Long-term balanced allocation via thinning

Ohad N. Feldheim, Ori Gurel-Gurevich, Jiange Li *

October 13, 2021

Abstract

We study the long-term behavior of the two-thinning variant of the classical balls-and-bins model. In this model, an overseer is provided with uniform random allocation of m balls into n bins in an on-line fashion. For each ball, the overseer could reject its allocation and place the ball into a new bin drawn independently at random. The purpose of the overseer is to reduce the *maximum load* of the bins, which is defined as the difference between the maximum number of balls in a single bin and m/n, i.e., the average number of balls among all bins.

We provide tight estimates for three quantities: the lowest maximum load that could be achieved at time m, the lowest maximum load that could be achieved uniformly over the entire time interval $[m] := \{1, 2, \cdots, m\}$, and the lowest *typical* maximum load that could be achieved over the interval $[m]$, where the typicality means that the maximum load holds for $1 - o(1)$ portion of the times in $[m]$.

We show that when m and n are sufficiently large, a typical maximum load of $(\log n)^{1/2+o(1)}$ can be achieved with high probability, asymptotically the same as the optimal maximum load that could be achieved at time m. However, for any strategy, the maximal load among all times in the interval $[m]$ is $\Omega\big(\frac{\log n}{\log\log n}\big)$ with high probability. A strategy achieving this bound is provided.

An explanation for this gap is provided by our optimal strategies as follows. To control the typical load, we restrain the maximum load for some time, during which we accumulate more and more bins with relatively high load. After a while, we have to employ for a short time a different strategy to reduce the number of relatively heavily loaded bins, at the expanse of temporarily inducing high load in a few bins.

Keywords: balls-and-bins, load balancing, two-choice, two-thinning.

1 Introduction

In the classical *balls-and-bins* model, m balls are independently and uniformly at random placed into n bins one after another. In this paper, we are interested in the following variant, which is called the *two-thinning* model. For each ball, after a uniformly random bin, which is called the *primary allocation*, has been suggested, an overseer has the choice of either accepting this bin, or placing the ball into a new bin selected independently and uniformly at random, which is called the *secondary allocation*. In this model, the overseer is oblivious to the secondary allocation before deciding whether to accept the primary allocation. In contrast, in the well-known *two-choice* model, which was introduced in the seminal work

*O. F. and O. G-G are with the Hebrew University of Jerusalem. J. L. is with the Harbin Institute of Technology. E-mail: ohad.feldheim@mail.huji.ac.il, ori.gurel-gurevich@mail.huji.ac.il, lijiange7@gmail.com.

[2], the overseer is aware of the secondary allocation and places the ball into the bin which contains fewer balls (break ties arbitrarily).

1.1 Main results

We define the *load* of a bin as the difference between the number of balls in this bin and the average number of balls among all bins. Given a two-thinning strategy f (see Section 2.1 for a formal definition), we denote by $\mathrm{MaxLoad}^f(m)$ the *single-time maximum load*, which is defined as the maximum load among all bins after allocating m balls using the strategy f, and denote by $\mathrm{MaxLoad}^f([m])$ the *all-time maximum load*, which is the maximum of $\mathrm{MaxLoad}^f(k)$ for all $k \in [m] := \{1, 2, \dots, m\}$. In general, we can replace $[m]$ by a subset $S \subseteq [m]$, and define $\mathrm{MaxLoad}^f(S)$ in a similar manner.

Theorem 1. *For all $m, n \in \mathbb{N}$, there exists an explicit two-thinning strategy $f := f_{m,n}$ such that, with high probability,*

$$\mathrm{MaxLoad}^f(m) = \begin{cases} \Theta\left(\sqrt{\frac{\log n}{\log\log n - 2\log(m/n)}}\right) & \Omega(n) \le m \le o(n\sqrt{\log n}), \\ \Theta(\sqrt{\log n}) & m = \Theta(n\sqrt{\log n}), \\ (\log n)^{1/2+o(1)} & m = \omega(n\sqrt{\log n}). \end{cases}$$

Moreover, in the first two cases the maximum loads are optimal up to some multiplicative constants, while in the third case we have a lower bound of $\Omega(\sqrt{\log n})$ for all two-thinning strategies.

Theorem 2. *There exists an explicit two-thinning strategy f such that, with high probability,*

$$\mathrm{MaxLoad}^f([m]) = \begin{cases} \Theta\left(\sqrt{\frac{\log n}{\log\log n - 2\log(m/n)}}\right) & \Omega(n) \le m \le O(n\sqrt{\log n}), \\ \Theta\big(\big(\frac{m\log n}{n}\big)^{1/3}\big) & \Omega(n\sqrt{\log n}) \le m \le o(n\log^2 n), \\ \Theta\big(\frac{\log n}{\log\log n}\big) & \Omega(n\log^2 n) \le m \le n^{O(1)}. \end{cases}$$

Moreover, the all-time maximum load achieved by f is optimal up to a multiplicative constant.

For $\varepsilon > 0$, we denote by $\mathrm{MaxLoad}^f_\varepsilon([m])$ the *ε-typical maximum load*, which is defined as the largest $\ell > 0$ such that $\mathrm{MaxLoad}^f(k) > \ell$ holds for at least εm many $k \in [m]$. Clearly, we have $\mathrm{MaxLoad}^f(m) \le \mathrm{MaxLoad}^f_\varepsilon([m]) \le \mathrm{MaxLoad}^f([m])$.

Theorems 1 and 2 show that for $m = O(n\sqrt{\log n})$, the difference between the optimal single-time and all-time maximum loads is at most a multiplicative constant and hence the optimal typical maximum load also has the same asymptotic behaviour. For $m = \omega(n\sqrt{\log n})$, however, there is a gap between the optimal single-time and all-time maximum loads. The next theorem shows that in this regime, the typical maximum load behaves like the single time maximum load, so is the gap between the optimal typical and all-time maximum loads.

Theorem 3. *Let $m, n \in \mathbb{N}$, and write $\varepsilon = e^{-\frac{1}{2}\sqrt{\log\log\log n}}$. There exists an explicit two-thinning strategy $f = f_{m,n}$ such that for n large enough and for all m,*

$$\mathrm{MaxLoad}^f_\varepsilon([m]) \le (\log n)^{\frac{1}{2}+o(1)},$$

holds with high probability.

It is worth pointing out that for $m \le n^{O(1)}$ our strategy actually governs the loads in some **predetermined**, large (i.e., $1-\varepsilon$ portion) set of times in $[m]$, with high probability (see Proposition 8.1).

1.2 Discussion

The classical balls-and-bins model and its two-choice variant have been extensively studied in probability theory, random graph theory, and computer science. Many applications have been found in various areas, such as hashing, load balancing and resource allocation in parallel and distributed systems (see e.g., [2], [3], [14], [22], [23]). In the balls-and-bins model, it is known that for $m = \Theta(n)$, the maximum load is $(1 + o(1))\frac{\log n}{\log \log n}$ with high probability, and for $m \gg n$, the maximum load is $\Theta\Big(\sqrt{\frac{m \log n}{n}}\Big)$ with high probability (see e.g. [21]). In the seminal paper [2], Azar, Broder, Karlin and Upfal showed that in the two-choice model, for $m = \Theta(n)$, the maximum load is $\frac{\log \log n}{\log 2} + O(1)$ with high probability – an exponential improvement over the balls-and-bins model. In fact, this phenomenon was first noticed by Karp, Luby and Meyer auf der Heide [14] in the context of PRAM simulations when switching from one hash function to two. In [2], the d-choice setting, where the overseer is given $d > 2$ choices, was also considered. In this setting, an optimal maximum load of $\frac{\log \log n}{\log d} + O(1)$ can be achieved with high probability; that is, compared with the case $d = 2$, the performance improves by merely a multiplicative factor for larger values of d. We refer the reader to the survey [18] for more details about the two-choice model.

The long-term behavior of the two-choice model, in which case the number of balls m can be super linear in n, proved to be more challenging. In the seminal paper [5], Berenbrink, Czumaj, Steger and Vöcking showed that for arbitrarily large m, one can achieve the maximum load of $\frac{\log \log n}{\log 2} + O(1)$ with high probability. A simpler proof of this result with a weaker lower order term was given by Talwar and Wieder [25]. Since this result is achieved via a single greedy strategy at all times, a simple union bound argument implies that this strategy also maintains this bound as the all-time and the typical maximum loads for m polynomially large in n.

Different variants of the two-choice model have been studied under weaker constraints from practical considerations. These include load balancing with limited memory [1, 6, 17], relaxation on the possible pairs the overseer may select from (known as two choices on graphs) [15, 20] and a hypergraph variant of it [13]. Other relaxations include bins with different selection probabilities [4] and balls with different weights [24]. An important purpose of this course of study is to understand the robustness of the load reduction achieved by the power of two choices, understanding the impact of constraints on memory, information and choice patterns. Particularly, Peres, Talwar and Wieder [20] studied the setting of two choices with errors, which is known as the $(1+\beta)$-choice model. In this setting, with probability β the ball is allocated using the two-choice model, and with probability $1 - \beta$ the ball is assigned to a random bin as in the balls-and-bins model. The authors showed that, irrespective of m, the gap between the maximum load and the average is $O\big(\frac{\log n}{\beta}\big)$. Since this result is irrespective of m, a simple union bound argument implies that this bound is also valid for the all-time and the typical maximum loads for m polynomially large in n.

The *two-thinning* variant is a different relaxation of the two-choice model which arises naturally in a statistical scenario, where one collects samples one-by-one and is allowed to decide whether to keep each sample or not, under the constraint of never discarding two consecutive samples. In [7], Dwivedi, Ramdas and the first two authors showed that two-thinning could reduce the discrepancy of a sequence of random points selected independently and uniformly at random from the interval $[0, 1]$ to be near optimal. The first two authors studied the two-thinning variant of the balls-and-bins model in [11]. They showed that for $m = \Theta(n)$, the optimal maximum load is $(2 + o(1))\sqrt{\frac{2 \log n}{\log \log n}}$ with high probability, a

polynomial improvement over the balls-and-bins model. Hence, this model is in some sense more powerful than the $(1+\beta)$-choice model. The authors also conjectured the upper bound $O\Big(\sqrt{\frac{\log n}{\log\log n}}\Big)$ for all $m \gg n$. Los and Sauerwald [16] recently disproved this conjecture by showing a lower bound of $\Omega(\sqrt{\log n})$ for $m = \Theta(n\sqrt{\log n})$, a bound which we show here holds for all $m = \Omega(n\sqrt{\log n})$. They also showed that a load of $\Omega(\log n/\log\log n)$ holds for at least $\Omega(n\log n/\log\log n)$ times in $[1, n\log^2 n]$. Our work sheds more light on this phenomenon. The results in [11] were extended by the first and third authors [12] to the d-thinning setting and the optimal maximum load of $(d+o(1))\big(\frac{d\log n}{\log\log n}\big)^{1/d}$ could be achieved with high probability.

Another relaxation of the two-choice model was recently studied by Los and Sauerwald [16]. They considered the situation that each ball is offered two random bins and is allowed to send up to k binary queries, each to one of the two bins. In one model, it inquires whether the absolute load crosses some threshold, and in the other model, it inquires whether the number of bins with loads higher than that of the queried bin is greater than some percentile. The $k = 1$ case is equivalent to our two-thinning model. They showed that in both models a maximum load of $O(k(\log n)^{1/k})$ can be achieved with high probability.

Here, we study the long-term behavior of the two-thinning model. Our discussions above and Theorems 1 and 3 show that, in the balls-and-bins and the two-choice models, the optimal single-time and the typical maximum loads are asymptotically nearly identical. However, in contrast with these two models, there is a big gap between the optimal typical and the all-time maximum loads in the two-thinning setting. We attribute this difference to the fact that in the two-thinning setting, short periods of relative high maximum loads are necessary for the process to "release steam" with the benefit of arriving at low maximum loads at the end of these periods. A comparison of the maximum loads in these three models is given in the following table.

	$m = \Theta(n\log^\alpha n)$	$\mathrm{MaxLoad}^f(m)$	$\mathrm{MaxLoad}^f_\varepsilon([m])$	$\mathrm{MaxLoad}^f([m])$
Balls-and-bins	$\alpha < 1$	$\Theta\big(\frac{\log n}{\log\log n}\big)$	$\Theta\big(\frac{\log n}{\log\log n}\big)$	$\Theta\big(\frac{\log n}{\log\log n}\big)$
	$\alpha \geq 1$	$\Theta\Big(\sqrt{\frac{m\log n}{n}}\Big)$	$\Theta\Big(\sqrt{\frac{m\log n}{n}}\Big)$	$\Theta\Big(\sqrt{\frac{m\log n}{n}}\Big)$
Two-thinning	$\alpha \in [0, \frac{1}{2})$	$\Theta\left(\sqrt{\frac{\log n}{\log\log n}}\right)$	$\Theta\left(\sqrt{\frac{\log n}{\log\log n}}\right)$	$\Theta\left(\sqrt{\frac{\log n}{\log\log n}}\right)$
	$\alpha = \frac{1}{2}$	$\Theta(\sqrt{\log n})$	$\Theta(\sqrt{\log n})$	$\Theta(\sqrt{\log n})$
	$\alpha \in (\frac{1}{2}, 2)$	$(\log n)^{1/2+o(1)}$	$(\log n)^{1/2+o(1)}$	$\Theta\big((\log n)^{\frac{1+\alpha}{3}}\big)$
	$\alpha \geq 2$	$(\log n)^{1/2+o(1)}$	$(\log n)^{1/2+o(1)}$	$\Theta\big(\frac{\log n}{\log\log n}\big)$
Two-choice	$\alpha \geq 0$	$\Theta(\log\log n)$	$\Theta(\log\log n)$	$\Theta(\log\log n)$

Table 1: A comparison of the single-time, all-time and typical maximum loads. Here, we write $m = \Theta(n\log^\alpha n)$ and select $\varepsilon = o(1)$, where α is allowed to depend on n, but some of the results require m to be at most polynomially large in n. In the two-thinning model, the results for $\alpha = 0$ appear in [11] and the lower bound for $\alpha = 1/2$ appears in [16] and the remaining results are new. In the two-choice model, the results follow from [5]. In the balls-and-bins model, the results are classical (see e.g. [21]).

1.3 Upper bound strategies and lower bound techniques

In the following, we give a brief description of our strategies that achieve the upper bounds in our main results as well as techniques for establishing the lower bounds. We write $m = nt$ for $t \in \mathbb{N}$. Different strategies are required for values of t in different ranges.

The single-time maximum load. For $t = O(\sqrt{\log n})$, our upper bound is achieved by the *threshold strategy* employed in [11], which retries a ball if the number of primary allocations accepted by the suggested bin reaches certain threshold. For $t \geq \omega(\sqrt{\log n})$, the threshold strategy alone is not sufficient since the optimal choice of the threshold would be $t + \Theta((t \log n)^{1/3})$ and this yields a maximum load of $O((t \log n)^{1/3})$, which is much larger than our desired upper bound $(\log n)^{1/2+o(1)}$. Instead, we divide the process into multiple shorter stages and, in each stage, apply the threshold strategy with a smaller threshold. It is likely that this will cause more retries and even a temporarily high maximum load. To prevent this from causing a high load at the end of the process, we always retry a ball if its primary allocation is a heavily loaded bin. The number of retries caused by this requirement is relatively small since the number of heavily loaded bins is small. This, together with a careful selection of time lengths of the stages, enables us to achieve the maximum load of $(\log n)^{1/2+o(1)}$ at the end of the process. We call this strategy the *multi-stage threshold strategy*. For $t \geq \omega(\log n)$, we need another ingredient in the form of a *drift strategy*. Under this strategy we retry a ball with positive probability if its primary allocation has a positive load, and surely if its load is very high. This creates a drift in the load of positively loaded bins towards zero, resulting in a load distribution with exponential tail and a maximum load of $\Theta(\frac{\log n}{\log \log n})$ (in some sense, this is an improvement of a similar strategy in [7]). For $t \geq \omega(\log n)$, we first apply this drift strategy up to $\Theta(\log n)$ time before the end, and then apply the aforementioned multi-stage threshold strategy to allocate the remaining $\Theta(n \log n)$ balls. Our lower bound follows from the simple observation that if we retry too many balls, the secondary allocations will cause a high maximum load, and if we retry too few balls, the primary allocations will cause a high maximum load.

The all-time maximum load. Our upper bound strategy is a time-adaptive version of the threshold strategy for the single-time maximum load, which we call a *relative threshold strategy*. We use a threshold strategy where the threshold after throwing tn balls, is $t + \ell$ for a fixed $\ell > 0$. This strategy results in a uniform control of the maximum load throughout the process. Our lower bound follows from the observation that a uniform bound on the maximum load in the process upper-bounds the number of retries in the allocation of each batch of n balls, and hence – the total number of retries in the entire process. Subject to this constraint, we consider the maximum load after all balls have been allocated and show it to be large.

The typical maximum load. As mentioned before Theorem 3, it suffices to consider the case $t \geq \omega(\sqrt{\log n})$. For $\omega(\sqrt{\log n}) \leq t \leq O(\log n)$, we apply a multi-scale strategy. Each scale consists of iterations of two strategies. In the first, longer part of each iteration, we apply the strategy of a smaller scale, while in the second, shorter part, we use a different regulating strategy. The strategy in the smallest scale is simply the relative threshold strategy, while the regulating strategy is the multi-stage threshold strategy used to control the single-time maximum load. These regulating segments play the role of "releasing steam" from the process – although they result in a high maximum load for a short period of time, they yield good control the maximum loads at the end of these segments, so that we can re-initiate the next iteration. For $t \geq \omega(\log n)$, we iterate over long segments of this strategy, separated by short segments of the drift strategy followed by the multi-stage threshold strategy.

1.4 Outline

This paper is organized as follows. In the next section, we introduce two-thinning strategies that are used to achieve the desired bounds on three types of maximum loads as stated in Theorems 1, 2 and 3. We provide some preliminary tools in Section 3, which are used in the analysis of different two-thinning strategies and the proofs of the main results. The proof of Theorem 1 on the single-time maximum load is provided in Section 4 (upper bound) and Section 5 (lower bound). The proof of Theorem 2 on the all-time maximum load is provided in Section 6 (upper bound) and 7 (lower bound). In the last section, we prove Theorem 3 on the ε-typical maximum load.

2 Strategies for taming the maximum loads

In this section, we provide two-thinning strategies that are used to control the maximum loads. We give the formal definition of a two-thinning strategy in Section 2.1, and provide an alternative, indirect way of describing a two-thinning strategy in Section 2.2. Several basic two-thinning strategies are given in Section 2.3, which are building blocks of more advanced strategies in Section 2.4. We provide an outline of how these strategies are used to obtain the main theorems in Section 2.5.

2.1 Two-thinning strategy

A *decision strategy* is a function

$$f : [n] \times [0,1] \to \{1,2\},$$

which, given a primary allocation and an external random number in $[0,1]$, decides whether to accept (denoted by 1) or reject (denoted by 2) the suggested allocation. Given Z_1, Z_2, a pair of independent random variables, uniform on $[n]$ and U uniform on $[0,1]$, we can consider the *output* of a decision strategy given by $Z_{f(Z_1,U)}$.

A *thinning strategy* f is a sequence of functions $\{f_k\}_{k\in\mathbb{N}}$, where the function

$$f_k : ([n] \times [n] \times \{1,2\})^{k-1} \times [n] \times [0,1] \to \{1,2\},$$

given the history $\mathcal{H} \in ([n] \times [n] \times \{1,2\})^{k-1}$ of the process up to time $k-1$ (that is, the primary allocations, the final allocations and the decisions of the first $k-1$ balls), the primary allocation at time k and an external random number in $[0,1]$, decides whether to accept or reject the suggested allocation. Hence, given the history of the process, the thinning strategy provides a decision strategy for the next allocation.

A thinning strategy f generates the *decisions* sequence $\{D_k\}_{k\in\mathbb{N}}$ and the *allocations* sequence $\{Z_k\}_{k\in\mathbb{N}}$ in the following way. We denote by $\{Z_k^1\}_{k\in\mathbb{N}}$ and $\{Z_k^2\}_{k\in\mathbb{N}}$ two independent sequences of random variables, which are independent and uniformly distributed in $[n]$. Here, Z_k^1 represents the primary allocation of the k-th ball, while $\{Z_k^2\}_{k\in\mathbb{N}}$ is used as a pool of secondary allocations. Set $R_0 = 0$ and we denote by R_k the number of rejections among the first k primary allocations. Let $\{U_k\}_{k\in\mathbb{N}}$ be a collection of uniform random variables on $[0,1]$.

For the k-th allocation, we can inductively define

$$\begin{aligned} D_k &= f_k\left(\{Z_j^1\}_{j\in[k-1]}, \{Z_j\}_{j\in[k-1]}, \{D_j\}_{j\in[k-1]}, Z_k^1, U_k\right), \\ R_k &= R_{k-1} + D_k - 1, \\ Z_k &= \begin{cases} Z_k^1 & \text{if } D_k = 1, \\ Z_{R_k}^2 & \text{if } D_k = 2. \end{cases} \end{aligned} \tag{2.1}$$

In other words, we look at the history $\mathcal{H}$ of the process up to time $k-1$ and at the primary allocation Z_k^1 at time k along with an additional source of randomness U_k and apply f to determine whether to accept Z_k^1 or not. If we reject Z_k^1, we will then allocate the k-th ball to the next unused secondary allocation $Z_{R_k}^2$ from our pool.

We allow bins to start with some initial loads $\{L_i(0)\}_{i\in[n]}$ satisfying $\sum_{i=1}^n L_i(0) = 0$, where $L_i(0)$ is the initial load of the i-th bin. Let $m \in \mathbb{N}$ and let $i \in [n]$. The *load* of bin i after allocating m balls using the thinning strategy f is defined as

$$L_i^f(m) = L_i(0) + \sum_{k=1}^{m} \mathbb{1}_{\{Z_k = i\}} - \frac{m}{n}. \tag{2.2}$$

For any $M \subseteq [m]$, we define

$$\begin{aligned} L_{1,i}^f(M) &= \left|\left\{k \in M : Z_k^1 = i \text{ and } D_k = 1\right\}\right|, \\ L_{2,i}^f(M) &= \left|\left\{k \in M : Z_{R_k}^2 = i\right\}\right|. \end{aligned} \tag{2.3}$$

Hence, $L_{1,i}^f([m])$ represents the number of primary allocations accepted by bin i after allocating m balls, and $L_{2,i}([m])$ represents the number balls that bin i receive from secondary allocations. It is clear that $L_i^f(m) = L_i(0) + L_{1,i}^f([m]) + L_{2,i}^f([m]) - m/n$. For any $S \subseteq [n]$ and $\ell \in \mathbb{R}$, we define

$$\phi_S^\ell(m) = \left|\left\{i \in S : L_i^f(m) \geq \ell\right\}\right|, \tag{2.4}$$

which is the number of bins in S with loads at least ℓ after allocating m balls using the thinning strategy f, and

$$\psi_S^\ell(M) = \left|\left\{i \in S : \sum_{k\in M} \mathbb{1}_{\{Z_k^0 = i\}} \geq \ell\right\}\right|, \tag{2.5}$$

which is the number of bins in S that are suggested as primary allocations at least ℓ times during the allocations of balls in M. The maximum load over a set of bins S after allocating m balls using the thinning strategy f is defined as

$$\mathrm{MaxLoad}_S^f(m) = \max_{i\in S} L_i^f(m). \tag{2.6}$$

We will omit the index S in these notations when $S = [n]$. For any $M \subseteq [m]$, we define the maximum load achieved during the allocation of balls in M as

$$\mathrm{MaxLoad}^f(M) = \max_{k\in M} \mathrm{MaxLoad}^f(k). \tag{2.7}$$

The ε-typical maximum load $\mathrm{MaxLoad}_\varepsilon^f(M)$ over the set M is defined as

$$\mathrm{MaxLoad}_\varepsilon^f(M) = \max\left\{\ell > 0 : \left|\left\{k \in M : \mathrm{MaxLoad}^f(k) \geq \ell\right\}\right| \geq \varepsilon|M|\right\}. \tag{2.8}$$

2.2 A realizability criterion

Under certain circumstances, instead of providing an explicit, formal description of a two-thinning strategy, we only show the realizability. The following result provides a criterion for a probability distribution to be realized by some two-thinning strategy.

Lemma 2.1. *Any probability distribution $\mathcal{P}$ on $[n]$ with probability mass function $\{p_i\}_{i\in[n]}$ for which*

$$\frac{c}{n} \le p_i \le \frac{1+c}{n}$$

for some $c > 0$ and for every $i \in [n]$, is the distribution of the output of a two-thinning decision strategy.

Proof. Let Z_1, Z_2, U be independent random variables uniformly distributed in $[n]$. Here, U is the external randomness. We define the two-thinning function $f : [n] \times [0,1] \to \{1,2\}$ as

$$f(i,u) = \begin{cases} 1, & np_i - c \ge u, \\ 2, & np_i - c < u. \end{cases}$$

Let $Z = Z_{f(Z_1,U)}$ be the output of f. For any $i \in [n]$, we have

$$\begin{aligned} \mathbb{P}(Z=i) &= \mathbb{P}(Z_1 = i, f(Z_1,U) = 1) + \mathbb{P}(Z_2 = i, f(Z_1,U) = 2) \\ &= \mathbb{P}(Z_1 = i)\mathbb{P}(f(i,U) = 1) + \mathbb{P}(Z_2 = i)\sum_{j=1}^{n} \mathbb{P}(Z_1 = j)\mathbb{P}(f(j,U) = 2) \\ &= \frac{1}{n} \cdot (np_i - c) + \frac{1}{n} \cdot \frac{1}{n} \sum_{j=1}^{n} (1 + c - np_j) \\ &= p_i. \end{aligned}$$

The second identity follows from the joint independence among Z_1, Z_2, U. □

2.3 The basic strategies

Here, we introduce some basic two-thinning strategies, which are building blocks of more advanced strategies in the next section. The first two thinning strategies are deterministic and rather natural.

The threshold strategy. The *ℓ-threshold strategy* accepts the primary allocation of a given ball whenver the suggested bin has accepted thus far less than ℓ primary allocations. In other words,

$$f_k(\mathcal{H}, i, u) = \begin{cases} 1 & \text{if } L^f_{1,i}(k) < \ell, \\ 2 & \text{if } L^f_{1,i}(k) \ge \ell. \end{cases}$$

This strategy is used to control the single-time maximum load of allocating $O(n\sqrt{\log n})$ balls.

The relative threshold strategy. The *ℓ-relative threshold strategy* accepts the k-th primary allocation if the suggested bin has accepted less than $\ell + \frac{k-1}{n}$ primary allocations or if the load of the suggested bin is below $-\log n$. In other words,

$$f_k(\mathcal{H}, i, u) = \begin{cases} 1 & \text{if } L^f_{1,i}(k) < \ell + \frac{k-1}{n} \text{ or } L^f_i(k) < -\log n, \\ 2 & \text{if } L^f_{1,i}(k) \ge \ell + \frac{k-1}{n} \text{ and } L^f_i(k) \ge -\log n. \end{cases}$$

This strategy is designed to control the all-time maximum load of allocating $o(n \log^2 n)$ balls.

The drift strategy. The third strategy relies on a coupling of the allocation process and a continuous time random process. This strategy can be used to achieve appropriate initial conditions for other strategies as it is very robust and can rather quickly reduce the load vector to a stationary distribution with an exponential tail. We denote by $\{X_i(t)\}_{i\in[n]}$ a collection of independent regular point processes with initial values $X_i(0) = L_i(0)$ and conditional intensity functions

$$\lambda_i(t) = \begin{cases} 1+\theta, & X_i(t) < t, \\ 1-\theta, & X_i(t) \geq t. \end{cases} \tag{2.9}$$

Write $X(t) = \sum_{i=1}^n X_i(t)$. We define the random process $\{Z_k\}_{k\in\mathbb{N}}$ as follows. For any $k \in \mathbb{N}$, we set

$$Z_k = i \text{ if the } k\text{-th point of } X(t) \text{ for } t > 0 \text{ is a point of } X_i(t). \tag{2.10}$$

We will show that, conditioned on $Z_1, \ldots, Z_{k-1}$, the variable Z_k meets the conditions of Lemma 2.1. Hence $\{Z_k\}_{k\in\mathbb{N}}$ is realizable as the output of a two-thinning strategy. We call this strategy the *θ-drift strategy*.

We write $\mathcal{F}_t$ for the natural filtration of $X(t)$ and denote by $T_k = \inf\{t : X(t) = k\}$. To see that the conditions of Lemma 2.1 are indeed satisfied, it suffices to show that there exists some $c > 0$ such that

$$\frac{c}{n} \leq \mathbb{P}(Z_k = i \mid Z_1, \cdots, Z_{k-1}, \mathcal{F}_{T_{k-1}}) \leq \frac{1+c}{n} \tag{2.11}$$

holds for all $k \in \mathbb{N}$ and all $i \in [n]$. By the definition of $\{Z_k\}_{k\in\mathbb{N}}$, we have

$$\frac{1-\theta}{n(1+\theta)} = \frac{\inf\limits_{t\geq 0} \lambda_i(t)}{n \max\limits_{j\in[n]} \sup\limits_{t\geq 0} \lambda_j(t)} \leq \mathbb{P}(Z_k = i \mid Z_1, \cdots, Z_{k-1}, \mathcal{F}_{T_{k-1}}) \leq \frac{\sup\limits_{t\geq 0} \lambda_i(t)}{n \min\limits_{j\in[n]} \inf\limits_{t\geq 0} \lambda_j(t)} \leq \frac{1+\theta}{n(1-\theta)}.$$

One can check that the criterion (2.11) holds for all $0 < \theta \leq \sqrt{5} - 2$.

