Electron. Commun. Probab. **26** (2021), article no. 34, 1–8.
https://doi.org/10.1214/21-ECP400
ISSN: 1083-589X

ELECTRONIC COMMUNICATIONS in PROBABILITY

The power of thinning in balanced allocation[*]

Ohad N. Feldheim[†] Ori Gurel-Gurevich[‡]

Abstract

Balls are sequentially allocated into n bins as follows: for each ball, an independent, uniformly random bin is generated. An overseer may then choose to either allocate the ball to this bin, or else the ball is allocated to a new independent uniformly random bin. The goal of the overseer is to reduce the load of the most heavily loaded bin after $\Theta(n)$ balls have been allocated. We provide an asymptotically optimal strategy yielding a maximum load of $(1+o(1))\sqrt{\frac{8\log n}{\log\log n}}$ balls.

Keywords: thinning; two-choices; one-retry; $(1+\alpha)$-choice; load balancing, balls and bins; balanced allocation; subsampling.
MSC2020 subject classifications: 60C05; 68W27.
Submitted to ECP on May 29, 2020, final version accepted on May 6, 2021.

1 Introduction and results

Fix $\rho > 0$ and consider an online model in which an overseer is monitoring the sequential allocation of $\lfloor \rho n \rfloor$ balls into n bins. Each ball is assigned a *primary allocation*, i.e., an independent, uniformly chosen random bin. Then, the overseer is given the choice to reject this primary allocation, in which case the ball is assigned a *secondary allocation* instead, that is, a new, independent, uniformly chosen random bin. The overseer's decision may depend on all past allocations, but not any future information. The resulting sequence allocations is called a *two-thinning* of the balls-and-bins process.

A *two-thinning strategy* is a function determining whether to accept or reject each suggested allocation, depending on all previous allocations (See formal definition in Section 2). Denote by $\mathrm{MaxLoad}_t^f([n])$ the load of the most heavily loaded bin after the player allocates $\lfloor t \rfloor$ balls into n bins, following the strategy f. A strategy f is *asymptotically optimal* if, for any strategy g with probability tending to one as $n \to \infty$ we have $\mathrm{MaxLoad}_{\rho n}^f([n]) \le (1+o(1))\,\mathrm{MaxLoad}_{\rho n}^g([n])$.

Here we describe and analyse an optimal two-thinning strategy which we call the *ℓ-threshold* strategy. This is the two-thinning strategy which rejects a ball whenever the number of primary allocations to the suggested bin at the allocation time is at least ℓ. Our main result is the following,

[*] Research was supported by the Israel Science Foundation (grant No. 1707/16).

[†] Hebrew University of Jerusalem, Israel. E-mail: Ohad.Feldheim@mail.huji.ac.il

[‡] Hebrew University of Jerusalem, Israel. E-mail: Ori.Gurel-Gurevich@mail.huji.ac.il

Theorem 1.1. *Let f be the $\sqrt{\frac{2\log n}{\log\log n}}$-threshold strategy for the allocation of $\lfloor \rho n \rfloor$ balls into n bins. Then f is asymptotically optimal and, with probability tending to one as $n \to \infty$,*

$$\mathrm{MaxLoad}^{f}_{\rho n}([n]) = (1+o(1))\sqrt{\frac{8\log n}{\log\log n}}.$$

1.1 Discussion

Balls-and-bins, two-choices and two-thinning. It is well known that if each of $\lfloor \rho n \rfloor$ balls is allocated independently to a uniformly chosen random bin in $[n] = \{1, \dots, n\}$, then the most heavily loaded bin contains $\frac{\log n}{\log\log n} + O(1)$ balls with high probability (see [8, Lemma 5.1 and Lemma 5.12]). In their seminal paper, Azar, Broder, Karlin and Upfal [3] have shown that a significantly lower maximum load of $\log_2 \log n + O(1)$ balls could be achieved, with high probability, in a *two-choices* setting, i.e., if the allocation of each ball is governed by an overseer who is offered a choice between two independent, uniformly chosen random bins. Moreover, the overseer can achieve this simply by following a naïve strategy of always selecting the less loaded of the two bins. Slight variations of this model, where the maximum load could be improved by a constant factor have also been considered, see [2, 12].

