Self-stabilizing Balls & Bins in Batches[*]
The Power of Leaky Bins

Petra Berenbrink[†]
Universität Hamburg
22527 Hamburg
Germany

Tom Friedetzky
Durham University
Durham, DH1 3LE, U.K.

Peter Kling
Universität Hamburg
22527 Hamburg
Germany

Frederik Mallmann-Trenn
École normale supérieure
75005 Paris, France
and
Simon Fraser University
Burnaby, B.C., V5A 1S6, Canada

Lars Nagel
Loughborough University
Loughborough, LE11 3TU, U.K.

Chris Wastell
Durham University
Durham, DH1 3LE, U.K.

March 16, 2018

Abstract

A fundamental problem in distributed computing is the distribution of requests to a set of uniform servers without a centralized controller. Classically, such problems are modeled as static balls into bins processes, where m balls (tasks) are to be distributed among n bins (servers). In a seminal work, Azar et al. [4] proposed the sequential strategy GREEDY$[d]$ for $n = m$. Each ball queries the load of d random bins and is allocated to a least loaded of them. Azar et al. showed that $d = 2$ yields an exponential improvement compared to $d = 1$. Berenbrink et al. [7] extended this to $m \gg n$, showing that for $d = 2$ the maximal load difference is independent of m (in contrast to the $d = 1$ case).

We propose a new variant of an *infinite* balls-into-bins process. In each round an expected number of λn new balls arrive and are distributed (in parallel) to the bins. Subsequently, each non-empty bin deletes one of its balls. This setting models a set of servers processing incoming requests, where clients can query a server's current load but receive no information

[*] A preliminary version of this paper was published in the proceedings of PODC'16

[†] The author did some of the research while affiliated with Simon Fraser University.

about parallel requests. We study the Greedy[d] distribution scheme in this setting and show a strong self-stabilizing property: for *any* arrival rate $\lambda = \lambda(n) < 1$, the system load is time-invariant. Moreover, for *any* (even super-exponential) round t, the maximum system load is (w.h.p.) $\mathrm{O}\bigl(\frac{1}{1-\lambda} \cdot \log \frac{n}{1-\lambda}\bigr)$ for $d = 1$ and $\mathrm{O}\bigl(\log \frac{n}{1-\lambda}\bigr)$ for $d = 2$. In particular, Greedy[2] has an exponentially smaller system load for high arrival rates.

keywords— balls-into-bins, self-stabilizing, 2-choice, positive recurrent, maximum load

Contents

1 Introduction

One of the fundamental problems in distributed computing is the distribution of requests, tasks, or data items to a set of uniform servers. In order to simplify this process and to avoid a single point of failure, it is often advisable to use a simple, randomized strategy instead of a complex, centralized controller to allocate the requests to the servers. In the most naïve strategy (*1-Choice*), each client sends its request to a server chosen uniformly at random. A more elaborate scheme (*2-Choice*) chooses two servers, queries their current loads, and sends the request to a least loaded of them. Both approaches are typically modeled as balls-into-bins processes [2, 4, 5, 7, 13, 20, 22], where requests are represented as balls and servers as bins. While the latter approach leads to considerably better load distributions [4, 7], it loses some of its power in synchronous settings, where requests arrive in parallel and cannot take each other into account [2, 22].

We propose and study a novel infinite batch-based balls-into-bins process to model the client-server scenario. In a round, each server (bin) consumes one of its current tasks (balls). Afterward, expectedly λn tasks arrive and are allocated using a given distribution scheme. The *arrival rate* λ is allowed to be a function of n (e.g., $\lambda = 1 - 1/\operatorname{poly}(n)$). Standard balls-into-bins results imply that, for high arrival rates, with high probability[1] (w.h.p.) in each round there is a bin that receives $\Theta(\log n/\log\log n)$ balls. Most other infinite balls-into-bins-type processes limit the total number of concurrent balls in the system by n [4, 5] and show a fast recovery. Since we do not limit the number of balls, our process can, in principle, result in an arbitrary high system load. In particular, if starting in a high-load situation (e.g., exponentially many balls), we cannot recover in a polynomial number of steps. Instead, we regard the system load as a Markov chain and adapt the following notion of *self-stabilization*: The system is positive recurrent (expected return time to a typical low-load situation is finite), and taking a snapshot of the load situation at an *arbitrary* (even super-exponential large) time step yields (w.h.p.) a time-independent maximum load. Positive recurrence is a standard notion for stability and basically states that the system load is time-invariant. For irreducible, aperiodic Markov chains it implies the existence of a unique stationary distribution (cf. Section 1.2). While this alone does not guarantee a good load in the stationary distribution, together with the snapshot property we can look at an arbitrary time window of polynomial size (even if it is exponentially far away from the start) and give strong load guarantees. In particular, we give the following bounds on the load in addition to showing positive recurrence:

1-Choice Process: The maximum load at an arbitrary time is (w.h.p.) bounded by $\mathrm{O}\bigl(\frac{1}{1-\lambda} \cdot \log \frac{n}{1-\lambda}\bigr)$. We also provide a lower bound which is asymptotically tight for $\lambda \leq 1 - 1/\operatorname{poly}(n)$. While this implies that already the simple 1-Choice process is self-stabilizing, the load properties in a "typical" state are poor: even an arrival rate of only $\lambda = 1 - 1/n$ yields a superlinear maximum load.

2-Choice Process: The maximum load at an arbitrary time is (w.h.p.) bounded by $\mathrm{O}\bigl(\log \frac{n}{1-\lambda}\bigr)$. This allows to maintain an exponentially better system load compared to the 1-Choice process; for any $\lambda \leq 1 - 1/\operatorname{poly}(n)$ the maximum load remains logarithmic.

Note that the resulting processes can be seen as queuing processes.

1.1 Related Work

We will continue with an overview of related work. We start with classical results for sequential and finite balls-into-bins processes, go over to parallel settings, and give an overview of infinite and batch-based processes similar to ours. We also briefly mention some results from queuing theory (which is related but studies slightly different quality of service measures and system models).

[1] An event $\mathcal{E}$ occurs *with high probability* (w.h.p.) if $\Pr(\mathcal{E}) = 1 - n^{-\Omega(1)}$.

Sequential Setting. There are many strong, well-known results for the classical, sequential balls-into-bins process. In the sequential setting, m balls are thrown one after another and allocated to n bins. For $m = n$, the maximum load of any bin is known to be (w.h.p.) $(1+\mathrm{o}(1)) \cdot \ln(n)/\ln\ln n$ for the 1-Choice process [13, 20] and $\ln\ln(n)/\ln d + \Theta(1)$ for the d-Choice process with $d \geq 2$ [4]. If $m \geq n \cdot \ln n$, the maximum load increases to $m/n + \Theta\big(\sqrt{m \cdot \ln(n)/n}\big)$ [20] and $m/n + \ln\ln(n)/\ln d + \Theta(1)$ [7], respectively. In particular, note that the number of balls above the average grows with m for $d = 1$ but is independent of m for $d \geq 2$. This fundamental difference is known as the *power of two choices*. A similar (if slightly weaker) result was shown by Talwar and Wieder [24] using a quite elegant proof technique (which we also employ and generalize for our analysis in Section 3). Czumaj and Stemann [10] study adaptive allocation processes where the number of a ball's choices depends on the load of queried bins. The authors subsequently analyze a scenario that allows reallocations.

Berenbrink et al. [9] adapt the threshold protocol from [2] (see below) to a sequential setting and $m \geq n$ bins. Here, ball i randomly chooses bins until it sees a load smaller than $1 + i/n$. While this is a relatively strong assumption on the balls, this protocol needs only $\mathrm{O}(m)$ choices in total (allocation time) and achieves an almost optimal maximum load of $\lceil m/n \rceil + 1$.

Parallel Setting. Several papers (e.g., [2, 22]) investigated parallel settings of multiple-Choice games for the case $m = n$. Here, all m balls have to be allocated in parallel, but balls and bins might employ some (limited) communication. Adler et al. [2] consider a trade-off between the maximum load and the number of communication rounds r the balls need to decide for a target bin. Basically, bounds that are close to the classical (sequential) processes can only be achieved if r is close to the maximum load [2]. The authors also give a lower bound on the maximum load if r communication rounds are allowed, and Stemann [22] provides a matching upper bound via a collision-based protocol.

Infinite Processes. In infinite processes, the number of balls to be thrown is not fixed. Instead, in each of infinitely many rounds, balls are thrown or reallocated and bins (possibly) delete old balls. Azar et al. [4] consider an infinite, sequential process starting with n balls arbitrarily assigned to n bins. In each round one random ball is reallocated using the d-Choice process. For any $t > cn^2 \log\log n$, the maximum load at time t is (w.h.p.) $\ln\ln(n)/\ln d + \mathrm{O}(1)$.

Adler et al. [1] consider a system where in each round $m \leq n/9$ balls are allocated. Bins have a FIFO queue, and each arriving ball is stored in the queue of two random bins. After each round, every non-empty bin deletes its frontmost ball (which automatically removes its copy from the second random bin). It is shown that the expected waiting time is constant and the maximum waiting time is (w.h.p.) $\ln\ln(n)/\ln d + \mathrm{O}(1)$. The restriction $m \leq n/9$ is the major drawback of this process. A further study of this process, based on differential methods and experiments, was conducted in [6]. The balls' arrival times are binomially distributed with parameters n and $\lambda = m/n$. Their results indicate a stable behavior for $\lambda \leq 0.86$. A similar model was considered by Mitzenmacher [18], who considers ball arrivals as a Poisson stream of rate λn for $\lambda < 1$. It is shown that the 2-Choice process reduces the waiting time exponentially compared to the 1-Choice process.

Czumaj [11] presents a framework to study the recovery time of discrete-time dynamic allocation processes. In each round one of n balls is reallocated using the d-Choice process. Two models are considered: in the first, the ball to be reallocated is chosen by taking a ball from a random bin. In the second, the ball to be reallocated is chosen by selecting a random ball. From an arbitrary initial assignment, the system is shown to recover to the maximum load from [4] within $\mathrm{O}(n^2 \ln n)$ rounds in the former and $\mathrm{O}(n \ln n)$ rounds in the latter case. Becchetti et al. [5] consider a similar (but parallel) process. In each round one ball is chosen from every non-empty bin and reallocated to a randomly chosen bin (one Choice per ball). The authors show that (w.h.p.) starting from an arbitrary configuration, it takes $\mathrm{O}(n)$ rounds to reach a configuration with maximum load $\mathrm{O}(\log n)$. Moreover, if the process starts in a configuration with maximum load $\mathrm{O}(\log n)$, then the maximum load stays in $\mathrm{O}(\log n)$ for $\mathrm{poly}(n)$ rounds. An interesting connection to our work is that the analysis of [5] is based on an auxiliary Tetris-process. This process can be seen a special version of our 1-Choice process and is defined as follows: starting from a state with at least $n/4$ empty bins, in each round every non-empty bin deletes one ball. Subsequently, exactly $3n/4$ new balls are allocated to the bins (one choice per ball).

Batch-Processes. Batch-based processes allocate m balls to n bins in batches of (usually) n balls each, where each batch is allocated in parallel. They lie between (pure) parallel and sequential processes. For $m = \tau \cdot n$, Stemann [22] investigates a scenario with n players each having m/n balls. To allocate a ball, every player independently chooses two bins and allocates copies of the ball to both of them. Every bin has two queues (one for first copies, one for second copies) and processes one ball from each queue per round. When a ball is processed, its copy is removed from the system and the player is allowed to initiate the allocation of the next ball. If $\tau = \ln n$, all balls are processed in $\mathrm{O}(\ln n)$ rounds and the waiting time is (w.h.p.) $\mathrm{O}(\ln \ln n)$. Berenbrink et al. [8] study the d-Choice process in a scenario where m balls are allocated to n bins in batches of size n each. The authors show that the load of every bin is (w.h.p.) $m/n \pm \mathrm{O}(\log n)$. As noted in Lemma 1, our analysis can be used to derive the same result by easier means.

Queuing Processes. Batch arrival processes have also been considered in the context of queuing systems. A key motivation for such models stems from the asynchronous transfer mode (ATM) in telecommunication systems. Tasks arrive in batches, are stored in a FIFO queue and served by a fixed number of servers which remove the tasks from the queue and process them. Several papers [3, 15, 16, 21] consider scenarios where the number of arriving tasks is determined by a finite state Markov chain. Results study steady state properties of the system to determine properties of interest (e.g., waiting times or queue lengths). Sohraby and Zhang [21] use spectral techniques to study a multi-server scenario with an infinite queue. Alfa [3] considers a discrete-time process for n identical servers and tasks with constant service time $s \geq 1$. To ensure a stable system, the arrival rate λ is assumed to be at most n/s and tasks are assigned cyclical, allowing to study an arbitrary server (instead of the complete system). Kamal [15] and Kim et al. [16] study a system with a finite capacity. The tasks which arrive when the buffer is full are lost. The authors study the steady state probability and give empirical results to show the decay of waiting times as n increases.

1.2 Model & Preliminaries

We model our load balancing problem as an infinite, parallel balls-into-bins process. Time is divided into discrete, synchronous rounds. There are n bins and n generators, and the initial system is assumed to be empty. At the start of each round, every non-empty bin deletes one ball. Afterward, every generator generates a ball with a probability of $\lambda = \lambda(n) \in [0, 1]$ (the *arrival rate*). This generation scheme allows us to consider arrival rates that are arbitrarily close to one (like $1 - 1/\operatorname{poly}(n)$). Generated balls are distributed in the system using a distribution process. In this paper we analyze two specific distribution processes:

- The 1-Choice process Greedy[1] assigns every ball to a random bin.
- The 2-Choice process Greedy[2] assigns every ball to a least loaded among two randomly chosen bins.

See Figure 1 for an illustration. It is worth mentioning, that the maximum load in Greedy[2] does not need to be smaller than in Greedy[1] as the following (artificial) example shows. Consider two bins ($n = 2$) with different initial loads and $\lambda = 1$. In Greedy[1] each bin receives $n/2 \pm c\sqrt{n}$ new balls for some constant c. On the other side, in Greedy[2] the bin with the smaller initial load receives $3n/4 \pm c\sqrt{n}$ new balls. However, as our results indicate, this effect becomes negligible when n grows.

Notation. The random variable $X_i(t)$ denotes the load (number of balls) of the i-th fullest bin at the end of round t. Thus, the load situation (configuration) after round t can be described by the load vector $\boldsymbol{X}(t) = (X_i(t))_{i \in [n]} \in \mathbb{N}^n$. We define $\varnothing(t) := \frac{1}{n} \sum_{i=1}^{n} X_i(t)$ as the average load at the end of round t. The value $\nu(t)$ denotes the fraction of non-empty bins after round t and $\eta(t) := 1 - \nu(t)$ the fraction of empty bins after round t. It will be useful to define $1_i(t) := \min\bigl(1, X_i(t)\bigr)$ and $\eta_i(t) := 1_i(t) - \nu(t)$ (which equals $\eta(t)$ if i is a non-empty bin and $-\nu(t)$ otherwise). For random variables X and Y we write $X \prec Y$ if X is stochastically dominated by Y. That is, if for all k we have $\Pr(X \geq k) \leq \Pr(Y \geq k)$.

