Balanced Allocations: A Simple Proof for the Heavily Loaded Case

Kunal Talwar and Udi Wieder

Microsoft Research, {kunal,uwieder}@microsoft.com

Abstract. We give a new proof for the fact that the expected gap between the maximum load and the average load in the two-choice process is bounded by $(1 + o(1)) \log\log n$, irrespective of the number of balls thrown. The original proof of this surprising result, due to Berenbrink et al. in [2], uses tools from Markov chain theory, and a sophisticated induction proof involving computer-aided calculations. We provide a significantly simpler and more elementary proof. The new technique allows us to generalize the result and derive new and often tight bounds for the case of weighted balls. The simplification comes at a cost of larger lower order terms and a weaker tail bound for the probability of deviating from the expectation.

1 Introduction

Balls-and-Bins processes are a name for randomized allocations processes, typically used to model the performance of hashing or more general load balancing schemes. Suppose there are m balls (think items) to be thrown into n bins (think hash buckets). We want a simple process that will keep the loads balanced, while allowing quick decentralized lookup. One of the simplest such process is the Balls-and-Bins process where balls are placed sequentially via some simple randomized allocation process, and not moved once placed. There are other approaches to the problem that we will not discuss here. A significant body of work had been amassed on the analysis of simple and natural versions of these Balls-and-Bins processes. In this work we present a simpler analysis for the heavily loaded case for many of these processes.

In the Greedy[d] process (sometimes called the d-choice process), balls are placed sequentially into n bins with the following rule: Each ball is placed by uniformly and independently sampling d bins and assigning the ball to the least loaded of the d bins[1]. In this paper we are interested in the *gap* of the allocation, which is the difference between the number of balls in the heaviest bin, and the average. The case $d = 1$, when balls are placed uniformly at random in the bins, is well understood. In particular when n balls are thrown, the bin with the largest number of balls will have $\Theta(\log n/ \log\log n)$ balls w.h.p. Since the average is 1, asymptotically this is also the gap. If $m \gg n$ balls are thrown the

[1] Assume for simplicity and w.l.o.g that ties are broken according to some fixed ordering of the bins.

heaviest bin will have $m/n + \Theta(\sqrt{m \log n/n})$ balls w.h.p. [9]. In other words $\text{Gap}(m) = \Theta(\sqrt{m \log n/n})$ w.h.p.

In an influential paper Azar et al. [1] showed that when n balls are thrown and $d > 1$ the gap[2] is $\log\log n/\log d + O(1)$ w.h.p. The case $d = 2$ is implicitly shown in Karp et al. [4]. The proof by Azar et al. uses a simple but clever induction; in our proof here we take the same approach. The proof in [1] breaks down once the number of balls is super-linear in the number of bins. Two other approaches to prove this result, namely, using differential equations or witness trees, also fail when the number of balls is large (see for example the survey [5]). A breakthrough was achieved by Berenbrink et al. in [2]:

Theorem 1 ([2]). *For every $c > 0$ there is a $\gamma = \gamma(c)$ so that for any given $m \in \mathbb{N}$,*

$$\Pr[Gap(m) \geq \frac{\log\log n}{\log d} + \gamma] \leq n^{-c}$$

Thus the additive gap remains at $\log\log n$ even after $m \gg n$ balls are thrown! Contrast this with the one choice case in which the gap diverges with the number of balls. At a (very) high level their approach was the following: first they show that the gap after m balls are thrown is distributed similarly to the gap after only *poly*(n) balls are thrown. This is done by bounding the mixing time of an underlying Markov chain. The second step is to extend the induction technique of [1] to the case of *poly*(n) balls. This turns out to be a major technical challenge which involves four inductive invariants and computer-aided calculations. Furthermore, whenever the model is tweaked, this technically difficult part of the proof needs to be redone, making such changes challenging. As such, finding a simpler proof has remained an interesting open problem [7].

1.1 Our Contributions

In this paper we present a simple proof for a bound similar to that of Theorem 1.

Theorem 2. *For any m, the expected gap between maximum and average of Greedy$[d]$ is at most $\frac{\log\log n}{\log d} + O(\log\log\log n)$. Moreover for an absolute constant c,*

$$\Pr[Gap(m) > \frac{\log\log n}{\log d} + c\log\log\log n] \leq c(\log\log n)^{-4}$$

Our proof builds on the *layered induction* approach of Azar et al. [1]. The basic layered induction approach bounds the number of bins containing at least h balls, by using an induction on h. This approach runs into several issues when trying to go beyond $O(n)$ balls, the most crucial of these is establishing the base case for the induction: When the number of balls is n it trivially holds that the number of bins that received at least 2 balls is at most $n/2$. When $m >> n$ there

[2] Unless otherwise stated, all logs in this paper are base 2.

is no straightforward argument to claim that the number of bins with load above the average is at most $n/2$. We show that the potential function approach of [8] allows us to surmount these hurdles: the bound from the potential function lets us restrict ourselves to the last $\tilde{O}(n \log n)$ balls, and also gives us a suitable base case for the layered induction.