A varying drift strategy. Our forth strategy is a modified drift strategy where the downwards drift is extremely strong for bins with loads above certain level ℓ. We denote by $\{X_i(t)\}_{i\in[n]}$ a collection of independent regular point processes with initial values $X_i(0) = 0$ and conditional intensity functions given by

$$\lambda_i(t) = \begin{cases} 1+\theta_1, & X_i(t) < t, \\ 1-\theta_2, & t \leq X_i(t) \leq t+\ell, \\ \theta_3, & X_i(t) > t+\ell. \end{cases} \tag{2.12}$$

Here, we set $\theta_1 = \theta_2 = \frac{1}{\sqrt{\log n}}$ and $\theta_3 = \frac{12}{\sqrt{\log n}}$. We write $X(t) = \sum_{i=1}^n X_i(t)$. For any $k \in \mathbb{N}$, we set $Z_k = i$ if the k-th point of $X(t)$ for $t > 0$ is a point of the process $X_i(t)$. We write $\mathcal{F}_t$ for the natural filtration of $X(t)$ and denote by $T_k = \inf\{t : X(t) = k\}$. Unlike in the case of the drift strategy, in certain situations, the distribution of Z_k given $Z_0, \ldots, Z_{k-1}$ is not the output of any two-thinning decision strategy. However, as the next lemma shows, this does not happen as long as the number of bins with very high load is not too large. We call the strategy which realizes Z_k for as long as possible (and, say, accepts all primary allocations from that time and on, for the sake of completion), the *ℓ-varying drift strategy*.

Lemma 2.2. *For sufficiently large n, for any $k \in \mathbb{N}$, if*

$$\left|\left\{i \in [n] : X_i(T_{k-1}) > T_{k-1} + \ell\right\}\right| \le \frac{n}{\sqrt{\log n}}, \tag{2.13}$$

then the distribution of Z_k given $Z_0, \ldots, Z_{k-1}$ can be realized by a two-thinning decision strategy.

Proof. We need to verify that the distribution of Z_k given $Z_0, \ldots, Z_{k-1}$ satisfies the condition of Lemma 2.1. To this end, it is enough to show that there exists some $c > 0$, which could depend on n, such that for sufficiently large n, for all $i \in [n]$ we have,

$$\frac{c}{n} \le \mathbb{P}(Z_k = i \mid Z_1, \cdots, Z_{k-1}, \mathcal{F}_{T_{k-1}}) \le \frac{1+c}{n}. \tag{2.14}$$

Denote $n_0 = |\{i \in [n] : X_i(T_{k-1}) > T_{k-1} + \ell\}|$. Then, the condition (2.13) says that $n_0 \le \frac{n}{\sqrt{\log n}}$. By the definition of Z_k, we have

$$\frac{\theta_3}{(n-n_0)(1+\theta_1)+n_0\theta_3} \le \mathbb{P}(Z_k = i \mid Z_1, \cdots, Z_{k-1}, \mathcal{F}_{T_{k-1}}) \le \frac{1+\theta_1}{(n-n_0)(1-\theta_2)+n_0\theta_3}$$

Using the fact that the denominators above are maximized when $n_0 = 0$ and are minimized when $n_0 = \frac{n}{\sqrt{\log n}}$, we obtain

$$\frac{6}{n\sqrt{\log n}} \le \mathbb{P}(Z_k = i \mid Z_1, \cdots, Z_{k-1}, \mathcal{F}_{T_{k-1}}) \le \frac{1+\frac{1}{\sqrt{\log n}}}{n\left(1-\frac{1}{\sqrt{\log n}}\right)^2} \le \frac{1}{n}\left(1+\frac{4}{\sqrt{\log n}}\right)$$

for all n sufficiently large. Thus, inequality (2.14) holds with $c = \frac{4}{\sqrt{\log n}}$. □

We shall see in Section 6.2 that for $\ell = \frac{2\log n}{\log\log n}$, the condition of Lemma 2.2 is indeed satisfied with high probability over polynomially long time in n.

2.4 Combinations of the basic strategies

In many scenarios, particularly when the number of balls is large, we need to adjust and combine the basic strategies in an appropriate way to obtain the upper bounds in our main results. The following are several such combinations.

The multi-stage (t, L_0, ℓ)-threshold strategy. Set $t_0 = 0$ and $k = \left\lfloor \frac{\log\log n}{3\log\log\log n} \right\rfloor$. We divide the process into k stages, where the i-th stage proceeds from time nt_{i-1} to time nt_i, where the definition of t_i as a function of t is given at the end of this description. We write H_0 for the set of bins with loads greater than L_0 at time t_0. We inductively define H_i as the set of bins in $(\cup_{j=1}^{i-1} H_j)^c$ (or in H_0^c in the case $i = 1$) whose loads at the end of the i-th stage are at least $L_0 + 2i\ell$. Then our strategy can be stated as follows. In the first stage, we retry a ball if its primary allocation bin has a load of at least $-\log n$ and either it is in H_0 or it has accepted $t_1 - t_0 + \ell$ primary allocations in the first stage so far. In i-th stage for $i \ge 2$, we retry a ball if its primary allocation bin has a load of at least $-\log n$ and either it is in $\cup_{j=1}^{i-1} H_j$, or it is a bin that has accepted $t_i - t_{i-1} + \ell$ primary allocations during the i-th stage so far.

Now we conclude the description with the definitions of $\{t_i\}_{i\in[k]}$. Denote $\alpha = \frac{\log t}{\log\log n}$. Given $\eta \in [0, \frac{\alpha-1/2}{4k-2}]$, we set $\beta = \alpha + \eta$, $\varepsilon = \frac{2\beta-1}{2(k+1)}$, and $\beta_i = \beta - \frac{(2\beta-1-\varepsilon)i}{2k+1}$. We then define $t_i = \lfloor t - \log^{\beta_i} n \rfloor$ for $1 \le i \le k-1$, and $t_k = t$.

Remark 2.3. *It might be worthwhile to point out that after the first stage, we do not retry primary allocations that are bins in H_0 unless they consist of bins with load at least $-\log n$ and already accepted ℓ primary allocations more than the average in the current stage. Hence, the initial set of heavily loaded bins H_0 will play the same role as any other bins from stage two and on.*

This multi-stage threshold strategy is designed to control the single-time maximum load for time $t \geq \omega(\sqrt{\log n})$, in which case the threshold strategy alone is not sufficient. Indeed, optimizing the choice of the threshold in the threshold strategy gives $t+\Theta((t\log n)^{1/3})$, which, in turn, yields a maximum load of $O((t\log n)^{1/3})$; much larger than the desired upper bound $(\log n)^{1/2+o(1)}$. Hence, we divide the process into multiple shorter stages and in each stage apply the threshold strategy with a smaller threshold. This is likely to cause more retries and even a temporarily higher maximum load. To prevent this from causing high load at the end of the process, we identify at the beginning of every stage heavily loaded bins (H_i) and from this time and on retry a ball if its primary allocation is one of these. The number of retries caused by this requirement is relatively small since the number of heavily loaded bins is small. This, together with a careful selection of time lengths of the stages, will effectively reduce the maximum load to $(\log n)^{1/2+o(1)}$ at the end of the process.

A sketch of the analysis of the strategy is as follows. We first control the maximum load after the first stage, and the number of relatively heavily loaded bins at the end of it (i.e., H_1). In every subsequent stage i there are two causes for retries: either the suggested bin already accumulated ℓ primary allocations more than the average in this stage, or it was marked as heavily loaded in previous stages (i.e., it is in $\cup_{j=1}^{i-1}H_j$). By inductive bounds on these, we are able to control the number of such retries. For a bin to be included into H_i, it must accumulate at least 2ℓ allocations above average, so that at least ℓ of them are secondary. Using binomial estimates we can control the number of such bins with high probability and establish our bound on H_i. Similar computations also allow us to control the maximum load in bins $\cup_{j=1}^{i}H_j$, taking advantage of the negative drift of the load in $\cup_{j=1}^{i-1}H_j$, caused by the fact that they are always rejected as primary allocations (except if the load is already lower than $-\log n$).

The drift-multi-stage $(\theta, t', t, L_0, \ell)$-threshold. This strategy is a combination of the drift strategy and the multi-stage threshold strategy. It is designed to control the single-time maximum load for $t \gg \log n$. This is simply done by applying the θ-drift strategy up to time t' followed by the multi-stage (t, L_0, ℓ)-threshold strategy starting at time t' and ending at time $t'+t$.

The Q-multi-scale strategy. This strategy is designed for controlling the typical maximum load for about $n(\log n)^{1+o(1)}$ time. The strategy is formed by multiple scales, each of which extends the previous one and consists of multiple iterations of the previous scale strategy separated by a different regulating strategy. Whenever we initiate a new strategy at some time, we treat this time point as the initiation time and the current loads as the initial loads for the new strategy. To avoid countless rounding operations, each strategy is applied for a not-necessarily integer time, and our policy is that if an integer point falls within the time domain of a strategy, then this strategy is applied to it.

We now give the exact description of the strategy, which is accompanied by an algorithmic description and a demonstration of the first three scales in Figure 1. We postpone the technical definitions of the parameters $L > 0, k \in \mathbb{N}, \{\alpha_i, \alpha_i', \ell_i\}_{i\in\mathbb{N}}$ after the description. We write $N_i = \lceil \frac{L}{3k\ell_i} \rceil$ and $Q^{i,j} = (2k+1)(j-1)\ell_i$. In the first scale, we simply apply the L-relative threshold strategy up to time $n\lfloor \log^{\alpha_1} n \rfloor$. In the second scale, we apply N_1 iterations of the

first scale strategy (the last iteration may be incomplete) and the j-th iteration is followed by the multi-stage $(\log^{\alpha'_1} n, Q + Q^{1,j} + \ell_1, \ell_1)$-threshold strategy. The value of Q in the j-th iteration of the first scale strategy is increased by $Q^{1,j}$. Generally, in the $(i+1)$-th scale, we apply N_i iterations of i-th scale strategy and the j-th iteration is followed by the multi-stage $(\log^{\alpha'_i} n, Q + Q^{i,j} + \ell_i, \ell_i)$-threshold strategy. In the j-th iteration, all values of Q in the nested multi-scale strategies are increased by $Q^{i,j}$ (in comparison with the value of Q in the current scale).

The technical definitions of the aforementioned parameters are given as follows. We set $\alpha_1 = \frac{1}{2} + \frac{2}{\lfloor\sqrt{\log\log\log n}\rfloor + 1/4}$, $L = (\log n)^{\frac{1+\alpha_1}{3}}$ and $k = \left\lfloor \frac{\log\log n}{3\log\log\log n} \right\rfloor$. We inductively define the sequences $\{\alpha_i, \alpha'_i, \ell_i\}_{i\in\mathbb{N}}$ via the following equations

$$\begin{aligned} \varepsilon_i &= \frac{2\alpha_i - 1}{2(k+1)}, \\ \ell_i &= (\log n)^{\frac{1}{2} + \frac{\alpha_i - 1/2 + k\varepsilon_i}{2k+1}}, \\ \alpha'_i &= \alpha_i - \frac{1}{5} \cdot \frac{2\alpha_i - 1 - \varepsilon_i}{2k+1}, \qquad (2.15) \\ \log^{\alpha_{i+1}} n &= N_i(\lfloor \log^{\alpha_i} n \rfloor + \lfloor \log^{\alpha'_i} n \rfloor). \qquad (2.16) \end{aligned}$$

According to the description of our strategy, the first part of each iteration runs for $n\log^{\alpha_i} n$ time, and the second part runs for $n\log^{\alpha'_i} n$ time, so that the $(i+1)$-th scale runs for $n\log^{\alpha_{i+1}} n$ time in total.

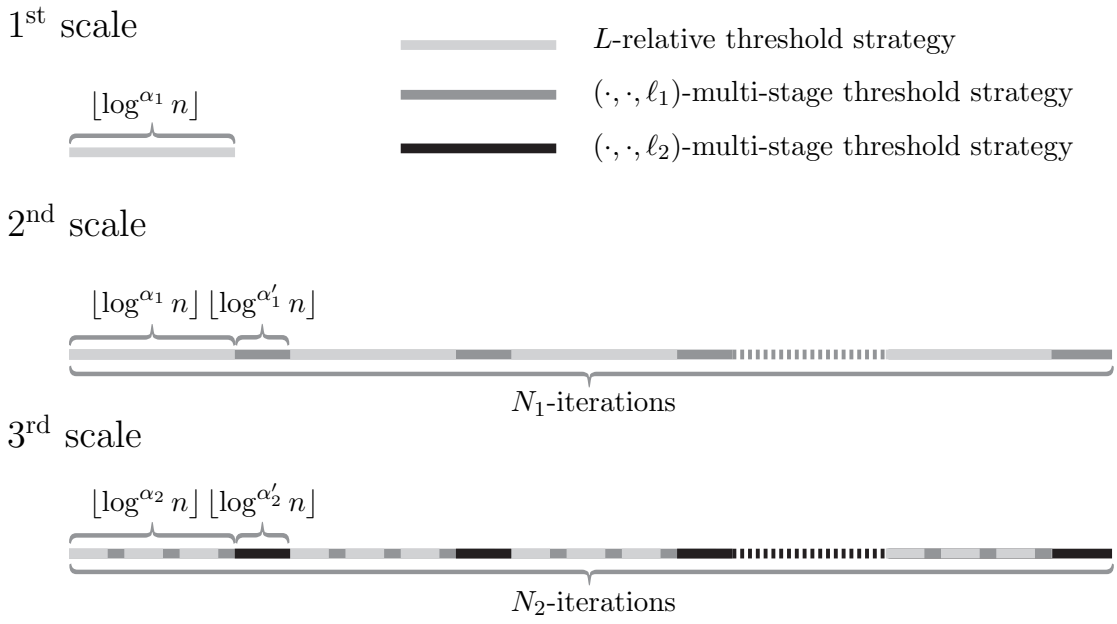
The idea behind this strategy is as follows. In each scale of the strategy, most of the time we apply the lower scale strategy, which yields a good control of the typical maximum load. However, the number of bins with loads close to the threshold will accumulate along the time. In order to mitigate this effect, we need to apply the multi-stage threshold strategy with a low threshold for a short period of time. This enables us to dramatically reduce the number of such relatively high loaded bins at the end of each regulating period, although it is possible that during these regulating periods, certain bins may temporarily accumulate very high loads. Once the regulating period is over, the small number of relatively high load bins allows us to iterate the lower scale strategy once again.

In the following figure, we provide an algorithmic description of the Q-multi-scale strategy and a demonstration of the first three scales of the strategy.

Algorithm 1 *Q-multi-scale* (Scale=$i+1$)

if $i = 0$ **then**
  Run *L-relative threshold* for $\log^{\alpha_1} n$ time
else
  for $j = 1$ to N_i **do**:
    Run $\big(Q + Q^{i,j}\big)$*-multi-scale*$\big(i\big)$ for $\log^{\alpha_i} n$ time
    Run *multi-stage* $(\log^{\alpha'_i} n, Q + Q^{i,j} + \ell_i, \ell_i)$*-threshold*
  end for
end if

Figure 1: **Above:** an algorithmic description of the Q-multi-scale strategy. **Below:** the first three scales of this strategy. The first scale is the L-relative threshold strategy. The second scale consists of N_1 iterations, the j-th of which incorporates the strategy of the first scale followed by the multi-stage $(\log^{\alpha'_1} n, Q + Q^{1,j} + \ell_1, \ell_1)$-threshold strategy. The third scale consists of N_2 iterations, each of which consists of the second scale strategy with its Q set to be $Q + Q^{2,j}$, followed by the multi-stage $(\log^{\alpha'_2} n, Q + Q^{2,j} + \ell_2, \ell_2)$-threshold strategy.



The d-multi-scale long-term combined strategy. This strategy is used to control the typical maximum load for arbitrarily long time and it consists of multiple iterations. As in the Q-multi-scale strategy, we set $\alpha_1 = \frac{1}{2} + \frac{2}{\lfloor\sqrt{\log\log\log n}\rfloor + 1/4}$, $L = (\log n)^{\frac{1+\alpha_1}{3}}$ and $k = \lfloor \frac{\log\log n}{3\log\log\log n} \rfloor$. The sequence $\{\alpha_i\}_{i\in\mathbb{N}}$ is defined in (2.15) and (2.16). We denote by $i_{\max} = \max\{i \in \mathbb{N} : \alpha_i \le 1\}$. We set

$$Q = L = (\log n)^{\frac{1+\alpha_1}{3}}, \quad A = \sqrt{6d(\log n)^{1+\alpha_{i_{\max}+1}}}, \tag{2.17}$$

$$m_0 = \lfloor 200dn\log n \rfloor, \quad m_1 = n(\log n)^{\alpha_{i_{\max}+1}}, \quad m_2 = \lceil 16nA \rceil. \tag{2.18}$$

$$L_0 = \lfloor (\log n)^{\frac{1}{2} + (2 - \frac{1}{2k+1})\frac{\alpha - 1/2}{2k+1}} \rfloor, \text{ where } \alpha = \frac{\log(m_0/n)}{\log\log n}. \tag{2.19}$$

In this strategy, a *standard iteration* consists of three phases: The first one consists of the allocation of m_0 balls according to the multi-stage $(m_0/n, L_0, L_0)$-threshold strategy defined in Section 2.4; the second phase consists of the allocation of m_1 balls using the Q-multi-scale strategy; the third phase consists of the allocation of balls according to the 1/5-drift strategy given in Section 2.3, until the first time m when the following three conditions are satisfied

- At least m_2 balls were allocated during this phase,
- $\max_{i\in[n]} \left|L_i^f(m)\right| \le 100d\log n$,

- $\left|\left\{i \in [n] : L_i^f(m) > L_0\right\}\right| < 4000ne^{-L_0/15}$.

The strategy itself consists of applying such iterations indefinitely, with the exception that we skip the first phase in the first iteration. The purpose of this exception is to make this strategy an extension of the Q-multi-scale strategy.

2.5 Optimal strategies

We summarize in Table 2 the strategies and the time intervals where these strategies are employed to control the single-time, all-time and typical maximum loads. Notice that strategies that work for larger values of m encapsulate those that work for smaller values so that the more advanced strategy could be also used for smaller values of m.

<table>
<tr><th></th><th>$m \le O(n\sqrt{\log n})$</th><th>$m \le O(n \log n)$</th><th>$m \le O(n \log^2 n)$</th><th>$m \le n^{O(1)}$</th><th>generic m</th></tr>
<tr><td>Maximum load at time m</td><td>Threshold strategy</td><td>Multi-stage threshold strategy</td><td colspan="3">Drift multi-stage threshold strategy</td></tr>
<tr><td>Maximum load up to time m</td><td colspan="2">Relative threshold strategy</td><td colspan="2">Varying drift strategy</td><td>-</td></tr>
<tr><td>Typical load up to time m</td><td>Relative threshold Strategy</td><td colspan="2">Q-multi-scale threshold strategy</td><td colspan="2">d-multi-scale long-term combined strategy</td></tr>
</table>

Table 2: Optimal strategies for the single-time, all-time and typical maximum loads.

3 Preliminaries

3.1 Poisson approximation

One difficulty of analyzing the balls-and-bins model is the correlation among the loads of different bins. The following result shows that the joint distribution of the loads of different bins can be well approximated by assuming that the loads of these bins are independent Poisson(m/n) random variables.

Let $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. Given $x, y \in (\mathbb{N}_0)^n$, we say that $x \le y$ if $x_i \le y_i$ for all $i \in [n]$. A subset $S \subset (\mathbb{N}_0)^n$ is called *monotone decreasing (resp. increasing)* if $x \in S$ implies that $y \in S$ for all $y \le x$ (resp. $x \le y$).

Lemma 3.1 ([19], Theorem 5.10)**.** *Let $\{X_i\}_{i\in[n]}$ be the number of balls in bins $i \in [n]$ when m balls are independently and uniformly placed into n bins. Let $\{Y_i\}_{i\in[n]}$ be independent* Poisson(m/n) *random variables. For any monotone set $S \subseteq [n]$, we have*

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

We borrow the following lemma from [11], which provides a concentration bound on the maximum load over a subset of bins.

Lemma 3.2 ([11], Lemma 2.2)**.** *Let $\{X_i\}_{i\in[n]}$ be the number of balls in bins $i \in [n]$ when $\lfloor \theta n \rfloor, 0 \le \theta \le 1$, balls are independently and uniformly placed into n bins. For $k \in \lfloor \theta n \rfloor$ and $S \subseteq [n]$, we have*

$$\mathbb{P}\left(\max_{i\in S} X_i < k\right) \le 2\exp\left(-\frac{\theta^k |S|}{ek!}\right).$$

3.2 Poisson tail estimate

Let X be a Poisson(λ) random variable. When λ is an integer, X can be seen as the sum of λ independent Poisson(1) random variables. As a consequence of Cramér's Theorem (e.g., [10], Theorem 2.2.3), $\lambda^{-1}X$ satisfies the Large Deviation Principle (LDP), namely, for any closed set $F \subset \mathbb{R}$,

$$\limsup_{\lambda\to\infty} \frac{1}{\lambda} \log \mathbb{P}(\lambda^{-1}X \in F) \le -\inf_{x\in F} \Lambda^*(x),$$

and for any open set $J \subset \mathbb{R}$,

$$\liminf_{\lambda\to\infty} \frac{1}{\lambda} \log \mathbb{P}(\lambda^{-1}X \in J) \ge -\inf_{x\in J} \Lambda^*(x),$$

where the rate function

$$\Lambda^*(x) = \begin{cases} 1 - x + x\log x, & x > 0 \\ +\infty, & \text{otherwise.} \end{cases}$$

The statement actually holds for general λ that is not necessarily an integer. This LDP readily implies the following Poisson tail bounds.

Lemma 3.3. *Let X be a* Poisson(λ) *random variable. For sufficiently large λ and any $\kappa > 0$,*

$$e^{-2\lambda I(\kappa/\lambda)} \le \mathbb{P}(X \ge \lambda + \kappa) \le e^{-\lambda I(\kappa/\lambda)}, \tag{3.1}$$

and for any $0 < \kappa < \lambda$,

$$e^{-2\lambda I(-\kappa/\lambda)} \le \mathbb{P}(X \le \lambda - \kappa) \le e^{-\lambda I(-\kappa/\lambda)}, \tag{3.2}$$

where $I(x) = \Lambda^*(1+x) = (1+x)\log(1+x) - x$ for $x \in (-1, \infty)$.

Remark 3.4. *In fact, the upper bounds hold for any $\lambda > 0$ and this readily follows from Chernoff's argument. As the name, LDP, indicates, Lemma 3.3 provides a good approximation of the Poisson tail when κ is larger than the standard deviation $\sqrt{\lambda}$. The following approximation of the rate function $I(x)$ will be repeatedly used. For $0 \le x \le 4$, we have*

$$\frac{x^2}{4} \le I(x) \le \frac{x^2}{2} \tag{3.3}$$

and, for $x \ge 4$, we have

$$x\log\frac{x}{e} \le I(x) \le 3x\log\frac{x}{e}. \tag{3.4}$$

The following result will be repeatedly used in later sections to estimate the number of retries in the allocation of balls using the threshold strategy.

Lemma 3.5. *Let $\{X_i\}_{i\in[n]}$ be independent* Poisson(λ) *random variables. Let $\ell > 0$. We define $Y_i = \max\{0, X_i - \lambda - \ell\}$ and $Y = \sum_{i=1}^n Y_i$. Set $r^* = 6ne^{-\lambda I(\ell/\lambda)}/\log(1+\ell/\lambda)$, where the function $I(x)$ is given in Lemma 3.3. Then we have*

$$\mathbb{P}(Y > r^*) < \exp\left(-ne^{-\lambda I(\ell/\lambda)}\right). \tag{3.5}$$

Proof. The statement follows from the classical Chernoff's argument. For any $u > 0$, we have

$$\begin{aligned}
\mathbb{E}e^{uY_1} &< 1 + e^{-u\ell} \sum_{k=\lceil\lambda+\ell\rceil}^{\infty} e^{u(k-\lambda)} \cdot \mathbb{P}(X_1 = k) \\
&= 1 + e^{-u\ell} \sum_{k=\lceil\lambda+\ell\rceil}^{\infty} e^{u(k-\lambda)} \Big(\mathbb{P}(X_1 \geq k) - \mathbb{P}(X_1 \geq k+1)\Big) \\
&= 1 + e^{-u\ell} \left(\sum_{k=\lceil\lambda+\ell\rceil}^{\infty} e^{u(k-\lambda)} \cdot \mathbb{P}(X_1 \geq k) - e^{-u} \sum_{k=\lceil\lambda+\ell\rceil+1}^{\infty} e^{u(k-\lambda)} \cdot \mathbb{P}(X_1 \geq k)\right) \\
&= 1 + e^{-u\ell} \left(e^{u(\lceil\lambda+\ell\rceil-\lambda)} \cdot \mathbb{P}(X_1 \geq \lceil\lambda+\ell\rceil) + (1-e^{-u}) \sum_{k=\lceil\lambda+\ell\rceil+1}^{\infty} e^{u(k-\lambda)} \cdot \mathbb{P}(X_1 \geq k)\right).
\end{aligned}$$

Write $\ell^* = \lceil\lambda+\ell\rceil - \lambda$ and $j_k = k - \lambda$. We obtain

$$\mathbb{E}e^{uY_1} < 1 + e^{-u\ell} \left(e^{u\ell^*} \cdot \mathbb{P}(X_1 \geq \lambda + \ell^*) + (1-e^{-u}) \sum_{k=\lambda+\ell^*+1}^{\infty} e^{uj_k} \cdot \mathbb{P}(X_1 \geq \lambda + j_k)\right). \tag{3.6}$$

For any $k > 0$, we apply Lemma 3.3 to obtain

$$e^{uk} \cdot \mathbb{P}(X_1 \geq \lambda + k) \leq e^{\lambda g_u(k/\lambda)}, \tag{3.7}$$

where $g_u(x) = (1+u)x - (1+x)\log(1+x)$. One can check that $g_u'(x) = u - \log(1+x)$ and that $g_u''(x) = -(1+x)^{-1} < 0$. Let $u^* = \frac{1}{2}\log(1+\ell/\lambda)$. Then, $g_{u^*}(x)$ is a decreasing and concave function for $x \geq \ell/\lambda$. Hence, we have for any $k \geq \ell$ that

$$\frac{e^{\lambda g_{u^*}((k+1)/\lambda)}}{e^{\lambda g_{u^*}(k/\lambda)}} = \exp\left(\frac{g_{u^*}((k+1)/\lambda) - g_{u^*}(k/\lambda)}{1/\lambda}\right) < e^{g'_{u^*}(k/\lambda)} \leq e^{g'_{u^*}(\ell/\lambda)} = e^{-u^*}, \tag{3.8}$$

where the second equality follows from the formula for $g'_{u^*}(x)$ and our choice of u^*. Combining (3.6), (3.7) and (3.8), we have

$$\begin{aligned}
\mathbb{E}e^{u^*Y_1} &< 1 + e^{-u^*\ell} \left(e^{\lambda g_{u^*}(\ell^*/\lambda)} + (1-e^{-u^*}) \sum_{k=\lambda+\ell^*+1}^{\infty} e^{\lambda g_{u^*}(j_k/\lambda)}\right) \\
&< 1 + e^{-u^*\ell} \left(e^{\lambda g_{u^*}(\ell^*/\lambda)} + e^{\lambda g_{u^*}((\ell^*+1)/\lambda)}\right) \\
&< 1 + 2e^{-u^*\ell} \cdot e^{\lambda g_{u^*}(\ell/\lambda)} = 1 + 2e^{-\lambda I(\ell/\lambda)} \\
&< \exp\left(2e^{-\lambda I(\ell/\lambda)}\right),
\end{aligned}$$

where the second last inequality follows from the fact that $\ell^* \geq \ell$ and that $g_{u^*}(x)$ is decreasing for $x \geq \ell/\lambda$. Then we apply Markov's inequality to obtain for any $r > 0$ that

$$\mathbb{P}(Y > r) \leq e^{-u^*r}\mathbb{E}e^{u^*Y} = e^{-u^*r}\left(\mathbb{E}e^{u^*Y_1}\right)^n < \exp\left(2ne^{-\lambda I(\ell/\lambda)} - u^*r\right).$$

Recall that $u^* = \frac{1}{2}\log(1+\ell/\lambda)$. In particular, for $r^* = 6ne^{-\lambda I(\ell/\lambda)}/\log(1+\ell/\lambda)$, we have

$$\mathbb{P}(Y > r^*) < \exp\left(-ne^{-\lambda I(\ell/\lambda)}\right).$$

This concludes the proof. □

3.3 Concentration bounds for the drift strategy

As our drift strategy is based on a coupling of the allocation process and a continuous time random process, our concentration bounds for the drift strategy rely on the study of a particular type of temporal point processes. We refer the interested readers to [8, 9] for more details of general temporal point processes.

