SIAM J. COMPUT.
Vol. 35, No. 6, pp. 1350–1385

BALANCED ALLOCATIONS: THE HEAVILY LOADED CASE*

PETRA BERENBRINK†, ARTUR CZUMAJ‡, ANGELIKA STEGER§, AND BERTHOLD VÖCKING¶

Abstract. We investigate balls-into-bins processes allocating m balls into n bins based on the multiple-choice paradigm. In the classical *single-choice* variant each ball is placed into a bin selected uniformly at random. In a *multiple-choice* process each ball can be placed into one out of $d \geq 2$ randomly selected bins. It is known that in many scenarios having more than one choice for each ball can improve the load balance significantly. Formal analyses of this phenomenon prior to this work considered mostly the lightly loaded case, that is, when $m \approx n$. In this paper we present the first *tight* analysis in the *heavily loaded* case, that is, when $m \gg n$ rather than $m \approx n$.

The best previously known results for the multiple-choice processes in the heavily loaded case were obtained using majorization by the single-choice process. This yields an upper bound of the maximum load of bins of $m/n + \mathcal{O}(\sqrt{m \ln n / n})$ with high probability. We show, however, that the multiple-choice processes are fundamentally different from the single-choice variant in that they have *"short memory."* The great consequence of this property is that the deviation of the multiple-choice processes from the optimal allocation (that is, the allocation in which each bin has either $\lfloor m/n \rfloor$ or $\lceil m/n \rceil$ balls) does not increase with the number of balls as in the case of the single-choice process. In particular, we investigate the allocation obtained by two different multiple-choice allocation schemes, the greedy scheme due to Azar et al. and the always-go-left scheme due to Vöcking. We show that these schemes result in a maximum load of only $m/n + \mathcal{O}(\ln \ln n)$ with high probability. All our detailed bounds on the maximum load are tight up to an additive constant.

Furthermore, we investigate the two multiple-choice algorithms in a comparative study. We present a majorization result showing that the always-go-left scheme obtains a better load balancing than the greedy scheme for any choice of n, m, and d.

Key words. occupancy problems, balls-into-bins processes, randomized resource allocation

AMS subject classifications. 60K30, 68W15, 68W20

DOI. 10.1137/S009753970444435X

1. Introduction.

One of the central topics in the area of randomized algorithms is the study of *occupancy problems* for *balls-into-bins processes*; see, e.g., [1, 2, 3, 5, 6, 8, 9, 10, 11, 12, 15, 17, 20, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 33]. We consider allocation processes in which a set of independent balls representing, e.g., tasks or jobs is assigned to a set of bins representing, e.g., servers or machines. Since the framework of balls-into-bins processes can be used to translate realistic problems into a formal mathematical model in a natural way, it has been frequently analyzed in probability theory [15, 17], random graph theory, and, most recently, in computer science. For example, in theoretical computer science, the balls-into-bins processes found many applications in hashing (see, e.g., [16]) or randomized rounding. They also

*Received by the editors June 9, 2004; accepted for publication (in revised form) August 15, 2005; published electronically March 24, 2006. This research was supported in part by SBR grant 421090, NSF grant CCR-0105701, an NSERC-Discovery grant, DFG grant SFB 342, and DFG grant Vo889/1. An extended abstract of this paper appeared in *Proceedings of the* 32*nd Annual ACM Symposium on Theory of Computing*, Portland, OR, 2000, ACM Press, New York, 2000, pp. 745–754.

http://www.siam.org/journals/sicomp/35-6/44435.html

†School of Computing Science, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada (petra@cs.sfu.ca).

‡Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102 (czumaj@cis.njit.edu).

§Department of Computer Science, ETH Zürich, 8092 Zürich, Switzerland (steger@inf.ethz.ch).

¶Department of Computer Science, RWTH Aachen, 52056 Aachen, Germany (voecking@cs.rwth-aachen.de).

play a crucial role in load balancing and resource allocation in parallel and distributed systems (see, e.g., [1, 10, 18, 31, 32]). The goal of our study is to derive improved upper and lower bounds on the maximum load or even the entire distribution of the load in all bins for balls-into-bins processes in which each ball is placed in a load-adaptive fashion into one out of a small number of randomly chosen bins.

In the classical *single-choice process* (see, e.g., [15, 17, 30]), each ball is placed into a bin chosen *independently and uniformly at random (i.u.r.)*. For the case of n bins and $m \geq n \ln n$ balls it is well known that there exists a bin receiving $m/n+\Theta(\sqrt{m \ln n\,/\,n})$ balls (see, e.g., [30]). This result holds not only in expectation but even with high probability (w.h.p.).[1] Let the *maximum load* denote the number of balls in the fullest bin and let the *max height above average* denote the difference between the maximum load and the average number of balls per bin (which is m/n in our notation). Then the max height above average of the single choice algorithm is $\Theta(\sqrt{m \ln n\,/\,n})$, w.h.p. Observe that the deviation between the randomized single-choice allocation and the optimal allocation increases with the number of balls.

In this paper we investigate randomized multiple-choice allocation schemes (see, e.g., [1, 3, 11, 22, 23, 33]). The idea of *multiple-choice algorithms* is to reduce the maximum load by choosing a small subset of the bins for each ball at random and placing the ball into one of these bins. Typically, the ball is placed into a bin with a minimum number of balls among the d alternatives. It is well known that having more than one choice for each ball can improve the load balancing significantly [1, 33]. Previous analyzes, however, are only able to deal with the *lightly loaded case*, i.e., $m = \mathcal{O}(n)$; the bounds for $m \gg n$ are far off. Our *main contribution* is the first tight analysis for the *heavily loaded case*, i.e., when $m = \omega(n)$. We investigate two different kinds of well-known multiple-choice algorithms, the *greedy scheme* and the *always-go-left scheme*.

- Algorithm Greedy[d] was introduced and analyzed by Azar et al. in [1]. Greedy[d] chooses $d \geq 2$ locations for each ball i.u.r. from the set of bins. The m balls are inserted sequentially, one by one, and each ball is placed into the least loaded among its d locations (if several locations have the same minimum load, then the ball is placed into an arbitrary one among them). Azar et al. [1] show that the max height above average (and the maximum load) is $\ln\ln n/\ln d + \Theta(m/n)$, w.h.p.
- Algorithm Left[d] was introduced and analyzed by Vöcking in [33]. Let n be a multiple of $d \geq 2$. This algorithm partitions the set of bins into d groups of equal size. These groups are ordered from left to right. Left[d] chooses for each ball d locations: one location from each group is chosen i.u.r. The m balls are inserted one by one and each ball is placed into the least loaded among its d locations. If there are several locations having the same minimum load, the ball is always placed into the leftmost group containing one of these locations. Vöcking [33] proved, rather surprisingly, that the use of this unfair tie breaking mechanism leads to a better load balancing than a fair mechanism that resolves ties at random. In particular, the max height above average (and the maximum load) produced by Left[d] is only $\ln\ln n/(d\ln\phi_d) + \Theta(m/n)$, w.h.p., where $1.61 \leq \phi_d \leq 2$.

In the lightly loaded case, the bounds above are tight up to additive constants. In the heavily loaded case, however, these bounds are not even as good as the bounds

[1] Throughout the entire paper we say an event A related to the process of allocating m balls into n bin occurs *with high probability (w.h.p.)* if $\mathbf{Pr}[A] \geq 1 - n^{-\kappa}$ for an arbitrarily chosen constant $\kappa \geq 0$. Notice that this probability does *not* depend on m, the number of balls.

known for the classical single-choice process. In fact, the best known bounds for the multiple-choice algorithms in the heavily loaded case are obtained using majorization from the single-choice process showing only that the multiple-choice algorithms do not behave worse than the single-choice process.

Unfortunately, the known methods for analyzing the multiple-choice algorithms do not allow us to obtain better results for the heavily loaded case. Both the techniques used in [1] ("layered induction") and [33] ("witness trees") inherently assume a load of $2m/n$ already in their base case and therefore they do not seem to be suitable to prove a bound better than $2\,m/n$. Alternative proof techniques using differential equations as suggested in [5, 22, 23, 34, 35] fail for the heavily loaded case, too. The reason is that the concentration results obtained by Kurtz's theorem hold only for a limited number of balls. Therefore, the analysis of the heavily loaded case requires new ideas. Before we proceed with the detailed statement of our results we first provide some terminology.

1.1. Basic definitions and notation. We model the state of the system by *load vectors*. A load vector $u = (u_1, \dots, u_n)$ specifies that the number of balls in the ith bin (the *load* of the ith bin) is u_i. If u is *normalized*, then the entries in the vector are sorted in decreasing order so that u_i describes the number of balls in the ith fullest bin. In the case of Greedy[d], the order among the bins does not matter so that we can restrict the state space to normalized load vectors. In the case of Left[d], however, we need to consider general load vectors.

Suppose X_t denotes the load vector at time t, i.e., after inserting t balls using Greedy[d] or Left[d]. Then the stochastic process $(X_t)_{t\in\mathbb{N}}$ corresponds to a Markov chain $\mathfrak{M} = (\mathbf{X}_t)_{t\in\mathbb{N}}$ whose transition probabilities are defined by the respective allocation process. In particular, X_t is a random variable obeying some probability distribution $\mathcal{L}(X_t)$ defined by the allocation scheme. (Throughout the paper we use the standard notation to denote the probability distribution of a random variable U by $\mathcal{L}(U)$.) We use a standard measure of discrepancy between two probability distributions ϑ and ν on a space Ω, the *variation distance*, defined as

$$\|\vartheta - \nu\| = \frac{1}{2}\sum_{\omega\in\Omega} |\vartheta(\omega) - \nu(\omega)| = \max_{A\subseteq\Omega}(\vartheta(A) - \nu(A)).$$

A basic technique used in this paper is *coupling* (cf., e.g., [4, 7]). A coupling for two (possibly identical) Markov chains $\mathfrak{M}_X = (\mathbf{X}_t)_{t\in\mathbb{N}}$ with state space Ω_X and $\mathfrak{M}_Y = (\mathbf{Y}_t)_{t\in\mathbb{N}}$ with state space Ω_Y is a stochastic process $(X_t, Y_t)_{t\in\mathbb{N}}$ on $\Omega_X \times \Omega_Y$ such that each of $(X_t)_{t\in\mathbb{N}}$ and $(Y_t)_{t\in\mathbb{N}}$ is a faithful copy of $\mathfrak{M}_X$ and $\mathfrak{M}_Y$, respectively.

Another basic concept that we use frequently is *majorization* (cf., e.g., [1]). We say that a vector $u = (u_1, \dots, u_n)$ is *majorized* by a vector $v = (v_1, \dots, v_n)$, written $u \le v$, if for all $1 \le i \le n$ it holds that

$$\sum_{1\le j\le i} u_{\pi(j)} \le \sum_{1\le j\le i} v_{\sigma(j)},$$

where π and σ are permutations of $1, \dots, n$ such that $u_{\pi(1)} \ge u_{\pi(2)} \ge \cdots \ge u_{\pi(n)}$ and $v_{\sigma(1)} \ge v_{\sigma(2)} \ge \cdots \ge v_{\sigma(n)}$. Given an allocation scheme $\mathcal{X}$ defining a Markov chain $\mathfrak{M}_X = (\mathbf{X}_t)_{t\in\mathbb{N}}$ and an allocation scheme $\mathcal{Y}$ defining a Markov chain $\mathfrak{M}_Y = (\mathbf{Y}_t)_{t\in\mathbb{N}}$, we say that $\mathcal{X}$ is *majorized* by $\mathcal{Y}$ if there is a coupling between the two Markov chains $\mathfrak{M}_X$ and $\mathfrak{M}_Y$ such that $X_t \le Y_t$ for all $t \in \mathbb{N}$.

In order to express our results of the always-go-left scheme we use *d-ary Fibonacci numbers*. Define $F_d(k) = 0$ for $k \le 0$, $F_d(1) = 1$, and $F_d(k) = \sum_{i=1}^{d} F_d(k-i)$ for $k \ge 2$.

Let $\phi_d = \lim_{k\to\infty} \sqrt[k]{F_d(k)}$. Then $F_d(k) = \Theta(\phi_d^{\,k})$. Notice that ϕ_2 corresponds to the golden ratio. ϕ_d is called the *d-ary golden ratio.* In general $1.61 < \phi_2 < \phi_3 < \cdots < 2$.

1.2. New results. Our main contribution is the first tight analysis for multiple-choice algorithms assuming an arbitrary number of balls. Our first result is a tight analysis of Greedy[d] in the heavily loaded case when the number of balls is upper-bounded by a polynomial in the number of bins.

LEMMA 1.1. *Let $\beta \geq 1$ be an arbitrary constant. Suppose we allocate m balls to n bins using* Greedy[d] *with $d \geq 2$ and $m \leq n^{\beta}$. Then the number of bins with load at least $\frac{m}{n} + i + \gamma$ is upper bounded by $n \cdot \exp(-d^{\,i})$, w.h.p., where γ denotes a suitable constant, $\gamma = \gamma(\beta)$.*

Even if Lemma 1.1 may seem to be a simple extension of the analysis of the greedy algorithm with $m = \mathcal{O}(n)$ from [1], our analysis is significantly more complicated. The main idea behind the proof is similar to the layered induction approach from [1]. However, the fact that the number of balls is only bounded by a polynomial in the number of bins requires many additional tricks to be applied. In particular, unlike in [1], we have to consider the entire distribution of the bins in our analysis (while in [1] the bins with load smaller than the average could be ignored).

The techniques used to prove Lemma 1.1 cannot be extended to deal with the case when m is unbounded. This is because the analysis in the proof of Lemma 1.1 is based on an inductive argument showing that the bound given in the lemma holds after throwing *any* number $m' \leq m$ of the balls. Of course, this approach cannot work if m is unbounded, because in that case we expect that for some $m' \leq m$ there will be some bins having a huge load, even w.h.p. Therefore, to extend the result of Lemma 1.1 to all values of m we will need other techniques.

Our next result is the main technical contribution of this paper and is central for extending Lemma 1.1 to all values of m. It shows that the multiple-choice processes are fundamentally different from the classical single-choice process in that they have "short memory."

LEMMA 1.2 (Short Memory Lemma). *Let $\varepsilon > 0$. Let $d \geq 2$ be any integer. Let X and Y be any two load vectors describing the allocation of m balls to n bins. Let X_t (Y_t) be the random variable that describes the load vector after allocating t further balls on top of X (Y, respectively) using protocol* Greedy[d]. *Then there is a $\tau = \mathcal{O}(m\,n^6 \ln^4(1/\varepsilon))$ such that $\|\mathcal{L}(X_\tau) - \mathcal{L}(Y_\tau)\| \leq \varepsilon$. In the special case when $d = 2$, this result holds even with $\tau = \mathcal{O}(m\,n^2 + n^4 \ln(m/\varepsilon))$.*

The proof of Lemma 1.2 is done by analyzing the mixing time of the underlying Markov chain describing the load distribution of the bins. Our study of the mixing time is via a new variant of the coupling method (called neighboring coupling (see Lemma 3.2 in section 3.1.2)) which may be of independent interest.

The Short Memory Lemma implies the following property of the Greedy[d] process (see Corollary 4.1 for a precise statement). Suppose that we begin in an arbitrary configuration with the load difference between any pair of bins being at most Δ. Then the Greedy[d] process "forgets" this unbalance in $\Delta \cdot \mathrm{polylog}(\Delta) \cdot \mathrm{poly}(n)$ steps. That is, the allocation after inserting further $\Delta \cdot \mathrm{polylog}(\Delta) \cdot \mathrm{poly}(n)$ balls is stochastically undistinguishable from an allocation obtained by starting from a totally balanced system. This is in contrast to the single-choice process that requires $\Delta^2 \cdot \mathrm{poly}(n)$ steps to "forget" a load difference of Δ. We show that this property implies a fundamental difference between the allocation obtained by the multiple- and the single-choice algorithms. While the allocation of the single-choice algorithm deviates more and more from the optimal allocation with an increasing number of balls, the deviation between the multiple-choice and the optimal allocation is independent of the number of balls. This allows us to concentrate the analysis for the large number of balls m on the case when m is only a polynomial of n.

Next, we show how to incorporate the results above, in Lemmas 1.1 and 1.2, to obtain our main result about the load distribution of Greedy$[d]$.

THEOREM 1.3. *Let γ denote a suitable constant. Suppose we allocate m balls to n bins using* Greedy$[d]$ *with $d \geq 2$. Then the number of bins with load at least $\frac{m}{n} + i + \gamma$ is upper bounded by $n \cdot \exp(-d^{\,i})$, w.h.p.*

This result is tight up to additive constants in the sense that, for $m \geq n$, the number of bins with load at least $\frac{m}{n} + i \pm \Theta(1)$ is also bounded below by $n \cdot \exp(-d^{\,i})$, w.h.p. In particular, Theorem 1.3 implies immediately the following corollary, which is tight up to a constant additive term.

COROLLARY 1.4. *If m balls are allocated into n bins using* Greedy$[d]$ *with $d \geq 2$, then the number of balls in the fullest bin is $\frac{m}{n} + \frac{\ln\ln n}{\ln d} \pm \Theta(1)$, w.h.p. (that is, the max height above average is $\frac{\ln\ln n}{\ln d} \pm \Theta(1)$, w.h.p.).*

Next, we analyze the always-go-left scheme. The load distribution is described in terms of Fibonacci numbers defined in the previous section.

THEOREM 1.5. *Let γ denote a suitable constant. Suppose we allocate m balls into n bins using* Left$[d]$ *with $d \geq 2$. Then the number of bins with load at least $\frac{m}{n} + i + \gamma$ is upper bounded by $n \cdot \exp(-\phi_d^{\,d \cdot i})$, w.h.p.*

Also this bound is tight up to additive constants because the number of bins with load at least $\frac{m}{n} + i \pm \Theta(1)$ is lower-bounded by $n \cdot \exp(-\phi_d^{\,d \cdot i})$, w.h.p., too. Similarly as in the case of the analysis of Greedy$[d]$, Theorem 1.5 immediately implies a tight bound for the maximum load when using Left$[d]$.

COROLLARY 1.6. *If m balls are allocated into n bins using* Left$[d]$ *with $d \geq 2$, then the number of balls in the fullest bin is $\frac{m}{n} + \frac{\ln\ln n}{d \cdot \ln \phi_d} \pm \Theta(1)$, w.h.p. (that is, the max height above average is $\frac{\ln\ln n}{d \cdot \ln \phi_d} \pm \Theta(1)$, w.h.p.).*

In addition to these quantitative results, we investigate the relationship between the greedy and the always-go-left scheme directly.

THEOREM 1.7. Left$[d]$ *is majorized by* Greedy$[d]$.

In other words, we show that the always-go-left scheme produces a (stochastically) better load balancing than the greedy scheme for any possible choice of d, n, and m. We notice also that Theorem 1.7 is the key part of our analysis in Theorem 1.5.

1.3. Outline. In the first part of the paper we present the analysis of the greedy process. We begin in section 2 with the analysis of Greedy$[d]$ for a polynomial number of balls (Lemma 1.1). Next, in section 3, we show that Greedy$[d]$ has short memory (Lemma 1.2). Based on this property, we show in section 4 that our analysis for a polynomial number of balls can be extended to the analysis of the allocation for an arbitrary number of balls (Theorem 1.3).

In the second part of the paper we analyze the always-go-left process. Here we do not prove the short memory property explicitly. Instead, our main tool is majorization of Left$[d]$ by Greedy$[d]$. In section 5, we show this majorization result, Theorem 1.7. In section 6, we analyze the allocation obtained by Left$[d]$ based on the knowledge about the allocation of Greedy$[d]$ to prove Theorem 1.5.