Our proof is relatively short and accessible. This simplicity comes at a price. Our bound is slightly weaker than Theorem 1 as it has larger lower order terms. We also have a weaker tail bound on the probability of deviation from expectation (see Section 2).

On the positive side the simple proof structure allows for easier generalization and we can obtain bounds on similar processes without much added difficulty. These include a bound on Vöcking's Left$[d]$ process (also shown in [2]) which we present in Section 3.2, as well as tight bounds on processes with weighted balls, which were previously unknown. For instance suppose that each ball has a weight sampled uniformly from the set $\{1, 2\}$. in Section 3.1, we show that the gap is upper bounded by $2 \log \log n$ up to lower order terms. This improves on the previously best known bound of $O(\log n)$.

We also show lower bounds for the weighted case that match our upper bounds for several interesting distributions. In particular, for the case of weights in $\{1, 2\}$, we show that the upper bound of $2 \log \log n$ is tight up to lower order terms.

Another way to characterize the d-choice process is by defining the probability a ball is placed in one of the i heaviest bins (at the time when it is placed) to be exactly $(i/n)^d$. We remark that using this characterization, there is no need to assume that d is a natural number. While the process is algorithmically simpler to describe when d is an integer, natural cases arise in which d is not an integer, c.f. [12]. Our approach, being based on layered induction, naturally extends to this setting for any $d > 1$.

2 The Main Proof

We define the normalized *load vector* X^t to be an n dimensional vector where X_i^t is the difference between the load of the i'th bin after tn balls are thrown and the average t, (so that a load of a bin can be negative and $\sum X_i = 0$). We also assume without loss of generality that the vector is sorted so that $X_1^t \geq X_2^t \geq ... \geq X_n^t$. We will consider a Markov chain with state space X^t, where one step of the chain consists of throwing n balls according to the d-choice scheme and then sorting and normalizing the load vector.

The main tool we use is the following Theorem proven in [8] using a potential function argument.

Theorem 3 ([8]). *For every $d > 1$ there exist positive constants a and b such that for all n and all t,*

$$\mathbb{E}\left[\sum_i \exp\left(a|X_i^t|\right)\right] \leq bn.$$

Let $G^t \overset{def}{=} X_1^t$ denote the gap between maximum and average when sampling from X^t. Applying Markov's inequality to Theorem 3 immediately implies the following:

Lemma 1. *For any t, any $c \geq 0$, $\Pr[G^t \geq \frac{c \log n}{a}] \leq \frac{bn}{n^c}$. Thus for every c there is a $\gamma = \gamma(c)$ such that $\Pr[G^t \geq \gamma \log n] \leq n^{-c}$.*

We remark that Theorem 3 is a statement about the absolute values of the X_i^t's and thus a version of Lemma 1 holds also for the gap between the *minimum* and the average. This bound is tight up to constant factors: the lightest bin indeed trails the average by a logarithmic number of balls (see e.g. [8]). The challenge is therefore to use a different technique to "sharpen" the bound on the gap between maximum and average. We do this next by showing that if the gap is indeed bounded by $\log n$, then after additional $n \log n$ balls are thrown the gap is reduced to $\log \log n$.

The crucial lemma, that we present next, says that if the gap at time t is L, then after throwing another nL balls, the gap becomes $\log \log n + O(\log L)$ with probability close to 1. Roughly speaking, our approach is to apply the lemma twice, first with $L = O(\log n)$ taken from Theorem 3. This reduces the bound to $O(\log \log n)$. A second application of the lemma with $L = O(\log \log n)$ implies Theorem 2. While Theorem 2 holds for any $d > 1$, for ease of exposition we assume in the following that $d = 2$. Generalizing for any $d > 1$ requires nothing more than choosing the constants appropriately and is done in the full version of the paper.

Lemma 2. *There is a universal constant γ such that the following holds: for any t, ℓ, L such that $1 \leq \ell \leq L \leq n^{\frac{1}{4}}$ and $\Pr[G^t \geq L] \leq \frac{1}{2}$,*

$$\Pr[G^{t+L} \geq \log \log n + \ell + \gamma] \leq \Pr[G^t \geq L] + \frac{16bL^3}{\exp(a\ell)} + \frac{1}{n^2},$$

where a, b are the constants froms Theorem 3.

Intuition: The lemma is relatively straightforward to prove using the layered induction technique: For a specific ball to increase the number of balls in a bin from i to $i+1$, it must pick two bins that already contain at least i balls. If we assume inductively that the fraction of bins with at least i balls when this ball is placed is at most β_i, then this probability is at most β_i^2 and thus there are (on expectation) at most $nL\beta_i^2$ bins with load at least $i+1$. Roughly speaking this implies that $\beta_{i+1} \approx L\beta_i^2$. While the β_i's are a function of time, they are monotonically increasing and using the final β_i value would give us an upper bound on the probability of increase. The main challenge is to obtain a base case for the induction. Theorem 3 provides us with such a base case, for bins with ℓ more balls than the average in X^{t+L}. For simplicity, the reader may think of L as $O(\log n)$ and ℓ as $O(\log \log n)$. With these parameters Theorem 3 implies that the fraction of bins with load at least $\ell = O(\log \log n)$ (at time $t + L$) is at most $\frac{1}{4 \log n}$, so the β's shrink in each induction step even though $n \log n$ balls are thrown. As mentioned above, we will use the lemma a second time for $L = O(\log \log n)$ and $\ell = O(\log \log \log n)$.