θ-standardizing point process. A temporal point process $X(t)$ is called *θ-standardizing* if the conditional intensity function $\lambda(t)$ satisfies

$$\lambda(t) < 1-\theta, \quad \text{if } X(t) \geq t, \tag{3.9}$$
$$\lambda(t) \geq 1+\theta, \quad \text{if } X(t) < t. \tag{3.10}$$

We say that $X(t)$ is *upper θ-standardizing* if (3.9) holds, and that $X(t)$ is *lower θ-standardizing* if (3.10) holds.

Lemma 3.6. *Let $\{X(t)\}_{t\geq 0}$ be a temporal point process adapted to the filtration $\{\mathcal{F}_t\}_{t\geq 0}$. Let $s \geq 0$ be a stopping time with respect to $\{\mathcal{F}_t\}_{t\geq 0}$ and let $\eta \in [0,1]$ be a $\mathcal{F}_s$ measurable random variable. Denote $Y(t) = X(t) - t$.*

1. *If $X(t)$ is upper 2θ-standarizing, then we have*

$$\mathbb{E}\left[e^{\theta Y(s+\eta)} \mid \mathcal{F}_s\right] \leq e^{-\theta^2\eta}\cdot e^{\theta Y(s)} + e^{2\theta}, \tag{3.11}$$

and for any λ satisfying $(1-2\theta)e^{\lambda} < \lambda/2$, we have

$$\mathbb{E}\left[e^{\lambda Y(s+\eta)} \mid \mathcal{F}_s\right] \leq e^{-\frac{\lambda}{2}\eta}\cdot e^{\lambda Y(s)} + e^{2\lambda}. \tag{3.12}$$

2. *If $X(t)$ is lower 2θ-standarizing, then we have*

$$\mathbb{E}\left[e^{-\theta Y(s+\eta)} \mid \mathcal{F}_s\right] \leq e^{-\theta^2\eta}\cdot e^{-\theta Y(s)} + e^{\theta}. \tag{3.13}$$

3. *If $X(t)$ is 2θ-standarizing, then we have*

$$\mathbb{E}\left[e^{\theta|Y(s+\eta)|} \mid \mathcal{F}_s\right] \leq e^{-\theta^2\eta}\cdot e^{\theta|Y(s)|} + 3e^{2\theta}. \tag{3.14}$$

Proof. We denote by $Z(\beta)$ a Poisson(β) random variable throughout the proof. We first prove inequalities (3.11) and (3.12). We need to estimate the Laplace transform of $Z(\alpha(1-2\theta))$ for any $\alpha > 0$ as follows

$$\begin{aligned}\mathbb{E}e^{\lambda[Z(\alpha(1-2\theta))-\alpha]} &= e^{\alpha(1-2\theta)(e^{\lambda}-1)-\alpha\lambda} \\ &\leq \begin{cases} e^{\alpha(1-2\theta)(\lambda+\lambda^2)-\alpha\lambda} \leq e^{\alpha(\lambda^2-2\lambda\theta)} & 0\leq\lambda\leq 1, \\ e^{\alpha(1-2\theta)e^{\lambda}-\alpha\lambda} \leq e^{-\lambda\alpha/2} & (1-2\theta)e^{\lambda} \leq \lambda/2. \end{cases}\end{aligned} \tag{3.15}$$

We define $s_* = \min\{t \in [s, s+\eta] \ : \ Y(t) \geq 1\}$ and set $s_* = s+\eta$ if the minimum is taken over an empty set. Then, s_* is a stopping time with respect to $\{\mathcal{F}_t\}_{t\geq 0}$. We have

$$\begin{aligned}
\mathbb{E}\big[e^{\lambda Y(s+\eta)} \mid \mathcal{F}_{s_*}\big] &= e^{\lambda Y(s_*)} \cdot \mathbb{E}\big[e^{\lambda[Y(s+\eta)-Y(s_*)]} \mid \mathcal{F}_{s_*}\big] \\
&\leq e^{\lambda Y(s_*)} \cdot \mathbb{E}\big[e^{\lambda[Z((1-2\theta)(s+\eta-s_*))-(s+\eta-s_*)]} \mid \mathcal{F}_{s_*}\big] \\
&\leq \begin{cases} e^{-\theta^2(s+\eta-s_*)} \cdot e^{\theta Y(s_*)} & \lambda = \theta, \\ e^{-\frac{\lambda}{2}(s+\eta-s_*)} \cdot e^{\lambda Y(s_*)} & (1-2\theta)e^{\lambda} \leq \lambda/2. \end{cases} \\
&\leq \begin{cases} e^{-\theta^2\eta} \cdot e^{\theta Y(s)} + e^{2\theta} & \lambda = \theta, \\ e^{-\frac{\lambda}{2}\eta} \cdot e^{\lambda Y(s)} + e^{2\lambda} & (1-2\theta)e^{\lambda} \leq \lambda/2. \end{cases}
\end{aligned} \tag{3.16}$$

To see the first inequality, observe that $Y(t) = X(t) - t \geq 0$ for $t \in [s_*, s+\eta]$. Since $X(t)$ is upper 2θ-standardizing, $Y(s+\eta) - Y(s_*) = X(s+\eta) - X(s_*) - (s+\eta-s_*)$ is dominated by $Z((1-2\theta)(s+\eta-s_*)) - (s+\eta-s_*)$. The second inequality follows from (3.15). In each case of (3.16), the first term is an upper bound for the case $s_* = s$, while the second term uses the fact that $Y(s_*) < 2$ when $s_* \neq s$. Inequalities (3.11) and (3.12) follow from the tower property of conditional expectation and (3.16).

Next we prove (3.13). Write $E = \{Y(t) \leq 0 \text{ for all } t \in [s, s+\eta]\}$. Observe that, whenever E^c occurs, we have $Y(s+\eta) \geq -1$. Hence,

$$\begin{aligned}
\mathbb{E}\big[e^{-\theta Y(s+\eta)} \mid \mathcal{F}_s\big] &= \mathbb{E}\big[e^{-\theta Y(s+\eta)}\mathbb{1}_E + e^{-\theta Y(s+\eta)}\mathbb{1}_{E^c} \mid \mathcal{F}_s\big] \\
&\leq \mathbb{E}\big[e^{-\theta Y(s+\eta)}\mathbb{1}_E \mid \mathcal{F}_s\big] + e^{\theta} \\
&= e^{-\theta Y(s)} \cdot \mathbb{E}\big[e^{-\theta[Y(s+\eta)-Y(s)]}\mathbb{1}_E \mid \mathcal{F}_s\big] + e^{\theta} \\
&\leq e^{-\theta Y(s)} \cdot \mathbb{E}\big[e^{-\theta[Z((1+2\theta)\eta)-\eta]} \mid \mathcal{F}_s\big] + e^{\theta} \\
&\leq e^{-\theta^2\eta} \cdot e^{-\theta Y(s)} + e^{\theta}.
\end{aligned}$$

To see the second inequality, observe that, whenever E occurs, we have $Y(t) = X(t) - t \leq 0$ for all $t \in [s, s+\eta]$. Since $X(t)$ is 2θ-standardizing, $Y(s+\eta) - Y(s) = X(s+\eta) - X(s) - \eta$ dominates $Z((1+2\theta)\eta) - \eta$. The last inequality follows from that for any $\alpha > 0$,

$$\mathbb{E}e^{-\theta[Z(\alpha(1+2\theta))-\alpha]} = e^{\alpha(1+2\theta)(e^{-\theta}-1)+\alpha\theta} < e^{\alpha(1+2\theta)(-\theta+\theta^2/2)+\alpha\theta} < e^{-\alpha\theta^2}.$$

When $X(t)$ is 2θ-standarizing, it is both upper and lower 2θ-standarizing. Hence, inequalities (3.11) and (3.13) hold. Observe that

$$\mathbb{E}\big[e^{\theta|Y(s+\eta)|} \mid \mathcal{F}_s\big] \leq \mathbb{E}\big[e^{\theta Y(s+\eta)} \mid \mathcal{F}_s\big] + \mathbb{E}\big[e^{-\theta Y(s+\eta)} \mid \mathcal{F}_s\big].$$

This, together with (3.11) and (3.13), yields (3.14). □

Corollary 3.7. *Let $\{X(t)\}_{t\geq 0}$ be a temporal point process adapted to the filtration $\{\mathcal{F}_t\}_{t\geq 0}$. Denote $Y(t) = X(t) - t$.*

1. *If $X(t)$ is upper 2θ-standarizing, we have for any $t \geq s$,*

$$\mathbb{E}\big[e^{\theta Y(t)} \mid \mathcal{F}_s\big] \leq e^{-\theta^2(t-s)} \cdot e^{\theta Y(s)} + \frac{2e^{2\theta}}{\theta^2}, \tag{3.17}$$

and for any λ satisfying $(1-2\theta)e^{\lambda} < \lambda/2$,

$$\mathbb{E}\big[e^{\lambda Y(t)} \mid \mathcal{F}_s\big] < e^{-\frac{\lambda}{2}(t-s)} \cdot e^{\lambda Y(s)} + \frac{2e^{2\lambda}}{1-e^{-\lambda/2}}. \tag{3.18}$$

2. If $X(t)$ is lower 2θ-standarizing, we have for any $t \geq s$,

$$\mathbb{E}\big[e^{-\theta Y(t)} \mid \mathcal{F}_s\big] \leq e^{-\theta^2(t-s)} \cdot e^{-\theta Y(s)} + \frac{2e^{\theta}}{\theta^2}. \tag{3.19}$$

3. If $X(t)$ is 2θ-standarizing, we have for any $t \geq s$,

$$\mathbb{E}\big[e^{\theta|Y(t)|} \mid \mathcal{F}_s\big] \leq e^{-\theta^2(t-s)} \cdot e^{\theta|Y(s)|} + \frac{6e^{2\theta}}{\theta^2}. \tag{3.20}$$

Proof. We only prove (3.17) and inequalities (3.18), (3.19), (3.20) can be proved in a similar manner. Lemma 3.6 yields that for any $k \in \mathbb{N}$,

$$\mathbb{E}\left[\left(e^{\theta Y(s+k)} - \frac{e^{2\theta}}{1-e^{-\theta^2}}\right) e^{\theta^2(s+k)} \;\middle|\; \mathcal{F}_{s+k-1}\right] \leq \left(e^{\theta Y(s+k-1)} - \frac{e^{2\theta}}{1-e^{-\theta^2}}\right) e^{\theta^2(s+k-1)}.$$

Hence, $\left\{\left(e^{\theta Y(s+k)} - \frac{e^{2\theta}}{1-e^{-\theta^2}}\right) e^{\theta^2(s+k)}\right\}_{k\in\mathbb{N}}$ is a supermartingale and for any $k \in \mathbb{N}$, we have

$$\mathbb{E}\big[e^{\theta Y(s+k)} \mid \mathcal{F}_s\big] \leq e^{-\theta^2 k} \cdot e^{\theta Y(s)} + \frac{e^{2\theta}(1-e^{-\theta^2 k})}{1-e^{-\theta^2}}. \tag{3.21}$$

For any $t \geq s$, we have

$$\begin{aligned}
\mathbb{E}\big[e^{\theta Y(t)} \mid \mathcal{F}_s\big] &= \mathbb{E}\big[\mathbb{E}\big[e^{\theta Y(t)} \mid \mathcal{F}_{s+\lfloor t-s\rfloor}\big] \mid \mathcal{F}_s\big] \\
&\leq e^{-\theta^2(t-s-\lfloor t-s\rfloor)} \cdot \mathbb{E}\big[e^{\theta Y(s+\lfloor t-s\rfloor)} \mid \mathcal{F}_s\big] + e^{2\theta} \\
&\leq e^{-\theta^2(t-s)} \cdot e^{\theta Y(s)} + \frac{2e^{2\theta}}{1-e^{-\theta^2}} \\
&\leq e^{-\theta^2(t-s)} \cdot e^{\theta Y(s)} + \frac{2e^{2\theta}}{\theta^2}.
\end{aligned}$$

In the first inequality, we use Lemma 3.6, and in the second inequality, we use (3.21). The last inequality follows from $e^{-x} > 1 - x$. □

Corollary 3.8. *We denote by $\{X_i(t)\}_{i\in[n]}$ independent 2θ-standarizing point processes with initial values $\{X_i(0)\}_{i\in[n]}$ such that $|X_i(0)| \leq L$ for all $i \in [n]$. For all $t \geq L/\theta$, we have*

$$\mathbb{E}e^{\theta|X_i(t)-t|} \leq \frac{20}{\theta^2}. \tag{3.22}$$

Write $Y(t) = \frac{1}{n}\sum_{i=1}^n X_i(t) - t$. For all $t \geq L/\theta$, we have

$$\mathbb{E}e^{\theta|Y(t)|} \leq \frac{20}{\theta^2} \quad \textit{and} \quad \mathbb{E}e^{\theta n|Y(t)|} \leq \left(\frac{20}{\theta^2}\right)^n. \tag{3.23}$$

In addition, for $0 \leq t < L/\theta$, we have

$$\mathbb{E}e^{\theta|X_i(t)-t|} \leq e^{\theta L} + \frac{20}{\theta^2} \quad \textit{and} \quad \mathbb{E}e^{\theta|Y(t)|} \leq e^{\theta L} + \frac{20}{\theta^2}. \tag{3.24}$$

Proof. Inequality (3.20) and the assumption that $|X_i(0)| \le L$ imply that

$$\mathbb{E}e^{\theta|X_i(t)-t|} \le e^{-\theta^2 t+\theta L} + \frac{6e^{2\theta}}{\theta^2}.$$

For $t \ge L/\theta$, the RHS of the above inequality is at most $1+\frac{6e^{2\theta}}{\theta^2} \le \frac{20}{\theta^2}$; for $0 \le t < L/\theta$, it can be trivially bounded above by $e^{\theta L}+\frac{20}{\theta^2}$. This proves inequality (3.22) and the first inequality of (3.24). Then we can use inequality (3.22) to obtain for $t \ge L/\theta$ that

$$\mathbb{E}e^{\theta|Y(t)|} \le \mathbb{E}e^{\frac{\theta}{n}\sum_{i=1}^n |X_i(t)-t|} = \left(\prod_{i=1}^n \mathbb{E}e^{\theta|X_i(t)-t|}\right)^{1/n} \le \frac{20}{\theta^2},$$

and

$$\mathbb{E}e^{\theta n|Y(t)|} \le \mathbb{E}e^{\theta\sum_{i=1}^n |X_i(t)-t|} = \prod_{i=1}^n \mathbb{E}e^{\theta|X_i(t)-t|} \le \left(\frac{20}{\theta^2}\right)^n.$$

Similarly, we can use the first inequality of (3.24) to obtain the second inequality of (3.24). □

Consider a collection independent regular point processes $\{X_i(t)\}_{i\in[n]}$ with the initial value $\{L_i(0)\}_{i\in[n]}$ and conditional intensity functions $\{\lambda_i(t)\}_{i\in[n]}$ given in (2.9). The process $\{Z_k\}_{k\in\mathbb{N}}$ defined in (2.10) is the output of the θ-drift strategy f as per Section 2.3. We show the following concentration bounds on the load vector $\{L_i^f(m)\}_{i\in[n]}$.

Lemma 3.9. *Suppose that $|L_i(0)| \le L$ for all $i \in [n]$. Set $\theta = 1/5$. The θ-drift strategy f satisfies that for any $m \ge (\frac{3L}{\theta}+\frac{10}{\theta}\log\frac{80}{\theta^2})n$, any $i \in [n]$ and any $k > 0$,*

$$\mathbb{P}\left(|L_i^f(m)| > k\right) \le \frac{320}{\theta^2}\exp\left(-\frac{\theta k}{5}\right). \tag{3.25}$$

Taking the union bound, we have

$$\mathbb{P}\left(\max_{i\in[n]}|L_i^f(m)| > k+\frac{5}{\theta}\log\frac{320n}{\theta^2}\right) \le \exp\left(-\frac{\theta k}{5}\right). \tag{3.26}$$

Proof. Set $t^* = m/n + k/2$ and $t_* = \max(m/n - k/2, 0)$. We denote by $E = \{X(t^*) \ge m\}$ and $F = \{X(t_*) \le m\}$. Using the law of total probability, we obtain

$$\begin{aligned}\mathbb{P}\left(|L_i^f(m)| > k\right) &= \mathbb{P}\left(L_i^f(m) > k\right) + \mathbb{P}\left(L_i^f(m) < -k\right)\\ &\le \mathbb{P}\left(L_i^f(m) > k, E\right) + \mathbb{P}(E^c) + \mathbb{P}\left(L_i^f(m) < -k, F\right) + \mathbb{P}(F^c).\end{aligned} \tag{3.27}$$

We now estimate the first two terms of (3.27). Since $X_i(t)$ given in (2.9) is θ-standardizing, we apply the first inequality of (3.23) and Markov's inequality to obtain

$$\mathbb{P}(E^c) = \mathbb{P}\left(\frac{X(t^*)}{n} < t^* - \frac{k}{2}\right) \le e^{-\frac{\theta k}{4}}\cdot\mathbb{E}\exp\left(\frac{\theta}{2}\Big|\frac{X(t^*)}{n} - t^*\Big|\right) \le \frac{80}{\theta^2}\exp\left(-\frac{\theta k}{4}\right). \tag{3.28}$$

Whenever E occurs, we have $L_i^f(m) \le X_i(t^*) - m/n$. This, together with inequality (3.22) and Markov's inequality, yields

$$\begin{aligned}\mathbb{P}\left(L_i^f(m) > k, E\right) &\le \mathbb{P}\left(X_i(t^*) > \frac{m}{n} + k\right) = \mathbb{P}\left(X_i(t^*) > t^* + \frac{k}{2}\right)\\ &\le e^{-\frac{\theta k}{4}}\cdot\mathbb{E}\exp\left(\frac{\theta}{2}|X_i(t^*) - t^*|\right) \le \frac{80}{\theta^2}\exp\left(-\frac{\theta k}{4}\right).\end{aligned} \tag{3.29}$$

We next estimate the last two terms of (3.27). We first estimate $\mathbb{P}(F^c)$. For $k \geq 2m/n$ we have $t_* = 0$ and $X(t_*) = \sum_{i\in[n]} L_i(0) = 0$. This yields $\mathbb{P}(F^c) = 0$. For $k < 2m/n$, we use the fact that $t_* = m/n - k/2$ to rewrite $F = \{X(t_*)/n \leq t_* + k/2\}$. Set $k_0 = 2m/n - 4L/\theta$. One can check that $t_* > 2L/\theta$ for $0 < k < k_0$ and that $0 < t_* < 2L/\theta$ for $k_0 < k < 2m/n$. We apply the first inequality of (3.23), the second inequality of (3.24) and Markov's inequality to obtain

$$\begin{aligned}
\mathbb{P}(F^c) = \mathbb{P}\left(\frac{X(t_*)}{n} > t_* + \frac{k}{2}\right) &\leq e^{-\frac{\theta k}{4}} \cdot \mathbb{E}\exp\left(\frac{\theta}{2}\Big|\frac{X(t_*)}{n} - t_*\Big|\right) \\
&\leq \begin{cases} \frac{80}{\theta^2} e^{-\theta k/4}, & 0 < k \leq k_0 \\ \left(e^{\theta L/2} + \frac{80}{\theta^2}\right) e^{-\theta k/4}, & k_0 < k < 2m/n \end{cases} \\
&\leq \begin{cases} \frac{80}{\theta^2} e^{-\theta k/4}, & 0 < k \leq k_0, \\ e^{-\theta k/5}, & k_0 < k < 2m/n, \end{cases}
\end{aligned} \tag{3.30}$$

where the second case of inequality (3.30) follows from $e^{\theta L/2} + 80/\theta^2 \leq e^{\theta k_0/20} < e^{\theta k/20}$. To see this, we observe that our assumption on m and our choice of $\theta = 1/5$ imply that

$$\frac{2L}{\theta} + \frac{10}{\theta}\log\left(e^{\theta L/2} + \frac{80}{\theta^2}\right) \leq \frac{2L}{\theta} + \frac{10}{\theta}\log\left(e^{\theta L/2}\right) + \frac{10}{\theta}\log\frac{80}{\theta^2} = \frac{3L}{\theta} + \frac{10}{\theta}\log\frac{80}{\theta^2} \leq \frac{m}{n}.$$

This can be rewritten as $k_0/2 \geq \frac{10}{\theta}\log\left(e^{\theta L/2} + \frac{80}{\theta^2}\right)$, which is equivalent to the desired statement.

We now estimate the third term of (3.27). For $k \geq 2m/n$, we derive from the assumption on m that

$$L_i^f(m) \geq -L - \frac{m}{n} > -\frac{4m}{3n} > -k.$$

In this case, we have $\mathbb{P}\big(L_i^f(m) < -k, F\big) = 0$. We now deal with the case that $k < 2m/n$. Whenever F occurs, we have $L_i^f(m) \geq X_i(t_*) - m/n$. Together with $t_* = m/n - k/2$, this yields

$$\mathbb{P}\left(L_i^f(m) < -k, F\right) \leq \mathbb{P}\left(X_i(t_*) < \frac{m}{n} - k\right) = \mathbb{P}\left(X_i(t_*) < t_* - \frac{k}{2}\right).$$

Recall that $k_0 = 2m/n - 4L/\theta$ and the fact that $t_* > 2L/\theta$ for $0 < k < k_0$ and that $t_* < 2L/\theta$ for $k_0 < k < 2m/n$. We apply inequality (3.22), the first inequality of (3.24) and Markov's inequality to obtain

$$\begin{aligned}
\mathbb{P}\left(L_i^f(m) < -k, F\right) &\leq e^{-\frac{\theta k}{4}} \cdot \mathbb{E}\exp\left(\frac{\theta}{2}|X_i(t_*) - t_*|\right) \\
&\leq \begin{cases} \frac{80}{\theta^2} e^{-\theta k/4}, & 0 < k \leq k_0 \\ \left(e^{\theta L/2} + \frac{80}{\theta^2}\right) e^{-\theta k/4}, & k_0 < k < 2m/n \end{cases} \\
&\leq \begin{cases} \frac{80}{\theta^2} e^{-\theta k/4}, & 0 < k \leq k_0, \\ e^{-\theta k/5}, & k_0 < k < 2m/n, \end{cases}
\end{aligned} \tag{3.31}$$

where the second case of inequality (3.31) again uses $e^{\theta L/2} + 80/\theta^2 \leq e^{\theta k_0/20} < e^{\theta k/20}$.

Combining (3.27)-(3.31), we obtain (3.25). □

Lemma 3.10. *Suppose that $|L_i(0)| \le L$ for all $i \in [n]$. Set $\theta = 1/5$ and $k_0 = 1 + \frac{2}{\theta}\log\frac{80}{\theta^2}$. The θ-drift strategy f satisfies that for any $m \ge 2nL/\theta$ and any $k \ge 3k_0$,*

$$\mathbb{P}\left(\left|\left\{i \in [n] : L_i^f(m) > k\right\}\right| \ge \frac{160}{\theta^2} n e^{-\frac{\theta k}{3}}\right) \le 2\exp\left(-2n\left(\frac{80}{\theta^2}\right)^2 e^{-\frac{2\theta k}{3}}\right). \tag{3.32}$$

Proof. Set $t^* = m/n + k_0$. Let $E = \{X(t^*) \ge m\}$. Denote $S_k = \left\{i \in [n] : L_i^f(m) \ge k\right\}$. By the law of total probability, we have

$$\mathbb{P}\left(|S_k| \ge \frac{160}{\theta^2} e^{-\frac{\theta k}{4}}\right) \le \mathbb{P}\left(|S_k| \ge \frac{160}{\theta^2} e^{-\frac{\theta k}{4}}, E\right) + \mathbb{P}(E^c). \tag{3.33}$$

The second inequality of (3.23), Markov's inequality and our choice of k_0 yield

$$\mathbb{P}(E^c) = \mathbb{P}(X(t^*) < nt^* - nk_0) \le \left(\frac{80}{\theta^2}\right)^n \exp\left(-\frac{n\theta k_0}{2}\right) = \exp\left(-\frac{\theta n}{2}\right). \tag{3.34}$$

To estimate the first term in (3.33), we introduce independent Bernoulli random variables W_i, which are indicator functions of the events that $X_i(t^*) > m/n + k$. Hence,

$$\begin{aligned}\mathbb{P}(W_i = 1) &= \mathbb{P}\left(X_i(t^*) > \frac{m}{n} + k\right) = \mathbb{P}\left(X_i(t^*) > t^* + k - k_0\right)\\ &\le \mathbb{P}\left(X_i(t^*) > t^* + \frac{2k}{3}\right) \le \frac{80}{\theta^2}\exp\left(-\frac{\theta k}{3}\right),\end{aligned}$$

where in the first inequality, we use the assumption that $k \ge 3k_0$, and in the second inequality, we use the fact that $X_i(t)$ is θ-standarizing and (3.22). Observe that, when the event E occurs, we have $L_i^f(m) \le X_i(t^*) - m/n$, which implies that $|S_k| \le \sum_{i=1}^n W_i$. This, together with Hoeffding's inequality, yields

$$\mathbb{P}\left(|S_k| \ge \frac{160}{\theta^2} n e^{-\frac{\theta k}{3}}, E\right) \le \mathbb{P}\left(\sum_{i=1}^n W_i \ge \frac{160}{\theta^2} n e^{-\frac{\theta k}{3}}\right) \le \exp\left(-2n\left(\frac{80}{\theta^2}\right)^2 e^{-\frac{2\theta k}{3}}\right).$$

This, along with (3.33) and (3.34), gives

$$\mathbb{P}\left(|S_k| \ge \frac{160}{\theta^2} e^{-\frac{\theta k}{4}}\right) \le \exp\left(-2n\left(\frac{80}{\theta^2}\right)^2 e^{-\frac{2\theta k}{3}}\right) + \exp\left(-\frac{\theta n}{2}\right).$$

This, together with the condition that $k \ge 3 + \frac{6}{\theta}\log\frac{80}{\theta^2}$, yields (3.32). □

We also provide a concentration bound on the time it takes the drift strategy to bring certain quantities close to stationarity.

Lemma 3.11. *Suppose that $|L_i(0)| \le L$ for all $i \in [n]$. Set $\theta = 1/5$. Denote*

$$\begin{aligned}A_m &= \left\{\max_{i\in[n]} \left|L_i^f(m)\right| \le k_a + \frac{5}{\theta}\log\frac{320n}{\theta^2}\right\},\\ B_m &= \left\{\left|\left\{i \in [n] : L_i^f(m) > k_b\right\}\right| < \frac{160}{\theta^2} n e^{-\frac{\theta k_b}{3}}\right\},\end{aligned}$$

and assume that

$$\exp\left(-\frac{\theta k_a}{5}\right) + 2\exp\left(-2n\left(\frac{80}{\theta^2}\right)^2 e^{-\frac{2\theta k_b}{3}}\right) < \frac{1}{2}.$$

If $T = \min\big\{m \in \mathbb{N} : A_m \cap B_m \text{ holds}\big\}$*, then under the* θ*-drift strategy, we have*

$$\mathbb{E}(T) < Cn(L + \log n)$$

for some absolute constant $C > 0$ *and all large enough* n*.*

Proof. Set $m_0 = 0$ and recursively define

$$m_{j+1} = m_j + \left\lceil \tfrac{3n}{\theta} \max_{i\in[n]} \left|L_i^f(m_j)\right| + \frac{10n}{\theta}\log\frac{80}{\theta^2}\right\rceil.$$

Denote

$$J = \min\{j : A_{m_j} \cap B_{m_j}\}.$$

It is obvious that $T \le m_J$. By Lemma 3.9, Lemma 3.10 and the union bound, we have, conditioned on the history of the process until m_j balls have been allocated, that

$$\mathbb{P}\Big(A^c_{m_{j+1}} \cup B^c_{m_{j+1}} \mid \mathcal{F}_{m_j}\Big) \le \exp\left(-\frac{\theta k_a}{5}\right) + 2\exp\left(-2n\left(\frac{80}{\theta^2}\right)^2 e^{-\frac{2\theta k_b}{3}}\right) < \frac{1}{2}.$$

Thus, we have $\mathbb{P}(J > j) \le 2^{-j}$ and hence

$$\mathbb{E}(J) \le 2.$$

For $j \ge 1$, we have by Lemma 3.9 that

$$\mathbb{P}\left(\max_{i\in[n]} \left|L_i^f(m_j)\right| > k_a + \frac{5}{\theta}\log\frac{320n}{\theta^2} \;\Big|\; \mathcal{F}_{m_{j-1}}\right) \le \exp\left(-\frac{\theta k_a}{5}\right),$$

which implies that

$$\mathbb{E}\big(m_{j+1} - m_j \mid \mathcal{F}_{m_{j-1}}\big) \le \left\lceil \frac{3n}{\theta}\left(\frac{5}{\theta}\log\frac{320n}{\theta^2} + \frac{1}{1-e^{-\theta/4}}\right) + \frac{10n}{\theta}\log\frac{80}{\theta^2}\right\rceil.$$

Putting all these together, we obtain

$$\mathbb{E}(T) \le \mathbb{E}(m_J) \le Cn(L + \log n)$$

for some $C > 0$ and n large enough. □

4 Single-time load discrepancy: upper bound

In this section, we investigate two-thinning strategies that can achieve the upper bounds on the single-time load discrepancy as stated in Theorem 1. Write $t = m/n$. Observe that for any thinning strategy f and any $m \in \mathbb{N}$,

$$\mathrm{MaxLoad}^f(\lfloor t\rfloor n) - 1 \le \mathrm{MaxLoad}^f(m) \le \mathrm{MaxLoad}^f(\lceil t\rceil n) + 1. \tag{4.1}$$

Hence, at the expense of an additive constant to the maximum load, we can always assume that m is divisible by n, and then it suffices to study $\mathrm{MaxLoad}^f(tn)$ for $t \in \mathbb{N}$.