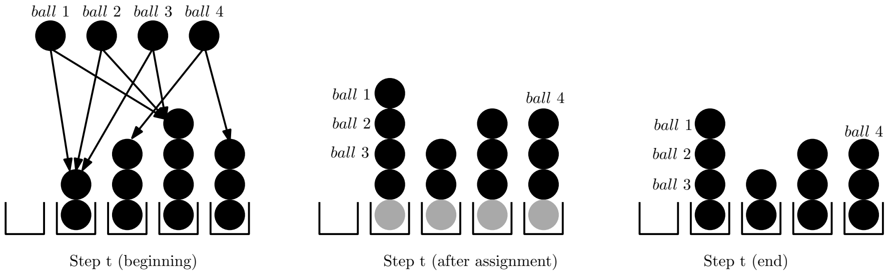

Figure 1: The figure depicts a typical round of Greedy[2]. In this example we have $n = 5$ and 4 balls arrive. Balls 1, 2, and 3 choose the same bin with a load of 2 and a bin with larger node and hence all move the same bin resulting in that bin having the highest load. Moreover, Ball 4 chooses two bins with equal load and chooses one of these uniformly at random. At the end of the round all non-empty bins delete one ball (marked gray).

Markov Chain Preliminaries. The random process $(\boldsymbol{X}(t))_{t\in\mathbb{N}}$ has the Markov property, since $\boldsymbol{X}(t)$ depends only on $\boldsymbol{X}(t-1)$ and the random choices during round t. We refer to this Markov chain as $\boldsymbol{X}$. Note that $\boldsymbol{X}$ is time-homogeneous (transition probabilities are time-independent), irreducible (every state is reachable from every other state[2]), and aperiodic (path lengths have no period; in fact, our chain is lazy). Recall that such a Markov chain is positive recurrent (or ergodic) if the probability to return to the start state is 1 and the expected return time is finite. In particular, this implies the existence of a unique stationary distribution. Positive recurrence is a standard formalization of the intuitive concept of stability. See [17] for an excellent introduction into Markov chains and the involved terminology.

2 The 1-Choice Process

We present two main results for the 1-Choice process: Theorem 1 states the stability of the system under the 1-Choice process for an arbitrary λ, using the standard notion of positive recurrence (cf. Section 1). In particular, this implies the existence of a stationary distribution for the 1-Choice process. Theorem 2 strengthens this by giving a high probability bound on the maximum load for an *arbitrary* round $t \in \mathbb{N}$. Together, both results imply that the 1-Choice process is self-stabilizing. That is, the system is positive recurrent and taking a snapshot of the load situation at an arbitrary time step yields (w.h.p.) a time-independent maximum load.

Theorem 1 (Stability)**.** *Let $\lambda = \lambda(n) < 1$. The Markov chain $\boldsymbol{X}$ of the 1-Choice process is positive recurrent.*

Theorem 2 (Maximum Load)**.** *Let $\lambda = \lambda(n) < 1$. Fix an arbitrary round t of the 1-Choice process. The maximum load of all bins is (w.h.p.) bounded by* $\mathrm{O}\left(\frac{1}{1-\lambda} \cdot \log \frac{n}{1-\lambda}\right)$.

[2] The state space includes all vectors with non-increasing entries over $\mathbb{N}^n$.

Note that for high arrival rates of the form $\lambda(n) = 1 - \varepsilon(n)$, the bound given in Theorem 2 is inversely proportional to $\varepsilon(n)$. For example, for $\varepsilon(n) = 1/n$ the maximal load is $\mathrm{O}(n \log n)$. Theorem 3 shows that this dependence is unavoidable: the bound given in Theorem 2 is tight for large values of λ.

Theorem 3. *Let n be sufficiently large. Let $\lambda = \lambda(n) \geq 3/4$ and consider step $t \coloneqq 9\lambda \log(n)/(64(1-\lambda)^2)$. With probability $1 - \mathrm{o}(1)$ there is a bin i in step t with load $\Omega\big(\frac{1}{1-\lambda} \cdot \log n\big)$.*

The proofs of these results can be found in the following subsections. We first prove a bound on the maximum load (Theorem 2). Afterward, we prove stability of the system (Theorem 1). Finally we prove the lower bound (Theorem 3).

2.1 Maximum Load – Proof of Theorem 2

The main idea of the proof is to bound the maximum load for any bin i and to take union bound of all resources. The load of bin i decreases whenever it is large and, thus, performs a biased random walk towards a load of zero. However, when the load is zero, it increases in expectation, such that standard drift theorems cannot not be applied directly. Nevertheless, the increase of the load for any given state has an exponential tail, which allows us to apply Hajek's Theorem (Theorem 7) to derive exponential tail bounds on the load of i at any (possibly super-exponential) number of time steps.

of Theorem 2. We prove Theorem 2 using a (slightly simplified) "drift theorem" from Hajek [14] (cf. Theorem 7 in Appendix A). As mentioned in Section 1.2, our process is a Markov chain, such that we need to condition only on the previous state (instead of the full filtration from Theorem 7). Our goal is to bound the load of a fixed bin i at time t using Theorem 7 and, subsequently, to use this with a union bound to bound the maximum load over all bins. To apply Theorem 7, we have to prove that the maximum load difference of bin i between two rounds is exponentially bounded (Majorization) and that, given a high enough load, the system tends to lose load (Negative Bias). We start with the majorization. Recall that for random variables X and Y we write $X \prec Y$ if X is stochastically dominated by Y, i.e., for all k it holds $\Pr(X \geq k) \leq \Pr(Y \geq k)$. The load difference $|X_i(t+1) - X_i(t)|$ is bounded by $\max(1, B_i(t)) \leq 1 + B_i(t)$, where $B_i(t)$ is the number of tokens bin i receives during round $t+1$. In particular, $\big(|X_i(t+1) - X_i(t)| \mid \boldsymbol{X}(t)\big) \prec 1 + B_i(t)$. Note that $B_i(t)$ is binomially distributed with parameters n and λ/n since each of the potential n balls has probability λ to spawn and, given that it spawned, with probability $1/n$ it ends up in bin i. Using standard inequalities we bound

$$\Pr(B_i(t) = k) \leq \binom{n}{k} \cdot \left(\frac{\lambda}{n}\right)^k \leq \left(\frac{e \cdot n}{k}\right)^k \cdot \left(\frac{1}{n}\right)^k = \frac{e^k}{k^k} \tag{1}$$

and calculate

$$\mathbb{E}\Big[e^{B_i(t)+1}\Big] = e \cdot \sum_{k=0}^{n} e^k \cdot \frac{e^k}{k^k} = e \cdot \sum_{k=0}^{\lceil e^3-1 \rceil} \frac{e^{2k}}{k^k} + e \cdot \sum_{k=e^3}^{\infty} \frac{e^{2k}}{k^k} \\ \leq \Theta(1) + \sum_{k=1}^{\infty} e^{-k} = \Theta(1). \tag{2}$$

This shows that the Majorization condition from Theorem 7 holds (with $\lambda' = 1$ and $D = \Theta(1)$). To see that the Negative Bias condition is also given, note that if bin i has non-zero load, it is guaranteed to delete one ball and receives in expectation $n \cdot \lambda / n = \lambda$ balls. We get $\mathbb{E}[X_i(t+1) - X_i(t) \mid X_i(t) > 0] \leq \lambda - 1 < 0$, establishing the Negative Bias condition (with $\varepsilon_0 = 1 - \lambda$). Thus, we can apply Theorem 7 with $\eta \coloneqq \min(1, (1-\lambda)/(2D), 1/(2-2\lambda)) = (1-\lambda)/(2D)$ and get for $b \geq 1$

$$\Pr(X_i(t) \geq b) \leq e^{-b \cdot \eta} + \frac{2D}{\eta \cdot (1-\lambda)} \cdot e^{\eta \cdot (-b)} \leq \frac{2 \cdot (2D)^2}{(1-\lambda)^2} \cdot e^{\frac{(1-\lambda) \cdot (-b)}{2D}} \\ \leq \frac{(4D)^2}{(1-\lambda)^2} \cdot e^{\frac{-b \cdot (1-\lambda)}{(4D)^2}} \leq \frac{c}{(1-\lambda)^2} \cdot e^{-\frac{b \cdot (1-\lambda)}{c}}, \tag{3}$$

where $c \geq (4D)^2$ denotes a suitable constant. Applying the Union bound to all n bins and choosing $b \coloneqq \frac{c}{1-\lambda} \cdot \ln\big(\frac{c \cdot n^{h+1}}{(1-\lambda)^2}\big)$, where $h > 2$ is a constant, yields $\Pr\big(\max_{i \in [n]} X_i(t) \geq b\big) \leq n^{-h}$. Since

$$b = \frac{c}{1-\lambda} \cdot \ln\left(\frac{c \cdot n^{h+1}}{(1-\lambda)^2}\right) \leq \frac{c^2 \cdot (h+1)}{1-\lambda} \cdot \ln\left(\frac{n}{1-\lambda}\right) \\ = \mathrm{O}\left(\frac{1}{1-\lambda} \cdot \ln\left(\frac{n}{1-\lambda}\right)\right), \tag{4}$$

we get the desired statement. □

2.2 Stability – Proof of Theorem 1

In the following, we provide an auxiliary result that will prove useful for deriving the stability of the 1-Choice process.

Corollary 1. *Let $\lambda = \lambda(n) < 1$. Fix an arbitrary round t of the 1-Choice process and a bin i. There is a constant $c > 1$ such that the expected load of bin i is bounded by $\frac{6c}{1-\lambda} \cdot \ln\big(\frac{n}{1-\lambda}\big)$.*

Proof. By Theorem 2, the maximum load of all bins is with high probability bounded by $c \cdot \frac{1}{1-\lambda} \cdot \log \frac{n}{1-\lambda}$, for a sufficiently large constant c. Let

$$\gamma \coloneqq \frac{c}{1-\lambda} \cdot \ln\left(\frac{e \cdot cn}{(1-\lambda)^2}\right). \tag{5}$$

Partitioning time into windows of γ rounds and with Equation (3), we calculate

$$\begin{aligned}\mathbb{E}[X_i(t)] &= \sum_{b=1}^{\gamma} b\cdot \Pr(X_i(t)=b) + \sum_{k=1}^{\infty}\sum_{b=k\cdot\gamma+1}^{(k+1)\gamma} b\cdot \Pr(X_i(t)=b)\\ &\le \gamma + \sum_{k=1}^{\infty}(k+1)\gamma\cdot \Pr(X_i(t) > k\cdot\gamma)\\ &\le \gamma + \sum_{k=1}^{\infty}(k+1)\gamma\cdot \frac{c}{(1-\lambda)^2}\cdot e^{-\frac{k\cdot\gamma\cdot(1-\lambda)}{c}} \\ &\le \gamma + \sum_{k=1}^{\infty}(k+1)\gamma\cdot \frac{c}{(1-\lambda)^2}\cdot e^{-k}\cdot e^{-\ln(cn/(1-\lambda)^2)}\\ &\le \gamma + \sum_{k=1}^{\infty}(k+1)\gamma\cdot e^{-k} \le 3\gamma \le \frac{6c}{1-\lambda}\cdot \ln\left(\frac{e\cdot cn}{1-\lambda}\right).\end{aligned} \tag{6}$$

This finishes the proof. □

of Theorem 1. We prove Theorem 1 using a result from Fayolle et al. [12] (cf. Theorem 6 in Appendix A). Note that $\boldsymbol{X}$ is a time-homogeneous irreducible Markov chain with a countable state space. In the following, let

$$\Delta \coloneqq \frac{12e^2\cdot c^2 n^2}{(1-\lambda)^3}, \tag{7}$$

where c is the constant from Corollary 1. For a configuration $\boldsymbol{x}$, we define the auxiliary potential $\Psi(\boldsymbol{x}) \coloneqq \sum_{i=1}^{n} x_i$ as the total system load of configuration $\boldsymbol{x}$. Consider the (finite) set $C \coloneqq \{\,\boldsymbol{x} \mid \Psi(\boldsymbol{x}) \le n\cdot\Delta\,\}$ of all configurations with not too much load. To prove positive recurrence, it remains to show that Condition 1 (expected potential drop if not in a high-load configuration) and Condition 2 (finite potential) of Theorem 6 hold. Let us start with Condition 1. Fix a round t and let $\boldsymbol{x} = \boldsymbol{X}(t) \notin C$. By definition of C, we have $\Psi(\boldsymbol{x}) > n\cdot\Delta$. Hence, there is at least one bin i with load $x_i \ge \Psi(\boldsymbol{x})/n > \Delta$. Thus, by definition of the process, during each of the next Δ rounds bin i deletes exactly one ball. On the other hand, bin i receives in expectation $\Delta\cdot\lambda n\cdot\frac{1}{n} = \lambda\Delta$ balls during the next Δ rounds. We get

$\mathbb{E}[X_i(t+\Delta) - x_i \mid \boldsymbol{X}(t)=\boldsymbol{x}] = \lambda\Delta - \Delta = -(1-\lambda)\cdot\Delta$. For any bin $j \ne i$, we assume pessimistically that no ball is deleted. Note that the expected load increase of each of these bins can be majorized by the load increase in an empty system running for Δ rounds. Thus, we can use Corollary 1 to bound the expected load increase in each of these bins by $\frac{6c}{1-\lambda}\cdot\ln\left(\frac{2\cdot cn}{1-\lambda}\right) \le \frac{6e^2\cdot c^2\cdot n}{(1-\lambda)^2} = \frac{(1-\lambda)\Delta}{2n}$, by definition of Δ. We get

$$\begin{aligned}\mathbb{E}[\Psi(\boldsymbol{X}(t+\Delta)) \mid \boldsymbol{X}(t)=\boldsymbol{x}] &\le -(1-\lambda)\cdot\Delta + (n-1)\cdot\frac{(1-\lambda)\Delta}{2n}\\ &\le -\frac{1-\lambda}{2}\cdot\Delta.\end{aligned} \tag{8}$$

This proves Condition 1 of Theorem 6. For Condition 2, assume $\boldsymbol{x} = \boldsymbol{X}(t) \in C$. We bound the system load after Δ rounds trivially by

$$\mathbb{E}[\Psi(\boldsymbol{X}(t+\Delta)) \mid \boldsymbol{X}(t) = \boldsymbol{x}] \leq \Psi(\boldsymbol{x}) + \Delta \cdot n \leq n \cdot \Delta + \Delta \cdot n < \infty \tag{9}$$

(note that the finiteness in Theorem 6 is with respect to time, not n). This finishes the proof. □

2.3 Lower Bound on Maximum Load – Proof of Theorem 3

In expectation, the load of any non-empty bin decreases. Thus, to derive a meaningful lower bound, we need to make use of the variance of the number of balls that are assigned to a bin over a period of suitable length. To do so, we make use of Theorem 8 (due to Raab and Steger [20]; see appendix), which lower-bounds the maximum number of balls a bin receives when m balls are allocated into n bins.

of Theorem 3. We assume that we start at an empty system and apply Theorem 8 to $m := \lambda t n$ many balls. The theorem states that, due to the variance, one of the bins is likely to get more than $c_1 \lambda t + c_2 \sqrt{t\lambda \log n}$ many balls for suitable constants c_1 and c_2. This allows us to show that the load of this bin is large, even if we assume, pessimistically, that it deletes a ball during each of the t time steps.