The two-thinning setting, considered in this paper, is intermediate between two-choices and no-choice, as it is equivalent to a two-choices setting in which the overseer is oblivious of the location of one of the two available bins. The name “two-thinning” is due to yet another point of view on this setting. According to this view, an infinite sequence of allocations has been drawn independently and uniformly at random, and the overseer is allowed to thin it on-line (i.e., delete some of the allocations depending only on the past), as long as at most one of every two consecutive entries is deleted (for a more thorough discussion of the model see joint work with Ramdas and Dwivedi [6], where the model was introduced).

From Theorem 1.1 we see that the optimal maximum load under two-thinning is indeed intermediate between the maximum load without thinning and the maximum load in the two-choices setting.

Threshold strategy in other settings. The threshold strategy has been considered in additional settings. In [1], Adler, Chakrabarti, Mitzenmacher and Rasmussen applied this strategy to a model of parallel allocation with limited communications in the k-choices model. Their upper bound is $O\left((\log n/\log\log n)^{1/(k+1)}\right)$ – of the same order of magnitude as in Theorem 1.1. We remark that it appears that neither model is stronger than the other.

More choice. Already in [3], Azar et al. showed that allowing the overseer choice between $k > 2$ choices, reduces the asymptotic maximal load by a factor of $\log(k)$. The expected counterpart in the k-thinning setting, where we allow the overseer to iteratively reject up to k suggested allocations for each ball, is the following conjecture.

Conjecture 1.2. *In the k-thinning setting, the asymptotically optimal maximum load is*

$$\Omega\left(\left(\frac{\log n}{\log\log n}\right)^{1/(k+1)}\right).$$

Indeed, a matching upper bound has already been verified in [1].

More balls. Berenbrink, Czumaj, Steger and Vöcking [4] have considered the power of two choices in the heavily loaded case of the balls and bins model, that is, when $\omega(n)$ balls are allocated into n bins. They showed that in this case under the power of k-choices, the deviation of the maximum load from the average load is asymptotically almost surely $\log_k \log n + O(1)$ (see Talwar and Wieder [11], for a simpler proof). We wonder whether the same phenomenon will occur for two-thinning. Namely,

Problem 1.3. In the two-thinning setting, where $m = \Omega(n)$ balls are two-thinned, is the asymptotically optimal maximum load $\frac{m}{n} + \Theta\left(\sqrt{\frac{\log n}{\log \log n}}\right)$?

1+β-thinning. In his thesis [7], Mitzenmacher suggested considering a variant of the power of two-choices in which, for each allocation independently, there is some small probability that a decision opposite to that made by the overseer will be executed. This notion was recently formulated and studied by Peres, Talwar and Wieder [9], viewing it as having two-choices with probability β and no-choice with probability $(1-\beta)$, independently for every ball. Once errors of this nature are introduced to the model, two-choices and one-retry are equivalent up to a parameter change, and in lightly loaded case of $\lfloor \rho n \rfloor$ balls allocated into n bins, both offer no improvement over having no-choice at all (see [6] for more details).

2 Definitions & notation

A strategy f is a collection of functions $\{f_t\}_{t\in\mathbb{N}}$ where

$$f_t : [n]^t \times [n]^{t-1} \times \{0,1\}^{t-1} \to \{0,1\}$$

which, given the sequence of primary allocations up to time t, the sequence of final allocations up to time $t-1$, and the sequence of previous decisions of whether to retry or not (up to time $t-1$), decides whether to accept the primary allocation (indicated by 0) or reject it (indicated by 1).

Given a thinning strategy f, generate $\{D_t\}_{t\in\mathbb{N}}$, $\{Z_t\}_{t\in\mathbb{N}}$, the sequence of allocations, in the following way. Let $\{Z_t^0\}_{t\in\mathbb{N}}$ and $\{Z_t^1\}_{t\in\mathbb{N}}$ be two independent sequences of independent random variables uniformly distributed in $[n]$. Here Z_t^0 represents the primary allocation of the t-th ball, while $\{Z_t^1\}_{t\in\mathbb{N}}$ is used as a pool of secondary allocations. Denote by r_t the number of rejections among the first t primary allocations.