2. The behavior of Greedy$[d]$ in the polynomially loaded case. In this section, we investigate the allocation obtained by Greedy$[d]$ in the polynomially loaded case. In particular, we prove Lemma 1.1. In this theorem it is assumed that the number of balls is polynomially bounded by the number of bins, that is, $m \leq n^{\delta}$ with

$\delta > 0$ denoting an arbitrary constant. The theorem states that there exists a constant $\gamma \geq 0$ such that the number of bins with load at least $\frac{m}{n}+i+\gamma$ is at most $n \cdot \exp(-d^{\,i})$, w.h.p. Recall that the term *w.h.p.* means that an event holds with probability at least $1 - n^{-\kappa}$ for any fixed $\kappa \geq 0$. Of course, the choice of γ has to depend on κ and δ. Observe that if $n < n_0$ for some constant n_0, then the theorem is trivially satisfied by setting $\gamma = n_0^{\delta}$. We will use this observation at several points in our analysis at which certain inequalities hold only under the assumption $n \geq n_0$ for a suitable choice of n_0. The choice of n_0 might depend only on δ and κ.

Without loss of generality, we assume that $m \leq n^{\delta}$ is a multiple of n. We prove the theorem by induction. For this purpose, we divide the set of balls into at most $n^{\delta-1}$ *batches* of size n each. The allocation *at time* t describes the number of balls in the bins after we have inserted the balls of the first t batches, i.e., after placing tn balls, starting with a set of empty bins at time 0. Our proof is by induction on $t \geq 0$. We provide several invariants characterizing a typical distribution of the balls among the bins at time t. We prove by induction that if these invariants hold before allocating a new batch, then they hold also after allocating the batch with probability at least $1 - n^{-\kappa}$ for any given $\kappa > 0$. This implies that the invariants hold with probability at least $1 - n^{-\kappa+(\delta-1)}$ after inserting the last batch because the number of batches is upper-bounded by $n^{\delta-1}$. In other words, we prove that each individual induction step holds w.h.p. which implies that the invariants hold w.h.p. over all steps because the number of induction steps is polynomially bounded.

2.1. Invariants for Greedy[d]. The average number of balls per bin at *time* t (that is, after allocating $t\,n$ balls) is t. The bins with less than t balls are called *underloaded bins* and the bins with more than t balls are called *overloaded bins*. The *number of holes* at time t is defined as the minimal number of balls one would need to add to the underloaded bins so that each bin has load at least t. The *height* of a ball in a bin is defined such that the ith ball allocated to a bin has height i. We investigate the number of holes in the underloaded bins and the number of balls in the overloaded bins. In particular, we show that the following invariants hold w.h.p.

- $L(t)$*:* At time t, there are at most $0.74\,n$ holes.
- $H(t)$*:* At time t, there are at most $0.27\,n$ balls of height at least $t + 3$.

We prove these invariants by an interleaved induction; that is, the proof for $L(t)$ assumes $H(t-1)$ and the proof for $H(t)$ assumes $L(t)$. Notice that since t is the average number of balls per bin at time t, the number of holes at time t corresponds to the number of balls above height t. Thus, invariant $L(t)$ implies that there are at most $0.74\,n$ balls with height $t + 1$ or larger at time t. This property will enable us to show the upper bound on the number of balls with large height given in $H(t)$. In turn, we will see that there is a way to translate the upper bound on the number of balls with large height given in $H(t-1)$ into an upper bound on the number of holes, which will enable us to prove invariant $L(t)$.

Observe that the two invariants above do not directly imply Lemma 1.1. However, they will allow us to split the analysis into two parts: one for the underloaded bins and one for the overloaded bins. These two parts depend on each other only through the invariants L and H. In both of these parts we will specify further invariants. Finally, the invariants for the overloaded bins will imply Theorem 1.1.

Throughout the analysis, we use the following notation. For $i, t \geq 0$, $\alpha_i^{(t)}$ denotes the fraction of bins with load at most $t - i$ at time t, and $\beta_i^{(t)}$ denotes the fraction of bins with load at least $t + i$ at the same time. Figure 1 illustrates this notation.

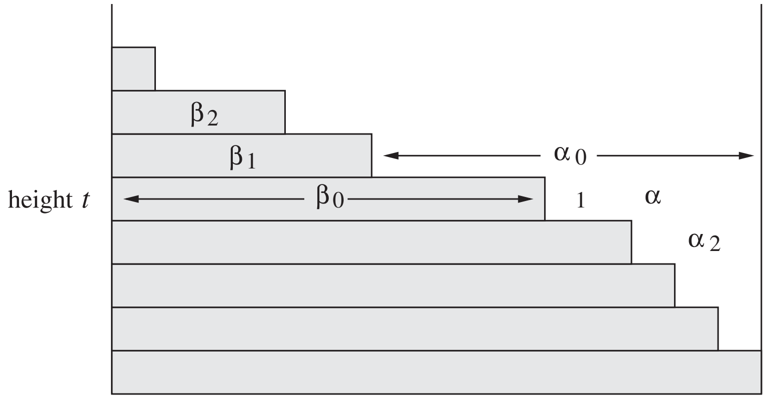

FIG. 1. *Illustration of the terms* $\alpha_i^{(t)}$ *and* $\beta_i^{(t)}$.

2.2. Analysis of the underloaded bins. In this section, we analyze the number of holes in the underloaded bins. We prove the following two invariants for time $t \geq 0$. Let c_1 and c_2 denote suitable constants with $c_1 \leq c_2$.

- $L_1(t)$: For $1 \leq i \leq c_1 \ln n$, $\alpha_i^{(t)} \leq 1.3 \cdot 2.8^{-i}$.
- $L_2(t)$: For $i \geq c_2 \ln n$, $\alpha_i^{(t)} = 0$.

Observe that the invariants $L_1(t)$ and $L_2(t)$ imply the invariant $L(t)$ as the number of holes at time t is at most

$$\sum_{i=1}^{\lfloor c_2 \ln n \rfloor} 1.3 \cdot 2.8^{-\min(i, \lceil c_1 \ln n \rceil)} \cdot n \leq 0.74\, n,$$

where the last inequality holds if $n \geq n_0$ for suitably chosen constant term n_0. In the following, we prove that $L_1(t)$ and $L_2(t)$ hold w.h.p. Our analysis is focused on Greedy[2]; that is, we explicitly prove that the invariants $L_1(t)$ and $L_2(t)$ hold for Greedy[2]. Given that the invariants hold for $d = 2$, a majorization argument [1, Theorem 3.5] implies that the same invariants hold for every $d \geq 2$. In fact, the same argument shows that the choice of the tie breaking mechanism of Greedy[2] is irrelevant. Therefore, without loss of generality, we can assume that Greedy[2] breaks ties among bins of the same height by flipping a fair coin. Under this assumption, we have the following lemma.

LEMMA 2.1. *Let ℓ be an arbitrary integer and assume that at some point in time there exist (at most) $a_\ell\, n$ bins with at most ℓ balls and (at most) $a_{\ell-1}\, n$ bins with at most $\ell - 1$ balls. Suppose that b is a bin with load exactly ℓ. Then, the probability that the next ball allocated by* Greedy[2] *will be placed into bin b is (at least) $(2 - a_\ell - a_{\ell-1})/n$.*

Proof. Since the term $(2 - a_\ell - a_{\ell-1})/n$ is decreasing in both a_ℓ and $a_{\ell-1}$, we can assume without loss of generality that there are exactly $a_\ell\, n$ bins with at most ℓ balls and exactly $a_{\ell-1}\, n$ bins with at most $\ell - 1$ balls. The probability that the ball goes to one of the bins with load ℓ is

$$(a_\ell - a_{\ell-1})^2 + 2(a_\ell - a_{\ell-1})(1 - a_\ell) \;=\; (a_\ell - a_{\ell-1}) \cdot (2 - a_\ell - a_{\ell-1}),$$

because this event happens if and only if either both random locations of the ball point to a bin with load ℓ or at least one of them points to a bin with load ℓ and the other to a bin with load larger than ℓ. Now, since we assume a random tie breaking mechanism, each of the bins with load ℓ is equally likely to receive the ball. Thus, given that the ball falls into one of these bins, the probability that b receives the ball is $\frac{1}{(a_\ell - a_{\ell-1})\,n}$ because the number of bins with load ℓ is $(a_\ell - a_{\ell-1})\,n$. Multiplying the two probabilities yields the lemma. ∎

Combining Lemma 2.1 with invariant L_1, we can now analyze the probability that a ball from batch t falls into a fixed bin b with a given number of holes. Applying invariant $L_1(t-1)$, there are at most $\alpha_i^{(t-1)} n \leq 1.3 \cdot 2.8^{-i} \cdot n$ bins with load at most $(t-1)-i$ at time $t-1$ for every $1 \leq i \leq c_1 \ln n$. This upper bound not only holds at the time immediately before the first ball of batch t is inserted but can be applied to any ball from this batch since the load of a bin is nondecreasing over time. Now, applying Lemma 2.1 yields that the probability that a ball from batch t is assigned to a bin with load at most $(t-1)-i$ is at least

$$\frac{2 - \alpha_i^{(t-1)} - \alpha_{i+1}^{(t-1)}}{n} \;\geq\; \frac{2 - 1.3 \cdot 2.8^{-i} - 1.3 \cdot 2.8^{-(i+1)}}{n}.$$

For $i \geq 3$, this probability is larger than $1.9/n$, which yields the following observation.

Observation 2.2. *The probability that a ball from batch t goes to any fixed bin with load at most $t-4$ at the ball's insertion time is at least $1.9/n$, unless invariant $L_1(t-1)$ fails.*

Thus bins with load $t-4$ or less have almost twice the probability of receiving a ball than the average. This might give an intuition as to why none of the bins falls far behind. The following analysis puts this intuition into formal arguments.

Lemma 2.3. *Let $t \geq 0$. Suppose the probability that one of the invariants $L_1(0), \ldots, L_1(t-1)$ fails is at most $n^{-\kappa'}$ for $\kappa' \geq 1$. For any fixed $\kappa > 0$, there exist constants c_0, c_1, c_2, c_3 (solely depending on κ) such that*

- *there are at most $n \cdot 0.18 \cdot 3^{-i+2}$ bins containing at most $t-i$ balls, for $c_0 < i \leq c_1 \ln n$, and*
- *every bin contains at least $t - c_2 \ln n$ balls,*

with probability at least $1 - n^{-\kappa} - n^{-\kappa'}$, provided $n \geq c_3$.

Proof. Consider a bin b. For any integer $s \geq 0$, let ℓ_s denote the load of bin b at time s, and let q_s be the number of holes in bin b at time s; that is, $q_s = s - \ell_s$. Now, suppose $q_t \geq i+4$, for $i \geq 0$. Since the number of holes can increase by at most one during the allocation of a batch, there exists an integer t', $0 \leq t' < t$, such that $q_{t'} = 4$ and for all $s \in \{t'+1, \ldots, t\}$ it holds that $q_s \geq 4$. Observe that by this definition of t', the number of balls from the batches $t'+1, \ldots, t$ that are assigned to bin b is at most $t - t' - i$.

The definition of t' implies that the bin b has at least four holes at any time during the allocation of the batches $t'+1, \ldots, t$. More formally, at the insertion time of any ball from batch $s \in \{t'+1, \ldots, t\}$, the bin b contains at most $s-4$ balls. Thus, by Observation 2.2, it follows that every ball from these batches has probability at least $1.9/n$ to be assigned to bin b or there exist $s < t$ such that invariant $L_1(s)$ fails. Ignoring the latter event, the number of balls from these batches that are placed into bin b is stochastically dominated by a binomially distributed random variable $\mathrm{B}((t-t')n, 1.9/n)$. However, we cannot simply condition on $L_1(0, \ldots, t-1)$ as this gives an unwanted bias to the random experiments under consideration. Instead we explicitly exclude the event $\neg L_1(0, \ldots, t-1)$ from our considerations. This way, we obtain

$$\begin{aligned}\mathbf{Pr}[(q_t \geq i+4) \wedge L_1(0,\ldots,t-1)] &\leq \sum_{t'=0}^{t-1} \mathbf{Pr}[\mathrm{B}((t-t')\,n, 1.9/n) \leq t-t'-i]\\ &\leq \sum_{\tau=1}^{\infty} \mathbf{Pr}[\mathrm{B}(\tau\,n, 1.9/n) \leq \tau - i]\\ &= \sum_{\tau=1}^{\infty}\sum_{k=i}^{\tau} \mathbf{Pr}[\mathrm{B}(\tau\,n, 1.9/n) = \tau - k].\end{aligned}$$

Next, for $0 \leq k < \tau$, we obtain

$$\begin{aligned}\mathbf{Pr}[\mathrm{B}(\tau\,n, 1.9/n) = \tau - k] &= \binom{\tau\,n}{\tau-k} \cdot \left(1 - \frac{1.9}{n}\right)^{\tau\,n-\tau+k} \cdot \left(\frac{1.9}{n}\right)^{\tau-k}\\ &\leq \frac{(1.9\,e\,\tau)^{\tau-k}}{(\tau-k)!}\ e^{-(\tau\,n-\tau+k)\,1.9/n}\\ &\leq \frac{(1.9\,\tau)^{\tau-k}}{(\tau-k)^{\tau-k}}\ e^{-0.9\,\tau-k+1.9\,\tau/n}\\ &\leq \left(\frac{1.9\,\tau}{\tau-k}\right)^{\tau-k}\ e^{-0.89\,\tau-k},\end{aligned}$$

where the last inequality holds for $n \geq 190$. Now, set $z = \tau/k > 1$. Then, we obtain for $k > 0$

$$\begin{aligned}\mathbf{Pr}[\mathrm{B}(\tau\,n, 1.9/n) = \tau - k] &\leq \left(\frac{1.9\,z\,k}{(z-1)\,k}\right)^{(z-1)\,k}\ e^{-0.79\,z\,k-0.1\,\tau-k}\\ &= \left(\frac{1.9\,z}{z-1}\right)^{(z-1)\,k}\ e^{-0.79\,(z-1)\,k-1.79\,k-0.1\,\tau}\\ &= \left(e^{-1.79}\ \left(\frac{1.9\,z}{e^{0.79}\,(z-1)}\right)^{(z-1)}\right)^{k}\ e^{-0.1\,\tau}.\end{aligned}$$

The term $\left(\frac{1.9\,z}{e^{0.79}\,(z-1)}\right)^{(z-1)}$, $z \geq 1$, takes its maximum at $z = 2.22\ldots$ and is bounded from above by 1.74. Therefore,

$$\mathbf{Pr}[\mathrm{B}(\tau\,n, 1.9/n) = \tau - k] \ \leq\ \left(1.74 \cdot e^{-1.79}\right)^{k} \cdot e^{-0.1\,\tau} \ \leq\ 3.4^{-k} \cdot e^{-0.1\,\tau}$$

for $0 < k < \tau$. The same inequality also holds for $k = 0$, because for $n \geq 13$ it holds that

$$\begin{aligned}\mathbf{Pr}[\mathrm{B}(\tau\,n, 1.9/n) = \tau - 0] &= \mathbf{Pr}[\mathrm{B}(\tau\,n, 1.9/n) = \tau]\\ &= \binom{\tau\,n}{\tau} \cdot \left(1 - \frac{1.9}{n}\right)^{\tau\,(n-1)} \cdot \left(\frac{1.9}{n}\right)^{\tau}\\ &\leq \left(\frac{e\,\tau\,n}{\tau}\right)^{\tau} e^{1.9\,\tau\,\frac{n-1}{n}} \cdot \left(\frac{1.9}{n}\right)^{\tau}\\ &= \left(\frac{1.9\,e}{e^{1.9\,\frac{n-1}{n}}}\right)^{\tau}\\ &\leq e^{-0.1\,\tau}.\end{aligned}$$

Finally, we can obtain the same bound for $k = \tau$. In this case,

$$\begin{aligned}\mathbf{Pr}[\mathrm{B}(\tau\, n, 1.9/n) = \tau - k] &= \mathbf{Pr}[\mathrm{B}(\tau\, n, 1.9/n) = 0]\\ &= \left(1 - \frac{1.9}{n}\right)^{\tau\, n}\\ &\le e^{-1.9\,\tau}\\ &\le 3.4^{-k}\cdot e^{-0.1\,\tau}.\end{aligned}$$

Substituting this bound into the upper bound for $\mathbf{Pr}[q_t \ge i+4]$ gives

$$\mathbf{Pr}[(q_t \ge i+4) \wedge L_1(0,\ldots,t-1)] \;\le\; \sum_{\tau\ge 1}\sum_{k\ge i} 3.4^{-k}\cdot e^{-0.1\tau} \;\le\; 13.5\cdot 3.4^{-i}.$$

Now let Q_t denote the maximum number of holes over all bins at time t. Recall that the probability for the event $\neg L_1(0,\ldots,t-1)$ is bounded from above by $n^{-\kappa'}$. Consequently,

$$\mathbf{Pr}[Q_t \ge i+4] \le n^{-\kappa'} + \mathbf{Pr}[(Q_t \ge i+4)\wedge L_1(0,\ldots,t-1)] \;\le\; n^{-\kappa'} + n\cdot 13.5\cdot 3.4^{-i}.$$

It follows $Q_t = \mathcal{O}(\log n)$, w.h.p. More specifically, for every $\kappa \ge 0$, there exists a constant c_2 such that every bin contains at least $t - c_2 \ln n$ balls, with probability $1 - n^{-\kappa'} - \frac{1}{2}n^{-\kappa}$. This yields the second statement given in the lemma. In the following, we show that the first statement holds with probability at least $1 - \frac{1}{2}n^{-\kappa}$. In particular, we prove that for any given $\kappa > 0$, there are constants c_0, c_1, c_3 such that, for $c_0 < i \le c_1 \ln n$ and $n \ge c_3$, there are at most $0.18\cdot 3^{-i+2}\cdot n$ bins containing $t-i$ or less balls at time t, with probability at least $1 - n^{-\kappa-1} \ge 1 - (2c_1\, n^{\kappa}\ln n)^{-1}$. This way, the probability that one of the statements listed in the lemma fails is at most $n^{-\kappa'} + n^{-\kappa}$, so that the lemma is shown.

Let $c_0 = 40$. For $i > c_0$, we obtain

$$\mathbf{Pr}[q_t \ge i] \;\le\; n^{-\kappa'} + 13.5\cdot 3.4^{-i+4} \;\le\; n^{-1} + 0.65\cdot 2.8^{-i}.$$

Now let X_b be an indicator random variable that is 1 if bin b holds at most $t-i$ balls, and that is 0, otherwise. Define $X = \sum X_b$. We have to prove that $X \;\le\; 1.3\cdot 2.8^{-i}\cdot n$, w.h.p.

Let us first notice that

$$\mathbf{E}[X] \;\le\; n\,\cdot \mathbf{Pr}[q_t \ge i] \;\le\; 0.65\cdot 2.8^{-i}\cdot n + 1.$$

The random variables X_b are "negatively associated" in the sense of [14, Proposition 7]. Hence, we can apply a Chernoff bound to these variables: For every $\mu \ge \mathbf{E}[X]$, $\mathbf{Pr}[X \ge 2\,\mu] \le e^{-\mu/2}$. We choose $\mu = 0.65\cdot 2.8^{-i}\cdot n + 1$, set $c_1 = 0.5$, and assume that c_3 is sufficiently large so that $\mu \ge 2\,(\kappa+1)\ln n$ for every $i \le c_1 \ln n$ and $n \ge c_3$. This yields

$$\mathbf{Pr}[X \ge 1.3\cdot 2.8^{-i}\cdot n + 2] \;\le\; \mathbf{Pr}[X \ge 2\,\mu] \;\le\; e^{-\mu/2} \;\le\; e^{-(\kappa+1)\,\ln n} \;=\; n^{-\kappa-1}.$$

This completes the proof of Lemma 2.3. $\square$

The second part of the lemma corresponds to invariant $L_2(t)$, and the first part corresponds to invariant $L_1(t)$, but only for $i > c_0$. Thus, it remains to show invariant $L_1(t)$ for $1 \le i \le c_0$; that is, we have to show $\alpha_i^t \le 1.3\cdot 2.8^{-i}$. We will solve this problem by specifying a recursive formula upper-bounding the term $\alpha_i^{(t)}$, $1 \le i \le c_0$, in terms of the vector $(\alpha_{i-1}^{(t-1)}, \ldots, \alpha_{i+3}^{(t-1)})$.