4.1 Case 1: $t \le O(\sqrt{\log n})$

In this case, we apply the $(t+\ell)$-threshold strategy introduced in [11] (see Section 2.3). Recall that this strategy retries a ball if its primary allocation is a bin which has accepted at least $t+\ell$ primary allocations.

Proposition 4.1. *Assume that $L_i(0) = 0$ for all $i \in [n]$ and that $t \le e^{-9}\sqrt{\log n}$. We set $\ell = \sqrt{\frac{3\log n}{\log\log n - 2\log t}}$. For any $\varepsilon > 0$ and sufficiently large n, the $(t+\ell)$-threshold strategy f satisfies*

$$\mathbb{P}\left(\mathrm{MaxLoad}^f(tn) > (2+\varepsilon)\ell\right) < 3n^{-\varepsilon}. \tag{4.2}$$

Proof. We write $r := R_{nt}$ for the total number of retries throughout the process. The strategy f guarantees that no bins accept more than $t+\ell$ primary allocations, i.e., $L^f_{1,i}([tn]) \le t+\ell$. This, together with the equation $L^f_i(tn) = L^f_{1,i}([tn]) + L^f_{2,i}([tn]) - t$, implies that

$$\mathbb{P}\left(\mathrm{MaxLoad}^f(tn) > (2+\varepsilon)\ell\right) \le \mathbb{P}\left(\max_{i\in[n]} L^f_{2,i}([tn]) > (1+\varepsilon)\ell\right), \tag{4.3}$$

where $L^f_{2,i}([tn])$ defined in (2.3) represents the number balls that bin i receives from secondary allocations. Set $r^* = 6ne^{-tI(\ell/t)}/\log(1+\ell/t)$. By the law of total probability, we have

$$\mathbb{P}\left(\max_{i\in[n]} L^f_{2,i}([tn]) > (1+\varepsilon)\ell\right) \le \mathbb{P}\left(\max_{i\in[n]} L^f_{2,i}([tn]) > (1+\varepsilon)\ell, r \le r^*\right) + \mathbb{P}(r > r^*). \tag{4.4}$$

First, we estimate the second term of (4.4). We write $\{X_i\}_{i\in[n]}$ for independent Poisson(t) random variables. Define $Y_i = \max\{0, X_i - t - \ell\}$ and $Y = \sum_{i=1}^n Y_i$. Lemmata 3.1 and 3.5 provide the following tail bound

$$\begin{aligned}\mathbb{P}(r > r^*) &\le 2\mathbb{P}(Y > r^*) < 2\exp\left(-ne^{-tI(\ell/t)}\right)\\ &< 2\exp\left(-n\left(\frac{et}{\ell}\right)^{3\ell}\right) = \exp\left(-n^{1-o(1)}\right),\end{aligned} \tag{4.5}$$

where the last inequality follows from the upper bound in (3.4) and the fact that $\ell \ge 4t$ for large enough n.

Next, we estimate the first term of (4.4). Again, using the lower bound in (3.4), we obtain $r^* < 6n(et/\ell)^\ell$ for n large enough. Set $\lambda = 6(et/\ell)^\ell$. We denote by $\{W_i\}_{i\in[n]}$ independent Poisson(λ) random variables. Lemma 3.1 and the union bound argument yield

$$\begin{aligned}\mathbb{P}\left(\max_{i\in[n]} L^f_{2,i}([tn]) > (1+\varepsilon)\ell, r \le r^*\right) &\le \mathbb{P}\left(\max_{i\in[n]} |\{s \le r^* \ :\ Z^2_s = i\}| > (1+\varepsilon)\ell\right)\\ &\le 2\mathbb{P}\left(\max_{i\in[n]} W_i > (1+\varepsilon)\ell\right)\\ &\le 2n\mathbb{P}(W_1 > (1+\varepsilon)\ell).\end{aligned} \tag{4.6}$$

Apply Lemma 3.3 and the lower bound of $I(x)$ in (3.4) to obtain

$$\mathbb{P}(W_1 > (1+\varepsilon)\ell) \le e^{-\lambda I((1+\varepsilon)\ell/\lambda)} < \left(\frac{6e}{(1+\varepsilon)\ell}\left(\frac{et}{\ell}\right)^\ell\right)^{(1+\varepsilon)\ell} < \left(\frac{et}{\ell}\right)^{(1+\varepsilon)\ell^2}. \tag{4.7}$$

One can check that

$$\left(\frac{et}{\ell}\right)^{(1+\varepsilon)\ell^2} = \exp\left(-(1+\varepsilon)\cdot\frac{3}{2}\left(1-\frac{\log(\log\log n-2\log t)+2-\log 3}{\log\log n-2\log t}\right)\log n\right).$$

Our assumption of t yields that $\log\log n-2\log t\geq 18$. This, together with the fact that $x^{-1}\log x$ is decreasing for $x>e$, yields that

$$\frac{\log(\log\log n-2\log t)+2-\log 3}{\log\log n-2\log t}\leq\frac{\log(18)+2-\log 3}{18}<\frac{1}{3}.$$

Hence, we obtain

$$\left(\frac{et}{\ell}\right)^{(1+\varepsilon)\ell^2}\leq n^{-(1+\varepsilon)}.$$

This, combined with (4.6), (4.7), yields

$$\mathbb{P}\left(\max_{i\in[n]}L^f_{2,i}([tn])>(1+\varepsilon)\ell, r\leq r^*\right)<2n^{-\varepsilon}. \tag{4.8}$$

The desired statement (4.2) follows from (4.3), (4.4), (4.5) and (4.8). □

Our next result complements the proof of the case $t\leq O(\sqrt{\log n})$. Moreover, it also provides a tight upper bound for the maximum load for $t=(\log n)^{1/2+o(1)}$.

Proposition 4.2. *Assume that $L_i(0)=0$ for all $i\in[n]$ and that $\Omega(\log^{1/2}n)\leq t\leq o(\log^2 n)$. We set $\ell=(ct\log n)^{1/3}$, where c is an absolute constant such that $\ell\leq t$. For any $\varepsilon>0$ and sufficiently large n, the $(t+\ell)$-threshold strategy f satisfies*

$$\mathbb{P}\left(\mathrm{MaxLoad}^f(tn)>\left(\frac{4(1+\varepsilon)}{c}+1\right)\ell\right)<3n^{-\varepsilon}. \tag{4.9}$$

Proof. We slightly modify the proof of Proposition 4.1. Set $r^*=6ne^{-tI(\ell/t)}/\log(1+\ell/t)$. As before, we define independent random variables $\{X_i\}_{i\in[n]}$, $\{Y_i\}_{i\in[n]}$ and $\{W_i\}_{i\in[n]}$ where $X_i\sim\mathrm{Poisson}(t)$, $Y_i=\max\{0,X_i-t-\ell\}$ and $W_i\sim\mathrm{Poisson}(\lambda)$ for $\lambda=r^*/n$. As before, we set $r:=R_{nt}$. Similar to (4.3), (4.4) and (4.6), we have

$$\begin{aligned}
\mathbb{P}\left(\mathrm{MaxLoad}^f(tn)>\left(\tfrac{4(1+\varepsilon)}{c}+1\right)\ell\right)&\leq\mathbb{P}\left(\max_{i\in[n]}L^f_{2,i}([tn])>\frac{4(1+\varepsilon)\ell}{c}\right)\\
&\leq\mathbb{P}\left(\max_{i\in[n]}L^f_{2,i}([tn])>\frac{4(1+\varepsilon)\ell}{c}, r<r^*\right)+\mathbb{P}(r>r^*)\\
&\leq 2\mathbb{P}\left(\max_{i\in[n]}W_i>\frac{4(1+\varepsilon)\ell}{c}\right)+\mathbb{P}(r>r^*)\\
&\leq 2n\mathbb{P}\left(W_1>\frac{4(1+\varepsilon)\ell}{c}\right)+\mathbb{P}(r>r^*).
\end{aligned}\tag{4.10}$$

Similar to (4.5), Lemmata 3.1 and 3.5 yield that

$$\begin{aligned}
\mathbb{P}(r>r^*)&<2\exp\left(-ne^{-tI(\ell/t)}\right)<2\exp\left(-n\exp\left(-\frac{\ell^2}{2t}\right)\right)\\
&<2\exp\big(-ne^{-\ell}\big)=\exp\big(-n^{1-o(1)}\big),
\end{aligned}\tag{4.11}$$

where the last two inequalities follow from the upper bound of $I(x)$ in (3.3) and the fact that $\ell \le t$. Using the lower bound of $I(x)$ in (3.3) and $\log(1+x) > x/2$ for $0 < x < 1$, one can check that $\lambda = r^*/n < \frac{12t}{\ell} \exp\left(-\frac{\ell^2}{4t}\right) = o(1)$. This, together with Lemma 3.3 and inequality $I(x) > x\log(x/e)$ for $x > 4$, yields

$$\begin{aligned}\mathbb{P}\left(W_1 > \frac{4(1+\varepsilon)\ell}{c}\right) &\le \exp\left(-\lambda I\left(\frac{3(1+\varepsilon)\ell}{c\lambda}\right)\right) \le \left(\frac{ce\lambda}{3(1+\varepsilon)\ell}\right)^{\frac{3(1+\varepsilon)\ell}{c}} \\ &\le \left(\frac{4cet}{(1+\varepsilon)\ell^2}\exp\left(-\frac{\ell^2}{3t}\right)\right)^{\frac{3(1+\varepsilon)\ell}{c}} \\ &< \exp\left(-\frac{(1+\varepsilon)\ell^3}{ct}\right) = n^{-(1+\varepsilon)}.\end{aligned}$$

Combining this with (4.10) and (4.11), we can obtain (4.9). □

4.2 Case 2: $\Omega(\sqrt{\log n}) \le t \le O(\log n)$

For $t = O\big((\log n)^{\frac{1}{2}+\frac{1}{\sqrt{\log\log\log n}}}\big)$, Theorem 1 follows from Proposition 4.2. Thus, here we treat

$$\Omega\big((\log n)^{\frac{1}{2}+\frac{1}{\sqrt{\log\log\log n}}}\big) \le t \le O(\log n).$$

In this subsection, we study the allocation problem in a more general setting. The initial loads are not necessarily perfectly balanced (i.e., allowing $L_i(0) \neq 0$). This will play an important role in Sections 4.3 and 8.

Recall that $k = \left\lfloor \frac{\log\log n}{3\log\log\log n} \right\rfloor$. Set $\ell = \lfloor \log^{\beta_k} n \rfloor$, where β_k is defined in Section 2.4. One can check that $\ell = \left\lfloor (\log n)^{\frac{1}{2}+\left(2-\frac{1}{2k+1}\right)\frac{\alpha+\eta-1/2}{2k+1}} \right\rfloor$. Then we have the following result.

Proposition 4.3. *Let $t > 0$ and $\alpha = \frac{\log t}{\log\log n}$ satisfying $\alpha \in \left[\frac{1}{2} + \frac{1}{\sqrt{\log\log\log n}}, 1 + \frac{\sqrt{\log\log\log n}}{\log\log n}\right]$. Suppose that for $L_0 \ge 0$ the following conditions hold:*

1. $\mathrm{MaxLoad}(0) < ct$ *for some constant* $0 < c < 1$,
2. $|H_0| \le 3n\exp\left(-\frac{\ell^2}{4\log^{\alpha+\eta} n}\right)$, *where* $H_0 = \{i \in [n] : L_i(0) > L_0\}$ *is the set of bins with load greater than* L_0.

Then the multi-stage (t, L_0, ℓ)-threshold strategy f (as defined in Section 2.4), with the parameters above, satisfies that

$$\mathbb{P}\left(\mathrm{MaxLoad}^f(tn) > L_0 + 2k\ell\right) \le n^{-e^{\sqrt{\log\log\log n}}}.$$

For $\omega(n\sqrt{\log n}) \le m \le O(\log n)$, Theorem 1 follows as an immediate consequence of the following corollary.

Corollary 4.4. *Let $t > 0$ and α as above, satisfying $\alpha \in \left[\frac{1}{2} + \frac{1}{\sqrt{\log\log\log n}}, 1 + \frac{\sqrt{\log\log\log n}}{\log\log n}\right]$. The multi-stage $(t, 0, \ell)$-threshold strategy f satisfies that*

$$\mathbb{P}\left(\mathrm{MaxLoad}^f(tn) > (\log n)^{\frac{1}{2}+o(1)}\right) \le n^{-e^{\sqrt{\log\log\log n}}}.$$

Proof. Apply Proposition 4.3 with $\eta = 0$ and $L(0) = L_0 = 0$ and observe that the two conditions of Proposition 4.3 trivially hold. Hence the corollary follows from the fact that $\ell = (\log n)^{\frac{1}{2}+o(1)}$ and $k = \log^{o(1)} n$. □

For $1 \le i \le k$, we denote by r_i be the number of retries in stage i of the multi-stage (t, L_0, ℓ)-threshold strategy. Recall our notation H_i for the set of bins in $(\cup_{j=0}^{i-1} H_j)^c$ whose loads after the i-th stage are at least $L_0 + 2i\ell$. To establish Proposition 4.3, we use the following lemma, to inductively bound the number of retries in every stage and the size of H_i, the set of heavily loaded bins.

Lemma 4.5. *Under the assumptions of Proposition 4.3, for all $1 \le i \le k$, we have*

$$\mathbb{P}(r_i > r_i^*) \le \exp\left(-n^{1/2-o(1)}\right), \tag{4.12}$$

where

$$r_i^* := \frac{20n \log^{\beta_{i-1}} n}{\ell} \exp\left(-\frac{\ell^2}{5 \log^{\beta_{i-1}} n}\right). \tag{4.13}$$

In addition, for $1 \le i \le k-1$, we have

$$\mathbb{P}\left(|H_i| > \frac{4n\lambda_i^\ell}{\ell!}\right) \le \exp\left(-n^{1/2-o(1)}\right), \tag{4.14}$$

where $\lambda_i := r_i^*/n$.

Before presenting the proof, we first make some technical observations. Using $\ell = \lfloor \log^{\beta_k} n \rfloor$ and $\beta_k = \beta - \frac{k(2\beta-1-\varepsilon)}{2k+1}$, it is easy to check that

$$\ell \cdot \log \ell \le (\log n)^{\frac{k(1+\varepsilon)+\beta}{2k+1}} \cdot \frac{k(1+\varepsilon)+\beta}{2k+1} \log\log n = (\log n)^{\frac{1}{2}+o(1)}. \tag{4.15}$$

For n large enough, we have

$$\begin{aligned} \frac{\ell^2}{\log^\beta n} &= \frac{\ell^2}{\log^{2\beta_k} n} \cdot (\log n)^{2\beta_k - \beta} > \frac{1}{2} (\log n)^{\frac{2k(1+\varepsilon)-(2k-1)\beta}{2k+1}} \\ &> \frac{1}{2} (\log n)^{\frac{1-(2k-1)(\beta-1)}{2k+1}} \ge (\log\log n)^{\frac{9}{8}-o(1)}, \end{aligned} \tag{4.16}$$

where the last equality follows from that $\beta - 1 < \frac{1/4+o(1)}{2k-1}$ and our choice of k. We also have

$$\begin{aligned} \frac{\ell^3}{\log^\beta n} &\le (\log n)^{2\beta - \frac{3k(2\beta-1-\varepsilon)}{2k+1}} = (\log n)^{1-\frac{2k-2}{2k+1}\left(\beta-\frac{1}{2}\right)+\frac{3k\varepsilon}{2k+1}} \\ &< (\log n)^{1-\frac{2k-2}{2k+1}\left(\beta-\frac{1}{2}\right)+\frac{3}{2k+1}\left(\beta-\frac{1}{2}\right)} \\ &= (\log n)^{1-\left(\beta-\frac{1}{2}\right)} = o(\log n), \end{aligned} \tag{4.17}$$

where the last equality uses $\beta \ge \alpha > \frac{1}{2} + \frac{1}{\sqrt{\log\log\log n}}$. For $1 \le i \le k$ and n large enough, we have

$$0 < \lambda_i < 1 \tag{4.18}$$

To see this, notice that $\{\beta_i\}_{i=0}^k$ is a decreasing arithmetic progression, hence, $\{\lambda_i\}_{i=1}^k$ is a decreasing sequence and it suffices to show that $0 < \lambda_1 < 1$. Observe that

$$\lambda_1 = \frac{20\log^\beta n}{\ell}\exp\left(-\frac{\ell^2}{5\log^\beta n}\right) = \exp\left(-\frac{\ell^2}{5\log^\beta n} + \log\frac{20\log^\beta n}{\ell}\right),$$

and

$$\log\frac{\log^\beta n}{\ell} < \beta\log\log n < 2\log\log n.$$

This, together with (4.16), yields that $0 < \lambda_1 < 1$ and hence (4.18).

Proof. We prove (4.12) and (4.14) inductively by establishing the i-th case of (4.12) on condition that (4.14) holds for all $j < i$, and by establishing the i-th case of (4.14) on condition that (4.12) holds for the same i. The case $i = 1$ is treated separately.

Bounding $\mathbb{P}(r_i > r_i^*)$ assuming that $|H_j| \le 4n\lambda_j^\ell/\ell!$ for $j < i$. We denote by $r_{i,1}$ the number of balls in the i-th stage whose primary allocations are bins that, at the time of the allocation, already accepted $t_i - t_{i-1} + \ell$ primary allocations during stage i. We write $r_{1,2}$ for the number balls in the first stage whose primary allocations are bins from H_0, and write $r_{i,2}$ for $i \ge 2$, for the number of balls in the i-th stage whose primary allocations are bins from $\cup_{j=1}^{i-1} H_j$. By the definition of the strategy, we thus have $r_i \le r_{i,1} + r_{i,2}$.

Estimating $r_{i,1}$. Recall that $t_i = \lfloor t - \log^{\beta_i} n\rfloor$ for $1 \le i \le k-1$, $t_k = t$, $\ell = \lfloor\log^{\beta_k} n\rfloor$, where $\beta_i = \beta - \frac{(2\beta-1-\varepsilon)i}{2k+1}$, and observe that $\ell < t_i - t_{i-1}$ for $1 \le i \le k$. Also, recall that $I(x)$ defined in Lemma 3.3 is the rate function of the large deviation bound of a Poisson random variable. We have

$$\frac{3}{5}r_i^* > \frac{12n(t_i - t_{i-1})}{\ell}\exp\left(-\frac{\ell^2}{4(t_i - t_{i-1})}\right) \ge \frac{6ne^{-(t_i-t_{i-1})I(\ell/(t_i-t_{i-1}))}}{\log(1+\ell/(t_i-t_{i-1}))} =: r^*, \tag{4.19}$$

where the first inequality follows from the definition of r_i^* in (4.13), and the second inequality follows from the lower bound of $I(x)$ in (3.3) and that $\log(1+x) \ge x/2$ for $0 < x < 1$ (indeed $\ell/(t_i - t_{i-1}) < 1$). Define $Y_j^{(i)} = \max\{0, X_j^{(i)} - (t_i - t_{i-1} + \ell)\}$, where $\{X_j^{(i)}\}_{j\in[n]}$ is a collection of independent Poisson$(t_i - t_{i-1})$ random variables, and write $Y = \sum_{j=1}^n Y_j^{(i)}$. By Lemmata 3.1, 3.5 and inequality (4.19), we have

$$\begin{aligned}
\mathbb{P}\left(r_{i,1} > \frac{3}{5}r_i^*\right) &\le 2\mathbb{P}\left(Y > \frac{3}{5}r_i^*\right) \le 2\mathbb{P}\left(Y > r^*\right) \\
&\le 2\exp\left(-n\exp\left(-(t_i - t_{i-1})I\left(\frac{\ell}{t_i - t_{i-1}}\right)\right)\right) \\
&\le 2\exp\left(-n\exp\left(-\frac{\ell^2}{2(t_i - t_{i-1})}\right)\right) \\
&\le 2\exp\left(-ne^{-\ell}\right) = \exp\left(-n^{1-o(1)}\right),
\end{aligned} \tag{4.20}$$

where the last two inequalities follow from the upper bound of $I(x)$ in (3.3) and the fact that $\ell/(t_i - t_{i-1}) < 1$. The last identity follows from the fact that $\ell = (\log n)^{\frac{1}{2}+o(1)}$.

Estimating $r_{1,2}$. Here we estimate the number of balls in the first stage whose primary allocations are bins from H_0. Using the assumption $|H_0| \le 3n\exp\left(-\frac{\ell^2}{4\log^\beta n}\right)$, we have for n large enough

$$2(t_1 - t_0)|H_0| \le 6n(t_1 - t_0)\exp\left(-\frac{\ell^2}{4\log^\beta n}\right) \le \frac{6n\log^\beta n}{\ell}\exp\left(-\frac{\ell^2}{5\log^\beta n}\right) < \frac{2}{5}r_1^*, \tag{4.21}$$

where r_1^* is given in (4.13), and the second inequality follows from $t_1 - t_0 < \log^\beta n$ and the observation that $\ell = o\big(\exp\big(\frac{\ell^2}{\log^\beta n}\big)\big)$ by (4.16). We denote by $\big\{X_j^{(i)}\big\}_{j\in[n]}$ independent Poisson$(t_i - t_{i-1})$ random variables, and write Z for a Poisson random variable with the parameter $3n(t_1 - t_0)\exp\Big(-\frac{\ell^2}{4\log^\beta n}\Big)$. Lemmata 3.1, 3.3 and inequality (4.21) yield

$$\begin{aligned}
\mathbb{P}\left(r_{1,2} > \frac{2}{5}r_1^*\right) &\le 2\mathbb{P}\left(\sum_{j\in H_0} X_j^{(1)} > \frac{2}{5}r_1^*\right) \le 2\mathbb{P}\left(Z > \frac{2}{5}r_1^*\right) \\
&\le 2\mathbb{P}\left(Z > 6n(t_1 - t_0)\exp\left(-\frac{\ell^2}{4\log^\beta n}\right)\right) \\
&\le 2\exp\left(-n(t_1 - t_0)\exp\left(-\frac{\ell^2}{4\log^\beta n}\right)\right) \\
&\le 2\exp\left(-n(t_1 - t_0)e^{-\ell}\right) \\
&= \exp\big(-n^{1-o(1)}\big). \qquad (4.22)
\end{aligned}$$

This, together with the $i = 1$ case of (4.20), implies the base case of (4.12), i.e., $i = 1$.

Estimating $r_{i,2}$ for $i \ge 2$, assuming that $|H_j| \le \frac{4n\lambda_j^\ell}{\ell!}$ for $1 \le j \le i-1$. Recall that $r_{i,2}$ is the number of balls in stage i whose primary allocations are bins from $\cup_{j=1}^{i-1} H_j$. Again, write $\big\{X_j^{(i)}\big\}_{j\in[n]}$ for independent Poisson$(t_i - t_{i-1})$ random variables, and write Z for a Poisson random variable with parameter $\frac{4n(t_i - t_{i-1})}{\ell!}\sum_{j=1}^{i-1}\lambda_j^\ell$. Define $E = \big\{|H_j| \le \frac{4n\lambda_j^\ell}{\ell!}, 1 \le j \le i-1\big\}$. Lemmata 3.1 and 3.3 imply that

$$\begin{aligned}
\mathbb{P}\left(r_{i,2} > \frac{8n(t_i - t_{i-1})}{\ell!}\sum_{j=1}^{i-1}\lambda_j^\ell,\ E\right) &\le 2\mathbb{P}\left(\sum_{m\in\cup_{j=1}^{i-1}H_j} X_m^{(i)} > \frac{8n(t_i - t_{i-1})}{\ell!}\sum_{j=1}^{i-1}\lambda_j^\ell,\ E\right) \\
&\le 2\mathbb{P}\left(Z > \frac{8n(t_i - t_{i-1})}{\ell!}\sum_{j=1}^{i-1}\lambda_j^\ell\right) \\
&\le \exp\left(-\frac{n(t_i - t_{i-1})}{\ell!}\sum_{j=1}^{i-1}\lambda_j^\ell\right) \le \exp\left(-\frac{n\lambda_1^\ell}{\ell!}\right) \\
&\le \exp\left(-\exp\left(\log n - \frac{\ell^3}{\log^\beta n} - \ell\log\ell\right)\right) \\
&= \exp\left(-n^{1-o(1)}\right), \qquad (4.23)
\end{aligned}$$

where the penultimate transition uses the fact that $\lambda_1 = r_1^*/n > \exp\big(-\frac{\ell^2}{4\log^\beta n}\big)$, where r_1^* is given in (4.13), and the bound $\ell! < \ell^\ell$, and the last transition uses (4.15) and (4.17). Using the fact that $0 < \lambda_i < 1$ and that $k < \ell$, we have

$$\begin{aligned}
\frac{8n(t_i - t_{i-1})}{\ell!}\sum_{j=1}^{i-1}\lambda_j^\ell &< \frac{8kn\log^{\beta_{i-1}} n}{\ell!} < \frac{8n\log^{\beta_{i-1}} n}{(\ell-1)!} < \frac{8n\log^{\beta_{i-1}} n}{\ell}\exp\left(-\frac{\ell\log\ell}{2}\right) \\
&\le \frac{8n\log^{\beta_{i-1}} n}{\ell}\exp\left(-\frac{\ell^2}{5\log^{\beta_{i-1}} n}\right) = \frac{2}{5}r_i^*, \qquad (4.24)
\end{aligned}$$

where the penultimate inequality uses Stirling's approximation and the last inequality follows the fact that $\ell < \log^{\beta_{i-1}} n$. Combining (4.23) and (4.24), we have

$$\mathbb{P}\left(r_{i,2} > \frac{2}{5}r_i^*,\ E\right) = \exp\left(-n^{1-o(1)}\right).$$

This, together with (4.14) for $1 \le j \le i-1$, implies that for $2 \le i \le k$,

$$\mathbb{P}\left(r_{i,2} > \frac{2}{5}r_i^*\right) \le \mathbb{P}\left(r_{i,2} > \frac{2}{5}r_i^*,\ E\right) + \sum_{j=1}^{i-1}\mathbb{P}\left(|H_j| > \frac{4n\lambda_j^\ell}{\ell!}\right) = \exp\left(-n^{1/2-o(1)}\right).$$

This, combined with (4.20) and (4.22), yields

$$\mathbb{P}(r_i > r_i^*) \le \mathbb{P}\left(r_{i,1} > \frac{3}{5}r_i^*\right) + \mathbb{P}\left(r_{i,2} > \frac{2}{5}r_i^*\right) = \exp\left(-n^{1/2-o(1)}\right).$$

This concludes the proof of the i-th case of (4.13) condition on that (4.14) holds for $j < i$.

Bounding $\mathbb{P}(|H_i| > 2p_i n)$ assuming that $r_i \le r_i^*$ for $i \ge 1$. Recall that

$$H_i = \left\{j \in [n] : L_j^f(t_i n) \ge L_0 + 2i\ell\right\} \setminus \bigcup_{i'<i} H_{i'}^c.$$

Let $j \in H_i$. We have $L_j^f(t_{i-1}n) < L_0 + 2(i-1)\ell$ (otherwise we would have $j \in H_{i-1}$). Let us show that j must have received at least ℓ secondary allocations in the i-th stage. During the i-th stage, if bin j accepted less than $t_i - t_{i-1} + \ell$ primary allocations, it clearly must have received at least ℓ secondary allocations in order to belong to H_i. Otherwise, once j accepted more than $t_i - t_{i-1} + \ell$ primary allocations (in the i-th stage), it rejects all further allocations unless its load is at most $-\log n$. Hence its load after accepting the last primary allocation must have been at most $-\log n$, so that in order to belong to H_i it must have received at least $L_0 + 2i\ell + \log n$ secondary allocations.