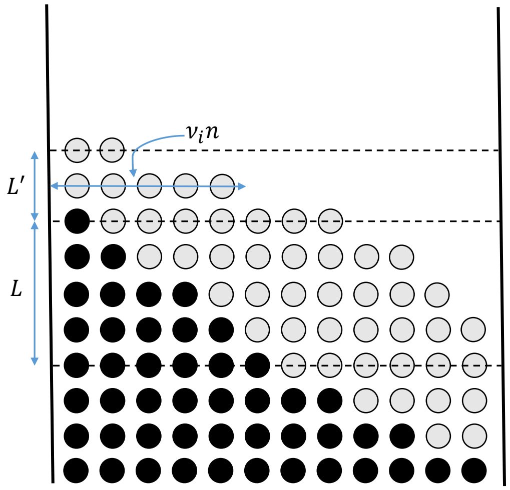

Fig. 1. Black balls are in X, nL white balls are thrown to obtain X'

Proof (Lemma 2). We sample an allocation X^t and let G^t be its gap. Now take an additional L steps of the Markov chain to obtain X^{t+L}: in other words, we throw an additional nL balls using the 2-choice process. For brevity, we will use X, G, X', G' to denote $X^t, G^t, X^{t+L}, G^{t+L}$ respectively. We condition on $G < L$ and we prove the bound for G'. Let $L' = \log\log n + \ell + \gamma$. Observe that:

$$\Pr[G' \geq L'] \leq \Pr[G' \geq L' \mid G < L] + \Pr[G \geq L] \tag{1}$$

It thus suffices to prove that $\Pr[G' \geq L' \mid G < L] \leq \frac{16bL^3}{\exp(a\ell)} + \frac{1}{n^2}$. We do this using a layered induction similar to the one in [1].

Let ν_i be the fraction of bins with normalized load at least i in X' (i.e. containing $t + L + i$ balls or more), we will define a series of numbers β_i such that $\nu_i \leq \beta_i$ with high probability. To convert an expectation bound to a high probability bound, we will use a Chernoff-Hoeffding tail bound as long as $\beta_i n$ is large enough (at least $\log n$). The case for larger i will be handled separately.

By Theorem 3 and Markov's inequality,

$$\Pr[\nu_\ell \geq \frac{1}{8L^3}] \leq \frac{8bL^3}{\exp(a\ell)},$$

which along with the assumption $\Pr[G < L] \geq \frac{1}{2}$ implies that

$$\Pr[\nu_\ell \geq \frac{1}{8L^3} \mid G < L] \leq \frac{16bL^3}{\exp(a\ell)}. \tag{2}$$

We will set $\beta_\ell = \frac{1}{8L^3}$ as the base of the layered induction. We next define the series β_i.

Let $i^* = \ell + \log\log n$. Recall that we set $\beta_\ell = \frac{1}{8L^3}$. For $i = \ell, \ldots, i^* - 1$ we set $\beta_{i+1} = \max(2L\beta_i^2, 18 \log n/n)$. It is easy to check that $\beta_{i^*} = 18 \log n/n$. Indeed suppsose that the truncation does not come into play until i^*. Then the recurrence

$$\begin{aligned}\log \beta_\ell &= -3\log(2L),\\ \log \beta_{i+1} &= 2\log \beta_i + \log(2L)\end{aligned}$$

solves to $\log \beta_{\ell+k} = (\log 2L)(-3 \cdot 2^k + (2^k - 1))$ so that $\log \beta_{i^*} = \log \beta_{\ell+\log\log n} \leq (\log 2L)(-2\log n)$. This is at most $-2\log n$ as $L \geq 1$ so that $\beta_{i^*} \leq \frac{1}{n^2}$ which is smaller than the truncation threshold, contradicting the assumption.

The inductive step is encapsulated in the next lemma. The proof is a simple expectation computation, followed by an application of the Chernoff-Hoeffding bound. We will use $B(n,p)$ to denote a binomially distributed random variable with parameters n and p.

Lemma 3. *For any $i \in [\ell, i^* - 1]$, we have*

$$\Pr[\nu_{i+1} > \beta_{i+1} \mid \nu_i \leq \beta_i, G < L] \leq \frac{1}{n^3}.$$

Proof. For convenience, let the balls existing in X be black, and let the new nL balls thrown be white. We define the *height* of a ball to be the load of the bin in which it was placed relative to X', that is, if the ball was the k'th ball to be placed in the bin, the ball's height is defined to be $k - (t + L)$. Notice that the conditioning that $G < L$ implies that all the black balls have a negative height. We use μ_i to denote the number of white balls with height $\geq i$. Thus for any $i \geq 0$, we have $\nu_i n \leq \mu_i$ and thus it suffices to bound μ_i.