LEMMA 2.4. *Let $\epsilon > 0$ and $a_0, \ldots, a_4$ be constant reals with $0 < a_4 < \cdots a_0 < 1$. Let k be a constant integer. Let ℓ denote any integer. Suppose for $i = 0, \ldots, 4$ there are at most $a_i\, n$ bins with load at most $\ell - i$ at time $t-1$. Then, at time t, the number of bins with load at most ℓ is at most $g_k(0) \cdot n$, w.h.p., where the function g is defined by*

$$g_j(i) \;=\; \begin{cases} a_i & \text{if } j = 0 \text{ or } i = 4, \\ (1+\epsilon) \cdot (g_{j-1}(i+1) + (g_{j-1}(i) - g_{j-1}(i+1)) \cdot E) & \text{otherwise,} \end{cases}$$

where

$$E \;=\; \exp\left(-\frac{2 - g_{j-1}(i+1) - g_{j-1}(i)}{k}\right).$$

Proof. For the time being, let us assume that n is a multiple of k. We divide the allocation of the n balls into k phases, into each of which we insert n/k balls using Greedy[2]. For $0 \le i \le 4$ and $1 \le j \le k$, we show that $n \cdot g_j(i)$ is an upper bound on the number of bins with load at most $\ell - i$ after phase j. Observe that this claim is trivially satisfied for $i = 4$.

We perform an induction on j, the index of the phase. Observe that for $0 \le i \le 4$, $n \cdot g_0(i)$ is an upper bound on the number of bins with load at most $\ell - i$ at the beginning of phase 1. In the inductive hypothesis we assume that $n \cdot g_{j-1}(i)$ is an upper bound on the number of bins with load at most $\ell - i$ at the beginning of phase $j \ge 1$. Now consider the allocation of the n/k balls in phase j. Suppose b is a bin having load $\ell - i$ $(0 \le i \le 3)$ at the beginning of that phase. Lemma 2.1 implies that the probability that b receives none of the next n/k balls is at most

$$\left(1 - \frac{2 - g_{j-1}(i) - g_{j-1}(i+1)}{n}\right)^{n/k} \;\le\; \exp\left(-\frac{2 - g_{j-1}(i+1) - g_{j-1}(i)}{k}\right) \;=\; E.$$

Thus, the expected number of bins with load exactly $\ell - i$ not receiving a ball in this phase is at most

$$n \cdot (g_{j-1}(i) - g_{j-1}(i+1)) \cdot E.$$

As a consequence, the expected fraction of bins containing at most $\ell - i$ balls at the end of phase k is upper-bounded by

$$g_{j-1}(i+1) + (g_{j-1}(i) - g_{j-1}(i+1)) \cdot E$$

for $0 \le i \le 3$. This term, however, by the definition, is equivalent to $g_j(i)/(1+\epsilon)$, and hence the expected number of bins containing at most $\ell - i$ balls is upper-bounded by $n\, g_j(i)/(1+\epsilon)$. Applying Azuma's inequality to this expectation, we can observe that the deviation is at most $o(n)$, w.h.p. Hence, since $n \cdot g_j(i) \ge a_4 \cdot n = \Theta(n)$, we conclude that the stochastic deviation can be bounded by a factor $(1+\epsilon)$ for any $\epsilon > 0$, provided n is sufficiently large. We conclude that the fraction of bins containing at most $\ell - i$ balls at the end of phase k is at most $g_j(i)$, w.h.p.

Now let us consider the case that n is not a multiple of k. In this case, we can upper-bound the probability that b receives none of the at least $n/k - 1$ balls from the next batch by

$$\left(1 - \frac{2}{n}\right)^{-1} \cdot E$$

instead of E as before. Obviously the leading factor in front of the E can be made arbitrarily small by choosing n sufficiently large so that it is covered by the $1+\epsilon$ factor that we already used above to cover stochastic deviations. This completes the proof of the lemma. $\square$

In the following, we apply the recursive formula given in Lemma 2.4 to prove that invariant $L_1(t)$ holds for every $i \in \{1, \ldots, c_0 - 1\}$. The recursion gives an upper bound for $\alpha_i^{(t)}$ in terms of the vector $(a_0, \ldots, a_4) = (\alpha_{i-1}^{(t-1)}, \ldots, \alpha_{i+3}^{(t-1)})$. For $i \geq 2$, the terms $\alpha_{i-1}^{(t-1)}, \ldots, \alpha_{i+3}^{(t-1)}$ can be estimated using invariant $L_1(t-1)$. Suppose this invariant is given. Fix any $i \in \{2, \ldots, c_0 - 1\}$. For $i' = 0, \ldots, 4$, we set $a_{i'} = 1.3 \cdot 2.8^{-(i+i'-1)}$. Then, $a_{i'}$ is an upper bound on the fraction of bins with load at most $i - i'$ at time $t-1$. Now we choose $k = 20$ and $\epsilon = \frac{1}{1000}$, and we numerically calculate $g_k(0)$ using Maple. For such a choice of parameters, we obtain $g_k(0) \leq 1.3 \cdot 2.8^{-i}$. By Lemma 2.4, $g_k(0)$ is an upper bound on $\alpha_i^{(t)}$, w.h.p. Thus, invariant $L(t)$ is shown for $i \geq 2$. Unfortunately, this approach does not work for $i = 1$, because in that case a_0 corresponds to $\alpha_0^{(t-1)}$, which is not covered by invariant $L_1(t-1)$. In what follows, we prove another lemma that gives an upper bound on $\alpha_0^{(t-1)}$ based on invariant $H(t-1)$.

LEMMA 2.5. *Suppose $H(t-1)$ is fulfilled. Let $(a_0, \ldots, a_4) := (\alpha_0^{(t-1)}, \ldots, \alpha_4^{(t-1)})$. Then*

$$a_0 \ \leq \ 1 - \frac{a_1 + a_2 + a_3 + a_4 - 0.27}{2}.$$

Proof. At any time $\tau \geq 0$, the number of holes at time τ is $A_\tau = \sum_{j \geq 1} \alpha_j^{(\tau)} n$. Since the number of balls above the average height is equal to the number of holes, we can conclude that A_τ also corresponds to the number of balls with height at least $\tau+1$ at time τ. Now, suppose invariant $H(\tau)$ holds. Then, there are at most $B_\tau = 0.27n$ balls of height at least $\tau + 3$ at time τ. Combining these two bounds, the number of balls with height either $\tau+1$ or $\tau+2$ is lower-bounded by $A_\tau - B_\tau$. This implies that at least $(A_\tau - B_\tau)/2$ bins contain more than τ balls at time τ. As a consequence, the number of bins containing at most τ balls is upper-bounded by $n - (A_\tau - B_\tau)/2$. Hence,

$$\alpha_0^{(\tau)} n \ \leq \ n - \frac{A_\tau - B_\tau}{2} \ \leq \ n \cdot \left(1 - \frac{\sum_{j=1}^{4} \alpha_j^{(\tau)} - 0.27}{2} \right).$$

Finally, setting $\tau = t-1$ and $\alpha_j^{(\tau)} = a_j$ gives the lemma. $\square$

Now, we are ready to prove invariant $L_1(t)$ by showing $g_k(0) \leq 1.3 \cdot 2.8^{-1}$ for all choices of $a_{i'} \in [0, 1.3 \cdot 2.8^{-i'}]$, $1 \leq i' \leq 4$, and $a_0 \in [0, 1 - \frac{1}{2}(a_1 + a_2 + a_3 + a_4 - 0.27)]$. Again, we check the condition on $g_k(0)$ numerically using Maple. For this purpose we need to discretize the domains of the a_i's. For the discretization, we use the monotonicity of $g_k(0)$: the term $g_k(0)$ is monotonically increasing in each of the terms $a_0, \ldots, a_4$. Therefore, it suffices to check the parameters $a_1, \ldots, a_4$ in discrete steps of a suitable size $\delta > 0$ while assuming

$$a_0 \ = \ 1 - \frac{a_1 + a_2 + a_3 + a_4 - 0.27 - 4\,\delta}{2}.$$

In fact, we choose $k = 20$, $\epsilon = \frac{1}{1000}$, and $\delta = \frac{1}{400}$. The numerical calculations with Maple show that the above condition on $g_k(0)$ is satisfied in all cases. Hence, the invariants $L_1(t)$, $L_2(t)$, and, thus, $L(t)$ are shown.

2.3. Analysis of the overloaded bins. In this section, we will analyze the distribution of load in the overloaded bins. In particular, we prove invariant $H(t)$ stating that the number of balls with height at least $t+3$ is at most $0.27\,n$. Our analysis is based on invariant $L(t)$ which we proved in the previous section based on $H(t-1)$. Thus $H(t-1)$ yields $L(t)$ and, in turn, $L(t)$ yields $H(t)$.

Our analysis of invariant $H(t)$ is obtained by the analysis of two further invariants that will imply invariant $H(t)$ and will yield the proof of Lemma 1.1. These invariants are defined as follows. Let

$$h(i) \;=\; 0.7 \cdot 0.53^{d^{i-2}}.$$

Let ℓ denote the smallest integer i such that $h(i) \le n^{-0.9}$. Let $b \ge 1$ denote a suitable constant, whose value will be specified later. For $i \ge 3$, define

$$f(i) = \begin{cases} h(i) & \text{for } 2 \le i < \ell, \\ \max\{h(i), \frac{1}{4}\, n^{-0.9}\} & \text{for } i = \ell, \\ b\, n^{-1} & \text{for } i = \ell + 1. \end{cases}$$

We will prove that the following invariants hold w.h.p.

- $H_1(t)$: $\beta_i^{(t)} \le f(i)$ *for* $2 \le i \le \ell$,
- $H_2(t)$: $\sum_{i>\ell} \beta_i^{(t)} \le b\, n^{-1}$.

Roughly speaking, invariant H_1 states that the sequence $\beta_2, \beta_3, \dots, \beta_\ell$ decreases doubly exponentially, and the number of balls on layer ℓ is upper-bounded by $\frac{1}{4}\, n^{0.1}$. Furthermore, invariant H_2 states that there is only a constant number of balls with a height larger than ℓ. These two invariants imply the bounds given in Lemma 1.1. Furthermore, these invariants imply the invariant $H(t)$ as they upper-bound the number of balls above layer $t+3$ by

$$\sum_{i=3}^{\ell-1} h(i)\, n + \frac{1}{4}\, n^{-0.1} + b \;\le\; 0.26\, n + \frac{1}{4}\, n^{-0.1} + b \;\le\; 0.27\, n,$$

where the last inequality holds for $n \ge 50b + 78$. We show the invariants H_1 and H_2 by induction. Our induction assumptions are $H_1(0), \dots, H_1(t-1)$, $H_2(t-1)$, and $L(t)$. We prove that these assumptions imply $H_1(t)$, $H_2(t)$, and $H(t)$, w.h.p. Our analysis starts by summarizing some properties of the function f. We assume that n is sufficiently large.

CLAIM 2.6.
A1. $f(2) = 0.371$;
A2. $f(i) \ge 0.53^{-2} f(i+1)$ *for* $3 \le i \le \ell$;
A3. $f(i) \ge 0.7^{-1} f(i-1)^d$ *for* $3 \le i \le \ell$;
A4. $f(i) \ge \frac{1}{4} n^{-0.9}$ *for* $3 \le i \le \ell$.

Proof. The properties A1 and A4 follow directly from the definition of f. Property A2 can be seen as follows. For $3 \le i \le \ell - 2$, $f(i) = h(i)$ as well as $f(i+1) = h(i+1)$. Thus,

$$f(i+1) \;=\; 0.7 \cdot 0.53^{d^{i-1}} \;=\; 0.7 \cdot 0.53^{d^{i-2} \cdot d} \;\le\; 0.7 \cdot 0.53^{d^{i-2}+2} \;\le\; 0.53^2 \cdot f(i).$$

If $i = \ell - 1$, then $f(i+1) = h(i+1)$ or $f(i+1) = \frac{1}{4} n^{-0.9}$. In the former case, we can apply the same argument as before. In the latter case, we apply $f(i) \ge n^{-0.9}$ which immediately yields $f(i)/f(i+1) > 4 \ge 0.53^{-2}$. Finally, in the case $i = \ell$, we need the assumption that n is sufficiently large so that $f(\ell)/f(\ell+1) \geq \frac{1}{4}n^{-0.9}/(bn) \geq 0.53^{-2}$. It remains to prove property A3. For $3 \leq i \leq \ell$,

$$f(i-1)^d \;=\; (0.7 \cdot 0.53^{d^{i-3}})^d \;= 0.7^d \cdot 0.53^{d^{i-2}} \leq 0.7 \cdot f(i).$$

Now we use these properties to show $H_1(t)$ using a "layered induction" on the index i, similar to the analysis presented in [1]. For the base case, i.e., $i = 2$, we apply invariant $L(t)$. This invariant yields that, at time t, there exist at most $0.74\,n$ balls of height larger than t. Consequently, the number of bins with $t+2$ or more balls is at most $0.74\,n/2 = 0.37\,n$. Applying property A1 yields $\beta_2^{(t)} \leq 0.37 \leq f(2)$. Thus, invariant $H_1(t)$ is shown for $i = 2$.

Next we prove $H_1(t)$ for $i \in \{3, \ldots, \ell\}$. Fix i, $3 \leq i \leq \ell$. We assume that $H_1(t)$ holds for $i-1$. Let q denote the number of bins that contain $t+i$ or more balls at time $t-1$, i.e., immediately before batch t is inserted, and let p denote the number of balls from batch t that are placed into a bin containing at least $t+i-1$ balls at time t. Observe that $\beta_i^{(t)} \cdot n \leq q + p$. Thus, it suffices to show $q + p \leq f(i) \cdot n$. Applying induction assumption $H_1(t-1)$, we obtain

$$q \;\leq\; \beta_{i+1}^{(t-1)} \cdot n \;\leq\; f(i+1) \cdot n \;\overset{\text{(A2)}}{\leq}\; 0.53^2 \cdot f(i) \cdot n.$$

Bounding p requires slightly more complex arguments. For $3 \leq i \leq \ell$, the probability that a fixed ball of batch t is allocated to height $t+i$ is at most $f(i-1)^d$. This is because each of its locations has to point to one of the bins with $t+i-1$ or more balls, and by our induction on i, the fraction of these bins is bounded from above by $f(i-1)$. Taking into account all n balls of batch t, we obtain

$$\mathbf{E}[p] \;\leq\; f(i-1)^d \cdot n \;\overset{\text{(A3)}}{\leq}\; 0.7 \cdot f(i) \cdot n.$$

Applying a Chernoff bound yields, for every $\epsilon \in (0, 1]$,

$$\begin{aligned} \mathbf{Pr}[p \;\geq\; (1+\epsilon) \cdot 0.7 \cdot f(i) \cdot n] \;&\leq\; \exp\left(\frac{-0.7\,\epsilon^2}{2} \cdot f(i) \cdot n\right) \\ &\overset{\text{(A4)}}{\leq} \exp\left(\frac{-0.7\,\epsilon^2}{8} \cdot n^{0.1}\right) \\ &\leq\; n^{-\kappa}, \end{aligned}$$

where the last inequality holds for any given κ and $\epsilon > 0$, provided n is sufficiently large. Hence, $p \leq (1+\epsilon) \cdot 0.7 \cdot f(i) \cdot n$, w.h.p. Consequently, $\beta_i^{(t)} \cdot n \leq q + p \leq (0.53^2 + 0.7 \cdot (1+\epsilon)) \cdot f(i) \cdot n$, w.h.p. We set $\epsilon = 0.02$ so that $(0.53^2 + 0.7 \cdot (1+\epsilon)) \leq 1$. This proves invariant $H_1(t)$ for $2 \leq i \leq \ell$.

Finally, we prove invariant $H_2(t)$. For $0 \leq \tau \leq t$, let x_τ denote a random variable which is one if at least one ball of batch τ is allocated into a bin with load larger than $\tau + \ell$, and zero, otherwise. Furthermore, let h_τ denote the number of balls from batch τ that are allocated into a bin with load larger than $\tau + \ell$. Because of the invariants $H_1(1), \ldots, H_1(t)$, the probability that a fixed ball from batch τ will fall into a bin with more than $\tau + \ell$ balls is at most $f(\ell)^d \leq (n^{-0.9})^d \leq n^{-1.8}$. Therefore, $\mathbf{Pr}[x_\tau = 1] \leq n \cdot n^{-1.8} = n^{-0.8}$. In particular, for any integer $j \geq 1$,

$$\mathbf{Pr}[h_\tau \geq j] \;\leq\; \binom{n}{j}\left(\frac{1}{n^{1.8}}\right)^j \;\leq\; n^{-0.8j}.$$

In other words, $h_\tau = O(1)$, w.h.p. Thus, we can assume that there exists a suitable constant j so that $h_\tau \le j$ for $1 \le \tau \le t$. A violation of $H_2(t)$ implies that the bins with load at least $t+\ell+1$ contain more than b balls of height at least $t+\ell+1$. Observe that these balls must have been placed during the last b rounds or one of the invariants $H_2(1), \dots, H_2(t-1)$ is violated. That is, if $H_2(1), \dots, H_2(t-1)$ hold, then a violation of $H_2(t)$ implies that $j \cdot \sum_{\tau=t-b}^{t} x_\tau \ge b$. The probability for this event is at most

$$\mathbf{Pr}\left[\sum_{\tau=t-b}^{t} x_\tau \ge b/j\right] \;\le\; \binom{b}{b/j} \cdot \left(\frac{1}{n^{0.8}}\right)^{b/j} \;\le\; \left(\frac{\mathrm{e}j}{n^{0.8}}\right)^{b/j} \;\le\; n^{-\kappa}$$

for any constant κ, provided that n is sufficiently large. Consequently, invariant H_2 holds, w.h.p., over all batches. This completes the proof of Lemma 1.1.

3. Greedy has short memory.

The goal of this section is to prove the Short Memory Lemma, Lemma 1.2. We will study the performance of Greedy[d] by analyzing the underlying Markov chain describing the load distribution of the bins and our analysis uses a new variant of coupling approach (neighboring coupling (see section 3.1.2)) to study the convergence time of Markov chains.

Remark 1. It is well known that the Short Memory Lemma does not hold for $d = 1$, that is, for the single-choice algorithm. In that case, in order to claim that $\mathcal{X}_\tau$ has stochastically almost the same distribution as $\mathcal{Y}_\tau$ one must have $\tau = \Omega(m^2 \cdot \mathrm{poly}(n))$.

3.1. Auxiliary lemmas.

In this subsection we state some auxiliary results that will be used in order to prove Lemma 1.2.

3.1.1. A simple load estimation.

We present here the following simple (and known) lemma (see, e.g., [15, 17, 30]) that we use in the proof of Lemma 3.9.