Let $\left\{X_j^{(i)}\right\}_{j\in[n]}$ be independent Poisson(λ_i) random variables. Let $Y_j^{(i)}$ be the indicator function of the event that $X_j^{(i)} \ge \ell$. Then, $\left\{Y_j^{(i)}\right\}_{j\in[n]}$ are independent Bernoulli(p_i) random variables, where $p_i = \mathbb{P}\big(X_1^{(i)} \ge \ell\big)$. Let $Y = \sum_{j=1}^n Y_j^{(i)}$. By Lemma 3.1 and Hoeffding's inequality,

$$\mathbb{P}(|H_i| > 2p_i n, r_i \le r_i^*) \le 2\mathbb{P}\left(Y > 2p_i n\right) \le 2e^{-2np_i^2}. \tag{4.25}$$

Using the fact that $0 < \lambda_i < 1$, we have

$$\frac{\lambda_i^\ell}{e\ell!} < p_i = e^{-\lambda_i}\sum_{j=\ell}^\infty \frac{\lambda_i^j}{j!} < \frac{2\lambda_i^\ell}{\ell!}.$$

This, together with (4.25), yields that, for $1 \le i \le k-1$,

$$\mathbb{P}\left(|H_i| > \frac{4n\lambda_i^\ell}{\ell!}, r_i \le r_i^*\right) < 2\exp\left(-\frac{2n}{e^2}\left(\frac{\lambda_i^\ell}{\ell!}\right)^2\right). \tag{4.26}$$

Since $\{\lambda_i\}_{i=1}^k$ is a decreasing sequence, we will upper bound the RHS of (4.26) for $i = k-1$. Using the fact that $\ell! \le e\sqrt{\ell}(\ell/e)^\ell$, by Stirling's approximation, we obtain

$$\frac{2n}{e^2}\left(\frac{\lambda_{k-1}^\ell}{\ell!}\right)^2 \ge \frac{2n}{e^4\ell}\left(\frac{e\lambda_{k-1}}{\ell}\right)^{2\ell} = \frac{2}{e^4}\exp\left(\log n - \log\ell + 2\ell\log\frac{e\lambda_{k-1}}{\ell}\right). \tag{4.27}$$

Using $\ell = \lfloor \log^{\beta_k} n \rfloor < \log^{\beta_{k-1}} n$, (4.13) and (4.14), we have $\lambda_{k-1} > \exp\left(-\frac{\ell^2}{5\log^{\beta_{k-2}} n}\right)$, and

$$\begin{aligned}\ell \log \lambda_{k-1} &> \frac{-\ell^3}{5\log^{\beta_{k-2}} n} = -\frac{\ell^3}{5\log^{3\beta_k} n} \cdot (\log n)^{4\beta_k - \beta_{k-2}} \\ &> -\frac{1}{4}(\log n)^{1+\frac{2(k+1)\varepsilon-(2\beta-1)}{2k+1}} = -\frac{\log n}{4}, \end{aligned} \tag{4.28}$$

where the second equality follows from that $\ell^3/\log^{3\beta_k} n = 1 - o(1)$ and $\beta_i = \beta - \frac{(2\beta-1-\varepsilon)i}{2k+1}$, and the last equality uses $\varepsilon = \frac{2\beta-1}{2(k+1)}$. Combining (4.26), (4.27), (4.28) and (4.15), we have

$$\mathbb{P}\left(|H_i| > \frac{4n\lambda_i^\ell}{\ell!}, r_i \le r_i^*\right) \le \exp\left(-n^{1/2-o(1)}\right).$$

This, together with (4.12), implies that

$$\mathbb{P}\left(|H_i| > \frac{4n\lambda_i^\ell}{\ell!}\right) \le \mathbb{P}\left(|H_i| > \frac{4n\lambda_i^\ell}{\ell!}, r_i \le r_i^*\right) + \mathbb{P}(r_i > r_i^*) \le \exp\left(-n^{1/2-o(1)}\right).$$

This concludes the proof of the i-th case of (4.14) given that the i-th case of (4.12) holds. This establishes the induction and thus the lemma. □

In the next lemma, we keep our notation r_i for the number of retries in the i-th stage, which proceeds from t_{i-1} to t_i and set $t_{k+1} := t_k + \ell$.

Lemma 4.6. *For $1 \le i \le k$ we have*

$$\mathbb{P}\left(\exists_{j\in[n]} L^f_{2,j}((t_{i-1}, t_i]) > t_{i+1} - t_i\right) \le 2n^{-e^{2\sqrt{\log\log\log n}}}. \tag{4.29}$$

Proof. Denote $E = \left\{\exists_{j\in[n]} L^f_{2,j}((t_{i-1}, t_i] > t_{i+1} - t_i\right\}$. Recall that r_i^* is defined in (4.13). Using the law of total probability, we have

$$\mathbb{P}(E) \le \mathbb{P}(E \mid r_i \le r_i^*) + \mathbb{P}(r_i > r_i^*). \tag{4.30}$$

We have already showed in Lemma 4.5 that

$$\mathbb{P}(r_i > r_i^*) \le \exp\left(-n^{1/2-o(1)}\right). \tag{4.31}$$

Next, we estimate the first term on the RHS of (4.30). Denote by $\{X_j^{(i)}\}_{j\in[n]}$ independent Poisson(λ_i) random variables, where λ_i is given in (4.14). By Lemma 3.1, we have

$$\mathbb{P}(E \mid r_i \le r_i^*) \le 2\mathbb{P}\left(\exists_{j\in[n]} X_j^{(i)} > t_{i+1} - t_i\right). \tag{4.32}$$

Using the face that $0 < \lambda_i < 1$ in (4.18), we have

$$\begin{aligned}\mathbb{P}\left(X_1^{(i)} \ge t_{i+1} - t_i\right) &< \frac{2\lambda_i^{t_{i+1}-t_i}}{(t_{i+1}-t_i)!} \le \left(\frac{e\lambda_i}{t_{i+1}-t_i}\right)^{t_{i+1}-t_i} \\ &\le \exp\left(-\frac{\ell^2(t_{i+1}-t_i)}{4\log^{\beta_{i-1}} n}\right) \le \exp\left(-\frac{\ell^2\log^{\beta_i} n}{5\log^{\beta_{i-1}} n}\right) \\ &\le \exp\left(-\frac{(\log n)^{2\beta_k+\beta_i}}{6\log^{\beta_{i-1}} n}\right) = \exp\left(-\frac{1}{6}\log^{1+\varepsilon} n\right).\end{aligned}$$

The second inequality follows from Stirling's approximation $n! \geq \sqrt{2\pi n}(n/e)^n$ for $n \in \mathbb{Z}_+$. The transition to the second line uses the definition of λ_i given in (4.14). In the penultimate inequality, we use $t_i = \lfloor t - \log^{\beta_i} n \rfloor$, where $\beta_i = \beta - \frac{(2\beta-1-\varepsilon)i}{2k+1}$, and that $\log^{\beta_{i+1}} n = o(\log^{\beta_i} n)$. The last inequality uses the fact that $\ell = \lfloor \log^{\beta_k} n \rfloor$. Taking into account of $\varepsilon = \frac{2\beta-1}{2(k+1)}$, $k = \lfloor \frac{\log\log n}{3\log\log\log n} \rfloor$ and $\beta > \frac{1}{2} + \frac{1}{\sqrt{\log\log\log n}}$, we have $\log^{\varepsilon} n \geq e^{3\sqrt{\log\log\log n}}$. Taking the union bound, we have for n large enough,

$$\mathbb{P}\left(\max_{j\in[n]} X_j^{(i)} > t_{i+1} - t_i\right) \leq n \exp\left(-\frac{1}{6}\log^{1+\varepsilon} n\right) \leq n^{-e^{2\sqrt{\log\log\log n}}}. \tag{4.33}$$

The desired statement (4.29) follows from (4.30)–(4.33). □

Now we are ready to prove Proposition 4.3.

Proof of Proposition 4.3. We will estimate the maximum loads after i stages for all $1 \leq i \leq k$. By the definition of H_i, we have

$$\mathrm{MaxLoad}^f_{(\cup_{j=0}^i H_j)^c}(t_i n) \leq L_0 + 2i\ell. \tag{4.34}$$

Next, we estimate the maximum load over $\cup_{j=1}^i H_j$ after i stages. For $1 \leq j \leq i \leq k$, we denote by $E_i^j = \left\{\mathrm{MaxLoad}^f_{H_j}(t_i n) > t_{i+1} - t_i + L_0 + (2j-1)\ell\right\}$, where $t_{k+1} = t_k + \ell$. We will show that

$$\mathbb{P}\big(E_i^j\big) \leq (i-j+1)\cdot 2n^{-e^{2\sqrt{\log\log\log n}}}. \tag{4.35}$$

We denote by r_i the number of retries in the i-th stage. In the i-th stage, for a bin in H_i to accept more than $t_i - t_{i-1} + \ell$ primary allocations, it is necessary that the load of this bin before accepting its last primary allocation is at most $-\log n$. Hence, we have

$$\begin{aligned}\mathrm{MaxLoad}^f_{H_i}(t_i n) &\leq \max\left\{\mathrm{MaxLoad}^f_{H_i}(t_{i-1} n) + \ell, -\log n\right\} + \max_{p\in H_i} L^f_{2,p}((t_{i-1}, t_i]) \\ &\leq L_0 + (2i-1)\ell + \max_{p\in H_i} L^f_{2,p}((t_{i-1}, t_i]),\end{aligned}$$

where the second inequality uses the fact that $H_i \subseteq (\cup_{j=0}^{i-1} H_j)^c$ and the $i-1$ case of (4.34). Using the inequalities above and Lemma 4.6, we obtain

$$\mathbb{P}\big(E_i^i\big) \leq \mathbb{P}\left(\max_{p\in H_i} L^f_{2,p}((t_{i-1}, t_i]) > t_{i+1} - t_i\right) \leq 2n^{-e^{2\sqrt{\log\log\log n}}}. \tag{4.36}$$

For $1 \leq j \leq i-1$ and $i \geq 2$, the strategy guarantees that in the i-th stage, each bin of H_j either accepts no primary allocations, or has a load at most $-\log n$ before accepting its last primary allocation. Hence, we have

$$\mathrm{MaxLoad}^f_{H_j}(t_i n) \leq \max\left\{\mathrm{MaxLoad}^f_{H_j}(t_{i-1} n) - (t_i - t_{i-1}), -\log n\right\} + \max_{p\in H_j} L^f_{2,p}((t_{i-1}, t_i]).$$

Hence, event E_i^j occurs only if one of the two conditions holds: $\max_{p\in H_j} L^f_{2,p}(r_i) > t_{i+1} - t_i$ or $\max\left\{\mathrm{MaxLoad}^f_{H_j}(t_{i-1} n) - (t_i - t_{i-1}), -\log n\right\} > L_0 + (2j-1)\ell$. The latter condition is equivalent to event E_{i-1}^j. This and Lemma 4.6 imply that

$$\mathbb{P}\big(E_i^j\big) \leq \mathbb{P}\big(E_{i-1}^j\big) + \mathbb{P}\left(\max_{p\in H_i} L^f_{2,p}((t_{i-1}, t_i]) > t_{i+1} - t_i\right) \leq \mathbb{P}\big(E_{i-1}^j\big) + 2n^{-e^{2\sqrt{\log\log\log n}}}.$$

Iterating this argument to obtain

$$\mathbb{P}\big(E_i^j\big) \le \mathbb{P}\big(E_j^j\big) + (i-j)\cdot n^{-e^{2\sqrt{\log\log\log n}}} \le (i-j+1)\cdot 2n^{-e^{2\sqrt{\log\log\log n}}},$$

where the second inequality follows from (4.36). This concludes the proof of (4.35).

Now, we estimate the maximum load over H_0. In the first stage, each bin in H_0 either accepts no primary allocations or has a load at most $-\log n$ before accepting its last primary allocation. Hence, we have

$$\mathrm{MaxLoad}^f_{H_0}(t_1 n) \le \max\Big\{\mathrm{MaxLoad}^f_{H_0}(t_0 n) - (t_1 - t_0), -\log n\Big\} + \max_{p\in H_0} L^f_{2,p}((t_0,t_1]). \quad (4.37)$$

In general, in the i-th stage for $2 \le i \le k$, for bin of H_0 to accept more than $t_i - t_{i-1} + \ell$ primary allocations, the load of this bin before accepting its last primary is at most $-\log n$. Hence, we obtain

$$\mathrm{MaxLoad}^f_{H_0}(t_i n) \le \max\Big\{\mathrm{MaxLoad}^f_{H_0}(t_{i-1} n) + \ell, -\log n\Big\} + \max_{p\in H_0} L^f_{2,p}((t_{i-1},t_i]). \quad (4.38)$$

Iteration of (4.38), together with (4.37), yields

$$\begin{aligned}
\mathrm{MaxLoad}^f_{H_0}(t_i n) &\le \max\Big\{\mathrm{MaxLoad}^f_{H_0}(t_1 n) + \ell, -\log n\Big\} + (i-2)\ell + \sum_{j=2}^{i} \max_{p\in H_0} L^f_{2,p}((t_{j-1},t_j]) \\
&\le \max\Big\{\mathrm{MaxLoad}^f_{H_0}(t_0 n) - (t_1 - t_0), -\log n\Big\} \\
&\quad + (i-1)\ell + \sum_{j=1}^{i} \max_{p\in H_0} L^f_{2,p}((t_{j-1},t_j]) \\
&\le \sum_{j=1}^{i} \max_{p\in H_0} L^f_{2,p}((t_{j-1},t_j]) + (i-1)\ell - \min\big\{(1-c-o(1))(t-t_0)), \log n\big\},
\end{aligned} \quad (4.39)$$

where the last inequality follows from the fact that $\mathrm{MaxLoad}^f_{H_0}(t_0 n) \le c(t-t_0)$ for some constant $0 < c < 1$, and that $t_1 - t_0 = (1-o(1))(t-t_0)$. Observe that $t_{i+1} - t_1 = o(t-t_0)$, $t_{i+1} - t_1 < \log n$ and $(i-1)\ell = o(t-t_0)$, $(i-1)\ell < \log n$. Hence, we have

$$(t_{i+1} - t_i) + (i-1)\ell < \min\big\{(1-c-o(1))(t-t_0)), \log n\big\}.$$

This, together with (4.39), implies that

$$\begin{aligned}
\mathbb{P}\left(\mathrm{MaxLoad}^f_{H_0}(t_i n) > 0\right) &\le \mathbb{P}\left(\sum_{j=1}^{i} \max_{p\in H_0} L^f_{2,p}((t_{j-1},t_j]) > t_{i+1} - t_1\right) \\
&\le \sum_{j=1}^{i} \mathbb{P}\left(\max_{p\in H_0} L^f_{2,p}((t_{j-1},t_j]) > t_{j+1} - t_j\right) \\
&\le i\cdot 2n^{-e^{2\sqrt{\log\log\log n}}},
\end{aligned} \quad (4.40)$$

where the last inequality follows from Lemma 4.6. Combine inequalities (4.34), (4.35), (4.40) to obtain

$$\mathbb{P}\left(\mathrm{MaxLoad}^f(tn) > L_0 + 2k\ell\right) \le k\cdot 2n^{-e^{2\sqrt{\log\log\log n}}} < n^{-e^{\sqrt{\log\log\log n}}}.$$

This concludes the proof. □

4.3 Case : $t \geq \omega(\log n)$

Proposition 4.7. *Denote by f the $\left(\frac{1}{5}, t - \frac{7}{\theta}\log n, \frac{7}{\theta}\log n, \ell, \ell\right)$-drift-threshold strategy with k and ℓ as in Proposition 4.3. Then, for n large enough, f has*

$$\mathbb{P}\left(\mathrm{MaxLoad}^f(tn) > (2k+1)\ell\right) < n^{-1/7}.$$

Proof. We employ the aforementioned concatenated strategy described in Section 2.4. Inequality (3.26) in Lemma 3.9 yield

$$\mathbb{P}\left(\mathrm{MaxLoad}^f(t_0 n) > \frac{6}{\theta}\log n\right) < n^{-1/6}.$$

Employing once again the notation

$$H_0 = \left\{i \in [n] : L_i^f(t_0 n) > \ell\right\},$$

we apply Lemma 3.10 to obtain

$$\mathbb{P}\left(|H_0| > \frac{160}{\theta^2} n e^{-\frac{\theta\ell}{3}}\right) \leq 2\exp\left(-2n\left(\frac{80}{\theta^2}\right)^2 e^{-\frac{2\theta\ell}{3}}\right) = \exp\left(-n^{1-o(1)}\right).$$

The inequalities above imply that, with probability at least $1 - \Theta(n^{-1/6})$, the conditions in Proposition 4.3 hold with $\eta = 0$ (observe that α there, satisfies $\alpha = 1 + \frac{\log 7 - \log\theta}{\log\log n}$. Hence, with high probability, we can apply the multi-stage (t, ℓ, ℓ)-threshold strategy in Section 4.2 from time t_0 to time t. Then we can apply Proposition 4.3 to conclude the proof. □

5 Single-time load discrepancy: lower bound

In this section, we show that no two-thinning strategy can achieve a maximum load better than that in Theorem 1. Due to inequality (4.1), we can again assume that $m = tn$ for $t \in \mathbb{N}$. The lower bound in Theorem 1 is an immediate consequence of the following statement applied on the process starting from time $\max\left\{\left\lfloor t - \frac{\sqrt{\log n}}{50}\right\rfloor, 0\right\}$.

Proposition 5.1. *Given $t \leq \frac{\sqrt{\log n}}{50}$, we set $\ell = \sqrt{\frac{\log n}{12(\log\log n - 2\log t)}}$. Then any two-thinning strategy f with any initial load vector $\{L_i(0)\}_{i\in[n]} \in \mathbb{Z}^n$ satisfies*

$$\mathbb{P}\left(\mathrm{MaxLoad}^f(tn) < \ell\right) < 3e^{-\sqrt{n}}. \tag{5.1}$$

Proof. If $\mathrm{MaxLoad}^f(0) \geq t + \ell$, we will have $\mathrm{MaxLoad}^f(tn) \geq \ell$ and inequality (5.1) trivially holds. Hence, we will assume that $\mathrm{MaxLoad}^f(0) < t + \ell$. We denote $S = \left\{i \in [n] : L_i(0) \geq 0\right\}$ and $S^c = [n] \setminus S$. We first show that

$$|S| \geq \frac{n}{t + \ell + 1}. \tag{5.2}$$

To see this, observe that

$$0 = \sum_{i\in[n]} L_i(0) = \sum_{i\in S} L_i(0) + \sum_{i\in S^c} L_i(0).$$

This, together with our assumptions that $\{L_i(0)\}_{i\in[n]} \in \mathbb{Z}^n$ and $\mathrm{MaxLoad}^f(0) < t+\ell$, yields

$$|S^c| \le \sum_{i\in S^c} |L_i(0)| = \sum_{i\in S} L_i(0) \le |S| \cdot (t+\ell).$$

Then inequality (5.2) readily follows from the inequality above and $|S^c| = n - |S|$.

Next, we set $r^* = \lfloor |S| e^{-2tI(\ell/t)}/2 \rfloor$, where $I(x)$ is given in Lemma 3.3. We denote by r the number of retries up to time tn. By the law of total probability, we have

$$\begin{aligned}
\mathbb{P}\left(\mathrm{MaxLoad}^f(tn) < \ell\right) &\le \mathbb{P}\left(\mathrm{MaxLoad}^f_S(tn) < \ell\right) \\
&= \mathbb{P}\left(\mathrm{MaxLoad}^f_S(tn) < \ell, r < r^*\right) \qquad (5.3)\\
&\quad + \mathbb{P}\left(\mathrm{MaxLoad}^f_S(tn) < \ell, r \ge r^*\right). \qquad (5.4)
\end{aligned}$$

We first estimate the probability in (5.3). Recall that $\psi_S^{t+\ell}(tn)$ defined in (2.5) represents the number of bins in S that are suggested as primary allocations at least $t+\ell$ times up to time tn. Observe that if we retry fewer than $\psi_S^{t+\ell}(tn)$ balls, the maximum load will be at least ℓ. Hence, we have

$$\mathbb{P}\left(\mathrm{MaxLoad}^f_S(tn) < \ell, r < r^*\right) \le \mathbb{P}\left(\psi_S^{t+\ell}(tn) < r^*\right). \qquad (5.5)$$

We denote by $\{X_i\}_{i\in[n]}$ independent Poisson(t) random variables. Let W_i be the indicator function of the event $\{X_i \ge t+\ell\}$. Hence, $\{W_i\}_{i\in[n]}$ are independent Bernoulli random variables such that

$$p := \mathbb{P}(W_i = 1) = \mathbb{P}(X_i \ge t+\ell) \ge e^{-2tI(\ell/t)} \ge \frac{2r^*}{|S|}, \qquad (5.6)$$

where the first inequality follows from Lemma 3.3. We then apply Lemma 3.1, inequality (5.6) and Hoeffding's inequality to obtain

$$\begin{aligned}
\mathbb{P}\left(\psi_S^{t+\ell}(tn) < r^*\right) &\le 2\mathbb{P}\left(\sum_{i\in S} W_i < r^*\right) \le 2\mathbb{P}\left(\sum_{i\in S} W_i < \frac{|S|p}{2}\right) \\
&\le 2\exp\left(-\frac{|S|p^2}{2}\right) \le 2\exp\left(-\frac{|S|}{2} e^{-4tI(\ell/t)}\right) \\
&\le 2\exp\left(-\frac{n}{2(t+\ell+1)}\left(\frac{et}{\ell}\right)^{12\ell}\right) \\
&= \exp\left(-n^{1-o(1)}\right), \qquad (5.7)
\end{aligned}$$

where the penultimate transition follows from the upper bound of $I(x)$ in (3.4) and the fact that $\ell > 4t$ for $t \le \frac{\sqrt{\log n}}{50}$.

Next we estimate the probability in (5.4). Recall that $L_{2,i}([tn])$ defined in (2.3) represents the number of balls that bin i receives from secondary allocations. Then we have

$$\mathbb{P}\left(\mathrm{MaxLoad}^f_S(tn) < \ell, r \ge r^*\right) \le \mathbb{P}\left(\max_{i\in S} L^f_{2,i}([tn]) < t+\ell, r \ge r^*\right). \qquad (5.8)$$

Apply Lemma 3.2 to obtain

$$\mathbb{P}\left(\max_{i\in S} L^f_{2,i}([tn]) < t+\ell, r \geq r^*\right) \leq 2\exp\left(-\frac{|S|(r^*/n)^{t+\ell}}{e(t+\ell)!}\right). \tag{5.9}$$

Using the upper bound of $I(x)$ in (3.4) and the fact that $\ell > 4t$ for $t \leq \frac{\sqrt{\log n}}{50}$, we obtain $r^* > \frac{|S|}{2}\left(\frac{et}{\ell}\right)^{6\ell}$. This, together with Stirling's approximation $k! \leq e\sqrt{k}(k/e)^k$, yields that for n large enough

$$\frac{(r^*/n)^{t+\ell}}{e(t+\ell)!} \geq \frac{1}{e^2\sqrt{t+\ell}}\left(\frac{e}{2(t+\ell)(t+\ell+1)}\right)^{t+\ell}\left(\frac{et}{\ell}\right)^{6\ell(t+\ell)} > \left(\frac{t}{\ell}\right)^{12\ell^2}.$$

This, together with (5.8) and (5.9), yields

$$\mathbb{P}\left(\mathrm{MaxLoad}^f(tn) < \ell, r \geq r^*\right) \leq 2\exp\left(-\frac{n}{(t+\ell+1)}\left(\frac{t}{\ell}\right)^{12\ell^2}\right) \leq 2e^{-\sqrt{n}}, \tag{5.10}$$

where the second inequality follows from the fact that

$$12\ell^2\log\frac{\ell}{t} = \frac{\log n}{2}\left(1 - \frac{\log(\log\log n - 2\log t) + \log 12}{\log\log n - 2\log t}\right) < \frac{\log n}{2}.$$

Then we can obtain (5.1) by combining (5.3), (5.4), (5.5), (5.7) and (5.10). □

6 All-time load discrepancy: upper bound

In the previous sections, we studied different thinning strategies which yield a good control of $\mathrm{MaxLoad}^f(m)$, the maximum load at the end of the process. Here we are interested in thinning strategies that can control $\mathrm{MaxLoad}^f([m])$, the maximum load throughout the entire process.

As before, we assume that $m = tn$ for $t \in \mathbb{N}$. Clearly, $\mathrm{MaxLoad}^f([m]) \geq \mathrm{MaxLoad}^f(m)$ and that $\mathrm{MaxLoad}^f([m])$ is monotone non-decreasing function of m. On the other hand, we also have $\mathrm{MaxLoad}^f([m]) \leq \mathrm{MaxLoad}^f(m) + t$, where the RHS is the maximum number of balls in a single bin at the end of the process. Hence, for $t = O(\sqrt{\log n})$, we can apply the $(t+\ell)$-threshold strategy as per the analysis in Section 4.1 and obtain an optimal all-time maximum load (up to some multiplicative constants). In the following couple of sections, we prove the upper bound in Theorem 2 for $t = \omega(\sqrt{\log n})$.

6.1 Case: $\omega(\sqrt{\log n}) \leq t \leq O(\log^2 n/(\log\log n)^3)$

Proposition 6.1. *Suppose that $\omega(\sqrt{\log n}) \leq t \leq \frac{\log^2 n}{(24\log\log n)^3}$. Set $\ell = (t\log n)^{1/3}$. We also assume that for all $i \in [n]$ the initial load satisfies $L_i(0) \leq L_0$ for some $L_0 > 0$. Then for any $c > 0$ and sufficiently large n, the ℓ-relative threshold strategy f satisfies*

$$\mathbb{P}\left(\mathrm{MaxLoad}^f([tn]) > L_0 + (12c+9)\ell\right) \leq n^{-c}. \tag{6.1}$$

Proof. Observe that for any $s \in [t]$ and any $(s-1)n < k \leq sn$,

$$\mathrm{MaxLoad}^f((s-1)n) - 1 \leq \mathrm{MaxLoad}^f(k) \leq \mathrm{MaxLoad}^f(sn) + 1.$$

Hence, it suffice to show that

$$\mathbb{P}\left(\max_{s\in[t]}\mathrm{MaxLoad}^f(sn) > L_0 + (12c+8)\ell\right) < n^{-c}. \tag{6.2}$$

For $s \in [t]$, we denote by r_s the number of retries in the s-th stage, i.e., in the time interval $(n(s-1), ns]$. On the one hand, if a bin $i \in [n]$ accepts more than $s+\ell$ primary allocations in the first s stages, the load of this bin before accepting the last primary allocation has to be at most $-\log n$. For such a bin $i \in [n]$, we have

$$L_i^f(sn) \le L_{2,i}^f([sn]) - \log n + 1,$$

where the function $L_{2,i}^f$ given in (2.3) is the number of balls bin i receive from secondary allocations. On the other hand, if a bin i accepts at most $s+\ell$ primary allocations in the first s stages, we have

$$L_i^f(sn) \le L_i(0) + \ell + L_{2,i}^f([sn]) \le L_{2,i}^f([sn]) + L_0 + \ell.$$

Write $E_s = \{r_k \le r_k^* \text{ for all } 1 \le k \le s\}$, where $r_k^* = 6ne^{-kI(\ell/k)}/\log(1+\ell/k)$ and $I(x)$ is given in Lemma 3.3. The inequalities above and the law of total probability imply that

$$\begin{aligned}\mathbb{P}\left(\mathrm{MaxLoad}^f(sn) > L_0 + (12c+8)\ell\right) &\le \mathbb{P}\left(\max_{i\in[n]} L_{2,i}^f([sn]) > (12c+7)\ell\right)\\ &\le \mathbb{P}\left(\max_{i\in[n]} L_{2,i}^f([sn]) > (12c+7)\ell,\ E_s\right) + \mathbb{P}\big(E_s^c\big).\end{aligned} \tag{6.3}$$

We first estimate $\mathbb{P}\big(E_s^c\big)$. The definition of our ℓ-relative threshold strategy given in Section 2.3 guarantees that if a retry occurs in the k-th stage, then it is necessary that the suggested bin has accepted at least $k-1+\ell$ primary allocations. Hence, for a single bin, the number of retries in the k-th stage is either 0 or the difference between the number of times this bin was suggested as a primary allocation up to stage k and $k-1+\ell$ provided that the difference is positive. We write $\{X_i^k\}_{i\in[n]}$ for independent Poisson(k) random variables. Define $Y_i^k = \max\big\{0, X_i^k - k - \ell + 1\big\}$ and $Y^k = \sum_{i=1}^n Y_i^k$. Lemmata 3.1 & 3.5 yield

$$\mathbb{P}(r_k > r_k^*) \le 2\mathbb{P}\big(Y^k > r_k^*\big) \le 2\exp\left(-ne^{-kI(\ell/k)}\right).$$

One can check that $I(x)/x$ is an increasing function. Then it is not hard to see that for any fixed $\ell > 0$, the function $e^{-kI(\ell/k)}$ is increasing with respect to k. Hence, for all $k \in [t]$, we have

$$\mathbb{P}(r_k > r_k^*) \le 2\exp\left(-ne^{-I(\ell)}\right) \le 2\exp\left(-n\left(\frac{e}{\ell}\right)^{3\ell}\right),$$

where the last inequality follows from the upper bound of $I(x)$ in (3.4). Our assumption of t and the choice of ℓ yield $\ell \le \frac{\log n}{24\log\log n}$ and hence

$$n\left(\frac{e}{\ell}\right)^{3\ell} = \exp\left(\log n - 3\ell\log\frac{\ell}{e}\right) > \sqrt{n}.$$

Take the union bound to obtain (for n large enough),

$$\mathbb{P}\big(E_s^c\big) \le \sum_{k=1}^{s}\mathbb{P}(r_k > r_k^*) \le 2se^{-\sqrt{n}} = e^{-(1-o(1))\sqrt{n}}. \tag{6.4}$$

Now, we estimate the first term of (6.3). Recall that $r_k^* = 6ne^{-kI(\ell/k)}/\log(1+\ell/k)$. We again use the fact that $I(x)/x$ is increasing to deduce that r_k^* is an increasing function. Hence, when E_s occurs, the total number of retries is no more than tr_t^*. We denote by $\{Z_i\}_{i\in[n]}$ independent Poisson(λ) random variables, where

$$\lambda = \frac{tr_t^*}{n} = \frac{6te^{-tI(\ell/t)}}{\log(1+\ell/t)} < \frac{12t^2}{\ell}\exp\left(-\frac{\ell^2}{4t}\right), \tag{6.5}$$

where the inequality follows from the lower bound of $I(x)$ in (3.3) and $\log(1+x) \geq x/2$ for $0 < x < 1$ and the fact that $\ell \leq t$. Using Lemma 3.1, we obtain

$$\mathbb{P}\left(\max_{i\in[n]} L_{2,i}^f([sn]) > (12c+7)\ell,\ E_s\right) \leq 2\mathbb{P}\left(\max_{i\in[n]} Z_i > (12c+7)\ell\right). \tag{6.6}$$

Apply Lemma 3.3 to obtain

$$\begin{aligned}\mathbb{P}(Z_1 > (12c+7)\ell) \leq \mathbb{P}(Z_1 > \lambda + (12c+6)\ell) \leq e^{-\lambda I((12c+6)\ell/\lambda)} &< \left(\frac{e\lambda}{(12c+6)\ell}\right)^{(12c+6)\ell} \\ &< \exp\left(-\frac{(12c+6)\ell^3}{4t} + (12c+6)\ell\log\frac{2et^2}{(2c+1)\ell^2}\right),\end{aligned}$$

where the first inequality follows from that $\lambda < \ell$, the third inequality follows from the lower bound of $I(x)$ in (3.4), and in the last inequality we use the upper bound on λ in (6.5). Our choice of ℓ and the assumption on t guarantees that $\ell^2 > 12t\log t$, which yields

$$\ell\log\frac{2et^2}{(2c+1)\ell^2} < \ell\log t < \frac{\ell^3}{12t}.$$

Combine the two inequalities above to obtain

$$\mathbb{P}(Z_1 > (12c+7)\ell) \leq \exp\left(-\frac{(2c+1)\ell^3}{t}\right) = n^{-(2c+1)}.$$

This, together with (6.6), yields that

$$\mathbb{P}\left(\max_{i\in[n]} L_{2,i}^f([sn]) > (12c+7)\ell,\ E_s\right) \leq 2n\mathbb{P}(Z_1 > (12c+7)\ell) \leq 2n^{-2c}.$$

Combining the inequality above with (6.3), (6.4), we obtain that for any $s \in [t]$ and n large enough,

$$\mathbb{P}\left(\mathrm{MaxLoad}^f(sn) > L_0 + (12c+8)\ell\right) \leq e^{-(1-o(1))\sqrt{n}} + 2n^{-2c} \leq 3n^{-2c}.$$

Taking a union bound, we can obtain for n large enough,

$$\mathbb{P}\left(\max_{s\in[t]} \mathrm{MaxLoad}^f(sn) > L_0 + (12c+8)\ell\right) \leq 3tn^{-2c} \leq n^{-c}.$$

This proves (6.2), and hence (6.1). □

6.2 Case: $\omega(\log^2 n/(\log\log n)^3)) \le t \le n^{O(1)}$

In this case, we utilize the varying drift strategy to control the all-time maximum load. We set $Z_k = i$ if the k-th point of $X(t)$ is a point of the process $X_i(t)$ define in Section 2.3. We will show that, with high probability, the random process $\{Z_k\}_{k\in\mathbb{N}}$ can be realized by some two-thinning strategy f and that it achieves the desired bound.