Let $M(t')$ be the number of balls allocated during the first $t' \in \mathbb{N}$ steps, and let $Y_{\max}(t')$ be the maximum number of balls allocated to any bin. Set

$$t := \frac{9\lambda \log(n)}{64(1-\lambda)^2} \tag{10}$$

and let $\varepsilon := (1-\lambda)/\lambda$. Since all balls are independent and $\mathbb{E}[M(t)] = t \cdot \lambda n \geq n \log n$ (due to $\lambda \geq 3/4$), it follows by Chernoff's inequality that

$$\Pr(M(t) \leq (1-\varepsilon) \cdot t \cdot \lambda n) \leq e^{-\varepsilon^2 \mathbb{E}[M(t)]/2} \leq \frac{1}{n^2}. \tag{11}$$

By Theorem 8 Cases 3 and 4 (depending on the size of $1-\lambda$) we get for $\alpha := \sqrt{8/9}$ (w.h.p.)

$$\begin{aligned} &Y_{\max}(t) \geq \\ &\geq (1-\varepsilon) \cdot t \cdot \lambda + \sqrt{2(1-\varepsilon) \cdot t \cdot \lambda \log n} \cdot \min\left\{ \alpha, \sqrt{1 - \frac{\log\log n}{2\alpha \log n}} \right\} \\ &= (1-\varepsilon) \cdot t \cdot \lambda + \alpha\sqrt{2(1-\varepsilon) \cdot t \cdot \lambda \log n}. \end{aligned} \tag{12}$$

Let $X_{\max}(t)$ denote the load of the bin of maximum load. We derive,

$$\begin{aligned}
X_{\max}(t) &\geq (1-\varepsilon)\cdot t\cdot\lambda + \sqrt{(1-\varepsilon)\cdot\frac{16}{9}t\cdot\lambda\log n} - t \\
&= (1-\varepsilon)\cdot t\cdot\lambda + \sqrt{\frac{1-\varepsilon}{4}\cdot\frac{\lambda\log n}{(1-\lambda)}} - t \\
&= \sqrt{\frac{1-\varepsilon}{4}\cdot\frac{\lambda\log n}{(1-\lambda)}} - 2(1-\lambda)t \qquad (13)\\
&= \sqrt{\frac{1-\varepsilon}{4}\cdot\frac{\lambda\log n}{(1-\lambda)}} - \frac{9\lambda\log(n)}{32(1-\lambda)} \\
&= \left(\sqrt{\frac{1-\frac{1-\lambda}{\lambda}}{4}} - \frac{9}{32}\right)\cdot\frac{\lambda\log n}{(1-\lambda)} = \Omega\left(\frac{\lambda\log n}{1-\lambda}\right),
\end{aligned}$$

where the last inequality holds since $\lambda \geq 3/4$. □

3 The 2-Choice Process

We continue with the study of the 2-Choice process. Here, new balls are distributed according to Greedy[2] (cf. description in Section 1.2). Our main results are the following theorems, which are equivalents to the corresponding theorems for the 1-Choice process.

Theorem 4 (Stability)**.** *Let $\lambda = \lambda(n) \in [1/4, 1)$. The Markov chain $\boldsymbol{X}$ of the 2-Choice process is positive recurrent.*

Theorem 5 (Maximum Load)**.** *Let $\lambda = \lambda(n) \in [1/4, 1)$. Fix an arbitrary round t of the 2-Choice process. The maximum load of all bins is (w.h.p.) bounded by $\mathrm{O}\big(\log\frac{n}{1-\lambda}\big)$.*

Note that Theorem 5 implies a much better behaved system than we saw in Theorem 2 for the 1-Choice process. In particular, it allows for an exponentially higher arrival rate: for $\lambda(n) = 1 - 1/\operatorname{poly}(n)$ the 2-Choice process maintains a maximal load of $\mathrm{O}(\log n)$. In contrast, for the same arrival rate the 1-Choice process results in a system with maximal load $\Omega(\operatorname{poly}(n))$.

Our analysis of the 2-Choice process relies to a large part on a good bound on the *smoothness* (the maximum load difference between any two bins). This is stated in the following proposition. This result is of independent interest, showing that even if the arrival rate is $\lambda(n) = 1-e^{-n}$, where we get a polynomial system load, the maximum load difference is still logarithmic.

Proposition 1 (Smoothness)**.** *Let $\lambda = \lambda(n) \in [1/4, 1]$. Fix an arbitrary round t of the 2-Choice process. The load difference of all bins is (w.h.p.) bounded by $\mathrm{O}(\ln n)$.*

Analysis Overview. To prove these results, we combine three different potential functions: For a configuration $\boldsymbol{x}$ with average load $\varnothing$ and for a suitable constant $\alpha < 1$ (to be fixed later), we define

$$\begin{aligned} &\Phi(\boldsymbol{x}) \coloneqq \sum_{i\in[n]} e^{\alpha\cdot(x_i-\varnothing)} + \sum_{i\in[n]} e^{\alpha\cdot(\varnothing-x_i)}, \qquad \Psi(\boldsymbol{x}) \coloneqq \sum_{i\in[n]} x_i, \quad \text{and} \\ &\Gamma(\boldsymbol{x}) \coloneqq \Phi(\boldsymbol{x}) + \tfrac{n}{1-\lambda}\cdot\Psi(\boldsymbol{x}). \end{aligned} \tag{14}$$

The potential Φ measures the *smoothness* (the maximum load difference to the average) of a configuration and is used to prove Proposition 1 (Section 3.1). The proof is based on the observation that whenever the load of a bin is far from the average load, it decreases in expectation. The potential Ψ measures the *total load* of a configuration and is used, in combination with our results on the smoothness, to prove Theorem 5 (Section 3.2). The potential Γ entangles the smoothness and total load, allowing us to prove Theorem 4 (Section 3.3). The proof is based on the fact that whenever Γ is large (i.e., the configuration is not smooth or it has a huge total load), it decreases in expectation.

Before we continue with our analysis, let us make a simple but useful observation concerning the smoothness: For any configuration $\boldsymbol{x}$ and value $b \geq 0$, the inequality $\Phi(\boldsymbol{x}) \leq e^{\alpha\cdot b}$ implies (by definition of Φ) $\max_i|x_i - \varnothing| \leq b$. That is, the load difference of any bin to the average is at most b and, thus, the load difference between any two bins is at most $2b$.

Observation 1. *Consider a configuration $\boldsymbol{x}$ with average load $\varnothing$ and let $b \geq 0$. If $\Phi(\boldsymbol{x}) \leq e^{\alpha\cdot b}$, then $|x_i - \varnothing| \leq b$ for all $i \in [n]$. In particular, $\max_i(x_i) - \min_i(x_i) \leq 2b$.*

3.1 Bounding the Smoothness – Proof of Proposition 1

The goal of this section is to prove Proposition 1. The key ingredient for its proof is the following statement: There are values $0 < c < 1$ and $\gamma > 0$ such that

$$\mathbb{E}[\Phi(\boldsymbol{X}(t+1)) \mid \boldsymbol{X}(t)] \leq c\cdot\Phi(\boldsymbol{X}(t)) + \gamma \tag{15}$$

holds for all rounds $t \geq 0$. Once Equation (15) is proven, taking the expected value on both sides yields $\mathbb{E}[\Phi(\boldsymbol{X}(t+1))] \leq c\cdot\mathbb{E}[\Phi(\boldsymbol{X}(t))] + \gamma$. This recursion is solved by $\mathbb{E}[\Phi(\boldsymbol{X}(t))] \leq \gamma\cdot(1-c)^{-1}$. In the rest of this section, we prove that Equation (15) holds for a constant c and $\gamma = \mathrm{O}(n)$, such that we immediately get the following bound on the expected smoothness (potential Φ) at an arbitrary time t:

Lemma 1. *Let $\lambda \in [1/4, 1]$. Fix an arbitrary round t of the 2-Choice process. There is a constant $\varepsilon > 0$ such that $\mathbb{E}[\Phi(\boldsymbol{X}(t))] \leq n/\varepsilon$.*

In Lemma 1, we chose $\lambda \in [1/4, 1]$ for convenience; the proof works with minor modifications for any $\lambda = \Theta(1)$ (i.e., for any constant λ, no matter whether $\lambda < 1$ or $\lambda > 1$). Also, our analysis easily adapts to the process without deletions by setting $\lambda = 1$ and $\eta_i(t) = 0$. This yields the same results as [8] using a simpler analysis.

Proposition 1 emerges by combining Observation 1, Lemma 1, and Markov's inequality:

$$\Pr\left(\max_i X_i(t) - \min_i X_i(t) \geq \frac{4}{\alpha} \cdot \ln\left(\frac{n}{\varepsilon}\right)\right) \leq \Pr\left(\Phi(\boldsymbol{X}(t)) \geq \frac{n^2}{\varepsilon^2}\right) \leq \frac{\varepsilon}{n}.$$

It remains to prove Equation (15). Our proof follows the lines of [19, 24][3]. We start by splitting the potential $\Phi(\boldsymbol{x})$ in two parts:

$$\Phi(\boldsymbol{x}) = \Phi_+(\boldsymbol{x}) + \Phi_-(\boldsymbol{x}), \tag{16}$$

with the *upper potential* $\Phi_+(\boldsymbol{x}) \coloneqq \sum_i e^{\alpha\cdot(x_i - \varnothing)}$ and with the *lower potential* $\Phi_-(\boldsymbol{x}) \coloneqq \sum_i e^{\alpha\cdot(\varnothing - x_i)}$. For a fixed bin i, we use $\Phi_{i,+}(\boldsymbol{x}) \coloneqq e^{\alpha\cdot(x_i - \varnothing)}$ and $\Phi_{i,-}(\boldsymbol{x}) \coloneqq e^{\alpha\cdot(\varnothing - x_i)}$ to denote i's contribution to the upper and lower potential, respectively. When we consider the effect of a fixed round $t+1$, we will sometimes omit the time parameter and use prime notation to denote the value of a parameter at the end of round $t+1$. For example, we write X_i and X_i' for the load of bin i at the beginning and at the end of round $t+1$, respectively.

Two simple but useful identities regarding the potential drops $\Delta_{i,+}(t+1) \coloneqq \Phi_{i,+}(\boldsymbol{X}(t+1)) - \Phi_{i,+}(\boldsymbol{X}(t))$ and $\Delta_{i,-}(t+1) \coloneqq \Phi_{i,-}(\boldsymbol{X}(t+1)) - \Phi_{i,-}(\boldsymbol{X}(t))$ due to a fixed bin i during round $t+1$ are as follows:

Observation 2. *Fix a bin i, let K denote the number of balls that are placed during round $t+1$ and let $k \leq K$ be the number of these balls that fall into bin i. Then,*

1. $\Delta_{i,+}(t+1) = \Phi_{i,+}(\boldsymbol{X}(t)) \cdot \left(e^{\alpha\cdot(k - \eta_i(t) - K/n)} - 1\right)$ *and*
2. $\Delta_{i,-}(t+1) = \Phi_{i,-}(\boldsymbol{X}(t)) \cdot \left(e^{-\alpha\cdot(k - \eta_i(t) - K/n)} - 1\right)$.

Proof. Remember that $\mathbf{1}_i$ is an indicator value which equals 1 if and only if the i-th bin is non-empty in configuration $\boldsymbol{X}$. Bin i looses exactly $\mathbf{1}_i$ balls and receives exactly k balls, such that $X_i' - X_i = -\mathbf{1}_i + k$. Similarly, we have $\varnothing' - \varnothing = -\nu + K/n$ for the change of the average load. With the identity $\eta_i = \mathbf{1}_i - \nu$ (see Section 1.2), this yields

$$\begin{aligned}\Delta_{i,+}(t+1) &= e^{\alpha\cdot\left(X_i' - \varnothing'\right)} - e^{\alpha\cdot\left(X_i - \varnothing\right)} \\ &= e^{\alpha\cdot\left(X_i - \varnothing\right)} \cdot \left(e^{\alpha\cdot\left(-\mathbf{1}_i + k + \nu - K/n\right)} - 1\right) = \Phi_{i,+} \cdot \left(e^{\alpha\cdot(k - \eta_i - K/n)} - 1\right),\end{aligned} \tag{17}$$

proving the first statement. The second statement follows similarly. □

[3] Talwar and Wieder [24] use the same potential function to analyze variants of the sequential d-Choice process without deletions. Our analysis turns out a bit more involved, since we have to consider deletions and argue over whole batches (of random size) instead of single balls.

3.1.1 Preliminaries to Bound the Potential Drop

We now derive the main technical lemma that states general bounds on the expected upper and lower potential change during one round. This will be used to derive different bounds on the potential change depending on the situation (Section 3.1.2). For this, let $p_i \coloneqq \left(\frac{i}{n}\right)^2 - \left(\frac{i-1}{n}\right)^2 = \frac{2i-1}{n^2}$ (the probability that a ball thrown with Greedy[2] falls into the i-th fullest bin). We also define

$$\hat{\alpha} \coloneqq e^{\alpha} - 1 \qquad \text{and} \qquad \check{\alpha} \coloneqq 1 - e^{-\alpha}. \tag{18}$$

Note that $\hat{\alpha} \in (\alpha, \alpha + \alpha^2)$ and $\check{\alpha} \in (\alpha - \alpha^2, \alpha)$ for $\alpha \in (0, 1.7)$. This follows from the Taylor approximation $e^x \leq 1 + x + x^2$, which holds for $x \in (-\infty, 1.7]$ (we will use this approximation several times in the analysis). Finally, let

$$\hat{\delta}_i \coloneqq \lambda n \cdot ({}^1\!/\!{}_n \cdot 1^- - p_i \cdot {}^{\hat{\alpha}}\!/\!{}_{\alpha}) \qquad \text{and} \qquad \check{\delta}_i \coloneqq \lambda n \cdot ({}^1\!/\!{}_n \cdot 1^+ - p_i \cdot {}^{\check{\alpha}}\!/\!{}_{\alpha}), \tag{19}$$

where $1^- \coloneqq 1 - \alpha/n < 1 < 1^+ \coloneqq 1 + \alpha/n$. These $\hat{\delta}_i$ and $\check{\delta}_i$ values can be thought of as upper/lower bounds on the expected difference in the number of balls that fall into bin i under the 1-Choice and 2-Choice process, respectively (note that 1^+, 1^-, $\hat{\alpha}/\alpha$, and $\check{\alpha}/\alpha$ are all close to 1).

