BALANCING SUMS OF RANDOM VECTORS

JUHAN ARU, BHARGAV NARAYANAN, ALEX SCOTT,
AND RAMARATHNAM VENKATESAN

ABSTRACT. We study a higher-dimensional 'balls-into-bins' problem. An infinite sequence of i.i.d. random vectors is revealed to us one vector at a time, and we are required to partition these vectors into a fixed number of bins in such a way as to keep the sums of the vectors in the different bins close together; how close can we keep these sums almost surely? This question, our primary focus in this paper, is closely related to the classical problem of partitioning a sequence of vectors into balanced subsequences, in addition to having applications to some problems in computer science.

1. INTRODUCTION

In this note, we consider the following partitioning problem. Let $\mathbf{V}(\mu) = (V_n)_{n \geq 1}$ be a sequence of independent random vectors, all distributed according to some common probability distribution μ on the d-dimensional Euclidean unit ball $\mathbb{B}^d \subset \mathbb{R}^d$. The elements of this i.i.d. sequence $\mathbf{V}(\mu)$ are revealed to us in order, one vector at a time. Each time a new vector is revealed to us, we are required to assign this vector to one of a fixed number of bins $\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_k$ before seeing the next vector in the sequence. By adopting a suitable strategy to assign vectors to bins, how 'close together' can we keep the sums of the vectors in the different bins?

A more precise formulation of this question is as follows. For each $1 \leq i \leq k$, let $B_i^0 = 0$, and for a positive integer $n \in \mathbb{N}$, let B_i^n denote the sum of the vectors in the bin $\mathbf{B}_i$ at time n; in other words,

$$B_i^n = \sum_{j=1}^{n} V_j \mathbb{1}_{\{V_j \in \mathbf{B}_i\}}.$$

A *partitioning strategy* is a (possibly randomised) map from $(\mathbb{R}^d)^{k+1}$ to the set of bins $\{\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_k\}$ which, given the vectors $B_1^n, B_2^n, \ldots, B_k^n$ and V_{n+1}, tells us which bin V_{n+1} should be assigned to; in the language of computer science, a

partitioning strategy is an *online algorithm* for assigning vectors to bins. We shall mainly be interested in partitioning strategies that minimise, for (large) $T \geq 0$, the quantity

$$D(T) = \max_{1 \leq n \leq T} \max_{1 \leq i,j \leq k} \left\| B_i^n - B_j^n \right\|,$$

the largest observed Euclidean distance between a pair of bins up to time T. In this paper, we shall mostly be concerned with the asymptotic behaviour of $D(T)$ as $T \to \infty$ while the dimension $d \geq 1$, the number of bins $k \geq 2$ and the distribution μ remain fixed, so it is perhaps worth emphasising that we choose to work with the Euclidean norm and to track the largest observed distance between any pair of bins purely for concreteness; indeed, all of our results hold as stated, albeit with different implied constants, for any choice of norm, and for any well-defined notion that tracks how close together the various bins are, such as the largest distance between any bin and the average of the bins, for example.

The fact that serves as the starting point for the work here is the classical result (see [2], for instance) that by assigning vectors to bins uniformly at random, one can always ensure that $D(T) = O(\sqrt{T \log T})$; we shall attempt to quantify by how much one can hope to improve on this. We discuss two other motivations for studying the problem at hand below.

First, the related problem of partitioning a *deterministic* sequence of vectors from the d-dimensional unit ball into a fixed number of 'balanced subsequences' has a rich history and may be traced back to an old question due to Riemann and Lévy that was subsequently answered by Steinitz [14]; various forms of this problem have since been investigated and we refer the interested reader to the survey of Bárány [6]. We mention one result in this area that is relevant to the problem at hand. Let $\mathbf{V} = (V_n)_{n \geq 1}$ be *any* sequence of vectors lying in the d-dimensional unit ball $\mathbb{B}^d \subset \mathbb{R}^d$, and consider the problem of assigning each element of this sequence of vectors to one of a fixed number of bins $\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_k$ as before, except using a *prescient partitioning strategy*: by a prescient partitioning strategy, we mean a strategy that is allowed to see the entire sequence $\mathbf{V}$ ahead of time. Improving on an earlier result of Doerr and Srivastav [11], Bárány and Doerr [8] proved that there exists a prescient partitioning strategy that ensures that $D(T) \leq Cd$ uniformly in T for any $d, k \in \mathbb{N}$ with $k \geq 2$ and any sequence $\mathbf{V}$ as above, where $C \approx 4.001$ is a universal constant (though it should be noted that while these prescient strategies do require 'knowledge of the future', they only need to look $\Theta(d)$ vectors ahead).

In the light of this fact, it is natural to ask what changes when we are required to partition $\mathbf{V}$ without any knowledge of the future.

Next, the question we study here arises naturally in the context of load-balancing, resource allocation and scheduling problems in large scale computation. In these settings, there is a finite set of servers and a sequence of incoming jobs. Jobs must be allocated to servers as and when they arrive, and each job consumes certain quantities of the different resources (memory or processing power, for example) available on the server it is assigned to. Ideally, one would like to allocate the jobs in a 'balanced' fashion where the total load on each server is roughly the same; see the survey of Azar [3] for a short introduction to this area. Partitioning strategies that perform well in the 'worst-case' can often be suboptimal in practice since the empirical distribution of the incoming jobs is typically random (and not adversarial). Therefore, various probabilistic models for these problems have been studied over the last twenty years and a number of results have been proved in different settings. These results, for the most part, deal with the one-dimensional case of the problem we study and address various questions about distributing (possibly weighted) balls into bins; for a small sample of the existing literature, see [4, 1, 15, 13]. The higher-dimensional problem we study here is not only inherently interesting, but also exhibits genuinely different behaviour compared to the one-dimensional problem, as we shall shortly see. Finally, let us make two remarks with practical applications in mind: first, in practice, we lose no generality by assuming that the 'job-vectors' come from the unit ball as our results remain valid after any suitable (finite) rescaling; second, fine tuning our partitioning strategies for specific distributions can make the strategies fragile, so we focus on results that either hold for all probability distributions on the unit ball, or are robust for a wide class of 'nice' distributions.

2. Our results

A few remarks about notation are in order before we state our results. In what follows, we write $[n]$ for the set $\{1, 2, \ldots, n\}$. We use $\langle \cdot, \cdot \rangle$ to denote the standard inner product in $\mathbb{R}^d$ and $\lVert \cdot \rVert$ to denote the associated Euclidean norm. Also, we write λ_d for the d-dimensional Lebesgue measure. We shall make use of standard asymptotic notation; in what follows, the variable tending to infinity will always be T unless we explicitly specify otherwise. Constants suppressed by the asymptotic notation are allowed to depend on the fixed parameters (d, k and μ) but not on T.

Finally, we use the term *with high probability* to mean with probability tending to 1 as $T \to \infty$.

Our first result, a strategy-agnostic lower bound on $D(T)$, will serve as a useful benchmark.

