and let $R_i := \hat{R}_i + \tau^{-j-1}$.

Denote the event

$$\mathcal{E}_i := \left\{d(x_i, \{x, y\}) \leqslant R_i \wedge \max\{d(x_i, x), d(x_i, y)\} > R_i\right\} ,$$

and let $i_* := \min\{i : d(x_i, \{x, y\}) \leqslant R_i\}$. Define $c := (1 - \frac{1}{\log k})\tau^{-j-1}$. Then:

$$\begin{aligned}
\mathbb{P}\left[P_t^j(x) \neq P_t^j(y)\right] &= \sum_{i=1}^{N} \mathbb{P}[i = i_*] \cdot \mathbb{P}[\mathcal{E}_i \mid i = i_*] \\
&\leqslant \sum_{i=1}^{N} \mathbb{P}[\mathcal{E}_i \wedge \{\hat{R}_i \geqslant c\}] + \sum_{i=1}^{N} \mathbb{P}[i = i_*] \cdot \mathbb{P}\left[\mathcal{E}_i \wedge \{\hat{R}_i < c\} \mid i = i_*\right].
\end{aligned}$$

For any $i = 1, \ldots, N$, we have

$$\begin{aligned}
\mathbb{P}[\mathcal{E}_i \wedge \{\hat{R}_i \geqslant c\}] &\leqslant \sup_{R \geqslant c} \left\{ \int_R^{R+d(x,y)} d\mu_{j+1}(r) \right\} \\
&\leqslant \int_c^{c+d(x,y)} d\mu_{j+1}(r) \\
&= \frac{k}{k-1}\left(1 - \exp\left(-d(x,y)\tau^{j+1}\log k\right)\right) e^{-c\tau^{j+1}\log k} \\
&\leqslant \frac{2e}{k} d(x,y)\tau^{j+1}\log k\,,
\end{aligned}$$

where the final line uses $1 - e^{-u} \leqslant u$ and $k \geqslant 2$. This yields $\sum_{i=1}^{N} \mathbb{P}[\mathcal{E}_i \wedge \{\hat{R}_i \geqslant c\}] \leqslant 4e d(x,y)\tau^{j+1}\log k$ since $N \leqslant 2k$.

Now analyze:

$$\begin{aligned}
\mathbb{P}[\mathcal{E}_i \wedge \{\hat{R}_i < c\} \mid i = i_*] &\leqslant \sup_{0 \leqslant R < c} \left\{ \frac{\int_R^{R+d(x,y)} d\mu_{j+1}(r)}{\int_R^{\tau^{-j-1}} d\mu_{j+1}(r)} \right\} \\
&\leqslant \frac{\int_c^{c+d(x,y)} d\mu_{j+1}(r)}{\int_c^{\tau^{-j-1}} d\mu_{j+1}(r)} \\
&= \frac{1 - \exp\left(-d(x,y)\tau^{j+1}\log k\right)}{1 - \exp(c\tau^{j+1}\log k)/k} \leqslant 2d(x,y)\tau^{j+1}\log k\,,
\end{aligned}$$

where in the last line we have used again $1 - e^{-u} \leqslant u$ and the value of c. □

5 Mirror descent

We now prove Theorem 2.2.

5.1 Preliminaries

Consider an $\mathbb{R}^n$-set-valued map F with domain $X \subseteq \mathbb{R}^n$. We will be interested in the following *viability problem* (i.e., a differential inclusion with a constraint set for the solution): Given a constraint set $K \subseteq X$ and initial point $x_0 \in K$, find an absolutely continuous solution $x : [0, \infty) \to K$ such that

$$\begin{aligned}
\partial_t x(t) &\in F(x(t))\,, \\
x(0) &= x_0\,.
\end{aligned}$$

The upshot is that, under appropriate continuity condition on F, this problem has a solution provided that F always contain *admissible directions*, that is $F(x) \cap T_K(x) \neq \emptyset$ where $T_K(x)$ is the tangent cone to K at x (see definition below). We now recall the needed definitions with some basic results, and state the general existence theorem from [AC84].