Lemma 2. *Consider a bin i after round t and a constant $\alpha \leq 1$.*

1. *For the expected change of i's upper potential during round $t+1$ we have*

$$\frac{\mathbb{E}[\Delta_{i,+}(t+1) \mid \boldsymbol{X}(t)]}{\Phi_{i,+}(\boldsymbol{X}(t))} \leq -\alpha \cdot \left(\eta_i + \hat{\delta}_i\right) + \alpha^2 \cdot \left(\eta_i + \hat{\delta}_i\right)^2. \tag{20}$$

2. *For the expected change of i's lower potential during round $t+1$ we have*

$$\frac{\mathbb{E}[\Delta_{i,-}(t+1) \mid \boldsymbol{X}(t)]}{\Phi_{i,-}(\boldsymbol{X}(t))} \leq \alpha \cdot (\eta_i + \check{\delta}_i) + \alpha^2 \cdot (\eta_i + \check{\delta}_i)^2. \tag{21}$$

Proof. For the first statement, we use Observation 2 to calculate

$$\begin{aligned}
&\mathbb{E}[\Delta_{i,+}(t) \mid \boldsymbol{X}]/\Phi_{i,+} = \\
&= \sum_{K=0}^{n} \sum_{k=0}^{K} \binom{n}{K} \binom{K}{k} (p_i\lambda)^k \cdot \big((1-p_i)\lambda\big)^{K-k} \cdot (1-\lambda)^{n-K} \cdot \left(e^{\alpha\cdot(k-\eta_i-K/n)} - 1\right) \\
&= \sum_{K=0}^{n} \binom{n}{K} (1-\lambda)^{n-K} \lambda^K \sum_{k=0}^{K} \binom{K}{k} \cdot p_i^k \cdot (1-p_i)^{K-k} \cdot \left(e^{\alpha\cdot(k-\eta_i-K/n)} - 1\right) \\
&= \sum_{K=0}^{n} \binom{n}{K} (1-\lambda)^{n-K} \lambda^K \cdot \left(e^{-\alpha(\eta_i+K/n)} \sum_{k=0}^{K} \binom{K}{k} (e^{\alpha} \cdot p_i)^k (1-p_i)^{K-k} - 1\right) \\
&= \sum_{K=0}^{n} \binom{n}{K} (1-\lambda)^{n-K} \lambda^K \cdot \left(e^{-\alpha(\eta_i+K/n)} \cdot (1 + \hat{\alpha} \cdot p_i)^K - 1\right),
\end{aligned}$$

where we first apply the law of total expectation together with Observation 2 and, afterward, twice the binomial theorem. Continuing the calculation using the aforementioned Taylor approximation $e^x \leq 1 + x + x^2$ (which holds for any $x \in (-\infty, 1.7]$), and the definition of $\hat{\delta}_i$ yields

$$
\begin{aligned}
&= e^{-\alpha\eta_i} \cdot \left(1 - \lambda + \lambda e^{-\alpha/n} \cdot (1 + \hat{\alpha} \cdot p_i)\right)^n - 1 \\
&\leq e^{-\alpha\eta_i} \cdot \left(1 - \lambda(1 - e^{-\alpha/n}) + \lambda \cdot \hat{\alpha} \cdot p_i\right)^n - 1 \\
&\leq e^{-\alpha\eta_i} \cdot \left(1 - \frac{\lambda \cdot \alpha}{n} \cdot (1 - \alpha/n) + \lambda \cdot \hat{\alpha} \cdot p_i\right)^n - 1 \\
&\leq e^{-\alpha\eta_i} \cdot \left(1 - \frac{\alpha}{n} \cdot \hat{\delta}_i\right)^n - 1 \\
&\leq e^{-\alpha \cdot \left(\eta_i + \hat{\delta}_i\right)} - 1.
\end{aligned}
$$

Now, the claim follows by another application of the Taylor approximation. The second statement follows similarly. □

Before we apply Lemma 2 to derive different bounds on the potential drop for various situations, we provide three auxiliary claims:

Claim 1. *Consider a bin i and the values $\hat{\delta}_i$ and $\check{\delta}_i$ as defined before Lemma 2. If $\alpha \leq \ln(10/9)$, then* $\max(|\hat{\delta}_i|, |\check{\delta}_i|) \leq 5\lambda/4$.

Proof. Remember that $\hat{\delta}_i = \lambda n \cdot ({}^1\!/\!_n \cdot 1^- - p_i \cdot {}^{\hat{\alpha}}\!/\!_{\alpha})$ and $\check{\delta}_i = \lambda n \cdot ({}^1\!/\!_n \cdot 1^+ - p_i \cdot {}^{\check{\alpha}}\!/\!_{\alpha})$, where $1^- = 1 - \alpha/n < 1 < 1 + \alpha/n = 1^+$ (see proof of Lemma 2). Note that if $\alpha \leq \ln(10/9)$, we have $1^+ < 5/4$ and $1^- > 8/9$. Since the p_i are non-decreasing in i, it is sufficient to consider the extreme cases $i = 1$ and $i = n$.

The claims hold trivially for $i = 1$, since $p_1 = 1/n^2$ and both $|{}^1\!/\!_n \cdot 1^- - p_i \cdot {}^{\hat{\alpha}}\!/\!_{\alpha}| \leq 1/n$ and $|{}^1\!/\!_n \cdot 1^+ - p_i \cdot {}^{\check{\alpha}}\!/\!_{\alpha}| \leq 1^+/n$. For the other extreme, $i = n$, we have $p_n \leq 2/n$. From this and the definition of $\hat{\alpha} = e^\alpha - 1$, we get $|\hat{\delta}_i| \leq \frac{5}{4}\lambda$, since $\frac{2}{n} \cdot \frac{\hat{\alpha}}{\alpha} - \frac{1}{n} \cdot 1^- \leq \frac{2}{n}\frac{10/9-1}{\ln(10/9)} - \frac{1}{n} \cdot 1^- < \frac{5}{4n}$. Similarly, $|\check{\delta}_i| \leq \frac{5}{4}\lambda$ follows together with $\frac{2}{n}\frac{\check{\alpha}}{\alpha} - \frac{1}{n} \cdot 1^+ < \frac{1}{n}$ (which holds for any $\alpha > 0$). □

Claim 2. *There is a constant $\varepsilon > 0$ such that*

1. $\sum_{i \leq \frac{3}{4}n} p_i \cdot \Phi_{i,+} \leq (1 - 2\varepsilon) \cdot \frac{\Phi_+}{n}$ *and*

2. $\sum_{i \in [n]} p_i \cdot \Phi_{i,-} \geq (1 + 2\varepsilon) \cdot \frac{\Phi_- - \sum_{i \leq \frac{n}{4}} \Phi_{i,-}}{n}$.

Proof. For Part 1, note that the $\Phi_{i,+}$ are non-increasing in i, that they sum up to Φ_+, and that the p_i are non-decreasing in i. Thus, the left hand side of the claim's first statement is maximized if $\Phi_{i,+} = \frac{4\Phi_+}{3n}$ for all i. Now note that there is a constant ε such that[4] $\sum_{i > 3n/4} p_i \geq \frac{1}{4} + \varepsilon$. We get $\sum_{i \leq 3n/4} p_i \leq \frac{3}{4} - \varepsilon$.

[4] This is easily verified by hand. Alternatively, [23, Appendix A] gives $\sum_{i \geq 3n/4} p_i \geq \frac{1}{4} + \varepsilon'$ and the statement follows by noting that $p_{3n/4} = \mathrm{o}(1)$.

With this, the result follows by

$$\sum_{i\le\frac{3}{4}n} p_i \cdot \Phi_{i,+} \le \left(\frac{3}{4}-\varepsilon\right)\frac{4\Phi_+}{3n} = \left(1-\frac{4\varepsilon}{3n}\right)\cdot\Phi_+ \le (1-2\varepsilon)\cdot\frac{\Phi_+}{n}. \tag{22}$$

Part 2 follows similarly. □

Claim 3. *Consider a round t and a constant $\alpha \ge 0$. Then:*

1. $\sum_{i\in[n]} \alpha\eta_i(\alpha\eta_i - 1)\cdot\Phi_{i,+}(\boldsymbol{X}(t)) \le \alpha^2\eta\nu\cdot\min\bigl(n, \Phi_+(\boldsymbol{X}(t))\bigr)$ *and*
2. $\sum_{i\in[n]} \alpha\eta_i(\alpha\eta_i + 1)\cdot\Phi_{i,-}(\boldsymbol{X}(t)) \le \alpha^2\eta\nu\cdot\Phi_-(\boldsymbol{X}(t))$.

Proof. For the first statement, we calculate

$$\begin{aligned}
&\sum_{i\in[n]} \alpha\eta_i(\alpha\eta_i - 1)\cdot\Phi_{i,+}(\boldsymbol{X}(t))\\
&= \sum_{i\le\nu n} \alpha\eta_i(\alpha\eta_i - 1)\cdot\Phi_{i,+}(\boldsymbol{X}(t)) + \sum_{i>\nu n} \alpha\eta_i(\alpha\eta_i - 1)\cdot\Phi_{i,+}(\boldsymbol{X}(t))\\
&= \alpha\eta(\alpha\eta - 1)\cdot\sum_{i\le\nu n}\Phi_{i,+}(\boldsymbol{X}(t)) + \alpha\nu(1+\alpha\nu)\cdot\sum_{i>\nu n}\Phi_{i,+}(\boldsymbol{X}(t))\\
&\le \alpha\eta(\alpha\eta-1)\cdot\nu\cdot\Phi_+(\boldsymbol{X}(t)) + \alpha\nu(1+\alpha\nu)\cdot\eta\cdot\min\bigl(n, \Phi_+(\boldsymbol{X}(t))\bigr)\\
&\le \alpha^2\eta\nu\cdot\min\bigl(n, \Phi_+(\boldsymbol{X}(t))\bigr),
\end{aligned} \tag{23}$$

where the first inequality uses that $\Phi_{i,+}(\boldsymbol{X}(t))$ is non-increasing in i and that $\Phi_{i,+}(\boldsymbol{X}(t)) \le 1$ for all $i > \nu n$. The claim's second statement follows by a similar calculation, using that $\Phi_{i,-}(\boldsymbol{X}(t))$ is non-decreasing in i (note that we cannot apply the same trick as above to get $\min\bigl(n, \Phi_-(\boldsymbol{X}(t))\bigr)$ instead of $\Phi_-(\boldsymbol{X}(t))$). □

3.1.2 Bounding the Potential Drop in Different Situations

With these tools in place, we can derive the bounds on the potential drop in different situations. We start with a relative bound on the upper potential change $\Delta_+(t+1) := \sum_{i\in[n]}\Delta_{i,+}(t+1)$ and lower potential change $\Delta_-(t+1) := \sum_{i\in[n]}\Delta_{i,-}(t+1)$ during round $t+1$, respectively.

Lemma 3. *Consider a round t and a constant $\alpha \le \ln(10/9)$ ($< 1/8$). Let $R \in \{+,-\}$ and $\lambda \in [1/4, 1]$. For the expected upper and lower potential drop during round $t+1$ we have*

$$\mathbb{E}[\Delta_R(t+1) \mid \boldsymbol{X}(t)] < 2\alpha\lambda\cdot\Phi_R(\boldsymbol{X}(t)). \tag{24}$$

Proof. We prove the statement for $R = +$. The case $R = -$ follows similarly. Using Lemma 2 and summing up over all $i \in [n]$ we get

$$\begin{aligned}\mathbb{E}[\Delta_+(t+1) \mid \boldsymbol{X}] &\leq \sum_{i\in[n]} \left(-\alpha \cdot (\eta_i + \hat{\delta}_i) + \alpha^2 \cdot (\eta_i + \hat{\delta}_i)^2\right) \cdot \Phi_{i,+} \\ &= \sum_{i\in[n]} \left(\eta_i\alpha(\eta_i\alpha - 1) + \alpha^2 \cdot (2\eta_i\hat{\delta}_i + \hat{\delta}_i^2) - \alpha \cdot \hat{\delta}_i\right) \cdot \Phi_{i,+} \\ &\leq \sum_{i\in[n]} \left(\eta_i\alpha(\eta_i\alpha - 1) + 5\alpha^2\lambda + \frac{5}{4}\alpha\lambda\right) \cdot \Phi_{i,+}.\end{aligned}$$

Here, the last inequality uses $\lambda \leq 1$ and $|\hat{\delta}_i| \leq \frac{5}{4}\lambda$ (Claim 1). We now apply Claim 3, $\nu\eta \leq 1/4 \leq \lambda$, and $\alpha < 1/8$ to get

$$\mathbb{E}[\Delta_+(t) \mid \boldsymbol{X}] \leq \left(\alpha^2\lambda + 5\alpha^2\lambda + \frac{5}{4}\alpha\lambda\right) \cdot \Phi_+ < 2\alpha\lambda \cdot \Phi_+, \tag{25}$$

the desired statement. □

The next two lemmas derive bounds that are used to bound the upper/lower potential change in reasonably balanced configurations.