LEMMA 3.1. *Suppose that m balls are allocated in n bins using* Greedy[2]. *Let p be any positive real. Then with probability at least $1-p$ the minimum load in any bin is larger than or equal to*

$$\frac{m}{n} - \sqrt{\frac{2\,m}{n} \cdot \ln \frac{n}{p}}.$$

Proof. The proof follows easily from the fact that the minimum load in Greedy[2] is (stochastically) not smaller than the minimum load in the process that allocates m balls in n bins i.u.r. (this follows, for example, from the majorization results presented in [1, Theorem 3.5]). The minimum load in the process that allocates m balls in n bins i.u.r. can be estimated by looking at each bin independently. Then, we apply the Chernoff bound[2] to estimate the probability that the load of the bin is smaller than the expected load (i.e., from $\frac{m}{n}$) by more than $\sqrt{\frac{2\,m}{n} \cdot \ln \frac{n}{p}}$. We apply the Chernoff bound to m independent random variables $X_1, \dots, X_m$ with X_i indicating the event that the ith ball is allocated in the given bin. Then, we obtain for every $t \in \mathbb{N}$

$$\mathbf{Pr}\left[\text{given bin has load } \le \frac{m}{n} - t\right] \le \exp\left(\frac{-n\,t^2}{2\,m}\right).$$

[2] We use the following form of the Chernoff bound (see, e.g., [21, Theorem 2.3 (c)]): If $X_1, \dots, X_m$ are binary independent random variables and if $X = \sum_{j=1}^{m} X_j$, then for any δ, $0 < \delta < 1$, it holds that $\mathbf{Pr}[X \le (1-\delta)\,\mathbf{E}[X]] \le \exp(-\delta^2\,\mathbf{E}[X]/2)$.

Therefore, by the union bound,

$$\mathbf{Pr}\left[\text{minimum load } \le \frac{m}{n} - t\right] \le n \cdot \exp\left(\frac{-n\,t^2}{2\,m}\right).$$

This implies our result by setting $t = \sqrt{\frac{2\,m}{n} \cdot \ln \frac{n}{p}}$. □

3.1.2. Neighboring coupling. Our main tool in the analysis of the convergence of Markov chains is a new variant of the path coupling arguments that extends the results from [7, 13] and which is described in the following lemma.

LEMMA 3.2 (Neighboring Coupling Lemma). *Let $\mathfrak{M} = (\mathbf{Y}_t)_{t\in\mathbb{N}}$ be a discrete-time Markov chain with a state space Ω. Let $\Omega^* \subseteq \Omega$. Let Γ be any subset of $\Omega^* \times \Omega^*$ (elements $(X, Y) \in \Gamma$ are called neighbors). Suppose that there is an integer D such that for every $(X, Y) \in \Omega^* \times \Omega^*$ there exists a sequence $X = \Lambda_0, \Lambda_1, \dots, \Lambda_r = Y$, where $(\Lambda_i, \Lambda_{i+1}) \in \Gamma$ for $0 \le i < r$, and $r \le D$.*

If there exists a coupling $(X_t, Y_t)_{t\in\mathbb{N}}$ for $\mathfrak{M}$ such that for some $T \in \mathbb{N}$, for all $(X, Y) \in \Gamma$, it holds that $\mathbf{Pr}[X_T \ne Y_T \mid (X_0, Y_0) = (X, Y)] \le \frac{\varepsilon}{D}$, then

$$\|\mathcal{L}(X_T|X_0 = X) - \mathcal{L}(Y_T|Y_0 = Y)\| \le \varepsilon$$

for every $(X, Y) \in \Omega^* \times \Omega^*$.

Proof. For any pair of neighbors $(\Lambda, \Lambda') \in \Gamma$,

$$\|\mathcal{L}(Z_T|Z_0 = \Lambda) - \mathcal{L}(Z_T|Z_0 = \Lambda')\| \;\le\; \frac{\varepsilon}{D}$$

by the well-known Coupling Lemma (see, e.g., [4, Lemma 3.6]). As a consequence, we obtain

$$\begin{aligned}
&\|\mathcal{L}(Z_T|Z_0 = X) - \mathcal{L}(Z_T|Z_0 = Y)\| \\
&\qquad \le \sum_{i=1}^{r} \|\mathcal{L}(Z_T|Z_0 = \Lambda_i) - \mathcal{L}(Z_T|Z_0 = \Lambda_{i-1})\| \le r \cdot \frac{\varepsilon}{D} \le \varepsilon. \qquad \square
\end{aligned}$$

Thus, if we can find a neighboring coupling, we obtain immediately a bound on the total variation distance in terms of the tail probabilities of the *coupling time*, i.e., a random time $\mathbb{T}$ for which $X_t = Y_t$ for all $t \ge \mathbb{T}$.

3.1.3. Random walks on $\mathbb{N}$. In this section we present some auxiliary results on the convergence rates of a random walk on the line $\mathbb{N}$ with the starting point D, with the absorbing barrier at 0, and with a drift $\beta \ge 0$ toward 0. We feel that these results might be known, but since we have not seen them in forms needed in our paper, we decided to present them here in detail for the sake of completeness.

We begin with the following result which can be viewed as a bound on the number of steps needed by a random walk on the integer line with positive drift toward 0 until the "barrier" 0 is hit.

LEMMA 3.3 (random walk on $\mathbb{N}$—positive drift toward 0). *Let ϵ and β be any positive reals. Let D be an arbitrary natural number. Let $c \ge 1$ be an arbitrary constant. Let $(\mathcal{X}_t)_{t\in\mathbb{N}}$ be a sequence of (not necessarily independent) random variables such that*

(1) $0 \le \mathcal{X}_0 \le D$,
(2) *for every $t \in \mathbb{N}$, $|\mathcal{X}_{t+1} - \mathcal{X}_t| \le c$,*
(3) *for every $t \in \mathbb{N}$, if $\mathcal{X}_t > 0$ then $\mathbf{E}[\mathcal{X}_{t+1} - \mathcal{X}_t \mid \mathcal{X}_0, \mathcal{X}_1, \dots, \mathcal{X}_t] \le -\beta$, and*

(4) *for every* $t \in \mathbb{N}$, *if* $\mathcal{X}_t \le 0$ *then* $\mathcal{X}_{t+1} = 0$.

Then, for certain $T = \Theta(D/\beta + \ln(1/\epsilon)/\beta^2)$, *if* $\tau \ge T$ *then* $\mathbf{Pr}[\mathcal{X}_\tau > 0] \le \epsilon$.

Proof. We eliminate the requirement that $\mathcal{X}_t$ is never negative by introducing a new sequence of random variables $(\mathcal{Y}_t)_{t\in\mathbb{N}}$. If $\mathcal{X}_t > 0$ then we set $\mathcal{Y}_t = \mathcal{X}_t$, and otherwise we define $\mathcal{Y}_t = \mathcal{Y}_{t-1} - \beta$.

Then, conditions (1)–(2) still hold for sequence $(\mathcal{Y}_t)_{t\in\mathbb{N}}$, condition (3) holds for $(\mathcal{Y}_t)_{t\in\mathbb{N}}$ even without the assumption that $\mathcal{Y}_t > 0$, and we have a new condition (4*) such that for every $t \in \mathbb{N}$, if $\mathcal{Y}_t \le 0$ then $\mathcal{Y}_{t+1} \le 0$. Additionally, it is easy to see that for every $t \in \mathbb{N}$ it holds that $\mathbf{Pr}[\mathcal{Y}_t > 0] = \mathbf{Pr}[\mathcal{X}_t > 0]$.

Next, since for every t we have $\mathbf{E}[\mathcal{Y}_{t+1} - \mathcal{Y}_t \mid \mathcal{Y}_0, \dots, \mathcal{Y}_t] \le -\beta$, we see that $\mathbf{E}[\mathcal{Y}_t] \le D - \beta\, t$. Therefore, we can apply the Hoeffding–Azuma inequality (see, e.g., [21, Theorem 3.10]) to obtain that for every $\alpha > 0$ it holds that

$$\mathbf{Pr}[\mathcal{Y}_t > \alpha + \mathbf{E}[\mathcal{Y}_t]] \le e^{-\alpha^2/(2tc)}.$$

Since for $\alpha = t\,\beta - D$ we have $\alpha + \mathbf{E}[\mathcal{Y}_t] \le 0$, we can conclude that

$$\mathbf{Pr}[\mathcal{X}_t > 0] = \mathbf{Pr}[\mathcal{Y}_t > 0] \;\le\; e^{-(t\,\beta - D)^2/(2tc)},$$

which yields the lemma. ∎

Next, we investigate random walks on the integers with "balanced" drift, or a drift which is very small; that is, β in Lemma 3.3 is tiny and hence the bound at that lemma is weak.

LEMMA 3.4 (random walk on $\mathbb{N}$—balanced case). *Let* ϵ *be any positive real. Let* D *be an arbitrary natural number. Let* $c \ge 1$ *be an arbitrary constant. Let* $(\mathcal{X}_t)_{t\in\mathbb{N}}$ *be a sequence of (not necessarily independent) integer random variables such that the following properties hold:*

1. $0 \le \mathcal{X}_t \le D$ *for every* $t \in \mathbb{N}$,
2. *for every* $t \in \mathbb{N}$, $|\mathcal{X}_{t+1} - \mathcal{X}_t| \le c$,
3. *for every* $t \in \mathbb{N}$, *if* $\mathcal{X}_t > 0$ *then* $\mathcal{X}_{t+1} \ne \mathcal{X}_t$ *and* $\mathbf{E}[\mathcal{X}_{t+1} - \mathcal{X}_t \mid \mathcal{X}_0, \mathcal{X}_1, \dots, \mathcal{X}_t] \le 0$, *and*
4. *for every* $t \in \mathbb{N}$, *if* $\mathcal{X}_t = 0$ *then* $\mathcal{X}_{t+1} = 0$.

Then, for certain $T = \Theta(D^2 \cdot \ln(1/\epsilon))$, *if* $\tau \ge T$ *then* $\mathbf{Pr}[\mathcal{X}_\tau > 0] \le \epsilon$.

Proof. We first observe that it is enough to prove the lemma only for $\mathbf{E}[\mathcal{X}_{t+1} - \mathcal{X}_t \mid \mathcal{X}_0, \dots, \mathcal{X}_t] = 0$. We follow here arguments given in [19, Lemma 4]. Recall that random variables $\mathcal{V}_0, \mathcal{V}_1, \dots,$ form a *submartingale* with respect to a sequence $(\mathcal{W}_t)_{t\in\mathbb{N}}$ if $\mathbf{E}[\mathcal{V}_{t+1} - \mathcal{V}_t \mid \mathcal{W}_0, \mathcal{W}_1, \dots, \mathcal{W}_t] \ge 0$ for every $t \in \mathbb{N}$. A random variable τ is a *stopping time* for the submartingale if for each t one can determine if $\tau \ge t$. We shall use the optional stopping time theorem (due to Doob) for submartingales which says that if $(\mathcal{U}_t)_{t\in\mathbb{N}}$ is a submartingale $(\mathcal{U}_t)_{t\in\mathbb{N}}$ with bounded $|\mathcal{U}_{t+1} - \mathcal{U}_t|$ for every $t \in \mathbb{N}$ and τ is a stopping time with finite expectation, then $\mathbf{E}[\mathcal{U}_\tau] \ge \mathbf{E}[\mathcal{U}_0]$.

Fix $\mathcal{X}_0$. Define the stochastic process $\mathcal{Z}_0, \mathcal{Z}_1, \dots$ such that for every $t \in \mathbb{N}$,

$$\mathcal{Z}_t = \begin{cases} (D - \mathcal{X}_t)^2 - t & \text{if either } t = 0 \text{ or } t > 0 \text{ and } \mathcal{X}_{t-1} > 0, \\ Z_{t-1} & \text{if } t > 0 \text{ and } \mathcal{X}_{t-1} = 0. \end{cases}$$

Let us first observe that $\mathcal{Z}_t$ is a submartingale with respect to the sequence $(\mathcal{X}_t)_{t\in\mathbb{N}}$. Indeed, conditioned on $\mathcal{X}_0, \mathcal{X}_1, \dots, \mathcal{X}_t$, if $\mathcal{X}_t = 0$ then $\mathcal{Z}_{t+1} - \mathcal{Z}_t = 0$. Otherwise, if $\mathcal{X}_t > 0$, then we obtain $\mathbf{E}[\mathcal{Z}_{t+1} - \mathcal{Z}_t] = \mathbf{E}[((D - \mathcal{X}_{t+1})^2 - (t+1)) - ((D - \mathcal{X}_t)^2 - t)] = \mathbf{E}[(\mathcal{X}_{t+1} - \mathcal{X}_t)^2 - 1 + 2 \cdot (\mathcal{X}_t - \mathcal{X}_{t+1}) \cdot (D - \mathcal{X}_t)] \ge \mathbf{E}[(\mathcal{X}_{t+1} - \mathcal{X}_t)^2] - 1 + 2 \cdot (D - \mathcal{X}_t) \cdot \mathbf{E}[\mathcal{X}_t -

$\mathcal{X}_{t+1}] \geq 1 - 1 + 0 \geq 0$. Here, we used the fact that if $\mathcal{X}_t > 0$ then $\mathcal{X}_{t+1} \neq \mathcal{X}_t$, and therefore since each of $\mathcal{X}_{t+1}$ and $\mathcal{X}_t$ is an integer, it holds that $\mathbf{E}[(\mathcal{X}_{t+1} - \mathcal{X}_t)^2] \geq 1$.

Next, we notice that the differences $|\mathcal{Z}_{t+1} - \mathcal{Z}_t|$ are bounded for every $t \in \mathbb{N}$ and the random time $\mathfrak{T}_{\mathcal{X}_0} = \min\{t : \mathcal{X}_t = 0\}$ is a stopping time for $\mathcal{Z}_t$ with finite expectation. Moreover, $\mathcal{Z}_0 = (D - \mathcal{X}_0)^2$ and $\mathcal{Z}_{\mathfrak{T}_{\mathcal{X}_0}} = D^2 - \mathfrak{T}_{\mathcal{X}_0}$. Since $\mathbf{E}[\mathcal{Z}_{\mathfrak{T}_{\mathcal{X}_0}}] \geq \mathbf{E}[\mathcal{Z}_0]$ by the optional stopping time theorem, we get $\mathbf{E}[\mathfrak{T}_{\mathcal{X}_0}] \leq \mathcal{X}_0 \cdot (2\,D - \mathcal{X}_0) \leq D^2$.

Take $t = e \cdot D^2$. The Markov inequality implies that $\mathbf{Pr}[\mathfrak{T}_{\mathcal{X}_0} \geq t] \leq e^{-1}$. If we run $\ln(1/\epsilon)$ independent trials of length t then for $T = t \cdot \ln(1/\epsilon)$ the probability that $\mathcal{X}_T \neq 0$ is at most ϵ^{-1}. $\square$

3.2. Greedy[d] has short memory: Analysis for d=2. In this section we prove Lemma 1.2 for $d = 2$. More precisely, we will prove various properties about Greedy[d]. For $d = 2$ these properties will immediately imply Lemma 1.2. For $d > 2$ we need some additional arguments which are then presented in section 3.3.

Throughout this section we deal only with *normalized* load vectors. For every $k \geq 0$, let Ω_k denote the set of normalized load vectors with k balls. In our analysis, we investigate the following Markov chain $\mathfrak{M}[d] = (\mathcal{M}_t)_{t\in\mathbb{N}}$, which models the behavior of protocol Greedy[d]:

> **Input:** $\mathcal{M}_0$ *is any load vector in* Ω_m
>
> **Transitions $\mathcal{M}_t \Rightarrow \mathcal{M}_{t+1}$:**
>
> *Pick* $q \in [n]$ *at random such that* $\mathbf{Pr}[q = k] = \frac{k^d - (k-1)^d}{n^d}$
>
> $\mathcal{M}_{t+1}$ *is obtained from* $\mathcal{M}_t$ *by adding a new ball to the* q*th fullest bin*

It is easy to see that the choice of q is equivalent to the choice obtained by the following simple randomized process: *Pick* $q_1, q_2, \ldots, q_d \in [n]$ *i.u.r. and set* $q = \max\{q_i : 1 \leq i \leq d\}$. This in turn, is equivalent to the choice of q obtained by Greedy[d]: *Pick d bins i.u.r. and let the least loaded among the chosen bins be the qth fullest bin in the system.*

Our proof of Lemma 1.2 is via the *neighboring coupling* method discussed in detail in section 3.1.2. Let X and Y denote two vectors from Ω_m. We study the process by which we add new balls on the top of each of the allocations described by these vectors. We analyze how many balls one has to add until the two allocations are almost indistinguishable.

3.2.1. Neighboring coupling. In order to apply the Neighboring Coupling Lemma to analyze the Markov chain $\mathfrak{M}[d] = (\mathcal{M}_t)_{t\in\mathbb{N}}$, we must first define the notion of neighbors. Let us fix m and n. Let us define $\Omega^* = \Omega_m$ and let Γ be the set of pairs of those load vectors from Ω_m which correspond to the balls' allocations that differ in exactly one ball (cf. Figure 2). In that case, if X can be obtained from Y by moving a ball from the ith fullest bin into the jth fullest bin, then we write $X = Y - \mathbf{e}_i + \mathbf{e}_j$. Thus,

$$\Gamma = \{(X, Y) \in \Omega_m \times \Omega_m \,:\, X = Y - \mathbf{e}_i + \mathbf{e}_j \quad \text{for certain } i, j \in [n], i \neq j\}.$$

Clearly, for each $X, Y \in \Omega_m$ there exists a sequence $X = Z^{\langle 0\rangle}, Z^{\langle 1\rangle}, \ldots, Z^{\langle l-1\rangle}, Z^{\langle l\rangle} = Y$, where l is the number of balls on which X and Y differ, $l \leq m$, and $(Z^{\langle i\rangle}, Z^{\langle i+1\rangle}) \in \Gamma$ for every i, $0 \leq i \leq l - 1$. Thus, we can apply the Neighboring Coupling Lemma with $D = m$.

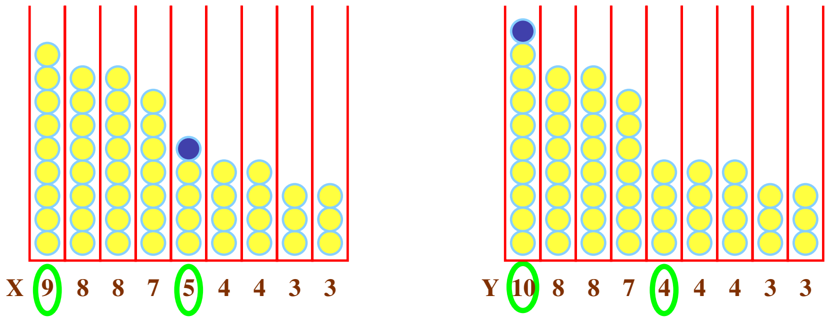

FIG. 2. *An example of neighboring load vectors X and Y with $X = Y - \mathbf{e}_1 + \mathbf{e}_5$ and $\Delta(X, Y) = 6$.*

3.2.2. Short memory lemma using neighboring coupling. The main result of this section is the following technical lemma.

LEMMA 3.5. *Let $\varepsilon > 0$ and let $d \geq 2$ be integer. Let $X, Y \in \Omega_m$. There exists a coupling $(X_t, Y_t)_{t \in \mathbb{N}}$ for $\mathfrak{M}[d]$ and there is $\mathbb{T} = \Theta(m \cdot n^d + n^{2\,d} \cdot \ln(m/\varepsilon))$ such that for any $\tau \geq \mathbb{T}$ it holds that $\mathbf{Pr}[X_\tau \neq Y_\tau \mid (X_0, Y_0) = (X, Y)] \leq \varepsilon$.*

Let us first observe that for $d = 2$ Lemma 3.5 immediately implies the Short Memory Lemma for Greedy[2]. However, Lemma 3.5 yields a weaker bound for $d \geq 3$. Therefore, a more specialized analysis for the case $d \geq 3$ is postponed to section 3.3.