Definition 5.1. The polar N° of a set $N \subseteq \mathbb{R}^n$ is

$$N^\circ := \{z \in \mathbb{R}^n : \langle z, y\rangle \leqslant 0 \text{ for all } y \in N\}.$$

The *normal cone to K at x* is

$$N_K(x) = (K - x)^\circ ,$$

and the *tangent cone to K at x* is

$$T_K(x) := N_K(x)^\circ .$$

Lemma 5.2 (Moreau's decomposition [HUL93, Thm III.3.2.5])**.** *Let N be a cone. Then for any $x \in \mathbb{R}^n$ there is a unique pair $(u, v) \in N \times N^\circ$ such that $\langle u, v\rangle = 0$ and $x = u + v$. Furthermore u (resp., v) is the projection of x onto N (resp., N°).*

Definition 5.3. F is *upper semicontinuous (u.s.c.)* if for any $x \in X$ and any open neighborhood $N \supset F(x)$ there exists a neighborhood M of x such that $F(M) \subseteq N$. F is *upper hemicontinuous (u.h.c.)* if, for any $\theta \in \mathbb{R}^n$, the support function $x \mapsto \sup_{y\in F(x)}\langle\theta, y\rangle$ is upper semicontinuous.

Lemma 5.4 ([AC84, Prop. 1, pg. 60; Cor. 2, pg. 63])**.** *For any F as above, u.s.c. implies u.h.c., and moreover if F takes compact, convex values then the two notions are equivalent.*

Lemma 5.5 ([AC84, Thm. 1, pg. 41])**.** *Let F and G be two set valued maps such that F is u.s.c., F takes compact values, and the graph of G is closed. Then the set-valued mapping $x \mapsto F(x) \cap G(x)$ is u.s.c.*

Theorem 5.6 ([AC84, Thm. 1, pg. 180])**.** *Assume that F is u.h.c. and takes compact, convex values, and that K is compact. Furthermore assume the tangential condition: For any $x \in X$,*

$$F(x) \cap T_K(x) \neq \emptyset .$$

Then the viability problem admits an absolutely continuous solution.

5.2 Existence

We prove the following theorem which can be viewed as a non-Euclidean extension of [AC84, pg. 217].

Theorem 5.7. *Let $K \subseteq \mathbb{R}^n$ be a compact convex set, let $H : K \to \{A \in \mathbb{R}^{n\times n} : A \succ 0\}$ be continuous, and let $f : [0, \infty) \times K \to \mathbb{R}^n$ be continuous. Then, for any $x_0 \in K$, there is an absolutely continuous solution $x : [0, \infty) \to K$ satisfying:*

$$\begin{aligned} \partial_t x(t) &\in H(x)\big(f(t, x(t)) - N_K(x(t))\big) , \\ x(0) &= x_0 . \end{aligned} \tag{5.1}$$

Furthermore, any solution to the viability problem satisfies

$$\partial_t x(t) = \operatorname{argmin}\left\{\|v - H(x)f(t, x(t))\|^2_{x,*} : v \in T_K(x(t))\right\}. \tag{5.2}$$

In particular, one has

$$\|\partial_t x(t)\|_{x,*} \leqslant \|f(t, x)\|_x . \tag{5.3}$$

Proof. It suffices to prove the existence on any time interval $[T, T + 1]$. We denote $\langle\cdot, \cdot\rangle_x$ for the inner product induced by $H(x)$ (i.e., $\langle\alpha, \beta\rangle_x := \langle\alpha, H(x)\beta\rangle$), $\|\cdot\|_x$ for the corresponding norm, and $\|\cdot\|_{x,*}$ for its dual norm. To apply Theorem 5.6, consider the following differential inclusion, with $\overline{K} = [T, T + 1] \times K$,

$$\widetilde{x}'(t) \in F(\widetilde{x}(t)) ,$$

where $F : \overline{K} \to 2^{\mathbb{R}^{n+1}}$ defined by

$$F(t,x) := \Big(1, H(x)\big(f(t,x) - N_K(x)\big) \cap \{v \in \mathbb{R}^n : \|v\|_{x,*} \leqslant \|f(t,x)\|_x\}\Big) \,.$$