Lemma 4. *Consider a round t and the constants ε (from Claim 2) and $\alpha \leq \min(\ln(10/9), \varepsilon/4)$. Let $\lambda \in [1/4, 1]$ and assume $X_{\frac{3}{4}n}(t) \leq \varnothing(t)$. For the expected upper potential drop during round $t+1$ we have*

$$\mathbb{E}[\Delta_+(t+1) \mid \boldsymbol{X}(t)] \leq -\varepsilon\alpha\lambda \cdot \Phi_+(\boldsymbol{X}(t)) + 2\alpha\lambda n. \tag{26}$$

Proof. To calculate the expected upper potential change, we use Lemma 2 and sum up over all $i \in [n]$ (using similar inequalities as in the proof of Lemma 3 and the definition of $\hat{\delta}_i$):

$$\begin{aligned}\mathbb{E}[\Delta_+(t+1) \mid \boldsymbol{X}] &\leq 6\alpha^2\lambda \cdot \Phi_+ - \sum_{i\in[n]} \alpha \cdot \hat{\delta}_i \cdot \Phi_{i,+} \\ &= \left(6\alpha^2\lambda - \alpha\lambda \cdot 1^-\right) \cdot \Phi_+ + \hat{\alpha}\lambda n \sum_{i\in[n]} p_i \cdot \Phi_{i,+}.\end{aligned} \tag{27}$$

We now use that $\Phi_{i,+} = e^{\alpha\cdot(X_i - \varnothing)} \leq 1$ for all $i > \frac{3}{4}n$ (by our assumption on $X_{\frac{3}{4}n}$). This yields

$$\mathbb{E}[\Delta_+(t+1) \mid \boldsymbol{X}] \leq \left(6\alpha^2\lambda - \alpha\lambda \cdot 1^-\right) \cdot \Phi_+ + \hat{\alpha}\lambda n \sum_{i\leq\frac{3}{4}n} p_i \cdot \Phi_{i,+} + 2\alpha\lambda n. \tag{28}$$

Finally, we apply Claim 2 and the definition of 1^- and $\hat{\alpha}$ to get

$$\begin{aligned}\mathbb{E}[\Delta_+(t+1) \mid \boldsymbol{X}] &\leq \left(6\alpha^2\lambda - \alpha\lambda \cdot 1^- + (1 - 2\varepsilon) \cdot \hat{\alpha}\lambda\right) \cdot \Phi_+ + 2\alpha\lambda n \\ &\leq \left(4\alpha^2\lambda - 2\varepsilon \cdot \alpha\lambda\right) \cdot \Phi_+ + 2\alpha\lambda n.\end{aligned} \tag{29}$$

Using $\alpha \leq \varepsilon/4$ yields the desired result. □

Lemma 5. *Consider a round t and the constants ε (from Claim 2) and $\alpha \leq \min(\ln(10/9), \varepsilon/8)$. Let $\lambda \in [1/4, 1]$ and assume $X_{\frac{n}{4}}(t) \geq \varnothing(t)$. For the expected lower potential drop during round t we have*

$$\mathbb{E}[\Delta_-(t+1) \mid \boldsymbol{X}(t)] \leq -\varepsilon\alpha\lambda \cdot \Phi_-(\boldsymbol{X}(t)) + \frac{\alpha\lambda n}{2}. \tag{30}$$

Proof. To calculate the expected lower potential change, we use Lemma 2 and sum up over all $i \in [n]$ (as in the proof of Lemma 4):

$$\begin{aligned} \mathbb{E}[\Delta_-(t+1) \mid \boldsymbol{X}] &\leq 6\alpha^2\lambda \cdot \Phi_- + \sum_{i \in [n]} \alpha \cdot \check{\delta}_i \cdot \Phi_{i,-} \\ &= \left(6\alpha^2\lambda + \alpha\lambda \cdot 1^+\right) \cdot \Phi_- - \check{\alpha}\lambda n \sum_{i \in [n]} p_i \cdot \Phi_{i,-}. \end{aligned} \tag{31}$$

We now use that $\Phi_{i,-} = e^{\alpha \cdot (\varnothing - X_i)} \leq 1$ for all $i \leq \frac{n}{4}$ (by our assumption on $X_{\frac{n}{4}}$) and apply Claim 2 to get

$$\begin{aligned} \mathbb{E}[\Delta_-(t) \mid \boldsymbol{X}] &\leq \left(6\alpha^2\lambda + \alpha\lambda \cdot 1^+\right) \cdot \Phi_- - (1+2\varepsilon) \cdot \check{\alpha}\lambda n \cdot \frac{\Phi_- - \frac{n}{4}}{n} \\ &= \left(6\alpha^2\lambda + \alpha\lambda \cdot 1^+ - (1+2\varepsilon) \cdot \check{\alpha}\lambda\right) \cdot \Phi_- + (1+2\varepsilon) \cdot \frac{\check{\alpha}\lambda n}{4} \\ &\leq \left(8\alpha^2\lambda - 2\varepsilon \cdot \alpha\lambda\right) \cdot \Phi_- + \frac{\alpha\lambda n}{2}, \end{aligned} \tag{32}$$

where the last inequality used the definitions of 1^+, $\check{\alpha}$, as well as $\check{\alpha} > \alpha - \alpha^2$. Using $\alpha \leq \varepsilon/8$ yields the desired result. □

The following two lemmas bound the potential drop in configurations with many balls far below the average to the right and with many balls far above the average to the left.

Lemma 6. *Consider a round t and constants $\alpha \leq 1/46$ ($< \ln(10/9)$) and $\varepsilon \leq 1/3$. Let $\lambda \in [1/4, 1]$ and assume $X_{\frac{3}{4}n}(t) \geq \varnothing(t)$ and $\mathbb{E}[\Delta_+(t+1) \mid \boldsymbol{X}(t)] \geq -\frac{\varepsilon\alpha\lambda}{4} \cdot \Phi_+(\boldsymbol{X}(t))$. Then, $\Phi_+(\boldsymbol{X}(t)) \leq \frac{\varepsilon}{4} \cdot \Phi_-(\boldsymbol{X}(t))$ or $\Phi(\boldsymbol{X}(t)) = \varepsilon^{-8} \cdot \mathrm{O}(n)$.*

Proof. Let $L \coloneqq \sum_{i \in [n]} \max(X_i - \varnothing, 0) = \sum_{i \in [n]} \max(\varnothing - X_i, 0)$ be the "excess load" above and below the average. First note that the assumption $X_{\frac{3}{4}n} \geq \varnothing$ implies $\Phi_- \geq \frac{n}{4} \cdot \exp(\frac{\alpha L}{n/4})$ (using Jensen's inequality). On the other hand, we can use the assumption $\mathbb{E}[\Delta_+(t+1) \mid \boldsymbol{X}] \geq -\frac{\varepsilon\alpha\lambda}{4} \cdot \Phi_+$ to show an upper bound on Φ_+. To this end, we use Lemma 2 and sum up over all $i \in [n]$ (as in the proof of Lemma 4):

$$\begin{aligned} \mathbb{E}[\Delta_+(t+1) \mid \boldsymbol{X}] &\leq 6\alpha^2\lambda \cdot \Phi_+ - \sum_{i \in [n]} \alpha \cdot \hat{\delta}_i \cdot \Phi_{i,+} \\ &= 6\alpha^2\lambda \cdot \Phi_+ - \sum_{i \leq \frac{n}{3}} \alpha \cdot \hat{\delta}_i \cdot \Phi_{i,+} - \sum_{i > \frac{n}{3}} \alpha \cdot \hat{\delta}_i \cdot \Phi_{i,+}. \end{aligned} \tag{33}$$

For $i \leq n/3$ we have $p_i = \frac{2i-1}{n^2} \leq \frac{2}{3n}$ and, using the definition of 1^- and $\hat{\alpha}$, $\hat{\delta}_i = \lambda n \cdot \left({}^1\!/\!{}_n \cdot 1^- - p_i \cdot {}^{\hat{\alpha}}\!/\!{}_{\alpha}\right) \geq (1-5\alpha)\lambda/3$. Setting $\Phi_{\leq n/3,+} \coloneqq \sum_{i \leq n/3} \Phi_{i,+}$ and $\Phi_{>n/3,+} \coloneqq \sum_{i>n/3} \Phi_{i,+}$, together with Claim 1 this yields

$$\begin{aligned}
&\mathbb{E}[\Delta_+(t+1) \mid \boldsymbol{X}] \leq \\
&\leq 6\alpha^2\lambda \cdot \Phi_+ - \frac{\alpha(1-5\alpha)\lambda}{3} \cdot \Phi_{\leq n/3,+} + \frac{5}{4}\alpha\lambda \cdot \Phi_{>n/3,+} \\
&= \left(6\alpha^2\lambda - \frac{\alpha(1-5\alpha)\lambda}{3}\right) \cdot \Phi_+ + \left(\frac{5}{4}\alpha\lambda + \frac{\alpha(1-5\alpha)\lambda}{3}\right) \cdot \Phi_{>n/3,+} \\
&\leq -\frac{\varepsilon\alpha\lambda}{2} \cdot \Phi_+ + 2\alpha\lambda \cdot \Phi_{>n/3,+},
\end{aligned} \tag{34}$$

where the last inequality uses $\alpha \leq 1/46 \leq \frac{1}{23} - \frac{3}{46}\varepsilon$. With this, the assumption $\mathbb{E}[\Delta_+(t+1) \mid \boldsymbol{X}] \geq -\frac{\varepsilon\alpha\lambda}{4} \cdot \Phi_+$ implies $\Phi_+ \leq \frac{8}{\varepsilon} \cdot \Phi_{>n/3,+} \leq \frac{8}{\varepsilon} \cdot \frac{2n}{3} e^{\frac{\alpha L}{n/3}} = \frac{16n}{3\varepsilon} e^{\frac{3\alpha L}{n}}$ (the last inequality uses that none of the $2n/3$ remaining bins can have a load higher than $L/(n/3)$). To finish the proof, assume $\Phi_+ > \frac{\varepsilon}{4} \cdot \Phi_-$ (otherwise the lemma holds). Combining this with the upper bound on Φ_+ and with the lower bound on Φ_-, we get

$$\frac{16n}{3\varepsilon} e^{\frac{3\alpha L}{n}} \geq \Phi_+ > \frac{\varepsilon}{4} \cdot \Phi_- \geq \frac{\varepsilon n}{16} \cdot e^{\frac{4\alpha L}{n}}. \tag{35}$$

Thus, the excess load can be bounded by $L < \frac{n}{\alpha} \cdot \ln\left(\frac{256}{3\varepsilon^2}\right)$. Now, the lemma's statement follows from $\Phi = \Phi_+ + \Phi_- < \frac{5}{\varepsilon} \cdot \Phi_+ \leq \frac{80n}{3\varepsilon^2} e^{\frac{3\alpha L}{n}} = \varepsilon^{-8} \cdot \mathrm{O}(n)$. □

Lemma 7. *Consider a round t and constants $\alpha \leq 1/32$ ($< \ln(10/9)$) and $\varepsilon \leq 1$. Let $\lambda \in [1/4, 1]$ and assume $X_{\frac{n}{4}}(t) \leq \varnothing(t)$ and $\mathbb{E}[\Delta_-(t+1) \mid \boldsymbol{X}(t)] \geq -\frac{\varepsilon\alpha\lambda}{4} \cdot \Phi_-(\boldsymbol{X}(t))$. Then, $\Phi_-(\boldsymbol{X}(t)) \leq \frac{\varepsilon}{4} \cdot \Phi_+(\boldsymbol{X}(t))$ or $\Phi(\boldsymbol{X}(t)) = \varepsilon^{-8} \cdot \mathrm{O}(n)$.*

Proof. Let $L \coloneqq \sum_{i \in [n]} \max(X_i - \varnothing, 0) = \sum_{i \in [n]} \max(\varnothing - X_i, 0)$ be the "excess load" above and below the average. First note that the assumption $X_{\frac{n}{4}} \leq \varnothing$ implies $\Phi_+ \geq \frac{n}{4} \cdot e^{\frac{\alpha L}{n/4}}$ (using Jensen's inequality). On the other hand, we can use the assumption $\mathbb{E}[\Delta_-(t+1) \mid \boldsymbol{X}] \geq -\frac{\varepsilon\alpha\lambda}{4} \cdot \Phi_-$ to show an upper bound on Φ_-. To this end, we use Lemma 2 and sum up over all $i \in [n]$ (as in the proof of Lemma 5):

$$\begin{aligned}
\mathbb{E}[\Delta_-(t+1) \mid \boldsymbol{X}] &\leq 6\alpha^2\lambda \cdot \Phi_- + \sum_{i \in [n]} \alpha \cdot \check{\delta}_i \cdot \Phi_{i,-} \\
&= 6\alpha^2\lambda \cdot \Phi_- + \sum_{i \leq \frac{2n}{3}} \alpha \cdot \check{\delta}_i \cdot \Phi_{i,-} + \sum_{i > \frac{2n}{3}} \alpha \cdot \check{\delta}_i \cdot \Phi_{i,-}.
\end{aligned} \tag{36}$$

For $i \geq 2n/3$ we have $p_i = \frac{2i-1}{n^2} \geq \frac{4}{3n} - \frac{1}{n^2}$. Using this with $p_i \leq p_n \leq 2/n$ and $\check{\alpha} \geq \alpha - \alpha^2$, we can bound $\check{\delta}_i = \lambda n \cdot \left({}^1\!/\!{}_n \cdot 1^+ - p_i \cdot {}^{\check{\alpha}}\!/\!{}_{\alpha}\right) \leq \lambda \cdot \left(-{}^1\!/\!{}_3 + \frac{1+\alpha}{n}\right) + 2\alpha\lambda \leq -\lambda/6 + 2\alpha\lambda$. Setting $\Phi_{\leq 2n/3,-} \coloneqq \sum_{i \leq 2n/3} \Phi_{i,-}$ and $\Phi_{>2n/3,-} \coloneqq \sum_{i>2n/3} \Phi_{i,-}$, together with Claim 1 this yields

$$
\begin{aligned}
&\mathbb{E}[\Delta_-(t+1) \mid \boldsymbol{X}] \leq \\
&\leq 6\alpha^2\lambda \cdot \Phi_- + \frac{5}{4}\alpha\lambda \cdot \Phi_{\leq 2n/3,-} - \frac{\alpha\lambda}{6} \cdot \Phi_{>2n/3,-} + 2\alpha^2\lambda \cdot \Phi_{>2n/3,-} \\
&\leq \left(8\alpha^2\lambda - \alpha\lambda/6\right) \cdot \Phi_- + \left(\frac{5}{4}\alpha\lambda + \alpha\lambda/6\right) \cdot \Phi_{\leq 2n/3,-} \\
&\leq -\frac{\varepsilon\alpha\lambda}{2} \cdot \Phi_- + 2\alpha\lambda \cdot \Phi_{\leq 2n/3,-},
\end{aligned}
\tag{37}
$$

where the last inequality uses $\alpha \leq 1/32 \leq \frac{1}{16} - \frac{1}{48}\varepsilon$. With this, the assumption $\mathbb{E}[\Delta_-(t+1) \mid \boldsymbol{X}] \geq -\frac{\varepsilon\alpha\lambda}{4} \cdot \Phi_-$ implies that $\Phi_- \leq \frac{8}{\varepsilon} \cdot \Phi_{\leq 2n/3,-} \leq \frac{8}{\varepsilon} \cdot \frac{2n}{3} e^{\frac{\alpha L}{n/3}} = \frac{16n}{3\varepsilon} e^{\frac{3\alpha L}{n}}$ (the last inequality uses that none of the $2n/3$ remaining bins can have a load higher than $L/(n/3)$). To finish the proof, assume $\Phi_- > \frac{\varepsilon}{4} \cdot \Phi_+$ (otherwise the lemma holds). Combining this with the upper bound on Φ_- and with the lower bound on Φ_+, we get

$$
\frac{16n}{3\varepsilon} e^{\frac{3\alpha L}{n}} \geq \Phi_- > \frac{\varepsilon}{4} \cdot \Phi_+ \geq \frac{\varepsilon n}{16} \cdot e^{\frac{4\alpha L}{n}}. \tag{38}
$$

Thus, the excess load can be bounded by $L < \frac{n}{\alpha} \cdot \ln\left(\frac{256}{3\varepsilon^2}\right)$. Now, the lemma's statement follows from $\Phi = \Phi_+ + \Phi_- < \frac{5}{\varepsilon} \cdot \Phi_- \leq \frac{80n}{3\varepsilon^2} e^{\frac{3\alpha L}{n}} = \varepsilon^{-8} \cdot \mathrm{O}(n)$. □

3.1.3 Proving Equation (15)

With the lemmas from Section 3.1.2, we are finally ready to prove Equation (15). More exactly, we argue that for the constant ε from Claim 2 and $\alpha \leq \min(1/32, \varepsilon/8)$, for any $\lambda \in [1/4, 1]$ we have

$$
\mathbb{E}[\Phi(\boldsymbol{X}(t+1)) \mid \boldsymbol{X}(t)] \leq \left(1 - \frac{\varepsilon\alpha\lambda}{4}\right) \cdot \Phi(\boldsymbol{X}(t)) + \varepsilon^{-8} \cdot \mathrm{O}(n). \tag{39}
$$

This follows via a case analysis analogously to [24]:

Case 1: $x_{\frac{n}{4}} \geq \varnothing$ and $x_{\frac{3n}{4}} \leq \varnothing$

The bound follows from Lemma 4 and Lemma 5.