Proposition 2.1. *Fix $d, k \geq 2$, and let μ be the uniform distribution on $\mathbb{B}^d \subset \mathbb{R}^d$. Regardless of the partitioning strategy used to partition $\mathbf{V}(\mu)$ into k bins, almost surely,*

$$\liminf_{T \to \infty} \left(\frac{D(T)(\log \log T)^{1/2}}{(\log T)^{1/2}} \right) = \Omega(1).$$

The above proposition immediately highlights the difference between the one-dimensional partitioning problem and the same problem in higher dimensions. Indeed, if we have a sequence of i.i.d. vectors distributed according to some common distribution μ on $\mathbb{B}^1 = [-1, 1]$, then the trivial partitioning strategy that assigns a vector V to the bin with the largest sum if $V < 0$ and to the bin with the smallest sum if $V > 0$ shows that in one dimension, we may uniformly ensure that $D(T) \leq 1$ for any number of bins k and any distribution μ.

In an attempt to match the lower bound in Proposition 2.1, we consider two different partitioning strategies below. Note that for any $d, k \in \mathbb{N}$ and any distribution μ on $\mathbb{B}^d$, we may, as discussed earlier, toss each element of $\mathbf{V}(\mu)$ into one of the k bins uniformly at random and thereby ensure with high probability that $D(T) = O(\sqrt{T \log T})$. To improve on this trivial bound, we shall have to work a bit harder.

The first partitioning strategy we propose, which we call the *inner product rule*, is as follows: simply assign V_{n+1} to the bin $\mathbf{B}_i$ for which $\langle V_{n+1}, B_i^n \rangle$ is minimal, breaking ties arbitrarily. Intuitively, this should keep the bins close together since we always add a vector to the bin it is 'most opposite' to. We shall show that the inner product rule is a near-optimal partitioning strategy for any reasonably well-behaved probability distribution. Recall that a measure μ on $\mathbb{R}^d$ is *Hölder continuous (with respect to the Lebesgue measure)* if there exist constants $K, \alpha > 0$ such that $\mu(S) \leq K\lambda_d(S)^\alpha$ for any measurable set $S \subset \mathbb{R}^d$; the following bounds for the inner product rule are essentially tight.

Theorem 2.2. *Fix $d, k \in \mathbb{N}$ with $k \geq 2$, let μ be a probability distribution on $\mathbb{B}^d \subset \mathbb{R}^d$, and suppose that we partition $\mathbf{V}(\mu)$ into k bins using the inner product*

4

rule. Then almost surely,

$$\limsup_{T \to \infty} \left(\frac{D(T)}{(\log T)^{1/2}} \right) = O(1); \tag{1}$$

if μ is additionally Hölder continuous, then almost surely,

$$\limsup_{T \to \infty} \left(\frac{D(T)(\log \log T)^{1/2}}{(\log T)^{1/2}} \right) = O(1). \tag{2}$$

In the light of Proposition 2.1, it is immediately clear that (2) is essentially best-possible. The discrepancy between the bound (2) for well-behaved distributions and the bound (1) for arbitrary distributions in Theorem 2.2 is not just an artefact of our proof: somewhat surprisingly, the next proposition demonstrates the existence of (slightly pathological) distributions which show that (1) is also nearly tight.

Proposition 2.3. *For every increasing function $\omega \colon \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ that grows without bound, there exists a probability distribution μ_ω on $\mathbb{B}^2 \subset \mathbb{R}^2$ for which the following holds. If we partition $\mathbf{V}(\mu_\omega)$ into two bins using the inner product rule, then almost surely,*

$$\limsup_{T \to \infty} \left(\frac{D(T)\omega(T)}{(\log T)^{1/2}} \right) \geq 1.$$

The other strategy we consider is motivated by more practical considerations: in applications, where the number of bins k is often large, it is usually too expensive to compute k inner products to make each decision. With this in mind, we investigate the following strategy, a higher dimensional analogue of the 'two random choices' strategy studied by Azar, Broder, Karlin and Upfal [5], which we call the *best-of-two rule*. Unlike the inner product rule, the best-of-two rule is a randomised strategy: given V_{n+1}, we choose two bins $\mathbf{B}_i$ and $\mathbf{B}_j$ randomly from the set of all bins (without replacement) and assign V_{n+1} to $\mathbf{B}_i$ if $\langle V_{n+1}, B_i^n \rangle \leq \langle V_{n+1}, B_j^n \rangle$ and to $\mathbf{B}_j$ otherwise, breaking ties arbitrarily. This strategy achieves a reduction in computational complexity, but this reduction comes at a price: the following estimate for the best-of-two rule is essentially tight.

Theorem 2.4. *Fix $d, k \in \mathbb{N}$ with $k \geq 2$, and let μ be any probability distribution on $\mathbb{B}^d \subset \mathbb{R}^d$. If we partition $\mathbf{V}(\mu)$ into k bins using the best-of-two rule, then almost surely,*

$$\limsup_{T \to \infty} \left(\frac{D(T)}{\log T} \right) = O(1).$$

5

Note that the best-of-two rule is identical to the inner product rule when $k = 2$. That Theorem 2.4 is essentially best-possible when $k \geq 3$ is evidenced by the simple observation that if μ is the uniform distribution on $\mathbb{B}^1 = [-1, 1]$, then with high probability, there exists an interval of length $\Omega(\log T)$ in the first T steps where we repeatedly choose the same pair of bins and only see numbers exceeding $1/2$. To prove Theorem 2.4, we shall show that the best-of-two rule enforces 'self-correction'. Similar methods based on self-correction have recently been used to answer some long-standing questions about random graph processes; see [9, 12, 10], for example.

This paper is organised as follows. We give the proof of Proposition 2.1 in Section 3. Section 4 is devoted to analysing the inner product rule. We then address the best-of-two rule in Section 5. We finally conclude this note with a discussion of some open problems in Section 6. For the sake of clarity of presentation, we systematically omit floor and ceiling signs whenever they are not crucial.

3. Lower bounds

This section is devoted to the proof of Proposition 2.1, our strategy-agnostic lower bound. For completeness, we first record the following fact about the size of 'slices' of the d-dimensional unit ball.

Proposition 3.1. *For any $d \in \mathbb{N}$, there exist constants $C, c > 0$ such that for any $e \in \mathbb{S}^{d-1}$ and any $0 \leq b \leq 1$,*

$$cb \leq \lambda_d(\{x : x \in \mathbb{B}^d \text{ and } |\langle x, e \rangle| \leq b\}) \leq Cb. \qquad \square$$

We now prove Proposition 2.1.

Proof of Proposition 2.1. First, suppose that $k = 2$ and consider any partitioning strategy that partitions $\mathbf{V}(\mu)$ into two bins. Let $\delta_n = B_1^n - B_2^n$ and write e_n for the unit vector in the direction of δ_n. Note that we have

$$\|\delta_{n+1}\|^2 = \|\delta_n\|^2 + \|V\|^2 + 2\|\delta_n\|\langle V, e_n \rangle,$$

where $V = V_{n+1}$ if the strategy assigns V_{n+1} to $\mathbf{B}_1$ and $V = -V_{n+1}$ otherwise.