Notice that since for any pair $X, Y \in \Omega_m$ there exists a sequence $\Lambda_0, \Lambda_1, \ldots, \Lambda_k$ such that $k \leq m$, $\Lambda_0 = X$, $\Lambda_k = Y$, and $(\Lambda_{i-1}, \Lambda_i) \in \Gamma$ for every $1 \leq i \leq k$, Lemma 3.5 follows immediately from the following lemma.

LEMMA 3.6. *Let $\varepsilon > 0$ and let $d \geq 2$ be integer. Let $X, Y \in \Gamma$. There exists a coupling $(X_t, Y_t)_{t \in \mathbb{N}}$ for $\mathfrak{M}[d]$ and there is $\mathbb{T} = \Theta(m \cdot n^d + n^{2\,d} \cdot \ln(m/\varepsilon))$ such that for any $\tau \geq \mathbb{T}$ it holds that $\mathbf{Pr}[X_\tau \neq Y_\tau \mid (X_0, Y_0) = (X, Y)] \leq \frac{\varepsilon}{m}$.*

The rest of this subsection is devoted to the proof of Lemma 3.6.

For any load vectors $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_n)$ with $X = Y - \mathbf{e}_i + \mathbf{e}_j$, $i, j \in [n]$, let us define the *distance* function $\Delta(X, Y)$ (cf. Figure 2),

$$\Delta(X, Y) \;=\; \max\{|x_i - x_j|, |y_i - y_j|\}.$$

Observe that $\Delta(X, Y)$ is always a nonnegative integer, it is zero only if $X = Y$and it never takes the value of 1. Let

$$\begin{aligned} \xi = \min\{&\mathbf{Pr}[\text{Greedy}[d] \text{ picks the } j\text{th fullest bin}] \\ &- \mathbf{Pr}[\text{Greedy}[d] \text{ picks the } i\text{th fullest bin}] \;:\; i, j \in [n],\; i < j\}. \end{aligned}$$

Then, clearly, $\xi \geq 1/n^d$. The following lemma describes main properties of the desired coupling.

LEMMA 3.7. *If $(X, Y) \in \Gamma$ then there exists a coupling $(X_t, Y_t)_{t \in \mathbb{N}}$ for $\mathfrak{M}[d]$ that, conditioned on $(X_0, Y_0) = (X, Y)$, satisfies the following properties for every $t \in \mathbb{N}$:*

- *if $X_t = Y_t$ then $X_{t+1} = Y_{t+1}$,*
- *if $X_t \neq Y_t$ then X_t and Y_t differ in at most one ball,*
- $\Delta(X_{t+1}, Y_{t+1}) - \Delta(X_t, Y_t) \in \{-2, -1, 0, 1\}$, *and*
- *if $X_t \neq Y_t$ then $\mathbf{E}[\Delta(X_{t+1}, Y_{t+1}) \mid X_t, Y_t] \leq \Delta(X_t, Y_t) - \xi$.*

Proof. We use the following natural coupling: each time we increase the vectors X and Y by one ball, we use the same random choice. That is, in each step the obtained load vectors will be obtained from X and Y, respectively, by allocating a new ball to the qth fullest bin for certain $q \in [n]$.

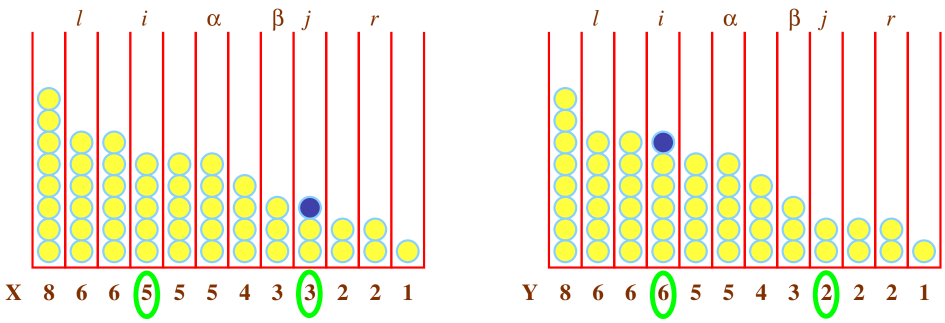

FIG. 3. *Illustration to the proof of Claim* 3.8. *In this case* $X = Y - \mathbf{e}_4 + \mathbf{e}_9$, $\Delta(X, Y) = 4$, $l = 2$, $\alpha = 6$, $\beta = 8$, *and* $r = 11$.

The lemma now follows directly from the properties of the coupling described in Claim 3.8 below. ∎

CLAIM 3.8. *Let* X, Y *be two load vectors from* Ω_m *that differ in one ball with* $X = Y - \mathbf{e}_i + \mathbf{e}_j$ *for certain* $i < j$. *Let* $X^{\langle q \rangle}$ *and* $Y^{\langle q \rangle}$ *be obtained from* X *and* Y, *respectively, by allocating a new ball to the* q*th fullest bin. Then, either*

1. $X^{\langle q \rangle} = Y^{\langle q \rangle}$ *and* $\Delta(X^{\langle q \rangle}, Y^{\langle q \rangle}) = \Delta(X, Y) - 2$, *or*
2. $X^{\langle q \rangle}$ *and* $Y^{\langle q \rangle}$ *differ in one ball and*

$$\Delta(X^{\langle q \rangle}, Y^{\langle q \rangle}) = \begin{cases} \Delta(X, Y) - 1 & \text{if and only if } q = j, \\ \Delta(X, Y) + 1 & \text{if and only if } q = i, \\ \Delta(X, Y) & \text{otherwise.} \end{cases}$$

Proof. The proof is by case analysis which is tedious but otherwise straightforward (see also Figure 3 for some intuition behind the coupling). We assume that $\Delta(X, Y) = y_i - y_j$; the case $\Delta(X, Y) = x_i - x_j$ can be done similarly. Let

$$\begin{aligned} l &= \min\{s \in [n] \ : \ y_s = y_i\}, \\ \alpha &= \max\{s \in [n] \ : \ x_i = x_s\}, \\ \beta &= \min\{s \in [n] \ : \ x_s = x_j\}, \\ r &= \max\{s \in [n] \ : \ y_j = x_s\}. \end{aligned}$$

Let us notice that $1 \le l \le i \le \alpha$, $l < \alpha$, $\beta \le j \le r \le n$, and $\beta < r$. We first consider six cases when either $1 \le q \le i$ or $j \le q \le n$.

Case (1) $1 \le q < l$. Since $q < l$ the same happens for both processes. For certain s, $s \le q$, we have $X^{\langle q \rangle} = X + \mathbf{e}_s$ and $Y^{\langle q \rangle} = Y + \mathbf{e}_s$. Therefore, after adding the ball to the qth fullest bin, we still have $X^{\langle q \rangle} = Y^{\langle q \rangle} - \mathbf{e}_i + \mathbf{e}_j$. Hence, $\Delta(X^{\langle q \rangle}, Y^{\langle q \rangle}) = y_i^{\langle q \rangle} - y_j^{\langle q \rangle} = y_i - y_j = \Delta(X, Y)$.

Case (2) $l \le q < i$. After reordering of the bins, the load of the lth largest bin has increased by one for both load vectors (note that all load vectors between the lth largest and the $i - 1$th largest bin have the same load in both processes). Hence, $X^{\langle q \rangle} = X + \mathbf{e}_l$ and $Y^{\langle q \rangle} = Y + \mathbf{e}_l$. Therefore, $X^{\langle q \rangle} = Y^{\langle q \rangle} - \mathbf{e}_i + \mathbf{e}_j$ and the rest of the case is similar to Case (1).

Case (3) $q = i$. In this case we have $X^{\langle q \rangle} = X + \mathbf{e}_i$. After reordering of the load vector of Y we have $Y^{\langle q \rangle} = Y + \mathbf{e}_l$ (see Case (2)). This yields $X^{\langle q \rangle} = Y^{\langle q \rangle} - \mathbf{e}_l + \mathbf{e}_j$ and $\Delta(X^{\langle q \rangle}, Y^{\langle q \rangle}) = y_l^{\langle q \rangle} - y_j^{\langle q \rangle} = (y_l + 1) - y_j = \Delta(X, Y) + 1$.

Case (4) $q = j$. The βth and the jth largest bins have the same number of elements after adding a ball to the jth largest bin of X. Hence, $X^{\langle q \rangle} = X + \mathbf{e}_\beta$ and

$Y^{\langle q \rangle} = Y + \mathbf{e}_j$ (in the case of Y the bins do not have to be reordered). Therefore $X^{\langle q \rangle} = Y^{\langle q \rangle} - \mathbf{e}_i + \mathbf{e}_\beta$.

Now, there are two possibilities. First, if $\Delta(X, Y) = 2$, then $i = \beta$ and therefore $X^{\langle q \rangle} = Y^{\langle q \rangle}$. Otherwise, $\Delta(X, Y) > 2$ and hence $i \le \alpha < \beta$, which implies that $\Delta(X^{\langle q \rangle}, Y^{\langle q \rangle}) = y_i^{\langle q \rangle} - y_\beta^{\langle q \rangle} = y_i - (y_j + 1) = \Delta(X, Y) - 1$.

Case (5) $j < q \le r$. After reordering we have $X^{\langle q \rangle} = X + \mathbf{e}_{j+1}$ and $Y^{\langle q \rangle} = Y + \mathbf{e}_j$. We get $X^{\langle q \rangle} = Y^{\langle q \rangle} - \mathbf{e}_i + \mathbf{e}_{j+1}$ and $\Delta(X^{\langle q \rangle}, Y^{\langle q \rangle}) = y_i^{\langle q \rangle} - y_{j+1}^{\langle q \rangle} = y_i - y_j = \Delta(X, Y)$.

Case (6) $r < q \le n$. This case is similar to Case (1). For certain s, $r < s \le q$, it holds that $X^{\langle q \rangle} = X + \mathbf{e}_s$ and $Y^{\langle q \rangle} = Y + \mathbf{e}_s$. Therefore, $X^{\langle q \rangle} = Y^{\langle q \rangle} - \mathbf{e}_i + \mathbf{e}_j$ and $\Delta(X^{\langle q \rangle}, Y^{\langle q \rangle})$ does not change.

Now it remains to consider the case when $i < q < j$. We distinguish here two main cases.

Case (A) $x_i = x_j$. In this case we have $\Delta(X, Y) = 2$. After reordering we have $X^{\langle q \rangle} = X + \mathbf{e}_i$ and $Y^{\langle q \rangle} = Y + \mathbf{e}_{i+1}$. This means $X^{\langle q \rangle} = Y^{\langle q \rangle} - \mathbf{e}_{i+1} + \mathbf{e}_j$ and $\Delta(X^{\langle q \rangle}, Y^{\langle q \rangle}) = y_{i+1}^{\langle q \rangle} - y_j^{\langle q \rangle} = y_i - y_j = 2 = \Delta(X, Y)$.

Case (B) $x_i > x_j$. In this case $\alpha < \beta$ and we distinguish three subcases:

- *Case* (B.1) $i < q \le \alpha$. After reordering we get $X^{\langle q \rangle} = X + \mathbf{e}_i$ and $Y^{\langle q \rangle} = Y + \mathbf{e}_{i+1}$. Therefore, $X^{\langle q \rangle} = Y^{\langle q \rangle} - \mathbf{e}_{i+1} + \mathbf{e}_j$ and $\Delta(X^{\langle q \rangle}, Y^{\langle q \rangle}) = y_{i+1}^{\langle q \rangle} - y_j^{\langle q \rangle} = y_i - y_j = \Delta(X, Y)$.
- *Case* (B.2) $\alpha < q < \beta$. Again, this case is similar to Case (1); for certain s, $\alpha < s \le q < \beta$, we get $X^{\langle q \rangle} = X + \mathbf{e}_s$ and $Y^{\langle q \rangle} = Y + \mathbf{e}_s$. Hence, $\Delta(X, Y)$ does not change.
- *Case* (B.3) $\beta \le q < j$. In this case $X^{\langle q \rangle} = X + \mathbf{e}_\beta$ and $Y^{\langle q \rangle} = Y + \mathbf{e}_\beta$. Therefore, $X^{\langle q \rangle} = Y^{\langle q \rangle} - \mathbf{e}_i + \mathbf{e}_j$ and $\Delta(X, Y)$ does not change. ◻

Now we are ready to present the proof of Lemma 3.6.

Proof. We use the coupling constructed in Lemma 3.7. Observe that if we define $\Delta_t = \Delta(X_t, Y_t)$, $t \ge 0$, then from Lemma 3.7 the random variable Δ_t behaves like a random walk on $\mathbb{N}$ with drift toward 0; see section 3.1.3. By our assumption, we can set $\xi = 1/n^d$. Given that, we can conclude the proof by applying Lemma 3.3 with $\mathcal{X}_t = \Delta(X_t, Y_t)$, $c_t = 2$ for every $t \in \mathbb{N}$, and with $D = m$ and $\beta = \xi = 1/n^d$. ◻

3.3. Short memory property of Greedy$[d]$ for large d.

The main problem with applying Lemma 3.7 for large d is that the value of ξ may be very small. Now we modify the analysis above to give a better bound for Greedy$[d]$ than the one of Lemma 3.5 for all $d > 2$.

3.3.1. Load difference reduction in Greedy$[d]$ for $d \ge 3$.

For any load vector $\mathcal{W}$ let us denote by $\textsc{Low}(\mathcal{W})$ ($\textsc{Upp}(\mathcal{W})$) the minimum load (respectively, the maximum load) in $\mathcal{W}$. We prove that independently of the initial difference between $\textsc{Low}(\mathcal{W})$ and $\textsc{Upp}(\mathcal{W})$ at some moment of the allocation process Greedy$[d]$, after allocating some new balls, this difference will be kept small.

LEMMA 3.9. *Let n and M be any positive integers and let ε be any positive real. Let $d \ge 2$ be any integer. Let $X \in \Omega_M$. Let $\mathfrak{M}[d] = (X_t)_{t \in \mathbb{N}}$ with $X_0 = X$; that is, $X_0, X_1, \ldots$ is the sequence of random variables describing the Markov chain $\mathfrak{M}[d]$ conditioned on the event $X_0 = X$. Then, there exist a certain $\mathbb{T} = \Theta\left(M\, n^2 + n^4 \cdot \ln(M/\varepsilon)\right)$ and a constant $c > 0$ such that the following hold.*

(1) *If $M = \mathcal{O}(n^3 \ln(n/\varepsilon))$, then the following two bounds hold:*

$$\mathbf{Pr}\left[\textsc{Low}(X_{\mathbb{T}}) \le \frac{M+\mathbb{T}}{n} - c \cdot \sqrt{(M + n^2 \ln(M/\varepsilon)) \cdot n \cdot \ln(n/\varepsilon)}\right] \le \varepsilon,$$

$$\mathbf{Pr}\left[\textsc{Upp}(X_{\mathbb{T}}) \ge \frac{M+\mathbb{T}}{n} + c \cdot \sqrt{(M + n^2 \ln(M/\varepsilon)) \cdot n \cdot \ln(n/\varepsilon)}\right] \le \varepsilon.$$

(2) *If $M = \Omega(n^3 \ln(n/\varepsilon))$, then the following two bounds hold:*

$$\mathbf{Pr}\left[\textsc{Low}(X_{\mathbb{T}}) \le \frac{M+\mathbb{T}}{n} - c \cdot n^{1.25} \cdot M^{0.25} \cdot (\ln(n/\varepsilon))^{0.75}\right] \le \varepsilon,$$

$$\mathbf{Pr}\left[\textsc{Upp}(X_{\mathbb{T}}) \ge \frac{M+\mathbb{T}}{n} + c \cdot n^{1.25} \cdot M^{0.25} \cdot (\ln(n/\varepsilon))^{0.75}\right] \le \varepsilon.$$

Proof. We prove the lemma only for M being the multiple of n; the general case can be handled similarly. Since the proofs for $\textsc{Low}(\mathcal{W})$ and $\textsc{Upp}(\mathcal{W})$ are almost the same, we will deal only with $\textsc{Low}(\mathcal{W})$. We also point out that our proof uses ideas similar to those discussed later in section 4.

Let $Y_0, Y_1, \dots$ be the sequence of random variables (normalized load vectors) describing the Markov chain $\mathfrak{M}[2]$ conditioned on the event $Y_0 = X$. It is known that for every $l \in \mathbb{N}$ the normalized load vector X_l is majorized by the normalized load vector Y_l (see, e.g., [1, Theorem 3.5]). Therefore, in particular, $\textsc{Low}(X_l)$ is stochastically larger than or equal to $\textsc{Low}(Y_l)$ (the minimum load in Y_l). Hence, it is enough to prove the lemma only for the load vectors $Y_0, Y_1, \dots$.

Let ς be any positive real. Let Z be the ideally balanced load vector in Ω_M (i.e., the loads of all bins in Z are the same). Let $Z_0, Z_1, \dots$ be the sequence of random variables (normalized load vectors) describing the Markov chain $\mathfrak{M}[2]$ conditioned on the event $Z_0 = Z$. Lemma 3.5 implies that for certain $T = \Theta(M\, n^2 + n^4 \cdot \ln(M/\varsigma))$, for any $t \ge T$ the load vectors Y_t and Z_t are almost indistinguishable (formally, $\|\mathcal{L}(Z_t) - \mathcal{L}(Y_t)\| \le \varsigma$). In particular, this means that the random variables $\textsc{Low}(Y_l)$ and $\textsc{Low}(Z_l)$ are stochastically the same with probability at least $1-\varsigma$. Furthermore, by Lemma 3.1, we know that for any $t \in \mathbb{N}$, it holds that

$$\mathbf{Pr}\left[\textsc{Low}(Z_t) \le \frac{M+t}{n} - \sqrt{\frac{2\,(M+t)}{n}\ \ln\frac{n}{\varsigma}}\right] \le \varsigma. \tag{1}$$

Therefore, since for any $t \ge T$ we have $\|\mathcal{L}(Z_t) - \mathcal{L}(Y_t)\| \le \varsigma$, we may conclude that for $t \ge T$ it holds that

$$\mathbf{Pr}\left[\textsc{Low}(Y_t) \le \frac{M+t}{n} - \sqrt{\frac{2\,(M+t)}{n}\ \ln\frac{n}{\varsigma}}\right] \le 2\,\varsigma. \tag{2}$$

With inequality (2) we immediately obtain the first estimation for $\textsc{Low}$ by setting $\varepsilon = 2\,\varsigma$ and $\mathbb{T} = t = T$.

In order to obtain the second estimation, we first fix the smallest $\tau \ge T$ such that τ is a multiple of n. Let $\mathcal{E}$ be the event that $\textsc{Low}(Y_\tau) \ge \frac{M+\tau}{n} - \sqrt{\frac{2\,(M+\tau)}{n}\ \ln\frac{n}{\varsigma}}$. Let us condition on this event for a moment.

For any $t \ge \tau$, let Y_t^* be the load vector obtained from Y_t after removing $r = \lfloor \frac{M+\tau}{n} - \sqrt{\frac{2\,(M+\tau)}{n}\ \ln\frac{n}{\varsigma}} \rfloor$ balls from each bin in Y_t. Clearly, since $\textsc{Low}(Y_t) \ge$

$\textsc{Low}(Y_\tau) \geq r$, Y_t^* is a proper normalized load vector in $\Omega_{t+M-r\cdot n}$. Notice further that there are $M^* = M + \tau - r \cdot n$ balls in the system described by Y_τ^*.