Set $r_0 = 0$ and, for the t-th allocation, inductively set

$$\begin{aligned} D_t &= f_t(\{Z_s^0\}_{s\in[t]}, \{Z_s\}_{s\in[t-1]}, \{D_s\}_{s\in[t-1]}), \\ r_t &= r_{t-1} + D_t, \\ Z_t &= \begin{cases} Z_t^0 & D_t = 0, \\ Z_{r_t}^1 & D_t = 1. \end{cases} \end{aligned}$$

In other words, we look at the history of the process up to time $(t-1)$ and at Z_t^0 and apply f to determine whether to accept or reject the primary allocation. If we reject it, we allocate the ball to the next unused secondary allocation from our pool, which is $\{Z_{r_t}^1\}$.

We introduce the following notation. For any $t \le \rho n$, $m \in [n]$ denote

$$\begin{aligned} L_t(m) &= |\{1 \le i \le t \ : \ Z_i = m\}|, \\ A_t(m) &= |\{1 \le i \le t \ : \ Z_i^0 = m\}|, \\ B_t(m) &= |\{1 \le i \le t \ : \ Z_i^1 = m\}|. \end{aligned}$$

Thus, $L_t(m)$ is the load of the m-th bin at time t, $A_t(m)$ describes how many times bin m is suggested as a primary allocation among the first t allocations and $B_t(m)$ describes how many of the first t secondary allocations are into bin m.

Finally, denote

$$\begin{aligned}\mathrm{MaxLoad}_t^f(S) &= \max_{m\in S} L_t(m)\\ \mathrm{MaxLoad}_t^f &= \mathrm{MaxLoad}_t^f([n]).\end{aligned}$$

3 Preliminaries

We take advantage of a comparison lemma of Mitzenmacher and Upfal, which we reproduce here, relating the balls-and-bins model with independent Poisson random variables. Denote by $\mathbb{N}_0$ the set of natural numbers together with 0. Given two vectors $x, y \in (\mathbb{N}_0)^n$ we write $x \le y$ if $x_i \le y_i$ for all $i \in [n]$. A set $S \subset (\mathbb{N}_0)^n$ is called *monotone decreasing (increasing)* if $x \in S$ implies $y \in S$ for all $y \le x$ $(y \ge x)$.

Lemma 3.1 (Mitzenmacher and Upfal [8, Corollary 5.11])**.** *Let $(X_m)_{m\in[n]}$ be the number of balls in the m-th bin when t balls are independently and uniformly allocated into n bins. Further let $(Y_m)_{m\in[n]}$ be independent Poisson$(\frac{t}{n})$ random variables, and let S be a monotone set (either increasing or decreasing). Then*

$$\mathbb{P}\Big((X_1,\dots,X_n)\in S\Big) \le 2\mathbb{P}\Big((Y_1,\dots,Y_n)\in S\Big).$$

We prove two useful corollaries of this lemma.

Lemma 3.2. *Let $(X_m)_{m\in[n]}$ be the number of balls in the m-th bin when $\lfloor\theta n\rfloor$-balls are independently and uniformly allocated into n-bins, for $\theta\in[0,1]$. Then, for any $a\in[\theta n]$ and $S\subset[n]$ we have*

$$\mathbb{P}\left(\max_{m\in S}(X_m) < a\right) \le 2\exp\left(-\frac{\theta^a|S|}{ea!}\right)$$

Proof. Let $\{Y_m\}_{m\in[n]}$ be i.i.d. Poisson(θ) random variables. By Lemma 3.1 we have

$$\begin{aligned}\mathbb{P}\left(\max_{m\in S}(X_m) < a\right) &\le 2\,\mathbb{P}\left(\max_{m\in S}(Y_m) < a\right) = 2\mathbb{P}\left(Y_1 < a\right)^{|S|} \le 2\left(1 - e^{-\theta}\frac{\theta^a}{a!}\right)^{|S|}\\ &\le 2\left(1-\frac{\theta^a}{ea!}\right)^{|S|} \le 2\exp\left(-\frac{\theta^a|S|}{ea!}\right)\ .\end{aligned}$$

□

Lemma 3.3. *Let $(X_m)_{m\in[n]}$ be the number of balls in the m-th bin when (θn)-balls are independently and uniformly allocated into n-bins, for $\theta\in[0,1]$. Then, for any $S\subset[n]$ we have*

$$\mathbb{P}\left(|\{m\in S : X_m > 0\}| \le \frac{\theta|S|}{2e}\right) \le 2\exp\left(-\frac{\theta^2|S|}{2e^2}\right)\ .$$