For a ball to have a height of at least $i+1$, it should pick two bins that have load at least i when the ball is placed, and hence at least as much in X'. Thus the probability that a ball has height at least $i+1$ is at most $\nu_i^2 \leq \beta_i^2 \leq \beta_{i+1}/2L$ under our conditioning. Since we place nL balls, the number of balls with height at least $i+1$ is dominated by a $B(nL, \beta_{i+1}/2L)$ random variable. Chernoff bounds (e.g. Theorem 1.1 in [**?**]) imply that the probability that $\Pr[B(n,p) \geq 2np] \leq \exp(-np/3)$. Thus

$$\begin{aligned}\Pr[\nu_i \geq \beta_{i+1}^2 \mid \nu_i \leq \beta_i] &\leq \Pr[B(nL, \beta_{i+1}/2L) \geq \beta_{i+1} n]\\ &\leq \exp(-\beta_{i+1} n/6)\\ &\leq 1/n^3.\end{aligned}$$

since $\beta_{i+1} n \geq 18 \log n$. □

It remains to bound the number of balls with height $\geq i^*$. To this end we condition on $\nu_{i^*} \leq \beta_{i^*}$, and let H be the set of bins of height at least i^* in X'. Once a bin reaches this height, an additional ball falls in it with probability at most $(2\beta_{i^*}n + 1)/n^2$. The probability that any specific bin in H gets at least 4 balls after reaching height i^* is then at most $\Pr[B(nL, (2\beta_{i^*}n + 1)/n^2) \geq 4]$. Recalling that $\Pr[B(n,p) \geq k] \leq \binom{n}{k}p^k \leq (enp/k)^k$. Using this estimate and applying a union bound over the bins in H, we conclude that

$$\begin{aligned} Pr[\nu_{i^*+4} > 0 \mid \nu_{i^*} \leq \beta_{i^*}, G < L] &\leq (18 \log n) \times (eL(32 \log n + 1)/4n)^4) \\ &\leq \frac{1}{2n^2}, \end{aligned} \tag{3}$$

as long as n exceeds an absolute constant n_0. On the other hand, Lemma 1 already implies that for $n \leq n_0$, Lemma 2 holds with $\gamma = O(\log n_0)$ so that this assumption is without loss of generality.

Finally a union bound using (2) and Lemma 3 and (3), we get that

$$\begin{aligned} &\Pr[\nu_{i^*+4} > 0 \mid G < L] \\ &\leq \Pr[\nu_\ell \geq \beta_\ell \mid G < L] + \sum_{i=\ell}^{i^*-1} \Pr[\nu_{i+1} > \beta_{i+1} \mid \nu_i \leq \beta_i, G < L] \\ &\quad + Pr[\nu_{i^*+4} > 0 \mid \nu_{i^*} \leq \beta_{i^*}, G < L] \\ &\leq \frac{16bL^3}{\exp(a\ell)} + \frac{\log\log n}{n^3} + \frac{1}{2n^2} \\ &\leq \frac{16bL^3}{\exp(a\ell)} + \frac{1}{n^2}. \end{aligned}$$

This concludes the proof of Lemma 2. □

Lemma 2 allows us to bound $\Pr[G^{t+L} \geq \log\log n + O(\log L)]$ by $\Pr[G^t \geq L] + \frac{1}{poly(L)}$. Since $\Pr[G^t \geq O(\log n)]$ is small, we can conclude that $\Pr[G^{t+O(\log n)} \geq O(\log\log n)]$ is small. Another application of the lemma, now with $L = O(\log\log n)$ then gives that $\Pr[G^{t+O(\log n)+O(\log\log n)} \geq \log\log n + O(\log\log\log n)]$ is small. We formalize these corollaries next.

Corollary 1. *There is a universal constant γ such that for any $t \geq (12\log n)/a$, $\Pr[G^t \geq (5 + \frac{10}{a}) \cdot \log\log n + \gamma] \leq \frac{2}{n^2} + \frac{1}{\log^4 n}$.*

Proof. Set $L = 12\log n/a$, and use Lemma 1 to bound $\Pr[G^{t-L} \geq L]$. Set $\ell = \log(16bL^3 \log^4 n)/a$ in Lemma 2 to derive the result. □

Corollary 2. *There are universal constants γ, α such that for any $t \geq \omega(\log n)$, $\Pr[G^t \geq \log\log n + \alpha \log\log\log n + \gamma] \leq \frac{3}{n^2} + \frac{1}{\log^4 n} + \frac{1}{(\log\log n)^4}$.*

Proof. Set $L = \log(16b(\frac{12\log n}{a})^3 \log^4 n)/a = \frac{7\log\log n}{a} + O_{a,b}(1)$ and use Corollary 1 to bound $\Pr[G^{t-L} \geq L]$. Set $\ell = \log(16bL^3(\log\log n)^4)/a$ to derive the result. □