Now we apply once again a similar procedure as we did above for the first estimation. Let V be the ideally balanced load vector in Ω_{M^*} (i.e., each bin in V has the same load). Let $V_0, V_1, \ldots$ be the sequence of random variables (normalized load vectors) describing the Markov chain $\mathfrak{M}[2]$ conditioned on the event $V_0 = V$. Proceeding similarly as above, we want to compare $Y_{t+\tau}^*$ with V_t for $t \geq 0$.

Lemma 3.5 implies that for certain $T^* = \Theta(M^*\, n^2 + n^4 \cdot \ln(M^*/\varsigma))$, for any $t \geq T^*$ it holds that $\|\mathcal{L}(V_t) - \mathcal{L}(Y_{t+\tau}^*)\| \leq \varsigma$. Therefore, in particular, the random variables $\textsc{Low}(Y_{t+\tau}^*)$ and $\textsc{Low}(V_t)$ are stochastically the same with probability at least $1 - \varsigma$. Furthermore, by Lemma 3.1, we know that for any $t \in \mathbb{N}$ it holds that

$$\mathbf{Pr}\left[\textsc{Low}(V_t) \leq \frac{M^* + t}{n} - \sqrt{\frac{2\,(M^* + t)}{n}\,\ln\frac{n}{\varsigma}}\right] \leq \varsigma.$$

Therefore, since $\|\mathcal{L}(V_t) - \mathcal{L}(Y_{t+\tau}^*)\| \leq \varsigma$ for any $t \geq T^*$, we may conclude that for any $t \geq T^*$ it holds that

$$\mathbf{Pr}\left[\textsc{Low}(Y_{t+\tau}^*) \leq \frac{M^* + t}{n} - \sqrt{\frac{2\,(M^* + t)}{n}\,\ln\frac{n}{\varsigma}}\right] \leq 2\,\varsigma.$$

Furthermore, since the load vector $Y_{t+\tau}^*$ is obtained from the load vector $Y_{t+\tau}$ by removing r balls from each bin, we obtain that (conditioned on $\mathcal{E}$) for any $t \geq T^*$,

$$\mathbf{Pr}\left[\textsc{Low}(Y_{t+\tau}) \leq \frac{M + \tau + t}{n} - \sqrt{\frac{2\,(M^* + t)}{n}\,\ln\frac{n}{\varsigma}}\;\Big|\; \mathcal{E}\right] \leq 2\,\varsigma.$$

Finally, since we have proved that event $\mathcal{E}$ holds with probability at least $1 - 2\,\varsigma$ (see inequality (2)), we can conclude that for any $t \geq T^*$ it holds that

$$\mathbf{Pr}\left[\textsc{Low}(Y_{t+\tau}) \leq \frac{M + \tau + t}{n} - \sqrt{\frac{2\,(M^* + t)}{n}\,\ln\frac{n}{\varsigma}}\right] \leq 4\,\varsigma.$$

Now, it remains to resolve this bound with respect to n, M, and ς. We observe that

$$\tau = \Theta(M\, n^2 + n^4\, \ln(M/\varsigma))$$

and

$$M^* = \Theta\left(\sqrt{\tau\, n\, \ln(n/\varsigma)}\right) = \Theta\left(n^{1.5}\,\sqrt{M\, \ln(n/\varsigma)} + n^{2.5}\,\sqrt{\ln(M/\varsigma)\, \ln(n/\varsigma)}\right).$$

Furthermore,

$$\begin{aligned} T^* &= \Theta(M^*\, n^2 + n^4 \cdot \ln(M^*/\varsigma)) \\ &= \Theta\Bigg(n^{3.5}\,\sqrt{M \cdot \ln(n/\varsigma)} \quad n^{4.5}\,\sqrt{\ln(M/\varsigma)\, \ln(n/\varsigma)} \\ &\qquad\qquad + n^4\, \ln\left(\frac{n\, M\, \ln(n/\varsigma)\, \ln(M/\varsigma)}{\varsigma}\right)\Bigg). \end{aligned}$$

Now, we use our assumption that $M = \Omega(n^3\, \ln(n/\varsigma))$ for $\varsigma = \Theta(\varepsilon)$. In this case, the first term dominates the other one in the bounds for τ and for M^* and the first term dominates the other two in the bound for T^*. Hence, $\tau = \Theta(M\,n^2)$, $M^* = \Theta(n^{1.5}\sqrt{M\,\ln(n/\varsigma)})$, $T^* = \Theta(n^{3.5}\sqrt{M\cdot\ln(n/\varsigma)})$, and $\tau + T^* = \Theta(\tau)$. Therefore, we can conclude with the following claim: There exists a positive constant c such that for $\mathbb{T} = \tau + T^*$ it holds that

$$\mathbf{Pr}\left[\mathrm{Low}(X_{\mathbb{T}}) \le \frac{M+\mathbb{T}}{n} - c\cdot n^{1.25}\cdot M^{0.25}\cdot(\ln(n/\varsigma))^{0.75}\right] \le 4\,\varsigma.$$

Now, the lemma follows by setting $\varepsilon = 4\,\varsigma$. ∎

3.3.2. Coupling arguments and the short memory lemma for $d > 2$. Observe that a trivial implication of Lemma 3.9 is that if we start with any pair of normalized load vectors $X_0, Y_0 \in \Omega_M$, then for certain $\tau = \Theta(M\,n^2 + n^4\,\ln(M/\varepsilon))$, w.h.p. (depending on ε and κ), it holds that for any integer $\kappa > 0$, and for *all* t, $0 \le t < \kappa$, the difference in the maximum load and the minimum load in $X_{\tau+t}$ (or $Y_{\tau+t}$) is upper-bounded by

$$\zeta = \begin{cases} \mathcal{O}\left(\sqrt{(M + n^2\,\ln(M/\varepsilon))\,n\,\ln(n/\varepsilon)}\right) & \text{for } M = \mathcal{O}(n^3\,\ln(n/\varepsilon)),\\ \mathcal{O}\left(n^{1.25}\,M^{0.25}\,(\ln(n/\varepsilon))^{0.75}\right) & \text{for } M = \Omega(n^3\,\ln(n/\varepsilon)).\end{cases}$$

From now on we shall fix τ and κ and shall condition on this event (which, by applying the union bound to Lemma 3.9, is satisfied with probability larger than $1 - 2\,\kappa\,\varepsilon$).

Now we proceed with coupling arguments. Let $X, Y \in \Gamma$. We use the same distance function $\Delta(\cdot,\cdot)$ as in section 3 and the same coupling as in Lemma 3.7 and Claim 3.8. Observe that by our discussion in section 3.2.2, for any $t \in \mathbb{N}$, either X_t and Y_t are identical or they differ by one ball.

Epochs. Let τ and κ be set as above. We divide the time into *epochs*. The 0th epoch starts at time step $l_0 = 0$ and ends in time step $r_0 = \tau$. Each following epoch corresponds to the time period when the value of Δ remains unchanged. That is, if the $(k-1)$st epoch, $k \ge 1$, ends in time step r_{k-1}, then, inductively, the kth epoch begins in time step $l_k = 1 + r_{k-1}$ and ends in the smallest time step $t \ge l_k$ for which $\Delta(X_{t-1}, Y_{t-1}) \ne \Delta(X_t, Y_t)$. Additionally, if $X_{r_{k-1}} = Y_{r_{k-1}}$, then we define $r_k = \infty$, and the kth epoch lasts until the infinity.

CLAIM 3.10. *Let μ be any positive integer. Let for every t, $\tau \le t \le \tau+\kappa$, the difference between the maximum load and the minimum load in each of X_t and Y_t be upper-bounded by μ. Then, for every $1 \le k \le \frac{\kappa}{2\,n\,\mu+1}$, if $X_{r_{k-1}} \ne Y_{r_{k-1}}$, then the kth epoch lasts at most $2\,n\,\mu + 1$ time steps.*

Proof. Let $X_t = Y_t - \mathbf{e}_i + \mathbf{e}_j$ with $i < j$ and let $X_{t+1} = Y_{t+1} - \mathbf{e}_{i^*} + \mathbf{e}_{j^*}$, where X_{t+1} and Y_{t+1} are obtained from X_t and Y_t, respectively, by allocating a ball to bin q. By the case analysis (cf. also the proof of Claim 3.8) one can show that one of the following two cases must hold:

- $q = i$ or $q = j$ in the transition $(X_t, Y_t) \to (X_{t+1}, Y_{t+1})$ of the coupling.
- The load of the i^*th fullest bin in Y_{t+1} is the same as the load of the ith fullest bin in Y_t.

Consider a kth epoch and suppose that $X_{r_{k-1}} \ne Y_{r_{k-1}}$ with $X_{r_{k-1}} = Y_{r_{k-1}} - \mathbf{e}_{i^*} + \mathbf{e}_{j^*}$, $i^* < j^*$, in time r_{k-1}. Let ℓ be the load of the i^*th fullest bin in $Y_{r_{k-1}}$. Then, by the observation above and by Claim 3.8, for every t with $r_{k-1} \le t \le r_k - 1$, if $X_t = Y_t - \mathbf{e}_i + \mathbf{e}_j$ with $i < j$, then the load of the ith fullest bin in Y_t is ℓ.

Now, we want to use the assumption that for every t, $\tau \le t \le \tau+\kappa$, the difference between the maximum load and the minimum load in Y_t is upper-bounded by μ. Therefore, if $\tau \le r_{k-1} \le \tau + \kappa$, then in time r_{k-1} the value of ℓ (which is the load of one of the bins in $Y_{r_{k-1}}$) is at most $\frac{r_{k-1}}{n} + \mu$. Furthermore, if $\tau \le r_k \le \tau + \kappa$, then in time $r_k - 1$ the value of ℓ (which is the load of one of the bins in Y_{r_k-1}) is at least $\frac{r_k-1}{n} - \mu$. Therefore, we obtain that

$$\tfrac{r_{k-1}}{n} + \mu \ge \ell \ge \tfrac{r_k-1}{n} - \mu.$$

This implies immediately that $r_k - l_k \le 2\,n\,\mu$. This also yields inductively that if $1 \le k \le \kappa/(2\,n\,\mu + 1)$ and $X_{r_{k-1}} \neq Y_{r_{k-1}}$, then the kth epoch lasts at most $2\,n\,\mu + 1$ time steps. $\square$

Let Δ_k be the value of $\Delta(X_t, Y_t)$ for $t = r_k$, i.e., for t being the last time step of the kth epoch. Clearly, $\Delta_{k-1} \neq \Delta_k$. Furthermore, by Claim 3.8 the following holds:

- If $\Delta_{k-1} \ge 3$ then $\Delta_k \in \{\Delta_{k-1} - 1, \Delta_{k-1} + 1\}$ and $\mathbf{E}[\Delta_k - \Delta_{k-1}] < 0$.
- If $\Delta_{k-1} = 2$ then $\Delta_k \in \{0, 3\}$ and $\mathbf{E}[\Delta_k - \Delta_{k-1}] < 0$.

Therefore, similarly as in the proof of Lemma 3.5, we can model sequence $\Delta_0, \Delta_1, \dots$ as a random walk on the line $\mathbb{N}$ with the starting point $D \le \zeta$, the absorbing barrier in 0, and with a positive drift toward 0. This time, however, the drift is very small and therefore we use a weaker bound for the convergence of this random walk. From Lemma 3.4 we obtain that

$$\mathbf{Pr}[\Delta_k > 0] \le 2\,\kappa\,\varepsilon \quad \text{for all } k \ge \gamma\,\zeta^2\,\ln(2\,\kappa\,\varepsilon)^{-1},$$

where γ is some absolute positive constant. Now, since by Claim 3.10 each epoch k with $\Delta_{k-1} > 0$ lasts at most $2\,n\,\zeta + 1$ time steps, the last inequality implies the following. For all t, $t \ge \tau + \gamma\,\zeta^2\,\ln(2\,\kappa\,\varepsilon)^{-1} \cdot (2\,n\,\zeta + 1)$, it holds that $\mathbf{Pr}[X_t \neq Y_t] \le 2\,\kappa\,\varepsilon$. Since, by Lemma 3.9, we have proven that with probability larger than or equal to $1 - 2\,\kappa\,\varepsilon$, for every t, $\tau \le t < \tau + \kappa$, the difference between the maximum load and the minimum load in each of X_t and Y_t is upper bounded by ζ, we can conclude with the following lemma.

LEMMA 3.11. *Let n and m be any positive integers and let ε be any positive real. Let $d \ge 2$ be any integer. Let $X_0, Y_0 \in \Omega_m$.*

(1) *If $m = \mathcal{O}(n^3\,\ln(n/\varepsilon))$, then there is a coupling $(X_t, Y_t)_{t \in \mathbb{N}}$ for $\mathfrak{M}[d]$ such that for certain $\mathbb{T}^*$,*

$$\mathbb{T}^* = \Theta\left(n^{2.5} \cdot (m + n^2\,\ln(n/\varepsilon))^{1.5} \cdot \ln(n/\varepsilon)^{1.5} \cdot \ln(1/\varepsilon)\right);$$

it holds for any $t \ge \mathbb{T}^*$ that

$$\mathbf{Pr}[X_t \neq Y_t \mid (X_0, Y_0) = (X, Y)] \le \varepsilon.$$

(2) *If $m = \Omega(n^3\,\ln(n/\varepsilon))$, then there is a coupling $(X_t, Y_t)_{t \in \mathbb{N}}$ for $\mathfrak{M}[d]$ such that for certain $\mathbb{T}^*$,*

$$\mathbb{T}^* = \Theta\left(m \cdot n^2 + m^{0.75} \cdot n^{4.75} \cdot (\ln(m/\varepsilon))^{2.25} \cdot \ln(1/\varepsilon)\right);$$

it holds for any $t \ge \mathbb{T}^*$ that

$$\mathbf{Pr}[X_t \neq Y_t \mid (X_0, Y_0) = (X, Y)] \le \varepsilon.$$

Now, Lemma 3.11 directly implies the Short Memory Lemma, Lemma 1.2, for all values of d.

4. A reduction to a polynomial number of balls for Greedy[d]. In this section, we discuss our main use of the Short Memory Lemma which is a reduction of the analysis of the problem of allocating an arbitrary number m of balls into n bins to the case when m is upper-bounded by a polynomial of n. If we combine this analysis with the analysis for a polynomial number of balls in section 2, we will immediately obtain Theorem 1.3.

Our arguments are similar to those used in section 3.3. We begin with the following corollary which follows directly from Lemma 1.2.

COROLLARY 4.1. *Let X_0 be any normalized load vector describing an arbitrary allocation of some number m of balls to n bins. Let Δ be the difference between the maximum and the minimum load in X_0. Let Y_0 be the normalized load vector describing the optimally balanced allocation of m balls into n bins (that is, each bin in Y_0 has either $\lfloor m/n \rfloor$ or $\lceil m/n \rceil$ balls). Let X_k and Y_k, respectively, denote the vectors obtained after inserting $k \geq 1$ further balls into both systems using* Greedy[d]. *For every constant α there is a constant c such that if $k \geq cn^7 \Delta \ln^4(n\Delta)$, then*

$$||\mathcal{L}(X_k) - \mathcal{L}(Y_k)|| \;\leq\; k^{-\alpha}.$$

Proof. Let ℓ denote the minimum load of any bin in X_0. We consider the scenario after removing ℓ balls from each bin in X_0 and Y_0; let X_0^* and Y_0^* be the respective load vectors. X_0^* and Y_0^* have an identical number of balls that we denote by m^*. Observe that since the maximum load in X_0 was $\ell + \Delta$, we have $m^* \leq n\Delta$.

Next, let X_t^* and Y_t^*, respectively, denote the vectors obtained after inserting $t \geq 1$ further balls to the systems corresponding to X_0^* and Y_0^*, where we use Greedy[d] to place the balls. We apply the Short Memory Lemma, Lemma 1.2, to the sequences X_t^* and Y_t^* to obtain that there is $\tau \leq c' m^* n^6 \ln^4(1/\varepsilon) \leq c'\Delta\, n^7 \ln^4(1/\varepsilon)$ for a suitable constant $c' > 0$, such that for every $t \geq \tau$ we have $||\mathcal{L}(X_t^*) - \mathcal{L}(Y_t^*)|| \leq \varepsilon$. Therefore, if we set $\varepsilon = k^{-\alpha}$ and choose k such that it satisfies $k = c'\Delta\, n^7\, \alpha \ln^4 k = O(n^7 \Delta \ln^4(n\Delta))$, we have $||\mathcal{L}(X_k^*) - \mathcal{L}(Y_k^*)|| \leq k^{-\alpha}$.

Now, the claim follows from the fact that for every $t \geq 0$ the distributions of X_t and Y_t^*, and Y_t and Y_t^*, respectively, differ only in that every bin corresponding to X_t (Y_t) has ℓ less balls than the corresponding bin in X_t^* (Y_t^*, respectively). ∎

Using Corollary 4.1, we present a general transformation which shows that the allocation obtained by an allocation process with "short memory" is more or less independent of the number of balls. The following theorem shows that the allocation (its distribution) is essentially determined after inserting a polynomial number of balls. In particular, this theorem together with Lemma 1.1 immediately imply Theorem 1.3. We assume that n is sufficiently large.

THEOREM 4.2. *Let us consider the process* Greedy[d], *$d \geq 2$, in which the balls are allocated into n bins. For any integer m, let $X_m = (x_1^{(m)}, \ldots, x_n^{(m)})$ be a load vector obtained after allocating m balls with* Greedy[d] *and let $\widetilde{X}_m = (x_1^{(m)} - \frac{m}{n}, \ldots, x_n^{(m)} - \frac{m}{n})$. Let $N = n^{36}$. Then, for every M, $M \geq N$, that is a multiple of n,*

$$||\mathcal{L}(\widetilde{X}_M) - \mathcal{L}(\widetilde{X}_N)|| \;\leq\; N^{-\alpha},$$

where α denotes an arbitrary constant.

Proof. Let us first begin with the claim that if M and m are multiples of n with $M \geq n^{36}$ and $M \geq m \geq M^{0.8}$, then

$$||\mathcal{L}(\widetilde{X}_M) - \mathcal{L}(\widetilde{X}_m)|| \;\leq\; M^{-\alpha}. \tag{3}$$

Let us set $m' = M - m$. We use the majorization from the single-choice process to estimate the distribution of the bins' loads after inserting m' balls. Since Greedy[d] is majorized by the single-choice process, each bin contains $\frac{m'}{n} \pm \mathcal{O}(\sqrt{m' \ln(n/p)/n})$ balls, with probability at least $1 - p$ for any $p \in [0, 1]$. We set $\Delta = M^{0.6}$ and $p = M^{-\alpha}/2$. Applying $M \geq m' \geq n$ yields that every bin contains between $\frac{m'}{n} - \Delta/2$ and $\frac{m'}{n} + \Delta/2$ balls, with probability at least $1 - p$, provided that n is sufficiently large. Let us now condition on this event and assume that the entries in $\tilde{X}_{m'}$ are in the Δ-range specified above.

Let Y describe another system in which the first m' balls are inserted in an optimal way; that is, $Y_{m'} = (\frac{m'}{n}, \ldots, \frac{m'}{n})$. Now, we add m balls using protocol Greedy[d] on top of $X_{m'}$ and $Y_{m'}$, respectively. Now, applying Corollary 4.1, we obtain $||\mathcal{L}(X_M) - \mathcal{L}(Y_M)|| \leq m^{-2\cdot\alpha} \leq M^{-\alpha}/2$ as $m \geq M^{0.8} \geq n^7 M^{0.6} \ln^4 M \geq c\, n^7 \Delta \ln^4(n\, \Delta)$, where c is the constant specified in the corollary. Thus, conditioned on the event that the values in $\tilde{X}_{m'}$ are in the interval $[\frac{m'}{n} - \Delta/2, \frac{m'}{n} + \Delta/2]$, we have $\mathcal{L}(\widetilde{Y}_M) = \mathcal{L}(\widetilde{X}_m)$, with probability at least $1 - M^{-\alpha}/2$. Therefore, since the condition is satisfied with probability at least $1 - M^{-\alpha}/2$ as well, we have

$$||\mathcal{L}(\widetilde{X}_M) - \mathcal{L}(\widetilde{X}_m)|| \ \leq \ M^{-\alpha}/2 + ||\mathcal{L}(X_M) - \mathcal{L}(Y_M)|| \ \leq \ M^{-\alpha}/2 + M^{-\alpha}/2 \ = \ M^{-\alpha},$$

which completes the proof of inequality (3).