Consider the event $E_{n+1} = \{1/2 \leq \|\delta_{n+1}\|^2 - \|\delta_n\|^2 \leq 5/4\}$. We claim that regardless of the partitioning strategy used, we have

$$\mathbb{P}(E_{n+1} \mid \delta_n) \geq \frac{c_d}{\|\delta_n\|}, \tag{3}$$

where $c_d > 0$ is a constant depending on the dimension d alone. Indeed, regardless of which bin we assign V_{n+1} to, if $\langle V_{n+1}, e_n \rangle \in [-(8\|\delta_n\|)^{-1}, (8\|\delta_n\|)^{-1}]$ and $\|V_{n+1}\| \geq 3/4$, then E_{n+1} holds. Since μ is the uniform distribution and $\mathbf{V}(\mu)$ is an i.i.d. sequence, the claimed bound (3) follows easily from Proposition 3.1.

Now, break the set $[T]$ into $r = T/m$ disjoint blocks $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_r$ each of length m for some $m = m(T)$ that grows slowly with T (and will be specified later). We say that a block $\mathcal{T}$ is *good* if $\|\delta_n\|^2 \geq m/2$ for some $n \in \mathcal{T}$. Now, for $i \in [r]$, consider the block $\mathcal{T}_i = \{t+1, t+2, \ldots, t+m\}$ with $t = (i-1)m$ and note, writing F_i for the event that $\|\delta_t\|^2 < m/2$, that

$$\mathbb{P}(\mathcal{T}_i \text{ is good} \mid F_i) \geq \mathbb{P}(E_{t+1} \cap E_{t+2} \cap \cdots \cap E_{t+m} \mid F_i).$$

Using (3) and the fact that $\mathbf{V}(\mu)$ is an i.i.d. sequence, we see that

$$\mathbb{P}(E_{t+j+1} \mid F_i \cap E_{t+1} \cap E_{t+2} \cap \cdots \cap E_{t+j}) \geq \frac{c_d}{\sqrt{m/2 + 5j/4}} \geq \frac{c_d}{2\sqrt{m}}.$$

It follows that

$$\mathbb{P}(\mathcal{T}_i \text{ is good} \mid F_i) \geq \left(\frac{c_d}{2\sqrt{m}}\right)^m.$$

Using the Markov property, we now deduce that

$$\mathbb{P}\big(D(T) < (m/2)^{1/2}\big) \leq \prod_{i=1}^{r} (1 - \mathbb{P}(\mathcal{T}_i \text{ is good} \mid F_i)) \leq \exp\left(-\frac{T}{m}\left(\frac{c_d}{2\sqrt{m}}\right)^m\right).$$

Applying the above bound with $m = \log T / \log \log T$, we conclude that

$$\mathbb{P}\left(D(T) < \left(\frac{\log T}{2 \log \log T}\right)^{1/2}\right) \leq T^{-2}$$

for all sufficiently large T; the proposition, in the case where $k = 2$, now follows from the Borel–Cantelli lemma.

In contrast to the situation with the arguments for upper bounds (that follow in subsequent sections), we may easily obtain a lower bound in the case where the number of bins exceeds two from the argument above that deals with the case of exactly two bins. Indeed when $k > 2$, we proceed by 'merging' the bins $\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_{k'}$ and the bins $\mathbf{B}_{k'+1}, \mathbf{B}_{k'+2}, \ldots, \mathbf{B}_k$ into two auxiliary bins $\mathbf{A}_1$ and $\mathbf{A}_2$ respectively, where $k' = \lfloor k/2 \rfloor$; in other words, we set $A_1^n = \sum_{i=1}^{k'} B_i^n$ and $A_2^n = \sum_{i=1}^{k} B_i^n - A_1^n$ for each $n \in \mathbb{N}$. If k is even, then we finish the proof as follows.

By the argument above, it is clear that regardless of the partitioning strategy used, there exists an $n \in [T]$ for which

$$\|A_1^n - A_2^n\| \geq \left(\frac{\log T}{2 \log \log T}\right)^{1/2}$$

with probability at least $1 - T^{-2}$; the result now follows from the triangle inequality. If k is odd on the other hand, then the result follows from an analogous argument where we track $\|(1 + 1/k')A_1^n - A_2^n\|$ instead of $\|A_1^n - A_2^n\|$. $\qquad\square$

4. The inner product rule

We shall analyse the inner product rule in this section. We need the following standard Chernoff-type bound; see [2] for a proof.

Proposition 4.1. *If $X_1, X_2, \ldots, X_n$ are independent random variables taking values in $\{0, 1\}$, then writing $X = \sum_{i=1}^n X_i$, we have*

$$\mathbb{P}(X \leq \mathbb{E}[X]/2) \leq \exp(-\mathbb{E}[X]/8). \qquad\square$$

We start by proving Theorem 2.2.

Proof of Theorem 2.2. Given $m \geq 0$, we wish to bound $\mathbb{P}(D(T) \geq m)$ from above. Somewhat surprisingly, this is harder to do in the case where $k \geq 3$ as opposed to when $k = 2$. Indeed, to control $D(T)$, we need to control the distance between each pair of bins; however, if we attempt to control these distances *individually*, we quickly run into difficulties because we cannot say much about how the distance between a *particular* pair of bins changes at each time (unless $k = 2$). The trick is to instead track the observable

$$S_n = \sum_{1 \leq i < j \leq k} \|\delta_n(i, j)\|^2,$$

where $\delta_n(i, j) = B_i^n - B_j^n$ for all $i, j \in [k]$.

First, writing $e_n(i, j)$ for the unit vector in the direction of $\delta_n(i, j)$, note that

$$\|\delta_{n+1}(i, j)\|^2 = \|\delta_n(i, j)\|^2 + \|V\|^2 + 2\|\delta_n(i, j)\|\langle V, e_n(i, j)\rangle,$$

where

(1) $V = V_{n+1}$ if V_{n+1} is assigned to $\mathbf{B}_i$,
(2) $V = -V_{n+1}$ if V_{n+1} is assigned to $\mathbf{B}_j$, and
(3) $V = 0$ otherwise.