Case 2: $x_{\frac{n}{4}} \geq x_{\frac{3n}{4}} > \varnothing$

For $\mathbb{E}[\Delta_+(t+1) \mid \boldsymbol{X}(t)] \leq \frac{-\varepsilon\alpha\lambda}{4} \cdot \Phi_+$ the results follows from Lemma 5. Otherwise, $\mathbb{E}[\Delta_+(t+1) \mid \boldsymbol{X}(t)] > \frac{-\varepsilon\alpha\lambda}{4} \cdot \Phi_+$ and Lemma 6 yields two subcases:

Case 2.1: $\Phi_+(\boldsymbol{X}(t)) \leq \frac{\varepsilon}{4} \cdot \Phi_-(\boldsymbol{X}(t))$

Using Lemma 3 and Lemma 5 we obtain

$$\begin{aligned}
&\mathbb{E}[\Delta(t+1) \mid \boldsymbol{X}(t)] \le \\
&\le 2\alpha\lambda \cdot \Phi_+(\boldsymbol{X}(t)) - \varepsilon\alpha\lambda \cdot \Phi_-(\boldsymbol{X}(t)) + \frac{\alpha\lambda n}{2} \\
&\le -\frac{\varepsilon\alpha\lambda}{2} \cdot \Phi_-(\boldsymbol{X}(t)) + \frac{\alpha\lambda n}{2} \\
&\le -\frac{\varepsilon\alpha\lambda}{4} \cdot \Phi(\boldsymbol{X}(t)) + \varepsilon^{-8} \cdot \mathrm{O}(n).
\end{aligned} \tag{40}$$

Case 2.2: $\Phi(\boldsymbol{X}(t)) = \varepsilon^{-8} \cdot \mathrm{O}(n)$

Using Lemma 3 we get $\mathbb{E}[\Delta(t+1) \mid \boldsymbol{X}(t)] \le 2\alpha\lambda\varepsilon^{-8} \cdot \mathrm{O}(n)$. Our choice of α ($< 1/8$), λ (< 1), and ε ($\ll 1$) yields $2\alpha\lambda \le (1 - \varepsilon\alpha\lambda/4)$. Using the case assumption, we compute

$$\begin{aligned}
\mathbb{E}[\Delta(t+1) \mid \boldsymbol{X}(t)] \le 2\alpha\lambda\varepsilon^{-8} \cdot \mathrm{O}(n) &\le \left(1 - \frac{\varepsilon\alpha\lambda}{4}\right) \cdot \varepsilon^{-8} \cdot \mathrm{O}(n) \\
&\le -\frac{\varepsilon\alpha\lambda}{4} \cdot \Phi(\boldsymbol{X}(t)) + \varepsilon^{-8} \cdot \mathrm{O}(n).
\end{aligned} \tag{41}$$

Case 3: $x_{\frac{3n}{4}} \le x_{\frac{n}{4}} \le \varnothing$

Similar to the previous case, for $\mathbb{E}[\Delta_-(t+1) \mid \boldsymbol{X}(t)] \le \frac{-\varepsilon\alpha n}{4} \cdot \Phi_-$ the result follows from Lemma 4. For $\mathbb{E}[\Delta_-(t+1) \mid \boldsymbol{X}(t)] \ge \frac{-\varepsilon\alpha n}{4} \cdot \Phi_-$, Lemma 7 yields two subcases that are proven analogously to Cases 2.1 and 2.2 (using Lemma 4 instead of Lemma 5).

Thus, all cases lead to Equation (39).

3.2 Maximum Load – Proof of Theorem 5

The goal of this section is to prove Theorem 5. Recall the definitions of $\Phi(\boldsymbol{x})$ and $\Psi(\boldsymbol{x})$ from Equation (14). For any fixed round t, we will prove that (w.h.p.) $\Psi(\boldsymbol{X}(t)) = \mathrm{O}(n \cdot \ln n)$ and that the average load is $\varnothing = \mathrm{O}(\ln n)$. Using Union bounds and Proposition 1, we see that (w.h.p.) the maximum load at the end of round t is bounded by $\varnothing + \mathrm{O}(\ln n) = \mathrm{O}(\ln n)$.

It remains to prove a high probability bound on $\Psi(\boldsymbol{X}(t))$ for arbitrary t. To get an intuition for our analysis, consider the toy case $t = \mathrm{poly}(n)$ and assume that exactly $\lambda \cdot n \le n$ balls are thrown each round. Here, we can combine Observation 1 and Lemma 1 to bound (w.h.p.) the load difference between any pair of bins and for all $t' < t$ by $\mathrm{O}(\ln n)$ (via a union bound over $\mathrm{poly}(n)$ rounds). Given this bound on load difference, we can use the following combinatorial observation (formally stated in Lemma 8). If the load distance to the average is bounded by some $b \ge 0$, the bound on the number of balls $\Psi \le 2b \cdot n$ is invariant under the 2-Choice process, since under our assumptions all bins are non-empty and thus at least as many balls are deleted as spawn. In particular, we get for $b = \mathrm{O}(\ln n)$ that $\Psi(\boldsymbol{X}(t)) \le 2b \cdot n = \mathrm{O}(n \cdot \ln n)$, as required.

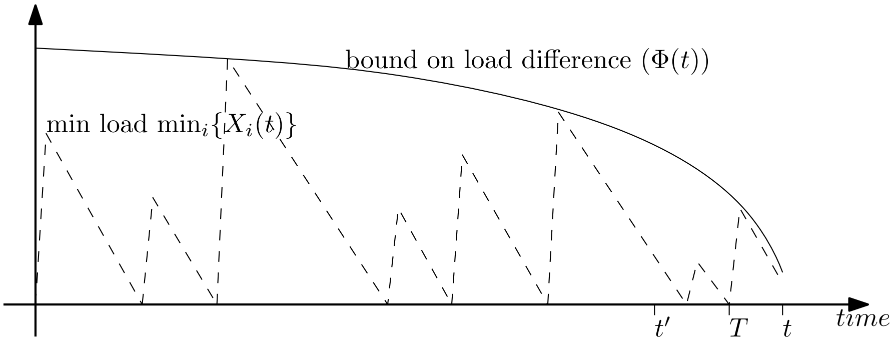
bound on load difference ($\Phi(t)$)
min load $\min_i\{X_i(t)\}$
t'
T
t
time

Figure 2: To bound the system load at time t, consider the minimum load and our bound on the load difference over time. We consider the last time T when there was an empty bin. The system load can only increase if there is an empty bin, and this increase is bounded by our bound on the load difference. Using that the system load decreases linearly in time while every increase is bounded by our logarithmic bound on the load difference, we find a small interval $[t', t]$ containing T. Due to the monotonic of our bound on Ψ, this will allow us to derive strong bounds on $\Psi(t)$. on the maximum load.

The case $t = \omega(\text{poly}(n))$ is considerably more involved. In particular, the fact that the number of balls in the system is only guaranteed to decrease when the total load is high *and* the load distance to the average is low makes it challenging to design a suitable potential function that drops fast enough when it is high. Thus, we deviate from this standard technique and elaborate on the idea of the toy case: Instead of bounding (w.h.p.) the load difference between any pair of bins by $\mathrm{O}(\ln n)$ for all $t' < t$ (which is not possible for $t \gg \text{poly}(n)$), we prove (w.h.p.) an *adaptive bound* of $\mathrm{O}(\ln(t - t') \cdot f(\lambda))$ for all $t' < t$, where f is a suitable function (Lemmas 9 and 10). Then we consider the last round $T < t$ with an empty bin. Observation 1 yields a bound of $\Psi(\boldsymbol{X}(T)) = 2 \cdot \mathrm{O}(\ln(t - T) \cdot f(\lambda)) \cdot n$ on the total load at time T. Using the same combinatorial observation as in the toy case, we get that (w.h.p.) $\Psi(\boldsymbol{X}(t)) \leq \Psi(\boldsymbol{X}(T)) = 2 \cdot \mathrm{O}(\ln(t - T) \cdot f(\lambda)) \cdot n$. The final step is to show that the load at time T (the load is is logarithmic in $t - T$) decreases *linearly* in $t - T$, showing that the time interval $[t - T, t]$ cannot be too large (or we would get a negative load at time t). Since the interval $[t - T, t]$ is short, we get a good bound on $\Psi(T)$. Using $\Psi(t) \leq \Psi(T)$ (due to the definition of T) together with the smoothness bounds of Lemma 9 yields the claim. See Figure 2 for an illustration.

Lemma 8. *Let $b \geq 0$ and consider a configuration $\boldsymbol{x}$ with $\Psi(\boldsymbol{x}) \leq 2b \cdot n$ and $\Phi(\boldsymbol{x}) \leq e^{\alpha \cdot b}$. Let $\boldsymbol{x'}$ denote the configuration after one step of the 2-Choice process. Then, $\Psi(\boldsymbol{x'}) \leq 2b \cdot n$.*

Proof. We distinguish two cases: if there is no empty bin, then all n bins delete one ball. Since the maximum number of new balls is n, the number of balls cannot increase. That is, we have $\Psi(\boldsymbol{x'}) \leq \Psi(\boldsymbol{x}) \leq 2b \cdot n$. Now consider the case that there is at least one empty bin. Let $\eta \in (0, 1]$ denote the fraction of empty bins (i.e., there are exactly $\eta \cdot n > 0$ empty bins). Since the minimal load is zero, Observation 1 implies $\max_i x_i \leq 2b$. Thus, the total number of balls in configuration $\boldsymbol{x}$ is at most $(1-\eta)n \cdot 2b$. Exactly $(1-\eta)n$ balls are deleted (one from each non-empty bin) and at most n new balls enter the system. We get $\Psi(\boldsymbol{x'}) \leq (1-\eta)n \cdot 2b - (1-\eta)n + n = (1-\eta)n \cdot (2b-1) + n \leq 2b \cdot n$. □

The next lemma bounds the probability of two events: First, it bounds Φ over an *arbitrary* time interval $[0, t)$ using a union bound over all past rounds $t' < t$. Note that t can be arbitrary large. Thus, in order to get a high probability bound, we must make the bound adaptive and allow for larger errors the further back in time we go. Second, the lemma shows that (w.h.p.) not too many balls are created.

Lemma 9. *Let $\lambda \in [1/4, 1)$. Fix a round t. For $i \in \mathbb{N}$ with $t - i \cdot \frac{8\log n}{1-\lambda} \geq 0$ define $\mathcal{I}_i := [t - i \cdot \frac{8\ln n}{1-\lambda}, t]$. Let Y_i be the number of balls which spawn in $\mathcal{I}_i$.*

1. *Define the (good) smooth event $\mathcal{S}_t := \bigcap_{t'<t} \{\, \Phi(\boldsymbol{X}(t')) \leq |t-t'|^2 \cdot n^2 \,\}$. Then, $\Pr(\mathcal{S}_t) = 1 - \mathrm{O}\big(n^{-1}\big)$.*
2. *Define the (good) bounded balls event $\mathcal{B}_t := \bigcap_i \{\, Y_i \leq \frac{1+\lambda}{2} \cdot |\mathcal{I}_i| \cdot n \,\}$. Then, $\Pr(\mathcal{B}_t) = 1 - \mathrm{O}\big(n^{-1}\big)$.*

Proof. Consider an arbitrary time $t' < t$. By Lemma 1 we have $\mathbb{E}[\Phi(t')] \leq n/\varepsilon$. Using Markov's inequality, this implies

$$\Pr(\Phi(t') \geq |t-t'|^2 \cdot n^2) \leq 1/(\varepsilon \cdot |t-t'|^2 \cdot n). \tag{42}$$

Using the union bound over all $t' < t$ we calculate

$$\Pr\big(\bar{\mathcal{S}_t}\big) \leq \sum_{t'<t} \Pr\big(\Phi(t') \geq |t-t'|^2 \cdot n^2\big) \leq \frac{1}{\varepsilon n} \cdot \sum_{t'<t} \frac{1}{|t-t'|^2} \leq \frac{\pi^2}{6\varepsilon \cdot n} = \mathrm{O}\big(n^{-1}\big),$$

where the last inequality uses the solution to the Basel problem. This proves the first statement.