Thanks to the restriction to velocities satisfying $\|v\|_{x,*} \leqslant \|f(t,x)\|_x$, it holds that $F(t,x)$ is compact (it is also clearly convex). Moreover, $\overline{K}$ is compact. Thus, besides the tangential condition, it remains to show that F is u.h.c. Since F is compact and convex valued, by Lemma 5.4 it suffices to show that F is u.s.c. For this we apply Lemma 5.5. Note that $\{(1, H(x)(f(t,x)-y)) : x \in K, y \in N_K(x)\}$ is closed (using again continuity of H and f), and thus it suffices to show that the mapping $x \mapsto \{v \in \mathbb{R}^n : \|v\|_{x,*} \leqslant \|f(t,x)\|_x\}$ is u.s.c., which is clearly true since its support function in direction θ is $x \mapsto \|f(t,x)\|_x\|\theta\|_x$ which is continuous by continuity of H and f.

It remains to check the tangential condition. We note that

$$N_{\overline{K}}(t,x) = \begin{cases} (-\infty,0] \times N_K(x) & t = 0\,, \\ \{0\} \times N_K(x) & \text{otherwise.} \end{cases}$$

$$T_{\overline{K}}(t,x) = \begin{cases} [0,\infty) \times T_K(x) & t = 0\,, \\ \mathbb{R} \times T_K(x) & \text{otherwise.} \end{cases}$$

Thus it suffices to show that there exists $u \in N_K(x)$ such that $\|H(x)(f(t,x)-u)\|_{x,*} \leqslant \|f(t,x)\|_x$ and $H(x)(f(t,x)-u) \in T_K(x)$.

We apply Moreau's decomposition (Lemma 5.2) in $\mathbb{R}^n$ equipped with the inner product $\langle\cdot,\cdot\rangle_x$ to the cone $N_K(x)$ and write $f(t,x) = u+v$ where $u \in N_K(x)$ and v is in the polar (w.r.t. $\langle\cdot,\cdot\rangle_x$) of $N_K(x)$. Since $\langle u, v\rangle_x = 0$, we have $\|f(t,x)-u\|_x \leqslant \|f(t,x)\|_x$. This gives $\|H(x)(f(t,x)-u)\|_{x,*} \leqslant \|f(t,x)\|_x$. Furthermore since v is in the polar, we have for all $y \in N_K(x)$, $\langle f(t,x)-u, y\rangle_x \leqslant 0$ which means that $H(x)(f(t,x)-u) \in T_K(x)$. This concludes the existence proof.

Since the objective function in (5.2) is strongly convex, the optimality condition of (5.2) shows that any $v \in T_K(x(t))$ satisfying $H(x)^{-1}(v - H(x)f(t,x(t))) \in -T_K(x(t))^\circ$ is the unique solution. Now, we note that $\partial_t x(t) \in T_K(x(t))$ because $x(t) \in K$ and

$$H(x)^{-1}(\partial_t x(t) - H(x)f(t,x(t))) \in -N_K(x(t)) = -T_K(x(t))^\circ\,,$$

where we used that $T_K(x(t))^\circ = N_K(x(t))^{\circ\circ} = N_K(x(t))$ (since $N_K(x(t))$ is a closed and convex cone). This establishes (5.2).