Proof. Let $\{Y_m\}_{m\in[n]}$ be i.i.d. Poisson(θ) random variables. By Lemma 3.1

$$\mathbb{P}\left(|\{m\in S : X_m > 0\}| \le \frac{\theta|S|}{2e}\right) \le 2\mathbb{P}\left(|\{m\in S : Y_m > 0| \le \frac{\theta|S|}{2e}\right).$$

We observe that $\mathbb{P}(Y_1 > 0) \ge \frac{\theta}{e}$. Using Chernoff-Okamoto bound for the tail of binomial distributions (see, e.g. [5, Equation 1.3.10]), we obtain,

$$\mathbb{P}\left(|\{m\in S : Y_m > 0\}| \le \frac{|S|\theta}{2e}\right) \le \exp\left(-2|S|\left(\frac{\theta}{e}-\frac{\theta}{2e}\right)^2\right) = \exp\left(-\frac{\theta^2|S|}{2e^2}\right).$$

□

4 Upper bound on $\mathrm{MaxLoad}^f_{\rho n}([n])$

For $n \geq 3$, denote $L = \left\lceil \sqrt{2\log n/\log\log n} \right\rceil$. Let f be the L-threshold strategy, i.e., that rejects the primary allocation of the t-th ball if and only if $A_{t-1}(Z^0_t) \geq L$. The main statement of this section is the following.

Proposition 4.1. *For any $n \geq n_0(\rho)$ sufficiently large and any $\eta > 0$ the strategy f satisfies*

$$\mathbb{P}\big(\mathrm{MaxLoad}^f_{\rho n} > (2+\eta)L\big) \leq 2n^{-\frac{\eta}{4}+\frac{2\log\log\log n}{\log\log n}} + 2e^{-\sqrt{n}}.$$

Let us begin by reducing the upper bound in Theorem 1.1 to this proposition.

Proof of the upper bound in Theorem 1.1. We apply Proposition 4.1 with $\eta = \frac{9\log\log\log n}{\log\log n}$. Observe that $\eta = o(1)$ and by the proposition we have

$$\mathbb{P}\big(\mathrm{MaxLoad}_{\rho n} > (2+\eta)L\big) \leq \exp\left(-\frac{\log n}{4\log\log n}\right) + 2e^{-\sqrt{n}} = o(1). \qquad \square$$

Proof of Proposition 4.1. Denote $r = r_{\lfloor \rho n \rfloor}$. Our strategy f guarantees that number of accepted primary allocations to bin m is at most L for all $m \in [n]$. Hence, under this strategy

$$\mathbb{P}\big(\mathrm{MaxLoad}_{\rho n} \geq (2+\eta)L\big) \leq \mathbb{P}\left(E_r\right) \ , \tag{4.1}$$

where E_k is the event that there is a bin m for which $B_k(m) \geq (1+\eta)L$. Notice that if E_r occurs then for any $0 \leq k \leq \rho n$, either $r > k$ or E_k occurs. Hence, for any $0 \leq k \leq \rho n$ we get

$$\mathbb{P}\left(E_r\right) \leq \mathbb{P}\left(r > k\right) + \mathbb{P}\left(E_k\right) \ . \tag{4.2}$$

We now bound the two probabilities on the right hand side.

To bound $\mathbb{P}(r > k)$, notice that under the L-threshold strategy, the number of rejections before time t is

$$r_t = \sum_{m\in[n]} \max(A_t(m) - L, 0).$$

This is clearly monotone in $A_t(m)$ so we can apply Lemma 3.1. Letting $\{Y_m\}_{m\in[n]}$ be i.i.d. Poisson(ρ) random variables and writing

$$Y := \sum_{m\in[n]} \max(Y_m - L, 0),$$

we have

$$\mathbb{P}(r > k) \leq 2\mathbb{P}\Big(Y > k\Big). \tag{4.3}$$

For a single Poisson(ρ) random variable and for n large enough so that $L > 2\rho e$, we have

$$\mathbb{E}\Big(e^{\max(Y_1 - L, 0)}\Big) \leq 1 + e^{-\rho}\sum_{\ell=1}^{\infty}\frac{\rho^{L+\ell}e^{\ell}}{(L+\ell)!} \leq 1 + \frac{\rho^L}{L!} < \exp\left(\frac{\rho^L}{L!}\right) \ .$$

Hence, by Markov's inequality, for $k \geq \frac{2n\rho^L}{L!}$ we have, for n sufficiently large,

$$\mathbb{P}\Big(Y > k\Big) = \mathbb{P}\Big(e^Y > e^k\Big) \leq \exp\left(\frac{n\rho^L}{L!} - k\right) < \exp\left(-\frac{n}{L!}\right) < e^{-\sqrt{n}}, \tag{4.4}$$

where the last inequality follows from the definition L and the fact that for sufficiently large n we have $L! < L^L < \sqrt{n}$.