For the second statement, let $Z_i := |\mathcal{I}_i| \cdot n - Y_i$ be the number of balls that did not spawn during $\mathcal{I}_i$. Note that Z_i is a sum of $|\mathcal{I}_i| \cdot n$ independent indicator variables with $\mathbb{E}[Z_i] = (1-\lambda) \cdot |\mathcal{I}_i| \cdot n = 8i \cdot \ln n$. Chernoff yields $\Pr(Z_i \leq (1-\lambda) \cdot |\mathcal{I}_i| \cdot n/2) \leq e^{-8i \cdot \ln n/8} = n^{-i}$. The desired statement follows from applying the identity $Z_i = |\mathcal{I}_i| \cdot n - Y_i$ and taking the union bound. □

Lemma 10. *Fix a round t and assume that both $\mathcal{S}_t$ and $\mathcal{B}_t$ hold. Then,*

$$\Psi(\boldsymbol{X}(t)) \leq \frac{9n}{\alpha} \cdot \ln\!\left(\frac{n}{1-\lambda}\right). \tag{43}$$

Proof. Let $T < t$ be the last time when there was an empty bin and set $\Delta := t - T$. Note that T is well defined, as we have $X_i(0) = 0$ for all $i \in [n]$. Since $\mathcal{S}_t$ holds, we have

$$\Phi(\boldsymbol{X}(T)) \leq \Delta^2 \cdot n^2 = \exp\!\big(\ln(\Delta^2 \cdot n^2)\big). \tag{44}$$

By definition of T, we have $\min_i X_i(T) = 0$. Together with Observation 1 we get

$$\max_i X_i(T) \leq 2\ln\bigl(\Delta^2 \cdot n^2\bigr)/\alpha. \tag{45}$$

Summing up over all bins (and pulling out the square), this implies that $\Psi(\boldsymbol{X}(T)) \leq 4n \cdot \ln\bigl(\Delta \cdot n\bigr)/\alpha$. Applying Lemma 8 yields

$$\Psi(\boldsymbol{X}(T+1)) \leq 4n \cdot \ln\bigl(\Delta \cdot n\bigr)/\alpha. \tag{46}$$

By the definition of T, is must be the case that there is no empty bin in $\boldsymbol{X}(t'')$ for all $t'' \in \{\, T+1, T+2, \dots, t-1 \,\}$. Thus, during each of these rounds exactly n balls are deleted. To bound the number of deleted balls, let i be maximal with $\mathcal{I}_i \subseteq [T, t]$ (as defined in Lemma 9). Recall that $\mathcal{I}_i = [t - i \cdot \frac{8\ln n}{1-\lambda}, t]$. Since $\mathcal{B}_t$ holds and using the maximality of i, the number of balls Y that spawn during $[T, t]$ is bounded by

$$(1+\lambda)|\mathcal{I}_i| \cdot n/2 + \frac{8\ln n}{1-\lambda} \cdot n \leq (1+\lambda)\Delta \cdot n/2 + \frac{8\ln n}{1-\lambda} \cdot n. \tag{47}$$

We calculate

$$\begin{aligned}
\Psi(\boldsymbol{X}(t)) &\leq \Psi(\boldsymbol{X}(T+1)) - \Delta \cdot n + Y \\
&\leq \frac{4n}{\alpha}\ln(\Delta \cdot n) - \frac{1-\lambda}{2}\Delta \cdot n + \frac{8\ln n}{1-\lambda} \cdot n \\
&= \frac{1-\lambda}{2} \cdot n \cdot \left(\frac{8}{\alpha(1-\lambda)} \cdot \ln(\Delta \cdot n) - \Delta + \frac{16\ln n}{(1-\lambda)^2}\right) \\
&\leq \frac{1-\lambda}{2} \cdot \Delta \cdot n \cdot \left(\frac{24}{\alpha(1-\lambda)^2} \cdot \frac{\ln(\Delta \cdot n)}{\Delta} - 1\right).
\end{aligned} \tag{48}$$

With $f = f(\lambda) \coloneqq 24/\bigl(\alpha(1-\lambda)^2\bigr)$ the last factor becomes $f \cdot \ln(\Delta \cdot n)/\Delta - 1$. It is negative if and only if $\Delta > f \cdot \ln(\Delta \cdot n)$. This inequality holds for any $\Delta > -f \cdot W_{-1}(-\frac{1}{f \cdot n})$, where W_{-1} denotes the lower branch of the Lambert W function[5]. This implies that $\Delta \leq -f \cdot W_{-1}(-\frac{1}{f \cdot n})$, since otherwise we would have $\Psi(\boldsymbol{X}(t)) < 0$, which is clearly a contradiction. Using the Taylor approximation $W_{-1}(x) = \ln(-x) - \ln\bigl(\ln(-1/x)\bigr) - \mathrm{o}(1)$ as $x \to -0$, we get

$$\Delta \leq -f \cdot W_{-1}\left(-\frac{1}{f \cdot n}\right) \leq f \cdot \ln(f \cdot n) + f \cdot \ln\bigl(\ln(f \cdot n)\bigr) + f \leq 2f \cdot \ln(f \cdot n). \tag{49}$$

Finally, we use this bound on Δ to get

$$\begin{aligned}
\Psi(\boldsymbol{X}(t)) \leq \Psi(\boldsymbol{X}(+1)) &\leq \frac{4n}{\alpha} \cdot \ln(\Delta \cdot n) \leq \frac{4n}{\alpha} \cdot \ln\bigl(2fn \cdot \ln(fn)\bigr) \\
&\leq \frac{4n}{\alpha} \cdot \ln\left(\frac{48n}{\alpha(1-\lambda)^2} \cdot \ln\left(\frac{24n}{\alpha(1-\lambda)^2}\right)\right) \leq \frac{9n}{\alpha} \cdot \ln\left(\frac{n}{1-\lambda}\right).
\end{aligned} \tag{50}$$

□

[5] Note that $-\frac{1}{f \cdot n} \geq -1/e$, so that $W_{-1}(-\frac{1}{f \cdot n})$ is well defined.

Now, by combining Lemma 10 with the fact that the events $\mathcal{S}_t$ and $\mathcal{B}_t$ hold with high probability (Lemma 9), we immediately get that (w.h.p.) $\Psi(\boldsymbol{X}(t)) = \mathrm{O}(n \cdot \ln n)$. As described at the beginning of this section, combining this with Proposition 1 proves Theorem 5.

3.3 Stability – Proof of Theorem 4

This section proves Theorem 4. In order to do so, we consider the potential Γ (defined in Equation 14) and show that, for a sufficiently high value of, this potential decreases (11).[6] To show this drop, we argue along the following lines. For the potential to be large and since the potential is the sum of two potentials Φ and Ψ, one of must have size at least $\Gamma(\boldsymbol{x})/2$. If $\Phi(\boldsymbol{x})$ is large, then we can even assume a worst-case increase of Ψ and invoke Equation (39) to show that Φ drops considerably resulting in an overall potential drop of Γ. Similarly, if $\Psi(\boldsymbol{x}) \geq \Gamma(\boldsymbol{x})/2$, then, due to the careful construction of Γ, we can show that all bins are non-empty, and the overall potential decreases in expectation. This overall potential decrease of Γ allows to apply Theorem 6 yielding stability.

Lemma 11 (Negative Bias Γ)**.** *Let $\lambda \in [1/4, 1)$. If $\Gamma(\boldsymbol{X}(t)) \geq 2\frac{n^4}{(1-\lambda)^2\lambda}$, then*

$$\mathbb{E}[\Gamma(\boldsymbol{X}(t+1)) - \Gamma(\boldsymbol{X}(t)) \mid \boldsymbol{X}(t)] \leq -1. \tag{51}$$

Proof. Assume $\boldsymbol{X}(t) = x$ is fixed. By definition of $\Gamma(\cdot)$, we have $\Phi(x) \geq \Gamma(x)/2$ or $\Psi(x) \geq \Gamma(x)/2$. We now show that in both cases

$$\mathbb{E}[\Gamma(\boldsymbol{X}(t+1)) - \Gamma(x) \mid \boldsymbol{X}(t) = \boldsymbol{x}] \leq -1. \tag{52}$$

1. If $\Phi(\boldsymbol{x}) \geq \Gamma(\boldsymbol{x})/2$, then we have, by Equation (39), a potential drop of

$$\begin{aligned}\mathbb{E}[\Phi(\boldsymbol{X}(t+1)) - \Phi(\boldsymbol{x}) \mid \boldsymbol{X}(t) = \boldsymbol{x}] &\leq -(\varepsilon\alpha\lambda/4) \cdot \Phi(\boldsymbol{x}) + n \log n \\ &\leq -(\varepsilon\alpha\lambda/8) \cdot \Gamma(\boldsymbol{x}) + n \log n.\end{aligned} \tag{53}$$

Note that, by definition of Ψ, $\Psi(\boldsymbol{X}(t+1)) - \Psi(\boldsymbol{x}) \leq n$. Together with $\Gamma(\boldsymbol{x}) \geq \frac{8(n \log n + n^2/(1-\lambda)+1)}{\epsilon\alpha\lambda}$,

$$\begin{aligned}&\mathbb{E}[\Gamma(\boldsymbol{X}(t+1)) - \Gamma(\boldsymbol{x}) \mid \boldsymbol{X}(t) = \boldsymbol{x}] \\ \leq\ &-\frac{\varepsilon\alpha\lambda}{8}\Gamma(\boldsymbol{x}) + n \log n + (n/(1-\lambda)) \cdot n \leq -1.\end{aligned} \tag{54}$$

2. Otherwise, i.e., if $\Phi(\boldsymbol{x}) < \Gamma(\boldsymbol{x})/2$, we have that

   (i) the load difference is, by Observation 1, bounded by $2\ln(\Gamma(\boldsymbol{x})/2)/\alpha$, and

[6] It might look tempting to use Γ together with Hajek's theorem to bound the maximum system load. However, this would require (exponentially) sharper bounds on Φ. Furthermore, it might be tempting to use the stability of Greedy[1] to prove stability of Greedy[2], however, as discussed earlier, it is not clear to achieve this, as it seems challenging to couple or majorize the processes.

(ii) $\Psi(\boldsymbol{x}) \geq \Gamma(\boldsymbol{x})/2$ must hold. This implies that $\varnothing \geq \frac{1}{n}\left(\frac{\Gamma(\boldsymbol{x})/2}{\frac{n}{1-\lambda}}\right) = \frac{(1-\lambda)\cdot\Gamma(\boldsymbol{x})}{2n^2}$.

From (i) and (ii) we have that the minimum load is at least $\frac{(1-\lambda)\cdot\Gamma(\boldsymbol{x})}{2n^2} - \ln(\Gamma(\boldsymbol{x})/2)/\alpha$. From Lemma 12 and $\Gamma(\boldsymbol{x}) \geq 2\frac{n^4}{(1-\lambda)^2\lambda}$, it follows that every bin has load at least load 1. Thus each bin will delete one ball and the number of balls arriving is λn in expectation. Hence,

$$\mathbb{E}[\Psi(\boldsymbol{X}(t+1)) - \Psi(\boldsymbol{x}) \mid \boldsymbol{X}(t) = \boldsymbol{x}] = -\frac{n}{1-\lambda}(1-\lambda)n. \tag{55}$$

Now,

$$\begin{aligned} &\mathbb{E}[\Gamma(\boldsymbol{X}(t+1)) - \Gamma(\boldsymbol{x}) \mid \boldsymbol{X}(t) = \boldsymbol{x}] \\ &= \mathbb{E}[\Phi(\boldsymbol{X}(t+1)) - \Phi(\boldsymbol{x}) \mid \boldsymbol{X}(t) = \boldsymbol{x}] - \frac{n}{1-\lambda}(1-\lambda)n \\ &\leq n\log n - \frac{n}{1-\lambda}(1-\lambda)n \leq -1. \end{aligned} \tag{56}$$

Thus, $\mathbb{E}[\Gamma(\boldsymbol{X}(t+1)) - \Gamma(\boldsymbol{x}) \mid \boldsymbol{X}(t) = \boldsymbol{x}] \leq -1$, which yields the claim. □

We now proceed with a technical result.

Lemma 12. *For all $x \geq 2\frac{n^4}{(1-\lambda)^2\lambda}$ it holds that $\frac{(1-\lambda)\cdot x}{2n^2} - 2\ln(x/2)/\alpha \geq 1$.*

Proof. Define $f(x) = \frac{(1-\lambda)\cdot x}{2n^2} - 2\ln(x/2)/\alpha$. We have $f\left(2\frac{n^4}{(1-\lambda)^2\lambda}\right) \geq \frac{n^2}{(1-\lambda)\lambda} - \frac{2}{\alpha}\ln\left(\frac{n^4}{(1-\lambda)^2\lambda}\right) \geq 1$, where the last inequality holds for large enough of n since α is a constant. Moreover, for all $x \geq 2\frac{n^4}{(1-\lambda)^2\lambda}$ we have $f'(x) = \frac{1-\lambda}{n^2} - \frac{2}{\alpha x} \geq 0$. Thus, the claim follows. □

We are ready to prove Theorem 4.

of Theorem 4. The proof proceeds by applying Theorem 6. We now define the parameters of Theorem 6. Let $\zeta(t) = \boldsymbol{X}(t)$ and hence Ω is the state space of X. First we observe that Ω is countable since there are a constant number of bins (n is consider a constant in this matter) each having a load which is a natural number. We define $\phi(\boldsymbol{X}(t))$ to be $\Gamma(\boldsymbol{X}(t))$. We define $C = \{\, x \mid \Gamma(x) \leq 2\frac{n^4}{(1-\lambda)^2\lambda} \,\}$. Define $\beta(x) = 1$ and $\eta = 1$. We now show that the preconditions (a) and (b) of Theorem 6 are fulfilled.

- Let $x \notin C$. By definition of C and $\phi(\boldsymbol{X}(t))$, and from Lemma 11 we have

$$\begin{aligned} &\mathbb{E}[\phi(X(t+1)) - \phi(x) \mid \boldsymbol{X}(t) = x] \\ &\leq \mathbb{E}[\Gamma(\boldsymbol{X}(t+1)) - \Gamma(x) \mid \boldsymbol{X}(t) = x] \leq -1. \end{aligned} \tag{57}$$

- Let $x \in C$. Recall that $\Gamma(\boldsymbol{X}(t)) = \Phi(\boldsymbol{X}(t)) + \Psi(\boldsymbol{X}(t))$. By Lemma 7 and the fact that the number of balls arriving in one round is bounded by n, we derive,

$$\begin{aligned}
&\mathbb{E}[\phi(X(t+1)) \mid \boldsymbol{X}(t) = x] = \\
&= \mathbb{E}[\Phi(\boldsymbol{X}(t+1)) \mid \boldsymbol{X}(t) = x] + \mathbb{E}[\Psi(\boldsymbol{X}(t+1)) \mid \boldsymbol{X}(t) = x] \\
&\leq \left(\left(1 - \frac{\varepsilon\alpha\lambda}{4}\right) 2 \frac{n^4}{(1-\lambda)^2\lambda}\right) + \frac{n}{1-\lambda} n < \infty.
\end{aligned} \tag{58}$$

The claim follows by applying Theorem 6 with Equations (57) and (58). □

4 Conclusion

Our results show that the power of two choices carries over to generalized setting with deletion: Similar to the classic setting without deletions, the maximum load under Greedy[[] 2] is exponentially smaller than the load under Greedy[[] 1]. Moreover, Greedy[2] can handle much larger arrival rates w.r.t. the maximum load difference.

One might assume that our (upper) bounds for Greedy[1] carry over to Greedy[2] (and, in general, to Greedy[d]) via a simple coupling (similar to [4]). However, we are not aware of such a coupling in the *parallel* setting. In fact, for naïve approaches to such a coupling, it is not hard to come up with situations where Greedy[2] behaves worse than Greedy[1] (in one step). It would be interesting to find arguments that, for example, for any $d \in \mathbb{N}$ Greedy[$d+1$] behaves "better" than Greedy[d].

Another open questions is concerned with arrival rates $\lambda \geq 1$ (this would require a slight reformulation of our model, which currently assumes the existence of n generators that generate balls with a probability of λ). As mentioned in Section 3.1, our assumptions on λ for proving bounds on the smoothness (Proposition 1) are merely for convenience. The corresponding proofs carry over (with minor modifications) to any constant λ, no matter whether $\lambda < 1$ or $\lambda > 1$. Thus, for Greedy[2] we know that the load difference between any two bins is still logarithmic, even for arrival rates > 1. Still, the maximum load obviously diverges for $\lambda \geq 1$. It would be interesting to quantify this divergence in terms of λ.

Compliance with Ethical Standards

Funding: Petra Berenbrink is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). Peter Kling is partly supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Pacific Institute for the Mathematical Sciences (PIMS). Lars Nagel is supported by the German Ministry of Education and Research under Grant 01IH13004. Christopher Wastell is supported by EPSRC.