Proposition 6.2. *Let $m, n \in \mathbb{N}$ sufficiently large and denote $d = \frac{\log m}{\log n}$. Let $\ell = \frac{2\log n}{\log\log n}$. The ℓ-varying drift strategy f defined above satisfies*

$$\mathbb{P}\left(\mathrm{MaxLoad}^f([m]) > (d+4)\ell\right) \le \frac{2\log^3 n}{n}. \tag{6.7}$$

Next, we provide an estimate of the probability that the realizability criterion (2.13) holds for a period of time, which implies that, with high probability, the process $\{Z_k\}_{k\in\mathbb{N}}$ can be realized by some two-thinning strategy f for quasi-exponential time.

Lemma 6.3. *For any $T > 0$ and sufficiently large n, we have*

$$\mathbb{P}\left(\exists t \in [0,T] \ : \ \left|\left\{i \in [n] : X_i(t) - t > \ell\right\}\right| > \frac{n}{\sqrt{\log n}}\right) \le T\exp\left(-\frac{n}{2\log n}\right). \tag{6.8}$$

Proof. We first estimate the probability $\mathbb{P}(\sup_{s\in[t,t+1]}(X_i(s)-s) > \ell)$ for all $0 \le t \le T-1$. We denote by $E = \{X_i(t) \le t + \ell/2\}$. By the law of total probability,

$$\mathbb{P}\left(\sup_{s\in[t,t+1]}(X_i(s)-s) > \ell\right) \le \mathbb{P}\left(\sup_{s\in[t,t+1]}(X_i(s)-s) > \ell,\ E\right) + \mathbb{P}(E^c). \tag{6.9}$$

Since $X_i(t)$, given in (2.12), is $\frac{1}{\sqrt{\log n}}$-standardizing, we can apply inequality (3.22) in Corollary 3.8 and Markov's inequality to obtain

$$\mathbb{P}(E^c) \le \mathbb{P}\left(|X_i(t) - t| \ge \frac{\ell}{2}\right) \le 80\log n \cdot \exp\left(-\frac{\sqrt{\log n}}{2\log\log n}\right) < \frac{1}{4\sqrt{\log n}}. \tag{6.10}$$

Next we bound $\mathbb{P}\left(\sup_{s\in[t,t+1]}(X_i(s)-s) > \ell,\ E\right)$. Let Y be a Poisson$(1+\theta_1)$ variable. Observe that, by (2.12), $X_i(t+1) - X_i(t)$ is stochastically dominated by Y. Hence, we have

$$\begin{aligned}\mathbb{P}\left(\sup_{s\in[t,t+1]}(X_i(s)-s) > \ell,\ E\right) &\le \mathbb{P}\left(Y > \frac{\ell}{2}\right) \le \exp\left(-(1+\theta_1)I\left(\frac{\ell}{2(1+\theta_1)}\right)\right) \\ &\le \exp\left(-\frac{\ell}{3}\log\frac{\ell}{3e}\right) = n^{-\frac{2}{3}+o(1)},\end{aligned} \tag{6.11}$$

where in the second inequality, the function I, appearing in Lemma 3.3, is the rate function of the deviation bound of Poisson random variables, and the last inequality follows from the fact that $I(x) > x\log(x/e)$ for $x > 4$. Combine (6.9), (6.10) and (6.11) to obtain

$$\mathbb{P}\left(\sup_{s\in[t,t+1]}(X_i(s)-s) > \ell\right) \le \frac{1}{4\sqrt{\log n}} + n^{-\frac{2}{3}+o(1)} \le \frac{1}{2\sqrt{\log n}}.$$

We denote by $S(t) = \{i \in [n] : \sup_{s\in[t,t+1]}(X_i(s) - s) > \ell\}$. Let W_i be the indicator function of the event $\{\sup_{s\in[t,t+1]}(X_i(s)-s) > \ell\}$. Hence, $\{W_i\}_{i\in[n]}$ are independent Bernoulli random variables such that

$$\mathbb{P}(W_i = 1) = \mathbb{P}\left(\sup_{s\in[t,t+1]}(X_i(s) - s) > \ell\right) \le \frac{1}{2\sqrt{\log n}}.$$

By Hoeffding's inequality,

$$\mathbb{P}\left(|S(t)| > \frac{n}{\sqrt{\log n}}\right) = \mathbb{P}\left(\sum_{i=1}^{n} W_i \ge \frac{n}{\sqrt{\log n}}\right) \le \exp\left(-\frac{n}{2\log n}\right).$$

The desired statement (6.8) follows by taking a union bound. □

We are now ready to establish Proposition 6.2.

Proof of Proposition 6.2. Set $T = m/n + \Delta$, where $\Delta = 1 + 2\sqrt{\log n}\log(80\log n)$. Let E be the event that $\{Z_i\}_{i\in\mathbb{N}}$ can be realized by some two-thinning strategy f. Lemma 6.3 yields

$$\mathbb{P}(E^c) \le T\exp\left(-\frac{n}{2\log n}\right). \tag{6.12}$$

For each fixed $1 \le k \le m$, we set $t^* = k/n + \Delta$. We write $F = \{X(t^*) \ge k\}$. The law of total probability yields

$$\mathbb{P}\left(L_i^f(k) > (d+4)\ell\right) \le \mathbb{P}\left(L_i^f(k) > (d+4)\ell,\ E \cap F\right) + \mathbb{P}(E^c) + \mathbb{P}(F^c). \tag{6.13}$$

Since $X_i(t)$ given in (2.12) is $\frac{1}{\sqrt{\log n}}$-standarizing, we can apply the second inequality of (3.23) in Corollary 3.8 and Markov's inequality to obtain

$$\mathbb{P}(F^c) = \mathbb{P}\left(X(t^*) < nt^* - n\Delta\right) \le (80\log n)^n \exp\left(-\frac{n\Delta}{2\sqrt{\log n}}\right) = \exp\left(-\frac{n}{2\sqrt{\log n}}\right), \tag{6.14}$$

where the last equality follows from our choice of Δ. The definition of $X_i(t)$ in (2.12) implies that $X_i(t) - \ell$ is upper $(1 - \frac{12}{\sqrt{\log n}})$-standardizing. One can check that the condition of inequality (3.18) in Corollary 3.7 holds for $2\theta = 1 - \frac{12}{\sqrt{\log n}}$, $\lambda = \frac{\log\log n}{2}$. Hence, we apply inequality (3.18) to obtain

$$\mathbb{E}\exp\left(\frac{\log\log n}{2}(X_i(t) - t - \ell)\right) < 1 + \frac{2e^{2\lambda}}{1 - e^{-\lambda/2}} < \log^3 n. \tag{6.15}$$

Whenever the event $E \cap F$ occurs, we have $L_i^f(k) + k/n \le X_i(t^*)$. Inequality (6.15) and Markov's inequality yield

$$\begin{aligned}
\mathbb{P}\left(L_i^f(k) > (d+4)\ell,\ E \cap F\right) &\le \mathbb{P}\left(X_i(t^*) > \frac{k}{n} + (d+4)\ell\right) \\
&= \mathbb{P}\left(X_i(t^*) - t^* - \ell > (d+3)\ell - \Delta\right) \\
&\le \mathbb{P}\left(X_i(t^*) - t^* - \ell > (d+2)\ell\right) \\
&\le (\log n)^3 \cdot n^{-(d+2)}.
\end{aligned}$$

This, together with (6.12), (6.13), (6.14), yields that, for sufficiently large n,

$$\mathbb{P}\left(L_i^f(k) > (d+4)\ell\right) \le 2(\log n)^3 \cdot n^{-(d+2)}.$$

Taking union bound over m and n, we obtain

$$\mathbb{P}\left(\mathrm{MaxLoad}^f([m]) > (d+4)\ell\right) \le 2(\log n)^3 m n^{-(d+1)}.$$

Then, inequality (6.7) follows from the fact that $m = n^d$ □

7 All-time load discrepancy: lower bound

Here we prove the lower bounds in Theorem 2. We again assume that m is divisible by n and write $m = tn$ for some $t \in \mathbb{Z}$. Observe that the lower bound of the single-time maximum load in Theorem 1 implies that of the all-time maximum load up to $t = O(\sqrt{\log n})$. Our next result covers the regime of $\sqrt{\log n} < t < \log^2 n/(24 \log\log n)^3$. This, together with the fact that the all-time maximum load is non-decreasing with respect to t, implies the lower bound of $\Theta\big(\frac{\log n}{\log\log n}\big)$ for $t \ge \log^2 n/(24\log\log n)^3$. This completes the proof of the lower bounds in Theorem 2.

Proposition 7.1. *Suppose that* $\sqrt{\log n} < t < \frac{\log^2 n}{(24\log\log n)^3}$. *Set* $\ell = \lfloor (t\log n)^{1/3} \rfloor$. *Any two-thinning strategy* f *satisfies that for* n *large enough,*

$$\mathbb{P}\left(\mathrm{MaxLoad}^f([tn]) < \ell\right) \le \exp\left(-n^{1/5}\right). \tag{7.1}$$

Proof. We denote by r the total number of retries and set $r^* = \frac{n}{2}e^{-\ell^2/t}$. Then we have

$$\mathbb{P}\left(\mathrm{MaxLoad}^f([tn]) < \ell\right) = \mathbb{P}\left(\mathrm{MaxLoad}^f([tn]) < \ell, r < r^*\right) \tag{7.2}$$

$$+ \mathbb{P}\left(\mathrm{MaxLoad}^f([tn]) < \ell, r \ge r^*\right). \tag{7.3}$$

We estimate (7.2). Recall that $\psi^{t+\ell}(tn)$ defined in (2.5) represents the number of bins that are suggested as primary allocations at least $t+\ell$ times after allocating tn balls. If we retry less than $\psi^{t+\ell}(tn)$ balls, then we will have $\mathrm{MaxLoad}^f(tn) \ge \ell$. Hence we obtain

$$\begin{aligned}\mathbb{P}\left(\mathrm{MaxLoad}^f([tn]) < \ell, r < r^*\right) &\le \mathbb{P}\left(\mathrm{MaxLoad}^f(tn) < \ell, r < r^*\right) \\ &\le \mathbb{P}\left(\psi^{t+\ell}(tn) < r^*\right). \end{aligned} \tag{7.4}$$

We denote by $\{X_i\}_{i\in[n]}$ independent Poisson(t) random variables. Write Y_i for the indicator function of the event $\{X_i > t+\ell\}$. Hence, $\{Y_i\}_{i\in[n]}$ are independent Bernoulli(p) random variables with

$$p = \mathbb{P}(X_1 > t+\ell) \ge e^{-2tI(\ell/t)} \ge e^{-\ell^2/t},$$

where the first inequality follows from Lemma 3.3 and the second inequality uses the upper bound of $I(x)$ in (3.3) and the fact that $\ell < t$. Apply Lemma 3.1 and Hoeffding's inequality

to obtain

$$\begin{aligned}\mathbb{P}\left(\psi^{t+\ell}(tn) < r^*\right) &\le \mathbb{P}\left(\psi^{t+\ell}(tn) < \frac{pn}{2}\right) \le 2\mathbb{P}\left(\sum_{i=1}^{n} Y_i < \frac{pn}{2}\right) \\ &\le 2\exp\left(-\frac{p^2 n}{2}\right) < \exp\left(-\frac{n}{2}e^{-2\ell^2/t}\right) \\ &= \exp\left(-n^{1-o(1)}\right). \end{aligned} \tag{7.5}$$

Next we estimate (7.3). Recall that R_k given in (2.1) is the number of retries after allocating k balls. Define $s_0 = \inf\left\{s \in [t] : R_{sn} - R_{(s-1)n} \ge r^*/t\right\}$. Whenever the event $\{r \ge r^*\}$ occurs, we have $s_0 < \infty$. Write $S = \left\{i \in [n] : L_i^f((s_0-1)n) \ge 0\right\}$. As per (5.2), we show that whenever the event $\left\{\mathrm{MaxLoad}^f((s_0-1)n) < \ell\right\}$ occurs, we have

$$|S| \ge \frac{n}{\ell+1}. \tag{7.6}$$

To see this, observe that

$$0 = \sum_{i\in[n]} L_i^f((s_0-1)n) = \sum_{i\in S} L_i^f((s_0-1)n) + \sum_{i\in S^c} L_i^f((s_0-1)n).$$

This, together the fact that $\left\{L_i^f((s_0-1)n)\right\}_{i\in[n]} \in \mathbb{Z}^n$ and $\mathrm{MaxLoad}^f((s_0-1)n) < \ell$, yields

$$|S^c| \le \sum_{i\in S^c} |L_i^f((s_0-1)n)| = \sum_{i\in S} L_i^f((s_0-1)n) \le |S|\cdot(\ell+1).$$

Then we can obtain (7.6) using $|S^c| = n - |S|$.

Apply Lemma 3.2 to obtain

$$\begin{aligned}\mathbb{P}\left(\mathrm{MaxLoad}^f([tn]) < \ell, r \ge r^*\right) &\le \mathbb{P}\big(s_0 < \infty,\ \mathrm{MaxLoad}^f(s_0 n) < \ell\big) \\ &\le \mathbb{P}\left(s_0 < \infty,\ \max_{i\in S} L_{2,i}^f\big(((s_0-1)n, s_0 n]\big) < \ell\right) \\ &\le 2\exp\left(-\frac{|S|}{e\ell!}\left(\frac{r^*}{tn}\right)^{\ell}\right). \end{aligned} \tag{7.7}$$

Recall that $r^* = \frac{n}{2}e^{-\ell^2/t}$, $\ell = \lfloor (t\log n)^{1/3}\rfloor$ and $|S| \ge n/(\ell+1)$. One can check that

$$\frac{|S|}{e\ell!}\left(\frac{r^*}{tn}\right)^{\ell} \ge \frac{\sqrt{n}}{e(\ell+1)!}\left(\frac{1}{2t}\right)^{\ell} > \frac{\sqrt{n}t^{-\ell}}{(\ell+1)^{\ell+3/2}} > \frac{\sqrt{n}}{t^{3\ell}} > n^{1/4}, \tag{7.8}$$

where the second inequality uses Stirling's approximation $(\ell+1)! \le e\sqrt{\ell+1}(\frac{\ell+1}{e})^{\ell+1}$; in the third inequality, we use the fact that $\ell < t$ and the last inequality follows from our choice of ℓ and the assumption on t. Combine (7.7) and (7.8) to obtain

$$\mathbb{P}\left(\mathrm{MaxLoad}^f([tn]) < \ell, r \ge r^*\right) \le 2\exp\left(-n^{1/4}\right).$$

This, together with (7.3), (7.4), (7.5), yields

$$\mathbb{P}\left(\mathrm{MaxLoad}^f([tn]) < \ell\right) \le \exp\left(-n^{1-o(1)}\right) + 2\exp\left(-n^{1/4}\right) < \exp\left(-n^{1/5}\right).$$

This concludes the proof of (7.1). □

8 Typical load discrepancy

In this section, we investigate two-thinning strategies for controlling the ε-typical maximum load $\mathrm{MaxLoad}^f_\varepsilon([m])$. The main technical statement in this section is the following Proposition, which implies Theorem 3.

Proposition 8.1. *Fix $d \geq 1$. Set $\ell = (\log n)^{\frac{1}{2}+\frac{1}{\sqrt{\log\log\log n}}}$ and $\varepsilon = e^{-\frac{1}{2}\sqrt{\log\log\log n}}$. For sufficiently large $n \in \mathbb{N}$ and $m \leq n^d$, there exists a set $S \subset [m]$ with $|S| \geq (1-\varepsilon)m$ such that the d-multi-scaled long-term combined strategy f satisfies*

$$\mathbb{P}\Big(\mathrm{MaxLoad}^f(S) > \ell\Big) \leq \frac{1}{n}. \tag{8.1}$$

For $d \geq 2$ and general values of m, the d-multi-scaled long-term combined strategy f satisfies

$$\mathbb{P}\Big(\mathrm{MaxLoad}^f_\varepsilon([m]) > \ell\Big) \leq \frac{1}{n}. \tag{8.2}$$

The proof of this result requires the following four propositions, each of which tells us certain property of the process after a phase of an iteration. The proofs of these propositions are given in the following subsections. Throughout this section we use the notations in (2.17), (2.18) and (2.19).

Proposition 8.2. *Fix $d \geq 1$. Let $n \in \mathbb{N}$ be sufficiently large. Suppose that the initial load vector $\{L_i(0)\}_{i\in[n]}$ satisfies that $|\{i \in [n] : L_i(0) > L_0\}| \leq 4000ne^{-L_0/15}$ and that $|L_i(0)| \leq 100d\log n$ for all $i \in [n]$. Then the multi-stage $(m_0/n, L_0, L_0)$-threshold strategy f satisfies that*

$$\mathbb{P}\Big(L^f_i(m_0) < -300d\log n \text{ or } L^f_i(m_0) > L \text{ for some } i \in [n]\Big) \leq n^{-e^{\sqrt{\log\log\log n}}}.$$

Proposition 8.3. *Fix $c > 0$. Let $m, n \in \mathbb{N}$ be sufficiently large. We write $\alpha = \frac{\log(m/n)}{\log\log n}$ and assume that $\alpha \in \big[\frac{1}{2}+\frac{3}{\sqrt{\log\log\log n}}, 1+\frac{1}{30\sqrt{\log\log\log n}}\big]$. Further, we denote by $\varepsilon = e^{-\frac{2}{3}\sqrt{\log\log\log n}}$ and $\ell = (\log n)^{\frac{1}{2}+\frac{1}{\sqrt{\log\log\log n}}}$. Suppose that the initial load vector $\{L_i(0)\}_{i\in[n]}$ satisfies that $L_i(0) \leq (\log n)^{\frac{1}{2}+\frac{1}{2\sqrt{\log\log\log n}}}$ for all $i \in [n]$. Then, there exists $A_m \subset [m]$ with $|A_m| \geq (1-\varepsilon)m$ such that the 0-multi-scale strategy f satisfies that*

$$\mathbb{P}\Big(\mathrm{MaxLoad}^f(A_m) > \ell\Big) \leq n^{-c}. \tag{8.3}$$

Proposition 8.4. *Fix $d \geq 1$. Let $n \in \mathbb{N}$ be sufficiently large. Suppose that the initial load vector $\{L_i(0)\}_{i\in[n]}$ satisfies that $-300d\log n \leq L_i(0) \leq Q$ for all $i \in [n]$. Then the Q-multi-scale strategy f satisfies that*

$$\mathbb{P}\left(\max_{i\in[n]} |L^f_i(m_1)| > A\right) \leq n^{-3d}. \tag{8.4}$$

Proposition 8.5. *Fix $d \geq 1$. Let $n \in \mathbb{N}$ be sufficiently large. Suppose that the initial load vector $\{L_i(0)\}_{i\in[n]}$ satisfies that $|L_i(0)| \leq A$ for all $i \in [n]$. Then the 1/5-drift strategy f satisfies that*

$$\mathbb{P}\left(\max_{i\in[n]} |L^f_i(m_2)| > 100d\log n \text{ or } \Big|\Big\{i \in [n] : L^f_i(m_2) > L_0\Big\}\Big| > 4000ne^{-L_0/15}\right) \leq 2n^{-3d}.$$

Proof of Proposition 8.1. Observe that Lemma 3.11 guarantees that the third phase of each iteration eventually terminates so that there are almost surely infinitely many iterations. Set $M_{1,0} = M_{1,1} = 0$, $M_{1,2} = m_1$, $M_{1,3} = m_1 + m_{2,1}$. For $j \geq 2$ and $k \in \{0,1,2,3\}$, we define

$$M_{j,0} = M_{j-1,3}, \quad M_{j,1} = M_{j,0} + m_0, \quad M_{j,2} = M_{j,1} + m_1, \quad M_{j,3} = M_{j,2} + m_{2,j}.$$

Hence, for $k \in \{0,1,2\}$, $M_{j,k}$ is the starting time of the $(k+1)$-th phase in the j-th iteration. For $j \in \mathbb{N}$, we define events

$$\begin{aligned} E_j &= \left\{ -300d \log n \leq L_i^f\left(M_{j,1}\right) \leq L \text{ for all } i \in [n] \right\}, \\ F_j &= \left\{ \max_{i \in [n]} \left| L_i^f\left(M_{j,2}\right) \right| \leq A \right\}, \\ G_j &= \{ m_{2,j} = m_2 \}. \end{aligned}$$

Our strategy guarantees that the load vector $\left\{L_i^f(M_{j,0})\right\}_{i\in[n]}$ at the beginning of the j-th iteration satisfies that

$$\max_{i\in[n]} \left|L_i^f(M_{j,0})\right| \leq 100d\log n \quad \text{and} \quad \left|\left\{ i \in [n] : L_i^f(M_{j,0}) > L_0 \right\}\right| \leq 4000ne^{-L_0/15}.$$

Hence, we apply Proposition 8.2 to obtain for all $j > 1$ that

$$\mathbb{P}(E_j^c) \leq n^{-e^{\sqrt{\log\log\log n}}}. \tag{8.5}$$

This inequality trivially holds for E_1^c. By Proposition 8.3, we have for all $j \geq 1$ that

$$\mathbb{P}\left(\mathrm{MaxLoad}^f\left(A_{m_1} + M_{j,1}\right) > \ell \mid E_j \right) \leq n^{-3d}. \tag{8.6}$$

By Proposition 8.4, we have for all $j \geq 1$ that

$$\mathbb{P}\left(F_j^c \mid E_j \right) \leq n^{-3d}. \tag{8.7}$$

By Proposition 8.5, we have for all $j \geq 1$ that

$$\mathbb{P}\left(G_j^c \mid F_j \right) \leq 2n^{-3d}. \tag{8.8}$$

Set $M = m_0 + m_1 + m_2$. On the event $\bigcap_{j\in[\kappa]} G_j$, we have

$$\bigcup_{j=1}^{\kappa} \left(A_{m_1} + M_{j,1}\right) = \bigcup_{j=0}^{\kappa-1} (A_{m_1} + jM).$$

Set $S = \bigcup_{j=0}^{\kappa-1}(A_{m_1} + jM)$. Putting together (8.5), (8.6), (8.7),(8.8) and taking the union bound, we now get

$$\mathbb{P}\left(\mathrm{MaxLoad}^f(S) > \ell \right) \leq 5\kappa n^{-3d}.$$

For $m \leq n^d$, we take $\kappa = \lfloor m/M \rfloor \leq n^d$, so that the probability above is less than $1/n$.

Next, we complete the proof of (8.1) by showing that $|S| \geq (1 - \varepsilon/4)m$. Notice that S is a disjoint union of copies of A_{m_1} shifted by multiples of M. Hence, it suffices to show that

$|A_{m_1}| > (1-\varepsilon/4)M$. Indeed, we have shown in Proposition 8.3 that $|A_{m_1}| > (1-\varepsilon')m_1$ with $\varepsilon' = e^{-\frac{2}{3}\sqrt{\log\log\log n}}$. By the definitions of m_0, m_1, m_2 in (2.18), we have

$$\frac{m_1}{M} = 1 - \frac{m_0+m_2}{M} = 1 - O\Big((\log n)^{\frac{1-\alpha_{i_{\max}+1}}{2}}\Big) > 1 - O\Big((\log n)^{-\frac{1}{60\sqrt{\log\log\log n}}}\Big), \tag{8.9}$$

where the equality follows from (8.21). These, together with $\varepsilon \gg \varepsilon'$, yield that for sufficiently large n we have $|A_{m_1}| > (1-2\varepsilon')M > (1-\varepsilon/4)M$ and hence that $|S| \geq (1-\varepsilon/4)m$.

We now prove (8.2). We say that the j-th iteration is *bad* if either E_j^c, F_j^c, G_j^c happened or $\mathrm{MaxLoad}^f(A_{m_1} + M_{j,1}) > \ell$; otherwise we say that it is *good*. We denote by J the set of bad iterations among the first κ iterations. By definition, each good iteration has length at most $M = m_0 + m_1 + m_2$ and the maximum load over $\cup_{j\in[\kappa]\setminus J}(M_{j,1} + A_{m_1})$ is bounded above by ℓ. Hence, we have

$$\begin{aligned}\Big|\Big\{m' < m : \mathrm{MaxLoad}^f(m') > \ell\Big\}\Big| &\leq \sum_{j\in J}(m_0+m_1+m_{2,j}) + (\kappa - |J|)(M - |A_{m_1}|) \\ &\leq \sum_{j\in J}(m_0+m_1+m_{2,j}) + \frac{\varepsilon m}{2}, \end{aligned} \tag{8.10}$$

where the second inequality follows from $\kappa \leq m/M$, $|A_{m_1}| > (1-\varepsilon/4)M$ and (8.9).