Since we exhibited a solution satisfying (5.3), the almost everywhere uniqueness of $\partial_t x(t)$ shows that (5.3) holds for any solution. □

5.3 Uniqueness

We prove here that the solution to the viability problem is in fact unique under slightly more restrictive assumptions than those in Theorem 5.7. In what follows, we denote, as in the proof of Theorem 5.7, $\langle\cdot,\cdot\rangle_x$ for the inner product induced by $H(x)$, $\|\cdot\|_x$ for its corresponding norm and $\|\cdot\|_{x,*}$ for its dual norm.

Lemma 5.8. *Let $x(t)$ be an absolutely continuous path with values in a convex set $K \subseteq \mathbb{R}^n$. Then, $\partial_t x(t) \in N_K(x(t))^\perp$ almost everywhere.*

Proof. For any t such that $\partial_t x(t)$ exists, one has

$$x(t+h) = x(t) + h\partial_t x(t) + o(h)\,.$$

In particular for any $y \in N_K(x(t))$, since $x(t+h) \in K$, one has

$$\big\langle y, x(t+h) - x(t)\big\rangle \leqslant 0.$$

Taking $h \to 0^+$, we obtain $\langle y, \partial_t x(t) \rangle \leqslant 0$. Taking $h \to 0^-$, we obtain $\langle y, \partial_t x(t) \rangle \geqslant 0$. Therefore, $\partial_t x(t) \in N_K(x(t))^\perp$. □

Lemma 5.9. *The solution $x(t)$ in Theorem 5.7 is unique provided that H is Lipschitz, and f is locally Lipschitz.*

Proof. Let $x(t)$ and $\tilde{x}(t)$ be two solutions to the viability problem. We show that for any $T \geqslant 0$ there are some constants $C, \varepsilon > 0$ such that for all $t \in [T, T+\varepsilon]$,

$$\partial_t \|x(t) - \tilde{x}(t)\|_{x(t),*}^2 \leqslant C\|x(t) - \tilde{x}(t)\|_{x(t),*}^2 , \tag{5.4}$$

which concludes the proof by a simple application of Gronwall's inequality (notice that since H is continuous and K is compact, there is some constant ε' such that $H^{-1}(x) \succeq \varepsilon' \mathrm{I}_n, \forall x \in K$).

Recall that we have $\partial_t x(t) = H(x(t))(f(t, x(t)) - u(t))$ for some $u(t) \in N_K(x(t))$ (and similarly for $\tilde{x}$ there is some path $\tilde{u}$ in the normal cones). In particular we get:

$$\frac{1}{2}\partial_t \|x(t) - \tilde{x}(t)\|_{x(t),*}^2 = \langle x(t) - \tilde{x}(t), H(x(t))(f(t, x(t)) - u(t)) - H(\tilde{x}(t))(f(t, \tilde{x}(t)) - \tilde{u}(t))\rangle_{x(t),*} \tag{5.5}$$

$$+ \frac{1}{2}(x(t) - \tilde{x}(t))^\top (\partial_t H(x(t))^{-1})(x(t) - \tilde{x}(t)) . \tag{5.6}$$

Denote $\|\cdot\|_{\mathrm{op}}$ for the spectral norm, and observe that by continuity of H and compactness of K, there exists $M > 0$ such that for all t, $\|H(x(t))\|_{\mathrm{op}}, \|H(x(t))^{-1}\|_{\mathrm{op}} \leqslant M$. Thus we can bound the term (5.6) as follows:

$$\begin{aligned}(x(t) - \tilde{x}(t))^\top (\partial_t H(x(t))^{-1})(x(t) - \tilde{x}(t)) &\leqslant \|x(t) - \tilde{x}(t)\|_{x(t),*}^2 \|H(x(t))^{1/2}(\partial_t H(x(t))^{-1})H(x(t))^{1/2}\|_{\mathrm{op}} \\ &\leqslant M\|x(t) - \tilde{x}(t)\|_{x(t),*}^2 \|(\partial_t H(x(t))^{-1})\|_{\mathrm{op}} .\end{aligned}$$