Conflict of Interest: The authors declare that they have no conflict of interest.

References

[1] Micah Adler, Petra Berenbrink, and Klaus Schröder. Analyzing an infinite parallel job allocation process. In *Proceedings of the 6th Annual European Symposium on Algorithms*, ESA, pages 417–428, London, UK, UK, 1998. Springer-Verlag. ISBN 3-540-64848-8. URL http://dl.acm.org/citation.cfm?id=647908.740138.

[2] Micah Adler, Soumen Chakrabarti, Michael Mitzenmacher, and Lars Rasmussen. Parallel randomized load balancing. *Random Structures & Algorithms*, 13(2):159–188, 1998. ISSN 1098-2418. doi: 10.1002/(SICI)1098-2418(199809)13:2⟨159::AID-RSA3⟩3.0.CO;2-Q. URL http://dx.doi.org/10.1002/(SICI)1098-2418(199809)13:2<159::AID-RSA3>3.0.CO;2-Q.

[3] Attahiru Sule Alfa. Algorithmic analysis of the bmap/d/k system in discrete time. *Adv. in Appl. Probab.*, 35(4):1131–1152, 12 2003. doi: 10.1239/aap/1067436338. URL http://dx.doi.org/10.1239/aap/1067436338.

[4] Yossi Azar, Andrei Z. Broder, Anna R. Karlin, and Eli Upfal. Balanced allocations. *SIAM Journal on Computing*, 29(1):180–200, 1999. doi: 10.1137/S0097539795288490.

[5] Luca Becchetti, Andrea E. F. Clementi, Emanuele Natale, Francesco Pasquale, and Gustavo Posta. Self-stabilizing repeated balls-into-bins. In Guy E. Blelloch and Kunal Agrawal, editors, *Proceedings of the 27th ACM on Symposium on Parallelism in Algorithms and Architectures, SPAA 2015, Portland, OR, USA, June 13-15, 2015*, pages 332–339. ACM, 2015. ISBN 978-1-4503-3588-1. doi: 10.1145/2755573.2755584. URL http://doi.acm.org/10.1145/2755573.2755584.

[6] Petra Berenbrink, Artur Czumaj, Tom Friedetzky, and Nikita D. Vvedenskaya. Infinite parallel job allocation (extended abstract). In *Proceedings of the Twelfth Annual ACM Symposium on Parallel Algorithms and Architectures*, SPAA, pages 99–108, New York, NY, USA, 2000. ACM. ISBN 1-58113-185-2. doi: 10.1145/341800.341813. URL http://doi.acm.org/10.1145/341800.341813.

[7] Petra Berenbrink, Artur Czumaj, Angelika Steger, and Berthold Vöcking. Balanced allocations: The heavily loaded case. *SIAM Journal on Computing*, 35(6):1350–1385, 2006. doi: 10.1137/S009753970444435X.

[8] Petra Berenbrink, Artur Czumaj, Matthias Englert, Tom Friedetzky, and Lars Nagel. Multiple-choice balanced allocation in (almost) parallel. In Anupam Gupta, Klaus Jansen, José Rolim, and Rocco Servedio, editors,

Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, volume 7408 of *Lecture Notes in Computer Science*, pages 411–422. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-32511-3. doi: 10.1007/978-3-642-32512-0_35. URL http://dx.doi.org/10.1007/978-3-642-32512-0_35.

[9] Petra Berenbrink, Kamyar Khodamoradi, Thomas Sauerwald, and Alexandre Stauffer. Balls-into-bins with nearly optimal load distribution. In *Proceedings of the Twenty-fifth Annual ACM Symposium on Parallelism in Algorithms and Architectures*, SPAA, pages 326–335, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1572-2. doi: 10.1145/2486159.2486191. URL http://doi.acm.org/10.1145/2486159.2486191.

[10] A. Czumaj and V. Stemann. Randomized allocation processes. In *Foundations of Computer Science, 1997. Proceedings., 38th Annual Symposium on*, pages 194–203, Oct 1997. doi: 10.1109/SFCS.1997.646108.

[11] Artur Czumaj. Recovery time of dynamic allocation processes. In *Proceedings of the Tenth Annual ACM Symposium on Parallel Algorithms and Architectures*, SPAA, pages 202–211, New York, NY, USA, 1998. ACM. ISBN 0-89791-989-0. doi: 10.1145/277651.277686. URL http://doi.acm.org/10.1145/277651.277686.

[12] G. Fayolle, V.A. Malyshev, and M.V. Menshikov. *Topics in the Constructive Theory of Countable Markov Chains*. Cambridge University Press, 1995. ISBN 9780521461979. URL https://books.google.ca/books?id=1TJ1tFEnnHcC.

[13] Gaston H. Gonnet. Expected length of the longest probe sequence in hash code searching. *J. ACM*, 28(2):289–304, April 1981. ISSN 0004-5411. doi: 10.1145/322248.322254. URL http://doi.acm.org/10.1145/322248.322254.

[14] B. Hajek. Hitting-time and occupation-time bounds implied by drift analysis with applications. *Advances in Applied Probability*, 14(3):502–525.

[15] A.E. Kamal. Efficient solution of multiple server queues with application to the modeling of atm concentrators. In *INFOCOM. Fifteenth Annual Joint Conference of the IEEE Computer Societies. Networking the Next Generation. Proceedings IEEE*, volume 1, pages 248–254 vol.1, Mar 1996. doi: 10.1109/INFCOM.1996.497900.

[16] Nam K Kim, Mohan L Chaudhry, Bong K Yoon, and Kilhwan Kim. A complete and simple solution to a discrete-time finite-capacity bmap/d/c queue. pages 2169–2173, 2012.

[17] David A. Levin and Yuval Perres. *Markov Chains and Mixing Times*. American Mathematical Society, December 2008. ISBN 978-0-8218-4739-8.

[18] Michael Mitzenmacher. The power of two choices in randomized load balancing. *IEEE Trans. Parallel Distrib. Syst.*, 12(10):1094–1104, 2001. doi: 10.1109/71.963420. URL http://doi.ieeecomputersociety.org/10.1109/71.963420.

[19] Yuval Peres, Kunal Talwar, and Udi Wieder. The $(1 + \beta)$-choice process and weighted balls-into-bins. In *Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, SODA, pages 1613–1619, Philadelphia, PA, USA, 2010. Society for Industrial and Applied Mathematics. ISBN 978-0-898716-98-6.

[20] Martin Raab and Angelika Steger. "balls into bins" - A simple and tight analysis. In Michael Luby, José D. P. Rolim, and Maria J. Serna, editors, *Randomization and Approximation Techniques in Computer Science, Second International Workshop, RANDOM, Barcelona, Spain, October 8-10, 1998, Proceedings*, volume 1518 of *Lecture Notes in Computer Science*, pages 159–170. Springer, 1998. ISBN 3-540-65142-X. doi: 10.1007/3-540-49543-6_13. URL http://dx.doi.org/10.1007/3-540-49543-6_13.

[21] Khosrow Sohraby and Ji Zhang. Spectral decomposition approach for transient analysis of multi-server discrete-time queues. In *INFOCOM. Eleventh Annual Joint Conference of the IEEE Computer and Communications Societies, IEEE*, pages 395–404. IEEE, 1992.

[22] Volker Stemann. Parallel balanced allocations. In *Proceedings of the Eighth Annual ACM Symposium on Parallel Algorithms and Architectures*, SPAA, pages 261–269, New York, NY, USA, 1996. ACM. ISBN 0-89791-809-6. doi: 10.1145/237502.237565. URL http://doi.acm.org/10.1145/237502.237565.

[23] Kunal Talwar and Udi Wieder. Balanced allocations: A simple proof for the heavily loaded case. *CoRR*, abs/1310.5367, 2013. URL http://arxiv.org/abs/1310.5367.

[24] Kunal Talwar and Udi Wieder. Balanced allocations: A simple proof for the heavily loaded case. In Javier Esparza, Pierre Fraigniaud, Thore Husfeldt, and Elias Koutsoupias, editors, *Automata, Languages, and Programming*, volume 8572 of *Lecture Notes in Computer Science*, pages 979–990. Springer Berlin Heidelberg, 2014. ISBN 978-3-662-43947-0. doi: 10.1007/978-3-662-43948-7_81. URL http://dx.doi.org/10.1007/978-3-662-43948-7_81.

A Auxiliary Results

The following theorem gives necessary and sufficient conditions for a Markov Chain to be positive recurrent. Roughly speaking, we want that the Markov

Chain returns to a finite number of states: Given the state x of the system at some time t, we need to show that there exists some potential $\zeta(\cdot)$ which is decreasing linearly (in $\beta(x)$, the length of a suitably chosen period) for "most states" (first condition). For a finite number of states (set C), whose size can be a function of n but not of t, the potential is not required to decrease but the potential is required to be finite after $\beta(x)$ time steps (second condition).

Theorem 6 (Fayolle et al. [12, Theorem 2.2.4])**.** *A time-homogeneous irreducible aperiodic Markov chain ζ with a countable state space Ω is positive recurrent if and only if there exists a positive function $\phi(x), x \in \Omega$, a number $\eta > 0$, a positive integer-valued function $\beta(x), x \in \Omega$, and a finite set $C \subseteq \Omega$ such that the following inequalities hold:*

1. $\mathbb{E}[\phi(\zeta(t+\beta(x))) - \phi(x) \mid \zeta(t) = x] \leq -\eta\beta(x)$, $x \notin C$
2. $\mathbb{E}[\phi(\zeta(t+\beta(x))) \mid \zeta(t) = x] < \infty$, $x \in C$

The following theorem gives a tail bound on a potential for which the following two properties hold: In increase in the potential has a tail bound (first condition) and whenever the potential is large, in decreases in expectation (second condition).

Theorem 7 (Simplified version of Hajek [14, Theorem 2.3])**.** *Let $(Y(t))_{t\geq 0}$ be a sequence of random variables on a probability space $(\Omega, \mathcal{F}, P)$ with respect to the filtration $(\mathcal{F}(t))_{t\geq 0}$. Assume the following two conditions hold:*

(i) (Majorization) There exists a random variable Z and a constant $\lambda' > 0$, such that $\mathbb{E}\left[e^{\lambda' Z}\right] \leq D$ for some finite D, and $(|Y(t+1) - Y(t)| \big| \mathcal{F}(t)) \prec Z$ for all $t \geq 0$; and

(ii) (Negative Bias) There exist $a, \varepsilon_0 > 0$, such for all t we have

$$\mathbb{E}[Y(t+1) - Y(t) \mid \mathcal{F}(t), Y(t) > a] \leq -\varepsilon_0.$$

Let $\eta = \min\{\, \lambda', \varepsilon_0 \cdot \lambda'^2/(2D), 1/(2\varepsilon_0) \,\}$. Then, for all b and t we have

$$\Pr(Y(t) \geq b \mid \mathcal{F}(0)) \leq e^{\eta(Y(0)-b)} + \frac{2D}{\varepsilon_0 \cdot \eta} \cdot e^{\eta(a-b)}.$$

Proof. The statement of the theorem provided in [14] requires besides (i) and (ii) to choose constants η, and ρ such that $0 < \rho \leq \lambda'$, $\eta < \varepsilon_0/c$ and $\rho = 1 - \varepsilon_0 \cdot \eta + c\eta^2$ where $c = \frac{\mathbb{E}\left[e^{\lambda' Z}\right] - (1+\lambda'\mathbb{E}[Z])}{\lambda'^2} = \sum_{k=2}^{\infty} \frac{\lambda'^{k-2}}{k!}\mathbb{E}\left[Z^k\right]$. With these requirements it then holds that for all b and t

$$\Pr(Y(t) \geq b \mid \mathcal{F}(0)) \leq \rho^t e^{\eta(Y(0)-b)} + \frac{1-\rho^t}{1-\rho} \cdot D \cdot e^{\eta(a-b)}. \tag{59}$$

In the following we bound Equation (59) by setting $\eta = \min\{\, \lambda', \varepsilon_0 \cdot \lambda'^2/(2D), 1/(2\varepsilon_0) \,\}$. The following upper and lower bound on ρ follow.

- $\rho = 1 - \varepsilon_0 \cdot \eta + c\eta^2 \le 1 - \varepsilon_0 \cdot \eta + \varepsilon_0 \cdot \eta \cdot c \cdot \lambda'^2/(2D) \le 1 - \varepsilon_0 \cdot \eta + \varepsilon_0 \cdot \eta/2 = 1 - \varepsilon_0 \cdot \eta/2$, where we used $c \le D/\lambda'^2$.
- $\rho = 1 - \varepsilon_0 \cdot \eta + c\eta^2 \ge 1 - \varepsilon_0/(2\varepsilon_0) \ge 0$.

We derive, from Equation (59) using that for any $t \ge 0$ we have $0 \le \rho^t \le 1$

$$\begin{aligned}\Pr(Y(t) \ge b \mid \mathcal{F}(0)) &\le \rho^t e^{\eta(Y(0)-b)} + \frac{1-\rho^t}{1-\rho} \cdot D \cdot e^{\eta(a-b)} \le e^{\eta(Y(0)-b)} + \frac{1}{1-\rho} \cdot D \cdot e^{\eta(a-b)} \\ &\le e^{\eta(Y(0)-b)} + \frac{2D}{\varepsilon_0 \cdot \eta} \cdot e^{\eta(a-b)},\end{aligned} \tag{60}$$

since $\frac{1}{(1-\rho)} \le \frac{2}{\varepsilon_0 \cdot \eta}$. This yields the claim. □

Theorem 8 (Raab and Steger [20, Theorem 1])**.** *Let M be the random variable that counts the maximum number of balls in any bin, if we throw m balls independently and uniformly at random into n bins. Then* $\Pr(M > k_\alpha) = \mathrm{o}(1)$ *if* $\alpha > 1$ *and* $\Pr(M > k_\alpha) = 1 - \mathrm{o}(1)$ *if* $0 < \alpha < 1$*, where*

$$k_\alpha = \begin{cases} \frac{\log n}{\log \frac{n \log n}{m}} \left(1 + \alpha \frac{\log\log \frac{n\log n}{m}}{\log \frac{n \log n}{m}}\right) & \textit{if } \frac{n}{\mathrm{polylog}(n)} \le m \ll n \log n \\ (d_c - 1 + \alpha)\log n & \textit{if } m = c \cdot n \log n \textit{ for some constant } c \\ \frac{m}{n} + \alpha\sqrt{2\frac{m}{n}\log n} & \textit{if } n \log n \ll m \le n\,\mathrm{polylog}(n) \\ \frac{m}{n} + \sqrt{2\frac{m}{n}\log n\left(1 - \frac{1}{\alpha}\frac{\log\log n}{2\log n}\right)} & \textit{if } m \gg n(\log n)^3, \end{cases}$$

where d_c is largest solution of $1 + x(\log c - \log x + 1) - c = 0$*. We have* $d_1 = e$ *and* $d_{1.00001} = 2.7183$*.*