This proves that with probability $(1 - o(1))$, the gap is at most $\log\log n + o(\log\log n)$. We can also use Lemma 2 to upper bound the expected gap. Towards this end, we prove slight generalizations of the above corollaries:

Corollary 3. *There is a universal constant γ such that for any $k \geq 0$, $t \geq (12\log n)/a$, $\Pr[G^t \geq (5 + \frac{10}{a}) \cdot \log\log n + k + \gamma] \leq \frac{2}{n^2} + \frac{\exp(-ak)}{\log^4 n}$.*

Proof. Set $L = 12\log n/a$, and use Lemma 1 to bound $\Pr[G^{t-L} \geq L]$. Set $\ell = k + \log(16bL^3 \log^4 n)/a$ to derive the result. □

Corollary 4. *There are universal constants γ, α such that for any $k \geq 0$, $t \geq \omega(\log n)$, $\Pr[G^t \geq \log\log n + \alpha\log\log\log n + k + \gamma] \leq \frac{3}{n^2} + \frac{1}{\log^4 n} + \frac{\exp(-ak)}{(\log\log n)^4}$.*

Proof. Set $L = \log(16b(\frac{12\log n}{a})^3 \log^4 n)/a = \frac{7\log\log n}{a} + O_{a,b}(1)$ and use Corollary 3 with k=0 to bound $\Pr[G^{t-L} \geq L]$. Set $\ell = k + \log(16bL^3(\log\log n)^4)/a$ to derive the result. □

Using the above results, we can now prove

Corollary 5. *There are universal constants γ, α such that for $t \geq \omega(\log n)$ $\mathbb{E}[G^t] \leq \log\log n + \alpha\log\log\log n + \gamma$.*

Proof. Let $\ell_1 = \log\log n + \alpha\log\log\log n + \gamma_1$ for α, γ_1 from Corollary 4, and let $\ell_2 = (5 + \frac{10}{a}) \cdot \log\log n + \gamma_2$ for γ_2 from Corollary 3. Finally, let $\ell_3 = 12\log n/a$. We bound

$$\mathbb{E}[G^t] \leq \ell_1 + \int_{\ell_1}^{\ell_2} \Pr[G^t \geq x]\,\mathrm{d}x + \int_{\ell_2}^{\ell_3} \Pr[G^t \geq x]\,\mathrm{d}x + \int_{\ell_3}^{\infty} \Pr[G^t \geq x]\,\mathrm{d}x$$

Each of the three integrals are bounded by constants, using Corollaries 4 and 3 and Lemma 1 respectively. □

All that remains to prove the $d = 2$ case of Theorem 2 is to show that the lower bound condition on t is unnecessary.

Lemma 4. *For $t \geq t'$, $G^{t'}$ is stochastically dominated by G^t. Thus $\mathbb{E}[G^{t'}] \leq \mathbb{E}[G^t]$ and for every k, $\Pr[G^{t'} \geq k] \leq \Pr[G^t \geq k]$.*

Proof (sketch). We use the notion of majorization, which is a variant of stochastic dominance. See for example [1] for definitions. Observe that trivially X^0 is majorized by $X^{t-t'}$. Now throw nt' balls using the standard coupling and get that $X^{t'}$ is majorized by X^t. The definition of majorization implies the stochastic dominance of the maximum and the bounds on the expectation and the tail follow.

3 Extensions

The technique we use naturally extends to other settings.

3.1 The Weighted Case

So far we assumed all balls are identical. Often balls-and-bins processes model scenarios where items are not necessarily of uniform size but are heterogenous. A natural way to extend the model is to assign weights to the balls drawn from some distribution. We use the model proposed in [10] and also used in [8]. Every ball comes with a weight W independently sampled from a non-negative weight distribution $\mathcal{W}$. The weight of a bin is the sum of weights of balls assigned to it. The *gap* is now naturally defined as the difference between the weight of the heaviest bin and the average bin. We observe that by multiplying all weights by the appropriate constant we can normalize the distribution so that $\mathbb{E}[\mathcal{W}] = 1$. In [10] it is shown that if $\mathcal{W}$ has a bounded second moment and satisfies some additional mild smoothness condition, then the expected gap does not depend on the number of balls. However, no explicit bounds on the gap are shown. In [8] it is shown that if $\mathcal{W}$ satisfies $\mathbb{E}[\exp(\lambda W)] < \infty$ for some $\lambda > 0$, then the gap is bounded by $O(\log n)$ (with λ effecting the hidden constant in O notation). For some distributions, such as the exponential distribution, this bound is tight. A bound of $O(\log n)$ does not necessarily remain tight as the distribution becomes more concentrated.

Consider for example the case where the size of each ball is drawn uniformly from $\{1, 2\}$. Previous techniques such as [2] fail to prove an $O(\log\log n)$ bound in this case, and the best bound prior to this work is the $O(\log n)$ via the potential function argument of [8].

The fact that Theorem 3 holds means that the techniques of this paper can be applied. The modifications needed are straightforward. The layered induction argument works as is, with the only change being that we go up in steps of size two instead of one. This shows a bound of $2\log\log n + O(\log\log\log n)$ for this distribution, which we soon show is tight up to lower order terms.