8

In particular, under the inner product rule, we have

$$\|\delta_{n+1}(i,j)\|^2 = \|\delta_n(i,j)\|^2 + \|V_{n+1}\|^2 - 2\|\delta_n(i,j)\||\langle V_{n+1}, e_n(i,j)\rangle|$$

if V_{n+1} is assigned to either $\mathbf{B}_i$ or $\mathbf{B}_j$, and $\|\delta_{n+1}(i,j)\| = \|\delta_n(i,j)\|$ otherwise. Hence, if V_{n+1} is assigned to some bin $\mathbf{B}_h$, then

$$S_{n+1} - S_n \le (k-1) - 2 \sum_{i\in[k]:i\neq h} \|\delta_n(h,i)\||\langle V_{n+1}, e_n(h,i)\rangle|. \tag{4}$$

Next, writing $\ell = km^2/2$, note that

$$\mathbb{P}(D(T) \ge m) \le \mathbb{P}\left(\left(\max_{1\le n\le T} S_n\right) \ge \ell\right);$$

indeed, as a consequence of the triangle inequality, we have

$$2\|\delta_n(h,i)\|^2 + 2\|\delta_n(h,j)\|^2 \ge \|\delta_n(i,j)\|^2$$

for any $h, i, j \in [k]$; summing this estimate over all $h \in [k]$, we deduce that

$$\frac{2S_n}{k} \ge \max_{1\le i,j\le k} \|\delta_n(i,j)\|^2.$$

Consequently, writing $E_n(\ell) = \{S_1 < \ell\} \cap \{S_2 < \ell\} \cap \cdots \cap \{S_{n-1} < \ell\} \cap \{S_n \ge \ell\}$, we have

$$\mathbb{P}(D(T) \ge m) \le \sum_{n=1}^T \mathbb{P}(E_n(\ell)).$$

Next, set $r = \ell/2k$. Note that for $n \le r$, we have $S_n \le n(k-1) < \ell$, so it follows that $\mathbb{P}(E_n(\ell)) = 0$. For $n \ge r+1$, we define

$$F_n(\ell) = \left\{\frac{\ell}{2} \le S_{n-r} < \ell\right\} \cap \left\{\frac{\ell}{2} \le S_{n-r+1} < \ell\right\} \cap \cdots \cap \left\{\frac{\ell}{2} \le S_{n-1} < \ell\right\} \cap \{S_n \ge \ell\}.$$

Under the inner product rule, we know (see (4)) that $S_{n+1} - S_n \le k-1$ for all $n \in \mathbb{N}$, so it is clear if $S_n \ge \ell$, then $S_{n-t} \ge \ell/2$ for each $t \in \{0, 1, \ldots, r\}$. Therefore, it is clear that $E_n(\ell) \subset F_n(\ell)$ for each $n \ge r+1$, so

$$\mathbb{P}(D(T) \ge m) \le \sum_{n=r+1}^T \mathbb{P}(F_n(\ell)). \tag{5}$$

We shall estimate $\mathbb{P}(F_n(\ell))$ by studying how our observable can change in a single step using (4), which in turn will allow us to bound $\mathbb{P}(D(T) \ge m)$ using (5). We need slightly different arguments depending on whether or not the underlying distribution μ is well-behaved. The key difference between the two cases is that the probability that the change in our observable in a single step is 'bad' decays

9

with ℓ in the case where μ is well-behaved (see Claim 4.4), but is merely bounded away from 1 in general (see Claim 4.3).

Case 1: Arbitrary distributions. We first establish (1) for an arbitrary probability distribution μ on $\mathbb{B}^d$. We proceed by induction over the dimension. The result is trivial in the case where $d = 1$ as the inner product rule coincides with the trivial one-dimensional strategy described in Section 2. Now, suppose that $d > 1$ and that we have established the required bound in dimension $d - 1$.

The starting point of our argument is the following observation.

Lemma 4.2. *For any probability distribution μ on $\mathbb{B}^d$, either there exists a hyperplane $\mathcal{H} \subset \mathbb{R}^d$ passing through the origin such that $\mu(\mathcal{H} \cap \mathbb{B}^d) = 1$, or there exist disjoint measurable sets $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_d \subset \mathbb{B}^d$ and constants $c, p > 0$ such that*

(1) *for every unit vector $e \in \mathbb{S}^{d-1}$, there exists $i \in [d]$ such that $|\langle x, e \rangle| \geq c$ for all $x \in \mathcal{A}_i$, and*

(2) *$\mu(\mathcal{A}_i) \geq p$ for each $1 \leq i \leq d$.*

Proof. We say that a point $x \in \mathbb{B}^d$ is *μ-heavy* if $\mu(\mathcal{U}) > 0$ for every open neighbourhood $\mathcal{U}$ of x. If the set of μ-heavy points is contained in some hyperplane $\mathcal{H}$ passing through the origin, then it follows by compactness that for any $\varepsilon > 0$, $\mu(\mathcal{H}_\varepsilon) = 1$, where $\mathcal{H}_\varepsilon$ is the set of points at distance less than ε from $\mathcal{H}$; as μ is a probability measure, it follows that $\mu(\mathcal{H}) = 1$. Therefore, we may suppose that there exist μ-heavy points $x_1, x_2, \ldots, x_d$ such that no hyperplane passing through the origin contains all of these points; in other words, we may assume that for every $e \in \mathbb{S}^{d-1}$, there exists an $i \in [d]$ such that $|\langle x_i, e \rangle| > 0$. This implies the conclusion of the lemma, once again by compactness. $\qquad\square$

We now apply Lemma 4.2 to μ: if $\mu(\mathcal{H} \cap \mathbb{B}^d) = 1$ for some hyperplane $\mathcal{H}$ passing through the origin, then we are done by induction; we may therefore assume that there exist $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_d$ and $c, p > 0$ as promised by Claim 4.2.

To bound $\mathbb{P}(F_n(\ell))$ from above, we first estimate, for each $t \geq 0$, the probability of the event

$$I_t(\ell) = \left\{ S_{t+1} \leq S_t - c\sqrt{\ell}/k \right\}.$$

Claim 4.3. *For each $t \geq 0$ and all sufficiently large ℓ, we have*

$$\mathbb{P}(I_t(\ell) \mid \{S_t \geq \ell/2\}) \geq p.$$

Proof. Relabelling the bins if necessary, suppose that the largest distance between a pair of bins at time t is the distance between the bins $\mathbf{B}_1$ and $\mathbf{B}_2$. If $S_t \geq \ell/2$, then it must be the case that $\|\delta_t(1,2)\|^2 \geq S_t/k^2 \geq \ell/2k^2$. We know from Lemma 4.2 that with probability at least p, we have

$$|\langle V_{t+1}, \delta_t(1,2)\rangle| \geq \frac{|\langle V_{t+1}, e_t(1,2)\rangle|\sqrt{\ell}}{\sqrt{2}k} \geq \frac{c}{k}\sqrt{\frac{\ell}{2}}.$$

Consider the bins $\mathbf{B}_\pm$ for which the inner products $\langle V_{t+1}, B_\pm^t\rangle$ are maximal and minimal. By definition, V_{t+1} gets assigned to $\mathbf{B}_-$ under the inner product rule. Now, since

$$\langle V_{t+1}, B_+^t - B_-^t\rangle \geq |\langle V_{t+1}, \delta_t(1,2)\rangle| \geq \frac{c}{k}\sqrt{\frac{\ell}{2}},$$

it follows from (4) that if $S_t \geq \ell/2$, then with probability at least p, we have

$$S_{t+1} \leq S_t + (k-1) - \frac{2c}{k}\sqrt{\frac{\ell}{2}} \leq S_t - \frac{c\sqrt{\ell}}{k},$$

where last inequality holds provide ℓ is sufficiently large; this proves the claim. $\square$