We now estimate the first term of (8.10). As we have just seen, the probability of an iteration being bad is bounded above by $5n^{-3d}$ and hence $\mathbb{E}|J| \leq 5\kappa n^{-3d}$. Then we apply Markov's inequality to obtain

$$\mathbb{P}\Big(|J| > 5\kappa n^{-2d}\Big) \leq n^{-d}. \tag{8.11}$$

This, together with $m_0 < m_1$ and $\kappa < m/m_1$, yields

$$\mathbb{P}\Big((m_0+m_1)|J| > 10mn^{-2d}\Big) \leq n^{-d}. \tag{8.12}$$

We now estimate $\sum_{j\in J} m_{2,j}$. Note that the load vector at the beginning of the third phase of each iteration satisfies

$$\max_{i\in[n]} \big|L_i^f(M_{j,2})\big| \leq 100d\log n + m_0 + m_1 = o(n^2).$$

We apply Lemma 3.11 to obtain $\mathbb{E}(m_{2,j}) \leq n^3$ and hence

$$\mathbb{E}\left(\sum_{j\in J} m_{2,j}\right) = \mathbb{E}\left(\sum_{j\in[\kappa]} m_{2,j}\mathbb{1}_{j\in J}\right) \leq 5\kappa n^{3-3d}.$$

Then we apply Markov's inequality to obtain

$$\mathbb{P}\left(\sum_{j\in J} m_{2,j} > 5mn^{3-2d}\right) \leq \frac{\kappa}{mn^d} \leq n^{-d}. \tag{8.13}$$

For $d \geq 2$ and sufficiently large n, we combine (8.10), (8.12) and (8.13) to obtain

$$\mathbb{P}\Big(\big|\big\{m' < m : \mathrm{MaxLoad}^f(m') > \ell\big\}\big| > \varepsilon m\Big) < \frac{1}{n}.$$

This concludes the proof of (8.2). □

8.1 Proof of Proposition 8.3

We first make some technical observations on the parameters used in the Q-multi-scale strategy given in Section 2.4. Recall that $\alpha_1 = \frac{1}{2} + \frac{2}{\lfloor\sqrt{\log\log\log n}\rfloor+1/4}$, $L = (\log n)^{\frac{1+\alpha_1}{3}}$, $k = \big\lfloor \frac{\log\log n}{3\log\log\log n}\big\rfloor$, $N_i = \lceil \frac{L}{3k\ell_i}\rceil$ and $Q^{i,j} = (2k+1)(j-1)\ell_i$. We first have for $i \in \mathbb{N}, j \in [N_i]$ that

$$Q^{i,j} < L. \tag{8.14}$$

Observing from (2.16) that $\{\alpha_i\}_{i\in\mathbb{N}}$ is a non-decreasing sequence, we have for $i \geq 1$ and sufficiently large n that

$$\begin{aligned}(\log n)^{\alpha_i'-\alpha_i} &= (\log n)^{-\frac{1}{5}\cdot\frac{2\alpha_i-1-\varepsilon_i}{2k+1}} = (\log n)^{-\left(\frac{1}{5}-o(1)\right)\frac{\alpha_i-1/2}{k+1/2}}\\ &\leq (\log n)^{-\left(\frac{1}{5}-o(1)\right)\frac{\alpha_1-1/2}{k+1/2}} \leq (\log n)^{-\frac{2/5-o(1)}{\sqrt{\log\log\log n}}\cdot\frac{1}{k+1/2}}\\ &= (\log n)^{-\left(\frac{6}{5}-o(1)\right)\frac{\sqrt{\log\log\log n}}{\log\log n}}\\ &< e^{-\sqrt{\log\log\log n}}.\end{aligned} \tag{8.15}$$

Recall that $i_{\max} = \max\{i \in \mathbb{N} : \alpha_i \leq 1\}$. Using (2.16), we have for all $i \leq i_{\max}$ that

$$N_i = \Big\lceil \frac{L}{3k\ell_i}\Big\rceil = \frac{(1+o(1))L}{3k\ell_i} = \frac{(1+o(1))}{3k}(\log n)^{\frac{2\alpha_1-1}{6}-\frac{\alpha_i-1/2+k\varepsilon_i}{2k+1}} = (\log n)^{\frac{2\alpha_1-1}{6}-O\left(\frac{1}{k}\right)}. \tag{8.16}$$

Using (2.16) and (8.15) we observe that $N_i = (1-o(1))(\log n)^{\alpha_{i+1}-\alpha_i}$. This, together with (8.16), yields the iteration formula

$$\alpha_{i+1} = \alpha_i + \frac{2\alpha_1-1}{6} - O\left(\frac{1}{k}\right). \tag{8.17}$$

This, along with the definition of $i_{\max}$, implies that

$$i_{\max} \leq (1+o(1))\frac{6(1-\alpha_1)}{2\alpha_1-1} = \frac{3+o(1)}{2\alpha_1-1} < \sqrt{\log\log\log n}. \tag{8.18}$$

In addition, we have

$$(\log n)^{\alpha_i-\alpha_{i+1}'} = (\log n)^{\alpha_i-\alpha_{i+1}+\frac{1}{5}\cdot\frac{2\alpha_{i+1}-1-\varepsilon_{i+1}}{2k+1}} = (\log n)^{-\frac{2\alpha_1-1}{6}+\Theta\left(\frac{1}{k}\right)} = o(1). \tag{8.19}$$

The main technical instrument for establishing Proposition 8.3 is the following lemma, the proof of which is provided in the next subsection.

Lemma 8.6. *Consider the Q-multi-scale strategy with the initial load vector $\{L_p(0)\}_{p\in[n]}$ satisfying $L_p(0) \leq Q \leq L$ for all $p \in [n]$. Fix $c > 0$. Set $\ell = (\log n)^{\frac{1}{2}+\frac{1}{\sqrt{\log\log\log n}}}$. For any $s = \sum_{i=1}^{i_{\max}} j_i n(\lfloor\log^{\alpha_i} n\rfloor + \lfloor\log^{\alpha_i'} n\rfloor)$ with $0 \leq j_i \leq N_i - 1$, we have*

$$\mathbb{P}\Big(\mathrm{MaxLoad}^f([s, s+n\lfloor\log^{\alpha_1} n\rfloor]) > \ell\Big) < n^{-c}. \tag{8.20}$$

Proof of Proposition 8.3. We will show that the Q-multi-scale strategy with $Q \leq L$ satisfies the statement in Proposition 8.3. Recall that $i_{\max} = \max\{i \in \mathbb{N} : \alpha_i \leq 1\}$. We first show for sufficiently large n that

$$\alpha_{i_{\max}+1} \geq 1 + \frac{1}{30\sqrt{\log\log\log n}}. \tag{8.21}$$

To this end, we iterate equation (8.17) to obtain

$$\alpha_{i_{\max}+1} = \alpha_1 + i_{\max}\left(\frac{2\alpha_1 - 1}{6} - O\left(\frac{1}{k}\right)\right) = \frac{1}{2} + (i_{\max}+3)\cdot\frac{2\alpha_1-1}{6} - O\left(\frac{i_{\max}}{k}\right). \quad (8.22)$$

The monotonicity of $\{\alpha_i\}_{i\in\mathbb{N}}$ and the definition of $i_{\max}$ implies that $\alpha_{i_{\max}+1} > 1$. This inequality, equation (8.22) and the fact that $i_{\max}$ is an integer yield that

$$i_{\max} + 3 \geq \left\lceil \frac{6}{2\alpha_1 - 1}\left(\frac{1}{2} + O\left(\frac{i_{\max}}{k}\right)\right)\right\rceil. \quad (8.23)$$

Recall that $\alpha_1 = \frac{1}{2} + \frac{2}{\lfloor\sqrt{\log\log\log n}\rfloor + 1/4}$, $k = \lfloor \frac{\log\log n}{3\log\log\log n}\rfloor$ and the bound $i_{\max} < \sqrt{\log\log\log n}$ given in (8.18). Then, for sufficiently large n, we can further write inequality (8.23) as

$$i_{\max} + 3 \geq \left\lceil \frac{3\lfloor\sqrt{\log\log\log n}\rfloor + 3/4 + o(1)}{4}\right\rceil \geq \frac{3\lfloor\sqrt{\log\log\log n}\rfloor + 1}{4}.$$

Plugging this into (8.22), we obtain

$$\alpha_{i_{\max}+1} \geq 1 + \frac{1}{24\lfloor\sqrt{\log\log\log n}\rfloor + 6} - O\left(\frac{i_{\max}}{k}\right) \geq 1 + \frac{1}{30\sqrt{\log\log\log n}}.$$

This proves (8.21).

We next prove the main statement (8.3). For $m \in \mathbb{N}$, we define the set

$$A_m = \cup_{s\in J_m}\left\{m' \in \mathbb{N} : s \leq m' \leq \max\left\{m, s + n\lfloor\log^{\alpha_1} n\rfloor\right\}\right\},$$

where J_m is defined as

$$J_m = \left\{ s = \sum_{i\in\mathbb{N}} j_i n(\lfloor\log^{\alpha_i} n\rfloor + \lfloor\log^{\alpha_i'} n\rfloor) : 0 \leq j_i \leq N_i - 1,\ s < m\right\}.$$

Observe that by the condition of Proposition 8.3, we have

$$\alpha = \frac{\log(m/n)}{\log\log n} \leq 1 + \frac{1}{30\sqrt{\log\log\log n}} \leq \alpha_{i_{\max}+1},$$

and hence

$$\frac{m}{n} = \log^{\alpha} n \leq (\log n)^{\alpha_{i_{\max}+1}} = N_{i_{\max}}(\lfloor(\log n)^{\alpha_{i_{\max}}}\rfloor + \lfloor(\log n)^{\alpha'_{i_{\max}}}\rfloor).$$

Together with (8.21) we thus have

$$J_m = \left\{ s = \sum_{i=1}^{i_{\max}} j_i n(\lfloor\log^{\alpha_i} n\rfloor + \lfloor\log^{\alpha_i'} n\rfloor) : 0 \leq j_i \leq N_i - 1,\ s < m\right\}.$$

For any fixed constant $c > 0$, we apply Lemma 8.6 and the union bound argument to obtain

$$\begin{aligned}\mathbb{P}\Big(\mathrm{MaxLoad}^f(A_m) \geq \ell\Big) &\leq \sum_{s\in J_m}\mathbb{P}\Big(\mathrm{MaxLoad}^f([s, s+n\lfloor\log^{\alpha_1} n\rfloor]) \geq \ell\Big)\\ &\leq |J_m|\cdot n^{-2c} < n^{-c}, \end{aligned} \quad (8.24)$$

where the last equality follows from that $|J_m| \le m/n < \log^2 n$.

We next show that $|A_m|/m \ge 1-\varepsilon$. We define $i^* = \max\{i \in \mathbb{N} : n\log^{\alpha_i} n < m\}$. It is clear from the condition of Proposition 8.3 and (8.21) that $i^* \le i_{\max}$. We further denote

$$\begin{aligned}\xi &= \min\left\{\xi' \in \mathbb{N} : \xi' n(\lfloor \log^{\alpha_{i^*}} n\rfloor + \lfloor \log^{\alpha'_{i^*}} n\rfloor) \ge m\right\},\\ m_\xi &= \xi n(\lfloor \log^{\alpha_{i^*}} n\rfloor + \lfloor \log^{\alpha'_{i^*}} n\rfloor).\end{aligned}$$

Observe from the definition of i^* that $\xi \le N_{i^*}$. This, along with (8.15), implies that $m_\xi < 2m$. Hence it suffices to show that

$$\frac{|A_{m_\xi}|}{m_\xi} \ge 1 - \frac{\varepsilon}{2}. \tag{8.25}$$

For $1 \le i \le i^*$, we define

$$\begin{aligned}J_i^* &= \left\{\sum_{i'=i}^{i^*} j_{i'} n(\lfloor \log^{\alpha_{i'}} n\rfloor + \lfloor \log^{\alpha'_{i'}} n\rfloor) : 0 \le j_{i'} \le N_{i'} - 1\right\},\\ B_i &= \bigcup_{j=0}^{N_i-1} \Big(jn(\lfloor \log^{\alpha_i} n\rfloor + \lfloor \log^{\alpha'_i} n\rfloor) + C_i\Big), \text{ where}\\ C_i &= (0, n\lfloor \log^{\alpha_i} n\rfloor].\end{aligned}$$

Observe that $B_i + (J_{i+1}^* \cap [m_\xi]) = C_i + (J_i^* \cap [m_\xi])$ consists of a disjoint union of shifted copies of B_i and that $C_{i+1} + (J_{i+1}^* \cap [m_\xi])$ consists of a disjoint union of shifted copies of C_{i+1}. We thus obtain

$$\frac{|C_i + (J_i^* \cap [m_\xi])|}{|C_{i+1} + (J_{i+1}^* \cap [m_\xi])|} = \frac{|B_i + (J_{i+1}^* \cap [m_\xi])|}{|C_{i+1} + (J_{i+1}^* \cap [m_\xi])|} = \frac{|B_i|}{|C_{i+1}|}.$$

By (2.16) and (8.15), we obtain

$$\frac{|B_i|}{|C_{i+1}|} = \frac{\lfloor \log^{\alpha_i} n\rfloor}{\lfloor \log^{\alpha_i} n\rfloor + \lfloor \log^{\alpha'_i} n\rfloor} = 1 - \frac{\lfloor \log^{\alpha'_i} n\rfloor}{\lfloor \log^{\alpha_i} n\rfloor + \lfloor \log^{\alpha'_i} n\rfloor} > 1 - (\log n)^{\alpha'_i - \alpha_i} > 1 - \delta,$$

where $\delta = e^{-\sqrt{\log\log\log n}}$. Moreover, we have

$$\frac{|C_{i^*} + (J_{i^*}^* \cap [m_\xi])|}{m_\xi} = \frac{\lfloor \log^{\alpha_{i^*}} n\rfloor}{\lfloor \log^{\alpha_{i^*}} n\rfloor + \lfloor \log^{\alpha'_{i^*}} n\rfloor} > 1 - \delta.$$

Iterating these observations we obtain

$$\begin{aligned}\frac{|A_{m_\xi}|}{m_\xi} &= \frac{|C_1 + (J_1^* \cap [m_\xi])|}{m_\xi} = \frac{|C_{i^*} + (J_{i^*}^* \cap [m_\xi])|}{m_\xi} \cdot \frac{|C_1 + (J_1^* \cap [m_\xi])|}{|C_{i^*} + (J_{i^*}^* \cap [m_\xi])|}\\ &= \frac{|C_{i^*} + (J_{i^*}^* \cap [m_\xi])|}{m_\xi} \cdot \prod_{i=1}^{i^*-1} \frac{|C_i + (J_i^* \cap [m_\xi])|}{|C_{i+1} + (J_{i+1}^* \cap [m_\xi])|}\\ &\ge (1-\delta)^{i_{\max}} > 1 - \sqrt{\log\log\log n}\; e^{-\sqrt{\log\log\log n}},\end{aligned}$$

where the inequalities follow from the fact that $i^* \le i_{\max} < \sqrt{\log\log\log n}$. This completes the proof of (8.25). □

8.1.1 Proof of Lemma 8.6

For $i \in \mathbb{N}, j \in [N_i]$, we write $s_i^j = (j-1)n(\lfloor \log^{\alpha_i} n \rfloor + \lfloor \log^{\alpha_i'} n \rfloor)$ and $t_i^j = s_i^j + n\lfloor \log^{\alpha_i} n \rfloor$. Hence, $(s_i^j, t_i^j]$ and $(t_i^j, s_i^{j+1}]$ are discrete time intervals in the $(i+1)$-th scale where we apply the j-th iteration of the i-th scale strategy and the j-th iteration of the regulating multi-stage threshold strategy, respectively. Fix $c > 0$. We set $\alpha_0 = \frac{1+\alpha_1}{3} + \frac{\log(12c+9)}{\log\log n}$ so that $\log^{\alpha_0} n = (12c+9)L$. One can check that $\alpha_0 < \alpha_1$ for n large enough. We introduce the following events

$$\begin{aligned} E_Q^{i,j} &= \left\{ \mathrm{MaxLoad}^f([s_i^j, t_i^j]) \le \log^{\alpha_{i-1}} n + Q^{i,j} + Q \right\}, \\ F_Q^{i,j} &= \left\{ \mathrm{MaxLoad}^f\left(s_i^{j+1}\right) \le Q^{i,j+1} + Q \right\}, \\ G_Q^{j,i} &= \left\{ |H_Q^{i,j}| \le 3n \exp\left(-\tfrac{\ell_i^2}{4\log^{\alpha_i} n}\right) \right\}, \text{ where} \\ H_Q^{i,j} &= \left\{ p \in [n] : L_p(t_i^j) \ge Q^{i,j} + \ell_i + Q \right\}. \end{aligned} \tag{8.26}$$

In this subsection, in order to simplify the notations, we denote by $\overline{E}$ the complement of the event E. The following result plays a key role in establishing Lemma 8.6.

Lemma 8.7. *Consider the Q-multi-scale strategy with the initial load vector $\{L_p(0)\}_{p\in[n]}$ satisfying $L_p(0) \le Q \le 2L\sqrt{\log\log\log n}$ for all $p \in [n]$. Fix $c > 0$. For sufficiently large n and all $i \in \mathbb{N}$ such that $\alpha_i \le 1$, we have*

$$\mathbb{P}\left(\bigcup_{j\in[N_i]} \overline{E_Q^{i,j} \cap F_Q^{i,j} \cap G_Q^{i,j}} \right) \le n^{-c}. \tag{8.27}$$

Proof of Lemma 8.6. Recall the notation $Q^{i,j} = (2k+1)(j-1)\ell_i$. The statement (8.20) is a consequence of the following stronger statement

$$\mathbb{P}\left(\mathrm{MaxLoad}^f([s, s+n\lfloor \log^{\alpha_1} n \rfloor]) > Q + \log^{\alpha_0} n + \sum_{i\in[i_{\max}]} Q^{i,j_i+1} \right) < n^{-c+o(1)}. \tag{8.28}$$

Recall that $L = (\log n)^{\frac{1}{2} + \frac{2}{3(\lfloor\sqrt{\log\log\log n}\rfloor + 1/4)}}$, $Q \le L$ and $\log^{\alpha_0} n = (12c+9)L$. These, together with (8.18) and (8.14), yield

$$Q + \log^{\alpha_0} n + \sum_{i\in[i_{\max}]} Q^{i,j_i+1} \le Q + (12c+9)L + i_{\max}L < 2L\sqrt{\log\log\log n} < \ell.$$

For $0 \le i \le i_{\max}$, we write

$$s_i = \sum_{h=i+1}^{i_{\max}} j_h n(\lfloor \log^{\alpha_h} n \rfloor + \lfloor \log^{\alpha_h'} n \rfloor)$$

so that $s_0 = s$ and $s_{i_{\max}} = 0$. We further denote $Q_i = Q + \sum_{h=i+1}^{i_{\max}} Q^{h,j_h+1}$ so that $Q_{i_{\max}} = Q$ and define $K_i = \left\{ \mathrm{MaxLoad}^f(s_i) \le Q_i \right\}$. Keeping the notations $P_{Q'}(\cdot)$ as in the proof of Lemma 8.7, we have

$$\mathbb{P}\Big(\mathrm{MaxLoad}^f([s, s+n\lfloor \log^{\alpha_1} n \rfloor]) > Q_0 + \log^{\alpha_0} n \,|\, K_1 \Big) < P_{Q_1}\left(\overline{E_{Q_1}^{1,j_1+1}} \right). \tag{8.29}$$

Since $Q_1 < 2L\sqrt{\log\log\log n}$, we may apply Lemma 8.7 to obtain

$$P_{Q_1}\left(\overline{E_{Q_1}^{1,j_1+1}}\right) < n^{-2c}. \tag{8.30}$$

Next, we estimate $\mathbb{P}(K_1)$. As mentioned in the proof of Lemma 8.7, the j-th iteration of the i-th scale of the Q-multi-scale strategy is identical to the first iteration of the i-th scale of the $(Q+Q^{i,j})$-multi-scale strategy. This self-similar property implies that

$$\mathbb{P}\big(\overline{K_i} \cap K_{i+1}\big) < P_{Q_{i+1}}\left(\overline{F_{Q_{i+1}}^{i+1,j_{i+1}}}\right).$$

Using this inequality and the fact $K_{i_{\max}} = \big\{\mathrm{MaxLoad}^f(0) \le Q\big\}$ which is trivially satisfied by the starting conditions, we obtain

$$\mathbb{P}\left(\overline{K_1}\right) = \mathbb{P}\left(\bigcup_{i=1}^{i_{\max}-1}(\overline{K_i}\cap K_{i+1})\right) \le \sum_{i=1}^{i_{\max}-1} P_{Q_{i+1}}\left(\overline{F_{Q_{i+1}}^{i+1,j_{i+1}}}\right) < (i_{\max}-1)\cdot n^{-2c}, \tag{8.31}$$

where the last inequality uses Lemma 8.7, which is applicable since $Q_i < 2L\sqrt{\log\log\log n}$. Combining (8.29), (8.30), (8.31) and the fact that $i_{\max} < \sqrt{\log\log\log n}$, we have for sufficiently large n that

$$\mathbb{P}\left(\mathrm{MaxLoad}^f([s, s+n\lfloor\log^{\alpha_1} n\rfloor]) > Q + \log^{\alpha_0} n\right) < i_{\max}\cdot n^{-2c} < n^{-c}.$$

This concludes the proof of (8.28), and hence (8.20). □

We now present a couple of auxiliary lemmata that are used in our proof of Lemma 8.7. The first lemma provides an upper bound on the number of bins with loads above certain level.

Lemma 8.8. *Let $t > \ell > 0$, $h, r^* > 0$ and $p \in [0,1]$. Let $\{L_i(0)\}_{i\in[n]}$ be an initial load vector such that $L_i(0) \le L_0$ for all $i \in [n]$. Let f be any two-thinning strategy, which satisfies that $\mathbb{P}\big(\mathrm{MaxLoad}^f([tn]) \le h\big) \ge 1-p$. Define $H = \big\{i \in [n] : L_i(tn) \ge L_0 + \ell\big\}$. Then we have*

$$\mathbb{P}\left(|H| > 2n\exp\left(-\frac{\ell^2}{4t}\right) + r^*\right) \le 2\exp\left(-2n\exp\left(-\frac{\ell^2}{2t}\right)\right) + 4\exp\left(-\frac{n(r^*/tn)^h}{e(h+1)!}\right) + p.$$

Given an event E and $Q \ge 0$, we write $P_Q(E)$ for the maximum probability of E under the Q-multi-scale strategy with the initial maximum load bounded above by Q. Then the second lemma is as follows.

Lemma 8.9. *Consider the Q-multi-scale strategy with the initial load vector $\{L_p(0)\}_{p\in[n]}$ satisfying $L_p(0) \le Q \le 3L\sqrt{\log\log\log n}$ for all $p \in [n]$. Fix $c > 0$. For sufficiently large n and all $i \in \mathbb{N}$ such that $\alpha_i \le 1$ and all $j \in [N_i]$, we have*

$$P_Q\left(\overline{E_Q^{1,1}}\right) < n^{-c}, \tag{8.32}$$

$$P_Q\left(\overline{F_Q^{i,j}}, E_Q^{i,j}, G_Q^{i,j}\right) < n^{-e^{\sqrt{\log\log\log n}}}, \tag{8.33}$$

$$P_Q\left(\overline{G_Q^{i,j}}, E_Q^{i,j}, F_Q^{i,j-1}\right) < 2\exp\left(-n^{1/2-o(1)}\right). \tag{8.34}$$

With these two lemmata at hand, we now prove Lemma 8.7.

Proof of Lemma 8.7. For $i \in \mathbb{N}, j \in [N_i]$, we write

$$U_Q^{i,j} = \bigcup_{j' \le j} \overline{E_Q^{i,j'} \cap F_Q^{i,j'} \cap G_Q^{i,j'}}$$

with $U_Q^{i,0} = \emptyset$. Observe that for $i \in \mathbb{N}, j \in [N_i]$ we have

$$U_Q^{i,j} = \left(\overline{F_Q^{i,j}} \cap E_Q^{i,j} \cap G_Q^{i,j}\right) \cup \left(\overline{G_Q^{i,j}} \cap E_Q^{i,j} \cap F_Q^{i,j-1}\right) \cup \left(\overline{E_Q^{i,j}} \cap F_Q^{i,j-1}\right) \cup U_Q^{i,j-1}.$$

Notice that this indeed holds for $j = 1$ since the initial load condition implies that $F_Q^{i,0} = \Omega$. This, along with (8.33), (8.34) from Lemma 8.9, yields

$$\begin{aligned} P_Q\left(U_Q^{i,j}\right) &\le P_Q\left(\overline{F_Q^{i,j}}, E_Q^{i,j}, G_Q^{i,j}\right) + P_Q\left(\overline{G_Q^{i,j}}, E_Q^{i,j}, F_Q^{i,j-1}\right) \\ &\quad + P_Q\left(\overline{E_Q^{i,j}}, F_Q^{i,j-1}\right) + P_Q\left(U_Q^{i,j-1}\right) \\ &\le P_Q\left(\overline{E_Q^{i,j}}, F_Q^{i,j-1}\right) + P_Q\left(U_Q^{i,j-1}\right) + n^{-e^{\sqrt{\log\log\log n}}} + 2\exp\left(-n^{1/2-o(1)}\right). \end{aligned} \tag{8.35}$$

Observe that the j-th iteration of the i-th scale of the Q-multi-scale strategy is identical to the first iteration of the i-th scale of the $(Q + Q^{i,j})$-multi-scale strategy. Recall that the event $F_Q^{i,j-1}$ asserts that the load at time s_i^j is at most $Q + Q^{i,j}$. Hence we have

$$P_Q\left(\overline{E_Q^{i,j}}, F_Q^{i,j-1}\right) \le P_{Q+Q^{i,j}}\left(\overline{E_{Q+Q^{i,j}}^{i,1}}\right). \tag{8.36}$$

Iteration of (8.35) and the above inequality yield

$$P_Q\left(U_Q^{i,j}\right) \le \sum_{j' \le j} P_{Q+Q^{i,j'}}\left(\overline{E_{Q+Q^{i,j'}}^{i,1}}\right) + n^{-\omega(1)}. \tag{8.37}$$

In order to iterate this inequality, we now show that for all $Q' > 0$ and $i \ge 2$ the following inclusion inequality holds

$$\overline{E_{Q'}^{i,1}} \subset U_{Q'}^{i-1,N_{i-1}}. \tag{8.38}$$

To see this, we define the event

$$\hat{E}_Q^{i,j} = \left\{\mathrm{MaxLoad}^f\left((s_i^j, s_i^{j+1}]\right) \le \log^{\alpha_i'} n + Q^{i,j+1} + Q\right\}.$$

The statement (8.38) follows from the monotonicity of $U_{Q'}^{i,j}$ and the following inclusion relations

$$\overline{E_{Q'}^{i,1}} \subset \left(\cup_{j \in [N_{i-1}]} \overline{\hat{E}_{Q'}^{i-1,j}}\right), \tag{8.39}$$

$$\overline{\hat{E}_{Q'}^{i,j}} \subset \left(\overline{E_{Q'}^{i,j}} \cup \overline{F_{Q'}^{i,j}}\right) \subset U_{Q'}^{i,j}. \tag{8.40}$$

To see (8.39), observe that $\overline{E_{Q'}^{i,1}}$ asserts that over $(0, t_i^1] = \cup_{j \in [N_{i-1}]}(s_{i-1}^j, s_{i-1}^{j+1}]$, the maximum load is greater than $\log^{\alpha_{i-1}} n + Q'$, while $\overline{\hat{E}_{Q'}^{i-1,j}}$ asserts that over $(s_{i-1}^j, s_{i-1}^{j+1}]$, the maximum load is greater than $\log^{\alpha_{i-1}'} n + Q^{i-1,j+1} + Q'$. Using (8.15), we have $\log^{\alpha_{i-1}'} n = o(\log^{\alpha_{i-1}} n)$ and by (8.14) we have $Q^{i-1,j+1} < L = o(\log^{\alpha_{i-1}} n)$. These observations yield (8.39). To see (8.40), observe that

$$\mathrm{MaxLoad}^f(s_i^{j+1}) \ge \mathrm{MaxLoad}^f\left((t_i^j, s_i^{j+1}]\right) - |(t_i^j, s_i^{j+1}]| = \mathrm{MaxLoad}^f\left((t_i^j, s_i^{j+1}]\right) - \log^{\alpha_i'} n.$$

Hence, whenever $F_{Q'}^{i,j}$ occurs, we have $\mathrm{MaxLoad}^f\left((t_i^j, s_i^{j+1}]\right) \le \log^{\alpha_i'} n + Q^{i,j+1} + Q'$. This, along with $E_{Q'}^{i,j}$, implies that $\left(E_{Q'}^{i,j} \cap F_{Q'}^{i,j}\right) \subset \hat{E}_{Q'}^{i,j}$, which is equivalent to the first inclusion inequality in (8.40). The second inclusion inequality in (8.40) is trivial.