Generalizing the argument, for a weight distribution W with a bounded exponential moment generating function, let M be the smallest value such that $\Pr[W \geq M] \leq \frac{1}{n\log n(\log\log n)^5}$ (the constant 5 here is somewhat arbitrary, and will only affect the probability of the gap exceeding the desired bound). Then carrying out a proof analogous to Lemma 2, with step size M gives us a bound of $M(\log\log n + O(\log\log\log n))$ with probability $(1 - \frac{3}{(\log\log n)^4})$. This follows since by the definition of M, the probability that any of the last $O(n\log n)$ exceeds size M is $O(\frac{1}{(\log\log n)^5})$, and conditioning on this event the proof goes through unchanged except for the fact that we go up in increments of M.

Indeed, when we use the lemma with $L = O(\log n)$, the base of the induction as before gives us for $\ell = O(\log\log n)$, the fraction of bins with load at least ℓ is at most $\frac{1}{L^3}$. By the argument in Lemma 3, no more than $\beta_{i_L+1}n$ balls will fall in bins that already have at least this load. Since we condition on the $O(n\log n)$ white balls being of size at most M, the number of bins of load $\ell + M$ is at most $\beta_{i_L+1}n$. Continuing in this fashion, with step size M in each step of the induction, we get that there are at most $O(\log n)$ bins of load larger than $O(\log\log n) + M\log_2\log n$. Finally, as before, we can complete the argument with an additional overhead of $O(M)$ as each of these bins is unlikely to get more than a constant number of balls. Finally, a second application of the Lemma gives us the claimed bound.

We next instantiate this bound for some specific distributions. As remarked above, for an exponential or a geometric distribution, the gap is $\Theta(\log n)$ and this induction approach will not help us prove a better bound. Consider a half-normal weight distribution with mean 1 (i.e. W is the absolute value of an $N(0, \frac{\pi}{2})$ random variable. Then $M = \sqrt{\pi} erf^{-1}(1 - \frac{1}{n \log n (\log\log n)^5}) = \Theta(\sqrt{\log n})$. This gives a bound of $O(\sqrt{\log n} \log\log n)$ instead of $O(\log n)$ that we get from [8]. On the other hand, as we show in the next section, a lower bound of $\Omega(\sqrt{\log n})$ is easily proved.

Similarly, if the weight distribution is uniform in $[a, b]$, normalizing the expectation to 1 makes $b = 2 - a \leq 2$. An upper bound of $b \log\log n \leq 2 \log\log n$ follows immediately.

We note that Lemma 4 does not hold when balls are weighted (c.f [10],[3]). As a result this proof leaves a “hole” between n and $n \log n$. It proves the bound on the gap when $O(n)$ or $\Omega(n \log n)$ balls are thrown but does not cover for example $\Theta(n\sqrt{\log n})$ balls.

Lower Bounds If weights are drawn uniformly from $\{1, 2\}$ one might hope the maximum load to be $3/2 \log\log n + O(1)$. It is true that $n/2$ balls of weight 2 already cause a gap of $2 \log\log n$ but one hopes that the balls of weight 1 would reduce this gap. Our first lower bound shows that this intuition is not correct and that the maximum load is indeed $2 \log\log - O(1)$.

Theorem 4. *Suppose that the weight distribution $\mathcal{W}$ satisfies $\Pr[W \geq s] \geq \epsilon$ for some $s \geq 1, \epsilon > 0$ and $\mathbb{E}[W] = 1$. For large enough n, for every $m \geq n/\epsilon$, the gap of Greedy$[d]$ is at least $s(\log\log n/\log d) - O(s)$ with constant probability.*

A similar lower bound is proven in [1] for the case $m = n$ and uniform weights. The proof in [1] uses a layered induction approach as is also outlined in the survey [5]. We note that in the uniform case, majorization (similar to Lemma 4) would extend the lower bound to any $m > n$. The same could not be said in the weighted case. For the $m = n$ case the weighted case is almost as simple as a variable change in the proof of [1]. Majorization however does not hold, so the extension of the lower bound to all $m \geq n$ is done, similarly to the upper bound, by using Theorem 3 to provide a base case for the inductive argument.

Proof. It is convenient to think of time m as time 0 and count both load and time with respect to the m'th ball, so when we say a bin has load i in time t it actually means it has load $w(m)/n + i$ at time $m + t$, where $w(m)$ is the total weight of the first m balls. The bound will be proven for time $m + n/\epsilon$ which is time n/ϵ in our notation. Intuitively, in this amount of time we will see about n balls of weight at least s which would cause the maximum load to increase by $s(\log\log n - O(1))$. The average however would increase only by $O(\frac{1}{\epsilon}) = O(s)$, hence the gap would be at least $s \log\log n - O(s)$.