Consider any interval of r steps in which our observable lies in the range $[\ell/2, \ell]$ and note that since our observable increases by at most $k-1$ at each step, there are at most $rk^2/c\sqrt{\ell}$ steps in this interval where our observable decreases by at least $c\sqrt{\ell}/k$. Consequently, if $F_n(\ell)$ holds, then there are at most $rk^2/c\sqrt{\ell} \leq rp/2$ different values of $t \in \{n-r, n-r+1, \ldots, n-1\}$ for which the event $I_t(\ell)$ holds, provided ℓ is sufficiently large. Using the Markov property, we deduce from Claim 4.3 and Proposition 4.1 that for all $n \in \mathbb{N}$, we have

$$\mathbb{P}(F_n(\ell)) \leq \exp(-rp/8).$$

We know that $\mathbb{P}(D(T) \geq m) \leq \sum_{n=1}^{T} \mathbb{P}(F_n(\ell))$, so it is now clear that for all $m \geq 0$, we have

$$\mathbb{P}(D(T) \geq m) \leq T\exp(-rp/8), \tag{6}$$

where $\ell = km^2/2$, $r = \ell/2k$ and $p > 0$ is a constant depending on d and μ alone. It follows from (6) that $D(T) = O((\log T)^{1/2})$ with probability at least $1 - T^{-2}$; the required bound (1) now follows from the Borel–Cantelli lemma.

Case 2: Hölder continuous distributions. We now show how we may improve on (1) for well-behaved distributions. It turns out that if μ is Hölder continuous, then it is possible to say a lot more about how our observable changes in a single step than in the general case.

The starting point in this case is to bound, for each $t \geq 0$, the probability of the event

$$J_t(\ell) = \{S_{t+1} - S_t \geq -k\}.$$

Claim 4.4. *If μ is Hölder continuous, then for each $t \geq 0$,*

$$\mathbb{P}(J_t(\ell) \mid \{S_t \geq \ell/2\}) \leq \frac{C}{\ell^c},$$

where $C, c > 0$ are constants depending on d, k and μ alone.

Proof. If V_{t+1} is assigned to some bin $\mathbf{B}_h$, then, by (4), we have

$$S_{t+1} - S_t \leq (k-1) - 2 \sum_{i \in [k]: i \neq h} \|\delta_t(h, i)\| |\langle V_{t+1}, e_t(h, i)\rangle|.$$

From the triangle inequality,

$$\sum_{i \in [k]: i \neq h} \|\delta_t(h, i)\| \geq \max_{1 \leq i,j \leq k} \|\delta_t(i, j)\| \geq \frac{\sqrt{S_t}}{k}.$$

Hence, if it so happens that

$$|\langle V_{t+1}, e_t(i, j)\rangle| \geq \frac{k^2}{\sqrt{S_t}}$$

for all $1 \leq i, j \leq k$, then

$$S_{t+1} - S_t \leq (k-1) - 2\left(\frac{\sqrt{S_t}}{k}\right)\left(\frac{k^2}{\sqrt{S_t}}\right) < -k.$$

Therefore, it follows that

$$\mathbb{P}(J_t(\ell) \mid \{S_t \geq \ell/2\}) \leq \sum_{1 \leq i,j \leq k} \mathbb{P}\left(|\langle V_{t+1}, e_t(i, j)\rangle| \leq \frac{k^2}{\sqrt{\ell/2}}\right).$$

As μ is a Hölder continuous probability distribution and $\mathbf{V}(\mu)$ is an i.i.d. sequence, the claim now follows from Proposition 3.1. $\qquad\square$

As before, since $S_{n+1} - S_n \leq k - 1$ for all $n \in \mathbb{N}$, it is clear that in any interval of r steps in which our observable lies in the range $[\ell/2, \ell]$, there must exist at least $r/2$ steps in this interval where our observable decreases by at most $k - 1$. Consequently, if $F_n(\ell)$ holds, then there must exist at least $r/2$ different values of $t \in \{n - r, n - r + 1, \ldots, n - 1\}$ for which the event $J_t(\ell)$ holds. Using the Markov property, we deduce from Claim 4.4 that if μ is Hölder continuous, then we have

$$\mathbb{P}(F_n(\ell)) \leq \binom{r}{r/2}\left(\frac{C}{\ell^c}\right)^{r/2}$$

12

for all $n \in \mathbb{N}$. It follows that for all $m \geq 0$, we have

$$\mathbb{P}(D(T) \geq m) \leq T \binom{r}{r/2} \left(\frac{C}{\ell^c}\right)^{r/2}, \tag{7}$$

where $\ell = km^2/2$, $r = \ell/2k$ and $C, c > 0$ are constants depending on d, k and μ alone; a simple calculation using (7) shows that $D(T) = O((\log T / \log \log T)^{1/2})$ with probability at least $1 - T^{-2}$; the required bound (2) in the case where μ is Hölder continuous now follows from the Borel–Cantelli lemma. $\qquad\square$

Note that under the inner product rule, the vector V_{n+1} is only ever assigned to a bin $\mathbf{B}_i$ if B_i^n lies on the convex hull of the set $\{B_1^n, B_2^n, \ldots, B_k^n\}$. Much is known about the convex hulls of random subsets of $\mathbb{R}^d$ (see [7], for example), and it seems possible to us that the Hölder condition in Theorem 2.2 could be relaxed by carefully tracking the convex hull of the bins. However, we cannot altogether do away with some sort of 'well-behavedness' condition: the inner product rule does not match the lower bound in Proposition 2.1 in general, as evidenced by Proposition 2.3 which we prove below.

Proof of Proposition 2.3. Given $\omega \colon \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ that is both increasing and unbounded, we first fix a fast-growing sequence of 'length-scales'. More precisely, we fix a sequence $\mathbf{L} = (L_s)_{s \geq 1}$ of positive reals such that for all $s \in \mathbb{N}$, we have $L_s \geq 2$ and

$$\omega\big(\exp(s^2 L_s^2) - 1\big) \geq 10s.$$

Writing $T_s = \lfloor \exp(s^2 L_s^2) \rfloor$, this construction ensures that we have

$$L_s \geq \frac{10(\log T_s)^{1/2}}{\omega(T_s)}$$

for each $s \in \mathbb{N}$. Having constructed $\mathbf{L}$, we define an atomic probability distribution $\mu_{\mathbf{L}}$ on $\mathbb{B}^2$ with weight $(6/\pi^2)s^{-2}$ on the vector $(1/L_s, -1/2)$ for each $s \in \mathbb{N}$.

We now define μ_ω as follows. A vector drawn from μ_ω is the vector $(0, 1/2)$ with probability $1/3$, uniformly distributed on $\mathbb{B}^2$ with probability $1/3$, and distributed according to $\mu_{\mathbf{L}}$ with probability $1/3$.