Putting together (4.3) and (4.4) we obtain

$$\mathbb{P}(r > k) < 2e^{-\sqrt{n}}. \tag{4.5}$$

Next we bound $\mathbb{P}(E_k)$. Let $\{Y_m\}_{m\in[n]}$ be i.i.d. Poisson(k/n) random variables. By Lemma 3.1 we have,

$$\mathbb{P}(E_k) \le 2\mathbb{P}\Big(\max_{m\in[n]}(Y_m) > (1+\eta)L\Big).$$

For $k \le \frac{3n\rho^L}{L!}$ and n sufficiently large we have

$$\mathbb{P}\Big(Y_1 > (1+\eta)L\Big) \le e^{-k/n} \sum_{\ell=\lceil(1+\eta)L\rceil}^{\infty} \frac{(k/n)^\ell}{\ell!} \le \frac{1}{2} \sum_{\ell=\lceil(1+\eta)L\rceil}^{\infty} \left(\frac{3\rho^L}{L!}\right)^\ell \le \left(\frac{3\rho^L}{L!}\right)^{\lceil(1+\eta)L\rceil}.$$

Taking a union bound, we obtain

$$\mathbb{P}(E_k) \le 2n \left(\frac{3\rho^L}{L!}\right)^{\lceil(1+\eta)L\rceil}. \tag{4.6}$$

By Stirling's approximation for all $L > 1$ we have $L! > 3\left(\frac{L}{e}\right)^L$. Hence,

$$\begin{aligned}
\mathbb{P}(E_k) &\le 2n \left(\frac{3\rho^L}{L!}\right)^{\lceil(1+\eta)L\rceil} \le 2n \left(\frac{L}{e\rho}\right)^{-(1+\eta)L^2} \\
&= 2\exp\Big(\log n - (1+\eta)L^2\big(\log L - 1 - \log\rho\big)\Big) \\
&\le 2\exp\left(\log n - (1+\eta)\frac{2\log n}{\log\log n}\left(\frac{1}{2}\log\log n - \frac{1}{2}\log\log\log n - 1 - \log\rho\right)\right) \\
&= 2\exp\left(-\eta\log n + (1+\eta)\frac{2\log n}{\log\log n}\left(\frac{1}{2}\log\log\log n + 1 + \log\rho\right)\right) \\
&\le 2\exp\left(-\eta\log n + (1+\eta)\frac{2\log n\log\log\log n}{\log\log n}\right) \\
&\le 2n^{-\frac{\eta}{4}+\frac{2\log\log\log n}{\log\log n}},
\end{aligned} \tag{4.7}$$

for any $n \ge n_0(\rho)$ sufficiently large. Putting (4.5) and (4.7) into (4.2), the proposition follows. □

5 Lower bound on $\mathrm{MaxLoad}^g_{\rho n}([n])$ for any strategy g

Let $\ell = \ell(n) = \left\lfloor \sqrt{2\log n/\log\log n}\right\rfloor$. In this section we prove the following proposition, from which the optimality of the threshold strategy in Theorem 1.1 is an immediate corollary.

Proposition 5.1. *Let $\varepsilon, \rho > 0$ and n sufficiently large (depending on ρ and ε). For any strategy g we have*

$$\mathbb{P}\big(\mathrm{MaxLoad}^g_{\rho n} < (2-\varepsilon)\ell\big) \le \exp\left(-n^{\varepsilon/5}\right).$$

To prove Proposition 5.1 we use the following lemma, whose contra-positive, roughly speaking, says the following. For any "large" set of bins S, under any strategy, after allocating $\rho n/2\ell$ balls, with relatively high probability the allocations will either include many bins in S (the event E^c) or quite a few of them will fall in the same bin in S (the event F^c). By iterating this lemma 2ℓ times we'll be able to guarantee that, with high probability, either one of these iterations caused high load, or they piled up to cause it.