We follow the notation set in [5], with appropriate changes. The variable $\nu_j(t)$ indicates the number of bins with load in $[(j-1)s, \infty)$ at time t. We will set a series of numbers γ_i and times t_i (to be specified later) and an event

$$\mathcal{F}_i := \{\nu_i(t_i) \geq \gamma_i\}.$$

For the base case of the induction we set $\gamma_0 = n/\log^2 n$ and $t_0 = 0$. We observe that Theorem 3 implies that for large enough n, $\Pr[\nu_0(0) \geq \gamma_0] \geq 1-1/n^2$, so $\mathcal{F}_0$ occurs with high probability. Indeed Theorem 3 implies that for the normalized load vector, $|X^t|_\infty \leq c\log n$ for an absolute constant c. If half the X_i^t's are at least $-s$, we are already done. If not then then $\sum_{i:X_i^t<-s}|X_i^t|$ is at least $\frac{ns}{2}$. Thus the sum $\sum_{i:X_i^t\geq 0}|X_i^t| = \sum_{i:X_i^t<0}|X_i^t| \geq \frac{ns}{2}$. The bound on $|X^t|_\infty$ then implies that at least $ns/c\log n$ X_i^t's are non-negative. Since $s \geq 1$, the base case is proved.

Our goal is to show that $\Pr[\mathcal{F}_{i+1}|\mathcal{F}_i]$ is large. To this end, we define $t_i = (1-2^{-i})\frac{n}{\epsilon}$ and the range $R_i := \left[(1-2^{-i})\frac{n}{\epsilon}, (1-2^{-(i+1)})\frac{n}{\epsilon}\right]$. Finally fix an $i > 0$ and for $t \in R_i$ define the binary random variable

$$Z_t = 1 \text{ iff ball } t \text{ pushes load of a bin above } is \text{ or } \nu_{(i+1)}(t-1) \geq \gamma_{i+1}.$$

As long as $\nu_{(i+1)}(t-1) < \gamma_{i+1}$ it holds that for $Z_t = 1$ it suffices that a ball of weight at least s is placed in a bin of load $h \in [s(i-1), si)$. Conditioned on $\mathcal{F}_i$, the probability of that is at least

$$\epsilon\left((\frac{\gamma_i}{n})^d - (\frac{\gamma_{i+1}}{n})^d\right) \geq \frac{\epsilon\gamma_i^d}{2n^d}$$

since we will set $\gamma_{i+1} \leq \gamma_i/2$. Denote $p_i := \frac{\epsilon\gamma_i^d}{2n^d}$ and by $B(n,p)$ a variable distributed according to the Binomial distribution. We have:

$$\Pr\left[\sum_{i\in R_i} Z_i \leq k \mid \mathcal{F}_i\right] \leq \Pr\left[B\left(\frac{n}{\epsilon 2^{i+1}}, p_i\right) \leq k\right].$$

We continue exactly as in [5] by choosing

$$\gamma_{i+1} = \frac{\gamma_i^d}{2^{i+3}n^{d-1}}.$$

Now Chernoff bounds imply that as long as $\frac{np_i}{\epsilon 2^{i+1}} \geq 17\log n$ it holds that

$$\Pr\left[B\left(\frac{n}{\epsilon 2^{i+1}}, p_i\right) \leq \gamma_{i+1}\right] = o(1/n^2).$$

The tail inequality holds as long as $i \leq \log\log n/\log d - O(1)$, at which point the load had increased by $s(\log\log n/\log d) - O(s)$. The average increased by at most $4/\epsilon \leq 4s$ with probability $3/4$, and the theorem follows. □

We note that the uniform distribution on $\{1,2\}$ (when normalized by a factor of $\frac{2}{3}$) satisfies the conditions of this Theorem with $s = 2, \epsilon = \frac{1}{2}$. Thus the gap is $2\log\log n - O(1)$.

Another, rather trivial lower bound applies to distributions with non-trivial tails.

Theorem 5. *Let $\mathcal{W}$ be a weight distribution with $\mathbb{E}_{W\sim\mathcal{W}}[W] = 1$. Let M be such that $\Pr_{W\sim\mathcal{W}}[W \geq M] \geq \frac{1}{n}$. Then for any allocation scheme, the gap is at least $M - O(1)$ with constant probability.*

Proof. After throwing n balls, the probability that we do not see a ball of weight M or more is at most $(1-\frac{1}{n})^n \leq \frac{1}{2}$. Moreover, by Markov's, the average is at most 4 except with probability $\frac{1}{4}$. Thus with probability at least $\frac{1}{4}$, the maximum is at least M and the average is at most 4. □

We note that this implies an $\Omega(\log n)$ lower bound for an exponential distribution, and an $\Omega(\sqrt{\log n})$ lower bound for the half normal distribution.