We shall show, using an argument analogous to the one used to prove Proposition 2.1, that if we partition $\mathbf{V}(\mu_\omega)$ into two bins $\mathbf{B}_1$ and $\mathbf{B}_2$ using the inner product rule, then

$$\mathbb{P}\left(D(T_s) < \frac{L_s}{10}\right) \leq s^{-2} \tag{8}$$

13

for all sufficiently large $s \in \mathbb{N}$. It is then clear that almost surely,

$$\limsup_{T \to \infty} \left(\frac{D(T)\omega(T)}{(\log T)^{1/2}} \right) \geq 1.$$

We now prove that (8) holds for all sufficiently large $s \in \mathbb{N}$. To this end, fix $s \in \mathbb{N}$ and write $L = L_s$ and $T = T_s = \lfloor \exp(s^2 L^2) \rfloor$. Also, let $\delta_n = B_1^n - B_2^n$ for all $n \in \mathbb{N}$.

As before, we break the set $[T]$ into $r = T/L^2$ disjoint blocks $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_r$ each of length L^2; in other words, for $i \in [r]$, we have $\mathcal{T}_i = \{t+1, t+2, \ldots, t+L^2\}$, where $t = (i-1)L^2$. We say that a block $\mathcal{T}$ is *good* if $\|\delta_n\| \geq L/10$ for some $n \in \mathcal{T}$. For $i \in [r]$, writing $t = (i-1)L^2$, we denote by F_i the event that $\|\delta_t\| < L$. We deduce (8) from the following claim.

Claim 4.5. *For each $i \in [r]$, we have*

$$\mathbb{P}(\mathcal{T}_i \text{ is good} \mid F_i) \geq s^{-L^2}.$$

Proof. We bound the probability of a block being good by showing that in the span of a block, there is a reasonably good chance of walking, using an alternating sequence of the vectors $(0, 1/2)$ and $(1/L, -1/2)$, a distance of about $L/10$ to the right starting from somewhere close to the origin. We make this precise below.

Writing $t = (i-1)L^2$, first consider the event E_1 that there exists a time $P \in \{t+1, t+2, \ldots, t+10L\}$ at which we have $\|\delta_P\| \leq 10$. We claim that

$$\mathbb{P}(E_1 \mid F_i) \geq (100)^{-10L};$$

indeed, this follows from the fact that for all $n \in \mathbb{N}$, we crudely have

$$\mathbb{P}(\{\|\delta_{n+1}\| \leq \|\delta_n\| - 1/10\} \mid \{\|\delta_n\| \geq 10\}) \geq 1/100$$

because V_{n+1} is sampled from the uniform distribution on $\mathbb{B}^2$ with probability $1/3$.

Next, let

$$\mathcal{S} = \{(x, y) : 0 \leq x \leq 1 \text{ and } -1/4 \leq y \leq 0\} \subset \mathbb{R}^2$$

and consider the event E_2 that there exists a time $Q \in \{P, P+1, \ldots, P+10^5\}$ at which we have $\delta_Q \in \mathcal{S}$. It is not hard to check that again, we crudely have

$$\mathbb{P}(E_2 \mid E_1 \cap F_i) \geq 100^{-100}.$$

To see this, note that if P exists, then it is possible to walk, while 'respecting the inner product rule' throughout, from δ_P to the set $\mathcal{S}$ using vectors of norm $1/2$ in

14

at most 100 steps; the claimed bound then follows by 'enlarging' such a walk and using the uniform component of μ_ω.

Finally, consider the event E_3 that the vectors $V_{Q+1}, V_{Q+2}, \ldots, V_{Q+L^2/5}$ are alternately the vectors $(0, 1/2)$ and $(1/L, -1/2)$. It is easy to see from the definition of μ_ω that

$$\mathbb{P}(E_3 \mid E_2 \cap E_1 \cap F_i) = \left(\frac{2}{3\pi^2 s^2}\right)^{\frac{L^2}{10}}.$$

Since $\delta_Q \in \mathcal{S}$ under $E_2 \cap E_1 \cap F_i$, if E_3 also holds, then a simple calculation shows that we have $\delta_{t+1} = \delta_t + V_{t+1}$ for each $t \in \{Q, Q+1, \ldots, Q+L^2/5 - 1\}$ under the inner product rule; consequently, under $E_3 \cap E_2 \cap E_1 \cap F_i$, we have $\delta_{Q+L^2/5} = \delta_Q + (L/10, 0)$. It is now clear, provided $s \in \mathbb{N}$ is sufficiently large, that we have

$$\mathbb{P}(\mathfrak{T}_i \text{ is good} \mid F_i) \geq \mathbb{P}(E_3 \mid E_2 \cap E_1 \cap F_i)\mathbb{P}(E_2 \mid E_1 \cap F_i)\mathbb{P}(E_1 \mid F_i) \geq s^{-L^2}$$

with room to spare. $\qquad \square$

Using the Markov property, we now deduce from Claim 4.5 that

$$\mathbb{P}(D(T) < L/10) \leq \prod_{i=1}^{r}(1 - \mathbb{P}(\mathfrak{T}_i \text{ is good} \mid F_i)) \leq \exp\left(-\frac{Ts^{-L^2}}{L^2}\right) \leq s^{-2},$$

where the last inequality holds provided s is sufficiently large as $T = \lfloor \exp(s^2 L^2) \rfloor$. It is now clear that (8) holds for all sufficiently large $s \in \mathbb{N}$; the proposition follows. $\qquad \square$

5. The best-of-two rule

In this section, we prove Theorem 2.4. Before turning to the proof, let us recall the following classical concentration inequality due to Azuma and Hoeffding.

Proposition 5.1. *Let $(X_t)_{t \geq 0}$ be a supermartingale such that $|X_t - X_{t-1}| \leq C$ for all $t \geq 1$. For all positive integers N and all $m \geq 0$, we have*

$$\mathbb{P}(X_N - X_0 \geq m) \leq \exp\left(\frac{-m^2}{2NC^2}\right). \qquad \square$$

Armed with the Azuma–Hoeffding inequality, we are ready to prove Theorem 2.4.

Proof of Theorem 2.4. To prove the result, we shall first show that the distance between any pair of bins is 'self-correcting' under the best-of-two rule, and then use martingale techniques to track these distances.

We proceed by induction over the dimension. Consider the function $f_\mu \colon \mathbb{S}^{d-1} \to [0,1]$ defined by

$$f_\mu(e) = \int_{\mathbb{B}^d} |\langle x, e \rangle| \, d\mu$$

and define

$$C_\mu = \inf_{e \in \mathbb{S}^{d-1}} f(e).$$

We claim that it suffices to prove the result in the case where $C_\mu > 0$. Indeed, if $C_\mu = 0$, then since f is continuous and $\mathbb{S}^{d-1}$ is compact, we have $f_\mu(e) = 0$ for some $e \in \mathbb{S}^{d-1}$. In other words, if $C_\mu = 0$, then there exists a hyperplane $\mathcal{H} \subset \mathbb{R}^d$ passing through the origin such that $\mu(\mathcal{H} \cap \mathbb{B}^d) = 1$. When $d = 1$, this is equivalent to saying that $\mu(\{0\}) = 1$, in which case the result is trivial. When $d > 1$, by identifying $\mathcal{H} \cap \mathbb{B}^d$ with $\mathbb{B}^{d-1}$, it is clear that μ may be identified with a probability distribution supported on the $(d-1)$-dimensional unit ball, in which case we may proceed inductively.