Lemma 5.2. *Let $\varepsilon, \rho > 0$ and n sufficiently large (depending on ρ and ε) and denote $\zeta = \rho/8e\ell$. For any $1 \le k \le 2\ell$, $t > \rho n/2\ell$ and $S \subset [n]$ such that $|S| \ge n\zeta^k$ and any strategy g, we have*

$$\mathbb{P}\big(E \cap F\big) \le \exp(-n^{\varepsilon/4}),$$

where

$$\begin{aligned}E &= \{|\{m \in S \ : \ L_t(m) \geq 1\}| < n\zeta^{k+1}\},\\ F &= \{\mathrm{MaxLoad}_t^g(S) < (2-\varepsilon)\ell - k\}.\end{aligned}$$

Proof. Write $T = n\zeta^{k+1}$ and denote

$$\begin{aligned}E' &= \{|\{m \in S \ : \ A_t(m) \geq 1\}| < 2T\},\\ F' &= \{\forall_{m\in S} \ : \ B_T(m) < (2-\varepsilon)\ell - k\},\end{aligned}$$

so that E' is the event that less than $2T$ bins in S were suggested as primary allocations until time T, while F' is the event that no bin in S would received $(2-\varepsilon)\ell - k$ secondary allocations among the first T secondary allocations.

By applying Lemma 3.3 with $\theta = \frac{\rho}{2\ell}$ and observing that $2T \leq 2\zeta|S| \leq \frac{\theta|S|}{2e}$, we obtain

$$\mathbb{P}\left(E'\right) \leq 2\exp\left(-\frac{\theta^2|S|}{2e^2}\right) \leq 2\exp\left(-\frac{8n\rho^{k+2}}{(8e\ell)^{k+2}}\right) \leq 2\exp\Big(-n^{1+o(1)}\Big),$$

Where in the rightmost inequality we used the fact that $k < 2\ell$. By applying Lemma 3.2 with $a = (2-\varepsilon)\ell - k$ and $\theta = \zeta^{k+1}$,

$$\mathbb{P}(F') \leq 2\exp\left(-\frac{\zeta^{(k+1)a}|S|}{ea!}\right) \leq 2\exp\left(-\frac{\zeta^{(k+1)a}\zeta^k n}{ea^a}\right) \leq 2\exp\left(-\frac{\zeta^{(k+1)(a+1)}n}{ea^a}\right).$$

Letting n be large enough, and observing that for such n we have $(a+1)(k+1) \leq (1-\varepsilon/2)\ell^2$ we obtain

$$\mathbb{P}(F') \leq 2\exp\left(-\frac{(\rho/8e\ell)^{(1-\varepsilon/2)\ell^2}n}{e(2\ell)^{2\ell}}\right) \leq 2\exp\left(-\frac{n}{\ell^{(1-\varepsilon/3)\ell^2}}\right) \leq 2\exp(-n^{\varepsilon/3}),$$

where the two rightmost inequalities use the fact that $\ell^{\ell^2} \geq n$, while c^{ℓ^2} and ℓ^ℓ are both sub-polynomial in n for any $c > 0$.

We claim that $E'^c \cap F'^c \subseteq E^c \cup F^c$. Indeed, notice that if the number of retries is at most T and there are at least $2T$ bins in S which were chosen as the primary allocation, then there are at least T nonempty bins in S. Thus, $\{r_t \leq T\} \cap E'^c \subset E^c$. Also, if the number of retries is more then T, and the first T secondary allocations cause a load of at least $(2-\varepsilon)\ell - k$ in some bin in S, then this bin has a load of at least $(2-\varepsilon)\ell - k$. Thus, $\{r_t > T\} \cap F'^c \subset F^c$. Hence, $E \cap F \subset E' \cup F'$. From our bounds on $\mathbb{P}(E')$ and $\mathbb{P}(F')$ the proposition follows. □

Proof of Proposition 5.1. Fix $\varepsilon, \rho > 0$ and let g be a thinning strategy. We divide our process into $s = \lceil(2-\varepsilon)\ell\rceil$ stages each consisting of the allocation of $w = \lceil\frac{\rho n}{2\ell}\rceil$ balls so that the k-th stage process consists of $Z_{(k-1)w+1}, \dots, Z_{kw}$. These are followed by a final stage in which the remaining balls are allocated.