Finally, we use inequality (3) to prove Theorem 4.2. Observe first that if $M \leq N^{1/0.8}$, then (3) directly implies Theorem 4.2. Otherwise, we have to apply inequality (3) repeatedly as follows. Let $m_0, m_1, \ldots, m_k$ denote a sequence of integers such that $m_0 = N$, $m_k = M$, $m_i^{0.8} \leq m_{i-1}$, and $m_i^{\alpha} \geq 2m_{i-1}^{\alpha}$. Then

$$||\mathcal{L}(\widetilde{X}_M) - \mathcal{L}(\widetilde{X}_N)|| \ \leq \ \sum_{i=1}^{k} ||\mathcal{L}(\widetilde{X}_{m_i}) - \mathcal{L}(\widetilde{X}_{m_{i-1}})|| \ \leq \ \sum_{i=1}^{k} m_i^{-\alpha} \ \leq \ N^{-\alpha},$$

where the last inequality follows from the fact that $m_i^{-\alpha} \leq 2^{-i} m_0^{-\alpha} = 2^{-i} N^{-\alpha}$. □

Remark 2. It is easy to see that the proof above does not use any of the properties of Greedy[d] but the following two: Corollary 4.1 and the fact that Greedy[d] is majorized by the single-choice process. Therefore, Theorem 4.2 holds for any allocation protocol $\mathcal{P}$ that has short memory (in the sense of Corollary 4.1) and that is majorized by the single-choice process.

5. Greedy[d] majorizes Left[d].

In this section, we will begin our analysis of the always-go-left allocation scheme and prove Theorem 1.7. Let $d \geq 2$, n be any multiple of d, and $m \geq 0$. We show that Left[d] is majorized by Greedy[d].

Our proof is by induction on the number of balls in the system. Let u denote the load vector obtained after inserting some number of balls with Left[d], and let v denote the load vector obtained after inserting the same number of balls with Greedy[d]. Without loss of generality, we assume that u and v are normalized, i.e., $u_1 \geq u_2 \geq \cdots \geq u_n$ and $v_1 \geq v_2 \geq \cdots \geq v_n$. Notice that the normalization of u jumbles the bins in the different groups used by Left[d] in some unspecified way so that it remains unclear which bin belongs to which group. Let u' and v' denote the load vectors obtained by adding another ball b with Left[d] and Greedy[d], respectively. To prove Theorem 1.7 by induction, we show that if $u \leq v$ then there is a coupling of Left[d] and Greedy[d] with respect to the allocation of b such that $u' \leq v'$, regardless of the unspecified mapping of the bins to the groups underlying u.

As a first step in the description of the coupling, we replace the original formulations of the allocation rules of the two random processes by alternative formulations that enable us to define an appropriate coupling.

- At first, we describe the alternative formulation of the allocation rules for Greedy[d]. For $1 \leq i \leq n$, let $\mathbf{e}_i$ denote the ith unit vector, and define $b_i = \mathbf{Pr}[v' = v + \mathbf{e}_i]$. For $0 \leq i \leq n$, let $B_i = \sum_{j=1}^{i} b_i$; that is, B_i denotes the probability that the next ball is added to a bin with index at most i with respect to the considered order of bins. Without loss of generality, we assume that Greedy[d] gives the ball to the bin with larger index in case of a tie. Then $B_i = (i/n)^d$ because Greedy[d] places the ball b in a bin with index smaller than or equal to i if and only if all of the d locations of b point to bins whose indices are at most i. Instead of inserting the next ball using the rules of Greedy[d], we now choose a continuous random variable x uniformly at random from the interval $[0, 1]$ and allocate b in the ith bin if $B_{i-1} < x \leq B_i$. By our construction, this results in the same distribution.
- Now we turn our attention to Left[d]. Given any allocation of the balls to bins corresponding to the load vector u, define $a_i = \mathbf{Pr}[u' = u + \mathbf{e}_i]$ for $1 \leq i \leq n$ and $A_i = \sum_{j=1}^{i} a_i$ for $0 \leq i \leq n$. Observe that the probabilities a_i and A_i do not depend only on the index i (as in the case of Greedy[d]) or the vector u, but also on the hidden mapping of the bins to the groups. Consequently, we cannot specify these terms as a functional of i or u. Nevertheless, for any given mapping of the bins to groups, the terms $A_0, \ldots, A_n$ are well defined so that we can replace the original allocation rules by the following rule that results in the same distribution: Choose a random variable x uniformly at random from the interval $[0, 1]$ and allocate the ball b into the ith bin if $A_{i-1} < x \leq A_i$.

For the coupling, we now assume that Left[d] and Greedy[d] use the same random number x to assign the ball b. By our construction, this coupling is faithful. Under the coupling, we have to show $u' \leq v'$. Let $u' = u + \mathbf{e}_i$ and $v' = v + \mathbf{e}_j$ for some i and j; that is, i and j specify the indices of the bins in the vectors u and v into which Left[d] and Greedy[d], respectively, put the ball b.

First, let us assume that the initial vectors u and v are equal. In this case, we have to show that $u + \mathbf{e}_i \leq u + \mathbf{e}_j$. Consider the *plateaus* of u, i.e., maximal index sets of bins with the same height. Suppose there are $k \geq 2$ plateaus $U_1, \ldots, U_k$ such that the load is decreasing from U_1 to U_k. Let I and J denote the indices of the plateaus that contain i and j, respectively. Observe that $I \geq J$ implies $u + \mathbf{e}_i \leq u + \mathbf{e}_j$ because adding a ball to different positions of the same plateau results in the same normalized vector. Thus, we have only to show that $J \leq I$. Let $\ell = \max\{U_I\}$. Since $i \leq \ell$, we have $x \leq A_\ell$. In the following lemma, we show that $A_\ell \leq (\ell/n)^d$. Above we have shown that $B_\ell = (\ell/n)^d$. Therefore, the lemma implies $x \leq B_\ell$, which shows that Greedy[d] places its ball in a bin with index at most ℓ; that is, $j \leq \ell$ and, hence, $J \leq I$. Consequently, $u + \mathbf{e}_i \leq u + \mathbf{e}_j$.

LEMMA 5.1. *For any mapping of the bins to the groups underlying the vector u, $A_\ell \leq (\ell/n)^d$.*

Proof. Recall that A_ℓ corresponds to the probability that Left[d] places the ball b in a location with index (with respect to u) smaller than or equal to ℓ. Let ℓ_k for $0 \leq k < d$ denote the number of bins in group k with load greater than or equal to u_ℓ. Then $A_\ell = \prod_{k=0}^{d-1} \frac{\ell_k}{n/d}$ because the kth location of b must be one of those ℓ_k bins among the d/n bins in group k that have load at least u_ℓ. Since $\sum_{k=0}^{d-1} \ell_k = \ell$, A_ℓ is maximized for $\ell_0 = \cdots = \ell_{d-1} = \ell/d$. Consequently, $A_\ell \leq (\ell/n)^d$. $\square$

Until now we have analyzed only the case when $u = v$ and have shown $u + \mathbf{e}_i \leq u + \mathbf{e}_j$. This, however, can be easily generalized to arbitrary normalized load vectors. Indeed, for any two normalized load vectors w and w', $w \leq w'$ implies $w + \mathbf{e}_i \leq w' + \mathbf{e}_i$, cf. [1, Lemma 3.4]. Consequently, we can conclude from $u + \mathbf{e}_i \leq u + \mathbf{e}_j$ that $u' = u + \mathbf{e}_i \leq u + \mathbf{e}_j \leq v + \mathbf{e}_j = v'$. Thus, Theorem 1.7 is shown.

6. Analysis of Left[d]. In this section, we investigate the allocation generated by Left[d]. In particular, we prove Theorem 1.5, that is, we show that the number of bins with load at least $\frac{m}{n} + i + \gamma$ is at most $n \cdot \exp(-\phi_d^{d \cdot i})$, w.h.p., where ϕ_d denotes the d-ary golden ratio (cf. section 1.1) and γ is a suitable constant. Similarly to the proof for Greedy[d], we divide the set of balls into batches of size n and we apply an induction on the number of batches. On one hand, the proof for Left[d] is slightly more complicated since we have to take into account that the set of bins is partitioned into d groups. On the other hand, we can avoid the detour through analyzing the holes below average height as we can instead make use of the majorization of Left[d] by Greedy[d].

For the time being, let us assume that $m \geq n \log_2 n$. We will use the majorization from Greedy[d] to estimate the allocation after allocating the first $m' = m - n \log_2 n$ balls. The special properties of Left[d] will only be taken into account for the remaining $n \log_2 n$ balls. Let us divide the set of these balls into $\log_2 n$ batches of size n each. Let time 0 denote the point of time before the first ball from batch 1 is inserted, that is, after inserting the first m' balls; and, for $1 \leq t \leq \log_2 n$, let time t denote the point of time after inserting the balls from batch t. Furthermore, set $\Gamma = \frac{m'}{n} + 7$ and, for $i \geq 0, 0 \leq j < d, 0 \leq t \leq \log_2 n$, let $\nu_{i,j}^{(t)}$ denote the number of bins with load at least $\Gamma + t + i$ in group j at time t. The following lemma gives an upper bound on the allocation of Left[d] obtained by the majorization from Greedy[d] at time 0. This upper bound is specified in terms of the function

$$h_0(i) \;=\; \frac{1}{4^i \cdot 64d}.$$

Later we will use the same lemma to estimate parts of the allocation also for other points of time, $t \geq 1$.

LEMMA 6.1. *Let ℓ denote the smallest integer such that $h_0(\ell) \leq n^{-0.9}$, i.e., $\ell = \lfloor 0.9 \log_4 n - \log_4 d \rfloor - 2$. For $0 \leq i < \ell, 0 \leq j < d, 0 \leq t \leq \log_2 n$, it holds $\nu_{i,j}^{(t)} \leq h_0(i) \cdot n/d$, w.h.p. For $i \geq \ell$, $\nu_{i,j}^{(t)} = 0$, w.h.p.*

Proof. Fix a time step t. Theorem 1.3 shows that, when using Greedy[d], the fraction of bins with load at least $\frac{m'}{n} + t + i$ is upper-bounded by a function that decreases doubly exponentially in i. Now, in order to simplify the subsequent calculations, we upper-bound this function by another function that decreases only exponentially in i, namely, the function h_0. With lots of room to spare, the analysis in section 2.3 yields that the fraction of bins with load at least $\Gamma + t + i$ can be upper-bounded by $h_0(i)/(2d)$, w.h.p., provided n is sufficiently large. This result holds for Greedy[d], and we want to apply it to Left[d] via majorization. In order to make use of the majorization of Left[d] by Greedy[d], we need a bound on the number of balls above some given height rather than a bound on the number of bins with load above the height. However, since the bound given above on the number of bins decreases geometrically in i, the number of balls of height at least $\Gamma + t + i$ when using Greedy[d] is bounded from above by $h_0(i) \cdot n/d$. Now, because of the majorization, this result holds for Left[d], too. In turn, the number of balls above height $\Gamma + t + i$ upper-bounds the number of bins with load at least $\Gamma + t + i$. Hence, when using Left$[d]$, the total number of bins with load at least $\Gamma + t + i$ is bounded from above by $h_0(i) \cdot n/d$. Of course, the same upper bound holds for the number of such bins in each individual group.

Finally, it remains to be shown that $\nu_{i,j}^{(t)} = 0$, w.h.p., for $i \geq \ell$ with $\ell = \lfloor 0.9 \log_4 n - \log_4 d \rfloor - 2$. Again this follows via majorization from Greedy$[d]$. Theorem 1.3 implies that the maximum load of Greedy$[d]$ and, hence, also of Left$[d]$ at time t is bounded from above by $\Gamma + t + O(\log_d \log n)$, w.h.p. Thus there is no ball with height $\Gamma + t + \ell$, w.h.p. $\square$

Before we turn to the technical details, let us explain the high-level idea behind the following analysis. We will use a function $f(k,t)$ as an upper bound for $\nu_{\lfloor k/d \rfloor, j \bmod d}^{(t)}$. For $t = 0$, $f(k,t)$ will be set equal to $h_0(\lfloor k/d \rfloor)$, and we will use the above lemma to show that $f(k,0)$ upper-bounds $\nu_{\lfloor k/d \rfloor, k \bmod d}^{(0)}$. When increasing t, the function $f(k,t)$ will become more similar to the function $h_1(k)$ defined by

$$h_1(k) = \frac{\exp(-F_d(k-d+1))}{64d},$$

where $F_d(k)$ denotes the kth d-ary Fibonacci number as defined in section 1.1. Let $i = \lfloor k/d \rfloor$ and $j = k \bmod d$. Then $h_1(k)$ will serve as an upper bound on the fraction of bins with height $\Gamma + t + i$ in group j. As explained in section 1.1, we use the d-ary golden ratio to upper-bound the d-ary Fibonacci numbers. This way,

$$h_1(k) = \frac{\exp\left(-\phi_d^{k \pm O(d)}\right)}{64d} = \frac{\exp\left(-\phi_d^{(i \pm O(1))d}\right)}{64d}.$$

Hence, for large t, the fraction of bins with some given height decreases "Fibonacci exponentially" with the height, exactly as described in Theorem 1.5.

Now we come to the technical details. We define

$$f(k,t) = \max\{h_0(\lfloor k/d \rfloor) \cdot 2^{-t}, h_1(k)\}.$$

Observe that f changes smoothly from h_0 into h_1 when increasing t. In particular, $f(k,0) = h_0(\lfloor k/d \rfloor)$ and $f(k, \log_2 n) \leq h_1(k) + \frac{1}{n}$. We need to refine the function f slightly. Intuitively, we truncate the function when the function values become "too small." Let c denote a sufficiently large constant term, whose value will be specified later. Let ℓ_t denote the smallest integer such that $f(\ell_t, t) \leq n^{-0.9}$. We set

$$f'(k,t) = \begin{cases} \max\left\{\dfrac{n^{-0.9}}{4}, f(k,t)\right\} & \text{if } 0 \leq k < \ell_t + d, \\ \dfrac{cd}{n} & \text{if } k \geq \ell_t + d. \end{cases}$$

The following properties of f' are crucial for our analysis. They hold only if n is sufficiently large.

LEMMA 6.2.

B1. $f'(k,t) = h_0(0)$ *for* $0 \leq k < d$, $t \geq 0$*;*

B2. $f'(k,t) \geq 2 \cdot f'(k+d, t-1)$ *for* $d \leq k < \ell_t + d$, $t \geq 1$*;*

B3. $f'(k,t) \geq (4d) \cdot \prod_{j=1}^{d} f'(k-j, t)$ *for* $d \leq k < \ell_t + d$, $t \geq 0$*;*

B4. $f'(k,t) \geq n^{-0.9}/4$ *for* $d \leq k < \ell_t$, $t \geq 0$.

Proof. We start with the proof of property B1. First, let us check the property for f instead of f'. For $0 \le k < d$, $F_d(k-d+1) = 0$ so that $h_1(k) = 1/(64d) = h_0(0)$. Hence, for every $t \ge 0$,

$$f(k,t) \;=\; \max\{h_0(\lfloor k/d \rfloor) \cdot 2^{-t}, h_1(k)\} \;=\; h_0(0).$$

If n is sufficiently large then the same is true for f'.

Next we show property B2, first for f and then for f'. The function $f(k,t)$ is defined by the maximum of the two terms $h_0(\lfloor k/d \rfloor) \cdot 2^{-t}$ and $h_1(k)$. We study these terms one after the other. The definition of h_0 immediately implies

$$h_0(\lfloor k/d \rfloor) \cdot 2^{-t} \;=\; 4h_0(\lfloor (k+d)/d \rfloor) \cdot 2^{-t} \;=\; 2h_0(\lfloor (k+d)/d \rfloor) \cdot 2^{-(t-1)}.$$

Furthermore, for $k \ge d$,

$$h_1(k) \;=\; \frac{\exp(-F_d(k-d+1))}{64d} \;\ge\; \frac{2\exp(-F_d(k+1))}{64d} \;=\; 2h_1(k+d).$$

As a consequence, $f(k,t) \ge 2f(k+d, t-1)$, that is, B2 is shown for f. The refinement from f to f' might raise the right-hand side of the inequality from $2f(k+d,t-1)$ to the value $2n^{-0.9}/4$, or the right-hand side might take the value $2cd/n$. At first, suppose $f'(k+d,t-1) = n^{-0.9}/4$. Then $k < \ell_{t-1}$ so that $f(k,t-1) \ge n^{-0.9}$. Now this implies $f(k,t) \ge n^{-0.9}/2$ as $f(k,t) \ge f(k,t-1)/2$. Consequently,

$$f'(k,t) \;=\; \max\left\{f(k,t), \frac{n^{-0.9}}{4}\right\} \;\ge\; \frac{n^{-0.9}}{2} \;=\; 2f'(k+d,t-1).$$

In the second case, $f'(k+d,t-1) = cd/n$. Observe that property B2 needs to be shown only for $k < \ell_t + d$. For this choice of k, $f'(k,t) \ge n^{-0.9}/4$ so that $f'(k,t) \ge 2f'(k+d,t-1)$ if n is sufficiently large. Hence, B2 is shown.

Property B3 is shown as follows. Again we first show the property for f and then for f'. Fix $k \ge d$. Depending on the outcome of the terms $f(k-d,t), \ldots, f(k-1,t)$, we distinguish two cases. First, suppose there exists $\delta \in \{1,\ldots,d\}$ such that $f(k-\delta,t) = h_0(\lfloor (k-\delta)/d \rfloor) \cdot 2^{-t}$. Observe that $h_0(\lfloor (k-\delta)/d \rfloor) \le h_0(\lfloor k/d \rfloor - 1) = 4\,h_0(\lfloor k/d \rfloor)$. We obtain

$$\begin{aligned}
\prod_{j=1}^{d} f(k-j,t) &= h_0(\lfloor (k-\delta)/d \rfloor) \cdot 2^{-t} \cdot \prod_{\substack{j=1\\ j\ne\delta}}^{d} f(k-j,t) \\
&\le 4\,h_0(\lfloor k/d \rfloor) \cdot 2^{-t} \cdot \left(\frac{1}{64d}\right)^{d-1} \\
&\le \frac{h_0(\lfloor k/d \rfloor) \cdot 2^{-t}}{4d} \\
&\le \frac{f(k,t)}{4d}.
\end{aligned}$$

Second, suppose $f(k-\delta,t) = h_1(k-\delta)$ for all $\delta \in \{1,\ldots,d\}$. Then

$$\begin{aligned}\prod_{j=1}^{d} f(k-j,t) &= \prod_{j=1}^{d} \frac{\exp\left(-F_d(k-j-d+1)\right)}{64d} \\ &= \frac{\exp\left(-\sum_{j=1}^{d} F_d(k-j-d+1)\right)}{(64d)^d} \\ &\le \frac{\exp\left(-F_d(k-d+1)\right)}{(4d)(64d)} \\ &= \frac{h_1(k)}{4d} \\ &\le \frac{f(k,t)}{4d}.\end{aligned}$$

The refinement from f to f' affects the above proof only if $f'(k,t) \neq f(k,t)$, since otherwise, $f(k-\delta,t) = f'(k-\delta,t)$ for all $0 \le \delta \le d$, so that the above arguments hold. If $f'(k,t) \neq f(k,t)$ then $f'(k,t) = n^{-0.9}/4$. In this case, either $f'(k-1,t)$ might take the value $n^{-0.9}/4$ as well or it takes the value $f(k-1,t)$. In the latter case, B3 follows by the same arguments as before, if we additionally apply $f(k,t) \le n^{-0.9}/4 = f'(k,t)$. If both $f'(k,t)$ and $f'(k-1,t)$ take the value $n^{-0.9}/4$, then

$$\prod_{j=1}^{d} f(k-j,t) \;\le\; \frac{n^{-0.9}}{4} \cdot \prod_{j=2}^{d} f(k-j,t) \;\le\; \frac{n^{-0.9}}{4} \cdot \frac{1}{4d} \;=\; \frac{f(k,t)}{4d}.$$

Thus, B1, B2, and B3 hold for f and f'. B4 does not hold for f. However, our refinement explicitly ensures this property for f'. ◻

Based on these properties we prove now that the following invariants hold w.h.p. for every $t \in \{0,\ldots,\log_2 n\}$. We say that a ball has *index k at time t* if the ball belongs to one of the batches $1,\ldots,t$ and it is placed in group $k \bmod d$ with height $\lfloor(\Gamma+t+k)/d\rfloor$.