Now, assume that $C_\mu > 0$. We shall show that with probability at least $1 - T^{-2}$, we have $\|B_1^n - B_2^n\| = O(\log T)$ for all $n \in [T]$; the result then follows from a union bound over all pairs of bins and the Borel–Cantelli lemma.

Let $\delta_n = B_1^n - B_2^n$ and let $A_n = \|\delta_n\|^2$. Our first task will be to estimate the conditional expectation $\mathbb{E}[A_{n+1} - A_n \,|\, A_n]$; we do this as follows.

Recall that given V_{n+1}, we choose two bins $\mathbf{B}_i$ and $\mathbf{B}_j$ uniformly at random from the set of all bins (without replacement) and assign V_{n+1} to $\mathbf{B}_i$, say, if $\langle V_{n+1}, B_i^n \rangle \le \langle V_{n+1}, B_j^n \rangle$. Let E^+ denote the event that the two bins chosen at time $n+1$ are precisely $\mathbf{B}_1$ and $\mathbf{B}_2$ and let E^- denote the event that neither of these two bins is chosen at time $n+1$. Also, for $i \in \{3, \dots, k\}$, let E_i denote the event that two bins chosen at time $n+1$ are $\mathbf{B}_i$ and one of $\mathbf{B}_1$ or $\mathbf{B}_2$.

First, writing e_n for the unit vector in the direction of δ_n, we have

$$\mathbb{E}[A_{n+1} - A_n \,|\, \delta_n, E^+] = \mathbb{E}\big[\|V_{n+1}\|^2\big] - 2 f_\mu(e_n)\sqrt{A_n}$$

from which it follows that

$$\mathbb{E}[A_{n+1} - A_n \,|\, A_n, E^+] \le 1 - 2 C_\mu \sqrt{A_n}. \tag{9}$$

16

Next, as the bins $\mathbf{B}_1$ and $\mathbf{B}_2$ are untouched at time $n+1$ under E^-, we also have $\mathbb{E}[A_{n+1} - A_n \mid A_n, E^-] = 0$. Finally, we observe the following.

Claim 5.2. *For each $3 \leq i \leq k$, we have*

$$\mathbb{E}[A_{n+1} - A_n \mid A_n, E_i] \leq 1.$$

Proof. To simplify notation, let $V = V_{n+1}$, $U_1 = B_1^n$, $U_2 = B_2^n$ and $U_i = B_i^n$. To prove the claim, we decompose E_i into the events

(1) $E_i(-,-) = E_i \cap \{\langle V, U_i - U_1 \rangle \leq 0\} \cap \{\langle V, U_i - U_2 \rangle \leq 0\}$,
(2) $E_i(-,+) = E_i \cap \{\langle V, U_i - U_1 \rangle \leq 0\} \cap \{\langle V, U_i - U_2 \rangle > 0\}$,
(3) $E_i(+,-) = E_i \cap \{\langle V, U_i - U_1 \rangle > 0\} \cap \{\langle V, U_i - U_2 \rangle \leq 0\}$, and
(4) $E_i(+,+) = E_i \cap \{\langle V, U_i - U_1 \rangle > 0\} \cap \{\langle V, U_i - U_2 \rangle > 0\}$.

First, as the vector V is deterministically assigned to the bin $\mathbf{B}_i$ under $E_i(-,-)$,

$$\mathbb{E}[A_{n+1} - A_n \mid A_n, E_i(-,-)] = 0.$$

Next, we claim that

$$\mathbb{E}[A_{n+1} - A_n \mid A_n, E_i(-,+)] \leq 1$$

and that

$$\mathbb{E}[A_{n+1} - A_n \mid A_n, E_i(+,-)] \leq 1.$$

Indeed, under $E_i(-,+)$ for example, it is clear that the best-of-two rule always assigns V to either $\mathbf{B}_i$ or $\mathbf{B}_2$ (but never to $\mathbf{B}_1$); since we also have $\langle V, U_1 - U_2 \rangle > 0$ under $E_i(-,+)$, the claim follows. Finally, under $E_i(+,+)$, the best-of-two rule never assigns V to $\mathbf{B}_i$, and V is equally like to be assigned to either $\mathbf{B}_1$ or $\mathbf{B}_2$ because each of these bins is equally likely to be the other bin selected in addition to $\mathbf{B}_i$. Therefore,

$$\mathbb{E}[A_{n+1} - A_n \mid \delta_n, E_i(+,+)] = \mathbb{E}\big[\|V\|^2\big] + (1/2)\mathbb{E}\Big[2\sqrt{A_n}\langle V, e_n \rangle \mid \delta_n, E_i(+,+)\Big]$$
$$+ (1/2)\mathbb{E}\Big[2\sqrt{A_n}\langle -V, e_n \rangle \mid \delta_n, E_i(+,+)\Big],$$

and consequently,

$$\mathbb{E}[A_{n+1} - A_n \mid A_n, E_i(+,+)] = \mathbb{E}\big[\|V\|^2\big] \leq 1.$$

Putting these facts together, it follows that

$$\mathbb{E}[A_{n+1} - A_n \mid A_n, E_i] \leq 1,$$

proving the claim. □

It is now clear that $\mathbb{E}[A_{n+1} - A_n \,|\, A_n, (E^+)^c] \leq 1$. As $\mathbb{P}(E^+) \geq 1/k^2$, we deduce from (9) that

$$\mathbb{E}[A_{n+1} - A_n \,|\, A_n] \leq 1 - C\sqrt{A_n}, \tag{10}$$

where $C = 2C_\mu/k^2 \leq 1$ is a positive constant depending on d, k and μ alone.

With the benefit of hindsight, let $m = (100 \log T/C)^2$ and denote by F the event that $A_n > 2m$ for some $n \in [T]$. To bound $\mathbb{P}(F)$ from above, we define a collection of stopping times as follows. Let $\mathcal{L}_0 = 0$ and for each $j \in \mathbb{N}$, let

(1) $\mathcal{U}_j = \inf\{n : n \geq \mathcal{L}_{j-1}$ and $A_n \geq m\}$, and
(2) $\mathcal{L}_j = \inf\{n : n \geq \mathcal{U}_j$ and $A_n < m\}$.

If F holds, then it is clear that there exists a $j \in \mathbb{N}$ such that $A_n > 2m$ for some $n \in [\mathcal{U}_j, \mathcal{L}_j \wedge T]$. Let F_j denote the event that there exists an $n \in [\mathcal{U}_j, \mathcal{L}_j]$ such that $A_n > 2m$ and note, by the union bound, that $\mathbb{P}(F) \leq \sum_{j=1}^{T} \mathbb{P}(F_j)$. Therefore, to complete the proof, it suffices to show that $\mathbb{P}(F_j) = o(T^{-3})$ for each $1 \leq j \leq T$.