Then we can use (8.38) to iterate (8.37) and obtain

$$P_Q\left(U_Q^{i,j}\right) \le \sum_{j_i \le j} \sum_{j_{i-1} \le N_{i-1}} \cdots \sum_{j_1 \le N_1} P_{Q+Q^{i,j_i}+\cdots+Q^{1,j_1}}\left(\overline{E_{Q+Q^{i,j_i}+\cdots+Q^{1,j_1}}^{1,1}}\right) + n^{-\omega(1)}.$$

One can use (8.18) and (8.14) to check that

$$Q + Q^{i,j_i} + \cdots + Q^{1,j_1} \le Q + iL \le Q + i_{\max} L \le 3L\sqrt{\log\log\log n}.$$

Then we apply (8.32) from Lemma 8.9, (8.16) and (8.18) to obtain for sufficiently large n that

$$\begin{aligned} P_Q\left(U_Q^{i,j}\right) &\le \left(\prod_{i'=1}^{i} N_{i'}\right) n^{-2c} + n^{-\omega(1)} \le (\log n)^{i_{\max}\cdot\frac{2\alpha_1-1}{6}} n^{-2c} + n^{-\omega(1)} \\ &= (\log n)^{O(1)} \cdot n^{-2c} + n^{-\omega(1)} \le n^{-c}. \end{aligned}$$

This concludes the proof. □

8.1.2 Proofs of Lemmata 8.8 and 8.9

Proof of Lemma 8.8. We denote by r the total number of retries up to time tn and by H' the set of bins which are suggested as primary allocations at least $t + \ell$ times by time tn. Then, we have

$$|H| \le |H'| + r.$$

Hence, we have

$$\mathbb{P}\left(|H| > 2n\exp\left(-\frac{\ell^2}{4t}\right) + r^*\right) \le \mathbb{P}\left(|H'| > 2n\exp\left(-\frac{\ell^2}{4t}\right)\right) + \mathbb{P}(r > r^*). \tag{8.41}$$

We now estimate the first term. We denote by $\{X_i\}_{i\in[n]}$ independent Poisson(t) random variables. Write Y_i for the indicator function of the event $\{X_i \ge t + \ell\}$ and $Y = \sum_{i=1}^n Y_i$. By Lemma 3.3, we have

$$\mathbb{P}(Y_i = 1) = \mathbb{P}(X_i \ge t + \ell) \le e^{-tI(\ell/t)} \le \exp\left(-\frac{\ell^2}{4t}\right),$$

where the second inequality follows from the lower bound of $I(x)$ in (3.3) and the assumption that $\ell/t < 1$. Lemma 3.1 and Hoeffding's inequality imply that

$$\mathbb{P}\left(|H'| > 2n\exp\left(-\frac{\ell^2}{4t}\right)\right) \le 2\mathbb{P}\left(Y > 2n\exp\left(-\frac{\ell^2}{4t}\right)\right) \le 2\exp\left(-2n\exp\left(-\frac{\ell^2}{2t}\right)\right). \tag{8.42}$$

Next, we estimate the second term. Set $E = \{\mathrm{MaxLoad}^f([tn]) \le h\}$. By the law of total probability,

$$\begin{aligned} \mathbb{P}(E) &= \mathbb{P}(E, r \ge r^*) + \mathbb{P}(E, r < r^*) \le \mathbb{P}(E, r \ge r^*) + \mathbb{P}(r < r^*) \\ &= \mathbb{P}(E, r \ge r^*) + 1 - \mathbb{P}(r \ge r^*). \end{aligned} \tag{8.43}$$

Recall that R_k given in (2.1) is the number of retries after allocating k balls. We denote by $s_0 = \inf\left\{s \in [t] : R_{sn} - R_{(s-1)n} \geq r^*/t\right\}$. Whenever $\{r \geq r^*\}$ occurs, we have $s_0 < \infty$. Write $S = \left\{i \in [n] : L_i^f((s_0-1)n) \geq 0\right\}$. As per (5.2), we show that whenever E occurs, then

$$|S| \geq \frac{n}{h+1}. \tag{8.44}$$

To see this, observe that

$$0 = \sum_{i\in[n]} L_i^f((s_0-1)n) = \sum_{i\in S} L_i^f((s_0-1)n) + \sum_{i\in S^c} L_i^f((s_0-1)n).$$

This, together with the fact that $\left\{L_i^f((s_0-1)n)\right\}_{i\in[n]} \in \mathbb{Z}^n$ and $\mathrm{MaxLoad}^f((s_0-1)n) < h$, yields

$$|S^c| \leq \sum_{i\in S^c} |L_i^f((s_0-1)n)| = \sum_{i\in S} L_i^f((s_0-1)n) \leq |S| \cdot (h+1).$$

Then we can obtain (8.44) using $|S^c| = n - |S|$.

Denote by $\{Z_i\}_{i\in[n]}$ independent $\mathrm{Poisson}\,(r^*/tn)$ random variables. By Lemma 3.1 and Lemma 3.2,

$$\mathbb{P}(r > r^*, E) \leq 2\mathbb{P}\left(\max_{i\in S} Z_i \leq h\right) \leq 4\exp\left(-\frac{n(r^*/tn)^h}{e(h+1)!}\right). \tag{8.45}$$

Inequalities (8.43), (8.45) and the fact that $\mathbb{P}(E) \geq 1-p$ imply that

$$\mathbb{P}(r > r^*) \leq 4\exp\left(-\frac{n(r^*/tn)^h}{e(h+1)!}\right) + p.$$

We can conclude the proof by combining this with (8.41) and (8.42). □

Proof of Lemma 8.9. **Proof of** (8.32)**.** The statement readily follows from the application of Proposition 6.1 with the parameters $L_0 := Q$, $t := \lfloor \log^{\alpha_1} n \rfloor$, $\ell := L = (\log n)^{\frac{1+\alpha_1}{3}}$ and our definition of α_0 such that $\log^{\alpha_0} n = (12c+9)L$.

Proof of (8.33)**.** The statement follows from the application of Proposition 4.3 with the parameters $t_0 := t_i^j$, $t := s_i^{j+1}$, $\alpha := \alpha_i'$, $\eta := \alpha_i - \alpha_i'$, $L_0 := Q^{i,j} + \ell_i + Q$. Hence it suffice to show that the conditions of Proposition 4.3 are satisfied.

We first verity the technical requirement $\eta \leq \frac{\alpha-1/2}{4k-2}$, which is assumed in our definition of the multi-stage threshold strategy in Section 2.4. Using $\eta = \alpha_i - \alpha_i', \alpha = \alpha_i'$ and (2.15), we can rewrite this requirement as

$$\frac{\alpha_i - 1/2 - \varepsilon_i/2}{5k+5/2} \leq \frac{\alpha_i - 1/2}{4k-1},$$

which clearly holds.

We next show that both assumptions in Proposition 4.3 hold when $E_Q^{i,j}$ and $G_Q^{i,j}$ occur. Given the event $G_Q^{i,j}$, the second assumption trivially holds. We now verify the first assumption that $\mathrm{MaxLoad}^f(t_i^j) = o(t-t_0)$. Assuming the event $E_Q^{i,j}$, we have

$$\mathrm{MaxLoad}^f(t_i^j) \leq \log^{\alpha_{i-1}} n + Q^{i,j} + Q \leq \log^{\alpha_{i-1}} n + L + Q,$$

where the last inequality follows from (8.14). Recall $\alpha_1 = \frac{1}{2} + \frac{2}{\lfloor\sqrt{\log\log\log n}\rfloor+1/4}$, $L = (\log n)^{\frac{1+\alpha_1}{3}}$, $Q \le 3L\sqrt{\log\log\log n}$ and $k = \lfloor\frac{\log\log n}{3\log\log\log n}\rfloor$. We have

$$L + Q < 4L\sqrt{\log\log\log n} = (\log n)^{\frac{1}{2}+\frac{2}{3\sqrt{\log\log\log n}}+O(\frac{\log\log\log\log n}{\log\log n})},$$

while

$$t - t_0 = \log^{\alpha_i'} n = (\log n)^{\alpha_i - O(\frac{1}{k})} > (\log n)^{\alpha_1 - O(\frac{1}{k})} = (\log n)^{\frac{1}{2}+\frac{2}{\sqrt{\log\log\log n}+1/4}-O(\frac{1}{k})}.$$

These, together with (8.19), verify the first assumption of Proposition 4.3. Hence, we can apply Proposition 4.3 to obtain (8.33).

Proof of (8.34)**.** Recall our definition $G_Q^{i,j} = \{|H_Q^{i,j}| \le 3n\exp\left(-\frac{\ell_i^2}{4\log^{\alpha_i} n}\right)\}$. We introduce

$$\tilde{G}_Q^{i,j} = \left\{|H_Q^{i,j}| \le 2n\exp\left(-\frac{\ell_i^2}{4\log^{\alpha_i} n}\right) + n\exp\left(-\frac{\log n}{2(\log^{\alpha_{i-1}} n + L + Q)}\right)\right\}.$$

We will show that $\tilde{G}_Q^{i,j} \subset G_Q^{i,j}$ and that

$$P\left(\overline{\tilde{G}_Q^{i,j}}, E_Q^{i,j}, F_Q^{i,j-1}\right) \le 2\exp\left(-n^{1/2-o(1)}\right), \tag{8.46}$$

which implies (8.34).

To see $\tilde{G}_Q^{i,j} \subset G_Q^{i,j}$, it suffice to show that $\frac{\ell_i^2}{\log^{\alpha_i} n} = o\left(\frac{\log n}{\log^{\alpha_{i-1}} n + L + Q}\right)$. We recall that $\alpha_0 = \frac{1+\alpha_1}{3} + \Theta(\frac{1}{\log\log n})$, $\alpha_1 = \frac{1}{2} + \Theta(\frac{1}{\sqrt{\log\log\log n}})$, $\ell_i = (\log n)^{\frac{1}{2}+O(\frac{1}{k})}$ and $k = \lfloor\frac{\log\log n}{3\log\log\log n}\rfloor$. Hence, using again $\log^{\alpha_0} n = (12c+9)L$, we have

$$\begin{aligned}
\frac{\ell_1^2}{\log^{\alpha_1} n}\Big/\frac{\log n}{\log^{\alpha_0} n + L + Q} &= \frac{\log^{\alpha_0} n + L + Q}{\log^{\alpha_0} n} \cdot \frac{\ell_1^2/\log^{\alpha_1} n}{\log n/\log^{\alpha_0} n} \\
&= \frac{(12c+10)L + Q}{(12c+9)L} \cdot (\log n)^{-\frac{2\alpha_1-1}{3}+O(\frac{1}{k})} \\
&< O\big(\sqrt{\log\log\log n}\big) \cdot (\log n)^{-\frac{2\alpha_1-1}{3}+O(\frac{1}{k})} = o(1),
\end{aligned}$$

where the inequality follows from that $Q \le 3L\sqrt{\log\log\log n}$. For $i \ge 2$, we use the fact that $\log^{\alpha_{i-1}} n + L + Q < 2\log^{\alpha_{i-1}} n$ to obtain

$$\begin{aligned}
\frac{\ell_i^2}{\log^{\alpha_i} n}\Big/\frac{\log n}{\log^{\alpha_{i-1}} n + L + Q} &< \frac{2\ell_i^2}{(\log n)^{1+\alpha_i-\alpha_{i-1}}} = 2(\log n)^{-(\alpha_i-\alpha_{i-1})+O(\frac{1}{k})} \\
&= (\log n)^{-\frac{2\alpha_1-1}{6}+O(\frac{1}{k})} = o(1),
\end{aligned}$$

where the second identity follows from (8.17).

Towards showing inequality (8.46), we observe that given $E_Q^{i,j}$ and $F_Q^{i,j-1}$, we can apply Lemma 8.8 to the process started at time s_i^j with $L_0 = Q^{i,j} + Q$, $t = \lfloor\log^{\alpha_i} n\rfloor$, $\ell = \ell_i$, $p = 0$, $h = \log^{\alpha_{i-1}} n + Q^{i,j} + Q$ and $r^* = n\exp\left(-\frac{\log n}{2(\log^{\alpha_{i-1}} n + L + Q)}\right)$ to obtain

$$\begin{aligned}
&P\left(\left\{|H_i^j| > 2n\exp\left(-\frac{\ell_i^2}{4\log^{\alpha_i} n}\right) + n\exp\left(-\frac{\log n}{2(\log^{\alpha_{i-1}} n + L + Q)}\right)\right\} \cap E_Q^{i,j} \cap F_Q^{i,j-1}\right) \\
&\quad \le 2\exp\left(-2n\exp\left(-\frac{\ell_i^2}{2\lfloor\log^{\alpha_i} n\rfloor}\right)\right) + 4\exp\left(-\frac{\sqrt{n}(\log n)^{-\alpha_i(\log^{\alpha_{i-1}} n + L + Q)}}{e\lceil\log^{\alpha_{i-1}} n + L + Q\rceil!}\right),
\end{aligned} \tag{8.47}$$

where, in the second term of (8.47), we use the fact that $h < \log^{\alpha_{i-1}} n + L + Q$. For the first term of (8.47), we have

$$2\exp\left(-2n\exp\left(-\frac{\ell_i^2}{2\lfloor\log^{\alpha_i} n\rfloor}\right)\right) = \exp\left(-n^{1-o(1)}\right). \tag{8.48}$$

The second term of (8.47) is increasing with respect to α_{i-1}, which, in turn, is increasing with respect to i. Hence, we can assume that $i \geq 2$ and use $\log^{\alpha_{i-1}} n + L + Q < 2\log^{\alpha_{i-1}} n$ to obtain

$$\begin{aligned}
4\exp\left(-\frac{\sqrt{n}(\log n)^{-\alpha_i(\log^{\alpha_{i-1}} n + L + Q)}}{e\lceil\log^{\alpha_{i-1}} n + L + Q\rceil!}\right) &\leq 4\exp\left(-\frac{\sqrt{n}(\log n)^{-2\alpha_i(\log n)^{\alpha_{i-1}}}}{(2\log^{\alpha_{i-1}} n)^{2(\log n)^{\alpha_{i-1}}}}\right) \\
&\leq 4\exp\left(-\frac{\sqrt{n}}{(2\log n)^{2(\alpha_i+\alpha_{i-1})(\log n)^{\alpha_{i-1}}}}\right) \\
&\leq 4\exp\left(-\frac{\sqrt{n}}{\exp(5\log\log n\cdot\log^{\alpha_{i-1}} n)}\right) \\
&= \exp\left(-n^{1/2-o(1)}\right),
\end{aligned} \tag{8.49}$$

where the first inequality follows from Stirling's approximation $n! \leq e\sqrt{n}(n/e)^n$, and the last inequality – from the observation that $\alpha_{i-1} < 1 - \frac{1}{4\sqrt{\log\log\log n}}$, which, in turn, follows from the fact that $\alpha_i \leq 1$ and (8.17). Plugging and (8.48), (8.49) into (8.47), inequality (8.46), and hence (8.34), follows. □

8.2 Proof of Proposition 8.4

Proof of Proposition 8.4. To establish equality (8.4), it would clearly suffice to show the following estimates

$$\mathbb{P}\Big(\mathrm{MaxLoad}^f(m_1) > A\Big) \leq n^{-4d}, \tag{8.50}$$

$$\mathbb{P}\left(\min_{i\in[n]} L_i^f(m_1) < -A\right) \leq n^{-4d}. \tag{8.51}$$

We first show that inequality (8.50) follows from Lemma 8.7. Our choice of m_1 in (2.18) guarantees that the allocation of m_1 balls using the Q-multi-scale strategy ends up with $N_{i_{\max}}$ complete iterations of the $i_{\max}$-th scale strategy followed by the regulating multi-stage threshold strategy. Recall the definition of $F_Q^{i,j}$ given in (8.26) and apply Lemma 8.7 to obtain

$$\mathbb{P}\left(\mathrm{MaxLoad}^f(m_1) > Q + Q^{i_{\max},N_{i_{\max}}+1}\right) = \mathbb{P}\left(\overline{F_Q^{i_{\max},N_{i_{\max}}}}\right) \leq n^{-4d}.$$

Using (8.14), we have $Q + Q^{i_{\max},N_{i_{\max}}+1} < Q + L < A$. The two inequalities above yield (8.50).

Next, we estimate $\mathbb{P}(L_i^f(m_1) < -A)$, which, together with the union bound argument, implies inequality (8.51). For each $i \in [n]$, we denote

$$k_i = \sup\{k \in [1, m_1] : L_i(k) \geq -300d\log n\}$$

and write F_i for the event $\{-\infty < k_i < m_1 - \log n\}$. Observe that, given our assumptions on $L_i(0)$, on F_i^c we have $L_i^f(m_1) \geq -A$ almost surely. We denote by $\mathcal{F}_k$ the filtration generated

by $\{L_i^f(p)\colon 1 \le p \le k, i \in [n]\}$. By Chernoff's argument, we thus have that for any $\lambda > 0$,

$$\mathbb{P}\big(L_i^f(m_1) < -A \,|\, F_i, \mathcal{F}_{k_i}\big) \le e^{-\lambda A} \cdot \mathbb{E}\big[e^{-\lambda L_i^f(m_1)} \,|\, F_i, \mathcal{F}_{k_i}\big]. \tag{8.52}$$

We write $p_{k,\mathcal{F}_{k-1}} = \mathbb{P}\big(L_i^f(k) - L_i^f(k-1) = 1 - 1/n \,|\, \mathcal{F}_{k-1}\big)$, i.e., the probability that the k-th ball has been allocated to the i-th bin conditioned on the load vector in time $k-1$. Our strategy never retries a ball if its primary allocation is a bin with load below $-\log n$. This and the definition of k_i imply that $p_{k,\mathcal{F}_{k-1}} \ge 1/n$ for all $k_i < k \le m_1$. We now compute

$$\begin{aligned}
\mathbb{E}\big[e^{-\lambda L_i^f(m_1)} \,|\, F_i, \mathcal{F}_{k_i}\big] &= \mathbb{E}\big[e^{-\lambda L_i^f(m_1-1)} \cdot \mathbb{E}\big[e^{-\lambda(L_i^f(m_1) - L_i^f(m_1-1))} \,|\, \mathcal{F}_{m_1-1}\big] \,|\, F_i, \mathcal{F}_{k_i}\big] \\
&\le \mathbb{E}\big[e^{-\lambda L_i^f(m_1-1)} \cdot \big(p_{m_1,\mathcal{F}_{m_1-1}} e^{-\lambda(1-1/n)} + (1 - p_{m_1,\mathcal{F}_{m_1-1}}) e^{\lambda/n}\big) \,|\, F_i, \mathcal{F}_{k_i}\big] \\
&= \mathbb{E}\big[e^{-\lambda L_i^f(m_1-1)} \cdot e^{\lambda/n} \cdot \big(1 - (1 - e^{-\lambda}) p_{m_1,\mathcal{F}_{m_1-1}}\big) \,|\, F_i, \mathcal{F}_{k_i}\big] \\
&\le \mathbb{E}\big[e^{-\lambda L_i^f(m_1-1)} \,|\, F_i, \mathcal{F}_{k_i}\big] \cdot e^{\lambda/n} \cdot \left(1 - \frac{1 - e^{-\lambda}}{n}\right),
\end{aligned}$$

Iterate this inequality to obtain

$$\begin{aligned}
\mathbb{E}\big[e^{-\lambda L_i^f(m_1)} \,|\, F_i, \mathcal{F}_{k_i}\big] &\le \mathbb{E}\big[e^{-\lambda L_i^f(k_i)} \,|\, F, \mathcal{F}_{k_i}\big] \cdot \left(e^{\lambda/n} \cdot \left(1 - \frac{1 - e^{-\lambda}}{n}\right)\right)^{m_1 - k_i} \\
&= e^{-\lambda L_i^f(k_i)} \cdot \left(e^{\lambda/n} \cdot \left(1 - \frac{1 - e^{-\lambda}}{n}\right)\right)^{m_1 - k_i} \\
&\le e^{300\lambda d \log n} \cdot \left(e^{\lambda/n} \cdot \left(1 - \frac{1 - e^{-\lambda}}{n}\right)\right)^{m_1},
\end{aligned} \tag{8.53}$$

where the second inequality follows from that $L_i^f(k_i) \ge -300 d \log n$ and that $e^{\lambda/n}\left(1 - \frac{1-e^{-\lambda}}{n}\right)$ is increasing for $\lambda > 0$. Combining (8.52) and (8.53), we obtain

$$\begin{aligned}
\mathbb{P}\big(L_i^f(m_1) < -A \,|\, F_i, \mathcal{F}_{k_i}\big) &\le e^{-\lambda A} \cdot e^{300\lambda d \log n} \cdot e^{\lambda m_1/n} \cdot \left(1 - \frac{1 - e^{-\lambda}}{n}\right)^{m_1} \\
&= \exp\left(-\lambda A + 300\lambda d \log n + \frac{\lambda m_1}{n} + m_1 \log\left(1 - \frac{1 - e^{-\lambda}}{n}\right)\right) \\
&\le \exp\left(-\lambda A + 300\lambda d \log n + \frac{\lambda m_1}{n} - m_1 \cdot \frac{1 - e^{-\lambda}}{n}\right) \\
&= \exp\left(-\lambda(A - 300 d \log n) + \frac{m_1}{n}(\lambda - 1 + e^{-\lambda})\right) \\
&\le \exp\left(-\lambda(A - 300 d \log n) + \frac{\lambda^2 m_1}{2n}\right),
\end{aligned}$$

where the second inequality uses $\log(1-x) \le -x$ for $0 \le x \le 1$, and the last inequality follows from that $e^{-x} < 1 - x + x^2/2$ for $x > 0$. We plug $\lambda = \frac{n(A - 300 d \log n)}{m_1}$ into the above inequality to obtain

$$\mathbb{P}\big(L_i^f(m_1) < -A \,|\, F_i, \mathcal{F}_{k_i}\big) \le \exp\left(-\frac{n(A - 300 d \log n)^2}{2 m_1}\right) = \exp\left(-\frac{(1 - o(1))A^2}{2 \log^{\alpha} n}\right) < n^{-5d}.$$

We recall that $\mathbb{P}\big(L_i^f(m_1) < -A\big) = \mathbb{P}(L_i^f(m_1) < -A, F) \le \mathbb{P}(L_i^f(m_1) < -A \,|\, F)$. Hence inequality (8.51) follows from taking a union bound of the above inequality over $i \in [n]$. $\square$

8.3 Proofs of Propositions 8.2 and 8.5

Proof of Proposition 8.2. The statement follows as easy consequence of Proposition 4.3 with the parameters $t = m_0/n, \alpha = \frac{\log(m_0/n)}{\log\log n}, \eta = 0$. We first show that Proposition 4.3 is applicable with the aforementioned parameters. Recall that $m_0 = \lfloor 200dn \log n \rfloor$. This, together with the assumption that $|L_i(0)| \le 100d \log n$ for all $i \in [n]$, implies that $\mathrm{MaxLoad}^f(0) \le t/2$. Hence, the first condition of Proposition 4.3 is satisfied. Notice that $\alpha \ge 1$ and that $L_0 = (\log n)^{\frac{1}{2}+\Theta(\frac{1}{k})}$, where $k = \lfloor \frac{\log\log n}{3\log\log\log n} \rfloor$. We thus have $L_0 \gg (L_0)^2/\log^\alpha n$. This, along with the assumption that $|\{i \in [n] : L_i(0) > L_0\}| \le 4000ne^{-L_0/15}$, guarantees the validity of the second condition of Proposition 4.3. Thus we can apply Proposition 4.3 to obtain that

$$\mathbb{P}\left(\mathrm{MaxLoad}^f(m_0) > (2k+2)L_0\right) \le n^{-e^{\sqrt{\log\log\log n}}}.$$

Observe that $(2k+2)L_0 = (\log n)^{\frac{1}{2}+O\left(\frac{\log\log\log n}{\log\log n}\right)}$ and that $L = (\log n)^{\frac{1}{2}+\Theta\left(\frac{1}{\sqrt{\log\log\log n}}\right)}$. Hence, we obtain

$$\mathbb{P}\left(\mathrm{MaxLoad}^f(m_0) > L\right) \le n^{-e^{\sqrt{\log\log\log n}}}. \tag{8.54}$$

Notice that the load of each bin can decrease by at most $m_0/n \le 200d \log n$ after the allocation of m_0 balls. Since $\max_{i\in[n]} |L_i^f(0)| \le 100d \log n$, we have

$$\min_{i\in[n]} L_i^f(m_0) \ge -300d \log n.$$

This, together with (8.54), concludes the proof of Proposition 8.2. □

Proof of Proposition 8.5. We apply Lemma 3.9 and Lemma 3.10 to obtain for sufficiently large n that

$$\mathbb{P}\left(\max_{i\in[n]} |L_i^f(m_2)| > 100d \log n\right) \le n^{-3d},$$

$$\mathbb{P}\left(\left|\left\{i \in [n] : L_i^f(m_2) > L_0\right\}\right| > 4000ne^{-L_0/15}\right) \le \exp\left(-n^{1-o(1)}\right).$$

Then we can conclude the proof by taking the union bound. □

References

[1] N. Alon, O. Gurel-Gurevich and E. Lubetzky. "Choice-memory tradeoff in allocations", *Ann. Appl. Probab.*, 20(4): 1470-1511, 2010.

[2] Y. Azar, A. Broder, A. Karlin and E. Upfal. "Balanced allocations", *SIAM J. Comput.*, 29(1): 180-200, 1999.

[3] M. Adler, S. Chakrabarti, M. Mitzenmacher and L. Rasmussen. "Parallel randomized load balancing", In *Proceedings of the 27th Annual ACM Symposium on Theory of Computing (STOC'95)*, pages 238-247, May 1995.

[4] P. Berenbrink, A. Brinkmann, T. Friedetzky and L. Nagel. "Balls into non-uniform bins", *J. Parallel Distributed Comput.*, 74(2): 2065-2076, 2014.

[5] P. Berenbrink, A. Czumaj, A. Steger, and B. Vöcking. "Balanced allocations: The heavily loaded case", *SIAM J. Comput.*, 35(6): 1350-1385, 2006.

[6] I. Benjamini and Y. Makarychev. "Balanced allocation: Memory performance tradeoffs", *Ann. Appl. Probab.*, 22(4): 1642-1649, 2012.

[7] R. Dwivedi, O. N. Feldheim, O. Gurel-Gurevich and A. Ramdas, "The power of thinning in reducing discrepancy", *Probab. Theory Related Fields.*, 174(1-2), 103–131, 2019.

[8] D. J. Daley and D. Vere-Jones, *An introduction to the theory of point processes, Volume I: Elementary theory and methods*, 2nd ed., Springer-Verlag, New York, 2003.

[9] D. J. Daley and D. Vere-Jones, *An introduction to the theory of point processes, Volume II: General theory and structure*, 2nd ed., Springer-Verlag, New York, 2008.

[10] A. Dembo and O. Zeitouni, *Large deviations techniques and applications*, 2nd ed., Springer-Verlag, Berlin, 2010.

[11] O. N. Feldheim and O. Gurel-Gurevich. "The power of thinning in balanced allocation", *Electron. Commun. Probab.* 26, 1-8, 2021.

[12] O. N. Feldheim and J. Li, "Load balancing under d-thinning", *Electron. Commun. Probab.* 25(1): 1-13, 2020.

[13] B. Godfrey. "Balls and bins with structure: balanced allocations on hypergraphs", In *Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithm (SODA'08)*, pages 511-517, January 2008.

[14] R. M. Karp, M. Luby and F. Meyer auf der Heide. "Efficient PRAM simulation on a distributed memory machine", In *Proceedings of the 24th Annual ACM Symposium on Theory of Computing (STOC'92)*, pages 318–325, May 1992.

[15] K. Kenthapadi and R. Panigrahy. "Balanced allocation on graphs", In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithm (SODA'06)*, pages 434-443, January 2006.

[16] D. Los and T. Sauerwaldy. "Balanced allocations with incomplete information: The power of two queries", *Available at*: `https://arxiv.org/abs/2107.03916`

[17] M. Mitzenmacher, B. Prabhakar and D. Shah. "Load balancing with memory", In *Proceedings of the 43rd Symposium on Foundations of Computer Science (FOCS'02)*, pages 799-808, November 2002.

[18] M. Mitzenmacher, A. W. Richa and R. Sitaraman. "The power of two random choices: A survey of techniques and results", In: Pardalos, P., Rajasekaran, S., Rolim, J., (eds) *Handbook of Randomized Computing*, Kluwer Academic Press, 2001.

[19] M. Mitzenmacher and E. Upfal. *Probability and computing: Randomized algorithms and probabilistic analysis*, 2nd ed. Cambridge University Press, 2005.

[20] Y. Peres, K. Talwar and U. Wieder, Graphical balanced allocations and the $(1+\beta)$-choice process, *Random Struct. Algor.*, 47(4): 760-775, 2015.

[21] 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 (RANDOM'98)*, pages 159-170, October 1998.

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

[23] V. Stemann. "Parallel balanced allocations", In *Proceedings of the 8th Annual ACM Symposium on Parallel Algorithms and Architectures (SPAA'96)*, pages 261–269, June 1996.

[24] K. Talwar and U. Wieder. "Balanced allocations: the weighted case", In *Proceedings of the 39th ACM Symposium on Theory of Computing (STOC'07)*, pages 256-265, June 2007.

[25] K. Talwar and U. Wieder. "Balanced allocations: A simple proof for the heavily loaded case." In: Esparza J., Fraigniaud P., Husfeldt T., Koutsoupias E. (eds) *Automata, Languages, and Programming. ICALP 2014*. Lecture Notes in Computer Science, vol 8572. Springer, Berlin, Heidelberg.