- $H_1(t)$: $\nu_{i,j}^{(t)} \le f'(id+j,t)\cdot n/d$ for $i \ge 0, 0 \le j < d$.
- $H_2(t)$: The number of balls with index at least $\ell_t + d$ at time t is bounded from above by a constant term c.

Observe that these invariants imply the bounds given in Theorem 1.5 as the function $f'(i,\log_2 n)$ decreases "Fibonacci exponentially" in i as discussed above.

We show the invariants H_1 and H_2 by an induction on the number of rounds t. Lemma 6.1 gives that the invariants hold at time 0. In the following, we prove that $H_1(t)$ and $H_2(t)$ hold w.h.p. assuming that $H_1(t-1)$ and $H_2(t-1)$ are given. Fix $t \in \{1,\ldots,\log_2 n\}$. First, we consider $H_1(t)$. We prove this invariant by a further induction on $k = id+j$. Observe that we need only to prove the invariant for $k < \ell_t+d$ as the upper bound given for $k \ge \ell_t + d$ is a direct consequence of invariant H_2. For $k \in \{0,\ldots,d-1\}$, property B1 gives $f'(k,t) = h_0(0)$. Hence, for $k < d$, invariant H_1 follows again directly from Lemma 6.1.

Now assume $d \le k < \ell_t + d$. Suppose $H_1(t)$ is shown for all $k' < k$. For $i = \lfloor k/d \rfloor$ and $j = k \bmod d$, let $q(k) = q(i,j)$ denote the number of bins of group j containing $\Gamma + t + i$ balls already at the beginning of round t, and let $p(k) = p(i,j)$ denote the number of balls from batch t that are placed into a bin of group j that contains at least $\Gamma + t + i - 1$ balls. Clearly,

$$\nu_{i,j}^{(t)} \;\le\; q(k) + p(k).$$

In the following, we calculate upper bounds for $q(k)$ and $p(k)$.

Observe that $q(k) = q(i,j)$ corresponds to $\nu^{(t-1)}_{i+1,j}$. Hence, invariant $H_1(t-1)$ gives

$$q(k) \;\leq\; f'((i+1)d+j,t-1)\cdot\frac{n}{d} \;\leq\; f'(k+d,t-1)\cdot\frac{n}{d} \overset{\text{(B2)}}{\leq} 0.5\cdot f'(k,t)\cdot\frac{n}{d}$$

for $d \leq k < \ell_t + d$.

The term $p(k) = p(i,j)$ can be estimated as follows. If a ball is placed into a bin of group j with $\Gamma + t + i - 1$ balls, the d possible locations for that ball fulfill the following conditions. The randomly picked location from group g, $0 \leq g < j$, points to a bin with load at least $\Gamma + t + i$. (Otherwise, the always-go-left scheme would assign the ball to that location instead of location j.) At the ball's insertion time the number of these bins is at most $\nu^{(t)}_{i,g}$. By the induction on k, $\nu^{(t)}_{i,g} \leq f'(i\cdot d+g,t)\cdot n/d$. Thus, the probability that the location points to a suitable bin is at most $f'(i\cdot d+g,t)$. Furthermore, the randomly picked location from group g, $j \leq g < d$, points to a bin with load at least $\Gamma+t+i-1$. At the ball's insertion time, the number of these bins is at most $\nu^{(t)}_{i-1,g}$. Thus, the probability for this event is at most $f'((i-1)\cdot d+g,t)$. Now multiplying the probabilities for all d locations yields that the probability that a fixed ball is allocated to group j with height $\Gamma + t + i$ or larger is at most

$$\prod_{g=0}^{j-1} f'(i\cdot d+g,t)\cdot\prod_{g=j}^{d-1} f'((i-1)\cdot d+g,t) \;=\; \prod_{g=1}^{d} f'(k-g,t) \overset{\text{(B3)}}{\leq} \frac{f'(k,t)}{4d}$$

for $d \leq k < \ell_t + d$. Taking into account all n balls of batch t, we obtain $\mathbf{E}[p(k)] \leq n\cdot f'(k,t)/(4d)$. Applying a Chernoff bound yields

$$\mathbf{Pr}\left[p(k) \geq 2n\cdot\frac{f'(k,t)}{4d}\right] \;\leq\; \exp\left(-n\cdot\frac{f'(k,t)}{8d}\right) \overset{\text{(B4)}}{\leq} \exp\left(-\frac{n^{0.1}}{32d}\right).$$

As a consequence, $p(k) \leq 0.5 f'(k,t)\cdot n/d$, w.h.p.

Combining the bounds on $q(k)$ and $p(k)$ gives

$$\nu^{(t)}_{i,j} \;\leq\; q(k)+p(k) \;\leq\; f'(k,t)\cdot\frac{n}{d} \;\leq\; f'(id+j,t)\cdot\frac{n}{d}$$

for $d \leq k = id + j < \ell_t + d$. Thus, invariant $H_1(t)$ is shown.

Now we turn to the proof of $H_2(t)$. For $s \geq 0$, let L_s denote the number of balls with index at least $\ell_s + d$ at time s. Using this notation, invariant $H_2(t)$ states that $L_t \leq c$. For $s \geq r \geq 1$, let $L_s(r)$ denote the number of balls from batch r with index at least $\ell_s + d$ at time s. We claim

$$L_t \;=\; \sum_{s=1}^{t} L_t(s) \;\leq\; \sum_{s=1}^{t} L_s(s).$$

The first equation follows directly from the definition. The second equation can be seen as follows. First, observe that ℓ_{s-1} might be larger than ℓ_s as the function f' decreases over time. However, property B2 combined with the fact that f' decreases by at most a factor of two from time $s-1$ to time s yields $\ell_{s-1} \leq \ell_s + d$ for every $s \geq 0$, which implies $L_s(r) \leq L_{s-1}(r)$ for every $r \leq s-1$. By induction, we obtain $L_t(r) \leq L_s(r)$ for $r \leq s \leq t$ and especially $L_t(s) \leq L_s(s)$.

Let us study the probability that a fixed ball from batch s has index at least $\ell_s + d$ at time s and, hence, contributes to $L_s(s)$. This event happens only if each of the d randomly selected locations of the ball points to a bin whose topmost ball has index at least $\ell_s + d - d = \ell_s$. invariant $H_1(s)$ yields that the fraction of such bins in each group is at most $f(\ell_s, s) \leq n^{-0.9}$. Thus, the probability that a ball from batch s contributes to $L(s)$ is at least $n^{-0.9d} \leq n^{-1.8}$. Now let us estimate the probability that there exist c balls from the batches 1 to t that fulfill this condition. This probability is at most

$$\binom{tn}{c} \cdot \left(\frac{1}{n^{1.8}}\right)^c \leq \left(\frac{\log_2 \mathrm{n}}{cn^{0.8}}\right)^c \leq n^{-c/2},$$

where the last inequality holds for sufficiently large n. Consequently, with probability at least $1 - n^{-c/2}$, $L_t \leq \sum_{s=1}^{t} L_s(s) \leq c$. Thus, we have shown that $H_1(0), \dots, H_1(t)$ imply $H_2(t)$, w.h.p. This completes the proof for the case $m \geq n \log_2 n$.

Finally, let us investigate the case $m < n \log_2 n$. We break the set of balls into at most $t \leq \log n$ batches. All batches except for the last one contain exactly n balls; only the last batch might contain less. We use a simplified variant of the above analysis. In particular, we define $f(k,t) = h_1(k)$ instead of $f(k,t) = \max\{h_0(\lfloor k/d \rfloor) \cdot 2^{-t}, h_1(k)\}$. The invariants H_1 and H_2 can be shown by the same arguments as before. The advantage is that the identity between $f(k,t)$ and $h_1(k)$ is given from the beginning on, so that one does not have to iterate for $\log_2 n$ batches until the two functions become similar. In other words, the invariants H_1 and H_2 imply the bounds described in the theorem already after the first batch as well as after all subsequent batches. This completes the proof of Theorem 1.5.

7. Conclusions. We have presented the first tight analysis of two balls-into-bins multiple-choice processes: the greedy protocol of Azar et al. [1] and the always-go-left scheme due to Vöcking [33]. We showed that these schemes result in a maximum load (w.h.p.) of only $\frac{m}{n} + \frac{\ln \ln n}{\ln d} + \Theta(1)$ and $\frac{m}{n} + \frac{\ln \ln n}{d \ln \phi_d} + \Theta(1)$, respectively. Both these bounds are tight up to additive constants. In addition, we have given upper bounds on the number of bins with any given load. Furthermore, we presented the first comparative study of the two multiple-choice algorithms and gave a majorization result showing that the always-go-left scheme obtains a stochastically better load balancing than the greedy scheme for any choice of d, n, and m.

Our important technical contribution is the Short Memory Lemma, which informally states that the multiple-choice processes quickly "forget" their initial distribution of balls. The great consequence of this property is that the deviation of the multiple-choice processes from the optimal allocation (that is, the allocation in which each bin has either $\lfloor m/n \rfloor$ or $\lceil m/n \rceil$ balls) does not increase with the number of balls as in the case of the single-choice process. This property played a fundamental role in our analysis. We also hope that it will find further applications in the analysis of allocation processes. In particular, we believe that the use of the Markov chain approach to estimate the mixing time of underlying stochastic processes will be an important tool in analyzing other balls-into-bins or similar processes.

Our calculations in section 2 use some help from computers; it would be interesting to come up with a more elegant analysis that would possibly provide more insight on the greedy process for a polynomial number of balls.

Acknowledgments. We thank anonymous referees for helpful suggestions. We are also grateful to Christian Scheideler for inspiring discussion about the paper.

REFERENCES

[1] Y. Azar, A. Z. Broder, A. R. Karlin, and E. Upfal, *Balanced allocations*, SIAM J. Comput., 29 (1999), pp. 180–200.

[2] M. Adler, P. Berenbrink, and K. Schröder, *Analyzing an infinite parallel job allocation process*, in Proceedings of the 6th Annual European Symposium on Algorithms, Venice, Italy, G. Bilardi, G. F. Italiano, A. Pietracaprina, and G. Pucci, eds., Lecture Notes in Comput. Sci. 1461, Springer-Verlag, Berlin, 1998, pp. 417–428.

[3] M. Adler, S. Chakrabarti, M. Mitzenmacher, and L. Rasmussen, *Parallel randomized load balancing*, Random Structures Algorithms, 13 (1998), pp. 159–188.

[4] D. Aldous, *Random walks of finite groups and rapidly mixing Markov chains*, in Séminaire de Probabilités XVII, 1981/82, J. Azéma and M. Yor, eds., Lecture Notes in Math. 986, Springer-Verlag, Berlin, 1983, pp. 243–297.

[5] P. Berenbrink, A. Czumaj, T. Friedetzky, and N. D. Vvedenskaya, *On the analysis of infinite parallel job allocation processes via differential equations*, in Proceedings of the 11th Annual ACM Symposium on Parallel Algorithms and Architectures, Bar Harbor, ME, 2000, ACM Press, New York, 2000, pp. 99–108.

[6] P. Berenbrink, F. Meyer auf der Heide, and K. Schröder, *Allocating weighted jobs in parallel*, Theory Comput. Syst., 32 (1999), pp. 361–386.

[7] R. Bubley and M. Dyer, *Path coupling: A technique for proving rapid mixing in Markov chains*, in Proceedings of the 38th IEEE Symposium on Foundations of Computer Science, Miami Beach, FL, 1997, IEEE Computer Society Press, Los Alamitos, CA, 1997, pp. 223–231.

[8] R. Cole, A. Frieze, B. M. Maggs, M. Mitzenmacher, A. W. Richa, R. K. Sitaraman, and E. Upfal, *On balls and bins with deletions*, in Proceedings of the 2nd International Workshop on Randomization and Approximation Techniques in Computer Science, Barcelona, Spain, 1998, Lecture Notes in Comput. Sci. 1518, M. Luby, J. Rolim, and M. Serna, eds., Springer-Verlag, Berlin, 1998, pp. 145–158.

[9] R. Cole, B. M. Maggs, F. Meyer auf der Heide, M. Mitzenmacher, A. W. Richa, K. Schröder, R. K. Sitaraman, and B. Vöcking, *Randomized protocols for low-congestion circuit routing in multistage interconnection networks*, in Proceedings of the 30th Annual ACM Symposium on Theory of Computing, Dallas, TX, 1998, ACM Press, New York, 1999, pp. 378–388.

[10] A. Czumaj, C. Riley, and C. Scheideler, *Perfectly balanced allocation*, in Proceedings of the 7th International Workshop on Randomization and Approximation Techniques in Computer Science, Princeton, NJ, 2003, Lecture Notes in Comput. Sci. 2764, S. Arora, K. Jansen, J. D. P. Rolim, and A. Sahai, eds., Springer-Verlag, Berlin, 2003, pp. 240–251.

[11] A. Czumaj and V. Stemann, *Randomized allocation processes*, Random Structures Algorithms, 18 (2001), pp. 297–331.

[12] A. Czumaj, *Recovery time of dynamic allocation processes*, Theory Comput. Syst., 33 (2000), pp. 465–487.

[13] A. Czumaj, P. Kanarek, M. Kutyłowski, and K. Loryś, *Delayed path coupling and generating random permutations via distributed stochastic processes*, in Proceedings of the 10th Annual ACM–SIAM Symposium on Discrete Algorithms, Baltimore, MD, 1999, ACM, New York, SIAM, Philadelphia, PA, 1999, pp. 271–280.

[14] D. Dubhashi and D. Ranjan, *Balls and bins: A study in negative dependence*, Random Structures Algorithms, 13 (1998), pp. 99–124.

[15] N. L. Johnson and S. Kotz, *Urn Models and Their Application: An Approach to Modern Discrete Probability Theory*, John Wiley and Sons, New York, 1977.

[16] R. M. Karp, M. Luby, and F. Meyer auf der Heide, *Efficient PRAM simulation on a distributed memory machine*, Algorithmica, 16 (1996), pp. 517–542.

[17] V. F. Kolchin, B. A. Sevast'yanov, and V. P. Chistyakov, *Random Allocations*, V. H. Winston and Sons, Washington, D.C., 1978.

[18] J. Korst, *Random duplicated assignment: An alternative to striping in video servers*, in Proceedings of the 5th ACM International Multimedia Conference, Seattle, WA, 1997, ACM Press, New York, 1997, pp. 219–226.

[19] M. Luby, D. Randall, and A. Sinclair, *Markov chain algorithms for planar lattice structures*, in Proceedings of the 36th IEEE Symposium on Foundations of Computer Science, Milwaukee, WI, 1995, IEEE Computer Society Press, Los Alamitos, CA, 1995, pp. 500–509.

[20] M. J. Luczak and E. Upfal, *Reducing network congestion and blocking probability through balanced allocation*, in Proceedings of the 40th IEEE Symposium on Foundations of Computer Science, New York, NY, 1999, IEEE Computer Society Press, Los Alamitos, CA, 1999, pp. 587–595.

[21] C. McDiarmid, *Concentration*, in Probabilistic Methods for Algorithmic Discrete Mathematics, M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed, eds., Algorithms and Combinatorics, Springer-Verlag, Berlin, 1998, pp. 195–247.

[22] M. Mitzenmacher, *Load balancing and density dependent jump Markov processes*, in Proceedings of the 37th IEEE Symposium on Foundations of Computer Science, Burlington, VT, 1996, IEEE Computer Society Press, Los Alamitos, CA, 1996, pp. 213–222.

[23] M. Mitzenmacher, *The Power of Two Choices in Randomized Load Balancing*, Ph.D. thesis, Computer Science Department, University of California at Berkeley, Berkeley, CA, 1996.

[24] M. Mitzenmacher, *Analyses of load stealing models based on differential equations*, in Proceedings of the 10th Annual ACM Symposium on Parallel Algorithms and Architectures, Puerto Vallarta, Mexico, 1998, ACM Press, New York, 1998, pp. 212–221.

[25] M. Mitzenmacher, *Studying balanced allocations with differential equations*, Combin. Probab. Comput., 8 (1999), pp. 473–482.

[26] M. Mitzenmacher, *On the analysis of randomized load balancing schemes*, Theory Comput. Syst., 32 (1999), pp. 292–301.

[27] M. Mitzenmacher, B. Prabhakar, and D. Shah, *Load balancing with memory*, in Proceedings of the 43rd IEEE Symposium on Foundations of Computer Science, Vancouver, BC, Canada, 2002, IEEE Computer Society Press, Los Alamitos, CA, 2002, pp. 799–808.

[28] M. Mitzenmacher, A. W. Richa, and R. Sitaraman, *The power of two random choices: A survey of techniques and results*, in Handbook on Randomized Algorithms, Volume I, S. Rajasekaran, P. M. Pardalos, J. H. Reif, and J. Rolim, eds., Kluwer, Dordrecht, The Netherlands, 2001, pp. 255–294.

[29] M. Mitzenmacher and B. Vöcking, *The asymptotics of selecting the shortest of two, improved*, in Proceedings of the 37th Allerton Conference on Communication, Control, and Computing, Urbana, IL, 1999, pp. 326–327.

[30] M. Raab and A. Steger, *"Balls into bins"—a simple and tight analysis*, in Proceedings of the 2nd International Workshop on Randomization and Approximation Techniques in Computer Science, Barcelona, Spain, 1998, M. Luby, J. Rolim, and M. Serna, eds., Lecture Notes in Comput. Sci. 1518, Springer-Verlag, Berlin, 1998, pp. 159–170.

[31] V. Stemann, *Parallel balanced allocations*, in Proceedings of the 8th Annual ACM Symposium on Parallel Algorithms and Architectures, Padua, Italy, 1996, ACM Press, New York, 1996, pp. 261–269.

[32] P. Sanders, S. Egner, and J. Korst, *Fast concurrent access to parallel disks*, Algorithmica, 35 (2003), pp. 21–55.

[33] B. Vöcking, *How asymetry helps load balancing*, J. ACM, 50 (2003), pp. 568–589.

[34] N. D. Vvedenskaya, R. L. Dobrushin, and F. I. Karpelevich, *Queueing system with selection of the shortest of two queues: An assymptotic approach*, Probl. Inform. Transm., 32 (1996), pp. 15–27.

[35] N. D. Vvedenskaya and Y. M. Suhov, *Dobrushin's Mean-Field Approximation for Queue with Dynamic Routing*, Technical report 3328, INRIA, Le Chesnay, France, 1997.