For concreteness, we show that $\mathbb{P}(F_1) = o(T^{-3})$; the same argument may be used to bound $\mathbb{P}(F_j)$ for any $j \leq T$. In what follows, all inequalities will hold provided T (and hence m) is sufficiently large. Writing $\mathcal{U} = \mathcal{U}_1$ and $\mathcal{L} = \mathcal{L}_1$, we define another stopping time

$$\mathcal{N} = \mathcal{L} \wedge \inf\{n : n \geq \mathcal{U} \text{ and } A_n > 2m\}.$$

Clearly, $\mathbb{P}(F_1) = \mathbb{P}(\mathcal{N} < \mathcal{L})$. Now, set $\ell = C\sqrt{m}/2$ and consider, for $t \geq 0$, the process

$$X_t = A_{t+\mathcal{U}} + t\ell.$$

First, note that for each $t \in [0, \mathcal{N} - \mathcal{U})$,

$$X_{t+1} - X_t = \ell + A_{t+1+\mathcal{U}} - A_{t+\mathcal{U}},$$

so we consequently have

$$|X_{t+1} - X_t| \leq \ell + 1 + 2\sqrt{2m},$$

where the inequality above is immediate from the definition of $\mathcal{N}$. Next, we also have

$$\mathbb{E}[X_{t+1} - X_t \,|\, X_t] = \ell + \mathbb{E}[A_{t+1+\mathcal{U}} - A_{t+\mathcal{U}} \,|\, A_{t+\mathcal{U}}] \leq \ell + 1 - C\sqrt{m}$$

for each $t \in [0, \mathcal{N} - \mathcal{U})$, where the last inequality follows from (10) and the definitions of $\mathcal{U}$ and $\mathcal{L}$. It is now clear that $(X_t)_{t \geq 0}$ with $t \in [0, \mathcal{N} - \mathcal{U}]$ is a supermartingale

with increments bounded by $4\sqrt{m}$. Therefore, for any $N \in [0, \mathcal{L} - \mathcal{U}]$, by the Azuma–Hoeffding inequality, we have

$$\mathbb{P}(\mathcal{N} = N + \mathcal{U}) \leq \mathbb{P}(X_N - X_0 \geq 2m + N\ell - X_0)$$
$$\leq \mathbb{P}(X_N - X_0 \geq m/2 + N\ell)$$
$$\leq \exp\left(\frac{-(m/2 + N\ell)^2}{16Nm}\right) = o(T^{-4}),$$

where the last inequality holds uniformly in N. By applying the union bound over the (at most T) possible values of N, we obtain that $\mathbb{P}(F_1) = o(T^{-3})$. This completes the proof of Theorem 2.4. $\qquad\square$

6. Conclusion

First, it would be nice to know under what conditions (2) holds in general. We have proved this estimate for probability distributions satisfying a Hölder condition. At the other end of the spectrum, the same bound also holds for probability distributions supported on a finite number of atoms; in fact, it can be shown in this case that under the inner product rule, we *deterministically* have $D(T) = O(1)$. We know from Proposition 2.3 that the inner product rule does not match the lower bound in Proposition 2.1 in general, however.

Next, it is worth mentioning that the construction in Proposition 2.3 was designed specifically to be 'bad' for the inner product rule; in particular, this construction does not improve on the strategy-agnostic lower bound in Proposition 2.1. It is therefore an intriguing problem to decide the following: given a probability distribution on the unit ball, does there exist a (distribution-specific) partitioning strategy that matches the lower bound in Proposition 2.1 to within a constant factor? Of course, one can also ask the following (perhaps more difficult) question: is there a universal strategy that matches the lower bound in Proposition 2.1 for every probability distribution on the unit ball?

Finally, it would also be good to improve the implicit constants in our results and quantitatively understand the influence of the number of bins on the problems at hand; indeed, it is natural to expect that the freedom to use more bins should offer better control. A careful analysis of our proofs shows that for the uniform distribution, the lower bound in Proposition 2.1 and the upper bound in Theorem 2.2 differ by a multiplicative factor of k, roughly; bridging this gap remains an interesting problem.

REFERENCES

1. M. Adler, S. Chakrabarti, M. Mitzenmacher, and L. Rasmussen, *Parallel randomized load balancing*, Random Structures Algorithms **13** (1998), 159–188.
2. N. Alon and J. H. Spencer, *The probabilistic method*, 3rd ed., Wiley-Interscience Series in Discrete Mathematics and Optimization, John Wiley & Sons, Inc., Hoboken, NJ, 2008.
3. Y. Azar, *On-line load balancing*, Online algorithms (Schloss Dagstuhl, 1996), Lecture Notes in Comput. Sci., vol. 1442, Springer, Berlin, 1998, pp. 178–195.
4. Y. Azar, A. Z. Broder, and A. R. Karlin, *On-line load balancing*, Theoret. Comput. Sci. **130** (1994), 73–84.
5. Y. Azar, A. Z. Broder, A. R. Karlin, and E. Upfal, *Balanced allocations*, SIAM J. Comput. **29** (1999), 180–200.
6. I. Bárány, *On the power of linear dependencies*, Building bridges, Bolyai Soc. Math. Stud., vol. 19, Springer, Berlin, 2008, pp. 31–45.
7. _____, *Random points and lattice points in convex bodies*, Bull. Amer. Math. Soc. **45** (2008), 339–365.
8. I. Bárány and B. Doerr, *Balanced partitions of vector sequences*, Linear Algebra Appl. **414** (2006), 464–469.
9. T. Bohman, A. Frieze, and E. Lubetzky, *Random triangle removal*, Adv. Math. **280** (2015), 379–438.
10. T. Bohman and P. Keevash, *Dynamic concentration of the triangle-free process*, Preprint, arXiv:1302.5963.
11. B. Doerr and A. Srivastav, *Multicolour discrepancies*, Combin. Probab. Comput. **12** (2003), 365–399.
12. G. Fiz Pontiveros, S. Griffiths, and R. Morris, *The triangle-free process and $r(3, k)$*, Preprint, arXiv:1302.6279.

13. Y. Peres, K. Talwar, and U. Wieder, *The $(1+\beta)$-choice process and weighted balls-into-bins*, Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, Philadelphia, PA, 2010, pp. 1613–1619.

14. E. Steinitz, *Bedingt konvergente Reihen und konvexe Systeme*, J. Reine Angew. Math. **146** (1916), 1–52.

15. B. Vöcking, *How asymmetry helps load balancing*, J. ACM **50** (2003), 568–589.

Departement Mathematik, ETH Zürich, Rämistrasse 101, 8092, Zürich, Switzerland

E-mail address: juhan.aru@math.ethz.ch

Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Wilberforce Road, Cambridge CB3 0WB, UK

E-mail address: b.p.narayanan@dpmms.cam.ac.uk

Mathematical Institute, University of Oxford, Andrew Wiles Building, Radcliffe Observatory Quarter, Woodstock Road, Oxford OX2 6GG, UK

E-mail address: scott@maths.ox.ac.uk

Mircosoft Research, Redmond WA 98052, USA

E-mail address: venkie@microsoft.com