3.2 The Left[*d*] Scheme

Next we sketch how this approach also proves a tight bound for Vöcking's Left[d] process [11]. The result had been shown in [2], though there they had to redo large sections of the proof, while here we only require minor changes. Recall that in Left[d], the bins are partitioned into d sets of n/d bins each (we assume n is divisible by d). When placing a ball, one bin is sampled uniformly from each set and the ball is placed in the least loaded of the d bins. The surprising feature of this process is that ties are broken according to a fixed ordering of the sets (we think of the sets as ordered from left to right and ties are broken "to the left", hence the name of the scheme). The surprising result is that the gap now drops from $\frac{\log\log n}{\log d}$ to $\frac{\log\log n}{d\ln\phi_d}$ where $\phi_d = \lim_{k\to\infty}(F_d(k))^{\frac{1}{k}} \in [1.61, 2)$ is the base of the order d Fibonacci number.

The key ingredient in the proof is Theorem 3 from [8]. The exponential potential function is Schur-Convex and therefore the theorem holds for any process which is majorized by the Greedy[d] process. It is indeed the case that Vöcking's Left[d] process [11] is majorized by Greedy[d] (see the proof in [2]). All that remains is to prove the analog of Lemma 2. For this we follow the analysis of Mitzenmacher and Vöcking in [6]. Let X_{jd+k} be the number of bins of load at least j from the k'th set, and set $x_i = X_i/n$. It is easy to verify the recursive equation

$$\mathbb{E}[x_i|x_{<i}] \leq d^d \prod_{j=i-d}^{i-1} x_j$$

From here the proof is similar to that of Lemma 2 and is left as an exercise to the reader.

4 Discussion

The main strength of our approach is that via Theorem 3 it effectively reduces the heavy loaded case to the simpler $m = n$ case. Thus known results for this case, whether it is a weighted upper and lower bounds or the Left[d] scheme follow in a rather similar way. A drawback of our proof technique is that it does not obtain the same tail inequality on deviation from expectation as [2] does. The theorem in [2] states that for every c there is a $\gamma = \gamma(c)$ so that $\Pr[G > \log\log n + \gamma] \leq n^{-c}$. The reason is that we do not ahve a small enough tail bound for the base case of the layered induction, i.e. on $\Pr[\nu_\ell \leq \beta_\ell]$. This is because the potential function used to prove Theorem 3 is not concentrated enough.

An interesting corollary from Theorem 3 is that the Markov chain X^t has a stationary distribution and that the bounds we prove hold also for the stationary distribution itself. In that sense, while in [2] the mixing of the chain was used to show that the interesting events happen at the beginning of the walk (and thus an induction on the first $poly(n)$ suffices), in our technique we look directly at the distant "future" and argue on the stationary distribution itself. When balls are unweighted a majorization based argument shows that moving closer in time can only improve the bounds on the gap (this is Lemma 4). Unfortunately, a similar Lemma does not hold when balls are weighted (see [3]), so we need to specify the time periods we look at. Indeed, while our results hold when considering a large number of balls, we curiously have a 'hole' for a number of balls that is smaller than $n \log n$.

References

1. Yossi Azar, Andrei Broder, Anna Karlin, and Eli Upfal. Balanced allocations. *SIAM J. Computing*, 29(1):180–200, 1999.
2. Petra Berenbrink, Artur Czumaj, Angelika Steger, and Berthold Vöcking. Balanced allocations: The heavily loaded case. *SIAM J. Computing*, 35(6):1350–1385, 2006.
3. Petra Berenbrink, Tom Friedetzky, Zengjian Hu, and Russell Martin. On weighted balls-into-bins games. *Theor. Comput. Sci.*, 409(3):511–520, December 2008.
4. Richard M. Karp, Michael Luby, and Friedhelm Meyer auf der Heide. Efficient pram simulation on a distributed memory machine. In *Proceedings of the twenty-fourth annual ACM symposium on Theory of computing (STOC)*, pages 318–326, 1992.
5. Michael Mitzenmacher, Andréa W. Richa, and Ramesh Sitaraman. The power of two random choices: A survey of techniques and results. In *in Handbook of Randomized Computing*, pages 255–312. Kluwer, 2000.
6. Michael Mitzenmacher and Berhold Vcking. The asymptotics of selecting the shortest of two, improved. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, pages 326–327, 1998.
7. Rasmus Pagh. Hashing 2. Slides for the *MADALGO Summer School on Data Structures* 2013. Available at `http://www.madalgo.au.dk/html_sider/2_5_Events/SS2013/Course_material2013.html`
8. Yuval Peres, Kunal Talwar, and Udi Wieder. The (1 + beta)-choice process and weighted balls-into-bins. In *SODA'10*, pages 1613–1619, 2010.
9. Martin Raab and Angelika Steger. "balls into bins" - a simple and tight analysis. In *Proceedings of the Second International Workshop on Randomization and Approximation Techniques in Computer Science (RANDOM)*, pages 159–170, 1998.
10. Kunal Talwar and Udi Wieder. Balanced allocations: the weighted case. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing (STOC)*, pages 256–265, 2007.

11. Berthold Vöcking. How asymmetry helps load balancing. *J. ACM*, 50(4):568–589, 2003.
12. Udi Wieder. Ballanced allocations with heterogenous bins. In *Proceedings of the nineteenth annual symposium on parallel algorithms and architectures (SPAA)*, pages 188–193, 2007.