Now since H is Lipschitz, there exists a constant M' such that for all $t \in [T, T+\varepsilon]$,

$$\begin{aligned}\|(\partial_t H(x(t))^{-1})\|_{\mathrm{op}} &= \|H(x(t))^{-1}(\partial_t H(x(t)))H(x(t))^{-1}\|_{\mathrm{op}} \\ &\leqslant M^2 M' \|\partial_t x(t)\|_2 \leqslant M^3 M' \|\partial_t x(t)\|_{x(t),*} \leqslant M^3 M' \|f(t, x(t))\|_{x(t)} ,\end{aligned}$$

where the last inequality follows from (5.3). Since $f(t,x)$ is uniformly bounded on $[T, T+\varepsilon]$, we finally get that for some $C > 0$, the term (5.6) is bounded from above by $C\|x(t) - \tilde{x}(t)\|_{x(t),*}^2$ for all $t \in [T + \varepsilon]$.

Next we consider the term (5.5) and decompose it into two terms:

(i) $(x(t) - \tilde{x}(t))^\top \left(f(t, x(t)) - u(t) - (f(t, \tilde{x}(t)) - \tilde{u}(t))\right)$, and
(ii) $\left\langle x(t) - \tilde{x}(t), (H(x(t)) - H(\tilde{x}(t)))\left(f(t, \tilde{x}(t)) - \tilde{u}(t)\right)\right\rangle_{x(t),*}$.

We further decompose (i) into

(iii) $(x(t) - \tilde{x}(t))^\top (f(t, x(t)) - f(t, \tilde{x}(t)))$, and
(iv) $(x(t) - \tilde{x}(t))^\top (\tilde{u}(t) - u(t))$.

To bound (iii), we now use that f is locally Lipschitz, which shows that this term is bounded from above by $C\|x(t) - \tilde{x}(t)\|_{x(t),*}$ for any $t \in [T, T+\varepsilon]$.

The term (iv) is nonpositive since $u(t) \in N_K(x(t))$ and $\tilde{u}(t) \in N_K(\tilde{x}(t))$. Finally, for (ii) one can combine Lipschitzness of H and (5.3) as above to obtain that it is bounded by $C\|x(t) - \tilde{x}(t)\|_{x(t),*}$ for all $t \in [T, T+\varepsilon]$, thus concluding the proof. □

Acknowledgements

We thank Anna Karlin, Manor Mendel, and Yuval Rabani for useful comments on an initial draft of this manuscript.

Part of this work was carried out while M. B. Cohen, Y. T. Lee, and J. R. Lee were at Microsoft Research in Redmond. We thank Microsoft for their hospitality. M. B. Cohen and A. Mądry are supported by NSF grant CCF-1553428 and an Alfred P. Sloan Fellowship. J. R. Lee is supported by NSF grants CCF-1616297 and CCF-1407779 and a Simons Investigator Award. Y. T. Lee is supported by NSF grants CCF-1740551.

References

[ABBS10] Jacob Abernethy, Peter Bartlett, Niv Buchbinder, and Isabelle Stanton. A regularization approach to metrical task systems. In *Algorithmic Learning Theory*, ALT 2010. Springer, 2010. 3

[AC84] Jean Pierre Aubin and A. Cellina. *Differential Inclusions: Set-Valued Maps and Viability Theory*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1984. 26, 27

[Bar96] Yair Bartal. Probabilistic approximations of metric spaces and its algorithmic applications. In *37th Annual Symposium on Foundations of Computer Science, FOCS '96, Burlington, Vermont, USA, 14-16 October, 1996*, pages 184–193, 1996. 2, 24

[Bar98] Yair Bartal. On approximating arbitrary metrices by tree metrics. In *STOC '98 (Dallas, TX)*, pages 161–168. ACM, New York, 1998. 2

[BBK99] Avrim Blum, Carl Burch, and Adam Kalai. Finely-competitive paging. In *40th Annual Symposium on Foundations of Computer Science (New York, 1999)*, pages 450–457. IEEE Computer Soc., Los Alamitos, CA, 1999. 2