Denote $S_k = \{m \in [n] \ : \ A_{kw}(m) \geq k\}$. For $\zeta = \rho/8e\ell$, we define $E_k = \{|S_k| < n\zeta^k\}$ and $L_k = \{\mathrm{MaxLoad}_{kw}^g < (2-\varepsilon)\ell\}$.

By applying Lemma 5.2 to the k-th stage process with $S = S_k$ we obtain that

$$\mathbb{P}(E_{k+1} \cap L_{k+1} \mid E_k^c) \leq \exp(-n^{\varepsilon/4}).$$

The see this, observe that the size of S_{k+1} is at least the number of bins in S_k which were allocated at least one ball in the k-th stage process and that $\mathrm{MaxLoad}_{(k+1)w}^g$ is at least k plus the maximum number of balls that were allocated in the k-th stage process to a single bin in S_k.

Observe that $L_{k+1} \subseteq L_k$ we use the law of total probability to obtain

$$\mathbb{P}(E_{k+1} \cap L_{k+1}) = \mathbb{P}(E_{k+1} \cap L_{k+1} \cap E_k) + \mathbb{P}(E_{k+1} \cap L_{k+1} \cap E_k^c) \leq \mathbb{P}(E_k \cap L_k) + \mathbb{P}(E_{k+1} \cap L_{k+1} \mid E_k^c)$$

Since $E_0 \cap L_0 = \emptyset$, we may use induction to deduce that for sufficiently large n we have,

$$\mathbb{P}\left(E_s \cap L_s\right) \leq \sum_{k=1}^{s} \mathbb{P}(E_k \cap L_k \mid E_{k-1}^c) \leq s \exp(-n^{\varepsilon/4}) \leq \exp(-n^{\varepsilon/5}).$$

Since $\{\mathrm{MaxLoad}_{\rho n}^{g} < (2-\varepsilon)\ell\} \subset E_s \cap L_s$, this concludes the proof. □

References

[1] M. Adler, S. Chakrabarti, M. Mitzenmacher and L. Rasmussen, *Parallel Randomized Load Balancing.* Random Structures & Algorithms 13, no. 2 (1998), pp. 159–188. MR-1642570

[2] J. Augustine, W. K. Moses Jr, A. Redlich and E. Upfal, *Balanced allocation: patience is not a virtue.* Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics, 2016, pp. 655–671. MR-3478424

[3] Y. Azar, A. Broder, A. Karlin and E. Upfal, *Balanced allocations.* SIAM Journal of Computing 29, no. 1 (1999), pp. 180–200. MR-1710347

[4] P. Berenbrink, A. Czumaj, A. Steger and B. Vöcking, *Balanced allocations: The heavily loaded case.* SIAM Journal on Computing 35, no. 6 (2006), pp. 1350–1385. MR-2217150

[5] R. M. Dudley, *Uniform central limit theorems.* Cambridge university press, 1999. MR-1720712

[6] R. Dwivedi, O. N. Feldheim, O. Gurel-Gurevich, and A. Ramdas, *The power of online thinning in reducing discrepancy.* Probability Theory and Related Fields 174, no. 1-2, pp. 103–131. MR-3947321

[7] M. Mitzenmacher, *The Power of Two Choices in Randomized Load Balancing.* PhD thesis, University of California, Berkeley, CA, 1996. MR-2695522

[8] M. Mitzenmacher and E. Upfal, *Probability and computing: Randomized algorithms and probabilistic analysis.* Cambridge University Press, 2005. MR-2144605

[9] Y. Peres, K. Talwar and U. Wieder, *Graphical Balanced Allocations and the $(1+\alpha)$-Choices Process.* Random Structures & Algorithms 47, no. 4 (2015), pp. 760–775, 157–163. MR-3418914

[10] A. W. Richa, M. Mitzenmacher and R. Sitarman, *The power of two random choices: A survey of techniques and results.* Combinatorial Optimization 9 (2001), pp. 255–304. MR-1966907

[11] K. Talwar and U. Wieder, *Balanced allocations: A simple proof for the heavily loaded case.* International Colloquium on Automata, Languages, and Programming. Springer Berlin Heidelberg, 2014. MR-3238686

[12] B. Vöcking, *How asymmetry helps load balancing.* Journal of the ACM (JACM) 50, no. 4 (2003), pp. 568–589. MR-2146887