[BBM06] Yair Bartal, Béla Bollobás, and Manor Mendel. Ramsey-type theorems for metric spaces with applications to online problems. *J. Comput. System Sci.*, 72(5):890–921, 2006. 2

[BBMN15] Nikhil Bansal, Niv Buchbinder, Aleksander Madry, and Joseph Naor. A polylogarithmic-competitive algorithm for the k-server problem. *J. ACM*, 62(5):Art. 40, 49, 2015. 3, 10

[BBN12] Nikhil Bansal, Niv Buchbinder, and Joseph Naor. A primal-dual randomized algorithm for weighted paging. *J. ACM*, 59(4):Art. 19, 24, 2012. 3

[BCN14] Niv Buchbinder, Shahar Chen, and Joseph (Seffi) Naor. Competitive analysis via regularization. In *Proceedings of the Twenty-fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '14, pages 436–444, Philadelphia, PA, USA, 2014. Society for Industrial and Applied Mathematics. 3

[BE98] Allan Borodin and Ran El-Yaniv. *Online computation and competitive analysis*. Cambridge University Press, New York, 1998. 2

[BKRS00] Avrim Blum, Howard Karloff, Yuval Rabani, and Michael Saks. A decomposition theorem for task systems and bounds for randomized server problems. *SIAM J. Comput.*, 30(5):1624–1661, 2000. 2

[BLMN05] Yair Bartal, Nathan Linial, Manor Mendel, and Assaf Naor. On metric Ramsey-type phenomena. *Ann. of Math. (2)*, 162(2):643–709, 2005. 2

[Bub15] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015. 3

[CKPV91] M. Chrobak, H. Karloff, T. Payne, and S. Vishwanathan. New results on server problems. *SIAM J. Discrete Math.*, 4(2):172–181, 1991. 3

[CL91] Marek Chrobak and Lawrence L. Larmore. An optimal on-line algorithm for k servers on trees. *SIAM J. Comput.*, 20(1):144–148, 1991. 3

[CMP08] Aaron Coté, Adam Meyerson, and Laura Poplawski. Randomized k-server on hierarchical binary trees. In *STOC'08*, pages 227–233. ACM, New York, 2008. 3

[FKL+91] A. Fiat, R. M. Karp, M. Luby, L. A. McGeoch, D. D. Sleator, and Young N. E. Competitive paging algorithms. *J. Algorithms*, 12(4):685–699, 1991. 2

[FRT04] Jittat Fakcharoenphol, Satish Rao, and Kunal Talwar. A tight bound on approximating arbitrary metrics by tree metrics. *J. Comput. Syst. Sci.*, 69(3):485–497, 2004. 2

[Haz16] Elad Hazan. Introduction to online convex optimization. *Foundations and TrendsÂő in Optimization*, 2(3-4):157–325, 2016. 3

[HUL93] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex analysis and minimization algorithms, Volume I: Fundamentals.* Springer-Verlag, 1993. 27

[Kou09] Elias Koutsoupias. The k-server problem. *Computer Science Review*, 3(2):105–118, 2009. 2

[KP95] Elias Koutsoupias and Christos H. Papadimitriou. On the k-server conjecture. *J. Assoc. Comput. Mach.*, 42(5):971–983, 1995. 2, 3

[KRR94] Howard Karloff, Yuval Rabani, and Yiftach Ravid. Lower bounds for randomized k-server and motion-planning algorithms. *SIAM J. Comput.*, 23(2):293–312, 1994. 2

[MMS90] Mark S. Manasse, Lyle A. McGeoch, and Daniel D. Sleator. Competitive algorithms for server problems. *J. Algorithms*, 11(2):208–230, 1990. 2

[MS91] Lyle A. McGeoch and Daniel D. Sleator. A strongly competitive randomized paging algorithm. *Algorithmica*, 6(6):816–825, 1991. 2

[NY83] A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley Interscience, 1983. 3